FOURIER TRANSFORMS

APPROACH TO SCIENTIFIC PRINCIPLES

Edited by **Goran S. Nikolić**

INTECH

1

Theoretical Description of the Fourier Transform of the Absolute Amplitude Spectra and Its Applications

Levente Csoka and Vladimir Djokovic
[1]University of West Hungary, Institute of Wood and Paper Technology,
Bajcsy Zs. E. u. 4, 9400 Sopron,
[2]Vinca, Institute of Nuclear Science, P.O. Box 522, 11001 Belgrade
[1]Hungary
[2]Serbia

1. Introduction

Speaking in a very broad sense, the Fourier transform (FT) can be treated as a systematic way to decompose arbitrary function into a superposition of harmonic ("symmetrical") functions. It is a fundamental tool for studying of various processes and for this reason it is present in basically every scientific discipline. In last decades, the Fourier transformation was used in distinctive fields such as geophysics (Maus 1999, Skianis et al. 2006), image decomposition in neuroscience (Guyader et al. 2004), imaging in medical applications (Lehmann et al., 1999) just to mention a few. Recently, FT was successfully applied in wood sciences (Fujita et al. 1996; Midorikawa et al. 2005, Midorikawa and Fujita 2005). For example, Fujita and co-workers (Midorikawa et al. 2005, Midorikawa and Fujita 2005) used two-dimensional Fourier transform method to analyze the cell arrangements within the xylem ground tissues. In our recent papers (Csoka et al. 2005, Csoka et al. 2007), we made one step forward and try to analyze the wood anatomy via FT of the density function of the tree. Method is based on a forwarded Fourier transformation of the absolute amplitude spectra. Since the comprehensive literature survey of the accessible studies did not reveal any similar results based on this method, in this chapter we will discuss the basic theorem of FT of an absolute amplitude spectrum and a possibility to generate higher order FT defined as,

$$\left\langle F\left[f(x)\right](k)\right\rangle^{th}. \qquad (1)$$

The discussion also includes a brief description of the theory of the forwarded FT of the complex and absolute amplitude spectrum found in the literature. In the second part of the chapter, it will be shown how the presented theory can be applied to the analysis of the wood anatomy, specifically to determination of the transition point between juvenile and mature wood.

2. Problem statement

We will start these theoretical considerations with familiar one-dimensional Fourier transform (FT) of a given function $f(x)$,

$$F(k) = F[f(x)](k) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi x k} dx, \tag{2}$$

where $F(k)$ is referred to as the spectrum of $f(x)$. The absolute amplitude spectrum of $F(k)$ is defined as,

$$\left|\{F(k)\}\right| = \sqrt{\Re\{F(k)\}^2 + \Im\{F(k)\}^2}\ . \tag{3}$$

Depending on the particular problem, the amplitude spectrum of a signal can be treated as complex or absolute function.

As it was stated in the introduction, the main topic of this chapter is the Fourier transform of the absolute amplitude spectrum and its application to analysis of the wood anatomy. For this reason, we will first consider two basic methods for calculation of the forwarded FT of the amplitude function. The first method is to transform the complex amplitude spectrum $F[f(x)]$ again according to Eq. (2). The second approach is to calculate Fourier transform on the absolute amplitude spectrum via so-called *Wiener-Khinchin* theorem.

2.1 Fourier transform of the complex amplitude spectrum

In the case when the complex amplitude spectrum is transformed the result is a time/space function which has been mirrored with respect to the y-axis, or,

$$f(x) \xrightarrow{FT} F(k) \text{ then } F(k) \xrightarrow{FT} f(-x)\ . \tag{4}$$

The theoretical exposition of Eq. (4) in discrete considerations is as follows. Let the basic finite interval be [0,1]. If we divide that interval in N equal parts, we will obtain $\left\{\frac{k}{N} : k = 0,\dots,N-1\right\}$ points. Let the value at k point be $f(k)$. From the practical reasons we will select the discrete basis $\left\{ e_j : j = 0,\dots,N-1\right\}$, where

$$e_j(k) = \frac{1}{\sqrt{N}} e^{2\pi i j \frac{k}{N}}\ . \tag{5}$$

The $\frac{1}{\sqrt{N}}$ coefficient is necessary, because of the normalization. Now, the discrete FT of f is,

$$F(f)(j) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} f(k) e^{-2\pi i j \frac{k}{N}} \qquad (j = 0,\dots,N-1)\,. \tag{6}$$

Performing the FT on the obtained $F(f)$:

$$F\big(F(f)\big)(l) = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} F(f)(j) e^{-2\pi i l \frac{k}{N}} = \tag{7}$$

$$= \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} \left(\frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} f(k) e^{-2\pi i l \frac{k}{N}} \right) e^{-2\pi i l \frac{k}{N}} = \tag{8}$$

$$= \frac{1}{N} \sum_{k=0}^{N-1} f(k) \sum_{j=0}^{N-1} e^{-2\pi i j \frac{k}{N}} e^{-2\pi i l \frac{k}{N}} = \tag{9}$$

$$= \frac{1}{N} \sum_{k=0}^{N-1} f(k) \sum_{j=0}^{N-1} e^{-2\pi i k \frac{j+l}{N}} \quad \left(j = 0, \ldots, N-1 \right) \tag{10}$$

The sum of a geometric series of the first N member,

$$\sum_{j=0}^{N-1} e^{-2\pi i k \frac{i+l}{N}} \tag{11}$$

is 0, if

$$j + l \neq 0. \tag{12}$$

If

$$j + l = 0, \tag{13}$$

that is

$$j = -l\ , \tag{14}$$

then,

$$F\big(F(f)\big)(l) = f(-l)\ . \tag{15}$$

2.2 Fourier transform of the absolute amplitude spectrum

The estimation of the Fourier transform of the absolute values of the amplitude spectrum, $\left|\{F(k)\}\right|$, requires different approach. In order to find the FT of the absolute spectrum,

$$F_k(\,|\,F(k)\,|\,)(\ell)\,, \tag{16}$$

it is necessary to use the *Wiener-Khinchin* theorem,

$$F_k[\,|\,F(k)\,|^2\,](\ell) = \int_{-\infty}^{\infty} \overline{f}(\tau) f(\tau + \ell) d\tau, \tag{17}$$

where $\overline{f}$ denotes the complex conjugate of f (by definition Eq. (17) is a relationship between FT and its autocorrelation function).

Using Eq. (17), $\left|F(k)\right|$ can be expressed as,

$$|\,F(k)\,| = \sqrt{\int_{-\infty}^{\infty} \Big[F[\,|\,F(k)\,|^2\,](\ell) \Big]\ e^{i 2\pi k \ell} d\ell} = \tag{18}$$

$$= \sqrt{\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \overline{f}(x) f(x+\ell) dx \right] e^{i2\pi k \ell} d\ell} \ . \tag{19}$$

Therefore, the Fourier transform of the absolute amplitude spectrum is

$$F[\,|F(k)|\,](\ell) = \int_{-\infty}^{\infty} |F(k)|\ e^{-i2\pi k \ell} dk \tag{20}$$

$$= \int\limits_{-\infty}^{\infty} \sqrt{\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \overline{f}(x) f(x+\tilde{\ell}) dx \right] e^{i2\pi k \tilde{\ell}} d\tilde{\ell}}\ \ e^{-i2\pi k \ell} dk \tag{21}$$

Spectrum presented by Eq. (21) is essentially different from that of Eq. (4). However, neither Eq. (21) nor Eq. (4) was suitable in our attempt to draw additional information from the experimentally determined density function of the tree stem e.g. to determine the transition point between juvenile and mature wood. As it will be seen below, it turned out that in this practical case it is necessary to perform additional forwarded FT to the positive half of the absolute amplitude spectrum only. In the following section we will consider the forwarded FT of the absolute amplitude spectrum which originates from the superposition of a multitude of harmonic signals.

3. The forwarded FT of the absolute amplitude spectrum which consists of a multitude of harmonic signals

We will start the analysis of the forwarded FT of the absolute amplitude spectrum by considering monochromatic functions obtained by Dirac delta segment sampling of a continuous signal. If $x(t)$ is a original continuous signal then the sampled discrete function, $x_s(t)$, is given by

$$x_s(t) = x(t)(T\ \Delta_T(t))\ , \tag{22}$$

where $\Delta_T(t)$ is the sampling Dirac delta operator and T is the period. Taking that Fourier series of $\Delta_T(t)$ is,

$$\sum_{k=-\infty}^{\infty} e^{i2\pi k f_s t}\ , \tag{23}$$

Eq. (22) can be written as:

$$x_s(t) = \sum_{k=-\infty}^{\infty} x(t) e^{i2\pi k f_s t}\ , \tag{24}$$

where f_s is the sampling frequency and the principle frequency of the periodicity of $\Delta_T(t)$. The amplitude spectrum of monochromatic function given by Eq. (24) can be represented by one dimensional Dirac delta function pair:

$$\delta(f - f_s) + \delta(f + f_s)\ . \tag{25}$$

If the signal is sampled at f_s samples per unit interval, the FT of the sampled function is periodic by a period of f_s.

Let us consider a finite length segment of $x_s(t)$ by performing an L length rectangle window function $\prod(x)$, which is 0 outside the L interval and unity inside it. The FT of a rectangle window function is given by

$$F_x[\prod(x)](k) = \int_{-\infty}^{\infty} e^{-i2\pi k x} \prod(x)\, dx = \sin c(\pi k) \cdot \tag{26}$$

Fourier transform of the $x_s(t)$ $\prod(x)$ product is a convolution operation, which allows us to calculate the spectrum of the windowed, finite function:

$$F[x_s(t) \prod(x)] = F[x_s(t)] * F[\prod(x)] = \delta(f \pm f_s) + \sin c(\pi k)\ . \tag{27}$$

It can be clearly seen that this convolution spectrum consists of a $\sin c(\pi k)$ set at the impulse-position of the Dirac delta function. If the $\prod(x)$ is positioned between $-L/2$ and $+L/2$ then the convolution's spectrum will contain real amplitude values only. Let us chose the length of the original $\prod(x)$ in such way that it contains the whole period of $x_s(t)$. In that case the convolution amplitude spectrum will be reduced to a Dirac delta function pair $\delta(f - f_s)$ and $\delta(f + f_s)$. For further considerations the positive frequency interval $[0, f_S/2]$ is taken which contains single Dirac delta function $\delta(f - f_s)$. That is achieved by multiplication of the amplitude spectrum in the frequency space with window $\prod(k)$ function. The $\prod(k)$ function is not symmetric at the centre; it is shifted to positive direction by one quarter of the original sampling frequency. Finally, the Fourier transform of the obtained $\delta(f - f_s)$ function is an exponential function:

$$F[\delta(f - f_s)] = \Sigma \delta(f - f_s) e^{-i2\pi f_s k} = e^{-i2\pi f_s k} \tag{28}$$

and its amplitude spectrum is unity,

$$\cos(2\pi f_0 x) \Rightarrow \left| e^{-i2\pi f_0 \ell} \right| = 1 \cdot \tag{29}$$

When, however, $x_s(t)$ is a superposition of more harmonic signals, the sum,

$$\sum_j \cos(2\pi f_{0j} x) \Rightarrow \left| \sum_j e^{-i2\pi f_{0j} \ell} \right|, \tag{30}$$

is generally not unity, but it exhibits oscillations. The former result suggests the presence of the complex interaction between amplitude waves which can be used in order to draw the additional information from the original signal. It should be noted that the performing of the FT on the absolute amplitude spectrum will give the spectrum with an argument that is expressed in the same dimensional units as the variable of the original spectrum. For this reason, we believe that the interference peaks in the forwarded FT of the absolute spectrum carry information about the specific positions where certain processes were activated, which, otherwise, can not be observed directly in the original spectrum. Reciprocate of Eq. (30) was further used to determine the FT spectrum of the absolute amplitude spectrum from a density function of a tree. Similarly to Eq. (30) we can generate formula for two dimensional signals (pictures) as,

$$\sum_m \sum_n \cos(2\pi f_{0m} x) \cos(2\pi f_{0n} x) \Rightarrow \left| \sum_m \sum_n e^{-i2\pi f_{0m} \ell} e^{-i2\pi f_{0n} \ell} \right|. \tag{31}$$

It should also be emphasized that if the sum in Eq. 30 is different from unity then it will be possible to generate higher order FT of the absolute amplitude spectrum.

4. Examples and discussion

Timber is a biosynthetic end product so the making of wood is a function of both gene expression and the catalytic rates of structural enzymes. Thus, to achieve a full understanding of wood formation, each component of the full set of intrinsic processes essential for diameter growth (i.e. chemical reactions and physical changes) must be known, investigate in complex form and information on how each one of those components is affected by other processes (Savidge et al., 2000).

The younger juvenile wood produced in the crown has features which distinguish it from the older, more mature wood of the bole (Zobel and Sprague, 1999). Variations within a species are caused by genetic differences and regional differences in growth rate. Differences also occur between the juvenile and mature wood within single trees, and between the earlywood (springwood) and latewood (summerwood) within each annual growth ring. Juvenile wood is an important wood quality attribute because depending on species, it can have lower density, has shorter tracheids, has thin-walled cells, larger fibrial angle, and high – more than 10% – lignin and hemicellulose content and slightly lower cellulose content than mature wood (Zobel and van Buijtenen, 1989, Zobel and Sprague, 1999). Wood juvenility can be established by examining a number of different physical or chemical properties.

Juvenile wood occupies the centre of a tree stem, varying from 5 to 20 growth rings in size, and the transition from juvenile to mature wood is supposed to gradual. This juvenile wood core extends the full tree height, to the uppermost tip (Myers et al., 1997). It is unsuitable for many applications and has great adverse economic impact. Juvenile wood is not desirable for solid wood products because of warpage during drying and low strength properties and critical factors in producing high stiffness veneer (Willits et al., 1997). In the other hand, in the pulp and paper industry juvenile wood has higher than mature wood in tear index, tensile index, zero-span tensile index, and compression strength. For the same chemical pulping conditions, pulp yield for juvenile wood is about 25 percent less than pulp yield for mature wood (Myers et al., 1997).

It is, therefore, important from scientific as well as from practical reasons to determine the demarcation line between juvenile and mature wood. The advantage of the present approach that this boundary line can be determined by analysis of density spectrum which was obtained by non-invasive X-ray densitometry method.

4.1 Materials and methods

Twelve selected trees were investigated, which were planted in Akita Prefectures, Japan. The name of the tree is sugi (Cryptomeria japonica D. Don). The trees were harvested in different ages between 71 and 214 years (Table 1). Tracheid lengths, annual ring structure, were also determined from those samples.

4.2 X-ray densitometry

Bark to bark radial strips of 5 mm thickness were prepared from the air-dried blocks cut from the sample disks. After conditioning at 20 °C and 65% RH, without warm water extraction, the strips were investigated by using X-ray densitometry technique, with 340 seconds of irradiation time. The current intensity and voltage were 14 mA and 17 kV, respectively. The distance between the X-ray source and the specimen was 250 cm. The developed films were scanned with a densitometer (JL Automation 3CS-PC) to obtain density values across the growth rings (Figure 1) and with a table scanner (HP ScanJet 4C) to obtain digital X-ray picture for image processing.

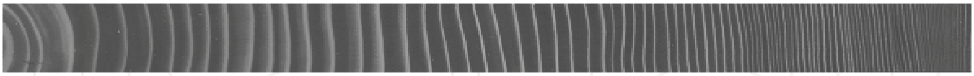


Fig. 1. X-ray image of a sugi sample

The growth ring parameters of ring width (RW), minimum density within a ring (D_{min}), maximum density within a ring (D_{max}) and ring density (RD: average density within a ring) were determined for each growth ring by a special computer software. The latewood is categorized by Mork's definition, as a region of the ring where the radial cell lumens are equal to, or smaller than, twice the thickness of radial double cell walls of adjacent tracheids (Denne, 1989). A threshold density, 0.55 g/cm^3 was used as the boundary between earlywood and latewood (Koizumi et al., 2003).

4.3 FT of the density function of the sugi tree

Figure 2 shows density function of the sugi tree obtained by laser scanning of the x-ray image. It can be seen that the signal is periodic and its amplitude FT spectrum is shown in Figure 3. The amplitude spectrum shows a strong peak at frequency 0.4 mm^{-1} which shows that the most frequent annual ring is about 2.5 mm. However, after reciprocate of the Eq. (34) was used in order to determine the spectrum of the absolute amplitude spectrum some additional information were obtained (Figure 4). While the amplitude spectrum shows the frequency structure of continuous or discrete signals, the forwarded FT of the absolute amplitude spectrum can provide the information about the complex effect of the interaction among these waves.

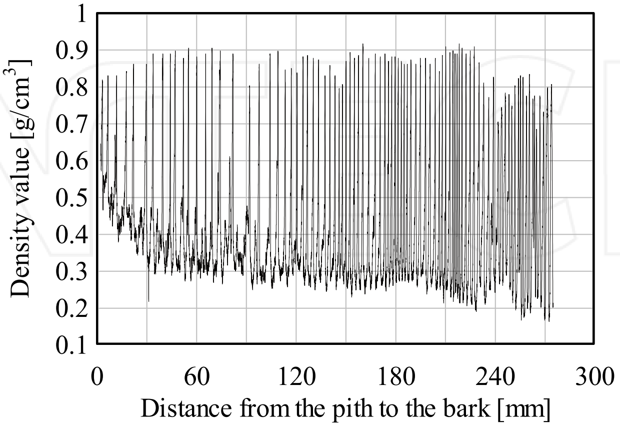


Fig. 2. The density function of a sugi tree (obtained by laser scanning of the X-ray image)

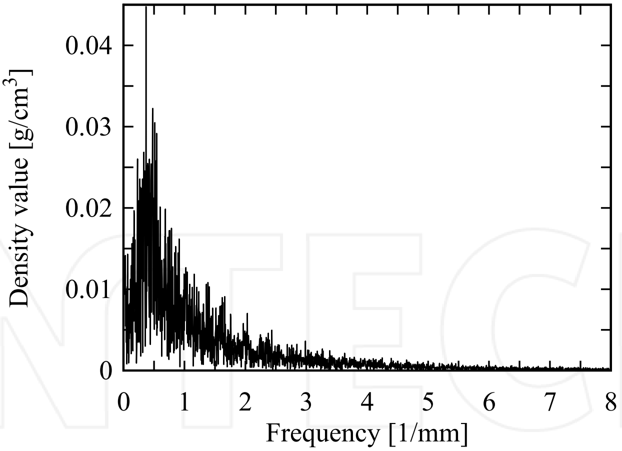


Fig. 3. The amplitude spectrum of the density function

As it can be noticed in Figure 4, the second FT spectrum shows spikes at certain positions. These peaks suggest the locations in the original complex function where the superposition of two or more periodic curves takes place. The highest peak has been assigned to the transition point between juvenile and mature wood (Csoka et al., 2007). Note that FT changes the dimension of the independent variable according to the input signals. The dimension of the variable of the second FT spectrum is the same as the dimension of the original variable. It should also be emphasized that the obtained values for the transition between juvenile and mature wood calculated from the second FT spectrum were in agreement with the values obtained from segmented model of tracheid lengths (Zhu et al., 2005).

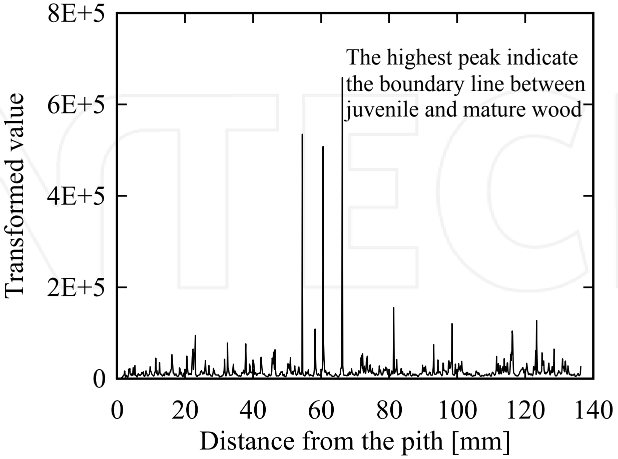


Fig. 4. The forwarded FT of the amplitude spectrum of a density function

The variables and typical parameters of the density function, its amplitude spectrum and the Fourier transform of the amplitude spectrum are given in *Table 1*. In that table *L* represents the actual length of the sample which depends of the age of the wood. However, increment between discrete points is kept constant to 0.015 mm. It should be also pointed out that one annual ring is represented by 200-400 points.

Properties of spectrums		
density function	*Amplitude spectrum*	*FT of amplitude spectrum*
length of x axis	length of the x axis	length of the x axis
$L \quad [mm]$	$f_s / 2 = 33.3\dot{3} \quad [1/mm]$	$L/2 \quad [mm]$
increment between points	increment between points	increment between points
$\Delta l = L/N = 0.015\,[mm]$	$\Delta l = L_1/N_1$	$\Delta l = L_2/N_2$
number of points	number of points	number of points
N	$N/2$	$N/4$

Table 1. The variables and typical parameters of the density function, its amplitude spectrum and the Fourier transform of the amplitude spectrum

4.4 FT of the X-ray image

X-ray image (Figure 1) was first processed by using a spatial grey level method. After the determination of the grey level at each point in the image, a 2D power spectrum that represents image in the frequency domain was calculated via Fourier transformation. Figure 5 shows the obtained power spectrum in a 3D representation. The amplitude spectrum of an X-ray image expresses a function (which is a point in some infinite dimensional vector space of functions) in terms of the sum of its projections onto a set of basis functions. The amplitude spectrum of the image carries information about the relative weights with which frequency components (projections) contribute to the spectrum, while the phase spectrum (not shown) localizes these frequency components in space (Fisher et al., 2002). It should be noted that in the Fourier domain image, the number of frequencies corresponds to the number of pixels in the spatial domain image, i.e. the image in the spatial and Fourier domains are of the same size (Castleman 1996).

The 3D representation of the power spectrum in Figure 5 is related to the rate at which gradual brightness in the X-ray image varies across the image. The frequency refers to the rate of repetitions per unit time i.e. the number of cycles per millimetre. Therefore, the intensive peaks observed in Figure 5 indicate the basic frequencies of the annual ring pattern in the frequency domain. The forwarded FT of the amplitude spectrum of the image is shown in Figure 6. With a closer look at the original image, a strong relationship between the annual ring texture and the spectrum in Figure 6 can be noticed, with could also justify our approach of using forwarded Fourier transformation of the absolute spectrum for determination of the demarcation zone between juvenile and mature wood. The texture of the 3D picture obtained from the forwarded FT of the absolute spectrum exhibit obvious annual ring pattern.

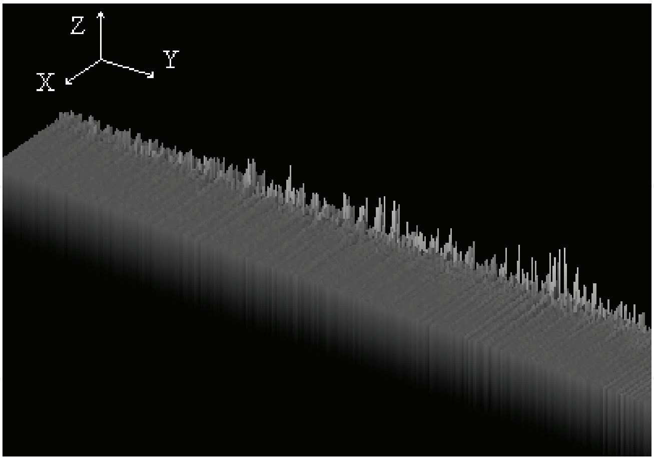


Fig. 5. The power spectrum of the X-ray image (Figure 1) in 3D representation

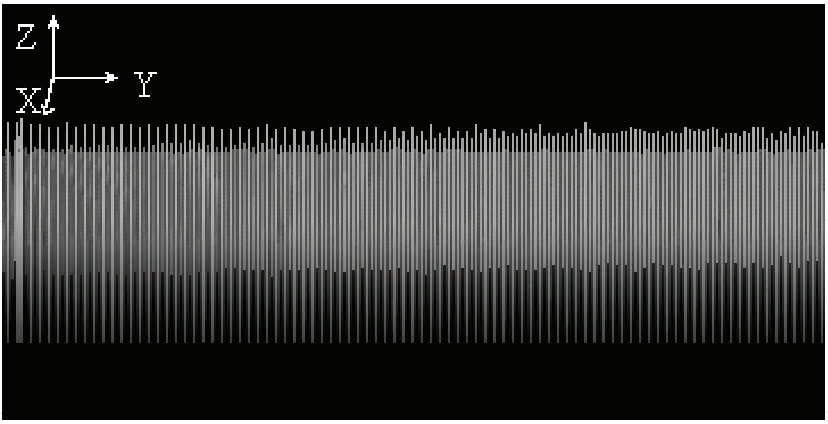


Fig. 6. The forwarded FT of the amplitude spectrum of the X-ray image

In order to analyze to transition between juvenile and mature wood from the forwarded FT of the amplitude spectrum of X-ray image, horizontal intensity line slices have been took through the spectrum in Figure 6. This pixel slice contains information about possible interactions between certain modes in the amplitude spectrum. The spectrum in Figure 7 was obtained by taking the sum of the slices from the bottom to the top of the image. The highest peak in the spectrum refers to the transition point of juvenile and mature wood.

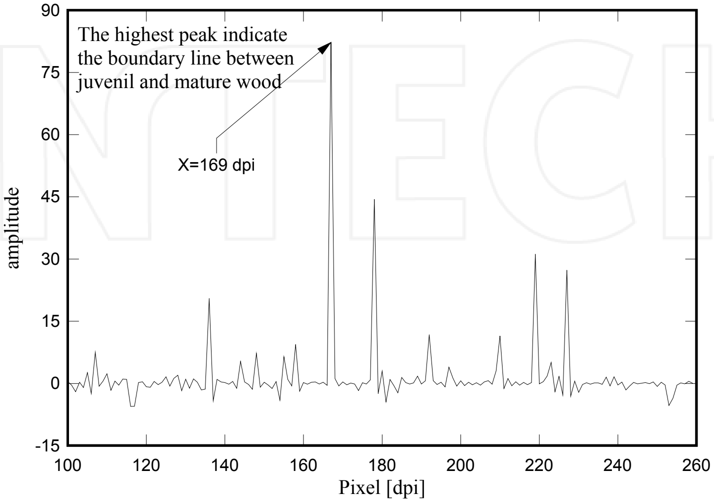


Fig. 7. The sum of pixel slices from the bottom to the top of the forwarded FT of the amplitude spectrum presented in Figure 6

5. Conclusions

Since the determination of the boundary zone between juvenile and mature wood is a subject of great practical importance in the area of wood anatomy, a various methods were suggested to address this problem. However, most of these methods considered only a limited number of characteristic features of the wood stem (one or two). For example, for a long time, the researchers were focused on the measurements of the annual ring width, the specific gravity, tracheid length and microfibril angle (Fujisaki 1985, Fukazawa 1967, Matyas and Peszlen 1997, Ota 1971, Yang et al. 1986, Zhu et al. 2000). The method based on the nonlinear, segmented regression method of tracheid length and microfibril angle (Cook and Barbour 1989, Zhu et al. 2005) has provided a common and simple tool for analyzing of the growth variation, while at the same time it was not restricted to certain groups of species or types of data. Unfortunately, all these different approaches did not take the complexity of the stem into account. The global nature of the above mentioned processes hides local density-distribution information, and makes the determination of the changes related to the distance from the pith impossible.

In this chapter we presented the FT of an amplitude spectrum theorem that can find direct application in studying of a wood anatomy. In spite of its simplicity, to our best knowledge there is no reference in the literature regarding the use of forwarded FT of the absolute amplitude spectra of an arbitrary vibration in the way we suggested. The suggested theoretical approach was used in order to determine the demarcation zone between juvenile and mature wood within a tree stem from the experimentally obtained density spectrum. The main advantage of the present method is that it enables simultaneous study of the changes in the density of annual rings and their distances from the pith, while they were, so far, studied as independent properties. The density function contains inherent information about changes in successive annual rings that may, after an appropriate mathematical analysis procedure, be used to describe the microstructure of the wood. It is assumed that the variation in the biological and physical characteristics of the cell (i.e. the cell dimension, the thickness of cell wall, the cellulose and lignin contents in the cell wall, and the growth rate) will be reflected in the sequences of wood density in the radial direction. The forwarded FT of the absolute amplitude spectrum provides information about the interaction of the amplitude waves, which can be further used to characterize the physical growth of the trees.

6. References

Denne, M.P. (1989). Definition of latewood according to Mork (1928). *International Association of Wood Anatomists Bulletin*. 10,1,59–62

Divos, F., Denes, L., Iniguez, G. (2005). Effect of cross-sectional change of a board specimen on stress wave velocity determination. *Holzforschung*. 59, 230-231

Cook, J.A., Barbour, R.J. (1989). The use of segmented regression analysis in the determination of juvenile and mature wood properties. *Report CFS No. 31. Forintek Canada*, Corp., Vancouver, BC

Csoka, L., Divos, F., Takata, K. (2007). Utilization of Fourier transform of the absolute amplitude spectrum in wood anatomy. *Applied Mathematics and Computation*. 193, 385-388

Csoka, L., Zhu, J., Takata, K. (2005). Application of the Fourier analysis to determine the demarcation between juvenile and mature wood. *J. Wood Sci.* 51, 309-311, 385-388

Fujisaki, K. (1985). On the relationship between the anatomical features and the wood quality in the sugi cultivars (1) on cv. Kumotoshi, cv. Yaichi, cv. Yabukuguri and cv. Measa (in Japanese). *Bull. Ehime Univ. Forest.* 23:47-58

Fujita, M., Ohyama, M., Saiki, H. (1996). Characterization of vessel distribution by Fourier transform image analysis. In: Lloyd AD, Adya PS, Brian GB, Leslie JW (eds) Recent advances in wood anatomy. Forest Research Institute, New Zealand, pp 36–44

Fukazawa, K. (1967). The variation of wood quality within a tree of Cryptomeria japonica – characteristics of juvenile and adult wood resulting from various growth conditions and genetic factors. *Res. Bull. Fac. Agric. Gifu Univ.* Japan, 25:47-127

Guyader, N., Chauvin, A., Peyrina, C., Hérault, J. Marendaz, Ch. (2004). Image phase or amplitude? Rapid scene categorization is an amplitude-based process *C. R. Biologies*. 327, 313–318

Koizumi, A., Takata, K., Yamashita, K., Nakada, R. (2003). Anatomical characteristics and mechanical properties of Larix Sibirica grown in South-central Siberia. *IAWA Journal*. 24, 4, 355-370

Lehmann, T.M., Gönner, C., Spitzer, K. (1999). Survey: Interpolation Methods in Medical Image Processing. IEEE *Transactions On Medical Imaging*. 18, 11, 1049-1075

Maus, S. (1999). Variogram analysis of magnetic and gravity data. *Geophysics*. 64, 3, 776-784

Midorikawa, Y., Ishida, Y., Fujita, M. (2005). Transverse shape analysis of xylem ground tissues by Fourier transform image analysis I: trial for statistical expression of cell arrangements with fluctuation. *J. Wood Sci*. 51, 201–208

Midorikawa, Y., Fujita, M. (2005). Transverse shape analysis of xylem ground tissues by Fourier transformimage analysis II: cell wall directions and reconstruction of cell shapes. J. Wood Sci. 51, 209–217

Matyas, C., Peszlen, I., (1997). Effect of age on selected wood quality traits of poplar clones. *Silvae Genetica*. 46, 64-72

Myers, G.C., Kumar, S., Gustafson, R.R., Barbour. R.J., Abubakr, S. (1997). Pulp quality from small-diameter trees. *Role of Wood Production in Ecosystem Management Proceedings of the Sustainable Forestry Working Group* at the IUFRO All Division 5 Conference, Pullman, Washington

Ota, S. (1971). Studies on mechanical properties of juvenile wood, especially of sugi-wood and hinoki-wood (in Japanese). *Bull. Kyushu Univ. Forests*. 45:1-80

Savidge, R.A., Barnett, J.R., Napier, R. /edited/ (2000). Cell and Molecular Biology of Wood Formation. BIOS, Biddles Ltd, Guilford, UK pp.2-7

Skianis, G.A., Papadopoulos, T.D., Vaiopoulos, D.A. (2006). Direct interpretation of self-potential anomalies produced by a vertical dipole. *Journal of Applied Geophysics*. 58, 130–143

Willits, S.A., Lowell, E.C., Christensen, G.A. (1997). Lumber and veneer yield from small-diameter trees. *Role of Wood Production in Ecosystem Management Proceedings of the Sustainable Forestry Working Group* at the IUFRO All Division 5 Conference, Pullman, Washington

Zhu, J., Nakano, T., Hirakwa, Y., Zhu, J.J. (2000). Effect of radial growth rate on selected indices of juvenile and mature wood of the Japanese larch. *J. Wood. Sci*. 46:417-422

Zhu, J., Tadooka, N., Takata, K., Koizumi, A. (2005). Growth and wood quality of sugi (*Cryptomeria japonica*) planted in Akita prefecture (II). Juvenile/mature wood determination of aged trees. *J. Wood Sci*. 51, 95-101

Zobel, B.J., van Buijtenen, J.R. (1989). *Wood variation, its cause and control*. p 375, Springer-Verlag, ISBN 354050298X, Berlin

Zobel, B., Sprague, J. R. (1998). *Juvenile wood in forest trees*. p 300, Springer-verlag, ISBN 3540640320, Berlin, Heidelberg

Yang, K. C., Benson, C., Wong J.K. (1986). Distribution of juvenile wood in two stems of Larix laricina. *Can J For Res*. 16:1041-1049

Yeh, T.F., Goldfarb, B., Chang, H.M., Peszlen, I., Braun, J.L., Kadla, J. F. (2005). Comparison of morphological and chemical properties between juvenile wood and compression wood of loblolly pine. *Holzforschung*. 59, 669-674

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

2

Gaussian and Fourier Transform (GFT) Method and Screened Hartree-Fock Exchange Potential for First-principles Band Structure Calculations

Tomomi Shimazaki[1,2] and Yoshihiro Asai[1]
[1]National Institute of Advanced Industrial Science and Technology (AIST),Umezono 1-1-1, Tsukuba Central 2, Tsukuba, Ibaraki 305-8568,
[2]Fracture and Reliability Research Institute (FRRI), Graduate School of Engineering, Tohoku University, 6-6-11-703 Aoba, Aramaki, Aoba-ku, Sendai, Miyagi 980-8579
Japan

1. Introduction

The Gaussian and Fourier Transform (GFT) method is a first-principles quantum chemistry approach based on the Gaussian basis set, which can take into account the periodic boundary condition (PBC).(Shimazaki et al. 2009 a) The quantum chemistry method has mainly concentrated on isolated molecular systems even if target system becomes large such as DNA molecules and proteins, and the periodic nature does not appear. However, chemists have been recently paying much attention to bulk materials and surface, which cover electrochemical reaction, photoreaction, and catalytic behaviour on the metal or semiconductor surfaces. The periodic boundary condition in the first principles (*ab-initio*) approach is a strong mathematical tool for handling those systems. In addition to this, the momentum (k-space) description for the electronic structure helps us to understand essential physical and chemical phenomena on those systems. Therefore, it is an inevitable desire to extend the ordinary quantum chemistry method toward the periodic boundary condition. The crystal orbital method is a straight-forward extension for the purpose.(Hirata et al. 2010; Ladik 1999; Pisani et al. 1988) However, the crystal orbital method naturally faces a challenging problem to calculate the Hartree term due to the long-range behavior of the Coulomb potential. The method requires for infinite lattice sum calculations with respect to two electron integral terms, which intensively takes CPU costs even if some truncation is employed. Therefore, several computational techniques, such as sophisticated cutoff-criteria and the fast multi-pole method (FMM), have been developed to cope with the problem. (Delhalle et al. 1980; Kudin et al. 2000; Piani et al. 1980) In this chapter, we explain an efficient method using Fourier transform technique and auxiliary pane wave, whose description is suitable for the periodic boundary condition, to calculate the periodic Hartree term. Our method is based on the Gaussian basis set and the Fourier (GFT) transform method, thus we refer to our method as the GFT method. In the GFT method, the Hartree (Coulomb) potential is represented by auxiliary plane waves, whose coefficients are obtained by solving Poisson's equation based on the Fourier transform technique. However, the matrix element of the Hartree term is determined in the real-space integration including Gaussian-based atomic orbtials and plane waves. We can employ a recursive relation to achieve the integration, as discussed later. Conversely, we can employ the effective core potential (ECP) instead of explicitly taking into account core electrons in the GFT method. This chapter will demonstrate several electronic band structures obtained from the GFT method to show the availability of our method for crystalline systems.

We try to develop the GFT method to become an extension of the ordinary quantum chemistry method, whereas our method employs the different integration algorithm for the Hartree term. Therefore, the various quantum chemistry techniques can be easily incorporated into the GFT method. In this chapter, we discuss the effect on the Hartree-Fock fraction term in the electronic structure calculation for solid-state materials. In first-principles calculations of crystalline and surface systems, local or semi-local density functional theory (DFT) is usually employed, but the use of the HF fraction can expand the possibility of DFT. The HF fraction is frequently adopted in the hybrid DFT functional, especially in the field of the quantum chemistry. It has been proved that the electronic structure description of the hybrid DFT method is superior compared with the local density (LDA) and generalized gradient approximations (GGA) in molecular systems. However, the hybrid DFT method is rarely adopted for crystalline and surface systems because of its larger computational cost.

When we discuss the hybrid DFT method in crystalline and surface systems, the concept of screening on the exchange term is imperative. The concept has been already taken into account in the HSE hybrid-DFT functional, which is proposed by Heyd et al. in 2003.(Heyd et al. 2003) On the other hand, the GW approximation handles the concept, whereas it is not in DFT framework.(Aryasetiawan et al. 1998; Hedin 1965) In the Coulomb hole plus screened exchange (COHSEX) approximation of the GW method, the screened exchange term is explicitly described. Thus, its importance has been recognized at early stage of first-principles calculations. However, the relationship between the hybrid-DFT method and the screening effect has not been paid attention so much. Recently, we propose a novel screened HF exchange potential, in which the inverse of the static dielectric constant represents the fraction of HF exchange term.(Shimazaki et al. 2010; Shimazaki et al. 2008; Shimazaki et al. 2009 b) The screened potential can be derived from a model dielectric function, which is discussed in Section III, and can give an interpretation how the screening effect behaves in semiconductors and metals. In addition, it will be helpful to present a physical explanation for the HF exchange term appeared in the hybrid-DFT method. In order to show the validity of our physical concept, we demonstrate several band structure calculations based on our screened HF exchange potential, and show that our concept on the screening effect is applicable to semiconductors. In this chapter, the screened HF exchange potential is incorporated with the GFT method, whereas it does not need to stick to the GFT method. The GFT method is based on the Gaussian-basis formalism, and therefore we can easily introduce the hybrid-DFT formalism for PBC calculations.

2. Gaussian and Fourier Transform (GFT) method

2.1 Crystal Orbital method

First we briefly review the crystal orbital method, which is a straight-forward extension of the quantum chemistry method to consider the periodic boundary condition.(Hirata et al. 2009; Hirata et al. 2010; Ladik 1999; Pisani et al. 1988; Shimazaki et al. 2009 c) The Bloch function (crystal orbital) for solid-state material is obtained from the linear combination of atomic orbitals (LCAO) expansion as follows:

$$b_j^{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{K}} \sum_{\alpha}^{M} \sum_{\mathbf{Q}}^{K} \exp(i\mathbf{k}\cdot\mathbf{Q})\, d_{\alpha,j}(\mathbf{k})\, \chi_\alpha^{\mathbf{Q}} \tag{1}$$

Where $\mathbf{Q}$ is the translation vector. The total number of cells is $K = K_1 K_2 K_3$, where K_1, K_2, and K_3 are the number of cells in the direction of each crystal axis, and $\mathbf{k}$ is the wave vector. $\chi_\alpha^{\mathrm{Q}} = \chi_\alpha(\mathbf{r} - \mathbf{Q} - \mathbf{r}_\alpha)$ is the α -th atomic orbital (AO), whose center is displaced from the origin of the unit cell at $\mathbf{Q}$ by $\mathbf{r}_\alpha$. $d_{\alpha,j}^{\mathbf{k}}$ is the LCAO coefficient, which is obtained from the Schrödinger equation as follows:

$$\mathbf{h}(\mathbf{k})\mathbf{d}_j^{\mathbf{k}} = \lambda_j^{\mathbf{k}} \mathbf{S}(\mathbf{k})\mathbf{d}_j^{\mathbf{k}} \tag{2-1}$$

$$\mathbf{d}_j^{\mathbf{k}} = \begin{pmatrix} d_{1,j}^{\mathbf{k}} & d_{2,j}^{\mathbf{k}} & \cdots & d_{\alpha,j}^{\mathbf{k}} & \cdots & d_{M,j}^{\mathbf{k}} \end{pmatrix}^T \tag{2-2}$$

$$\mathbf{h}(\mathbf{k}) = \sum_{\mathbf{Q}}^{K} \exp(i\mathbf{k}\cdot\mathbf{Q})\mathbf{h}(\mathbf{Q}) \tag{2-3}$$

$$\mathbf{S}(\mathbf{k}) = \sum_{\mathbf{Q}}^{K} \exp(i\mathbf{k}\cdot\mathbf{Q})\mathbf{S}(\mathbf{Q}) \tag{2-4}$$

$$\mathbf{d}_j^{\mathbf{k}*T} \mathbf{S}(\mathbf{k}) \mathbf{d}_{j'}^{\mathbf{k}} = \delta_{j,j'}\,. \tag{2-5}$$

Here, the Hamiltonian and the overlap matrices are given by $[\mathbf{h}(\mathbf{Q})]_{\alpha\beta} = \langle \chi_\alpha^{\mathrm{Q}_1} | \hat{h} | \chi_\beta^{\mathrm{Q}_2} \rangle$ and $[\mathbf{S}(\mathbf{Q})]_{\alpha\beta} = \langle \chi_\alpha^{\mathrm{Q}_1} | \chi_\beta^{\mathrm{Q}_2} \rangle$, respectively. $\hat{h}$ is the one-electron Hamiltonian operator, and $\mathbf{Q} = \mathbf{Q}_2 - \mathbf{Q}_1$. The Hamiltonian matrix is composed of the following terms:

$$\mathbf{h}(\mathbf{Q}) = \mathbf{T}(\mathbf{Q}) + \mathbf{V}_{NA}(\mathbf{Q}) + \mathbf{V}_{Hartree}(\mathbf{Q}) + \mathbf{V}_{XC}(\mathbf{Q}) \tag{3}$$

Here, $\mathbf{T}(\mathbf{Q})$ represents the kinetic term, whose matrix element is obtained from $[\mathbf{T}(\mathbf{Q})]_{\alpha\beta} = \langle \chi_\alpha^{\mathbf{0}} | -(1/2)\nabla^2 | \chi_\beta^{\mathbf{Q}} \rangle$. $\mathbf{V}_{NA}(\mathbf{Q})$ is the nuclear attraction term, which is obtained from $[\mathbf{V}_{NA}(\mathbf{Q})]_{\alpha\beta} = \langle \chi_\alpha^{\mathbf{0}} | \sum_A -(Z_A / |\mathbf{r} - \mathbf{R}_A|) | \chi_\beta^{\mathbf{Q}} \rangle$, $\mathbf{V}_{Hartree}(\mathbf{Q})$ is the Hartree term, and $\mathbf{V}_{XC}(\mathbf{Q})$ is the exchange-correlation term. For example, the exchange-correlation term in the HF approximation is expressed using $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$ as follows,

$$[\mathbf{V}^{Fock}(\mathbf{Q})]_{\alpha\beta} = -\sum_{\gamma}\sum_{\delta}\sum_{\mathbf{Q}_1,\mathbf{Q}_2} \mathbf{D}_{\gamma\delta}(\mathbf{Q}_1 - \mathbf{Q}_2) \iint \chi_\alpha^0(\mathbf{r}_1)\chi_\gamma^{\mathbf{Q}}(\mathbf{r}_1)\frac{1}{r_{12}}\chi_\beta^{\mathbf{Q}_1}(\mathbf{r}_2)\chi_\delta^{\mathbf{Q}_2}(\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \tag{4}$$

Here the AO-basis density matrix $\mathbf{D}$ is obtained from the following equation.

$$D_{\alpha\beta}(\mathbf{Q}) = \frac{1}{K}\sum_{\mathbf{k}}\sum_{j} f_{FD}\left(E_F - \lambda_j^{\mathbf{k}}\right) d_{\alpha,j}^{\mathbf{k}*} d_{\beta,j}^{\mathbf{k}} \exp(i\mathbf{k}\cdot\mathbf{Q}) \tag{5}$$

Where $f_{FD}\left(E_F-\lambda_j^{\mathbf{k}}\right)$ and E_F are the Fermi–Dirac distribution function and the Fermi energy, respectively.

2.2 Gaussian and Fourier Transform (GFT) method

In the crystal orbital method, the calculation of the Hartree term is the most time-consuming part due to the long-range behavior of the Coulomb potential. The electron-electron repulsion integrals need to be summed up to achieve numerical convergence. In order to avoid the time-consuming integrations, we employ the Hartree (Coulomb) potential with the plane-wave description and the Fourier transform technique.(Shimazaki et al. 2009 a) In the method, we divided the nuclear attraction and Hartree terms into core and valence contributions as follows.

$$\mathbf{V}_{Hartree}\left(\mathbf{Q}\right)=\mathbf{V}_{Hartree}^{core}\left(\mathbf{Q}\right)+\mathbf{V}_{Hartree}^{valence}\left(\mathbf{Q}\right) \tag{6-1}$$

$$\mathbf{V}_{NA}\left(\mathbf{Q}\right)=\mathbf{V}_{NA}^{core}\left(\mathbf{Q}\right)+\mathbf{V}_{SR-NA}^{valence}\left(\mathbf{Q}\right)+\mathbf{V}_{LR-NA}^{valence}\left(\mathbf{Q}\right) \tag{6-2}$$

The above equation is obtained by simply dividing the terms into core and valence contributions, where $\mathbf{V}_{NA}^{core}\left(\mathbf{Q}\right)$ and $\mathbf{V}_{Hartree}^{core}\left(\mathbf{Q}\right)$ are the nuclear attraction and Hartree terms for the core contribution, respectively. $\mathbf{V}_{SR-NA}^{valence}\left(\mathbf{Q}\right)$ and $\mathbf{V}_{LR-NA}^{valence}\left(\mathbf{Q}\right)$ are the short-range (SR) and long-range (LR) nuclear attraction terms, respectively, for the valence contribution. $\mathbf{V}_{Hartree}^{valence}\left(\mathbf{Q}\right)$ is the Hartree term for the valence contribution. The electron-electron and electron-nuclear interactions of the core contribution are directly determined based on the conventional quantum chemical (direct lattice sum) calculations. However, this lattice sum calculations does not intensively consume CPU-time, because core electrons are strongly localized, and therefore its potential-tail rapidly decays to cancel the core nuclear charges. We will discuss the effective core potential (ECP) for core electrons in the next section. On the other hand, the contribution of valence electrons is considered by using the Poisson's equation and the Fourier transform. In order to divide the terms into core and valence contributions, we introduce the following core and valence electron densities.

$$\rho\left(\mathbf{r}\right)=\sum_{\alpha}\sum_{\beta}\sum_{\mathbf{Q}_1,\mathbf{Q}_2}\mathbf{D}_{\alpha\beta}\left(\mathbf{Q}_2-\mathbf{Q}_1\right)\chi_{\beta}^{\mathbf{Q}_2}\left(\mathbf{r}\right)\chi_{\alpha}^{\mathbf{Q}_1}\left(\mathbf{r}\right)\equiv\rho^{core}\left(\mathbf{r}\right)+\rho^{valence}\left(\mathbf{r}\right) \tag{7-1}$$

$$\rho^{valence}\left(\mathbf{r}\right)=\sum_{\alpha}^{valence}\sum_{\beta}^{valence}\sum_{\mathbf{Q}_1,\mathbf{Q}_3}\mathbf{D}_{\alpha\beta}\left(\mathbf{Q}_3\right)\chi_{\alpha}^{\mathbf{Q}_1}\left(\mathbf{r}\right)\chi_{\beta}^{\mathbf{Q}_1+\mathbf{Q}_3}\left(\mathbf{r}\right) \tag{7-2}$$

$$\rho^{core}\left(\mathbf{r}\right)=\rho\left(\mathbf{r}\right)-\rho^{valence}\left(\mathbf{r}\right) \tag{7-3}$$

Here, $\rho\left(\mathbf{r}\right)$ is the total electron density, and $\rho^{core}\left(\mathbf{r}\right)$ and $\rho^{valence}\left(\mathbf{r}\right)$ are the core and valence electron densities, respectively. The Hartree potential is divided into core and valence components on the basis of eq. (7) as follows:

$$\begin{aligned}V_{Hartree}\left(\mathbf{r}\right)&=\int\frac{\rho\left(\mathbf{r}'\right)}{\left|\mathbf{r}-\mathbf{r}'\right|}d\mathbf{r}'=\int\frac{\rho^{core}\left(\mathbf{r}'\right)}{\left|\mathbf{r}-\mathbf{r}'\right|}d\mathbf{r}'+\int\frac{\rho^{valence}\left(\mathbf{r}'\right)}{\left|\mathbf{r}-\mathbf{r}'\right|}d\mathbf{r}'\\&\equiv V_{Hartree}^{core}\left(\mathbf{r}\right)+V_{Hartree}^{valence}\left(\mathbf{r}\right)\end{aligned} \tag{8}$$

The "core" Hartree term in the GFT method is obtained from the "core" contribution of the density matrix as follows:

$$\begin{aligned}\left[\mathbf{V}_{Hartree}^{core}(\mathbf{Q})\right]_{\alpha\beta} &= \left\langle \chi_\alpha^{\mathbf{0}} \middle| V_{Hartree}^{core} \middle| \chi_\beta^{\mathbf{Q}} \right\rangle \\ &= \sum_\gamma^{core}\sum_\delta^{core}\sum_{\mathbf{Q}_1,\mathbf{Q}_2} D_{\gamma\delta}(\mathbf{Q}_1-\mathbf{Q}_2)\left\langle \chi_\alpha^{\mathbf{0}}\chi_\gamma^{\mathbf{Q}_1} \middle| \chi_\beta^{\mathbf{Q}}\chi_\delta^{\mathbf{Q}_2}\right\rangle \\ &+ \sum_\gamma^{core}\sum_\delta^{valence}\sum_{\mathbf{Q}_1,\mathbf{Q}_2} D_{\gamma\delta}(\mathbf{Q}_1-\mathbf{Q}_2)\left\langle \chi_\alpha^{\mathbf{0}}\chi_\gamma^{\mathbf{Q}_1} \middle| \chi_\beta^{\mathbf{Q}}\chi_\delta^{\mathbf{Q}_2}\right\rangle \\ &+ \sum_\gamma^{valence}\sum_\delta^{core}\sum_{\mathbf{Q}_1,\mathbf{Q}_2} D_{\gamma\delta}(\mathbf{Q}_1-\mathbf{Q}_2)\left\langle \chi_\alpha^{\mathbf{0}}\chi_\gamma^{\mathbf{Q}_1} \middle| \chi_\beta^{\mathbf{Q}}\chi_\delta^{\mathbf{Q}_2}\right\rangle\end{aligned} \tag{9-1}$$

$$\left\langle \chi_\alpha^{\mathbf{0}}\chi_\gamma^{\mathbf{Q}_1} \middle| \chi_\beta^{\mathbf{Q}}\chi_\delta^{\mathbf{Q}_2}\right\rangle = \iint \chi_\alpha^{\mathbf{0}}(\mathbf{r}_1)\chi_\beta^{\mathbf{Q}}(\mathbf{r}_1)\frac{1}{r_{12}}\chi_\gamma^{\mathbf{Q}_1}(\mathbf{r}_2)\chi_\delta^{\mathbf{Q}_2}(\mathbf{r}_2)\,d\mathbf{r}_1 d\mathbf{r}_2 \tag{9-2}$$

The lattice sum over a small number of sites is required since the core electrons are strongly localized around the center of the nucleus and their electron charges are thus perfectly compensated by the core nuclear charges. On the other hand, the "valence" Hartree term in the GFT method can be taken into account through the following Poisson's equation, where we can employ Fourier transform technique to solve the equation.

$$\nabla^2 V_{Hartree}^{valence}(\mathbf{r}) = -4\pi\rho^{valence}(\mathbf{r}) \tag{10-1}$$

$$V_{Hartree}^{valence}(\mathbf{r}) = \frac{4\pi}{N_{FT}}\sum_{\mathbf{G}}\frac{\rho(\mathbf{G})}{G^2}\exp(i\mathbf{G}\cdot\mathbf{r}) \tag{10-2}$$

$$\rho(\mathbf{G}) \equiv \sum_{\mathbf{r}_g}\rho(\mathbf{r}_g)\exp(-i\mathbf{G}\cdot\mathbf{r}_g) \tag{10-3}$$

We can employ a fast Fourier transform (FFT) method, and $\mathbf{r}_g$ represents a grid point for the FFT calculations. Thus, the "valence" Hartree term is obtained as follows,

$$\left[\mathbf{V}_{Hartree}^{valence}(\mathbf{Q})\right]_{\alpha\beta} = \left\langle \chi_\alpha^{\mathbf{0}} \middle| V_{Hartree}^{valence} \middle| \chi_\beta^{\mathbf{Q}} \right\rangle = \frac{4\pi}{N_{FT}}\sum_{\substack{\mathbf{G}\\ G\neq 0}}\frac{\rho(\mathbf{G})}{G^2}\left\langle \chi_\alpha^{\mathbf{0}} \middle| \exp(i\mathbf{G}\cdot\mathbf{r}) \middle| \chi_\beta^{\mathbf{Q}} \right\rangle \tag{11}$$

In the above equation, we omit the term of $G = 0$. The term will be discussed later.
The nuclear attraction potential is also divided into core and valence components:

$$V_{NA}(\mathbf{r}) = \sum_A \frac{Z_A}{|\mathbf{r}-\mathbf{R}_A|} = \sum_A \frac{Z_A^{core}}{|\mathbf{r}-\mathbf{R}_A|} + \sum_A \frac{Z_A^{valence}}{|\mathbf{r}-\mathbf{R}_A|} \equiv V_{NA}^{core}(\mathbf{r}) + V_{NA}^{valence}(\mathbf{r}) \tag{12}$$

Here, Z_A is the nuclear charge for atom number A. The "core" nuclear charge Z_A^{core} is defined as follows.

$$Z_A^{core} = \sum_{\alpha\in A}^{core}\sum_\beta^{core}\sum_{\mathbf{Q}} D_{\alpha\beta}(\mathbf{Q})S_{\beta\alpha}(\mathbf{Q}) + \sum_{\alpha\in A}^{core}\sum_\beta^{valence}\sum_{\mathbf{Q}} D_{\alpha\beta}(\mathbf{Q})S_{\beta\alpha}(\mathbf{Q}) + \sum_{\alpha\in A}^{valence}\sum_\beta^{core}\sum_{\mathbf{Q}} D_{\alpha\beta}(\mathbf{Q})S_{\beta\alpha}(\mathbf{Q}) \tag{13}$$

The remaining charge is assigned as the "valence" nuclear $Z_A^{valence}$ charge as follows.

$$Z_A^{valence} = Z_A - Z_A^{core} \tag{14}$$

The "core" nuclear attraction term is obtained as follows.

$$\left[\mathbf{V}_{NA}^{core}(\mathbf{Q})\right]_{\alpha\beta} = -\left\langle \chi_\alpha^{\mathbf{0}} \middle| V_{NA}^{core} \middle| \chi_\beta^{\mathbf{Q}} \right\rangle = -\left\langle \chi_\alpha^{\mathbf{0}} \middle| \sum_A \frac{Z_A^{core}}{|\mathbf{r}-\mathbf{R}_A|} \middle| \chi_\beta^{\mathbf{Q}} \right\rangle \tag{15}$$

Note that the "core" and "valence" nuclear charges are renewed in each self-consistent field (SCF) cycle. In the GFT method, the "valence" nuclear attraction potential is divided into short-range (SL) and long-range (LR) contributions, where $V_{NA}^{valence} = V_{SR-NA}^{valence} + V_{LR-NA}^{valence}$ by using the following error function (erf) and complementary error function (erfc).

$$\frac{1}{r} = \frac{erf(wr)}{r} + \frac{erfc(wr)}{r} \tag{16-1}$$

$$erf(wr) = \int_0^{wr} \exp\left(-t^2\right) dt \tag{16-2}$$

The short range (SR) "valence" nuclear attraction term is determined from the complementary error function (erfc) and the "valence" nuclear charges $Z_A^{valence}$ as follows,

$$\left[\mathbf{V}_{SR-NA}^{valence}(\mathbf{Q})\right]_{\alpha\beta} = -\left\langle \chi_\alpha^{\mathbf{0}} \middle| V_{SR-NA}^{valence} \middle| \chi_\beta^{\mathbf{Q}} \right\rangle = -\left\langle \chi_\alpha^{\mathbf{0}} \middle| \sum_A \frac{Z_A^{valence} erfc\left(\sqrt{\eta}|\mathbf{r}-\mathbf{R}_A|\right)}{|\mathbf{r}-\mathbf{R}_A|} \middle| \chi_\beta^{\mathbf{Q}} \right\rangle \tag{17}$$

The long range (LR) "valence" nuclear attraction term is obtained as follows,

$$\begin{aligned} \left[\mathbf{V}_{LR-NA}^{valence}(\mathbf{Q})\right]_{\alpha\beta} &= -\left\langle \chi_\alpha^{\mathbf{0}} \middle| \frac{Z_A^{valence} erf\left(\sqrt{\eta}|\mathbf{r}-\mathbf{R}_A|\right)}{|\mathbf{r}-\mathbf{R}_A|} \middle| \chi_\beta^{\mathbf{Q}} \right\rangle \\ &= -\frac{4\pi}{V_{cell}} \sum_A^{cell} \sum_{\substack{\mathbf{G} \\ (G\neq 0)}} \left(\frac{Z_Z^{valence}}{G^2}\right) \exp\left(-\frac{G^2}{4\eta}\right) \exp\left(-i\mathbf{G}\cdot\mathbf{R}_A\right) \int \chi_\alpha^{\mathbf{0}}(\mathbf{r}) \chi_\beta^{\mathbf{Q}}(\mathbf{r}) \exp\left(i\mathbf{G}\cdot\mathbf{r}\right) d\mathbf{r} \end{aligned} \tag{18}$$

In order to derive the above representation, we use the following equation.

$$\int \frac{erf\left(\sqrt{\eta}r\right)}{r} \exp\left(-i\mathbf{G}\cdot\mathbf{r}\right) d\mathbf{r} = \frac{4\pi}{G^2} \exp\left(-\frac{4\eta}{G^2}\right) \tag{19}$$

Equation (18) does not include the term of $G = 0$. The term will be discussed with the corresponding term of the Hartree potential in the section 2.4.

2.3 Effective Core Potential (ECP) and total energy fourmula

If the effective core potential (ECP) is employed together with the GFT method, the core electron density, $\rho^{core}(\mathbf{r})$, and nuclear charges, Z_A^{core}, become zero, and the ECP term of

$\mathbf{V}_{ECP}$ is used for the Fock matrix instead of core contributions of the nuclear attraction and Hartree terms as follows,

$$\begin{aligned}\mathbf{h}(\mathbf{Q}) &= \mathbf{T}(\mathbf{Q}) + \mathbf{V}_{ECP}(\mathbf{Q}) + \mathbf{V}_{SR-NA}(\mathbf{Q}) \\ &+ \mathbf{V}_{LR-NA}(\mathbf{Q}) + \mathbf{V}_{Hartree}(\mathbf{Q}) + \mathbf{V}_{XC}(\mathbf{Q})\end{aligned} \tag{20}$$

The total energy per unit cell in this scheme is obtained as follows,

$$\begin{aligned}E_{total} &= \sum_{\alpha,\beta}\sum_{\mathbf{Q}} D_{\alpha\beta}(\mathbf{Q})\left\{\left[\mathbf{T}(\mathbf{Q})\right]_{\alpha\beta} + \left[\mathbf{V}_{ECP}(\mathbf{Q})\right]_{\alpha\beta} + \left[\mathbf{V}_{SR-NA}(\mathbf{Q})\right]_{\alpha\beta}\right\} \\ &+ E_{XC} + \sum_{\substack{\mathbf{G}\\(G\neq 0)}} \frac{1}{2}\frac{4\pi V_{cell}}{N_{FT}^2}\frac{\left|\rho(\mathbf{G})\right|^2}{G^2} + \frac{\pi}{\eta V_{cell}}\left(\sum_{A}^{cell} Z_A\right)\sum_{\mathbf{Q}}\sum_{\alpha\beta} D_{\alpha\beta}(\mathbf{Q}) S_{\beta\alpha}(\mathbf{Q}) \\ &- \frac{4\pi}{N_{FT}}\sum_{A}^{cell}\sum_{\substack{\mathbf{G}\\(G\neq 0)}} \frac{Z_A\rho(\mathbf{G})}{G^2}\exp\left(-\frac{G^2}{4\eta}\right)\exp\left(i\mathbf{G}\cdot\mathbf{R}_A\right) + \frac{1}{2}\sum_{\mathbf{Q}}\sum_{\substack{A,B\\(\mathbf{R}_A-\mathbf{R}_B-\mathbf{Q}\neq\mathbf{0})}}^{cell} \frac{Z_A Z_B erfc\left(\sqrt{\eta}\left|\mathbf{R}_A-\mathbf{R}_B-\mathbf{Q}\right|\right)}{\left|\mathbf{R}_A-\mathbf{R}_B-\mathbf{Q}\right|} \\ &+ \frac{1}{2}\frac{4\pi}{V_{cell}}\sum_{\substack{\mathbf{G}\\(G\neq 0)}}\sum_{A,B}^{cell} Z_A Z_B \frac{1}{G^2}\exp\left(-\frac{G^2}{4\eta}\right)\exp\left(i\mathbf{G}\cdot\left(\mathbf{R}_B-\mathbf{R}_A\right)\right) \\ &- \left(\sum_{A} Z_A^2\right)\sqrt{\frac{\eta}{\pi}} - \frac{1}{2}\left[\sum_{A}^{cell} Z_A\right]^2\frac{\pi}{\eta V_{cell}}\end{aligned} \tag{21}$$

Here, $\mathbf{G}$ is the reciprocal lattice vector. E_{XC} is the exchange-correlation energy of the unit cell, which is written as $E_{XC} = 0.5\sum_{\mathbf{Q}} Tr\left[\mathbf{D}(\mathbf{Q})\mathbf{V}_X^{Fock}(\mathbf{Q})\right]$ in the Hartree-Fock approximation. N_{FT} is the number of grids for the Fourier transform. $\rho(\mathbf{G}) = \sum_{\mathbf{r}_g}\rho(\mathbf{r}_g)\exp(-i\mathbf{G}\cdot\mathbf{r}_g)$, where $\mathbf{r}_g$ is the grid point for FFT calculations. Last four terms of the total energy come from the Ewald-type representation for the nuclear-nuclear repulsion energy.

2.4 Constant term

In this section, we discuss the terms of $G = 0$, which appears in the nuclear attraction and Hartree terms, and the nuclear-nuclear repulsion. The total energy formula of eq. (21) includes the constant terms, which are derived from considerations of $G = 0$. The Fourier coefficient of the electron density behaves in the limit of $G = 0$ as follows,

$$\begin{aligned}\lim_{G\to 0}\rho(G) &= \lim_{G\to 0}\sum_{\mathbf{r}_g}\rho(\mathbf{r}_g)\exp(-i\mathbf{G}\cdot\mathbf{r}_g) \\ &\simeq \lim_{G\to 0}\left[\sum_{\mathbf{r}_g}\rho(\mathbf{r}_g)\left(1 - iGr_g\cos\theta + \frac{1}{2}G^2 r_g^2\cos^2\theta\right)\right] \\ &\approx \sum_{\mathbf{r}_g}\rho(\mathbf{r}_g) + \beta G^2 = N_{FT}\bar{\rho} + \beta G^2\end{aligned} \tag{22}$$

Here, we use $\lim_{G\to 0}\sum_{\mathbf{r}_g}\rho\left(\mathbf{r}_g\right)\left(-iGr_g\cos\theta\right)=0$ and $\bar{\rho}=\left(1/N_{FT}\right)\sum_{\mathbf{r}_g}\rho\left(\mathbf{r}_g\right)$. β is a constant, however it disappears in the final form of the total energy, as shown bellow. On the other hand, the Hartree energy per unit cell is determined from the following equation:

$$\begin{aligned}E_{Hartree}&=\frac{1}{2}\int\limits_{V_{cell}}V_{ele}\left(\mathbf{r}\right)\rho\left(\mathbf{r}\right)d\mathbf{r}\\&=\int\limits_{V_{cell}}\frac{4\pi}{N_{FT}}\sum_{\mathbf{G}}\frac{\rho\left(\mathbf{G}\right)}{G^2}\exp\left(i\mathbf{G}\cdot\mathbf{r}\right)\frac{1}{N_{FT}}\sum_{\mathbf{G}'}\rho\left(\mathbf{G}'\right)\exp\left(i\mathbf{G}'\cdot\mathbf{r}\right)d\mathbf{r}\\&=\frac{4\pi}{N_{FT}^2}\sum_{\mathbf{G},\mathbf{G}'}\frac{1}{G^2}\rho\left(\mathbf{G}\right)\rho\left(\mathbf{G}'\right)V_{cell}\delta_{\mathbf{G}',-\mathbf{G}}=\frac{4\pi}{N_{FT}^2}V_{cell}\sum_{\mathbf{G}}\frac{\left|\rho\left(\mathbf{G}\right)\right|^2}{G^2}\end{aligned}\tag{23}$$

Here, we use $\left(1/V_{cell}\right)\int_{V_{cell}}\exp\left(i\left(\mathbf{G}-\mathbf{G}'\right)\cdot\mathbf{r}\right)d\mathbf{r}=\delta_{\mathbf{G},\mathbf{G}'}$ and $\rho^*\left(\mathbf{G}\right)=\rho\left(-\mathbf{G}\right)$. If we consider the limit of $G=0$, the Hartree energy can be described as follows,

$$\begin{aligned}E_{Hartree}&=\frac{1}{2}\frac{4\pi}{N_{FT}^2}V_{cell}\sum_{\mathbf{G},(G\neq 0)}\frac{\left|\rho\left(\mathbf{G}\right)\right|^2}{G^2}+\lim_{G\to 0}\frac{1}{2}\frac{4\pi}{N_{FT}^2}V_{cell}\frac{\left|\rho\left(\mathbf{G}\right)\right|^2}{G^2}\\&\approx\sum_{\mathbf{G},(G\neq 0)}\frac{1}{2}\frac{4\pi}{N_{FT}^2}V_{cell}\frac{\left|\rho\left(\mathbf{G}\right)\right|^2}{G^2}+\lim_{G\to 0}\frac{1}{2}\frac{4\pi}{N_{FT}^2}V_{cell}\frac{\left(N_{FT}\bar{\rho}+\beta G^2\right)\left(N_{FT}\bar{\rho}+\beta G^2\right)}{G^2}\\&\approx\sum_{\mathbf{G},(G\neq 0)}\frac{1}{2}\frac{4\pi V_{cell}}{N_{FT}^2}\frac{\left|\rho\left(\mathbf{G}\right)\right|^2}{G^2}+\lim_{G\to 0}\frac{1}{2}4\pi V_{cell}\bar{\rho}^2\frac{1}{G^2}+\frac{4\pi}{N_{FT}}V_{cell}\beta\bar{\rho}\end{aligned}\tag{24}$$

Next, we discuss the long-range nuclear attraction energy.

$$\begin{aligned}E_{LR-NA}&=-\int_{V_{cell}}V_{LR-NA}\left(\mathbf{r}\right)\rho\left(\mathbf{r}\right)d\mathbf{r}\\&=-\int\limits_{V_{cell}}\frac{4\pi}{V_{cell}}\sum_{\mathbf{G}}\sum_{A}^{cell}Z_A\frac{1}{G^2}\exp\left(-\frac{G^2}{4\eta}\right)\exp\left(i\mathbf{G}\cdot\left(\mathbf{r}-\mathbf{R}_A\right)\right)\frac{1}{N_{FT}}\sum_{\mathbf{G}'}\rho\left(\mathbf{G}'\right)\exp\left(i\mathbf{G}'\cdot\mathbf{r}\right)d\mathbf{r}\\&=-\frac{4\pi}{N_{FT}}\sum_{A}^{cell}\sum_{\mathbf{G}}Z_A\frac{1}{G^2}\exp\left(-\frac{G^2}{4\eta}\right)\exp\left(i\mathbf{G}\cdot\mathbf{R}_A\right)\rho\left(\mathbf{G}\right)\end{aligned}\tag{25}$$

Then, the following asymptotic equation is obtained from the above term in the $G\to 0$ case.

$$\begin{aligned}-\lim_{G\to 0}\frac{4\pi}{N_{FT}}\sum_{A}^{cell}\frac{Z_A\rho\left(\mathbf{G}\right)}{G^2}\exp\left(-\frac{G^2}{4\eta}\right)\exp\left(i\mathbf{G}\cdot\mathbf{R}_A\right)&\approx-\lim_{G\to 0}\frac{4\pi}{N_{FT}}\sum_{A}^{cell}\frac{Z_A\left(1-\frac{G^2}{4\eta}\right)\left(N_{FT}\bar{\rho}+\beta G^2\right)}{G^2}\\&\approx-\lim_{G\to 0}4\pi\left(\sum_{A}^{cell}Z_A\right)\bar{\rho}\frac{1}{G^2}-\frac{4\pi}{N_{FT}}\left(\sum_{A}^{cell}Z_A\right)\beta+4\pi\left(\sum_{A}^{cell}Z_A\right)\bar{\rho}\frac{1}{4\eta}\\&=-\lim_{G\to 0}4\pi V_{cell}\bar{\rho}^2\frac{1}{G^2}-\frac{4\pi}{N_{FT}}V_{cell}\beta\bar{\rho}+\frac{\pi}{\eta V_{cell}}\left(\sum_{A}^{cell}Z_A\right)\sum_{\mathbf{Q}}\sum_{\alpha\beta}\mathbf{D}_{\alpha\beta}^{\mathbf{Q}}\mathbf{S}_{\alpha\beta}^{\mathbf{Q}}\end{aligned}\tag{26}$$

Here, $\sum_A^{cell} Z_A = V_{cell}\bar{\rho}$ and $\sum_A Z_A = \sum_{\mathbf{r}_g} \rho(\mathbf{r}_g)$. The second term of eq. (26), which includes β, is cancelled out by the corresponding constant term of the Hartree energy of eq. (24).
The asymptotic description for nucleus-nucleus repulsion term is obtained from the Ewald-type long range interaction as follows,

$$\begin{aligned}
&\lim_{G\to 0}\frac{1}{2}\frac{4\pi}{V_{cell}}\sum_{A,B}^{cell} Z_A Z_B \frac{1}{G^2}\exp\left(-\frac{G^2}{4\eta}\right)\exp\left(i\mathbf{G}\cdot(\mathbf{R}_B-\mathbf{R}_A)\right) \\
&\simeq \lim_{G\to 0}\frac{1}{2}\frac{4\pi}{V_{cell}}\sum_{A,B}^{cell} Z_A Z_B \frac{1}{G^2}\left(1-\frac{G^2}{4\eta}\right) \\
&= \lim_{G\to 0}\frac{1}{2}4\pi V_{cell}\bar{\rho}^2\frac{1}{G^2} - \frac{1}{2}\left[\sum_A^{cell} Z_A\right]^2\frac{\pi}{\eta V_{cell}} \qquad (27)
\end{aligned}$$

The first term of eq. (27) is cancelled by the corresponding terms of eqs (24) and (26). The compensation of the self-interaction of the long-range nucleus-nucleus interaction also brings in the constant term, and the term can be obtained from $A = B$ in the long-range nucleus-nucleus interaction as follows,

$$\begin{aligned}
&\frac{1}{2}\frac{4\pi}{V_{cell}}\sum_{\mathbf{G}}\sum_A^{cell} Z_A^2 \frac{1}{G^2}\exp\left(-\frac{G^2}{4\eta}\right)\exp\left(i\mathbf{G}\cdot\mathbf{0}\right) \\
&= \frac{1}{2}\sum_A^{cell} Z_A^2 \lim_{r\to\infty}\frac{erf\left(\sqrt{\eta}r\right)}{r} = \sum_A^{cell} Z_A^2\sqrt{\frac{\eta}{\pi}}
\end{aligned} \qquad (28)$$

Here, we use $\lim_{r\to 0} erf\left(\sqrt{\eta}r\right)/r \approx 2\sqrt{\eta/\pi}$.

2.5 Recursion relation

In order to obtain the integrations in eqs (11) and (18), we can use the following recursion relation.

$$\begin{aligned}
&\langle \mathbf{a}+\mathbf{1}_\xi|\exp(i\mathbf{G}\cdot\mathbf{r})|\mathbf{b}\rangle = \left(P_\xi - R_\xi^A + \frac{iG_\xi}{2p}\right)\langle \mathbf{a}|\exp(i\mathbf{G}\cdot\mathbf{r})|\mathbf{b}\rangle \\
&+\frac{1}{2p}N_\xi(\mathbf{a})\langle \mathbf{a}-\mathbf{1}_\xi|\exp(i\mathbf{G}\cdot\mathbf{r})|\mathbf{b}\rangle + \frac{1}{2p}N_\xi(\mathbf{b})\langle \mathbf{a}|\exp(i\mathbf{G}\cdot\mathbf{r})|\mathbf{b}-\mathbf{1}_\xi\rangle \qquad (29\text{-}1)
\end{aligned}$$

$$\langle \mathbf{0}|\exp(i\mathbf{G}\cdot\mathbf{r})|\mathbf{0}\rangle = \exp\left(-\mu|\mathbf{R}_A-\mathbf{R}_B|^2\right)\left(\frac{\pi}{p}\right)^{\frac{3}{2}}\exp\left(-\frac{G^2}{4p}\right)\exp\left(i\mathbf{G}\cdot\mathbf{P}\right) \qquad (29\text{-}2)$$

$$|\mathbf{a}\rangle = \left(x-R_x^A\right)^{a_x}\left(y-R_y^A\right)^{a_y}\left(z-R_z^A\right)^{a_z}\exp\left(-g_a|\mathbf{r}-\mathbf{R}_A|^2\right) \qquad (29\text{-}3)$$

Here, $p = g_a + g_b$, $\mu = g_a g_b/(g_a+g_b)$, $\mathbf{a} = \begin{pmatrix} a_x & a_y & a_z\end{pmatrix}$, $N_\xi(\mathbf{a}) = a_\xi$, and $\mathbf{1}_\xi = \begin{pmatrix}\delta_{x\xi} & \delta_{y\xi} & \delta_{z\xi}\end{pmatrix}$ using Kronecker's delta. $\mathbf{P} = (g_a\mathbf{R}_A + g_b\mathbf{R}_b)/(g_a+g_b)$. ξ represents one of x, y, or z. The recursion relation is an expansion of the Obara and Saika (OS) technique for atomic orbital (AO) integrals.(Obara et al. 1986)

3. Screened Hartree-Fock exchange potential

3.1 Dielectric function and screened exchange potential

The screening effect caused by electron correlations is an important factor in determining the electronic structure of solid-state materials. The Fock exchange term can be represented as a bare interaction between electron and exchange hole in the Hartree–Fock approximation.(Parr et al. 1994) The electron correlation effect screens the interaction. In this section, we discuss the screening effect for bulk materials, especially semiconductors.

The screening effect is closely related to the electric part of the dielectric function. The Thomas–Fermi model is a well-known dielectric model function for free electron gas.(Yu et al. 2005; Ziman 1979)

$$\varepsilon^{TF}(\mathbf{k}) = 1 + \left[\left(\frac{k^2}{k_{TF}^2}\right)\right]^{-1} \tag{30}$$

Here, k_{TF} is the Thomas–Fermi wave number. Although the Thomas–Fermi model is applicable for metallic system, it is not suitable to semiconductors because it diverges when $k = 0$. The dielectric constant of semiconductors must take a finite value at $k = 0$. Therefore, a number of different dielectric function models for semiconductors have been proposed for semiconductors,(Levine et al. 1982; Penn 1962) and Bechstedt et al. proposed the following model to reproduce the property of semiconductors.(Bechstedt et al. 1992; Cappellini et al. 1993)

$$\varepsilon^{Bechstedt}(\mathbf{k}) = 1 + \left[(\varepsilon_s - 1)^{-1} + \alpha\left(\frac{k^2}{k_{TF}^2}\right) + \frac{k^4}{4k_F^2 k_{TF}^2/3}\right]^{-1} \tag{31}$$

Here, k_F is the Fermi wave number whose value depends on the average electron density. ε_s is the electric part of the dielectric constant. The value of coefficient α is determined in such way that it fits to the random phase approximation (RPA) calculation, and Bechstedt et al. reported that the values of α do not display a strong dependence on the material type. In this paper, we employ $\alpha = 1.563$ according to their suggestion. The most important point is that the Bechstedt's model does not diverge at $k = 0$; $\varepsilon^{Bechstedt}(k = 0) = \varepsilon_s$. In this paper, we simplify Bechstedt's model, and employ the following dielectric function model.

$$\varepsilon(\mathbf{k}) = 1 + \left[(\varepsilon_s - 1)^{-1} + \alpha\left(\frac{k^2}{k_{TF}^2}\right)\right]^{-1} \tag{32}$$

The following equation is obtained from the above equation through the inversion.

$$\frac{1}{\varepsilon(\mathbf{k})} = \left(1 - \frac{1}{\varepsilon_s}\right)\frac{k^2}{k^2 + \tilde{k}_{TF}^2} + \frac{1}{\varepsilon_s} \tag{33-1}$$

$$\tilde{k}_{TF}^2 = \frac{k_{TF}^2}{\alpha}\left(\frac{1}{\varepsilon_s - 1} + 1\right) \tag{33-2}$$

Then, we obtain the following screened potential from these equations and the Fourier transform.

$$V(\mathbf{r}) = \frac{1}{(2\pi)^3}\int \frac{4\pi}{k^2\varepsilon(\mathbf{k})}\exp(i\mathbf{k}\cdot\mathbf{r})d\mathbf{k} = \left(1 - \frac{1}{\varepsilon_s}\right)\frac{\exp\left(-\tilde{k}_{TF}r\right)}{r} + \frac{1}{\varepsilon_s}\frac{1}{r} \tag{34}$$

The first term in the above equation represents the short range screened potential. However, since the screening effect is not complete in the semiconductors, the partial bare interaction appears in the second term. Conversely, for metallic systems, i.e., $\varepsilon_s \to \infty$, complete screening is achieved, and the second term disappears.

The Yukawa type potential, $\exp(-q_Y r)/r$, is difficult to handle with Gaussian basis sets, and therefore we employ $erfc(wr)/r$ instead of the Yukawa potential because the both functions behaves similarly if the relation $q_Y = 3w/2$ holds true.(Shimazaki et al. 2008) The use of a complementary error function provides a highly efficient algorithm for calculating Gaussian-based atomic orbital integrals. Thus, we obtain the following approximation.

$$V(\mathbf{r}) \approx \left(1 - \frac{1}{\varepsilon_s}\right)\frac{erfc\left(2\tilde{k}_{TF}r/3\right)}{r} + \frac{1}{\varepsilon_s}\frac{1}{r} \tag{35}$$

Based on the above discussions, we employ the following screened Fock exchange in this paper.

$$\begin{aligned}\left[\mathbf{V}_X^{screened}(w,\varepsilon_s)\right]_{\alpha\beta} &= -\sum_{\gamma,\delta} D_{\gamma\delta}\iint \chi_\alpha(\mathbf{r}_1)\chi_\gamma(\mathbf{r}_1) \\ &\times\left[\left(1 - \frac{1}{\varepsilon_s}\right)\frac{erfc(wr_{12})}{r_{12}} + \frac{1}{\varepsilon_s}\frac{1}{r_{12}}\right]\chi_\beta(\mathbf{r}_2)\chi_\delta(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \\ &\equiv \left(1 - \frac{1}{\varepsilon_s}\right)\left[\mathbf{V}_{SR-X}^{erfc}(w)\right]_{\alpha\beta} + \frac{1}{\varepsilon_s}\left[\mathbf{V}_X^{Fock}\right]_{\alpha\beta}\end{aligned} \tag{36}$$

Here, $w = 2\tilde{k}_{TF}/3$. The screened Fock exchange includes parameters such as $\tilde{k}_{TF}$ and ε_s, which strongly depend on the material. When $\varepsilon_s = 1$, eq. (36) reduces to the ordinary Fock exchange; $\mathbf{V}_X^{screened}\left(\tilde{k}_{TF},1\right) = \mathbf{V}_X^{HF}$.

3.2 Local potential approximation

The semiconductors discussed in this paper have a large Thomas-Fermi wave vector; thus, the screening length becomes small and the first term of eq. (36) mainly takes into account short-range interactions and small non-local contributions. This potentially allows the first term to be approximated by a local potential and to neglect its non-local contribution. In this paper, we examine the LDA functional as a replacement for the first term of eq. (36). Although the LDA functional is not the same as the local component of first term of eq. (36), this replacement can expand the scope of eq. (36), because electron correlations other than the screening effect can be taken into account through the LDA functional. We should note that Bylander et al. employed a similar strategy.(Bylander et al. 1990) In this paper, we examine the following potentials:

$$\mathbf{V}^{screened}\left(\varepsilon_s\right) = \left(1 - \frac{1}{\varepsilon_s}\right)\mathbf{V}^{Slater} + \frac{1}{\varepsilon_s}\mathbf{V}^{Fock} \tag{37}$$

$$\mathbf{V}^{screened}\left(\varepsilon_s\right) = \left(1 - \frac{1}{\varepsilon_s}\right)\mathbf{V}^{Slater} + \frac{1}{\varepsilon_s}\mathbf{V}^{Fock} + \mathbf{V}^{VWN} \tag{38}$$

Here, $\mathbf{V}^{Slater}$ is the Slater exchange term, and $\mathbf{V}^{VWN}$ is the Vosko-Wilk-Nusair (VWN) correlation term. The above potentials appear to be types of hybrid-DFT functional. The relation between the screened HF exchange potential and the hybrid-DFT functional is discussed later. Equations (37) and (38) have a system-dependent HF exchange fraction, which is unusual for the ordinal hybrid-DFT method. In metal systems, that is, $\varepsilon_s \to \infty$, the above potential reduces to the ordinal LDA functional. It is worth noting that eqs. (37) and (38) depend on only ε_s, although eq. (36) depends on two parameters, namely, $\tilde{k}_{TF}$ and ε_s.

3.3 Self consistent scheme for dielectric constant

In eqs (36), (37), and (38), the fraction of the Fock exchange term is proportional to the inverse of the dielectric constant. Consequently, in order to use these equations, we must know the value of the dielectric constant for the target semiconductor. Although an experimentally obtained value is a possible candidate, here we discuss a self-consistent scheme for theoretically considering the dielectric constant. In this scheme, the static dielectric constant is assumed to be obtained from the following equations: (Ziman 1979)

$$\varepsilon_s = 1 + \left({}^{\omega_p}\!\big/_{\bar{E}_{gap}} \right)^2 \tag{39-1}$$

$$\bar{E}_{gap} = \frac{1}{K}\sum_{\mathbf{k}}^{K}\left(\lambda_{LUMO}^{\mathbf{k}} - \lambda_{HOMO}^{\mathbf{k}}\right) \tag{39-2}$$

Here, $\lambda_{HOMO}^{\mathbf{k}}$ and $\lambda_{LUMO}^{\mathbf{k}}$ are the HOMO and LUMO energies, respectively, at wave vector $\mathbf{k}$. $\bar{E}_{gap}$ is the average energy gap, and ω_p is the plasma frequency, the value of which is obtained from $\omega_p = \sqrt{4\pi n_e^{valence}}$. Equation (39-1) depends on the averaged energy gap because the dielectric constant reflects overall responses of k-space. The equation is combined with equations (36), (37), or (38) in the self-consistent-field (SCF) loop, and ε_s and $\bar{E}_{gap}$ are calculated and renewed in each SCF step. Here, the fraction of the HF exchange term, which is proportional to ε_s^{-1}, is not constant throughout the SCF cycle. We obtain the self-consistent dielectric constant and the energy band structure after the iterative procedure. Notably, this self-consistent scheme does not refer to any experimental results.

4. Application for GFT method and screened HF exchange potential

4.1 Diamond, silicon, AlP, AlAs, GaP, and GaAs

In this section, we present the energy band structures of the following semiconductors: diamond (C), silicon (Si), AlP, AlAs, GaP, and GaAs. We discuss the electronic structures of these semiconductors on the basis of the HF method, the local density approximation (LDA), and the hybrid-DFT method. The Slater–Vosko–Wilk–Nusair (SVWN) functional (Slater 1974; Vosko et al. 1980) and the B3LYP functional,(Becke 1993 b) the latter of which includes 80% of the Slater local density functional $\mathbf{V}^{Slater}$, 72% of the Becke88 (B88)-type gradient correction $\Delta\mathbf{V}^{B88}$ (Becke 1988) , and 20% of the HF exchange term, are employed for the LDA and the hybrid-DFT calculations, respectively. The B3LYP functional is shown below:(Becke 1993 b)

$$\mathbf{V}_{XC}^{B3LYP} = 0.8\mathbf{V}^{Slater} + 0.72\boldsymbol{\Delta}\mathbf{V}^{B88} + 0.2\mathbf{V}^{Fock} + \mathbf{V}_{C}^{B3LYP} \tag{40-1}$$

$$\mathbf{V}_{C}^{B3LYP} = 0.19\mathbf{V}_{C}^{VWN} + 0.81\mathbf{V}_{C}^{LYP} \tag{40-2}$$

Here, $\mathbf{V}_{C}^{LYP}$ is the Lee–Yang–Parr correlation functional.(Lee et al. 1988) We used the 6-21G* basis set, which was proposed by Catti et al.(Catti et al. 1993), for diamond calculations. On the other hand, we employ the effective core potential proposed by Stevens et al. for silicon, AlP, AlAs, GaP, and GaAs.(Stevens et al. 1984; Stevens et al. 1992) The exponents and contraction coefficients listed in our previous paper are employed for the atomic orbitals for Si, Al, P, As, and Ga.(Shimazaki et al. 2010) We employ $24 \times 24 \times 24$ k-points for the Fourier transform technique, and $25 \times 25 \times 25$ mesh grid points are used, to calculate the valence electron contribution of the Hartree term. In addition to these, we employ the truncation condition of third neighboring cells. It should be noted that the truncation affect only the HF exchange and "core" Hartree (Coulomb) terms. In this section, we also present calculation results obtained from the screening HF exchange potential.

	C	Si	AlP	AlAs	GaP	GaAs
Lattice constant	6.74	10.26	10.30	10.70	10.30	10.68
V_{cell}	76.76	207.11	273.10	305.90	273.10	304.29
r_s	1.32	2.01	2.01	2.09	2.01	2.09
k_{Fermi}	1.46	0.96	0.95	0.92	0.95	0.92
k_{TF}	1.36	1.10	1.10	1.08	1.10	1.08
ε_s	5.65	12.1	7.54	8.16	10.75	12.9
$\tilde{k}_{TF}$	1.2	0.92	0.95	0.92	0.93	0.90
ε_s^{-1}	0.18	0.083	0.13	0.12	0.093	0.078

Table 1. Parameters for semiconductors in atomic unit [a.u.]

Table 1 presents the lattice constants and the dielectric constants of those semiconductors; the lattice constant, the volume of the unit cell V_{cell}, and the screening parameters are given in atomic units (a.u.). Here, the eight valence electrons in the unit cell are considered for calculating the screening parameters. These parameters are used in the screened HF exchange potential, for example ε_s in Table 1 is used for eqs (36), (37), and (38). On the other hand, calculation results based on the self-consistent procedure are presented in Section 3.2.

Table 2 presents the direct and indirect bandgaps calculated by the SVWN, HF, and B3LYP methods for semiconductors, here we also show experimental bandgap values (Yu et al. 2005). The direct bandgap of GaAs is the same as the minimum energy difference. The SVWN functional underestimates the bandgaps in comparison with the experimental ones. The kind of underestimation is a well-known problem of LDA. On the other hand, the HF method overestimates the bandgap properties, and the B3LYP method yields better calculation results. However, the calculation results of B3LYP are more complex than the LDA and HF; for example, the B3LYP functional gives calculation results that are close to the experimental bandgap in diamond case, but the same functional overestimates the bandgaps for AlAs, AlP, and GaP. The B3LYP functional yields the indirect bandgap of 3.3 eV for AlAs, whereas the experimental property is 2.2 eV. While the B3LYP functional gives 3.6 eV for the indirect bandgap of AlP, the experimental one is 2.5 eV. In the case of GaP, the B3LYP and experimental bandgaps are 3.3 eV and 2.4 eV, respectively. The B3LYP functional can reproduce the experimental band structure of diamond well, however the results are poorer for other semiconductors such as AlAs, AlP, and GaP.

		HF [ev]	SVWN [eV]	B3LYP [eV]	Exp.[eV]
C	Direct	14.6	5.9	7.4	7.3
	Indirect	12.6	4.2	6.0	5.48
Si	Direct	8.0	2.0	3.3	3.48
	Indirect	6.1	0.50	1.8	1.11
AlP	Direct	11.4	4.0	5.6	3.6
	Indirect	8.5	2.0	3.6	2.5
AlAs	Direct	9.6	2.7	4.0	3.13
	Indirect	7.8	1.7	3.3	2.23
GaP	Direct	9.1	2.2	3.5	2.89
	Indirect	8.0	1.9	3.3	2.39
GaAs	Direct	6.8	0.86	1.91	1.52

Table 2. Theoretical and experimental bandgaps of semiconductors [eV]

Next, we discuss the screened HF exchange potential discussed in Section 3. The direct and indirect bandgaps calculated by the screened HF exchange potential are presented in Table 3. The overall calculation results are better than those from the SVWN, HF, and B3LYP methods. Equation (36) tends to underestimate the indirect bandgap, however eqs (37) and (38), which use the Slater functional instead of $\mathbf{V}_{SR-X}^{erfc}(w)$, show good agreement with the experimental results. The underestimation obtained from eq. (36) may cause that the equation takes into account only the screening effect. On the other hand, the VWN correlation functional of eq. (38) slightly improves the calculation results.

However, it should be noted that there is a lager gap between the experimental direct bandgap of AlP and our calculation result. The experimental value determined by photoluminescence spectroscopy is 3.62 eV (Monemar 1973), and our calculation result of eq. (38) is 4.9 eV. Zhu et al. noted that the experimentally obtained spectrum was broad and poorly defined due to a high concentration of defects in the AlP sample.(Zhu et al. 1991)

They also contended that the transition from Γ_{15v} to X_{3c} was assigned in error. They calculated 4.38 eV as the direct bandgap by the GW method, which is closer to our calculation result.

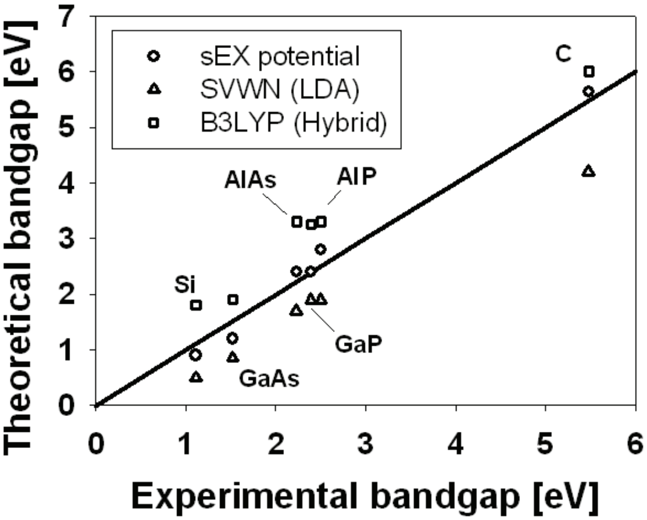


Fig. 1. Experimental and theoretical bandgap properties. Circles indicate calculation results based on screened HF exchange method (sEX) of eq. (38) with experimental dielectric constant, and triangles and squares denote SVWN and B3LYP results, respectively

		Eq. (36) [eV]	Eq. (37) [eV]	Eq. (38) [eV]
C	Direct	7.1	6.9	7.0
	Indirect	5.0	5.5	5.6
Si	Direct	2.4	2.4	2.5
	Indirect	0.44	0.77	0.90
AlP	Direct	5.1	4.9	5.0
	Indirect	2.1	2.6	2.8
AlAs	Direct	3.6	3.4	3.4
	Indirect	1.6	2.2	2.4
GaP	Direct	2.8	2.6	2.7
	Indirect	1.9	2.2	2.4
GaAs	Direct	1.4	1.2	1.2

Table 3. Bandgaps obtained from screened HF exchange potential with experimental ε_s

The theoretical bandgaps of diamond, silicon, AlP, AlAs, GaP, and GaAs, which are obtained from SVWN, B3LYP, and eq. (38), are shown with the experimental bandgaps in Figure 1. From the figure, we can easily confirm that the LDA (SVWN) functional underestimates the experimental bandgap. On the other hand, the B3LYP method reproduces the experimental results for diamond well, but overestimates AlP, AlAs, and GaP. The screened HF exchange potential shows good agreements with experiment (Yu et al. 2005). The energy band structures of diamond, silicon, AlP, AlAs, GaP, and GaAs, which are obtained from eq. (38), are presented in Figure 2.

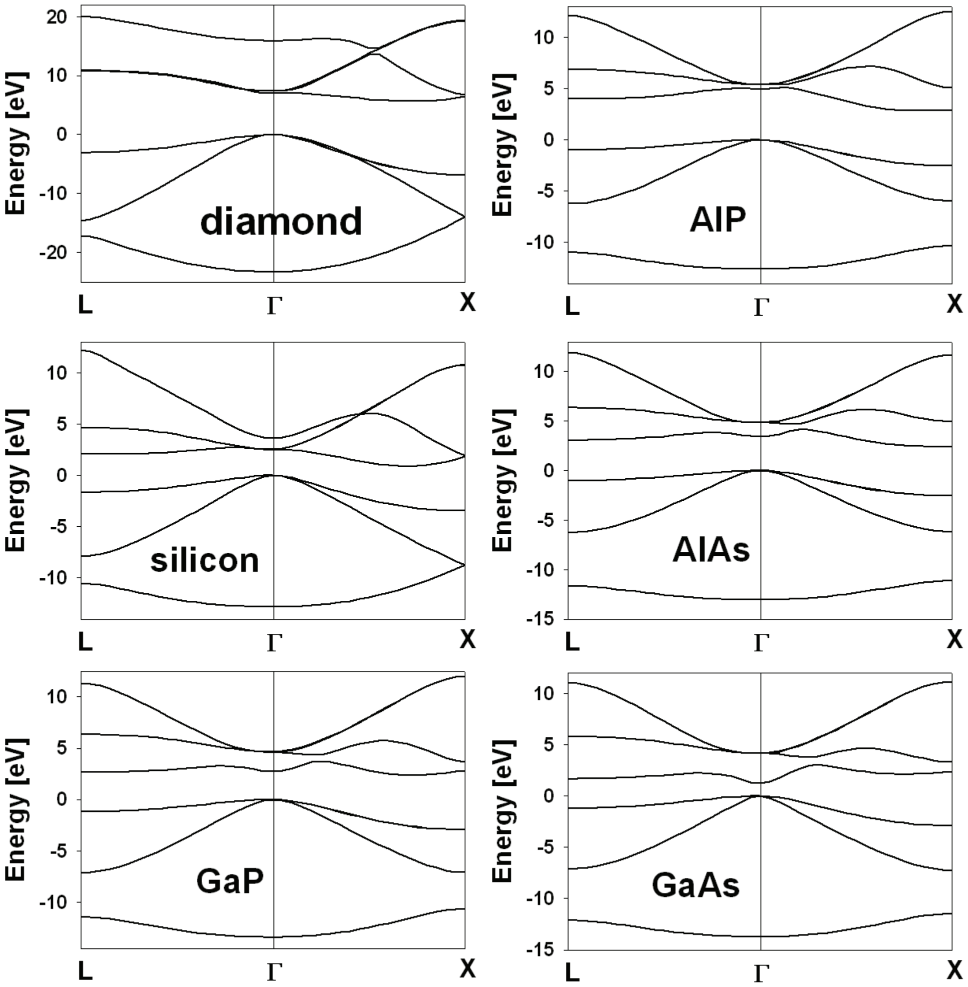


Fig. 2. Energy band structure calculated using eq. (38) with experimental dielectric constant: diamond, silicon, AlP, AlAs, GaP, and GaAs

4.2 Self-consistent calculation for dielectric constant

We summarize the calculation results with the self-consistent dielectric constant, the scheme of which is discussed in Section 2.4, in Table 4. The self-consistent scheme brings in calculation results that are similar to those obtained by using an experimental dielectric constant; for example, eq. (36) based on the self-consistent dielectric constant yields 5.5 eV for the indirect bandgap of diamond, and the use of the experimental dielectric constant yields 5.6 eV.

We demonstrate change of ε_s in the SCF cycle of eq. (36) combined with eq. (39) for diamond in Figure 3. In the figure, we prepare for two different starting (initial) electronic structures; one is the HF electronic structure, and the other is the LDA-SVWN one. In the HF reference calculation, ε_s is underestimated at the early stages of iterative calculations, and then converged to the final value. Conversely, the procedure started from the LDA-SVWN overestimates the dielectric constant at the early stages. There are differences in the initial steps of the self-consistent (SC) cycles, however those dielectric constants are converged to the same value through the iterative calculations. Thus, the same energy band structure is obtained from the SCF cycles even if the initial electronic structures are different. In other words, the self-consistent method does not depend on the starting (initial) electronic structure. On the other hand, the single-shot method, in which the SCF loop is only once calculated, strongly depends on the reference electronic structure. The HF-referenced single-shot calculation underestimates the dielectric constant, $\varepsilon = 2.9$, and it overestimates the bandgap property; the direct and indirect bandgap are 8.5 eV and 6.3 eV, respectively, because the HF method tends to overestimate the bandgap property. On the other hand, the SVWN-referenced single-shot method overestimates the dielectric constant, $\varepsilon_s = 8.1$, and underestimates the bandgap property; the direct and indirect bandgaps are 6.6 eV and 4.4 eV, respectively. Thus, the single-shot calculations yield different results.

Table 5 lists the theoretically determined dielectric constants based on eqs (36), (37), and (38). These calculation results present slight underestimations of the dielectric constant.

		Eq. (36) [eV]	Eq. (37) [eV]	Eq. (38) [eV]
C	Direct	7.0	6.7	6.8
	Indirect	4.9	5.3	5.5
Si	Direct	2.5	2.5	2.6
	Indirect	0.48	0.8	0.95
AlP	Direct	4.9	4.8	4.9
	Indirect	2.0	2.5	2.7
AlAs	Direct	3.5	3.3	3.4
	Indirect	1.5	2.1	2.4
GaP	Direct	2.9	2.8	2.8
	Indirect	2.0	2.3	2.5
GaAs	Direct	1.5	1.3	1.3

Table 4. Bandgaps obtained from the screened HF exchange potential with self-consistent dielectric constant

	Eq. (36)	Eq. (37)	Eq. (38)
C	6.31	6.55	6.51
Si	10.90	11.14	10.83
AlP	8.97	8.69	8.33
AlAs	9.14	8.92	8.58
GaP	8.65	8.87	8.63
GaAs	9.27	9.75	9.55

Table 5. Dielectric constants calculated from self-consistent scheme based on eq. (39)

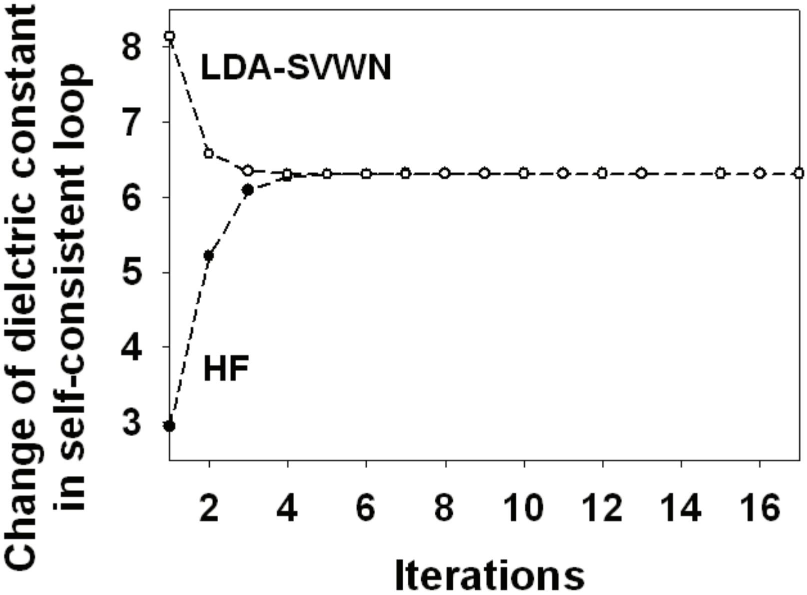


Fig. 3. The change of the dielectric constant in the self-consistent (SC) method for diamond. The black filled circles indicate the SC procedure which starts from the HF energy band structure. The white ones represent the SC procedure started from the LDA-SVWN energy band structure. Two different initials yield the same electronic structure through the iterative procedure. (Shimazaki et al. 2009 b)

5. Discussion

We summarize the inverse of the dielectric constant in Table 1. Those values represent the fraction of the HF exchange term incorporated into the screened HF exchange potential; for example, about 18% and about 8% of the HF exchange terms are used in the screened HF exchange potential for diamond and silicon, respectively. On the other hand, 20% is used for all material in the B3LYP functional. In the case of diamond, the fraction in the proposed method takes a value similar to the fraction in the B3LYP method. However, the HF fraction of the B3LYP functional is larger compared with those of other semiconductors. The B3LYP functional potentially overestimates the bandgap values, other than that of diamond, because a larger fraction of the HF exchange term causes a larger energy bandgap. The HF fraction of the B3LYP functional is set to reproduce the properties of the G1 basis set, which mainly covers light elements such as N, C, and O, and small molecules such as methane, ammonia, and silane.(Curtiss et al. 1990; Pople et al. 1989) Thus, the parameter set of the B3LYP functional is especially suitable for organic molecules. However, the B3LYP functional is not designed for solid-state materials. In order to employ the hybrid-DFT method to solid-state materials, the fraction of the HF exchange term must be decided appropriately.

Here, we emphasize the similarity between the screened HF exchange potential and the hybrid-DFT method. While eqs (37) and (38) are derived from the model dielectric function of eq. (32) and the local potential approximation, these equations appear to be a type of hybrid-DFT functional. The hybrid-DFT method was introduced by Becke in 1993 by using the adiabatic connection,(Becke 1993 a) and some empirical justifications, such as compensation of the intrinsic self-interaction error (SIE) of semi-local exchange-correlation functional, have been discussed.(Janesko et al. 2009) On the other hand, from the careful observation of actual behaviors of HF and semi-local DFT calculation, the mixing of the HF fraction is reported to bring in useful cancelation, because the semi-local DFT functional have a tendency to overestimate the strength of covalent bonds, and the HF method has the opposite feature.(Janesko et al. 2009) Now, we have proposed an interpretation that the HF fraction represents the incompleteness of the screening effect in semiconductors. Besides, its incompleteness can be described by the inverse of the electronic component of the dielectric constant. This discussion will be helpful to determine an appropriate HF exchange fractions for the target solid-state material.

The screened HF exchange method can be regard as a type of the generalized Kohn-Shan (GKS) method.(Seidl et al. 1996) In the GKS framework, the screened-exchange LDA (sX-LDA) method, which is proposed by Seidl et al., can reproduce eigenvalue gaps in good agreement with experimental bandgaps of several semiconductors. They also presented a calculation result for germanium, employing a semiconductor dielectric function model proposed by Bechstedt et al., and reported that the screening effect of the Bechstedt model is weaker than the Thomas-Fermi model. This feature should correspond to the incompleteness of the screening effect of semiconductors discussed in Section 3.1 because our dielectric function can be derived from a simplification of the Bechstedt model. We should note that the true quasi-particle bandgap is different from the band gap of the GKS method due to the derivative discontinuity of the exchange-correlation potential. However, the discontinuity is, to some extent, incorporated in the GKS single-particle eigenvalues. This fundamental feature of the GKS formalism brings in the improvements of the bandgap calculations of the screened HF exchange method.

Next, we discuss the HSE functional including a splitting parameter ω. The splitting parameter is used to divide the potential into short- and long- range interactions, where the relation of $1/r = erfc(\omega r)/r + erf(\omega r)/r$ is used. The HSE functional has a form similar to our screened HF exchange potential due to the use of $erfc(\omega r)/r$. However, we need to pay attention to the value of ω. In the HSE functional refined by Krukau et al. in 2006, $\omega = 0.11$ is recommended for the parameter.(Krukau et al. 2006) Conversely, in our screened HF exchange potential, the corresponding parameter takes about 0.8. Thus, in the HSE functional, the term of $erfc(\omega r)/r$ can take into account a longer interaction than ours. On the other hand, in our method, the term including $erfc(\omega r)/r$ can represent only short-range interaction because of a large ω value. The long-range interaction in our method is incorporated by the bare HF exchange interaction represented by the second term of eq. (36). The HSE functional has a different theoretical background from our method. Therefore, even if the similar term appears in both methods, the physical meaning is different.

Although the screened HF exchange method and the GW method are taken into account in real space and momentum space, respectively, the both theoretical concepts may be similar, especially in the Coulomb hole plus screened exchange (COHSEX) approximation, because the dielectric function plays an important role in both methods. Gygi et al. have reported that the diagonal-COHSEX approximation has a tendency to underestimate the indirect bandgap property.(Gygi et al. 1986) This feature of diagonal-COHSEX approximation resembles calculation results determined by eq. (36). The both neglect the energy dependence of the self-energy, and this simplification possibly causes the underestimation of the indirect bandgap property.

In order to describe the screened HF exchange method, we adopt the Gaussian-based formalism; however, our method is not restricted to Gaussian basis sets, and can be used together with other basis set such as the plane-wave basis set. Conversely, the linear muffin-tin orbital (LMTO) and linearized augmented plane wave (LAPW) methods can taken into account the HF exchange term,(Martin 2004) thus our methodology can be easily introduced and implemented in these methods.

6. Summary

This chapter explains the GFT method, which is based on the Gaussian-basis formalism. In the GFT method, the periodic Hartree potential is expanded by auxiliary plane waves, and those expansion coefficients can be calculated by Fourier transform method. We discuss that this simple approach enables us to estimate the Hartree term efficiently. In addition to this, we discuss the screened HF exchange potential, which has a close relationship to the hybrid-DFT method and the GW approximation. In the screened HF exchange potential, the fraction of the HF exchange term is proportional to the inverse of the static dielectric constant, and therefore it depends on the target material. In this chapter, we present not only experimental values but also a self-consistent scheme for the estimation of the dielectric constant. We also discuss that the local potential approximation can expand the possibility of the screened HF exchange method, and it is useful to speculation between the screening effect and the HF fraction term appeared in the hybrid DFT functional.

We have demonstrated the energy band structure of diamond, silicon, AlP, AlAs, GaP, and GaAs from the GFT method and the screened HF exchange potential. The combination of these methodologies can reproduce the experimental bandgap property well. On the other hand, the HF method overestimates the bandgap, while the local DFT (SVWN) method underestimate the bandgap. These kinds of discrepancy between theory and experiment cause the manipulation of the HF exchange term. The fraction of the HF exchange term is closely related to the screening effect, and thus we need to determine the fraction appropriately according to the target system. The discussion in this chapter will be a helpful guideline to determine the fraction.

7. Acknowledgement

This work was partly supported by a Grant-in-Aid for Young Scientists (Start-up) (No. 21850002) from the Japan Society for the Promotion of Science (JSPS).

8. References

Aryasetiawan, F. and O. Gunnarsson. (1998). The GW method, *Rep. Prog. Phys.*, 61, 237-312.

Bechstedt, F., R. D. Sole, G. Cappellini and L. Reining. (1992). AN EFFICIENT METHOD FOR CALCULATING QUASIPARTICLE ENERGIES IN SEMICONDUCTORS, *Solid State Comm.*, 84, 765-770.

Becke, A. D. (1988). Density-functional exchange-energy approximation with correct asymptoic behavior, *Phys. Rev. A*, 38, 3098-3100.

Becke, A. D. (1993 b). Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.*, 98, 5648-5652.

Becke, A. D. (1993 a). A new mixing of Hartree-Fock and local density-functional theories, *J. Chem. Phys.*, 98, 1372-1377.

Bylander, D. M. and L. Kleinman. (1990). Good semiconductor band gaps with a modified local-density approximation, *Phys. Rev. B*, 41, 7868-7871.

Cappellini, G., R. D. Sole, L. Reining and F. Bechstedt. (1993). Model dielectric function for semiconductors, *Phys. Rev. B*, 47, 9892-9895.

Catti, M., A. Pavese, R. Dovesi and V. R. Saunders. (1993). Static lattice and electron properties of $MgCO_3$ (magnesite) calculated by *ab initio* periodic Hartree-Fock methods, *Phys. Rev. B*, 47, 9189-9198.

Curtiss, L. A., C. Jones, G. W. Trucks, K. Raghavachari and J. A. Pople. (1990). Gaussian-1 theory of molecular energies for second-row compounds, *J. Chem. Phys.*, 93, 2537-2545.

Delhalle, J., L. Piela, J.-L. Bredas and J.-M. Andre. (1980). Multiple expansion in tight-binding Hartree-Fock calculations for infinite model polymers, *Phys. Rev. B*, 22, 6254-6267.

Gygi, F. and A. Baldereschi. (1986). Self-consistent Hartree-Fock and screened-exchange calculations in solids: Application to silicon, *Phys. Rev. B*, 34, 4405-4408(R).

Hedin, L. (1965). New Method for Calculating the One-Particle Green's Function with Applicaton to the Electron-Gas Problem, *Phys. Rev.*, 139, A796-A823.

Heyd, J., G. E. Scuseria and M. Ernzerhof. (2003). Hybrid functionals based on a screened Coulomb potential, *J. Chem. Phys.*, 118, 8207.

Hirata, S. and T. Shimazaki. (2009). Fast second-order many-body perturbation method for extended systems, *Phys. Rev. B*, 80, 085118.

Hirata, S., O. Sode, M. Keceli and T. Shimazaki (2010). Electron correlation in solids: Delocalized and localized orbital approaches, *Accurate Condensed-Phase Quantum Chemistry*, F. R. Manby (Ed.), 129-161, CRC Press, Boca Raton.

Janesko, B. G., T. M. Henderson and G. E. Scuseria. (2009). Screened hybrid density functionals for solid-state chemistry and physics, *Phys. Chem. Chem. Phys.*, 11, 443-454.

Krukau, A. V., O. A. Vydrov, A. F. Izmaylov and G. E. Scuseria. (2006). Influence of the exchange screening parameter on the performance of screened hybrid functionals, *J. Chem. Phys.*, 125, 224106.

Kudin, K. and G. E. Scuseria. (2000). Linear-scaling density-functional theory with Gaussian orbitals and periodic boundary conditions: Efficient evaluation of energy and forces via the fast multipole method, *Phys. Rev. B*, 61, 16440-16453.

Ladik, J. J. (1999). *Phys. Rep.*, 313, 171.

Lee, C., W. Yang and R. G. Parr. (1988). Development of the Colle-Savetti correlation-energy formula into a functional of the electron density, *Phys. Rev. B*, 37, 785-789.

Levine, Z. H. and S. G. Louie. (1982). New model dielectric function and exchange-correlation potential for semiconductors and insulators, *Phys. Rev. B*, 25, 6310-6316.

Martin, R. M. (2004). *Electronic Structure, Basic Theory and Practical Methods*, Cambrige University Press, Cambridge.

Monemar, B. (1973). Fundamental Energy Gaps of AlAs and AlP from Photoluminescence Excitation Spectra, *Phys. Rev. B*, 8, 5711-5718.

Obara, S. and A. Saika. (1986). Efficient recursive computation of molecular integrals over Cartesian Gaussian functions, *J. Chem. Phys.*, 84, 3963-3974.

Parr, R. G. and W. Yang (1994). *Density-functional Theory of Atoms and Molecules*, Oxford Univ. Press, New York.

Penn, D. R. (1962). Wave-Number-Dependent Dielectric Function of Semiconductors, *Phys. Rev.*, 128, 2093-2097.

Piani, C. and R. Dovesi. (1980). Exact-exchange Hartree-Fock calculations for periodic systems. I. Illustration of the method, *Int. J. Quantum Chem.*, 17, 501-516.

Pisani, C., R. Dovesi and C. Roetti (1988). *Hartree-Fock Ab Initio Treatment of Crystalline Systems*, Springer-Verlag, Berlin.

Pople, J. A., M. Head-Gordon, D. J. Fox, K. Raghavachari and L. A. Curtiss. (1989). Gaussian-1 theory: A general procedure for prediction of molecular energies, *J. Chem. Phys.*, 90, 5622-5629.

Seidl, A., A. Görling, P. Vogl, J. A. Majewski and M. Levy. (1996). Generalized Kohn-Sham schemes and the band-gap problem, *Phys. Rev. B*, 53, 3764-3774.

Shimazaki, T. and Y. Asa. (2010). Energy band structure calculations based on screened Hartree–Fock exchange method: Si, AlP, AlAs, GaP, and GaAs, *J. Chem. Phys.*, 132, 224105.

Shimazaki, T. and Y. Asai. (2008). Band structure calculations based on screened Fock exchange method, *Chem. Phys. Lett.*, 466, 91.

Shimazaki, T. and Y. Asai. (2009 a). Electronic Structure Calculations under Periodic Boundary Conditions Based on the Gaussian and Fourier Transform (GFT) Method, *J. Chem. Theory Comput.*, 5, 136-143.

Shimazaki, T. and Y. Asai. (2009 b). First principles band structure calculations based on self-consistent screened Hartree–Fock exchange potential, *J. Chem. Phys.*, 130, 164702.

Shimazaki, T. and S. Hirata. (2009 c). On the Brillouin-Zone Integrations in Second-Order Many-Body Perturbation Calculations for Extended Systems of One-Dimensional Periodicity, *Int. J. Quan. Chem.*, 109, 2953.

Slater, J. C. (1974). *The Self-Consistent Field for Molecules and Solids, Quantum Theory of Molecules and Solids*, McGraw-Hill, New York.

Stevens, W. J., H. Basch and M. Krauss. (1984). Compact effective potentials and efficient shared-exponent basis sets for the first- and second-row atoms, *J. Chem. Phys.*, 81, 6026-6033.

Stevens, W. J., M. Krauss, H. Basch and P. G. Jasien. (1992). Relativistic compact effective potentials and efficient, shared-exponent basis sets for the third-, fourth-, and fifth-row atoms, *Can. J. Chem.*, 70, 612-630.

Vosko, S. H., L. Wilk and M. Nusair. (1980). Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis, *Canadian J. Phys.*, 58, 1200-1211.

Yu, P. Y. and M. Cardona (2005). *Fundamental of Semiconductors*, Springer, New York.

Zhu, X. and S. G. Louie. (1991). Quasiparticle band structure of thirteen semiconductors and insulators, *Phys. Rev. B*, 43, 14142-14156.

Ziman, J. M. (1979). *Principles of the Theory of Solids*, Cambridge University Press, Cambridge.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

Tomomi Shimazaki and Yoshihiro Asai (2011). Gaussian and Fourier Transform (GFT) Method and Screened Hartree-Fock Exchange Potential for First-principles Band Structure Calculations, Fourier Transforms - Approach to Scientific Principles, Prof. Goran Nikolic (Ed.), ISBN: 978-953-307-231-9, InTech, Available from: http://www.intechopen.com/books/fourier-transforms-approach-to-scientific-principles/gaussian-and-fourier-transform-gft-method-and-screened-hartree-fock-exchange-potential-for-first-pri

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Low Complexity Fourier Transforms using Multiple Square Waves

Khoirul Anwar[1] and Minoru Okada[2]
[1]School of Information Science,
Japan Advanced Institute of Science and Technology (JAIST),
[2]Graduate School of Information Science,
Nara Institute of Science and Technology (NAIST)
Japan

1. Introduction

Fourier Transform (FT) is widely applied in digital mobile cellular radio systems. The implementation requires low power consumption and smaller chip size. The primary factor of the FT applications is its chip complexity. The complexity is typically expressed in terms of number of adders, the number of multiplier, data storage and control complexity rather than the speed of operation.

The current divide and conquer technique in fast Fourier transform (FFT) reduces the number of operations in conventional discrete Fourier transform (DFT) by utilizing the advantage of complex twiddle factors instead of matrix multiplications (Oppenheim, 1990). The computation of DFT is decomposed into nested smaller DFTs which are computed separately and combined to give the final results. FFT reduces the number of multiplier which account of much of the chip area and power consumption in digital hardware design.

However, a pipeline FFT processor is characterized by real time continuous processing of an input data sequence. It is difficult to initiate the FFT operation until all of the N sampled data are taken. Another complexity issue is the arithmetic unit, especially multipliers, that requires larger area than a digital register. To meet real-time processing in FFT with size of N, the multiplicative complexity of $N \log_r N$ is required (r is generally the radix). It contributes the complexity of the processor and power consumption.

Another consideration of FFT is the data storage or memory for buffering the data and intermediate results of the real time computations. The butterfly at the first stage has to take the input data elements separated by N/r from the sequence. The required memory becomes another major chip area issue especially for large Fourier transform.

The facts expressed above need to be improved so that the amounts of power consumption, chip area and complexity are suitable especially for handheld transceiver. Since the power consumption is directly related to the number of complex multiplications, an algorithm to reduce or replace these multiplications is important.

In (Shattil and Nassar, 2002), a simple computation of Fourier transform using a square-wave is introduced. A mathematical derivation shows that it is possible to replace the complex multiplication in Fourier transforom by additions. However, the performance evaluation of the method in (Shattil and Nassar, 2002) is not available to make sure the effectiveness of the method.

In this paper, we propose double square-waves (DS) that completely replace complex multiplications by sampling and additions for Fourier and inverse Fourier transforms called DSFT. The proposed method only requires sampler, multiplier and filter to remove the harmonic components of square-wave. Our results confirm that DS-FT is applicable to any system that requires Fourier transforms such as orthogonal frequency division multiplexing (OFDM) (Nee and Prasad, 2000), multicarrier code division multiple access (MC-CDMA), FFT-based carrier interferometry spreading (Anwar and Yamamoto, 2006) and other techniques that requires FFT.

2. Important

This chapter presents a simple computation method for Fourier Transform (FT) and its inverse (IFT) by employing multiple square waves (MSW), whose complex multiplications are replaced by simple additions. Since the square wave is superposition of harmonic sinusoids, a simple mathematical derivation shows that fast Fourier transform (FFT) and its inverse can be performed by MSW with low computational complexity. MSW replaces the complex twiddle factor multiplications in FFT/IFFT by simple adding operation. The main parts of this chapter is adapted from (Takahashi et. al., 2007).

The orthogonality of FFT/IFFT is still kept, by which the bit-error-rate (BER) performance is satisfactory. Compared to the standard single square wave (SSW), our results confirm that excellent BER performance is achievable without error floor. Furthermore, the proposed multiple square wave for Fourier transform (MSW-FT) is free from restriction in its size (e.g. power of two, etc.) and is useful for signal processing of multi-carrier system, such as orthogonal frequency division multiplexing (OFDM), and multi-carrier code division multiple access (MC-CDMA), WiMAX, single carrier frequency division multiple access (SC-FDMA) and other frequency domain processing such as frequency domain turbo equalization. The proposed MSW-FT and MSW-IFT are less complex than FFT or IFFT, which is suitable to digital communication systems, where the power consumption constraint is considered.

3. System model

We consider an OFDM system as the model to evaluate the effectiveness of the proposed DS-FT. Fig. 1 describes the transceiver structure of OFDM system where its FFT is replaced by DS-FT. Inverse DS-FT, called DS-IFT, is located at the transmitter while DS-FT is located at the receiver. The N incoming data symbols are converted from serial to parallel. Then $(L - 1)N$ zeros are added to the center of the parallel data to obtain the oversampled signal, where L is the oversampling factor. The LN data symbols (with zero padding) are converted to time-domain signals using DS-IFT. After filtering, guard interval (GI) is inserted. The OFDM signals are then transmitted to the channel.

At the receiver, first GI is removed, then the signals are converted to frequency domain signals by DS-FT. From the frequency-domain signals, padded zeros are removed. Finally, we obtain the data.

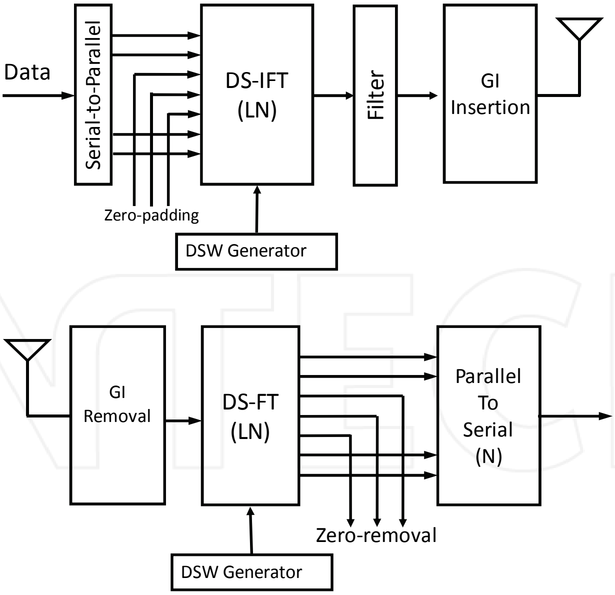


Fig. 1. System Model of OFDM WLAN where its FFT is replaced by DS-FT

4. Proposed double square-waves for Fourier transform

4.1 Square-wave model

The frequency-domain signal $X(f_n)$ converted from timedomain symbol x_k by discrete Fourier transform (DFT) is expressed by

$$\begin{aligned} X(f_n) &= \sum_{k=0}^{K-1} x_k e^{-j2\pi f_n k t_0} \\ &= \sum_{k=0}^{K-1} x_k \Psi(f_n,t) \end{aligned} \tag{1}$$

where f_n is an n-th frequency component, K is the number of time-domain samples and t_0 is the interval of time-domain samples. The exponential function $\Psi(f_n,t)$ is expressed by Euler's theorem as

$$\begin{aligned} \Psi(f_n,t) &= e^{-j2\pi f_n t} \\ &= \cos 2\pi f_n t - j\sin 2\pi f_n t, \end{aligned} \tag{2}$$

From (1) and (2), at least $K \times K$ complex multiplications are required. In addition, a lot of phases should be restored when performing the multiplications or additions. To replace a large number of multiplications, we propose to use square-waves which consist of only 2 levels of amplitude as a substitute for exponential function in DFT.

The single square-wave function for n-th frequency can be expressed as a sum of harmonic sinusoids as (Kreyzig, 1993)

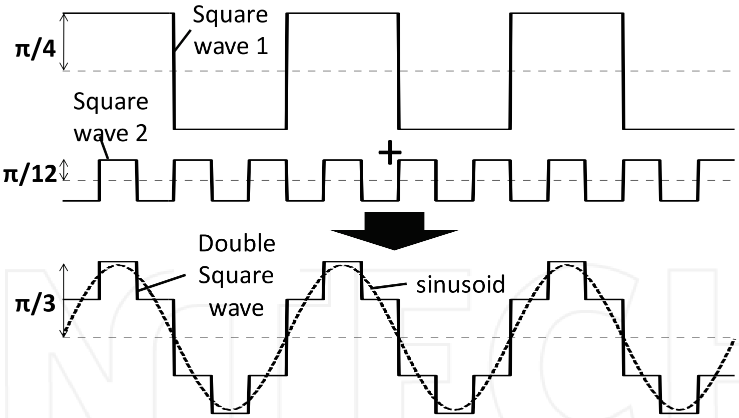


Fig. 2. Double Square-waves consists of $\pi/4$ and $\pi/12$ single square-waves

$$
\begin{aligned}
x_{ss}(2\pi f_n t) &= \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{\sin((2k-1)2\pi f_n t)}{(2k-1)} \\
&= \frac{4}{\pi}\left\{\sin 2\pi f_n t + \frac{1}{3}\sin(3\cdot 2\pi f_n t) + \cdots\right\}.
\end{aligned}
\tag{3}
$$

Here, (3) can be rewrittten as

$$
\begin{aligned}
\sin 2\pi f_n t &= \frac{\pi}{4}x_{ss}(2\pi f_n t) - \frac{1}{3}\sin(3\cdot 2\pi f_n t) \\
&\quad - \frac{1}{5}\sin(5\cdot 2\pi f_n t) - \frac{1}{7}\sin(7\cdot 2\pi f_n t) - \cdots \\
&= \frac{\pi}{4}x_{ss}(2\pi f_n t) \\
&\quad - \frac{1}{3}\left\{\sin(3\cdot 2\pi f_n t) + \frac{1}{3}\sin(9\cdot 2\pi f_n t) + \cdots\right\} \\
&\quad - \frac{1}{5}\sin(5\cdot 2\pi f_n t) - \frac{1}{7}\sin(7\cdot 2\pi f_n t) - \cdots \\
&= \frac{\pi}{4}x_{ss}(2\pi f_n t) - \frac{1}{3}\left\{\frac{\pi}{4}x_{ss}(3\cdot 2\pi f_n t)\right\} \\
&\quad - \frac{1}{5}\sin(5\cdot 2\pi f_n t) - \frac{1}{7}\sin(7\cdot 2\pi f_n t) - \cdots .
\end{aligned}
\tag{4}
$$

4.2 Order of truncation

Due to the hardware limitation, truncation is required in performing sin $2\pi f_n t$ in (4). This subsection discusses errors caused by the truncation of (4). We construct a sinusoid by some square-waves and measure the average error that shows how the signal is similar with the perfect sinusoid signal. The order of number of square-waves is 1, 2, 3, · · ·, 12. The result is plotted in Fig. 3.

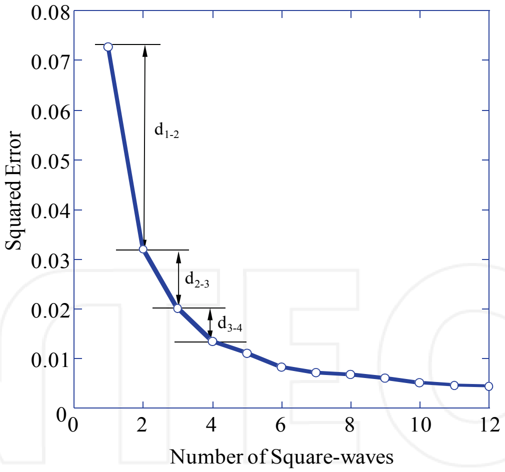


Fig. 3. Average squared-error of square-waves based sinusoid compared to a perfect sinusoid-waves

It is shown when we have a single square-wave, the average error is about 0.073 while it is 0.032 with a double squarewaves. The difference between a single square-wave and double square-waves, $d_{1\text{-}2}$, is about 0.041. That means double square-waves improve about 0.041 of average error. Increasing the number of square-waves more than 3 does not significantly reduce the average error, i.e. $d_{1\text{-}2} > d_{2\text{-}3} > d_{3\text{-}4} >$.

On the other hand, using square-waves more than 3 will increase the computational complexity in hardware. We conclude that double square-waves is enough to keep lower error and hardware complexity.

4.3 Double square-wave transform

As a consequence of result in Subsection 4.2, it is reasonable to assume

$$-\frac{1}{5}\sin(5\cdot 2\pi f_n t) - \frac{1}{7}\sin(7\cdot 2\pi f_n t) - \cdots \cong 0 \tag{5}$$

such that we obtain

$$\sin 2\pi f_n t \cong \frac{\pi}{4}\left\{x_{ss}(2\pi f_n t) - \frac{1}{3}x_{ss}(3\cdot 2\pi f_n t)\right\}. \tag{6}$$

From (6), it is shown that the sinusoid can be composed by combining two square-waves of different amplitudes and different periods. We call it as double square-wave (DS) signal and it is noted as $x_{ds}(2\pi f_n t)$.

Because additions require less computational complexity than subtractions, we modify the phase of second wave by π to prevent the subtraction. Then double square-wave function $x_{ds}(2\pi f_n t)$ is expressed by

$$x_{ds}(2\pi f_n t) = \frac{\pi}{4}\left\{x_{ss}(2\pi f_n t) + \frac{1}{3}x_{ss}(3\cdot 2\pi f_n t + \pi)\right\}.. \tag{7}$$

The number of samples is 96.

Now, we can obtain the function $\Psi_{ds-ft}(f_n,t)$ for DS-FT and $\Psi_{ds-ift}(f_n,t)$ for DS-IFT as substitution of exponential function $\Psi(f_n,t)$ in (2) as

$$\Psi_{ds-ft}(f_n,t) = x_{ds}(2\pi f_n t + \frac{\pi}{2}) - j \cdot x_{ds}(2\pi f_n t), , \tag{8}$$

$$\Psi_{ds-ift}(f_n,t) = x_{ds}(2\pi f_n t + \frac{\pi}{2}) + j \cdot x_{ds}(2\pi f_n t). . \tag{9}$$

Finally, we can express the frequency-domain signal $X(f_n)$ as

$$X(f_n) \cong \sum_{k=0}^{K-1} x_k \Psi_{ds-ft}(f_n, kt_0), , \tag{10}$$

and the time-domain signal $x(f_n)$ as

$$x(f_n) \cong \sum_{k=0}^{K-1} X_k \Psi_{ds-ift}(f_n, kt_0). . \tag{11}$$

4.4 Computational complexity

The square-wave generator is simpler than the sinusoid generator because it uses digital logic. It doesn’t need the complex analog multiplier and can be replaced by a simple hardware.

An inverter can be used to multiply the data by −1, while multiplication of +1 is possible by copying the signal. Compared to the conventional single square-wave method, our proposed method needs an addtional multiplication by 1/3. However, multiplication by a constant is not too complex in a hardware. Therefore, multiplication by double square-waves is easier in hardware implementation than multiplying by a sinusoid.

5. Performance evaluation

This section evaluates signal resolution and BER performances using the proposed DS-FT compared to that of single square-wave Fourier transform (SS-FT) (Shattil and Nassar, 2002) (Bates et. al., 1970).

5.1 Signal resolution

Figures 4(a) and (b) show the signal resolution of a sinc function. The sinc waveform is represented by the dashed line that has been sampled in 96 samples with normalized amplitude. The sinc waveform by SS-FT is shown in Fig. 4(a), while that by DS-FT is shown in Fig. 4(b). It is shown that the resolution of SS-FT can not reach the maximum while the left and right parts of signals are too high. The sinc waveform represented by the proposed DS-FT has better quality than that of SS-FT.

5.2 BER performances evaluation

In this subsection, to confirm the effectiveness of the proposed method, we evaluate the BER performances of an OFDM system where its FFT and IFFT are replaced by DSFT and

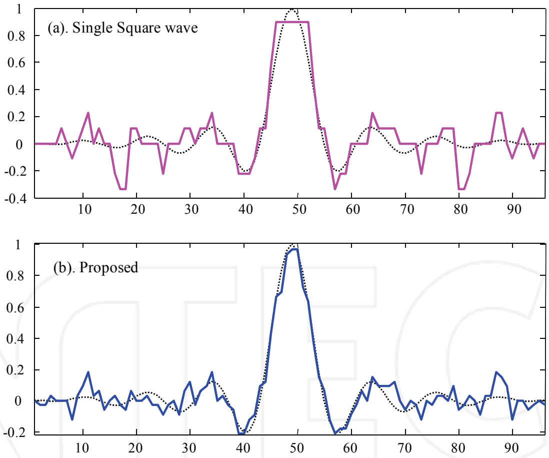


Fig. 4. Sinc waveform using (a) single square-wave and (b) double squarewaves

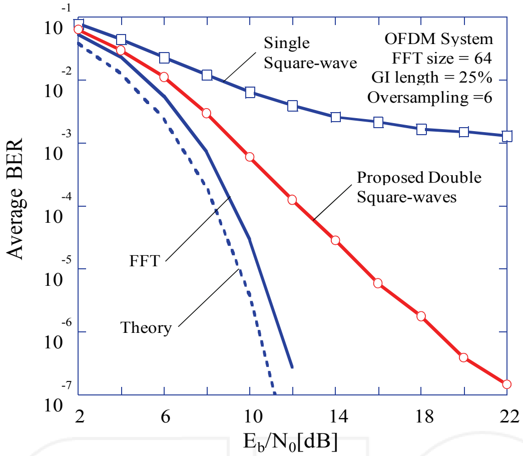


Fig. 5. BER performances of WLAN using FFT, SS-FT and the proposed DS-FT method

DS-IFT. Evaluation of DS-FT using signal resolution is not enough. Thus, evaluation of the BER is important to make sure the effectiveness of sampling and orthogonality guarantee. The parameters used for BER performance evaluation are shown in Table I which expected to be the condition of IEEE802.11a/g Wireless LAN system.

The modulation is QPSK for OFDM system with number of subcarrier is 52, as in Wireless LAN system. We use oversampling factor of 6 to observe efficiently the signal resolution. GI length is 25% of the symbol length. The overall simulation is performed in additive white Gaussian noise (AWGN) channel without error correction coding.

The BER performancess are plotted in Fig. 5. The dashed line is a theoretical BER performance of QPSK symbol for reference. The BER of OFDM with FFT has degradation by about 1 dB as a consequence of guard interval (GI) insertion with length of 1/4 or 25% of the length of OFDM symbol. SS-FT has residual bit error at 1.5 × 10^{-3}. Increasing the number of oversampling does not change the BER performance of SS-FT. The reason is that the orthogonality can not be kept by the SS-FT.

The proposed DS-FT does not have residual bit error (up to 10^{-7}) though it has BER degradation by about 2dB at BER level of 10^{-3}. Increasing the oversampling factor will increase the BER performance. But the oversampling factor enhancement should consider a practical reason related to the additional complexity.

Parameters	**Value(s)**
Modulation	QPSK
Number of Subcarriers	52
FFT size	64
GI Length	16 (25%)
Oversampling factor (*L*)	6
Channel	AWGN

Table 1. Simulation Parameters

6. Conclusion

In this paper, we propose DS-FT and evaluate it in the OFDM system. The DS-FT comprises double square-waves to simplify the Fourier transform computation with better signal resolution and BER performance compared to the Fourier transform using single square-wave. The double square-waves can be easily generated by two weighted single square waves with different periods. DS-FT contributes lower computational complexity of Fourier transform by replacing the complex multiplication with sampling, addition and filtering (only at the transmitter). Therefore, power consumption (related to the number of multiplication) and chip area (related to the memory) can be reduced by DS-FT with allowable performance degradation.

In our future work, we will consider the filter at the output of DS-FT to obtain a better signal resolution by completely removing the harmonic frequency components.

7. References

Anwar, K. & Yamamoto, H. (2006). A New Design of Carrier Interferometry OFDM with FFT as spreading code, *Proceedings of IEEE Radio and Wireless Symposium (RWS 2006),* pp. 543-546.

Bates, R. H. T; Napier, P. J. & Chang, Y. P. (1970). Square-Wave Fourier Transform. *Electronic Letters,* Vol. 6, No. 23, pp. 741-742.

Kreyzig, E. (1993). *Advanced Engineering Mathematics,* John Wiley & Sons, 7th Edition.

Nee, R. V. & Prasad, R. (2000). *OFDM for Wireless Multimedia Communications,* Artech House Publishing.

Oppenheim, A. V. & Schaffer, R. W. (1999). *Discrete Time Signal Processing,* Prentice Hall

Shattil, S. & Nassar, C. R. (2002). Improved Fourier Transform for Multicarrier Processing, *Proceedings of SPIE, Emerging Technologies for Future Generation Wireless Communications,* Vol. 4869, pp. 35-41.

Takahashi, H. ; Anwar, K. ; Saito, M. & Okada, M. (2007). Low-Complexity Fourier Transform using Double Square-waves. *Proceesings of IEEE ISCIT 2007*, Australia.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Orbital Stability of Periodic Traveling Wave Solutions

Jaime Angulo Pava[1] and Fábio Natali[2]
[1]Department of Mathematics, IME-USP - São Paulo - SP
[2]Department of Mathematics, Universidade Estadual de Maringá - Maringá - PR
Brazil

1. Introduction

The theory of stability of periodic traveling waves associated with evolution partial differential equations of dispersive type has increased significantly in the last five years. A good number of researchers are interested in solving a rich variety of new mathematical problems due to physical importance related to them. This subject is often studied in relation to perturbations of symmetric classes, e.g., the class of periodic functions with the same minimal period as the underlying wave. However, it is possible to consider a stability study with general non-periodic perturbations, e.g., by the class of spatially localized perturbations $L^2(\mathbb{R})$ or by the class of bounded uniformly continuous perturbations $C_b(\mathbb{R})$ (see Mielke (1997), Gardner (1993)-(1997) and Gallay&Hărăguş (2007)).

Here our purpose is to consider the nonlinear stability and linear instability of periodic traveling waveforms. From our experience with nonlinear dispersive equations we know that traveling waves, when they exist, are of fundamental importance in the development of a broad range of disturbance. Then we expect the issue of stability of periodic waves to be of interest and it inspires future developments in this fascinating subject.

It is well known that such theory has started with the pioneering work of Benjamin (1972) regarding the periodic steady solutions called *cnoidal waves*. Its waveform profile was found first by Korteweg&de-Vries (1895) for the currently called Korteweg-de Vries equation (KdV henceforth)

$$u_t + uu_x + u_{xxx} = 0, \tag{1}$$

where $u = u(x,t)$ is a real-valued function of two variables $x, t \in \mathbb{R}$. The cnoidal traveling wave solution, $u(x,t) = \varphi_c(x - ct)$, has a profile determined in the form

$$\varphi_c(\xi) = \beta_2 + (\beta_3 - \beta_2)\text{cn}^2\left(\sqrt{\frac{\beta_3 - \beta_1}{12}}\xi; k\right), \tag{2}$$

where β_i's are real constants and *cn* represents the Jacobi elliptic function *cnoidal*. Among the physical application associated with equation (1) we can mention the propagation of shallow-water waves with weakly non-linear restoring forces, long internal waves in a density-stratified ocean, ion-acoustic waves in a plasma, acoustic waves on a crystal lattice, and so on. Thus, the study of qualitative properties of these nonlinear periodic waves represents a fundamental piece for the understanding of the dynamic associated to this equation. A first stability approach for the cnoidal wave profile as in (2) was determined by Benjamin in (1972). But only years later a complete study was carry out by Angulo et al. (2006). Indeed, by extending the general stability theory due to Grillakis et al. (1987) to the periodic case it was obtained that the orbit generated by the solution φ_c,

$$\Omega_{\varphi_c} = \{\varphi_c(\cdot + y) : y \in \mathbb{R}\}, \tag{3}$$

remains stable by the periodic flow of the KdV equation, more specifically, for initial data close enough to Ω_{φ_c} the solution of the KdV starting in this point will remain close enough to to Ω_{φ_c} for all time. Many ingredients are basic for obtaining this remarkable behavior of the cnoidal waves. One of the cornerstones is the spectral structure associated to the self-adjoint operator on $L^2_{per}([0, L])$ (here L represents the minimal period of φ_c)

$$\mathcal{L}_{kdv} = -\frac{d^2}{dx^2} + c - \varphi_c, \tag{4}$$

which is a Schrödinger operator with a periodic potential.
In this case, the existence of a unique negative eigenvalue and simple and the non-degeneracy of the eigenvalue zero is required. We recall that $\mathcal{L}_{kdv}$ has a compact resolvent and so zero is an isolated eigenvalue. It is well known that the determination of these spectral informations in the periodic case is not an easy task. By taking advantage of the cnoidal profile of φ_c, the eigenvalue problem for $\mathcal{L}_{kdv}$ is reduced to study the classical Lamé problem

$$\frac{d^2}{dx^2}\psi + [\rho - n(n+1)k^2 sn^2(x;k)]\psi = 0, \tag{5}$$

on the space $L^2_{per}([0, 2K(k)])$, for $n \in \mathbb{N}$, $sn(\cdot;k)$ denoting the Jacobi elliptic function *snoidal* and K representing the complete elliptic integral of first kind. Therefore the Floquet theory arises in a crucial form in the stability analysis. The existence of a finite number of instability intervals associated to (5) and an oscillation Sturm analysis will imply the required spectral structure for $\mathcal{L}_{kdv}$. Next, by supposing that φ_c has mean zero property we consider the manifold $M = \{f : \int f^2 dx = \int \varphi_c^2 dx, \int f dx = 0\}$. Then the condition $\frac{d}{dc}\int \varphi_c^2(x)dx > 0$ will imply that

$$\langle \mathcal{L}_{kdv} f, f\rangle \geqq \beta \|f\|^2_{H^1_{per}} \quad \text{for every } f \in T_{\varphi_c}M \cap [\frac{d}{dx}\varphi_c]^{\perp}, \tag{6}$$

where $T_{\varphi_c}M$ represents the tangent space to M in φ_c and $\beta > 0$. Then, from the continuity of the functional $E(f) = \int (f')^2 - \frac{1}{3}f^3 dx$ and from the Taylor theorem we have the following stability property of Ω_{φ_c}: *there is $\eta > 0$ and $D > 0$ such that*

$$E(u) - E(\varphi_c) \geqq D \inf_{g \in \Omega_{\varphi_c}} \|u - g\|^2_{H^1_{per}} \tag{7}$$

for u satisfying that $\inf_{g\in\Omega_{\varphi_c}} \|u - g\|_{H^1_{per}} < \eta$ *and* $F(u) \equiv \frac{1}{2}\int u^2 dx = \frac{1}{2}\int \varphi_c^2 dx, \int u dx = 0.$
In other words, φ_c is a constraint local minimum of E. Then, since E and F are conserved quantities by the continuous KdV-flow, $t \to u(t)$, we obtain from (7) that the orbit Ω_{φ_c} is stable by initial perturbation in the manifold M. For general perturbations of Ω_{φ_c} we need to have the existence of a smooth curve of traveling waves, $c \to \varphi_c$, and to use the triangular inequality. We call attention that mean zero constraint can be eliminated in the definition of

the manifold M, because the KdV equation is invariant under the Galilean transformation $v(x,t) = u(x+\gamma t, t) - \gamma$, where γ is any real number. That is, if u solves (1), then so does v. In this point some comments on the speed-wave associated to the cnoidal wave solution φ_c deserves to be hold. If we are looking for φ_c having mean zero, we obtain a curve $c \in (0,\infty) \to \varphi_c \in H^1_{per}([0,L])$. But, for instance, if φ_c is positive we obtain a curve $c \in (4\pi^2/L^2, \infty) \to \varphi_c \in H^1_{per}([0,L])$ for any $L > 0$ (see Angulo (2009)).

From analysis above we see that spectral information about the operator in (4) is fundamental for a stability study. Indeed, the second order differential operator appearing in equation (4) is the key to apply the Floquet theory, but from our experience with nonlinear dispersive evolution equations we know that depending of the periodic potential the study can be tricky. Moreover, the Floquet theory is not useful for more general linear operators that arise in the study of nonlinear dispersive equations. For instance, a general kind of dispersive equations can be

$$u_t + u^p u_x - (\mathcal{M}u)_x = 0, \tag{8}$$

where $p \in \mathbb{N}$ and $\mathcal{M}$ is a Fourier multiplier operator defined by

$$\widehat{\mathcal{M}f}(n) = \beta(n)\widehat{f}(n), \quad n \in \mathbb{Z}, \tag{9}$$

with β being a measurable, locally bounded, even function on $\mathbb{R}$, and satisfying the conditions, $A_1|n|^{m_1} \le \beta(n) \le A_2(1+|n|)^{m_2}$, for $m_1 \le m_2$, $|n| \ge k_0$, $\beta(n) > b$ for all $n \in \mathbb{Z}$, and $A_i > 0$. Then, the following unbounded linear self-adjoint operator $\mathcal{L}_{\mathcal{M}} : D(\mathcal{L}_{\mathcal{M}}) \to L^2_{per}([0,L])$

$$\mathcal{L}_{\mathcal{M}} = (\mathcal{M} + c) - \varphi_c^p, \tag{10}$$

arises in the study of traveling wave solutions of the form $u(x,t) = \varphi_c(x - ct)$ for equation (8). Here the profile $\varphi = \varphi_c$ must satisfy the following nonlinear equation

$$(\mathcal{M} + c)\varphi - \frac{1}{p+1}\varphi^{p+1} = A_\varphi, \tag{11}$$

where A_φ is a constant of integration which can be assumed to be zero and the wave-speed c is chosen such that $\mathcal{M} + c$ is a positive operator. Equation (8) with $p = 1$, contains two important models in internal water-wave research: The Benjamin-Ono equation (BO henceforth), $\mathcal{M} = \mathcal{H}\partial_x$, where $\mathcal{H}$ denotes the periodic Hilbert transform defined via the Fourier transform as $\widehat{\mathcal{H}f}(n) = -isgn(n)\widehat{f}(n)$, $n \in \mathbb{Z}$. So, we have that $\mathcal{M}$ has associated the symbol $\beta(n) = |n|$. The other model is the Intermediate Long Wave equation (ILW henceforth), where the pseudo-differential operator $\mathcal{M}$ has associated the symbol $\beta_h(n) = n\coth(nh) - \frac{1}{h}$, $h \in (0, +\infty)$.

Recently, Angulo&Natali (2008) established a new approach for studying the general linear operator $\mathcal{L}_{\mathcal{M}}$ in (10) within the framework of the theory of stability for even and positive periodic traveling waves (see Section 3). Indeed, by using Fourier techniques associated to positive linear operators was obtained that the positivity of the Fourier coefficients associated to φ_c together with a specific positivity property called $PF(2)$ for the Fourier coefficients of the power function φ_c^p, will imply the existence of a unique negative eigenvalue and simple and the non-degeneracy of the eigenvalue zero. Therefore, one of the advantage of Angulo&Natali's approach is the possibility of studying non-local linear operators such as that associated to the BO equation (see Section 5)

$$\mathcal{L}_{bo} = \mathcal{H}\partial_x + c - \varphi_c. \tag{12}$$

We also note that in the case of the critical KdV equation ($p = 4$ and $\mathcal{M} = -\partial_x^2$ in (8)) Angulo&Natali's approach was applied successfully for obtaining the relevant result that there is a family of periodic traveling waves, $c \to \varphi_c$, such that they are stable if the wave-speed $c \in (\pi^2/L^2, r_0/L^2)$ and unstable if $c \in (r_0/L^2, +\infty)$, where $r_0 > 0$ does not depend on L (see Angulo&Natali (2009)). In the case of operators of type Schrödinger,

$$\mathcal{L} = -\frac{d^2}{dx^2} + c - \varphi_c^p, \tag{13}$$

Neves (2009) (see Section 6) and Johnson (2009) have obtained other criterium for obtaining the required spectral information in a stability study. Their approach are different from those ones that we shall establish in this chapter. For instance, Johnson uses tools from ordinary differential equations and Evans function methods. Its stability approach works for perturbations restricted to the manifold of initial data u_0 such that $\int u_0^2(x)dx = \int \varphi_c^2(x)dx$ and $\int u_0(x)dx = \int \varphi_c(x)dx$.

Other important piece of information in a stability study is the existence of solutions for the nonlinear equation (11). For $\mathcal{M} = -\partial_x^2$ is obvious that the quadrature method is the most natural tool to be used (see subsection 5.2). Therefore, the theory of elliptic integrals and Jacobian elliptic functions arise in a very natural way. For $\mathcal{M}$ being a non-local operator the existence problem is not an easy task. In this point the use of Fourier methods can be very useful. Indeed, suppose that φ_c represents a solitary wave solution for equation (11) ($A_\varphi = 0$) with $\widehat{\varphi_c}^{\mathbb{R}}$ representing its Fourier transform on the line, then the Poisson Summation Theorem produces a periodic function ψ given by formula

$$\psi(\xi) = \sum_{n\in\mathbb{Z}} \varphi_c(\xi + Ln) = \frac{1}{L} \sum_{n\in\mathbb{Z}} \widehat{\varphi_c}^{\mathbb{R}} \left(\frac{n}{L}\right) e^{\frac{2\pi i n\xi}{L}}. \tag{14}$$

Note that ψ has a minimal period L. Now, from our experience with dispersive evolution equations we know that the profile ψ does not give for every c a solution for equation (11). Indeed, we have only that for a specific range of the solitary wave-speed, c, it will produce that ψ is in fact a periodic traveling wave solution. In other words, there are an interval I and a smooth wave-speed mapping, $c \in I \to v(c)$, such that ψ satisfies $(\mathcal{M} + v(c))\psi - \frac{1}{p+1}\psi^{p+1} = 0$. An example where equality (14) can be used is in obtaining the well-know Benjamin's periodic traveling wave solution for the BO equation (see subsection 5.1). We note from formula (14) that a good knowledge of the Fourier transform $\widehat{\varphi_c}^{\mathbb{R}}$ is necessary for obtaining an explicit profile of ψ and that the Fourier coefficients of ψ are depending of the discretization of $\widehat{\varphi_c}^{\mathbb{R}}$ to the enumerable set $\{n/L\}_{n\in\mathbb{Z}}$.

We note that in our approach we consider the minimal period associated to the periodic traveling wave solutions completely arbitrary. Our analysis is not restricted to small or large wavelength. We also note that the stability theory to be established here it can be applied to a sufficiently wide range of non-linear dispersive models, such as the nonlinear Schrödinger equation

$$iu_t + u_{xx} + |u|^p u = 0 \tag{15}$$

with $u = u(x,t) \in \mathbb{C}$ and $p = 2, 3, 4, ...$, and for the generalized Benjamin-Bona-Mahony equations

$$u_t + u_x + u^p u_x + \mathcal{M} u_t = 0, \tag{16}$$

for $p \geqq 1$, $p \in \mathbb{N}$, and $\mathcal{M}$ given by (9) (see Angulo et al. (2010)).

We will also be interested in this chapter in the *linear instability* of periodic traveling wave solutions. By using the theoretical framework of Weinstein (1986) and Grillakis (1988) we show that there is a family of periodic traveling wave for the cubic Schrödinger equation ($p = 2$ in (15)) with a minimal period L which are orbitally stable in $H^1_{per}([0,L])$ but linearly unstable in $H^1_{per}([0,jL])$, for $j \geqq 2$ (see subsection 5.3.1). In the general case of equations in (8) we establish a criterium of linear instability developed recently by Angulo&Natali (2010) (see Section 7).

In the last section of this chapter, we establish some results about the existence and stability of periodic-peakon for the following nonlinear Schrödinger equation (NLS-δ equation),

$$iu_t + u_{xx} + \gamma\delta(x)u + |u|^2u = 0, \tag{17}$$

defined for functions on the torus $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$. Here the symbol δ denotes the Dirac delta distribution, $(\delta,\psi) = \psi(0)$, and $\gamma \in \mathbb{R}$ is denominated the coupling constant or strength attached to the point source located at $x = 0$.

2. Notation

For any complex number $z \in \mathbb{C}$, we denote by $\Re(z)$ and $\Im(z)$ the real part and imaginary part of z, respectively. For $s \in \mathbb{R}$, the Sobolev space $H^s_{\text{per}}([0,L])$ consists of all periodic distributions f such that $\|f\|^2_{H^s} = L\sum\limits_{k=-\infty}^{\infty}(1+n^2)^s|\widehat{f}(n)|^2 < \infty$. For simplicity, we will use the notation H^s_{per} and $H^0_{\text{per}} = L^2_{\text{per}}$. The Fourier transform of a periodic distribution Ψ is the function $\widehat{\Psi}: \mathbb{Z} \to \mathbb{C}$ defined by the formula $\widehat{\Psi}(n) = \frac{1}{L}\langle \Psi, \Theta_{-n}\rangle$, $n \in \mathbb{Z}$, for $\Theta_n(x) = \exp(2\pi inx/L)$. So, if Ψ is a periodic function with period L, we have $\widehat{\Psi}(n) = \frac{1}{L}\int_0^L \Psi(x)e^{-\frac{2n\pi xi}{L}}dx$. The normal elliptic integral of first type (see Byrd&Friedman (1971)) is defined by

$$\int_0^y \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} = \int_0^\phi \frac{d\theta}{\sqrt{1-k^2\sin^2\theta}} = F(\phi,k)$$

where $y = \sin\phi$ and $k \in (0,1)$. k is called the modulus and ϕ the argument. When $y = 1$, we denote $F(\pi/2,k)$ by $K = K(k)$. The Jacobian elliptic functions are denoted by $sn(u;k)$, $cn(u;k)$ and $dn(u;k)$ (called, snoidal, cnoidal and dnoidal, respectively), and are defined via the previous elliptic integral. More precisely, let $u(y;k) := u = F(\phi,k)$, then $y = sin\phi := sn(u;k)$, $cn(u;k) = \sqrt{1-sn^2(u;k)}$ and $dn(u;k) = \sqrt{1-k^2sn^2(u;k)}$. We have the following asymptotic formulas: $sn(x;1) = tanh(x)$, $cn(x;1) = sech(x)$ and $dn(x;1) = sech(x)$.

3. Positivity properties of the Fourier transform in the nonlinear stability theory

The approach contained in Angulo& Natali (2008) introduces a new criterium for obtaining that the self-adjoint operator $\mathcal{L}_{\mathcal{M}}$ in (10) possesses exactly one negative eigenvalue which is simple and the eigenvalue zero is simple with eigenfunction $\frac{d}{dx}\varphi$. These specific spectral properties are obtained provided that φ is an even positive periodic function with *a priori* minimal period, and such that $\widehat{\varphi}(n) > 0$ for every $n \in \mathbb{Z}$ and $(\widehat{\varphi^p}(n))_{n\in\mathbb{Z}} \in PF(2)$-discrete class which we shall define below.

We start our approach by defining for all $\theta \geq 0$, the convolution operator $S_\theta : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ by

$$S_\theta \alpha(n) = \frac{1}{\omega_\theta(n)} \sum_{j=-\infty}^{\infty} \mathcal{K}(n-j)\alpha_j = \frac{1}{\omega_\theta(n)} (\mathcal{K} * \alpha)_n,$$

where $\omega_\theta(n) = \beta(n) + \theta + c$, $\mathcal{K}(n) = \widehat{\varphi_c^p}(n)$, $n \in \mathbb{Z}$. Here we have chosen c such that $c > -b$ where $b \in \mathbb{R}$ satisfies $\beta(n) > b$ for all $n \in \mathbb{Z}$. Then we have $\omega_\theta(n) > 0$ for all $n \in \mathbb{Z}$. It follows that the space X defined by

$$X = \{\alpha \in \ell^2(\mathbb{Z}); ||\alpha||_{X,\theta} := (\sum_{n=-\infty}^{\infty} |\alpha_n|^2 \omega_\theta(n))^{\frac{1}{2}} < \infty\},$$

is a Hilbert space with norm $||\alpha||_{X,\theta}$ and inner product $< \alpha^1, \alpha^2 >_{X,\theta} = \sum_{n=-\infty}^{\infty} \alpha_n^1 \overline{\alpha_n^2} \omega_\theta(n)$.

The next Proposition is a consequence of the theory of self-adjoint operators with a compact resolvent.

Proposition 3.1. *For every $\theta \geq 0$, we have the following*

(a) If $\alpha \in \ell^2$ is an eigensequence of S_θ for a non-zero eigenvalue, then $\alpha \in X$.

(b) The restriction of S_θ to X is a compact, self-adjoint operator with respect to the norm $|| \cdot ||_{X,\theta}$.

(c) 1 is an eigenvalue of S_θ (as an operator of X) if and only if $-\theta$ is an eigenvalue of $\mathcal{L}_\mathcal{M}$ (as an operator of L^2_{per}). Furthermore, both eigenvalues have the same multiplicity.

(d) S_θ has a family of eigensequences $(\psi_{i,\theta})_{i=0}^\infty$ forming an orthonormal basis of X with respect to the norm $|| \cdot ||_{X,\theta}$. The eigensequences correspond to real eigenvalues $(\lambda_i(\theta))_{i=0}^\infty$ whose only possible accumulation point is zero. Moreover, $|\lambda_0(\theta)| \geq |\lambda_1(\theta)| \geq |\lambda_2(\theta)| \geq \cdots$.

Proof. See Angulo & Natali (2008). □

Definition 3.1. *We say that a sequence $\alpha = (\alpha_n)_{n\in\mathbb{Z}} \subseteq \mathbb{R}$ is in the class $PF(2)$ discrete if*

i) $\alpha_n > 0$, for all $n \in \mathbb{Z}$,

ii) $\alpha_{n_1-m_1}\alpha_{n_2-m_2} - \alpha_{n_1-m_2}\alpha_{n_2-m_1} \geqq 0$, for $n_1 < n_2$ and $m_1 < m_2$,

iii) $\alpha_{n_1-m_1}\alpha_{n_2-m_2} - \alpha_{n_1-m_2}\alpha_{n_2-m_1} > 0$, if $n_1 < n_2$, $m_1 < m_2$, $n_2 > m_1$, and $n_1 < m_2$.

Example: The sequence $a_n = e^{-\eta|n|}$, $n \in \mathbb{Z}$, $\eta > 0$, belongs to $PF(2)$ discrete class. Indeed, the conditions ii) and iii) in Definition 3.1 are equivalents to

$$\begin{array}{l} 1)\ |n_1 - m_1| + |n_2 - m_2| \leqq |n_1 - m_2| + |n_2 - m_1|, \text{ if } n_1 < n_2 \text{ and } m_1 < m_2, \text{and} \\ 2)\ |n_1 - m_1| + |n_2 - m_2| < |n_1 - m_2| + |n_2 - m_1|, \text{ if } n_1 < n_2,\ m_1 < m_2, \\ \quad n_2 > m_1 \text{ and } n_1 < m_2, \end{array} \tag{18}$$

which are immediately verified. In section 4 we will use this example in the stability theory of periodic traveling wave solutions for the BO equation.

The next result will also be useful in section 4.

Theorem 3.1. *Let α^1 and α^2 be two even sequences in the class $PF(2)$ discrete, then the convolution $\alpha^1 * \alpha^2 \in PF(2)$ discrete (if the convolution makes sense).*

Proof. See Karlin (1968). □

We present the main result of this section.

Theorem 3.2. *Let φ_c be an even positive solution of (11) with $A_\varphi = 0$. Suppose that $\widehat{\varphi_c}(n) > 0$ for every $n \in \mathbb{Z}$, and $(\widehat{\varphi_c^p}(n))_{n\in\mathbb{Z}} \in PF(2)$ discrete. Then $\mathcal{L}_{\mathcal{M}}$ in (10) possesses exactly a unique negative eigenvalue which is simple, and zero is a simple eigenvalue with eigenfunction $\frac{d}{dx}\varphi_c$.*

Proof. The complete proof of this theorem is very technical and long (see Angulo&Natali (2008) for details), so we only give a sketch of it divided in three basic steps as follows.

I- Since S_θ is a compact-self-adjoint operator on X, it follows that

$$\lambda_0(\theta) = \pm \sup_{||\alpha||_X=1} | < S_\theta\alpha, \alpha >_X |. \tag{19}$$

Let $\psi(\theta) := \psi$ be an eigensequence of S_θ corresponding to $\lambda_0(\theta) := \lambda_0$. We will show that ψ is one-signed, that is, either $\psi(n) \leq 0$ or $\psi(n) \geq 0$. By contradiction, suppose ψ takes both negative and positive values. By hypotheses the kernel $\mathcal{K} = (\mathcal{K}(n)) = (\widehat{\varphi_c^p}(n))$ *is positive*, then

$$\begin{aligned} S_\theta|\psi|(n) &= \tfrac{1}{\omega_\theta(n)} \textstyle\sum_{j=-\infty}^{\infty} \mathcal{K}(n-j)\psi^+(j) + \tfrac{1}{\omega_\theta(n)} \textstyle\sum_{j=-\infty}^{\infty} \mathcal{K}(n-j)\psi^-(j) \\ &> \left| \tfrac{1}{\omega_\theta(n)} \textstyle\sum_{j=-\infty}^{\infty} \mathcal{K}(n-j)\psi^+(j) - \tfrac{1}{\omega_\theta(n)} \textstyle\sum_{j=-\infty}^{\infty} \mathcal{K}(n-j)\psi^-(j) \right|, \end{aligned}$$

where ψ^+ e ψ^- are the positive and negative parts of ψ respectively. It follows that

$$< S_\theta(|\psi|), |\psi| >_{X,\theta} > \sum_{n=-\infty}^{\infty} |\lambda_0||\psi(n)|^2\omega_\theta(n) = |\lambda_0|||\psi||^2_{X,\theta}.$$

Hence, if we assume that $||\psi||_X = 1$, we obtain $< S_\theta(|\psi|), |\psi| >_X > |\lambda_0|$, which contradicts (19). Then, there is an eigensequence ψ_0 which is nonnegative. Now, since $\mathcal{K}$ *is a positive sequence* and $S_\theta(\psi_0) = \lambda_0\psi_0$, we have $\psi_0(n) > 0$, $\forall n \in \mathbb{Z}$. Therefore, ψ_0 can not be orthogonal to any non-trivial one-signed eigensequence in X, which implies that λ_0 is a simple eigenvalue. Notice that the preceding argument also shows that $-\lambda_0$ can not be an eigenvalue of S_θ, therefore it follows that $|\lambda_1| < \lambda_0$.

II- The next step will be to study the behavior of the eigenvalue $\lambda_1(\theta)$. In fact, it considers the following set of indices,

$$\triangle = \{(n_1, n_2) \in \mathbb{Z} \times \mathbb{Z};\ n_1 < n_2\}.$$

Denoting $\overline{n} = (n_1, n_2)$ and $\overline{m} = (m_1, m_2)$, we define for $\overline{n}, \overline{m} \in \triangle$ the following sequence

$$\mathcal{K}_2(\overline{n}, \overline{m}) := \mathcal{K}(n_1 - m_1)\mathcal{K}(n_2 - m_2) - \mathcal{K}(n_1 - m_2)\mathcal{K}(n_2 - m_1).$$

By hypothesis $\mathcal{K} \in PF(2)$ discrete, hence $\mathcal{K}_2 > 0$. Let $\ell^2(\triangle)$ be defined as

$$\ell^2(\triangle) = \left\{ \alpha = (\alpha_{\overline{n}})_{\overline{n}\in\triangle}; \textstyle\sum\sum_\triangle |\alpha_{\overline{n}}|^2 := \displaystyle\sum_{n_1\in\mathbb{Z}} \sum_{\substack{n_1<n_2 \\ n_2\in\mathbb{Z}}} |\alpha(n_1, n_2)|^2 < +\infty \right\},$$

and define the operator $S_{2,\theta} : \ell^2(\triangle) \to \ell^2(\triangle)$ by

$$S_{2,\theta} g(\overline{n}) = \sum\sum_{\triangle} G_{2,\theta}(\overline{n}, \overline{m}) g(\overline{m}),$$

where $G_{2,\theta}(\overline{n}, \overline{m}) = \frac{\mathcal{K}_2(\overline{n},\overline{m})}{\omega_\theta(n_1)\omega_\theta(n_2)}$. It also consider, the space

$$W = \left\{ \alpha \in \ell^2(\triangle); ||\alpha||_{W,\theta} := \left(\sum\sum_{\triangle} |\alpha(\overline{n})|^2 \omega_\theta(n_1)\omega_\theta(n_2) \right)^{\frac{1}{2}} < \infty \right\}.$$

Then W is a Hilbert space with norm $|| \cdot ||_{W,\theta}$ given above and with inner product

$$< \alpha^1, \alpha^2 >_{W,\theta} = \sum\sum_{\triangle} \alpha^1(\overline{n}) \overline{\alpha^2(\overline{n})} \omega_\theta(n_1)\omega_\theta(n_2).$$

Remark 3.1. 1) *We can show, analogous to Proposition 3.1, that* $S_{2,\theta}|_W$ *is a self-adjoint, compact operator. Therefore, the eigenvalues associated to it operator can be enumerated in order of decreasing absolute value, that is,* $|\mu_0(\theta)| \geq |\mu_1(\theta)| \geq |\mu_2(\theta)| \geq \cdots$.
2) *We also obtain that* $\mu_0(\theta) := \mu_0$ *is positive, simple and* $|\mu_1| < \mu_0$.

Definition 3.2. *Let* α^1, $\alpha^2 \in \ell^2(\mathbb{Z})$, *we define the wedge product* $\alpha^1 \wedge \alpha^2$ *in* $\triangle$ *by* $(\alpha^1 \wedge \alpha^2)(n_1, n_2) = \alpha^1(n_1)\alpha^2(n_2) - \alpha^1(n_2)\alpha^2(n_1)$.

We have the following results from Definition 3.2.

Lemma 3.1. *1) Let* $A = \{\alpha^1 \wedge \alpha^2; \text{for } \alpha^1, \alpha^2 \in X, \alpha^1 \wedge \alpha^2 \in \ell^2(\triangle)\}$. *Then* A *is dense in* W.
2) Let $\alpha^1, \alpha^2 \in \ell^2(\mathbb{Z})$. *Then* $S_{2,\theta}(\alpha^1 \wedge \alpha^2) = S_\theta \alpha^1 \wedge S_\theta \alpha^2$.

Proof. See Karlin (1964), Karlin (1968) and Albert (1992) . □

The following Lemma is the key to characterize the second eigenvalue λ_1.

Lemma 3.2. *For all* $\theta \geqq 0$ *we have:*
a) $\mu_0(\theta) = \lambda_0(\theta)\lambda_1(\theta)$, *and then* $\lambda_1(\theta) > 0$.
b) $\lambda_1(\theta)$ *is simple.*

Proof. See Angulo&Natali (2008). □

III- Final step. For $i = 0, 1$, we have that the differentiable curve $\theta \to \lambda_i(\theta)$ satisfies $\frac{d}{d\theta}\lambda_i(\theta) < 0$ and $\lim_{\theta\to\infty} \lambda_0(\theta) = 0$. From $\widehat{\varphi_c}(n) > 0$ for all $n \in \mathbb{Z}$, it follows $\lambda_1(0) = 1$. Since $\lambda_0(0) > \lambda_1(0) = 1$, there is a unique $\theta_0 \in (0, +\infty)$ such that $\lambda_0(\theta_0) = 1$. From Proposition 3.1, we obtain that $\kappa \equiv -\theta_0$ is a negative eigenvalue of $\mathcal{L}_{\mathcal{M}}$ which is simple. For $i \geq 2$ and $\theta > 0$ we have that $\lambda_i(\theta) \leq \lambda_1(\theta) < \lambda_1(0) = 1$, so 1 can not be eigenvalue of S_θ for all $\theta \in (0, +\infty) \setminus \{\theta_0\}$, since 1 is an eigenvalue only for $\theta = 0$ and $\theta = \theta_0$. Then $\mathcal{L}_{\mathcal{M}}$ has a unique negative eigenvalue which is simple. Finally, since $\lambda_1(0) = 1$ and λ_1 is a simple eigenvalue it follows that $\theta = 0$ is a simple eigenvalue of $\mathcal{L}_{\mathcal{M}}$ by Proposition 3.1. This shows the theorem. □

Remark 3.2. *In Theorem 3.2 the Fourier transform of* φ_c *and* φ_c^p *must be calculated in the minimal period* L *of* φ_c.

3.1 Construction of periodic functions in $PF(2)$ discrete class

In this subsection we show a method for building non-trivial periodic functions such that its Fourier transform belongs to the $PF(2)$ discrete class. We start with the $PF(2)$ continuous class.

Definition 3.3. *We say that a function $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ is in the $PF(2)$ continuous class if*

i) $\mathcal{K}(x) > 0$ *for* $x \in \mathbb{R}$,

ii) $\mathcal{K}(x_1 - y_1)\mathcal{K}(x_2 - y_2) - \mathcal{K}(x_1 - y_2)K(x_2 - y_1) \geqq 0$, *for* $x_1 < x_2$ *and* $y_1 < y_2$; *and*

iii) strict inequality holds in ii) whenever the intervals (x_1, x_2) *and* (y_1, y_2) *intersect.*

The following result is immediate.

Proposition 3.2. *Suposse $\mathcal{K}$ is in the $PF(2)$ continuous class. Then for $\alpha(n) \equiv \mathcal{K}(n)$ we have that the sequence $(\alpha(n))_{n\in\mathbb{Z}}$ is in the $PF(2)$ discrete class.*

Next, we have the following Theorem (see Albert&Bona (1991)).

Theorem 3.3. *Suppose f is a positive, twice-differentiable function on $\mathbb{R}$ satisfying*

$$\frac{d^2}{dx^2}(log\ f(x)) < 0 \qquad \textit{for}\ \ x \neq 0, \qquad (\textit{logarithmically concave}) \tag{20}$$

then $f \in PF(2)$.

Now we illustrate Theorem 3.3. Indeed, let us consider the solitary wave solution associated to the KdV and modified KdV equation ($p = 2$ and $\mathcal{M} = -\partial_x^2$ in (8)),

$$\phi_{c,p}(\xi) = \Big[\frac{(p+1)(p+2)c}{2}\Big]^{1/p} \operatorname{sech}^{2/p}\Big(\frac{p\sqrt{c}}{2}\xi\Big), \qquad c > 0, \ \ p = 1, 2. \tag{21}$$

Then the Fourier transforms are given by

$$\widehat{\phi_{c,1}}(\xi) = 12\pi\ \frac{\xi}{\sinh(\pi\xi/\sqrt{c})}, \quad \widehat{\phi_{c,2}}(\xi) = \sqrt{\frac{3}{2}}\pi \operatorname{sech}\Big(\frac{\pi\xi}{2\sqrt{c}}\Big). \tag{22}$$

Hence, since $\widehat{\phi_{c,i}}$, $i = 1, 2$, are logarithmically concave functions it follows from Theorem 3.3 that they belong to $PF(2)$. Moreover, from Proposition 3.2 we have that the sequences $(\widehat{\phi_{c,i}}(n))_{n\in\mathbb{Z}}$, $i = 1, 2$, belong to $PF(2)$ discrete class.

Next, for one better convenience of the reader, we establish the Poisson Summation Theorem.

Theorem 3.4. *Let $\widehat{f}^{\mathbb{R}}(\xi) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i x\xi}dx$ and $f(x) = \int_{-\infty}^{\infty} \widehat{f}^{\mathbb{R}}(\xi)e^{2\pi i x\xi}d\xi$ satisfy*

$$|f(x)| \leq \frac{A}{(1+|x|)^{1+\delta}} \quad \textit{and} \quad |\widehat{f}^{\mathbb{R}}(\xi)| \leq \frac{A}{(1+|\xi|)^{1+\delta}},$$

where $A > 0$ and $\delta > 0$ (then f and $\widehat{f}$ can be assumed continuous functions). Thus, for $L > 0$

$$\sum_{n=-\infty}^{\infty} f(x + Ln) = \frac{1}{L}\sum_{n=-\infty}^{\infty} \widehat{f}^{\mathbb{R}}\left(\frac{n}{L}\right) e^{\frac{2\pi i n x}{L}}.$$

The two series above converge absolutely.

Proof. See for example Stein&Weiss (1971). □

From Theorem 3.4 and formulas in (22) we have that the periodization of the solitary wave solutions in (21), $p = 1, 2$, produces the following periodic functions

$$\psi_i(\xi) = \frac{1}{L}\sum_{n=-\infty}^{\infty} \widehat{\phi_{c,i}}\left(\frac{n}{L}\right) e^{\frac{2\pi i n\xi}{L}}, \qquad i = 1, 2, \tag{23}$$

such that its Fourier transform belongs to the $PF(2)$ discrete class.

4. Orbital stability definition and main theorem

In this section we establish the definition of stability which we are interested in this chapter and a general stability theorem for periodic traveling waves.

Definition 4.1. *Let φ be a periodic traveling wave solution of (11) with minimal period L and consider $\tau_r\varphi(x) = \varphi(x + r)$, $x \in \mathbb{R}$ and $r \in \mathbb{R}$. We define the set $\Omega_\varphi \subset H_{per}^{\frac{m_2}{2}}$, the orbit generated by φ, as $\Omega_\varphi = \{g;\ g = \tau_r\varphi,$ for some $r \in \mathbb{R}\}$. For any $\eta > 0$, let us define the set $U_\eta \subset H_{per}^{\frac{m_2}{2}}$ by $U_\eta = \{f;\ \inf_{g\in\Omega_\varphi} ||f - g||_{H_{per}^{\frac{m_2}{2}}} < \eta\}$. With this terminology, we say that φ is (orbitally) stable in $H_{per}^{\frac{m_2}{2}}$ by the flow generated by equation (8) if,*

(i) there is s_0 such that $H_{per}^{s_0} \subseteq H_{per}^{\frac{m_2}{2}}$ and the initial value problem associated to (8) is globally well-posed in $H_{per}^{s_0}$.

(ii) For every $\varepsilon > 0$, there is $\delta > 0$ such that for all $u_0 \in U_\delta \cap H_{per}^{s_0}$, the solution u of (8) with $u(0, x) = u_0(x)$ satisfies $u(t) \in U_\varepsilon$ for all $t \in \mathbb{R}$.

Remark 4.1. *We have some comments about Definition 4.1:*

1. *Definition 4.1 is based on the translation symmetry associated to model (8).*
2. *In Definition 4.1 we are introducing other space, $H_{per}^{s_0}$, because to obtain a global well-posed theory in the energy space $H_{per}^{\frac{m_2}{2}}$ can not be an easy task. For instance, in the case of the regularized Benjamin-Ono equation (equation (16) with $p = 1$ and $\mathcal{M} = \mathcal{H}\partial_x$) it is possible to have a global well-posed theory in the space $H_{per}^{s_0}$ with $s_0 > \frac{1}{2}$, but global well-posed in $H_{per}^{\frac{1}{2}}$ remains an open problem (see Angulo et al. (2010)).*
3. *Definition 4.1 was given for equations in (8), but naturally it is also valid for those ones in (16). Stability definition for the Schrödinger models (15) and (17) is different to that given in definition 4.1, since we have two symmetries (translations and rotations) and one symmetry (rotations) for that models, respectively (see Theorem 5.5 and Theorem 8.1 below).*

The proof of the following general stability theorem can be shown by using techniques due to Benjamin (1972), Bona (1975), Weinstein (1986) or Grillakis *et al.* (1987) (see also Angulo (2009))

Theorem 4.1. *Let φ_c be a periodic traveling wave solution of (11) and suppose that part* (i) *of the Definition 4.1 holds. Suppose also that the operator $\mathcal{L}_\mathcal{M}$ in (10) possesses exactly a unique negative eigenvalue which is simple, and zero is a simple eigenvalue with eigenfunction $\frac{d}{dx}\varphi_c$. Choose $\chi \in L_{per}^2$ such that $\mathcal{L}_\mathcal{M}\chi = \varphi_c$, and define $I = (\chi, \varphi_c)_{L_{per}^2}$. If $I < 0$, then φ_c is stable in $H_{per}^{\frac{m_2}{2}}$.*

Remark 4.2. *In our cases the function χ in Theorem 4.1 will be chosen as $\chi = -\frac{d}{dc}\varphi_c$. Then, $I < 0$ if and only if $\frac{d}{dc}\int \varphi_c^2(\xi)d\xi > 0$.*

5. Stability of periodic traveling wave solutions for some dispersive models

In this section we are interested in applying the theory obtained in Section 3 to obtain the stability of specific periodic traveling waves associated to the following models: the BO equation, the modified KdV and the cubic Schrödinger equation.

5.1 Stability for the BO equation

We start by finding a smooth curve $c \to \varphi_c$ of solutions associated with the following non-local differential equation

$$\mathcal{H}\varphi_c' + c\varphi_c - \frac{1}{2}\varphi_c^2 = 0. \tag{24}$$

Here we present an approach based on the Poisson Summation Theorem for obtaining an explicit solution to equation (24). Indeed, it we consider, the solitary wave solution associated to BO equation, namely, $\phi_\omega(x) = \dfrac{4\omega}{1+\omega^2x^2}$, with $\omega > 0$. Since its Fourier transform is given by $\widehat{\phi_\omega}^{\mathbb{R}}(x) = 4\pi e^{\frac{-2\pi}{\omega}|x|}$, we obtain from Theorem 3.4 the following periodic wave of minimal period L

$$\psi_\omega(x) = \frac{4\pi}{L}\sum_{n=0}^{+\infty}\varepsilon_n e^{\frac{-2\pi n}{\omega L}}\cos\left(\frac{2n\pi x}{L}\right) = \frac{4\pi}{L}\frac{\sinh\left(\frac{2\pi}{\omega L}\right)}{\cosh\left(\frac{2\pi}{\omega L}\right) - \cos\left(\frac{2\pi x}{L}\right)}, \tag{25}$$

where $\varepsilon_n = 1$ for $n = 0$, and $\varepsilon_n = 2$ for $n \geqq 1$. Next we see that the profile ψ_ω represents a periodic solution for (24) with $\omega = \omega(c)$ and $c > 2\pi/L$. Let φ_c, $c > 0$, be a smooth periodic solution of (24) with minimal period L, then φ_c can be expressed as a Fourier series

$$\varphi_c(x) = \sum_{n=-\infty}^{+\infty} a_n e^{\frac{2n\pi i x}{L}}. \tag{26}$$

Now, from (24), we get

$$\left[\frac{2\pi|n|}{L} + c\right]a_n = \frac{1}{2}\sum_{m=-\infty}^{+\infty} a_{n-m}a_m.$$

We consider $a_n \equiv 4\pi e^{-\gamma|n|/L}$, $n \in \mathbb{Z}$, $\gamma \in \mathbb{R}$. Substituting a_n in the last identity we obtain

$$\sum_{m=-\infty}^{+\infty} a_{n-m}a_m = \frac{16\pi^2}{L^2}e^{-\gamma|n|}\left[|n| + 1 + 2\sum_{k=1}^{+\infty} e^{-2\gamma k}\right] = \frac{16\pi^2}{L^2}e^{-\gamma|n|}(|n| + \coth\gamma).$$

Then,

$$c + \frac{2\pi|n|}{L} = \frac{4\pi}{L}\cdot\frac{1}{2}(|n| + \coth\gamma). \tag{27}$$

Consider $\gamma = 2\pi/(\omega L)$. Then for $c > 2\pi/L$ we choose the solitary wave-speed $\omega = \omega(c) > 0$ such that

$$\tanh(\gamma) = \frac{2\pi}{cL}. \tag{28}$$

Therefore, from uniqueness of the Fourier series we obtain $\varphi_c = \psi_{\omega(c)}$. Hence, since the mapping $c \to \gamma(c) = \tanh^{-1}(2\pi/(cL))$ is a differentiable function for $c > 2\pi/L$, it follows that $c \in \left(\frac{2\pi}{L}, +\infty\right) \mapsto \varphi_c \in H^n_{per}([0,L])$, $n \in \mathbb{N}$, is a smooth curve of periodic traveling wave solutions for the BO equation. From our analysis we have then the following Fourier expansion for φ_c

$$\varphi_c(x) = \frac{4\pi}{L} \sum_{n=-\infty}^{+\infty} e^{-\gamma|n|} e^{\frac{2\pi i n x}{L}}, \tag{29}$$

with γ satisfying (28). Then, we obtain immediately that $(\widehat{\varphi_c}(n))_{n\in\mathbb{Z}} \in PF(2)$ discrete class (see (18)) and from Theorem 3.2, that the operator in (12) possesses exactly a unique negative eigenvalue which is simple and whose kernel is generated by $\frac{d}{dx}\varphi_c$. Next we calculate the sign of the quantity $I = -\frac{1}{2}\frac{d}{dc}||\varphi_c||^2_{L^2_{per}}$. Indeed, from (29) and Parseval Theorem it follows

$$I = -\frac{L}{2}\frac{d}{dc}||\widehat{\varphi_c}||^2_{\ell^2} = -\frac{8\pi^2}{L}\frac{d}{dc}\sum_{n=-\infty}^{\infty} e^{-2\gamma|n|} = -\frac{32\pi^3}{c^2L^2}\frac{1}{1-\left(\frac{2\pi}{cL}\right)^2}\sum_{n=-\infty}^{\infty}|n|e^{-2\gamma|n|} < 0. \tag{30}$$

Hence, from Theorem 4.1 we obtain the orbital stability of the periodic solutions (29) in $H^{\frac{1}{2}}_{per}$ by the periodic flow of the BO equation.

Remark 5.1. *The periodic global well-posed theory for the BO in $H^{\frac{1}{2}}_{per}$ has been shown by Molinet (2008) and Molinet&Ribaut (2009).*

5.2 Stability for the mKdV equation

Next we study the modified KdV equation written as

$$u_t + 3u^2u_x + u_{xxx} = 0. \tag{31}$$

In this case, the periodic traveling wave solution $u(x,t) = \varphi_c(x-ct)$ satisfies the equation

$$\varphi_c'' - c\varphi_c + \varphi_c^3 = 0. \tag{32}$$

On this time we are going to use the quadrature method to determine a profile for φ_c (see Angulo et al. (2010) for the use of the Poisson Summation Theorem). Thus, multiplying equation (32) by φ_c' and integrating once we deduce the following differential equation in quadrature form

$$[\varphi_c']^2 = \frac{1}{2}\left[-\varphi_c^4 + 2c\varphi_c^2 + 4B_{\varphi_c}\right], \tag{33}$$

where B_{φ_c} is a nonzero constant of integration. The periodic solutions arise of the specific form of the roots associated with the polynomial $F(t) = -t^4 + 2ct^2 + 4B_{\varphi_c}$. We start by considering F with four real roots such that $-\eta_1 < -\eta_2 < 0 < \eta_2 < \eta_1$, then we obtain

$$[\varphi_c']^2 = \frac{1}{2}(\eta_1^2 - \varphi_c^2)(\varphi_c^2 - \eta_2^2). \tag{34}$$

By looking for positive solutions we have $\eta_2 \leqq \varphi_c \leqq \eta_1$ and from (34), $2c = \eta_1^2 + \eta_2^2$ and $4B_{\varphi_c} = -\eta_1^2\eta_2^2$. Next, for $\phi_c \equiv \varphi_c/\eta_1$ and $\phi_c^2 = 1 - k^2\sin^2\psi$ we obtain from (34) the following elliptic integral equation $F(\psi(\xi), k) = \eta_1\xi/\sqrt{2}$, with $k^2 = (\eta_1^2 - \eta_2^2)/\eta_1^2$. Therefore, from the definition of the Jacobi elliptic function snoidal, sn, it follows that for $l = \eta_1/\sqrt{2}$, $\sin(\psi(\xi)) = \mathrm{sn}(l\xi;k)$, and hence $\phi_c(\xi) = \sqrt{1-k^2\mathrm{sn}^2(l\xi;k)} = \mathrm{dn}(l\xi;k)$. Then if we return back to the initial variable φ_c, we obtain the so-called **dnoidal wave** solutions:

$$\varphi_c(\xi) \equiv \varphi_c(\xi;\eta_1,\eta_2) = \eta_1\,\mathrm{dn}\Big(\frac{\eta_1}{\sqrt{2}}\,\xi;k\Big) \tag{35}$$

with

$$k^2 = \frac{\eta_1^2-\eta_2^2}{\eta_1^2}, \qquad \eta_1^2+\eta_2^2 = 2c, \qquad 0<\eta_2<\eta_1. \tag{36}$$

Next we study the fundamental period associated to φ_c. Indeed, since $\mathrm{dn}(u+2K) = \mathrm{dn}\ u$, it follows that φ_c has the fundamental period (wavelength) T_{φ_c}, given by

$$T_{\varphi_c} \equiv \frac{2\sqrt{2}}{\eta_1}\,K(k). \tag{37}$$

Now, by using (36) we have for $c>0$ that $0<\eta_2<\sqrt{c}<\eta_1<\sqrt{2c}$. Hence one can consider (37) as a function of η_2, namely

$$T_{\varphi_c}(\eta_2) = \frac{2\sqrt{2}}{\sqrt{2c-\eta_2^2}}\,K(k(\eta_2)) \qquad \text{with} \qquad k^2(\eta_2) = \frac{2c-2\eta_2^2}{2c-\eta_2^2}. \tag{38}$$

Then, since for $\eta_2 \to 0$ we have $K(k(\eta_2)) \to +\infty$, it follows that $T_{\varphi_c}(\eta_2) \to +\infty$ as $\eta_2 \to 0$. Now, for $\eta_2 \to \sqrt{c}$ we obtain $K(k(\eta_2)) \to \pi/2$. Therefore, $T_{\varphi_c}(\eta_2) \to \pi\sqrt{2}/\sqrt{c}$ as $\eta_2 \to \sqrt{c}$. Finally, since $\eta_2 \to T_{\varphi_c}(\eta_2)$ is a strictly decreasing function we obtain $T_{\varphi_c} > \frac{\pi\sqrt{2}}{\sqrt{c}}$. Then the implicit function theorem implies the following result (see Angulo (2007)).

Theorem 5.1. *Let $L>0$ be arbitrary but fixed. Then there exists a smooth mapping curve $c \in J_0 = \left(\frac{2\pi^2}{L^2}, +\infty\right) \to \varphi_c \in H^n_{per}([0,L])$, such that φ_c satisfies equation (32) and it has the dnoidal profile*

$$\varphi_c(\xi) = \eta_1 dn\Big(\frac{\eta_1}{\sqrt{2}}\xi;k\Big), \ \ \xi \in [0,L]. \tag{39}$$

Here, $c \in J_0 \to \eta_1(c) \in (\sqrt{c},\sqrt{2c})$, $c \in J_0 \to k(c) \in (0,1)$ are smooth.

Our next step is the study of the following periodic eigenvalue problem,

$$\begin{cases} \mathcal{L}_{mkdv}\psi \equiv \left(-\dfrac{d^2}{dx^2}+c-3\varphi_c^2\right)\psi = \lambda\psi \\ \psi(0) = \psi(L),\ \psi'(0) = \psi'(L). \end{cases} \tag{40}$$

Then, we have the following theorem.

Theorem 5.2. *Let φ_c be the dnoidal wave solution given by Theorem 5.1. Then problem (40) defined on $H^2_{per}([0,L])$ has exactly its three first eigenvalues simple, being the eigenvalue zero, the second one with eigenfunction φ_c'. Moreover, the remainder of the spectrum is constituted by a discrete set of eigenvalues which are double.*

Remark 5.2. *The periodic global well-posed theory for the mKdV in H^1_{per} can be found in Colliander et al. (2003).*

The proof of Theorem 5.2 is based on the Floquet theory. So, we have the following classical theorem (see Magnus&Winkler (1976))

Theorem 5.3. *Consider the Lamé's equation*

$$-\chi'' + m(m+1)k^2 sn^2(x;k)\chi = \rho\chi, \tag{41}$$

where m is a real parameter. Then we guarantee the existence of two linearly independent periodic solutions to (41) *with period 2K or 4K if and only if m is an integer. By letting $l = m$ if m is a non-negative integer and $l = -m-1$ if m is a negative integer then Lamé's equation* (41) *has, at most, $l+1$ intervals of instability (including the interval $(-\infty, \rho_0)$ with ρ_0 being the first eigenvalue). In addition, if m is a non-negative integer then* (41) *has exactly $m+1$ intervals of instability.*

Proof. (Theorem 5.2) Since operator $\mathcal{L}_{mkdv}$ has a compact resolvent its spectrum is a countable infinity set of eigenvalues $\{\lambda_n; n = 0, 1, 2, ...\}$, with

$$\lambda_0 < \lambda_1 \leqq \lambda_2 < \lambda_3 \leqq \lambda_4 < \cdots. \tag{42}$$

We denote by ψ_n the eigenfunction associated to the eigenvalue λ_n. The eigenvalue distribution in (42) is a consequence of the following Oscillation Sturm-Liouville result:

i) ψ_0 has no zeros in $[0, L]$,

ii) ψ_{2n+1} and ψ_{2n+2} have exactly $2n+2$ zeros in $[0, L)$.

Next, since $\mathcal{L}_{mkdv}\varphi_c' = 0$ and φ_c' has two zeros in $[0, L)$ we have that the eigenvalue zero is either λ_1 or λ_2. For determining that $0 = \lambda_1 < \lambda_2$ we will use Theorem 5.3. Indeed, the transformation $Q(x) = \psi(x\sqrt{2}/\eta_1)$ implies from (40) the following Lamé problem for Q,

$$\begin{cases} Q'' + [\rho - 6k^2 sn^2(x,k)]Q = 0 \\ Q(0) = Q(2K(k)), \;\; Q'(0) = Q'(2K(k)), \end{cases} \tag{43}$$

with

$$\rho = 2(\lambda + 3\eta_1^2 - c)/\eta_1^2. \tag{44}$$

Therefore, since problem (43) has exactly 3 intervals of instability we have that the eigenvalues $\{\rho_n; n = 0, 1, 2, ...\}$ will satisfy that ρ_0, ρ_1 and ρ_2 are simples and $\rho_3 = \rho_4$, $\rho_5 = \rho_6, \cdots$. Next, we establish the values of the eigenvalues ρ_i, $i = 1, 2, 3$. Indeed, $\rho_0 = 2[1 + k^2 - \sqrt{1 - k^2 + k^4}]$, $\rho_1 = 4 + k^2$, $\rho_2 = 2[1 + k^2 + \sqrt{1 - k^2 + k^4}]$. Therefore relation (44) implies that for $i = 1, 2, 3$, ρ_i determine the eigenvalues λ_i, respectively. Hence, zero is the second eigenvalue for (40) and it is simple. This shows the theorem. □

Remark 5.3. *We note that a part of the conclusion of Theorem 5.2 can be also obtained via Theorem 3.2. Indeed, the Fourier transform of the dnoidal profile φ_c is given by $\widehat{\varphi_c}(n) = \frac{\sqrt{2}\pi}{L} sech\left(\frac{\pi n}{\sqrt{\omega(c)}L}\right)$, where for $c > \frac{2\pi^2}{L^2}$ we have $\omega(c) = c/(16(2-k^2)K^2(\sqrt{1-k^2}))$. Then since the function $f(x) = \mu sech(\nu x)$ belongs to $PF(2)$ continuous for μ, ν positive (see (22)), it follows $(\widehat{\varphi_c}(n))_{n\in\mathbb{Z}} \in PF(2)$ discrete. Finally, since the convolution of even sequences in $PF(2)$ discrete is in $PF(2)$ discrete (see Theorem 3.1) we obtain that $\widehat{\varphi_c^2} \in PF(2)$ discrete.*

Finally, we calculate the sign of the quantity $D = \frac{1}{2}\frac{d}{dc}||\varphi_c||^2_{L^2_{per}}$. From integral elliptic theory we have

$$||\varphi_c||^2_{L^2_{per}} = \sqrt{2}\eta_1 \int_0^{\frac{\eta_1}{\sqrt{2}}L} dn^2(x;k)dx = \frac{8K(k)}{L}\int_0^K dn^2(x;k)dx = \frac{8}{L}K(k)E(k). \tag{45}$$

Then, since the maps $k \in (0,1) \to K(k)E(k)$ and $c \to k(c)$ are strictly increasing functions, it follows immediately from (45) that $D > 0$. Hence, Theorem 4.1 implies stability in $H^1_{per}([0,L])$ of the dnoidal solutions (39) by the periodic flow of the mkdV equation.

Remark 5.4. *Recently Johnson (2009) has proposed an approach to determine the nonlinear stability of periodic traveling wave for models of KdV type. Next, we would like to show that this theory can not be applied to the smooth curve of dnoidal wave, $c \to \varphi_c$, established by Theorem 5.1. Indeed, from analysis in subsection 5.2 we have for $L > 0$ fixed that $B := B_{\varphi_c}(c) = -\frac{-16(1-k(c)^2)K^2(k(c))}{L^2}$. Now, we note that $\frac{dB}{dc} > 0$ and so φ_c can be seen as a function of the parameter B. In the proof of Lemma 4.2 in Johnson (2009) is deduced that $\frac{d}{dB}\varphi_c$ is a periodic function if and only if the kernel of the operator $\mathcal{L}_{mkdv}$ is double. So, from the equality $\frac{d}{dB}\varphi_c = \frac{dc}{dB}\frac{d\varphi_c}{dc}$, we deduce that $\frac{d}{dB}\varphi_c$ is periodic since $\frac{dc}{dB} > 0$ and $\frac{d\varphi_c}{dc}$ is a periodic function by construction. Therefore from Johnson's Lemma we obtain that $ker(\mathcal{L}_{mkdv})$ is double which is a contradiction.*

5.3 Stability and instability for the cubic Schrödinger equation

In this subsection we are interested in studying stability properties of two specific families de traveling wave solutions for the cubic Schrödinger equation (NLS henceforth)

$$iu_t + u_{xx} + |u|^2 u = 0, \tag{46}$$

namely, the dnoidal and cnoidal solutions.

5.3.1 The dnoidal case

The stability analysis associated to the mKdV equation (31) gives us the basic information to study the stability of periodic standing-wave solutions for the NLS equation (46) in the form $u(x,t) = e^{ict}\varphi_c(x)$. Indeed, Theorem 5.1 implies the existence of a smooth curve $c \to e^{ict}\varphi_c$ of periodic traveling wave solutions for the NLS with a profile given by the dnoidal function. Now, since the NLS has two basic symmetries, rotations and translations, we have that the orbit to be studied here will be $\mathcal{O}_{\varphi_c} = \{e^{i\theta}\varphi_c(\cdot + y) : y \in \mathbb{R}, \theta \in [0, 2\pi)\}$. Therefore, from Weinstein (1986) and Grillakis et. al (1987) we need to study the spectrum of the following linear operators: $\mathcal{L}_{mkdv}$ in (40), which henceforth we denote by $\mathcal{L}^-$, and the operator $\mathcal{L}^+$ defined by

$$\mathcal{L}^+ = -\frac{d^2}{dx^2} + c - \varphi_c^2. \tag{47}$$

The following theorem is related to the specific structure of $\mathcal{L}^+$.

Theorem 5.4. *The self-adjoint operator $\mathcal{L}^+$ defined on $H^2_{per}([0,L])$ is a nonnegative operator. The eigenvalue zero is simple with eigenfunction associated φ_c and the remainder of the eigenvalues are double.*

Proof. Since $\varphi_c > 0$ and $\mathcal{L}^+\varphi_c = 0$, it follows from the theory of self-adjoint operators that zero is simple and it is the first eigenvalue. Now, by Theorem 5.3 we obtain that $\mathcal{L}^+$ has exactly two instability intervals and so the remainder of the spectrum of $\mathcal{L}^+$ is constituted by eigenvalues which are double. This finishes the Theorem. □

Now, from Angulo (2007) we have the following stability theorem for the NLS equation.

Theorem 5.5. *Let φ_c be the dnoidal wave solution given by Theorem 5.1. Then the orbit $\mathcal{O}_{\varphi_c}$ is stable in $H^1_{per}([0,L])$ by the periodic flow of the NLS equation. Indeed, for every $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that if the initial data u_0 satisfies*

$$\inf_{(y,\theta)\in\mathbb{R}\times[0,2\pi)} \|u_0 - e^{i\theta}\varphi_c(\cdot + y)\|_1 < \delta,$$

then the solution $u(t)$ of the NLS equation with $u(0) = u_0$ satisfies

$$\inf_{(y,\theta)\in\mathbb{R}\times[0,2\pi)} \|u(t) - e^{i\theta}\varphi_c(\cdot + y)\|_1 < \epsilon,$$

for all $t \in \mathbb{R}$, $\theta = \theta(t)$ and $y = y(t)$.

Remark 5.5. *The periodic global well-posed theory for the NLS equation in H^1_{per} has been shown by Bourgain (1999).*

Theorem 5.5 establishes that the dnoidal solutions are stable by periodic perturbations of the same minimal period L in $H^1_{per}([0,L])$. Now, since φ_c is also a periodic traveling wave solution for the NLS equation in every interval $[0,jL]$, for $j \in \mathbb{N}$ and $j \geqq 2$, it is natural to ask about its stability in $H^1_{per}([0,jL])$. As we see below they will be nonlinearly unstable (in fact, they are linearly unstable). We start our analysis with the following elementary result.

Lemma 5.1. *Define P_j and Q_j as the number of negatives eigenvalues of $\mathcal{L}^-$ and $\mathcal{L}^+$, respectively, with periodic boundary condition in $[0,jL]$ and $j \geqq 2$. Then $Q_j = 0$ and $P_j = 2j$ or $P_j = 2j-1$.*

Proof. Since $\varphi_c > 0$ and $\mathcal{L}^+\varphi_c(x) = 0$, $x \in [0,jL]$, by the Oscillation Sturm-Liouville result for Hill equations, we obtain that zero must be the first eigenvalue and therefore for all $j \geqq 2$, $Q_j = 0$. Next, since $\mathcal{L}^-\varphi_c'(x) = 0$, $x \in [0,jL]$, and the number of zeros (nodes) of φ_c' in the semi-open interval $[0,jL)$ is $2j$, the Oscillation Theorem implies that the eigenvalues corresponding to the zero eigenvalue are λ_{2j} or λ_{2j-1}. Therefore, we have $P_j = 2j$ or $P_j = 2j-1$. This finishes the Lemma. □

A theoretical framework for proving nonlinear instability from a linear instability result for nonlinear Schrödinger type's equation was developed in Grillakis (1988), Jones (1988) and Grillakis et al. (1990). In those approach one deduces instability when the number of negative eigenvalues of $\mathcal{L}^-$ exceeds the number of negative eigenvalues of $\mathcal{L}^+$ by more than one (see Theorem 5.8 below). The parts of the instability theorems in Grillakis (1988) that are needed for obtaining a linear instability of φ_c, connect P_j, Q_j and the existence of real eigenvalues of the operator (the linearized Hamiltonian)

$$N = \begin{pmatrix} 0 & \mathcal{L}^+ \\ -\mathcal{L}^- & 0 \end{pmatrix}. \tag{48}$$

First, define: 1)K_j - the orthogonal projection on $(\ker \mathcal{L}^-)^\perp$; 2) R_j - the operator $R_j = K_j\mathcal{L}^-K_j$; 3) S_j - the number of negative eigenvalues of R_j; 4) $I_{real}(N_j)$ - the number of pairs of nonzero real eigenvalues of N considered on $[0,jL]$.

Theorem 5.6. *[Grillakis (1988), Jones (1988)] For $j \geqq 1$ we have*

1) If $|S_j - Q_j| = m_j > 0$, then $I_{real}(N_j) \geqq m_j$.

2) If $S_j = Q_j$ and $\{f \in L^2_{per}([0, jL]) : (R_i f, f) < 0$ and $((\mathcal{L}^+)^{-1} f, f) < 0\} = \emptyset$, then $I_{real}(N_j) \geqq 1$.

The following result gives a condition for obtaining the number S_j.

Theorem 5.7. *[Grillakis (1988)] If $\frac{d}{dc}\int_0^{jL} \varphi_c^2(x)dx > 0$ then $S_j = P_j - 1$.*

The following theorem is the main result of this section.

Theorem 5.8. *[Instability] Consider the dnoidal solution φ_c given by Theorem 5.1. Then the orbit $\mathcal{R}_{\varphi_c} = \{e^{iy}\varphi_c : y \in \mathbb{R}\}$ is $H^1_{per}([0, jL])$-unstable for all $j \geqq 2$, by the flow of the periodic NLS equation.*

Proof. The strategy of the proof is initially to show that the orbit $\mathcal{R}_{\varphi_c}$ is *linearly unstable*. For this, we rewrite equation (15) in the Hamiltonian form

$$\frac{d}{dt}\mathbf{u}(t) = JG'(\mathbf{u}(t)), \tag{49}$$

where $\mathbf{u} = (\Re(u), \Im(u))^t$, J is the skew-symmetric, one-one and onto matrix given by

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{50}$$

and

$$G(\mathbf{u}) = \int \left[\frac{1}{2}|\mathbf{u}'|^2 - \frac{1}{4}|\mathbf{u}|^4\right]dx, \tag{51}$$

which is a conservation law to (46). Next, for the linearization of (49) around the orbit $\mathcal{R}_{\varphi_c}$ we proceed as follows: we write $\Psi_c = (0, \varphi_c)^t$ and define

$$\mathbf{v}(t) = T_p(-ct)\mathbf{u}(t) - \Psi_c. \tag{52}$$

Here $T_p(s)$ acts as a rotation matrix. Hence, if $T_p'(0)$ denotes the infinitesimal generator of $T_p(s)$, from the properties: $T_p(s)T_p'(0) = T_p'(0)T_p(s)$, $T_p(-s)JT_p(s) = J$, $G'(T_p(s)\mathbf{u}) = T_p(s)H'(\mathbf{u})$, $J^{-1}T_p'(0)\mathbf{u} = -F'(\mathbf{u})$ ($F(\mathbf{u}) = \frac{1}{2}\int |\mathbf{u}|^2 dx$) we obtain, via Taylor Theorem,

$$\begin{aligned} \frac{d\mathbf{v}}{dt} &= J[G'(\mathbf{v} + \Psi_c) + cF'(\mathbf{v} + \Psi_c)] \\ &= J[G''(\Psi_c)\mathbf{v} + cF''(\Psi_c)\mathbf{v} + G'(\Psi_c) + cF'(\Psi_c) + O(\|\mathbf{v}\|^2)] = N\mathbf{v} + O(\|\mathbf{v}\|^2), \end{aligned} \tag{53}$$

where in the last equality we are taking into account that Ψ_c is a critical point of $G + cF$ and J is a bounded linear operator. Here, N is the linear operator defined in (48). Therefore, we are interested in the problem of determining a growing mode solution $\mathbf{v}(t) = e^{\lambda t}\Phi(x)$ with $\Re(\lambda) > 0$ of the *linearized problem*

$$\frac{d\mathbf{v}}{dt} = N\mathbf{v}. \tag{54}$$

We note that the eigenvalues of N appear in conjugate pairs. Now, since $\frac{d}{dc}\int_0^L \varphi_c^2(x)dx > 0$ (see (45)) it follows from Lemma 5.1, Theorem 5.6 and Theorem 5.7 that $m_j = 2j - 1$, or, $2j - 2$. Therefore for $j \geqq 2$, the number $I_{real}(N_j) \geqq m_j > 0$. Then the zero solution of (54) is unstable, which implies that the orbit $\mathcal{R}_{\varphi_c}$ is nonlinearly unstable (see Grillakis (1988) and Grillakis et al. (1990)). □

Remark 5.6. *1. The Instability criterium in Grillakis et al. (1990) can not be used for studying the instability of the orbit $\mathcal{R}_{\varphi_c}$. In fact, we denote by $n(H_c)$ the number of negative eigenvalues of the diagonal linear operator $H_c \equiv J^{-1}N$. Since the kernel of H_c is generated by φ_c and $\frac{d}{dx}\varphi_c$, and $\frac{d}{dc}\int_0^{jL}\varphi_c^2(x)dx > 0$, we have that if $n(H_c) - 1$ is odd then the orbit $\mathcal{R}_{\varphi_c}$ is nonlinearly unstable in $H^1_{per}([0, jL])$. Now, from Lemma 5.1 we have for $j \geqq 2$ that $n(H_c) - 1 = P_j - 1 = 2j - 1, or, 2j - 2$. But as we will see latter $P_j = 2j - 1$ (see Section 6), therefore we have that $n(H_c)$ is always even.*

2. *The analysis in this subsection can be applied to study the following forced periodic nonlinear Schrödinger equation*

$$iu_t - u_{xx} - \gamma|u|^2u = \varepsilon exp(-i\Omega^2 t + i\alpha) - i\delta u,$$

where ε is the small forcing amplitude, δ is the small damping coefficient, Ω^2 is the forcing frequency and α is an arbitrary phase (see Shlizerman&Rom-Kedar (2010)) .

5.3.2 The cnoidal case

Equation (32) has other family of periodic solutions determined by the Jacobi elliptic function *cnoidal*. Indeed, now suppose that equation (33) now is written in the quadrature form

$$[\varphi_c']^2 = \frac{1}{2}(a^2 + \varphi_c^2)(b^2 - \varphi_c^2).$$

For $b > 0$ we have that $-b \leqq \varphi_c \leqq b$ and so $2c = b^2 - a^2$ and $4B_{\varphi_c} = a^2b^2 > 0$. Therefore, it is possible to obtain (see Angulo (2007)) that the profile

$$\varphi_c(\xi) = bcn(\beta\xi; k) \tag{55}$$

is a solution of (32) with $k^2 = b^2/(a^2 + b^2)$ and $\beta = \sqrt{(a^2 + b^2)/2}$. Then by using the implicit function theorem we have the following result (see Angulo (2007)).

Theorem 5.9. *Let $L > 0$ arbitrary but fixed. Then we have two branches of cnoidal wave solutions for the NLS equation. Indeed,*

1) *there are a strictly increasing smooth function $c \in (0, +\infty) \to b(c) \in (\sqrt{2c}, +\infty)$ and a smooth curve $c \in (0, +\infty) \to \varphi_{c,1} \in H^1_{per}([0, L])$ of solutions for equation (32) with*

$$\varphi_{c,1}(\xi) = b\ cn\Big(\sqrt{b^2 - c}\ \xi; k\Big). \tag{56}$$

Here the modulus $k = k(c)$ satisfies $k^2 = b^2/(2b^2 - 2c)$ and $k'(c) > 0$,

2) *there are a strictly decreasing smooth function $c \in (-\frac{4\pi^2}{L^2}, 0) \to a(c) \in (\sqrt{-2c}, +\infty)$ and a smooth curve $c \in \Big(-\frac{4\pi^2}{L^2}, 0\Big) \to \varphi_{c,2} \in H^1_{per}([0, L])$ of solutions for equation (32) with*

$$\varphi_{c,2}(\xi) = \sqrt{a^2 + 2c}\ cn\Big(\sqrt{a^2 + c}\ \xi; k\Big). \tag{57}$$

Here the modulus $k = k(c)$ satisfies $k^2 = (a^2 + 2c)/(2a^2 + 2c)$ and $k'(c) > 0$.

Now, for the cnoidal case it makes necessary to study the behavior of the first eigenvalues related to the following self-adjoint operators

$$\mathcal{L}_i^- = -\frac{d^2}{dx^2} + c - 3\varphi_{c,i}^2, \quad \mathcal{L}_i^+ = -\frac{d^2}{dx^2} + c - \varphi_{c,i}^2, \quad i = 1, 2. \tag{58}$$

The next theorem gives the specific structure for $\mathcal{L}_i^{\pm}$.

Theorem 5.10. *Let $L > 0$ and $\varphi_{c,i}$, $i = 1, 2$, the cnoidal wave solution given by Theorem 5.9. Then,*

1) *The operators $\mathcal{L}_i^+$ defined on $H^2_{per}([0, L])$ have exactly one negative eigenvalue which is simple, the eigenvalue zero is also simple with eigenfunction $\varphi_{c,i}$. Moreover, the remainder of the spectrum is a discrete set of eigenvalues converging to infinity.*
2) *The operators $\mathcal{L}_i^-$ defined on $H^2_{per}([0, L])$ have exactly two negative eigenvalue which are simple. The eigenvalue zero is the third one, which is simple with eigenfunction $\varphi'_{c,i}$. Moreover, the remainder of the eigenvalues are double and converging to infinity.*

Proof. The proof is a consequence of Theorem 5.3 and the Oscillation Sturm-Liouville Theorem (see Angulo (2007)). □

We note that the stability or instability of the cnoidal solutions $\varphi_{c,i}$ can not be determined by using the same techniques mentioned above for the case of dnoidal solution (see Angulo (2007) for discussion). Indeed, since $\frac{d}{dc}\|\varphi_{c,i}\|^2 > 0$ and for

$$H_{c,i} = \begin{pmatrix} \mathcal{L}_i^- & 0 \\ 0 & \mathcal{L}_i^+ \end{pmatrix} \tag{59}$$

we have $n(H_{c,i}) - 1 = 2$, it follows that the Grillakis et. al (1990) stability approach is not applicable in this case. A similar situation occurs with Grillakis (1988) and Jones (1988) instability theories.

Remark 5.7. *Recently, Natali&Pastor (2008) have determined that the cnoidal wave solution in (55) is orbitally unstable by the periodic flow of the Klein-Gordon equation*

$$u_{tt} - u_{xx} + u - |u|^2 u = 0, \tag{60}$$

by using the abstract theory due to Grillakis et al. (1990). In fact, if one considers a standing wave solution for (60) of the form $u(x, t) = e^{ict}\varphi_c(x)$, $|c| < 1$, we conclude from Theorem 5.10 that the operator

$$\mathcal{L}_{kg} = \begin{pmatrix} \mathcal{L}_{kg}^- & 0 \\ 0 & \mathcal{L}_{kg}^+ \end{pmatrix} \tag{61}$$

has three negative eigenvalues which are simple and the eigenvalue zero is double. Here, $\mathcal{L}_{kg}^- = \begin{pmatrix} -\frac{d^2}{dx^2} + 1 - 3\varphi_c^2 & -c \\ -c & 1 \end{pmatrix}$ and $\mathcal{L}_{kg}^+ = \begin{pmatrix} -\frac{d^2}{dx^2} + 1 - \varphi_c^2 & c \\ c & 1 \end{pmatrix}$. So, since $D = -\frac{d}{dc}\left(c\int_0^L \varphi_c(x)^2 dx\right)$ is negative it follows that $n(\mathcal{L}_{kg}) - 0 = 3$ is an odd number. Then, the approach in Grillakis et al. (1990) can be applied in order to conclude the instability result.

In Section 7 we establish a new criterium for the instability of periodic traveling wave solutions for general nonlinear dispersive equations. An application of this technique shows that the cnoidal wave profile associated to the mKdV is actually unstable.

6. Hill's operators and the stability of periodic waves.

As we have seen in previous sections the study of the spectrum associated to the Hill operator

$$\mathcal{L}_Q = -\frac{d^2}{dx^2} + Q(x), \tag{62}$$

with Q being a periodic potential, is of relevance in the stability's study of periodic traveling wave solutions for nonlinear evolution equations. Recently, Neves (2009), have presented a new technique to establish a characterization of the nonpositive eigenvalues of $\mathcal{L}_Q$ by knowing one of its eigenfunctions. Next, we will give the main points of his theory and we apply it to a specific situation. Indeed, let us consider the Hill equation related to the operator in (62),

$$y''(x) + Q(x)y(x) = 0, \tag{63}$$

where we assume that the potential Q is a $\pi-$periodic function. Denote by y_1 and y_2 two normalized solutions of (63), that is, solutions uniquely determined by the initial conditions $y_1(0) = 1$, $y_1'(0) = 0$, $y_2(0) = 0$, $y_2'(0) = 1$. The characteristic equation associated with (63) is given by

$$\rho^2 - [y_1(\pi) + y_2'(\pi)]\rho + 1 = 0, \tag{64}$$

and the characteristic exponent is a number α which satisfies the equation $e^{i\alpha\pi} = \rho_1$, $e^{-i\alpha\pi} = \rho_2$, where ρ_1 and ρ_2 are the roots of the characteristic equation (64). It is well known from Floquet's Theorem (see Magnus&Winkler (1976)) that if $\rho_1 = \rho_2 = 1$ equation (63) has a nontrivial $\pi-$periodic smooth solution. So, if one considers p such periodic solution and y be another solution which is linearly independent of p, then $y(x+\pi) = \rho_1 y(x) + \theta p(x)$, for θ constant. The case $\theta = 0$ is equivalent to say that y is also a $\pi-$periodic solution. Next, suppose that $z_1 < z_2 < \cdots < z_{2n}$ are the simple zeros of p in the interval $[0, \pi)$. Then from Taylor's formula, p can be written as

$$p(x) = (x - z_i)p'(z_i) + O((x - z_i)^3), \tag{65}$$

and therefore, for x near z_i we deduce, $\frac{x-z_i}{p(x)} = \frac{1}{p'(z_i)+O((x-z_i)^2)}$. Next, we choose in each interval (z_{i-1}, z_i) one point x_i such that $p'(x_i) = 0$. Thus, the zeros z_i of p and x_i of p' intercalated as follows $z_1 < x_1 < z_2 < x_2 < \cdots < z_{2n} < x_{2n}$ and, of course, they shall repeat to the right and to the left by the periodicity of the functions.

Define, for $[x_1, x_1 + \pi)$

$$q(x) = \frac{x - z_i}{p(x)} = \frac{1}{p'(z_i) + O((x - z_i)^3)}, \quad x \in [x_{i-1}, x_1), \tag{66}$$

where $i = 2, \cdots, 2n+1$, $z_{2n+1} = z_1 + \pi$ and $x_{2n+1} = x_1 + \pi$. Next, it is possible to extend q to whole line by periodicity. Moreover we guarantee that function q is a piecewise smooth with jump discontinuities in the points x_i, q is continuous to the right and $\pi-$periodic with $q'(z_i) = 0$, $q'(x_i) = \frac{1}{p(x_i)}$, that is, q' is continuous on whole real line.

Then, we can state the following result which is a new version of the Floquet Theorem for the case $\rho_1 = \rho_2 = 1$.

Theorem 6.1. *If p is a $\pi-$periodic solution of (63), q is the function defined in (66) and*

$$j(x_i) = \frac{q(x_i^+) - q(x_i^-)}{p(x_i)} = -\frac{z_{i+1} - z_i}{p^2(x_i)}.$$

Then, the solution y linearly independent with p such that the Wronskian $W(p, y) = 1$ satisfies

$$y(x + \pi) = y(x) + \theta p(x), \tag{67}$$

where θ is given by

$$\theta = \sum_{x_i \in (0,\pi]} j(x_i) + 2\int_0^\pi \frac{q'(x)}{p(x)} dx. \tag{68}$$

In particular, y is $\pi-$periodic if and only if $\theta = 0$.

Proof. See Neves (2009). □

Now, we turn back to the linear operator in (62). We have from Oscillation Theorem (see Magnus&Winkler (1976)) that the spectrum of $\mathcal{L}_Q$ under periodic conditions is formed by an unbounded sequence of real numbers, $\lambda_0 < \lambda_1 \leq \lambda_2 < \lambda_3 \leq \lambda_4 \cdots < \lambda_{2n-1} \leq \lambda_{2n} \cdots$, where $\lambda_n' s$ are the roots of the characteristic equation

$$\Delta(\lambda) = y_1(\pi, \lambda) + y_2'(\pi, \lambda) = 2, \tag{69}$$

and $y_1(\cdot, \lambda)$ and $y_2(\cdot, \lambda)$ are the solutions of the differential equation $-y'' + (Q(x) - \lambda)y = 0$ determined by the initial conditions $y_1(0, \lambda) = 1$, $y_1'(0, \lambda) = 0$, $y_2(0, \lambda) = 0$ and $y_2'(0, \lambda) = 1$. We recall that the mapping $\lambda \to \Delta(\lambda)$ is an analytic function.

Now, we know that the spectrum of $\mathcal{L}_Q$ is also characterized by the number of zeros of the eigenfunctions. So, if p is an eigenfunction associated to the eigenvalues λ_{2n-1} or λ_{2n}, then p has exactly $2n$ zeros in the interval $[0, \pi)$. We can enunciate the converse of the previous result with the following statement.

Theorem 6.2. *If p is the eigenfunction of $\mathcal{L}_Q$ associated with the eigenvalue λ_k, $k \geq 1$, and θ is the constant given by Theorem 6.1, then λ_k is simple if and only if $\theta \neq 0$. In addition, if p has $2n$ zeros in the interval $[0, \pi)$, then $\lambda_k = \lambda_{2n-1}$ if $\theta < 0$, and $\lambda_k = \lambda_{2n}$ if $\theta > 0$.*

Proof. See Neves (2009). □

Remark 6.1. *We note that the main idea in the proof of Theorem 6.2 is to determine the sign of $\Delta'(\lambda_k)$, this fact can be obtained from the equality*

$$\Delta'(\lambda_k) = -\theta\Big[\|y_1\|^2 p^2(0) + 2 < y_1, y_2 > p(0)p'(0) + \|y_2\|^2 (p'(0))^2\Big].$$

Indeed, since $\Delta'(\lambda_k)\theta < 0$ we have for $\theta < 0$ that $\lambda_k = \lambda_{2n-1}$ and for $\theta > 0$ that $\lambda_k = \lambda_{2n}$.

As an application of Theorem 6.2 we obtain the spectral information for the linear operator $\mathcal{L}_{mkdv}$ in (40) with φ_c being the dnoidal profile determined by Theorem 5.1. Initially, we write $\mathcal{L}_{mkdv} = -\frac{d^2}{dx^2} + c - 3\varphi_c^2 = -\frac{d^2}{dx^2} + Q(x, c)$, then for this kind of operators we already know that the nonpositive spectrum is invariant with respect to parameter c (see Neves (2008)), so, it is sufficient to establish the spectral condition contained in Theorem 4.1 for a fixed value of

$c \in I = (2\pi^2/L^2, +\infty)$. Then, if one considers $L = \pi$ and the unique value of $c \in I$ such that $k(c) = \frac{1}{2}$, the value of θ in Theorem 6.1 will be $\theta \approx -0.5905625$. Now, we known that φ_c' is an eigenfunction of $\mathcal{L}_{mkdv}$ with eigenvalue $\lambda = 0$ and such that it has two zeros in the interval $[0, \pi)$. We conclude from Theorem 6.2 that $\mathcal{L}_{mkdv}$ possesses one negative eigenvalue which is simple. Moreover, since $\theta \neq 0$ it follows that $\lambda = 0$ is a simple eigenvalue.

Remark 6.2. *Theorem 6.1 and Theorem 6.2 can also be used to show that the operator $\mathcal{L}_{mkdv}$ with periodic boundary conditions on $[0, jL]$, $j \geqq 2$, has the zero eigenvalue as being simple and it is the $2j$-nth eigenvalue. So, the number of negative eigenvalue of $\mathcal{L}_{mkdv}$ is $P_j = 2j - 1$.*

7. Instability of periodic waves

This section is devoted to establish sufficient conditions for the linear instability of periodic traveling wave solutions, $u(x,t) = \varphi_c(x - ct)$, for the general class of dispersive equation in (8). We shall extend the asymptotic perturbation theories in Vock&Hunziker (1982) and Lin (2008) (see also Hislop&Sigal (1996)) to the periodic case.

We start by denoting $f(u) = u^{p+1}/(p+1)$, then the linearized equation associated to (8) in the traveling frame $(x + ct, t)$ is given by

$$(\partial_t - c\partial_x)u + \partial_x(f'(\varphi_c)u - \mathcal{M}u) = 0. \tag{70}$$

As mentioned in Subsection 5.3, the central point in this type of problems is the existence of a growing mode solution $e^{\lambda t}u(x)$, with $\Re(\lambda) > 0$, for (70). Hence, the function u must satisfy

$$(\lambda - c\partial_x)u + \partial_x(f'(\varphi_c)u - \mathcal{M}u) = 0. \tag{71}$$

Equation (71) gives us the family of operators $\mathcal{A}^\lambda : H^{m_2}_{per}([0,L]) \to L^2_{per}([0,L])$ given by,

$$\mathcal{A}^\lambda u = cu + \frac{c\partial_x}{\lambda - c\partial_x}(f'(\varphi_c)u - \mathcal{M}u). \tag{72}$$

Hence the existence of a growing mode solution is reduced to find $\lambda > 0$ such that $\mathcal{A}^\lambda$ has a nontrivial kernel. For $\mathcal{A}^0 = \mathcal{M} + c - f'(\varphi_c)$ (see (10)), we have the following results:

1) For $\lambda > 0$, $\mathcal{A}^\lambda \to \mathcal{A}^0$ strongly in $L^2_{per}([0,L])$ when $\lambda \to 0^+$.
2) The compact embedding $H^{m_2}_{per}([0,L]) \hookrightarrow L^2_{per}([0,L])$ give us $\sigma_{ess}(\mathcal{A}^\lambda) = \emptyset$ for all $\lambda > 0$.
3) There exists $\Lambda > 0$ such that for all $\lambda > \Lambda$, $\mathcal{A}^\lambda$ has no eigenvalues $z \in \mathbb{C}$ satisfying $\Re(z) \leq 0$.

Definition 7.1. *An eigenvalue $\mu_0 \in \sigma_p(\mathcal{A}^0)$ is stable with respect to the family of perturbations $\mathcal{A}^\lambda$ defined in (72) if the following two conditions hold:*
(i) there is $\delta > 0$ such that the annular region $\mathcal{Q}_\delta := \{z \in \mathbb{C};\ 0 < |z - \mu_0| < \delta\}$ is contained in the $\rho(\mathcal{A}^0)$ and in the region of boundedness for the family $\{\mathcal{A}^\lambda\}$, Δ_b, defined by

$$\Delta_b := \{z \in \mathbb{C};\ ||R_\lambda(z)||_{B(L^2_{per})} \leq M,\ \forall\, 0 < \lambda \ll 1\}.$$

Here $M > 0$ does not depend on λ and $R_\lambda(z) = (\mathcal{A}^\lambda - z)^{-1}$.
(ii) Let Γ be a simple closed curve about $\mu_0 \in \sigma_p(\mathcal{A}^\lambda)$ contained in the resolvent set of $\mathcal{A}^\lambda$ and define the Riesz projector $P_\lambda = \frac{1}{2\pi i}\int_\Gamma R_\lambda(z)dz$. Then

$$\lim_{\lambda \to 0^+} ||P_\lambda - P_{\mu_0}||_{B(L^2_{per})} = 0. \tag{73}$$

Remark 7.1. *It follows from Definition 7.1 that for all* $0 < \lambda \ll 1$, $\mathcal{A}^\lambda$ *has total algebraic multiplicity equal to the* μ_0 *inside* $\mathcal{Q}_\delta$.

The next lemma is the cornerstone of our analysis.

Lemma 7.1. *The following three conditions are equivalent:*

(i) the number $z \in \Delta_b$;

(ii) for all $u \in C^\infty_{per}([0,L])$ *we have* $||(\mathcal{A}^\lambda - z)u||_{L^2_{per}} \geq \varepsilon ||u||_{L^2_{per}} > 0$ *for all* $0 < \lambda \ll 1$;

(iii) the number $z \in \rho(\mathcal{A}^0)$.

Proof. See Angulo&Natali (2010). □

Lemma 7.1 enable us to prove the following result.

Theorem 7.1. *Let* $\mathcal{A}^\lambda$ *be the linear operator defined in (72). Suppose that* μ_0 *is a discrete eigenvalue of* $\mathcal{L}_{\mathcal{M}}$. *Then* μ_0 *is stable in the sense of the Definition 7.1.*

Proof. See Angulo&Natali (2010). □

Then, we can enunciate the following instability criteria (see Lin (2008) for the solitary wave case).

Theorem 7.2. *Let* φ_c *be a periodic traveling wave solution related to equation (11). We assume that* $\ker(\mathcal{A}^0) = [\varphi_c']$. *Denote by* $n^-(\mathcal{A}^0)$ *the number (counting multiplicity) of negative eigenvalues of the operator* $\mathcal{A}^0$. *Then there is a purely growing mode* $e^{\lambda t}u(x)$ *with* $\lambda > 0$, $u \in H^{m_2}_{per}([0,L])$ *to the linearized equation (70), if one of the following two conditions is true:*

(i) $n^-(\mathcal{A}^0)$ *is even and* $\frac{d}{dc}\int_0^L \varphi_c^2(x)dx > 0$.

(ii) $n^-(\mathcal{A}^0)$ *is odd and* $\frac{d}{dc}\int_0^L \varphi_c^2(x)dx < 0$.

Proof. See Angulo&Natali (2010). □

7.1 Nonlinear instability of cnoidal waves for the mKdV equation.

The arguments presented in Subsection 5.3.2 and from Theorem 7.2 enable us to determine that the cnoidal wave solutions $\varphi_{c,i}$ defined by Theorem 5.9 are linearly unstable for the mKdV equation. Now, we sketch the proof that linear instability implies nonlinear instability of cnoidal waves for the mKdV equation. In fact, we have that the linearized equation (70) takes the form $u_t = J\mathcal{L}_i^- u$, $i = 1,2$, where $J = \partial_x$ and $\mathcal{L}_i^-$ are defined in (58). So, $J\mathcal{L}_i^-$ has a positive real eigenvalue. Next, we define $S : H^1_{per}([0,L]) \to H^1_{per}([0,L])$ as $S(u) = u_\phi(1)$ where $u_\phi(t)$ is the solution of the Cauchy problem,

$$\begin{cases} u_t + 3u^2u_x - cu_x + u_{xxx} = 0, \\ u(x,0) = \phi(x). \end{cases} \tag{74}$$

Then, it follows that the cnoidal waves $\varphi_{c,i}$ are stationary solutions for (74). Now, from Colliander et al. (2003) follows that the mapping data-solution related to the mKdV equation (74), $\Upsilon_c : H^1_{per}([0,L]) \to C([0,T]; H^1_{per}([0,L]))$ is smooth. Furthermore $S(\varphi_{c,i}) = \varphi_{c,i}$ for $i = 1,2$. Thus, since S is at least a $C^{1,\alpha}$ map defined on a neighborhood of the fixed point $\varphi_{c,i}$, we have from Henry et al. (1982) that there is an element $\mu \in \sigma(S'(\varphi_{c,i}))$ with $|\mu| > 1$ which implies the nonlinear instability in $H^1_{per}([0,L])$ of the cnoidal wave solutions $\varphi_{c,i}$.

8. Stability of periodic-peakon waves for the NLS-δ

Recently Angulo&Ponce (2010) have established a theory of existence and stability of periodic-peakon solutions for the cubic NLS-δ equation in (17) ($p = 2$). More precisely, it was shown the existence of a smooth branch of periodic solutions, $(\omega, Z) \to \varphi_{\omega,Z} \in H^1_{per}([0,2L])$, for the semi-linear elliptic equation

$$-\varphi''_{\omega,Z} + \omega\varphi_{\omega,Z} - Z\delta(x)\varphi_{\omega,Z} = \varphi^3_{\omega,Z}, \tag{75}$$

such that

$$\begin{aligned} &(1)\ -\varphi''_{\omega,Z}(x) + \omega\varphi_{\omega,Z}(x) = \varphi^3_{\omega,Z}(x) \quad \text{for } x \neq \pm 2nL,\ n \in \mathbb{N}.\\ &(2)\ \varphi'_{\omega,Z}(0+) - \varphi'_{\omega,Z}(0-) = -Z\varphi_{\omega,Z}(0),\\ &(3)\ \lim_{Z\to 0}\varphi_{\omega,Z} = \varphi_\omega, \end{aligned} \tag{76}$$

where φ_ω is the dnoidal profile in (39). We note that if $\varphi_{\omega,Z}$ is a solution of (75) then $\varphi_{\omega,Z}(\cdot + y)$ **is not necessarily a solution of** (75). Therefore the stability study for the *"periodic-peakon"* $\varphi_{\omega,Z}$ is for the orbit,

$$\Omega_{\varphi_{\omega,Z}} = \{e^{i\theta}\varphi_{\omega,Z} : \theta \in [0,2\pi]\}. \tag{77}$$

The profile of $\varphi_{\omega,Z}$ is based in the Jacobi elliptic function *dnoidal* and determined for $\omega > Z^2/4$ by the patterns:

$$\begin{aligned} &(1)\text{ for } Z > 0,\ \varphi_{\omega,Z}(\xi) = \eta_{1,Z}dn\left(\tfrac{\eta_{1,Z}}{\sqrt{2}}|\xi| + a; k\right),\\ &(2)\text{ for } Z < 0,\ \varphi_{\omega,Z}(\xi) = \eta_{1,Z}dn\left(\tfrac{\eta_{1,Z}}{\sqrt{2}}|\xi| - a; k\right), \end{aligned} \tag{78}$$

where $\eta_{1,Z}$ and k depend of ω and Z. The shift-function a satisfies that $\lim_{Z\to 0} a(\omega, Z) = 0$. So, since the basic symmetry for the NLS-δ equation is the phase-invariance we have the following stability definition for $\Omega_{\varphi_{\omega,Z}}$.

Definition 8.1. *For $\eta > 0$ we put $U_\eta = \{v \in H^1_{per}([0,2L]);\ \inf_{\theta\in\mathbb{R}} ||v - e^{i\theta}\varphi_{\omega,Z}||_{H^1_{per}} < \eta\}$. The periodic standing wave $e^{i\omega t}\varphi_{\omega,Z}$ is stable if for $\epsilon > 0$ there exists $\eta > 0$ such that for $u_0 \in U_\eta$, the solution $u(t)$ of the NLS-δ equation with $u(0) = u_0$ satisfies $u(t) \in U_\epsilon$ for all $t \in \mathbb{R}$. Otherwise, $e^{i\omega t}\varphi_{\omega,Z}$ is said to be unstable in $H^1_{per}([0,2L])$.*

The stability result established in Angulo&Ponce (2010) for the family of periodic-peakon in (78) is the following;

Theorem 8.1. *Let $\omega > \frac{\pi^2}{2L^2}$, $\omega > \frac{Z^2}{4}$ and ω large. Then we have:*

1. *For $Z > 0$ the dnoidal-peakon standing wave $e^{i\omega t}\varphi_{\omega,Z}$ is stable in $H^1_{per}([-L,L])$.*
2. *For $Z < 0$ the dnoidal-peakon standing wave $e^{i\omega t}\varphi_{\omega,Z}$ is unstable in $H^1_{per}([-L,L])$.*
3. *For $Z < 0$ the dnoidal-peakon standing wave $e^{i\omega t}\varphi_{\omega,Z}$ is stable in $H^1_{per,even}([-L,L])$.*

9. References

Albert, J.P. (1992) Positivity properties and stability of solitary-wave solutions of model equations for long waves. *Comm PDE*, Vol. 17, 1-22.

Albert, J. P.& Bona, J. L. (1991) Total positivity and the stability of internal waves in stratified fluids of finite depth. *The Brooke Benjamin special issue (University Park, PA, 1989). IMA J. Appl. Math.*, Vol 46, 1-19.

Angulo, J. (2009) Nonlinear Dispersive Equations: Existence and Stability of Solitary and Periodic Travelling Wave Solutions. *Mathematical Surveys and Monographs (SURV)*, Vol. 156, AMS.

Angulo, J. (2007) Nonlinear stability of periodic traveling wave solutions to the Schrödinger and the modified Korteweg-de Vries equations. *J. Diff. Equations*, Vol. 235, 1-30.

Angulo, J. & Banquet, C. & Scialom, M. (2010) The regularized Benjamin-Ono and BBM equations: Well-posedness and nonlinear stability. *To appear in J. Diff. Equation*.

Angulo, J. & Bona, J.L. & Scialom, M. (2006) Stability of cnoidal waves. *Adv. Diff. Equations*, Vol. 11, 1321-1374.

Angulo, J. & Natali, F. (2009) Stability and instability of periodic traveling waves solutions for the critical Korteweg-de Vries and non-linear Schrödinger equations. *Physica D*, Vol. 238, 603-621.

Angulo, J. & Natali, F. (2008) Positivity properties of the Fourier transform and stability of periodic travelling waves solutions. *SIAM J. Math. Anal.*, Vol. 40, 1123-1151.

Angulo, J. & Ponce, G. (2010) The Non-linear Schrödinger equation with a periodic δ–interaction. *pre-print*.

Benjamin, T.B. (1972) The stability of solitary waves. *Proc. R. Soc. Lond. Ser. A*, Vol. 338, 153-183.

Bona, J.L. (1975) On the stability theory of solitary waves. *Proc Roy. Soc. London Ser. A* , Vol. 344, 363-374.

Bourgain, J. (1999) Global Solutions of Nonlinear Schrödinger Equations. *Amer. Math. Soc. Coll. Publ.*, Vol. 46, AMS, Providence, RI.

Byrd, P. F. & Friedman, M. D. (1971) *Handbook of Elliptic Integrals for Engineers and Scientists*, 2^{nd} ed., Springer-Verlag: New York and Heidelberg.

Colliander, J. & Keel, M. & Staffilani, G. & Takaoka, H. & Tao, T. (2003) Sharp global well-posedness for KdV and modified KdV on $\mathbb{R}$ and $\mathbb{T}$. *J. Amer. Math. Soc.*, Vol. 16, 705-749.

Gallay, T. & Hărăgus, M. (2007) Stability of small periodic waves for the nonlinear Schrödinger equation. *J. Differential Equations*, Vol. 234, 544-581.

Gardner, R. A. (1997) Spectral analysis of long wavelength periodic waves and applications. *J. Für Die Reine und Angewandte Mathematik*, Vol. 491, 149-181.

Gardner, R. A. (1993) On the structure of the spectra of periodic travelling waves. *J. Math. Pures Appl.*, Vol. 72, 415-439.

Grillakis, M. (1988) Linearized instability for nonlinear Schrödinger and Klein-Gordon equations. *J. Comm. Pure Appl. Math.*, Vol. XLI, 747-774.

Grillakis, M. & Shatah, J. & Strauss, W. (1990) Stability theory of solitary waves in the presence of symmetry II. *J. Functional Anal.*, Vol. 94, 308-348.

Grillakis, M.& Shatah, J. & Strauss, W. (1987) Stability theory of solitary waves in the presence of symmetry I. *J. Functional Anal.*, Vol. 74, 160-197.

Henry, D. & Perez, J. F. & Wreszinski, W. (1982), Stability theory for solitary-wave solutions of scalar field equations. *Comm. Math. Phys.*, Vol. 85, 351-361.

Hislop, P. D. & Sigal, I. M. (1996) *Introduction to spectral theory. With appplications to Schrödinger operators*, Springer-Verlag, New York.

Johnson, M. (2009) Nonlinear stability of periodic traveling wave solutions of the generalized Korteweg-de Vries equation. *SIAM J. Math. Anal.* Vol. 41, 1921-1947.

Jones, C. K. R. T. (1988) An instability mechanism for radially symmetric standing waves of a nonlinear Schrödinger equation. *Journal of Differential Equations*, Vol. 71, 34-62.

Karlin, S. (1968) *Total Positivity*, Stanford University Press.

Karlin, S. (1964) The existence of eigenvalues for integral operator. *Trans. Am. Math. Soc.*, Vol. 113, 1-17.

Korteweg, D.J. & de Vries, G. (1895) On the change of form of long wave advancing in a retangular canal, and on a nem type of long stationary waves. *Philos. Mag.*, Vol. 39, No. 5, 422-443.

Lin, Z. (2008) Instability of nonlinear dispersive solitary wave. *J. Funct. Anal.*, Vol. 255, 1091-1124.

Molinet, L. & Ribaud, F. (2009) Well-posedness in H^1 for generalized Benjamin-Ono equations on the circle. *Discrete Contin. Dyn. Syst.*, Vol. 23, No. 4, 1295-1311.

Molinet, L. (2008) Global well-posedness in L^2 for the periodic Benjamin-Ono equation. *Amer. J. Math.*, Vol. 130, No. 5, 635-683

Mielke, A. (1997) Instability and stability of rolls in the Swift-Hohenberg equation. *Comm. Math. Phys.*, Vol. 189, 829-853.

Natali, F. & Pastor, A. (2008) Stability and instability of periodic standing wave solutions for some Klein-Gordon equations. *J. Math. Anal. Appl.*, Vol. 347, 428-441.

Neves, A. (2009) Floquet's theorem and stability of periodic solitary waves. *J. Dynam. Differential Equations*, Vol 21, 555-565.

Shlizerman, E. & Rom-Kedar, V. (2010) Classification of solutions of the forced periodic nonlinear Schrödinger equation. *Nonlinearity*, Vol 23 , 2183-2218.

Stein, E. M.&Weiss, G. (1971) Introduction to Fourier analysis on Euclidean spaces. *Princeton Mathematical Series*, No. 32. Princeton University Press, Princeton, N.J.

Vock, E. & Hunziker, W. (1982) Stability of Schrödinger eigenvalue problems. *Comm. Math. Phys.*, Vol. 83, 281-302.

Weinstein, M.I. (1986) Liapunov stability of ground states of nonlinear dispersive evolution equations. *Comm. Pure Appl. Math.*, Vol. 39, 51-68.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jaime Angulo Pava and Fábio Natali (2011). Orbital Stability of Periodic Traveling Wave Solutions, Fourier Transforms - Approach to Scientific Principles, Prof. Goran Nikolic (Ed.), ISBN: 978-953-307-231-9, InTech, Available from: http://www.intechopen.com/books/fourier-transforms-approach-to-scientific-principles/orbital-stability-of-periodic-traveling-wave-solutions

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

5

Approach to Fundamental Properties of the Henstock-Fourier Transform

Fco. Javier Mendoza Torres, J. Alberto Escamilla Reyna
and Ma. Guadalupe Raggi Cárdenas
Benemérita Universidad Autónoma de Puebla
México

1. Introduction

The Riemann integral was designed to solve different problems in different areas of mathematics. Unfortunately, the Riemann integral has some shortcomings: the derivative of a function is not necessarily Riemann integrable, it lacks of “good" convergence theorems,. . . . To correct these defects, in the year 1902, H. Lebesgue designed an integral (Lebesgue integral) which is more general than Riemann’s, it has better convergence theorems, and it allows integration over other type of sets different from intervals. However, the derivative of a function does not need to be Lebesgue-integrable. On the other hand, every function which is Improper-Riemann-Integrable is not necessarily Lebesgue-integrable.

A. Denjoy (1912) and O. Perron (1914) developed more general integrals than Lebesgue’s. In both integrals, any derivative of a differentiable function is integrable. Both integrals are equivalent but they are difficult to construct. (Gordon, 1994)

Jaroslav Kurzweil (1957), a Czech mathematician, and Ralph Henstock built independently equivalent integrals (Gordon, 1994) which generalize the Lebesgue integral, and it has as “good" convergence theorems as Lebesgue and the derivative of a differentiable function is Henstock-Kurzweil integrable, including the improper Riemann integral. In addition, the construction follows the same pattern as the construction of the Riemann integral. This also facilitates its teaching.

This new integral provided new research lines:

- Construction of new types of integrals by following the Riemann approach.
- Generalization of this concept for functions of several variables, and for functions with values within a Banach space.

In addition, this integral (Henstock-Kurzweil) can be applied to the differential equations theory, integral equations theory, Fourier analysis, probability, statistics, etc.

In the Lebesgue-integrable functions space, we can define a norm with which this space becomes a Banach space with good properties.

Today, Lebesgue integral is the main integral used in various areas of mathematics, for example Fourier analysis. However, many functions (e.g. functions that have a “bad" oscillatory behavior) which are not Lebesgue-integrable are Henstock-Kurzweil-integrable. Therefore, it seems a natural way to study Fourier analysis by using this integral.

Recall that if f is integrable "in some sense", on $\mathbb{R}$, its Fourier transform in $s \in \mathbb{R}$, is defined as

$$\widehat{f}(s) = \int_{-\infty}^{\infty} e^{-ixs} f(x)dx. \tag{1}$$

In the Lebesgue space on $\mathbb{R}$, $L(\mathbb{R})$, the Fourier transform is a bounded linear transformation, whose codomain is the space of continuous functions on $\mathbb{R}$ which "vanish at infinity". It has good algebraic and analytical properties, which have wide applications in mathematics and other sciences.

Four important properties of the Fourier transform in space $L(\mathbb{R})$ are:

i For all $s \in \mathbb{R}$, the Fourier transform exists , because the function $\exp(-ixs)$ is a bounded measurable function.

ii $\widehat{f}$ is continuous on $\mathbb{R}$.

iii Riemann-Lebesgue Lemma: $\lim_{s\to\pm\infty} \widehat{f}(s) = 0$.

iv The Dirichlet-Jordan theorem is valid. This theorem provides us the pointwise inversion for functions of bounded variation on $\mathbb{R}$.

The first study of the Fourier transform using the Henstock-Kurzweil integral was made by E. Talvila in 2002, (Talvila, 2002). He shows important properties of the Fourier transform in the space of Henstock-Kurzweil integrable functions on $\mathbb{R}$, $HK(\mathbb{R})$. However, this study is incomplete, our purpose is to study other properties. We will call *Henstock-Fourier transform* to the Fourier transform definite on $HK(\mathbb{R})$.

Two important differences between the Henstock-Fourier transform and the Fourier transform are:

- This transform does not always exist. For example, the function $f : \mathbb{R} \to \mathbb{R}$ defined as

$$f(x) = \begin{cases} \dfrac{\sin x}{x}, & x \neq 0, \\ 1, & x = 0 \end{cases}$$

  belongs to $HK(\mathbb{R})$, but its Henstock-Fourier transform is not defined in $s = 1$.

- The Riemann-Lebesgue Lemma is not always valid. For example, the function $g(x) = \exp(ix^2)$ (Talvila, 2002) is such that $\widehat{g}(s) = \sqrt{\pi}\exp(i(\pi - s^2)/4)$, however, this later function is not tend to zero when s tend to infinity.

We begin the chapter exposing some fundamental concepts concerning the Henstock-Kurzweil integral, after we show that the intersection of $HK(\mathbb{R})$ and the space of bounded variation functions over $\mathbb{R}$, $HK(\mathbb{R}) \cap BV(\mathbb{R})$, does not have inclusion relations with $L(\mathbb{R})$, for this, we exhibit a wide family of functions in $HK(\mathbb{R}) \cap BV(\mathbb{R})$, which is not in $L(\mathbb{R})$. This makes the study of the Henstock-Fourier transform in this space interesting. Subsequently, in base of $HK(\mathbb{R}) \cap BV(\mathbb{R})$, we prove fundamental properties such as continuity, the Riemann-Lebesgue Lemma, and the Dirichlet-Jordan Theorem.

2. The Henstock-Kurzweil integral

For compact intervals in $\mathbb{R}$, the Henstock-Kurzweil integral is defined in the following way:

Definition 2.1. *Let $f : [a,b] \to \mathbb{R}$ be a function, we will say that f is* ***Henstock-Kurzweil integrable*** *if there exists $A \in \mathbb{R}$, which satisfies the following:*
for each $\varepsilon > 0$ exists a function $\gamma_\varepsilon : [a,b] \to (0,\infty)$ such that for every partition labeled as $P = \{([x_{i-1},x_i],t_i)\}_{i=1}^{n}$, where $t_i \in [x_{i-1},x_i]$, if

$$[x_{i-1},x_i] \subset [t_i - \gamma_\varepsilon(t_i), t_i + \gamma_\varepsilon(t_i)] \quad \textit{for } i = 1,2,...,n, \tag{2}$$

then

$$|\Sigma_{i=1}^{n} f(t_i)(x_i - x_{i-1}) - A| < \varepsilon.$$

The function γ_ε is commonly called **gauge,** and if the partition P complies with the condition (2), we will say that it is γ_ε-**fine**. The number A is named as the integral of f over $[a,b]$ and it is denoted as

$$A = \int_a^b f = \int_a^b f(x)\,dx.$$

If f is defined over an interval of the way $[a,\infty]$, we condition it to $f(\infty) = 0$. In this case, given a gauge function $\gamma_\varepsilon : [a,\infty] \to (0,\infty)$, where $\gamma_\varepsilon(\infty) \in \mathbb{R}^+$, we will say that the labeled partition $P = \{([x_{i-1},x_i],t_i)\}_{i=1}^{n+1}$ is γ_ε-**fine** if:

a) $x_0 = a,\ x_{n+1} = \infty.$

b) $[x_{i-1},x_i] \subset [t_i - \gamma_\varepsilon(t_i), t_i + \gamma_\varepsilon(t_i)]$, for $i = 1,2,...,n$

c) $[x_n,\infty] \subset [1/\gamma_\varepsilon(\infty),\infty]$.

Thus, the function will be integrable if it satisfies Definition 2.1, and also the condition of that the partition P be γ_ε-**fine** according to the previous incises. In addition, these conditions cause that the last term of $\Sigma_{i=1}^{n+1} f(t_i)(x_i - x_{i-1})$ is zero and thus this sum is finite. For functions defined over intervals $[-\infty,a]$ and $[-\infty,+\infty]$ we do similar considerations.
Through the theory of this integral we have that $f : [-\infty,\infty] \to \mathbb{R}$ is an integrable function, if and only if, f is an integrable function over the intervals $[a,\infty]$ y $[-\infty,a]$. In this case

$$\int_{-\infty}^{\infty} f = \int_{-\infty}^{a} f + \int_a^{\infty} f. \tag{3}$$

We denote as

$$HK(I) = \{f : I \to \mathbb{R} \mid f \text{ is Henstock-Kurzweil integrable on } I\}.$$

Some features of $HK(I)$ are the following:

1. It is a vector space, i.e.: the sum of functions and the product by scalars of Henstock-Kurzweil integrable functions are integrable. The integral is a linear functional over this space.
2. It contains the Riemann and Lebesgue integrable functions. Also, the functions whose Riemann or Lebesgue improper integrals exist, and their values coincide.
3. It generalizes the Fundamental Theorem of Calculus, in the sense that every derivative function is integrable. This does not happen with Riemann and Lebesgue integrals. In this case we have:

$$\int_a^b f' = f(b) - f(a).$$

4. Since we know, in Riemann's integral, if two functions are integrable, then their product is also integrable. In the case of the integral of HK, this is not true. Nevertheless, the product of a HK-integrable function by a bounded variation function, is in fact integrable.
5. The HK integral is not an absolute integral. This asseveration is in the sense that if f is HK-integrable, it does not imply that $|f|$ is so. When $|f|$ and f are integrable, we say that f is absolutely HK-integrable.
6. The space of the absolutely HK-integrable lebesgue is $L(I)$, the space of the functions integrable functions.

As we shall see, the properties (4) and (5) produce important differences in the behavior of the Henstock-Fourier transform with respect to the Fourier transform.

2.1 Notation and important theorems for Henstock-Kurzweil integral

Let I be a finite or infinite close interval. We work on the following subspaces:

- $HK(I) = \{f \mid f$ is Henstock-Kurzweil integrable on $I\}$.
- $HK_{loc}(\mathbb{R}) = \{f \mid f \in \mathcal{HK}(I),$ for each finite close interval $I\}$.
- $BV(I) = \{f \mid f$ is of bounded variation on $I\}$.
- If $f \in BV(I)$, $V_I f$ is the total variation of f on I.
- $BV(\,[\pm\infty]\,) = \{f \mid f \in BV(\,[a,\infty]\,) \cap BV(\,[-\infty,b]\,),$ for some $a,b \in \mathbb{R}\}$.
- $BV_0(\,[\pm\infty]\,) = \{f \in BV(\,[\pm\infty]\,) \mid \lim_{|x|\to\infty} f(x) = 0\}$.
- $L(I) = \{f \mid f$ is Lebesgue integrable on $I\}$.

Some of the most important theorems of the Henstock-Kurzweil integral will be used in the proof of our results are as follows.

Theorem 2.1 (Fundamental Theorem I.). *(Bartle, 2001) If $f : [a,b] \to \mathbb{R}$ has a primitive F on $[a,b]$, then $f \in HK([a,b])$ and*

$$\int_a^b f = F(b) - F(a).$$

This theorem guarantees that the derivative of any function on $[a,b]$ is always Henstock-Kurzweil integrable. This result is not valid for Lebesgue integral.

Theorem 2.2 (Fundamental Theorem II.). *(Bartle, 2001) Let a I be a finite o infinite interval. If $f \in HK([a,b])$ then any indefinite integral F is continuous on I and exists a null $Z \subset [a,b]$ such that*

$$F'(x) = f(x) \qquad \textit{for all } x \in I - Z.$$

Theorem 2.3 (Multiplier Theorem.). *(Bartle, 2001) Let $[a,b]$ a finite interval, $f \in HK([a,b])$, $\varphi \in BV([a,b])$ and $F(x) = \int_a^x f(t)$, for $x \in [a,b]$, then, the product $f\varphi \in HK([a,b])$ and*

$$\int_a^b f\varphi = \int_a^b \varphi\, dF = F(b)\varphi(b) - \int_a^b F\, d\varphi, \tag{4}$$

where the second and third integrals are Riemann-Stieltjes integrals.

If $a \in \mathbb{R}$ and $b = \infty$, (4) has the following form

$$\int_a^{\infty} f\varphi = \lim_{b\to\infty}\left[F(b)\varphi(b) - \int_a^b F d\varphi\right]. \tag{5}$$

Similary, if the integration is over the intervals $[-\infty, a]$ or $[-\infty, \infty]$, we have the respective limits in (4).

Theorem 2.4 (Dominated Convergence Theorem.). *(Bartle, 2001) Let $[a,b]$ a interval (finite or infinite), let $\{f_n\}$ be sequence in $HK([a,b])$ such that $f(x) = \lim_{n\to\infty} f_n(x)$ a.e. on $[a,b]$. Suppose that there exist functions $\alpha, \omega \in HK([a,b])$ such that*

$$\alpha(x) \leq f_n(x) \leq \omega(x) \text{ a.e. on } [a,b], \text{ and for all } n \in \mathbb{N}.$$

Then $f \in HK([a,b])$ and

$$\int_a^b f(x)\,dx = \lim_{n\to\infty}\int_a^b f_n(x)\,dx.$$

This theorem is an extension to the Henstock Kurzweil integral of a Dominated Convergence Theorem (DCT) for the Lebesgue integral.

Theorem 2.5 (Hake Theorem.). *(Bartle, 2001) $f \in HK([a,\infty])$, if and only if, $f \in HK([a,c])$ for every compac interval $[a,c]$ with $c \in [a,\infty)$, and there exist $A \in \mathbb{R}$ such that $\lim\limits_{c\to\infty}\int_a^c f(t)dt = A$. In this case, $\int_a^{\infty} f(t)dt = A$.*

There are versions of the Hake's Theorem for functions on $[-\infty,\infty]$ and $[-\infty,a]$.

Theorem 2.6 (Chartier-Dirichlet's Test.). *(Bartle, 2001) Let $f, \varphi : [a,\infty] \to \mathbb{R}$ and suppose that:*

- *$f \in HK([a,c])$ for all $c \geq a$ and $F(x) := \int_a^x f$ is bounded on $[a,\infty)$.*
- *φ is monotone on $[a,\infty]$ and $\lim_{x\to\infty}\varphi(x) = 0$.*

Then $f\varphi \in HK([a,\infty])$.

Theorem 2.7 (Characterization of Absolute Integrability.). *(Bartle, 2001) Let $f \in HK([a,b])$. Then $|f|$ is Henstock-Kurzweil integrable, if and only if, the indefinite integral $F(x) = \int_a^x f$ has bounded variation on $[a,b]$, in this case,*

$$\int_a^b |f| = V_{[a,b]}F.$$

Theorem 2.8 (Comparison Test for Absolute Integrability.). *(Bartle, 2001) If $f, g \in HK([a,b])$ and $|f(x)| \leq g(x)$ for $x \in [a,b]$, then $f \in L([a,b])$. More over, we have*

$$\left|\int_a^b f\right| \leq \int_a^b |f| \leq \int_a^b g.$$

3. The $HK(I) \cap BV(I)$ subspace

If I is a finite interval, we know that:

$$BV(I) \subset L(I) \subset HK(I),$$

and consequently $HK(I) \cap BV(I) \subset L(I)$.
Now, if I is unbounded, the first two observations which we have are

$$BV(I) \nsubseteq L(I), \tag{6}$$

and

$$L(I) \nsubseteq HK(I) \cap BV(I). \tag{7}$$

Really it is easy to demonstrate that the function $f(x) = 1/x$ defined in $[1, \infty]$, is of bounded variation, with

$$V_{[1,\infty]} f = 1,$$

and

$$\int_1^\infty \frac{1}{x}\, dx = \infty.$$

This implies that (6) is true.
To verify (7), we consider the function $f : [0, \infty] \to \mathbb{R}$ defined like

$$f(x) = \begin{cases} \sqrt{x} \sin(1/x), & \text{si } x \in (0, 1], \\ 0, & \text{si } x = 0,\ x \in (1, \infty] \end{cases}$$

which is in $L(\,[0, \infty]\,) \setminus BV(\,[0, \infty]\,)$.

Next, we will prove that: $HK(I) \cap BV(I) \nsubseteq L(I)$.

Proposition 3.1. *(Mendoza et al., 2008) [Theorem 2.1] Let $\varphi : [a, \infty] \to \mathbb{R}$ be a non-negative function, which is decreasing to zero when $x \to \infty$. If $\varphi \notin HK([a, \infty]\,)$, then the functions: $\varphi(t)\sin(t)$ and $\varphi(t)\cos(t)$ are in $HK(\,[a, \infty]\,) \setminus L(\,[a, \infty]\,)$.*

Proof: We will demonstrate that $\varphi(t)\sin(t) \notin L(\,[a, \infty]\,)$. The proof that $\varphi(t)\cos(t) \notin L(\,[a,\ \infty]\,)$ can be done in a similar way.
Suppose that n_0 is the first natural number for which $a < (1 + 4n_0)\pi/4$. For $t \in [a, \infty]$ we have

$$|\sin t| \geq \frac{1}{\sqrt{2}} \text{ if and only if } t \in \cup_{k=n_0}^{\infty} [\,(1 + 4k)\pi/4,\ (3 + 4k)\pi/4].$$

Let $n \in \mathbb{N}$ with $n \geq n_0$, since $(3 + 4n)\pi/4 < (1 + n)\pi$, we have that:

$$\begin{aligned} \int_a^{(1+n)\pi} \varphi(t)|\sin t| dt &\geq \frac{1}{\sqrt{2}} \sum_{k=n_0}^{n} \int_{(1+4k)\pi/4}^{(3+4k)\pi/4} \varphi(t)\, dt \\ &\geq \frac{1}{\sqrt{2}} \sum_{k=n_0}^{n} \int_{(1+4k)\pi/4}^{(3+4k)\pi/4} \varphi((3 + 4k)\pi/4)\, dt \\ &= \frac{\pi}{2\sqrt{2}} \sum_{k=n_0}^{n} \varphi((3 + 4k)\pi/4) \\ &\geq \frac{\pi}{2\sqrt{2}} \sum_{k=n_0}^{n} \varphi((1 + k)\pi). \end{aligned} \tag{8}$$

On the other hand,

$$\begin{aligned}\int_a^{(1+n)\pi}\varphi(t)\,dt &= \int_a^{n_0\pi}\varphi(t)\,dt+\int_{n_0\pi}^{(1+n)\pi}\varphi(t)\,dt\\ &= \int_a^{n_0\pi}\varphi(t)\,dt+\sum_{k=n_0}^{n}\int_{k\pi}^{(1+k)\pi}\varphi(t)\,dt\\ &\leq \int_a^{n_0\pi}\varphi(t)\,dt+\pi\sum_{k=n_0}^{n}\varphi(k\pi). \end{aligned}\tag{9}$$

Since $\varphi \notin HK([a,\infty])$, then $\int_a^\infty \varphi(t)\,dt = \infty$ and from (9) it follows

$$\sum_{k=n_0}^{\infty}\varphi(k\pi)=\infty. \tag{10}$$

Using (10) and approaching $n \to \infty$ in (8), we conclude that $\varphi(t)\sin(t) \notin L([a,\infty])$.
For any $x \in [a,\infty)$,

$$\left|\int_a^x \sin(t)\,dt\right| \leq 2 \ \text{ and } \ \left|\int_a^x \cos(t)\,dt\right| \leq 2.$$

Then according to Chartier-Dirichlet Test (2.6), we have that: $\varphi(t)\sin(t)$ and $\varphi(t)\cos(t)$ are in $HK[a,\infty]$. ■

Example 3.1. *For any $a > 0$,*

$$\frac{\sin(t)}{t} \in HK([a,\infty]) \setminus L([a,\infty]).$$

Proposition 3.2. *(Mendoza et al., 2008) [Corollary 2.2,Theorem 2.2] Let $1 > \alpha > 0$. The function $f_\alpha : [\pi^{1/\alpha},\infty] \to \mathbb{R}$ defined as*

$$f_\alpha(t)=\frac{\sin(t^\alpha)}{t}$$

satisfies:

(a) $f_\alpha \in HK[\pi^{1/\alpha},\infty] \setminus L([\pi^{1/\alpha},\infty])$.

(b) $f_\alpha \in BV([\pi^{1/\alpha},\infty])$.

Proof: **(a)** Let $c > \pi^{1/\alpha}$. Doing a change of variable $u = t^\alpha$ we have that

$$\int_{\pi^{1/\alpha}}^{c}\frac{\sin(t^\alpha)}{t}dt=\frac{1}{\alpha}\int_{\pi}^{c^\alpha}\frac{\sin(u)}{u}du.$$

Since $\sin(u)/u \in HK[\pi,\infty]$, we have that:

$$\lim_{c\to\infty}\int_{\pi^{1/\alpha}}^{c}\frac{\sin(t^\alpha)}{t}dt \ \text{ exists,}$$

thus $f_\alpha \in HK[\pi^{1/\alpha},\infty]$. Moreover since

$$\int_{\pi^{1/\alpha}}^{c}\left|\frac{\sin(t^\alpha)}{t}\right|dt=\frac{1}{\alpha}\int_{\pi}^{c^\alpha}\left|\frac{\sin(u)}{u}\right|du.$$

and $\sin(u)/u \notin L(\,[\pi,\infty]\,)$, then $f_\alpha \notin L[\pi^{1/\alpha},\infty]$.
(b) Let $x \in (\pi^{1/\alpha},\infty)$. We know that $f'_\alpha \in HK(\,[\pi^{1/\alpha},x]\,)$. Now since

$$f'_\alpha(t) = \frac{\alpha\cos(t^\alpha)}{t^{2-\alpha}} - \frac{\sin(t^\alpha)}{t^2},$$

we have that

$$|f'_\alpha(t)| \leq \frac{\alpha}{t^{2-\alpha}} + \frac{1}{t^2}. \tag{11}$$

The function $g(t) = \dfrac{\alpha}{t^{2-\alpha}} + \dfrac{1}{t^2} \in HK(\,[\pi^{1/\alpha},x]\,)$, then by (11) and Theorem 2.8, we conclude that: $f'_\alpha \in L(\,[\pi^{1/\alpha},x]\,)$ and

$$\begin{aligned}\int_{\pi^{1/\alpha}}^{x} |f'_\alpha| &\leq \int_{\pi^{1/\alpha}}^{x} \left(\frac{\alpha}{t^{2-\alpha}} + \frac{1}{t^2}\right) dt \\ &= \left(\frac{1}{\alpha-1}\right)\left[x^{\alpha-1} - \pi^{\frac{\alpha-1}{\alpha}}\right] - \frac{1}{x} + \frac{1}{\pi^{1/\alpha}}.\end{aligned}$$

Consequently by Theorem 2.7,

$$V_{[\pi^{1/\alpha},x]} f_\alpha \leq \left(\frac{1}{\alpha-1}\right)\left[x^{\alpha-1} - \pi^{\frac{\alpha-1}{\alpha}}\right] - \frac{1}{x} + \frac{1}{\pi^{1/\alpha}}.$$

Therefore, as $1-\alpha > 0$ we have that

$$V_{[\pi^{1/\alpha},\infty]} f \leq \frac{1}{(1-\alpha)\pi^{(1-\alpha)/\alpha}} + \frac{1}{\pi^{1/\alpha}}.$$

Thus, $f_\alpha \in BV(\,[\pi^{1/\alpha},\infty]\,)$. ■
Analogy, we can to prove that for $1 > \alpha > 0$, the function $g_\alpha : [-\infty, -\pi^{1/\alpha}] \to \mathbb{R}$ defined as

$$g_\alpha(t) = \frac{\sin(-t)^\alpha}{-t}$$

belongs to $HK([-\infty,-\pi^{1/\alpha}]) \cap BV([-\infty,-\pi^{1/\alpha}]) \setminus L([-\infty,-\pi^{1/\alpha}])$.
Let $h \in BV([-\pi^{1/\alpha},\, \pi^{1/\alpha}])$. For $1 > \alpha > 0$, the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} h(x), & \text{if } x \in (-\pi^{1/\alpha},\, \pi^{1/\alpha}), \\ \dfrac{\sin|t|^\alpha}{|t|}, & \text{if } x \in (-\infty,\, -\pi^{1/\alpha}] \cup [\pi^{1/\alpha},\, \infty) \end{cases}$$

is in $HK(\mathbb{R}) \cap BV(\mathbb{R}) \setminus L(\mathbb{R})$. With this example and Proposition 3.1, we have the following theorem.

Theorem 3.1. *(Mendoza et al., 2009) [Theorem 2.4] There exists a function f in $HK(\mathbb{R}) \cap BV(\mathbb{R}) \setminus L(\mathbb{R})$.*

Now, since $BV(\mathbb{R}) \subset BV(\,[\pm\infty]\,)$, we have immediately the next corollary.

Corollary 3.1. *(Mendoza et al., 2009) [Corollary 2.5] $HK(\mathbb{R}) \cap BV(\,[\pm\infty]\,) \not\subseteq L(\mathbb{R})$.*

We observe that $BV(\mathbb{R}) \subset BV([\pm\infty])$ properly, because instead of the function h in $BV([-\pi^{1/\alpha}, \pi^{1/\alpha}])$ we can to take a function in $HK([-\pi^{1/\alpha}, \pi^{1/\alpha}]) \setminus BV([-\pi^{1/\alpha}, \pi^{1/\alpha}])$.
Also we observe that if $f \in BV(\mathbb{R})$ then, by Multiplier Theorem (2.3), $f(t)\sin t/t \in HK([0,\infty])$.
To conclude this section, we know that $f \in HK(\mathbb{R})$ implies that $f(\pm\infty) = 0$, If in addition $f \in BV(\mathbb{R})$ then $\lim_{|x|\to\infty} f(x)$ exists. Therefore, we have the following lemma.

Lemma 3.1. *If $f \in HK(\mathbb{R}) \cap BV(\mathbb{R})$, then $\lim_{|x|\to\infty} f(x) = 0$ and f is bounded.*

4. Existence and continuity of $\widehat{f}(s)$

4.1 Existence

A part from the Proposition 2.1 b) in (Talvila, 2002), say us that: If $f \in HK_{loc}(\mathbb{R}) \cap BV_0([\pm\infty])$, then $\widehat{f}(s)$ exists for all $s \in \mathbb{R}$. If $s \neq 0$, then the result is true. However with these conditions, it is not necessarily true the existence of $\widehat{f}(0)$. For example, the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1, & \text{if } x \in (-1, 1), \\ \frac{1}{x}, & \text{if } x \in (-\infty, -1] \cup [1, \infty) \end{cases}$$

is in $HK_{loc}(\mathbb{R}) \cap BV_0([\pm\infty])$ but $\widehat{f}(0)$ does not exist.
In order to have the existence of $\widehat{f}(0)$, we need that $f \in HK(\mathbb{R})$.
We will demonstrate that the Henstock-Fourier transform exist in $HK(\mathbb{R}) \cap BV([\pm\infty])$, for every $s \in \mathbb{R}$.

Theorem 4.1. *(Mendoza et al., 2009) [Theorem 3.1] If $f \in HK(\mathbb{R}) \cap BV([\pm\infty])$, then $\widehat{f}(s)$ exists for all $s \in \mathbb{R}$.*

Proof: The result is true for $s = 0$ because $f \in HK(\mathbb{R})$. Now let $s \neq 0$, since $HK(\mathbb{R}) \cap BV([\pm\infty]) \subset HK_{loc}(\mathbb{R}) \cap BV_0([\pm\infty])$ then by (Talvila, 2002) [Proposition **2.1 (b)**] it follows that $\widehat{f}(s)$ exists. However for the sake of completes, here we will give proof of it:
The condition $f \in BV_0([\pm\infty])$ implies that $\lim_{|x|\to\infty} f(x) = 0$ and there exists $a < 0$, $b > 0$ such that f is of bounded variation on $(-\infty, a] \cup [b, \infty)$.
Let us prove that $f(x)e^{-ixs} \in HK([b, \infty))$. The functions φ_1, φ_2 defined as

$$\varphi_1(x) = V_{[b,x]}f - V_{[b,\infty)}f, \qquad \varphi_2(x) = [V_{[b,x]}f - f(x)] - V_{[b,\infty)}f$$

are increasing on $[b, \infty)$ and satisfies that $\lim_{x\to\infty} \varphi_1(x) = \lim_{x\to\infty} \varphi_2(x) = 0$ and $f = \varphi_1 - \varphi_2$. Therefore, since

$$\left| \int_b^x e^{-ius} du \right| = \left| \frac{1}{is}(e^{-ibs} - e^{-ixs}) \right| \leq \frac{2}{s} \qquad \text{for all } x \in [b, \infty),$$

we have by the Chartier-Dirichlet Test (2.6), that $\varphi_1(x)e^{-ixs}$, $\varphi_2(x)e^{-ixs} \in HK([b, \infty))$. Thus $f(x)e^{-ixs} \in HK([b, \infty))$.
In the same way we can to prove that $f(x)e^{-ixs} \in HK((-\infty, a])$. ■

4.2 Continuity

We know that the continuity of the Lebesgue-Fourier transform, on $\mathbb{R}$, is consequence of the dominated convergence Theorem and that the Lebesgue integral is absolute. Now for to prove the continuity of the Henstock-Fourier transform we don't can use the same arguments, because the Henstock - Kurzweil integral is not absolute.

Theorem 4.2. *(Mendoza et al., 2009) [Theorem 4.1] Let f be a function with support in a compact interval such that $f \in HK(\mathbb{R})$. Then $\widehat{f}$ is continuous on $\mathbb{R}$.*

Proof: We consider $[a, b] \subseteq \mathbb{R}$ such that $f(x) = 0$ for all $x \in \mathbb{R} \setminus [a, b]$. Take $t \in \mathbb{R}$ and let $\{t_n\}_{n\in\mathbb{N}} \subseteq (t-1, t+1)$ such that $t_n \to t$. For every $n \in \mathbb{N}$, define $\alpha_n(x) = e^{-ixt_n}$. Then

$$\lim_{n\to\infty} \alpha_n(x) \;=\; \lim_{n\to\infty} e^{-ixt_n} \;=\; e^{-ixt} \qquad \text{for all } x \in [a, b].$$

On the other hand, for every $n \in \mathbb{N}$, α_n is of bounded variation on $[a, b]$ and $V_{[a,b]}\alpha_n \leq 2\max\{|t-1|, |t+1|\}(b-a)$.
Thus according to (Talvila, 1999) [Corollary **3.2**],

$$\lim_{n\to\infty} \int_a^b f(x)e^{-ixt_n}dx \;=\; \lim_{n\to\infty} \int_a^b f(x)\alpha_n(x)dx \;=\; \int_a^b f(x)e^{-ixt}dx.$$

Hence $\lim_{n\to\infty} \widehat{f}(t_n) \;=\; \widehat{f}(t)$. ■

Theorem 4.3. *(Mendoza et al., 2009) [Theorem 4.2] If $f \in HK(\mathbb{R}) \cap BV([\pm\infty])$, then $\widehat{f}$ is continuous on $\mathbb{R} \setminus \{0\}$.*

Proof: Let $t_0 \in \mathbb{R} \setminus \{0\}$ and consider $a < 0$ and $b > 0$ such that $f \in BV(-\infty, a] \cap BV[b, \infty)$. If we show that $\widehat{f\chi}_{(-\infty,a]}$, $\widehat{f\chi}_{[a,b]}$ and $\widehat{f\chi}_{[b,\infty)}$ are continuous in t_0, then $\widehat{f}$ is continuous in t_0, because

$$\widehat{f}(t) = \widehat{f\chi}_{(-\infty,a]}(t) + \widehat{f\chi}_{[a,b]}(t) + \widehat{f\chi}_{[b,\infty)}(t) \quad \text{for all } t \in \mathbb{R}.$$

By the Theorem 4.2, $\widehat{f\chi}_{[a,b]}$ is continuous in t_0. To prove that $\widehat{f\chi}_{(-\infty,a]}$ and $\widehat{f\chi}_{[b,\infty)}$ are continuous in t_0 we will use (Talvila, 2002) [Proposition **6(a)**]. The conditions f is Henstock - Kurzweil integrable on $\mathbb{R}$ and f is of bounded variation on $(-\infty, a] \cup [b, \infty)$ implies that $\lim_{|x|\to\infty} f(x) = 0$. Now since $t_0 \neq 0$, there exists $K > 0$ and $\delta > 0$ such that if $|t - t_0| < \delta$, then $\frac{1}{|t|} < K$. Thus for all $|t - t_0| < \delta$,

$$\left|\int_u^v e^{-ixt}dx\right| \leq \frac{2}{|t|} \;<\; 2K \qquad \text{for all } [u, v] \subseteq \mathbb{R}.$$

Therefore, by (Talvila, 2002) [Proposition **6(a)**], $\widehat{f\chi}_{(-\infty,a]}$ and $\widehat{f\chi}_{[b,\infty)}$ are continuous in t_0. ■

5. The Riemann-Lebesgue lemma

A generalization of the Riemann-Lebesgue Lemma was given, still in the context of the Lebesgue integral, by Bachman (Bachman et al., 1991) [Theorem 4.4.1], assuring that for any $-\infty \leq a < b \leq \infty$,

$$\lim_{|s|\to\infty} \int_a^b h(xs)f(x)dx = 0, \qquad \text{for each } f \in L^1(\mathbb{R}), \tag{12}$$

provided $h : \mathbb{R} \to \mathbb{R}$ is a bounded measurable function satisfying

$$\lim_{|r|\to\infty} \frac{1}{r} \int_0^r h(s)ds = 0.$$

In this section, we show a similar generalization for the Henstock-Fourier transform. In the case of a compact interval, Talvila (Talvila, 2001) showed that the Fourier transform $\hat{f}$ of a function $f \in HK(I) \setminus L^1(I)$ has the asymptotic behavior:

$$\hat{f}(s) = o(s), \text{ as } |s| \to \infty.$$

Titchmarsh (Titchmarsh, 1999) proved it is the best possible approximation for improper Riemann integrable functions. Next, we show too a generalization from this result for the Henstock-Fourier transform.

5.1 The case of a compact interval

The following theorem implies as a corollary the result in (Talvila, 2001).

Theorem 5.1. *(Mendoza et al., 2010) Let $[a,b]$ be a compact interval. Suppose $\varphi : \mathbb{R} \to \mathbb{R}$ is everywhere differentiable with bounded derivative, and such that $\varphi(w) - \varphi(0) = o(w)$, as $|w| \to \infty$. Then,*

$$\int_a^b \varphi(wt)f(t)dt = o(w) \quad \text{as } |w| \to \infty,$$

for each $f \in HK([a,b])$.

Proof: For $w \in \mathbb{R}$, we define $\varphi_w : \mathbb{R} \to \mathbb{R}$ with $\varphi_w(t) = \varphi(wt)$. Moreover, $F(x) := \int_a^x f(t)dt$. Being F continuous and φ' a bounded measurable function, then $F\varphi'^1([a,b]) \subset HK([a,b])$. Also, $f \in HK([a,b])$ and $\varphi_w \in BV([a,b])$, implying $f\varphi_w \in HK([a,b])$. Furthermore, from the Multiplier Theorem (2.3),

$$\int_a^b f(t)\varphi_w(t)dt = F(b)\varphi_w(b) - \int_a^b F(t)\frac{d\varphi_w(t)}{dt}dt.$$

Therefore, for $w \neq 0$,

$$\left| \frac{1}{w} \int_a^b f(t)\varphi(wt)dt \right| \leq \left| \frac{F(b)\varphi(wb)}{w} \right| + \left| \int_a^b F(t)\varphi'(wt)dt \right|. \tag{13}$$

The Fundamental Theorem I (2.1), and the hypotheses for φ imply

$$\lim_{|w|\to\infty} \frac{1}{w} \int_0^w \varphi'(t)dt = \lim_{|w|\to\infty} \frac{\varphi(w) - \varphi(0)}{w} = 0.$$

In consequence,

$$\lim_{|w|\to\infty} \frac{F(b)\varphi(wb)}{w} = 0. \tag{14}$$

Seeing also that $F \in L^1([a,b])$, it follows that equation (12) is valid with f and h substituted for F and φ', respectively. This together with equations (13) and (14) give the result. ■

A direct consequence of the previous theorem is the result of Talvila (Talvila, 2001).

Corollary 5.1. *Let $[a,b]$ be a compact interval. For each $f \in HK([a,b]) \setminus L^1([a,b])$ the Fourier transform has the asymptotic behavior $\hat{f}(s) = o(s)$, as $|s| \to \infty$.*

5.2 The unbounded interval case

Theorem 5.2. *(Mendoza et al., 2010) Let $\varphi \in HK_{loc}(\mathbb{R})$ be fixed. Suppose in addition that $\Phi(x) = \int_0^x \varphi(t)dt$ is bounded on $\mathbb{R}$. Then, for each $f \in HK(\mathbb{R}) \cap BV(\mathbb{R})$,*

$$\lim_{|w|\to\infty} \int_{-\infty}^{\infty} f(t)\varphi(wt)dt = 0.$$

Proof: For $\omega \in \mathbb{R}$, we define $\varphi_w(t) = \varphi(wt)$. Since $\varphi \in HK_{loc}(\mathbb{R})$ then φ and φ_w are in $HK([0,b]$, for $b > 0$. Because $f \in HK(\mathbb{R}) \cap BV(\mathbb{R})$, f is the sum of two monotone functions with limit 0 in infinity. As Φ is bounded in $[0,\infty)$, by the above and from the Chartier-Dirichlet Test (2.6), we have that $f\varphi_w \in HK([0,\infty])$.
For $w \neq 0$, $\Phi_w(t) = (1/w)\Phi(wt)$ is a primitive of φ_w, bounded and continuous on $[0,\infty)$. Because $f \in BV([0,b])$, for $b > 0$, it follows from the Multiplier Theorem (2.3) that

$$\int_0^b f(t)\varphi(wt)dt = \frac{f(b)}{w}\Phi(wb) - \frac{1}{w}\int_0^b \Phi(wt)df(t) \tag{15}$$

The hypotheses for φ imply that $|\Phi(x)| \leq M$, for each $x > 0$, for some constant M.
Now we use theorems (Rudin, 1987) [Theorem 3.8] and Theorem 2.7 to obtain,

$$\left|\int_0^b \Phi(wt)df(t)\right| \leq MV(f;[0,b]),$$

implying, from (15), that

$$\left|\int_0^b f\,(t)\,\varphi\,(\omega t)\,dt\right| \leq \frac{M}{|\omega|}\left(|f(b)| + V\,(f;[0,b])\right). \tag{16}$$

Since $f \in HK([0,\infty)) \cap BV([0,\infty))$, $\lim_{b\to\infty} V(f;[0,b])) = V(f;[0,\infty])$ and $\lim_{b\to\infty} f(b) = 0$. From (16) and Hake's Theorem (2.5) it follows that

$$\left|\int_0^\infty f(t)\varphi(wt)dt\right| \leq \frac{M}{|w|}V(f;[a,\infty].$$

Taking $|w| \to \infty$, we get

$$\lim_{|w|\to\infty} \int_0^\infty f(t)\varphi(wt)dt = 0.$$

A similar argument is valid for the interval $[-\infty,0]$, which yield the result. ■
The trigonometric functions $sin(t)$ and $cos(t)$ obeys the hypotheses the Theorem 5.2. Thus, the result of Mendoza-Escamilla-Sánchez (Mendoza et al., 2009) is a particular case of this theorem.

Corollary 5.2. *For each $f \in HK(\mathbb{R}) \cap BV(\mathbb{R})$, $\lim_{|s|\to\infty} \hat{f}(s) = 0$.*

6. The Dirichlet-Jordan theorem

A fundamental problem for the Fourier Transform is its pointwise inversion, which means to recover the function at given points from its Fourier transform. As is known, the Dirichlet-Jordan Theorem in $L(\mathbb{R})$ solves the pointwise inversion for functions of bounded variation. This theorem tells us that if $f \in L(\mathbb{R}) \cap BV(\mathbb{R})$ then, for each $x \in \mathbb{R}$,

$$\lim_{M\to\infty} \frac{1}{2\pi} \int_{-M}^{M} e^{ixs} \widehat{f}(s)ds = \frac{1}{2}\{f(x+0)+f(x-0)\}. \tag{17}$$

We demonstrate a similar result to (17) for the Henstock-Fourier transform at $HK(\mathbb{R}) \cap BV(\mathbb{R})$. We will use the Sine Integral function, wich is defined as $Si(x) = \frac{2}{\pi}\int_0^x \frac{\sin t}{t}dt$, and has the properties:

1. $Si(0) = 0$, $\lim_{x\to\infty} Si(x) = 1$ and
2. $Si(x) \leq Si(\pi)$ for all $x \in [0,\infty]$.
3. If $b > a > 0$ and $M > 0$, then $\left|\int_a^b \frac{\sin Mt}{t}dt\right| \leq \frac{2}{M}(\frac{1}{a}+\frac{1}{b})$.

Lemma 6.1. *(Mendoza, 2011) Let $\delta > 0$. If $f \in HK([\delta,\infty]) \cap BV([\delta,\infty])$ then*

$$\lim_{M\to\infty} \int\limits_{\delta}^{\infty} \frac{f(t)}{t} sinMtdt = 0$$

Proof: By the Multiplier Theorem (2.3) and the property 3 of the Sine Integral function, it is easy to see that

$$\left|\int_{\delta}^{\infty} \frac{\sin Mt}{t} f(t)dt\right| \leq \frac{2}{M\delta} + \frac{4}{M\delta} V_f([\delta,\infty]).$$

Therefore, making tend M to infinity, we have the result. ■

Lemma 6.2. *(Mendoza, 2011) Let $\delta > 0$. If $f \in HK(\mathbb{R}) \cap BV(\mathbb{R})$, then*

$$\lim_{\varepsilon\to 0} \int_{\delta}^{\infty} f(t)\frac{\sin \varepsilon t}{t} dt = 0.$$

Proof: By the Multiplier Theorem 2.3 and by Lemma 3.1, we have

$$\begin{aligned} \left|\int_{\delta}^{\infty} \frac{\sin \varepsilon t}{t} f\right| &\leq \lim_{b\to\infty} \left\{ \left| f(b) \int_{\delta}^{b} \frac{\sin \varepsilon t}{t} dt\right| + \left|\int_{\delta}^{b} \left(\int_{\delta}^{u} \frac{\sin \varepsilon t}{t}\right) df\right| \right\} \\ &\leq \left|\int_{\delta}^{\infty} \left(\int_{\delta\varepsilon}^{u\varepsilon} \frac{\sin t}{t}\right) df\right|. \end{aligned}$$

How for each $u \in [a,\infty)$: $\lim_{\varepsilon\to\infty} \int_{\delta\varepsilon}^{u\varepsilon} \frac{\sin t}{t}dt = 0$; $\left|\int_{\delta\varepsilon}^{u\varepsilon} \frac{\sin t}{t}dt\right| \leq \pi Si(\pi)$ for all $\varepsilon > 0$; and $\pi(Si)(\pi) \in L(df)$, then, by the Lebesgue Dominated Convergence Theorem 2.4, we obtain the result. ■

Lemma 6.3. *(Mendoza, 2011) Suppose that $f \in HK(\mathbb{R}) \cap BV(\mathbb{R})$ and $\beta, \gamma \in \mathbb{R}$ are such that $[\beta,\gamma] \cap (\mathbb{R} \setminus \{0\} = [\beta,\gamma]$. For all $s \in [\beta,\gamma]$ we have*

$$\lim_{\substack{a\to-\infty \\ b\to\infty}} \int_{\beta}^{\gamma} e^{ixs} \int_{a}^{b} f(t)e^{-ist}dtds = \int_{\beta}^{\gamma} e^{ixs} \int_{-\infty}^{\infty} f(t)e^{-ist}dt\, ds. \tag{18}$$

Proof: For c fixed, let $\widehat{f}_{cb}(s)=\int_c^b f(t)e^{-ist}dt$, $\widehat{f}_c(s)=\int_c^\infty f(t)e^{-ist}dt$, wich are continuous at $\mathbb{R}\setminus\{0\}$. We know that there exists $S>0$ such that $|f(t)|\leq S$ for all $t\in\mathbb{R}$ and that for any $b>c$: $(V_f([c,b])\leq(V_f([c,\infty])$ and $f\in L([c,b])$. By the Multiplier Theorem (2.3), for each $s\in[\beta,\gamma]$, we have

$$\begin{aligned}\left|\int_c^b f(t)e^{-ist}dt\right| &\leq \left|f(b)\left\{\frac{e^{-isb}-e^{-isc}}{-is}\right\}\right|+\left|\int_c^b\left\{\frac{e^{-ist}-e^{-isc}}{-is}\right\}df(t)\right|\\ &\leq \frac{2}{|\beta|}\left\{S+\left|\int_c^b df(t)\right|\right\}\\ &\leq \frac{2}{|\beta|}\left\{S+V_f([c,\infty])\right\}=N_c.\end{aligned}$$

The previous inequality tells us that for any $b>c$ and all $s\in[\beta,\gamma]$: $\left|e^{ixs}\widehat{f}_{cb}(s)\right|\leq N_c$. Applying the Theorem of Hake (2.5): $\lim_{b\to\infty}\widehat{f}_{cb}(s)=\widehat{f}_c(s)$. Then, by the Dominated Convergence Theorem 2.4

$$\lim_{b\to\infty}\int_\beta^\gamma e^{ixs}\int_c^b f(t)e^{-ist}dt\,ds=\int_\beta^\gamma e^{ixs}\int_c^\infty f(t)e^{-ist}dt\,ds.$$

To get the result, we conducted a similar process, now taking the interval $[a,c]$ and making tend a to minus infinity. ■

Because we do not know if $e^{ixs}\widehat{f}$ is integrable around 0, our theorem is as follows:

Theorem 6.1 (Dirichlet-Jordan Theorem for $HK(\mathbb{R})$.)**.** *(Mendoza, 2011) If $f\in HK(\mathbb{R})\cap BV(\mathbb{R})$ then, for each $x\in\mathbb{R}$*

$$\lim_{\substack{M\to\infty\\ \varepsilon\to 0}}\frac{1}{2\pi}\int\limits_{\varepsilon<|s|<M} e^{ixs}\widehat{f}(s)ds=\frac{1}{2}\{f(x+0)+f(x-0)\}. \tag{19}$$

In terms of the Henstock-Kurzweil integral, by the Hake's Theorem (2.5), the above expression (19), shall be equal to

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{ixs}\widehat{f}(s)ds=\frac{1}{2}\{f(x+0)+f(x-0)\}.$$

Proof: Suppose that $\delta>0$ and let $F(x,t)=f(x-t)+f(x+t)$. By the Fubini Theorem for the Lebesgue integral (Apostol, 1974) [Theorem 15.7] at $[-M,-\varepsilon]\times[a,b]$ and $[\varepsilon,M]\times[a,b]$ and by Lemma 6.3

$$\begin{aligned}\int\limits_{\varepsilon<|s|<M} e^{ixs}\int_{-\infty}^{\infty}f(t)e^{-ist}dt\,ds &= \int_{-\infty}^{\infty}f(t)\left(\int_{-M}^{-\varepsilon}+\int_{\varepsilon}^{M}\right)e^{is(x-t)}ds\,dt\\ &= 2\int_0^\infty\frac{F(x,t)}{t}(\sin Mt-\sin\varepsilon t)dt\\ &= 2\int_\delta^\infty\frac{F(x,t)}{t}(\sin Mt-\sin\varepsilon t)dt\\ &+ 2\int_0^\delta\frac{F(x,t)}{t}(\sin Mt-\sin\varepsilon t)dt\end{aligned}$$

In $[\delta, \infty]$ by Lemma 6.1 and Lemma 6.2, we obtain

$$\lim_{\substack{M\to\infty,\\ \varepsilon\to 0}} \int_\delta^\infty \frac{F(x,t)}{t}(\sin Mt - \sin \varepsilon t)dt = 0. \tag{20}$$

In $[0, \delta]$, the DCT (2.4) implies that

$$\lim_{\varepsilon\to 0} \int_0^\delta \frac{F(x,t)}{t} \sin \varepsilon t \, dt = 0. \tag{21}$$

Integrating by parts

$$\begin{aligned} \int_0^\delta [F(x,t)] \frac{\sin Mt}{t} dt &= [F(x,\delta)] \left(\int_0^{\delta M} \frac{\sin t}{t} dt \right) \\ &- \int_0^\delta \left(\int_0^{tM} \frac{\sin u}{u} du \right) d\left[F(x,t)\right]. \end{aligned}$$

Since $\lim_{M\to\infty} \left(\int_0^{Mt} \frac{\sin u}{u} du \right) = \frac{\pi}{2}$ and applying the CDT (2.4) to the last integral, we infer that

$$\begin{aligned} \lim_{M\to\infty} \int_0^\delta [F(x,t)] \frac{\sin Mt}{t} dt &= \frac{\pi}{2} F(x,\delta) \\ &- \frac{\pi}{2} \{ (F(x,\delta)) - (F(x,0)) \} \\ &= \pi \frac{[f(x-0) + f(x+0)]}{2}. \end{aligned}$$

Combining (20), (21) and the above expression, we obtain the result. ■

7. References

Apostol Tom M. (1974). *Mathematical Analysis*, Addison-Wesley Publishing Company, ISBN 0-201-00288-4, Reading, Massachusetts.

Bachman George, Narici Lawrence, Beckenstein Edward. (1991). *Fourier and Wavelet Analysis*, Springer-Verlang, ISBN 0-387-98899-8, New York. Inc.

Bartle Robert G., (2001). *A Modern Theory of Integration*, Graduate Studies in Mathematics, Vol. 32, American Mathematical Society, ISBN 0-8218-0845-1, Providence, Rhode Island.

Gordon R. A., (1994). *The Integrals of Lebesgue, Denjoy, Perron, and Henstock,* Graduate Studies in Mathematics, Vol 4, American Mathematical Society, ISBN 0-8218-3805-9, Providence, Rhode Island.

Mendoza Torres Francisco J., Escamilla Reyna Juan A. y Raggi Cárdenas María G., (2008). About an Existence Theorem of the Henstock-Fourier Transform. *Proyecciones*, Vol. 27, No. 3, (Dec-2008) (307-318), ISSN 0716-0917.

Mendoza Torres Francisco J., Escamilla Reyna Juan A., Sánchez Perales Salvador, Some results about the Henstock-Kurzweil Fourier Transform. *Mathematica Bohemica*, Vol. 134, No. 4, (2009), 379-386, ISSN 0862-7959.

Mendoza Torres Francisco J., Escamilla Reyna Juan A., Morales Macías M. Gpe., Arredondo Ruiz J. Héctor. On the Generalized Riemann-Lebesgue Lemma, submitted to review for publication.

Mendoza Torres Francisco J. The Dirichlet-Jordan Theorem for the Henstock-Fourier Transform, accepted for publication in Annals of Functional Analysis.

Pinsky Mark A., (2002). *Introduction to Fourier Analysis and Wavelets*, Books/Cole, Pacific Grove, ISBN 0-53-37660-6, C.A.

Rudin Walter, (1987). *Real and Complex Analysis*, third edition, Mc. Graw-Hill Companies, ISBN 0-07-054234-1.

Talvila E. (1999). Limits and Henstock integrals of products , *Real Anal. Exchange*, Vol. 25, No. 2, (1999) 907-918, ISSN 0147-1937.

Talvila E., (2001). Rapidly Growing Fourier Integrals, *American Mathematical Monthly*, Vol. 108, No. 7, (August-2001), 636-641, ISSN 0002-9890.

Talvila E., (2002). Henstock-Kurzweil Fourier Transforms, *Illinois Journal of Mathematics*, Vol. 46, No. 4, (2002), 1207-1226, ISSN 0019-2082.

Titchmarsh E. C., The Order of Magnitude of the Coefficients in a Generalized Fourier Series, *Proc. London Math. Soc.*, Vol. 2, No. 22, 1923-24.

This chapter was sponsored by VIEP-BUAP, through of the research projects: *“Properties of the Space of Henstock-Kurzweil integrable functions and the Henstock-Fourier Transform”*, and *“Fourier analysis in the space of integrable functions in the sense of Henstock-Kurzweil”*.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Three Dimensional Reconstruction Strategies Using a Profilometrical Approach based on Fourier Transform

Pedraza-Ortega Jesus Carlos, Gorrostieta-Hurtado Efren,
Aceves-Fernandez Marco Antonio, Sotomayor-Olmedo Artemio,
Ramos-Arreguin Juan Manuel, Tovar-Arriaga Saul
and Vargas-Soto Jose Emilio
Facultad de Informatica – Universidad Autonoma de Queretaro
Mexico

1. Introduction

In the past 3 decades, there is an idea to extract the 3D information of a scene from its 2D images ant it has been a research interest in many fields. The main idea is to extract the useful depth information from an image or set of images in an efficient and automatic way. The result of the process (depth information) can be used to guide various tasks such as synthetic aperture radar (SAR), magnetic resonance imaging (MRI), automatic inspection, reverse engineering, 3D robot navigation, interferometry and so on. The obtained information can be used to guide various processes such as robotic manipulation, automatic inspection, inverse engineering, 3D depth map for navigation and virtual reality applications (Gokstorp, 1995). Depending on the application, a simple 3D description is necessary to understand the scene and perform the desired task, while in other cases a dense map or detailed information of the object's shape is necessary. Furthermore, in some cases a complete 3D description of the object may be required.

In 3D machine vision, the three-dimensional shape can be obtained by using two different methodologies; Active and Passive Methods, which are also classified as contact and non contact methods. The active methods project energy in the scene and detect the reflected energy; some examples of these methods are sonar, laser ranging, fringe projection and structured method.

The fringe processing methods are widely used in non-destructive testing, optical metrology and 3D reconstruction systems. Some of the desired characteristics in these methods are high accuracy, noise-immunity and processing speed.

In the spatial and temporal fringe pattern analysis, the main characteristics are the number of fringes, and the intensity variation due temporal and spatial measurements.

A few commonly used fringe processing methods are well-known like Fourier Transform Profilometry (FTP) method (Malacara, 2006) and phase-shifting interferometry (Takeda et al., 1992). The main problem to overcome in these methods is the wrapped phase, where the depth information is included.

Despite the fact that most of the previous cited works are proved and tested in previous research, the present work present two strategies as an overview to the Fourier Transform Profilometry. The first one is a modified algorithm to the Fourier Transform Profilometry Method, where an additional pre-processing filter plus a data analysis in the unwrapping step is presented. The second strategy presented here is the use of the local and global phase unwrapping in the Modified Fourier Transform Profilometry. Both proposed methods present some advantages and the simplicity of the algorithm could be considered for implementation in real time 3D reconstruction.

2. Fourier transform profilometry

The image of a projected fringe pattern and an object with projected fringes can be represented by the following equations:

$$g(x,y) = a(x,y) + b(x,y) * \cos[2 * \pi f_0 x + \varphi(x,y)] \tag{1}$$

$$g_0(x,y) = a(x,y) + b(x,y) * \cos[2 * \pi f_0 x + \varphi_0(x,y)] \tag{2}$$

where *g(x,y)* *g_0(x,y)* are the intensity of the images at *(x,y)* point, *a(x,y)* represents the background illumination, *b(x,y)* is the contrast between the light and dark fringes, *f_0* is the spatial-carrier frequency and *φ(x,y)* and *φ_0(x,y)* are the corresponding phase to the fringe and distorted fringe pattern, observed from the camera.

The phase *φ(x,y)* contains the desired information, and *a(x,y)* and *b(x,y)* are unwanted irradiance variations. In most cases *φ(x,y)*, *a(x,y)* and *b(x,y)* vary slowly compared with the spatial-carrier frequency f0. Then, the angle *φ(x,y)* is the phase shift caused by the object surface end the angle of projection, and its expressed as:

$$\varphi(x,y) = \varphi_0(x,y) + \varphi_z(x,y) \tag{3}$$

Where *φ_0 (x,y)* is the phase caused by the angle of projection corresponding to the reference plane, and φ_z(x,y) is the phase caused by the object's height distribution.

Considering the figure 1, we have a fringe which is projected from the projector, the fringe reaches the object at point H and will cross the reference plane at the point C. By observation, the triangles DpHDc and CHF are similar and

$$\frac{CD}{-h} = \frac{d_0}{l_0} \tag{4}$$

Leading us to the next equation:

$$\varphi_z(x,y) = \frac{h(x,y) 2\pi f_0 d_0}{h(x,y) - l_0} \tag{5}$$

Where the value of *h(x,y)* is measured and considered as positive to the left side of the reference plane. The previous equation can be rearranged to express the height distribution as a function of the phase distribution:

$$h(x,y) = \frac{l_0 \phi_z(x,y)}{\phi_z(x,y) - 2\pi f_0 d_0} \tag{6}$$

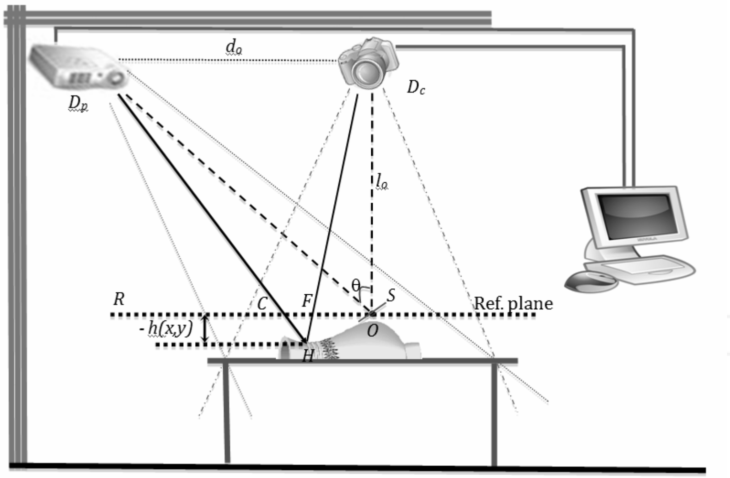


Fig. 1. Experimental setup

2.1 Fringe analysis

The fringe projection equation 1 can be rewritten as:

$$g(x,y) = \sum_{n=-\infty}^{\infty} A_n r(x,y) \exp(in\varphi(x,y)) * \exp(i2\pi n f_0 x) \tag{7}$$

Where r(x,y) is the reflectivity distribution on the diffuse object (Taketa et al., 1992) (Berryman et al., 2003). Then, a FFT (Fast Fourier Transform) is applied to the signal for in the x direction only. Notice that even y is considered as fix, the same procedure will be applied for the number of y lines in both images. Therefore, we obtain the next equation:

$$G(f,y) = \sum_{-\infty}^{\infty} Q_n(f - nf_0, y) \tag{8}$$

Now, we can observe that φ(x,y) and r(x,y) vary very slowly in comparison with the fringe spacing, then the Q peaks in the spectrum are separated each other. Also it is necessary to consider that if we choose a high spatial fringe pattern, the FFT will have a wider spacing among the frequencies. The next step is to remove all the signals with exception of the positive fundamental peak f_0. The obtained filtered image is then shifted by f_0 and centered. Later, the IFFT (Inverse Fast Fourier Transform) is applied in the *x* direction only, same as the FFT. The obtained equations for the reference and the object are given by:

$$\hat{g}(x,y) = A_1 r(x,y) \exp\{i(2\pi f_0 x + \varphi(x,y))\} \tag{9}$$

$$\hat{g}_0(x,y) = A_1 r_0(x,y) \exp\{i(2\pi f_0 x + \varphi_0(x,y))\} \tag{10}$$

By multiplying the *φ(x,y)* with the conjugate of *φ_0(x,y),* and separating the phase part of the result from the rest we obtain:

$$\begin{aligned}\varphi_z(x,y) &= \varphi(x,y) + \varphi_0(x,y) \\ &= \mathrm{Im}\{\log(\hat{g}(x,y)\hat{g}_0^*(x,y))\}\end{aligned} \tag{11}$$

From the above equation, we can see that the phase map can be obtained by applying the same process for each horizontal line. The values of the phase map are wrapped at some specific values. Those phase values range between π and -π.

To recover the true phase it is necessary to restore to the measured wrapped phase of an unknown multiple of $2\pi f_0$.The phase unwrapping process is not a trivial problem due to the presence of phase singularities (points in 2D, and lines in 3D) generated by local or global sub-sampling. The correct 2D branch cut lines and 3D branch cut surfaces should be placed where the gradient of the original phase distribution exceeded π rad value. However, this important information is lost due to undersampling and cannot be recovered from the sampled wrapped phase distribution alone. Also, is important to notice that finding a proper surface, or obtaining a minimal area or using a gradient on a wrapped phase will not work and could not find the correct branch in cut surfaces. From here, it can be observed that some additional information must be added in the branch cut placement algorithm.

Therefore, the next step is to apply some improved phase unwrapping algorithms. The whole methodology is described in figure 2.

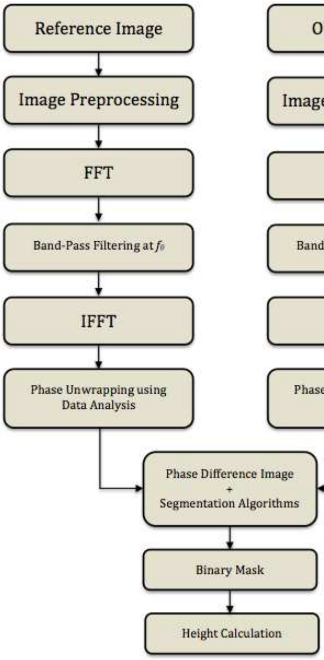

Reference Image
Object Image
Image Preprocessing
Image Preprocessing
FFT
FFT
Band-Pass Filtering at f_0
Band-Pass Filtering at f_0
IFFT
IFFT
Phase Unwrapping using Data Analysis
Phase Unwrapping using Data Analysis
Phase Difference Image + Segmentation Algorithms
Binary Mask
Height Calculation

Fig. 2. Firstly Proposed Methodology

2.2 Phase unwrapping in the modified Fourier transform profilometry

As was early mentioned, the unwrapping step consists of finding discontinuities of magnitude close to 2π, and then depending on the phase change we can add or take 2π to the shape according to the sign of the phase change. There are various methods for doing the phase unwrapping, and the important thing to consider here is the abrupt phase changes in the neighbouring pixels. There are a number of 2π phase jumps between 2 successive wrapped phase values, and this number must be determined. This number depends on the spatial frequency of the fringe pattern projected at the beginning of the process.

This step is the modified part in the Fourier Transform Profilometry originally proposed by Takeda [3], and represents the major contribution of this work. Another thing to consider is to carry out a smoothing before the doing the phase unwrapping, this procedure will help to reduce the error produced by the unwanted jump variations in the wrapped phase map. Some similar methods are described in (Pramod, 2003)(Wu, 2006). Moreover, a modified Fourier Transform Profilometry method was used in (Pedraza et al, 2007) that include some extra analysis which considers local and global properties of the wrapped phase image.

Moreover, a second modification to the Fourier Transform Profilometry is proposed and presented in figure 3.

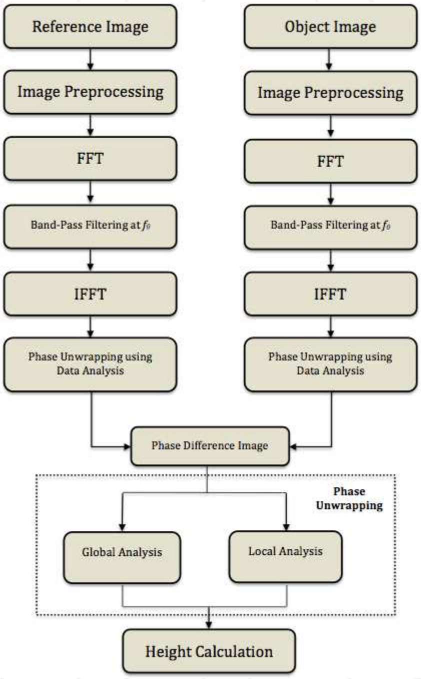


Fig. 3. Second Proposed Methodology

3. Phase unwrapping

Since two decades ago, phase unwrapping has been a research area and many papers have been published, presenting some ideas that solves the problem. Several phase unwrapping algorithms have been proposed, implemented and tested.

The phase unwrapping process is not a trivial problem due to the presence of phase singularities (points in 2D, and lines in 3D) generated by local or global undersampling. The correct 2D branch cut lines and 3D branch cut surfaces should be placed where the gradient of the original phase distribution exceeded π rad value. However, this important information is lost due to undersampling and cannot be recovered from the sampled wrapped phase distribution alone. Also, is important to notice that finding a proper surface, or obtaining a minimal area or using a gradient on a wrapped phase will not work and one could not find the correct branch in cut surfaces.

The phase unwrapping has many applications in applied optics that require an unwrapping process, and hence many phase unwrapping algorithms has been developed specifically for data with a particular application. Moreover, there is no universal phase unwrapping algorithm that can solve wrapped phase data from any application. Therefore, phase unwrapping algorithms are considered as a trade-off problem between accuracy of solution and computational requirements. However, even the most robust and complete phase unwrapping algorithm cannot guarantee in giving successful or acceptable unwrapped results without a good set of initial parameters. Unfortunately, there is no standard or technique to define the parameters that guarantee a good performance on phase unwrapping.

In literature, exist several phase unwrapping algorithms, a general review of the most widely used algorithms used started with the single phase unwrapping algorithm proposed by (Takeda et al, 1982), later the continuous phase map was proposed by (Giglia & Pritt, 1998) and more recently (Kian et al., 2005) proposed a windowed fourier transform as a filter to approach the phase unwrapping, later the local and global analysis was proposed by (Pedraza et al, 2007). Broadly speaking, the local phase unwrapping algorithms can be divided in two main subcategories *quality guided and residue balancing or branch cuts.* On the other hand are the global phase unwrapping algorithms that deal with the problem of phase unwrapping in a minimum-norm (or minimization) approach, example of this phase unwrapping algorithms are unweighted least squares, weighted least squares, etc.

Generally, in order to face the phase unwrapping problems, algorithms can be divided in two categories: local and global phase unwrapping. Local phase unwrapping algorithms find the unwrapped phase values by integrating the phase along a certain path. This is called path-following algorithms. Another way to classify the phase unwrapping algorithms are temporal (mention some algorithms) or spatial (add some algorithms too) phas unwrapping according with the appropriate fringe pattern analysis. The post processing of the unwrapped phase is needed in order to improve the 3D Reconstruction results, and some analysis that were carried out by (Kian Q, 2007).

Global phase unwrapping algorithms locate the unwrapped phase by minimizing a global error function and are also called local phase unwrapping algorithm and a global phase unwrapping algorithm, by following the methodology proposed by us in (Pedraza et al., 2009). The unwrapped phase values and the wrapped phase can be related with each other according with the Shannon's sampling theorem:

$$\Psi(n) = \varphi(\pi) + 2\pi k(n) \qquad -\pi < \Psi(n) \le \pi \tag{12}$$

$$\varphi(n) = \Psi(n) + 2\pi v(n) \qquad -\infty < \varphi(n) \le \infty \tag{13}$$

here *Ψ(n)* holds the wrapped phase values and *φ(n)* holds the unwrapped phase values, *k(n)* is the function containing the integers that must be added to the wrapped phase *φ* to be

unwrapped, n is an integer and *v(n)* is the function containing a set of integers that must be added to the wrapped phase *Ψ*.
Noticing that;

$$v(n) = -k(n) \tag{14}$$

The wrapping operation *ω* which converts the unwrapped phase is defined by:

$$\varpi\{\varphi(n)\} = \arctan\left[\frac{\sin(\varphi(n))}{\cos(\varphi(n))}\right] \tag{15}$$

3.1 Local phase unwrapping

Local phase unwrapping algorithms finds the unwrapped phase values by integrating the phase along certain path that covers the whole wrapped phase map. The local phase unwrapping defines the quality of each pixel in the phase map to unwrap the highest quality pixels first and the lowest quality pixels last (quality- guided phase unwrapping). The second type is known as residue-balancing methods, which attempts to prevent error propagation by identifying residues (the source of noise in the wrapped phase). The residues must be balanced and isolated by using barriers (branch-cuts), therefore, it aims to produce a path-independent wrapped phase map. Path-dependency occurs to the existence of residues.

Residue-balancing algorithms search for residues in a wrapped-phase map and attempt to balance positive and negative residues by placing cut lines between them to prevent the unwrapping path breaking the mesh created. The residue is identified for each pixel in the phase map by estimating the wrapped gradients in a 2 × 2 closed loop, as shown in Figure 4.

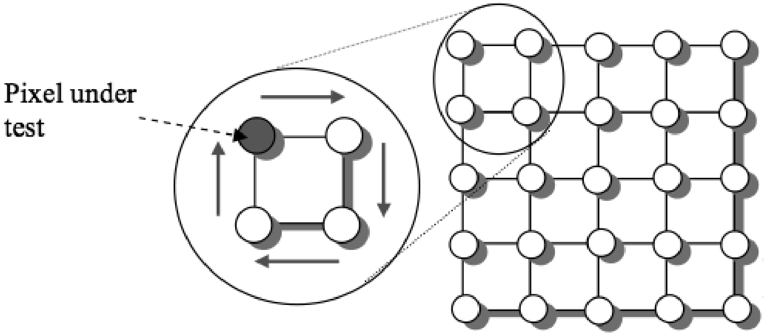


Fig. 4. Identifying residues in a 2 × 2 closed path

This is carried out using the following equation:

$$r = \Re\left[\frac{\Psi_{i,j} - \Psi_{i+1,j}}{2\pi}\right] + \Re\left[\frac{\Psi_{i+1,j} - \Psi_{i+1,j+1}}{2\pi}\right] + \Re\left[\frac{\Psi_{i,j+1} - \Psi_{i,j}}{2\pi}\right] \tag{26}$$

Where **ℜ[]** rounds its argument to the nearest integer, *Ψx,y* is the wrapped pixel. The equation of interest; $f(x) = a(x) + b(x)\cos(2\pi f_0 x + \varphi(x))$ can only take three possible results: 0,

+1, and -1. A pixel under test is considered to be a positive residue if the value of r is +1, and it is considered to be a negative residue if the value is -1. Conversely, the pixel is not a residue if the value of *r* is zero. After identifying all residues in the wrapped phase map, these residues have to be balanced by means of branch cuts. Branch-cuts act as barriers to prevent the unwrapping path going thorough them. If these branch cuts are avoided during the unwrapping process, no errors propagate and the unwrapping path is considered to be path independent. On the other hand, if these branch cuts are penetrated during the unwrapping, errors propagate throughout the whole phase map, and in this case the unwrapping path is considered to be path dependent.

3.2 Global phase unwrapping

In the previous section, it was stated that local phase unwrapping algorithms follow a certain unwrapping path in order to unwrap the phase. They begin at a grid point and integrate the wrapped phase differences over that path, which ultimately covers the entire phase map. Local phase unwrapping algorithms (residue-balancing algorithms) generate branch cuts and define the unwrapping path around these cuts in order to minimize error propagation.

In contrast, global phase unwrapping algorithms formulate the phase unwrapping problem in a generalized minimum-norm sense (Ichioka & Inuiya, 1972). Global phase unwrapping algorithms attempt to find the unwrapped phase by minimizing the global error function as:

$$\varepsilon^2 = ||\text{solution – problem }||^2 \tag{17}$$

Global phase unwrapping algorithms seek the unwrapped phase whose local gradients in the x and y direction match, as closely as possible.

$$\varepsilon^2 = \sum_{i=0}^{M-2}\sum_{j=0}^{N-1}\left|\Delta^x\varphi(i,j) - \hat{\Delta}^x\psi(i,j)\right|^p + \sum_{i=0}^{M-1}\sum_{j=0}^{N-2}\left|\Delta^y\varphi(i,j) - \hat{\Delta}^y\psi(i,j)\right|^p \tag{18}$$

Where $\Delta^x\varphi(i,j)$ and $\Delta^y\varphi(i,j)$ are unwrapped phase gradients in the x and y directions respectively, which are given by:

$$\Delta^x\varphi(i,j) = \varphi(i+1,j) - \varphi(i,j) \tag{19}$$

$$\Delta^y\varphi(i,j) = \varphi(i,j+1) - \varphi(i,j) \tag{20}$$

$\hat{\Delta}^x\psi(i,j)$ and $\hat{\Delta}^y\psi(i,j)$ are the wrapped values of the phase gradients in the x and y directions respectively, and they are given by:

$$\hat{\Delta}^x\psi(i,j) = \omega\{\psi(i+1,j) - \psi(i,j)\} \tag{21}$$

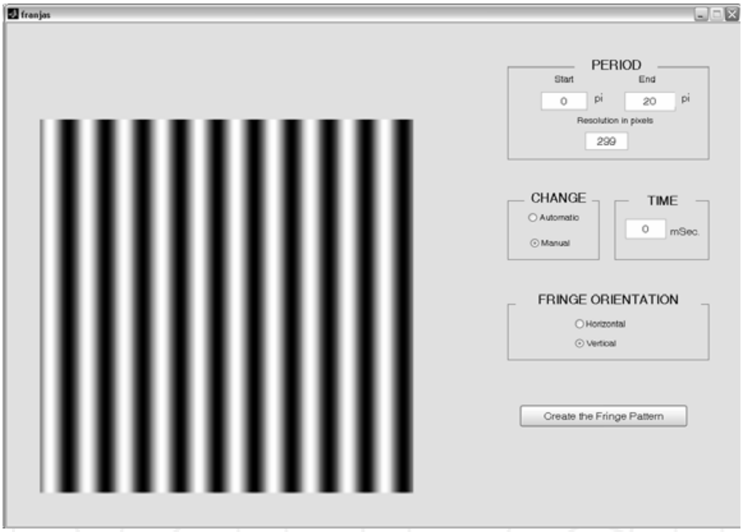
$$\hat{\Delta}^y\psi(i,j) = \omega\{\psi(i,j+1) - \psi(i,j)\} \tag{22}$$

Finally the wrapping operator is defined by the equation 15.

4. Experimental results

An experimental setup such as the one shown on figure 1 is suitable to apply the proposed methodology. In figure 1, a high resolution digital projector is used to create the structured light fringe pattern, and a mega-pixel digital CCD camera is used as a sensor to acquire the images. Also, a high-resolution digital CCD camera can be used instead of the mega-pixel camera. The reference plane can be any flat surface like a plain wall, or a whiteboard. In the reference plane is important to notice that the surface is a non-reflecting one in order to avoid the unwanted reflections that may affect or distort the image acquisition process. The object of interest can be any object and for this project, 2 objects are considered; the first one is an oval with a symmetrical shape and also a pyramid.

To create different fringe patterns, a GUI was developed. The GUI is capable to create several patterns by modifying the spatial frequency (number of fringes per unit area), and resolution (number of levels to create the sinusoidal pattern) of the fringe pattern. The GUI has the capability to do phase shifting if necessary and also the projection of the fringe pattern can have a horizontal or vertical orientation.



Start 0 pi End 20 pi

Resolution in pixels 299

○ Automatic

● Manual

0 mSec.

○ Horizontal

● Vertical

Fig. 3. Fringe Pattern GUI in MATLAB

As an example of one object, we can see on figure 4 the reference pattern projected on a plane and the same pattern projected on the object.

Applying the modified Fourier Transform Profilometry we can obtain the Fourier spectra corresponding to the images on figure 5.

On the left part of the figure 6 we can observe the wrapped depth map before applying the unwrapped algorithm. Usually, in doing phase unwrapping, linear methods are used [5-7]. These methods fail due to the fact the in the wrapped direction of the phase, a high frequency can be present and a simple unwrapping algorithm can generate errors in the mentioned direction. That is the main reason why a more complete analysis should be performed. In this research, a local discontinuity analysis together with the use of the global analysis is also implemented.

The main algorithm for the local discontinuity analysis [8] is described as; a) first, divide the wrapped phase map in regions and give a different weights (w1, w2, .., wn) to each region, b) the modulation unit is defined and helps to detect the fringe quality and divides the fringes into regions, c) regions are grouped from the biggest to the smallest modulation value, d) next, the unwrapping process is started from the biggest to the smallest region, e) later, an evaluation of the phase changes can be carried out to avoid variations smaller than f0.

After the local analysis, an unwrapping algorithm is applied considering punctual variations in the phase difference image, which will lead us to the desired phase distribution profile, that is, the object's form. Also, to help the selection of the proper value, a binary mask is used, like the one showed on figure 6 on the right hand. This binary mask gives an extra parameter, which has the value of "1" in the pixel where a phase jump is bigger than 2π. From the figure 6, it can be shown that there is more than 1 jump in a frequency higher than f0.

All the phase unwrapping was carried out in the y direction.

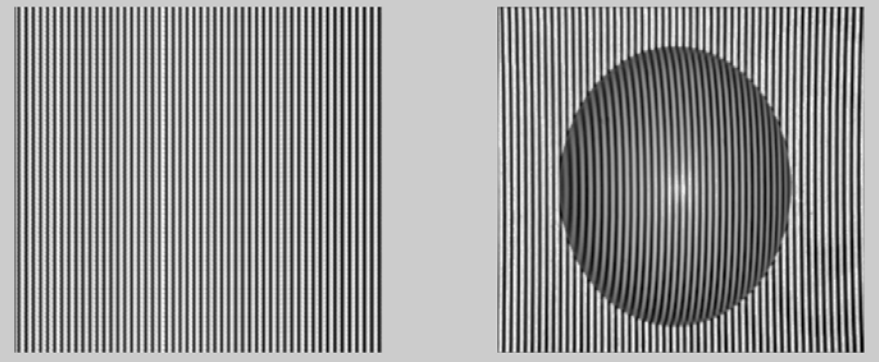


Fig. 4. Fringe Pattern projected on a plane and object to digitize respectively

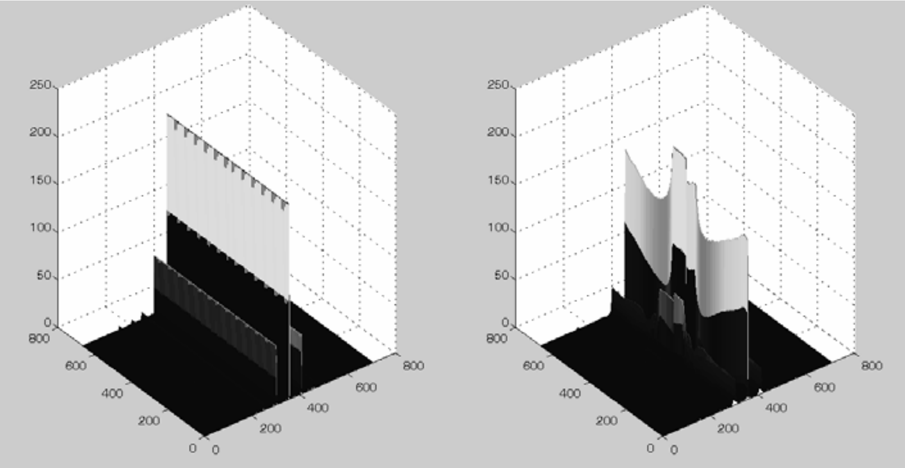
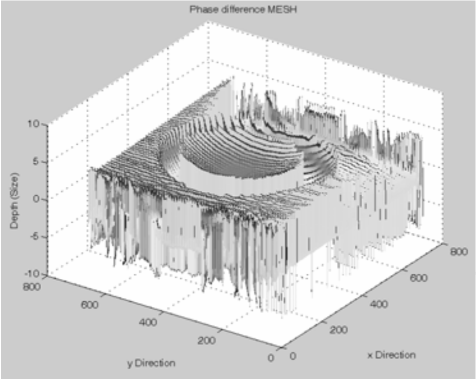


Fig. 5. Phase of projected fringe pattern on a plane and object to digitize respectively



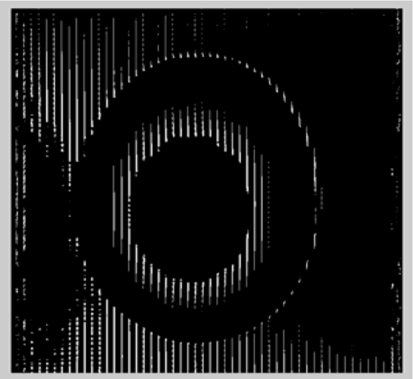


Fig. 6. Wrapped mesh and binary mask

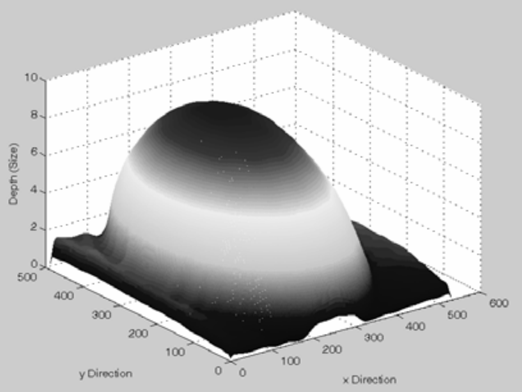


Fig. 7. The final 3D reconstruction mesh after the proposed methodology was applied

On figure 7, the final 3D reconstruction mesh is obtained after the proposed methodology is applied.

For the experimentation, a CCD camera SONY TRV-30 1.3 Mega-pixels was used. As a reference frame, a wooden made plane was used, and it was painted with black opaque paint to avoid glare. The digitized objects were an oval and a pyramid.

On Figure 8, a pyramid, the second object used in this work is presented. The projected fringe pattern and its Fourier spectra is observed, where a nearly 50 pixel spatial frequency is observed. Figure 9 shows the wrapped phase of the pyramid, and after applying the phase difference algorithm and binary mask, the 3D mesh of the object is presented.

As a second test, some computer based simulations using virtual created objects were carried out. On figure 10, a computer created hand was created. In this figure, the mesh visualization, as well as the projected pattern are presented Later, on figure 11, the wrapped phase image as well as the phase mesh are presented. Finally, after applying the proposed strategy, the reconstructed object is presented by using the global phase unwrapping and the local phase unwrapping. Two more object

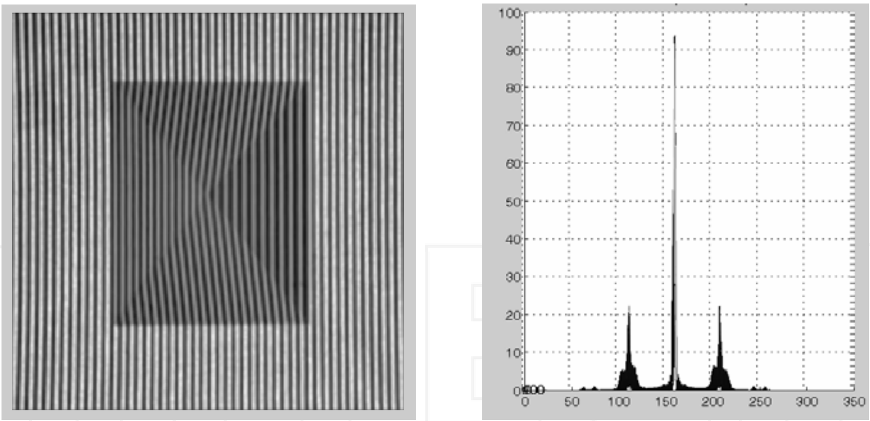


Fig. 8. Fringe pattern projected on a pyramid object and its phase

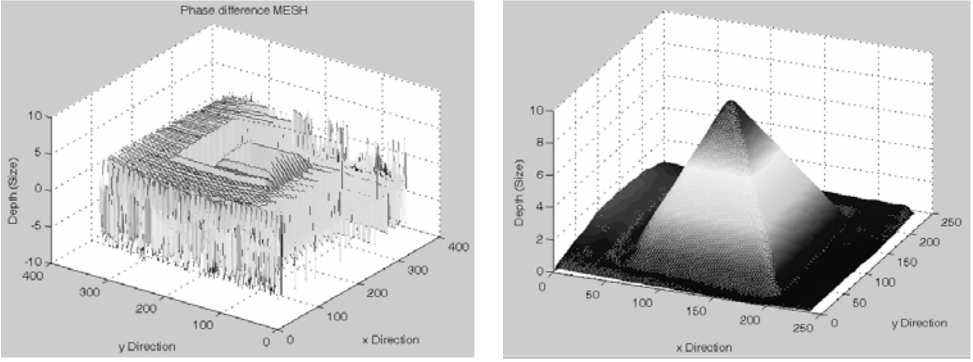


Fig. 9. Pyramid wrapped phase and its 3D reconstruction after the proposed method is applied

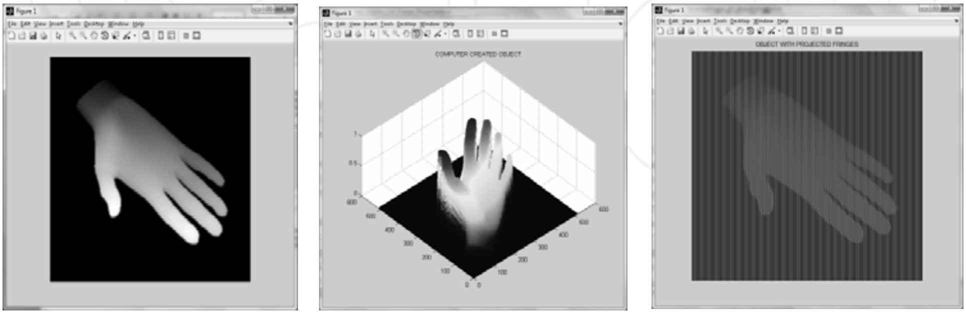
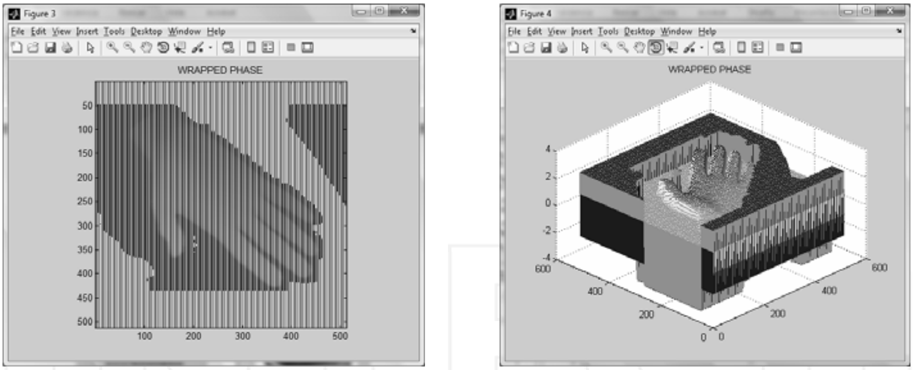


Figure 1
File Edit View Insert Tools Desktop Window Help
COMPUTER CREATED OBJECT
OBJECT WITH PROJECTED FRINGES

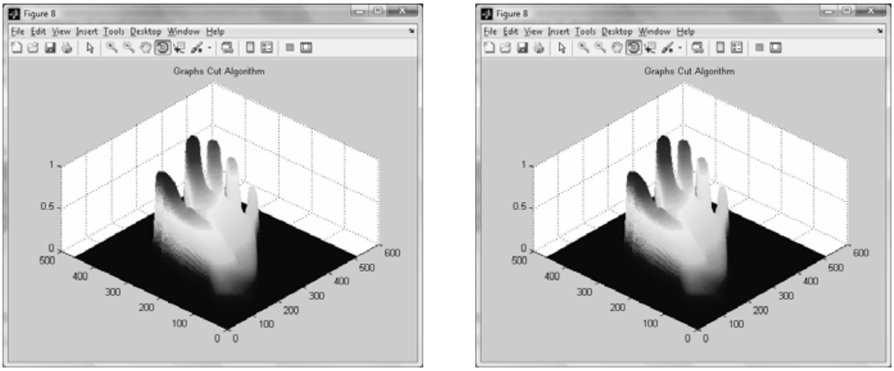
(a) (b) (c)

Fig. 10. Computer created object: (a) Hand, (b) Mesh Visualization and (c) Projected fringes



(a) (b)

Fig. 11. Computer created object: (a) Wrapped phase image, (b) Wrapped phase Mesh



(a) (b)

Fig. 12. After applying the second strategy: (a) using local phase unwrapping, (b) using global phase unwrapping

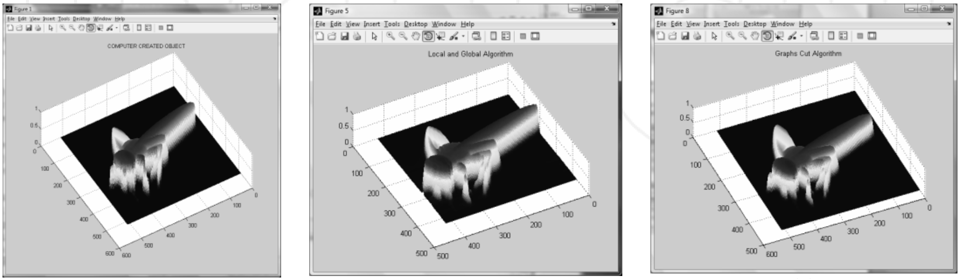
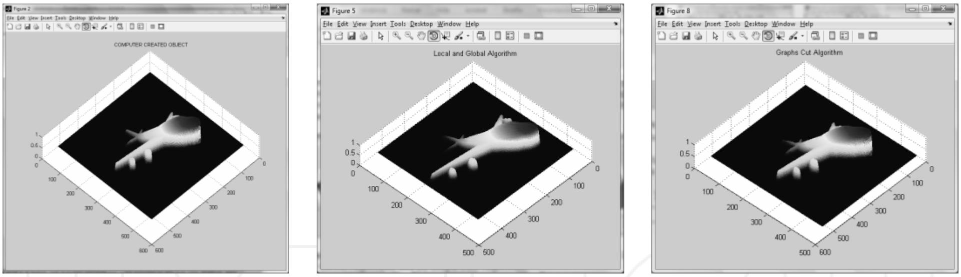


Figure 1
File Edit View Insert Tools Desktop Window Help
COMPUTER CREATED OBJECT
Figure 5
File Edit View Insert Tools Desktop Window Help
Local and Global Algorithm
Figure 8
File Edit View Insert Tools Desktop Window Help
Graphs Cut Algorithm

(a) (b) (c)

Fig. 13. Dragonfly digitizing using the Second Strategy: (a) Virtual object to digitize (b) using local phase unwrapping, (c) using global phase unwrapping



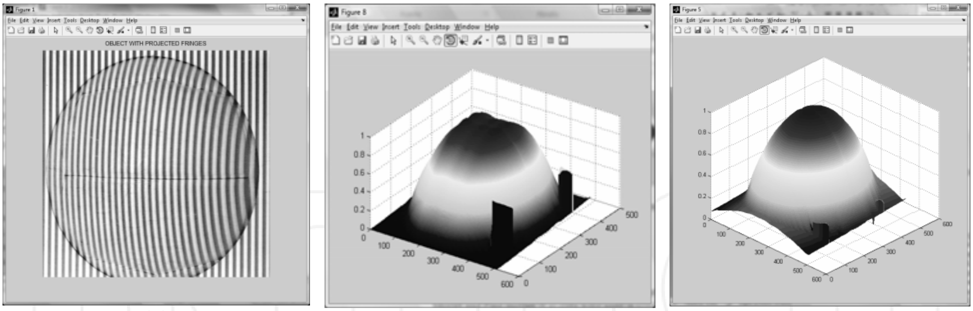
(a) (b) (c)

Fig. 13. Plane digitizing using the Second Strategy: (a) Virtual object to digitize (b) using local phase unwrapping, (c) using global phase unwrapping

Object	Local	Global
Hand	4.45	3.51
Dragonfly	4.67	3.81
Airplane	4.58	3.67

Table 1. Comparison between the phase unwrapping algorithms in the second strategy

Finally, the methodology was applied to a real object, which is presented on figure 14(a). The object is a volley ball, and the results of the local and global phase unwrapping algorithms are presented on the same figure 14(b) and (c) respectively.



(a) (b) (c)

Fig. 13. Object to digitize: (a) Volley ball (a) using local phase unwrapping, (b) using global phase unwrapping

5. Conclusions and future work

There are several methods and techniques to made three-dimensional reconstruction of virtual and real objects. Among all these methods the fringe pattern analysis had been widely used since it provides a non-destructive approach, to optical metrology and 3D reconstruction systems.

As an option to get the 3D reconstruction of objects, two basic strategies were introduced. A modified Fourier Transform Profilometry methodology and the Modified Fourier Transform Profilometry including the local and global phase unwrapping were described.

The fringe pattern analysis has two main phases, the phase extraction and the phase unwrapping; in this chapter a two modified profilometry methodologies are presented in order to perform this two phases.

The phase extraction basically consists in analyse a distorted fringe pattern image by using Fourier transform and filtering the undesired noise and frequency. The result of this phase is commonly a phase map wrapped into π and -π range. Therefore a phase unwrapping algorithm is applied to recover the accurate phase map from the wrapped phase map. In literature the phase unwrapping algorithms have been classified in local and global phase unwrapping algorithms.

These phase unwrapping algorithms have been reviewed in this chapter. Local phase unwrapping algorithms are reasonably swift and had low computational requirements although it implies a decreasing in the quality of three-dimensional reconstruction. Conversely the global phase unwrapping algorithms are high time-consuming and had elevated computational requirements and commonly perform a superior quality of 3-D reconstruction.

This methodology could be widely used to digitize diverse objects for reverse engineering, virtual reality, 3D navigation, and so on.

Notice that the method can reconstruct only the part of the object that can be seen by the camera, if a full 3D reconstruction (360 degrees) is needed, a rotating table is can be used and the methodology will be applied n times, where n is the rotation angle of the table.

One big challenge is to obtain the 3D reconstruction in real time. As a part of the solution, an optical filter to obtain the FFT directly can be used. Moreover, the algorithm can be implemented into a FPGA to carry out a parallel processing and minimize the processing time. Some other tools include the testing of the algorithm performance and doing a comparison with a wavelet or neural networks.

6. References

Berryman, F; Pynsent, P.; Cubillo, J.; A theoretical Comparison of three fringe analysis methods for determining the three-dimensional shape of an object in the presence of noise, Optics and Lasers in Engineering,Vol. 39, pp. 35-50, 2003, ISSN 0143-8166 4

Giglia, D.C.; Prittm M.D., *Two dimensional phase unwrapping theory, algorithm and software*, Edit. Wiley, New York, 1998, ISBN 9780471249351.

Gokstorp, M.(1995). *Depth Computation in Robot Vision*, Ph.D. Thesis, Department of Electrical Engineering, Linkoping University, Linkoping, Sweden.

Ichioka, Y.; Inuiya, M.; Direct Phase Detecting System, *Applied Optics*, Vol. 11, Issue 7, pp. 1507-1514 , 1972, ISSN 1559-128X.

Itoh, K.; Analysis of the phase unwrapping algorithm, *Applied Optics*, 21(14): 2470-2486, 1982, ISSN 1559-128X.

Kian, Q.; Soon, S.H.; Asundi, A.; A simple phase unwrapping approach based on filtering by windowed Fourier transform, *Optics & Laser Technology*, Vol. 37, Issue 6, September 2005, pp. 458-462, ISSN 0030-3992.

Kian, Q.; Two dimensional windowed Fourier transform for fringe pattern analysis: Principles, applications and implementations, *Optics & Laser in Engineering*, Vol. 45, pp. 304-317, 2007, ISSN 0143-8166.

Malacara, D.; *Optical Shop Testing*, D Malacara, Ed. Wiley, New York, 2006.

Pedraza, J.C.; Rodriguez W.; Barriga, L.; Ramos, J.; Gorrostieta, E.; Rivas, A.; Image Processing for 3D Reconstruction using a Modified Fourier Transform Profilometry Method, *Lecture Notes on Artificial Intelligence (Advances in Artificial Intelligence)*, pp. 705–712, Springer-Verlag Berlin Heidelberg, 2007, ISSN 0302-9743.

Pedraza, J.C.; Canchola S.L.; Gorrostieta E.; Aceves M.A.; Ramos, J.M.; Delgado M.: Three-Dimensional Reconstruction System based on a Segmentation Algorithms and a Modified Fourier Transform Profilometry, *Proceedings of the IEEE Electronics, Robotics and Automotive Mechanics Conference CERMA 2009*, Morelos Mexico, 2009., ISBN 9780769537993.

Pedraza, J.C.; Gorrostieta E.; Ramos, J.M.; Canchola S.L.; Aceves M.A.; Delgado M.; Rico, R.A., A Profilometric Approach for 3D Reconstruction Using Fourier and Wavelet Transforms, *Lecture Notes on Artificial Intelligence (Advances in Artificial Intelligence)*, pp. 313-323, 2009, ISSN 0302-9743.

Pedraza, J.C.; Gorrostieta E.; Delgado M.; Canchola, S.L.; Ramos, J.M.; Aceves, M.A., Sotomayor, A.; A 3D Sensor Based on a Profilometric Approach, *Sensors Journal (www.mdpi.com/journal/sensors)*, pp. 10326-10340, ISSN 1424-8220,2009. 11Pramod, K; Digital Speckle Pattern Interferometry and related Techniques. Edit. Wiley, 2001, ISBN 9780471490524 5.

Sotomayor A.; Pedraza, J.C; Aceves, M.A.; Gorrostieta E.; Canchola, S.L.; Ramos, J.M.; Qintanar, M.E.; (2010). A Comparison between Local and Global Phase Unwrapping Algorithms in a Modified Fourier Transform Profilometry Method, *Proceedings of the IEEE, 20th International Conference on Electronics, communications and computers CONIELECOMP 2010*, pp.301-306, ISBN 978-1-4244-5353-5-10, Cholula Puebla Mexico, February 2010. IEEE, Puebla.

Takeda, M; Ina, H. Kobayashi, S; Fourier-Transform method of fringe pattern analysis for computed-based topography and interferometry. J.Opt. Soc.Am. Vol. 72, No.1, pp. 156-160, January 1982, ISSN 1084-7529. 3

Wu, L.; Research and development of fringe projection-based methods in 3D shape reconstruction, *Journal of Zhejiang University SCIENCE A*, pp. 1026-1036, 2006, ISSN 1673-565X 7

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

7

Quadratic Discrete Fourier Transform and Mutually Unbiased Bases

Maurice R. Kibler
Université de Lyon, Université Claude Bernard et CNRS (IPNL et IN2P3)
France

1. Introduction

The use of the discrete Fourier transform (DFT) is quite spread in many fields of physical sciences and engineering as for instance in signal theory. This chapter deals with a quadratic extension of the DFT and its application to quantum information.

From a very general point of view, the DFT can be defined as follows. Let us denote (x_0, x_1, ..., x_{d-1}) a collection of d complex numbers. The transformation

$$x \equiv (x_0, x_1, \dots, x_{d-1}) \mapsto \tilde{x} \equiv (\tilde{x}_0, \tilde{x}_1, \dots, \tilde{x}_{d-1}) \tag{1}$$

defined by

$$\tilde{x}_\alpha = \frac{1}{\sqrt{d}} \sum_{n=0}^{d-1} e^{i2\pi\alpha n/d} x_n, \quad \alpha = 0, 1, \dots, d-1 \tag{2}$$

will be referred to as the DFT of x.

Equation (2) can be transcribed in finite quantum mechanics. In that case, x is often replaced by an orthonormal basis $\{|n\rangle : n = 0, 1, \dots, d-1\}$ of the Hilbert space $\mathbb{C}^d$ (with an inner product noted $\langle\,|\,\rangle$ in Dirac notations). The analog of (2) reads

$$|\tilde{\alpha}\rangle = \frac{1}{\sqrt{d}} \sum_{n=0}^{d-1} e^{i2\pi\alpha n/d} |n\rangle, \quad \alpha = 0, 1, \dots, d-1 \tag{3}$$

Equation (3) makes it possible to pass from the orthonormal basis $\{|n\rangle : n = 0, 1, \dots, d-1\}$ to another orthonormal basis $\{|\tilde{\alpha}\rangle : \alpha = 0, 1, \dots, d-1\}$ and vice versa since

$$\langle n|n'\rangle = \delta(n, n') \Leftrightarrow \langle\tilde{\alpha}|\tilde{\alpha'}\rangle = \delta(\alpha, \alpha') \tag{4}$$

The transformation (3) defines a quantum DFT. In the last twenty years, the notion of quantum DFT has received a considerable attention in connection with finite quantum mechanics and quantum information (Vourdas, 2004).

As an interesting property, the two bases $\{|n\rangle : n = 0, 1, \dots, d-1\}$ and $\{|\tilde{\alpha}\rangle : \alpha = 0, 1, \dots, d-1\}$, connected via a quantum DFT, constitute a couple of unbiased bases. Let us recall that two distinct orthonormal bases

$$B_a = \{|a\alpha\rangle : \alpha = 0, 1, \dots, d-1\} \tag{5}$$

and

$$B_b = \{|b\beta\rangle : \beta = 0, 1, \ldots, d-1\} \tag{6}$$

of the space $\mathbb{C}^d$ are said to be unbiased if and only if

$$\forall \alpha = 0, 1, \ldots, d-1, \ \forall \beta = 0, 1, \ldots, d-1 \ : \ |\langle a\alpha | b\beta \rangle| = \frac{1}{\sqrt{d}} \tag{7}$$

The unbiasedness character of the bases $\{|n\rangle : n = 0, 1, \ldots, d-1\}$ and $\{|\tilde{\alpha}\rangle : \alpha = 0, 1, \ldots, d-1\}$ then follows from

$$\langle n|\tilde{\alpha}\rangle = \frac{1}{\sqrt{d}} e^{i2\pi\alpha n/d} \ \Rightarrow \ |\langle n|\tilde{\alpha}\rangle| = \frac{1}{\sqrt{d}} \tag{8}$$

which is evident from (3).

The determination of sets of mutually unbiased bases (MUBs) in $\mathbb{C}^d$ is of paramount importance in the theory of information and in quantum mechanics. Such bases are useful in classical information (Calderbank et al., 1997), quantum information (Cerf et al., 2002) as well as for the construction of discrete Wigner functions (Gibbons et al., 2004), the solution of the mean King problem (Englert & Aharonov, 2001) and the understanding of the Feynman path integral formalism (Tolar & Chadzitaskos, 2009). It is well known that the number N_{MUB} of MUBs in $\mathbb{C}^d$ is such that $3 \leq N_{MUB} \leq d+1$ (Durt et al., 2010). Furthermore, the maximum number $N_{MUB} = d+1$ is reached when d is a prime number or a power of a prime number (Calderbank et al., 1997; Ivanović, 1981; Wootters & Fields, 1989). However, when d is not a prime number or more generally a power of a prime number, it is not known if the limiting value $N_{MUB} = d+1$ is attained. In this respect, in the case $d = 6$, in spite of an enormous number of works it was not possible to find more than three MUBs (see for example (Bengtsson et al., 2007; Brierley & Weigert, 2009; Grassl, 2005)).

The main aim of this chapter is to introduce and discuss a generalization of the DFTs defined by (2) and (3) in order to produce other couples of MUBs. The generalization will be achieved by introducing quadratic terms in the exponentials in (2) and (3) through the replacement of the linear term αn by a quadratic term $\xi n^2 + \eta n + \zeta$ with ξ, η and ζ in $\mathbb{R}$. The resulting generalized DFT will be referred to as a quadratic DFT.

The material presented in this chapter is organized in the following way. Section 2 is devoted to the study of those aspects of the representation theory of the group $SU(2)$ in a nonstandard basis which are of relevance for the introduction of the quadratic DFT. The quadratic DFT is studied in section 3. Some applications of the quadratic DFT to quantum information are given in section 4.

Most of the notations in this chapter are standard. Some specific notations shall be introduced when necessary. As usual, $\delta_{a,b}$ stands for the Kronecker delta symbol of a and b, i for the pure imaginary, $\bar{z}$ for the complex conjugate of the number z, $A^\dagger$ for the adjoint of the operator A, and I for the identity operator. We use $[A, B]_q$ to denote the q-commutator $AB - qAB$ of the operators A and B; the commutator $[A, B]_{+1}$ and anticommutator $[A, B]_{-1}$ are noted simply $[A, B]$ and $\{A, B\}$, respectively, as is usual in quantum mechanics. Boldface letters are reserved for squared matrices ($\mathbf{I_d}$ is the d-dimensional identity matrix). We employ a notation of type $|\psi\rangle$, or sometimes $|\psi)$, for a vector in an Hilbert space and we denote $\langle\phi|\psi\rangle$ and $|\phi\rangle\langle\psi|$ respectively the inner and outer products of the vectors $|\psi\rangle$ and $|\phi\rangle$. The symbols $\oplus$ and $\ominus$ refer to the addition and subtraction modulo d or $2j+1$ (with $d = 2j+1 = 2, 3, 4, \ldots$

depending on the context) while the symbol $\otimes$ serves to denote the tensor product of two vectors or of two spaces. Finally $\mathbb{N}$, $\mathbb{N}^*$ and $\mathbb{Z}$ are the sets of integers, strictly positive integers and relative integers; $\mathbb{R}$ and $\mathbb{C}$ the real and complex fields; and $\mathbb{Z}/d\mathbb{Z}$ the ring of integers $0, 1, \ldots, d-1$ modulo d.

2. A nonstandard approach to $SU(2)$

2.1 Quon algebra

The idea of a quon takes its origin in the replacement of the commutation (sign $-$) and anticommutation (sign $+$) relations

$$a_- a_+ \pm a_+ a_- = 1 \tag{9}$$

of quantum mechanics by the relation

$$a_- a_+ - q a_+ a_- = f(N) \tag{10}$$

where q is a constant and $f(N)$ an arbitrary fonction of a number operator N. The introduction of q and $f(N)$ yields the possibility to replace the harmonic oscillator algebra by a deformed oscillator algebra. For $f(N) = I$, the case $q = -1$ corresponds to fermion operators (describing a fermionic oscillator) and the case $q = +1$ to boson operators (describing a bosonic oscillator). The other possibilities for q and $f(N) = I$ correspond to quon operators. We shall be concerned here with a quon algebra or q-deformed oscillator algebra for q a root of unity.

Definition 1. *The three linear operators a_-, a_+ and N_a such that*

$$[a_-, a_+]_q = I, \quad [N_a, a_+] = a_+, \quad [N_a, a_-] = -a_-, \quad (a_+)^k = (a_-)^k = 0, \quad (N_a)^\dagger = N_a \tag{11}$$

where

$$q = \exp\left(\frac{2\pi i}{k}\right), \quad k \in \mathbb{N} \setminus \{0, 1\} \tag{12}$$

define a quon algebra or q-deformed oscillator algebra denoted $A_q(a_-, a_+, N_a)$ or simply $A_q(a)$. The operators a_- and a_+ are referred to as quon operators. The operators a_-, a_+ and N_a are called annihilation, creation and number operators, respectively.

Definition 1 differs from the one by Arik and Coon (Arik & Coon, 1976) in the sense that we take q as a primitive kth root of unity instead of $0 < q < 1$. In Eq. (12), the value $k = 0$ is excluded since it would lead to a non-defined value of q. The case $k = 1$ must be excluded too since it would yield trivial algebras with $a_- = a_+ = 0$. We observe that for $k = 2$ (i.e., for $q = -1$), the algebra $A_{-1}(a)$ corresponds to the ordinary fermionic algebra and the quon operators coincide with the fermion operators. On the other hand, we note that in the limiting situation where $k \to \infty$ (i.e., for $q = 1$), the algebra $A_1(a)$ is nothing but the ordinary bosonic algebra and the quon operators are boson operators. For k arbitrary, N_a is generally different from a_+a_-; it is only for $k = 2$ and $k \to \infty$ that $N_a = a_+a_-$. Note that the nilpotency relations $(a_+)^k = (a_-)^k = 0$, with k finite, are at the origin of k-dimensional representations of $A_q(a)$ (see section 2.2).

For arbitrary k, the quon operators a_- and a_+ are not connected via Hermitian conjugation. It is only for $k = 2$ or $k \to \infty$ that we may take $a_+ = (a_-)^\dagger$. In general (i.e., for $k \neq 2$ or $k \not\to \infty$), we have $(a_\pm)^\dagger \neq a_\mp$. Therefore, it is natural to consider the so-called k-fermionic algebra Σ_q with the generators a_-, a_+, $a_+^+ = (a_+)^\dagger$, $a_-^+ = (a_-)^\dagger$ and N_a (Daoud et al., 1998). The defining relations for Σ_q correspond to the ones of $A_q(a_-, a_+, N_a)$ and $A_{\bar{q}}(a_+^+, a_-^+, N_a)$ complemented by the relations

$$a_- a_+^+ - q^{-\frac{1}{2}} a_+^+ a_- = 0, \quad a_+ a_-^+ - q^{\frac{1}{2}} a_-^+ a_+ = 0 \tag{13}$$

Observe that for $k = 2$ or $k \to \infty$, the latter relation corresponds to an identity. The operators a_-, a_+, a_+^+ and a_-^+ are called k-fermion operators and we also use the terminology k-fermions in analogy with fermions and bosons. They clearly interpolate between fermions and bosons. In passing, let us mention that the k-fermions introduced in (Daoud et al., 1998) share some common properties with the parafermions of order $k-1$ discussed in (Beckers & Debergh, 1990; Durand, 1993; Khare, 1993; Klishevich & Plyushchay, 1999; Rubakov & Spiridonov, 1988). The k-fermions can be used for constructing a fractional supersymmetric algebra of order k (or parafermionic algebra of order $k-1$). The reader may consult (Daoud et al., 1998) for a study of the k-fermionic algebra Σ_q and its application to supersymmetry.

2.2 Quon realization of $su(2)$

Going back to quons, let us show how the Lie algebra $su(2)$ of the group $SU(2)$ can be generated from two quon algebras. We start with two commuting quon algebras $A_q(a)$ with $a = x, y$ corresponding to the same value of the deformation parameter q. Their generators satisfy Eqs. (11) and (12) with $a = x, y$ and $[X, Y] = 0$ for any X in $A_q(x)$ and any Y in $A_q(y)$. Then, let us look for Hilbertian representations of $A_q(x)$ and $A_q(y)$ on k-dimensional Hilbert spaces $\mathcal{F}_x$ and $\mathcal{F}_y$ spanned by the bases $\{|n_1) : n_1 = 0, 1, \ldots, k-1\}$ and $\{|n_2) : n_2 = 0, 1, \ldots, k-1\}$, respectively. These two bases are supposed to be orthonormal, i.e.,

$$(n_1|n_1') = \delta(n_1, n_1'), \quad (n_2|n_2') = \delta(n_2, n_2') \tag{14}$$

We easily verify the following result.
Proposition 1. *The relations*

$$\begin{aligned} x_+|n_1) &= |n_1 + 1), \quad x_+|k-1) = 0 \\ x_-|n_1) &= [n_1]_q \, |n_1 - 1), \quad x_-|0) = 0 \\ N_x|n_1) &= n_1|n_1) \end{aligned} \tag{15}$$

and

$$\begin{aligned} y_+|n_2) &= [n_2 + 1]_q \, |n_2 + 1), \quad y_+|k-1) = 0 \\ y_-|n_2) &= |n_2 - 1), \quad y_-|0) = 0 \\ N_y|n_2) &= n_2|n_2) \end{aligned} \tag{16}$$

define k-dimensional representations of $A_q(x)$ and $A_q(y)$, respectively. In (15) and (16), we use the notation

$$\forall n \in \mathbb{N}^* \ : \ [n]_q = \frac{1 - q^n}{1 - q} = 1 + q + \ldots + q^{n-1}, \quad [0]_q = 1 \tag{17}$$

which is familiar in q-deformations of algebraic structures.
Definition 2. *The cornerstone of the quonic approach to su(2) is to define the two linear operators*

$$h = \sqrt{N_x \left(N_y + 1\right)}, \quad v_{ra} = s_x s_y \tag{18}$$

with

$$s_x = q^{a(N_x+N_y)/2} x_+ + e^{i\phi_r/2} \frac{1}{[k-1]_q!} (x_-)^{k-1} \tag{19}$$

$$s_y = y_- q^{-a(N_x-N_y)/2} + e^{i\phi_r/2} \frac{1}{[k-1]_q!} (y_+)^{k-1} \tag{20}$$

In (19) and (20), we take

$$a \in \mathbb{Z}/d\mathbb{Z}, \quad \phi_r = \pi(k-1)r, \quad r \in \mathbb{R} \tag{21}$$

and the q-deformed factorials are defined by

$$\forall n \in \mathbb{N}^* \ : \ [n]_q! = [1]_q [2]_q \dots [n]_q, \quad [0]_q! = 1 \tag{22}$$

Note that the parameter a might be taken as real. We limit ourselves to a in $\mathbb{Z}/d\mathbb{Z}$ in view of the applications to MUBs.

The operators h and v_{ra} act on the states

$$|n_1, n_2) = |n_1) \otimes |n_2) \tag{23}$$

of the k^2-dimensional Fock space $\mathcal{F}_x \otimes \mathcal{F}_y$. It is straightforward to verify that the action of v_{ra} on $\mathcal{F}_x \otimes \mathcal{F}_y$ is governed by

$$\begin{aligned} v_{ra}|k-1, n_2) &= e^{i\phi_r/2}|0, n_2 - 1), \quad n_2 \neq 0 \\ v_{ra}|n_1, n_2) &= q^{n_2 a}|n_1 + 1, n_2 - 1), \quad n_1 \neq k-1, \quad n_2 \neq 0 \\ v_{ra}|n_1, 0) &= e^{i\phi_r/2}|n_1 + 1, k-1), \quad n_1 \neq k-1 \end{aligned} \tag{24}$$

and

$$v_{ra}|k-1, 0) = e^{i\phi_r}|0, k-1) \tag{25}$$

As a consequence, we can prove the identity

$$(v_{ra})^k = e^{i\phi_r} I \tag{26}$$

The action of h on $\mathcal{F}_x \otimes \mathcal{F}_y$ is much simpler. It is described by

$$h|n_1, n_2) = \sqrt{n_1(n_2+1)}|n_1, n_2) \tag{27}$$

which holds for $n_1, n_2 = 0, 1, \dots, k-1$. Finally, the operator v_{ra} is unitary and the operator h Hermitian on the space $\mathcal{F}_x \otimes \mathcal{F}_y$.

We are now in a position to introduce a realization of the generators of the non-deformed Lie algebra $su(2)$ in terms of the operators v_{ra} and h. As a preliminary step, let us adapt the trick used by Schwinger in his approach to angular momentum via a coupled pair of harmonic oscillators (Schwinger, 1965). This can be done by introducing two new quantum numbers J and M

$$J = \frac{1}{2}(n_1 + n_2), \quad M = \frac{1}{2}(n_1 - n_2) \tag{28}$$

and the state vectors

$$|J, M\rangle = |n_1, n_2) = |J+M, J-M) \quad \Rightarrow \quad \langle J, M|J', M'\rangle = \delta_{J,J'}\delta_{M,M'} \tag{29}$$

Note that

$$j = \frac{1}{2}(k-1) \tag{30}$$

is an admissible value for J. We may thus have $j = \frac{1}{2}, 1, \frac{3}{2}, \ldots$ (since $k = 2, 3, 4, \ldots$). For the value j of J, the quantum number M can take the values $m = j, j-1, \ldots, -j$. Then, let us consider the $(2j+1)$-dimensional subspace $\epsilon(j)$ of the k^2-dimensional space $\mathcal{F}_x \otimes \mathcal{F}_y$ spanned by the basis

$$B_{2j+1} = \{|j, m\rangle : m = j, j-1, \ldots, -j\} \tag{31}$$

with the orthonormality property

$$\langle j, m|j, m'\rangle = \delta_{m,m'} \tag{32}$$

We guess that $\epsilon(j)$ is a space of constant angular momentum j. As a matter of fact, we can check that $\epsilon(j)$ is stable under h and v_{ra}.

Proposition 2. *The action of the operators h and v_{ra} on $\epsilon(j)$ is given by*

$$\begin{aligned} h|j, m\rangle &= \sqrt{(j+m)(j-m+1)}|j, m\rangle & (33)\\ v_{ra}|j, m\rangle &= \delta_{m,j}e^{i2\pi jr}|j, -j\rangle + (1-\delta_{m,j})q^{(j-m)a}|j, m+1\rangle & (34) \end{aligned}$$

where q is given by (12) with $k = 2j+1$, $r \in \mathbb{R}$ and $a \in \mathbb{Z}/(2j+1)\mathbb{Z}$.

It is sometimes useful to use the Dirac notation by writing

$$\begin{aligned} h &= \sum_{m=-j}^{j} \sqrt{(j+m)(j-m+1)}|j, m\rangle\langle j, m| & (35)\\ v_{ra} &= e^{i2\pi jr}|j, -j\rangle\langle j, j| + \sum_{m=-j}^{j-1} q^{(j-m)a}|j, m+1\rangle\langle j, m| & (36)\\ (v_{ra})^{\dagger} &= e^{-i2\pi jr}|j, j\rangle\langle j, -j| + \sum_{m=-j+1}^{j} q^{-(j-m+1)a}|j, m-1\rangle\langle j, m| & (37) \end{aligned}$$

It is understood that the three preceding relations are valid as far as the operators h, v_{ra} and $(v_{ra})^{\dagger}$ act on the space $\epsilon(j)$. It is evident that h is an Hermitian operator and v_{ra} a unitary operator on $\epsilon(j)$.

Definition 3. *The link with $su(2)$ can be established by introducing the three linear operators j_+, j_- and j_z through*

$$j_+ = hv_{ra}, \quad j_- = (v_{ra})^{\dagger} h, \quad j_z = \frac{1}{2}\left[h^2 - (v_{ra})^{\dagger} h^2 v_{ra}\right] \tag{38}$$

For each couple (r, a) we have a triplet (j_+, j_-, j_z). It is clear that j_+ and j_- are connected via Hermitian conjugation and j_z is Hermitian.

Proposition 3. *The action of j_+, j_- and j_z on $\epsilon(j)$ is given by the eigenvalue equation*

$$j_z|j,m\rangle = m|j,m\rangle \tag{39}$$

and the ladder equations

$$j_+|j,m\rangle = q^{(j-m+s-1/2)a}\sqrt{(j-m)(j+m+1)}|j,m+1\rangle \tag{40}$$

$$j_-|j,m\rangle = q^{-(j-m+s+1/2)a}\sqrt{(j+m)(j-m+1)}|j,m-1\rangle \tag{41}$$

where $s = 1/2$.

For $a = 0$, Eqs. (39), (40) and (41) give relations that are well known in angular momentum theory. Indeed, the case $a = 0$ corresponds to the usual Condon and Shortley phase convention used in atomic and nuclear spectroscopy. As a corollary of Proposition 3, we have the following result.

Corollary 1. *The operators j_+, j_- and j_z satisfy the commutation relations*

$$[j_z, j_+] = j_+, \quad [j_z, j_-] = -j_-, \quad [j_+, j_-] = 2j_z \tag{42}$$

and thus span the Lie algebra of $SU(2)$.

The latter result does not depend on the parameters r and a. The writing of the ladder operators j_+ and j_- in terms of h and v_{ra} constitutes a two-parameter polar decomposition of the Lie algebra $su(2)$. Thus, from two q-deformed oscillator algebras we obtained a polar decomposition of the non-deformed Lie algebra of $SU(2)$. This decomposition is an alternative to the polar decompositions obtained independently in (Chaichian & Ellinas, 1990; Lévy-Leblond, 1973; Vourdas, 1990).

2.3 The $\{j^2, v_{ra}\}$ scheme

Each vector $|j,m\rangle$ is a common eigenvector of the two commuting operators j_z and

$$j^2 = \frac{1}{2}(j_+j_- + j_-j_+) + j_3^2 = j_+j_- + j_3(j_3-1) = j_-j_+ + j_3(j_3+1) \tag{43}$$

which is known as the Casimir operator of $su(2)$ in group theory or as the square of a generalized angular momentum in angular momentum theory. More precisely, we have the eigenvalue equations

$$j^2|j,m\rangle = j(j+1)|j,m\rangle, \quad j_z|j,m\rangle = m|j,m\rangle, \quad m = j, j-1, \ldots, -j \tag{44}$$

which show that j and m can be interpreted as angular momentum quantum numbers (in units such that the rationalized Planck constant $\hbar$ is equal to 1). Of course, the set $\{j^2, j_z\}$ is a complete set of commuting operators. It is clear that the two operators j^2 and v_{ra} commute. As a matter of fact, the set $\{j^2, v_{ra}\}$ provides an alternative to the set $\{j^2, j_z\}$ as indicated by the next result.

Theorem 1. *For fixed j (with $2j \in \mathbb{N}^*$), r (with $r \in \mathbb{R}$) and a (with $a \in \mathbb{Z}/(2j+1)\mathbb{Z}$), the $2j+1$ common eigenvectors of the operators j^2 and v_{ra} can be taken in the form*

$$|j\alpha; ra\rangle = \frac{1}{\sqrt{2j+1}} \sum_{m=-j}^{j} q^{(j+m)(j-m+1)a/2 - jmr + (j+m)\alpha}|j,m\rangle, \quad \alpha = 0, 1, \ldots, 2j \tag{45}$$

where

$$q = \exp\left(\frac{2\pi i}{2j+1}\right) \tag{46}$$

The corresponding eigenvalues are given by

$$j^2|j\alpha; ra\rangle = j(j+1)|j\alpha; ra\rangle, \quad v_{ra}|j\alpha; ra\rangle = q^{j(r+a)-\alpha}|j\alpha; ra\rangle, \quad \alpha = 0, 1, \ldots, 2j \tag{47}$$

so that the spectrum of v_{ra} is nondegenerate and $\{j^2, v_{ra}\}$ does form a complete set of commuting operators. The inner product

$$\langle j\alpha; ra|j\beta; ra\rangle = \delta_{\alpha,\beta} \tag{48}$$

shows that

$$B_{ra} = \{|j\alpha; ra\rangle : \alpha = 0, 1, \ldots, 2j\} \tag{49}$$

is a nonstandard orthonormal basis for the irreducible matrix representation of $SU(2)$ associated with j. For fixed j, there exists a priori a $(2j+1)$-multiple infinity of orthonormal bases B_{ra} since r can have any real value and a, which belongs to the ring $\mathbb{Z}/(2j+1)\mathbb{Z}$, can take $2j+1$ values ($a = 0, 1, \ldots, 2j$).
Equation (45) defines a unitary transformation that allows to pass from the standard orthonormal basis B_{2j+1}, quite well known in angular momentum theory and group theory, to the nonstandard orthonormal basis B_{ra}. For fixed j, r and a, the inverse transformation of (45) is

$$|j,m\rangle = q^{-(j+m)(j-m+1)a/2+jmr}\frac{1}{\sqrt{2j+1}}\sum_{\alpha=0}^{2j} q^{-(j+m)\alpha}|j\alpha; ra\rangle, \quad m = j, j-1, \ldots, -j \tag{50}$$

which looks like an inverse DFT up to phase factors. For $r = a = 0$, Eqs. (45) and (50) lead to

$$|j\alpha; 00\rangle = \frac{1}{\sqrt{2j+1}}\sum_{m=-j}^{j} q^{(j+m)\alpha}|j,m\rangle, \quad \alpha = 0, 1, \ldots, 2j \tag{51}$$

$$\Leftrightarrow \quad |j,m\rangle = \frac{1}{\sqrt{2j+1}}\sum_{\alpha=0}^{2j} q^{-(j+m)\alpha}|j\alpha; 00\rangle, \quad m = j, j-1, \ldots, -j \tag{52}$$

Equations (51) and (52) correspond (up to phase factors) to the DFT of the basis B_{2j+1} and its inverse DFT, respectively.
Note that the calculation of $\langle j\alpha; ra|j\beta; sb\rangle$ is much more involved for ($r \neq s$, $a = b$), ($r = s$, $a \neq b$) and ($r \neq s$, $a \neq b$) than the one of $\langle j\alpha; ra|j\beta; ra\rangle$ (the value of which is given by (48)). For example, the overlap between the bases B_{ra} and B_{sa}, of relevance for the case ($r \neq s$, $a = b$), is given by

$$\langle j\alpha; ra|j\beta; sa\rangle = \frac{1}{2j+1}\frac{\sin[j(s-r)+\alpha-\beta]\pi}{\sin[j(s-r)+\alpha-\beta]\frac{\pi}{2j+1}} \tag{53}$$

The cases ($r = s$, $a \neq b$) and ($r \neq s$, $a \neq b$) need the use of Gauss sums as we shall see below.
The representation theory and the Wigner-Racah algebra of the group $SU(2)$ can be developed in the $\{j^2, v_{ra}\}$ quantization scheme. This leads to Clebsch-Gordan coefficients and $(3-$

$j\alpha)_{ra}$ symbols with properties very different from the ones of the usual $SU(2) \supset U(1)$ Clebsch-Gordan coefficients and $3-jm$ symbols corresponding to the $\{j^2, j_z\}$ quantization scheme. For more details, see Appendix which deals with the case $r = a = 0$.

The nonstandard approach to the Wigner-Racah algebra of $SU(2)$ and angular momentum theory in the $\{j^2, v_{ra}\}$ scheme is especially useful in quantum chemistry for problems involving cyclic symmetry. This is the case for a ring-shape molecule with $2j+1$ atoms at the vertices of a regular polygon with $2j+1$ sides or for a one-dimensional chain of $2j+1$ spins of $\frac{1}{2}$-value each (Albouy & Kibler, 2007). In this connection, we observe that the vectors of type $|j\alpha; ra\rangle$ are specific symmetry-adapted vectors. Symmetry-adapted vectors are widely used in quantum chemistry, molecular physics and condensed matter physics as for instance in ro-vibrational spectroscopy of molecules (Champion et al., 1977) and ligand-field theory (Kibler, 1968). However, the vectors $|j\alpha; ra\rangle$ differ from the symmetry-adapted vectors considered in (Champion et al., 1977; Kibler, 1968; Patera & Winternitz, 1976) in the sense that v_{ra} is not an invariant under some finite subgroup (of crystallographic interest) of the orthogonal group $O(3)$. This can be clarified as follows.

Proposition 4. *From (36), it follows that the operator v_{ra} is a pseudo-invariant under the cyclic group C_{2j+1}, a subgroup of $SO(3)$, whose elements are the Wigner operators $P_{R(\varphi)}$ associated with the rotations $R(\varphi)$, around the quantization axis Oz, with the angles*

$$\varphi = p\frac{2\pi}{2j+1}, \quad p = 0, 1, \ldots, 2j \tag{54}$$

More precisely, v_{ra} transforms as

$$P_{R(\varphi)} v_{ra} \left(P_{R(\varphi)}\right)^{\dagger} = e^{-i\varphi} v_{ra} \tag{55}$$

Thus, v_{ra} belongs to the irreducible representation class of C_{2j+1} of character vector

$$\chi^{(2j)} = (1, q^{-1}, \ldots, q^{-2j}) \tag{56}$$

In terms of vectors of $\epsilon(j)$, we have

$$P_{R(\varphi)}|j\alpha; ra\rangle = q^{jp}|j\beta; ra\rangle, \quad \beta = \alpha \ominus p \tag{57}$$

so that the set $\{|j\alpha; ra\rangle : \alpha = 0, 1, \ldots, 2j\}$ is stable under $P_{R(\varphi)}$. The latter set spans the regular representation of C_{2j+1}.

2.4 Examples

Example 1: The $j = \frac{1}{2}$ case. The eigenvectors of v_{ra} are

$$|\frac{1}{2}\alpha; ra\rangle = \frac{1}{\sqrt{2}} e^{i\pi(a/2 - r/4 + \alpha)} |\frac{1}{2}, \frac{1}{2}\rangle + \frac{1}{\sqrt{2}} e^{i\pi r/4} |\frac{1}{2}, -\frac{1}{2}\rangle, \quad \alpha = 0, 1 \tag{58}$$

where $r \in \mathbb{R}$ and a can take the values $a = 0, 1$. In the case $r = 0$, Eq. (58) gives the two bases

$$B_{00} \ : \ |\frac{1}{2}0; 00\rangle = \frac{1}{\sqrt{2}}\left(|\frac{1}{2}, \frac{1}{2}\rangle + |\frac{1}{2}, -\frac{1}{2}\rangle\right), \quad |\frac{1}{2}1; 00\rangle = -\frac{1}{\sqrt{2}}\left(|\frac{1}{2}, \frac{1}{2}\rangle - |\frac{1}{2}, -\frac{1}{2}\rangle\right) \tag{59}$$

and

$$B_{01} \ : \ |\frac{1}{2}0; 01\rangle = \frac{i}{\sqrt{2}}\left(|\frac{1}{2}, \frac{1}{2}\rangle - i|\frac{1}{2}, -\frac{1}{2}\rangle\right), \quad |\frac{1}{2}1; 01\rangle = -\frac{i}{\sqrt{2}}\left(|\frac{1}{2}, \frac{1}{2}\rangle + i|\frac{1}{2}, -\frac{1}{2}\rangle\right) \tag{60}$$

The bases (59) and (60) are, up to phase factors, familiar bases in quantum mechanics for $\frac{1}{2}$-spin systems.

Example 2: The $j = 1$ case. The eigenvectors of v_{ra} are

$$|1\alpha; ra\rangle = \frac{1}{\sqrt{3}} q^r \left(q^{a+2\alpha-2r}|1,1\rangle + q^{a+\alpha-r}|1,0\rangle + |1,-1\rangle \right), \quad \alpha = 0,1,2 \tag{61}$$

where $r \in \mathbb{R}$ and a can take the values $a = 0,1,2$. In the case $r = 0$, Eq. (61) gives the three bases

$$\begin{aligned} B_{00} \ : \ |10;00\rangle &= \frac{1}{\sqrt{3}} \left(|1,-1\rangle + |1,0\rangle + |1,1\rangle\right) \\ |11;00\rangle &= \frac{1}{\sqrt{3}} \left(|1,-1\rangle + q|1,0\rangle + q^2|1,1\rangle\right) \\ |12;00\rangle &= \frac{1}{\sqrt{3}} \left(|1,-1\rangle + q^2|1,0\rangle + q|1,1\rangle\right) \end{aligned} \tag{62}$$

$$\begin{aligned} B_{01} \ : \ |10;01\rangle &= \frac{1}{\sqrt{3}} \left(|1,-1\rangle + q|1,0\rangle + q|1,1\rangle\right) \\ |11;01\rangle &= \frac{1}{\sqrt{3}} \left(|1,-1\rangle + q^2|1,0\rangle + |1,1\rangle\right) \\ |12;01\rangle &= \frac{1}{\sqrt{3}} \left(|1,-1\rangle + |1,0\rangle + q^2|1,1\rangle\right) \end{aligned} \tag{63}$$

$$\begin{aligned} B_{02} \ : \ |10;02\rangle &= \frac{1}{\sqrt{3}} \left(|1,-1\rangle + q^2|1,0\rangle + q^2|1,1\rangle\right) \\ |11;02\rangle &= \frac{1}{\sqrt{3}} \left(|1,-1\rangle + |1,0\rangle + q|1,1\rangle\right) \\ |12;02\rangle &= \frac{1}{\sqrt{3}} \left(|1,-1\rangle + q|1,0\rangle + |1,1\rangle\right) \end{aligned} \tag{64}$$

It is worth noting that the vectors of the basis B_{00} exhibit all characters

$$\chi^{(\alpha)} = \left(1, q^{\alpha}, q^{2\alpha}\right), \quad \alpha = 0,1,2 \tag{65}$$

of the three vector representations of C_3. On another hand, the bases B_{01} and B_{02} are connected to projective representations of C_3 because they are described by the pseudo-characters

$$\chi_1^{(\alpha)} = \left(1, q^{1+\alpha}, q^{1-\alpha}\right), \quad \alpha = 0,1,2 \tag{66}$$

and

$$\chi_2^{(\alpha)} = \left(1, q^{2+\alpha}, q^{2-\alpha}\right), \quad \alpha = 0,1,2 \tag{67}$$

respectively.

3. Quadratic discrete Fourier transforms

We discuss in this section two quadratic extensions of the DFT, namely, a quantum quadratic DFT that connects state vectors in a finite-dimensional Hilbert space, of relevance in quantum information, and a quadratic DFT that might be of interest in signal analysis.

3.1 Quantum quadratic discrete Fourier transform

Relations of section 2 concerning $SU(2)$ can be transcribed in a form more adapted to the Fourier transformation formalism and to quantum information. In this respect, let us introduce the change of notations

$$d = 2j+1, \quad n = j+m, \quad |n\rangle = |j, -m\rangle \tag{68}$$

and

$$|a\alpha; r\rangle = |j\alpha; ra\rangle \tag{69}$$

so that (49) becomes

$$B_{ra} = \{|a\alpha; r\rangle : \alpha = 0, 1, \ldots, d-1\} \tag{70}$$

(Note that d coincides with the dimension k of the spaces $\mathcal{F}_x$ and $\mathcal{F}_y$ of section 2.) Then from Eq. (45), we have

$$|a\alpha; r\rangle = q^{(d-1)^2 r/4} \frac{1}{\sqrt{d}} \sum_{n=0}^{d-1} q^{n(d-n)a/2 + n[\alpha - (d-1)r/2]} |d-1-n\rangle, \quad \alpha = 0, 1, \ldots, d-1 \tag{71}$$

or equivalently

$$|a\alpha; r\rangle = q^{(d-1)^2 r/4} \frac{1}{\sqrt{d}} \sum_{n=0}^{d-1} q^{(d-1-n)(n+1)a/2 + (d-1-n)[\alpha - (d-1)r/2]} |n\rangle, \quad \alpha = 0, 1, \ldots, d-1 \tag{72}$$

where

$$q = \exp\left(\frac{2\pi i}{d}\right) \tag{73}$$

The inversion of (71) gives

$$|d-1-n\rangle = q^{-n(d-n)a/2 - (d-1)^2 r/4 + n(d-1)r/2} \frac{1}{\sqrt{d}} \sum_{\alpha=0}^{d-1} q^{-n\alpha} |a\alpha; r\rangle, \quad n = 0, 1, \ldots, d-1 \tag{74}$$

By introducing

$$(\mathbf{F_{ra}})_{n\alpha} = \frac{1}{\sqrt{d}} q^{n(d-n)a/2 + (d-1)^2 r/4 + n[\alpha - (d-1)r/2]}, \quad n, \alpha = 0, 1, \ldots, d-1 \tag{75}$$

equations (71) and (74) can be rewritten as

$$|a\alpha; r\rangle = \sum_{n=0}^{d-1} (\mathbf{F_{ra}})_{n\alpha} \, |d-1-n\rangle, \quad \alpha = 0, 1, \ldots, d-1 \tag{76}$$

and

$$|d-1-n\rangle = \sum_{\alpha=0}^{d-1} \overline{(\mathbf{F_{ra}})_{n\alpha}}\, |a\alpha;r\rangle, \quad n = 0,1,\ldots,d-1 \tag{77}$$

respectively. For $r = a = 0$, Eqs. (76) and (77) yield

$$\begin{aligned} |0\alpha;0\rangle &= \frac{1}{\sqrt{d}} \sum_{n=0}^{d-1} e^{i2\pi\alpha n/d} |d-1-n\rangle, \quad \alpha = 0,1,\ldots,d-1 \\ \Leftrightarrow \quad |d-1-n\rangle &= \frac{1}{\sqrt{d}} \sum_{\alpha=0}^{d-1} e^{-i2\pi n\alpha/d} |0\alpha;0\rangle, \quad n = 0,1,\ldots,d-1 \end{aligned} \tag{78}$$

which corresponds (up to a change of notations) to the DFT described by (3). For $a \neq 0$, Eq. (76) can be considered as a quadratic extension (quadratic in n) of the DFT of the basis $\{|n\rangle : n = 0,1,\ldots,d-1\}$ and Eq. (77) thus appears as the corresponding inverse DFT. This can be summed up by the following definition.

Definition 4. *Let $\mathbf{H_{ra}}$ be the $d \times d$ matrix defined by the matrix elements*

$$(\mathbf{H_{ra}})_{n\alpha} = \frac{1}{\sqrt{d}} q^{(d-1-n)(n+1)a/2+(d-1)^2 r/4+(d-1-n)[\alpha-(d-1)r/2]}, \quad n,\alpha = 0,1,\ldots,d-1 \tag{79}$$

where, for a fixed value of d (with $d \in \mathbb{N} \setminus \{0,1\}$), r and a may have values in $\mathbb{R}$ and $\mathbb{Z}/d\mathbb{Z}$, respectively. In compact form

$$(\mathbf{H_{ra}})_{n\alpha} = \frac{1}{\sqrt{d}} e^{2\pi i\nu/d} \tag{80}$$

with

$$\nu = -\frac{1}{4}(d-1)^2 r + \frac{1}{2}(d-1)a + (d-1)\alpha - \frac{1}{2}[2\alpha + 2a - da - (d-1)r]n - \frac{1}{2}an^2 \tag{81}$$

The expansion

$$|a\alpha;r\rangle = \sum_{n=0}^{d-1} (\mathbf{H_{ra}})_{n\alpha}\, |n\rangle, \quad \alpha = 0,1,\ldots,d-1 \tag{82}$$

defines a quadratic quantum DFT of the orthonormal basis

$$B_d = \{|n\rangle : n = 0,1,\ldots,d-1\} \tag{83}$$

This transformation produces another orthonormal basis, namely, the basis B_{ra} (see Eq. (70)). The inverse transformation

$$|n\rangle = \sum_{\alpha=0}^{d-1} \overline{(\mathbf{H_{ra}})_{n\alpha}}\, |a\alpha;r\rangle, \quad n = 0,1,\ldots,d-1 \tag{84}$$

gives back the basis B_d.

For fixed d, r and a, each of the d vectors $|a\alpha;r\rangle$, with $\alpha = 0, 1, \ldots, d-1$, is a linear combination of the vectors $|0\rangle, |1\rangle, \ldots, |d-1\rangle$. The vector $|a\alpha;r\rangle$ is an eigenvector of the operator

$$v_{ra} = e^{i\pi(d-1)r}|d-1\rangle\langle 0| + \sum_{n=0}^{d-2} q^{(d-1-n)a}|d-2-n\rangle\langle d-1-n| \tag{85}$$

or

$$v_{ra} = e^{i\pi(d-1)r}|d-1\rangle\langle 0| + \sum_{n=1}^{d-1} q^{na}|n-1\rangle\langle n| \tag{86}$$

(cf. Eq. (36)). The operator v_{ra} can be developed as

$$v_{ra} = e^{i\pi(d-1)r}|d-1\rangle\langle 0| + q^{a}|0\rangle\langle 1| + q^{2a}|1\rangle\langle 2| + \ldots + q^{(d-1)a}|d-2\rangle\langle d-1| \tag{87}$$

Then, the action of v_{ra} on the state $|n\rangle$ is described by

$$v_{ra}|n\rangle = \delta_{n,0}e^{i\pi(d-1)r}|d-1\rangle + (1-\delta_{n,0})q^{na}|n-1\rangle \tag{88}$$

(cf. Eq. (34)). Its eigenvalues are given by

$$v_{ra}|a\alpha;r\rangle = q^{(d-1)(r+a)/2-\alpha}|a\alpha;r\rangle, \quad \alpha = 0, 1, \ldots, d-1 \tag{89}$$

(cf. Eq. (47)).

3.1.1 Diagonalization of v_{ra}

Let $\mathbf{V_{ra}}$ be the $d \times d$ unitary matrix that represents the linear operator v_{ra} (given by (87)) on the basis B_d. Explicitly, we have

$$\mathbf{V_{ra}} = \begin{pmatrix} 0 & q^{a} & 0 & \ldots & 0 \\ 0 & 0 & q^{2a} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & q^{(d-1)a} \\ e^{i\pi(d-1)r} & 0 & 0 & \ldots & 0 \end{pmatrix} \tag{90}$$

where the lines and columns are arranged in the order $0, 1, \ldots, d-1$. Note that the nonzero matrix elements of V_{0a} are given by the irreducible character vector

$$\chi^{(a)} = (1, q^{a}, \ldots, q^{(d-1)a}) \tag{91}$$

of the cyclic group C_d.

Proposition 5. *The matrix* $\mathbf{H_{ra}}$ *reduces the endomorphism associated with the operator* v_{ra}. *In other words*

$$(\mathbf{H_{ra}})^{\dagger}\,\mathbf{V_{ra}}\mathbf{H_{ra}} = q^{(d-1)(r+a)/2}\begin{pmatrix} q^{0} & 0 & \ldots & 0 \\ 0 & q^{-1} & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & q^{-(d-1)} \end{pmatrix} \tag{92}$$

in agreement with Eq. (47).

Concerning the matrices in (90) and (92), it is important to note the following convention. According to the tradition in quantum mechanics and quantum information, all the matrices in this chapter are set up with their lines and columns ordered from left to right and from top to bottom in the range $0, 1, \ldots, d-1$. Different conventions were used in some previous works by the author. However, the results previously obtained are equivalent to those of this chapter.

The eigenvectors of the matrix $\mathbf{V_{ra}}$ are

$$\phi(a\alpha; r) = \sum_{n=0}^{d-1} (\mathbf{H_{ra}})_{n\alpha}\, \phi_n, \quad \alpha = 0, 1, \ldots, d-1 \tag{93}$$

where the ϕ_n with $n = 0, 1, \ldots, d-1$ are the column vectors

$$\phi_0 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \phi_1 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots, \quad \phi_{d-1} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \tag{94}$$

representing the state vectors $|0\rangle, |1\rangle, \ldots, |d-1\rangle$, respectively. These eigenvectors are the column vectors of the matrix $\mathbf{H_{ra}}$. They satisfy the eigenvalue equation (cf. 89)

$$\mathbf{V_{ra}}\phi(a\alpha; r) = q^{(d-1)(r+a)/2-\alpha}\phi(a\alpha; r) \tag{95}$$

with $\alpha = 0, 1, \ldots, d-1$.

3.1.2 Examples

Example 3: The $d = 2$ case. For $d = 2$, there are two families of bases B_{ra}: the B_{r0} family and the B_{r1} family (a can take the values $a = 0$ and $a = 1$). In terms of matrices, we have

$$\mathbf{H_{ra}} = \frac{1}{\sqrt{2}} \begin{pmatrix} q^{a/2-r/4} & -q^{a/2-r/4} \\ q^{r/4} & q^{r/4} \end{pmatrix}, \quad \mathbf{V_{ra}} = \begin{pmatrix} 0 & q^a \\ q^r & 0 \end{pmatrix}, \quad q = e^{i\pi} \tag{96}$$

The matrix $\mathbf{V_{ra}}$ has the eigenvectors (corresponding to the basis B_{ra})

$$\phi(a\alpha; r) = \frac{1}{\sqrt{2}}(q^{a/2-r/4+\alpha}\phi_0 + q^{r/4}\phi_1), \quad \alpha = 0, 1 \tag{97}$$

where

$$\phi_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \phi_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{98}$$

For $r = 0$, we have

$$V_{00} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad V_{01} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{99}$$

the eigenvectors of which are (cf. (97))

$$\phi(00; 0) = \frac{1}{\sqrt{2}}(\phi_1 + \phi_0) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \phi(01; 0) = \frac{1}{\sqrt{2}}(\phi_1 - \phi_0) = -\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{100}$$

and

$$\phi(10;0) = \frac{1}{\sqrt{2}}(\phi_1 + i\phi_0) = \frac{i}{\sqrt{2}}\begin{pmatrix}1\\-i\end{pmatrix}, \quad \phi(11;0) = \frac{1}{\sqrt{2}}(\phi_1 - i\phi_0) = -\frac{i}{\sqrt{2}}\begin{pmatrix}1\\i\end{pmatrix} \tag{101}$$

which correspond to the bases B_{00} and B_{01}, respectively. Note that (100) and (101) are, up to unimportant multiplicative phase factors, qudits used in quantum information.

Example 4: The $d = 3$ case. For $d = 3$, we have three families of bases, that is to say B_{r0}, B_{r1} and B_{r2}, since a can be 0, 1 and 2. In this case

$$\mathbf{H_{ra}} = \frac{1}{\sqrt{3}}\begin{pmatrix}q^{a-r} & q^{a+2-r} & q^{a+1-r}\\ q^{a} & q^{a+1} & q^{a+2}\\ q^{r} & q^{r} & q^{r}\end{pmatrix}, \quad \mathbf{V_{ra}} = \begin{pmatrix}0 & q^{a} & 0\\ 0 & 0 & q^{2a}\\ q^{3r} & 0 & 0\end{pmatrix}, \quad q = e^{i2\pi/3} \tag{102}$$

and $\mathbf{V_{ra}}$ admits the eigenvectors (corresponding to the basis B_{ra})

$$\phi(a\alpha;r) = \frac{1}{\sqrt{3}}q^{r}\left(q^{a+2\alpha-2r}\phi_0 + q^{a+\alpha-r}\phi_1 + \phi_2\right), \quad \alpha = 0,1,2 \tag{103}$$

where

$$\phi_0 = \begin{pmatrix}1\\0\\0\end{pmatrix}, \quad \phi_1 = \begin{pmatrix}0\\1\\0\end{pmatrix}, \quad \phi_2 = \begin{pmatrix}0\\0\\1\end{pmatrix} \tag{104}$$

In the case $r = 0$, we get

$$V_{00} = \begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 0 & 0\end{pmatrix}, \quad V_{01} = \begin{pmatrix}0 & q & 0\\ 0 & 0 & q^2\\ 1 & 0 & 0\end{pmatrix}, \quad V_{02} = \begin{pmatrix}0 & q^2 & 0\\ 0 & 0 & q\\ 1 & 0 & 0\end{pmatrix} \tag{105}$$

The eigenvectors of V_{00}, V_{01} and V_{02} follow from Eq. (103). This yields

$$\begin{aligned}\phi(00;0) &= \frac{1}{\sqrt{3}}(\phi_2 + \phi_1 + \phi_0)\\ \phi(01;0) &= \frac{1}{\sqrt{3}}\left(\phi_2 + q\phi_1 + q^2\phi_0\right)\\ \phi(02;0) &= \frac{1}{\sqrt{3}}\left(\phi_2 + q^2\phi_1 + q\phi_0\right)\end{aligned} \tag{106}$$

or

$$\phi(00;0) = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}, \quad \phi(01;0) = \frac{1}{\sqrt{3}}\begin{pmatrix}q^2\\q\\1\end{pmatrix}, \quad \phi(02;0) = \frac{1}{\sqrt{3}}\begin{pmatrix}q\\q^2\\1\end{pmatrix} \tag{107}$$

corresponding to B_{00},

$$\begin{aligned}\phi(10;0) &= \frac{1}{\sqrt{3}}(\phi_2 + q\phi_1 + q\phi_0)\\ \phi(11;0) &= \frac{1}{\sqrt{3}}\left(\phi_2 + q^2\phi_1 + \phi_0\right)\\ \phi(12;0) &= \frac{1}{\sqrt{3}}\left(\phi_2 + \phi_1 + q^2\phi_0\right)\end{aligned} \tag{108}$$

or

$$\phi(10;0) = \frac{1}{\sqrt{3}} \begin{pmatrix} q \\ q \\ 1 \end{pmatrix}, \quad \phi(11;0) = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ q^2 \\ 1 \end{pmatrix}, \quad \phi(12;0) = \frac{1}{\sqrt{3}} \begin{pmatrix} q^2 \\ 1 \\ 1 \end{pmatrix} \tag{109}$$

corresponding to B_{01}, and

$$\begin{aligned} \phi(20;0) &= \frac{1}{\sqrt{3}} \left(\phi_2 + q^2\phi_1 + q^2\phi_0\right) \\ \phi(21;0) &= \frac{1}{\sqrt{3}} (\phi_2 + \phi_1 + q\phi_0) \\ \phi(22;0) &= \frac{1}{\sqrt{3}} (\phi_2 + q\phi_1 + \phi_0) \end{aligned} \tag{110}$$

or

$$\phi(20;0) = \frac{1}{\sqrt{3}} \begin{pmatrix} q^2 \\ q^2 \\ 1 \end{pmatrix}, \quad \phi(21;0) = \frac{1}{\sqrt{3}} \begin{pmatrix} q \\ 1 \\ 1 \end{pmatrix}, \quad \phi(22;0) = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ q \\ 1 \end{pmatrix} \tag{111}$$

corresponding to B_{02}. Note that (107), (109) and (111) are, up to unimportant multiplicative phase factors, qutrits used in quantum information.

3.1.3 Decomposition of V_{ra}

The matrix $\mathbf{V_{ra}}$ can be decomposed as

$$\mathbf{V_{ra}} = \mathbf{P_r X Z}^a \tag{112}$$

where

$$\mathbf{P_r} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & e^{i\pi(d-1)r} \end{pmatrix} \tag{113}$$

and

$$\mathbf{X} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & q & 0 & \dots & 0 \\ 0 & 0 & q^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & q^{d-1} \end{pmatrix} \tag{114}$$

The matrices $\mathbf{X}$ and $\mathbf{Z}$ can be derived from particular $\mathbf{V_{ra}}$ matrices since

$$\mathbf{X} = \mathbf{V_{00}}, \quad \mathbf{Z} = (\mathbf{V_{00}})^\dagger \, \mathbf{V_{01}} \tag{115}$$

which emphasize the important role played by the matrix $\mathbf{V_{ra}}$.

The matrices $\mathbf{P_r}$, $\mathbf{X}$ and $\mathbf{Z}$ (and thus $\mathbf{V_{ra}}$) are unitary. They satisfy

$$\mathbf{V_{ra}Z} = q\mathbf{ZV_{ra}} \tag{116}$$

$$\mathbf{V_{0a}X} = q^{-a}\mathbf{XV_{0a}} \tag{117}$$

Equation (116) can be iterated to give the useful relation

$$(\mathbf{V_{ra}})^m\mathbf{Z}^n = q^{mn}\mathbf{Z}^n(\mathbf{V_{ra}})^m \tag{118}$$

where $m, n \in \mathbb{Z}/d\mathbb{Z}$. Furthermore, we have the trivial relations

$$e^{-i\pi(d-1)r}(\mathbf{V_{r0}})^d = \mathbf{Z}^d = \mathbf{I_d} \tag{119}$$

More generally, we can show that

$$\forall n \in \mathbb{Z}/d\mathbb{Z} \ : \ (\mathbf{V_{ra}})^n = q^{-n(n-1)a/2}(\mathbf{V_{r0}})^n\mathbf{Z}^{an} \tag{120}$$

Consequently

$$(\mathbf{V_{ra}})^d = e^{i\pi(d-1)(r+a)}\mathbf{I_d} \tag{121}$$

in agreement with the obtained eigenvalues for $\mathbf{V_{ra}}$ (see Eq. (95)).

3.1.4 Weyl pairs

The relations in sections 3.1.1 and 3.1.3 can be particularized in the case $r = a = 0$. For example, Eq. (118) gives the useful relation

$$\mathbf{X}^m\mathbf{Z}^n = q^{mn}\mathbf{Z}^n\mathbf{X}^m, \quad (m,n) \in \mathbb{N}^2 \tag{122}$$

The fundamental relationship between the matrices $\mathbf{X}$ and $\mathbf{Z}$ is emphasized by the following proposition.

Proposition 6. *The unitary matrices* $\mathbf{X}$ *and* $\mathbf{Z}$ *satisfy the q-commutation relation*

$$[\mathbf{X},\mathbf{Z}]_q = \mathbf{XZ} - q\mathbf{ZX} = 0 \tag{123}$$

and the cyclicity relations

$$\mathbf{X}^d = \mathbf{Z}^d = \mathbf{I_d} \tag{124}$$

In addition, they are connected through

$$(\mathbf{F_{00}})^\dagger\,\mathbf{XF_{00}} = \mathbf{Z} \tag{125}$$

that indicates that $\mathbf{X}$ *and* $\mathbf{Z}$ *are related by an ordinary DFT transform.*

According to Proposition 6, the matrices $\mathbf{X}$ and $\mathbf{Z}$ constitute a Weyl pair $(\mathbf{X},\mathbf{Z})$. Weyl pairs were introduced at the beginning of quantum mechanics (Weyl, 1931) and used for building operator unitary bases (Schwinger, 1960). We shall emphasis their interest for quantum information and quantum computing in section 4.

Let x and z be the linear operators associated with $\mathbf{X}$ and $\mathbf{Z}$, respectively. They are given by

$$x = v_{00}, \quad z = (v_{00})^\dagger v_{01} \ \Rightarrow \ xz = v_{01} \tag{126}$$

as functions of the operator v_{ra}. Each of the relations involving $\mathbf{X}$ and $\mathbf{Z}$ can be transcribed in terms of x and z.

The properties of x follow from those of v_{ra} with $r = a = 0$. The unitary operator x is a shift operator when acting on $|j,m\rangle$ or $|n\rangle$ (see (34) and (88)) and a phase operator when acting on $|j\alpha;00\rangle = |0\alpha;0\rangle$ (see (47) and (89)). More precisely, we have

$$x|j,m\rangle = |j,m \oplus 1\rangle \;\Leftrightarrow\; x|n\rangle = |n \ominus 1\rangle \tag{127}$$

and

$$x|0\alpha;0\rangle = q^{-\alpha}|0\alpha;0\rangle \tag{128}$$

The unitary operator z satisfies

$$z|j,m\rangle = q^{j-m}|j,m\rangle \;\Leftrightarrow\; z|n\rangle = q^{n}|n\rangle \tag{129}$$

and

$$z|a\alpha;0\rangle = q^{-1}|a\alpha_1;0\rangle, \quad \alpha_1 = \alpha \ominus 1 \tag{130}$$

It thus behaves as a phase operator when acting on $|j,m\rangle$ or $|n\rangle$ and a shift operator when acting on $|a\alpha;0\rangle$.

In view of (128) and (129), the two cyclic operators x and z (cf. $x^d = z^d = I$) are isospectral operators. They are connected via a discrete Fourier transform operator (see Eq. (125)).

Let us now define the operators

$$u_{ab} = x^a z^b, \quad a,b = 0,1,\ldots,d-1 \tag{131}$$

The d^2 operators u_{ab} are unitary and satisfy the following trace relation

$$\operatorname{tr}\left((u_{ab})^{\dagger}u_{a'b'}\right) = d\,\delta_{a,a'}\,\delta_{b,b'} \tag{132}$$

where the trace is taken on the d-dimensional space $\epsilon(d) = \epsilon(2j+1)$. This trace relation shows that the d^2 operators u_{ab} are pairwise orthogonal operators so that they can serve as a basis for developing any operator acting on the Hilbert space $\epsilon(d)$. Furthermore, the commutator and the anticommutator of u_{ab} and $u_{a'b'}$ are given by

$$[u_{ab}, u_{a'b'}] = \left(q^{-ba'} - q^{-ab'}\right)u_{a''b''}, \quad a'' = a \oplus a', \quad b'' = b \oplus b' \tag{133}$$

and

$$\{u_{ab}, u_{a'b'}\} = \left(q^{-ba'} + q^{-ab'}\right)u_{a''b''}, \quad a'' = a \oplus a', \quad b'' = b \oplus b' \tag{134}$$

Consequently, $[u_{ab}, u_{a'b'}] = 0$ if and only if $ab' \ominus ba' = 0$ and $\{u_{ab}, u_{a'b'}\} = 0$ if and only if $ab' \ominus ba' = (1/2)d$. Therefore, all anticommutators $\{u_{ab}, u_{a'b'}\}$ are different from 0 if d is an odd integer. From a group-theoretical point of view, we have the following result.

Proposition 7. *The set $\{u_{ab} = x^a z^b : a,b = 0,1,\ldots,d-1\}$ generates a d^2-dimensional Lie algebra. This algebra can be seen to be the Lie algebra of the general linear group $GL(d,\mathbb{C})$. The subset $\{u_{ab} : a,b = 0,1,\ldots,d-1\} \setminus \{u_{00}\}$ thus spans the Lie algebra of the special linear group $SL(d,\mathbb{C})$.*

A second group-theoretical aspect connected with the operators u_{ab} concerns a finite group, the so-called finite Heisenberg-Weyl group $WH(\mathbb{Z}/d\mathbb{Z})$, known as the Pauli group P_d in quantum information (Kibler, 2008). The set $\{u_{ab} : a, b = 0, 1, \ldots, d-1\}$ is not closed under multiplication. However, it is possible to extend the latter set in order to have a group as follows.

Proposition 8. *Let us define the operators w_{abc} via*

$$w_{abc} = q^a u_{bc}, \quad a, b, c = 0, 1, \ldots, d-1 \tag{135}$$

Then, the set $\{w_{abc} = q^a x^b z^c : a, b, c = 0, 1, \ldots, d-1\}$, endowed with the multiplication of operators, is a group of order d^3 isomorphic with the Heisenberg-Weyl group $WH(\mathbb{Z}/d\mathbb{Z})$. This group, also referred to as the Pauli group P_d, is a nonabelian (for $d \geq 2$) nilpotent group with nilpotency class equal to 3. It is isomorphic with a finite subgroup of the group $U(d)$ for d even or $SU(d)$ for d odd.

Proposition 8 easily follows from the composition law

$$w_{abc}w_{a'b'c'} = w_{a''b''c''}, \quad a'' = a \oplus a' \ominus cb', \quad b'' = b \oplus b', \quad c'' = c \oplus c' \tag{136}$$

Note that the group commutator of the two elements w_{abc} and $w_{a'b'c'}$ of the group $WH(\mathbb{Z}/d\mathbb{Z})$ is

$$w_{abc}w_{a'b'c'}(w_{abc})^{-1}(w_{a'b'c'})^{-1} = w_{a''00}, \quad a'' = bc' \ominus cb' \tag{137}$$

which can be particularized as

$$u_{ab}u_{a'b'}(u_{ab})^{-1}(u_{a'b'})^{-1} = q^{ab' \ominus ba'} I \tag{138}$$

in terms of the operators u_{ab}.

All this is reminiscent of the group $SU(2)$, the generators of which are the well-known Pauli matrices. Therefore, the operators u_{ab} shall be referred as generalized Pauli operators and their matrices as generalized Pauli matrices. This will be considered further in section 4.

3.1.5 Link with the cyclic group C_d

There exists an interesting connection between the operator v_{ra} and the cyclic group C_d (see section 2.3). The following proposition presents another aspect of this connection.

Proposition 9. *Let R be a generator of C_d (e.g., a rotation of $2\pi/d$ around an arbitrary axis). The application*

$$R^n \mapsto \mathbf{X}^n \ : \ n = 0, 1, \ldots, d-1 \tag{139}$$

defines a d-dimensional matrix representation of C_d. This representation is the regular representation of C_d.

Thus, the reduction of the representation $\{\mathbf{X}^n : n = 0, 1, \ldots, d-1\}$ contains once and only once each (one-dimensional) irreducible representation

$$\chi^{(a)} = (1, q^a, \ldots, q^{(d-1)a}), \quad a = 0, 1, \ldots, d-1 \tag{140}$$

of C_d.

3.1.6 Link with the W_∞ algebra

Let us define the matrix

$$\mathbf{T}_{(\mathbf{n_1},\mathbf{n_2})} = q^{\frac{1}{2}n_1 n_2} \mathbf{Z}^{n_1} \mathbf{X}^{n_2}, \quad (n_1, n_2) \in \mathbb{N}^2 \tag{141}$$

It is convenient to use the abbreviation

$$(n_1, n_2) \equiv n \quad \Rightarrow \quad \mathbf{T}_{(\mathbf{n_1},\mathbf{n_2})} \equiv \mathbf{T_n} \tag{142}$$

The matrices $\mathbf{T_n}$ span an infinite-dimensional Lie algebra. This may be precised as follows.
Proposition 10. *The commutator* $[\mathbf{T_m}, \mathbf{T_n}]$ *is given by*

$$[\mathbf{T_m}, \mathbf{T_n}] = -2i \sin\left(\frac{\pi}{d} m \times n\right) \mathbf{T_{m+n}} \tag{143}$$

where

$$m \times n = m_1 n_2 - m_2 n_1, \quad m + n = (m_1 + n_1, m_2 + n_2) \tag{144}$$

The matrices $\mathbf{T_m}$ *can be thus formally viewed as the generators of the infinite-dimensional Lie algebra* W_∞.
The proof of (143) is easily obtained by using (122). This leads to

$$\mathbf{T_m} \mathbf{T_n} = q^{-\frac{1}{2} m \times n} \mathbf{T_{m+n}} \tag{145}$$

which implies (143). Thus, we get the Lie algebra W_∞ (or sine algebra) investigated in (Fairlie et al., 1990).

3.2 Quadratic discrete Fourier transform

3.2.1 Generalities

We are now prepared for discussing analogs of the transformations (82) and (84) in the language of classical signal theory.
Definition 5. *Let us consider the transformation*

$$x = \{x_m \in \mathbb{C} : m = 0, 1, \ldots, d-1\} \quad \leftrightarrow \quad y = \{y_n \in \mathbb{C} : n = 0, 1, \ldots, d-1\} \tag{146}$$

defined by

$$y_n = \sum_{m=0}^{d-1} (\mathbf{F_{ra}})_{mn} \, x_m \quad \Leftrightarrow \quad x_m = \sum_{n=0}^{d-1} \overline{(\mathbf{F_{ra}})_{mn}} \, y_n \tag{147}$$

where

$$(\mathbf{F_{ra}})_{nm} = \frac{1}{\sqrt{d}} q^{n(d-n)a/2 + (d-1)^2 r/4 + n[m-(d-1)r/2]}, \quad n, m = 0, 1, \ldots, d-1 \tag{148}$$

For $a \neq 0$, *the bijective transformation* $x \leftrightarrow y$ *can be thought of as a quadratic DFT.*
In Eq. (147), we choose the matrix $\mathbf{F_{ra}}$ as the quadratic Fourier matrix instead of the matrix $\mathbf{H_{ra}}$ because the particular case $r = a = 0$ corresponds to the ordinary DFT (see also (Atakishiyev et al., 2010)). Note that the matrices $\mathbf{F_{ra}}$ and $\mathbf{H_{ra}}$ are interrelated via

$$(\mathbf{F_{ra}})_{nm} = (\mathbf{H_{ra}})_{n'm}, \quad n' = d - 1 - n \tag{149}$$

Therefore, the lines of $\mathbf{F_{ra}}$ in the order $0, 1, \ldots, d-1$ coincide with those of $\mathbf{H_{ra}}$ in the reverse order $d-1, d-2, \ldots, 0$.

The analog of the Parseval-Plancherel theorem for the ordinary DFT can be expressed in the following way.

Theorem 2. *The quadratic transformations $x \leftrightarrow y$ and $x' \leftrightarrow y'$ associated with the same matrix $\mathbf{F_{ra}}$, with $r \in \mathbb{R}$ and $a \in \mathbb{Z}/d\mathbb{Z}$, satisfy the conservation rule*

$$\sum_{n=0}^{d-1} \overline{y_n}\, y'_n = \sum_{m=0}^{d-1} \overline{x_m}\, x'_m \tag{150}$$

where both sums do not depend on r and a.

3.2.2 Properties of the quadratic DFT matrix

In order to get familiar with the quadratic DFT defined by (147), we now examine some of the properties of the quadratic DFT matrix $\mathbf{F_{ra}}$.

Proposition 11. *For d arbitrary, the matrix elements of $\mathbf{F_{ra}}$ satisfies the useful symmetry properties*

$$(\mathbf{F_{ra}})_{d-1\alpha} = q^{(d-1)(r+a)/2-\alpha} e^{-i\pi(d-1)r} (\mathbf{F_{ra}})_{0\alpha}, \quad \alpha = 0, 1, \ldots, d-1 \tag{151}$$

$$(\mathbf{F_{ra}})_{n-1\alpha} = q^{(d-1)(r+a)/2-\alpha+na} (\mathbf{F_{ra}})_{n\alpha}, \ n = 1, 2, \ldots, d-1, \ \alpha = 0, 1, \ldots, d-1 \tag{152}$$

which can be reduced to the sole symmetry relation

$$(\mathbf{F_{0a}})_{n\ominus 1\alpha} = q^{(d-1)a/2-\alpha+na} (\mathbf{F_{0a}})_{n\alpha}, \quad n, \alpha = 0, 1, \ldots, d-1 \tag{153}$$

when $r = 0$.

Proposition 12. *For d arbitrary, the matrix $\mathbf{F_{ra}}$ is unitary.*

The latter result can be checked from a straightforward calculation. It also follows in a simple way from

$$\langle j\alpha; ra | j\beta; sb \rangle = \langle a\alpha; r | b\beta; s \rangle = ((\mathbf{F_{ra}})^{\dagger} \mathbf{F_{sb}})_{\alpha\beta} \tag{154}$$

It is sufficient to put $s = r$ and $b = a$ in (154) and to use (48).

For d arbitrary, in addition to be unitary the matrix $\mathbf{F_{ra}}$ is such that the modulus of each of its matrix elements is equal to $1/\sqrt{d}$. Thus, $\mathbf{F_{ra}}$ can be considered as a generalized Hadamard matrix (we adopt here the normalization of Hadamard matrices generally used in quantum information and quantum computing (Kibler, 2009)). In the case where d is a prime number, we shall prove in section 4 from (154) that the matrix $(\mathbf{F_{ra}})^{\dagger}\mathbf{F_{rb}}$ is another Hadamard matrix for $b \neq a$. Similar results hold for the matrix $\mathbf{H_{ra}}$.

Proposition 13. *For d arbitrary, the matrix $\mathbf{F_{ra}}$ can be factorized as*

$$\mathbf{F_{ra}} = \mathbf{D_{ra}}\mathbf{F}, \quad \mathbf{F} = \mathbf{F_{00}} \tag{155}$$

where $\mathbf{D_{ra}}$ is the $d \times d$ diagonal matrix with the matrix elements

$$(\mathbf{D_{ra}})_{mn} = q^{m(d-m)a/2+(d-1)^2r/4-m(d-1)r/2} \delta_{m,n} \tag{156}$$

and $\mathbf{F}$ is the well-known ordinary DFT matrix.

For fixed d, there is one d-multiple infinity of Gaussian matrices $\mathbf{D_{ra}}$ (and thus $\mathbf{F_{ra}}$) distinguished by $a \in \mathbb{Z}/d\mathbb{Z}$ and $r \in \mathbb{R}$. The matrix $\mathbf{F}$ was the object of a great number of studies. The main properties of the ordinary DFT matrix **F** are summarized in (Atakishiyev et al., 2010). Let us simply recall here the fundamental property

$$\mathbf{F}^4 = \mathbf{I_d} \tag{157}$$

of interest for obtaining the eigenvalues and eigenvectors of **F**.
Proposition 14. *The determinant of* $\mathbf{F_{ra}}$ *reads*

$$\det \mathbf{F_{ra}} = e^{i\pi(d^2-1)a/6} \det \mathbf{F} \tag{158}$$

where the value of $\det \mathbf{F}$ *is well known (Atakishiyev et al., 2010; Mehta, 1987).*
Proposition 15. *The trace of* $\mathbf{F_{ra}}$ *reads*

$$\operatorname{tr} \mathbf{F_{ra}} = e^{i\pi(d-1)^2 r/(2d)} \frac{1}{\sqrt{d}} S(u,v,w) \tag{159}$$

where $S(u,v,w)$ *is*

$$S(u,v,w) = \sum_{k=0}^{|w|-1} e^{i\pi(uk^2+vk)/w} \tag{160}$$

with

$$u = 2-a, \quad v = d(a-r)+r, \quad w = d \tag{161}$$

(note that v *is not necessarily an integer).*
Let us recall that the sum defined by (160) is a generalized quadratic Gauss sum. It can be calculated easily in the situation where u, v and w are integers such that u and w are mutually prime, uw is not zero, and $uw+v$ is even (Berndt et al., 1998).
Note that the case $a = 2$ deserves a special attention. In this case, the quadratic character of $\operatorname{tr} \mathbf{F_{ra}}$ disappears. In addition, if $r = 0$ we get

$$\operatorname{tr} \mathbf{F_{02}} = \sqrt{d} \tag{162}$$

as can be seen from a direct calculation.
Example 5: In order to illustrate the preceding properties, let us consider the matrix

$$\mathbf{F_{02}} = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ q^5 & 1 & q & q^2 & q^3 & q^4 \\ q^2 & q^4 & 1 & q^2 & q^4 & 1 \\ q^3 & 1 & q^3 & 1 & q^3 & 1 \\ q^2 & 1 & q^4 & q^2 & 1 & q^4 \\ q^5 & q^4 & q^3 & q^2 & q & 1 \end{pmatrix} \tag{163}$$

corresponding to $d = 6$ ($\Rightarrow q = e^{i\pi/3}$), $r = 0$ and $a = 2$. It is a simple matter of trivial calculation to check that the properties given above for $\mathbf{F_{ra}}$ are satisfied by the matrix $\mathbf{F_{02}}$.

4. Application to quantum information

4.1 Computational basis and standard $SU(2)$ basis

In quantum information science, we use qubits which are indeed state vectors in the Hilbert space $\mathbb{C}^2$. The more general qubit

$$|\psi_2\rangle = c_0|0\rangle + c_1|1\rangle, \quad c_0 \in \mathbb{C}, \quad c_1 \in \mathbb{C}, \quad |c_0|^2 + |c_1|^2 = 1 \tag{164}$$

is a linear combination of the vectors $|0\rangle$ and $|1\rangle$ which constitute an orthonormal basis

$$B_2 = \{|0\rangle, |1\rangle\} \tag{165}$$

of $\mathbb{C}^2$. These two vectors can be considered as the basis vectors for the fundamental irreducible representation class of $SU(2)$, in the $SU(2) \supset U(1)$ scheme, corresponding to $j = 1/2$ with

$$|0\rangle \equiv |1/2, 1/2\rangle, \quad |1\rangle \equiv |1/2, -1/2\rangle \tag{166}$$

More generally, in d dimensions we use qudits of the form

$$|\psi_d\rangle = \sum_{n=0}^{d-1} c_n|n\rangle, \quad c_n \in \mathbb{C}, \quad n = 0, 1, \ldots, d-1, \quad \sum_{n=0}^{d-1} |c_n|^2 = 1 \tag{167}$$

where the vectors $|0\rangle, |1\rangle, \ldots, |d-1\rangle$ span an orthonormal basis of $\mathbb{C}^d$ with

$$\langle n|n'\rangle = \delta_{n,n'} \tag{168}$$

By introducing

$$j = \frac{1}{2}(d-1), \quad m = n - \frac{1}{2}(d-1), \quad |j, m\rangle = |d-1-n\rangle \tag{169}$$

(a change of notations equivalent to (68)), the qudits $|n\rangle$ can be viewed as the basis vectors $|j, m\rangle$ for the irreducible representation class associated with j of $SU(2)$ in the $SU(2) \supset U(1)$ scheme. More precisely, the correspondence between angular momentum states and qudits is

$$|0\rangle \equiv |j, j\rangle, \quad |1\rangle \equiv |j, j-1\rangle, \quad \ldots, \quad |d-1\rangle \equiv |j, -j\rangle \tag{170}$$

where $|j, j\rangle, |j, j-1\rangle, \ldots, |j, -j\rangle$ are common eigenvectors of angular momentum operators j^2 and j_z. In other words, the basis B_d (see (83)), known in quantum information as the computational basis, may be identified to the $SU(2) \supset U(1)$ standard basis or angular momentum basis B_{2j+1} (see (31)). We shall see in section 4.2 that such an identification is very useful when d is a prime number and does not seem to be very interesting when d is not a prime integer. Note that the qudits $|0\rangle, |1\rangle, \ldots, |d-1\rangle$ are often represented by the column vectors $\phi_0, \phi_1, \ldots, \phi_{d-1}$ (given by (94)), respectively.

4.2 Mutually unbiased bases

The basis B_{ra} given by (70) can serve as another basis for qudits. For arbitrary d, the couple (B_{ra}, B_d) exhibits an interesting property. For fixed d, r and a, Eq. (71) gives

$$\forall n \in \mathbb{Z}/d\mathbb{Z},\ \forall \alpha \in \mathbb{Z}/d\mathbb{Z}\ :\ |\langle n|a\alpha;r\rangle| = \frac{1}{\sqrt{d}} \tag{171}$$

Equation (171) shows that B_{ra} and B_d are two unbiased bases.
Other examples of unbiased bases can be obtained for $d = 2$ and 3. We easily verify that the bases B_{r0} and B_{r1} for $d = 2$ given by (58) are unbiased. Similarly, the bases B_{r0}, B_{r1} and B_{r2} for $d = 3$ given by (61) are mutually unbiased. Therefore, by combining these particular results with the general result implied by (171) we end up with three MUBs for $d = 2$ and four MUBs for $d = 3$, in agreement with $N_{MUB} = d + 1$ when d is a prime number. The results for $d = 2$ and 3 can be generalized in the case where d is a prime number. This leads to the following theorem (Albouy & Kibler, 2007; Kibler, 2008; Kibler & Planat, 2006).
Theorem 3. *For $d = p$, with p a prime number, the bases B_{r0}, B_{r1}, ..., B_{rp-1}, B_p corresponding to a fixed value of r form a complete set of $p + 1$ MUBs. The p^2 vectors $|a\alpha;r\rangle$ or $\phi(a\alpha;r)$, with $a, \alpha = 0, 1, \ldots, p-1$, of the bases $B_{r0}, B_{r1}, \ldots, B_{rp-1}$ are given by a single formula, namely, Eq. (72) or (93). The index r makes it possible to distinguish different complete sets of $p + 1$ MUBs.*
The proof is as follows. First, according to (171), the computational basis B_p is unbiased with any of the p bases $B_{r0}, B_{r1}, \ldots, B_{rp-1}$. Second, we get

$$\langle a\alpha;r|b\beta;r\rangle = \frac{1}{p}\sum_{k=0}^{p-1} q^{k(p-k)(b-a)/2+k(\beta-\alpha)} \tag{172}$$

or

$$\langle a\alpha;r|b\beta;r\rangle = \frac{1}{p}\sum_{k=0}^{p-1} e^{i\pi\{(a-b)k^2+[p(b-a)+2(\beta-\alpha)]k\}/p} \tag{173}$$

The right-hand side of (173) can be expressed in terms of a generalized quadratic Gauss sum. This leads to

$$\langle a\alpha;r|b\beta;r\rangle = \frac{1}{p}S(u,v,w) \tag{174}$$

where the Gauss sum $S(u,v,w)$ is given by (160) with the parameters

$$u = a - b, \quad v = -(a-b)p - 2(\alpha - \beta), \quad w = p \tag{175}$$

which ensure that $uw + v$ is even. The generalized Gauss sum $S(u,v,w)$ in (174)-(175) can be calculated from the methods described in (Berndt et al., 1998). We thus obtain

$$|\langle a\alpha;r|b\beta;r\rangle| = \frac{1}{\sqrt{p}} \tag{176}$$

for all a, b, α, and β in $\mathbb{Z}/p\mathbb{Z}$ with $b \neq a$. This completes the proof.
Theorem 3 renders feasible to derive in one step the $(p+1)p$ qupits (i.e., qudits with $d = p$ a prime integer) of a complete set of $p + 1$ MUBs in $\mathbb{C}^p$. The single formula (72) or (93), giving the p^2 vectors $|a\alpha;r\rangle$ or $\phi(a\alpha;r)$, with $a, \alpha = 0, 1, \ldots, p-1$, of the bases $B_{r0}, B_{r1}, \ldots, B_{rp-1}$, is easily codable on a classical computer.

Example 6: The $p = 2$ case. For $r = 0$, the $p + 1 = 3$ MUBs are

$$\begin{aligned} B_{00} &: \quad \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \quad -\frac{|0\rangle - |1\rangle}{\sqrt{2}} \\ B_{01} &: \quad i\frac{|0\rangle - i|1\rangle}{\sqrt{2}}, \quad -i\frac{|0\rangle + i|1\rangle}{\sqrt{2}} \\ B_2 &: \quad |0\rangle, \quad |1\rangle \end{aligned} \tag{177}$$

cf. (59), (60), (100) and (101). The global factors -1 in B_{00} and $\pm i$ in B_{01} arise from the general formula (72); they are irrelevant for quantum information and can be omitted.

Example 7: The $p = 3$ case. For $r = 0$, the $p + 1 = 4$ MUBs are

$$\begin{aligned} B_{00} &: \quad \frac{|0\rangle + |1\rangle + |2\rangle}{\sqrt{3}}, \quad \frac{q^2|0\rangle + q|1\rangle + |2\rangle}{\sqrt{3}}, \quad \frac{q|0\rangle + q^2|1\rangle + |2\rangle}{\sqrt{3}} \\ B_{01} &: \quad \frac{q|0\rangle + q|1\rangle + |2\rangle}{\sqrt{3}}, \quad \frac{|0\rangle + q^2|1\rangle + |2\rangle}{\sqrt{3}}, \quad \frac{q^2|0\rangle + |1\rangle + |2\rangle}{\sqrt{3}} \\ B_{02} &: \quad \frac{q^2|0\rangle + q^2|1\rangle + |2\rangle}{\sqrt{3}}, \quad \frac{q|0\rangle + |1\rangle + |2\rangle}{\sqrt{3}}, \quad \frac{|0\rangle + q|1\rangle + |2\rangle}{\sqrt{3}} \\ B_3 &: \quad |0\rangle, \quad |1\rangle, \quad |2\rangle \end{aligned} \tag{178}$$

with $q = e^{i2\pi/3}$, cf. (62), (63), (64), (106), (108) and (110).

As a simple consequence of Theorem 3, we get the following corollary which can be derived by combining Theorem 3 with Eq. (154).

Corollary 2. *For $d = p$, with p a prime number, the $p \times p$ matrix $(\mathbf{F_{ra}})^\dagger \mathbf{F_{rb}}$ with $b \neq a$ $(a, b = 0, 1, \ldots, p-1)$ is a generalized Hadamard matrix.*

Going back to arbitrary d, it is to be noted that for a fixed value of r, the $d + 1$ bases B_{r0}, B_{r1}, ..., B_{rd-1}, B_d do not provide in general a complete set of $d + 1$ MUBs even in the case where d is a power p^e with $e \geq 2$ of a prime integer p. However, it is possible to show (Kibler, 2009) that the bases B_{ra}, $B_{ra\oplus 1}$ and B_d are three MUBs in $\mathbb{C}^d$, in agreement with $N_{MUB} \geq 3$. Therefore for d arbitrary, given two Hadamard matrices $\mathbf{F_{ra}}$ and $\mathbf{F_{sb}}$, the product $\mathbf{F_{ra}}^\dagger \mathbf{F_{sb}}$ is not in general a Hadamard matrix.

In the case where d is a power p^e with $e \geq 2$ of a prime integer p, tensor products of the unbiased bases B_{r0}, B_{r1}, ..., B_{rp-1} can be used for generating $p^e + 1$ MUBs in dimension $d = p^e$. This can be illustrated with the following example.

Example 8: The $d = 2^2$ case. This case corresponds to a spin $j = 3/2$. The application of (45) or (72) yields four bases B_{0a} $(a = 0, 1, 2, 3)$. As a point of fact, the five bases B_{00}, B_{01}, B_{02}, B_{03} and B_4 do not form a complete set of $d + 1 = 5$ MUBs ($d = 4$ is not a prime number). Nevertheless, it is possible to find five MUBs because $d = 2^2$ is the power of a prime number. This can be achieved by replacing the space $\epsilon(4)$ spanned by

$$B_4 = \{|3/2, m\rangle : m = 3/2, 1/2, -1/2, -3/2\} \quad \text{or} \quad \{|n\rangle : n = 0, 1, 2, 3\} \tag{179}$$

by the tensor product space $\epsilon(2) \otimes \epsilon(2)$ spanned by the canonical or computational basis

$$B_2 \otimes B_2 = \{|0\rangle \otimes |0\rangle, |0\rangle \otimes |1\rangle, |1\rangle \otimes |0\rangle, |1\rangle \otimes |1\rangle\} \tag{180}$$

The space $\epsilon(2) \otimes \epsilon(2)$ is associated with the coupling of two spin angular momenta $j_1 = 1/2$ and $j_2 = 1/2$ or two qubits (in the vector $u \otimes v$, u and v correspond to j_1 and j_2, respectively). Four of the five MUBs for $d = 4$ can be constructed from the direct products

$$|ab : \alpha\beta\rangle = |a\alpha; 0\rangle \otimes |b\beta; 0\rangle \tag{181}$$

which are eigenvectors of the operators

$$w_{ab} = v_{0a} \otimes v_{0b} \tag{182}$$

(the operators v_{0a} and v_{0b} refer to the two spaces $\epsilon(2)$, the vectors of type $|a\alpha; 0\rangle$ and $|b\beta; 0\rangle$ are given by the master formula (72) for $d = 2$). Obviously, the set

$$B_{0a0b} = \{|ab : \alpha\beta\rangle : \alpha, \beta = 0, 1\} \tag{183}$$

is an orthonormal basis in $\mathbb{C}^4$. It is evident that B_{0000} and B_{0101} are two unbiased bases since the modulus of the inner product of $|00 : \alpha\beta\rangle$ by $|11 : \alpha'\beta'\rangle$ is

$$|\langle 00 : \alpha\beta|11 : \alpha'\beta'\rangle| = |\langle 0\alpha; 0|1\alpha'; 0\rangle\langle 0\beta; 0|1\beta'; 0\rangle| = \frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}} = \frac{1}{\sqrt{4}} \tag{184}$$

A similar result holds for the two bases B_{0001} and B_{0100}. However, the four bases B_{0000}, B_{0101}, B_{0001} and B_{0100} are not mutually unbiased. A possible way to overcome this no-go result is to keep the bases B_{0000} and B_{0101} intact and to re-organize the vectors inside the bases B_{0001} and B_{0100} in order to obtain four MUBs. We are thus left with four bases

$$W_{00} \equiv B_{0000}, \quad W_{11} \equiv B_{0101}, \quad W_{01}, \quad W_{10} \tag{185}$$

which together with the computational basis B_4 give five MUBs. Specifically, we have

$$W_{00} = \{|00 : \alpha\beta\rangle : \alpha, \beta = 0, 1\} \tag{186}$$
$$W_{11} = \{|11 : \alpha\beta\rangle : \alpha, \beta = 0, 1\} \tag{187}$$
$$W_{01} = \{\lambda|01 : \alpha\beta\rangle + \mu|01 : \alpha \oplus 1\beta \oplus 1\rangle : \alpha, \beta = 0, 1\} \tag{188}$$
$$W_{10} = \{\lambda|10 : \alpha\beta\rangle + \mu|10 : \alpha \oplus 1\beta \oplus 1\rangle : \alpha, \beta = 0, 1\} \tag{189}$$

where

$$\lambda = \frac{1-i}{2}, \quad \mu = \frac{1+i}{2} \tag{190}$$

As a résumé, only two formulas are necessary for obtaining the $d^2 = 16$ vectors for the bases W_{ab}, namely

$$W_{00}, W_{11} \quad : \quad |aa : \alpha\beta\rangle \tag{191}$$
$$W_{01}, W_{10} \quad : \quad \lambda|aa \oplus 1 : \alpha\beta\rangle + \mu|aa \oplus 1 : \alpha \oplus 1\beta \oplus 1\rangle \tag{192}$$

for all a, α and β in $\mathbb{Z}/2\mathbb{Z}$. The five MUBs are listed below as state vectors and column vectors with

$$|0\rangle \equiv |j = 1/2, m = 1/2\rangle \text{ or } \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |1\rangle \equiv |j = 1/2, m = -1/2\rangle \text{ or } \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{193}$$

The canonical basis:

$$|0\rangle \otimes |0\rangle, \quad |0\rangle \otimes |1\rangle, \quad |1\rangle \otimes |0\rangle, \quad |1\rangle \otimes |1\rangle$$

or in column vectors

$$\begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \quad \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \quad \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \quad \begin{pmatrix}0\\0\\0\\1\end{pmatrix} \tag{194}$$

The W_{00} basis:

$$|00:00\rangle = +\frac{|0\rangle + |1\rangle}{\sqrt{2}} \otimes \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \quad |00:01\rangle = -\frac{|0\rangle + |1\rangle}{\sqrt{2}} \otimes \frac{|0\rangle - |1\rangle}{\sqrt{2}}$$
$$|00:10\rangle = -\frac{|0\rangle - |1\rangle}{\sqrt{2}} \otimes \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \quad |00:11\rangle = +\frac{|0\rangle - |1\rangle}{\sqrt{2}} \otimes \frac{|0\rangle - |1\rangle}{\sqrt{2}}$$

or in developed form

$$|00:00\rangle = +\frac{1}{2}(|0\rangle \otimes |0\rangle + |0\rangle \otimes |1\rangle + |1\rangle \otimes |0\rangle + |1\rangle \otimes |1\rangle)$$
$$|00:01\rangle = -\frac{1}{2}(|0\rangle \otimes |0\rangle - |0\rangle \otimes |1\rangle + |1\rangle \otimes |0\rangle - |1\rangle \otimes |1\rangle)$$
$$|00:10\rangle = -\frac{1}{2}(|0\rangle \otimes |0\rangle + |0\rangle \otimes |1\rangle - |1\rangle \otimes |0\rangle - |1\rangle \otimes |1\rangle)$$
$$|00:11\rangle = +\frac{1}{2}(|0\rangle \otimes |0\rangle - |0\rangle \otimes |1\rangle - |1\rangle \otimes |0\rangle + |1\rangle \otimes |1\rangle)$$

or in column vectors

$$\frac{1}{2}\begin{pmatrix}1\\1\\1\\1\end{pmatrix}, \quad -\frac{1}{2}\begin{pmatrix}1\\-1\\1\\-1\end{pmatrix}, \quad -\frac{1}{2}\begin{pmatrix}1\\1\\-1\\-1\end{pmatrix}, \quad \frac{1}{2}\begin{pmatrix}1\\-1\\-1\\1\end{pmatrix} \tag{195}$$

The W_{11} basis:

$$|11:00\rangle = -\frac{|0\rangle - i|1\rangle}{\sqrt{2}} \otimes \frac{|0\rangle - i|1\rangle}{\sqrt{2}}, \quad |11:01\rangle = +\frac{|0\rangle - i|1\rangle}{\sqrt{2}} \otimes \frac{|0\rangle + i|1\rangle}{\sqrt{2}}$$
$$|11:10\rangle = +\frac{|0\rangle + i|1\rangle}{\sqrt{2}} \otimes \frac{|0\rangle - i|1\rangle}{\sqrt{2}}, \quad |11:11\rangle = -\frac{|0\rangle + i|1\rangle}{\sqrt{2}} \otimes \frac{|0\rangle + i|1\rangle}{\sqrt{2}}$$

or in developed form

$$|11:00\rangle = -\frac{1}{2}(|0\rangle \otimes |0\rangle - i|0\rangle \otimes |1\rangle - i|1\rangle \otimes |0\rangle - |1\rangle \otimes |1\rangle)$$
$$|11:01\rangle = +\frac{1}{2}(|0\rangle \otimes |0\rangle + i|0\rangle \otimes |1\rangle - i|1\rangle \otimes |0\rangle + |1\rangle \otimes |1\rangle)$$
$$|11:10\rangle = +\frac{1}{2}(|0\rangle \otimes |0\rangle - i|0\rangle \otimes |1\rangle + i|1\rangle \otimes |0\rangle + |1\rangle \otimes |1\rangle)$$
$$|11:11\rangle = -\frac{1}{2}(|0\rangle \otimes |0\rangle + i|0\rangle \otimes |1\rangle + i|1\rangle \otimes |0\rangle - |1\rangle \otimes |1\rangle)$$

or in column vectors

$$-\frac{1}{2}\begin{pmatrix}1\\-i\\-i\\-1\end{pmatrix},\quad \frac{1}{2}\begin{pmatrix}1\\i\\-i\\1\end{pmatrix},\quad \frac{1}{2}\begin{pmatrix}1\\-i\\i\\1\end{pmatrix},\quad -\frac{1}{2}\begin{pmatrix}1\\i\\i\\-1\end{pmatrix} \tag{196}$$

The W_{01} basis:

$$\begin{aligned}
\lambda|01:00\rangle+\mu|01:11\rangle &= +\mu\frac{|0\rangle+|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle-i|1\rangle}{\sqrt{2}}-\lambda\frac{|0\rangle-|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle+i|1\rangle}{\sqrt{2}}\\
\mu|01:00\rangle+\lambda|01:11\rangle &= -\lambda\frac{|0\rangle+|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle-i|1\rangle}{\sqrt{2}}+\mu\frac{|0\rangle-|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle+i|1\rangle}{\sqrt{2}}\\
\lambda|01:01\rangle+\mu|01:10\rangle &= -\mu\frac{|0\rangle+|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle+i|1\rangle}{\sqrt{2}}+\lambda\frac{|0\rangle-|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle-i|1\rangle}{\sqrt{2}}\\
\mu|01:01\rangle+\lambda|01:10\rangle &= +\lambda\frac{|0\rangle+|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle+i|1\rangle}{\sqrt{2}}-\mu\frac{|0\rangle-|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle-i|1\rangle}{\sqrt{2}}
\end{aligned}$$

or in developed form

$$\begin{aligned}
\lambda|01:00\rangle+\mu|01:11\rangle &= +\frac{i}{2}(|0\rangle\otimes|0\rangle-|0\rangle\otimes|1\rangle-i|1\rangle\otimes|0\rangle-i|1\rangle\otimes|1\rangle)\\
\mu|01:00\rangle+\lambda|01:11\rangle &= +\frac{i}{2}(|0\rangle\otimes|0\rangle+|0\rangle\otimes|1\rangle+i|1\rangle\otimes|0\rangle-i|1\rangle\otimes|1\rangle)\\
\lambda|01:01\rangle+\mu|01:10\rangle &= -\frac{i}{2}(|0\rangle\otimes|0\rangle+|0\rangle\otimes|1\rangle-i|1\rangle\otimes|0\rangle+i|1\rangle\otimes|1\rangle)\\
\mu|01:01\rangle+\lambda|01:10\rangle &= -\frac{i}{2}(|0\rangle\otimes|0\rangle-|0\rangle\otimes|1\rangle+i|1\rangle\otimes|0\rangle+i|1\rangle\otimes|1\rangle)
\end{aligned}$$

or in column vectors

$$\frac{i}{2}\begin{pmatrix}1\\-1\\-i\\-i\end{pmatrix},\quad \frac{i}{2}\begin{pmatrix}1\\1\\i\\-i\end{pmatrix},\quad -\frac{i}{2}\begin{pmatrix}1\\1\\-i\\i\end{pmatrix},\quad -\frac{i}{2}\begin{pmatrix}1\\-1\\i\\i\end{pmatrix} \tag{197}$$

The W_{10} basis:

$$\begin{aligned}
\lambda|10:00\rangle+\mu|10:11\rangle &= +\mu\frac{|0\rangle-i|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle+|1\rangle}{\sqrt{2}}-\lambda\frac{|0\rangle+i|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle-|1\rangle}{\sqrt{2}}\\
\mu|10:00\rangle+\lambda|10:11\rangle &= -\lambda\frac{|0\rangle-i|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle+|1\rangle}{\sqrt{2}}+\mu\frac{|0\rangle+i|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle-|1\rangle}{\sqrt{2}}\\
\lambda|10:01\rangle+\mu|10:10\rangle &= -\mu\frac{|0\rangle-i|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle-|1\rangle}{\sqrt{2}}+\lambda\frac{|0\rangle+i|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle+|1\rangle}{\sqrt{2}}\\
\mu|10:01\rangle+\lambda|10:10\rangle &= +\lambda\frac{|0\rangle-i|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle-|1\rangle}{\sqrt{2}}-\mu\frac{|0\rangle+i|1\rangle}{\sqrt{2}}\otimes\frac{|0\rangle+|1\rangle}{\sqrt{2}}
\end{aligned}$$

or in developed form

$$\begin{aligned}
\lambda|10:00\rangle + \mu|10:11\rangle &= +\frac{i}{2}(|0\rangle \otimes |0\rangle - i|0\rangle \otimes |1\rangle - |1\rangle \otimes |0\rangle - i|1\rangle \otimes |1\rangle) \\
\mu|10:00\rangle + \lambda|10:11\rangle &= +\frac{i}{2}(|0\rangle \otimes |0\rangle + i|0\rangle \otimes |1\rangle + |1\rangle \otimes |0\rangle - i|1\rangle \otimes |1\rangle) \\
\lambda|10:01\rangle + \mu|10:10\rangle &= -\frac{i}{2}(|0\rangle \otimes |0\rangle + i|0\rangle \otimes |1\rangle - |1\rangle \otimes |0\rangle + i|1\rangle \otimes |1\rangle) \\
\mu|10:01\rangle + \lambda|10:10\rangle &= -\frac{i}{2}(|0\rangle \otimes |0\rangle - i|0\rangle \otimes |1\rangle + |1\rangle \otimes |0\rangle + i|1\rangle \otimes |1\rangle)
\end{aligned}$$

or in column vectors

$$\frac{i}{2}\begin{pmatrix}1\\-i\\-1\\-i\end{pmatrix}, \quad \frac{i}{2}\begin{pmatrix}1\\i\\1\\-i\end{pmatrix}, \quad -\frac{i}{2}\begin{pmatrix}1\\i\\-1\\i\end{pmatrix}, \quad -\frac{i}{2}\begin{pmatrix}1\\-i\\1\\i\end{pmatrix} \tag{198}$$

The five preceding bases are of central importance in quantum information for expressing any ququart or quartic (corresponding to $d = 4$) in terms of qudits (corresponding to $d = 2$). It is to be noted that the vectors of the W_{00} and W_{11} bases are not intricated (i.e., each vector is the direct product of two vectors) while the vectors of the W_{01} and W_{10} bases are intricated (i.e., each vector is not the direct product of two vectors). To be more precise, the degree of intrication of the state vectors for the bases W_{00}, W_{11}, W_{01} and W_{10} can be determined in the following way. In arbitrary dimension d, let

$$|\Phi\rangle = \sum_{k=0}^{d-1}\sum_{l=0}^{d-1} a_{kl}|k\rangle \otimes |l\rangle \tag{199}$$

be a double qudit state vector. Then, it can be shown that the determinant of the $d \times d$ matrix $A = (a_{kl})$ satisfies

$$0 \leq |\det A| \leq \frac{1}{\sqrt{d^d}} \tag{200}$$

as discussed in (Albouy, 2009). The case $\det A = 0$ corresponds to the absence of *global* intrication while the case

$$|\det A| = \frac{1}{\sqrt{d^d}} \tag{201}$$

corresponds to a maximal intrication. As an illustration, we obtain that all the state vectors for W_{00} and W_{11} are not intricated and that all the state vectors for W_{01} and W_{10} are maximally intricated.

Generalization of (191) and (192) can be obtained in more complicated situations (two qupits, three qubits, ...). The generalization of (191) is immediate. The generalization of (192) can be achieved by taking linear combinations of vectors such that each linear combination is made of vectors corresponding to the same eigenvalue of the relevant tensor product of operators of type v_{0a}.

4.3 Mutually unbiased bases and Lie agebras

4.3.1 Generalized Pauli matrices

We now examine the interest for quantum information of the Weyl pair $(\mathbf{X}, \mathbf{Z})$ introduced in section 3.1.4. The linear operators x and z corresponding to the matrices $\mathbf{X}$ and $\mathbf{Z}$ are known in quantum information and quantum computing as shift and clock operators, respectively. (Note however that for each of the operators x and z, the *shift* or *clock* character depends on which state the operator acts. The qualification adopted in quantum information and quantum computing corresponds to the action of x and z on the computational basis B_d.) For d arbitrary, they are at the root of the Pauli group P_d, a finite subgroup of $U(d)$ (see section 3.1.4). The normaliser of P_d in $U(d)$ is a Clifford-type group in d dimensions noted Cl_d. More precisely, Cl_d is the set $\{\mathbf{U} \in U(d) | \mathbf{U} P_d \mathbf{U}^\dagger = P_d\}$ endowed with matrix multiplication (the elements of P_d being expressed in terms of the matrices $\mathbf{X}$ and $\mathbf{Z}$). The Pauli group P_d, as well as any other invariant subgroup of Cl_d, is of considerable importance for describing quantum errors and quantum fault tolerance in quantum computing (see (Havlíček & Saniga, 2008; Planat, 2010; Planat & Kibler, 2010) and references therein for recent geometrical approaches to the Pauli group). These concepts are very important in the case of n-qubit systems (corresponding to $d = 2^n$).

The Weyl pair $(\mathbf{X}, \mathbf{Z})$ turns out to be an integrity basis for generating the set $\{\mathbf{X}^a \mathbf{Z}^b : a, b \in \mathbb{Z}/d\mathbb{Z}\}$ of d^2 generalized Pauli matrices in d dimensions (see for instance (Bandyopadhyay et al., 2002; Gottesman et al., 2001; Kibler, 2008; Lawrence et al., 2002; Pittenger & Rubin, 2004) in the context of MUBs and (Balian & Itzykson, 1986; Patera & Zassenhaus, 1988; Šťovíček & Tolar, 1984) in group-theoretical contexts). As seen in section 3.1.4, the latter set constitutes a basis for the Lie algebra of the linear group $GL(d, \mathbb{C})$ (or its unitary restriction $U(d)$) with respect to the commutator law. Let us give two examples of these important generalized Pauli matrices.

Example 9: The $d = 2$ case. For $d = 2 \Leftrightarrow j = 1/2$ ($\Rightarrow q = -1$), the matrices of the four operators u_{ab} with $a, b = 0, 1$ are

$$\mathbf{I}_2 = \mathbf{X}^0 \mathbf{Z}^0, \quad \mathbf{X} = \mathbf{X}^1 \mathbf{Z}^0, \quad \mathbf{Y} = \mathbf{X}^1 \mathbf{Z}^1, \quad \mathbf{Z} = \mathbf{X}^0 \mathbf{Z}^1 \tag{202}$$

or in terms of the matrices $\mathbf{V_{0a}}$

$$\mathbf{I}_2 = (\mathbf{V_{00}})^2, \quad \mathbf{X} = \mathbf{V_{00}}, \quad \mathbf{Y} = \mathbf{V_{01}}, \quad \mathbf{Z} = (\mathbf{V_{00}})^\dagger \mathbf{V_{01}} \tag{203}$$

In detail, we get

$$\mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{204}$$

Alternatively, we have

$$\mathbf{I}_2 = \sigma_0, \quad \mathbf{X} = \sigma_x, \quad \mathbf{Y} = -i\sigma_y, \quad \mathbf{Z} = \sigma_z \tag{205}$$

in terms of the usual (Hermitian and unitary) Pauli matrices σ_0, σ_x, σ_y and σ_z

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{206}$$

The approach developed here leads to generalized Pauli matrices in dimension 2 that differ from the usual Pauli matrices. This is the price one has to pay in order to get a systematic generalization of Pauli matrices in arbitrary dimension. It should be observed that the commutation and anti-commutation relations given by (133) and (134) with $d = 2$ correspond to the well-known commutation and anti-commutation relations for the usual Pauli matrices transcribed in the normalization $\mathbf{X}^1\mathbf{Z}^0 = \sigma_x$, $\mathbf{X}^1\mathbf{Z}^1 = -i\sigma_y$, $\mathbf{X}^0\mathbf{Z}^1 = \sigma_z$.

From a group-theoretical point of view, the matrices $\mathbf{I}_2$, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ can be considered as generators of the group $U(2)$. On the other hand, the Pauli group P_2 contains eight elements; due to the factor $-i$ in $\mathbf{Y} = -i\sigma_y$, the group P_2 is isomorphic to the group of hyperbolic quaternions rather than to the group of ordinary quaternions.

In terms of column vectors, the vectors of the bases B_{00}, B_{01} and B_2 (see (177)) are eigenvectors of σ_x, σ_y and σ_z, respectively (for each matrix the eigenvalues are 1 and -1).

Example 10: The $d = 3$ case. For $d = 3 \Leftrightarrow j = 1$ ($\Rightarrow q = e^{i2\pi/3}$), the matrices of the nine operators u_{ab} with $a, b = 0, 1, 2$, viz.,

$$\begin{array}{l} \mathbf{X}^0\mathbf{Z}^0 = \mathbf{I}_3, \quad \mathbf{X}^1\mathbf{Z}^0 = \mathbf{X}, \quad \mathbf{X}^2\mathbf{Z}^0 = \mathbf{X}^2 \\ \mathbf{X}^0\mathbf{Z}^1 = \mathbf{Z}, \quad \mathbf{X}^0\mathbf{Z}^2 = \mathbf{Z}^2, \quad \mathbf{X}^1\mathbf{Z}^1 = \mathbf{X}\mathbf{Z} \\ \mathbf{X}^2\mathbf{Z}^2, \quad \mathbf{X}^2\mathbf{Z}^1 = \mathbf{X}^2\mathbf{Z}, \quad \mathbf{X}^1\mathbf{Z}^2 = \mathbf{X}\mathbf{Z}^2 \end{array} \tag{207}$$

are

$$\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}^2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & q & 0 \\ 0 & 0 & q^2 \end{pmatrix}, \quad \mathbf{Z}^2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & q^2 & 0 \\ 0 & 0 & q \end{pmatrix}, \quad \mathbf{X}\mathbf{Z} = \begin{pmatrix} 0 & q & 0 \\ 0 & 0 & q^2 \\ 1 & 0 & 0 \end{pmatrix} \tag{208}$$

$$\mathbf{X}^2\mathbf{Z}^2 = \begin{pmatrix} 0 & 0 & q \\ 1 & 0 & 0 \\ 0 & q^2 & 0 \end{pmatrix}, \quad \mathbf{X}^2\mathbf{Z} = \begin{pmatrix} 0 & 0 & q^2 \\ 1 & 0 & 0 \\ 0 & q & 0 \end{pmatrix}, \quad \mathbf{X}\mathbf{Z}^2 = \begin{pmatrix} 0 & q^2 & 0 \\ 0 & 0 & q \\ 1 & 0 & 0 \end{pmatrix}$$

The generalized Pauli matrices (208) differ from the Gell-Mann matrices used in elementary particle physics. They constitute another extension of the Pauli matrices in dimension $d = 3$ of interest for the Lie group $U(3)$ and the Pauli group P_3.

In terms of column vectors, the vectors of the bases B_{00}, B_{01}, B_{02} and B_3 (see (178)) are eigenvectors of $\mathbf{X}$, $\mathbf{X}\mathbf{Z}$, $\mathbf{X}\mathbf{Z}^2$ and $\mathbf{Z}$, respectively (for each matrix the eigenvalues are 1, q and q^2).

4.3.2 MUBs and the special linear group

In the case where d is a prime integer or a power of a prime integer, it is known that the set $\{\mathbf{X}^a\mathbf{Z}^b : a, b = 0, 1, \ldots, d-1\} \backslash \{\mathbf{X}^0\mathbf{Z}^0\}$ of cardinality $d^2 - 1$ can be partitioned into $d + 1$ subsets containing each $d - 1$ commuting matrices (cf. (Bandyopadhyay et al., 2002)). Let us give an example before going to the case where d is an arbitrary prime number.

Example 11: The $d = 5$ case. For $d = 5$, we have the six following sets of four commuting matrices

$$\begin{array}{ll} \mathcal{V}_0 = \{01, 02, 03, 04\}, & \mathcal{V}_1 = \{10, 20, 30, 40\} \\ \mathcal{V}_2 = \{11, 22, 33, 44\}, & \mathcal{V}_3 = \{12, 24, 31, 43\} \\ \mathcal{V}_4 = \{13, 21, 34, 42\}, & \mathcal{V}_5 = \{14, 23, 32, 41\} \end{array} \tag{209}$$

where ab is used as an abbreviation of $\mathbf{X}^a\mathbf{Z}^b$.

Proposition 16. *For $d = p$ with p a prime integer, the $p+1$ sets of $p-1$ commuting matrices are easily seen to be*

$$\begin{aligned}
\mathcal{V}_0 &= \{\mathbf{X}^0\mathbf{Z}^a : a = 1, 2, \dots, p-1\} \\
\mathcal{V}_1 &= \{\mathbf{X}^a\mathbf{Z}^0 : a = 1, 2, \dots, p-1\} \\
\mathcal{V}_2 &= \{\mathbf{X}^a\mathbf{Z}^a : a = 1, 2, \dots, p-1\} \\
\mathcal{V}_3 &= \{\mathbf{X}^a\mathbf{Z}^{2a} : a = 1, 2, \dots, p-1\} \qquad (210) \\
&\vdots \\
\mathcal{V}_{p-1} &= \{\mathbf{X}^a\mathbf{Z}^{(p-2)a} : a = 1, 2, \dots, p-1\} \\
\mathcal{V}_p &= \{\mathbf{X}^a\mathbf{Z}^{(p-1)a} : a = 1, 2, \dots, p-1\}
\end{aligned}$$

Each of the $p+1$ sets $\mathcal{V}_0, \mathcal{V}_1, \dots, \mathcal{V}_p$ can be put in a one-to-one correspondance with one basis of the complete set of $p+1$ MUBs. In fact, $\mathcal{V}_0$ is associated with the computational basis B_p; furthermore, in view of

$$\mathbf{V_{0a}} \in \mathcal{V}_{a+1} = \{\mathbf{X}^b\mathbf{Z}^{ab} : b = 1, 2, \dots, p-1\}, \quad a = 0, 1, \dots, p-1 \qquad (211)$$

it follows that $\mathcal{V}_1$, $\mathcal{V}_2$, …, $\mathcal{V}_p$ are associated with the p remaining MUBs B_{00}, B_{01}, …, B_{0p-1}, respectively.

Keeping into account the fact that the set $\{\mathbf{X}^a\mathbf{Z}^b : a, b = 0, 1, \dots, p-1\} \setminus \{\mathbf{X}^0\mathbf{Z}^0\}$ spans the Lie algebra of the special linear group $SL(p, \mathbb{C})$, we have the next theorem.

Theorem 4. *For $d = p$ with p a prime integer, the Lie algebra $sl(p, \mathbb{C})$ of the group $SL(p, \mathbb{C})$ can be decomposed into a sum (vector space sum indicated by $\uplus$) of $p+1$ abelian subalgebras each of dimension $p-1$, i.e.,*

$$sl(p, \mathbb{C}) \simeq v_0 \uplus v_1 \uplus \ldots \uplus v_p \qquad (212)$$

where the $p+1$ subalgebras $v_0, v_1, \dots, v_p$ are Cartan subalgebras generated respectively by the sets $\mathcal{V}_0, \mathcal{V}_1, \dots, \mathcal{V}_p$ containing each $p-1$ commuting matrices.

The latter result can be extended when $d = p^e$ with p a prime integer and e an integer ($e \geq 2$): there exists a decomposition of $sl(p^e, \mathbb{C})$ into $p^e + 1$ abelian subalgebras of dimension $p^e - 1$ (cf. (Boykin et al., 2007; Kibler, 2009; Patera & Zassenhaus, 1988)).

5. Conclusion

The quadratic discrete Fourier transform studied in this chapter can be considered as a two-parameter extension, with a quadratic term, of the usual discrete Fourier transform. In the case where the two parameters are taken to be equal to zero, the quadratic discrete Fourier transform is nothing but the usual discrete Fourier transform. The quantum quadratic discrete Fourier transform plays an important role in the field of quantum information. In particular, such a transformation in prime dimension can be used for obtaining a complete set of mutually unbiased bases. It is to be mentioned that the quantum quadratic discrete Fourier transform also arises in the determination of phase operators for the groups $SU(2)$ and $SU(1,1)$ in connection with the representations of a generalized oscillator algebra (Atakishiyev et al., 2010; Daoud & Kibler, 2010). As an open question, it should be worth investigating the relation between the quadratic discrete Fourier transform and the Fourier transform on a finite ring or a finite field.

6. Acknowledgements

The author is greatly indebted to M. Daoud, M. Planat, O. Albouy, M. Saniga, N.M. Atakishiyev and K.B. Wolf for collaboration on various facets of the material contained in this chapter.

7. Appendix: Wigner-Racah algebra of $SU(2)$ in the $\{j^2, x\}$ scheme

In this self-contained Appendix, the bar does not indicate complex conjugation. Here, complex conjugation is denoted with a star.

The Wigner-Racah algebra of the group $SU(2)$ in the $SU(2) \supset U(1)$ or $\{j^2, j_z\}$ scheme is well known. It corresponds to the use of bases of type B_{2j+1} resulting from the simultaneous diagonalization of the Casimir operator j^2 and of the Cartan generator j_z of $SU(2)$. Any change of basis of type

$$|j, \mu\rangle = \sum_{m=-j}^{j} |j, m\rangle \langle j, m|j, \mu\rangle \tag{213}$$

(where for fixed j the elements $\langle j, m|j, \mu\rangle$ define a $(2j+1) \times (2j+1)$ unitary matrix) leads to another acceptable scheme for the Wigner-Racah algebra of $SU(2)$. In this scheme, the matrices of the irreducible representation classes of $SU(2)$ take a new form as well as the coupling coefficients (and the associated $3 - jm$ symbols). For instance, the Clebsch-Gordan or coupling coefficients $(j_1 j_2 m_1 m_2 | jm)$ are simply replaced by

$$(j_1 j_2 \mu_1 \mu_2 | j\mu) = \sum_{m_1=-j_1}^{j_1} \sum_{m_2=-j_2}^{j_2} \sum_{m=-j}^{j} (j_1 j_2 m_1 m_2 | jm) \\ \langle j_1, m_1 | j_1, \mu_1\rangle^* \, \langle j_2, m_2 | j_2, \mu_2\rangle^* \, \langle j, m | j, \mu\rangle \tag{214}$$

when passing from the $\{jm\}$ quantization to the $\{j\mu\}$ quantization while the recoupling coefficients, and the corresponding $3(n-1) - j$ symbols, for the coupling of n ($n \geq 3$) angular momenta remain invariant. The adaptation to the $\{j\mu\}$ quantization scheme afforded by Eq. (213) is transferable to $SU(2)$ irreducible tensor operators. This yields the Wigner-Eckart theorem in the $\{j\mu\}$ scheme.

We give here the basic ingredients for developing the Wigner-Racah algebra of $SU(2)$ in the $\{j^2, v_{00}\}$ or $\{j^2, x\}$ scheme. For such a scheme, the vector $|j, \mu\rangle$ is of the form $|j\alpha; 00\rangle$ so that the label μ can be identified with α. Thus, the inter-basis expansion coefficients $\langle j, m|j, \mu\rangle$ are

$$\langle j, m | j\alpha; 00\rangle = \frac{1}{\sqrt{2j+1}} q^{(j+m)\alpha} = \frac{1}{\sqrt{2j+1}} \exp\left[\frac{2\pi \mathrm{i}}{2j+1}(j+m)\alpha\right] \tag{215}$$

with $m = j, j-1, \ldots, -j$ and $\alpha = 0, 1, \ldots, 2j$. Equation (215) corresponds to the unitary transformation (45) with $r = a = 0$, that allows to pass from the standard basis B_{2j+1} to the nonstandard basis B_{00}. Then, the Clebsch-Gordan coefficients in the $\{j^2, v_{00}\}$ scheme are

$$(j_1 j_2 \alpha_1 \alpha_2 | j_3 \alpha_3) = \frac{1}{\sqrt{(2j_1+1)(2j_2+1)(2j_3+1)}} \sum_{m_1=-j_1}^{j_1} \sum_{m_2=-j_2}^{j_2} \sum_{m_3=-j_3}^{j_3} \\ (q_1)^{-(j_1+m_1)\alpha_1} \, (q_2)^{-(j_2+m_2)\alpha_2} \, (q_3)^{(j_3+m_3)\alpha_3} \, (j_1 j_2 m_1 m_2 | j_3 m_3) \tag{216}$$

where the various q_k are given in terms of j_k by

$$q_k = \exp\left(\frac{2\pi \mathrm{i}}{2j_k+1}\right), \quad k = 1, 2, 3 \tag{217}$$

The symmetry properties of the coupling coefficients $(j_1 j_2 \alpha_1 \alpha_2 | j_3 \alpha_3)$ cannot be expressed in a simple way (except the symmetry under the interchange $j_1\alpha_1 \leftrightarrow j_2\alpha_2$). Therefore, it is interesting to introduce the following $\overline{f}$ symbol through

$$\overline{f}\begin{pmatrix} j_1 & j_2 & j_3 \\ \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix} = \frac{1}{\sqrt{(2j_1+1)(2j_2+1)(2j_3+1)}} \sum_{m_1=-j_1}^{j_1} \sum_{m_2=-j_2}^{j_2} \sum_{m_3=-j_3}^{j_3}$$
$$(q_1)^{-(j_1+m_1)\alpha_1} (q_2)^{-(j_2+m_2)\alpha_2} (q_3)^{-(j_3+m_3)\alpha_3} \begin{pmatrix} j_1 & j_2 & j_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \tag{218}$$

where the $3 - jm$ symbol on the right-hand side of (218) is an ordinary Wigner symbol for the group $SU(2)$ in the $\{j^2, j_z\}$ scheme. (The $\overline{f}$ symbol is to the $\{j^2, x\}$ scheme what the $\overline{V}$ symbol of Fano and Racah is to the $\{j^2, j_z\}$ scheme, up to a permutation.) The $\overline{f}$ symbol exhibits the same symmetry properties under permutations of its columns as the $3 - jm$ Wigner symbol (identical to the $\overline{V}$ symbol up to a phase factor): Its value is multiplied by $(-1)^{j_1+j_2+j_3}$ under an odd permutation and does not change under an even permutation. In contrast to the $3 - jm$ symbol, not all the values of the $\overline{f}$ symbol are real. In this respect, the $\overline{f}$ symbol behaves under complex conjugation as

$$\overline{f}\begin{pmatrix} j_1 & j_2 & j_3 \\ \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix}^* = (-1)^{j_1+j_2+j_3} (q_1)^{\alpha_1} (q_2)^{\alpha_2} (q_3)^{\alpha_3} \, \overline{f}\begin{pmatrix} j_1 & j_2 & j_3 \\ \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix} \tag{219}$$

Other properties (e.g., orthogonality properties, connection with the Clebsch-Gordan coefficients and the Herring-Wigner tensor, etc.) of the $\overline{f}$ symbol and its relations with $3(n-1) - j$ symbols for $n \geq 3$ can be derived along the lines developed in (Kibler, 1968).

8. References

Albouy, O. (2009). Discrete algebra and geometry applied to the Pauli group and mutually unbiased bases in quantum information theory. *Thesis*, (2009), Université de Lyon

Albouy, O. & Kibler, M.R. (2007). SU(2) nonstandard bases: case of mutually unbiased bases. *SIGMA*, 3, (2007) 076

Arik, M. & Coon, D.D. (1976). Hilbert spaces of analytic functions and generalized coherent states. *J. Math. Phys.*, 17, (1976) 524–527

Atakishiyev, N.M.; Kibler, M.R. & Wolf, K.B. (2010). SU(2) and SU(1,1) approaches to phase operators and temporally stable phase states: applications to mutually unbiased bases and discrete Fourier transforms. *Symmetry*, 2, (2010) 1–24

Balian, R. & Itzykson, C. (1986). Observations sur la mécanique quantique finie. *C. R. Acad. Sci. (Paris)*, 303, (1986) 773–778

Bandyopadhyay, S.; Boykin, P.O.; Roychowdhury, V. & Vatan, F. (2002). A new proof for the existence of mutually unbiased bases. *Algorithmica*, 34, (2002) 512–528

Beckers, J. & Debergh, N. (1990). Parastatistics and supersymmetry in quantum-mechanics. *Nucl. Phys. B*, 340, (1990) 767–776

Bengtsson, I.; Bruzda, W.; Ericsson, Å.; Larsson, J.-Å.; Tadej, W. & Życkowski, K. (2007). Mutually unbiased bases and Hadamard matrices of order six. *J. Math. Phys.*, 48, (2007) 052106

Berndt, B.C.; Evans, R.J. & Williams, K.S. (1998). *Gauss and Jacobi Sums*, Wiley, New York

Boykin, P.O.; Sitharam, M.; Tiep, P.H. & Wocjan, P. (2007). Mutually unbiased bases and orthogonal decompositions of Lie algebras. *Quantum Inf. Comput.*, 7, (2007) 371–382

Brierley, S. & Weigert, S. (2009). Constructing mutually unbiased bases in dimension six. *Phys. Rev. A*, 79, (2009) 052316

Calderbank, A.R.; Cameron, P.J.; Kantor, W.M. & Seidel, J.J. (1997). Z4-Kerdock codes, orthogonal spreads, and extremal Euclidean line-sets. *Proc. London Math. Soc.*, 75, (1997) 436–480

Cerf, N.J.; Bourennane, M.; Karlsson, A. & Gisin, N. (2002). *Phys. Rev. Lett.*, 88, (2002) 127902

Chaichian, M. & Ellinas, D. (1990). On the polar decomposition of the quantum SU(2) algebra. *J. Phys. A: Math. Gen.*, 23, (1990) L291–L296

Champion, J.P.; Pierre, G.; Michelot, F. & Moret-Bailly, J. (1977). Composantes cubiques normales des tenseurs sphériques, *Can. J. Phys.*, 55, (1977) 512–520

Daoud, M.; Hassouni, Y. & Kibler, M. (1998). The k-fermions as objects interpolating between fermions and bosons. In: *Symmetries in Science X*, B. Gruber & M. Ramek (Eds.), 63–77, Plenum Press, New York

Daoud, M. & Kibler, M.R. (2010). Phase operators, temporally stable phase states, mutually unbiased bases and exactly solvable quantum systems. *J. Phys. A: Math. Theor.*, 43, (2010) 115303

Durand, S. (1993). Fractional superspace formulation of generalized mechanics. *Modern Phys. Lett. A*, 8, (1993) 2323–2334

Durt, T.; Englert, B.-G.; Bengtsson, I. & Życkowski, K. (2010). On mutually unbiased bases. *Internat. J. Quantum Info.*, 8, (2010) 535–640

Englert, B.-G. & Aharonov, Y. (2001). The mean king's problem: prime degrees of freedom. *Phys. Lett. A*, 284, (2001) 1–5

Fairlie, D.B.; Fletcher, P. & Zachos, C.K. (1990). Infinite-dimensional algebras and a trigonometric basis for the classical Lie-algebras. *J. Math. Phys.*, 31, (1990) 1088–1094

Gibbons, K.S.; Hoffman, M.J. & Wootters, W.K. (2004). Discrete phase space based on finite fields. *Phys. Rev. A*, 70, (2004) 062101

Gottesman, D.; Kitaev, A. & Preskill, J. (2001). Encoding a qubit in an oscillator. *Phys. Rev. A*, 64, (2001) 012310

Grassl, M. (2005). Tomography of quantum states in small dimensions. *Elec. Notes Discrete Math.*, 20, (2005) 151–164

Havlíček, H. & Saniga, M. (2008). Projective ring line on an arbitrary single qudit. *J. Phys. A: Math. Theor.*, 41, (2008) 015302

Ivanović, I.D. (1981). Geometrical description of quantum state determination. *J. Phys. A: Math. Gen.*, 14, (1981) 3241–3245

Khare, A. (1993). Parasupersymmetry in quantum mechanics. *J. Math. Phys.*, 34, (1993) 1277–1294

Kibler, M. (1968). Ionic and paramagnetic energy levels: algebra. *J. Molec. Spectrosc.*, 26, (1968) 111–130

Kibler, M.R. (2008). Variations on a theme of Heisenberg, Pauli and Weyl. *J. Phys. A: Math. Theor.*, 41, (2008) 375302

Kibler, M.R. (2009). An angular momentum approach to quadratic Fourier transform, Hadamard matrices, Gauss sums, mutually unbiased bases, unitary group and Pauli group. *J. Phys. A: Math. Theor.*, 42, (2009) 353001

Kibler, M.R. & Planat, M. (2006). A SU(2) recipe for mutually unbiased bases. *Internat. J. Modern Phys. B*, 20, (2006) 1802–1807

Klishevich, S. & Plyushchay, T. (1999). Supersymmetry of parafermions. *Modern Phys. Lett. A*, 14, (1999) 2739–2752

Lawrence, J.; Brukner, Č. & Zeilinger, A. (2002). Mutually unbiased binary observable sets on N qubits. *Phys. Rev. A*, 65, (2002) 032320

Lévy-Leblond, J.-M. (1973). Azimuthal quantization of angular momentum. *Rev. Mex. Física*, 22, (1973) 15–23

Mehta, M.L. (1987). Eigenvalues and eigenvectors of the finite Fourier transform. *J. Math. Phys.*, 28, (1987) 781–785

Patera, J. & Winternitz, P. (1976). On bases for irreducible representations of O(3) suitable for systems with an arbitrary finite symmetry group. *J. Chem. Phys.*, 65, (1976) 2725–2731

Patera, J. & Zassenhaus, H. (1988). The Pauli matrices in n dimensions and finest gradings of simple Lie algebras of type A_{n-1}. *J. Math. Phys.*, 29, (1988) 665–673

Pittenger, A.O. & Rubin, M.H. (2004). Mutually unbiased bases, generalized spin matrices and separability. *Linear Algebr. Appl.*, 390, (2004) 255–278

Planat, M. (2010). Pauli graphs when the Hilbert space dimension contains a square: why the Dedekind psi function ? arXiv:1009.3858

Planat, M. & Kibler, M. (2010). Unitary reflection groups for quantum fault tolerance. *J. Comput. Theor. Nanosci.*, 7, (2010) 1–12

Rubakov, V.A. & Spiridonov, V.P. (1988). Parasupersymmetric quantum-mechanics. *Modern Phys. Lett. A*, 3, (1988) 1337–1347

Schwinger, J. (1960). Unitary operator bases. *Proc. Nat. Acad. Sci. USA*, 46, (1960) 570–579

Schwinger, J. (1965). On angular momentum. In: *Quantum Theory of Angular Momemtum*, L.C. Biedenharn & H. van Dam (Eds.), 55–66, Academic Press, New York

Šťovíček, P. & Tolar, J. (1984). Quantum mechanics in a discrete space-time. *Rep. Math. Phys.*, 20, (1984) 157–170

Tolar, J. & Chadzitaskos, G. (2009). Feynman's path integral and mutually unbiased bases. *J. Phys. A: Math. Theor.*, 42, (2009) 245306

Vourdas, A. (1990). SU(2) and SU(1,1) phase states. *Phys. Rev. A*, 41, (1990) 1653–1661

Vourdas, A. (2004). Quantum systems with finite Hilbert space. *Rep. Prog. Phys.*, 67, (2004) 267–320

Weyl, H. (1931). *The Theory of Groups and Quantum Mechanics*, Dover Publications, New York

Wootters, W.K. & Fields, B.D. (1989). Optimal state-determination by mutually unbiased measurements. *Ann. Phys. (N.Y.)*, 191, (1989) 363–381

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

8

Orthogonal Discrete Fourier and Cosine Matrices for Signal Processing

Daechul Park[1] and Moon Ho Lee[2]
[1]Hannam University,
[2]Chonbuk National University
Korea

1. Introduction

The A DFT (Discrete Fourier Transform) has seen studied and applied to signal processing and communication theory. The relation between the Fourier matrix and the Hadamard transform was developed in [Ahmed & Rao, 1975; Whelchel & Guinn, 1968] for signal representation and classification and the Fast Fourier-Hadamand Transform(FFHT) was proposed. This idea was further investigated in [Lee & Lee, 1998] as an extension of the conventional Hadamard matrix. Lee et al [Lee & Lee, 1998] has proposed the Reverse Jacket Transform(RJT) based on the decomposition of the Hadamard matrix into the Hadamard matrix(unitary matrix) itself and a sparse matrix. Interestingly, the Reverse Jacket(RJ) matrix has a strong geometric structure that reveals a circulant expansion and contraction properties from a basic 2x2 sparse matrix.

The discrete Fourier transform (DFT) is an orthogonal matrix with highly practical value for representing signals and images [Ahmed & Rao, 1975; Lee, 1992; Lee, 2000]. Recently, the Jacket matrices which generalize the weighted Hadamard matrix were introduced in [Lee, 2000], [Lee & Kim, 1984, Lee, 1989, Lee & Yi, 2001; Fan & Yang, 1998]. The Jacket matrix[1] is an abbreviated name of a reverse Jacket geometric structure. It includes the conventional Hadamard matrix [Lee, 1992; Lee, 2000; Lee et al., 2001; Hou et. al., 2003], but has the weights, ω, that are j or 2^k, where k is an integer, and $j=\sqrt{-1}$, located in the central part of Hadamard matrix. The weighted elements' positions of the forward matrix can be replaced by the non-weighted elements of its inverse matrix and the signs of them do not change between the forward and inverse matrices, and they are only as element inverse and transpose. This reveals an interesting complementary matrix relation.

Definition 1: If a matrix $[J]_m$ of size $m \times m$ has nonzero elements

$$[J]_m = \begin{bmatrix} j_{0,0} & j_{0,1} & \cdots & j_{0,m-1} \\ j_{1,0} & j_{1,1} & \cdots & j_{1,m-1} \\ \vdots & \vdots & & \vdots \\ j_{m-1,0} & j_{m-1,1} & \cdots & j_{m-1,m-1} \end{bmatrix}, \tag{6-1}$$

$$\left[J\right]_m^{-1} = \frac{1}{C}\begin{bmatrix} 1/j_{0,0} & 1/j_{0,1} & \cdots & 1/j_{0,m-1} \\ 1/j_{1,0} & 1/j_{1,1} & \cdots & 1/j_{1,m-1} \\ \vdots & \vdots & & \vdots \\ 1/j_{m-1,0} & 1/j_{m-1,1} & \cdots & 1/j_{m-1,m-1} \end{bmatrix}^T \tag{6-2}$$

where C is the normalizing constant, and T is of matrix transposition, then the matrix $\left[J\right]_m$ is called a Jacket matrix [Whelchel & Guinn, 1968],[Lee et al., 2001],[Lee et al, 2008; Chen et al., 2008]. Especially orthogonal matrices, such as Hadamard, DFT, DCT, Haar, and Slant matrices belong to the Jacket matrices family [Lee et al., 2001]. In addition, the Jacket matrices are associated with many kind of matrices, such as unitary matrices, and Hermitian matrices which are very important in communication (e.g., encoding), mathematics, and physics.

In section 2 DFT matrix is revisited in the sense of sparse matrix factorization. Section 3 presents recursive factorization algorithms of DFT and DCT matrix for fast computation. Section 4 proposes a hybrid architecture for implentation of algorithms simply by adding a switching device on a single chip module. Lastly, conclusions were drawn in section 4.

2. Preliminary of DFT presentation

The discrete Fourier transform (DFT) is a Fourier representation of a given sequence *x(m),* $0 \le m \le N-1$ and is defined as

$$X(n) = \sum_{m=0}^{N-1} x(m)W^{nm},\ 0 \le n \le N-1\ , \tag{6-3}$$

where $W = e^{-j\frac{2\pi}{N}}$. Let's denote N -point DFT matrix as $F_N = \left[W^{nm}\right]_N$, $n,m = \{0,1,2,\ldots,N-1\}$,

where $W = e^{-j\frac{2\pi}{N}}$ (see about DFT in appendix), and the $N \times N$ Sylvester Hadamard matrix as $\left[H\right]_N$, respectively. The Sylvester Hadamard matrix is generated recursively by successive Kronecker products,

$$\left[H\right]_N = \left[H\right]_2 \otimes \left[H\right]_{N/2}\ , \tag{6-4}$$

for N=4, 8, 16, … and $\left[H\right]_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. For the remainder of this chapter, analysis will be concerned only with *N=2^k^, k=1,2,3,* … as the dimensionality of both the *F* and *H* matrices.

Definition 2: A sparse matrix $\left[S\right]_N$, which relates $\left[F\right]_N$ and $\left[H\right]_N$, can be computed from the factorization of *F* based on *H*.

The structure of the *S* matrix is rather obscure. However, a much less complex and more appealing relationship will be identified for *S* [Park et al., 1999].

To illustrate the DFT using direct product we alter the denotation of *W* to lower case $w = e^{-j2\pi}$, so that $w^{n/N}$ becomes the n-th root of unit for N -point W. For instance, the DFT matrix of dimension 2 is given by:

$$\left[F\right]_2 = \begin{bmatrix} w^{0/2} & w^{0/2} \\ w^{0/2} & w^{1/2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \left[H\right]_2 \ . \tag{6-5}$$

Let’s define

$$\left[W\right]_2 = \begin{bmatrix} w^{0/4} & 0 \\ 0 & w^{1/4} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -j \end{bmatrix} \text{ and } \left[E\right]_2 = \left[F\right]_2 \left[W\right]_2 = \begin{bmatrix} 1 & -j \\ 1 & j \end{bmatrix}, \tag{6-6}$$

and in general, We define

$$\left[W\right]_N = diag(w^{0/2N}, w^{1/2N}, w^{2/2N}, ..., w^{(N-1)/2N})$$

and

$$\left[E\right]_N = \left[F\right]_N \left[W\right]_N = [P]_N^T [\tilde{F}]_N [W]_N \,. \tag{6-7}$$

where $[P]_N$ is a permutation matrix and $[\tilde{F}]_N = [P]_N [F]_N$ is a permuted version of DFT matrix $[F]_N$.

3. A sparse matrix factorization of orthogonal transforms

3.1 A sparse matrix analysis of discrete Fourier transform

Now we will present the Jacket matrix from a direct product of a sparse matrix computation and representation given by [Lee, 1989], [Lee & Finlayson, 2007]

$$\left[J\right]_m = \frac{1}{m} \left[H\right]_m \left[S\right]_m , \tag{6-8}$$

where $m = 2^{k+1}, k \in \{1,2,3,4,...\}$ and $\left[S\right]_m$ is sparse matrix of $\left[J\right]_m$. Thus the inverse of the Jacket matrix can be simply written as

$$\left(\left[J\right]_m\right)^{-1} = \left(\left[S\right]_m\right)^{-1} \left[H\right]_m . \tag{6-9}$$

As mentioned previously, the DFT matrix is also a Jacket matrix. By considering the sparse matrix for the 4-piont DFT matrix $\left[F\right]_4$,

$$\left[F\right]_4 = \begin{bmatrix} W^0 & W^0 & W^0 & W^0 \\ W^0 & W^1 & W^2 & W^3 \\ W^0 & W^2 & W^4 & W^6 \\ W^0 & W^3 & W^6 & W^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & e^{-j\frac{\pi}{2}} & e^{-j\frac{\pi}{2}\times 2} & e^{-j\frac{\pi}{2}\times 3} \\ 1 & e^{-j\frac{\pi}{2}\times 2} & e^{-j\frac{\pi}{2}\times 4} & e^{-j\frac{\pi}{2}\times 6} \\ 1 & e^{-j\frac{\pi}{2}\times 3} & e^{-j\frac{\pi}{2}\times 6} & e^{-j\frac{\pi}{2}\times 9} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & j \\ 1 & -1 & 1 & -1 \\ 1 & j & -1 & -j \end{bmatrix}.$$

we can rewrite $\left[F\right]_4$ by using permutations as

$$\left[\tilde{F}\right]_4 = \left[\Pr\right]_4 \left[F\right]_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & j \\ 1 & -1 & 1 & -1 \\ 1 & j & -1 & -j \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -j & -1 & j \\ 1 & j & -1 & -j \end{bmatrix} = \begin{bmatrix} [\tilde{F}]_2 & [\tilde{F}]_2 \\ E_2 & -E_2 \end{bmatrix}$$

$$= \left(\begin{bmatrix} I_2 & I_2 \\ I_2 & -I_2 \end{bmatrix} \begin{bmatrix} [\tilde{F}]_2 & 0 \\ 0 & E_2 \end{bmatrix} \right)^T , \qquad (6\text{-}10)$$

where $E_2 = \begin{bmatrix} 1 & -j \\ 1 & j \end{bmatrix}$, its inverse matrix is from element-inverse, such that

$$\left(E_2\right)^{-1} = \begin{bmatrix} 1 & 1 \\ j & -j \end{bmatrix} = \left(\begin{bmatrix} 1/1 & -1/j \\ 1/1 & 1/j \end{bmatrix} \right)^T . \qquad (6\text{-}11)$$

In general, we can write that

$$\left[\tilde{F}\right]_N = \left[\Pr\right]_N \left[F\right]_N = \begin{bmatrix} [\tilde{F}]_{N/2} & [\tilde{F}]_{N/2} \\ E_{N/2} & -E_{N/2} \end{bmatrix} = \left(\begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \begin{bmatrix} [\tilde{F}]_{N/2} & 0 \\ 0 & E_{N/2} \end{bmatrix} \right)^T , \qquad (6\text{-}12)$$

where $\left[F\right]_2 = \left[\tilde{F}\right]_2$. And the submatrix E_N could be written from (6-7) by

$$\left[E\right]_N = [F]_N [W]_N = \left[\Pr\right]_N \left[\tilde{F}\right]_N \left[W\right]_N , \qquad (6\text{-}13)$$

where $\left[W\right]_N = \begin{bmatrix} W^0 & 0 & \cdots & 0 \\ 0 & W^1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & W^{N-1} \end{bmatrix}$, and W is the complex unit for $2N$ point DFT matrix.

For example, $\left[E\right]_2 = \begin{bmatrix} 1 & -j \\ 1 & j \end{bmatrix}$ can be calculated by using

$$\left[E\right]_2 = \left[\Pr\right]_2 \left[\tilde{F}\right]_2 \left[W\right]_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} W^0 & 0 \\ 0 & W^1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -j \end{bmatrix} = \begin{bmatrix} 1 & -j \\ 1 & j \end{bmatrix} . \qquad (6\text{-}14)$$

By using the results from (6-12) and (6-13), we have a new DFT matrix decomposition as

$$\left[\tilde{F}\right]_N = \left[\Pr\right]_N \left[F\right]_N = \left(\begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \begin{bmatrix} [\tilde{F}]_{N/2} & 0 \\ 0 & E_{N/2} \end{bmatrix} \right)^T$$

$$= \begin{bmatrix} [\tilde{F}]_{N/2} & 0 \\ 0 & E_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix}$$

$$
\begin{aligned}
\left[\tilde{F}\right]_N &= \begin{bmatrix} [\tilde{F}]_{N/2} & 0 \\ 0 & \Pr_{N/2}[\tilde{F}]_{N/2}W_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \\
&= \begin{bmatrix} I_{N/2} & 0 \\ 0 & \Pr_{N/2} \end{bmatrix} \begin{bmatrix} [\tilde{F}]_{N/2} & 0 \\ 0 & [\tilde{F}]_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & 0 \\ 0 & W_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix}.
\end{aligned} \tag{6-15}
$$

Finally, based on the recursive form we have

$$
\begin{aligned}
\left[F\right]_N &= \left(\left[\Pr\right]_N\right)^{-1}\left[\tilde{F}\right]_N = \left(\left[\Pr\right]_N\right)^{T} \begin{bmatrix} I_{N/2} & 0 \\ 0 & \Pr_{N/2} \end{bmatrix} \begin{bmatrix} [\tilde{F}]_{N/2} & 0 \\ 0 & [\tilde{F}]_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & 0 \\ 0 & W_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \\
&= \left(\left[\Pr\right]_N\right)^{T} \begin{bmatrix} I_{N/2} & 0 \\ 0 & \Pr_{N/2} \end{bmatrix} \cdots \left[I_{N/4} \otimes \begin{bmatrix} I_2 & 0 \\ 0 & \Pr_2 \end{bmatrix} \right] \left[I_{N/2} \otimes F_2 \right] \\
&\quad \left[I_{N/4} \otimes \begin{bmatrix} I_2 & 0 \\ 0 & W_2 \end{bmatrix} \right] \left[I_{N/4} \otimes \begin{bmatrix} I_2 & I_2 \\ I_2 & -I_2 \end{bmatrix} \right] .. \begin{bmatrix} I_{N/2} & 0 \\ 0 & W_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix}.
\end{aligned} \tag{6-16}
$$

Using (6-16) butterfly data flow diagram for DFT transform is drawn from left to right to perform $X = [F]_N \mathbf{x}$.

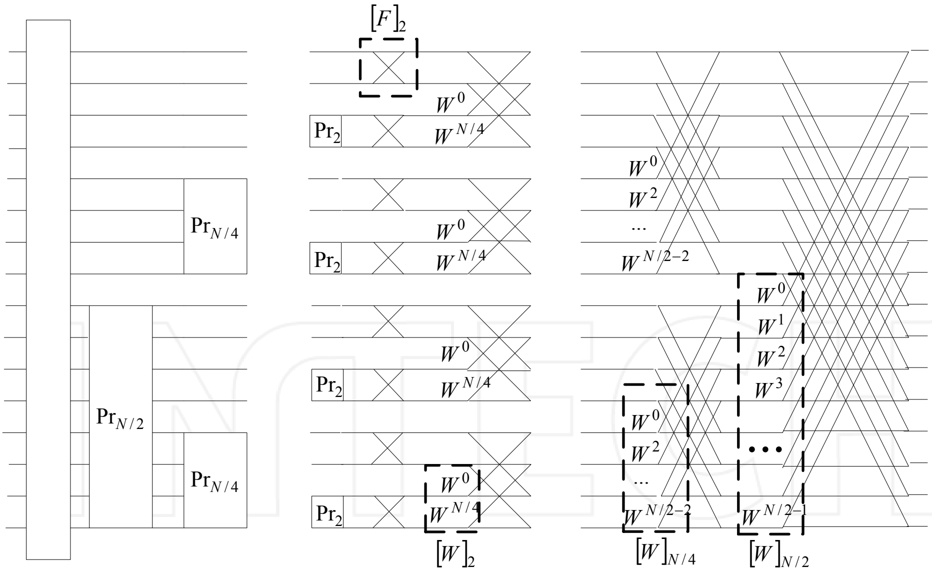


Fig. 1. Butterfly data flow diagram of proposed DFT matrix with order N

3.2 A Sparse matrix analysis of discrete cosine transform

Similar to the section 3.1, we will present the DCT matrices by using the element inverse or block inverse Jacket like sparse matrix [Lee, 2000; Park et al., 1999] decomposition. In this section, the simple construction and fast computation for forward and inverse calculations and analysis of the sparse matrices, was very useful for developing the fast algorithms and orthogonal codes design.

Discrete Cosine Transform (DCT) is widely used in image processing, and orthogonal transform. There are four typical DCT matrices [Rao & Yip, 1990; Rao & Hwang, 1996],

$$\text{DCT - I: } \left[C_{N+1}^{I}\right]_{m,n} = \sqrt{\frac{2}{N}} k_m k_n \cos\frac{mn\pi}{N}, \quad m,n = 0,1,\ldots,N \quad ; \tag{6-17}$$

$$\text{DCT - II: } \left[C_{N}^{II}\right]_{m,n} = \sqrt{\frac{2}{N}} k_m \cos\frac{m(n+\frac{1}{2})\pi}{N}, \quad m,n = 0,1,\ldots,N-1 \quad ; \tag{6-18}$$

$$\text{DCT - III: } \left[C_{N}^{III}\right]_{m,n} = \sqrt{\frac{2}{N}} k_n \cos\frac{(m+\frac{1}{2})n\pi}{N}, \quad m,n = 0,1,\ldots,N-1 \quad ; \tag{6-19}$$

$$\text{DCT - IV: } \left[C_{N}^{IV}\right]_{m,n} = \sqrt{\frac{2}{N}} \cos\frac{(m+\frac{1}{2})(n+\frac{1}{2})\pi}{N}, \quad m,n = 0,1,\ldots,N-1 \quad , \tag{6-20}$$

where

$$k_j = \begin{cases} 1, & j = 1,2,\ldots,N-1 \\ \dfrac{1}{\sqrt{2}}, & j = 0,N \end{cases}.$$

To describe the computations of DCT, in this chapter, we will focus on the DCT - II algorithm, and introduce the sparse matrix decomposition and fast computations.

The 2-by-2 DCT - II matrix can be simply written as

$$\left[C\right]_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ C_4^1 & C_4^3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \frac{1}{\sqrt{2}}, \tag{6-21}$$

where $1/\sqrt{2}$ can be seen as a special element inverse matrix of order 1, its inverse is $\sqrt{2}$, and $C_l^i = \cos(i\pi / l)$ is the cosine unit for DCT computations.

Furthermore, 4-by-4 DCT - II matrix is of the form

$$\left[C\right]_4 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ C_8^1 & C_8^3 & C_8^5 & C_8^7 \\ C_8^2 & C_8^6 & C_8^6 & C_8^2 \\ C_8^3 & C_8^7 & C_8^1 & C_8^5 \end{bmatrix}, \tag{6-22}$$

we can write

$$\left[P\right]_4\left[C\right]_4=\begin{bmatrix}1&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\C_8^1&C_8^3&C_8^5&C_8^7\\C_8^2&C_8^6&C_8^6&C_8^2\\C_8^3&C_8^7&C_8^1&C_8^5\end{bmatrix}=\begin{bmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\C_8^2&C_8^6&C_8^6&C_8^2\\C_8^1&C_8^3&C_8^5&C_8^7\\C_8^3&C_8^7&C_8^1&C_8^5\end{bmatrix}, \qquad (6\text{-}23)$$

where $\left[P\right]_4=\begin{bmatrix}1&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&1\end{bmatrix}$ is permutation matrix. $\left[P\right]_N$ permutation matrix is a special case which has the form

$$\left[P\right]_2=\left[I\right]_2\text{, and }\left[P\right]_N=\begin{bmatrix}1&0&0&\cdots&0&0&\cdots&0\\0&0&0&\cdots&1&0&\cdots&0\\0&1&0&&0&0&\cdots&0\\0&0&0&&0&1&&0\\0&0&1&&0&0&&0\\0&0&0&&0&0&\ddots&0\\0&0&0&\ddots&0&0&&0\\0&0&0&0&0&0&&1\end{bmatrix}, \quad N\geq 4,$$

where

$$\left[P\right]_N=\left[pr_{i,j}\right]_N,$$

with

$$\begin{cases}pr_{i,j}=1, & if \quad i=2j, \quad 0\leq j\leq \frac{N}{2}-1,\\ pr_{i,j}=1, & if \quad i=(2j+1)\operatorname{mod}N, \quad \frac{N}{2}\leq j\leq N-1,\\ pr_{i,j}=0, & others.\end{cases}$$

where $i,j\in\{0,1,...,N-1\}$.
Since $C_8^1=-C_8^7, C_8^2=-C_8^6, C_8^3=-C_8^5$, we rewrite (6-23) as

$$\left[P\right]_4\left[C\right]_4=\begin{bmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\C_8^2&C_8^6&C_8^6&C_8^2\\C_8^1&C_8^3&C_8^5&C_8^7\\C_8^3&C_8^7&C_8^1&C_8^5\end{bmatrix}=\begin{bmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\C_8^2&-C_8^2&-C_8^2&C_8^2\\C_8^1&C_8^3&-C_8^3&-C_8^1\\C_8^3&-C_8^1&C_8^1&-C_8^3\end{bmatrix}, \qquad (6\text{-}24)$$

and let us define a column permutation matrix $\left[Pc\right]_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, and $\left[Pc\right]_N$ is a reversible permutation matrix which is defined by

$$\left[Pc\right]_2 = \left[I\right]_2 \text{, and}$$

$$\left[Pc\right]_N = \begin{bmatrix} I_{N/4} & 0 & 0 & 0 \\ 0 & I_{N/4} & 0 & 0 \\ 0 & 0 & 0 & I_{N/4} \\ 0 & 0 & I_{N/4} & 0 \end{bmatrix}, \; N \geq 4 .$$

Thus we have

$$\left[P\right]_4 \left[C\right]_4 \left[Pc\right]_4 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ C_8^2 & C_8^6 & C_8^6 & C_8^2 \\ C_8^1 & C_8^3 & C_8^5 & C_8^7 \\ C_8^3 & C_8^7 & C_8^1 & C_8^5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ C_8^2 & -C_8^2 & C_8^2 & -C_8^2 \\ C_8^1 & C_8^3 & -C_8^1 & -C_8^3 \\ C_8^3 & -C_8^1 & -C_8^3 & C_8^1 \end{bmatrix} = \begin{bmatrix} C_2 & C_2 \\ B_2 & -B_2 \end{bmatrix}, \qquad \text{(6-25)}$$

where $\left[C\right]_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$, the 2-by-2 DCT - II matrix and $\left[B\right]_2 = \begin{bmatrix} C_8^1 & C_8^3 \\ C_8^3 & -C_8^1 \end{bmatrix}$. Thus we can write that

$$\left[P\right]_4 \left[C\right]_4 \left[Pc\right]_4 = \begin{bmatrix} C_2 & C_2 \\ B_2 & -B_2 \end{bmatrix} = \left(\begin{bmatrix} I_2 & I_2 \\ I_2 & -I_2 \end{bmatrix} \begin{bmatrix} C_2 & 0 \\ 0 & B_2 \end{bmatrix} \right)^T, \qquad \text{(6-26)}$$

it is clear that $\begin{bmatrix} C_2 & 0 \\ 0 & B_2 \end{bmatrix}$ is a block inverse matrix, which has

$$\begin{bmatrix} C_2 & 0 \\ 0 & B_2 \end{bmatrix}^{-1} = \begin{bmatrix} \left(C_2\right)^{-1} & 0 \\ 0 & \left(B_2\right)^{-1} \end{bmatrix}. \qquad \text{(6-27)}$$

The (6-27) is Jacket –like sparse matrix with block inverse.

In general the permuted DCT - II matrix $\left[\tilde{C}\right]_N$ can be constructed recursively by using

$$\left[\tilde{C}\right]_N = \left[P\right]_N \left[C\right]_N \left[Pc\right]_N = \begin{bmatrix} C_{N/2} & C_{N/2} \\ B_{N/2} & -B_{N/2} \end{bmatrix} = \left(\begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \begin{bmatrix} C_{N/2} & 0 \\ 0 & B_{N/2} \end{bmatrix} \right)^T . \tag{6-28}$$

where $\left[C\right]_{N/2}$ denotes the $\frac{N}{2} \times \frac{N}{2}$ DCT - II matrix, and $\left[B\right]_{N/2}$ can be calculated by using

$$\left[B\right]_{N/2} = \left[\left(C_{2N}^{f(m,n)} \right)_{m,n} \right]_{N/2} , \tag{6-29}$$

where

$$\begin{cases} f(m,1) = 2m - 1, \\ f(m,n+1) = f(m,n) + f(m,1) \times 2, \end{cases} \quad m,n \in \{1,2,\ldots,N/2\} . \tag{6-30}$$

For example, in the 4-by-4 permuted DCT - II matrix $\left[\tilde{C}\right]_4$, B_2 could be calculated by using $f(1,1) = 1$, $f(2,1) = 3$, $f(1,2) = f(1,1) + f(1,1) \times 2 = 3$, and $f(2,2) = f(2,1) + f(2,1) \times 2 = 9$,

$$\left[B\right]_2 = \left[\left(C_8^{f(m,n)} \right)_{m,n} \right]_4 = \begin{bmatrix} \left(C_8^{f(1,1)} \right)_{1,1} & \left(C_8^{f(1,2)} \right)_{1,2} \\ \left(C_8^{f(2,1)} \right)_{2,1} & \left(C_8^{f(2,2)} \right)_{2,2} \end{bmatrix} . \tag{6-31}$$

and its inverse is of $\left[\tilde{C}\right]_N$ can be simply computed from the block inverse

$$\begin{aligned} \left(\left[P\right]_N \left[C\right]_N \left[Pc\right]_N \right)^{-1} &= \left(\left(\begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \begin{bmatrix} C_{N/2} & 0 \\ 0 & B_{N/2} \end{bmatrix} \right)^T \right)^{-1} \\ &= \left(\begin{bmatrix} \left(C_{N/2} \right)^{-1} & 0 \\ 0 & \left(B_{N/2} \right)^{-1} \end{bmatrix} \begin{bmatrix} \frac{2}{N} I_{N/2} & \frac{2}{N} I_{N/2} \\ \frac{2}{N} I_{N/2} & -\frac{2}{N} I_{N/2} \end{bmatrix} \right)^T . \\ &= \frac{2}{N} \left(\begin{bmatrix} \left(C_{N/2} \right)^{-1} & 0 \\ 0 & \left(B_{N/2} \right)^{-1} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \right)^T \end{aligned} \tag{6-32}$$

For example, the 8×8 DCT - II matrix has

$$[C]_8 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ C_{16}^{1} & C_{16}^{3} & C_{16}^{5} & C_{16}^{7} & C_{16}^{9} & C_{16}^{11} & C_{16}^{13} & C_{16}^{15} \\ C_{16}^{2} & C_{16}^{6} & C_{16}^{10} & C_{16}^{14} & C_{16}^{18} & C_{16}^{22} & C_{16}^{26} & C_{16}^{30} \\ C_{16}^{3} & C_{16}^{9} & C_{16}^{15} & C_{16}^{21} & C_{16}^{27} & C_{16}^{33} & C_{16}^{39} & C_{16}^{45} \\ C_{16}^{4} & C_{16}^{12} & C_{16}^{20} & C_{16}^{28} & C_{16}^{36} & C_{16}^{44} & C_{16}^{52} & C_{16}^{60} \\ C_{16}^{5} & C_{16}^{15} & C_{16}^{25} & C_{16}^{35} & C_{16}^{45} & C_{16}^{55} & C_{16}^{65} & C_{16}^{75} \\ C_{16}^{6} & C_{16}^{18} & C_{16}^{30} & C_{16}^{42} & C_{16}^{54} & C_{16}^{66} & C_{16}^{78} & C_{16}^{90} \\ C_{16}^{7} & C_{16}^{21} & C_{16}^{35} & C_{16}^{49} & C_{16}^{63} & C_{16}^{77} & C_{16}^{91} & C_{16}^{105} \end{bmatrix},$$

and it can be represented by

$$[P]_8[C]_8[Pc]_8 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ C_{16}^{4} & -C_{16}^{4} & C_{16}^{4} & -C_{16}^{4} & C_{16}^{4} & -C_{16}^{4} & C_{16}^{4} & -C_{16}^{4} \\ C_{16}^{2} & C_{16}^{6} & -C_{16}^{2} & -C_{16}^{6} & C_{16}^{2} & C_{16}^{6} & -C_{16}^{2} & -C_{16}^{6} \\ C_{16}^{6} & -C_{16}^{2} & -C_{16}^{6} & C_{16}^{2} & C_{16}^{6} & -C_{16}^{2} & -C_{16}^{6} & C_{16}^{2} \\ C_{16}^{1} & C_{16}^{3} & -C_{16}^{7} & -C_{16}^{5} & -C_{16}^{1} & -C_{16}^{3} & C_{16}^{7} & C_{16}^{5} \\ C_{16}^{5} & -C_{16}^{1} & -C_{16}^{3} & -C_{16}^{7} & -C_{16}^{5} & C_{16}^{1} & C_{16}^{3} & C_{16}^{7} \\ C_{16}^{3} & -C_{16}^{7} & C_{16}^{5} & C_{16}^{1} & -C_{16}^{3} & C_{16}^{7} & -C_{16}^{5} & -C_{16}^{1} \\ C_{16}^{7} & -C_{16}^{5} & C_{16}^{1} & -C_{16}^{3} & -C_{16}^{7} & C_{16}^{5} & -C_{16}^{1} & C_{16}^{3} \end{bmatrix} = \begin{bmatrix} C_4 & C_4 \\ B_4 & -B_4 \end{bmatrix}.$$

Additionally, it is clearly that the function (6-28) also can be recursively constructed by using different permutations matrices $\left[\tilde{P}\right]_N$ and $\left[\tilde{P}c\right]_N$, as

$$\left[\tilde{P}\right]_N\left[\tilde{C}\right]_N\left[\tilde{P}c\right]_N = \begin{bmatrix} \tilde{C}_{N/2} & \tilde{C}_{N/2} \\ \tilde{B}_{N/2} & -\tilde{B}_{N/2} \end{bmatrix} = \left(\begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix}\begin{bmatrix} \tilde{C}_{N/2} & 0 \\ 0 & \tilde{B}_{N/2} \end{bmatrix}\right)^T, \tag{6-33}$$

where $\left[\tilde{C}\right]_{N/2}$ and $\left[\tilde{B}\right]_{N/2}$ are the permutated cases of $\left[C\right]_{N/2}$ and $\left[B\right]_{N/2}$, respectively. The new permutation matrices have the form

$$\left[\tilde{P}\right]_N = \begin{bmatrix} [P]_{N/2} & 0 \\ 0 & I_{N/2} \end{bmatrix}, \text{and } \left[\tilde{P}c\right]_N = \begin{bmatrix} [Pc]_{N/2} & 0 \\ 0 & [Pc]_{N/2} \end{bmatrix}, \ N > 4\ . \tag{6-34}$$

Easily, we can check that

$$\left[\tilde{P}\right]_N\left[\tilde{C}\right]_N\left[\tilde{P}c\right]_N = \begin{bmatrix} P_{N/2} & 0 \\ 0 & I_{N/2} \end{bmatrix}\begin{bmatrix} C_{N/2} & C_{N/2} \\ B_{N/2} & -B_{N/2} \end{bmatrix}\begin{bmatrix} Pc_{N/2} & 0 \\ 0 & Pc_{N/2} \end{bmatrix}$$

$$= \begin{bmatrix} P_{N/2}C_{N/2} & \Pr_{N/2}C_{N/2} \\ I_{N/2}B_{N/2} & -I_{N/2}B_{N/2} \end{bmatrix}\begin{bmatrix} Pc_{N/2} & 0 \\ 0 & Pc_{N/2} \end{bmatrix}$$

$$\left[\tilde{P}\right]_N\left[\tilde{C}\right]_N\left[\tilde{Pc}\right]_N=\begin{bmatrix}P_{N/2}C_{N/2}Pc_{N/2} & P_{N/2}C_{N/2}Pc_{N/2}\\ I_{N/2}B_{N/2}Pc_{N/2} & -I_{N/2}B_{N/2}Pc_{N/2}\end{bmatrix}=\begin{bmatrix}\tilde{C}_{N/2} & \tilde{C}_{N/2}\\ \tilde{B}_{N/2} & -\tilde{B}_{N/2}\end{bmatrix}, \qquad (6\text{-}35)$$

where $\left[\tilde{B}\right]_{N/2}=\left[B\right]_{N/2}\left[Pc\right]_{N/2}$.

For example, the 4-by-4 DCT - II case has

$$\left[\tilde{B}\right]_{4/2}=\left[B\right]_{4/2}\left[Pc\right]_{4/2}=\begin{bmatrix}C_8^1 & C_8^3\\ C_8^3 & -C_8^1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}=\begin{bmatrix}C_8^1 & C_8^3\\ C_8^3 & -C_8^1\end{bmatrix}. \qquad (6\text{-}36)$$

Moreover, the matrix $\left[B\right]_2=\begin{bmatrix}C_8^1 & C_8^3\\ C_8^3 & -C_8^1\end{bmatrix}$ can be decomposed by using the 2-by-2 DCT - II matrix as

$$\left[B\right]_2=\begin{bmatrix}C_8^1 & C_8^3\\ C_8^3 & -C_8^1\end{bmatrix}=\left[K\right]_2\left[C\right]_2\left[D\right]_2=\begin{bmatrix}\sqrt{2} & 0\\ -\sqrt{2} & 2\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ C_8^2 & -C_8^2\end{bmatrix}\begin{bmatrix}C_8^1 & 0\\ 0 & C_8^3\end{bmatrix}, \qquad (6\text{-}37)$$

where $\left[K\right]_2=\begin{bmatrix}\sqrt{2} & 0\\ -\sqrt{2} & 2\end{bmatrix}$ is a upper triangular matrix, $\left[D\right]_2=\begin{bmatrix}C_8^1 & 0\\ 0 & C_8^3\end{bmatrix}$ is a diagonal matrix, and we use the cosine related function

$$\cos(2k+1)\phi_m=2\cos(2k\phi_m)\cos\phi_m-\cos(2k-1)\phi_m \quad . \qquad (6\text{-}38)$$

where ϕ_m is m-th angle.

In a general case, we have

$$\left[B\right]_N=\left[K\right]_N\left[C\right]_N\left[D\right]_N, \qquad (6\text{-}39)$$

where

$$\left[K\right]_N=\begin{bmatrix}\sqrt{2} & 0 & 0 & \cdots\\ -\sqrt{2} & 2 & 0 & \cdots\\ \sqrt{2} & -2 & 2 & \cdots\\ \vdots & \vdots & & \ddots\end{bmatrix}, \left[D\right]_N=\begin{bmatrix}C_{4N}^{\Phi_0} & 0 & \cdots & 0\\ 0 & C_{4N}^{\Phi_1} & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & C_{4N}^{\Phi_{N-1}}\end{bmatrix}, \qquad (6\text{-}40)$$

and $\Phi_i=2i+1$, $i\in\{0,1,2,...,N-1\}$. See appendix 2 for proof of (6-39).

By using the results from (6-28) and (6-39), we have a new form for DCT - II matrix

$$\left[\tilde{C}\right]_N=\left[\Pr\right]_N\left[C\right]_N\left[Pc\right]_N=\left(\begin{bmatrix}I_{N/2} & I_{N/2}\\ I_{N/2} & -I_{N/2}\end{bmatrix}\begin{bmatrix}C_{N/2} & 0\\ 0 & B_{N/2}\end{bmatrix}\right)^T=\begin{bmatrix}C_{N/2} & 0\\ 0 & B_{N/2}\end{bmatrix}\begin{bmatrix}I_{N/2} & I_{N/2}\\ I_{N/2} & -I_{N/2}\end{bmatrix}$$

$$
\begin{aligned}
\left[\tilde{C}\right]_N &= \begin{bmatrix} C_{N/2} & 0 \\ 0 & K_{N/2}C_{N/2}D_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \\
&= \begin{bmatrix} I_{N/2} & 0 \\ 0 & K_{N/2} \end{bmatrix} \begin{bmatrix} C_{N/2} & 0 \\ 0 & C_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & 0 \\ 0 & D_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix}.
\end{aligned} \quad (6\text{-}41)
$$

Given the recursive form of (6-41), we can write

$$
\begin{aligned}
\left[C\right]_N = &\left(\left[\mathrm{Pr}\right]_N\right)^{-1} \begin{bmatrix} I_{N/2} & 0 \\ 0 & K_{N/2} \end{bmatrix} \cdots \left[I_{N/4} \otimes \left(\left[\mathrm{Pr}\right]_4\right)^{-1} \right] \left[I_{N/4} \otimes \begin{bmatrix} I_2 & 0 \\ 0 & K_2 \end{bmatrix} \right] \left[I_{N/2} \otimes C_2 \right] \\
&\left[I_{N/4} \otimes \begin{bmatrix} I_2 & 0 \\ 0 & D_2 \end{bmatrix} \right] \left[I_{N/4} \otimes \begin{bmatrix} I_2 & I_2 \\ I_2 & -I_2 \end{bmatrix} \right] \left[I_{N/4} \otimes \left(\left[Pc\right]_4\right)^{-1} \right] \\
&\cdots \begin{bmatrix} I_{N/2} & 0 \\ 0 & D_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \left(\left[Pc\right]_N\right)^{-1}
\end{aligned} . \quad (6\text{-}42)
$$

By taking all permutation matrices outside, we can rewrite (6-42) as

$$
\begin{aligned}
\left[C\right]_N = &\left(\left[\tilde{\mathrm{P}}\mathrm{r}\right]_N\right)^{-1} \begin{bmatrix} I_{N/2} & 0 \\ 0 & K_{N/2} \end{bmatrix} \cdots \left[I_{N/4} \otimes \begin{bmatrix} I_2 & 0 \\ 0 & K_2 \end{bmatrix} \right] \left[I_{N/2} \otimes C_2 \right] \\
&\left[I_{N/4} \otimes \begin{bmatrix} I_2 & 0 \\ 0 & D_2 \end{bmatrix} \right] \left[I_{N/4} \otimes \begin{bmatrix} I_2 & I_2 \\ I_2 & -I_2 \end{bmatrix} \right] .. \begin{bmatrix} I_{N/2} & 0 \\ 0 & D_{N/2} \end{bmatrix} \begin{bmatrix} I_{N/2} & I_{N/2} \\ I_{N/2} & -I_{N/2} \end{bmatrix} \left(\left[\tilde{P}c\right]_N\right)^{-1}
\end{aligned} . \quad (6\text{-}43)
$$

Using (6-43) butterfly data flow diagram for DCT-II transform is drawn as Fig.2 from left to right to perform X=[C]$_N$ x.

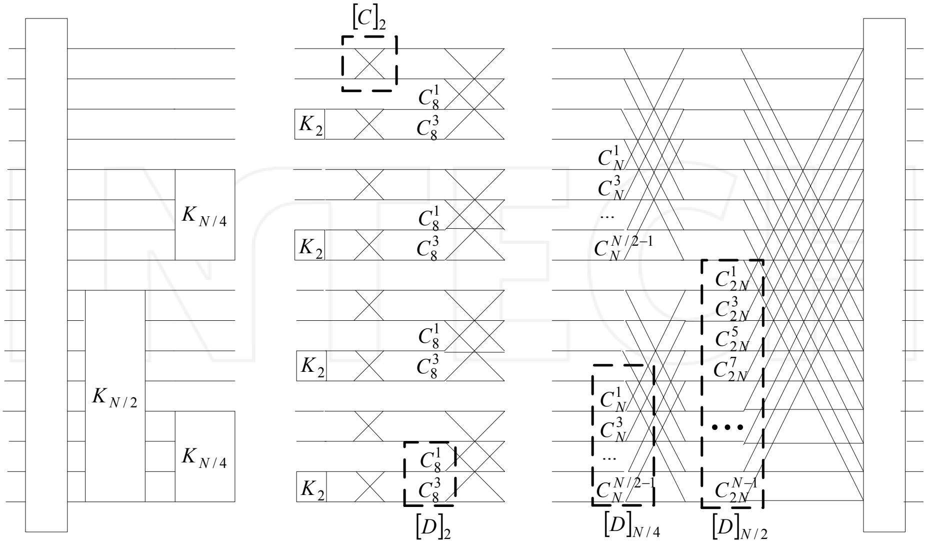


Fig. 2. Butterfly data flow diagram of proposed DCT - II matrix with order N

3.3 Hybrid DCT/DFT architecture on element inverse matrices

It is clear that the form of (6-43) is the same as that of (6-16), where we only need change K_l to $\Pr_l$ and D_l to W_l, with $l \in \{2,4,8,...,N/2\}$. Consequently, the results show that the DCT - II and DFT can be unified by using same algorithm and architecture within some characters changed. As illustrated in Fig.1, and Fig.2, we find that the DFT calculations can be obtained from the architecture of DCT by replacing the $[D]_N$ to $[W]_N$, and a permutation matrix $[\Pr]_N$ to $[K]_N$. Hence a unified function block diagram for DCT/DFT hybrid architecture algorithm can be drawn as Fig.3. In this figure, we can joint DCT and DFT in one chip or one processing architecture, and use one switching box to control the output data flow. It will be useful to developing the unified chip or generalized form for DCT and DFT together.

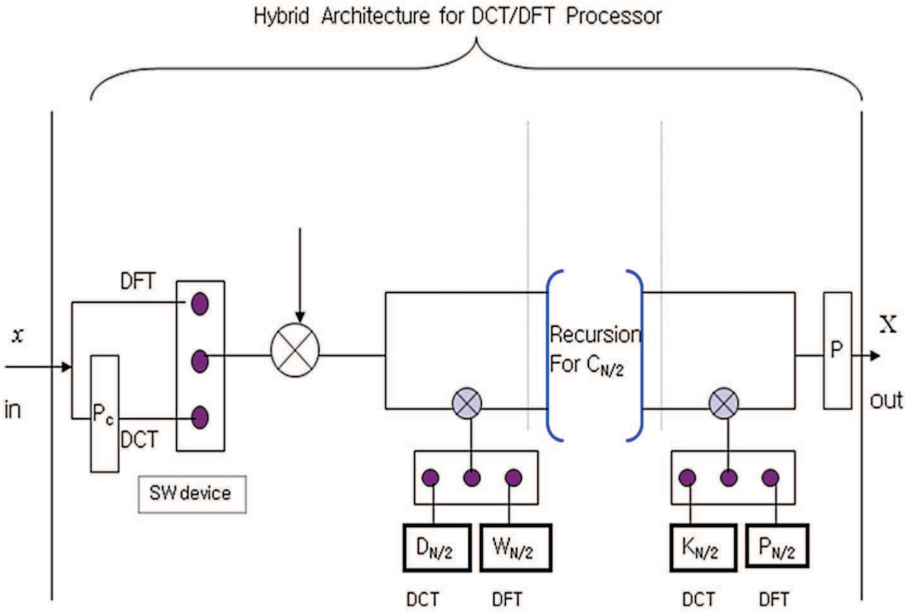


Fig. 3. A unified function block diagram for proposed DCT/DFT hybrid architecture algorithm

4. Conclusion

We propose a new representation of DCT/DFT matrices via the Jacket transform based on the block-inverse processing. Following on the factorization method of the Jacket transform, we show that the inverse cases of DCT/DFT matrices are related to their block inverse sparse matrices and the permutations. Generally, DCT/DFT can be represented by using the same architecture based on element inverse or block inverse decomposition. Linking between two transforms was derived based on matrix recursion formula.

Discrete Cosine Transform (DCT) has applications in signal classification and representation, image coding, and synthesis of video signals. The DCT-II is a popular structure and it is usually accepted as the best suboptimal transformation that its performance is very close to that of the statistically optimal Karhunen-Loeve transform for picture coding. Further, the discrete Fourier transform (DFT) is also a popular algorithm for signal processing and communications, such as OFDM transmission and orthogonal code designs. Being combined these two different transforms, a unified fast processing module to implement DCT/DFT hybrid architecture algorithm can be designed by adding switching device to control either DCT or DFT processing depending on mode of operation.

Further investigation is needed for unified treatment of recursive decomposition of orthogonal transform matrices exploiting the properties of Jacket-like sparse matrix architecture for fast trigonometric transform computation.

5. Acknowledgement

This work was supported by World Class Univ. R32-2008-000-20014-0, NRF , Korea.

6. Appendix

6.1 Appendix 1

The DFT matrix brings higher powers of w , and the problem turns out to be

$$\begin{bmatrix} 1 & 1 & 1 & . & 1 \\ 1 & w & w^2 & . & w^{n-1} \\ 1 & w^2 & w^4 & . & w^{2(n-1)} \\ . & . & . & . & . \\ 1 & w^{n-1} & w^{2(n-1)} & . & w^{(n-1)^2} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ . \\ c_{n-1} \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ . \\ y_{n-1} \end{bmatrix}. \tag{A-1}$$

and the inverse form

$$F^{-1} = \frac{1}{n} \begin{bmatrix} 1 & 1 & 1 & . & 1 \\ 1 & w^{-1} & w^{-2} & . & w^{-(n-1)} \\ 1 & w^{-2} & w^{-4} & . & w^{-2(n-1)} \\ . & . & . & . & . \\ 1 & w^{-(n-1)} & w^{-2(n-1)} & . & w^{-(n-1)^2} \end{bmatrix}. \tag{A-2}$$

Then, we can define the Fourier matrix as follows.

Definition A.1: An $n \times n$ matrix $F = \left[a_{ij} \right]$ is a Fourier matrix if

$$a_{ij} = w^{(i-1)(j-1)}, w = e^{2\pi i/n}, \text{ and } i, j \in \{1, 2, ..., n\}. \tag{A-3}$$

and the inverse form $F^{-1} = \frac{1}{n}\left[\left(a_{ij} \right)^{-1} \right] = \frac{1}{n}\left[\left(w^{-(i-1)(j-1)} \right) \right]$.

For example, in the cases $n = 2$ and $n = 3$, and inverse is an element-wise inverse like Jacket matrix, then, we have

$$F_2 = \begin{bmatrix} w^0 & w^0 \\ w^0 & w^1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \; F_2^{-1} = \frac{1}{2} \begin{bmatrix} w^0 & w^0 \\ w^0 & w^{-1} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

and

$$F_3 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & e^{\frac{i2\pi}{3}} & e^{\frac{i4\pi}{3}} \\ 1 & e^{\frac{i4\pi}{3}} & e^{\frac{i8\pi}{3}} \end{bmatrix},\ F_3^{-1} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & e^{\frac{-i2\pi}{3}} & e^{\frac{-i4\pi}{3}} \\ 1 & e^{\frac{-i4\pi}{3}} & e^{\frac{-i8\pi}{3}} \end{bmatrix}.$$

We need to confirm that FF^{-1} equals the identity matrix. On the main diagonal that is clear. Row j of F times column j of F^{-1} is $(1/n)(1+1+...+1)$, which is 1. The harder part is off the diagonal, to show that row j of F times column k of F^{-1} gives zero:

$$, \text{if } j \neq k . \tag{A-4}$$

The key is to notice that those terms are the powers of :

$$1 + W + W^2 + ... + W^{n-1} = 0 . \tag{A-5}$$

6.2 Appendix 2

In a general case, we have

$$[B]_N = [K]_N [C]_N [D]_N ,$$

where

$$[K]_N = \begin{bmatrix} \sqrt{2} & 0 & 0 & \cdots \\ -\sqrt{2} & 2 & 0 & \cdots \\ \sqrt{2} & -2 & 2 & \cdots \\ \vdots & \vdots & & \ddots \end{bmatrix},\ [D]_N = \begin{bmatrix} C_{4N}^{\Phi_0} & 0 & \cdots & 0 \\ 0 & C_{4N}^{\Phi_1} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & C_{4N}^{\Phi_{N-1}} \end{bmatrix},$$

and $\Phi_i = 2i+1$, $i \in \{0,1,2,...,N-1\}$.

Proof: In case of $N \times N$ DCT - II matrix, $[C]_N$, it can be represented by using the form as

$$[C]_N = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ C_{4N}^{2k_0\Phi_0} & C_{4N}^{2k_0\Phi_1} & C_{4N}^{2k_0\Phi_2} & \cdots & C_{4N}^{2k_0\Phi_{N-2}} & C_{4N}^{2k_0\Phi_{N-1}} \\ C_{4N}^{2k_1\Phi_0} & C_{4N}^{2k_1\Phi_1} & C_{4N}^{2k_1\Phi_2} & \cdots & C_{4N}^{2k_1\Phi_{N-2}} & C_{4N}^{2k_1\Phi_{N-1}} \\ C_{4N}^{2k_2\Phi_0} & C_{4N}^{2k_2\Phi_1} & C_{4N}^{2k_2\Phi_2} & \cdots & C_{4N}^{2k_2\Phi_{N-2}} & C_{4N}^{2k_2\Phi_{N-1}} \\ \vdots & & & \vdots & & \vdots \\ C_{4N}^{2k_{N-2}\Phi_0} & C_{4N}^{2k_{N-2}\Phi_1} & C_{4N}^{2k_{N-2}\Phi_2} & \cdots & C_{4N}^{2k_{N-2}\Phi_{N-2}} & C_{4N}^{2k_{N-2}\Phi_{N-1}} \end{bmatrix}, \tag{A-6}$$

where $k_i = i+1$, $i \in \{0,1,2,...\}$.

According to (6-37), a $N \times N$ matrix $[B]_N$ from $[C]_{2N}$ can be simply presented by

$$\left[B\right]_N=\begin{bmatrix} C_{4N}^{\Phi_0} & C_{4N}^{\Phi_1} & C_{4N}^{\Phi_2} & \cdots & C_{4N}^{\Phi_{N-1}} \\ C_{4N}^{(2k_0+1)\Phi_0} & C_{4N}^{(2k_0+1)\Phi_1} & C_{4N}^{(2k_0+1)\Phi_2} & \cdots & C_{4N}^{(2k_0+1)\Phi_{N-1}} \\ C_{4N}^{(2k_1+1)\Phi_0} & C_{4N}^{(2k_1+1)\Phi_1} & C_{4N}^{(2k_1+1)\Phi_2} & \cdots & C_{4N}^{(2k_1+1)\Phi_{N-1}} \\ \vdots & & \vdots & & \vdots \\ C_{4N}^{(2k_{N-2}+1)\Phi_0} & C_{4N}^{(2k_{N-2}+1)\Phi_1} & C_{4N}^{(2k_{N-2}+1)\Phi_2} & \cdots & C_{4N}^{(2k_{N-2}+1)\Phi_{N-1}} \end{bmatrix}. \qquad \text{(A-7)}$$

And based on (6-78), we have the formula

$$C_{4N}^{(2k_i+1)\Phi_m}=2C_{4N}^{2k_i\Phi_m}C_{4N}^{\Phi_m}-C_{4N}^{(2k_i-1)\Phi_m}=-C_{4N}^{(2k_i-1)\Phi_m}+2C_{4N}^{2k_i\Phi_m}C_{4N}^{\Phi_m}, \qquad \text{(A-8)}$$

where $m\in\{0,1,2,....\}$.
Thus we can calculate that

$$\left[K\right]_N\left[C\right]_N\left[D\right]_N$$

$$=\begin{bmatrix} \sqrt{2} & 0 & 0 & 0 & \cdots & 0 \\ -\sqrt{2} & 2 & 0 & 0 & & 0 \\ \sqrt{2} & -2 & 2 & 0 & & 0 \\ -\sqrt{2} & 2 & -2 & 2 & \cdots & \vdots \\ \sqrt{2} & -2 & 2 & -2 & 2 & \\ \vdots & & & & & \ddots \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ C_{4N}^{2k_0\Phi_0} & C_{4N}^{2k_0\Phi_1} & C_{4N}^{2k_0\Phi_2} & \cdots & C_{4N}^{2k_0\Phi_{N-2}} & C_{4N}^{2k_0\Phi_{N-1}} \\ C_{4N}^{2k_1\Phi_0} & C_{4N}^{2k_1\Phi_1} & C_{4N}^{2k_1\Phi_2} & \cdots & C_{4N}^{2k_1\Phi_{N-2}} & C_{4N}^{2k_1\Phi_{N-1}} \\ C_{4N}^{2k_2\Phi_0} & C_{4N}^{2k_2\Phi_1} & C_{4N}^{2k_2\Phi_2} & \cdots & C_{4N}^{2k_2\Phi_{N-2}} & C_{4N}^{2k_2\Phi_{N-1}} \\ \vdots & & & \vdots & & \vdots \\ C_{4N}^{2k_{N-2}\Phi_0} & C_{4N}^{2k_{N-2}\Phi_1} & C_{4N}^{2k_{N-2}\Phi_2} & \cdots & C_{4N}^{2k_{N-2}\Phi_{N-2}} & C_{4N}^{2k_{N-2}\Phi_{N-1}} \end{bmatrix},$$

$$\begin{bmatrix} C_{4N}^{\Phi_0} & 0 & \cdots & 0 \\ 0 & C_{4N}^{\Phi_1} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & C_{4N}^{\Phi_{N-1}} \end{bmatrix}$$

$$=\begin{bmatrix} 1 & 1 & ... & 1 \\ -1+2C_{4N}^{2k_0\Phi_0} & -1+2C_{4N}^{2k_0\Phi_1} & ... & -1+2C_{4N}^{2k_0\Phi_{N-1}} \\ 1-2C_{4N}^{2k_0\Phi_0}+2C_{4N}^{2k_1\Phi_0} & 1-2C_{4N}^{2k_0\Phi_1}+2C_{4N}^{2k_1\Phi_1} & \cdots & 1-2C_{4N}^{2k_0\Phi_{N-1}}+2C_{4N}^{2k_1\Phi_{N-2}} \\ \vdots & & & \vdots \end{bmatrix}$$

$$\begin{bmatrix} C_{4N}^{\Phi_0} & 0 & \cdots & 0 \\ 0 & C_{4N}^{\Phi_1} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & C_{4N}^{\Phi_{N-1}} \end{bmatrix},$$

$$=\begin{bmatrix} C_{4N}^{\Phi_0} & C_{4N}^{\Phi_1} & ... & C_{4N}^{\Phi_{N-1}} \\ -C_{4N}^{\Phi_0}+2C_{4N}^{2k_0\Phi_0}C_{4N}^{\Phi_0} & -C_{4N}^{\Phi_1}+2C_{4N}^{2k_0\Phi_1}C_{4N}^{\Phi_1} & \cdots & -C_{4N}^{\Phi_{N-1}}+2C_{4N}^{2k_0\Phi_{N-1}}C_{4N}^{\Phi_{N-1}} \\ C_{4N}^{\Phi_0}-2C_{4N}^{2k_0\Phi_0}C_{4N}^{\Phi_0}+2C_{4N}^{2k_1\Phi_0}C_{4N}^{\Phi_0} & C_{4N}^{\Phi_1}-2C_{4N}^{2k_0\Phi_1}C_{4N}^{\Phi_1}+2C_{4N}^{2k_1\Phi_1}C_{4N}^{\Phi_1} & \cdots & \\ \vdots & & & \vdots \end{bmatrix}, \qquad \text{(A-9)}$$

Since $k_0 = 1$, we get

$$-C_{4N}^{\Phi_m} + 2C_{4N}^{2k_0\Phi_m} C_{4N}^{\Phi_m} = -C_{4N}^{(2k_0-1)\Phi_m} + 2C_{4N}^{2k_0\Phi_m} C_{4N}^{\Phi_m} = C_{4N}^{(2k_0+1)\Phi_m}, \tag{A-10}$$

and $$C_{4N}^{\Phi_m} - 2C_{4N}^{2k_0\Phi_m} C_{4N}^{\Phi_m} = -(-C_{4N}^{(2k_0-1)\Phi_m} + 2C_{4N}^{2k_0\Phi_m} C_{4N}^{\Phi_m}) = -C_{4N}^{(2k_0+1)\Phi_m}. \tag{A-11}$$

In case of $k_i = i+1$, we have $(2k_{i-1}+1)\Phi_m = (2(k_i-1)+1)\Phi_m = \left(2k_i - 1\right)\Phi_m$, then we get

$$\begin{aligned} &C_{4N}^{\Phi_m} - 2C_{4N}^{2k_{i-1}\Phi_m} C_{4N}^{\Phi_m} + 2C_{4N}^{2k_i\Phi_m} C_{4N}^{\Phi_m} = -C_{4N}^{(2k_{i-1}+1)\Phi_m} + 2C_{4N}^{2k_i\Phi_m} C_{4N}^{\Phi_m} \\ &= -C_{4N}^{(2k_i-1)\Phi_m} + 2C_{4N}^{2k_i\Phi_m} C_{4N}^{\Phi_m} = C_{4N}^{(2k_i+1)\Phi_m} \end{aligned}. \tag{A-12}$$

Taking the (A-10)-(A-12) to (A-9), we can rewrite that

$$[K]_N [C]_N [D]_N$$

$$= \begin{bmatrix} C_{4N}^{\Phi_0} & C_{4N}^{\Phi_1} & \cdots & C_{4N}^{\Phi_{N-1}} \\ -C_{4N}^{\Phi_0} + 2C_{4N}^{2k_0\Phi_0} C_{4N}^{\Phi_0} & -C_{4N}^{\Phi_1} + 2C_{4N}^{2k_0\Phi_1} C_{4N}^{\Phi_1} & \cdots & -C_{4N}^{\Phi_{N-1}} + 2C_{4N}^{2k_0\Phi_{N-1}} C_{4N}^{\Phi_{N-1}} \\ C_{4N}^{\Phi_0} - 2C_{4N}^{2k_0\Phi_0} C_{4N}^{\Phi_0} + 2C_{4N}^{2k_1\Phi_0} C_{4N}^{\Phi_0} & C_{4N}^{\Phi_1} - 2C_{4N}^{2k_0\Phi_1} C_{4N}^{\Phi_1} + 2C_{4N}^{2k_1\Phi_1} C_{4N}^{\Phi_1} & \cdots & \\ \vdots & & & \vdots \end{bmatrix}$$

$$= \begin{bmatrix} C_{4N}^{\Phi_0} & C_{4N}^{\Phi_1} & \cdots & C_{4N}^{\Phi_{N-1}} \\ C_{4N}^{(2k_0+1)\Phi_0} & C_{4N}^{(2k_0+1)\Phi_1} & \cdots & C_{4N}^{(2k_0+1)\Phi_{N-1}} \\ C_{4N}^{(2k_1+1)\Phi_0} & C_{4N}^{(2k_1+1)\Phi_1} & \cdots & C_{4N}^{(2k_1+1)\Phi_{N-1}} \\ \vdots & & & \vdots \end{bmatrix} = [B]_N. \tag{A-13}$$

The proof is completed.

7. References

Ahmed, N. & K.R. Rao(1975), *Orthogonal Transforms for Digital Signal Processing*, 0387065563, Berlin, Germany: Springer-Verlag

Chen, Z.; M.H. Lee, & Guihua Zeng(2008), Fast Cocylic Jacket Transform, *IEEE Trans. on Signal Proc.* Vol. 56, No.5, 2008, (2143-2148) ,1053-587X

Fan, C.-P. & Jar-Ferr Yang(1998), Fast Center Weighted Hadamard Tranform Algorithm, *IEEE Trans.on CAS-II* Vol. 45, No. 3, (1998) (429-432), 1057-7130

Hou, Jia; M.H. Lee, & Ju Yong Park(2003), New Polynomial Construction of Jacket Transform, *IEICE Trans. Fund.* Vol. E86A, No. 3, (2003) (652-660) , 0916-8508

Lee, M.H.(1989), The Center Weighted Hadamard Transform, *IEEE Trans. on Circuits and Syst.* 36(9), (1989) (1247- 1249), 0098-4094

Lee, M.H.(1992), High Speed Multidimensional Systolic Arrays for Discrete Fourier Transform, *IEEE Trans. on .Circuits and Syst.II* 39 (12) , (1992) (876-879), 1057-7130

Lee, M.H.(2000) , A New Reverse Jacket Transform and Its Fast Algorithm, *IEEE, Trans. on Circuit and System* 47(1) ,2000, (39-47), 1057-7130

Lee, M.H. & Ken Finlayson(2007), A simple element inverse Jacket transform coding, *IEEE Signal Processing Lett.* 14 (5) , 2007, (325-328), 1070-9908

Lee, M. H. & D.Y.Kim(1984), Weighted Hadamard Transform for S/N Ratio Enhancement in Image Transform, *Proc. of IEEE* ISCAS'84 , 1984 , pp. 65-68.

Lee, S.R. & M.H. Lee(1998), On the Reverse Jacket Matrix for Weighted Hadamard Transform, *IEEE Trans. On CAS-II*, Vol. 45, No. 3, (1998) (436-441) , 1057-7130

Lee, M.H.; N.L.Manev & Xiao-Dong Zhang (2008), Jacket transform eigenvalue decomposition, *Appl. Math.Comput*.198 , (May 2008). (858-864)

Lee, M.H.; B. S. Rajan, & Ju Yong Park(2001), A Generalized Reverse Jacket Transform, *IEEE Trans. on Circuits and Syst. II* 48 (7) ,(2001) (684-690), 1057-7130

Lee, S. R. & J. H. Yi(2002), Fast Reverse Jacket Transform as an Altenative Representation of N point Fast Fourier Transform, *Journal of Mathematical Imaging and Vision* Vol. 16, No. 1,,(2002) (1413-1420), 0924-9907

Park, D.; M.H. Lee & Euna Choi(1999), Revisited DFT matrix via the reverse jacket transform and its application to communication, *The 22nd symposium on Information theory and its applications* (SITA 99), 1999, Yuzawa, Niigata, Japan

Rao, K.R. & P. Yip(1990), *Discrete Cosine Transform Algorithms, Advantages, Applications,*. 0-12-580203-X, Academic Press, San Diego, CA, USA

Rao, K.R. & J.J. Hwang(1996), *Techniques & Standards for Image Video & Audio Coding*, 0-13-309907-5, Prentice Hall , Upper Saddle River, NJ, USA

Whelchel, J.E. & D.F. Guinn(1968), The fast fourier hadamard transform and its use in signal representation and classification, *Proc. EASCON'68*, 1968 , pp. 561-564

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

9

Optimized FFT Algorithm and its Application to Fast GPS Signal Acquisition

Lin Zhao, Shuaihe Gao, Jicheng Ding and Lishu Guo
College of Automation,
Harbin Engineering University
China

1. Introduction

The essence of GPS signal acquisition is a two-dimensional search process for the carrier Doppler and code phase, generally including correlator, signal capture device, and logic control module. The disposal efficiency of correlator would affect the capture speed of the whole acquisition process. Because of the corresponding relation between frequency-domain multiplication and time-domain convolution, the Discrete Fourier Transform (DFT) could be applied to play the role of the correlator, which is suitable to implement for computers (Akopian, 2005) (Van Nee & Coenen, 1991).

However, with the performance requirements of GPS receivers increasing, especially in the cold start and the long code acquisition, such as P code, the acquisition time should be furtherly reduced. Therefore, the fast discrete Fourier transform processing approach, that is, Fast Fourier Transform (FFT), is described, including radix-2, radix-4, split-radix algorithm, Winograd Fourier Transform Algorithm (WFTA) which is suitable for a small number of treatment points, and Prime Factor Algorithm (PFA) in which the treatment points should be the product of some prime factors.

According to the actual needs of GPS signal acquisition, an optimized FFT algorithm was put forward, which comprehensively utilize the advantages of different FFT algorithms. Applying optimized FFT algorithm to GPS signal acquisition, the results of simulations indicate that the improved processing could reduce the acquisition time significantly and improve the performance of GPS baseband processing.

2. Analysis on GPS signal acquisition

2.1 Basic characteristics of GPS L1 signal

There are three basic components in GPS signal: carrier waves, pseudo-random numbers (PRN) codes and navigation message (D code). Among them, the carrier waves are located at the L-band, including L1-band (1575.42MHz) which is with the most common application, L2-band (1227.6MHz) and L5-band (1176.45MHz). There are two basic types of PRN codes, the coarse/ acquisition (C/ A) code and the precise (P) code. The simplified structure of GPS L1 signal is shown in Fig. 1. The frequency of carrier wave is 1575.42MHz. The code rate of C/ A code and P code are 1.023MHz and 10.23MHz respectively. The data rate of D code is 50Hz.

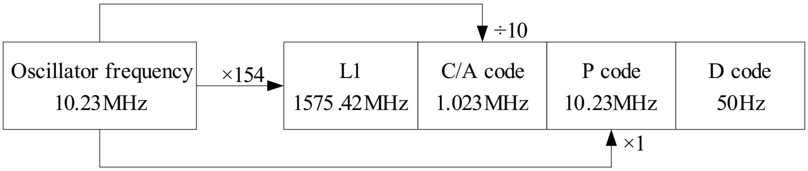


Fig. 1. The simplified structure of GPS L1 signal

C/ A code is used to achieve spread spectrum of D code. Compound code C/ A⊕D could be gained when the D code is spreading, and the code frequency is extended to 1.023MHz. Compound code P⊕D could be gained when the P code is spreading, and the code frequency is extended to 10.23MHz. And then multiply the spread spectrum signal with the L1 carrier wave to complete the modulation. Quaternary phase shift keying (QPSK) is applied in the modulation of L1 signal, where the in-phase carrier component $\cos(2\pi f_{L1}t+\theta_{L1})$ is modulated with compound code C/ A⊕D, and the orthogonal carrier components $\sin(2\pi f_{L1}t+\theta_{L1})$ is modulated with compound code P⊕D, so the L1 signal transmitted by satellites could be expressed as:

$$s_{L1}(t)=\sqrt{2P_x}D(t)x(t)\cos(2\pi f_{L1}t+\theta_{L1})+\sqrt{2P_y}D(t)y(t)\sin(2\pi f_{L1}t+\theta_{L1}) \tag{1}$$

Where P_x and P_y are powers of different signal components respectively, $D(t)$, $x(t)$ and $y(t)$ are D code, C/ A code and P code of satellite respectively, f_{L1} is the frequency of carrier wave, and θ_{L1} is the initial phase.
The process of spread spectrum and modulation for GPS transmitted L1 signal could be shown in Fig. 2 (Kaplan, & Hegarty, 2006).

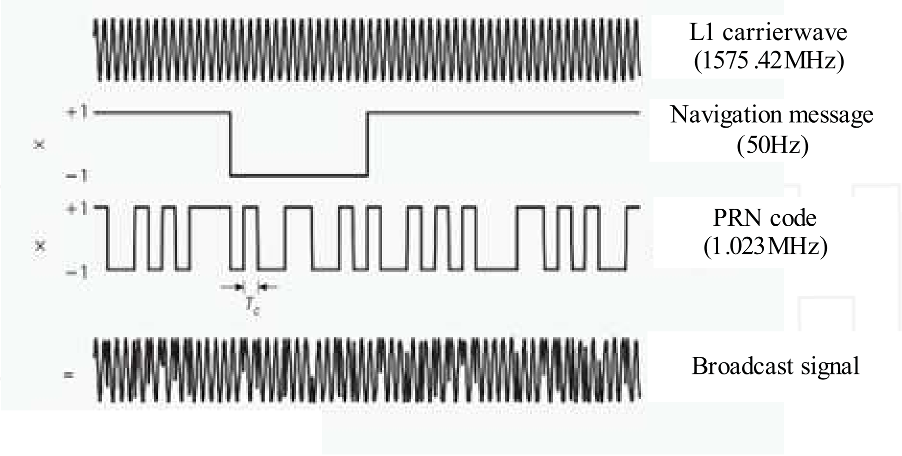


Fig. 2. The spread spectrum and modulation for GPS L1 signal

In fact, P code mainly used for the military, and is not open to civilian use, so the received L1 signal could be simplified to include carrier wave, C/ A code and D code in the research of GPS C/ A code acquisition.

2.2 Signal acquisition in GPS receiver

The original GPS signal, which is interfered and attenuated in the transmission path, is gained by the receiver antenna. Radio Frequency (RF) front-end completes the frequency conversion and analog-digital conversion for the weak signal, where the high frequency analogy signal is transformed into Intermediate Frequency (IF) digital signal which is beneficial to computer processing. The coarse and precision estimates of carrier Doppler frequency shift and PRN code phase are achieved by acquisition and tracking in the baseband processing, and then dispreading and then demodulation of navigation message should be completed. Actually it is the opposite process of spreading and modulation mentioned above. Position calculation could be realized by adequate information gained by baseband module (Michael, & Dierendonck, 1999). The workflow of typical GPS software receiver is shown in Fig. 3.

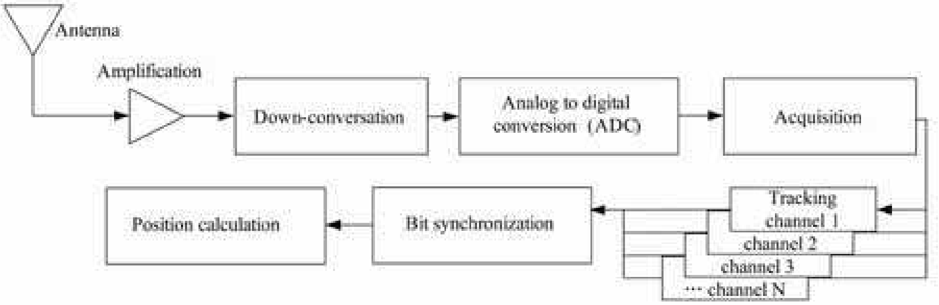


Fig. 3. The workflow of GPS receiver

As a part of baseband signal processing in GPS receiver, the main aim of signal acquisition is to find visible satellites, and estimate their C/ A code phase and carrier Doppler frequency shift respectively. The essence of GPS signal capture is a two-dimensional search process for the carrier Doppler and code phase. The PRN code phase and carrier Doppler frequency could be considered respectively. As shown in Fig. 4, each C/ A code contains 1023 code elements, and search step with one code element would be commonly selected. In high dynamic environment, the Doppler frequency ranges from -10kHz to +10kHz, and 1kHz search step is generally selected.

In the above search process, once the code phase and the carrier Doppler frequency shift generated by local oscillator are close to the receiving code phase and Doppler frequency shift, there will be a correlation peak for the randomness of C/ A code. Generally, the code phase error is less than half a symbol, and the Doppler frequency shift error is within [-500Hz, 500Hz]. At this time, the code phase and carrier Doppler frequency parameters which the peak point corresponds could be the acquisition results, and then the baseband processing enters the second stage, signal tracking.

When the predetermined frequency points are tested one by one in time-domain, the large computation would cause great time-consuming, unless there are a lot of hardware resources as the supplement. Fortunately, we can use another method to get all results the 1023 possible code phases corresponding for each frequency point, which is the acquisition method based on FFT.

3. GPS signal acquisition method based on FFT

Processing speed is constrained in the traditional acquisition method based on time domain, so the acquisition method based on Fourier transform is always used in current software receiver (Li, Zhang, Li, & Zhang, 2008).

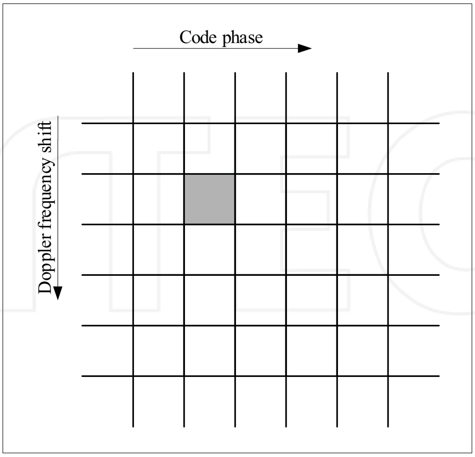


Fig. 4. Two-dimensional search for Doppler frequency shift and C/ A code phase

3.1 Corresponding relation between frequency-domain and time-domain

Using circular correlation theory, convert the correlation of received signal and local generated signal in the time-domain to spectrum multiplying in the frequency-domain (Akopian, 2005) (Van Nee & Coenen, 1991). Assumed $c(\tau)$ is the circulating moving local code, N is the number of corresponding sampling points, the output of correlator would be expressed as:

$$y(t) = \sum_{\tau=0}^{N-1} S_{IF}(t)c(t+\tau) \tag{2}$$

Do discrete Fourier transform (DFT) of $y(t)$,

$$Y(k) = \sum_{\tau=0}^{N-1} \sum_{t=0}^{N-1} s_{IF}(t)c(t+\tau)e^{-2\pi jk\tau/N} \tag{3}$$

And then, $Y(k)$ could be transformed to

$$Y(k) = \sum_{t=0}^{N-1} s_{IF}(t)(\sum_{\tau=0}^{N-1} c(t+\tau)e^{-2\pi j(t+\tau)k/N})e^{2\pi jtk/N} \tag{4}$$

If $S(k)$ and $C(k)$ are DFT forms of $x(t)$ and $c(\tau)$,

$$Y(k) = C(k) \sum_{t=0}^{N-1} s_{IF}(t)e^{2\pi jtk/N} = C(k) \times S(k)^{-1} \tag{5}$$

As the local code $c(\tau)$ is real signal, the complex conjugates of $S(k)$ and $C(k)$ could be expressed by $S^*(k)$ and $C^*(k)$ respectively. The amplitude output is

$$|Y(k)| = |S^*(k) \times C(k)| = |S(k) \times C^*(k)| \tag{6}$$

According to the correlation of original signal and local code, adjudicate the relevant results. The number of visible satellites, and the estimation of code phase and Doppler frequency shift could be drawn. So the process of signal acquisition based on FFT could be expressed as follows:

$$Y(k) = f_{IFFT}(f_{FFT}(s(k)) \times f^*_{FFT}(c(k))) \tag{7}$$

Where f_{IFFT} is the inverse fast Fourier transform (IFFT) operation, f_{FFT} is the FFT operation, and f^*_{FFT} is the conjugate form of f_{FFT}.

3.2 GPS signal acquisition based on FFT

The signal received by antenna would go though amplification, mixing, filtering, and analog-digital conversion in RF front-end, and its output is the IF digital signal. The local carrier wave numerical controlled oscillator (NCO) would generate two-way mutually orthogonal signal $\sin(\omega t)$ and $\cos(\omega t)$, which would be utilized to multiply the IF digital signal respectively. As shown in Fig. 5, the value of branch I and branch Q are regarded as real part and imaginary part respectively. Construct a new complex sequence with the form of

$$x(n) = I(n) + jQ(n) \tag{8}$$

Do FFT of this new sequence, and do FFT of the local generated C/ A codes at the same time. Then complex multiplications are carried out between these two FFT values. After correlation, IFFT operations are carried out for it. Calculate the modulus of IFFT results one by one, and find the maximum value, which is shown in Fig. 6. Comparing the maximum value and the pre-set threshold, if the maximum value is less than the threshold, it means there is no effective signal. But if the maximum value is higher than the threshold, it means the acquisition is successful, and the received signal code phase and Doppler frequency shift would appear in the location of peak.

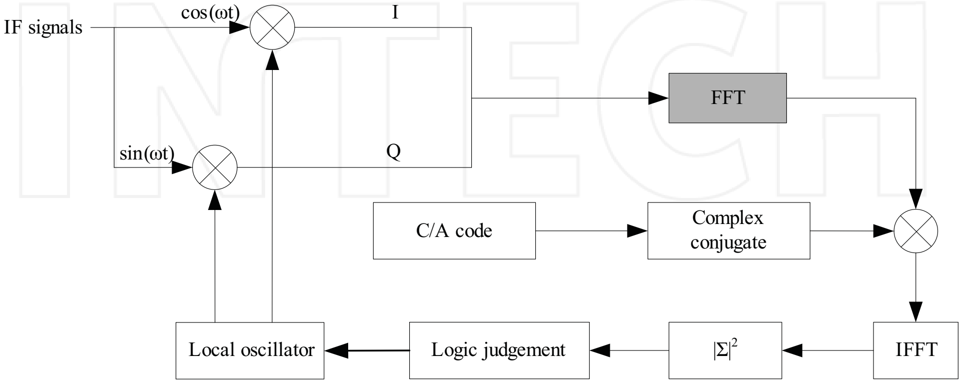


Fig. 5. Signal acquisition process based on FFT

The corresponding C/ A code and Doppler phase shift could be obtained in the same time, so it is obviously that the FFT operations play a crucial role in the acquisition, especially in the quick acquisition for high dynamic environments. The efficiency of FFT computation would determine the capture speed and whole performance of the receiver.

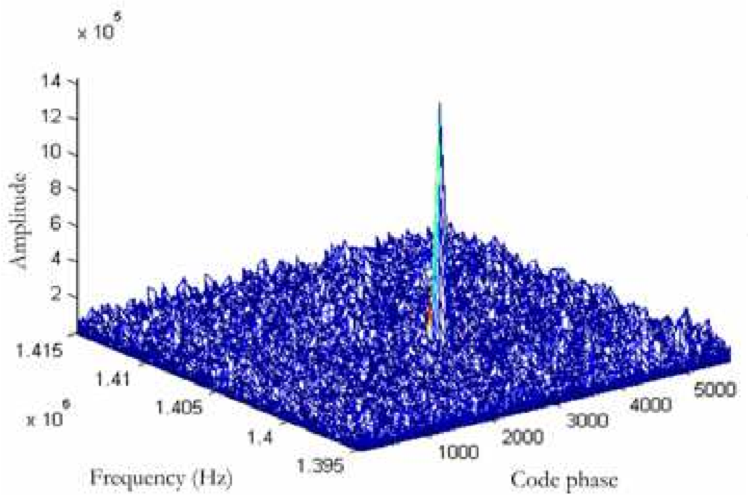


Fig. 6. The peak of signal acquisition (PRN=13)

4. Description on FFT algorithms

DFT of $x(n)$ could be defined as:

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{nk} \tag{9}$$

According to the definition of equation (9), its computation of complex multiplication is N^2, and its computation of complex addition is $N(N-1)$, whose operation is quite large. To calculate DFT rapidly, FFT algorithms came into being in nearly half a century with the purpose to reduce the calculation of DFT (Duhamel, & Vetterli, 1990).

Since Cooley and Tukey proposed FFT algorithm, the new algorithms have emerged constantly. In general, there are two basic directions. One is that the length of sequence equals to an integral power of 2, with the form of $N = 2^l$, such as radix-2 algorithm, radix-4 algorithm and split-radix algorithm. The other is that the number of points does not equal to an integer power of 2, with the form of $N \neq 2^l$, which is represented by a class of Winograd algorithm, such as prime factor algorithm and WFTA algorithm (Burrus, & Eschenbacher, 1981).

But the basic idea of various FFT algorithms is to divide the long sequence to short sequences successively, and then make full use of the periodicity, symmetry and reducibility of rotation factors to decompose DFT with a large number N into DFT with a combination of small number of points to reduce the computation. The property of the rotation factor W_N^{km} is as follows:

- Periodicity：

$$W_N^{(k+N)m} = W_N^{k(m+N)} = W_N^{km} \tag{10}$$

- Symmetry：

$$W_N^{mk+N/2} = -W_N^{mk} \tag{11}$$

$$(W_N^{km})^* = W_N^{-mk} \tag{12}$$

- Reducibility ：

$$W_N^{mk} = W_{nN}^{nmk} \tag{13}$$

$$W_N^{mk} = W_{N/n}^{mk/n} \tag{14}$$

Where N/n is an integer.

4.1 Radix-2 FFT algorithm

Radix-2 FFT algorithm is commonly used, which is described in detail in the literatures (Jones, & Watson, 1990) (Sundararajan, 2003). Its basic requirement is the length of the sequence N should satisfy $N = 2^l$, where l is an integer. If N could not satisfy $N = 2^l$, zeros-padding method is always applied. There are two categories in radix-2 FFT algorithm: one is to decompose the time sequence $x(n)$ (n is time label) successively which is called decimation-in-time algorithm, and the other one is to decompose the Fourier transform sequence $X(k)$ (k is frequency label) which is called decimation-in-frequency algorithm. To some extent these two algorithms are consistent, so only decimation-in- time algorithm would be described in detail here.

Divide the sequence $x(n)$ with the length $N = 2^l$ into two groups according to parity,

$$x(n) \implies \begin{cases} x_1(r) = x(2r) \\ x_2(r) = x(2r+1) \end{cases} \tag{15}$$

Where $n = 0, 1, ..., N/2 - 1$. Therefore, DFT could be transformed into

$$X(k) = DFT(x(n)) = \sum_{n=0}^{N-1} x(n) W_N^{nk} \tag{16}$$

Separate according to its parity,

$$X(k) = \sum_{r=0}^{N/2-1} x(2r) W_N^{2rk} + \sum_{r=0}^{N/2-1} x(2r+1) W_N^{(2r+1)k} \tag{17}$$

Simplify,

$$X(k) = \sum_{n=0}^{N-1} x_1(n) W_{N/2}^{rk} + W_N^k \sum_{n=0}^{N-1} x_2(n) W_{N/2}^{rk} \tag{18}$$

If there exist

$$X_1(k) = \sum_{r=0}^{N/2-1} x(2r) W_{N/2}^{rk} = \sum_{r=0}^{N/2-1} x_1(n) W_{N/2}^{nk} \tag{19}$$

$$X_2(k) = \sum_{r=0}^{N/2-1} x(2r+1) W_{N/2}^{rk} = \sum_{r=0}^{N/2-1} x_2(n) W_{N/2}^{nk} \tag{20}$$

Where $k = 0, 1, ..., N/2 - 1$. And

$$X(k) = X_1(k) + W_N^k X_2(k) \tag{21}$$

Utilize the property of rotating factor, and formula (22), (23), (24) and (25) could be got.

$$X_1(k + N/2) = X_1(k) \tag{22}$$

$$X_2(k + N/2) = X_2(k) \tag{23}$$

$$X(k) = X_1(k) + W_N^k X_2(k) \tag{24}$$

$$X(k + N/2) = X_1(k + N/2) + W_N^{k+N/2} X_2(k + N/2) = X_1(k) - W_N^k X_2(k) \tag{25}$$

In the process of decimation-in-time radix-2 FFT, the N points DFT needs to convert to two groups with even and odd serial numbers, and each group has $N/2$ points. Then the periodicity, symmetry and reducibility would be used. The operations of formula (24) and (25) could be described by butterfly unit as shown in Fig. 7. The transmission coefficient $+1$ and -1 in the figure means the multiplication with $+W_N^k$ and $-W_N^k$.

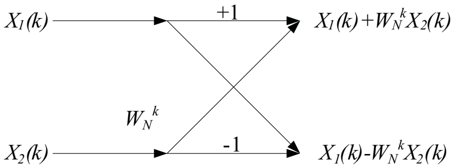


Fig. 7. Butterfly operation in decimation-in-time

Supposed that $N = 2^3 = 8$, now the decomposition process could be shown in Fig. 8.

Obviously, each butterfly operation requires one complex multiplication and two complex additions. If N points DFT is divided into two $N/2$ points DFT, calculating each $N/2$ points DFT directly, its computation of complex multiplication and complex addition are $N/2$ and $(N/2-1)N/2$ respectively. So theses two $N/2$ points DFT requires $N^2/2$ complex multiplications and $(N/2-1)N$ complex additions. Considering the existing $N/2$ butterfly operations in synthesis of N points DFT, there would be $N/2$ complex multiplications and N complex additions. So the calculation of complex multiplication would be reduced to $N^2/2 + N/2 \approx N^2/2$, and the calculation of complex multiplication would be reduced to $N(N/2-1) + N \approx N^2/2$ with the first step decomposition. Therefore, when N equals to the integer power of 2, there would be $N \log_2 N/2$ complex multiplications and $N \log_2 N$ complex additions.

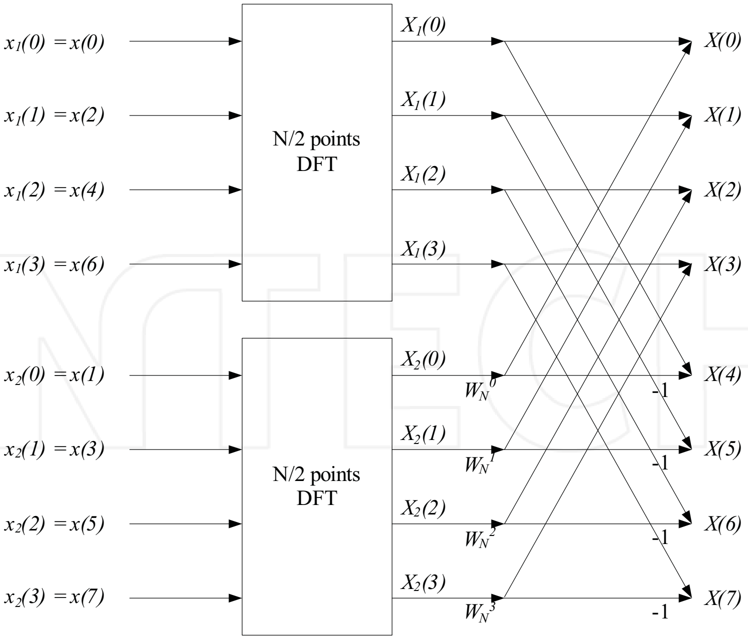


Fig. 8. Decomposition process of radix-2 FFT

4.2 Radix-4 FFT algorithm

Similar to the thought of radix-2 FFT, basic requirement of radix-4 FFT is the length of sequence N should satisfy $N = 4^l$, and it has been described in detail in the literatures (Jones, & Watson, 1990) (Sundararajan, 2003). It is worth to mention that each separate 4 points DFT would not require multiplication, and complex multiplication only appears in multiplying rotation factors operation. Rotation factor $W_N^0 = 1$, which need no multiplying, so each 4 points needs three multiplying rotation factors. And each step has $N/4$ 4 points DFT, so there would be $3N/4$ complex multiplications in each step. For N equals to 4^l , having l steps, the whole calculations of complex multiplications is

$$\frac{3}{4}N(l-1) = \frac{3}{4}N((\log_2 N)/2 - 1) \approx \frac{3}{8}N\log_2 N \tag{26}$$

There is no multiplying rotation factor in first step operation. Compared to the calculation of radix-2 FFT , the multiplications operation is much less. The number of butterfly unit is the same, so the calculation of complex additions in radix-4 FFT is $N\log_2 N$, which equals to the calculation of radix-2 FFT.

4.3 Split-radix FFT algorithm

Split-radix FFT algorithm was proposed in 1984, whose basic idea is to use radix-2 FFT algorithm in even-number DFT, and use radix-4 FFT algorithm in odd-number DFT (Jones, & Watson, 1990) (Sundararajan, 2003).

Radix-2 algorithm is applied to process the DFT of even numbers, and then the DFT of even sample points could be:

$$X(2k) = \sum_{n=0}^{N/2-1} [x(n) + x(n + N/2)W_{N/2}^{nk}] \tag{27}$$

Where $k = 0, 1, ..., N/2 - 1$. DFT of these points could be obtained by calculating the DFT of $N/2$ points without using any additional multiplication.

Similarly, radix-4 algorithm is applied to process the DFT of odd serial numbers. It is that rotation factor W_N^n should be multiplied for calculating the DFT of $\{X(2k+1)\}$. For these sample points, the efficiency would be improved to use radix-4 decomposition, because the multiplication of 4 points butterfly operation is the least. Appling radix-4 decimation-in-frequency algorithm to calculate the DFT of odd sample points, the following $N/4$ points DFT could be obtained.

$$X(4k+1) = \sum_{n=0}^{N/4-1} [(x(n) - x(n+N/2)) - j(x(n+N/4) - x(n+3N/4))]W_N^n W_{N/4}^{kn} \tag{28}$$

$$X(4k+3) = \sum_{n=0}^{N/4-1} [(x(n) - x(n+N/2)) + j(x(n+N/4) - x(n+3N/4))]W_N^{3n} W_{N/4}^{kn} \tag{29}$$

So the N points DFT could be decomposed to one $N/2$ points DFT with no rotation factor and two $N/4$ points DFT with rotation factor. Use this strategy repeatedly until there is no decomposition.

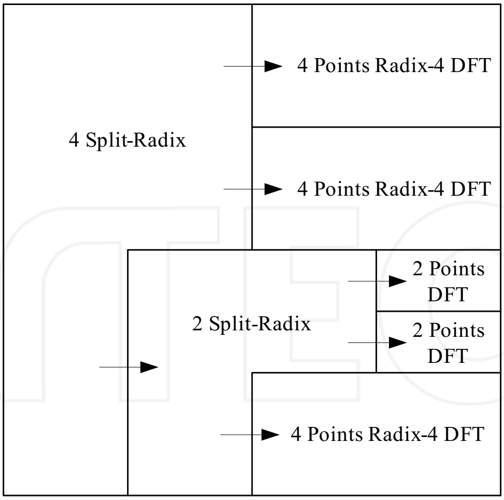


Fig. 9. 16 points DFT utilizing split-radix FFT

Taking $N = 4^2 = 16$ for example, there would be four split radixes in the first step decomposition for $X(k)$. There would be two split radixes in the second step decomposition for $X_1(k)$, including a 4 points DFT utilizing radix-4 FFT and two 2 points DFT utilizing radix-2 FFT. And the second step decomposition for $X_2(k)$ and $X_3(k)$ are both 4 points DFT utilizing radix-2 FFT. The decomposition could be shown in Fig. 9.

Considering the efficiency and application conditions for radix-4 and radix-2 algorithm comprehensively, split-radix algorithm is one of the most ideal methods to process the DFT with the length of $N = 2^l$ (Lin, Mao, Tsao, Chang, & Huang, 2006) (Mao, Lin, Tseng, Tsao, & Chang, 2007) (Nagaraj, Andrew, & Chris, 2009).

4.4 PFA FFT

In the FFT calculation, the number of points N could not usually be approximated to the integer power of 2, where traditional radix-2, radix-4 and split-radix algorithms could not be used. PFA was proposed by Kolba and Park in 1977, which alleviates the conflict between computation and the structure of algorithm (Chu, & Burrus, 1982) (Liu, & Zhang, 1997). But when N equals to the product of a number of prime factors, that is $N = N_1 \times N_2 ... \times N_i$, and most of them are odd items, its computational complexity would be slightly increased relative to the radix-2 FFT algorithm. Therefore the length of the decomposition factors should be better to be even, reducing the computations, whose basic idea is to transform one-dimensional DFT to two-dimensional or multi-dimensional small number of points DFT, and to get some superiors in calculation. However, it is provided that N_1, N_2 … and N_i are prime to each other, so there could be only one even factor. Taking the whole efficiency of operation and computer resources cost into account, the application of PFA method in this section would be converted to the form of formula (30).

$$N = N_1 \times N_2 \tag{30}$$

Where N_1 and N_2 are prime factors to each other, and $N_1 = 2^m$, m is an integer.

4.5 WFTA FFT

The expression of DFT is:

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{nk} \tag{31}$$

In the process of WFTA, the x and X in formula (31) could be expressed as the vector forms.

$$x = [x(0), x(1), ..., x(N-1)]^T \tag{32}$$

$$X = [X(0), X(1), ..., X(N-1)]^T \tag{33}$$

If D_N is a N-by-N matrix,

$$D_N(n,k) = [W_N^n k] \tag{34}$$

Where $n, k = 0, 1, ..., N-1$. Here the DFT could be the matrix form of

$$X = D_N x \tag{35}$$

Winograd draw the decomposition of D_N, and

$$D_N = S_N C_N T_N \tag{36}$$

Where T_N is a J-by-N incidence matrix, S_N is a N-by-J incidence matrix, C_N a J-by-J diagonal matrix, J is a positive integer. According to different values of N, J is to be determined. Therefore,

$$X = S_N C_N T_N x \tag{37}$$

For the smaller number of points DFT, WFTA could obtain D_N by calculating S_N, C_N and T_N whose computations are less. When $N = 2, 3, 4, 5, 7, 8, 9, 16$, the results of DFT are defined as a smaller factor DFT, which were presented in the literatures (Winogard, 1976). It could be substituted to formula (36), and their computations could be shown in table 1, which is relatively less. For the larger number of points DFT, the structure and program are complex, which restrict the application of WFTA (Liu, & Zhang, 1997).

Length of sequence (N)	Multiplication computation	Addition computation
3	4	12
4	0	16
6	10	34
7	16	72
8	4	52
9	20	88
16	20	144

Table 1. The computation of smaller number of points DFT with WFTA

5. Optimized FFT algorithm for GPS signal acquisition

5.1 Preprocess for FFT

As for GPS receivers, the best sampling rate is an integer power of 2 (Jin, Wu, & Li, 2005), but actually the points in GPS receivers could not always meet the best sampling rate. So data pre-processing is needed. When $N \neq 2^l$, pretreatment would be used to transform the number of points satisfying $N = 2^l$, which includes following means (Zhao, Gao, & Hao, 2009).

1. Zeros-padding method

For arbitrary sampling rate f_s in RF front-end, the C/ A code sequences are filled to the needed points. However, this process would change the cyclical properties of the C/ A codes. Because of decreasing the correlation peak, the cross-correlation increases, and the signal to noise ratio (SNR) output diminishes, but it is easy to achieve.

2. Average correlation method

Divide the data into average packets, receive appropriate points and process FFT. This method could reduce the consumption of hardware resources, but also decrease the ratio of two peaks.

3. The linear interpolation method

If radix-2 FFT processing is applied, and linear interpolation methods are used, such as Lagrange interpolation algorithm, the input data would be interpolated to an integer power of 2.

4. Sinc interpolation method

First of all, apply Sinc filter interpolate the input data, and the original continuous signal would be recovered. Then it would be resampled with a new sampling frequency. The advantage is that the distortion of PRN code is smaller which makes receivers can still normally work in the low SNR environment, but the realization is complexity. Meanwhile, the volume of calculating is larger.

5. Double – Length Zero –Padding method

The method was proposed by Stockham in 1966, the main idea is to extend the calculation points from N to N_1,

$$N_1 = (2^k)_{min} \geq 2N - 1 \tag{38}$$

Where k is an integer. Add $N_1 - N$ zeros after the input data directly. The first N local C/ A codes and the final N local C/ A codes are in the same cycle. So fill $N_1 - 2N$ zeros in the intermediate and treat the extended data with FFT. The DFT of former N points is the required correlation results, and there is no loss of the correlation peak.

In the fast GPS signal acquisition process based on FFT, the commonly pretreatment is zeros-padding method.

5.2 Optimized FFT algorithm

As mentioned above, if the points of sequence meets $N = 2^l$, split-radix algorithm is one of the most effective approaches. In this article, an improved FFT method for the sequence with $N = 2^l$ points is proposed, which is called optimized FFT algorithm for integer power of 2 (OFFTI). Its specific operation is to maintain the split-radix algorithm until decomposed into the final 16 points DFT, and then utilize smaller points DFT with WFTA.

A smaller amount of zeros could be added to transform the type of $N \neq 2^l$ into the type of $N = 2^l$, and then OFFTI algorithm could be utilize. But for the sequence which could not be converted to the type of $N = 2^l$ with few zeros. The specific processing of this condition is as follows. Firstly, PFA is utilized to decompose $N = N_1 \times N_2$ points DFT to the nested form with N_1 ($N_1 = 2^m$) groups N_2 points DFT and N_2 groups N_1 points DFT. As each layer calculates relatively independence in the PFA method, so it will be still decomposed with PFA method in process of N_2 points DFT. Until decomposing to less points DFT, the WFTA would be considered. For $N_1 = 2^m$ points DFT, the OFFTI algorithm could be used. We call this method against the type not satisfying $N = 2^l$ OFFTN algorithm.

Taking GPS C/ A code acquisition for example, if the digital rate is 5MHz, there would be 5000 data points in 1msec data. Zeros-padding method is applied in the data preprocessing. According to the common radix-2 FFT, add zeros to 8192 points based on the 5000 data points. Do N=8192 points DFT, and then discard the later 3192 points, which would bring extra computing and increase computation. Besides, there would be more errors brought in a certain extent. But if the OFFTN algorithm is used here, only several zeros should be added. Extending the points N to 5120 (45 × 5), the computation and errors would be much smaller than the traditional radix-2 FFT approach. The structure of the algorithm is shown in Fig. 10.

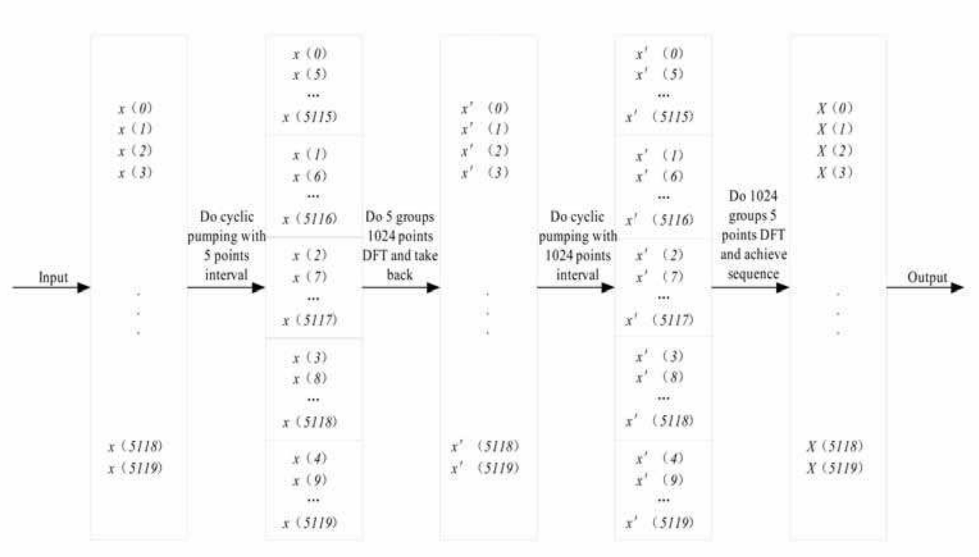


Fig. 10. The OFFTN structure of 5000 points

6. Improved acquisition method and simulation analysis

Apply the optimized methods mentioned above to GPS signal acquisition method based on FFT, the specific process is shown in Fig. 11.

The IF digital signal provided by GPS RF front-end is sent to baseband processing module, and then achieve pre-processing by few zeros padding. Multiply baseband signal with local generated carrier wave, and I channel signal is obtained. Multiply baseband signal with local generated carrier wave with 90° phase shift, and also Q channel signal is obtained. Then take the complex signal formed by I and Q channel signal to the FFT processing. Considering the character of the signal length, if N equals to the integer power of 2 approximately, OFFTI algorithm would be used, otherwise, OFFTN would be selected. The peak would generate by the correlation operations, and then compare it with the predetermined threshold. If the value is greater than the threshold, there would be GPS signal captured. Otherwise, no useful signal exists. Repeat the process until all of the available satellites are searched.

Signal acquisition base on radix-2 algorithm and the improved method are compared. As shown in Fig. 12, the correlation peaks have little difference for various methods and the acquisition results are mostly the same.

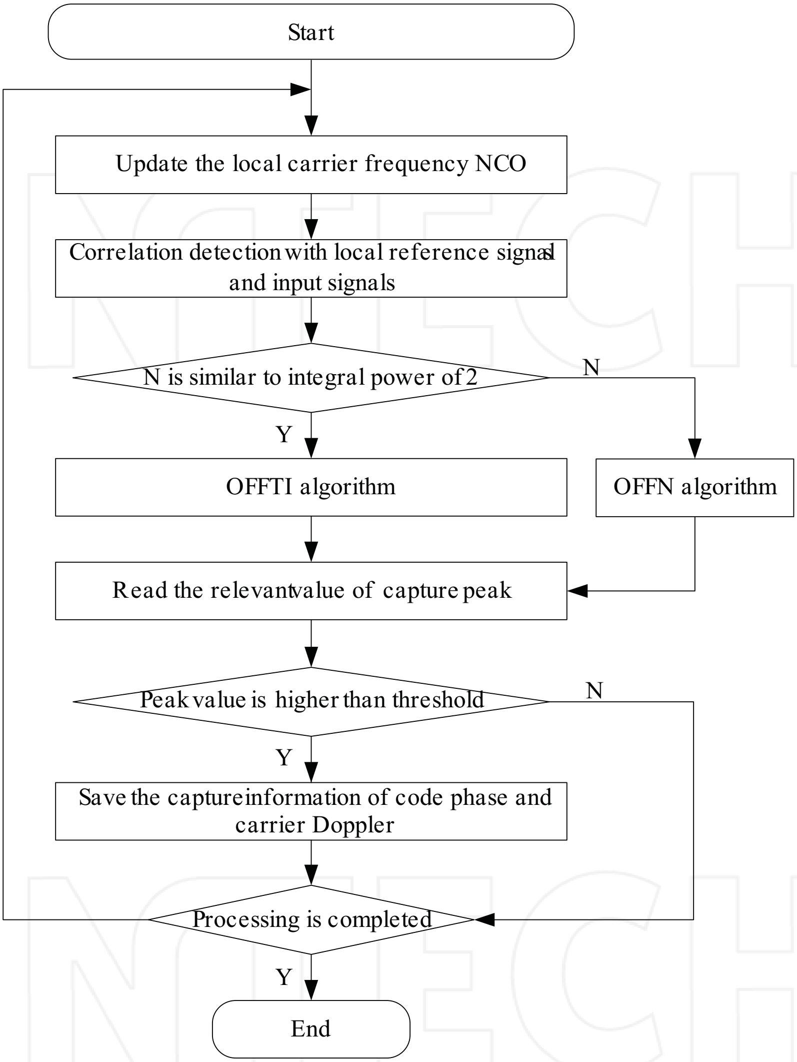
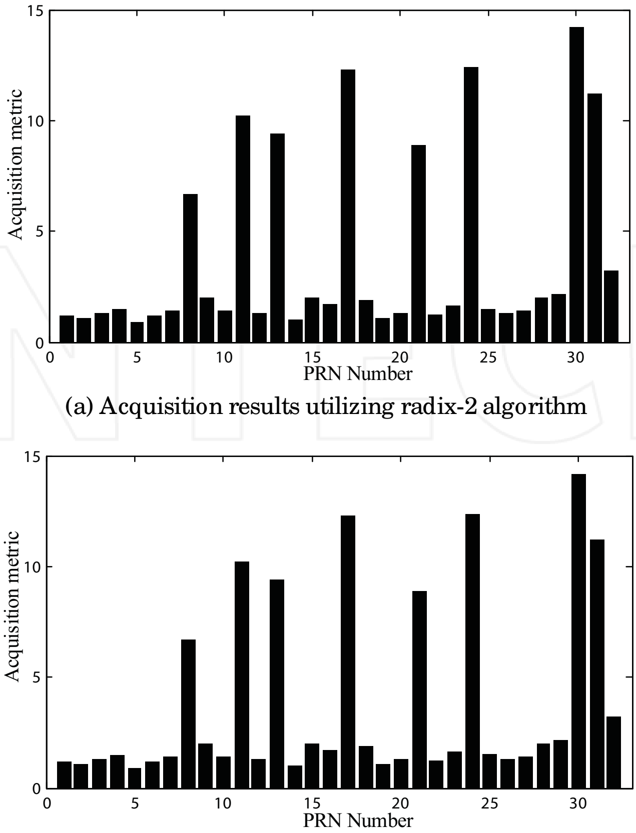

Start
Update the local carrier frequency NCO
Correlation detection with local reference signal and input signals
N is similar to integral power of 2
N
Y
OFFTI algorithm
OFFN algorithm
Read the relevant value of capture peak
Peak value is higher than threshold
N
Y
Save the capture information of code phase and carrier Doppler
Processing is completed
Y
End

Fig. 11. GPS signal acquisition utilizing optimized FFT

To further verify the advantages of processing efficiency with utilizing optimized FFT, compare the signal acquisition time in various Doppler shifts. The results are shown in Fig. 13, which indicate the processing efficiency of improved method has a significant superiority to the traditional methods.



(a) Acquisition results utilizing radix-2 algorithm

(b) Acquisition results utilizing improved algorithm

Fig. 12. Acquisition results utilizing different algorithms

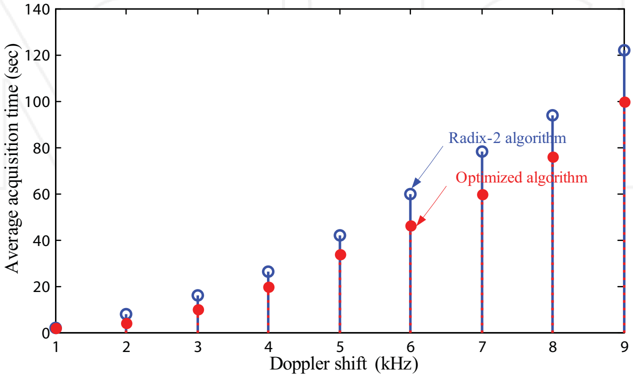


Fig. 13. Comparison of average acquisition time

7. Conclusion and future work

Apply an optimized FFT algorithm which integrates the traditional radix-2, radix-4, split-radix, PFA and WFTA to GPS C/ A code acquisition processing, and the primary results of simulation and experiment indicate that the optimized FFT algorithm could improve the average acquisition time and operation efficiency significantly. It is believed that this method could also be utilized in the long code acquisition, such as P code. Future work is to research and develop a more efficient and flexible processing platform to satisfy the demands of fast DFT calculation.

8. References

D. Akopian. (2005). Fast FFT based GPS satellite acquisition methods, *IEE Proceedings Radar, Sonar and Navigation*, vol. 152, no. 4, pp. 277-286.

D.J.R. Van Nee & A.J.R.M. Coenen. (1991). New fast GPS code-acquisition technique using FFT, *Electronic Letters*, vol. 27,no. 4, pp. 158-160.

J. Jin, S. Wu, & J. Li. (2005). Implementation of a new fast PN code-acquisition using radix-2 FFT, *Journal of Systems Engineering and Electronics*, vol. 27, no. 11, pp.1957-1960.

Elliott D. Kaplan, & Christopher J. Hegarty. (2006). Understanding GPS Principles and Applications (Second Edition), ARTECH HOUSE, INC., pp.113-152, Norwood.

Michael S. Braasch , & A. J. Van Dierendonck. (1999). GPS receiver architectures and measurements, *Proceedings of the IEEE*, vol. 87, no. 1, pp. 48-64.

C. Li, X. Zhang, H. Li, & Z. Zhang. (2008). Frequency-domain tracking method of GPS signals, *Chinese Journal of Electronics*, vol. 17, no.4, pp. 665-668.

P. Duhamel, & M. Vetterli. (1990). Fast fourier transforms: A tutorial review and a state of the art, *Signal Processing*, vol. 19, no. 4, pp. 259-299.

C. S. Burrus, & P. W. Eschenbacher. (1981). An In-Place, In-Order Prime Factor FFT Algorithm*, IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 4, pp. 806-817.

N. B. Jones, & J. D. McK. Watson. (1990). Digital signal processing: principles, devices, and applications, *Institution of Engineering And Technology*, pp. 43-77, United Kingdom.

D. Sundararajan. (2003). Digital signal processing: theory and practice, World Scientific Publishing Company, pp. 43-58, Singapore.

W. L. Mao, W. H. Lin, Y. F. Tseng, H. W. Tsao, & F. R. Chang. (2007). New Acquisition Method in GPS Software Receiver with Split-Radix FFT Technique, *9th International Conference on Advanced Communication Technology*, 722-727, February 2007, Phoenix Park, Korea.

C. S. Nagaraj, G. D. Andrew, & R. Chris. (2009). Application of Mixed-radix FFT Algorithms in Multi-band GNSS Signal Acquisition Engines, *Journal of Global Positioning Systems*, Vol. 8, No. 2, pp. 174-186.

S. Chu, & C. S. Burrus. (1982). A Prime Factor FFT Algorithm using Distributed Arithmetic, IEEE Transactions on Acoustics, *Speech, and Signal Processing*, vol. 30, no. 2, pp. 217-227.

B. Liu, & L.J. Zhang. (1997). An Improved Algorithm For Fast Fourier Transform, *Journal of Northeast Heavy Machinery Institute*, vol. 21, no. 4, pp. 296-300.

Winogard S. (1976). On computing the discrete Fourier transform, *Mathematics of Computation*, vol. 73, no. 4, pp. 1005-1006.

W. H. Lin, W. L. Mao, H. W. Tsao, F. R. Chang, & W. H. Huang. (2006). Acquisition of GPS Software Receiver Using Split-Radix FFT, *2006 IEEE Conference on Systems, Man, and Cybernetics*, pp. 4608-4613, October 2006, Taipei, Taiwan.

L. Zhao, S. Gao, & Y. Hao. (2009). Improved Fast Fourier Transform Processing on Fast Acquisition Algorithms for GPS Signals, *The Ninth International Conference on Electronic Measurement & Instruments*, vol. 4, pp. 221-224, August 2009, Beijing, China.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

10

Homogenization of Nonlocal Electrostatic Problems by Means of the Two-Scale Fourier Transform

Niklas Wellander
Swedish defence research agency (FOI) and Lund University
Sweden

1. Introduction

Multiple scales phenomena are ubiquitous, ranging from mechanical properties of wood, turbulent flow in gases and fluids, combustion, remote sensing of earth to wave propagation or heat conduction in composite materials. The obstacle with multi-scale problems is that they, due to limited primary memory even in the largest computational clusters, can not easily be modeled in standard numerical algorithms. Usually we are not even interested in the fine scale information in the processes. However, the fine scale properties are important for the macroscopic, effective, properties of for example a fiber composite. Attempts to find effective properties of composites dates back more than hundred years, e.g. see Faraday (1965); Maxwell (1954a;b); Rayleigh (1892). One way to find effective properties is to introduce a fine scale parameter, $\varepsilon > 0$, in the corresponding governing equations (modeling fast oscillating coefficients) and then study the asymptotic behavior of the sequence of solutions, and equations, when the fine scale parameter tends to zero. The limit yields the homogenized equations, that have constant coefficients (corresponding to homogeneous material properties). The discipline of partial differential equations dealing with such issues is called homogenization theory.

The foundation of homogenization theory was started by Spagnolo (1967) who introduced *G-convergence*, followed by Γ-*convergence* by Dal Maso (1993); De Giorgi (1975); De Giorgi & Franzoni (1975); De Giorgi & Spagnolo (1973), and *H-convergence* Tartar (1977). The *two-scale convergence* concept introduced by Nguetseng (1989) and developed by Allaire (1992); Allaire & Briane (1996) simplified many proofs. Floquet-Bloch expansion Bloch (1928); Floquet (1883) provides a method to find dispersion relations in the case the fine scales are on the same order as for example the wavelength of a propagating wave. The technique of Floquet-Bloch expansion can also be used to find the classical homogenized properties Allaire & Conca (1996); Bensoussan et al. (1978); Conca et al. (2002); Conca & Vanninathan (1997; 2002). *Two-scale transforms* have been introduced in different settings, Arbogast et al. (1990); Brouder & Rossano (2002); Cioranescu et al. (2002); Griso (2002); Laptev (2005); Nechvátal (2004). The general idea with the two-scale transform is to map bounded sequences of functions defined on $L^2(\Omega)$ to sequences defined on the product space $L^2(\Omega \times T^n)$ and then taking the weak limit in $L^2(\Omega \times T^n)$. Besides finding the effective material properties, one can also establish easily computed bounds of these. The bounds may be as simple as the arithmetic and harmonic averages, or more complex. For further reading we recommend the monograph by Milton (2002) as an introduction to the theory of composites.

In this paper we return to a two-scale Fourier transform, which belongs to the class of two-scale transforms, presented in Wellander (2004; 2007; 2009). The transform is applied to nonlocal constitutive relations in electrostatic applications for periodic composites. The current density is given as a spatial convolution of the electric field with a conductivity kernel. It turns out that the homogenized equation also posse's a nonlocal constitutive relation if we do not scale the non-localness. However, if we decrease the neighborhood which influence the current density simultaneously as we make the fine structure finer and finer then we are ending up with a constitutive relation which is local. To be strict, this is a three-scale problem. The finest scale is the variation of material properties. The second scale is the non-localness in the constitutive relation, and the third scale is the global equation, containing only the scales of the domain, boundary conditions and internal body forces.
The paper is organized in the following way. In Section 2 we give some basic definitions, mainly to do with two-scale convergence. In Section 3 we define and explore the two-scale Fourier transform and its application to homogenization of PDEs. In Section 4 we present the main assumptions and give some basic existence, uniqueness and a priori estimates. Section 5 is devoted to the main homogenization results. Some concluding remarks are given in Section 6.

2. Preliminaries

We begin to state the weak and two-scale convergence concepts. A bounded sequence $\{u^\varepsilon\}$ in $L^2(\Omega)$, where Ω is an open bounded set in $\mathbb{R}^n$, $n \geq 1$, with a Lipschitz continuous boundary $\partial\Omega$, has a subsequence which converges weakly in $L^2(\Omega)$, still denoted $\{u^\varepsilon\}$. That is,

$$\int_\Omega u^\varepsilon(x)\varphi(x)\,\mathrm{d}x \to \int_\Omega u(x)\varphi(x)\,\mathrm{d}x, \tag{1}$$

for all test functions $\varphi \in L^2(\Omega)$. We call u the weak limit of $\{u^\varepsilon\}$. Bounded sequences in $L^2(\Omega)$ does not imply strong convergence, i.e.,

$$\|u^\varepsilon - u\|_{L^2(\Omega)} \to 0$$

To study convergence of sequences with fast oscillations Nguetseng (1989) extended the class of test functions to functions with two scales, $\varphi \in C_0^\infty(\Omega; C^\infty(T^n))$, where T^n is the unit torus in $\mathbb{R}^n$. We will refer to two-scale convergence using smooth test functions as *distributional two-scale convergence*.

Definition 1. *A sequence $\{u^\varepsilon\}$ in $L^2(\Omega)$ is said to two-scale converge in a distributional sense to a function $u_0 = u_0(x,y)$ in $L^2(\Omega \times T^n)$ if*

$$\lim_{\varepsilon\to 0}\int_\Omega u^\varepsilon(x)\varphi\left(x,\frac{x}{\varepsilon}\right)\,\mathrm{d}x = \int_\Omega\int_{T^n} u_0(x,y)\varphi(x,y)\,\mathrm{d}y\mathrm{d}x, \tag{2}$$

for all test functions $\varphi \in C_0^\infty(\Omega; C^\infty(T^n))$.

The extension of weak convergence to *weak two-scale convergence* reads,

Definition 2. *A sequence $\{u^\varepsilon\}$ in $L^2(\Omega)$ is said to weakly two-scale converge to a function $u_0 = u_0(x,y)$ in $L^2(\Omega \times T^n)$ if*

$$\lim_{\varepsilon\to 0}\int_\Omega u^\varepsilon(x)\varphi\left(x,\frac{x}{\varepsilon}\right)\,\mathrm{d}x = \int_\Omega\int_{T^n} u_0(x,y)\varphi(x,y)\,\mathrm{d}y\mathrm{d}x, \tag{3}$$

for all test functions $\varphi \in L^2(\Omega; C(T^n))$.

A more general class of *admissible test functions* are those that two-scale converge strongly, *i.e.*, functions defined as

Definition 3. *If a sequence $\{u^\varepsilon\}$ in $L^2(\Omega)$ weakly two-scale converge to $u_0 \in L^2(\Omega \times T^n)$ and*

$$\lim_{\varepsilon\to 0} \|u^\varepsilon\|_{L^2(\Omega)} = \|u_0\|_{L^2(\Omega\times T^n)}, \tag{4}$$

then it is said to two-scale converge strongly to $u_0 \in L^2(\Omega \times T^n)$.

Strongly two-scale converging functions are called admissible test functions. Some examples are functions in $L^2(\Omega; C(T^n))$, or for Ω bounded, $C(\overline{\Omega}; C(T^n))$ or $L^2(T^n; C(\overline{\Omega}))$. See Allaire (1992) for more details regarding this issue. The basic compactness results, Nguetseng (1989), reads

Theorem 1. *For every bounded sequence $\{u^\varepsilon\}$ in $L^2(\Omega)$ there exists a subsequence and a function u_0 in $L^2(\Omega \times T^n)$ such that u^ε two-scale converges weakly to u_0.*

Theorem 2. *Assume that $\{u^\varepsilon\}$ is a bounded sequence in $H^1(\Omega)$. Then there exists a subsequence, still denoted $\{u^\varepsilon\}$, which two-scale converges weakly to $u_0 = u$, and ∇u^ε two-scale converges weakly to $\nabla_x u + \nabla_y u_1$. Here u is the weak $L^2(\Omega)$-limit in (1) and $u_1 \in L^2(\Omega; H^1(T^n))$.*

By the Rellich theorem, u is the strong L^2-limit of the sequence $\{u^\varepsilon\}$. We close this section by definition of some nonstandard function spaces.

$$\begin{aligned}
&H(\mathrm{div}, \Omega) := \{\boldsymbol{u} \in L^2(\Omega; \mathbb{R}^n) : \mathrm{div}\, \boldsymbol{u} \in L^2(\Omega)\}\\
&H(\mathrm{curl}, \Omega) := \{\boldsymbol{u} \in L^2(\Omega; \mathbb{R}^3) : \mathrm{curl}\, \boldsymbol{u} \in L^2(\Omega; \mathbb{R}^3)\}\\
&L^p(\mathrm{div}, \Omega) := \{\boldsymbol{u} \in L^p(\Omega; \mathbb{R}^n) : \mathrm{div}\, \boldsymbol{u} \in L^p(\Omega)\}\\
&L^p(\mathrm{curl}, \Omega) := \{\boldsymbol{u} \in L^p(\Omega; \mathbb{R}^3) : \mathrm{curl}\, \boldsymbol{u} \in L^p(\Omega; \mathbb{R}^3)\}\\
&l^{1,2}(\mathbb{Z}^n) := \{\phi \in l^2(\mathbb{Z}^n) : 2\pi i \boldsymbol{m} \phi(\boldsymbol{m}) \in l^2(\mathbb{Z}^n; \mathbb{C}^n) \forall \boldsymbol{m} \in \mathbb{Z}^n\}\\
&l^2(\mathrm{div}, \mathbb{Z}^n; \mathbb{C}^n) := \{\boldsymbol{\phi} \in l^2(\mathbb{Z}^n; \mathbb{C}^n) : 2\pi i \boldsymbol{m} \cdot \boldsymbol{\phi}(\boldsymbol{m}) \in l^2(\mathbb{Z}^n) \forall \boldsymbol{m} \in \mathbb{Z}^n\}\\
&l^2(\mathrm{curl}, \mathbb{Z}^3; \mathbb{C}^3) := \{\boldsymbol{\phi} \in l^2(\mathbb{Z}^3; \mathbb{C}^3) : 2\pi i \boldsymbol{m} \times \boldsymbol{\phi}(\boldsymbol{m}) \in l^2(\mathbb{Z}^3; \mathbb{C}^3) \forall \boldsymbol{m} \in \mathbb{Z}^3\}
\end{aligned}$$

3. The two-scale fourier transform

We define the *two-scale Fourier transform*, which is nothing but the standard Fourier transform evaluated at $\boldsymbol{\xi} + \varepsilon^{-1}\boldsymbol{m}$ where $\boldsymbol{\xi}$ is restricted to a cube in $\mathbb{R}^n$ with sidelength $1/\varepsilon$, Wellander (2009).

Definition 4 (Two-scale Fourier transform)**.** *For any function f in $L^1(\mathbb{R}^n)$ and every $0 < \varepsilon$ the* two-scale Fourier transform at the ε-scale *of f is defined by*

$$\mathcal{F}_\varepsilon\{f\}(\boldsymbol{\xi}, \boldsymbol{m}) = \widehat{f}_\varepsilon(\boldsymbol{\xi}, \boldsymbol{m}) = \int_{\mathbb{R}^n} f(\boldsymbol{x}) e^{-2\pi i \boldsymbol{x} \cdot (\boldsymbol{\xi} + \frac{\boldsymbol{m}}{\varepsilon})} \, \mathrm{d}\boldsymbol{x},$$

for all $\boldsymbol{\xi} \in \left]-\frac{1}{2\varepsilon}, \frac{1}{2\varepsilon}\right[^n$, $\boldsymbol{m} \in \mathbb{Z}^n$. The inverse is given by

$$\mathcal{F}_\varepsilon^{-1}\{\widehat{f}_\varepsilon\}(\boldsymbol{x}) = \sum_{\boldsymbol{m} \in \mathbb{Z}^n} \int_{\boldsymbol{\xi} \in \left]-\frac{1}{2\varepsilon}, \frac{1}{2\varepsilon}\right[^n} \widehat{f}_\varepsilon(\boldsymbol{\xi}, \boldsymbol{m}) e^{2\pi i \boldsymbol{x} \cdot (\boldsymbol{\xi} + \frac{\boldsymbol{m}}{\varepsilon})} \, \mathrm{d}\boldsymbol{\xi}.$$

The forward transform is well defined for any $\boldsymbol{\xi}$ in $\mathbb{R}^n$. For the inverse we only need the ones in the cube $]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}[^n$. For fixed ε, the transform is the usual Fourier transform, where we for each $\boldsymbol{m}$ integrate over the cube $]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}[^n$ with respect to $\boldsymbol{\xi}$ in the inner loop, see Figure 1 for the one dimensional case. It is a question of cutting the frequency space into n-dimensional cubes of side length $1/\varepsilon$ centered at the points $\boldsymbol{m}/\varepsilon$ and summing up the contribution from each cube. When $\varepsilon \to 0$ then $\boldsymbol{\xi}$ belongs to the whole real space, $\mathbb{R}^n$. The standard Fourier transform is recovered if we let $\boldsymbol{m} = \boldsymbol{0}$ and permit $\boldsymbol{\xi}$ to take any value in $\mathbb{R}^n$ for all $\varepsilon > 0$. The cube $]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}[^n$ corresponds precisely to the first Brillouin zone appearing in the Floquet-Bloch theory Bloch (1928); Floquet (1883) which is extensively used in solid state physics.

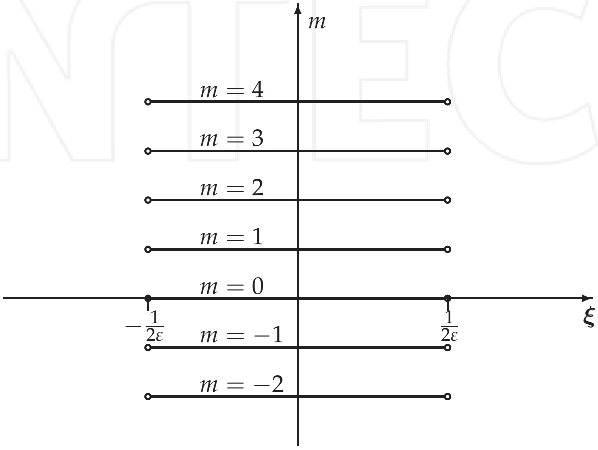


Fig. 1. The Fourier indices used in the inverse transform. Pieces of length $1/\varepsilon$ centered at m/ε are cut from the ζ-axis and stacked along the m-axis. Hence, the pieces labeled $m = 0$ and $m = 1$ corresponds to the intervals $]-1/2\varepsilon, 1/2\varepsilon[$ and $]1/2\varepsilon, 3/2\varepsilon[$ on the $\boldsymbol{\xi} - axis$, respectively.

The transform can be defied as a mapping $L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n; l^2(\mathbb{Z}^n))$, and then by interpolation to $L^p(\mathbb{R}^n)$, $p \in [1,2]$ with values in $L^q(\mathbb{R}^n; l^q(\mathbb{Z}^n))$, $\frac{1}{p} + \frac{1}{q} = 1$. Here $l^p(\mathbb{Z}^n)$ is the space of all sequences indexed by the n-tuple of integers, equipped with the usual p-norm. The transform extends to the Fourier theory for tempered distributions Taylor (1996), but in this case we have to include one of the boundaries of the Brillouin zone in the definition of the inverse transform. For example, the semi open cube $[-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}[^n$ would be suitable. By doing this modification we will not exclude Dirac functions with support on the boundary of the Brillouin zone. Keeping the same notation as in the L^1-case we define the L^p-version of the two-scale Fourier transform.

Definition 5. *For any function f in $L^p(\mathbb{R}^n)$, $p \in [1,2]$, and every $0 < \varepsilon$ the* two-scale Fourier transform at the ε-scale *of f is defined by*

$$\mathcal{F}_\varepsilon\{f\}(\boldsymbol{\xi},\boldsymbol{m}) = \widehat{f}_\varepsilon(\boldsymbol{\xi},\boldsymbol{m}) = \int_{\mathbb{R}^n} f(x)e^{-2\pi i\boldsymbol{x}\cdot\left(\boldsymbol{\xi}+\frac{\boldsymbol{m}}{\varepsilon}\right)}\,\mathrm{d}\boldsymbol{x}$$

for all $\boldsymbol{\xi} \in \left]-\frac{1}{2\varepsilon}, \frac{1}{2\varepsilon}\right[^n, \boldsymbol{m} \in \mathbb{Z}^n$. *The inverse is given by*

$$\mathcal{F}_\varepsilon^{-1}\{\widehat{f}_\varepsilon\}(\boldsymbol{x}) = \sum_{\boldsymbol{m}\in\mathbb{Z}^n} \int_{\boldsymbol{\xi}\in\left]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}\right[^n} \widehat{f}_\varepsilon(\boldsymbol{\xi}, \boldsymbol{m}) e^{2\pi i \boldsymbol{x}\cdot\left(\boldsymbol{\xi}+\frac{\boldsymbol{m}}{\varepsilon}\right)} \, \mathrm{d}\boldsymbol{\xi}.$$

We have Parseval-Plancherel's relations, which holds because of the corresponding identities for the usual Fourier transform.

Theorem 3. *(Parseval-Plancherel) Suppose that* $\boldsymbol{f}$ *and* $\boldsymbol{g}$ *belong to* $L^2(\mathbb{R}^n; \mathbb{R}^p)$. *Then for every* $\varepsilon > 0$

$$\int_{\mathbb{R}^n} \boldsymbol{f}(\boldsymbol{x}) \cdot \boldsymbol{g}(\boldsymbol{x}) \, dx = \sum_{\boldsymbol{m}\in\mathbb{Z}^n} \int_{\boldsymbol{\xi}\in\left]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}\right[^n} \widehat{\boldsymbol{f}}_\varepsilon(\boldsymbol{\xi}, \boldsymbol{m}) \cdot \overline{\widehat{\boldsymbol{g}}_\varepsilon(\boldsymbol{\xi}, \boldsymbol{m})} \, \mathrm{d}\boldsymbol{\xi},$$

$$\|\boldsymbol{f}\|_{L^2(\mathbb{R}^n;\mathbb{R}^p)} = \left(\sum_{\boldsymbol{m}\in\mathbb{Z}^n} \int_{\boldsymbol{\xi}\in\left]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}\right[^n} |\widehat{\boldsymbol{f}}_\varepsilon(\boldsymbol{\xi}, \boldsymbol{m})|^2 \, \mathrm{d}\boldsymbol{\xi} \right)^{1/2} = \|\widehat{\boldsymbol{f}}_\varepsilon\|_{L^2\left(\left]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}\right[^n \times\mathbb{Z}^n;\mathbb{R}^p\right)}.$$

The following properties of the two-scale Fourier transform follow at once from the usual Fourier transform.

Proposition 1. *The two-scale Fourier transform has the following properties,*

(i) $\mathcal{F}_\varepsilon\{fg\} = \mathcal{F}_\varepsilon\{f\} * \mathcal{F}_\varepsilon\{g\}$, *for* $f, g \in L^2(\mathbb{R}^n)$.

(ii) $\mathcal{F}_\varepsilon\{f * g\} = \mathcal{F}_\varepsilon\{f\}\mathcal{F}_\varepsilon\{g\}$ *for* $f \in L^1(\mathbb{R}^n)$, $g \in L^p(\mathbb{R}^n)$, $p \in [1, 2]$.

(iii)$\mathcal{F}_\varepsilon\{\nabla \boldsymbol{u}\}(\boldsymbol{\xi}, \boldsymbol{m}) = 2\pi i(\boldsymbol{\xi} + \varepsilon^{-1}\boldsymbol{m})\mathcal{F}_\varepsilon\{\boldsymbol{u}\}(\boldsymbol{\xi}, \boldsymbol{m})$ *for* $\boldsymbol{u} \in W^{1,p}(\mathbb{R}^n)$, $p \in [1, 2]$.

(iv)$\mathcal{F}_\varepsilon\{\nabla \cdot \boldsymbol{u}\}(\boldsymbol{\xi}, \boldsymbol{m}) = 2\pi i(\boldsymbol{\xi} + \varepsilon^{-1}\boldsymbol{m}) \cdot \mathcal{F}_\varepsilon\{\boldsymbol{u}\}(\boldsymbol{\xi}, \boldsymbol{m})$ *for* $\boldsymbol{u} \in L^p(\mathrm{div}, \mathbb{R}^n)$, $p \in [1, 2]$.

(v) $\mathcal{F}_\varepsilon\{\nabla \times \boldsymbol{u}\}(\boldsymbol{\xi}, \boldsymbol{m}) = 2\pi i(\boldsymbol{\xi} + \varepsilon^{-1}\boldsymbol{m}) \times \mathcal{F}_\varepsilon\{\boldsymbol{u}\}(\boldsymbol{\xi}, \boldsymbol{m})$ *for* $\boldsymbol{u} \in L^p(\mathrm{curl}, \mathbb{R}^3)$, $p \in [1, 2]$.

(vi)$\mathcal{F}_\varepsilon\{ue^{-2\pi i\boldsymbol{x}\cdot\left(\boldsymbol{\eta}+\frac{\boldsymbol{s}}{\varepsilon}\right)}\}(\boldsymbol{\xi}, \boldsymbol{m}) = \mathcal{F}_\varepsilon\{u\}(\boldsymbol{\xi} + \boldsymbol{\eta}, \boldsymbol{m} + \boldsymbol{s})$ *for* $u \in L^p(\mathbb{R}^n)$, $p \in [1, 2]$,
$\boldsymbol{\xi} \in]-1/2\varepsilon, 1/2\varepsilon[^n$.

The convolution in Fourier space ($*$) is defined as

$$\mathcal{F}_\varepsilon\{f\} * \mathcal{F}_\varepsilon\{g\}(\boldsymbol{\xi}, \boldsymbol{m}) = \sum_{\boldsymbol{s}\in\mathbb{Z}^n} \int_{\boldsymbol{\eta}\in\left]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}\right[^n} \mathcal{F}_\varepsilon\{f\}(\boldsymbol{\eta}, \boldsymbol{s})\mathcal{F}_\varepsilon\{g\}(\boldsymbol{\xi} - \boldsymbol{\eta}, \boldsymbol{m} - \boldsymbol{s}) \, \mathrm{d}\boldsymbol{\eta},$$

for $\boldsymbol{\xi} \in]-1/2\varepsilon, 1/2\varepsilon[^n$. Translated functions like $\mathcal{F}_\varepsilon\{g\}(\cdot, \boldsymbol{m} - \boldsymbol{s})$ are extended by zero outside $]-1/2\varepsilon, 1/2\varepsilon[^n$ for all $\boldsymbol{m}$ and $\boldsymbol{s}$ in $\mathbb{Z}^n$.
The admissible test functions (as in Definition 3) converge strongly in Fourier space.

Proposition 2. *Assume sequence* $\{\phi^\varepsilon\}$ *two-scale converges strongly to* ϕ. *Extend* $\widehat{\phi}^\varepsilon_\varepsilon(\cdot, \boldsymbol{m})$ *by zero outside* $\left]-\frac{1}{2\varepsilon}, \frac{1}{2\varepsilon}\right[^n$ *for all* $\boldsymbol{m} \in \mathbb{Z}^n$ *then*

$$\widehat{\phi}^\varepsilon_\varepsilon(\boldsymbol{\xi}, \boldsymbol{m}) \to \widehat{\phi}(\boldsymbol{\xi}, \boldsymbol{m}) \text{ strongly in } L^2(\mathbb{R}^n \times \mathbb{Z}^n).$$

Proof: By assumption $\{\phi^\varepsilon\}$ is bounded in $L^2(\Omega)$. It follows that the two-scale Fourier transformed sequence $\widehat{\phi}^\varepsilon_\varepsilon(\boldsymbol{\xi}, \boldsymbol{m})$ is bounded in $L^2\left(\left]-\frac{1}{2\varepsilon}, \frac{1}{2\varepsilon}\right[^n \times \mathbb{Z}^n\right)$. The extended function is bounded in $L^2(\mathbb{R}^n \times \mathbb{Z}^n)$ and converges weakly in $L^2(\mathbb{R}^n \times \mathbb{Z}^n)$. Further, Parseval-Plancherel (Theorem 3) yields

$$\lim_{\varepsilon\to 0} \|\widehat{\phi}^{\varepsilon}_{\varepsilon}\|_{L^2(\left]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}\right[^n \times \mathbb{Z}^n)} = \lim_{\varepsilon\to 0} \|\widehat{\phi}^{\varepsilon}_{\varepsilon}\|_{L^2(\mathbb{R}^n\times\mathbb{Z}^n)} =$$
$$\lim_{\varepsilon\to 0} \|\phi^{\varepsilon}_{\varepsilon}\|_{L^2(\Omega)} = \|\phi\|_{L^2(\Omega\times T^n)} = \|\widehat{\phi}\|_{L^2(\mathbb{R}^n\times\mathbb{Z}^n)}.$$

The statement follows since the sequence converges weakly and in norm. □

We continue by restating Nguetseng's two-scale compactness theorem (Theorems 1 and 2) in Fourier space.

Proposition 3. *Let $\{u^\varepsilon\}$ be a uniformly bounded sequence in $L^2(\mathbb{R}^n)$ and $\boldsymbol{\xi} \in \mathbb{R}^n$, $\boldsymbol{m} \in \mathbb{Z}^n$ arbitrary.*

(i) *If ϕ^ε two-scale converge strongly to ϕ (as in Proposition 2) then there exists a subsequence and $u^0 \in L^2(\Omega \times T^n)$ such that*

$$\lim_{\varepsilon\to 0} \widehat{(u^\varepsilon\phi^\varepsilon)}_\varepsilon(\boldsymbol{\xi},\boldsymbol{m}) = \widehat{(u^0\phi)}(\boldsymbol{\xi},\boldsymbol{m}).$$

(ii) *The sequence $\widehat{u}^\varepsilon_\varepsilon(\,\cdot\,,\boldsymbol{m})$ extended by zero outside $\left]-\frac{1}{2\varepsilon},\frac{1}{2\varepsilon}\right[^n$*

$$\widehat{u}^\varepsilon_\varepsilon \rightharpoonup \widehat{u}^0$$

weakly in $L^2(\mathbb{R}^n \times \mathbb{Z}^n)$.

(iii) *If there exists a compact set K in $\mathbb{R}^n$ and a positive number ε_0 such that $\operatorname{supp} u^\varepsilon \subset K$, for all $\varepsilon < \varepsilon_0$, then*

$$\lim_{\varepsilon\to 0} \widehat{u}^\varepsilon_\varepsilon(\boldsymbol{\xi},\boldsymbol{m}) = \widehat{u}^0(\boldsymbol{\xi},\boldsymbol{m}).$$

pointwise in $\mathbb{R}^n \times \mathbb{Z}^n$.

Remark 1. *If $\{u^\varepsilon\}$ in Proposition 3 (iii) is uniformly bounded in $L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n)$ (or just bounded in $L^1(\mathbb{R}^n)$) then the convergence is pointwise in Fourier space. That is due to the fact that a subsequence of $\{u^\varepsilon\}$ converges weakly in $L^1(\mathbb{R}^n)$ and $e^{-2\pi i(x\cdot\boldsymbol{\xi}+y\cdot m)}$ is a function in $L^\infty(\mathbb{R}^n \times T^n)$.*

We have the following corollary which follows from Proposition 3

Corollary 1. *If $\{u^\varepsilon\}$ is a bounded sequence in $L^2(\mathbb{R}^n)$ and if there exists a compact set K in $\mathbb{R}^n$ and a positive number ε_0 such that $\operatorname{supp} u^\varepsilon \subset K$, for all $\varepsilon < \varepsilon_0$ (or if $\{u^\varepsilon\}$ is bounded in $L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n)$), then there exists a subsequence such that,*

$$\mathcal{F}_\varepsilon\{u^\varepsilon\}(\boldsymbol{\xi},0) \to \widehat{u}^0(\boldsymbol{\xi},0),$$

as $\varepsilon \to 0$, for all $\boldsymbol{\xi} \in \mathbb{R}^n$. Here $\widehat{u}^0(\boldsymbol{\xi},0)$ is the Fourier transform of the weak limit of $\{u^\varepsilon\}$ in $L^2(\Omega)$.

Remark 2. *Proposition 3 (i) can be illustrated by the following commutative diagram*

$$\begin{tikzcd}
u^\varepsilon\phi^\varepsilon \arrow[r, "2-s"] \arrow[d, "\mathcal{F}_\varepsilon"] & u^0\phi \arrow[d, "\mathcal{F}"] \\
\widehat{(u^\varepsilon\phi^\varepsilon)}_\varepsilon(\boldsymbol{\xi},\boldsymbol{m}) \arrow[r, "pointwise"] & \widehat{(u^0\phi)}(\boldsymbol{\xi},\boldsymbol{m})
\end{tikzcd}$$

Assertions (ii) and (iii) are illustrated by

$$\begin{tikzcd} u^{\varepsilon} \arrow[r, "2-s"] \arrow[d, "\mathcal{F}_{\varepsilon}"] & u^{0} \arrow[d, "\mathcal{F}"] \\ \widehat{u}^{\varepsilon}_{\varepsilon}(\boldsymbol{\xi}, \boldsymbol{m}) \arrow[r, "weakly/pointwise"] & \widehat{u}^{0}(\boldsymbol{\xi}, \boldsymbol{m}) \end{tikzcd}$$

which indicates that for sequences defined on bounded domains the two-scale convergence becomes (by considering the exponential function as a test function) pointwise convergence in Fourier space.

Next we give some compactness results for the two-scale Fourier transform. The first one asserts that we recover the standard Fourier transform of any function in L^2 as the limit of the two-scale Fourier transformed function.

Proposition 4. *Let $u \in L^2(\mathbb{R}^n)$ and $\widehat{u}$ be the standard Fourier transform of u. Then,*

$$\lim_{\varepsilon \to 0} \mathcal{F}_{\varepsilon}\{u\}(\boldsymbol{\xi}, \boldsymbol{m}) = \widehat{u}(\boldsymbol{\xi})\delta_{\boldsymbol{m}\mathbf{0}},$$

pointwise for all $\boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{m} \in \mathbb{Z}^n$.

Here $\mathbf{0}$ is the n-dimensional null vector and $\delta_{\boldsymbol{kl}}$ is the Kronecker delta defined by

$$\delta_{\boldsymbol{kl}} = \begin{cases} 1, & \boldsymbol{k} = \boldsymbol{l}, \\ 0, & \boldsymbol{k} \neq \boldsymbol{l} \end{cases}$$

We find that a sequence of scaled periodic functions are recovered as the Fourier transform of the unscaled function.

Proposition 5. *Let $u \in L^2(T^n)$, and define $u^{\varepsilon}(\boldsymbol{x}) = u(\boldsymbol{x}/\varepsilon)$. Then,*

$$\mathcal{F}_{\varepsilon}\{u^{\varepsilon}\}(\mathbf{0}, \boldsymbol{m}) = \widehat{u}(\boldsymbol{m})$$

for all $0 < \varepsilon$ such that $1/\varepsilon$ is an integer, $\boldsymbol{m} \in \mathbb{Z}^n$, and

$$\lim_{\varepsilon \to 0} \mathcal{F}_{\varepsilon}\{u^{\varepsilon}\}(\boldsymbol{\xi}, \boldsymbol{m}) = \widehat{u}(\boldsymbol{m})$$

for all $\boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{m} \in \mathbb{Z}^n$. Here,

$$\widehat{u}(\boldsymbol{m}) = \int_{T^n} u(\boldsymbol{x}) e^{-2\pi i \boldsymbol{x} \cdot \boldsymbol{m}} \, \mathrm{d}\boldsymbol{x},$$

Proof: The definition of the two-scale Fourier transform, Definition 5, yields

$$\mathcal{F}_{\varepsilon}\{u^{\varepsilon}\}(\mathbf{0}, \boldsymbol{m}) = \int_{T^n} u(\boldsymbol{x}/\varepsilon) e^{-2\pi i \boldsymbol{x} \cdot (\mathbf{0} + \frac{\boldsymbol{m}}{\varepsilon})} \, \mathrm{d}\boldsymbol{x} = \varepsilon^n \int_{T^n_{1/\varepsilon}} u(\boldsymbol{s}) e^{-2\pi i \boldsymbol{s} \cdot (\mathbf{0} + \boldsymbol{m})} \, \mathrm{d}\boldsymbol{s} =$$

$$= \varepsilon^n \sum_{1}^{\varepsilon^{-n}} \int_{T^n} u(\boldsymbol{s}) e^{-2\pi i \boldsymbol{s} \cdot (\mathbf{0} + \boldsymbol{m})} \, \mathrm{d}\boldsymbol{s} = \int_{T^n} u(\boldsymbol{s}) e^{-2\pi i \boldsymbol{s} \cdot (\mathbf{0} + \boldsymbol{m})} \, \mathrm{d}\boldsymbol{s} = \widehat{u}(\boldsymbol{m})$$

for all $0 < \varepsilon$ such that $1/\varepsilon$ is an integer, $\boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{m} \in \mathbb{Z}^n$. Here, $T^n_{1/\varepsilon}$ is the $1/\varepsilon$-torus in $\mathbb{R}^n$. The second statement follows by similar arguments. □

In the next three propositions we will assume that there exists an $\varepsilon_0 > 0$ such that for all $\varepsilon < \varepsilon_0$ the support of all sequences are contained in a compact set K in $\mathbb{R}^n$. It follows that $e^{-2\pi i \boldsymbol{x}\cdot(\boldsymbol{\xi}+\frac{\boldsymbol{m}}{\varepsilon})}$ belongs to $L^2(K)$ and is an admissible test function in the two-scale convergence sense. If the support is not compact then the convergence in Fourier space will be weak in L^2, as in Proposition 3 (iii). The proofs will be similar in these cases, just multiply with test functions in $L^2(\mathbb{R}^n \times \mathbb{Z}^n)$ before taking the limits. Alternatively, we can localize the sequence first by multiplying with functions $\phi \in C_0(\Omega)$.

Proposition 6. *If $\{u^\varepsilon\}$ is a bounded sequence in $H^1(\mathbb{R}^n)$ then there exists a subsequence such that,*

(i) $\mathcal{F}_\varepsilon\{u^\varepsilon\}(\boldsymbol{\xi},\boldsymbol{m}) \to \widehat{u}(\boldsymbol{\xi})\delta_{\boldsymbol{m0}}$,

(ii) $\mathcal{F}_\varepsilon\{\nabla u^\varepsilon\}(\boldsymbol{\xi},\boldsymbol{m}) \to 2\pi i\boldsymbol{\xi}\widehat{u}(\boldsymbol{\xi})\delta_{\boldsymbol{m0}} + 2\pi i\boldsymbol{m}\widehat{u}^1(\boldsymbol{\xi},\boldsymbol{m})$

as $\varepsilon \to 0$, for all $\boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{m} \in \mathbb{Z}^n$. Here $\widehat{u}$ is the standard Fourier transform of u which is the weak limit of u^ε in $L^2(\mathbb{R}^n)$, and $\widehat{u}^1 \in L^2(\mathbb{R}^n; l^{1,2}(\mathbb{Z}^n))$ is the Fourier transform of a function $u^1 \in L^2(\mathbb{R}^n; H^1(T^n))$.

Proposition 7. *If $\{\boldsymbol{u}^\varepsilon\}$ is a bounded sequence in $H(\mathrm{div}, \mathbb{R}^n)$ then there exists a subsequence such that,*

(i) $\mathcal{F}_\varepsilon\{\boldsymbol{u}^\varepsilon\}(\boldsymbol{\xi},\boldsymbol{m}) \to \widehat{\boldsymbol{u}}^0(\boldsymbol{\xi},\boldsymbol{m})$, *with* $2\pi i\boldsymbol{m}\cdot\widehat{\boldsymbol{u}}^0(\boldsymbol{\xi},\boldsymbol{m}) = 0$

(ii) $\mathcal{F}_\varepsilon\{\nabla\cdot\boldsymbol{u}^\varepsilon\}(\boldsymbol{\xi},\boldsymbol{m}) \to 2\pi i\boldsymbol{\xi}\cdot\widehat{\boldsymbol{u}}(\boldsymbol{\xi})\delta_{\boldsymbol{m0}} + 2\pi i\boldsymbol{m}\cdot\widehat{\boldsymbol{u}}^1(\boldsymbol{\xi},\boldsymbol{m})$

as $\varepsilon \to 0$, for all $\boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{m} \in \mathbb{Z}^n$, where $\widehat{\boldsymbol{u}}(\boldsymbol{\xi}) = \widehat{\boldsymbol{u}}^0(\boldsymbol{\xi}, 0)$ is the standard Fourier transform of $\boldsymbol{u} \in H(\mathrm{div}, \mathbb{R}^n)$, $\boldsymbol{u}(x) = \int_{T^n} u^0(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}$ and $\widehat{\boldsymbol{u}}^1 \in L^2(\mathbb{R}^3; l^2(\mathrm{div}, \mathbb{Z}^3; \mathbb{C}^n))$ is the Fourier transform of a function $\boldsymbol{u}^1 \in L^2(\mathbb{R}^n; H(\mathrm{div}, T^n))$.

Proposition 8. *If $\{\boldsymbol{u}^\varepsilon\}$ is a bounded sequence in $H(\mathrm{curl}, \mathbb{R}^3)$ then there exists a subsequence such that,*

(i) $\mathcal{F}_\varepsilon\{\boldsymbol{u}^\varepsilon\}(\boldsymbol{\xi},\boldsymbol{m}) \to \widehat{\boldsymbol{u}}^0(\boldsymbol{\xi},\boldsymbol{m}) = \widehat{\boldsymbol{u}}(\boldsymbol{\xi})\delta_{\boldsymbol{m0}} + 2\pi i\boldsymbol{m}\widehat{\phi}(\boldsymbol{\xi},\boldsymbol{m})$, *with* $2\pi i\boldsymbol{m}\times\widehat{\boldsymbol{u}}^0(\boldsymbol{\xi},\boldsymbol{m}) = 0$

(ii) $\mathcal{F}_\varepsilon\{\nabla\times\boldsymbol{u}^\varepsilon\}(\boldsymbol{\xi},\boldsymbol{m}) \to 2\pi i\boldsymbol{\xi}\times\widehat{\boldsymbol{u}}(\boldsymbol{\xi})\delta_{\boldsymbol{m0}} + 2\pi i\boldsymbol{m}\times\widehat{\boldsymbol{u}}^1(\boldsymbol{\xi},\boldsymbol{m})$

as $\varepsilon \to 0$, for all $\boldsymbol{\xi} \in \mathbb{R}^3, \boldsymbol{m} \in \mathbb{Z}^3$. Here $\widehat{\boldsymbol{u}}(\boldsymbol{\xi}) = \widehat{\boldsymbol{u}}^0(\boldsymbol{\xi}, 0)$ is the Fourier transform of $\boldsymbol{u}(\boldsymbol{x}) = \int_{T^n} u^0(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}$, $\boldsymbol{u} \in H(\mathrm{curl}, \mathbb{R}^3)$, $\widehat{\phi} \in L^2(\mathbb{R}^3; l^2(\mathbb{Z}^3))$ is the Fourier transform of a function $\phi \in L^2(\mathbb{R}^3; H^1(T^3))$ and $\widehat{\boldsymbol{u}}^1 \in L^2(\mathbb{R}^3; l^2(\mathrm{curl}, \mathbb{Z}^3; \mathbb{C}^3))$ is the Fourier transform of a function $\boldsymbol{u}^1 \in L^2(\mathbb{R}^3; H(\mathrm{curl}, T^3))$.

4. The non-local homogenization problems

We will consider two non-local elliptic problems. The physical problem in mind is a nonlocal electrostatic equation for a periodic composite. This is an elliptic problem with spatial convolution of the electric field with a conductivity, which consists of a periodic part multiplied with a localizing function. The localizer gives a finite contribution to the current density when convoluted with the electric fields in the neighborhood of the observation point.

4.1 Assumptions and weak formulation

The domain, Ω, is assumed to be a bounded subset of $\mathbb{R}^n$, $n \in \mathbb{N}$ with a Lipshitz boundary $\partial\Omega$. We assume the current density is given by a spatial convolution of the electric field with a nonlocal kernel $\mathbf{K}$ which gives the current density contribution at a point due to the electric field in the neighborhood of $\boldsymbol{x}$,

$$\boldsymbol{J}(\boldsymbol{x}, \nabla\phi) = \boldsymbol{J}(\boldsymbol{x}) = \int_\Omega \mathbf{K}(\boldsymbol{x}-\boldsymbol{\xi})\nabla\phi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}. \tag{5}$$

The kernel maps electric fields to current densities ($\mathbb{R}^n \to \mathbb{R}^n$) and decays monotonically for large arguments. To model the fine scale structure in a heterogeneous material we introduce the fine scale parameter $\varepsilon > 0$. The scaled current density is given by

$$\boldsymbol{J}^\varepsilon(\boldsymbol{x}) = \int_\Omega \mathbf{K}^\varepsilon(\boldsymbol{x}-\boldsymbol{\xi})\nabla\phi^\varepsilon(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} \tag{6}$$

where ϕ^ε is the electric potential. We integrate over the support of $\mathbf{K}^\varepsilon$ which overlaps Ω, which has to be taken into account close to the boundary $\partial\Omega$. The static equation reads

$$\begin{cases} -\nabla\cdot\boldsymbol{J}^\varepsilon(\boldsymbol{x}) = f^\varepsilon(\boldsymbol{x}) & \boldsymbol{x}\in\Omega \\ \phi^\varepsilon|_{\partial\Omega} = 0 \end{cases} \tag{7}$$

where f^ε is some given current density source bounded in $L^2(\Omega)$ which converges strongly to f in $H^{-1}(\Omega)$ when $\varepsilon \to 0$. Equation (7) is to be understood in the weak sense, *i.e.*,

$$\int_\Omega \boldsymbol{J}^\varepsilon(\boldsymbol{x})\cdot\nabla\psi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_\Omega f^\varepsilon(\boldsymbol{x})\psi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \forall\psi\in H_0^1(\Omega) \tag{8}$$

We introduce the scaled bilinear form

$$a^\varepsilon(\phi,\psi) = \int_\Omega\int_\Omega \mathbf{K}^\varepsilon(\boldsymbol{x}-\boldsymbol{\xi})\nabla\phi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}\cdot\nabla\psi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{9}$$

Equation (8) can now be restated in the following weak formulation. Find $\phi^\varepsilon \in H_0^1(\Omega)$ such that

$$a^\varepsilon(\phi^\varepsilon,\psi) = \int_\Omega f^\varepsilon(\boldsymbol{x})\psi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \forall\psi\in H_0^1(\Omega) \tag{10}$$

We will assume that the kernel $\mathbf{K}$ is such that the following boundedness and coercivity properties follows

Theorem 4. *There exist constants $C_1, C_2 > 0$ such that*

$$|a^\varepsilon(\phi,\psi)| \le C_1\|\nabla\phi\|_{L^2(\Omega;\mathbb{R}^n)}\|\nabla\psi\|_{L^2(\Omega;\mathbb{R}^n)} \tag{11}$$

$$C_2\|\nabla\phi\|^2_{L^2(\Omega;\mathbb{R}^n)} \le a^\varepsilon(\phi,\phi) \tag{12}$$

for all $\phi,\psi \in H_0^1(\Omega)$

The precise form of the kernel $\mathbf{K}$ will be given in the next sections.

4.2 Existence of unique solution

For the existence of solution we need the Lax-Milgram theorem (e.g. see Evans (1998))

Theorem 5 (Lax-Milgram)**.** *Assume that*

$$B : H\times H \to \mathbb{R}$$

is a bilinear mapping, for which there exist constants $\alpha, \beta > 0$ such that

$$|B[u,v]| \le \alpha\|u\|\|v\| \ \ (u,v\in H)$$

and

$$\beta\|u\|^2 \le |B[u,u]| \ \ (u\in H).$$

Finally, let $f : H \to \mathbb{R}$ be a bounded linear functional on H. Then there exists a unique element $u \in H$ such that

$$B[u,v] = \langle f, v \rangle$$

for all $v \in H$.

Here, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between H and its dual H'.

Theorem 6 (Existence and uniqueness)**.** *Equation* (10) *has a unique solution $\phi^\varepsilon \in H_0^1(\Omega)$ for each $\varepsilon > 0$.*

Proof: The result follows from Theorems 4 and 5. □

The main question to be answered is: Which equation with constand coefficients has a solution that is the best possible approximation of the solution of equation (10) when ε is small? To be able to answer this question we need to find the limit of the bilinear form when $\varepsilon \to 0$. The first step is to establish a priori estimates of the sequence of solutions.

4.3 A priori estimates

We have the standard a priori estimate

Theorem 7 (A priori estimate)**.** *The solutions of* (10) *satisfies*

$$\|\phi^\varepsilon\|_{H_0^1(\Omega)} \leq C \tag{13}$$

uniformly with respect to $\varepsilon > 0$.

Proof: Letting $\psi = \phi^\varepsilon$ in (10), the coercivity property in equation (12) and Hölder's inequality yields

$$C\|\nabla\phi^\varepsilon\|^2_{L^2(\Omega;\mathbb{R}^n)} \leq \|f^\varepsilon\|_{L^2(\Omega)}\|\phi^\varepsilon\|_{L^2(\Omega)} \tag{14}$$

The Poincare inequality and the boundedness of $\|f^\varepsilon\|_{L^2(\Omega)}$ gives

$$\|\nabla\phi^\varepsilon\|_{L^2(\Omega;\mathbb{R}^n)} \leq C \tag{15}$$

$$\|\phi^\varepsilon\|_{L^2(\Omega)} \leq C \tag{16}$$

The assertion is proved. □

5. Homogenization

5.1 Case I, Non-vanishing non-localness

Let us consider a non-vanishing convolution kernel. Assume that $\mathbf{K}$ is an admissible test function in the two-scale sense, as in Definition 3, *i.e.*, satisfying Proposition 2. As a model let us use

$$\mathbf{K}(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} C\boldsymbol{\sigma}(\boldsymbol{y})\exp\left(\frac{1}{\left|\frac{\boldsymbol{x}}{r}\right|^2-1}\right) & , \quad |\boldsymbol{x}| < r. \\ 0 & , \quad |\boldsymbol{x}| \geq r \end{cases} \tag{17}$$

where r is the radius of the non-local influence zone, $\boldsymbol{\sigma}$ is the conductivity associated with the non-locality, it is assumed to be Y-periodic, *i.e.*, $\boldsymbol{\sigma}(\boldsymbol{y}+\boldsymbol{e}) = \boldsymbol{\sigma}(\boldsymbol{y})$ for all $\boldsymbol{y} \in]0,1[^n$, and $C > 0$ is a constant.

The scaled kernel reads

$$\mathbf{K}^\varepsilon(\boldsymbol{x}) = \mathbf{K}\left(\boldsymbol{x}, \frac{\boldsymbol{x}}{\varepsilon}\right) = \begin{cases} C\boldsymbol{\sigma}\left(\frac{\boldsymbol{x}}{\varepsilon}\right)\exp\left(\frac{1}{\left|\frac{\boldsymbol{x}}{r}\right|^2-1}\right) & , \quad |\boldsymbol{x}| < r. \\ 0 & , \quad |\boldsymbol{x}| \geq r \end{cases} \tag{18}$$

The conductivity $\boldsymbol{\sigma}$ satisfies the coercivity condition

$$\boldsymbol{\sigma}\boldsymbol{\xi}\cdot\boldsymbol{\xi} \geq c_1|\boldsymbol{\xi}|^2 \tag{19}$$

for all $\boldsymbol{\xi} \in \mathbb{R}^n$, $\boldsymbol{x} \in \Omega$ a.e., and is bounded, *i.e.*, $\boldsymbol{\sigma} \in L^\infty(\Omega;\mathbb{R}^{n\times n})$

Theorem 8 (Homogenization, non-vanishing non-localness)**.** *Let $\{\phi^\varepsilon\}$ be a sequence of solutions to* (10) *where the kernel in the bilinear form* (9) *is given by* (18)*. The sequence $\{\phi^\varepsilon\}$ converges weakly in $H^1_0(\Omega)$ to $\phi \in H^1_0(\Omega)$, the unique solution of the Homogenized Problem*

$$-\nabla\cdot\int_{\Omega\cap\operatorname{supp}\boldsymbol{\sigma}_h(\boldsymbol{x}-\,\cdot\,)}\boldsymbol{\sigma}_h(\boldsymbol{x}-\boldsymbol{z})\,\nabla\phi(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = f(\boldsymbol{x}), \tag{20}$$

a.e. in Ω, where the homogenized conductivity is given by

$$\boldsymbol{\sigma}_h(\boldsymbol{x}) = \int_{T^n}\mathbf{K}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \int_{T^n} C\boldsymbol{\sigma}(\boldsymbol{y})\exp\left(\frac{1}{|\frac{\boldsymbol{x}}{r}|^2-1}\right)\mathrm{d}\boldsymbol{y} \tag{21}$$

Proof: Since $\phi^\varepsilon \in H^1(\mathbb{R}^n)$ and $f^\varepsilon \in L^2(\mathbb{R}^n)$ we can apply the two-scale Fourier transform to (7). The a priori estimate in Theorem 7 and Definition 5 gives

$$2\pi i(\boldsymbol{\xi}+\varepsilon^{-1}\boldsymbol{m})\cdot\widehat{\mathbf{K}}^\varepsilon_\varepsilon(\boldsymbol{\xi},\boldsymbol{m})2\pi i(\boldsymbol{\xi}+\varepsilon^{-1}\boldsymbol{m})\widehat{\phi}^\varepsilon_\varepsilon(\boldsymbol{\xi},\boldsymbol{m}) = \widehat{f}^\varepsilon_\varepsilon(\boldsymbol{\xi},\boldsymbol{m}), \tag{22}$$

for all $\varepsilon < 0$, $\boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{m} \in \mathbb{Z}^n$. Next we multiply with ε and send a subsequence (still denoted by ε) to zero. Taking Propositions 6 and 2 into account will give us the Fourier transform of the local problem as the L^2-weak limit in Fourier space,

$$2\pi i\boldsymbol{m}\cdot\widehat{\mathbf{K}}(\boldsymbol{\xi},\boldsymbol{m})2\pi i\left(\boldsymbol{\xi}\widehat{\phi}(\boldsymbol{\xi})\delta_{\boldsymbol{m}\mathbf{0}} + \boldsymbol{m}\widehat{\phi}^1(\boldsymbol{\xi},\boldsymbol{m})\right) = 0, \tag{23}$$

for a.e. $\boldsymbol{\xi} \in \mathbb{R}^n$, and all $\boldsymbol{m} \in \mathbb{R}^n$. It has a trivial solution $\widehat{\phi}^1(\boldsymbol{\xi},\boldsymbol{m}) = 0$ for all $\boldsymbol{m} \neq \mathbf{0}$. To get the homogenized problem we let $\boldsymbol{m} = \mathbf{0}$ in (22), extract another subsequence and send $\varepsilon \to 0$ which yields the standard Fourier transform of the weak $L^2(\Omega)$-limit,

$$2\pi i\boldsymbol{\xi}\cdot\widehat{\mathbf{K}}(\boldsymbol{\xi},0)2\pi i\boldsymbol{\xi}\widehat{\phi}(\boldsymbol{\xi}) = \widehat{f}(\boldsymbol{\xi}), \tag{24}$$

for a.e. $\boldsymbol{\xi} \in \mathbb{R}^n$. Apparently we do not need $\widehat{\phi}^1$ in the homogenized equation. The homogenized equation (24) is the Fourier transform of

$$-\nabla\cdot\int_{\Omega\cap\operatorname{supp}\boldsymbol{\sigma}_h(\boldsymbol{x}-\,\cdot\,)}\int_{T^n}\mathbf{K}(\boldsymbol{x}-\boldsymbol{z},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\,\nabla\phi(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = f(\boldsymbol{x}), \tag{25}$$

Here $\boldsymbol{\sigma}_h$, is the mean value of $\mathbf{K}$ with respect to the local variable. Indeed, it is the homogenized conductivity

$$\boldsymbol{\sigma}_h(\boldsymbol{x}) = \int_{T^n}\mathbf{K}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \tag{26}$$

The homogenized equation has a unique solution (see Theorem 10 below) which implies that the whole sequence converges. □

5.2 Case II, Vanishing non-localness

In this case we will use the same kernel, but we will scale both variables, *i.e.*, let

$$\mathbf{K}(\boldsymbol{y}) = \begin{cases} C\boldsymbol{\sigma}(\boldsymbol{y}) \exp\left(\frac{1}{\left|\frac{\boldsymbol{y}}{r}\right|^2 - 1}\right) & , \qquad |\boldsymbol{y}| < r. \\ 0 & , \qquad |\boldsymbol{y}| \geq r \end{cases} \tag{27}$$

where $r > 1$ is the radius of the non-local influence zone, and $C > 0$ is a constant and scale the kernel as

$$\mathbf{K}^{\varepsilon}(\boldsymbol{x}) = \varepsilon^{-n}\mathbf{K}\left(\frac{\boldsymbol{x}}{\varepsilon}\right) = \begin{cases} \varepsilon^{-n}C\boldsymbol{\sigma}\left(\frac{\boldsymbol{x}}{\varepsilon}\right) \exp\left(\frac{1}{\left|\frac{\boldsymbol{x}}{\varepsilon r}\right|^2 - 1}\right) & , \qquad |\boldsymbol{x}| < \varepsilon r. \\ 0 & , \qquad |\boldsymbol{x}| \geq \varepsilon r \end{cases} \tag{28}$$

The assumptions for Case I applies. Note that the scaling implies

$$\lim_{\varepsilon \to 0} \widehat{\mathbf{K}}^{\varepsilon}_{\varepsilon}(\boldsymbol{\xi}, \boldsymbol{m}) = \widehat{\mathbf{K}}(\boldsymbol{m}) \tag{29}$$

for all $\boldsymbol{\xi} \in \mathbb{R}^n$. The support of $\widehat{\mathbf{K}}(\boldsymbol{m})$ is continuous ($\mathbb{R}^n$) in contrast to what is indicated by the argument $\boldsymbol{m}$. The reason is that the mollifier has a compact support and larger than the unit cell $]0,1[^n$, since $r > 1$. Actually, this case asks for a modified definition of the two-scale Fourier transform, e.g. in one dimension we let $|\boldsymbol{\xi}| < \delta/2\epsilon$, and $m = \pm\delta, \pm 2\delta, \ldots$. Then we send $\delta \to 0$, but slower than ε, e.g. $\delta = \sqrt{\varepsilon}$ should work.

Theorem 9 (Homogenization, vanishing non-localness)**.** *Let* $\{\phi^{\varepsilon}\}$ *be a sequence of solutions to* (10) *where the kernel in the bilinear form* (9) *is given by* (28)*. The sequence* $\{\phi^{\varepsilon}\}$ *converges weakly in* $H^1_0(\Omega)$ *to* $\phi \in H^1_0(\Omega)$*, the unique solution of the Homogenized Problem*

$$-\nabla \cdot \boldsymbol{\sigma}_h \nabla \phi(\boldsymbol{x}) = f(\boldsymbol{x}), \tag{30}$$

a.e. in Ω*, where the homogenized conductivity is given as the mean value*

$$\boldsymbol{\sigma}_h = \int_{|\boldsymbol{y}|<r} \mathbf{K}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \int_{|\boldsymbol{y}|<r} C\boldsymbol{\sigma}(\boldsymbol{y}) \exp\left(\frac{1}{\left|\frac{\boldsymbol{y}}{r}\right|^2 - 1}\right) \mathrm{d}\boldsymbol{y} \tag{31}$$

Proof: Since ϕ^{ε} is bounded in $\in H^1(\mathbb{R}^n)$ and $f^{\varepsilon} \in L^2(\mathbb{R}^n)$ we can apply the (modified) two-scale Fourier transform in Definition 5 to (10),

$$2\pi i(\boldsymbol{\xi} + \varepsilon^{-1}\boldsymbol{m}) \cdot \widehat{\mathbf{K}}^{\varepsilon}_{\varepsilon}(\boldsymbol{\xi}, \boldsymbol{m}) 2\pi i(\boldsymbol{\xi} + \varepsilon^{-1}\boldsymbol{m}) \widehat{\phi}^{\varepsilon}_{\varepsilon}(\boldsymbol{\xi}, \boldsymbol{m}) = \widehat{f}^{\varepsilon}_{\varepsilon}(\boldsymbol{\xi}, \boldsymbol{m}), \tag{32}$$

for all $\boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{m} \in \mathbb{Z}^n$. Next we multiply with ε and send a subsequence (still denoted by ε) to zero. Taking Proposition 6 and limit (29) into account will give us the Fourier transform of the local problem as the L^2-weak limit in Fourier space,

$$2\pi i\boldsymbol{m} \cdot \widehat{\mathbf{K}}(\boldsymbol{m}) 2\pi i \left(\boldsymbol{\xi}\widehat{\phi}(\boldsymbol{\xi})\delta_{\boldsymbol{m}\mathbf{0}} + \boldsymbol{m}\widehat{\phi}^1(\boldsymbol{\xi}, \boldsymbol{m})\right) = 0, \tag{33}$$

for a.e. $\boldsymbol{\xi} \in \mathbb{R}^n$, and all $\boldsymbol{m} \in \mathbb{R}^n$. It has a trivial solution $\widehat{\phi}^1(\boldsymbol{\xi}, \boldsymbol{m}) = 0$ for all $\boldsymbol{m} \neq \mathbf{0}$. To get the homogenized problem we let $\boldsymbol{m} = \mathbf{0}$ in (32) and send another subsequence $\varepsilon \to 0$ which yields the usual Fourier transform of the weak $L^2(\Omega)$-limit in (30), once again using Proposition 6,

$$2\pi i\boldsymbol{\xi} \cdot \widehat{\mathbf{K}}(\mathbf{0}) 2\pi i\boldsymbol{\xi}\widehat{\phi}(\boldsymbol{\xi}) = \widehat{f}(\boldsymbol{\xi}), \tag{34}$$

for a.e. $\boldsymbol{\xi} \in \mathbb{R}^n$. Applying (29) gives the homogenized equation. The homogenized equation reads in real space

$$-\nabla \cdot \int_{|\boldsymbol{y}|<r} \mathbf{K}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\,\nabla\phi(\boldsymbol{x}) = f(\boldsymbol{x}), \tag{35}$$

Inspection of equation (35) yields the homogenized conductivity as

$$\boldsymbol{\sigma}_h = \int_{|\boldsymbol{y}|<r} \mathbf{K}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}, \tag{36}$$

The whole sequence converges since the homogenized equation has a unique solution in $H_0^1(\Omega)$, see Theorem 10. □

Theorem 10 (Existence of solution)**.** *The homogenized problems* (20) *and* (30) *has each a unique solution in* $H_0^1(\Omega)$*.*

Proof: It follows from the assumptions that the homogenized conductivity $\boldsymbol{\sigma}_h$ inherits the properties of **K**. The statement follows from Theorem 5. □

6. Remarks and conclusions

The localization of the constitutive relation for Case II in (30) can equally be obtained by multiplying the kernel in (17) with r^{-1} and sending $r \to 0$ either before sending $\varepsilon \to 0$ or after. Introducing a spatially local contribution in the constitutive relations will somewhat complicate the analysis, but it is doable. An effect that we have not taken into account is the influence of the boundary $\partial\Omega$. In real life, e.g. for wave propagation in cases the wavelength is on the same order as the material periodicity, we expect the nonlocal constitutive relation to depend on the distance to the boundary. We will return to these issues in forthcoming papers. We conclude that spatially nonlocal constitutive relations are particularly easy to homogenize since we need only to integrate the kernel over the fast variable. In retrospective, this is to some degree expected since spatial convolution is an averaging procedure which smoothers fast oscillations.

7. References

Allaire, G. (1992). Homogenization and two-scale convergence, *SIAM J. Math. Anal.* 23(6): 1482–1518.

Allaire, G. & Briane, M. (1996). Multiscale convergence and reiterated homogenisation, *Proc. Roy. Soc. Edinburgh Sect. A* 126(2): 297–342.

Allaire, G. & Conca, C. (1996). Bloch-wave homogenization for a spectral problem in fluid-solid structures, *Arch. Rational Mech. Anal.* 135(3): 197–257.
URL: *http://dx.doi.org/10.1007/BF02198140*

Arbogast, T., Douglas, Jr., J. & Hornung, U. (1990). Derivation of the double porosity model of single phase flow via homogenization theory, *SIAM J. Math. Anal.* 21(4): 823–836.
URL: *http://dx.doi.org/10.1137/0521046*

Bensoussan, A., Lions, J. L. & Papanicolaou, G. (1978). *Asymptotic Analysis for Periodic Structures*, Vol. 5 of *Studies in Mathematics and its Applications*, North-Holland, Amsterdam.

Bloch, F. (1928). Über die Quantenmechanik der Electronen in Kristallgittern, *Z. Phys.* 52: 555–600.

Brouder, C. & Rossano, S. (2002). Microscopic calculation of the constitutive relations, ArXiv Physics e-prints.

Cioranescu, D., Damlamian, A. & Griso, G. (2002). Periodic unfolding and homogenization, *C. R. Math. Acad. Sci. Paris* 335: 99–104.

Conca, C., Orive, R. & Vanninathan, M. (2002). Bloch approximation in homogenization and applications, *SIAM J. Math. Anal.* 33(5): 1166–1198 (electronic).
URL: *http://dx.doi.org/10.1137/S0036141001382200*

Conca, C. & Vanninathan, M. (1997). Homogenization of periodic structures via Bloch decomposition, *SIAM J. Appl. Math.* 57(6): 1639–1659.
URL: *http://dx.doi.org/10.1137/S0036139995294743*

Conca, C. & Vanninathan, M. (2002). Fourier approach to homogenization problems, *ESAIM Control Optim. Calc. Var.* 8: 489–511 (electronic). A tribute to J. L. Lions.

Dal Maso, G. (1993). *An introduction to Γ-convergence*, Progress in Nonlinear Differential Equations and their Applications, 8, Birkhäuser Boston Inc., Boston, MA.

De Giorgi, E. (1975). Sulla convergenza di alcune successioni d'integrali del tipo dell'area, *Rend. Mat. (6)* 8: 277–294. Collection of articles dedicated to Mauro Picone on the occasion of his ninetieth birthday.

De Giorgi, E. & Franzoni, T. (1975). Su un tipo di convergenza variazionale, *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8)* 58(6): 842–850.

De Giorgi, E. & Spagnolo, S. (1973). Sulla convergenza degli integrali dell'energia per operatori ellittici del secondo ordine, *Boll. Un. Mat. Ital. (4)* 8: 391–411.

Evans, L. C. (1998). *Partial Differential Equations*, American Mathematical Society, Providence, Rhode Island.

Faraday, M. (1965). *Experimental researches in electricity : in three volumes bound as two*, Dover Publications, New York.

Floquet, G. (1883). Sur les équations différentielles linéaries à coefficients périodique, *Ann. École Norm. Sup.* 12: 47–88.

Griso, G. (2002). Estimation d'erreur et éclatement en homogénéisation périodique, *C. R. Math. Acad. Sci. Paris* 335(4): 333–336.
URL: *http://dx.doi.org/10.1016/S1631-073X(02)02477-9*

Laptev, V. (2005). Two-scale extensions for non-periodic coefficients, arXiv:math.AP/0512123.

Maxwell, J. C. (1954a). *A Treatise on Electricity and Magnetism*, Vol. 1, Dover Publications, New York.

Maxwell, J. C. (1954b). *A Treatise on Electricity and Magnetism*, Vol. 2, Dover Publications, New York.

Milton, G. W. (2002). *The Theory of Composites*, Cambridge University Press, Cambridge, U.K.

Nechvátal, L. (2004). Alternative approaches to the two-scale convergence, *Appl. Math.* 49(2): 97–110.

Nguetseng, G. (1989). A general convergence result for a functional related to the theory of homogenization, *SIAM J. Math. Anal.* 20(3): 608–623.

Rayleigh, L. (1892). On the influence of obstacles arranged in rectangular order upon the properties of the medium, *Philosophical Magazine* 34: 481–502.

Spagnolo, S. (1967). Sul limite delle soluzioni di problemi di Cauchy relativi all'equazione del calore, *Ann. Sc. Norm. Sup. Pisa* 21: 657–699.

Tartar, L. (1977). Cours peccot au Collège de France. Unpublished.

Taylor, M. E. (1996). *Partial Differential Equations I: Basic Theory*, Springer-Verlag, New York.

Wellander, N. (2004). Review of contemporary homogenization methods, *2004 URSI EMTS International Symposium on Electromagnetic Theory*, Pisa, Italy.

Wellander, N. (2007). Periodic homogenization in Fourier space, ICIAM07, 6th International Congress on Industrial and Applied Mathematics.

Wellander, N. (2009). The two-scale Fourier transform approach to homogenization; periodic homogenization in Fourier space, *Asymptot. Anal.* 62(1-2): 1–40.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Time-Resolved Fourier Transform Infrared Emission Spectroscopy: Application to Pulsed Discharges and Laser Ablation

Svatopluk Civiš[1]*and Vladislav Chernov[2]
[1]*J. Heyrovský Institute of Physical Chemistry*
[2]*Voronezh State University*
[1]*Czech Republic*
[2]*Russia*

1. Introduction

1.1 Time-resolved Fourier transform infrared spectroscopy

Time-resolved spectroscopy (TRS) is a wide-spectrum technique used for studying the dynamics of chemical reactions, or the dynamic properties of molecules, radicals and ions in liquid, gas and solid states. In the infrared spectral range it can be achieved by using lasers (Smith & Palmer, 2002), grating spectrometers (Rödig & Siebert, 2002) or by interferometers (Masutani, 2002). The presented report is focused on the development and application of a time resolved system based on commercially available continuously scanning high resolution interferometer and its modification for time resolved Fourier transform spectroscopy (TR-FTS) (Kawaguchi et al., 2003).

The main advantage of TR-FTS lies in obtaining spectra in wide wavenumber intervals. The speed of data acquisition is limited by the duration of the acquisition process and by the band width of the used detector.

There are basically two ways of obtaining the time-resolved spectra: the continuous scan and the non-continuous, step scan (Masutani, 2002; Rödig & Siebert, 2002; Smith & Palmer, 2002). The continuous scan is best used when the duration of the observed phenomenon is longer than the time needed for carrying out one scan, *i. e.* for obtaining an interferogram up to the maximum trajectory difference (Rapid and ultrarapid scanning FT). Time-shifted individual scans provide a sequence of interferograms from which a conventional spectrum can be calculated. When using the rapid scanning and short distance mirror traversing, a time resolution from 1000 s to 1 ms can be reached.

A special approach to the time-resolved spectra of phenomena lasting from milliseconds to microseconds is the synchronous scanning FT technique (Kawaguchi et al., 2005). This method, as well as the methods mentioned below, requires the possibility of initiating the reaction in a pulse mode, *e. g.* using a laser, electric discharge, electron bombardment, a UV discharge lamp, *etc*. (Civiš et al., 2006). The apparatus carries out a continuous scan and, during the pulse, it reads the signal from the detector corresponding to the position

*Corresponding author. FAX: +420286591766; e-mail: civis@jh-inst.cas.cz of the mirror and to the time from the beginning of the pulse reaction using the He–Ne laser fringe signals generated by the interferometer. This method is called stroboscopic interferometry (Smith & Palmer, 2002). After accumulating a sufficient amount of data and scans, the time-shifted interferograms are composed. The time mode is usually from 10 ms to 1 μs.

A more favorable method of non-continuous scanning in steps (step-scan) is achieved in discrete jumps and the time-resolved data from each position can be recorded after each transient event (Rödig & Siebert, 2002). Such a system is easy to couple with a pulse laser or a pulsed discharge. The step-scan spectrometers are commercially available and are used mainly for photolytic experiments in biology. The resolution of commercial step-scan type interferometers is limited to 0.1 cm^{-1}. A high resolution measurement with a step-scan type interferometer has been reported: a Connes type interferometer (CNRS Orsay) was used for the measurement of N_2 spectra with a resolution of 0.03 cm^{-1} (Durry & Guelachvili, 1994).

1.2 Continuous scan systems: Synchronous triggering and data sampling

Continuously scanning spectrometers have been applied for time-resolved spectroscopy by several teams following the first report by Mantz (1976). Berg & Sloan (1993) developed a compact data acquisition system for submicrosecond time-resolved FTS. Nakanaga et al. (1993) applied a pulse discharge system to a continuously scanning interferometer without any modification of the system's software. The pulsed discharge was triggered by a He–Ne laser fringe signal with an appropriate delay time. The system was applied to the measurement of the time profiles of a vibration-rotation absorption spectrum of discharged CO. Recently, Kawaguchi et al. (2005) reviewed the methods of time-resolved Fourier transform infrared spectroscopy and its application to pulsed discharges and demonstrated the technique of FTS using a high-resolution Bruker IFS 120 HR supported by a microcontroller SX or Field Programmable Gate Array processor (FPGA) on He_2, ArH and ArH^+ spectra. The same system was used for studding the products of ArF excimer laser ablation products (Civiš et al., 2010; Civiš et al., 2010; Kawaguchi et al., 2008).

The continuous scanning principle was the basis for data acquisition by a modified (Bruker IFS 120) spectrometer in our laboratory at the J. Heyrovský Institute of Physical Chemistry, and a similarly modified spectrometer was used in Okayama (Japan).

The data acquisition system can be described as follows:

The position of the traversing mirror of the Michelson interferometer is detected by reading the interference maxima of the He–Ne laser emission. The input signal in a cosine function shape is digitally processed into rectangular pulses and becomes the internal standard of the interferometer. The frequency of these rectangular pulses depends on the mirror speed. In the classic measurement mode, the frequency is usually 10 kHz with a pulse duration of 100 μs. An external processor monitors the beginning of the He–Ne laser digital pulse, its order and the zero position of the mirror. During one pulse, the signal from the detector is read (30 or up to 64 readings), this is the so-called AD trigger Kawaguchi et al. (2008). These signals are shifted in time by Δt, where $\Delta t = 1$ or $2, 3 \ldots$ μs.

In this way, a matrix $I(t_k, \delta_i)$ of intensity I in times t_k is acquired for the given optical path difference δ_i (i being the index of the selected optical path difference, from its zero to maximum values). A discharge pulse of variable length can be arbitrarily inserted into the data acquisition process (AD trigger). This process results in 30 to 64 reciprocally time-shifted interferograms.

Time resolved spectra are obtained by collecting data at various points between the zero-crossings and calculating the FT transformation for each such point. This system was utilized using a FPGA processor. The main role of the FPGA processor in our experiment was to create a discharge or laser pulse and AD trigger signals (the signal for data collection from the detector) synchronously with the He–Ne laser fringe signals from the spectrometer (see Figure 1 and Figure 2). The FPGA processor also controls the data transmission from the digital input board to the PC.

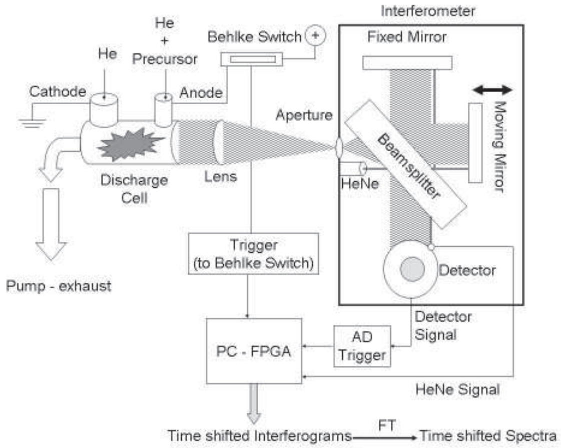


Fig. 1. A diagram of the time resolved Fourier transform spectrometer with FPGA microcontroller

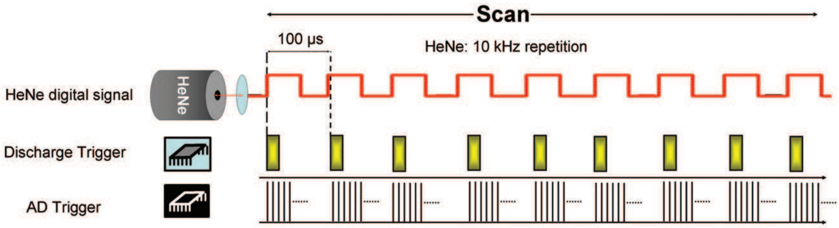


Fig. 2. Timing diagram for time resolved FT measurement. The scan and He–Ne fringe signals are supplied from Bruker 120 HR spectrometer. The velocity of the scanner is 10 kHz 100 μs time intervals are produced. The discharge trigger is programmed using FPGA microcontroller. Maximum 64 interferograms (64 time shifted spectra) can be obtained during one scan.

2. Continuous scan systems: application with discharges

Figure 1 depicts the experimental arrangement used in presented study. Infrared emission was observed from a pulsed discharge of FT time-resolved measurements. The parent compound hydrogen or $(CN)_2$ was entrained in an inert carrier gas (He, Ar) and entered in the 20 cm long positive column discharge (or hollow cathode) tube with an inner diameter of

12 mm. The pulsed discharge was induced by a high voltage transistor switch HTS 81 (Behlke electronic GmbH, Frankfurt, Germany) between the stainless steel anode and grounded cathode. The plasma produced from the reaction mixture was cooled by flowing water in the outer jacket of the cell. The best conditions for the generation of radicals or ions were found to be $p(\text{He}, Ar) = 2$–10 Torr and 50 mTorr of parent molecules. The voltage drop across the discharge was 1000 V, with a pulse width of 20 or 40 μs and 0.5 A peak-to-peak current. The scanner velocity of the FT spectrometer was set to produce a 10 kHz He–Ne laser fringe frequency which was used to trigger the pulsed discharge. The recorded spectral range was 1800–4000 cm^{-1} with an optical filter, and a unapodized resolution of 0.07 or 0.025 cm^{-1}. The 32–100 scans were coadded so as to obtain a reasonable signal-to-noise ratio. The observed wavenumbers were calibrated using CO ground state rotation-vibration lines presenting in the spectra (Guelachvili & Rao, 1986) as impurities.

2.1 He discharge plasma

2.1.1 Introduction

The He$_2$ molecule is known as the first Rydberg molecule, since its spectrum was reported in 1913. Many spectroscopic studies have been carried out as compiled in a book of Huber & Herzberg (1979) and in the DiRef web site (Bernath & McLeod, 2001). Most of the spectra of He$_2$ molecule have been observed in the visible and ultraviolet regions. Ginter & Ginter (1988); Ginter et al. (1984) compiled and analyzed the energy levels of Rydberg states originating from the electronic configurations $(1\sigma_g)^2(1\sigma_u)np\lambda(^3\Pi_g, {}^3\Sigma_g^+)$ and $(1\sigma_g)^2(1\sigma_u)ns\sigma, nd\lambda(^3\Sigma_u^+, {}^3\Sigma_u^+, {}^3\Pi_u, {}^3\Pi_u)$ by multichannel quantum defect theory, where n is the principal quantum number in the united atom molecular orbital designation. According to the energy levels listed in Refs. (Ginter & Ginter, 1988; Ginter et al., 1984), many electronic transitions are expected in the infrared region. However, observations of the infrared spectra so far have been limited to the three band systems below 8000 cm^{-1}:

(1) $b^3\Pi_g$–$a^3\Sigma_u^+$ with the 0–0 band origin at 4750cm^{-1}, studied by Hepner (1956), Gloersen & Dieke (1965), and (Rogers et al., 1988),

(2) $B^1\Pi_g$–$A^1\Sigma_u^+$ with the 0–0 band origin at 3501cm^{-1}, studied by Solka et al. (1987) and

(3) the $4f$–$3d$ band in 5100–5800 cm^{-1} spectral region, studied by Herzberg & Jungen (1986).

The assignment of $4f$–$3d$ band was the first example concerning electronic states originating from the f-orbital electron.

The time-resolved Fourier transform spectroscopic system was applied for the observations of He$_2$ emission spectra produced by a pulsed discharge (Hosaki et al., 2004). This method has enabled us to observe many electronic transitions in the infrared region, including the previously reported bands. The spectroscopic analysis of newly observed three bands and their time profiles are briefly reported.

2.1.2 Experimental

The spectra of He$_2$ were observed in emission from a hollow cathode discharge plasma. The hollow cathode stainless steel tube was 20 cm long with an inner diameter equal to 12 mm. The ac discharge was maintained by a high voltage transistor switch applied between the stainless steel anode and grounded cathode. The emission of He$_2$ has been also observed from a positive column, where lines from vibrationally excited states of $b^3\Pi_g$ were found to be more intense, compared with the case of the hollow cathode discharge. Here we report only the spectra obtained from the hollow cathode discharge, because of its higher efficiency in the production of the highly excited electronic states of He_2.

The plasma made from a pure helium was cooled down by flowing water or by liquid nitrogen in outer jacket of the cell. The best conditions for the generation of the He_2 were found to be $p(He) = 1.33$ kPa (10 Torr).

2.1.3 Observed spectra and analysis

Figure 3 shows a part of the observed time-resolved emission spectrum from a discharge in He. The discharge was initiated at time zero and turned off at 20 μs. For AD-converter triggers, we used 3 μsec for the zero offset and interval values, that is, AD conversion occurs every 3 μsec from the start of the discharge and all together 30 pulses cover 90 μsec. The strong line (5880 cm^{-1}) in Figure 3 belongs to the He atomic line (4d–3p) and is observed as two intense peaks. It may be noted that the second peak appears after the discharge is off, that is, it is due to the afterglow plasma. The other spectral lines in Figure 3 pertain to the 4f–3d transitions of He_2 which have been analyzed by Herzberg & Jungen (1986).

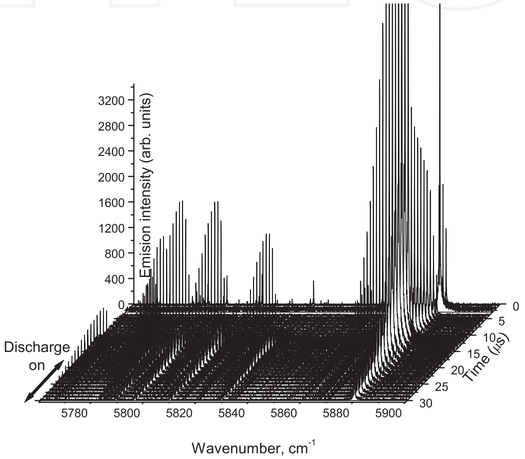


Fig. 3. A portion of the time-resolved spectrum observed by a pulse discharge in He with a pressure of 1.33 kPa (10 Torr). The discharge was applied in the interval of 0–20 μsec with a peak current of 0.5 A. The strongest peak belongs to atomic He line (4d–3p). Other lines pertain to 4d–3p transitions of He_2.

Figure 4 shows an observed spectrum in the 2750–5600 cm^{-1} region, where we averaged all 30 spectra obtained by the time-resolved method. In the figure, the $b^3\Pi_g$–$a^3\Sigma_u^+$ $\nu = 0$–0 band is strongly observed in 4800 cm^{-1} region. Most of spectral lines in the 5200–5900 cm^{-1} region could be attributed to the 4f–3d band (Solka et al., 1987). In the 3300 cm^{-1} region, the $B^1\Pi_g$–$A^1\Sigma_u^+$ $\nu = 0$–0 band found to be weak. From the time profile, it appears that the population in the singlet $B^1\Pi_g$ state decreased during the discharge period and increased in the afterglow, similarly to that observed for high-energy triplet states.

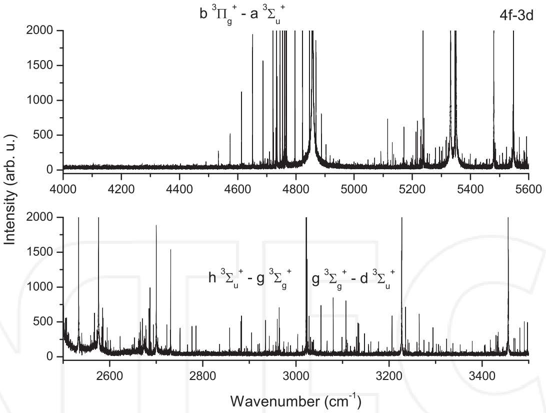


Fig. 4. An observed spectrum of He_2 in the 2750–5600 cm^{-1}, where 30 time-resolved spectra (90 μsec) are averaged. The discharge condition is given in the caption of Figure 3

In addition to these already reported bands, some new bands were observed. In the 3200 cm^{-1} region, two series of lines have been observed with no Q-branch transitions. Rotational assignments are listed in Table 1 with the observed wavenumbers.

N	$P(N)$	o.-c.	$R(N)$	o.-c.
$h^3\Sigma_u^+ – g^3\Sigma_g^+$				
0	3177.5569	0.0006		
2	3134.9969	-0.0005	3206.4133	0.0018
4	3107.2556	-0.0001	3235.5447	0.0000
6	3080.1473	-0.0010	3264.8626	-0.0064
8	3053.7916	0.0000	3294.3028	0.0089
10	3028.3000	0.0012	3323.7108	-0.0054
12	3003.7698	-0.0002	3353.0067	0.0013
14	2980.2783	-0.0002		
$g^3\Sigma_g^+ – d^3\Sigma_u^+$				
1	3190.4064	-0.0019	3232.9664	-0.0013
3	3160.7770	-0.0003	3259.9345	0.0005
5	3130.2548	0.0017	3285.6522	0.0016
7	3098.93068	0.0021	3130.0069	0.0006
9	3066.8917	-0.0002	3332.8839	-0.0017
11	3034.2243	-0.0031	3354.1697	0.0005
13	3001.0169	0.0015		

Table 1. Observed transitions of He_2 (cm^{-1}). N denotes the rotational quantum number neglecting spin in lower electronic states.

The analysis using the standard energy level expressions gave the rotational and centrifugal distortion constants, and the band origin (term energy) as listed in Table 2.

	Present	Previous
$(h^3\Sigma_u^+)$		(a)
B	7.14853(24)	7.149
$D \times 10^3$	0.5053(24)	0.574
$H \times 10^7$	−0.686(87)	
E		
$(g^3\Sigma_g^+)$		(b)
B	7.096458(94)	7.0968(1)
$D \times 10^3$	0.53071(44)	0.538(7)
E	3204.8746(11)	
$(d^3\Sigma_u^+)$		(c)
B	7.226364(88)	7.2286(15)[a]
$D \times 10^3$	0.51991(37)	0.532(3)
E	0.0	0.0

(a) Huber & Herzberg (1979)
(b) Orth & Ginter (1976)
(c) Ginter (1965)

Table 2. Molecular constants of He$_2$ in the $h^3\Sigma_u^+$, $g^3\Sigma_g^+$ and $d^3\Sigma_u^+$ states (cm^{-1} units). Numbers in parentheses denote one standard deviation and applied to the last significant digits.

Spin splitting was not observed in these bands. The magnitude of the rotational constants was useful for identification of the electronic state. The band origin frequencies 3204.9 cm^{-1} and 3194 cm^{-1} of the two bands were consistent with those of the $g^3\Sigma_g^+$–$d^3\Sigma_u^+$ ($v = 0$–0) and $h^3\Sigma_u^+$–$g^3\Sigma_g^+$ ($v = 0$–0) transitions, respectively. Both the bands were identified for the first time in the infrared region, although electronic states involved have been observed by other electronic transitions in the visible region. Molecular constants determined in the previous studies are also listed in Table 2 for comparison.

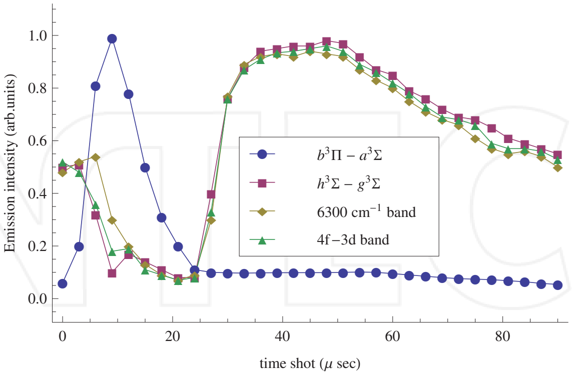


Fig. 5. Observed time profiles of emission intensities of He$_2$

Time profiles of observed spectral lines are depicted in Figure 5 for several bands. Except for the transitions from the $b^3\Pi$ state, the high-energy Rydberg states are produced strongly in the afterglow plasma. This means those states are more fragile during the discharge period and may not be observed strongly in a normal DC discharge.

The pulsed discharge and multi-sampling system produce an interesting spectral feature of He_2 in the infrared region. Especially, when the data sampling is carried out after turning off the discharge, intense emissions from many electronic bands are strongly observed. The analysis may provide information about high-energy Rydberg states including states originating from f-orbital electrons.

Figure 6 shows the energy level diagram of He_2, where the energy values are represented relative to the $a^3\Sigma_u^+(v = 0)$ state, which is located 144952 cm^{-1} (18 eV) above the repulsive ground $X^1\Sigma_g^+$) state. The observed transitions are demonstrated by arrows.

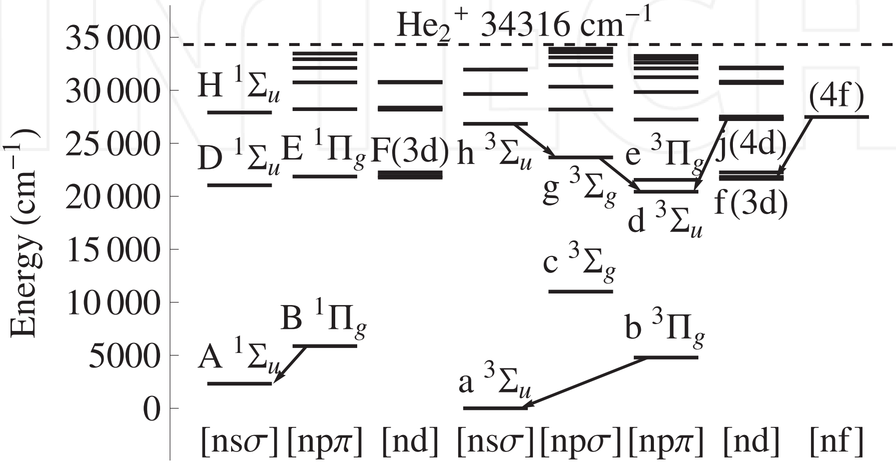


Fig. 6. Energy level diagram of He_2. The transitions observed in the present study are shown by arrows. The energy value is measured from the $a^3\Sigma_u^+(v = 0)$ state; $n(> 1)$ is the principal quantum number in a united atom molecular orbital designation. The ionization limit to $He_2{}^+$ is 34316 cm^{-1}.

2.2 Hydrogen containing discharge

Hydrogen and helium are the two most abundant elements in the universe. The hydrogen molecule and its various hydrides are the first source of aggregation and formation of interstellar matter. This process occurs in dense interstellar clouds, in star-forming regions and also in the atmosphere of some heavy planets (*e. g.*, Jupiter, Saturn,Uranus). The $H_3{}^+$ ion plays a dominant role in all these cases (Drossart et al., 1989; Herbst & Klemperer, 1973; McCall et al., 1998). After the first laboratory spectroscopic detection of $H_3{}^+$ (Oka, 1980), a large number of laboratory studies have been published, of which about 20 were concerned with measuring new infrared spectra, describing about 800 transitions from a variety of vibrational bands in the spectral range between 1800 and 9000 cm^{-1}. It is apparent from the comprehensive evaluation and compilation study of $H_3{}^+$ spectroscopy by Lindsay & McCall (2001) that most laboratory studies were carried out using absorption measurements. The only exception consists of the pioneering experiments of Majewski et al. (1987; 1994), who used a combination of a water-cooled, high-pressure, high-current emission hollow cathode together with an FT spectrometer. Majewski *et al.* used a high pressure of hydrogen gas (5–50 Torr) and high discharge current (up to 2.5 A) for production of H_3^+. They obtained a very dense spectrum containing, in addition to H_3^+, a large number of H atomic, H_2 valence, H_2 Rydberg, H_3 neutral, and also other unidentified transitions.

In this part we describe the application to the observation of H_3^+, He and H emission produced by a pulsed discharge in a He/H_2 mixture in the infrared spectral range. The use of time-resolved FT spectroscopy opens new pathways and new points of view in the study of the formation and decay processes inside the discharge plasma and permits description of the dynamics of the formation and decay of excited states of the H_3^+ ion (Civiš et al., 2006).

2.2.1 Experimental

The emission spectra from a hollow cathode discharge plasma in He/H_2 mixtures were observed with the time-resolved Fourier transform high-resolution Bruker IFS 120 HR interferometer (Civiš et al., 2006). The hollow cathode stainless steel tube, covered with an outer glass jacket, was 25 cm long with an inner diameter of 12 mm. The He/H_2 plasma was cooled by liquid nitrogen in the outer jacket of the cell. The voltage drop across the discharge was 800 V, with a pulse width of 20 μs and 0.6 A peak-to-peak current The recorded spectral range was 1800–4000 cm^{-1} with an optical filter, at an unapodized resolution of 0.1 cm^{-1}. Sixty-four scans were averaged to obtain a reasonable signal-to-noise ratio. The initial pressure of H_2 was 0.35 Torr and the He pressure was changed from 2 to 10.8 Torr.

The experiments were carried out with pulsed discharge with a width of 20 μs. Because of the high pressures (up to 10 Torr) required for generation of H_3^+ and thus, subsequently, the short relaxation times of the H_3^+ ions, the measurement was carried out with maximum time resolution of 1 μs, which is currently limited by the response time of a preamplifier of the InSb detector. Compared to previous measurements (Hosaki et al., 2004; Kawaguchi et al., 2003), the data acquisition system was modified for enabling recording of 64 time-shifted spectra in a single scan.

2.2.2 Results and discussion

Helium has a higher ionization potential than H_2 (24.6 eV and 15.4 eV correspondingly; see Huntress, 1977), so that the electron temperatures are higher when He predominates in the discharge. Because of the low proton affinity of He (1.9 eV) compared to H_2 (4.5 eV), the He buffer is chemically quite inert.

The low temperature emission spectra from a hollow cathode discharge in a He/H_2 mixture were found to contain only several of the low J and K transitions of H_3^+, together with the atomic lines of He and H. No further lines of He^+, H^+ or molecular lines of H_2 or of neutral H_3 which also absorb in this area, were found (Davies et al., 1990; Vervloet & Watson, 2003). Figure 7 shows part of the observed time-resolved emission spectra of He (2129.83 cm^{-1}) and H (2148.79 cm^{-1}). Figure 7(b) depicts the H_3^+ line $Q(1,0)$ at 2529.724 cm^{-1}, which belongs to $\nu_2 = 1 \to 0$ band.

Figure 8 shows time profiles of emission lines of He (a,b), H_3^+ (c) and H (d). The absolute energy levels and Einstein coefficients A_{ij} for He, H and H_3 were taken from the NIST database (Ralchenko et al., 2008) and from (Kao et al., 1991). The vertical axis shows observed intensity divided by the Einstein coefficients A_{ij} and corresponds to the abundance in the upper state of the transition. It is noted that the emission from $n = 5$ of H was observed through the $n = 5$–4 transition, but other transitions from $n = 5$ are not observed because of the limited frequency range. Therefore, the abundance in the initial state of atomic transitions should be multiplied by a factor of 6 in the case of hydrogen. This sequence of processes was

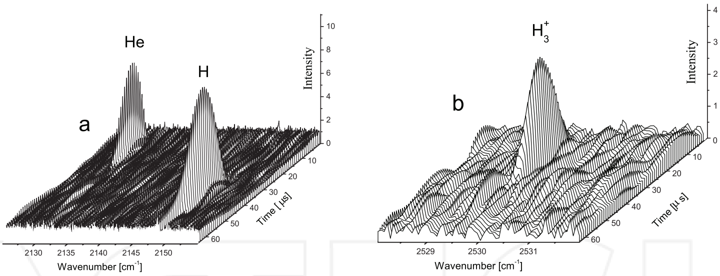


Fig. 7. Time-resolved infrared spectrum at 2125–2155 cm^{-1} (a) and 2528–2532 cm^{-1} (b), observed in a pulsed discharge in a He (10.8 Torr)/H_2 (0.4 Torr) mixture. The discharge was applied in a time interval of 0–20 μs with a peak current of 0.5 A.

observed for all the lines in the spectrum; the time in which the maximum emissions were measured for the individual transitions of the same species (He, $H_3{}^+$ and H) did not differ by more than ± 2 μs.

Chemical processes in He/H_2 plasma that can lead to the formation of $H_3{}^+$ are discussed extensively in literature (Eqns. 1–4; see Plašil et al. (2002)):

$$\mathrm{He} + \mathrm{e}_{\mathrm{fast}} \longrightarrow \mathrm{He}^* + \mathrm{e}_{\mathrm{slow}} \qquad \sigma = 4 \times 10^{-18}\ \mathrm{cm}^2 \quad \text{for He}(2^3S) \tag{1}$$

$$\mathrm{He}^* + \mathrm{H}_2 \longrightarrow \mathrm{H}_2{}^+ + \mathrm{e} \qquad \sigma = 2.6 \times 10^{-16}\ \mathrm{cm}^2 \quad \text{for He}(2^3S) \tag{2}$$

$$\mathrm{H}_2 + \mathrm{e}_{\mathrm{fast}} \longrightarrow \mathrm{H}_2{}^+ + \mathrm{e}_{\mathrm{slow}} + \mathrm{e}_{\mathrm{slow}} \qquad \sigma = 10^{-16}\ \mathrm{cm}^2 \quad \text{for electron energy of 70 eV} \tag{3}$$

$$\mathrm{H}_2{}^+ + \mathrm{H}_2 \longrightarrow \mathrm{H}_3{}^+ \ + \mathrm{H} \qquad k = 2.0 \times 10^{-9}\ \mathrm{cm}^3\,\mathrm{s}^{-1}, \quad \Delta H = -1.65\ \mathrm{eV} \tag{4}$$

In all our experiments no emission from $H_3{}^+$ has been detected in pure hydrogen discharge. Reaction (4) is exothermic to have enough energy to produce $H_3{}^+$ in $\nu_2 = 1$ state. However, since H_2 is known to be efficient for relaxing the vibrational excited state, the vibrationally excited $H_3{}^+$ in pure H_2 discharge will be relaxed by the collision with H_2. It should be noted that the obtained radiative lifetime of $H_3{}^+$ is about 10 ms.

By adding a large amount of He the vibrational relaxation will be suppressed, and the $H_3{}^+$ emission becomes strong. At low helium pressures, $H_3{}^+$ is formed directly in the discharge through direct ionization of H_2 and subsequent processes (Eqns. 1,4). Formation and decay of excited $H_3{}^+$ occurs in μs time range. The calculated values for the deactivation rate $k_{ij} = (2..5) \times 10^5\ \mathrm{s}^{-1}$ are several orders of magnitude higher than the Einstein coefficients A_{ij} for spontaneous emission of $H_3{}^+$, and recombination process described later, which clearly demonstrates the efficiency of collision processes for deactivation of the $H_3{}^+$ excited states. In Figure 8, the second weaker maximum of the $H_3{}^+$ line was found at 45 ± 2 μs. This appears as a consequence of collisions of H_2 and metastable He with a long lived lifetime, as shown in Eqns. (2) and (4).

The entire process in the afterglow is concluded by formation of H, which could be observed on two lines (2148.7 and 2467.8 cm^{-1}) at times of 40–60 μs (Figure 8d). The two lines corresponding to transitions in excited atomic hydrogen were observed with two peaks (Figure 8d): a small peak at 12 μs and a high-intensity peak with a maximum at 39 ± 1 μs. The former peak may be explained by direct dissociation of H_2 through collision with electron. We

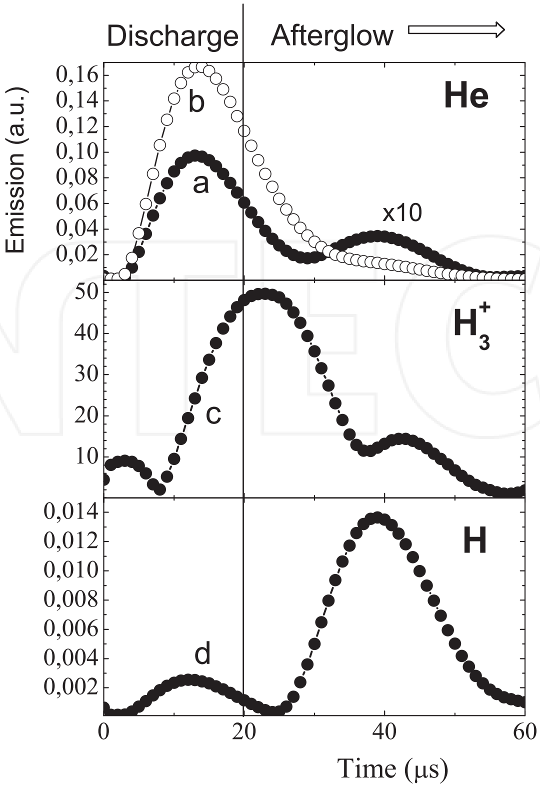


Fig. 8. Time-resolved emission profiles: (a) excited He $5f^1F^o$–$4d^1D$ transition at 2474.64 cm^{-1} (magnified by ten), (b) He $3p^3P^o$–$3s^3S$ transition at 2327.77 cm^{-1} (strongest line in the FTIR spectrum), (c) $H_3{}^+$ transition $Q(1,0)$ at 2529.72 cm^{-1} and (d) H Brackett α ($n = 5$–4) at 2467.60 cm^{-1}. The absolute intensity of emission was corrected by the corresponding Einstein coefficients A_{ij}.

expect that the second peak is due to hydrogen atom produced by recombination of $H_3{}^+$ with electron.

When the discharge is turned off, the electron energy becomes lower through collision, and the recombination rate will be increased. The recombination products of the reaction of $H_3{}^+$ with an electron are either three neutral hydrogen atoms or a hydrogen molecule together with a hydrogen atom. For the most important case of near-zero initial energy, *i. e.*, low-energy electrons and small internal excitation of the $H_3{}^+$ ion, the two reaction channels can be expected (Eqns. 5,6) (Motret et al., 1985; Strasser et al., 2002).

$$H_3{}^+ + e^- \longrightarrow H_2(X^1\Sigma_g^+, \nu J) + H(1s), \quad \Delta E = 9.2 \text{ eV} \tag{5}$$

$$\longrightarrow H(1s) + H(1s) + H(1s), \quad \Delta E = 4.8 \text{ eV} \tag{6}$$

We observed two atomic H transitions in our spectra after electron recombination of H_3^+ (Figure 8); both originated from energy levels above 13 eV. This is more than the kinetic energy released in an energetically more advantageous process (Eq. 6). Because of the high pressure of He, excitation of H atoms can be expected in collisions with excited He atoms in the early afterglow.

2.3 $(CN)_2$-containing discharge

The CN free radical is observed in interstellar molecular clouds and the atmospheres of stars, planets and comets. It is also significant in numerous laboratory processes at high temperatures (flames, chemical reactions, discharges) where it is often formed from trace amounts of carbon and nitrogen. It is a very strong absorber/emitter of radiation and its spectra, extending from the vacuum UV far into the infrared without significant gaps, provide a very useful tool for its detection and monitoring. A vast proportion of the available spectral data arises from the $A^2\Pi$–$X^2\Sigma^+$ and $B^2\Sigma^+$–$X^2\Sigma^+$ electronic transitions (Prasad & Bernath, 1992; Ram et al., 2006) and the infrared transitions in the $X^2\Sigma^+$ ground electronic state (Cerny et al., 1978; Horká et al., 2004). In our previous paper (Horká et al., 2004) we concentrated primarily on the measuring and analysis of $^{12}C^{14}N$ vibration-rotation bands for the sequences $\nu = (1–0)$ through (9–8) which were observed in the spectral region 1800–2200 cm^{-1} with Fourier Transform spectroscopy. From the point of view of the vibrational excitation, the most important information is obtained from vibronic data involving vibrational levels up to $\nu = 18$ (Ram et al., 2006). Such high vibrational excitation corresponds to temperatures well above 45000 K thus indicating the potential use of CN in high temperature monitoring and the possibility of experimental determination of the molecular potential energy function (Horká et al., 2004).

Cerny et al. (1978) analyzed fourteen vibronic bands of the $\Delta\nu = 1, 0, -1, -2$ spread out in the near infrared spectral range with $\nu' = 0$ to 4 for $A^2\Pi$ electronic state. Kotlar et al. (1980) carried out a perturbation analysis of data taken at the University of Berkeley, to give a deperturbed set of the constants for the $\nu = 0$ to $\nu = 12$ vibrational levels of the $A^2\Pi$ state. Prasad & Bernath (1992) measured and analyzed the red system of CN by using a jet-cooled corona excited supersonic expansion in a spectral range of 16500–22760 cm^{-1}. They measured a total of 27 bands with $\nu' = 8$ to 21 for $A^2\Pi$ electronic state. Furio et al. (1989) used the laser fluorescence excitation spectra for the measurement of the $B^2\Sigma^+$–$A^2\Pi(\nu = 8, 7)$ band in the 20400 cm^{-1} spectral range and derived the constants for $\nu = 7$ of the $A^2\Pi$ state. Rehfuss et al. (1992) used an FT spectrometer in the ultraviolet, visible and infrared region for a measurement of the CN spectrum. A total of 54 bands were observed throughout the red and infrared region from 16000 to 2500 cm^{-1}. The observed sequences include $\Delta\nu = +4$, $+3$, $+2$, $+1$, 0, -1, -2 and -3 with vibrational levels up to $\nu = 14$, where some sequences were not observed, due to small Franck–Condon factors and/or sensitivity of the spectrometer.

The 0–0 band of the $A^2\Pi$–$X^2\Sigma^+$ system appears at 9117 cm^{-1}. Since the vibrational frequency of CN is about 2042 cm^{-1} and 1813 cm^{-1} in the $X^2\Sigma^+$ and $A^2\Pi$ respectively, the $\Delta\nu = -1$, -2 and -3 sequences occur near 7000, 5000 and 3000 cm^{-1} respectively. In the region between 5000 and 2000 cm^{-1}, the vibronic transitions are rather unfavorable due to the Franck-Condon factors of 0.15–0.05 Prasad & Bernath (1992); Sharp (1984). Furthermore one loses, compared with the 0–0 band, at least an additional factor of 20 due to the ν^3 dependence in the Einstein A coefficient. Thus a high resolution vibronic CN spectrum with a good signal-to-noise ratio for the $\nu = 3$ sequence band region has not been reported until now. Only a low resolution spectrum was weakly observed by Rehfuss et al. (1992).

There is still a gap for the high resolution measurement and detailed analysis of the spectral bands concerning $\nu = 5$ to 6 of the $A^2\Pi$ state. The turning point in this measurement of CN in the infrared spectral range was the introduction of time resolved FT spectroscopy (Civiš et al., 2008). This method makes it possible to distinguish the weak emission (or absorption) bands from strong bands appearing in the spectrum if the time profiles are different. In the case of CN, weak vibronic bands in the 5 μm region were separated from strong long lived vibration-rotation bands. In this part a spectroscopic analysis of 7 newly observed $\Delta\nu = -3$ sequences bands: 0–3, 1–4, 2–5, 3–6, 4–7, 5–8 and 6–9 of the $A^2\Pi$–$X^2\Sigma^+$ transition is reported.

2.3.1 Observed CN spectra and their analysis

Figure 9 shows a part of the time-resolved FT spectra of emission from a discharge in a $(CN)_2$ and He mixture, where the discharge pulse width was 20 μs. Thirty time-resolved spectra were obtained in one measurement with a time-interval of 3 μs, and 6 spectra are shown in Figure 9.

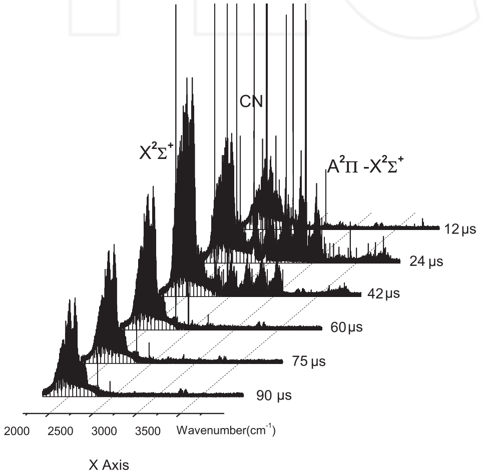


Fig. 9. The time-resolved emission FT spectrum from a pulsed discharge in a $(CN)_2$ and He mixture. The discharge pulse duration was 20 μs. The 30 time-resolved spectra were collected from $t = 0$–90 μs with a step of 3 μs. The spectra of C_2H_2 and C_2 were observed at 3300 and 3600 cm^{-1}.

The variation in intensity of the vibrational bands in the $X^2\Sigma^+$ state is low. On the other hand, relaxation of electronic transitions is as fast as expected from a short radiative lifetime. The wavenumber resolution of Figure 9 was 0.07 cm^{-1}, which was found to be insufficient for the analysis of the majority of the electronic transitions, because in the band-head region, the lines remained unresolved and the fit of the spectra was unsatisfactory.

In another time-resolved measurement we used a 0.025 cm^{-1} resolution with a long discharge pulse (40 μs) in order to reach the maximum excitation and to set the system into a “steady state”. The data collection system was set using the offset time of 5 μs before the end of the pulse and the spectra were taken in 1 μs intervals. A series of experiments was carried out under this condition, while the basic parameters of the discharge, He pressure and discharge current were varied. A series of time-resolved FT spectra was measured in time intervals of 1–30 μs, providing the time profile of CN relaxation from the $A^2\Pi$ state to the ground electronic state. The time-scale was short for the study of the relaxation of the vibration-rotation transitions in the ground electronic state, but is enough for observations of relaxation of atomic He, N, C lines, and the C_2 radical. From this vast complex spectra, the spectrum No. 15, as shown in Figure 10, was chosen for the present spectroscopic analysis, which was obtained 10 μs after the end of the discharge. Figure 11 shows typical time-profiles of a vibration-rotation transition in the $X^2\Sigma^+$ state and a vibronic transition of CN, together with the atomic lines. The lifetime of the $A^2\Pi$–$X^2\Sigma^+$ transition is an order of 10 μs. However, the vibrational relaxation of CN in its ground state is significantly longer; even at a time of 25 μs after the discharge, the intensity of the vibrational fundamental band is still rising. The intensity of the He atomic line (2469.7 cm^{-1}) shows a fast decay, and lines of C and N atoms also relax with a speed comparable with atomic helium.

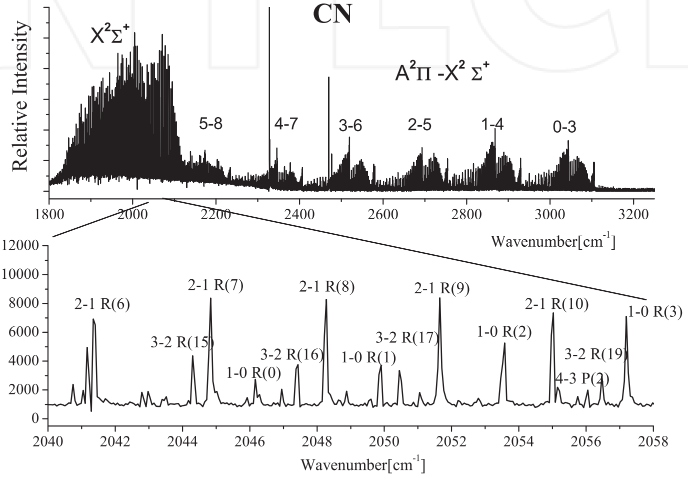


Fig. 10. The emission spectrum of the CN radical in the spectral range 3–5 μm. The overall view of the CN $A^2\Pi$–$X^2\Sigma^+(\Delta\nu = -3)$ sequence with the 1–0 fundamental electronic vibration–rotation band present with a band origin at 2042 cm^{-1} and the other hot bands.

The rotational assignments of the $A^2\Pi$–$X^2\Sigma^+$: 0–3, 1–4, 2–5, 3–6, 4–7 bands were carried out according to the transition frequency calculations using molecular constants reported by Cerny et al. (1978). The $v = 5$–8 and 6–9 bands were assigned by molecular constants from Kotlar et al. (1980) who reported the Dunham parameters. The standard deviation of the fitting was 0.0009 cm^{-1}.

The time-resolved experiment itself was carried out in a wide range of time scales and with various discharge pulse lengths, thus enabling a complex study of the relaxation processes of the CN system in helium. Such a pulsed discharge gives a stronger emission for the $\Delta\nu = -3$ bands, compared with that of a DC discharge. In this study emission from the $\nu = 5$ and 6 vibrational levels of the $A^2\Pi$ state was observed for the first time. The observed intensity of each band was plotted against the energy value of the upper state of the transition to give a vibrational temperature of 6700 K for the $A^2\Pi$. The value does not change significantly after 10 μsec, because the vibrational relaxation is not fast. Similarly the vibrational temperature is

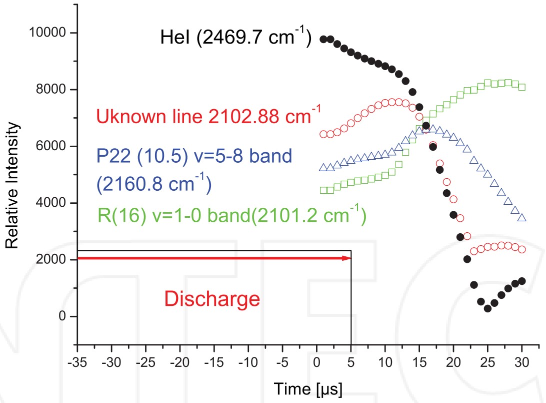


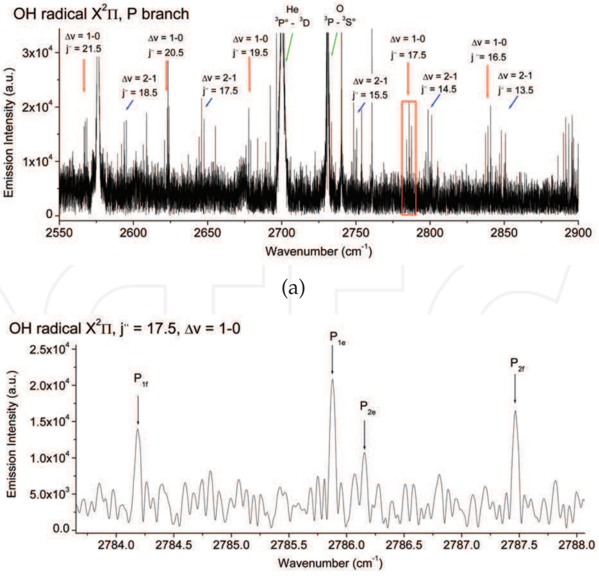
Fig. 11. Time-profiles of the He(I) atomic line, P22 (10.5) line of the A–X, 5–8 band and R(16) line of the 1–0 fundamental vibration-rotation band, and nitrogen atomic line. The discharge pulse was 40 μs long. The time-resolved spectra were collected after 35 μs (5 μs before the end of the discharge pulse) with a step of 1 μs.

found to be 6757±534 K for the $X^2\Sigma^+$. Using the vibrational temperatures, we estimated the abundance of the $A^2\Pi$ to be 0.58% of that of the $X^2\Sigma^+$.

2.4 Other examples

A discharge in a hydrogen-containing mixture can produce hydrides whose spectra can be registered using Fourier transform spectroscopy. One of then goals of such studies is simulation of potential high energy processes in early Earth's atmosphere (as meteorite impact, lightning), which could lead to more complex compounds generated from simple molecular gases. (Babánková et al., 2006). Large-scale plasma was created in molecular gases (O_2, N_2, C_2H_4) and their mixtures by high-power laser-induced dielectric breakdown (LIDB). Compositions of the mixtures used are those suggested for the early Earth's atmosphere. Time-integrated as well as time-resolved optical emission spectra emitted from the laser spark have been measured and analyzed. The spectra of the plasma generated in the above mixtures are dominated by emission of diatomic radicals which are precursors of stable products as acetylene and hydrogen cyanide. Occurrence of these species was confirmed in irradiated gaseous mixture by FTIR spectroscopy. The figures below illustrate spectra of some hydrides formed in reactions due to discharge in different hydrogen-containing mixtures.

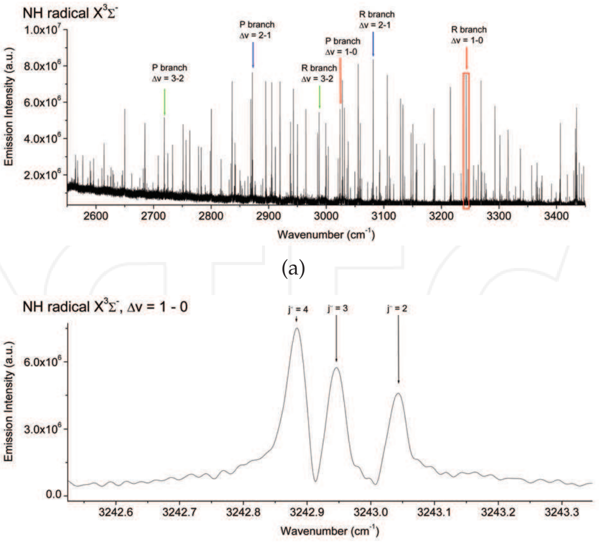
Baskakov et al. (2005) applied the Fourier transform spectroscopy to study a dc glow discharge in a mixture of argon and hydrogen. Several strong emission bands of ^{40}ArH were observed in the 2500–8500 cm^{-1} region. Rotational-electronic transitions of the two previously unstudied $4p$–$5s$ and $5p$–$6s$ (ν = 0–0) bands of ArH were measured and assigned in the 6060 and 3770 cm^{-1} regions, respectively. An overview spectrum of the $4p$–$5s$ and $5p$–$6s$ band is shown in Figure 16. There are still many unassigned bands in the observed spectra including higher vibrationally excited ν–ν bands.



(a)

(b)

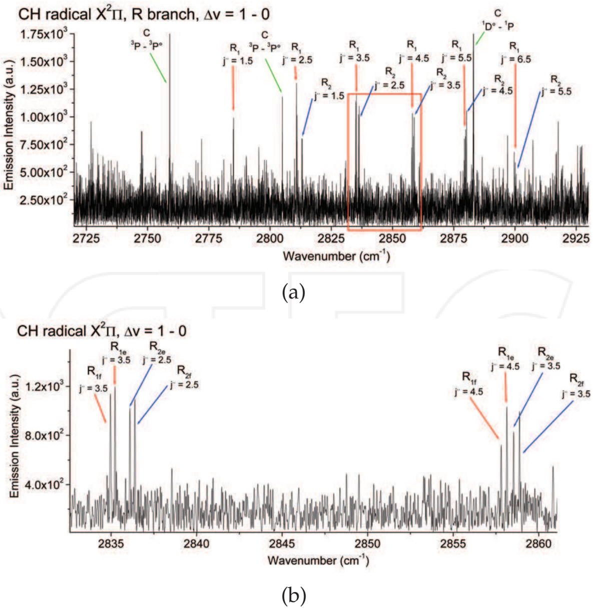
Fig. 12. Spectra of OH radical formed in $H_2 + O_2 + He$ mixture discharge



(a)

(b)

Fig. 13. Spectra of NH radical formed in $H_2 + N_2 + He$ mixture discharge



(a)

(b)

Fig. 14. Spectra of CH radical formed in CH_4 + He mixture discharge

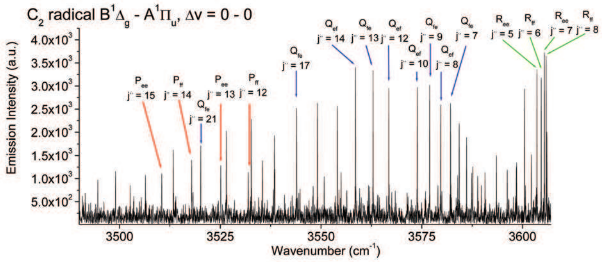


Fig. 15. Spectra of C_2 radical formed in CH_4 + He mixture discharge

3. Continuous scan systems: application with lasers

3.1 Interleaved sampling $1/n$

In the case of the measurement of time-resolved spectra in combination with a laser whose maximum repetition rate is slower than the interferometer mirror speed, there is no possibility of sampling at each individual trigger point of the He–Ne laser.

The lowest scanning speed of the interferometer is limited to the He–Ne laser fringe frequency of about 3 kHz. However, by utilizing the under-sampling condition, sampling several times more slowly becomes possible. Figure 17 shows the clock pattern for sampling in the present experiments where triggers for the pulse event and for the sampling are produced with a period of $1/n$ times the He–Ne laser fringe frequency. Complete interferograms are then obtained with n scans if the trigger point is changed for each scan. The time sequence shown in Figure 17 corresponds to the case of $n = 3$.

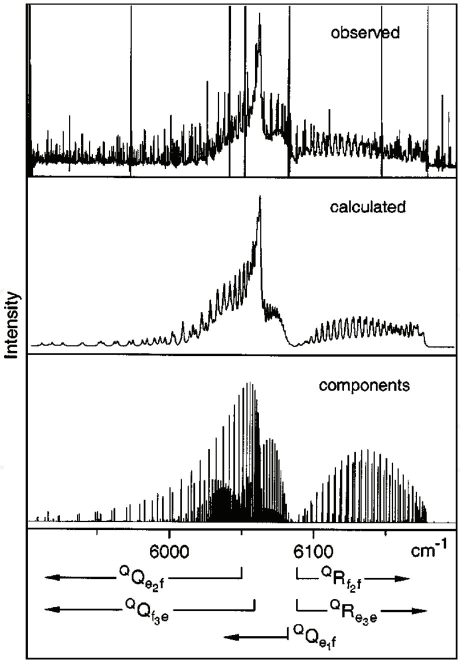


Fig. 16. Observed and calculated emission spectrum of the 4*p*–5*s* band of ArH. The strongest lines in the upper drawing originated from the Ar atom. Sharp lines belong to the 4*f*–3*d* band and probably to other unnassigned bands of ArH.

The maximum frequency of the used ArF laser is 1 kHz. The laser pulse is therefore repeated every 1000 μs. The minimum speed of the interferometer mirror is 3 kHz, then the digital signal produced by the He–Ne laser is repeated every 333.33 μs. In order to obtain data in the maximum density, *i. e.* for every trajectory difference defined by the He–Ne laser, the complete record is taken during three scans (for the mirror speed 3 kHz and laser repetition frequency 1 kHz). The complete set of the time-resolved spectra (one complete interferogram) is acquired by three time-shifted scans.

3.2 Synchronous triggering and data sampling

FT data are taken at the zero crossing points of the He–Ne laser fringe signals, while the wavelength of the He–Ne laser is used for the measurement of path differences. The data are sampled at time intervals which correspond to a mirror movement of either one wavelength or half a wavelength, depending on the frequency range of the measurements (8000 or 16000 cm^{-1}). Time resolved spectra are obtained by collecting data at various points between the zero-crossings and calculating the FT transformation for each such point. This system was utilized using a Field Programmable Gate Array processor (FPGA). The main role of the FPGA processor in our experiment was to create a discharge or laser pulse and AD trigger signals (the signal for data collection from the detector) synchronously with the He–Ne laser fringe signals from the spectrometer. The FPGA processor also controls the data transmission from the digital input board to the PC.

Figure 17 depicts the timing chart (clock pattern) produced by the FPGA for the laser-pulse ablation method. The scan signal and He–Ne laser fringe signals are supplied by the Bruker 120 HR spectrometer and used as the system time standard. A discharge trigger pulse is produced at a width which is preset by the FPGA. AD triggers are also produced by the FPGA with a time offset value between the beginning of the laser pulse and the interval between pulses. In the present experiments we used a 60 μs offset and interval values covering, 30 μs when 30 AD triggers were supplied. A series of data signals corresponding to the AD triggers are stored and Fourier-transformed.

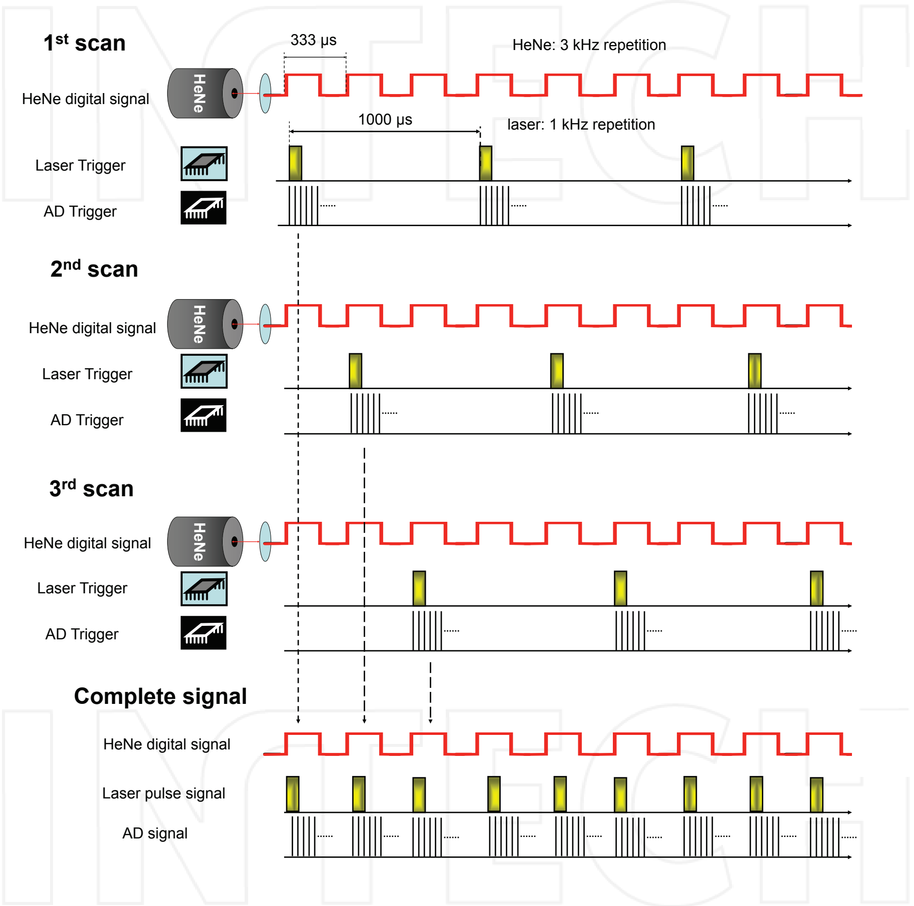


Fig. 17. Timing diagram for the interleaved sampling. During the scan, the laser pulse and the AD trigger sampling are induced with a rate of $1/n$ times of the He–Ne laser fringe frequency. The complete interferograms are obtained after n scans ($n = 3$ here).

The maximum number of spectra taken by this method is 64. The time resolution is about 1 μs, which is limited by the band width of the detector amplifier. The present system collects 64 times more data in comparison to the original Bruker system. This is possible because of rapid development in the field of PCs, their memory size and the writing speed of the hard disk. In the case of our current data collection program, we are able to store 64 interferograms in a single scan when the resolution is up to 0.03 cm^{-1}. For resolutions higher than 0.03 cm^{-1}, 30 time-resolved data are recorded simultaneously, which means the number of time-resolved data points can be varied according to the type of experiment and the memory capacity.

3.3 The experimental setup

The time resolution FTIR spectra were measured using the modified Bruker IFS 120 HR spectrometer (modified for the time-resolution scan of emission data) in a spectral range of 1800–6000 cm^{-1} using a CaF_2 beam splitter,and an InSb detector. The aperture size was 4 mm, the preamplifier gain was 3. The spectra were measured at a resolution of 0.1 cm^{-1} with a MID IR filter for the range of 1538–3500 cm^{-1} (number of scans from 1 to 10, zero filling 2, trapezoid apodization function).

The Bruker system was equipped with an analog-digital converter (ADC 4322: Analogic, USA), which was connected to a PC containing a programmable control processor of Field Programmable Gate Array – FPGA, (ACEX 1K: Altera, USA) set up at a frequency of 33 MHz, and digital input board PCI (2172C: Interface, Japan). The data collection process and synchronization with the laser were controlled by the FPGA processor programmed by QUARTUS II 7.1, Altera. The computer programs for data acquisition and fast FT transformation and displaying of the data were written in C $^{++}$ language.

Time-resolved FTIR spectroscopy was applied for observations of the emission arising after the irradiation of metals with a pulsed nanosecond ArF ($\lambda = 193$ nm) laser. A high repetition rate ArF laser ExciStar S-Industrial V2.0 1000 (193 nm, laser pulse width 12 ns, frequency 1 kHz) with 15 mJ output power was focused on a rotating and linearly traversing gold target inside a vacuum chamber (average pressure 10^{-2} Torr). The infrared emission (axial distance from the target 10 mm) was focused into the spectrometer using a CaF_2 (100 mm) lens (see Figure 18). The emission was observed in the 1800–3500 cm^{-1} spectral region with a time profile showing maximum emission intensity at 9–11 μs after the laser shot.

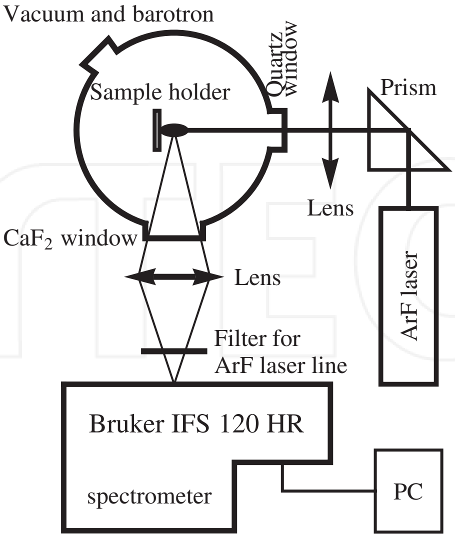


Fig. 18. Experimental set-up of the metal emission measurement.

For data sampling we used the so-called 1/3 sampling, where the scanner rate was set to produce a 3 kHz HeNe laser interference signal, the ArF laser oscillation was triggered, and 30 sets of time-resolved data were recorded with a preset time interval of 1 μs. Three scans were needed for a complete interferogram, and only 5 scans were coadded to improve the signal-to-noise ratio. The acquired spectra were post-zerofilled in the OPUS program and subsequently corrected by subtracting the blackbody background spectrum. The wavenumbers, line widths and their intensities were then obtained using the peak picking method (OPUS).

4. TR FTIR emission spectroscopy

4.1 Introduction

Pulsed laser ablation and depositing processes are currently frequently used techniques. Laser induced plasma at low fluence (typically 10 J/cm^2) has numerous applications, *e. g.* Pulsed Laser Deposition (PLD) or multi-elemental analysis. The latter technique, known as Laser Induced Plasma Spectroscopy (LIPS) or Laser-Induced Breakdown Spectroscopy (LIBS) consists of analyzing the light spectrum emitted from a plasma created on the sample surface by laser pulses. LIPS has many practical advantages over the conventional methods of chemical analysis of elements and is consequently being considered for a growing number of applications (Babánková et al., 2006; Barthélemy et al., 2005; Gomes et al., 2004; Lee et al., 2004; Radziemski & Cremers, 1989).

Excimer lasers operating in near-ultraviolet regions with typical laser fluences of 1–30 J/cm^2 are used for many types of ablation (Claeyssens et al., 2003). The ablation plume arising after irradiation with fluences of nanosecond duration pulses is governed by a great number of very complex physical processes. During the laser pulse (with typical duration of 20 ns), the laser photons heat the sample and bring a part of its surface to the critical temperature. The heated material starts to boil explosively (Miotello & Kelly, 1999) and creates an emission plume consisting of ejected particles, atoms and ions. The particles inside the plume can themselves interact with the laser photons, which leads to a subsequent rise in the temperature of the ablation plume and to photochemical and photodissociation processes (Rubahn, 1999, p. 219). The population of Rydberg states responsible for IR emission lines is governed mainly by collisional processes. The electrons created in the photodissociation processes can interact with the laser pulse via the electron-ion inverse bremsstrahlung, which again causes additional heating of the plume (Vertes et al., 1994) and leads to the fast transition of the plume from ionized gas to plasma. The electrons escaping from the corona region cause a separation of charges, thereby inducing the ionized part of the plasma to accelerate. After the end of the laser pulse, the plume expands adiabatically. The electron-ion collision inside the plume can create excited ions. The electron-ion collision in the presence of a third body can results in their recombination leading to formation of atoms in highly excited Rydberg states (Claeyssens et al., 2001). A radiative cascade of these Rydberg states is then observed as the optical emission of the ablation plume.

The investigation of such emission is complicated by nonequilibrium and nonstationary conditions of the plasma for the excited states (Aragon & Aguilera, 2008), so the information on population dynamics is only scarcely available for these states Rossa et al. (2009). As an example of such data we report temporal evolution dynamics for each IR atomic line of the recorded spectra of metal atoms.

The properties of the observed plumes obtained by the ablation of different materials can eventually reflect the superposition of the ensemble processes described above. Here we report some results of a study focused on time-resolved spectra arising from 193 nm pulsed laser ablation of metallic (Au, Ag and Cu) targets in a 10^{-3} Torr vacuum. The atomic metal spectra were measured by a high resolution Fourier Transform infrared spectrometer specially modified for time-resolved measurements (Civiš et al., 2010; Civiš et al., 2010).

4.2 Results for Au

The observed IR emission spectra of the Au atom are presented in Figure 19 at 10 μs after the laser shot, when the time profile of the emission intensity is maximum for all the observed lines.

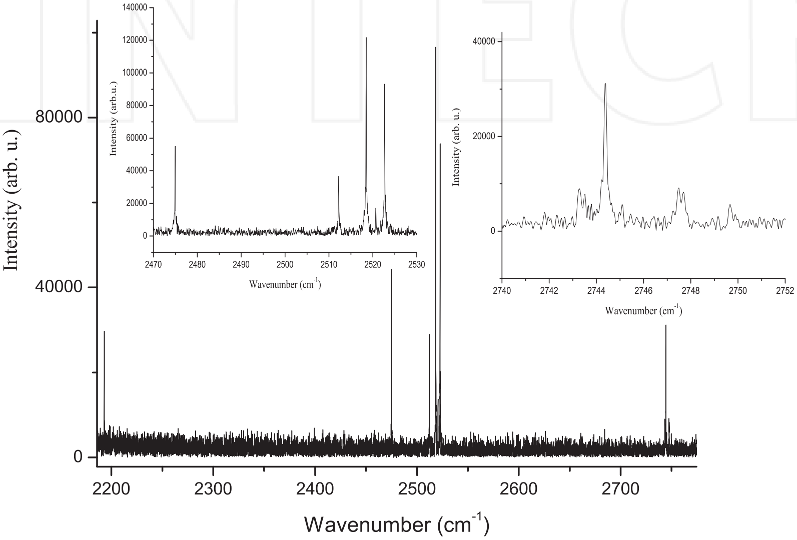


Fig. 19. Some parts of the observed IR emission spectra of Au. The 2743.358 cm^{-1} and 2747.567 cm^{-1} values are given to the centers of gravity of the hyperfine patterns clearly seen in the second graph as double peaks.

Although the Au spectrum has been studied in various spectral domains for several decades (Brown & Ginter, 1978; Ding et al., 1989; Dyubko et al., 2005; Ehrhardt & Davis, 1971; George et al., 1988; Jannitti et al., 1979; Platt & Sawyer, 1941), to our knowledge only one experimental study George et al. (1988) concerning the studied 3–5 μm IR range is reported. As compared to George et al. (1988) we observed several strong new Au lines in the 1800–4000 cm^{-1} domain. The most prominent IR lines observed for Au are listed in Table 3. Their half-widths at half-maxima (HWHM) are calculated from fitting to the Lorentzian shape. The decay time, τ given in the Table 3 was calculated by fitting of the measured time profiles of the corresponding lines. These profiles are given in Figure 20. The time decay of most of the strong lines is well described by exponential fitting, excepting the 2156.484 cm^{-1} line which demonstrates a non-constant decay rate during the 30 μs after the laser shot. Some weaker lines demonstrate such behavior more clearly, their decay is not exponential (and is

Wavenumber (cm^{-1})	Intensity (arb.u)	HWHM (cm^{-1})	Decay time (μs)	Identification
2156.484	12679	0.098	$5.24 \pm 1.7^*$	$8f_{\frac{7}{2}} \to 8d_{\frac{5}{2}}$
2193.030	38690	0.12	6.56 ± 0.61	$8s_{\frac{1}{2}} \to 8p_{\frac{1}{2}}$
2428.358	8024	0.39	$6.36 \pm 1.4^*$	$12s_{\frac{1}{2}} \to 9p_{\frac{3}{2}}$
2474.954	53951	0.13	5.25 ± 0.21	$7d_{\frac{5}{2}} \to 6f_{\frac{7}{2}}$
2512.219	36631	0.14	5.73 ± 0.25	$9p_{\frac{1}{2}} \to 10d_{\frac{3}{2}}$
2518.489	121588	0.13	5.56 ± 0.22	$7d_{\frac{3}{2}} \to 6f_{\frac{5}{2}}$
2520.684	16574	0.16	6.83 ± 0.58	$8p_{\frac{3}{2}} \to 8d_{\frac{5}{2}}$
2522.683	91622	0.13	5.70 ± 0.36	$5f_{\frac{7}{2}} \to 8d_{\frac{5}{2}}$
2743.370	8780	0.39	$5.41 \pm 1.2^*$	?
2744.380	41786	0.12	5.96 ± 0.68	$8s_{\frac{1}{2}} \to 8p_{\frac{3}{2}}$
2747.567	10249	0.13	$8.52 \pm 2.6^*$	$12s_{\frac{1}{2}} \to 9p_{\frac{1}{2}}$
2749.6	6453	0.27	$5.51 \pm 1.8^*$	$23p_{\frac{1}{2},\frac{3}{2}} \to 9d_{\frac{3}{2}}$
3187.811	13199	0.093	$6.03 \pm 0.91^*$	$(5d^9 6s)6p_{\frac{3}{2}} \to 6d_{\frac{5}{2}}$
3828.96	9540	0.14	$5.47 \pm 2.0^*$	$7f_{\frac{7}{2}} \to 7d_{\frac{5}{2}}$
3862.41	9370	0.13	$2.98 \pm 1.1^*$	$9d_{\frac{3}{2}} \to 5f_{\frac{5}{2}}$
3866.606	7738	0.11	$6.84 \pm 0.83^*$	$9d_{\frac{3}{2}} \to 8p_{\frac{3}{2}}$

Table 3. Experimental Au lines and their identification. The decay time, τ, was calculated by exponential fitting of the measured time profiles of the corresponding lines. Time profile of the lines denoted by asterisk demonstrates significant deviation from the exponential decay; τ values are roughly approximate

even non-monotonic), so their τ values are estimated in Table 3 in a rough approximation. Such a non-monotonic decay can be due to more complex population kinetics of the atomic Au states in the ablation plasma. We consider most of the observed lines to be due to transitions between the Rydberg $n = 5..10$ states of the valence electron outside the closed-shell $5d^{10}$ core.

4.3 Results for Ag

Some parts of the observed IR emission spectra of the Ag atom are presented in Figure 21 at 11 μs after the laser shot, when the time profile of the emission intensity is maximum for all the observed lines. The most prominent IR lines observed for Ag are listed in Table 4. Their full widths at half-maxima (FWHM) are calculated from fitting to a Voigt profile, but under our conditions this profile does not differ much from the Lorentzian shape (see Ref. Civiš et al. (2010)).

We measured the emission spectrum at a different delay time, from 0 to 30 μs, after the laser shot. This allows us to measure the time profiles of the observed Ag lines. Some such profiles are shown in Figure 22. The temporal decay of some lines is well described by exponential fitting, while some lines display non-exponential (including some “plateaux” at 20–25 μs after the laser shot) and even non-monotonic behavior. Therefore their decay time, τ, values are estimated in Table 4 in a rough approximation; so, for essentially non-exponential decays, the standard deviation $\Delta\tau$ is of order of τ.

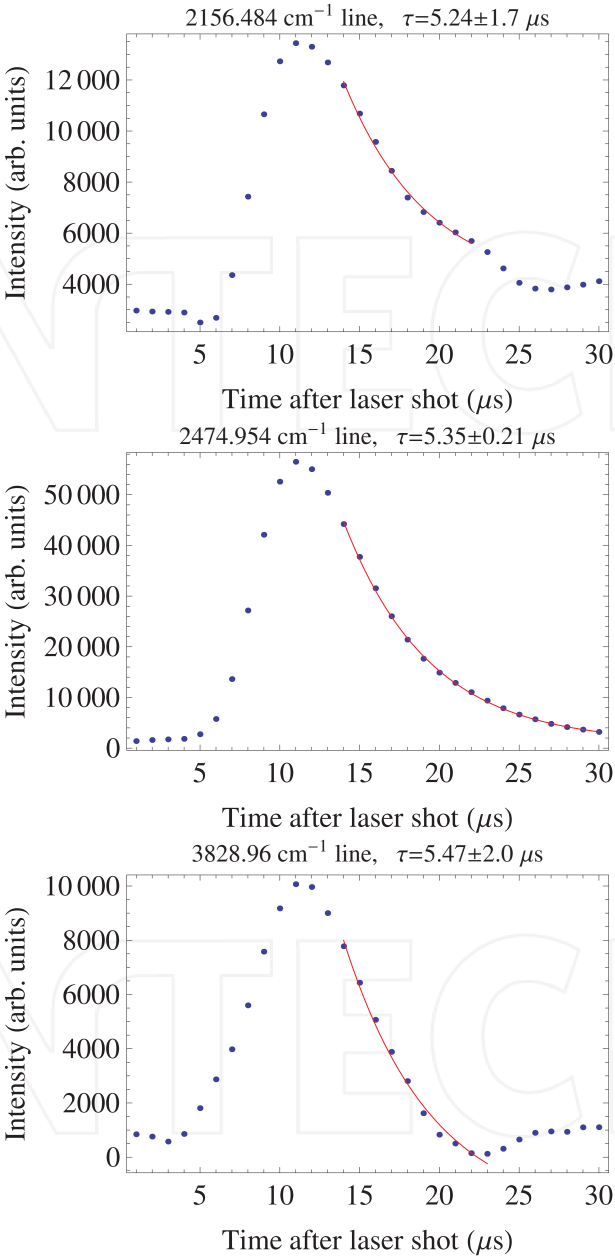


Fig. 20. The time profiles of some observed lines (dots) and their fit with exponential decay (solid curves)

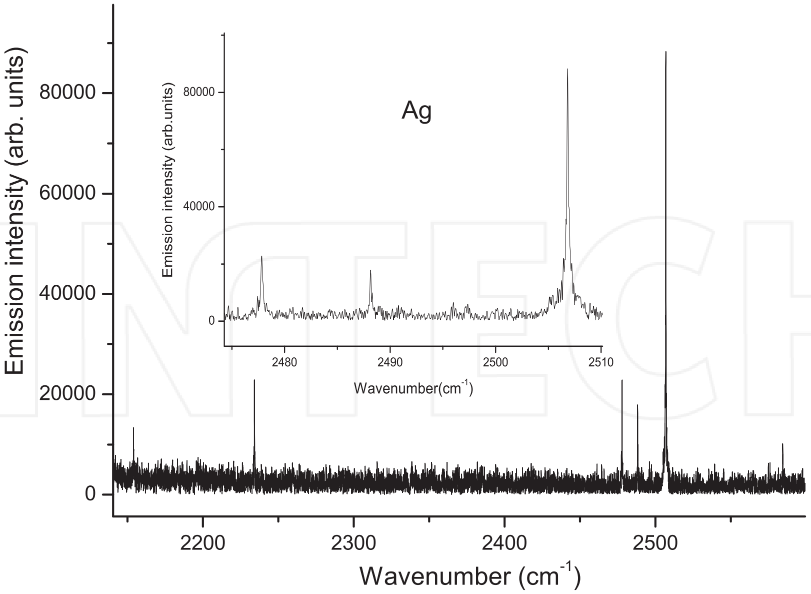


Fig. 21. A section of the the observed IR emission spectra of Ag.

It should be noted that the decay times $\tau \simeq$ 1–10 μs given in Table 4 are not related to the radiative lifetimes of Ag atom levels which are at least two orders shorter (Bengtsson et al., 1991; 1990; Zhankui et al., 1990). The temporal dynamics shown in Figure 22 is due to a complex combination of the collisional cascade repopulation of the emitting levels (Civiš et al., 2010) and the transfer processes in ablation products (Kawaguchi et al., 2008).

The revised energy values E_i of some Ag terms are presented in Table 5. For the levels with $n \leq 6$ our values coincides with the reference values within the uncertainty intervals, but it is not the case for $n > 6$. However we consider our values preferable since they are extracted from spectra recorded with 0.02 cm^{-1} resolution while the reference values were obtained from spectra with resolution of 0.035–0.045 cm^{-1} Pickering & Zilio (2001) and 0.06 cm^{-1} Brown & Ginter (1977)

4.4 Results for Cu

Figure 23 shows some parts of the observed IR emission spectra of Cu I at 20 μs after the laser shot, when the emission intensity is maximal for almost all of the observed lines. The list of the IR lines observed for Cu I is presented in Table 6. Their full widths at half-maxima (FWHM) are calculated from fitting to the Lorentzian shape (Civiš et al., 2010; Civiš et al., 2010).

As in the cases of Au and ag, we measured the time profiles of the observed Cu lines, *i. e.* their emission intensities as functions of the delay time, from 0 to 60 μs, after the laser shot. Some such profiles are shown in Figure 24. While the temporal decay of some lines can be fitted, at least roughly, by an exponential function, several lines display essentially non-exponential behavior including some "plateaux" or even secondary maxima at 35–50 μs after the laser

Wavenumber (cm^{-1})	Intensity (arb. units)	SNR	FWHM (cm^{-1})	Decay time (μs)	Identification
1345.570(6)	14889	14.	0.125(55)	4.17(104)(b)	$(4d^{10})7d_{\frac{3}{2}} \leftarrow (4d^{10})6f_{\frac{5}{2}}$
1363.326(14)	3910	4.3	0.146(150)	5.90(609)(b)	$(4d^{10})5f \leftarrow (4d^{10})6g$
1460.115(6)	4530	5.6	0.111(035)	3.34(170)(b)	$(4d^{10})7p_{\frac{3}{2}} \leftarrow (4d^{10})8s_{\frac{1}{2}}$
2149.320(2)	6956	2.4	0.025(19)	7.17(206)(b)	$(4d^{10})7d_{\frac{5}{2}} \leftarrow (4d^{10})7f_{\frac{7}{2}}$
2153.988(4)	32525	9.4	0.031(21)	3.07(15)	$(4d^{10})7s_{\frac{1}{2}} \leftarrow (4d^{10})7p_{\frac{1}{2}}$
2154.599(2)	5958	2.2	0.029(11)	12.8(78)(b)	$(4d^{10})7d_{\frac{3}{2}} \leftarrow (4d^{10})7f_{\frac{5}{2}}$
2234.168(1)	74835	13.	0.035(6)	3.63(34)	$(4d^{10})7s_{\frac{1}{2}} \leftarrow (4d^{10})7p_{\frac{3}{2}}$
2417.757(2)	4880	5.9	0.030(14)	5.03(214)(b)	$(4d^{10})6d_{\frac{3}{2}} \leftarrow (4d^{10})8p_{\frac{1}{2}}$
2447.045(2)	5463	6.6	0.026(24)	4.88(81)(b)	$(4d^{10})6d_{\frac{5}{2}} \leftarrow (4d^{10})8p_{\frac{3}{2}}$
2477.833(1)	113278	10.	0.044(4)	4.27(29)	$(4d^{10})6d_{\frac{5}{2}} \leftarrow (4d^{10})5f_{\frac{7}{2}}$
2488.1560(9)	78947	7.0	0.048(4)	4.20(35)(a)	$(4d^{10})6d_{\frac{3}{2}} \leftarrow (4d^{10})5f_{\frac{5}{2}}$
2506.8196(5)	460526	41.	0.056(2)	3.87(20)	$(4d^{10})4f \leftarrow (4d^{10})5g$
2579.210(3)	5252	6.2	0.036(21)	5.14(202)(b)	$(4d^{10})7p_{\frac{3}{2}} \leftarrow (4d^{10})7d_{\frac{3}{2}}$
2584.3387(6)	48450	31.	0.041(3)	4.20(37)(b)	$(4d^{10})7p_{\frac{3}{2}} \leftarrow (4d^{10})7d_{\frac{5}{2}}$
2658.7911(9)	22693	14.	0.038(5)	4.15(39)(a)	$(4d^{10})7p_{\frac{1}{2}} \leftarrow (4d^{10})7d_{\frac{3}{2}}$
3386.152(1)	98370	9.2	0.059(6)	3.35(28)	$(4d^{10})6p_{\frac{3}{2}} \leftarrow (4d^{10})7s_{\frac{1}{2}}$
3589.552(5)	13825	15.	0.065(4)	3.15(11)(a)	$(4d^{10})6p_{\frac{1}{2}} \leftarrow (4d^{10})7s_{\frac{1}{2}}$

Table 4. Experimental Ag lines and their identification. The decay time, τ, was calculated by exponential fitting of the measured time profiles of the corresponding lines. The profiles denoted as ((a)) demonstrates significant deviation from the exponential decay; those denoted by (b) demonstrate the decay curves of essentially non-exponential form with a plateau or a second maximum; τ value is roughly approximate

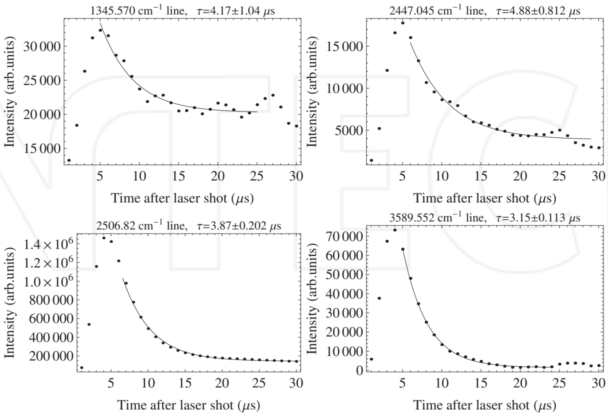


Fig. 22. The time profiles of some observed Ag lines (dots) and their fit with exponential decay (solid curves)

Term	Energy (cm^{-1})	Other sources
$(4d^{10})5f_{\frac{5}{2}}$	56691.275(2)	56691.4 Shenstone (1940), 56692.5 Safronova et al. (2003)
$(4d^{10})5f_{\frac{7}{2}}$	56691.397(4)	56691.4 Shenstone (1940), 56694.4 Safronova et al. (2003)
$(4d^{10})6p_{\frac{1}{2}}$	48297.402(2)	48297.402(3) Pickering & Zilio (2001)
$(4d^{10})6p_{\frac{3}{2}}$	48500.804(1)	48500.804(2) Pickering & Zilio (2001)
$(4d^{10})6d_{\frac{3}{2}}$	54203.119(2)	54203.119(2) Pickering & Zilio (2001)
$(4d^{10})6d_{\frac{5}{2}}$	54213.564(3)	54213.570(3) Pickering & Zilio (2001)
$(4d^{10})6f_{\frac{5}{2}}$	58045.481(7)	This work
$(4d^{10})6g$	58054.723(16)	This work
$(4d^{10})7s_{\frac{1}{2}}$	51886.954(1)	51886.971(2) Pickering & Zilio (2001)
$(4d^{10})7p_{\frac{1}{2}}$	54041.087(2)	54040.99(6) Brown & Ginter (1977)
$(4d^{10})7p_{\frac{3}{2}}$	54121.059(2)	54121.129(5) Pickering & Zilio (2001)
$(4d^{10})8s_{\frac{1}{2}}$	55581.246(3)	55581.258(3) Pickering & Zilio (2001)
$(4d^{10})8p_{\frac{1}{2}}$	56620.876(3)	56620.72(6) Brown & Ginter (1977)
$(4d^{10})8p_{\frac{3}{2}}$	56660.596(6)	56660.559(17) Pickering & Zilio (2001)
$(4d^{10})7d_{\frac{3}{2}}$	56699.911(2)	56699.768(3) Pickering & Zilio (2001)
$(4d^{10})7d_{\frac{5}{2}}$	56705.435(2)	56705.498(3) Pickering & Zilio (2001)
$(4d^{10})7f_{\frac{5}{2}}$	58854.510(3)	This work
$(4d^{10})7f_{\frac{7}{2}}$	58854.755(3)	This work

Table 5. Revised values of some levels of Ag I

shot. Their decay time, τ, values are therefore estimated in Table 6 in a rough approximation; it is seen from this table that for essentially non-exponential decays the uncertainity $\Delta\tau$ is of the same order of magnitude as τ itself. Note that that the decay times τ given in Table 6 are due to a complex combination of the collisional cascade repopulation of the emitting levels (Civiš et al., 2010) and the transfer processes in ablation products (Kawaguchi et al., 2008) and by no means related to the radiative lifetimes of the atomic levels. The temporal dynamics of some lines is shown in Figure 24.

After the assignment we refined the energy values for some levels involved into the classified transitions; the revised values of these energies are presented in Table 7. It is interesting to note that the fine-structure $5p$ doublet (fine splitting is about 0.3 cm^{-1}) is well resolved in our experiment unlike the previous measurements Longmire et al. (1980); Shenstone (1948) where only a single line was observed. The ratio of the $5p_{\frac{3}{2}} \leftarrow 6s_{\frac{1}{2}}$ and $5p_{\frac{1}{2}} \leftarrow 6s_{\frac{1}{2}}$ transition intensities is close to theoretical nonrelativistic value 2:1.

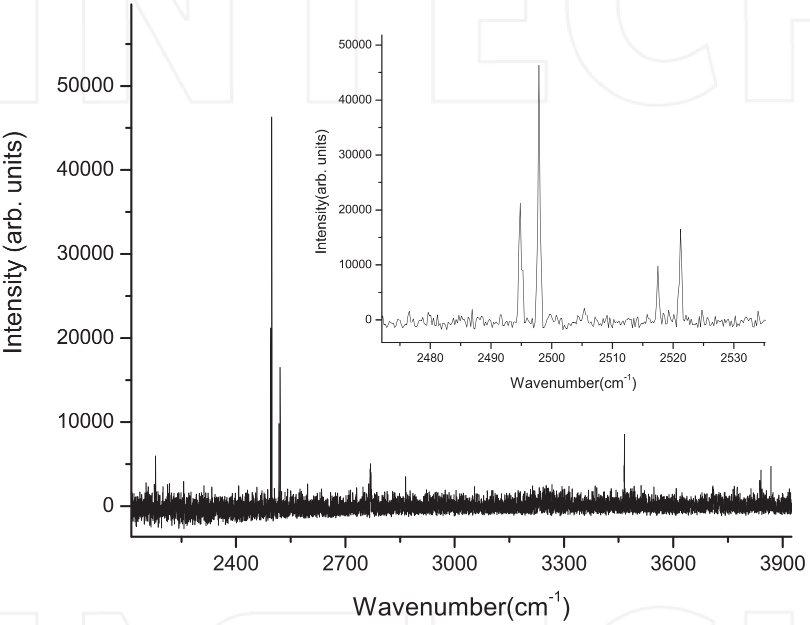


Fig. 23. Some parts of the observed IR emission spectra of Cu I

Wavenumber (cm^{-1})	Intensity (arb. units)	SNR	FWHM (cm^{-1})	Decay time (μs)	Identification
1887.307(6)	1.30×10^4	17.	0.083(16)	12.1(44) (b)	$6p_{\frac{3}{2}} \leftarrow 7s_{\frac{1}{2}}$
1935.313(2)	6.58×10^4	63.	0.106(5)	5.88(221) (b)	$6s_{\frac{1}{2}} \leftarrow 6p_{\frac{3}{2}}$
2163.890(16)	3.42×10^2	6.2	0.104(46)	16.2(41) (a)	$5f_{\frac{7}{2}} \leftarrow 7g_{\frac{9}{2}}$
2171.118(18)	1.64×10^2	4.6	0.074(55)	9.50(266) (a)	$5f_{\frac{5}{2}} \leftarrow 7g_{\frac{7}{2}}$
2176.426(16)	2.16×10^2	5.6	0.077(43)	12.8(34) (b)	$6d_{\frac{5}{2}} \leftarrow 7f_{\frac{7}{2}}$
2179.011(3)	7.68×10^4	58.	0.065(10)	9.06(479) (a)	$6s_{\frac{1}{2}} \leftarrow 6p_{\frac{1}{2}}$
2494.8098(3)	2.98×10^5	63.	0.038(1)	11.7(42) (b)	$4f_{\frac{5}{2}} \leftarrow 5g_{\frac{7}{2}}$
2497.7750(3)	3.77×10^5	121.	0.041(1)	11.3(36) (b)	$4f_{\frac{7}{2}} \leftarrow 5g_{\frac{9}{2}}$
2513.814(4)	4.65×10^3	8.4	0.048(13)	11.5(41) (b)	$5d_{\frac{5}{2}} \leftarrow 5f_{\frac{5}{2}}$
2517.4511(3)	1.03×10^5	78.	0.051(1)	15.0(57) (b)	$5d_{\frac{3}{2}} \leftarrow 5f_{\frac{5}{2}}$
2521.0550(3)	1.58×10^5	129.	0.049(1)	14.9(57) (b)	$5d_{\frac{5}{2}} \leftarrow 5f_{\frac{7}{2}}$
2865.233(2)	2.32×10^4	37.	0.062(5)	7.93(167) (b)	$6p_{\frac{1}{2}} \leftarrow 6d_{\frac{3}{2}}$
3110.955(4)	2.06×10^4	27	0.087(11)	8.53(214) (b)	$6p_{\frac{3}{2}} \leftarrow 6d_{\frac{5}{2}}$
3465.481(4)	8.46×10^4	10.	0.089(13)	5.06(87) (b)	$5p_{\frac{1}{2}} \leftarrow 6s_{\frac{1}{2}}$
3465.8044(7)	1.70×10^5	23.	0.048(2)	5.25(90) (b)	$5p_{\frac{3}{2}} \leftarrow 6s_{\frac{1}{2}}$
3837.402(12)	7.49×10^3	9.5	0.193(36)	8.51(209) (a)	$4f_{\frac{5}{2}} \leftarrow 6g_{\frac{7}{2}}$
3840.376(15)	9.64×10^3	11.	0.235(47)	14.7(190) (a)	$4f_{\frac{7}{2}} \leftarrow 6g_{\frac{9}{2}}$

Table 6. Experimental Cu I lines and their identification. The decay time, τ, was calculated by exponential fitting of the measured time profiles of the corresponding lines. See the caption to Table 4

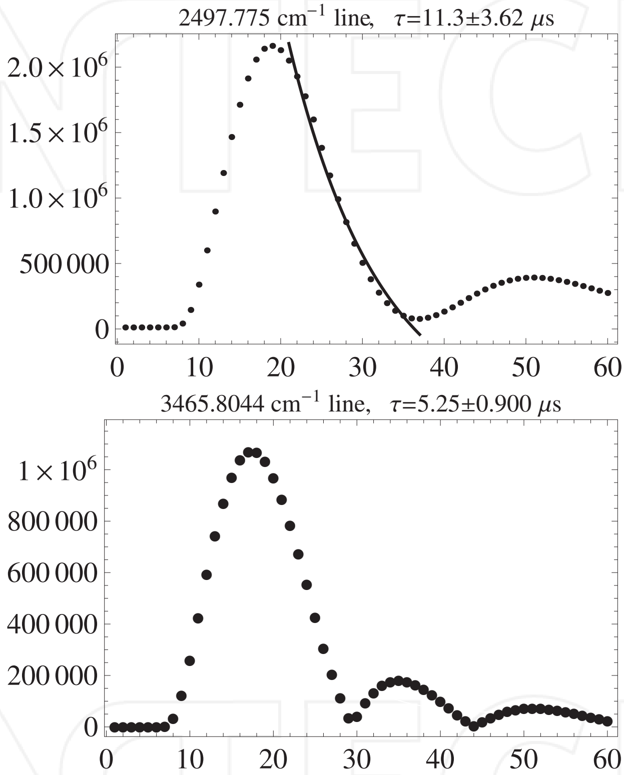


Fig. 24. The time profiles of some observed Cu I lines (dots) and their fit with exponential decay (solid curves)

Term	Energy (cm^{-1})	Other sources
$(3d^{10})5p_{\frac{1}{2}}$	49383.263(23)	49383.26 Shenstone (1948)
$(3d^{10})5p_{\frac{3}{2}}$	49382.949(14)	49382.95 Shenstone (1948)
$(3d^{10})5d_{\frac{3}{2}}$	55387.621(11)	55387.668 Shenstone (1948)
$(3d^{10})5d_{\frac{5}{2}}$	55390.569(9)	55391.292 Shenstone (1948)
$(3d^{10})5f_{\frac{5}{2}}$	57905.041(14)	57905.2 Shenstone (1948), 57905.23 Longmire et al. (1980)
$(3d^{10})5f_{\frac{7}{2}}$	57911.090(12)	57908.7 Shenstone (1948)
$(3d^{10})5g_{\frac{7}{2}}$	57924.610(30)	This work
$(3d^{10})5g_{\frac{9}{2}}$	57924.075(30)	This work
$(3d^{10})6s_{\frac{1}{2}}$	52848.752(9)	52848.749 Shenstone (1948)
$(3d^{10})6p_{\frac{1}{2}}$	55027.763(26)	55027.74 Shenstone (1948), 55027.713 Longmire et al. (1980)
$(3d^{10})6p_{\frac{3}{2}}$	54784.081(21)	54784.06 Shenstone (1948), 54784.073 Longmire et al. (1980)
$(3d^{10})6d_{\frac{3}{2}}$	57893.028(24)	57893.05 Shenstone (1948)
$(3d^{10})6d_{\frac{5}{2}}$	57895.084(24)	57895.1 Shenstone (1948)
$(3d^{10})6g_{\frac{7}{2}}$	57267.202(33)	This work
$(3d^{10})6g_{\frac{9}{2}}$	57266.676(34)	This work
$(3d^{10})7f_{\frac{7}{2}}$	60071.510(30)	This work
$(3d^{10})7g_{\frac{7}{2}}$	60076.159(23)	This work
$(3d^{10})7g_{\frac{9}{2}}$	60074.980(20)	This work

Table 7. Revised values of some levels of Cu I

5. Conclusion

The longstanding interests of our laboratory are the spectroscopic investigations of molecular ions, radicals or atoms which play a fundamental role in many plasma chemical processes, as well as in the reactions taking place inside of interstellar clouds or in stellar envelopes of giant stars of our universe.

The presented report is focused on the development and application of a time resolved system based on commercially available continuously scanning high resolution interferometer and its modification for time resolved Fourier transform spectroscopy.

The use of time-resolved FT spectroscopy opens new pathways and new points of view in study of the formation and decay processes inside the discharge or laser plasma.

Here, we are able to study the individual processes using atomic or molecular lines in a very wide spectral range of the high resolution FT technology, which has been simultaneously extended into the time dimension.

The only limitation is the sensitivity of the infrared FT technique, together with the considerable time required for the acquisition of the spectral data.

6. References

Aragon, C. & Aguilera, J. A. (2008). Characterization of laser induced plasmas by optical emission spectroscopy: A review of experiments and methods, *Spectrochim. Acta, Part B* **63**(9): 893–916.

Babánková, D., Civiš, S. & Juha, L. (2006). Chemical consequences of laser-induced breakdown in molecular gases, *Progress in Quantum Electronics* **30**(2-3): 75 – 88.

Barthélemy, O., Margot, J., Chaker, M., Sabsabi, M., Vidal, F., Johnston, T. W., Laville, S. & Drogoff, B. L. (2005). Influence of the laser parameters on the space and time characteristics of an aluminum laser-induced plasma, *Spectrochim. Acta, Part B* **60**(7-8): 905 – 914. 3rd International Conference on Laser Induced Plasma Spectroscopy and Applications (LIBS04), Torremolinos, SPAIN, SEP 28-OCT 01, 2004.

Baskakov, O. I., Civiš, S. & Kawaguchi, K. (2005). High resolution emission fourier transform infrared spectra of the 4p-5s and 5p-6s bands of arh, *The Journal of Chemical Physics* **122**(11): 114314.

Bengtsson, G. J., Jönsson, P., Larsson, J. & Svanberg, S. (1991). Time-resolved spectroscopic studies of the $7p\ ^2P$ states of neutral silver following vuv excitation, *ZPD* **22**(1): 437–439.

Bengtsson, J., Larsson, J. & Svanberg, S. (1990). Hyperfine structure and radiative-lifetime determination for the $4d^{10}6p\ ^2P$ states of neutral silver using pulsed laser spectroscopy, *Phys. Rev. A* **42**(9): 5457–5463.

Berg, P. A. & Sloan, J. J. (1993). Compact standalone data acquisition system for submicrosecond time-resolved fourier transform spectroscopy, *Rev. Sci. Instrum.* **64**(9): 2508–2514.

Bernath, P. F. & McLeod, S. (2001). DiRef, a database of references associated with the spectra of diatomic molecules, *J. Mol. Spectrosc.* **207**(2): 287.

Brown, C. M. & Ginter, M. L. (1977). Absorption spectrum of Ag I between 1540 and 1850 Å, *J. Opt. Soc. Am.* **67**(10): 1323–1327.

Brown, C. M. & Ginter, M. L. (1978). Absorption spectrum of Au I between 1300 and 1900 Å, *J. Opt. Soc. Am.* **68**(2): 243–246.

Cerny, D., Bacis, R., Guelachvili, G. & Roux, F. (1978). Extensive analysis of the red system of the CN molecule with a high resolution fourier spectrometer, *J. Mol. Spectrosc.* **73**(1): 154 – 167.

Chalmers, J. & Griffiths, P. (eds) (2002). *Handbook of Vibrational Spectroscopy (5 Volume Set)*, 1 edn, Wiley.

Civiš, S., Kubát, P., Nishida, S. & Kawaguchi, K. (2006). Time-resolved fourier transform infrared emission spectroscopy of H_3^+ molecular ion, *Chem. Phys. Lett.* **418**(4-6): 448–453.

Civiš, S., Matulková, I., Cihelka, J., Kawaguchi, K., Chernov, V. E. & Buslov, E. Y. (2010). Time-resolved fourier-transform infrared emission spectroscopy of Au in the 1800–4000-cm^{-1} region: Rydberg transitions, *Phys. Rev. A* **81**(1): 012510.

Civiš, S., Šedivcová-Uhliková, T., Kubelík, P. & Kawaguchi, K. (2008). Time-resolved fourier transform emission spectroscopy of $A^2\Pi$–$X^2\Sigma^+$ infrared transition of the CN radical, *Journal of Molecular Spectroscopy* **250**(1): 20 – 26.

Civiš, S., Matulková, I., Cihelka, J., Kubelík, P., Kawaguchi, K. & Chernov, V. E. (2010). Time-resolved fourier-transform infrared emission spectroscopy of Ag in the (1300–3600)-cm^{-1} region: Transitions involving f and g states and oscillator strengths, *Phys. Rev. A* **82**(2): 022502.

Claeyssens, F., Henley, S. J. & Ashfold, M. N. R. (2003). Comparison of the ablation plumes arising from arf laser ablation of graphite, silicon, copper, and aluminum in vacuum, *J. Appl. Phys.* **94**(4): 2203–2211.

Claeyssens, F., Lade, R. J., Rosser, K. N. & Ashfold, M. N. R. (2001). Investigations of the plume accompanying pulsed ultraviolet laser ablation of graphite in vacuum, *J. Appl. Phys.* **89**(1): 697–709.

Davies, P. B., Guest, M. A. & Stickland, R. J. (1990). Infrared laser spectroscopy of H_2 and D_2 Rydberg states. I. Application of the polarization model, *J. Chem. Phys.* **93**(8): 5408–5416.

Ding, G. J., Shang, R. C., Chang, L. F., Wen, K. L., Hui, Q. & Chen, D. Y. (1989). Experimental study of Au atom Rydberg states, *J. Phys. B* **22**(10): 1555–1561.

Drossart, P., Maillard, J.-P., Caldwell, J., Kim, S. J., Watson, J. K. G., Majewski, W. A., Tennyson, J., Miller, S., Atreya, S. K., Clarke, J. T., Waite, J. H. & Wagener, R. (1989). Detection of H_3^+ on Jupiter, *Nature* **340**(6234): 539–541.

Durry, G. & Guelachvili, G. (1994). N_2 (B–A) time-resolved fourier transform emission spectra from a pulsed microwave discharge, *J. Mol. Spectrosc.* **168**(1): 82–91.

Dyubko, S. F., Efremov, V. A., Gerasimov, V. G. & MacAdam, K. (2005). Millimetre-wave spectroscopy of Au I Rydberg states: S, P and D terms, *J. Phys. B* **38**(8): 1107–1118.

Ehrhardt, J. C. & Davis, S. P. (1971). Precision wavelengths and energy levels in gold, *J. Opt. Soc. Am.* **61**(10): 1342–1349.

Furio, N., Ali, A., Dagdigian, P. J. & Werner, H.-J. (1989). Laser excitation of the overlapping CN B–$A(8,7)$ and B–$X(8,11)$ bands: The relative phase of the B–A and B–X transition moments, *J. Mol. Spectrosc.* **134**(1): 199 – 213.

George, S., Grays, A. & Engleman, Jr., R. (1988). Spectrum of Au I in the infrared using a fourier-transform spectrometer, *J. Opt. Soc. Am. B* **5**(7): 1500–1502.

Ginter, D. S. & Ginter, M. L. (1988). Multichannel interactions in the $(1\sigma_g)^2(1\sigma_u)ns\sigma, nd\lambda$ ($^3\Sigma_u^+$, $^3\Sigma_u^+$, $^3\Pi_u$, $^3\Pi_u$) Rydberg structures of He_2, *J. Chem. Phys.* **88**(6): 3761–3774.

Ginter, D. S., Ginter, M. L. & Brown, C. M. (1984). Multichannel interactions in the $(1\sigma_g)^2(1\sigma_u)np\lambda(^3\Pi_g,^3\Sigma_g^+)$ Rydberg structures of He_2, *J. Chem. Phys.* **81**(12): 6013–6025.

Ginter, M. L. (1965). The spectrum and structure of the He_2 molecule: Part III. characterization of the triplet states associated with the UAO's 3*s* and $2p\pi$, *J. Mol. Spectrosc.* **18**(3): 321 – 343.

Gloersen, P. & Dieke, G. H. (1965). Molecular spectra of hydrogen and helium in the infrared, *J. Mol. Spectrosc.* **16**(1): 191–204.

Gomes, A., Aubreton, A., Gonzalez, J. J. & Vacquié, S. (2004). Experimental and theoretical study of the expansion of a metallic vapour plasma produced by laser, *J. Phys. D* **37**(5): 689.

Guelachvili, G. & Rao, K. R. (1986). *Handbook of Infrared Standards: With Spectral Maps and Transition Assignments Between 3 and 2600 μm (v. 1)*, Academic Press.

Hepner, G.; Herman, L. (1956). Nouveau système de bandes d'émission de la molécule He_2 vers 4700 cm^{-1}, *C. R. Acad. Sci. Paris, Ser. B* **243**(20): 1504–1506.

Herbst, E. & Klemperer, W. (1973). The formation and depletion of molecules in dense interstellar clouds, *Astrophys. J.* **185**: 505–534.

Herzberg, G. & Jungen, C. (1986). The $4f$ states of He_2: A new spectrum of He_2 in the near infrared, *J. Chem. Phys.* **84**(3): 1181–1192.

Horká, V., Civiš, S., Špirko, V. & Kawaguchi, K. (2004). The infrared spectrum of CN in its ground electronic state, *Collection of Czechoslovak Chemical Communications* **69**(1): 73–89.

Hosaki, Y., Civiš, S. & Kawaguchi, K. (2004). Time-resolved Fourier transform infrared emission spectroscopy of He_2 produced by a pulsed discharge, *Chem. Phys. Lett.* **383**(3-4): 256–260.

Huber, K. & Herzberg, G. (1979). *Molecular Spectra and Molecular Structure. IV. Constants of Diatomic Molecules.*, 1 edn, Van Nostrand.

Huntress, Jr., W. T. (1977). Laboratory studies of bimolecular reactions of positive ions in interstellar clouds, in comets, and in planetary atmospheres of reducing composition, *Astrophys. J., Suppl. Ser.* **33**(4): 495–514.

Jannitti, E., Cantù, A. M., Grisendi, T., Pettini, M. & Tozzi, G. P. (1979). Absorption spectrum of Au I in the vacuum ultraviolet, *Phys. Scr.* **20**(2): 156–162.

Kao, L., Oka, T., Miller, S. & Tennyson, J. (1991). A table of astronomically important ro-vibrational transitions for the H_3^+ molecular ion, *Astrophys. J., Suppl. Ser.* **77**(2): 317–329.

Kawaguchi, K., Baskakov, O., Hosaki, Y., Hama, Y. & Kugimiya, C. (2003). Time-resolved Fourier transform spectroscopy of pulsed discharge products, *Chem. Phys. Lett.* **369**(3-4): 293–298.

Kawaguchi, K., Hama, Y. & Nishida, S. (2005). Time-resolved Fourier transform infrared spectroscopy: Application to pulsed discharges, *J. Mol. Spectrosc.* **232**(1): 1–13.

Kawaguchi, K., Sanechika, N., Nishimura, Y., Fujimori, R., Oka, T. N., Hirahara, Y., Jaman, A. & Civiš, S. (2008). Time-resolved fourier transform infrared emission spectroscopy of laser ablation products, *Chem. Phys. Lett.* **463**(1–3): 38–41.

Kotlar, A. J., Field, R. W., Steinfeld, J. I. & Coxon, J. A. (1980). Analysis of perturbations in the $A^2\Pi\text{˘}X^2\Sigma^+$ "red" system of CN, *J. Mol. Spectrosc.* **80**(1): 86 – 108.

Lee, W.-B., Wu, J.-Y., Lee, Y.-I. & Sneddon, J. (2004). Recent applications of laser-induced breakdown spectrometry: A review of material approaches, *Appl. Spectrosc. Rev.* **39**(1): 27–97.

Lindsay, C. M. & McCall, B. J. (2001). Comprehensive evaluation and compilation of H_3^+ spectroscopy, *J. Mol. Spectrosc.* **210**(1): 60 – 83.

Longmire, M. S., Brown, C. M. & Ginter, M. L. (1980). Absorption spectrum of Cu I between 1570 Å and 2500 Å, *J. Opt. Soc. Am.* **70**(4): 423–429.

Majewski, W. A., Marshall, M. D., McKellar, A. R. W., Johns, J. W. C. & Watson, J. K. G. (1987). Higher rotational lines in the ν_2 fundamental of the H_3^+ molecular ion, *J. Mol. Spectrosc.* **122**(2): 341 – 355.

Majewski, W. A., McKellar, A. R. W., Sadovskii, D. & Watson, J. K. G. (1994). New observations and analysis of the infrared vibrationŰrotation spectrum of H_3^+, *Can. J. Phys.* **72**(11-12): 1016–1027.

Mantz, A. W. (1976). Infrared multiplexed studies of transient species, *Appl. Spectrosc.* **30**(4): 459–461.

Masutani, K. (2002). *Time-resolved Mid-infrared Spectrometry Using an Asynchronous Fourier Transform Infrared Spectrometer*, Vol. 1 of Chalmers & Griffiths (2002), 1 edn, pp. 655–665.

McCall, B. J., Geballe, T. R., Hinkle, K. H. & Oka, T. (1998). Detection of H_3^+ in the Diffuse Interstellar Medium Toward Cygnus OB2 No. 12, *Science* **279**(5358): 1910–1913.

Miotello, A. & Kelly, R. (1999). Laser-induced phase explosion: new physical problems when a condensed phase approaches the thermodynamic critical temperature, *Appl. Phys. A* **69**(Suppl. S): S67–S73. 5th International Conference on Laser Ablation COLA'99, GOTTINGEN, GERMANY, JUL 19-23, 1999.

Motret, O., Pouvesle, J. M. & Stevefelt, J. (1985). Spectroscopic study of the afterglow excited by intense electrical discharges in high-pressure helium hydrogen mixtures, *J. Chem. Phys.* **83**(3): 1095–1100.

Nakanaga, T., Ito, F. & Takeo, H. (1993). Time-resolved high-resolution ftir absorption spectroscopy in a pulsed discharge, *Chem. Phys. Lett.* **206**(1-4): 73 – 76.

Oka, T. (1980). Observation of the infrared spectrum of H_3^+, *Phys. Rev. Lett.* **45**(7): 531–534.

Orth, F. B. & Ginter, M. L. (1976). The spectrum and structure of the He_2 molecule: Characterization of the triplet states associated with the UAO's $4p\sigma$, $5p\sigma$, and $6p\sigma$, *J. Mol. Spectrosc.* **61**(2): 282 – 288.

Pickering, J. C. & Zilio, V. (2001). New accurate data for the spectrum of neutral silver, *Eur. Phys. J. D* **13**(2): 181–185.

Plašil, R., Glošik, J., Poterya, V., Kudrna, P., Rusz, J., Tichý, M. & Pysanenko, A. (2002). Advanced integrated stationary afterglow method for experimental study of recombination of processes of H_3^+ and D_3^+ ions with electrons, *Int. J. Mass Spectrom.* **218**(2): 105 – 130.

Platt, J. R. & Sawyer, R. A. (1941). New classifications in the spectra of Au I and Au II, *Phys. Rev.* **60**(12): 866–876.

Prasad, C. V. V. & Bernath, P. F. (1992). Fourier transform jet-emission spectroscopy of the $A^2\Pi_i$–$X^2\Sigma^+$ transition of CN, *J. Mol. Spectrosc.* **156**(2): 327 – 340.

Radziemski, L. J. & Cremers, D. A. (eds) (1989). *Lasers-induced Plasmas and Applications*, Marcel Dekker, New York. Chap. 7, pp. 295–325.

Ralchenko, Y., Kramida, A., Reader, J. & Team, N. (2008). NIST atomic spectra database (version 3.1.5).

Ram, R., Davis, S., Wallace, L., Engleman, R., Appadoo, D. & Bernath, P. (2006). Fourier transform emission spectroscopy of the $B^2\Sigma^+–X^2\Sigma^+$ system of CN, *J. Mol. Spectrosc.* **237**(2): 225 – 231.

Rehfuss, B. D., Suh, M.-H., Miller, T. A. & Bondybey, V. E. (1992). Fourier transform UV, visible, and infrared spectra of supersonically cooled CN radical, *J. Mol. Spectrosc.* **151**(2): 437 – 458.

Rödig, C. & Siebert, F. (2002). *Fast Time-resolved Mid-infrared Spectroscopy Using an Interferometer*, Vol. 1 of Chalmers & Griffiths (2002), 1 edn, pp. 625–Ű640.

Rogers, S. A., Brazier, C. R., Bernath, P. F. & Brault, J. W. (1988). Fourier transform emission spectroscopy of the $b^3\Pi_g–a^3\Sigma_u^+$ transition of He_2, *Mol. Phys.* **63**(5): 901–908.

Rossa, M., Rinaldi, C. A. & Ferrero, J. C. (2009). Internal state populations and velocity distributions of monatomic species ejected after the 1064 nm laser irradiation of barium, *J. Appl. Phys.* **105**(6): 063306.

Rubahn, H.-G. (1999). *Laser Applications in Surface Science and Technology*, Wiley, New York.

Safronova, U. I., Savukov, I. M., Safronova, M. S. & Johnson, W. R. (2003). Third-order relativistic many-body calculations of energies and lifetimes of levels along the silver isoelectronic sequence, *Phys. Rev. A* **68**(6): 062505.

Sharp, C. M. (1984). The computation of Franck–Condon factors, *r*-centroids and associated quantities in the electronic transitions of diatomic molecules, *Astron. Astrophys. Suppl. Ser.* **55**: 33–50.

Shenstone, A. G. (1940). The arc spectrum of silver, *Phys. Rev.* **57**(10): 894–898.

Shenstone, A. G. (1948). The first spectrum of copper (Cu I), *Philos. Trans. R. Soc. London, Ser. A* **241**(832): 297–322.

Smith, G. D. & Palmer, R. A. (2002). *Instrumental Aspects of Time-resolved Spectra Generated Using Step-scan Interferometers*, Vol. 1 of Chalmers & Griffiths (2002), 1 edn, pp. 641–Ű654.

Solka, H., Zimmermann, W., Stahn, A., Reinert, D. & Urban, W. (1987). Observation of the $B^3\Pi_u–A^3\Sigma_g^+$ band of He_2, *Mol. Phys.* **60**(5): 1179–1182.

Strasser, D., Lammich, L., Kreckel, H., Krohn, S., Lange, M., Naaman, A., Schwalm, D., Wolf, A. & Zajfman, D. (2002). Breakup dynamics and the isotope effect in H_3^+ and D_3^+ dissociative recombination, *Phys. Rev. A* **66**(3): 032719.

Vertes, A., Dreyfus, R. W. & Platt, D. E. (1994). Modeling the thermal-to-plasma transitions for Cu photoablation, *IBM J. Res. Dev.* **38**(1): 3–10.

Vervloet, M. & Watson, J. K. (2003). Improved infrared and visible emission spectra of the H_3 and D_3 molecules, *Journal of Molecular Spectroscopy* **217**(2): 255 – 277.

Zhankui, J., Jönsson, P., Larsson, J. & Svanberg, S. (1990). Studies on radiative lifetimes in the $4d^{10}ns\ ^2S$ and $4d^{10}nd\ ^2D$ sequences of neutral silver, *Z. Phys. D* **17**(1): 1–14.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Weighting Iterative Fourier Transform Algorithm for Kinoform Implemented with Liquid-Crystal SLM

Alexander Kuzmenko[1], Pavlo Iezhov[2] and Jin-Tae Kim[3]
[1]*Institute of Applied Optics, 04053, Kyiv*
[2]*Institute of Physics, 680028, Kyiv*
[3]*Chosun University, 501-759, Gwangju*
[1,2]*Ukraine*
[3]*South Korea*

1. Introduction

One of the most important trends in digital holography is the synthesis (calculation and fabrication) of diffraction optical elements (DOEs) serving for the transformation of a given light distribution into another light distribution with the desired characteristics (Bryngdahl & Wyrowski, 1990). In the case where both the amplitude and the phase of output emission are of importance, the DOE is a digital hologram which can be binary amplitude, phase, or amplitude-phase (Lohmann & Paris, 1967; Lee, 1979; Wyrowski & Bryngdahl, 1988; Wyrowski, 1990-1991). But if we are interested only in the output emission intensity, then DOEs are synthesized as a purely phase structure of the kinoform type (Lesem et al., 1967; Hirsch et al., 1971; Gallagher & Liu, 1973; Akahori, 1986; Aagebal & Wyrowski, 1997; Skeren et al., 2002).

As distinct from an ordinary optical or digital hologram, a kinoform has a rather high diffraction efficiency attaining at least 90 per cent for a continuous kinoform. Therefore, it attracts a significant attention of experts in the applied and calculation-theoretic aspects. Among a lot of uses of the kinoform, there are particularly three interesting applications such as beam splitting (fan-out), beam shaping, and pattern or image generation. Optical fan-out elements split a single laser beam into a one- or two-dimensional array of beams and are key components in many applications of modern optics such as parallel optical processing, free-space communication in optical computing (Herzig et al., 1990; Gale et al., 1992, 1993; Ehbets et al., 1992; Prongue et al., 1992; Mait & Brenner, 1988), and fiber optic communication (Wyrowski & Zuidema, 1994). Fan-out elements with a smooth periodic phase structure have a theoretical limit by diffraction efficiency which is close to 100 % (Herzig et al., 1990). Beam shaping is most commonly used in high-energy laser applications to the processing of various materials and in the laser branding or photolithographic illumination. These applications often require minimal energy losses, implying the use of phase-only elements such as a kinoform (Dixit et al., 1994; Leger et al., 1994; Chen et al., 1994; Duparre et al., 1995; Xin Tan et al., 1995; Johansson & Bengtsson, 2000; Liu & Taghizaden, 2002). Playing the role of a generator of images, a kinoform serves for the reproduction, in the form of a light intensity distribution, of a real binary or half-tone function stored in a computer in the discrete form.

The calculation of the phase structure of a kinoform is a partial case of the solution of a phase problem in the so-called "two-intensity" statement which is formulated for a Fourier-kinoform as follows. Let the input data such as the real function of an object $f_o(x,y)$ and the modulus of some spectrum equal to $1(u,v)$ be given. It is necessary to determine such phase distributions $\varphi(x,y)$ and $\psi(u,v)$ which together with the input data form a Fourier-pair

$$f_0(x,y)exp[i\varphi(x,y)] \Leftarrow \Im^{\pm 1} \Rightarrow exp[i\psi(u,v)], \tag{1}$$

where $\Im^{+1}, \Im^{-1}$ are the direct and inverse Fourier transformation, respectively. The obtained solutions φ and ψ describe, respectively, the object-oriented phase scatterer (diffuser) and a spectral distribution of phases which is registered then on the phase medium in the form of a kinoform. We note that the solution of the phase problem for a kinoform has a specific feature. In the classical two-intensity statement of the phase problem, a solution exists always, because the true amplitude of the spectrum of an object function $f_o(x,y)$ is used in the Fourier-plane (though the determination of a solution can be not an easy task). For a kinoform, we require that the spectrum amplitude be equal to $1(u,v)$ in all the cases irrespective of the form of a function $f_o(x,y)$. In other words, we set the spectrum amplitude. Therefore, strictly speaking, the frequency-bounded phase structure $\psi(u,v)$, whose Fourier-transformation will form the given distribution of intensities $f_o(x,y)|^2$, should not obligatorily exist. Nevertheless, an approximate (and sufficiently exact) solution of the kinoform problem exists practically always, which is supported by the practice of calculations.

Many algorithms of solution of phase problems are available. Prior to the beginning of the 1970s, the solution of inverse problems (which include the phase problem as well) was mainly a prerogative of professional mathematicians, because it requires to use a complicated mathematical apparatus and to construct high-complexity calculation's algorithms (Tikhonov & Arsenin, 1977; Inverse Source Problems in Optics, 1978). The situation was changed, when the mathematically simple and physically transparent projective iterative Fourier-transform (IFT) algorithms were developed in the 1970-1980s (Lesem et al., 1967; Gerchberg & Saxton, 1972; Gallagher & Liu, 1973; Fienup, 1980, 1982), and after the clarification of the mathematical nature of these algorithms (Youla & Webb, 1982; Levi & Stark, 1987; Catino et al., 1997). The family of IFT-algorithms follows the philosophy of the Gerchberg-Saxton algorithm known as the error-reduction (ER) algorithm. All of these algorithms incorporate a similar idea – to iterate between the spatial and frequency domains, while successively satisfying a set of constraints in both. We start with an arbitrary phase-only filter in the object domain multiplying the input object (the original image). After the Fourier transformation, we obtain a Fourier domain image and set the required Fourier intensity (actually, the magnitude), leaving the phase, as it is. The inverse Fourier transformation brings us back to the object domain. Since we demand a phase-only filter, we impose the intensity of the input object in this plane. Then we calculate the Fourier transform and return to the Fourier domain, and so on. Earlier and now, this simple efficient idea of ER-iterations gives a conceptual basis for the development of most iterative methods of solution of phase problems in such fields as coherent optics, optical astronomy, electron and X-ray microscopy, biophysics, etc., where one deals with the diffraction of a coherent emission.

In Section 2 we describe algorithms we have tested. Section 3 presents the results of computer simulations, and Section 4 gives the results of optical experiments. The final section is devoted to the conclusion of this chapter.

2. Algorithms

2.1 Weighting IFT-algorithm

It is known that the phase calculation for a kinoform with the help of the IFT-algorithm in its classical ER-version described above gives no satisfactory results. The algorithm's convergence stagnates rapidly, and the mean square error in a reconstructed image remains large. In the present chapter, we will discuss a modified IFT-algorithm of synthesis of a kinoform (Kuzmenko, 2006, 2008), whose principal single distinction from the classical ER-algorithm consists in the use of a new nonlinear operation of the processing of the field amplitude in the object plane. To clarify its application, the work of the algorithm is illustrated in Fig. 1.

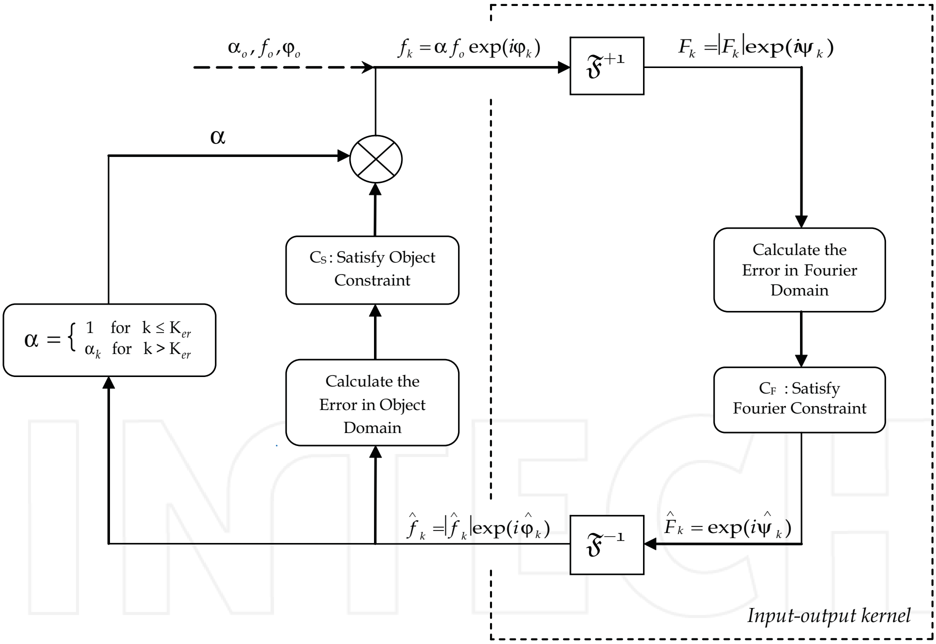

α_o, f_o, φ_o
$f_k = \alpha f_o \exp(i\varphi_k)$
$\mathfrak{F}^{+1}$
$F_k = |F_k| \exp(i\psi_k)$
α
Cs: Satisfy Object Constraint
Calculate the Error in Fourier Domain
$\alpha = \begin{cases} 1 & \text{for } k \le K_{er} \\ \alpha_k & \text{for } k > K_{er} \end{cases}$
Calculate the Error in Object Domain
CF : Satisfy Fourier Constraint
$\hat{f}_k = |\hat{f}_k| \exp(i\hat{\varphi}_k)$
$\mathfrak{F}^{-1}$
$\hat{F}_k = \exp(i\hat{\psi}_k)$
Input-output kernel

Fig. 1. Weighting IFT algorithm

First, one or several iterations (K_{er}) are realized by the classical ER scheme, in which the function f_o is the ideal object, and φ_o is an input phase scatterer imposed on it. Then, in all iterations with $k > K_{er}$ at the formation of an input, the amplitude f_o will be replaced by a new amplitude defined as

$$f_k = \alpha_k f_o \tag{2}$$

where the weight coefficients α_k are determined by the recurrence relation

$$\alpha_k = \alpha_{k-1}\beta_{k-1}, \qquad (k > 1), \tag{3}$$

where

$$\beta_{k-1} = f_o / \left[|\hat{f}_{k-1}| + \varepsilon \right]. \tag{4}$$

Here, $|\hat{f}_{k-1}|$ is the reconstructed amplitude on the $(k-1)-th$ iteration, and ε is a small number ~ 10^{-10}, excluding the division by zero. It should be mentioned that f_o is real. The processing of the phases φ_k and ψ_k remains the same as that in the classical ER-algorithm. The phase of a kinoform ψ_k can be quantized on each iteration with a required number of quantization levels in that case where it is necessary to study the quantization effects in the reconstructed image or to register a kinoform on a recording medium with a finite number of gradations of the phase.

Operations (2) - (4) are heuristic and have no strict mathematical justification. Their efficiency is established by extensive model and optical experiments. The physical sense of the coefficients α becomes clear if we consider the block of the algorithm separated by a dashed line in Fig. 1, according to Fienup (Fienup, 1980, 1982), as a nonlinear unit with the input f_k, output $\hat{f}_k$, and action operator $\Im^{+1}C_F\Im^{-1}$. Then, from the viewpoint of the theory of systems, the coefficients α is nothing but the matrix of coefficients of a negative feedback "output-input": if the amplitude $|\hat{f}_{k-1}|$ on the $(k-1)-th$ iteration at some point (x,y) of the plane of images is more than a given value f_o , then, on the next $k-th$ iteration, the input f_o at the corresponding point will be corrected. Namely, it will be decreased by $\alpha(x,y)$ times, and *vice versa*. At the same time, from the viewpoint of optics, the system of coefficients α normalized to one can be interpreted as some object-dependent amplitude filter which acts on the initial object f_o and varies in the process of iterations. It is clear that, for all ER-iterations $\alpha(x,y) = \alpha_o(x,y) = 1(x,y)$.

2.2 Input-output algorithm

In the course of experiments, we compared the weighting IFT-algorithm with the kinoform version of the Fienup input-output (IO) algorithm (Fienup, 1980). It is well-known and is one of the best at its estimation from the viewpoint of simplicity of the algorithm and the quality of a reconstructed image. Like the weighting algorithm, it differs from the ER-algorithm only by the mean of the processing of a field in the object plane. In the IO algorithm, after several preliminary ER-iterations, the input for the input-output kernel (see Fig. 1) for all subsequent iterations is taken in the form

$$f_{k+1} = f_k + \mu\{2|f_o|\exp[i\hat{\varphi}_k] - \hat{f}_k - |f_o|\exp[i\varphi_k]\}, \tag{5}$$

or

$$f_{k+1} = \hat{f}_k + \mu\{2|f_0|\exp[i\hat{\varphi}_k] - \hat{f}_k - |f_0|\exp[i\varphi_k]\} \tag{6}$$

where μ is a free parameter, whose optimum value is selected experimentally. It is close to unity for half-tone objects and is usually in the interval 1.5-3.5 for binary objects. It should be noted that, in Eqs. (5) and (6), the previous input f_k and the previous output $\hat{f}_k$ serve, respectively, as a reference for the next input f_{k+1}. The term in the braces in both relations is a correction which must turn to zero, if the algorithm converges. Later on (Fienup, 1982), these two versions of the algorithm were named the input-output (IO) and output-output (OO) algorithms.

3. Computer simulation

A number of model experiments with various objects was realized with the purpose to study the potentialities of the weighting IFT-algorithm. Analogous experiments were performed also with the use of the IO and OO algorithms. In all the cases, the same phase starting diffuser φ_0 with a uniform distribution of phases in the interval *(0-2π)* is used. The variance of the amplitudes of images reconstructed in the process of iterations was evaluated as

$$\sigma_f(k) = \frac{\sum\limits_{l,m} [(f_o)_{l,m} - \chi(k)|\hat{f}_{l,m}(k)|]^2}{\sum\limits_{l,m} (f_o)^2_{l,m}}, \tag{7}$$

where

$$\chi(k) = \frac{\sum\limits_{i,j} (f_o)^2_{i,j}}{\sum\limits_{i,j} |\hat{f}_{i,j}(k)|^2} \tag{8}$$

is the scale factor, the indices *l*, *m* and *i*, *j* run over the points, where the amplitude of an initial object f_o is nonzero, and *k* is the iteration number.

In the experiments involving the IO and OO algorithms, we used the optimum value of the object-depended coefficient μ_{opt} in the equations (5) and (6), which provides the best convergence. The value of μ_{opt} was determined by means of the cyclic repetition of the procedure of synthesis for various values of μ (from the interval 0.1-5.0 with a step of 0.1). In Figs. 2 to 6, the results of model experiments on the synthesis of the kinoforms of binary and half-tone objects with a dimension of 64×64 counts are presented.

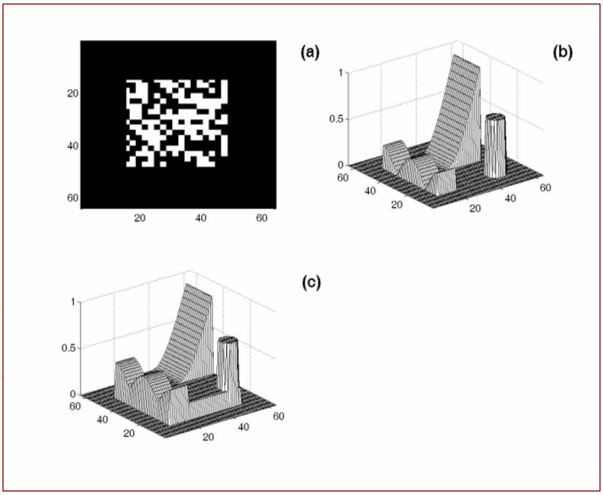


Fig. 2. Objects 64x64: (a) binary; (b), (c) half-tone without and with a base (equal to 0.17)

3.1 Binary object (the beam splitting)

In Fig. 3, we present the plots characterizing the quality of the image of a binary object (Fig. 2a) reconstructed by a kinoform. As seen from Fig. 3a, the weighting algorithm allows one to decrease the dispersal of the one-bit-intensity $\Delta I_{one-bit}$ given by the ER-algorithm practically to zero, i.e., the algorithm does not reveal the effect of stagnation for binary objects. In our example, 180 weighting-iterations reduce $\Delta I_{one-bit}$ from 0.008 to 7.6×10^{-7} (Fig. 3b), whereas 1500 such iterations result in $\Delta I_{one-bit} = 2\times10^{-12}$. At the same time, the IO algorithm (with optimized μ) "stops" at the value $\Delta I_{one-bit} = 2.5\times10^{-5}$. That is, it falls in a minimum of $\sigma_f(k)$ which is sufficiently deep, but, nevertheless, is local. We note that the ratios of the minimum one-bit-intensity to the maximum zero-bit intensity for three algorithms are equal to, respectively, 4 (ER), 4.53 (IO), and 7 (weighting algorithm). Figure 3b demonstrates the effect of a diminution of the variance $\sigma_f(k)$ at the transition from one algorithm to another one. Analogous results were obtained also for other binary objects with dimensions of 64×64 and 128×128.

3.2 Half-tone object (the image generation)

We observe a somewhat more complicated situation for half-tone objects, one of which is presented in Fig. 2b. As was shown by model experiments, the kinoforms of such objects calculated with the help of the weighting algorithm reconstruct a high-quality image only in the range of amplitudes from ~ 0.15 to 1 (at the normalization of the image to 1). The rest amplitudes are distorted to various degrees. We can reach the proper reconstruction of all amplitudes, including those close to zero, if the initial object is positioned on a pedestal (Fig. 2c), whose height is ~ 15-20% of its maximum amplitude, and if the reconstructed image amplitude (the intensity in an optical experiment) is cut off by the pedestal level. It is obvious that, in this case, the useful diffraction efficiency of a kinoform decreases. The dependences of $\sigma_f(k)$ for both compared algorithms given in Fig. 4, as well as the visual

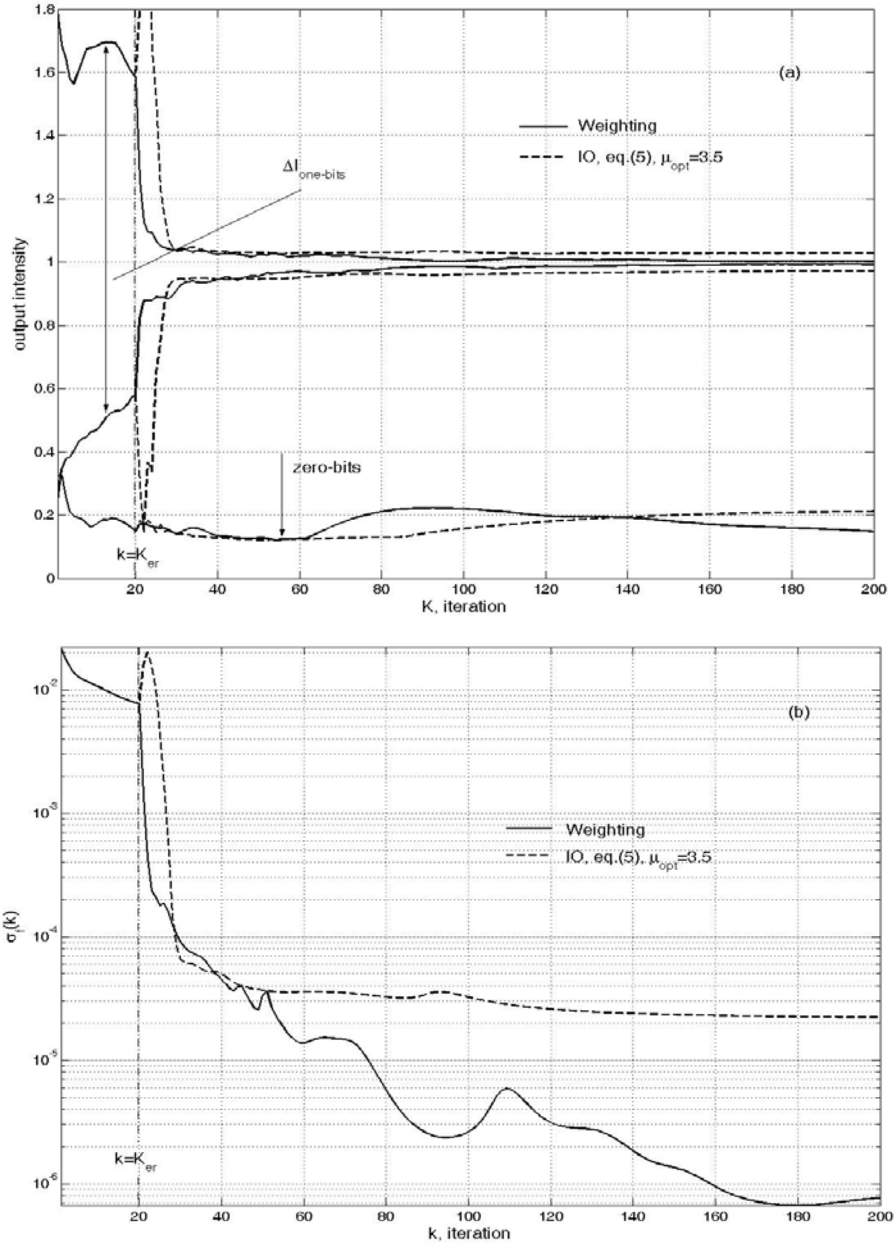


Fig. 3. The kinoform of the binary object (Fig. 2a): (a) range of output intensities, (b) variance of the amplitude of reconstructed image *vs* the iteration number

observations of reconstructed images, indicate that, in the case where a base is supplemented to an object, the weighting algorithm begins to surpass the IO algorithm in convergence after a certain number of iterations. In our example with the object in Fig. 2c, the advantage of the weighting algorithm over the IO algorithm begins to reveal itself after 100 iterations, increases with the number of iterations, and is almost four orders of magnitude by the 2000-th iteration (5.8×10^{-9} against 2.5×10^{-5} for $\sigma_f(2000)$). But if the base is absent, the IO algorithm has some advantage.

The calculated efficiencies of kinoforms (in parentheses, the values obtained within the IO algorithm are given) are as follows: 91.39 (91.25)% for the object in Fig. 2a, 94.82 (92.48)% for the object in Fig. 2b, and 96.91 (95.02)% for the object in Fig. 2c. In the course of calculations, the criticality of the weighting algorithm with respect to a value of the parameter ε in formula (4) is verified. By varying ε from 1×10^{-22} to 1×10^{-6} , the deviation of $\sigma_f(\varepsilon)$ from $\sigma_f(\varepsilon_{10})=10^{-10}$ is determined as

$$\Delta_{\sigma}=100\%[\sigma_f(\varepsilon)-\sigma_f(\varepsilon_{10})]/\sigma_f(\varepsilon_{10}) \qquad (9)$$

for various objects with the fixed number of iterations equal to 50. On the average, Δ_{σ} was (0.002-0.05)%. Thus, the variation of ε in the indicated limits did not influence practically the exactness of the calculation of a kinoform and, at the same time, excluded the situation where one should divide the numerator in formula (4) by zero.

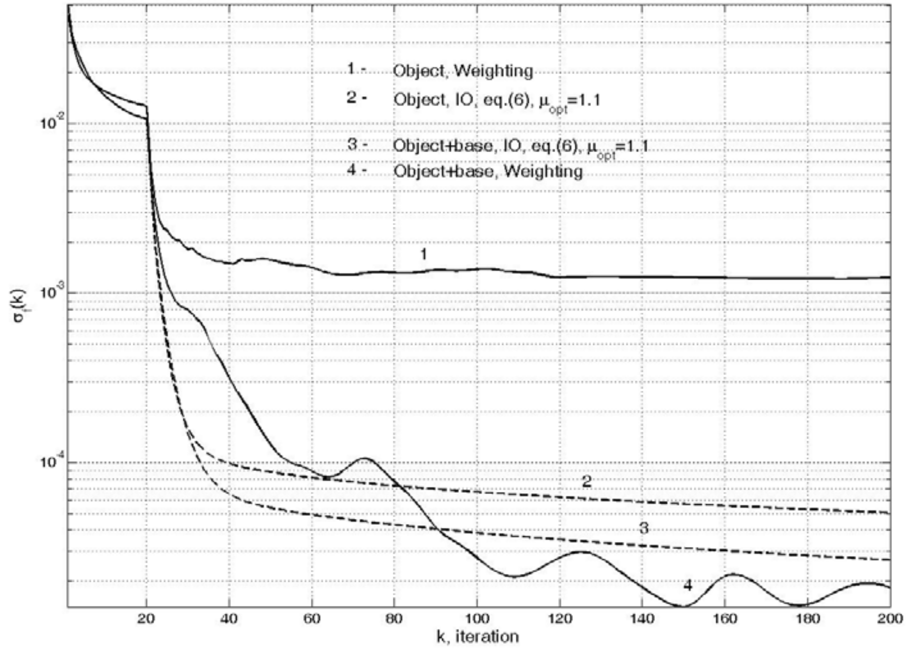


Fig. 4. Variance of the amplitude of reconstructed image for a half-tone object without and with a base (Fig. 2b, c)

3.3 Super-Gaussian (SG) beam shaping

Within the weighting and IO algorithms, the calculations of kinoforms that are the transducers of the intensity of a Gauss beam of the form $\exp[-(u^2+v^2)/2r_o^2]$ in a SG beam of the form $\exp[-(x^2/2r_o^{'2})^M-(y^2/2r_o^{'2})^M]$ are performed, where r_o and $r_o^{'}$ are the inflection radii of the Gauss curves, and M is the SG order (as known, the calculation of a kinoform involves the square root of the both indicated intensities). In Fig. 5, the input (r_o = 70) and output ($r_o^{'}$ = 25) intensity profiles for M = 4 and M = 100 with a dimension of the object 256×256 counts are presented. The iteration process with K_{er} = 10 was truncated at the 100-th iteration. With regard for the separation of the working part of a SG beam so as shown in Fig. 5, the intensity variance σ_I and the output efficiency η are as follows: $\sigma_I = 2.9\times10^{-4}$ (6.59×10^{-4}), $\eta = 96.46$ $(89.72)\%$ for M = 4; $\sigma_I = 3.7\times10^{-5}$ (1.98×10^{-4}), $\eta = 93.63$ $(91.2)\%$ for M = 100. The calculation of σ_I was performed by a form analogous to (4), but for intensities. It should be noted that, in the calculation of a kinoform-former of a SG, a special attention should be paid to a choice of r_o defining the effective width of a beam illuminating the kinoform. For small r_o (in our example, $r_o = 35$), the kinoform is illuminated by a narrow Gauss beam, which means the actual nullification of light amplitudes on the edges of the kinoform. This is equivalent to a reduction of the band of space frequencies forming a SG. As a result, the pattern of a SG will be covered by a speckle irrespective of the value of $r_o^{'}$ (see Fig. 6). In more details, the problem of restriction of the frequency band and its relation to the quality of images are considered formerly (Wyrowski and Bryngdahl, 1988).

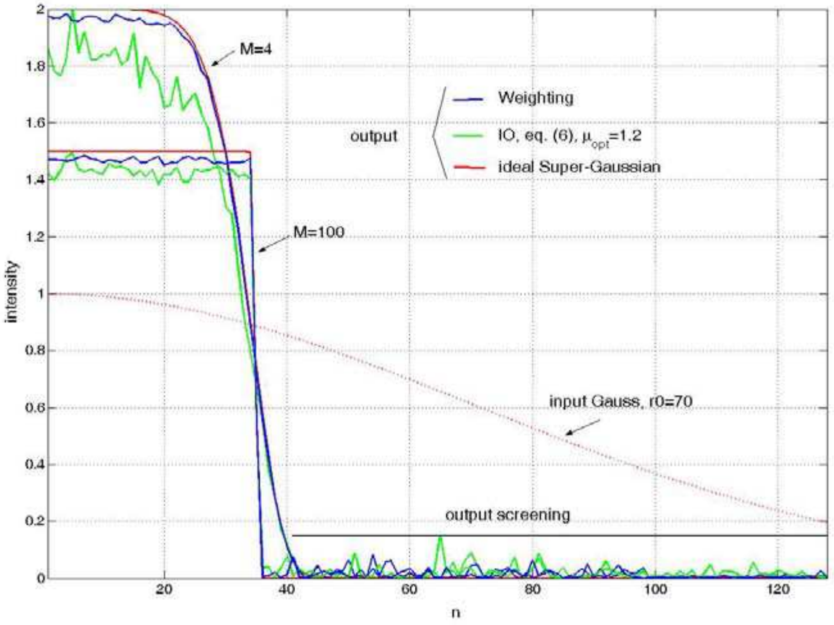


Fig. 5. The profiles of intensities of super-Gaussian beams with $r_o^{'} = 25$ of the 4th and 100th orders within the weighting and IO algorithms. Curves for M=4 and M=100 are vertically shifted up for clarity

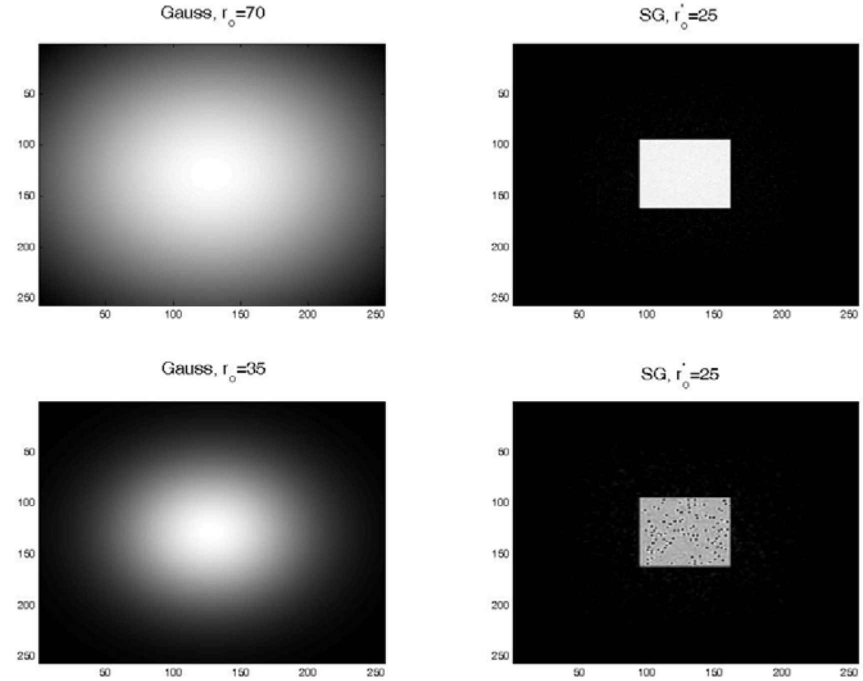


Fig. 6. Proper ($r_o = 70$) and erroneous ($r_o = 35$) choices of the effective width of an illuminating beam for the kinoform-former of a super-Gaussian ($r_o^{'} = 25$). In the second case, the cross-section of a super-Gaussian is covered by a speckle

3.4 Off-axis kinoform

Irrespective of the mean and the accuracy of calculations of the phase function ψ of a kinoform, the quality of a reconstructed image depends eventually on the accuracy of the representation of a microrelief of this phase on a recording medium. It is obvious that this accuracy depends on the technical potentialities of a registering unit and the characteristics of a recording medium. As for a programmed SLM, the accuracy is determined by its physico-technical parameters. In the synthesis of an axial Fourier-kinoform which reconstructs the image in the zero order of diffraction, the inaccuracy of the representation of the phase ψ leads to the appearance of a bright spot surrounded by noises at the center of the image. This effect can be eliminated in the single way due to a displacement of the image, as a whole, to the side from the optical axis of a reconstruction system. This can be achieved by the synthesis of a kinoform which reconstructs the image in the nonzero order. Such a kinoform is called the off-axis one.

One knows the mean of synthesis (S1) of an off-axis kinoform (Wyrowski, 1990) with the reconstruction of the image in a nonzero order with components P_x, P_y along the axes x, y of the plane of images. The components P_x and P_y are defined identically. Therefore, in what follows, we will write all relations only for the axis x . The admissible linear displacement x_o of an image from the optical axis of the Fourier-system of reconstruction in mean S1 is defined as

$$x_o = P_x L^p, \tag{10}$$

where

$$P_x \le \frac{1}{2}(1 - \frac{L^i}{L^p}), \tag{11}$$

L^i – linear sizes of the image along the axis x, and L^p – linear sizes of the diffraction order. According to (10) and (11), the value of x_o decreases with increase in L^i. At the standard values $L^i \approx L^p / 2$, we get $0 \le P_x \le 1/4$. Respectively, the image can be displaced in the interval $0 \le x_o \le L^p / 4$. From the viewpoint of practice, the principal drawbacks of mean S1 are both a small interval of admissible displacements of the image and the dependence on the image size. As a positive feature of the mean, we mention the invariance of the total number of pixels of a kinoform at the transition from the axial to off-axis variant of the synthesis.

One knows also the mean of synthesis (S2) of an off-axis kinoform (Turunen et al., 1990) with the possibility of the reconstruction of an image in the order $P_x \ne 0$ which can vary in the limits $0 \le P_x \le 1$ for the same limitations on L^i as in mean S1. Mean S2 ensures a wide interval of displacements x_o $(0 \le x_o \le L^p)$ which does not depend on the image size. However, this is attained due to the increase in the total number of pixels necessary for the registration of a kinoform by K times (practically, 2 ≤ K ≤ 8). The calculation and the registration of such kinoforms (e.g., by the methods of laser or electron lithography) are quite complicated (Turunen et al., 1990). The use of programmed SLMs (the total number of pixels is $\approx 10^3 \times 10^3$ on the average) for their representation becomes problematic already for the dimension of a kinoform of 256×256 and $K > 4$.

We propose a mean of synthesis of an off-axis kinoform which ensures a significantly greater interval of admissible displacements of the reconstructed image as compared with mean S1. In this case, we conserve the main advantage of the latter, namely the invariability of the total number of pixels of a kinoform at the transition from the axial to off-axis variant of the synthesis. The essence of the mean is simple: in order to make a kinoform to be off-axis, we have to introduce the spatial carrier frequency to it. To this end, we propose to supplement of any IFT-method of calculation of the kinoform (including weighting algorithm) by one more operation – to add the linear phase $2\pi(x_o u + y_o v)$ $(x_o, y_o \ge 0 \text{ or } \le 0)$ to the phase $\psi(u,v)$ of an on-axis kinoform at the last iteration. As a result, the calculated kinoform will reconstruct the image

$$\begin{aligned} f_{off}(x,y) &= \Im^{-1}\{\exp[i(\psi(u,v) + 2\pi(x_o u + y_o v)]\} = \\ &= f(x,y) \otimes \delta(x - x_o, y - y_o) = f(x - x_o, y - y_o). \end{aligned} \tag{12}$$

Here, $f_{off}(x,y)$ - off-axis image, $f(x,y)$ – axial image, δ - delta-function, and the symbol $\otimes$ stands for the operation of convolution. It follows from (12) that f_{off} is nothing but the axial image displaced along the axes x, y by x_o, y_o. The values of x_o, y_o (and also the order P$_x$, P$_y$) are independent of the ratio L^i/ L^p in this case, as distinct from mean S1, and can be, in principle, arbitrary. However, the optical and model experiments executed by us have shown that x_o, y_o should be chosen in the limits $0 \le x_o, y_o \le L^p / 2$, which corresponds to

$0 \le P_x, P_y \le 1/2$. Thus, the image can be displaced in the proposed mean in the interval $0 \le x_o, y_o \le L^p / 2$ which exceeds at least twice the analogous interval in mean S1. In Fig. 7a, we show the example of the reconstruction into a fractional order ($P_x = 0.45$, $P_y = 0$) (capture of 3x3 diffraction orders). The off-axis kinoform was calculated using the weighting IFT-algorithm with the introduction of a spatial carrier frequency (a deflecting grating) along the axis x. In Fig. 7b, we give a fragment of the cross-section of a given grating. An analogous complicated structure of the grating is observed for the remaining values of P_x, P_y except for P_x, P_y=0.25 and 0.5 (for them, the grating periods are, respectively, *0, π/2, π, 3π/2,* and *0, π*).

It is obvious that, in order that such complicated grating have a sufficient number of periods T_g of oscillations in the structure of a kinoform (and thus could manifest the deflecting properties), the kinoform format must be sufficiently great (≥500×500 pixels). It is worth noting that the diffraction efficiency of an off-axis kinoform decreases with increase in a displacement of the image. In Section 4, we present the quantitative results of measurements of the diffraction efficiency of off-axis kinoforms.

3.5 Iterative quantization

For practical reasons, the phase structure of a kinoform $\psi(u,v)$ is usually quantized. This simplifies the production step. We investigated both the direct iterative quantization and the stepwise (soft) iterative quantization. The former is given by the standard operator

$$\bar{\psi}(u,v) = \begin{cases} 0, & \psi(u,v) \le 0.5\Delta\psi \\ \vdots & \vdots \\ m\Delta\psi, & (m-0.5)\Delta\psi \le \psi(u,v) \le (m+0.5)\Delta\psi \\ \vdots & \vdots \\ 2\pi, & \psi(u,v) \ge (M-0.5)\Delta\psi \end{cases} \tag{13}$$

and the latter is presented by the operator (Wyrowski, 1990)

$$\bar{\psi}(u,v) = \begin{cases} 0, & \psi(u,v) \le 0.5\varepsilon^{(p)}\Delta\psi \\ \vdots & \vdots \\ m\Delta\psi, & (m-0.5\varepsilon^{(p)})\Delta\psi \le \psi(u,v) \le (m+0.5\varepsilon^{(p)})\Delta\psi \\ \vdots & \vdots \\ 2\pi, & \psi(u,v) \ge (M-0.5\varepsilon^{(p)})\Delta\psi \\ \psi(u,v), & \text{otherwise} \end{cases} \tag{14}$$

where

$$\begin{aligned} & m = 0,1,\ \cdots, M,\ \Delta\psi = 2\pi/M,\ M \text{ - number of quantization levels,} \\ & 0 < \varepsilon^{(1)} < \varepsilon^{(2)} < \ \cdots\ \varepsilon^{(p)}\ \cdots\ < \varepsilon^{(P)} = 1, \\ & p = 1,2,\ \cdots\ P,\ P - \text{number of stages of quantization.} \end{aligned} \tag{15}$$

The principle of stepwise quantization consists in the following. The whole process of iteration is divided into P cycles, each of which (except for the last one) includes Q iterations. In the course of the implementation of a cycle, only

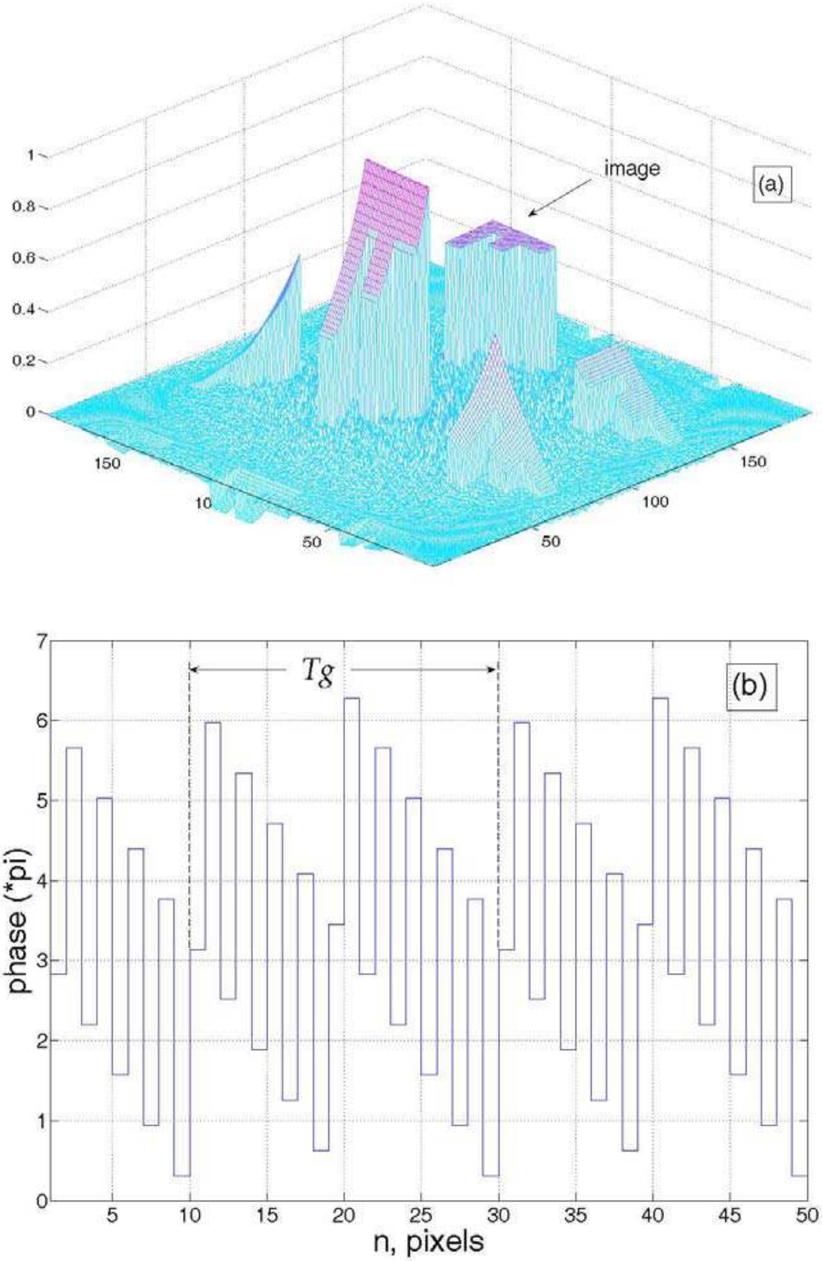


Fig. 7. (a) reconstruction into the order $P_x = 0.45$, $P_y = 0$; (b) fragment of the spatial carrier frequency

a part of the phases of a kinoform, which belong to the interval $\varepsilon^{(p)}\Delta\psi < \Delta\psi$ is quantized, rather than all phases falling in the interval $\Delta\psi$ (relate to some *m*-level of quantization). The remaining phases remain invariable. The quantity *ε* increases with the index *p*. Respectively, the interval of quantized phases is extended, by attaining eventually a value close to $\Delta\psi$. At the last step (*P*), the direct quantization operator is reached, and only one iteration is performed. Thus, the total number of iterations *K=Q(P-1)+1*. Values of *P* and *Q* can be, in principle, arbitrary, as well as the values of elements of the sequence $\varepsilon^{(p)}$. Their optimization is attained experimentally. Some versions of the choice of *Q*, *P* and *ε* for *K=const* were considered by Skeren et al. (2002). But we used, in our experiments with binary objects, the collection of values

$$\begin{aligned} &P = 10, \\ &\varepsilon^{(1)} = 0.3,\ \varepsilon^{(2)} = 0.5,\ \varepsilon^{(3)} = 0.6,\ \varepsilon^{(4)} = 0.7,\ \varepsilon^{(5)} = 0.75, \\ &\varepsilon^{(6)} = 0.8,\ \varepsilon^{(7)} = 0.85,\ \varepsilon^{(8)} = 0.9,\ \varepsilon^{(9)} = 0.95,\ \varepsilon^{(10)} = 1, \end{aligned} \tag{16}$$

which was proposed and approved by Wyrowski (1990) and allows one to improve, in dependence on the number of quantization levels and the form of an object, the signal/noise ratio for the reconstructed image by 3 to 10 times as compared with that for the direct quantization.

4. Experiment

4.1 Optical-digital system

Model experiments (see Section 3) have shown the high efficiency of the weighting IFT-algorithm just for binary objects. Therefore, we investigated kinoforms acting as a beam splitter and the generation of patterns of binary objects. A typical optical-digital Fourier system (Fig. 8) with a He-Ne laser (543 nm) is used to investigate the kinoform reconstruction characteristics. Here, we use a reflection-type phase-only SLM HEO 1080 Pluto produced by the HOLOEYE Inc. The reconstructed images were recorded and processed with the use of a SP620-USB CCD-camera of Spiricon Inc. with a high dynamic range.

As known (Oton et al., 2007), the spatial calibration of reflective LCoS SLM is essential for the correct use of the modulator in applications with high requirements of the wavefront control. We determined the additional phase 2D-distribution compensating the distortions of a wave front which appear due to the imperfection of SLM (backplane curvature, thickness variations of the liquid crystal layer across the aperture of the SLM, and so on) and elements of the optical system, by using the interference-based method proposed by Oton et al. (2007). At the implementation of experiments, we sum the obtained distribution with the calculated phase of a kinoform. This allowed us to exclude, to a significant degree, the hardware-based effect at the measurement of characteristics of reconstructed images. The size of objects and kinoforms was 1000x1000 pixels in format. As objects, we took letter F and a 14 × 14 two-dimensional one-bits array occupying, respectively, 250 × 150 and 200 × 200 counts of the input plane which are used in the study of the output intensities, diffraction efficiency, and effects of quantization of the kinoform phase in the reconstructed images.

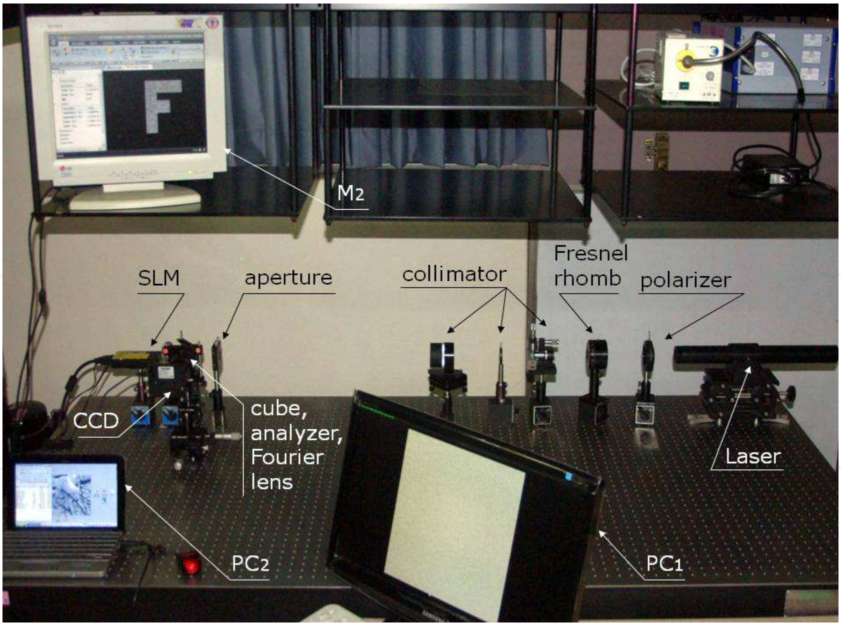


Fig. 8. Optical-digital Fourier system. Notations: SLM - SLM HEO 1080 PLUTO, CCD - SPU620 CCD with BeamGage software, PC1, PC2+M2 – computers for the control, respectively, SLM and CCD-camera, Laser - He-Ne laser (543 nm)

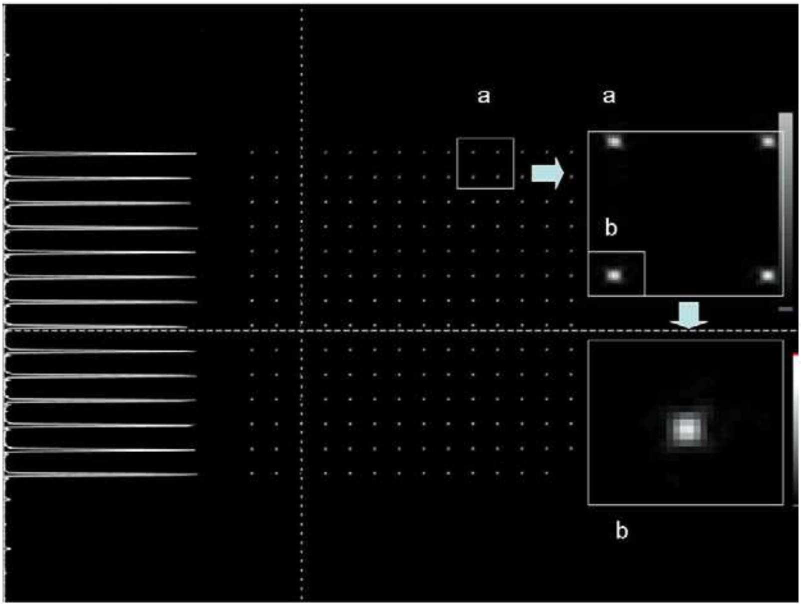


Fig. 9. Kinoform as a beam-splitter: reconstructed image of 14 × 14 spots with the intensity profile (third column) and the structure (a, b) of spots

4.2 Kinoform as a beam-splitter

In Fig. 9, we present the results of experiments with the use of a kinoform for the two-dimensional spot array generation. In calculations, we applied the stepwise quantization with parameters given in relation (16), the number of quantization levels M=256, and the ratio of iterations ER/weighting = 10/15. The measured profile of intensities demonstrates a high homogeneity of light spots. Each spot of an image was registered by an area including 9×9 pixels of a CCD-camera, which allows us to control the regularity of arrangement of intensity maxima of spots in the output plane. No deviations from the regularity were observed.

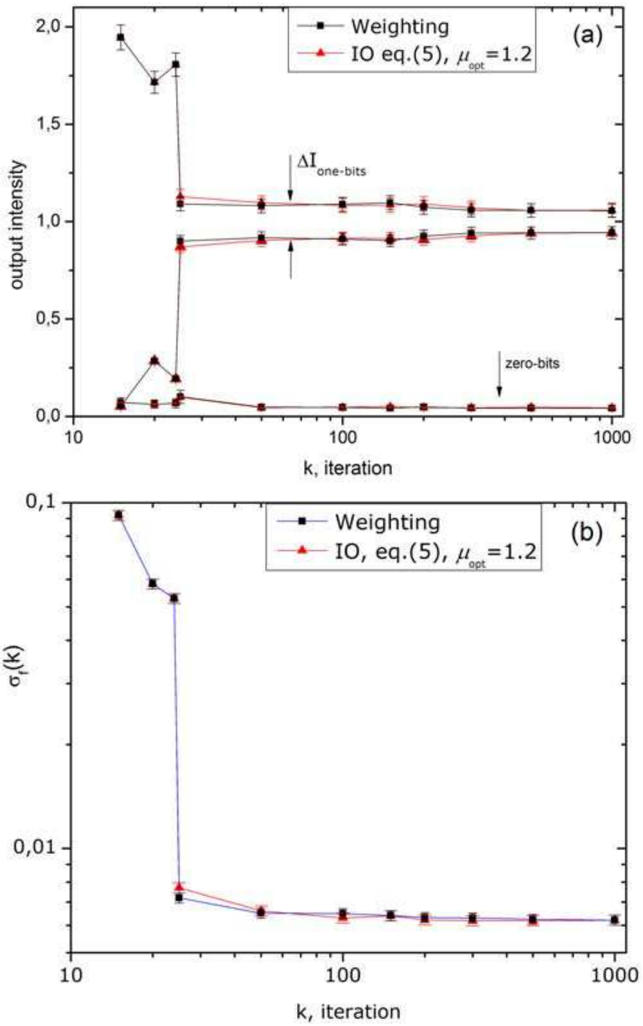


Fig. 10. Kinoform as a beam-splitter: (a) range of output intensities, and (b) variance of spots - intensities *vs* the iteration number

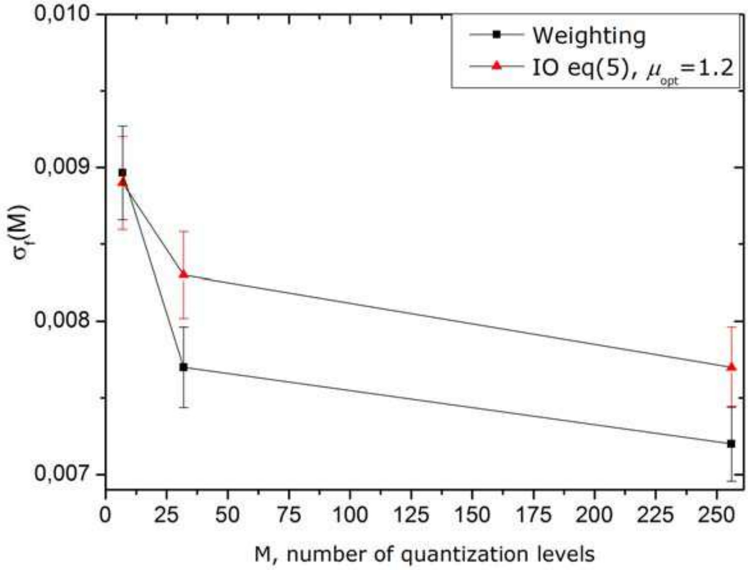


Fig. 11. Kinoform as a beam-splitter: variance of a reconstructed image *vs* the number of quantization levels of the kinoform phase

In Fig. 10, the plots characterizing the quality of the image (Fig. 9) reconstructed by a kinoform are presented. As seen, both methods give practically the same experimental results as distinct from the model experiments with binary objects (see Fig. 3), in which the weighting method has had the obvious advantage over the IO method. The reasons for this situation will be discussed in Conclusion in more details. In Fig. 11, we show the variance of a reconstructed image as a function of the number of quantization levels. In the obtaining of curves in Figs. 10 and 11, kinoforms calculation was performed with use the stepwise quantization at the ratio of iterations ER/weighting=10/15.

4.3 Off-axis kinoform

In Figs. 12 and 13, we give the results of experiments with off-axis kinoforms for object-letter F. Figure 12 demonstrates the example of the reconstruction into partial orders P_x = 0, 0.25, and 0.50. As expected, the 0th and 1st orders remain immobile, and the image shifts between them. It is seen from Fig. 13 how the diffraction efficiency (DE) of a kinoform varies at the successive shift of the image. Curve 1a corresponds to a shift along the axis X, and curve 2a does to a shift along the diagonal in the X,Y-plane (P_x, P_y = 0, 0.1, 0.15, …, 0.50). In this case, the synthesis of kinoforms is realized with the help of the weighting algorithm at the ratio of the numbers of iterations ER/Weighting = 20/200. Curves 1b and 2b represent analogous dependences, but for the synthesis of a kinoform with addition of amplitude freedom iterations (Wyrowski, 1990). While applying the amplitude freedom, the zero-noise on the format of a valid image decreases practically to zero, however, the DE decreases significantly in this case. In the second case, the ratio of iterations ER/Weighting/Amplitude freedom = 20/50/150. In Table 1, we present the more detailed quantitative data. In measurements, we used a Newport Dual Channel Power meter, Model No. 2832-C. The error of measurements was 2.3 % on the average. A decrease in DE at a shift if the image is explained by the finite size of the SLM pixel aperture, in the meanwhile the character of this decrease does not depend on the method of calculation of a kinoform. The synthesis within the input-output method gives close results. The aperture distortion of an image is compensated by multiplication, at the beginning of iterations, of the input function of an object f_o by the inverse function of the pixel aperture $\sin c^{-1}$ (Lohmann & Paris, 1967; Kuzmenko & Yezhov, 2007). We note that the upper bound of DE for object-letter F, covered by the optimized scatterer, calculated by the method Wyrowski (1991) is 90.36%. Thus, the energy losses due to the presence of the zero and higher orders of diffraction are, in the best case (curve 1a, Px=0) of the order of ~ 27% .

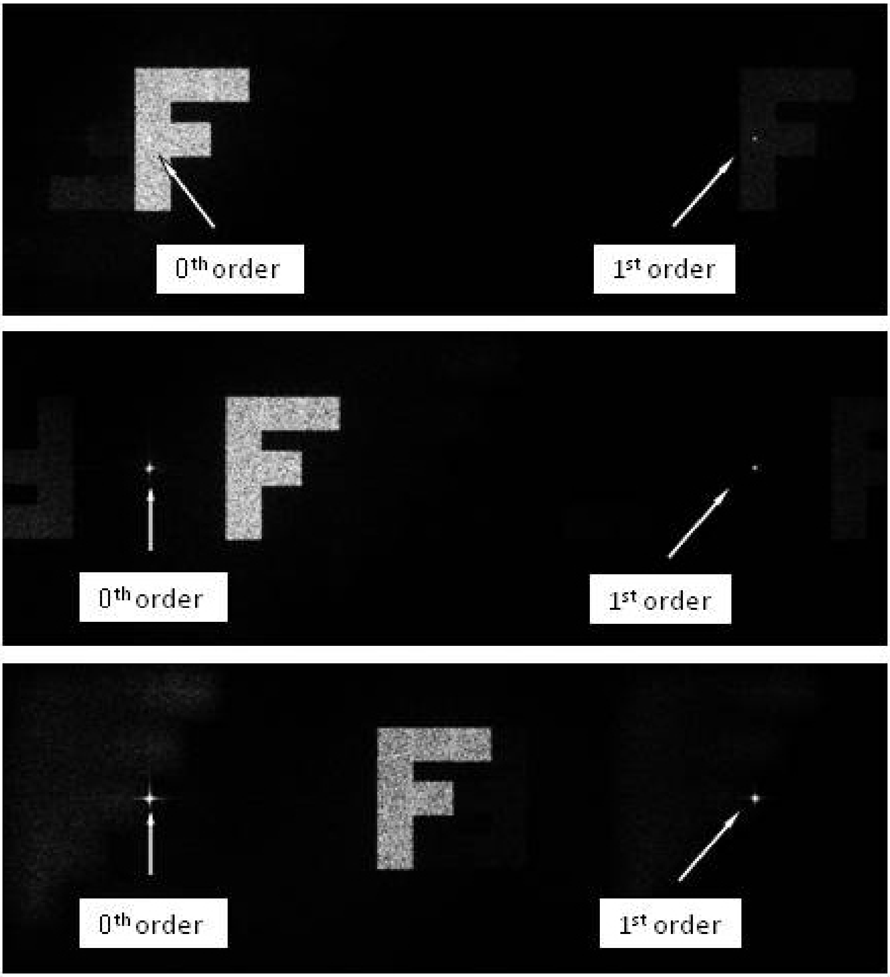


Fig. 12. Off-axis kinoform. Reconstruction into partial diffraction orders, P_x = 0, 0.25, and 0.5

	Diffraction efficiency $\eta = (E_{img} / E_{whole}) \times 100\%$											
	W-algorithm				W-algorithm with A-freedom				IO-algorithm			
	X		X, Y		X		X, Y		X		X, Y	
Order	Exper	Theor	Exper	Theor	Exper	Theor	Exper	Theor	Exper	Theor	Exper	Theor
0	63	89.20	63	89.20	45	63.99	45	63.99	63	88.60	63	88.60
0.1	56	89.70	56	88.40	41	64.00	41	63.30	56	89.00	56	87.82
0.15	52	87.70	51	83.80	39	62.07	37	59.97	53	87.90	53	83.21
0.20	50	84.40	44	76.70	37	60.35	33	55.02	51	83.90	45	76.20
0.25	47	79.00	39	67.9	34	57.06	29	48.87	48	79.40	40	67.50
0.30	42	74.30	35	58.15	30	52.94	26	41.70	44	73.85	36	57.75
0.35	40	--	31	47.9	28	48.3	23	34.47	39	67.45	32	47.5
0.40	35	60.84	26	37.98	26	43.32	20	27.30	36	60.42	26	37.70
0.45	31	53	19	28.6	22	37.90	19	20.75	33	53.02	20	28.60
0.50	29	45	16	--	20	32.64	12	15.08	29	45.54	16	20.81

Table 1. Diffraction efficiency *vs* the partial diffraction order. X – means the image in a partial order along the axis X; X,Y – means the image in a partial order along the diagonal of the X-Y plane. A-freedom means the use of amplitude-freedom iterations at the final stage of calculations. The relative error of measurements Δη/η ≈ 2.3%

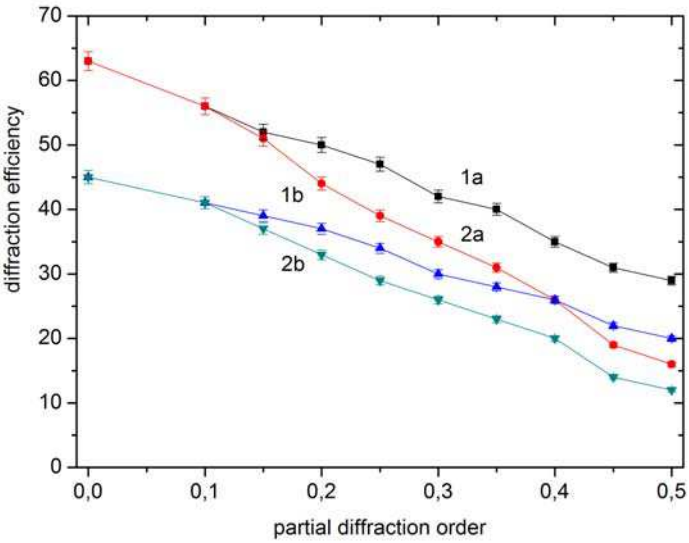


Fig. 13. Diffraction efficiency of the off-axis kinoform *vs* the partial diffraction order (for object-letter F)

5. Conclusion

Summarizing, we may assert that the weighting algorithm has high efficiency in the synthesis of the kinoforms of binary objects. It is worth noting that the effect of stagnation of the algorithm is absent in this case, i.e., the one-bits variance $\sigma_f(k)$ in a reconstructed image tends to zero with increase in the number of iterations, and the noise level (zero-bits) is the same as that of other algorithms. The weighting algorithm is also efficient in calculations of kinoforms as the formers of super-gaussian laser beams. It must be emphasized that the weighting algorithm contains no parameters requiring the optimization (like the feedback parameter μ in the IO algorithm), which essentially accelerates the counting rate.

However, as noted in the literature (Skeren et al., 2002), the methods of synthesis of kinoforms, which differ in accuracy, can give practically identical results. The matter is in that the physico-technical parameters of the available SLM do not allow one to completely realize the potentialities of high-accuracy algorithms. This is indicated by the above-presented results of experiments, where the comparison of the weighting and IO methods is performed. It is possible to assert that all algorithms ensuring $\sigma_f(k) \approx 1{\times}10^{-4}$ or less at the realization of a kinoform on SLM of the type used by us give images of the approximately identical quality.

6. References

Aagebal, H. & Wyrowski, F. (1997). Paraxial beam splitting and shaping. In *Diffractive Optics for Industrial and Commercial Applications*, J. Turunen and F. Wyrowski, ed. (Academie Verlag, Berlin 1997), Chapter 6, pp. 165-188.

Akahori, H. (1986). Spectrum leveling by an iterative algorithm with a dummy area for synthesizing the kinoform. *Appl. Opt.*, vol. 25, pp. 802-811.

Bryngdahl, O. & Wyrowski, F. (1990). Digital holography – computer-generated holograms. In *Progress in Optics*, E. Wolf, ed. (North-Holland, Amsterdam, 1990), vol. 28, pp. 1-86.

Catino, W. C.; LoCicero J. L. & Stark H. (1997). Design of continuous and quantized phase holograms by generalized projections. *JOSA A*, vol. 14, pp. 2715-2725.

Chen, W.; Roychoudhuri C. S. & Banas C. M. (1994). Design approaches for laser-diode material-processing systems using fibers and micro-optics. *Opt. Eng.,* vol. 33, pp. 3662-3669.

Dixit, S. N.; Lawson J. K.; Manes K. R.; Powell H. T. & Nugent K. A. (1994). Kinoform phase plates for focal plane irradiance profile control. *Opt. Lett.*, vol. 19, pp. 417-419.

Duparre, M.; Golub M. A.; Ludge B.; Pavelyev V. S.; Soifer V. A. & Uspleniev G. V. (1995). Investigation of computer-generated diffraction beam shapers for flattening of single-modal CO_2 laser beams. *Appl. Opt.*, vol. 34, pp.2489-2497.

Ehbets, P.; Herzig H. P.; Prongue D. & Gale M. T. (1992). High-efficiency continuous surface-relief gratings for two-dimensional array generation. *Opt. Lett.*, vol. 17, pp. 908-910.

Fienup, J. (1980). Iterative method applied to image reconstruction and to computer-generated holograms. *Opt. Eng*., vol. 19, pp. 297-306.

Fienup, J. (1982). Phase retrieval algorithms: a comparison. *Appl. Opt.*, vol. 21, pp. 2758-2769.

Gale, M. T.; Lang G. K.; Raynor J. M.; Schutz H. & Prongue D. (1992). Fabrication of kinoform structures for optical computing. *Appl. Opt*., vol. 31, pp. 5712-5715.

Gale, M. T.; Rossi M.; Schutz H.; Ehbets P.; Herzig H. P. & Prongue D. (1993). Continuous-relief diffractive optical elements for two-dimensional array generation. *Appl. Opt.*, vol. 32, pp. 2526-2533.

Gallagher, N. C. & Liu, B. (1973). Method for computing kinoforms that reduces image reconstruction error. *Appl. Opt.*, vol. 12, pp. 2328-2335.

Gerchberg, R. W. & Saxton, W. O. (1972). A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, vol. 35, pp. 237-246.

Herzig, H. P.; Prongue D. & Dandliker R. (1990). Design and fabrication of highly efficient fan-out elements. *Japanese J. of Appl. Phys.*, vol. 29, pp. L1307-L1309.

Johansson, M. & Bengtsson, J. (2000). Robust design method for highly efficient beam-shaping diffractive optical elements using an iterative-Fourier transform algorithm with soft operations. *J. Mod. Opt.*, vol. 47, pp. 1385-1398.

Kuzmenko, A. V. (2006). Method of kinoform synthesis. UA Patent No. 65295 (priority from 03.07.2003), UkrPatent, Bulletin of Inventions No. 1, 2006.

Kuzmenko, A. V. & Yezhov, P. V. (2007). Iterative algorithms for off-axis double-phase computer-generated holograms implemented with phase-only spatial light modulator. *Appl. Opt.*, vol. 46, pp. 7392-7400.

Kuzmenko, A. V. (2008). Weighting iterative Fourier transform algorithm of the kinoform synthesis. *Opt. Lett.*, vol. 33, pp. 1147-1149.

Kuzmenko, A. V. & Yezhov, P. V (2008). Iterative Fourier-transform algorithm of synthesis of a kinoform with the use of the operation of predistortion of an object. *Proc. of SPIE*, vol. 7008, 70081W-1-9.

Kuzmenko, A. V. & Yezhov, P. V (2009). Method of kinoform synthesis. UA Patent No. 86245 (priority from 05.02.2007), UkrPatent, Bulletin of Inventions No. 7, 2009.

Lee, W. H. (1979). Binary computer-generated holograms. *Appl. Opt.*, vol. 18, pp. 3661-3668.

Levi, A. & Stark, H. (1987). Restoration from phase and magnitude by generalized projections. Chapter 8 in Image Recovery: Theory and Application, ed. H. Stark, Academic Press, INC., 1987.

Lesem, L. B.; Hirsch P. M. & Jordan J. A. Jr. (1967). Computer generation and reconstruction of holograms. *Proc. Symp. Modern Optics*, vol. 17 (New York: Polytechnic Institute of Brooklyn) p. 681-690.

Leger, J. R.; Chen D. & Wang Z. (1994). Diffractive optical element for mode shaping of a Nd: YAG laser. *Opt. Lett.*, vol. 19, pp.108-110.

Liu, J. S. & Taghizaden, M. R. (2002). Iterative algorithm for the design of diffractive phase elements for laser beam shaping. *Opt. Lett.*, vol. 27, pp. 1463-1465.

Lohmann, A. W. & Paris, D. P. (1967). Binary Fraunhofer holograms generated by computer. *Appl. Opt.*, vol. 6, pp. 1739-1748.

Mait, J. N. & Brenner, K.-H. (1988). Optical symbolic substitution: system design using phase-only holograms. *Appl. Opt.*, vol. 27, pp. 1692-1700.

Oton, J.; Ambs P.; Millan M. S. & Perez-Cabre E. (2007). Multipoint phase calibration for improved compensation of inherent wavefront distortion in parallel aligned liquid crystal on silicon displays. *Appl. Opt.*, vol. 46, pp. 5667-5679.

Prongue, D.; Herzig H. P.; Dandliker R. & Gale M. T. (1992). Optimized kinoform structures for highly efficient fan-out elements. *Appl. Opt.*, vol. 31, pp. 5706-5711.

Skeren, M.; Richter I. & Fiala P. (2002). Iterative Fourier transform algorithm: comparison of various approaches. *J. Mod. Opt.*, vol. 49, pp. 1851–1870.

Tikhonov, A. N. & Arsenin, V. IA. (1977). Solution of ill-posed problems, Halsted, New York.

Wyrowski, F. & Bryngdahl, O. (1988). Iterative Fourier-transform algorithm applied to computer holography. *JOSA A*, vol. 5, pp. 1058-1065.

Wyrowski, F. (1990). Diffraction efficiency of analog and quantized digital amplitude holograms: analysis and manipulation. *JOSA A*, vol. 7, pp.383-393.

Wyrowski, F. (1990). Diffractive optical elements: iterative calculation of quantized, blazed phase structures. *JOSA A*, vol. 7, pp. 961-969.

Wyrowski, F. (1991). Upper bound of the diffraction efficiency of diffractive phase elements. *Opt. Lett.*, vol. 16, pp. 1915-1917.

Wyrowski, F. & Zuidema, R. (1994). Diffractive interconnection between a high-power ND:YAG laser and a fiber bundle. *Appl. Opt.*, vol. 33, pp. 6732-6740.

Xin, T.; Gu B.-Y.; Yang G.-Z. & Dong B.-Z. (1995). Diffractive phase elements for beam shaping: a new design method. *Appl. Opt.*, vol. 34, pp.1314-1320.

Youla, D. C. & Webb, H. (1982). Image restoration by the method of convex projections: Part 1 – Theory. *IEEE Trans. On Medical Imaging*, vol. MI-1, pp. 81-94.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

Alexander Kuzmenko, Pavlo Iezhov and Jin-Tae Kim (2011). Weighting Iterative Fourier Transform Algorithm for Kinoform Implemented with Liquid-Crystal SLM, Fourier Transforms - Approach to Scientific Principles, Prof. Goran Nikolic (Ed.), ISBN: 978-953-307-231-9, InTech, Available from: http://www.intechopen.com/books/fourier-transforms-approach-to-scientific-principles/weighting-iterative-fourier-transform-algorithm-for-kinoform-implemented-with-liquid-crystal-slm

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Two-Dimensional Quaternionic Windowed Fourier Transform

Mawardi Bahri[1] and Ryuichi Ashino[2]
[1]*Department of Mathematics, Hasanuddin University*
Tamalanrea Makassar
[2]*Mathematical Sciences, Osaka Kyoiku University*
Kashiwara, Osaka, 582-8582
[1]*Indonesia*
[2]*Japan*

1. Introduction

Signal processing is a fast growing area today and the desired effectiveness in utilization of bandwidth and energy makes the progress even faster. Special signal processors have been developed to make it possible to implement the theoretical knowledge in an efficient way. Signal processors are nowadays frequently used in equipment for radio, transportation, medicine, and production, etc.

One of the basic problems encountered in signal representations using conventional Fourier transform (FT) is the ineffectiveness of the Fourier kernel to represent and compute location information. One method to overcome such a problem is the windowed Fourier transform (WFT). Recently, Gröchenig (2001); Gröchenig & Zimmermann (2001); Weisz (2008) have extensively studied the WFT and its properties from a mathematical point of view. Kemao (2007); Zhong & Zeng (2007) applied the WFT as a tool of spatial-frequency analysis, which is able to characterize the local frequency at any location in a fringe pattern.

On the other hand the quaternion Fourier transform (QFT), which is a nontrivial generalization of the real and complex Fourier transform (FT) using quaternion algebra, has been of interest to researchers, for example, Hitzer (2007); Mawardi et al. (2008); Sangwine & Ell (2007). It was found that many FT properties still hold but others have to be modified. Based on the (right-sided) QFT, one can extend the classical windowed Fourier transform (WFT) to quaternion algebra while enjoying the same properties as in the classical case.

In this paper, by using the adjoint operator of the (right-sided) QFT, we derive the Plancherel theorem for the QFT. We apply it to prove the orthogonality relation and reconstruction formula of the two-dimensional quaternionic windowed Fourier transform (QWFT). Our results can be considered as an extension and continuation of the previous work of Mawardi et al. (2008). We then present several examples to show the differences between the QWFT and the WFT. Finally, we present a generalization of the QWFT to higher dimensions.

2. Basics

For convenience of further discussions, we briefly review some basic facts on quaternions. The quaternion algebra over $\mathbb{R}$, denoted by

$$\mathbb{H} = \{q = q_0 + \boldsymbol{i}q_1 + \boldsymbol{j}q_2 + \boldsymbol{k}q_3 \mid q_0, q_1, q_2, q_3 \in \mathbb{R}\}, \tag{1}$$

is an associative non-commutative four-dimensional algebra, which obeys Hamilton's multiplication rules:

$$\boldsymbol{ij} = -\boldsymbol{ji} = \boldsymbol{k}, \quad \boldsymbol{jk} = -\boldsymbol{kj} = \boldsymbol{i}, \quad \boldsymbol{ki} = -\boldsymbol{ik} = \boldsymbol{j}, \quad \boldsymbol{i}^2 = \boldsymbol{j}^2 = \boldsymbol{k}^2 = \boldsymbol{ijk} = -1. \tag{2}$$

The quaternion conjugate of a quaternion q is defined by

$$\bar{q} = q_0 - \boldsymbol{i}q_1 - \boldsymbol{j}q_2 - \boldsymbol{k}q_3, \qquad q_0, q_1, q_2, q_3 \in \mathbb{R}, \tag{3}$$

and it is an anti-involution, i.e.

$$\overline{qp} = \bar{p}\bar{q}. \tag{4}$$

From (3), we obtain the norm of $q \in \mathbb{H}$ defined as

$$|q| = \sqrt{q\bar{q}} = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2}. \tag{5}$$

It is not difficult to see that

$$|qp| = |q||p|, \qquad \forall p, q \in \mathbb{H}. \tag{6}$$

Using the conjugate (3) and the modulus of q, we can define the inverse of $q \in \mathbb{H} \setminus \{0\}$ as

$$q^{-1} = \frac{\bar{q}}{|q|^2}, \tag{7}$$

which shows that $\mathbb{H}$ is a normed division algebra.

It is convenient to introduce the inner product $(f, g)_{L^2(\mathbb{R}^2;\mathbb{H})}$ valued in $\mathbb{H}$ of two quaternion functions f and g as follows:

$$(f, g)_{L^2(\mathbb{R}^2;\mathbb{H})} = \int_{\mathbb{R}^2} f(\boldsymbol{x})\overline{g(\boldsymbol{x})}\, d^2\boldsymbol{x}. \tag{8}$$

The associated norm is defined by

$$\|f\|_{L^2(\mathbb{R}^2;\mathbb{H})} = (f, f)^{1/2}_{L^2(\mathbb{R}^2;\mathbb{H})} = \left(\int_{\mathbb{R}^2} |f(\boldsymbol{x})|^2\, d^2\boldsymbol{x}\right)^{1/2}. \tag{9}$$

As a consequence of the inner product (8), we obtain the *quaternion Cauchy-Schwarz* inequality:

$$\left|\int_{\mathbb{R}^2} f\bar{g}\, d^2\boldsymbol{x}\right| \leq \left(\int_{\mathbb{R}^2} |f|^2 d^2\boldsymbol{x}\right)^{1/2} \left(\int_{\mathbb{R}^2} |g|^2 d^2\boldsymbol{x}\right)^{1/2}, \qquad \forall f, g \in L^2(\mathbb{R}^2; \mathbb{H}). \tag{10}$$

3. Quaternionic Fourier Transform (QFT)

Let us introduce the continuous (right-sided) QFT. For more details, we refer the reader to Hitzer (2007); Mawardi et al. (2008); Sangwine & Ell (2007).

3.1 Definition of QFT

Definition 3.1 (Right-sided QFT). *The QFT of $f \in L^1(\mathbb{R}^2;\mathbb{H})$ is the function $\mathcal{F}_q\{f\}:\mathbb{R}^2 \to \mathbb{H}$ given by*

$$\mathcal{F}_q\{f\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^2} f(\boldsymbol{x}) e^{-\boldsymbol{i}\omega_1 x_1} e^{-\boldsymbol{j}\omega_2 x_2}\, d^2\boldsymbol{x}, \tag{11}$$

where $\boldsymbol{x} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2$, $\boldsymbol{\omega} = \omega_1\boldsymbol{e}_1 + \omega_2\boldsymbol{e}_2$, and the quaternion exponential product $e^{-\boldsymbol{i}\omega_1 x_1} e^{-\boldsymbol{j}\omega_2 x_2}$ is the quaternion Fourier kernel.

Theorem 3.1 (Inverse QFT). *Suppose that $f \in L^2(\mathbb{R}^2;\mathbb{H})$ and $\mathcal{F}_q\{f\} \in L^1(\mathbb{R}^2;\mathbb{H})$. Then the QFT of f is an invertible transform and its inverse is given by*

$$\mathcal{F}_q^{-1}[\mathcal{F}_q\{f\}](\boldsymbol{x}) = f(\boldsymbol{x}) = \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} \mathcal{F}_q\{f\}(\boldsymbol{\omega}) e^{\boldsymbol{j}\omega_2 x_2} e^{\boldsymbol{i}\omega_1 x_1}\, d^2\boldsymbol{\omega}, \tag{12}$$

where the quaternion exponential product $e^{\boldsymbol{j}\omega_2 x_2} e^{\boldsymbol{i}\omega_1 x_1}$ is called the inverse (right-sided) quaternion Fourier kernel.

4. Linear Operators on Quaternionic Hilbert Spaces

In this section, we will briefly introduce the notation of linear operator on quaternionic Hilbert spaces. In fact, it is a natural generalization of the idea of an operator on a real and complex Hilbert space.

Definition 4.1. *Let X and Y be two $\mathbb{H}$-vector spaces. The operator $T : X \longrightarrow Y$ is called a left $\mathbb{H}$-linear space if*

$$T(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha T(\boldsymbol{x}) + \beta T(\boldsymbol{x}), \tag{13}$$

for all quaternion constants $\alpha, \beta \in \mathbb{H}$ and for all $\boldsymbol{x}, \boldsymbol{y} \in X$.

Definition 4.2. *The adjoint of $\mathbb{H}$-linear operator $T : X \longrightarrow X$ is the unique $\mathbb{H}$-linear operator $T^* : X \longrightarrow X$ such that*

$$(T\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, T^*\boldsymbol{y}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X. \tag{14}$$

This gives the following result.

Theorem 4.1. *The adjoint of the QFT is inverse of the QFT multiplied by $(2\pi)^2$, i.e.*

$$(\mathcal{F}_q\{f\}, g)_{L^2(\mathbb{R}^2;\mathbb{H})} = (2\pi)^2 (f, \mathcal{F}_q^{-1}\{g\})_{L^2(\mathbb{R}^2;\mathbb{H})}. \tag{15}$$

Proof. For $f, g \in L^2(\mathbb{R}^2;\mathbb{H})$ we calculate the inner product (8) to get

$$\begin{aligned}
(\mathcal{F}_q\{f\}, g)_{L^2(\mathbb{R}^2;\mathbb{H})} &= \int_{\mathbb{R}^2} \mathcal{F}_q\{f\}(\boldsymbol{\omega})\, \overline{g(\boldsymbol{\omega})}\, d^2\boldsymbol{\omega} \\
&\overset{(11)}{=} \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} f(\boldsymbol{x})\, e^{-\boldsymbol{i}\omega_1 x_1} e^{-\boldsymbol{j}\omega_2 x_2}\, d^2\boldsymbol{x}\overline{g(\boldsymbol{\omega})}\, d^2\boldsymbol{\omega} \\
&\overset{(4)}{=} \int_{\mathbb{R}^2} f(\boldsymbol{x}) \left(\int_{\mathbb{R}^2} \overline{g(\boldsymbol{\omega})\, e^{\boldsymbol{j}x_1\omega_1} e^{\boldsymbol{i}x_2\omega_2}} d^2\boldsymbol{\omega} \right) d^2\boldsymbol{x} \\
&= \int_{\mathbb{R}^2} f(\boldsymbol{x})\, (2\pi)^2 \overline{\mathcal{F}_q^{-1}\{g\}(\boldsymbol{x})}\, d^2\boldsymbol{x} \\
&= (2\pi)^2 (f, \mathcal{F}_q^{-1}\{g\})_{L^2(\mathbb{R}^2;\mathbb{H})},
\end{aligned} \tag{16}$$

which completes the proof. □

Remark 4.1. *Note that Theorem 4.1 is not valid for the (two-sided) QFT. This fact implies that the Plancherel theorem can not be established.*

Theorem 4.2 (Plancherel formula). *Suppose that $f, g \in L^2(\mathbb{R}^2; \mathbb{H})$. Then*

$$(\mathcal{F}_q\{f\}, \mathcal{F}_q\{g\})_{L^2(\mathbb{R}^2;\mathbb{H})} = (2\pi)^2 (f, g)_{L^2(\mathbb{R}^2;\mathbb{H})} \tag{17}$$

and

$$(\mathcal{F}_q^{-1}[\mathcal{F}_q\{f\}], \mathcal{F}_q^{-1}[\mathcal{F}_q\{g\}])_{L^2(\mathbb{R}^2;\mathbb{H})} = (2\pi)^2 (f, g)_{L^2(\mathbb{R}^2;\mathbb{H})}. \tag{18}$$

Proof. A simple calculation gives for every $f, g \in L^2(\mathbb{R}^2; \mathbb{H})$

$$\begin{aligned}(\mathcal{F}_q\{f\}, \mathcal{F}_q\{g\})_{L^2(\mathbb{R}^2;\mathbb{H})} &\overset{(15)}{=} (2\pi)^2 (f, \mathcal{F}_q^{-1}[\mathcal{F}_q\{g\}])_{L^2(\mathbb{R}^2;\mathbb{H})} \\ &\overset{(12)}{=} (2\pi)^2 (f, g)_{L^2(\mathbb{R}^2;\mathbb{H})},\end{aligned} \tag{19}$$

as desired. Equation (18) can be established in a similar manner. □

4.1 Discrete QFT

Similar to the discrete Fourier transform, the discrete quaternionic Fourier transform (DQFT) and the inverse discrete quaternionic Fourier transform (IDQFT) are defined as follows.

Definition 4.3. *Let $f(m, n)$ be a two-dimensional quaternion discrete-time sequence. The DQFT of $f(m, n)$ is defined by $F(u, v) \in \mathbb{H}^{M\times N}$, where*

$$F(u, v) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m, n)\, e^{-\boldsymbol{i}\frac{um}{M}} e^{-\boldsymbol{j}\frac{vn}{N}}. \tag{20}$$

Definition 4.4. *The IDQFT is defined by*

$$f(m, n) = \frac{1}{(2\pi)^2 MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u, v)\, e^{\boldsymbol{j}\frac{vn}{N}} e^{\boldsymbol{i}\frac{um}{M}}. \tag{21}$$

4.2 Application of DQFT

In the following, we introduce an application of the DQFT to study two-dimensional discrete linear time-varying (TV) systems. For this purpose, let us introduce the following definition.

Definition 4.5. *Consider a two-dimensional discrete linear TV system with the quaternion impulse response of the filter denoted $h(\cdot, \cdot, \cdot, \cdot)$. The output $r(\cdot, \cdot)$ of the system to the input $f(\cdot, \cdot)$ is defined by*

$$r(m, n) = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} f(u, v)\, h(m, n, m-u, n-v). \tag{22}$$

The transfer function of the TV filter h can be obtained by

$$R(m, n, \omega_1, \omega_2) = \sum_{m'=-\infty}^{\infty} \sum_{n'=-\infty}^{\infty} h(m, n, m', n')\, e^{-\boldsymbol{i}m'\omega_1} e^{-\boldsymbol{j}n'\omega_2}. \tag{23}$$

The following simple theorem relates the DQFT to the output of a discrete linear TV band-pass filter.

Theorem 4.3. *Consider a linear TV system with the quaternion impulse response h defined by*

$$h(m,n,m',n') = e^{-\boldsymbol{i}\frac{m(m-m')}{M}} e^{-\boldsymbol{j}\frac{n(n-n')}{N}}, \quad \text{for } 0 \le m \le M-1,\, 0 \le n \le N-1. \tag{24}$$

If the input to this system is the quaternion signal $f(u,v)$, then its output $r(\cdot,\cdot)$ is equal to the DQFT of $f(u,v)$.

Proof. Using Definition 4.5, we obtain

$$\begin{aligned} r(m,n) &= \sum_{u=-\infty}^{\infty}\sum_{v=-\infty}^{\infty} f(u,v)\, h(m,n,m-u,n-v) \\ &= \sum_{u=0}^{M-1}\sum_{v=0}^{N-1} f(u,v)\, e^{-\boldsymbol{i}\frac{m(m-(m-u))}{M}}\, e^{-\boldsymbol{j}\frac{n(n-(n-v))}{N}} \\ &= \sum_{u=0}^{M-1}\sum_{v=0}^{N-1} f(u,v)\, e^{-\boldsymbol{i}\frac{um}{M}} e^{-\boldsymbol{j}\frac{vn}{N}}, \end{aligned} \tag{25}$$

which completes the proof by Definition 4.3. □

If the quaternion impulse response h is given by

$$h(m,n,m',n') = \frac{1}{(2\pi)^2 MN} e^{\boldsymbol{j}\frac{n(n-n')}{N}} e^{\boldsymbol{i}\frac{m(m-m')}{M}}, \tag{26}$$

then (22) implies

$$\begin{aligned} r_2(m,n) &= \sum_{u=-\infty}^{\infty}\sum_{v=-\infty}^{\infty} f(u,v)\, h(m,n,m-u,n-v) \\ &= \frac{1}{(2\pi)^2 MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} F(u,v)\, e^{\boldsymbol{j}\frac{n(n-(n-v))}{N}}\, e^{\boldsymbol{i}\frac{m(m-(m-u))}{M}} \\ &= \frac{1}{(2\pi)^2 MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} F(u,v)\, e^{\boldsymbol{j}\frac{nv}{N}}\, e^{\boldsymbol{i}\frac{um}{M}}, \end{aligned} \tag{27}$$

where the input to the system is quaternion signal $F(u,v)$.
Equations (24) and (26) show that the choice of the quaternion impulse response of the filter determines output characteristics of the discrete linear TV systems.

5. Quaternionic windowed Fourier Transform

In section, we introduce the QWFT presented in Mawardi et al. (2010). As we will see, not all properties of the WFT can be established for the QWFT.

5.1 2-D WFT

Although the FT is a powerful tool for the analysis of stationary signals, the FT is not well suited for the analysis of non-stationary signals. Because the FT is a global transformation with poor spatial localization Zhong & Zeng (2007). However, in practice, most natural signals are non-stationary. In order to characterize a non-stationary signal properly, the WFT is commonly used.

Definition 5.1 (WFT). *The WFT of a two-dimensional real signal $f \in L^2(\mathbb{R}^2;\mathbb{R})$ with respect to the window function $g \in L^2(\mathbb{R}^2) \setminus \{0\}$ is given by*

$$\mathcal{G}_g f(\boldsymbol{\omega}, \boldsymbol{b}) = \int_{\mathbb{R}^2} f(\boldsymbol{x})\, \overline{g_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x})}\, d^2\boldsymbol{x}, \tag{28}$$

where the window daughter function $g_{\boldsymbol{\omega},\boldsymbol{b}}$ is defined by

$$g_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x}) = g(\boldsymbol{x} - \boldsymbol{b}) e^{\sqrt{-1}\,\boldsymbol{\omega}\cdot\boldsymbol{x}}. \tag{29}$$

The window daughter function $g_{\boldsymbol{\omega},\boldsymbol{b}}$ is also called the windowed Fourier kernel.

Most applications make use of the Gaussian window function g, which is non-negative and well localized around the origin both in spatial and frequency domains. The Gaussian window function can be represented as

$$g(\boldsymbol{x}, \sigma_1, \sigma_2) = e^{-\left[(x_1/\sigma_1)^2 + (x_2/\sigma_2)^2\right]/2}, \tag{30}$$

where σ_1 and σ_2 are the standard deviations of the Gaussian function. For fixed $\boldsymbol{\omega}_0 = u_0\boldsymbol{e}_1 + v_0\boldsymbol{e}_2$,

$$g_{c,\boldsymbol{\omega}_0}(\boldsymbol{x}, \sigma_1, \sigma_2) = e^{\sqrt{-1}\,(u_0x_1 + v_0x_2)} e^{-\left[(x_1/\sigma_1)^2 + (x_2/\sigma_2)^2\right]/2} \tag{31}$$

is called a *complex Gabor filter*.

5.2 Quaternionic Gabor filters

Bülow (1999; Felsberg & Sommer) extended the complex Gabor filter $g_{c,\boldsymbol{\omega}_0}(\boldsymbol{x}, \sigma_1, \sigma_2)$ to quaternions by replacing the complex kernel $e^{\sqrt{-1}(u_0x_1+v_0x_2)}$ with the inverse (two-sided) quaternion Fourier kernel $e^{\boldsymbol{i}u_0x_1}\, e^{\boldsymbol{j}v_0x_2}$. He proposed the extension form

$$g_q(\boldsymbol{x}, \sigma_1, \sigma_2) = e^{\boldsymbol{i}u_0x_1} e^{\boldsymbol{j}v_0x_2} e^{-\left[(x_1/\sigma_1)^2 + (x_2/\sigma_2)^2\right]/2}, \tag{32}$$

which he called *quaternionic Gabor filter*, and applied it to get the local quaternionic phase of a two-dimensional real signal. Bayro-Corrochano et al. (2007) also used quaternionic Gabor filters for the preprocessing of 2D speech representations. Based on (32), the quaternionic windowed Fourier kernel can be written in the form

$$\Phi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x}) = e^{\boldsymbol{i}u_0x_1} g(\boldsymbol{x} - \boldsymbol{b}) e^{\boldsymbol{j}v_0x_2}. \tag{33}$$

The extension of the WFT to quaternions using the quaternionic windowed Fourier kernel (33) is rather complicated, due to the non-commutativity of quaternion functions. Alternatively, we use the kernel of the (right-sided) QFT to define the quaternionic windowed Fourier kernel which enables us to extend the WFT to quaternions.

Definition 5.2. *For a non-zero quaternion window function $\phi \in L^2(\mathbb{R}^2;\mathbb{H}) \setminus \{0\}$, its quaternionic window daughter function is defined by*

$$\phi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x}) = e^{\boldsymbol{j}\omega_2 x_2} e^{\boldsymbol{i}\omega_1 x_1} \phi(\boldsymbol{x}-\boldsymbol{b}). \tag{34}$$

For fixed $\boldsymbol{\omega}_0 = u_0\boldsymbol{e}_1 + v_0\boldsymbol{e}_2$, our quaternionic Gabor filter is defined by

$$G_q(\boldsymbol{x}, \sigma_1, \sigma_2) = e^{\boldsymbol{j}v_0 x_2} e^{\boldsymbol{i}u_0 x_1} e^{-\left[(x_1/\sigma_1)^2 + (x_2/\sigma_2)^2\right]/2}. \tag{35}$$

Lemma 5.1. *For $\phi_{\omega,\boldsymbol{b}} \in L^2(\mathbb{R}^2;\mathbb{H})$, we have*

$$\|\phi_{\omega,\boldsymbol{b}}\|^2_{L^2(\mathbb{R}^2;\mathbb{H})} = \|\phi\|^2_{L^2(\mathbb{R}^2;\mathbb{H})}. \tag{36}$$

5.3 Definition of QWFT

Definition 5..3 (QWFT). *Let $\phi \in L^2(\mathbb{R}^2;\mathbb{H}) \setminus \{0\}$ be a non-zero quaternion window function. Denote by G_ϕ, the QWFT on $L^2(\mathbb{R}^2;\mathbb{H})$. The QWFT of $f \in L^2(\mathbb{R}^2;\mathbb{H})$ with respect to ϕ is defined by*

$$\begin{aligned} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) &= \int_{\mathbb{R}^2} f(\boldsymbol{x})\, \overline{\phi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x})}\, d^2\boldsymbol{x} \\ &= \int_{\mathbb{R}^2} f(\boldsymbol{x})\, \overline{\phi(\boldsymbol{x}-\boldsymbol{b})} e^{-\boldsymbol{i}\omega_1 x_1} e^{-\boldsymbol{j}\omega_2 x_2} d^2\boldsymbol{x}. \end{aligned} \tag{37}$$

The quaternionic window daughter function

$$\phi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x}) = e^{\boldsymbol{j}\omega_2 x_2} e^{\boldsymbol{i}\omega_1 x_1} \phi(\boldsymbol{x}-\boldsymbol{b}) \tag{38}$$

is also called the quaternionic windowed Fourier kernel.

These lead to the following observations:

- Equation (37) shows that it is generated using the inverse (right-sided) QFT kernel. Note that the definition is not valid using the kernel of the (two-sided) QFT.
- If we fix $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, and $b_1 = b_2 = 0$, and take the Gaussian function as the window function of (38), then we get the quaternionic Gabor filter

$$G_q(\boldsymbol{x}, \sigma_1, \sigma_2) = e^{\boldsymbol{j}v_0 x_2} e^{\boldsymbol{i}u_0 x_1} e^{-\left[(x_1/\sigma_1)^2 + (x_2/\sigma_2)^2\right]/2}. \tag{39}$$

- Since the modulation property does not hold for the QFT, equations (37) and (38) can not be expressed in terms of the QFT.

It is easy to see that

$$G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) = \mathcal{F}_q\{f \cdot T_{\boldsymbol{b}}\bar{\phi}\}(\boldsymbol{\omega}), \tag{40}$$

where the translation operator is defined by

$$T_{\boldsymbol{b}} f = f(\boldsymbol{x}-\boldsymbol{b}). \tag{41}$$

Equation (40) clearly shows that the QWFT can be regarded as the (right-sided) QFT of the product of a quaternion-valued signal f and a quaternion conjugated and shifted quaternion window function, or as an inner product (8) of f and the quaternionic window daughter function. In contrast to the QFT basis $e^{-\boldsymbol{i}\omega_1 x_1} e^{-\boldsymbol{j}\omega_1 x_2}$, which has an infinite spatial extension, the QWFT basis $\phi(\boldsymbol{x}-\boldsymbol{b})\, e^{-\boldsymbol{i}\omega_1 x_1} e^{-\boldsymbol{j}\omega_1 x_2}$ has a limited spatial extension due to the locality of the quaternion window function $\phi(\boldsymbol{x}-\boldsymbol{b})$.

5.4 Properties of QWFT

The following proposition describes the elementary properties of the QWFT. Its proof is straightforward.

Proposition 5.2. *Let $\phi \in L^2(\mathbb{R}^2;\mathbb{H})$ be a quaternion window function.*

(i). (Left linearity)

$$[G_\phi(\lambda f + \mu g)](\boldsymbol{\omega}, \boldsymbol{b}) = \lambda G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) + \mu G_\phi g(\boldsymbol{\omega}, \boldsymbol{b}), \tag{42}$$

for arbitrary quaternion constants $\lambda, \mu \in \mathbb{H}$.

(ii). (Parity)

$$G_{P\phi}(Pf)(\boldsymbol{\omega}, \boldsymbol{b}) = G_\phi f(\boldsymbol{\omega}, -\boldsymbol{b}), \tag{43}$$

where P is the parity operator defined by $Pf(x) = f(-x)$.

(iii). (Specific shift) *Assume that $f = f_0 + \boldsymbol{i} f_1$ and $\phi = \phi_0 + \boldsymbol{i}\phi_1$.*

$$G_\phi(T_{\boldsymbol{x}_0} f)(\boldsymbol{\omega}, \boldsymbol{b}) = e^{-\boldsymbol{i}\omega_1 x_0} \left(G_\phi f(\boldsymbol{\omega}, \boldsymbol{b} - \boldsymbol{x}_0)\right) e^{-\boldsymbol{j}\omega_2 y_0}. \tag{44}$$

Let us give alternative proofs of the orthogonality relation and reconstruction formula. We follow the idea of Gröchenig (2001) to prove the theorems.

Theorem 5..3 (Orthogonality relation)**.** *Let ϕ, ψ be quaternion window functions and $f, g \in L^2(\mathbb{R}^2;\mathbb{H})$ arbitrary. Then we have*

$$\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b})\, \overline{G_\psi g(\boldsymbol{\omega}, \boldsymbol{b})}\, d^2\boldsymbol{\omega}\, d^2\boldsymbol{b} = (2\pi)^2 (f(\bar{\phi}, \bar{\psi})_{L^2(\mathbb{R}^2;\mathbb{H})}, g)_{L^2(\mathbb{R}^2;\mathbb{H})}. \tag{45}$$

Proof. We notice that

$$G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) = \mathcal{F}_q\{f \cdot T_{\boldsymbol{b}}\bar{\phi}\}(\boldsymbol{\omega}), \tag{46}$$

for fixed $\boldsymbol{b}$. We have known that the Plancherel theorem is valid for the (right-sided) QFT. So, applying it into the left-hand side of (45), we get

$$\begin{aligned}\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b})\, \overline{G_\psi g(\boldsymbol{\omega}, \boldsymbol{b})}\, d^2\boldsymbol{\omega} &= (\mathcal{F}_q\{f \cdot T_{\boldsymbol{b}}\bar{\phi}\}, \mathcal{F}\{f \cdot T_{\boldsymbol{b}}\bar{\psi}\})_{L^2(\mathbb{R}^2;\mathbb{H})} \\ &= (2\pi)^2 (f \cdot T_{\boldsymbol{b}}\bar{\phi}, f \cdot T_{\boldsymbol{b}}\bar{\psi})_{L^2(\mathbb{R}^2;\mathbb{H})} \\ &= (2\pi)^2 \int_{\mathbb{R}^2} f(\boldsymbol{x})\overline{\phi(\boldsymbol{x} - \boldsymbol{b})}\psi(\boldsymbol{x} - \boldsymbol{b})\overline{g(\boldsymbol{x})}\, d^2\boldsymbol{x}. \end{aligned} \tag{47}$$

If we assume that $f\bar{\phi}$ and $\psi\bar{g}$ are in $L^2(\mathbb{R}^2;\mathbb{H})$, then integrating (47) with respect to $d^2\boldsymbol{b}$ yields

$$\begin{aligned}\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b})\, \overline{G_\psi g(\boldsymbol{\omega}, \boldsymbol{b})}\, d^2\boldsymbol{\omega}\, d^2\boldsymbol{b} &= (2\pi)^2 \int_{\mathbb{R}^2} f(\boldsymbol{x}) \int_{\mathbb{R}^2} \overline{\phi(\boldsymbol{x} - \boldsymbol{b})}\psi(\boldsymbol{x} - \boldsymbol{b})\overline{g(\boldsymbol{x})}\, d^2\boldsymbol{x}\, d^2\boldsymbol{b} \\ &= (2\pi)^2 \int_{\mathbb{R}^2} f(\boldsymbol{x}) \int_{\mathbb{R}^2} \overline{\phi(\boldsymbol{x}')}\psi(\boldsymbol{x}')\, d^2\boldsymbol{x}'\, \overline{g(\boldsymbol{x})}\, d^2\boldsymbol{x}, \end{aligned} \tag{48}$$

which proves the theorem. □

From the above theorem, we obtain the following consequences.

(i). If $\phi = \psi$, then

$$\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \, \overline{G_\phi g(\boldsymbol{\omega}, \boldsymbol{b})} \, d^2\boldsymbol{b} \, d^2\boldsymbol{\omega} = (2\pi)^2 \|\phi\|_{L^2(\mathbb{R}^2;\mathbb{H})} (f, g)_{L^2(\mathbb{R}^2;\mathbb{H})}. \tag{49}$$

This formula is quite similar to the orthogonality relation of the classical WFT, for example, see Gröchenig (2001). However, we must remember that equation (49) is a quaternion valued function.

(ii). If $f = g$, then

$$\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \, \overline{G_\psi f(\boldsymbol{\omega}, \boldsymbol{b})} \, d^2\boldsymbol{b} \, d^2\boldsymbol{\omega} = (2\pi)^2 (f(\bar{\phi}, \bar{\psi})_{L^2(\mathbb{R}^2;\mathbb{H})}, f)_{L^2(\mathbb{R}^2;\mathbb{H})}. \tag{50}$$

(iii). If $f = g$ and $\phi = \psi$, then

$$\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \left| G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \right|^2 d^2\boldsymbol{b} \, d^2\boldsymbol{\omega} = (2\pi)^2 \|f\|^2_{L^2(\mathbb{R}^2;\mathbb{H})} \|\phi\|^2_{L^2(\mathbb{R}^2;\mathbb{H})}. \tag{51}$$

(iv). If the quaternion window function is normalized so that $\|\phi\|_{L^2(\mathbb{R}^2;\mathbb{H})} = 1$, then (51) becomes

$$\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \left| G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \right|^2 d^2\boldsymbol{b} \, d^2\boldsymbol{\omega} = (2\pi)^2 \|f\|^2_{L^2(\mathbb{R}^2;\mathbb{H})}. \tag{52}$$

Equation (52) shows that the QWFT is an *isometry* from $L^2(\mathbb{R}^2; \mathbb{H})$ into $L^2(\mathbb{R}^2; \mathbb{H})$. In other words, up to the factor $(2\pi)^2$, the *total energy* of a quaternion-valued signal computed in the spatial domain is equal to the total energy computed in the quaternionic windowed Fourier domain.

Theorem 5..4 (Reconstruction formula)**.** *Let $\phi, \psi \in L^2(\mathbb{R}^2; \mathbb{H})$ be two quaternion window functions. Assume that $(\phi, \psi)_{L^2(\mathbb{R}^2;\mathbb{H})} \neq 0$. Then, every 2-D quaternion signal $f \in L^2(\mathbb{R}^2; \mathbb{H})$ can be fully reconstructed by*

$$f(\boldsymbol{x}) = (2\pi)^{-2} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \psi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x}) \, (\bar{\phi}, \bar{\psi})^{-1}_{L^2(\mathbb{R}^2;\mathbb{H})} d^2\boldsymbol{b} \, d^2\boldsymbol{\omega}. \tag{53}$$

Under the same assumptions as in (49)*, we obtain*

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^2 \|\phi\|^2_{L^2(\mathbb{R}^2;\mathbb{H})}} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \phi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x}) \, d^2\boldsymbol{b} \, d^2\boldsymbol{\omega}. \tag{54}$$

Proof. By direct calculation, we obtain

$$\begin{aligned} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \, \overline{G_\psi g(\boldsymbol{\omega}, \boldsymbol{b})} \, d^2\boldsymbol{\omega} \, d^2\boldsymbol{b} &= \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \, \psi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x}) \bar{g}(\boldsymbol{x}) \, d^2\boldsymbol{\omega} \, d^2\boldsymbol{b} \, d^2\boldsymbol{x} \\ &= (\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \psi_{\boldsymbol{\omega},\boldsymbol{b}} \, d^2\boldsymbol{\omega} \, d^2\boldsymbol{b}, g)_{L^2(\mathbb{R}^2;\mathbb{H})}, \end{aligned} \tag{55}$$

for every $g \in L^2(\mathbb{R}^2; \mathbb{H})$. Applying (45) of Theorem 5.3 to the left-hand side of (55), we have

$$(2\pi)^2 (f(\bar{\phi}, \bar{\psi})_{L^2(\mathbb{R}^2;\mathbb{H})}, g)_{L^2(\mathbb{R}^2;\mathbb{H})} = \left(\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) \psi_{\boldsymbol{\omega},\boldsymbol{b}} \, d^2\boldsymbol{\omega} \, d^2\boldsymbol{b}, g \right)_{L^2(\mathbb{R}^2;\mathbb{H})}, \tag{56}$$

for every $g \in L^2(\mathbb{R}^2;\mathbb{H})$. Since the inner product identity (56) holds for every $g \in L^2(\mathbb{R}^n;\mathbb{H})$, we conclude that

$$(2\pi)^2 f(\bar{\phi},\bar{\psi})_{L^2(\mathbb{R}^2;\mathbb{H})} = \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega},\boldsymbol{b})\psi_{\boldsymbol{\omega},\boldsymbol{b}}\, d^2\boldsymbol{\omega}\, d^2\boldsymbol{b}. \tag{57}$$

Multiplying both sides of (57) from the right side by $(2\pi)^{-2}(\bar{\phi},\bar{\psi})^{-1}_{L^2(\mathbb{R}^2;\mathbb{H})}$, we immediately obtain

$$f = (2\pi)^{-2}\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega},\boldsymbol{b})\psi_{\boldsymbol{\omega},\boldsymbol{b}}\,(\bar{\phi},\bar{\psi})^{-1}_{L^2(\mathbb{R}^2;\mathbb{H})}\, d^2\boldsymbol{\omega}\, d^2\boldsymbol{b}. \tag{58}$$

Notice also that if $\phi = \psi$, then $(\bar{\phi},\bar{\psi})_{L^2(\mathbb{R}^2;\mathbb{H})} = \|\bar{\phi}\|^2_{L^2(\mathbb{R}^2;\mathbb{H})} = \|\phi\|^2_{L^2(\mathbb{R}^2;\mathbb{H})}$. This proves (54). □

Theorem 5.5 (Reproducing kernel). *Let be $\phi \in L^2(\mathbb{R}^2;\mathbb{H})$ be a quaternion window function. If*

$$\mathbb{K}_\phi(\boldsymbol{\omega},\boldsymbol{b};\boldsymbol{\omega}',\boldsymbol{b}') = \frac{1}{(2\pi)^2\|\phi\|^2_{L^2(\mathbb{R}^2;\mathbb{H})}}(\phi_{\boldsymbol{\omega},\boldsymbol{b}},\phi_{\boldsymbol{\omega}',\boldsymbol{b}'})_{L^2(\mathbb{R}^2;\mathbb{H})}, \tag{59}$$

then $\mathbb{K}_\phi(\boldsymbol{\omega},\boldsymbol{b};\boldsymbol{\omega}',\boldsymbol{b}')$ is a reproducing kernel, i.e.

$$G_\phi f(\boldsymbol{\omega}',\boldsymbol{b}') = \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega},\boldsymbol{b})\mathbb{K}_\phi(\boldsymbol{\omega},\boldsymbol{b};\boldsymbol{\omega}',\boldsymbol{b}')\, d^2\boldsymbol{\omega}\, d^2\boldsymbol{b}. \tag{60}$$

Proof. By inserting (53) into the definition of the QWFT (37), we obtain

$$\begin{aligned}
G_\phi f(\boldsymbol{\omega}',\boldsymbol{b}') &= \int_{\mathbb{R}^2} f(\boldsymbol{x})\,\overline{\phi_{\boldsymbol{\omega}',\boldsymbol{b}'}(\boldsymbol{x})}\, d^2\boldsymbol{x} \\
&= \int_{\mathbb{R}^2}\left(\frac{1}{(2\pi)^2\|\phi\|^2_{L^2(\mathbb{R}^2;\mathbb{H})}}\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega},\boldsymbol{b})\,\phi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x})d^2\boldsymbol{b}\, d^2\boldsymbol{\omega}\right)\overline{\phi_{\boldsymbol{\omega}',\boldsymbol{b}'}(\boldsymbol{x})}\, d^2\boldsymbol{x} \\
&= \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega},\boldsymbol{b})\frac{1}{(2\pi)^2\|\phi\|^2_{L^2(\mathbb{R}^2;\mathbb{H})}}\left(\int_{\mathbb{R}^2}\phi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x})\overline{\phi_{\boldsymbol{\omega}',\boldsymbol{b}'}(\boldsymbol{x})}\, d^2\boldsymbol{x}\right) d^2\boldsymbol{b}\, d^2\boldsymbol{\omega} \\
&= \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} G_\phi f(\boldsymbol{\omega},\boldsymbol{b})\mathbb{K}_\phi(\boldsymbol{\omega},\boldsymbol{b};\boldsymbol{\omega}',\boldsymbol{b}')\, d^2\boldsymbol{b}\, d^2\boldsymbol{\omega},
\end{aligned} \tag{61}$$

which was to be proved. □

5.5 Examples of the QWFT

For illustrative purposes, we will give examples of the QWFT. Let us begin with a straightforward example given in Mawardi et al. (2010).

Example 5.1. *Consider the two-dimensional first order B-spline window function defined by*

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & \text{if } -1 \le x_1 \le 1 \text{ and } -1 \le x_2 \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{62}$$

Obtain the QWFT of the function defined as follows:

$$f(\boldsymbol{x}) = \begin{cases} e^{x_1+x_2}, & \text{if } -\infty < x_1 < 0 \text{ and } -\infty < x_2 < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{63}$$

By applying the definition of the QWFT, we have

$$G_\phi f(\boldsymbol{\omega},\boldsymbol{b}) = \frac{1}{(2\pi)^2}\int_{-1+b_1}^{m_1}\int_{-1+b_2}^{m_2} e^{x_1+x_2}e^{-\boldsymbol{i}\omega_1 x_1}e^{-\boldsymbol{j}\omega_2 x_2}dx_1dx_2,$$
$$m_1 = \min(0,1+b_1), \quad m_2 = \min(0,1+b_2). \tag{64}$$

Simplifying (64) yields

$$\begin{aligned}G_\phi f(\boldsymbol{\omega},\boldsymbol{b}) &= \frac{1}{(2\pi)^2}\int_{-1+b_1}^{m_1}\int_{-1+b_2}^{m_2} e^{x_1(1-\boldsymbol{i}\omega_1)}e^{x_2(1-\boldsymbol{j}\omega_2)}\,d^2\boldsymbol{x}\\ &= \frac{1}{(2\pi)^2}\int_{-1+b_1}^{m_1} e^{x_1(1-\boldsymbol{i}\omega_1)}dx_1\int_{-1+b_2}^{m_2} e^{x_2(1-\boldsymbol{j}\omega_2)}\,dx_2\\ &= \frac{1}{(2\pi)^2}\,e^{x_1(1-\boldsymbol{i}\omega_1)}(1-\boldsymbol{i}\omega_1)\Big|_{-1+b_1}^{m_1}\frac{e^{x_2(1-\boldsymbol{j}\omega_2)}}{(1-\boldsymbol{j}\omega_2)}\Big|_{-1+b_2}^{m_2}\\ &= \frac{(e^{m_1(1-\boldsymbol{i}\omega_1)}-e^{(-1+b_1)(1-\boldsymbol{i}\omega_1)})(e^{m_2(1-\boldsymbol{j}\omega_2)}-e^{(-1+b_2)(1-\boldsymbol{j}\omega_2)})}{(2\pi)^2(1-\boldsymbol{i}\omega_1-\boldsymbol{j}\omega_2+\boldsymbol{k}\omega_1\omega_2)}.\end{aligned} \tag{65}$$

Using the properties of quaternions, we obtain

$$G_\phi f(\boldsymbol{\omega},\boldsymbol{b}) = \frac{(e^{m_1(1-\boldsymbol{i}\omega_1)}-e^{(-1+b_1)(1-\boldsymbol{i}\omega_1)})(e^{m_2(1-\boldsymbol{j}\omega_2)}-e^{(-1+b_2)(1-\boldsymbol{j}\omega_2)})(1+\boldsymbol{i}\omega_1+\boldsymbol{j}\omega_2-\boldsymbol{k}\omega_1\omega_2)}{(2\pi)^2(1+\omega_1^2+\omega_2^2+\omega_1^2\omega_2^2)}. \tag{66}$$

Example 5.2. *Let the window function be the two-dimensional Haar function defined by*

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & \text{for } 0\le x_1<1/2 \text{ and } 0\le x_2<1/2,\\ -1, & \text{for } 1/2\le x_1<1 \text{ and } 1/2\le x_2<1,\\ 0, & \text{otherwise.}\end{cases} \tag{67}$$

Find the QWFT of the Gaussian function $f(\boldsymbol{x}) = e^{-(x_1^2+x_2^2)}$.

From Definition 5.3, we obtain

$$\begin{aligned}G_\phi f(\boldsymbol{\omega},\boldsymbol{b}) &= \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} f(\boldsymbol{x})\overline{\phi(\boldsymbol{x}-\boldsymbol{b})}e^{-\boldsymbol{i}\omega_1x_1}e^{-\boldsymbol{j}\omega_2x_2}d^2\boldsymbol{x}\\ &= \frac{1}{(2\pi)^2}\int_{b_1}^{1/2+b_1} e^{-x_1^2}e^{-\boldsymbol{i}\omega_1x_1}\,dx_1\int_{b_2}^{1/2+b_2} e^{-x_2^2}e^{-\boldsymbol{j}\omega_2x_2}\,dx_2\\ &\quad-\frac{1}{(2\pi)^2}\int_{1/2+b_1}^{1+b_1} e^{-x_1^2}e^{-\boldsymbol{i}\omega_1x_1}\,dx_1\int_{1/2+b_2}^{1+b_2} e^{-x_2^2}e^{-\boldsymbol{j}\omega_2x_2}\,dx_2.\end{aligned} \tag{68}$$

By completing squares, we have

$$\begin{aligned}G_\phi f(\boldsymbol{\omega},\boldsymbol{b}) &= \frac{1}{(2\pi)^2}\int_{b_1}^{1/2+b_1} e^{-(x_1+\boldsymbol{i}\omega_1/2)^2-\omega_1^2/4}dx_1\int_{b_2}^{1/2+b_2} e^{-(x_2+\boldsymbol{j}\omega_2/2)^2-\omega_2^2/4}dx_2\\ &\quad-\frac{1}{(2\pi)^2}\int_{1/2+b_1}^{1+b_1} e^{-(x_1+\boldsymbol{i}\omega_1/2)^2-\omega_1^2/4}dx_1\int_{1/2+b_2}^{1+b_2} e^{-(x_2+\boldsymbol{j}\omega_2/2)^2-\omega_2^2/4}dx_2.\end{aligned} \tag{69}$$

Making the substitutions $y_1 = x_1 + \boldsymbol{i}\,\frac{\omega_1}{2}$ and $y_2 = x_2 + \boldsymbol{j}\,\frac{\omega_2}{2}$ in the above expression, we immediately obtain

$$\begin{aligned}
G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) &= \frac{e^{-(\omega_1^2+\omega_2^2)/4}}{(2\pi)^2} \int_{b_1+\boldsymbol{i}\omega_1/2}^{1/2+b_1+\boldsymbol{i}\omega_1/2} e^{-y_1^2}\, dy_1 \int_{b_2+\boldsymbol{j}\omega_2/2}^{1/2+b_2+\boldsymbol{j}\omega_2/2} e^{-y_2^2}\, dy_2 \\
&\quad - \frac{e^{-(\omega_1^2+\omega_2^2)/4}}{(2\pi)^2} \int_{1/2+b_1+\boldsymbol{i}\omega_1/2}^{1+b_1+\boldsymbol{i}\omega_1/2} e^{-y_1^2}\, dy_1 \int_{1/2+b_2+\boldsymbol{j}\omega_2/2}^{1+b_2+\boldsymbol{j}\omega_2/2} e^{-y_2^2}\, dy_2 \\
&= \frac{e^{-(\omega_1^2+\omega_2^2)/4}}{(2\pi)^2} \left[\left(\int_0^{b_1+\boldsymbol{i}\omega_1/2} (-e^{-y_1^2})\, dy_1 + \int_0^{1/2+b_1+\boldsymbol{i}\omega_1/2} e^{-y_1^2}\, dy_1 \right) \right. \\
&\quad \times \left(\int_0^{b_2+\boldsymbol{j}\omega_2/2} (-e^{-y_2^2})\, dy_2 + \int_0^{1/2+b_2+\boldsymbol{j}\omega_2/2} e^{-y_2^2}\, dy_2 \right) \\
&\quad - \left(\int_0^{1/2+b_1+\boldsymbol{i}\omega_1/2} (-e^{-y_1^2})\, dy_1 + \int_0^{1+b_1+\boldsymbol{i}\omega_1/2} e^{-y_1^2}\, dy_1 \right) \\
&\quad \left. \times \left(\int_0^{1/2+b_2+\boldsymbol{j}\omega_2/2} (-e^{-y_2^2})\, dy_2 + \int_0^{1+b_2+\boldsymbol{j}\omega_2/2} e^{-y_2^2}\, dy_2 \right) \right]. \qquad (70)
\end{aligned}$$

Denote $\operatorname{erf}(x) = \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^x e^{-t^2}\, dt$. Equation (70) can be written in the form

$$\begin{aligned}
G_\phi f(\boldsymbol{\omega}, \boldsymbol{b}) &= \frac{e^{-(\omega_1^2+\omega_2^2)/4}}{(2\sqrt{\pi})^3} \left\{ \left[-\operatorname{erf}\left(b_1 + \frac{\boldsymbol{i}}{2}\omega_1 \right) + \operatorname{erf}\left(\frac{1}{2} + b_1 + \frac{\boldsymbol{i}}{2}\omega_1 \right) \right] \right. \\
&\quad \times \left[-\operatorname{erf}\left(b_2 + \frac{\boldsymbol{j}}{2}\omega_2 \right) + \operatorname{erf}\left(\frac{1}{2} + b_2 + \frac{\boldsymbol{j}}{2}\omega_2 \right) \right] \\
&\quad - \left[-\operatorname{erf}\left(\frac{1}{2} + b_1 + \frac{\boldsymbol{i}}{2}\omega_1 \right) + \operatorname{erf}\left(1 + b_1 + \frac{\boldsymbol{i}}{2}\omega_1 \right) \right] \\
&\quad \left. \times \left[-\operatorname{erf}\left(\frac{1}{2} + b_2 + \frac{\boldsymbol{j}}{2}\omega_2 \right) + \operatorname{erf}\left(1 + b_2 + \frac{\boldsymbol{j}}{2}\omega_2 \right) \right] \right\}. \qquad (71)
\end{aligned}$$

6. Clifford windowed Fourier Transform

In this section, we introduce the Clifford windowed Fourier transform as a generalization of two-dimensional quaternionic Fourier transform to higher dimensions. Let us start with the following definition.

Definition 6.1. *The Clifford windowed Fourier transform of a multivector function $f \in L^2(\mathbb{R}^n; Cl_{0,n})$ with respect to the non-zero Clifford window function $\phi \in L^2(\mathbb{R}^n; Cl_{0,n})$ is given by*

$$\begin{aligned}
G_\phi^c f(\boldsymbol{\omega}, \boldsymbol{b}) &= \int_{\mathbb{R}^n} f(\boldsymbol{x})\, \overline{\phi_{\boldsymbol{b},\boldsymbol{\omega}}(\boldsymbol{x})}\, d^n\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} f(\boldsymbol{x})\, \overline{\phi(\boldsymbol{x}-\boldsymbol{b})} \prod_{k=1}^{n} e^{-\boldsymbol{e}_k \omega_k x_k}\, d^n\boldsymbol{x}, \qquad (72)
\end{aligned}$$

where $\boldsymbol{\omega}, \boldsymbol{b} \in \mathbb{R}^n$ *and* $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \cdots, \boldsymbol{e}_n$ *are the orthonormal vector basis of Clifford algebra* $Cl_{0,n}$ *which satisfy the following rules:*

$$\begin{aligned} \boldsymbol{e}_i\, \boldsymbol{e}_j &= -\boldsymbol{e}_j\, \boldsymbol{e}_j \qquad \textit{for} \quad i \neq j, \quad i, j = 1, 2, 3, \cdots, n \\ \boldsymbol{e}_i^2 &= -1 \qquad\quad \textit{for} \qquad\qquad i = 1, 2, 3, \cdots, n. \end{aligned}$$

We call

$$\phi_{\boldsymbol{\omega},\boldsymbol{b}}(\boldsymbol{x}) = \prod_{k=0}^{n-1} e^{\boldsymbol{e}_{n-k}\omega_{n-k}x_{n-k}} \phi(\boldsymbol{x} - \boldsymbol{b}), \tag{73}$$

a *Clifford windowed Fourier kernel*. Notice that the Clifford windowed Fourier transform for $n = 2$ is identical with the QWFT and that for $n = 1$ is identical with the classical windowed Fourier transform.

7. Conclusion

Using the basic concepts of quaternion algebra and its Fourier transform, we have introduced 2-D quaternionic windowed Fourier transform. Since the multiplication in quaternions is non-commutative, some properties of the classical windowed Fourier transform, such as the shift property, orthogonality relation and reconstruction formula, needed to be modified. We have shown that the construction formula can be extended to higher dimensions using the Clifford Fourier transform. Like quaternion wavelets, which are successfully applied to optical flow, it will be possible to apply the QWFT to optical flow, image features and image fusion in the future.

8. References

Bülow, T. (1999). Hypercomplex spectral signal representations for the processing and analysis of images, Ph.D. thesis, University of Kiel, Germany.

Bülow, T., Felsberg, M. & Sommer, G. (2001). Non-commutative hypercomplex Fourier transforms of multidimensional signals, In: *Geometric Computing with Clifford Algebras: Theoretical Foundations and Applications in Computer Vision and Robotics*, Sommer, G. (Ed.), pp. 187-207, Springer.

E. Bayro-Corrochano, E.; Trujillo, N.; Naranjo, M. (2007). Quaternion Fourier descriptors for the preprocessing and recognition of spoken words using images of spatiotemporal representations, *Journal of Mathematical Imaging and Vision*, Vol. 28, No. 62, pp. 179-190.

Gröchenig, K. (2001). *Foundation of Time-Frequency Analysis*, Birkhäuser, Boston.

Gröchenig, K. & Zimmermann, G. (2001). Hardy's theorem and the short-time Fourier transform of Schwartz functions, *Journal of the London Mathematical Society*, Vol. 2, No. 62, pp. 205–214.

Ghosh, P. K. & Sreenivas, T. V. (2006). Time-varying filter interpretation of Fourier transform and its variants, *Signal Processing*, Vol. 11, No. 86, pp. 3258–3263.

Hitzer, E. (2007). Quaternion Fourier transform on quaternion fields and generalizations, *Advances in Applied Clifford Algebras*, Vol. 17, No. 3, pp. 497–517.

Kemao, Q. (2007). Two-dimensional windowed Fourier transform for fringe pattern analysis: principles, applications, and implementations, *Optics and Laser Engineering*, Vol. 45, pp. 304–317.

Kuipers, J. B. (1999). *Quaternions and Rotation Sequences*, Princeton University Press, New Jersey.

Mawardi, B.; Hitzer, E.; Hayashi, A.; Ashino, R. (2008). An uncertainty principle for quaternion Fourier transform, *Computers and Mathematics with Applications*. Vol. 56, No. 9, pp. 2411–2417.

Mawardi, B.; Hitzer, E.; Ashino, R.; Vaillancourt, R. (2010). Windowed Fourier transform of two-dimensional quaternionic signals, *Applied Mathematics and Computation*. Vol. 28, No. 6, pp. 2366–2379.

Mawardi, B. (2010). A generalized windowed Fourier transform for quaternions, *Far East Journal of Applied Mathematics*, Vol. 42, No. 1, pp. 35–47.

Mawardi, B.; Adji, S.; Zhao, J. (2011). Real Clifford windowed Fourier transform, *Acta Mathematica Sinica*, in press, 2011.

Gröchenig, K. & Zimmermann, G. (2001). Hardy's theorem and the short-time Fourier transform of Schwartz functions, *Journal of the London Mathematical Society*, Vol. 2, No. 62, pp. 205–214.

Sangwine, S. J. & Ell, T. A. (2007). Hypercomplex Fourier transforms of color images, IEEE Transactions on Image Processing, Vol. 16, No. 1, pp. 22–35.

Weisz, F. (2008). Multiplier theorems for the short-time Fourier transform, *Integral Equation and Operator Theory*, Vol. 60, No. 1, pp. 133–149.

Zhong, J. & Zeng, H. (2007). Multiscale windowed Fourier transform for phase extraction of fringe pattern, *Applied Optics*, Vol. 46, No. 14, pp. 2670–2675.

FOURIER
TRANSFORMS
APPROACH TO SCIENTIFIC PRINCIPLES
Edited by Goran S. Nikolić
INTECH

14

High Frame Rate Ultrasonic Imaging through Fourier Transform using an Arbitrary Known Transmission Field

Hu Peng
University of Science and Technology of China
China

1. Introduction

Based on the study of limited array diffraction beams, a High Frame Rate (HFR) imaging method which uses a broadband pulsed plane wave, or array beams transmission field from a linear transducer array to illuminate the area to be imaged has been developed by Jian-yu LU. Echoes from the objects are received with the same transducer as is used in transmission. For each array beam parameter in a certain range, the received signals are weighted with that array beam and are summed up. The summations are Fourier transformed from time domain to frequency domain, and then processed further with the so called ‘‘parameter match’’ to produce the spectrum of the imaging. 2D and 3D images are constructed with inverse Fourier transformer respectively. In this way, the frame per transmission imaging rate is achieved.

Despite its advantages of high frame rate and high signal to noise ratio, the original HFR method has several drawbacks. It can only use the plane wave or the array beam transmission field, and is difficult to be ported for a non-array beam field, such as a cylindrical or spherical wave. Moreover, since the plane wave transmission field illuminates only a narrow area of its own width, the imaged area is quite small, and the only way to widen it up is to steer the transmission beams several times from different angles, which lowers the frame rate. Besides, the array beam field demands a linear transducer and a very complex weighting process.

Therefore, the HFR method needs to allow diffraction wave transmission fields in order to be practically useful. For example, it may use a cylindrical or spherical field and output sector format images like the conventional sector B mode ones, which have contributed a lot in diagnosing myocardial diseases.

In this chapter, an extended HFR method for 2D imaging is proposed. It allows all kinds of transmission field, including the cylindrical one and the spherical one, as well as the plane wave one. It is more general than the original HFR method.

The extended HFR method works mostly like the original one, except that 1, it implements the weight-and-sum process through the Fourier transform; and 2, it iterates for each frequency in a certain range to obtain firstly a coarse image component at that frequency and then the refined one with the information of the transmission field removed. After the iteration the image components are summed up and that is the final image.

In ultrasonic imaging systems, the cylindrical transducer, circle or curve transducer and linear transducer are commonly used. The advantages and disadvantages among them are different. One characteristic of the cylindrical or circle transducer is they can illuminate a sector or pyramidal area of the object. Therefore, in the following section, we extend the HFR method by using a cylindrical wave to illuminate an object. Mathematical formulas are derived and computer simulations are performed to verify the method. The method allows to increase the illumination area by using a transducer of a small footprint, which is important for applications such as cardiac imaging where acoustic window sizes are limited.

This Chapter is organized as follows. Firstly, the HFR ultrasonic imaging system based on the angular spectrum principle is introduced. In the flowing section, this system is extended. The extended HFR method allows all kinds of transmission field. Finally, a high frame rate 2D and 3D imaging system with a curved or cylindrical array is proposed.

2. High Frame Rate ultrasonic imaging system based on the angular spectrum principle

A kind of high frame rate (HFR) 2D and 3D imaging method was developed by Jianyu Lu in 1997. Because only one transmission is required to construct a frame of image, this method can reach an ultra high frame rate (about 3750 volumes or frames per second for biological soft tissues at a depth of 200 mm). In this section, a new HFR method is presented in the view of angular spectrum. Compared with conventional dynamic focusing method which uses delay-and-sum processing and Lu’s HFR method, which uses a kind of special weighting on the received signal, the new method only use the Fourier transform algorithm to construct image. So the system implementation of the method could be greatly simplified. During constructing image, several array beams with different parameters are used as transmitted signals, and the spectra of a frame of image is obtained by synthesizing the image spectrums related to different transmit event. The simulation result shows that the solution not only suppresses the sidelobe of system greatly and obtains the high quality image but also still keeps high frame rate to some extent.

2.1 Theory

The HFR method is based on one transmission event. In order to get the image of the object, the transducer transmits the limited diffraction beams to the object then the same transducer receives the echo signals reflected by the scatters and constructs image by Fourier transform. Fig.2.1 is diagram of the linear array used by HFR method. As the transducer emits the limited diffraction beams, the distribution of the field is

$$p(x,y,z,k) = A(k)e^{ikz} \tag{2.1}$$

where $p(x,y,z,k)$ means the acoustic pressure at the position (x,y,z) under the certain wave numbered k, and $k = \omega / c = 2\pi f / c$, f is frequency and c is acoustic speed. $A(k)$ is the frequency spectrum of exciting signal.

If there are some scatters in the $z = z'$ plane, the pressure for the scatters is

$$\begin{aligned} s(x,y,z_i,k) &= f(x,y,z_i)p(x,y,z_i,k) \\ &= f(x,y,z_i)A(k)e^{jkz_i} \end{aligned} \tag{2.2}$$

where $f(x,y,z_i)$ is the reflective coefficient function of the scatters. Using Fourier transform, we can get another expression for equation (2.2) in the angular spectrum domain:

$$\begin{aligned} S(k_x,k_y,k,z_i) &= \int_{x,y} s(x,y,z_i,k)e^{ik_x x+ik_y y}dxdy \\ &= \int_{x,y} f(x,y,z_i)A(k)e^{ikz_i}e^{ik_x x+ik_y y}dxdy \end{aligned} \tag{2.3}$$

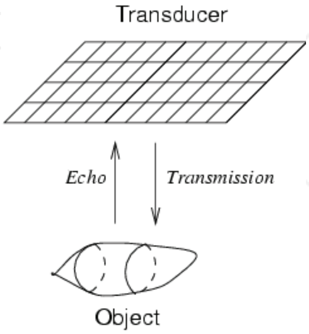


Fig. 2.1. Linear transducer array used in HFR

Because of the reflection of scatters, the echo signal represented by equation (2.2) or equation (2.3) propagates to the surface of the transducer. Based on the angular spectrum, it is easy to get the signal received by the transducer located in the plane $z=0$ in angular spectrum:

$$R(k_x,k_y,k) = T(k)S(k_x,k_y,k,z_i)e^{i\sqrt{k^2-k_x^2-k_y^2}} \tag{2.4}$$

Where $T(k)$ is the frequency response of the transducer. For simplicity, we assume

$$k_z = \sqrt{k^2-k_x^2-k_y^2} \tag{2.5}$$

From equation (2.4) and equation (2.3), the received signal can be represented further as follow

$$R(k_x,k_y,k) = \int_{x,y} f(x,y,z_i)A(k)T(k)e^{ik_z^{'}z_i}e^{ik_x x+ik_y y}dxdy \tag{2.6}$$

where

$$k_z^{'} = k + \sqrt{k^2-k_x^2-k_y^2} \tag{2.7}$$

The equation (2.6) means the signal, which is received by transducer located in the plane $z=0$, is come from echo signal produced by the scatters at the plane $z=z^{'}$. In fact the received signal comes from a lot of planes in the acoustic field. So it should be the summation of $R(k_x,k_y,k)$ over different depth, namely

$$R^{'}(k_x,k_y,k)=\int_z R(k_x,k_y,k_z)dz=$$
$$\iiint_{x,y,z} f(x,y,z_i)A(k)T(k)e^{ik_x x+ik_y y+ik_z^{'} z_i}dxdydz \tag{2.8}$$

Assume

$$F_{BL}(k_x,k_y,k_z^{'})=R^{'}(k_x,k_y,k) \tag{2.9}$$

where the subscript "BL" means "band-limited". From the spectrum $F_{BL}(k_x,k_y,k_z^{'})$, the useful information of the object can be obtained by the inverse Fourier transform:

$$f_{BL}(x,y,z)=F^{-1}(F_{BL}(k_x,k_y,k_z^{'})) \tag{2.10}$$

$F^{-1}(\cdot)$ is a inverse Fourier transform. From equation (2.10), the relationship between image $f_{BL}(x,y,z)$ and the object $f(x,y,z)$ is expressed as

$$\begin{aligned} f_{BL}(x,y,z) &= \int_{x,y,x} dxdydzf(x,y,z) \int_{k_x,k_y,k_z} A(k)T(k)e^{-ik_x(x^{'}-x)-ik_y(y^{'}-y)-ik_z(z^{'}-z)}dk_x dk_y dk_z \\ &= \int_{x,y,z} dxdydzf(x,y,z)p(x-x^{'},y-y^{'},z-z^{'}) \\ &= f(x,y,z)\otimes p(x,y,z) \end{aligned} \tag{2.11}$$

Where the function $p(x,y,z)$ is defined as

$$\begin{aligned} &p(x,y,z)=F^{-1}(P(k_x,k_y,k_z^{'})) \quad and \\ &P(k_x,k_y,k_z^{'})=[A(k)T(k)]_{k_x,k_y,k_z^{'}} \end{aligned} \tag{2.12}$$

From the equation (2.11) and (2.12), we can see that if the size of the aperture is infinite, the image is the result of the convolution between object reflection coefficient and the function $p(x,y,z)$. So $p(x,y,z)$ is the point spread function (PSF) of the imaging system, which is determined by the excited signal and the frequency response of transducer. Obviously under the condition that k is infinite and $A(k)T(k)$ is equal to one, $p(x,y,z)$ turns to be Dirac delta function and $f_{BL}(x,y,z)$ is the object $f(x,y,z)$. Generally, The bandwidth of $T(k)$ and $A(k)$ is limited and $p(x,y,z)$ is pulse in three dimension. So $f(x,y,z)$ only presents some useful information of the object.

2.2 Simulation results

In equation (2.8), function $R^{'}(k_x,k_y,k)$ is the received signal by transducer in the domain of frequency spectrum k and the domain of space spectrum (k_x,k_y). In practice, the received signal is expressed by $r^{'}(x_e,y_e,t)$ in the domain of time t and the domain of space (x,y). So in the first step the algorithm 3D Fourier transform is used in order to change $r^{'}(x_e,y_e,t)$ to $R^{'}(k_x,k_y,k)$. It means the weighting process can be realized by Fourier transform over transducer surface and the time parameter.

But there is still a little difference between Fourier transform and the weighting process. First we know the number of wave k is positive for constructing imaging, but the result of

Fourier transform contains information of positive and negative k. Secondly the results of Fourier transform include the information part which corresponds to $|k_x| > k_{\max}$ and $|k_y| > k_{\max}$, and obviously the part has no physics meaning for the weighting result $R'(k_x, k_y, k)$.
Considering the two condition, $R'(k_x, k_y, k)$ can be obtained from the modified Fourier transform results of the received signal $r'(x, y, t)$ under the condition below:

$$R'(k_x, k_y, k) = \begin{cases} 0 & k < 0 \quad or \quad |k_x| > k_{\max} \quad or \quad |k_y| > k_{\max} \\ F(r(x, y, t)) & otherwise \end{cases} \tag{13}$$

$F^{-1}(\cdot)$ is a inverse Fourier transform. Based on the study above, the system of HFR method can be simplified into Fig.2.2.

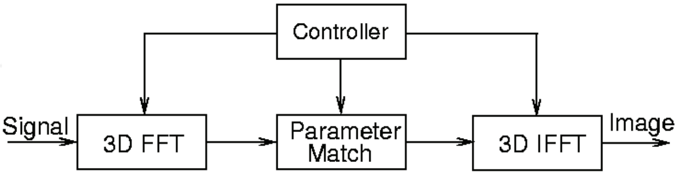


Fig. 2.2. The new solution to the realization of HFR method, which consists of three parts mainly, two 3D FFT and parameter match

In Fig.2.2 the system consists of three parts, two FFT chips and one parameter match chip. The received signal is imputed into the first FFT chip to get the signal $R'(k_x, k_y, k)$, then processed by parameter match unit which changes $R'(k_x, k_y, k)$ to the spectrum $F_{BL}(k_x, k_y, k_z')$ of the image, and at last step the image $f_{BL}(x, y, z)$ is obtained from $F_{BL}(k_x, k_y, k_z')$ by the second FFT chip.

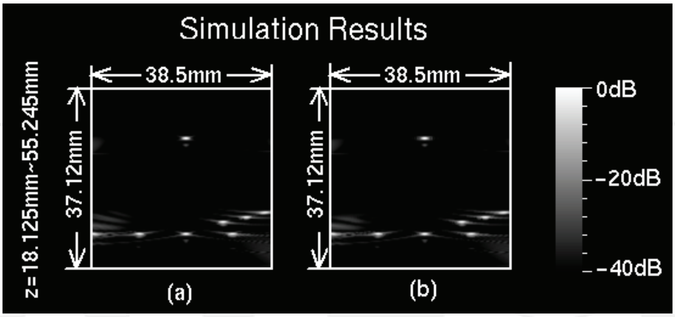


Fig. 2.3. Simulation of 2D B-mode image with log compressed over 40db scale. (a) is obtained by original HFR method, and (b) is obtained by the new method

To verify the new process, a simulation in two dimension was performed to construct image by the HFR method. In the simulation, the phantom consists of eight point scatters objects. The linear array transducers is a 2.5 MHz array of 64 elements and a dimension of 38.4 mm with an inter-element space of 0.6 mm. Two-way (pulse-echo) spectra of the arrays are assumed to be proportional to the Blackman window function with a fractional bandwidth of about 81% that is typical for a modern array. The simulation results shows in Fig.2.3 and Fig.2.4. In the figures, Fig.2.3a is obtained by original HFR process (*IEEE Trans on UFFC*, 44(4), 1997, pp. 839-856) and Fig.2.3b is obtained by the new process. It can be seen that the two results are the same.

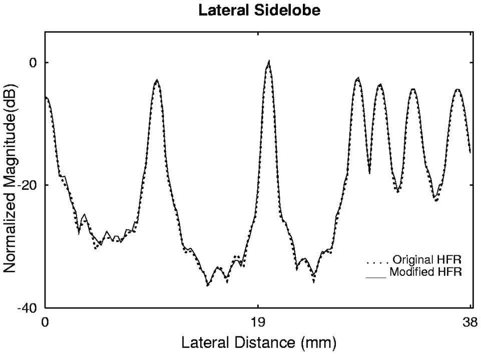


Fig. 2.4. The sidelobe along lateral direction

2.3 Conclusion

This section presents a theory analysis, which is based on angular spectrum principle, to simplify the HFR imaging system presented by Lu. Besides a new imaging mode is proposed, which use several transmission events to synthesize the image. In every transmission event, array beam with different parameters is used as excited signal. The constructed image has very high resolution and contrast, and meanwhile the imaging system still hold high frame rate to some extent.

3. Construction of High Frame Rate ultrasonic images with Fourier transform in any kind of acoustic field

In HFR method, a plane wave was used to illuminate an area for either 2D or 3D imaging. The drawback of this method is that the area illuminated is only as wide as the size of the aperture of the array transducer. In this section, a generic HFR method is developed. 2D high frame rate images can be constructed using the Fourier transform with a single transmission of an ultrasound pulse from an array under different transmission filed as long as the transmission filed is known. To verify our theory, computer simulations have been performed in the non-plane wave field. The field is cylindrical field defined by zero order Hankle function and produced by a linear array. The image with sector format and lower sidelobes is obtained. The simulation results are consistent with our theory.

3.1 Theories

For simplicity, we discuss our new method in the two dimension (2D). Let us assume that there is a linear array at the position *z*=0 (Fig.3.1), and the transmitted field is $p(x,z,k)$ where $k = 2\pi f / c$. f means frequency and c is acoustic speed. If there is a scatter at the position which the coordination is (*x*,*z*), and the reflection coefficient is $f(x,z)$, The echo signal, which object scatter reflects, is as follow:

$$f^{'}(x,z,k)=f(x,z)p(x,z,k)T(k) \tag{3.1}$$

where $T(k)$ is spectra, which is related to the spectrum of excited signal and the frequency response of the transducer.

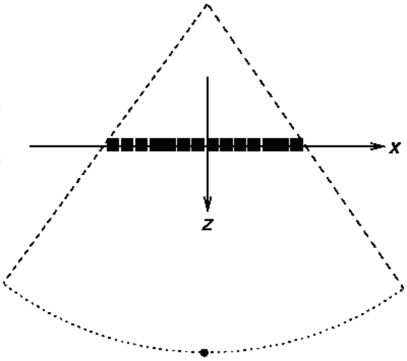


Fig. 3.1. The diagram of the transducer

The received signal to an element of the transducer can be obtained by the equation (3.2)

$$r(x_e,z_e=0,k)=\int_{x,z} f^{'}(x,z,k)H(x,z;x_e,z_e=0,k)dxdz \tag{3.2}$$

Here $(x_e,z_e=0)$ is the position's coordination of an element of the transducer, k is wave number, $H(x,z;x_e,z_e=0,k)$ is transmission function, which is determined by Rayleigh-Sommerfield diffraction theory and presents the relationship between source point (x,y) and observed point (x_e,y_e). The function $r(x_e,z_e=0,k)$ means the received signal echoed by the object to be imaged.

Under a certain frequency component k, Using signal $e^{jk_xx_e}$ to weight the received signal

$$\begin{aligned} R(k_x,k_z) &= \int_l r(x_e,z_e=0,k)e^{jk_xx_e}dx_e \\ &= \int_l dxdzf^{'}(x,z,k)\int_l e^{jk_xx_e}H(x,z;x_e,z_e=0,k)dx_e \end{aligned} \tag{3.3}$$

If the *l*, the size of the transducer is infinite, the result of the integrate over *l* is

$$\int e^{jk_xx_e}H(x,z;x_e,z_e=0,k)dx_e=e^{jk_xx+jk_yy} \tag{3.4}$$

Here

$$k_z^2=\sqrt{k^2-k_x^2} \tag{3.5}$$

This means that the transducer at the position $z=0$, which is excited by the weight signal $e^{jk_xx_e}$, produces the plane wave $e^{jk_xx+k_yy}$ in the field. From the equation (3.4) and (3.4), we have

$$R(k_x,k_z)=\int_{x,z} f^{'}(x,z,k)e^{ik_x x+ik_z z}dxdz \tag{3.6}$$

Make inverse Fourier transform for the weighted signal $R(k_x,k_z)$ to get the imaging under the frequency component k and the transmitted filed $p(x,z,k)$;

$$f^{'}(x^{'},z^{'},k)=F\{R(k_x,k_z)\} \tag{3.7}$$

where $F\{\cdot\}$ is Fourier transform. Remove the information of transmitted filed.

$$f^{''}(x^{'},z^{'},k)=f^{'}(x^{'},z^{'},k)p^{-1}(x^{'},z^{'},k) \tag{3.8}$$

Sum the imaging $f^{''}(x^{'},z^{'},k)$ over all frequency components to get final imaging.

$$f^{'''}(x^{'},z^{'})=\sum_k f^{''}(x^{'},z^{'},k) \tag{3.9}$$

We can prove that equation (3.9) is a good approximation of the object function $f(x,z)$. Especially when the transmitted filed is plane wave, it is the original HFR method.

3.2 Simulation results

From the theoretical analysis above, the simulation process is divided into several steps as follows:

1. According to the transmitted signal and the boundary of the transducer, calculate the distribution of the acoustic filed $p(x,z,k)$;
2. According to the equation (3.3) and (3.5), using signal $e^{jk_x x_e}$ to weight the received signal $r(x_e,z_e=0,k)$ to get the spectrum signal $R(k_x,k_z)$;
3. Using equation (3.7) to get imaging $f^{'}(x,z,k)$, which is at the frequency component *k*;
4. Remove the imaging's phase caused by $p(x,z,k)$ according to equation (3.8) to get signal $f^{''}(x,z,k)$;

Sum the imaging $f^{''}(x,z,k)$ for all frequency components to get final imaging $f^{'''}(x,y)$ by equation (3.9);

Fig.3.1 shows the block diagram of the experiment. Assume the transmitted field is cylindrical function determined by zero order Hankel function. By adjusting the phase and amplitude of excited signal over the linear transducer according to the equation (3.10), the linear transducer produces the transmission filed as follows:

$$p(x,z,k)=\frac{e^{jk\sqrt{x^2+z^2}}}{\sqrt{k}\sqrt{x^2+z^2}} \tag{3.10}$$

The number of the transducer arrays is 128. The length of the transducer is 37mm. The central frequency of the transducer is 2.5MHz and the bandwidth is about 80 percent of the central frequency. The frequency response function *T*(*k*) is assumed to be Blackman window, which is adopted in most literature's simulation condition. The image is produced at the depth which *z* is equal to 50mm. Fig.3.2 is the result of the simulation for one scatters located at (0,50). Fig3.2.a is the images of the object, which is Log compressed over 40db. In order to observe the sidelobes, Fig3.2.b gives the plot line along lateral direction. Fig.3.2c gives the plot line along axial direction. From the figures, we can see that the sidelobes are about below -40db, which the resolution is about 1.2mm in the lateral direction, and 0.8 in the axial direction. Fig.3.3 is another result of the simulation for seven scatters. The scatters are on the part of a circle, among which the central point scatters is equal to 50mm far away from the surface of transducer. Fig3.3.a is the imaging of the object, which is Log compressed over 40db. Fig3.3.b shows the sidelobes along *x* direction. Though the size of array is only 37mm, the distance in the images along lateral direction from left point scatter to the right scatter is about 58mm, which is larger than the size of the transducer. Obviously the imaging is impossible to be obtained for original HFR method.

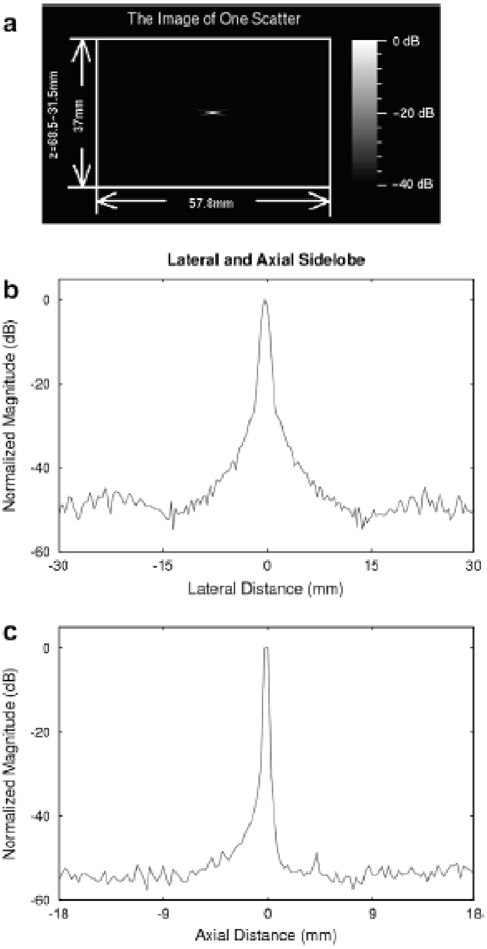


Fig. 3.2. Simulation results of Fig1. 1's one scatterer. (a). The 40db log compressed image and (b) and (c) sidelobes along the lateral and the axial direction

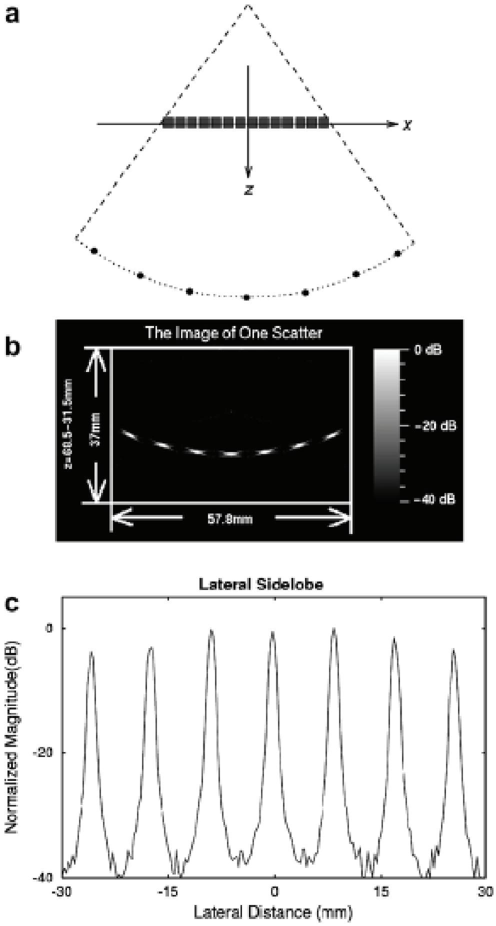


Fig. 3.3. Simulation results of seven scatters. (a). The diagram of the transducer and seven scatterers, (b). the 40 db log compressed image, (c). the sidelobes along x direction

3.3 Conclusion

Though the method is analyzed in the two dimensions, it can be obviously used in the three dimensions. So the method gives an effective way to construct images with sector form (2D) and pyramidal form (3D) by the linear array based on the Fourier transformer. Like the original HFR method, the system can construct images with only one time transmission, and the quality of the imaging is high. Compared to the original Fourier transform method, it is effective in any kind of acoustic field, though the principle of the method and the original HFR method is the same. In original HFR method, the kernel function of the Fourier transform contains the information of the transmission field because the transmission waves and weighting waves are the same kind beams, which belong to array beams. In our new method we extend the kind of transmission field from plane wave or array beams to other kind wave, such as cylindrical wave. As the weighting signal is not the same form as the transmission filed, it is difficult to combine the transmission filed and weighted signal together in the kernel function of the Fourier transform. As a result we have to repeat the Fourier transform process under different frequency component and make the summation over different frequency results. So the shortcoming of the method is obvious compared to the original HFR method, namely its quantity of the computation is high. If some kind of quick arithmetic is found, the method will be more effective in practice.

Because the original HFR method assumes the transmission filed to be the plane wave or array beams although it is impossible actually due to diffraction property in physics, the assumption makes results obtained by original HFR method a little disturbed when the distance between object and transducer is some large. For our method if the transmission filed is pre-known exactly by some method, such as simulation or measurement, the better results can be obtained because the new method can cancel the effects of transmission filed.

4. High Frame Rate 2D and 3D imaging with a curved or cylindrical array

The cylindrical transducer, circle or curve transducer and linear transducer are commonly used in ultrasonic imaging system. The advantages and disadvantages between them are different. One characteristic of the cylindrical or circle transducer is the transducer can illuminate a sector or pyramidal area of the object. The scanning format is primarily useful for cardiac imaging to avoid interference from the ribs. Since this kind of transducer is nonlinear transducer, the method of constructed images is a little different from the linear transducer's method.

4.1 Theoretical preliminaries

In the section, a new imaging method (Fourier method and radial matched filter) for a pulse system will be developed and formulas for construction of 2D and 3D images will be derived with zero order Hankle function.

4.1.1 3D images construction

To simplify the analysis, we assume that the sampling of the array along each direction is regarded as continuous, our results, based on this assumption, should closely approximate that of a sampled aperture as long as the Nyquist criterion is met to avoid spatial aliasing . A sufficient condition for this criterion to be satisfied is a half-wavelength spacing of elements along the arrays. For simplicity, we will also neglect the diffraction patterns of the individual elements; they are assumed to be behaving as point sources and receivers. Although we assume continuously sampled, the simulation results shows similar principles can be applied to arrays of discrete elements of finite size.

For the generality, we discuss the method in three-dimension in the cylindrical coordinate system. Fig.4.1 shows a cylindrical transducer. Though the filed produced by cylindrical transducer is much more complex than the plane wave, we still can get simple form under some reasonable assumption. The simplest mode of the filed form produced by the cylindrical transducer is zero order Hankle function. If the *kr* is relatively large, the acoustic pressure, which is presented by zero order Hankle function, can be estimated by:

$$p(r,k) \approx A(k)\frac{e^{ikr}}{\sqrt{kr}} \tag{4.1}$$

Where k is wave number, $r=\sqrt{x^2+y^2}$ represents radial coordinate, $A(k)$ is related to the spectrum of the signal and the response of the transducer frequency and can be presented by the Blackman windows .

Based on the Rayleigh-Sommerfeld diffraction theory, the received signal for an element for all scatterers is easily given by

$$R(k,\theta_e,z_e)=\frac{1}{i\lambda}\times$$
$$\iiint\limits_V rdrd\theta dzf(r,\theta,z)p(r,k)T(k)\frac{e^{ik\sqrt{r^2+r_e^2-2rr_e\cos(\theta_e-\theta)+(z_e-z)^2}}}{\sqrt{r^2+r_e^2-2rr_e\cos(\theta_e-\theta)+(z_e-z)^2}}\cos(\vec{n}_1,\vec{n}_2) \tag{4.2}$$

where λ is wavelength, and $\lambda=2\pi/k$. θ is azimuthal angle, z is axial axis, which is perpendicular to the plane defined by r and θ. $\sqrt{r^2+r_e^2-2rr_e\cos(\theta_e-\theta)+(z_e-z)^2}=|\vec{r}_e-\vec{r}|$ is the distance between the scatterer and the transducer element, where $\vec{r}_e=(r_e,\theta_e,z_e)$ is the coordinate of transducer element, $\vec{r}=(r,\theta,z)$ is the coordinate of the scatters. $f(r,\theta,z)$ is the function of reflection coefficient of the object, $R=(k,\theta_e,z_e)$ means the received signal for the element at $\vec{r}_e$, $\vec{n}_1$ is an unit vector which direction is from $(0,\theta_e,z_e)$ to (r_e,θ_e,z_e), and $\vec{n}_2$ is another unit vector which the direction is from (r,θ,z) to (r_e,θ_e,z_e). Our objective is to image the reflectivity function $f(r,\theta,z)$, which is the inverse problem of equation (4.2).

After some mathematical manipulations, one can easily find the following

$$R(k,\theta_e,z_e)=\frac{\sqrt{k}}{i2\pi}A(k)T(k)\times$$
$$\iiint\limits_V drd\theta dzf(r,\theta,z)\frac{e^{ikr+ik\sqrt{r^2+r_e^2-2rr_e\cos(\theta_e-\theta)+(z_e-z)^2}}}{r^2+r_e^2-2rr_e\cos(\theta_e-\theta)+(z_e-z)^2}\sqrt{r}(r\cos(\theta_e-\theta)-r_e) \tag{4.3}$$

Even if Equation.(4.3) is similar to Equation.(4.2), it is still different because the equation (4.3) includes cylindrical field information, and based on which the image can be constructed by only one transmission. …. It is obvious that equation (4.3) is the convolution form over θ,z , so we have

$$R(k,\theta_e,z_e)=\int_{r_e}^{\infty}drf(r,\theta,z)*_{\theta,z}h(k,r,\theta,z) \tag{4.4}$$

where $*_{\theta,z}$ means convolution operator over θ,z. $h(k,r,\theta,z)$ is defined as

$$h(k,r,\theta,z)=\frac{\sqrt{k}}{i2\pi}A(k)T(k)\frac{e^{ikr+ik\sqrt{r^2+r_e^2-2rr_e\cos\theta+z^2}}}{r^2+r_e^2-2rr_e\cos\theta+z^2}\sqrt{r}(r\cos\theta-r_e) \tag{4.5}$$

The equation (4.5) is system transform function, which transforms the object function $f(r,\theta,z)$ to the received signal $R(k,\theta_e,z_e)$. In the study, the exact form is used to construct image instead of an approximate form.

From equation (4.4), using Fourier transform theory, we have another expression in spectrum k_θ,k_z domain.

$$\tilde{R}(k,k_\theta,k_z) = F_{\theta,z}\{R(k,\theta,z)\} = \int_{r_e}^{\infty} dr\tilde{F}(r,k_\theta,k_z)\tilde{H}(k,r,k_\theta,k_z) \tag{4.6}$$

where $\tilde{R}(k,k_\theta,k_z)$ is the Fourier transform of $R(k,\theta,z)$ in terms of θ,z, $\tilde{H}(k,r,k_\theta,k_z)$ is the Fourier transform of $h(k,r,\theta,z)$. It is clear that Eq. (4.6) establishes a relationship between the Fourier transform of measured signal and the Fourier transform of object function. However, this relationship is established through the integration over *r*, which is the radial axis of the object function in the cylindrical coordinates. In the following section, we will use some mathematical manipulation to find and establish a more direct relationship between the Fourier transforms of these two functions.

From (4.5), it is clear that the filter function $\tilde{H}(k,r,k_\theta,k_z)$ contains an oscillating term of k and r. This term may play a role of decreasing the integration in terms of either k or r. If such oscillation term can be removed under some conditions, we may be able to construct images. Multiplying the conjugate of $\tilde{H}(k,r,k_\theta,k_z)$, $\tilde{H}^*(k,r',k_\theta,k_z)$, to both sides of (4.6), we have:

$$\tilde{R}(k,k_\theta,k_z)\tilde{H}^*(k,r',k_\theta,k_z) = \int_{r_e}^{\infty} dr\tilde{F}(r,k_\theta,k_z)\tilde{H}(k,r,k_\theta,k_z)\tilde{H}^*(k,r',k_\theta,k_z) \tag{4.7}$$

Integrating over wave number k for on both side of (4.7), one obtains:

$$\begin{aligned}\tilde{R}'(r',k_\theta,k_z) &= \int_{-\infty}^{\infty} dk\tilde{R}(k,k_\theta,k_z)\tilde{H}^*(k,r',k_\theta,k_z)\\ &= \int_{r_e}^{\infty} dr\tilde{F}(r,k_\theta,k_z)G(r,r',k_\theta,k_z)\end{aligned} \tag{4.8}$$

where

$$G(r,r',k_\theta,k_z) = \int_{-\infty}^{\infty} dk\tilde{H}(k,r,k_\theta,k_z)\tilde{H}^*(k,r',k_\theta,k_z) \tag{4.9}$$

Equation (4.8) establishes a relationship between the measured signal, $\tilde{R}'(r',k_\theta,k_z)$, which is known, and the Fourier transform of the object function, $\tilde{F}(r,k_\theta,k_z)$. After inverse Fourier transform over k_θ,k_z for equation (4.8), we have

$$f'(r,\theta,z) = F^{-1}_{k_\theta,k_z}\left\{\tilde{R}'(r,k_\theta,k_z)\right\} = \int_{r_e}^{\infty} drf(r,\theta,z) *_{\theta,z} g(r,r',\theta,z) \tag{4.10}$$

What is the relationship between $f'(r,\theta,z)$ and $f(r,\theta,z)$? In order to answer the question, let us consider the function $g(r,r',\theta,z)$, which is inverse Fourier transform of $G(r,r',k_\theta,k_z)$ in (4.9) and see its role in the constructed images. Fig4.2 and Fig4.3 shows the simulation of the distribution of the function $g(r,r',\theta,z)$. From the results of the simulation, two important properties can be seen. First it is clear that $g(r,r',\theta,z)$ peaks sharply only when r is equal to r'. Second, $g(r,r',\theta,z)$ is symmetric to both r' and r. This gives us the following approximate relationship:

$$\left|g(r,r',\theta,z)\right| \approx \left|g'(r-r',\theta,z)\right| \tag{4.11}$$

From equation (4.10) and (4.11), we can reasonably assume:

$$f^{'}(r,\theta,z)=\int_{r_e}^{\infty}drf(r,\theta,z)*_{\theta,z}g^{'}(r-r^{'},\theta,z)=f(r,\theta,z)*_{r,\theta,z}g^{'}(r,\theta,z) \tag{4.12}$$

Obviously the function $g^{'}(r,\theta,z)$ can be treated as the point spread function of the imaging system. For a perfect imaging system, $g^{'}(r,\theta,z)$ is a Dirac-Delta function in space domain (r,θ,z). So the image $f^{'}(r,\theta,z)$ is object function $f^{'}(r,\theta,z)$. From the simulation in Figs.4.2 and 4.3 we see that $f^{'}(r,\theta,z)$ has a sharp point spread function in the space domain (r,θ,z) (please note that only one-way PSF is shown in Figs. 4.2 and 4.3 for high frame rate imaging), which is similar to the Derac-Delta function. Based on the analysis above, we have the answer for the question above:

$$f^{'}(r,\theta,z)\approx f(r,\theta,z) \tag{4.13}$$

4.1.2 2D image construction

A 2D image in any orientation (including both B-mode and C-mode images) can be readily obtained from 3D images with equation (4.6), (4.8) and (4.10). However, 3D imaging is more complex and generally requires more computation. In the following, the formulas that are simplified from (4.6), (4.8) and (4.4.10) and are suitable for conventional B-mode imaging and a C-mode imaging will be derived. In B-mode imaging, objects are assumed to be independent of z (along the axial direction) and in C-mode imaging, objects are assumed to be a thin layer located at a radial direction $r=r_0$ away from the transducer, where r_0 is a constant.

C-mode imaging assumes the object function $f(r,\theta,z)$ represents a thin layer that is in parallel with the surface of the cylindrical transducer. This is indicated mathematically as follows:

$$f^{(C)}(\theta,z)=f(r,\theta,z)\delta(r-r_0) \tag{4.14}$$

where δ is the Dirac-Delta function and $f^{(C)}(\theta,z)$ is a transverse object function. Thus (4.8) can be simplified as:

$$\tilde{R}^{'}(r_0,k_\theta,k_z)=\int_{-\infty}^{\infty}\tilde{R}(k,k_\theta,k_z)\tilde{H}^{*}(k,r_0,k_\theta,k_z)dk \tag{4.15}$$

Following the discussion in 3D case above, we obtain the constructed image:

$$f^{'}(r_0,\theta,z)=F^{-1}_{k_\theta,k_z}\left\{\tilde{R}^{'}(r_0,k_\theta,k_z)\right\}\approx f(r_0,\theta,z) \tag{4.16}$$

To summarize, the steps to construct a C-mode image are as follows: Perform a 2D Fourier transform of received echo signals to get the spectrum in terms of θ and z, multiply the results with $\tilde{H}^{*}(k,r_0,k_\theta,k_z)$ and integrate over k, and then the images is constructed with an inverse Fourier transform over both k_θ and k_z.

A similar approach can be used to construct a 2D B-mode image, i.e., an image along both r and θ dimension with a fixed z (or object is uniform along z). Under this condition,

$f(r,\theta,z)$ can be replaced with $f(r,\theta)$. For simplicity, without loosing generality, we assume $z=0$. From (4.4), we have:

$$R(k,\theta_e)=\int_{r_e}^{\infty} f(r,\theta)\otimes_\theta h(k,r,\theta)dr \tag{4.17}$$

After the Fourier transform of $R(k,\theta_e)$ in terms of θ_e, from (8), we obtain:

$$\tilde{R}'(r',k_\theta)=\int_{-\infty}^{\infty}\tilde{R}(k,k_\theta)\tilde{H}^*(k,r',k_\theta)dk \tag{4.18}$$

Instead of 2D, 1D inverse Fourier transform is used to construct the image:

$$f'(r,\theta)=F_{k_\theta}^{-1}\left\{\tilde{R}'(r,k_\theta)\right\}\approx f(r,\theta) \tag{4.19}$$

4.2 Simulation results

The simulation of pulse-echo imaging is performed in the three-dimension. In the simulations, Rayleigh-Sommerfeld diffraction formula is used. The parameters of the cylindrical transducer for the simulation are as follows (Fig.4.1): The transducer is broadband and its center frequency is 1.5MHz. The bandwidth of the transducer is about 81% of the center frequency. [assume that the combined transmit and receive transfer function is proportional to the Blackman window function], The background medium is assumed to be water that has a speed of sound of 1500m/s given the wavelength of 1 mm at the central frequency. The objects are assumed to be composed of point scatters. The radius of the cylindrical transducer is 40mm; the range of the angle is from -45^0 to 45^0, and the range of transducer along the *z*-axis is from –25mm to 25mm. The inter-element distance of the array transducer is assumed to be 0.7087^0 along angle direction and 0.3927 mm along z direction. So the element number of the discrete transducer is 128×128. In transmission, all the array elements are connected electronically to transmit the cylindrical wave, which is approximated by zero order Hankle function. Echoes from object (Fig.4.1) are received with the same array and processed to construct imaging by the several steps below based on the previous analysis.

1. Do Fourier transform of received signal (see (4.6)) in terms of θ and z, i.e., $\tilde{R}(k,k_\theta,k_z)=F_{\theta,z}\left\{R(k,\theta,z)\right\}$
2. Multiply the results with the known function, $\tilde{H}^*(k,r',k_\theta,k_z)$ (see (4.7)).
3. Integrate the result over k according to (4.8).
4. Performing an inverse Fourier transform according to (4.10) to construct image.

The objects used for the construction are shown in Fig.4.1 and Fig.4.4 and are composed of either a single point scatter or nine point scatters which form a cross shape in the plane defined by $r-\theta$ and the plane defined by $z-\theta$. The geometry center of the object is located at $(r,\theta,z)=(r_0,0,0)$. Results of the pulse-echo images are given in Fig.4.2, Fig.4.3 and Fig.4.5, Fig.4.6

Fig.4.2d, e and f show the image for one scatter which position is $(r,\theta,z)=(90mm,0,0)$. Because the radius of the transducer is 40*mm*, the nearest distance between the scatter and the surface of transducer is 50*mm*. To see the sidelobes of the constructed images, line plots of the single point scatter along *r*, *z* and θ direction in the $z-\theta$ plane, $r-\theta$ plane and $z-r$ plane are shown in Fig.4.3. From the results it can be seen that imaging is very similar to the result of PSF. This means that the approximations (4.10) due to a finite temporal bandwidth and limited spatial Fourier-domain coverage that are typical in medical ultrasonic imaging do not significantly affect the equality of constructed images in terms of spatial resolutions, sidelobes, and contrast.

Fig.4.5 shows the images for nine scatters. The nearest distance between surface of the transducer and the geometry center of the object is chosen to be 50mm (r_0 =90mm), 100mm (r_0 =140mm) and 200mm (r_0 =240mm), respectively. An interesting phenomenon is that simulation shows the sidelobes and resolution of the images of the object, of which the geometry center is at different distance, is nearly the same (Fig.4.6). The reason is that the transmitted field keeps the same form as the theory prediction along *r* direction if the filed is cylindrical wave. Though the side lobe and resolution in the *r* and θ direction is the same for the larger area, the sidelobe rises and the resolution is lower in the regular coordinate system when *r* becomes larger because of the relationships $z = r\cos(\theta), x = r\sin(\theta)$.

4.3 Conclusion

In this section a new 3D images system in cylindrical coordinate has been developed with cylindrical wave beams (zero order Hankle function). This computation is much less than conventional delay and sum method, so the method has a potential to achieve a high image frame rate and can be implemented with relatively simple inexpensive hardware because the FFT and IFFT algorithm can be used. Computer simulation with the new method has been carried out to construct 3D images. Though the aperture geometry of the transducer is only part of a cylinder, and the transmitted filed is not exact zero order Hankle function, the results of the simulation still match theoretical prediction. So the new imaging method is robust and is not sensitive to various limitations imposed by practical system. In addition, though the discussion above is mainly for 2D cylindrical transducer in three-dimension in the cylindrical coordinate system, the method can be used directly for the 1D curve transducer in two-dimension in the polar coordinate system obviously.

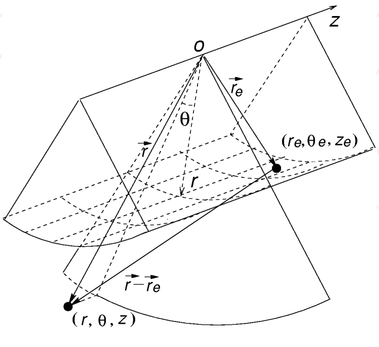


Fig. 4.1. Transducer in the cylindrical coordinate system. The radius of the transducer is $r_e = 40mm$, and the range of axial axis z_e is from –25mm~25mm, the range of azimuthal angle θ_e is from -45^0 to 45^0 . There are $N_z N_\theta = 128\times 128$ elements

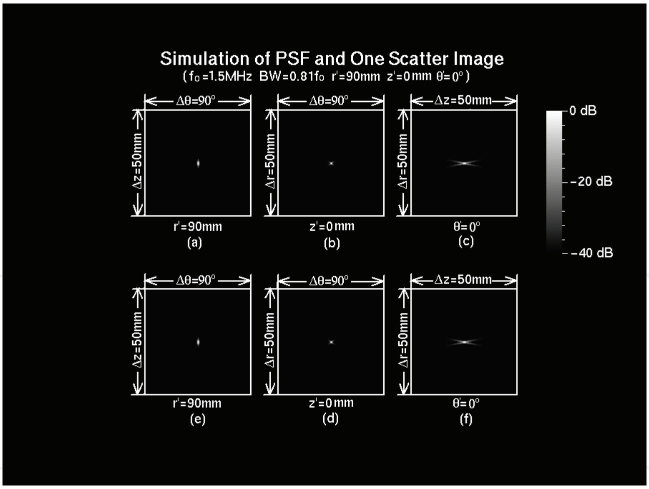


Fig. 4.2. Calculated PSF, $g^{'}(r-r^{'},\theta,z)=F^{-1}_{k_\theta,k_z}\left\{G^{'}(r-r^{'},k_\theta,k_z)\right\}$, and the image constructed from one scatterer (Fig. 4.1) where $r^{'}=90mm$. (a) shows the distribution of PSF in the plane $(\theta,z)_{r=90mm}$.(b) shows the distribution of PSF in the plane $(r,\theta)_{z=0}$. (c) shows the distribution of PSF in the plane $(r,z)_{\theta=0}$. (d) shows the constructed image of the scatter in the plane $(\theta,z)_{r=90}$. (e) shows the image of the scatter in the plane $(r,\theta)_{z=0}$. (f) shows the image of the scatter in the plane $(r,z)_{\theta=0}$. The images are log compressed over 40db

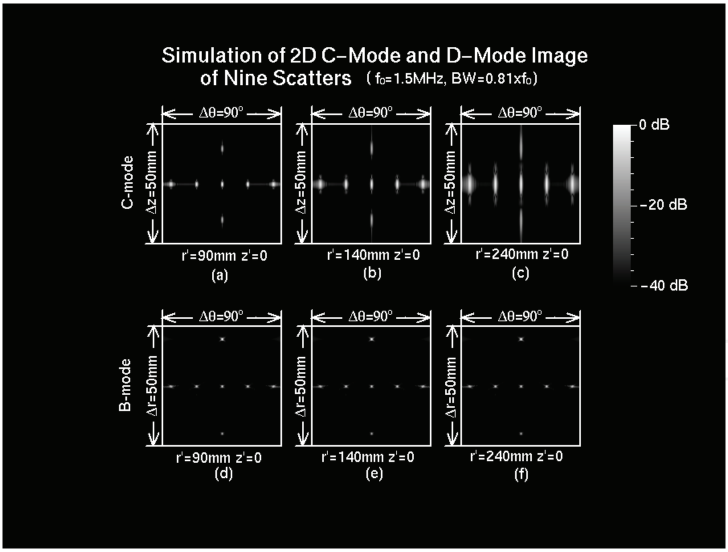


Fig. 4.5. The images for Fig.4.4. (a), (b) and (c) show images where the geometry center is equal to 90mm, 140mm and 240mm by C-mode($z-\theta$ plane) respectively. (d),(e) and (f) show images where the geometry center is equal to 90mm, 140mm and 240mm by B-mode($r-\theta$ plane), respectively. The images are log compressed over 40db

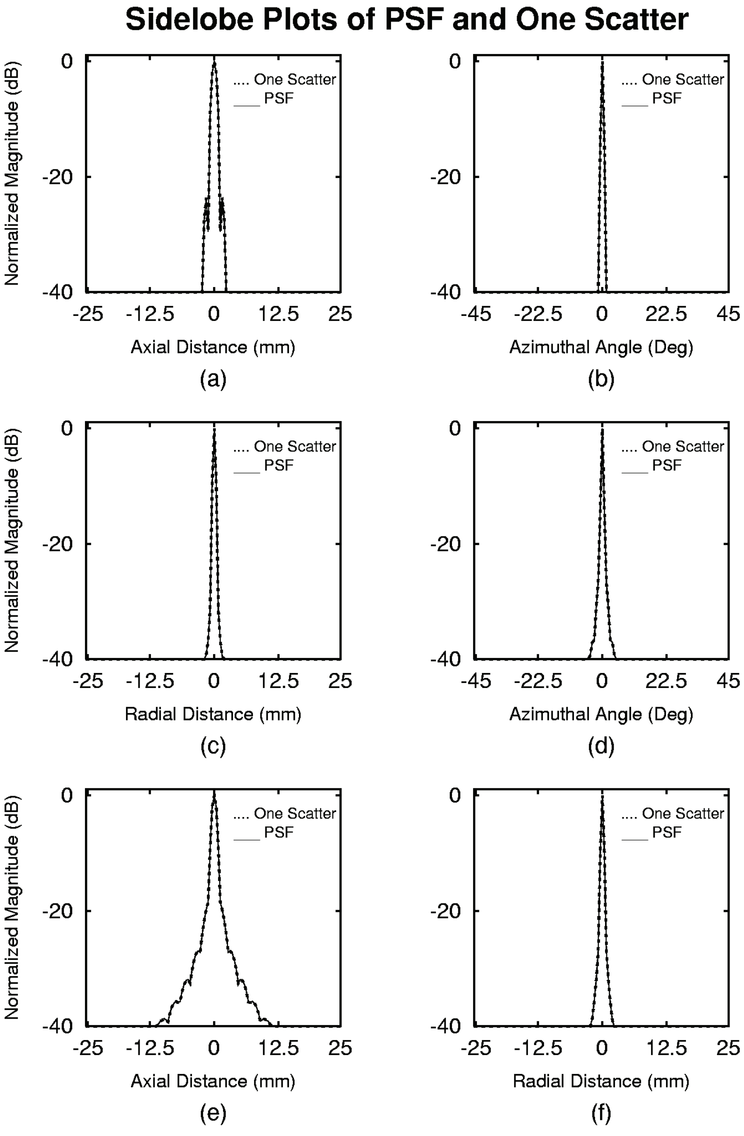


Fig. 4.3. Sidelobe of PSF and constructed image of one scatter in Fig.4.2. (a) and (b) show sidelobe in the plane $(\theta,z)_{r=90}$ along z direction and θ direction, respectively. (c) and (d) show sidelobe in the plane $(r,\theta)_{z=0}$ along r direction and θ direction, respectively. (e) and (f) show sidelobe in the plane $(r,\theta)_{z=0}$ along z direction and r direction, respectively

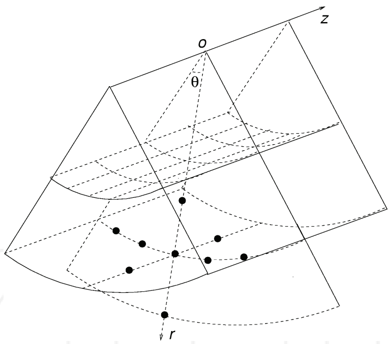


Fig. 4.4. The objects used for construction of images, which contains nine scatters and the geometry center is $(r,\theta,z) = (90,0,0)$, $(140,0,0)$ or $(240,0,0)$

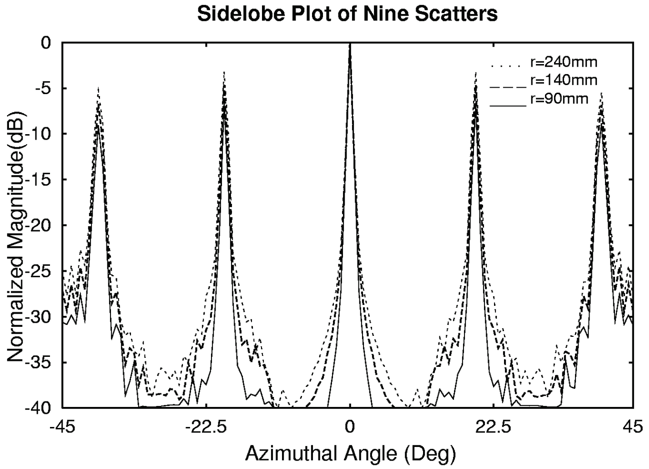


Fig. 4.6. Plots line shows sidelobe for Fig.4.5 along θ direction in the B-mode image

5. References

Jian-yu Lu. (1997). 2D and 3D high frame rate imaging with limited diffraction beams, *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, Vol.44, No.4, (1997), pp. 839-856, ISSN 0885-3010

Hu, Peng; Angning Yu. (2008). High Frame Rate Ultrasonic Imaging through Fourier Transform Using an Arbitrary Known Transmission, *Elsevier—Computers and Electrical Engineering*, Vol.34(2008), pp. 141-147, ISSN 0045-7906

Hu, Peng (2007). A new model of ultrasonic imaging system based on plane wave transmission and angular spectrum propagation principle, *2007 IEEE International Conference on Integration Technology,* pp. 307-310, ISBN 1-4244-1092-4, Shenzhen, China, March 20-24, 2007

Hu, Peng; Jianyu Lu & Xue Mei Han. (2006). High Frame Rate Ultrasonic Imaging System Based on the Angular Spectrum Principle, *Elsevier —Ultrasonics*, Vol.44 (2006), pp. 97-99, ISSN 0041-624X

Hu, Peng; Jianyu Lu (2002). High Frame Rate 2D and 3D Imaging with a Curved or Cylindrical Array, *2002 IEEE Ultrasonics Symposium,* pp. 1725-1728, ISBN 0-7803-7582-3, Munich, Germany, October 8-11, 2002

FOURIER
TRANSFORMS
APPROACH TO SCIENTIFIC PRINCIPLES
Edited by Goran S. Nikolić
INTECH

15

High-Accuracy and High-Security Individual Authentication by the Fingerprint Template Generated Using the Fractional Fourier Transform

Reiko Iwai and Hiroyuki Yoshimura
Graduate School of Engineering, Chiba University
Japan

1. Introduction

The personal identities by the biological information have been increasing everywhere, e.g. on ATM of the bank, at the airport, and so on. In particular, the fingerprint authentication or handwriting analysis has been used as the simplest way, because the biological information does not have to be remembered and there is no worry to be lost like a password. Most of the methods, whole fingerprint images or handwriting data were stored as templates on paper without modification. When the authentication is needed, the biological information and the template were matched manually.

Recently, the fingerprint images or handwriting data have been stored as templates on the database of the computer. The authentication method, where the biological information and the template on the database are matched automatically, is becoming mainstream on the background that the computer technology has rapidly been developed.

The generation methods of the templates of the fingerprint images are classified into two major categories. One generation method is to use a priori extracted features of the images, such as minutiae (Maltoni et al., 2003). The other is to use the spatial frequency data of the one-dimensional (1D) data extracted from the two-dimensional (2D) original fingerprint image in a specific direction (Takeuchi et al., 2007).

However, there are some problems related to storing the templates; 1) the unfair use is possible when the information leaks out; 2) the biological information cannot keep the same condition forever so that it could not be always verified accurately when matching.

We have to consider the following points to solve these problems. For the former, the information on the fingerprint images should be hidden in order not to be used unfairly by unauthorized persons when the information leaks out. For the latter, the high accuracy of the authentication should be demanded even if there are some hurt and dirty on the fingerprint images.

To solve these problems, in this manuscript, the templates are generated using the fractional Fourier transform (FRT) (Ozaktas et al., 2001) which is the generalization of the conventional Fourier transform (FT). The FRT has a feature that the FRT's orders can be changed to arbitrary real numbers. Therefore, we could generate the templates solving the above-mentioned problems when the FRT is applied to the 1D data extracted from the 2D original fingerprint image in a specific direction. In addition, recently, research on a high-speed optical arithmetic processing of the FRT has been developed (Lohmann, 1993; Moreno et al., 2003; Ozaktas et al., 2001). Therefore, under the assumption of realizing the high-speed optical arithmetic processing of the FRT in the near future, the templates are stored as the intensity FRT in our study.

In this manuscript, we introduce the templates generated by the FRT of the fingerprint images. Moreover, we indicate the authentication accuracy by use of the templates and the robustness for unauthorized third persons. Specifically, we analyze from the following three perspectives: 1) the behavior of peak value of cross-correlation function between the original fingerprint image and the generated template expressed in terms of the intensity FRT; 2) the behavior of peak value of cross-correlation function between the original fingerprint image and the intensity inverse FRT (IFRT) of generated template; 3) derivation of the minimum error rate (MER) and authentication threshold on the basis of the false acceptance rate (FAR) and the false rejection rate (FRR) (Mansfield et al., 2001).

These analyses allow us to show the difference between the template and the original fingerprint image and that between the intensity IFRT of the template and the original image, quantitatively. This fact means that we cannot identify the original fingerprint image as the difference between them becomes greater and greater. In addition, the high authentication accuracy can be obtained by the analysis using the FAR and FRR which are the criterion of authentication accuracy.

2. Definition of the Fractional Fourier Transform (FRT)

The FRT is the generalization of a conventional FT. The FRT of 1D input data $u(x)$ is defined as (Ozaktas et al., 2001; Bultheel & Martinez Sulbaran, 2004a)

$$u_p(x_p) = F^{(p)}\left[u(x)\right] = \int u(x)\exp[i\pi(x_p^2 + x^2) / s^2 \tan\phi]$$

$$\times\exp[-2i\pi x_p x / s^2 \sin\phi]dx, \tag{1}$$

where a constant factor has been dropped; $\phi = p\pi / 2$, where p is the FRT's order; s is a constant. In particular, in the optical FRT, s is called a scale parameter expressed in terms of $s = \sqrt{\lambda f_s}$ where λ is the wavelength and f_s is an arbitrarily fixed focal length (Ozaktas et al., 2001). In this manuscript, the value of s was fixed at 1.0.

When p takes a value of $4n$+1, n being any integer, the FRT corresponds to the conventional FT. The intensity distribution of the FRT, $I_p(x_p)$, which is named intensity FRT in our study, is obtained by calculating $|u_p(x_p)|^2$. In addition, $u_p(x_p)$ can be decoded to $u(x)$ by the IFRT with the order $-p$ as follows:

$$u(x) = F^{(-p)}\left[u_p(x_p)\right]. \tag{2}$$

In this manuscript, we call p in Eq. (2) the IFRT's order. "Disfrft.m" (Bultheel & Martinez Sulbaran, 2004b) was used in our numerical calculation of the FRT.

2.1 Modeling waveform pattern of the fingerprint

In this subsection, as a 1D modeled fingerprint image, we used the finite rectangular wave which is regarded as the simplification of the grayscale distribution in an arbitrary scanned line of the 2D original fingerprint images. We make clear the charactererisc of the amplitude, phase and intensity distributions of the FRT.

First, Fig. 1 (a) shows the cross-sectional waveform that isn't modeled in an arbitrary scanned line of the 2D original fingerprint images. Although the grayscale levels are composed of intermediate values between 0 and 255 at the actual scanned lines in the case of 2D black and white image of 8 bits, in order to highlight the FRT as our method together with its feasibility, a finite rectangular wave is assumed to be the simplification of the grayscale distribution of the fingerprint image as shown in Fig. 1 (b). Horizontal axis is intentionally composed of 1024 (2^{10}) pixels to be smoothly illustrated the results of the FT and the FRT. We premise the application of the FRT to the 2D original fingerprint image which has multiple lines with random FRT's orders. In addition, the FRT's orders can be used as arbitrary real numbers.

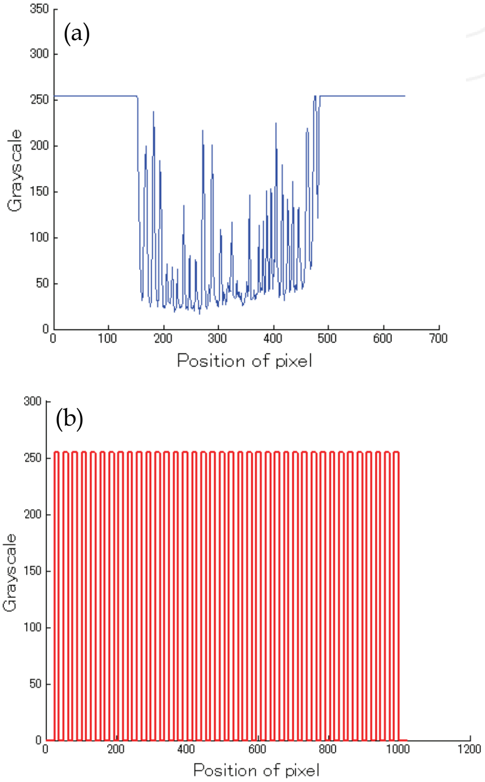


Fig. 1. (a) Cross-sectional waveform of a 2D original fingerprint image and (b) the finite rectangular wave as a modeled fingerprint image

2.2 Application of the FRT

The algorithm of the FRT has been intensively studied (Marinho & Bernardo, 1998; Yang et al., 2004; Bailey & Swarztrauber, 1991). Alternatively, the FRT was also applied to the fake finger detection (Lee et al., 2009).

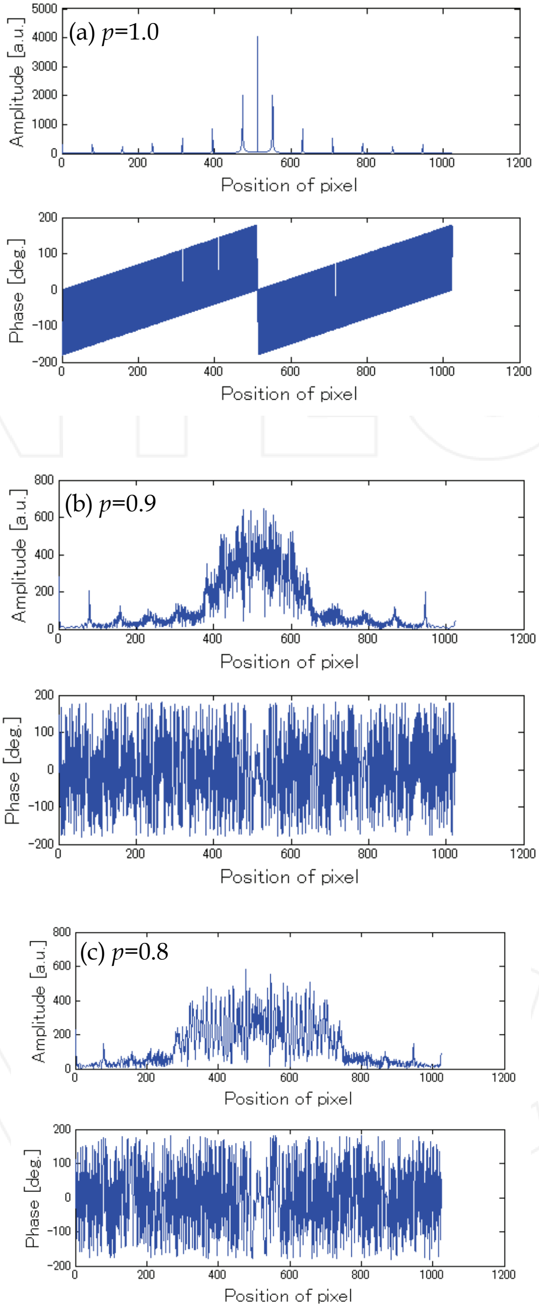


Fig. 2. Examples of the amplitude and phase distributions of the FRTs applied to the finite rectangular wave, when ps= (a) 1.0, (b) 0.9 and (c) 0.8

In this subsection we apply the FRT to the 1D finite rectangular wave data shown in Fig. 1 (b) as a modeled fingerprint image. Basically, the FRT with the order *p* is applied to the finite rectangular wave in Eq. (1). The FRT with the order *p* can be decoded to the finite rectangular wave by the IFRT with the same order *p* as already explained in Eq. (2). Fig. 2 demonstrates the results of the FRTs in comparison with the conventional FT (i.e., the FRT with *p*=1.0). Namely, Fig. 2 (a) shows the result of the FT as the amplitude distribution at the upper portion and the phase distribution at the lower portion. Figs. 2 (b) and 2 (c) are the results of the FRTs with *p*s=0.9 and 0.8, respectively. As a result, the peak values of the amplitude distributions in Figs. 2 (a), 2 (b) and 2 (c) are 4.04×10^3, 6.59×10^2 and 5.80×10^2, respectively.

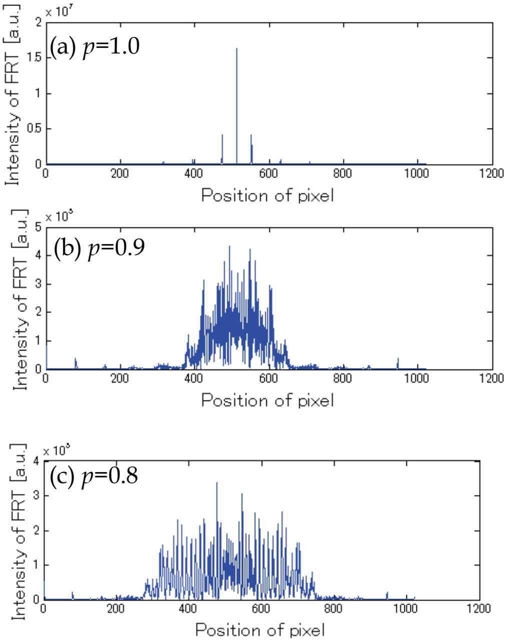


Fig. 3. The intensity distributions of the FRTs of the finite rectangular wave shown in Fig. 1, when *p*s= (a) 1.0, (b) 0.9 and (c) 0.8

It is found that the peak value of the amplitude distribution falls remarkably and the width of spread increases when the value of the FRT's order *p* decreases. It is also found that there is little difference in phase distributions between Figs. 2 (b) and 2 (c). In the case of FT shown in Fig. 2 (a), the order *p* can be identified through the waveforms of the amplitude and phase distributions. However, in the case of FRT, the order *p* may not be identified through them. In particular, it is difficult to identify the FRT's orders *p*s through the waveforms of the phase distributions shown in Figs. 2 (b) and 2 (c). Therefore, this fact led us the new method safer than the conventional method using the FT, because the FRT's order has highly-confidential in the applied FRT condition.

In this way, we focused on the intensity distribution of the FRT from a viewpoint of the security of individual information, because the intensity FRT may not be completely decoded to the original fingerprint image by the IFRT. Fig. 3 depicts the intensity FT of Fig. 2 (a) and the intensity FRTs of Figs. 2 (b) and 2 (c). The peak values of the intensity distributions in Figs. 3 (a), 3 (b) and 3 (c) are 1.63×10^7, 4.34×10^5 and 3.37×10^5, respectively. It is found from the comparison between Figs. 2 and 3 that the peak value of the wave pattern of the intensity distribution is very high.

3. How to generate the fingerprint template by use of the FRT and its characteristics

Fingerprint images provided by the Biometric System Laboratory (Maltoni & Maio., 2004) were used as original raw data. As an example, the data in the TIF format with 480 vertical and 640 horizontal pixels (480×640 pixels) is visualized in Fig. 4. In this manuscript, as shown in Fig. 4, height and width of the images are called 'line' and 'column,' respectively. The templates were generated by the FRT of the cross-sectional waveform with an arbitrary random order in every longitudinally (or transversally) scanned line of the original fingerprint images.

Fig. 4 illustrates an example where the FRTs with the random orders of p_1, p_2, p_3, ... p_m and p_n are conducted in transversally-scanned lines from the top to the bottom of the fingerprint image. Therefore, the information on the FRT's order in every transversally-scanned line is needed to be decoded to the original fingerprint image by use of the IFRT. For that reason, there is almost no possibility of the unfair use by the unauthorized third persons.

Fig. 5 depicts an example of the template expressed in terms of the intensity FRT. The reason why we used the intensity FRT as the template is that we would use a high-speed optical processing system of the FRT (Lohmann, 1993; Moreno et al., 2003; Ozaktas et al., 2001) to generate the templates in the near future. In this case, the template can be produced at higher speed because of no need of calculation by a computer. As shown in Fig. 5, the information on the original fingerprint image cannot be known from the template which was generated by the FRT with a random order in every transversally-scanned line.

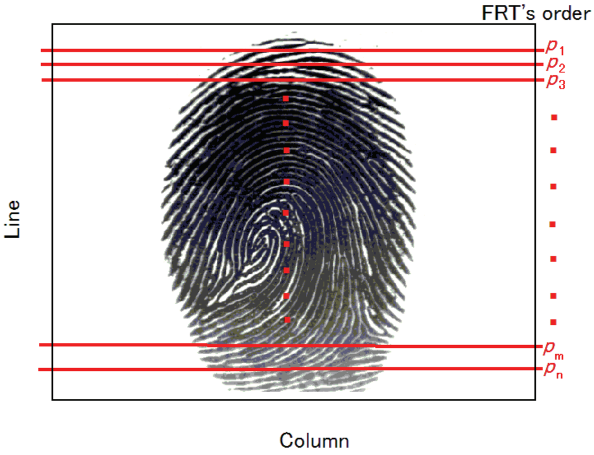


Fig. 4. Fingerprint image

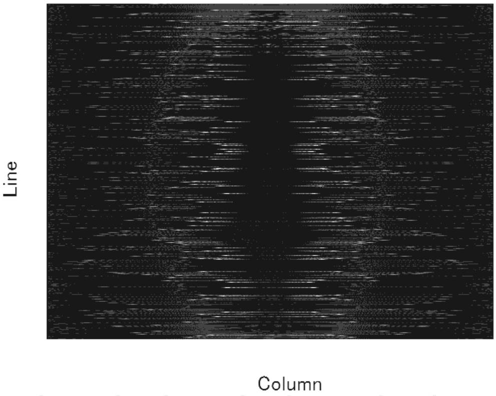


Fig. 5. Template

4. Characteristics of the fingerprint authentication

Next, 100 kinds of fingerprint images were prepared, and the fingerprint images with 200 lines and 200 columns (200×200 pixels) were extracted from a central part of the original fingerprint images. Two examples of extracted fingerprint images are depicted in Fig. 6 and the real size is 10.2 mm by 10.2 mm. The blank space was deleted from the original fingerprint images so that more accurate authentication could be conducted by use of the extracted fingerprint images. For this reason, the matching speed can be expected to be faster because the matching range is small.

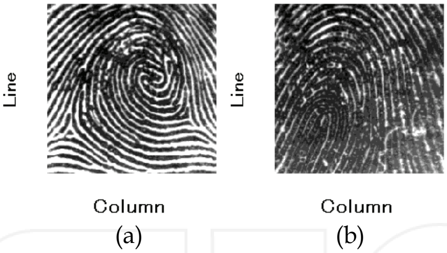


Fig. 6. Fingerprint images with 200 lines and 200 columns extracted at a center of the original fingerprint images

4.1 Difference between the template and the extracted fingerprint image

We analyzed the behavior of the peak value of the normalized cross-correlation function between the template generated by the FRT with a different order in every line and the extracted fingerprint image shown in Fig. 6. The templates were generated for 100 kinds of extracted fingerprint images with 200 lines and 200 columns.

The behavior was analyzed for the FRT's order ranges of 4 kinds of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9. Fig. 7 gives the result. In Fig. 7, the vertical and horizontal axes denote the peak value of the normalized cross-correlation function and the FRT's order range, respectively.

In the figure, the symbol of circle and the bar denote the averaged peak value and the standard deviation of the peak value, respectively. The averaged peak values for the FRT's order ranges of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9 are 0.636, 0.420, 0.443 and 0.429, respectively. Additionally, the standard deviations of the peak values for the FRT's order ranges of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9 are 0.0484, 0.0474, 0.0592 and 0.0533, respectively. The averaged peak value of 0.420 is the smallest of them when the FRT's order range is 0.1-1.9. It is found that the template has a great difference between the extracted fingerprint image and the template under the condition that the FRT's order range is 0.1-1.9, 0.1-2.9 or 0.1-3.9.

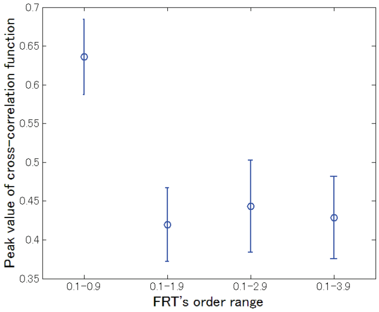


Fig. 7. Peak value of the normalized cross-correlation function between the template and the extracted fingerprint image for every FRT's order range

4.2 Robustness of the template for the IFRT

Next, we analyzed the behavior of the peak value of the normalized cross-correlation function between the intensity IFRT of the template and the extracted fingerprint image shown in Fig. 6. The IFRT of the template was generated by the IFRT with a different order in every line of the template generated in Subsection 4.1. In the analysis, 100 kinds of fingerprint images with 200 lines and 200 columns were used.

The behavior was analyzed for the IFRT's order ranges of 4 kinds of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9. Fig. 8 gives the result. In Fig. 8, the vertical and horizontal axes denote the peak value of the normalized cross-correlation function and the IFRT's order range, respectively.

In the figure, the symbol of square and the bar denote the averaged peak value and the standard deviation of the peak value, respectively. The averaged peak values for the IFRT's order ranges of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9 are 0.340, 0.215, 0.211 and 0.205, respectively. Additionally, the standard deviations of the peak values for the IFRT's order ranges of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9 are 0.0406, 0.0365, 0.0436 and 0.0351, respectively. The averaged peak value of 0.205 is the smallest of them when the IFRT's order range is 0.1-3.9. It is found that the template has a great difference between the extracted fingerprint image and the intensity IFRT of the template under the condition that the order range is 0.1-1.9, 0.1-2.9 or 0.1-3.9. Therefore, the unauthorized third persons who are unapprised of the information on the FRT's order in every line cannot retrieve the extracted fingerprint data from the template.

5. Authentication accuracy based on the FAR and FRR

Fig. 9 illustrates the basic concept of the FAR and FRR. In the figure, the left-hand curve is the imposter distribution and the right-hand curve is the authentic distribution. The authentication threshold is decided by a value satisfied with the condition that the FAR and

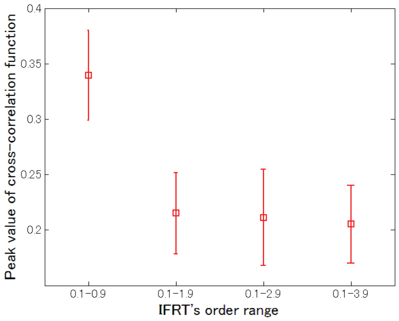


Fig. 8. Peak value of the normalized cross-correlation function between the intensity IFRT of the template and the extracted fingerprint image for every IFRT's order range

FRR take the same value corresponding to the MER. The FAR is the probability of accepting other person erroneously. As shown in the figure, it corresponds to an area of the impostor distribution higher than the authentication threshold. On the other hand, the FRR is the probability of rejecting identical person and corresponds to the area of the authentic distribution lower than the authentication threshold. In our analysis, the horizontal axis in Fig. 9 corresponds to the peak value of the 2D normalized cross-correlation function of the intensity FRTs for the two sets of fingerprint images.

In order to obtain the imposter and authentic distributions, 100 kinds of templates were used. For each of them, 10 kinds of templates were prepared to obtain the imposter distribution. On the other hand, for each of the templates, 10 kinds of templates, which were produced by the FRT of the extracted fingerprint images superimposed by random noise (average μ=0, standard deviation σ=25.5), were prepared to obtain the authentic distribution. Figs. 10 and 11 are the results showing the behavior of peak value of the normalized cross-correlation function of the FRT intensity by changing the FRT's order range for the impostor distribution and the authentic distribution, respectively. As same as Fig. 7 in the Subsection 4.1, in Figs. 10 and 11, the vertical and horizontal axes denote the peak value of normalized cross-correlation function and the FRT's order range, respectively.

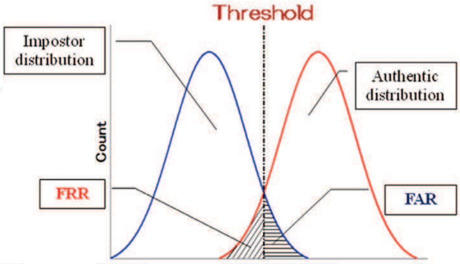


Fig. 9. Basic concept of the FAR and FRR

In Fig. 10 related to the impostor distribution, the symbol of cross and the bar denote the averaged peak value and the standard deviation of the peak value, respectively. The averaged peak values for the FRT's order ranges of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9 are 0.658, 0.735, 0.764 and 0.732, respectively. Additionally, the standard deviations of the peak values for the FRT's order ranges of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9 are 0.0928, 0.0868, 0.0650 and 0.0861, respectively. On the other hand, in Fig. 11 related to the authentic

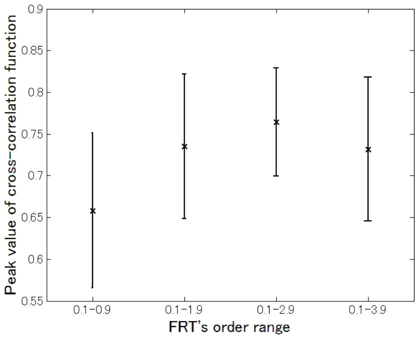


Fig. 10. Behavior of peak value of the normalized cross-correlation function related to the impostor distribution

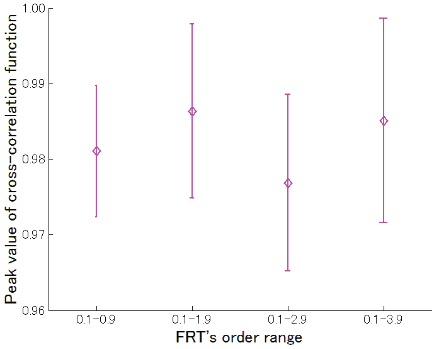


Fig. 11. Behavior of peak value of the normalized cross-correlation function related to the authentic distribution

distribution, the symbol of diamond shape and the bar denote the averaged peak value and the standard deviation of the peak value, respectively. The averaged peak values for the FRT's order ranges of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9 are 0.981, 0.986, 0.977 and 0.985, respectively. Additionally, the standard deviations of the peak values for the FRT's order ranges of 0.1-0.9, 0.1-1.9, 0.1-2.9 and 0.1-3.9 are 0.00869, 0.0115, 0.0117 and 0.0135, respectively.

Moreover, Fig. 12 depicts histograms that correspond to the impostor and authentic distributions, when the FRT's order range is 0.1-0.9. The left-hand curve in Fig. 12 corresponds to the impostor distribution related to Fig. 10, when the FRT's order range is 0.1-0.9. The right-hand curve in Fig. 12 corresponds to the authentic distribution related to Fig. 11, when the FRT's order range is 0.1-0.9. In this case, the MER is 7.36×10^{-4}% and the authentication threshould is 0.95.

Furthermore, Fig. 13 illustrates histograms that correspond to the impostor and authentic distributions, when the FRT's order range is 0.1-1.9. The MER is 5.31×10^{-3}% and the authentication threshould is 0.96. As we can see from the comparison between Figs. 12 and 13, the peak of the imposter distribution shifts to right and the peak of the authentic distribution becomes high, when the FRT's range is changed from 0.1-0.9 to 0.1-1.9.

The recent available specification sheets of major fingerprint authentication systems in the market indicate that the matching accuracy is from 0.001 % to 1.0 % in the FAR and from 0.0001 % to 0.1 % in the FRR. As summarized in Table 1, the MER takes a value of 7.36×10^{-4}% when *p*=0.1-0.9, 5.31×10^{-3}% when *p*=0.1-1.9, 2.80×10^{-3}% when *p*=0.1-2.9, and 5.51×10^{-3}% when *p*=0.1-3.9. As a result, we found that the fingerprint authentication by use of the FRT has the high matching accuracy.

From the results shown in Figs. 7, 8, 10 and 11 and Table 1 and our final objective to realize the FRT by the optical system, we can say that the suitable FRT's order range is 0.1-1.9 in our method.

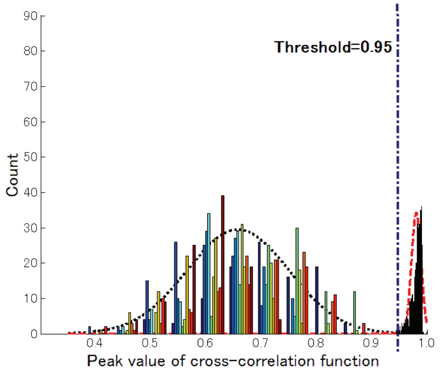


Fig. 12. A set of histograms corresponding to the impostor and authentic distributions (FRT's order range=0.1-0.9)

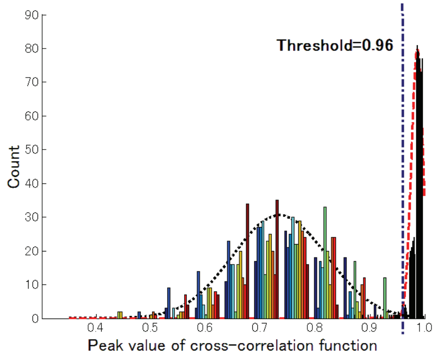


Fig. 13. A set of histograms corresponding to the impostor and authentic distributions (FRT's order range=0.1-1.9)

FRT's order range	MER (FAR / FRR)	Threshold
0.1～0.9	7.36×10^{-4}	0.95
0.1～1.9	5.31×10^{-3}	0.96
0.1～2.9	2.80×10^{-3}	0.94
0.1～3.9	5.51×10^{-3}	0.95

Table 1. MERs and authentication thresholds for various FRT's order ranges

6. Effects of size reduction of the extracted fingerprint image on the authentication

In Section 5, we analyzed the authentication accuracy by use of the templates generated by the FRT of the extracted fingerprint images with the size of 200×200 pixels. In this section, the authentication accuracy is analyzed by changing the size of the extracted fingerprint image, for example, 50×200, 100×200 and 150×200 pixels, when the FRT's order range is 0.1-1.9. The analysis method is the same as that in Section 5.

First, Fig. 14 illustrates the result related to the impostor distribution which is the behavior of peak value of the normalized cross-correlation function of the intensity FRTs of two different extracted fingerprint images, by changing the extracted line number. The vertical and horizontal axes denote the peak value of normalized cross-correlation function and the extracted line number, respectively. In the figure, the symbols of diamond shape, cross and circle denote the averaged peak values when the FRT's orders are 1.0, 0.0 and random between 0.1 and 1.9, respectively. Additionally, the bar denotes the standard deviation of the peak value.

When the extracted line numbers are 50, 100, 150 and 200, the averaged peak values for *p*=1.0 are 0.967, 0.949, 0.931 and 0.916, respectively. For *p*=0.0, the averaged peak values are 0.749, 0.751, 0.757 and 0.764, respectively, and for *p*=random, they are 0.734, 0.732, 0.732 and 0.735, respectively.

From these results, it is found that the probability of the accepting other person erroneously is low when *p*=random in comparison with those when *p*s=1.0 and 0.0. Moreover, there is little effect for the variation of the extracted line number when *p*=random in comparison with that when *p*=1.0.

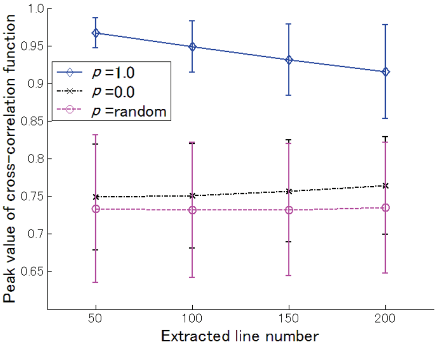


Fig. 14. Peak value of the normalized cross-correlation function of the intensity FRTs by changing the extracted line number (Impostor distribution)

Next, Fig. 15 illustrates the result related to the authentic distribution which is the behavior of peak value of the normalized cross-correlation function of the intensity FRTs of the extracted fingerprint images with and without random noise, by changing the extracted line number. The vertical and horizontal axes denote the peak value of normalized cross-correlation function and the extracted line number, respectively. In the figure, the symbols of diamond shape, circle and cross denote the averaged peak values when the FRT's orders are 1.0, random between 0.1 and 1.9, and 0.0, respectively. Additionally, the bar denotes the standard deviation of the peak value.

When the extracted line numbers are 50, 100, 150 and 200, the averaged peak values for *p*=1.0 are 0.989, 0.993, 0.995 and 0.995, respectively. For *p*=random, the averaged peak values are 0.976, 0.982, 0.984 and 0.986, respectively, and for *p*=0.0, they are 0.967, 0.972, 0.975 and 0.977, respectively.

From these results, it is found that the probabilities of the rejecting identical person erroneously are more-or-less identical for *p*s=1.0, 0.0 and random. Moreover, there is little effect for the variation of the extracted line number for *p*s=1.0, 0.0 and random.

Table 2 illustrates the MERs for the variation of the extracted line number, which were obtained from Figs. 14 and 15. From the table, it is found that the effect of the variation of the extracted line number on the authentication accuracy is very little because the values of MERs are fully small as shown in Table 2.

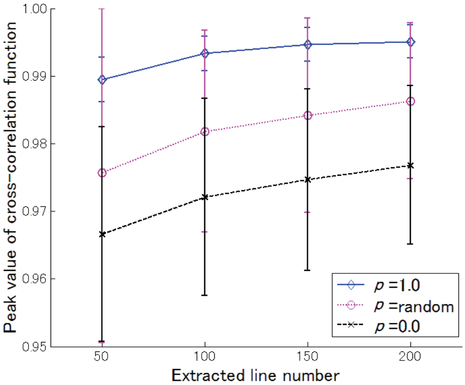
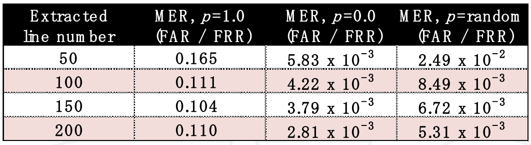


Fig. 15. Peak value of the normalized cross-correlation function of the intensity FRTs by changing the extracted line number (Authentic distribution)



Extracted line number	MER, *p*=1.0 (FAR / FRR)	MER, *p*=0.0 (FAR / FRR)	MER, *p*=random (FAR / FRR)
50	0.165	5.83×10^{-3}	2.49×10^{-2}
100	0.111	4.22×10^{-3}	8.49×10^{-3}
150	0.104	3.79×10^{-3}	6.72×10^{-3}
200	0.110	2.81×10^{-3}	5.31×10^{-3}

Table 2. MERs for variations of the extracted line number and the FRT's order

7. Conclusions

First, we generated the templates of many original fingerprint images by use of the FRT. As a result from comparisons between the generated templates and the original fingerprint images, it was found that the templates are fully different from the original fingerprint images when the templates were generated by changing randomly the FRT's order in every line of the original fingerprint images. It was also found that the generated templates are very high secure, because the templates could not be decoded to the original fingerprint images by the unauthorized third persons who are unapprised of the information on the FRT's order in every line.

Additionally, the authentication accuracy of the templates generated by the FRT of the extracted fingerprint images with 200×200 pixels was analyzed by changing the FRT's order range. We found that the suitable FRT's order range for the generation of the template in our method is 0.1-1.9.

The authentication accuracy was also analyzed by changing the size of the extracted fingerprint image, concretely, 150×200, 100×200 and 50×200 pixels. As a result, it was found that the authentication accuracy is fully high even if the size of the extracted fingerprint image is small, so that the authentication is possible at higher speed.

8. References

Bailey, D. H. & Swarztrauber, P. N. (1991). The fractional Fourier transform and applications, *SIAM Review*, Vol. 33, No. 3, pp. 389-404, ISSN: 0036-1445

Bultheel, A. & Martinez Sulbaran, H. E. (2004a). Computation of the fractional Fourier transform, *Applied and Computational Harmonic Analysis*, Vol. 16, No. 3, pp. 182-202, ISSN: 1063-5203

Bultheel, A. & Martinez Sulbaran, H. E. (2004b). http://nalag.cs.kuleuven.be/research/software/FRFT/

Lee, H.; Maeng, H. & Bae, Y. (2009). Fake finger detection using the fractional Fourier transform, In: *Biometric ID Management and Multimodal Communication*, Fierrez, J.; Ortega-Garcia, J.; Esposito, A.; Drygajlo, A. & Faundez-Zanuy, M. (Eds.), pp. 318-324, Springer, ISBN: 978-3-642-04390-1, Heidelberg

Lohmann, A. W. (1993). Image rotation, Wigner rotation, and the fractional Fourier transform, *Journal of the Optical Society of America A*, Vol. 10, No. 10, pp.2181-2186, ISSN: 0740-3232

Maltoni, D.; Maio, D.; Jain, A.K. & Prabhakar, S. (2003). *Handbook of Fingerprint Recognition*, Springer, 2003, ISBN : 0-387-95431-7, New York

Maltoni, D. & Maio, D.; (2004). Download Page of FVC2004, Biometric System Laboratory, University of Bologna, Italy (http://bias.csr.unibo.it/fvc2004/ download.asp)

Mansfield, T.; Kelly, G.; Chandler, D. & Kane, J. (2001). Biometric Product Testing Final Report, Issue 1.0, In: *CESG/BWG Biometric Test Programme,* Centre for Mathematics and Scientific Computing, National Physical Laboratory

Marinho, F. J. & Bernardo, L. M. (1998). Numerical calculation of fractional Fourier transforms with a single fast-Fourier-transform algorithm, *Journal of the Optical Society of America A*, Vol. 15, No. 8, pp. 2111-2116, ISSN: 0740-3232

Moreno, I.; Davis, J. A. & Crabtree, K. (2003). Fractional Fourier transform optical system with programmable diffractive lenses, *Applied Optics*, Vol. 42, No. 32, pp. 6544-6548, ISSN: 0003-6935

Ozaktas, H. M. ; Zalevsky, Z. & Kutay, M. A. (2001). *The Fractional Fourier Transform*, John Wiley & Sons., 2001, ISBN: 0-471-96346-1, Chichester

Takeuchi, H. ; Umezaki, T. ; Matsumoto, N. & Hirabayashi, K.. (2007). Evaluation of low-quality images and imaging enhancement methods for fingerprint verification, *Electronics and Communications in Japan, Part 3*, Vol. 90, No. 10, pp. 40-53, Online ISSN : 1520-6424

Yang, X.; Tan, Q.; Wei, X.; Xiang, Y.; Yan, Y. & Jin, G. (2004). Improved fast fractional-Fourier-transform algorithm, *Journal of the Optical Society of America A*, Vol. 21, No. 9, pp. 1677-1681, ISSN: 1084-7529

FOURIER TRANSFORMS
APPROACH TO SCIENTIFIC PRINCIPLES
Edited by Goran S. Nikolić
INTECH

16

Fourier Transform Mass Spectrometry for the Molecular Level Characterization of Natural Organic Matter: Instrument Capabilities, Applications, and Limitations

Rachel L. Sleighter and Patrick G. Hatcher
Old Dominion University
United States of America

1. Introduction

All living matter in the environment (i.e., animals, plants, microorganisms, etc.) eventually dies and decomposes into what is known as natural organic matter (NOM). NOM is formed from a vast variety of sources that have been chemically or microbially degraded in the environment where they arise, and NOM can be generally described as a complex mixture of organic compounds (Stevenson, 1994). Within this mixture, some compounds retain their individual reactivity and characteristics, while others tend to aggregate together and act as a polymeric unit. Overall, NOM can encompass a variety of natural biomolecules, such as lipids, peptides/protein, amino-sugars, carbohydrates, lignins, tannins, and condensed aromatics. Because NOM is a random assortment of organic constituents, its size, shape, concentration, and other physico-chemical properties vary greatly with location and season. For these reasons, the molecular level characterization of NOM continues to be one of the greatest challenges to modern analytical chemists.

NOM is ubiquitously present in all natural waters, soils, sediments, and air, giving NOM a central role in numerous environmental processes. These processes are linked together by the global carbon cycle, which describes the storage and flux of carbon sources and sinks throughout the environment (Thurman, 1985; Eglinton and Repeta, 2003; Perdue and Ritchie, 2003). Special attention is generally paid to land-sea interfaces, atmosphere-sea interfaces, and long-term carbon burial/storage. NOM in soils affects the cation exchange capacity and water retention of soils, which has triggered studies by the agricultural communities. Furthermore, NOM in soils/sediments influences carbon sequestration and burial, and this carbon is altered over long periods of time and can be transformed to petroleum precursors. NOM in soils and rivers can affect the solubility, transport, and eventual fate of anthropogenic pollutants. These hydrophobic organic contaminants can interact and bind with NOM in the environment, making it difficult to trace throughout the river systems that eventually lead to the ocean. The amount of carbon in dissolved organic matter (DOM) in the ocean is approximately the same as that of atmospheric CO_2 (Hedges, 1992; Eglinton and Repeta, 2003) and this exchange has been directly linked to climate change (Canadell et al., 2007; Sabine and Feely, 2007). NOM in the atmosphere can exist as an aerosol or particulate, which impacts human health, climate, and overall air quality. The brief explanation of NOM, along with the short list of NOM implications given here, barely touches upon the large variety of important research that is ongoing within the NOM community. The significance of NOM in the environment and the numerous roles that NOM plays in the biogeochemical processes that mediate Earth's ecosystems highlight the necessity for a more fundamental comprehension of NOM chemistry and composition.

The characterization of NOM from different sources is a difficult task, simply because NOM does not have an exact composition or structure and occurs at aqueous concentrations that can vary by 3 orders of magnitude, depending on NOM type and location. Bulk characteristics of NOM can be identified by elemental analysis, ultraviolet and infrared spectroscopy, and traditional one-dimensional nuclear magnetic resonance (Hatcher et al., 2001). Much has been learned about NOM chemistry from these techniques, such as elemental compositions and ratios (%C, %N, %O, %S, C/N, etc.), specific functional groups that primarily exist in NOM, and general trends that occur amongst different NOM samples. Chromatographic techniques, such as gas and liquid chromatography coupled to various detectors (i.e., flame ionization detection, mass spectrometry, photodiode array detection) has also provided a wealth of structural information on various types of NOM. Two-dimensional nuclear magnetic resonance has more recently been utilized to characterize and understand the nature of both soluble and insoluble NOM beyond that of the one-dimensional approach, and this has led to the ability to link NOM to well known biopolymer classes. However, despite the abundance of data that has been acquired using the aforementioned techniques, NOM remains as an analytical challenge. Because NOM exists amongst a background matrix, it can be difficult to separate from water or inorganic matter, without losing or altering the NOM (Mopper et al., 2007; Dittmar et al., 2008). Sample preparation for NOM is an important consideration. Furthermore, NOM is not amenable to most instrumental analyses, because it is a low concentration of highly functionalized polymeric substances that do not have uniform behaviour. NOM has a wide size and volatility range, since portions are hydrophilic, allowing them to be water soluble, while other parts retain their hydrophobic nature. Overall, the goal of molecular level characterization of NOM continues be a daunting task.

The advent of atmospheric pressure ionization (API) sources and Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) has revolutionized our ability to analyze NOM. These methods were originally employed by the biochemical communities to elucidate the structure of biological macromolecules (proteins, metabolic products, DNA, etc.). The knowledge that has evolved from these studies has been applied to NOM, where the goal is to transfer polar analytes in solution to molecular ions that can be detected by mass spectrometry. With the exception of petroleum and crude oil samples, the first application and use of electrospray ionization (ESI) for studying NOM was by McIntyre et al. (1997), utilizing a triple quadrupole mass spectrometer to analyze organic acids. Not long after, Fievre et al. (1997) utilized ESI-FTICR-MS to investigate the size and composition of humic and fulvic acids. In the last 15 years, the utilization of FTICR-MS for the analysis of NOM samples has increased from year to year, and from 2000 to 2010, the number of publications has increased by nearly 900% (Fig. 1). Clearly, FTICR-MS is a powerful technique that has great promise in the field of NOM chemistry, with a rise in its use progressing every year. In this chapter, the spotlight lies in utilizing FTICR-MS specifically for the characterization of NOM, concentrating on particular instrumental capabilities, its application to a variety of different types of NOM, and limitations that exist for the acquisition and analysis of data.

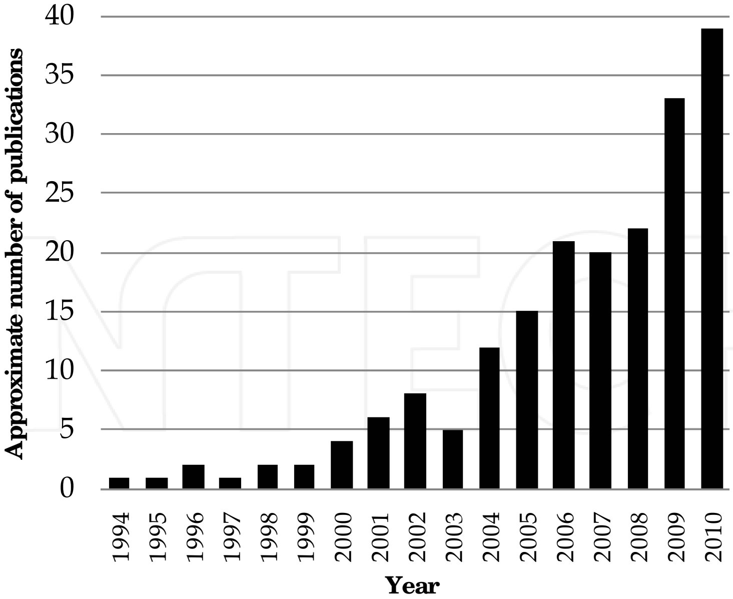


Fig. 1. Approximate number of publications on the topic of natural organic matter (water, soil, sediment, aerosol, petroleum, crude oil, etc.) analysis using Fourier transform ion cyclotron resonance mass spectrometry. Search conducted in September, 2010

2. Introduction to Fourier transform ion cyclotron resonance mass spectrometry

Over the course of the last decade, FTICR-MS has emerged as an invaluable tool for the characterization of NOM by providing details about its composition. While previous chemical and instrumental analyses (e.g., gas and liquid chromatography, ultraviolet and infrared spectroscopy, fluorescence excitation emission matrix spectroscopy, nuclear magnetic resonance (NMR), elemental and isotopic analyses, etc.) have revealed vital information about NOM, these techniques are biased for certain compound classes and fail to resolve the numerous constituents in NOM (Hatcher et al., 2001; Leenheer and Croué, 2003). Other mass spectrometers have been employed for the analysis of NOM, such as quadrupoles, ion traps, and quadrupole time of flights. However, FTICR-MS has abilities and advantages over these systems and has achieved a more in depth analysis of NOM where other mass spectrometers have been marginally or less successful (Sleighter and Hatcher, 2007).

The detailed theory and instrumental parameters of FTICR-MS are expertly reviewed by Marshall et al. (1998). Here, we give a brief overview of the instrument. Ions are produced in the ion source region (via a variety of available ion sources discussed in section 3) that is maintained at atmospheric pressure. These ions are focused, using ion funnels and skimmers, into differential pumping regions that vary in pressure from atmospheric pressure, to low vacuum (10^{-4} - 10^{-6} mbar) just after the ion source region, to high vacuum

(10^{-9} - 10^{-10} mbar) in the ICR cell. Ions are steered through these regions and typically pass through a mass analyzer in the lower vacuum area where an initial sort and storage occurs. Commercial instruments vary in which type of mass analyzer is used; some are a combination of hexapoles and/or quadrupoles, while others are linear ion traps. In this first mass analyzer region, only ions of a certain mass to charge (m/z) range (typically 100-2000 m/z) are allowed to be accumulated for a designated time prior to their transfer to the detector. Once past this accumulation stage, the ions are guided through more pumping stages of the ion transfer optics region and are eventually transferred into the horizontal bore of a cryogenic magnet, where they are trapped in the ion cyclotron resonance cell. Modern commercial FTICR-MS instruments employ cryomagnets of various strengths (usually 7-15 T). Once ions are trapped in the ICR cell under the influence of a homogeneous magnetic field, they circulate at a frequency characteristic of their m/z value. Cyclotron frequency is inversely proportional to m/z, as shown below:

$$f_c = B_0(z/m) \tag{1}$$

In equation 1, f_c is the cyclotron frequency, B_0 is the magnetic field strength, z is the charge of the ion, and m is the mass of the ion. Ions in orbit in the ICR cell are excited by a broadband RF pulse, which increases the radius of their orbit but not their frequency. The ions are now circulating closer to the detector plates, where the ion packets can induce an image current on the receiving electrode. Field inhomogeneities cause the ions to lose coherence and orbit radius, which leads to collapse of the ions to the central core of the ICR cell. This produces an image current trace that is called a free induction decay (FID), similar to what is observed for NMR signals. The time-domain FID signal (Fig. 2a) is digitized and Fourier transformed into a frequency domain signal (Fig. 2b), after acquiring and summing multiple FID spectra to build up signal-to-noise. The frequency domain data are then converted into mass spectra (Fig. 2c) by use of equation 1 and calibrated with compound mixtures having components with known m/z values. This sequence of data detection and conversion is shown in Fig. 2 for a sample of Suwannee River NOM.

In our opinion, FTICR-MS is the only type of mass spectrometer that can achieve the resolving powers that are necessary (those exceeding 10^5) for mass deconvolution of the thousands of compounds that are present in a single NOM sample (Fig. 2c). Orbitrap mass spectrometers (providing resolving powers of approximately 60,000) provide sufficient resolution for evaluating elemental formulas of mixtures containing only C, H, and O molecules (see section 5 and Fig. 8), but the inclusion of heteroatoms in the molecules (i.e., N, S, and P) requires the use of FTICR-MS to achieve assignment of exact elemental formulas. Fig. 2d shows that the most intense peaks are detected at odd nominal masses, indicating that they are composed of compounds with either 0 or an even number of nitrogens (based on the nitrogen rule). Ions detected at even nominal masses contain an odd number of nitrogens or are ^{13}C isotopologues of the ^{12}C compounds detected (Fig. 2e). Based on the identification of the ^{12}C and ^{13}C isotopologues, one can determine the charge state of the compound. If a compound is singly charged, then the ^{13}C isotope peak will be observed at 1.0034 mass units higher than the ^{12}C peak, which is the mass difference between ^{12}C and ^{13}C. Doubly charged peaks have isotopes that appear at 0.5017 mass units higher, but doubly charged peaks are rarely detected in NOM samples (Kujawinski et al. 2002; Stenson et al., 2002; Kim et al., 2003). Generally, at least 10 peaks are detected at each individual nominal mass, with upwards of 20-30 being observed in some cases.

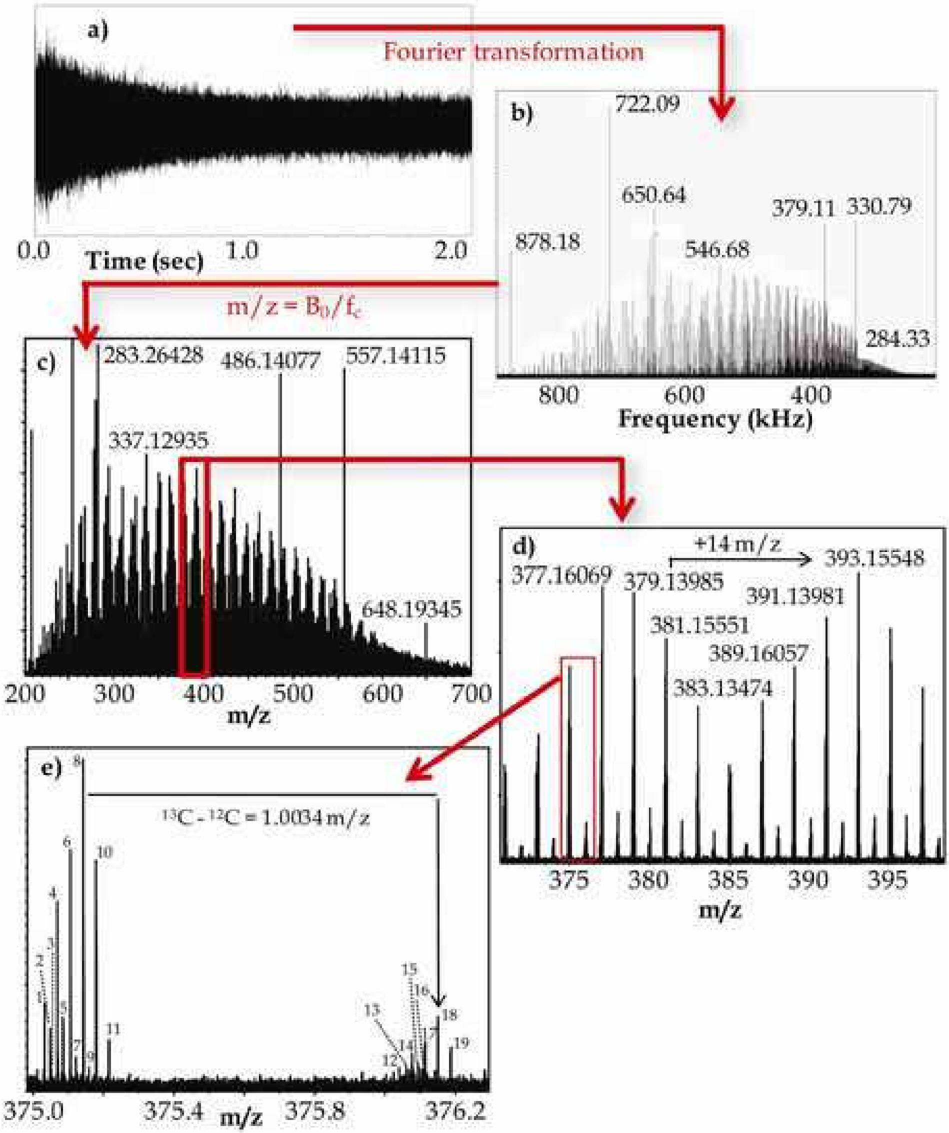


Fig. 2. Suwannee River NOM analyzed by negative ion mode ESI-FTICR-MS: a) time domain free induction decay (FID), b) FID Fourier transformed into a frequency domain spectrum, c) mass spectrum after frequency is converted to m/z, d) expanded region of mass spectrum at 371-398 m/z, e) nominal masses 375 and 376 of the mass spectrum. Numbered peaks in (e) correspond to Table 1

The ultrahigh resolving powers that FTICR-MS at 12 T routinely achieves is highlighted in Fig. 2e and is the main reason why FTICR-MS is preferred over other mass spectrometers that exhibit only nominal mass resolution (see examples: Fig. 3 in Kujawinski et al., 2002;

Fig. 1 in Sleighter and Hatcher, 2007). The equation for calculating resolving power is shown below in equation 2, where RP is resolving power, m is mass, and FWHM is the full width at half maximum of the peak.

$$P = m/(FWHM) \tag{2}$$

Peak Number	measured m/z	S/N	FWHM	Resolving Power	proposed formula	calculated m/z	Error (ppm)
1	375.03585	15.0	0.00078	479014	$C_{17}H_{11}O_{10}^-$	375.035770	0.21
2	375.05110	10.7	0.00074	508681	$C_{21}H_{11}O_7^-$	375.051026	0.20
3	375.05444	3.1	0.00052	721130	$C_{18}H_{15}O_7S_1^-$	375.054397	0.11
4	375.07227	33.4	0.00076	493629	$C_{18}H_{15}O_9^-$	375.072156	0.30
5	375.08755	12.4	0.00074	503478	$C_{22}H_{15}O_6^-$	375.087412	0.37
6	375.10864	43.2	0.00077	487137	$C_{19}H_{19}O_8^-$	375.108541	0.26
7	375.12390	5.4	0.00077	484204	$C_{23}H_{19}O_5^-$	375.123797	0.27
8	375.14491	59.2	0.00083	454710	$C_{20}H_{23}O_7^-$	375.144927	-0.05
9	375.16033	3.5	0.00115	326330	$C_{24}H_{23}O_4^-$	375.160183	0.39
10	375.18132	40.8	0.00082	459226	$C_{21}H_{27}O_6^-$	375.181312	0.02
11	375.21784	8.7	0.00060	625389	$C_{22}H_{31}O_5^-$	375.217698	0.38
12	376.03904	3.3	0.0006	629102	$^{13}C_1{}^{12}C_{16}H_{11}O_{10}^-$	376.039125	-0.23
13	376.06734	3.2	0.00048	783961	$C_{17}H_{14}N_1O_9^-$	376.067405	-0.17
14	376.07554	5.9	0.00066	569024	$^{13}C_1{}^{12}C_{17}H_{15}O_9^-$	376.075510	0.08
15	376.09090	4.2	0.00057	662373	$^{13}C_1{}^{12}C_{21}H_{15}O_6^-$	376.090767	0.35
16	376.10381	3.2	0.0007	538763	$C_{18}H_{18}N_1O_8^-$	376.103790	0.05
17	376.11201	8.1	0.00065	576557	$^{13}C_1{}^{12}C_{18}H_{19}O_8^-$	376.111896	0.30
18	376.14839	12.7	0.00085	443845	$^{13}C_1{}^{12}C_{19}H_{23}O_7^-$	376.148282	0.29
19	376.18463	7.1	0.00085	444104	$^{13}C_1{}^{12}C_{20}H_{27}O_6^-$	376.184667	-0.10

FWHM: full width at half maximum

Error units = parts per million deviation of calculated m/z from the measured m/z

Table 1. Details of the peaks shown in Fig. 1e

Resolving power calculations are demonstrated in Table 1, which shows the molecular formula assignments to the peaks in Fig. 2e, along with measured details of each. Because the frequencies, at which ions orbit within the ICR cell, can be measured very accurately, m/z can also be calculated very accurately, usually to the fifth decimal place. With careful external and internal calibration (Sleighter et al., 2008), accurate m/z values can be calculated and utilized for the determination of unique molecular formulas. These can be confidently assigned with an error difference (between the measured m/z and the calculated exact m/z) of less than 0.5 ppm (or 500 ppb). Once molecular formulas are assigned to the majority of the peaks in the mass spectrum, compositional make-up of the sample can be established.

3. FTICR-MS instrument capabilities and data acquisition

While FTICR-MS is a particularly impressive instrument and has the ability to provide molecular level details about NOM samples, there are many factors to take into account during data acquisition. It is important to consider the specific instrument's design and capabilities, in order to optimize certain parameters and decide how the data should be acquired so that high quality, meaningful mass spectra are obtained.

The first concern is for sample composition. There are several inherent difficulties with obtaining publishable mass spectra of NOM, particularly when the organic matter exists at a low concentration in the presence of a much higher concentration of inorganic matrix components. Salty samples are especially problematic, as emphasized in Fig. 3, which shows a NOM sample from the Elizabeth River that was analyzed by ESI-FTICR-MS before and after desalting by electrodialysis. An expanded region of the mass spectra acquired for the Elizabeth River NOM before desalting (Fig. 3a) shows high magnitude peaks with a high mass defect (0.7-0.8) and lower magnitude peaks at mass defects that are typical of organic matter (0.0-0.5) in this size range. Once the riverine NOM has been desalted (Fig. 3b), the high mass defect peaks attributed to salts are absent and the lower mass defect peaks are enhanced. Mass defect refers to the deviation of an m/z value from the exact nominal mass, and it is indicative of the type of compound present, based on the mass defect of atoms in organic compounds. The exact masses (in amu) of ^{12}C, ^{1}H, and ^{16}O are 12.000000, 1.007825, and 15.994915, respectively. Hydrogen has a positive mass defect, while oxygen has a negative mass defect. Thus compounds that are oxygen-rich and/or hydrogen-poor will display peaks at a lower mass defect (ca. 0.0-0.2), while compounds that are hydrogen-rich and/or oxygen-poor give peaks with a higher mass defect (ca. 0.2-0.5). The composition of peaks detected at high mass defect in Fig. 3a has not been confidently determined, but they are inorganic compounds present amongst the NOM sample. Their magnitudes are intense because they have higher ionization efficiencies than the OM constituents. Compounds that exist as ions in solution will ionize much more readily by ESI than compounds that do not. ESI is a competitive ionization process, and the OM components simply cannot out-compete the inorganic compounds for the negative charge, hence explaining why OM peaks are detected at lower magnitudes than the inorganic compounds. Because the OM compounds are the analytes of interest, desalting techniques are utilized to remove the background of inorganic matrix.

There is constant research being performed to determine the best method for desalting, isolating, and concentrating NOM samples. Traditionally, NOM was extracted from water, soils, and sediments as humic substances, which can be further categorized as fulvic acid, humic acid, and humin, depending on the pH at which they are soluble (Stevenson, 1994). Humic substances are extracted by well established acid/base laboratory protocols (Schnitzer and Khan, 1978; Thurman and Malcolm, 1981). More recently, DOM has been the focus of many studies interested in carbon cycling through groundwater, porewater, rivers, estuaries, and the ocean. Methods that are commonly employed to isolate the NOM from water are ultrafiltration, solid phase extraction, electrodialysis, and combined reverse osmosis electrodialysis. Each of these methods has certain problems associated, such as irreversible NOM sorption (lowering the NOM recovery of that technique), breakthrough

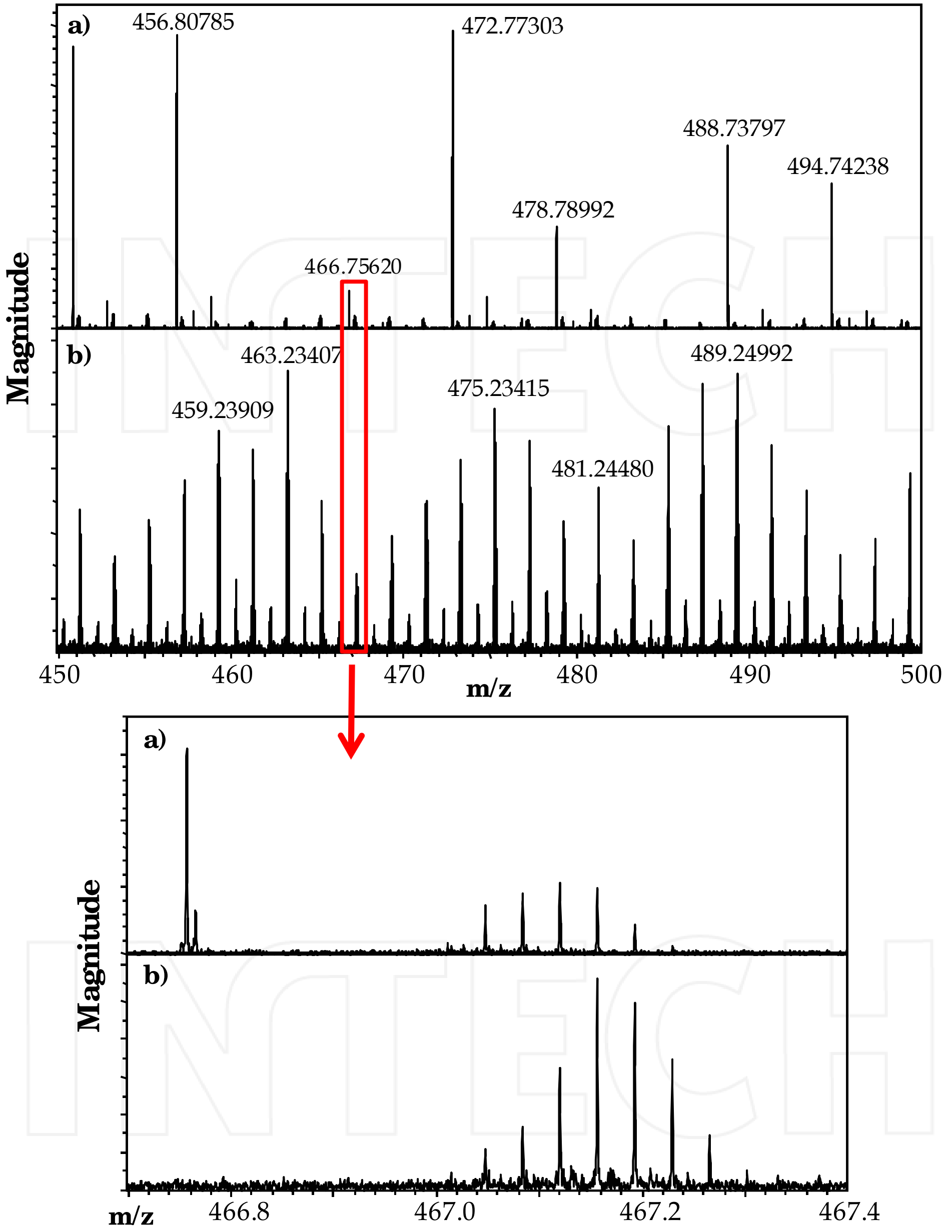


Fig. 3. Elizabeth River NOM analyzed by negative ion mode ESI-FTICR-MS before (a) and after (b) complete desalting by electrodialysis (expanded ranges are 450-500 m/z and 466.7-467.4 m/z). Peaks with high mass defects (0.7-0.9) are from incomplete desalting. More NOM peaks are detected when these compounds are not competing for a charge

contamination/bleeding, resin/membrane contamination of the NOM, and typically time consuming cleaning requirements (Simjouw et al., 2005; Mopper et al., 2007; Dittmar et al., 2008). These methods are biased by the chemical or physical properties that regulate the extraction procedure, and readers are referred to more in depth discussions of these desalting techniques, along with the pros and cons of each, in the references given above. In general, desalting and sample cleanup are important issues to consider when preparing samples for FTICR-MS.

After the appropriate sample preparation is performed, the optimal ionization source should be identified. There are many different commercially available API sources for the application to NOM (Hoffmann and Stroobant, 2003), such as ESI (Bruins, 1991; Gaskell et al., 1997; Cech and Enke, 2001), chemical ionization (CI; Bruins, 1991; Harrison, 1992), and atmospheric pressure photoionization (APPI; Raffaelli and Saba, 2003; Bos et al., 2006; Purcell et al., 2007). Each method varies in its ionization mechanism, and, consequently, the analytical window for each is quite different. It is important to mention that non-ionizable compounds that exist in the NOM will be invisible to the mass spectrometer. This is important, because each ion source has its own innate bias and each can give a different resulting mass spectrum for the same NOM sample (Hockaday et al., 2009). The ionization methods mentioned above are known as 'soft', meaning that compounds are not fragmented (as they are in electron ionization) and molecular ions ($M^{\bullet+}$, $M^{\bullet-}$) or pseudomolecular ions, also known as molecular ion adducts [$(M+H)^+$, $(M+Na)^+$, $(M-H)^-$, $(M-Cl)^-$], are predominantly observed in the mass spectrum. Mechanistic studies of each of the ion sources are referenced above, but brief explanations of each are described here.

CI introduces a reagent gas (methane is quite common, CH_4) into the ion source to produce primary ions of the reagent gas (i.e., $CH_4^{\bullet+}$) to collide with the molecule of interest. Through ion-molecule collisions, where the reagent gas ion acts as a Brønsted-Lowry acid and the analyte is a Brønsted-Lowry base, and proton transfer reactions, the analyte is ionized with minimal fragmentation. During ESI, the liquid sample is sprayed through a needle, and a high voltage difference between the spray needle and metal inlet induces a charge on the sprayed droplets. The charged droplet diminishes in size as the solvent is evaporated (by aid of heat or a drying gas), concentrating the charges held on the droplet. As charge-charge repulsions occur, the Rayleigh limit is exceeded, making the Coulombic repulsions greater than the surface tension of the droplet. The result is that the droplet bursts into many smaller droplets that can be completely desolvated, leaving only charged analyte ions in the gas phase for further introduction into the mass spectrometer. ESI operates in either positive or negative mode, depending on the functional group composition of the analyte. Functional groups that will readily lose a proton (such as alcohols, carboxylic acids, cyanides, peptides, nitric- and sulfonic- acids, and phosphates) are analyzed in negative ion mode. Basic functional groups that can easily gain a proton (i.e., amines, amides, peptides, and thiols) are analyzed in positive ion mode. By changing the pH of the sample solution (slightly basic for negative ion and slightly positive for positive ion), one can increase the ionization efficiencies for ESI. Another ion source of choice for analyzing less polar compounds is APPI, and APPI does not tend to suffer from charge competition with inorganic matrices, which is commonly observed in ESI. In APPI, ionization is initiated by supplying UV photons to the analyte molecule (typically via a krypton lamp). The analyte absorbs the photons and enters into an excited state. The analyte becomes ionized when the energy of the UV photons is greater than the ionization energy of the analyte. Dopants (such as toluene or tetrahydrofuran) are usually employed to act as intermediates between the photons and analytes, so that charge exchange and proton transfer reactions can occur more readily, increasing the ionization of NOM molecules.

Overall, before selecting an ion source, it is beneficial to know the bulk functional group composition of the sample (by previously obtained FTIR or NMR data), because then an informed decision can be made on which ion source(s) to employ for that specific sample. ESI and APPI are the most commonly used ion sources for the analysis of NOM, and Hockaday et al. (2009) performed a study to investigate which appeared to be optimal for a terrestrial sample obtained in the Dismal Swamp (Suffolk, VA, USA). They found that little overlap existed between the formulas assigned to APPI(+) and ESI(+ and -) mass spectral peaks, suggesting that data acquired from the two ion sources complemented each other greatly. Furthermore, APPI yielded formulas that were more aromatic and less polar than those from ESI, which is expected due to the ionization mechanism for each. Based on Hockaday et al. (2009) and the discussion of the types of compounds ionized by various ionization sources, many investigations of NOM do multiple-source FTICR-MS analyses to produce data that supplement each other, furthering the overall characterization of the NOM sample.

Once the best ion source has been determined for a particular sample, the mass spectrometric parameters can be examined. The two main ways to acquire mass spectral data are either broadband mode or narrow scan mode. The vast majority of publications that analyze NOM by FTICR-MS utilize broadband mode, where all m/z values, across a wide range of generally 200-2000 m/z, are detected during the analysis. Narrow scan mode is commonly referred to as sequential selective ion accumulation (SSIA; Sleighter et al., 2009) or as selected ion monitoring (SIM; Kido Soule et al., 2010). During SSIA (or SIM), the initial mass analyzer [quadrupole for Sleighter et al. (2009) and linear ion trap for Kido Soule et al. (2010)] isolates ions within a narrow range of m/z values before transferring the ion packet to the ICR cell. The operator selects the range, and ions outside of this range will be eliminated by the initial mass analyzer, thus decreasing the total number of ions in the ICR cell simultaneously. Because there are fewer ions in the cell at the same time, space-charge effects are minimized, generally increasing the resolving power and selectively enhancing the S/N of the peaks. Using SSIA, the sample is analyzed multiple times, incrementally increasing the m/z range, so that the entire mass range is eventually covered. By this SSIA method, the m/z ranges are acquired in sequential 'slices', and these slices can be merged together to assemble the entire mass spectrum.

Whole water from the Dismal Swamp (Suffolk, VA, USA) was sterile filtered (0.2 μm) to remove particulates and bacteria and analyzed directly using both broadband and SSIA modes, as shown in Fig. 4. The colors in Fig. 4b show each 'slice' that was acquired, and the 'slices' overlap by approximately 30-40 m/z to ensure that no area is missed. The nominal mass region in Fig. 4 is shown in order to highlight the increase in S/N and resolving power. An S/N threshold of 3 is commonly used for peak-picking, and the increase in overall S/N using SSIA leads to a larger number of peaks detected. Table 2 gives details for the mass spectra shown in Fig. 4, parsed according to the m/z range that was acquired. This table shows that the increase in the number of peaks, S/N, and generally resolving power exists for SSIA across the entire mass range and is more substantial and pronounced at higher m/z values. The broadband analysis requires 20 minutes (1.0 sec ion accumulation and 200 co-added scans), while the SSIA analysis requires 30 minutes (1.0 sec ion accumulation and 50 co-added scans, for 6 separate m/z ranges). While the SSIA mode involves more time to acquire data, 2880 more peaks were detected during that time. It is important to note that while nearly 3000 more peaks were detected with SSIA, this does not necessarily translate to 3000 more molecular formulas being assigned, because some of the extra peaks detected by SSIA are isotopic peaks.

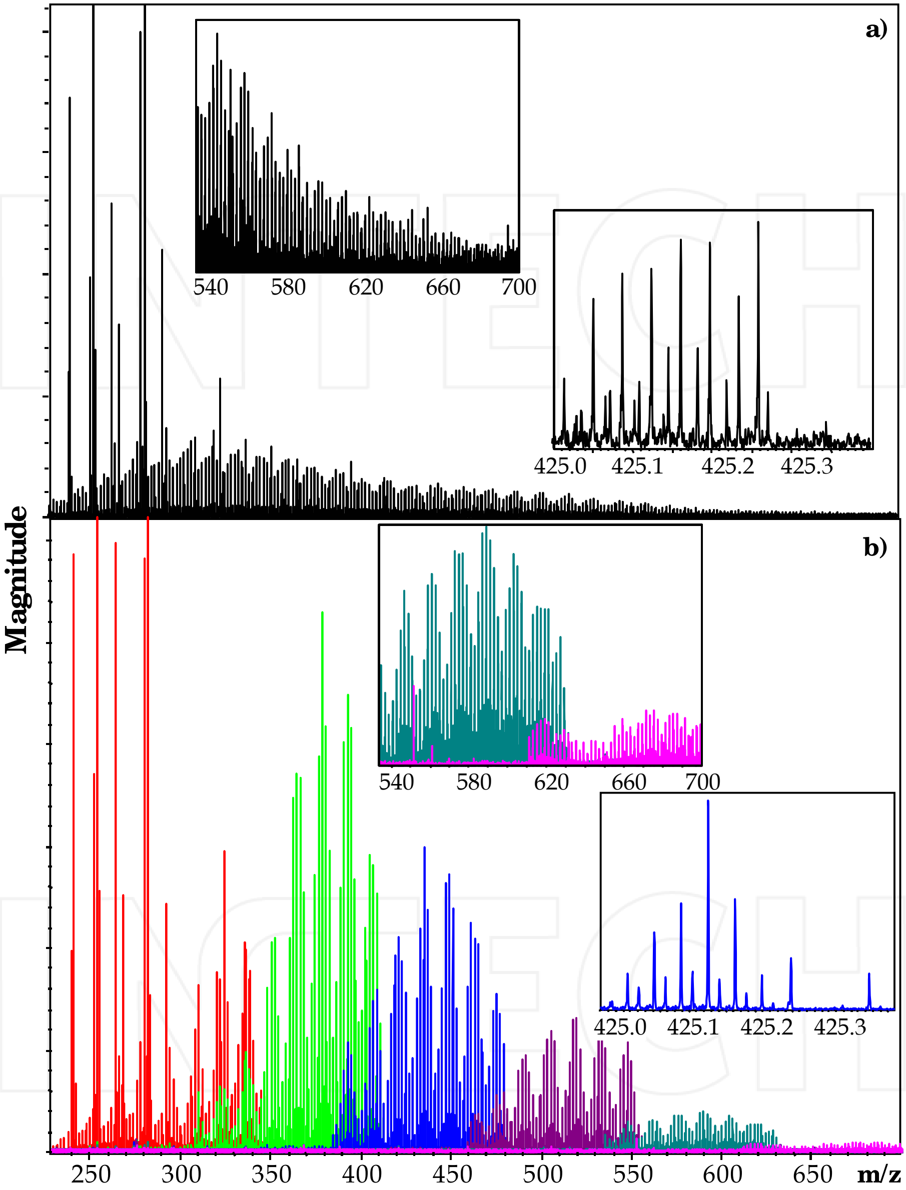


Fig. 4. Dismal Swamp whole water analyzed by negative ion mode ESI-FTICR-MS in broadband mode (a) and using sequential selective ion accumulation (b). The insets show expanded mass spectra at m/z 540-700 and 425.0-425.35, to show the peaks at higher m/z more clearly and to highlight the enhanced S/N achieved using SSIA

Broadband Mode					Sequential Selective Ion Accumulation			
m/z range	Number of Peaks	S/N [a]	Resolving Power [a]		m/z range	Number of Peaks	S/N [a]	Resolving Power [a]
200-300	651	25.2	491690		200-300	761	28.5	522998.6
300-400	1441	15.6	401404		300-400	2197	28.8	354605.1
400-500	959	11.8	302398		400-500	1707	27.6	273525.8
500-600	571	7.9	252256		500-600	1205	18.4	229191.6
600-700	81	4.9	215090		600-700	713	8.9	216097.5
200-700	3703	14.9	364562		200-700	6583	24.4	315089.0

[a] The average value is given for the m/z range specified.

Table 2. The number of peaks detected, average S/N, and average resolving power for each m/z range of the broadband mass spectra and SSIA mass spectra shown in Fig. 4

The discussion in this section of sample preparation, ion source selection, and optimizing FTICR-MS acquisition modes emphasizes just a few factors that are of paramount importance when analyzing NOM. There are countless other mass spectral parameters that can also be optimized (ion source voltages, ion optics and transmission parameters, ion accumulation times, trapping parameters within the ICR cell, etc.) in order to obtain the highest quality data. However, these parameters are instrument specific, and each individual instrument requires tuning before each sample set is analyzed. Further discussion and advice on these parameters is generally reviewed in detail by the manufacturer and is beyond the scope of this chapter.

4. FTICR-MS applications to NOM

As discussed above, molecular formulas can be assigned to the multitude of peaks detected in mass spectra of NOM, and this capability is the main justification for use of FTICR-MS for the molecular characterization of various NOM samples. The molecular formulas provide meaningful compositional information that is associated with groups of natural biopolymers. This practice of correlating assigned formulas to sample composition has been applied to various types of NOM, and several methods exist to assist in this correlation. Because it can be difficult, tedious, and labor-intensive to compare the thousands of assigned molecular formulas for a single NOM sample, generally visualization diagrams are called upon to assist in displaying the formulas in a chemically representative manner. The two-dimensional van Krevelen diagram has been the most commonly used approach, which plots H/C values vs. O/C values (van Krevelen, 1950; Kim et al., 2003). Each molecular formula aligns on the diagram in a location that can typically be correlated to that commonly associated with natural biomolecules, as shown in Fig. 5 for riverine DOM isolated from the Elizabeth River in south-eastern Virginia, USA. The circles overlain on the plot highlight the types of molecules that are commonly detected in NOM samples, as well as their position on the van Krevelen diagram, based on the compound's elemental ratios (Kim et al., 2003; Sleighter and Hatcher, 2007; Hockaday et al., 2009). It should be noted that these circles are not strictly representative of all similar molecules, but rather approximate guidelines for identifying compounds of similar composition. Relating formulas to compound classes in this manner has been exploited in many studies of various types of NOM over the years, and interested readers are referred to the literature for more details (Sleighter and Hatcher, 2007; Reemtsma, 2009; references therein).

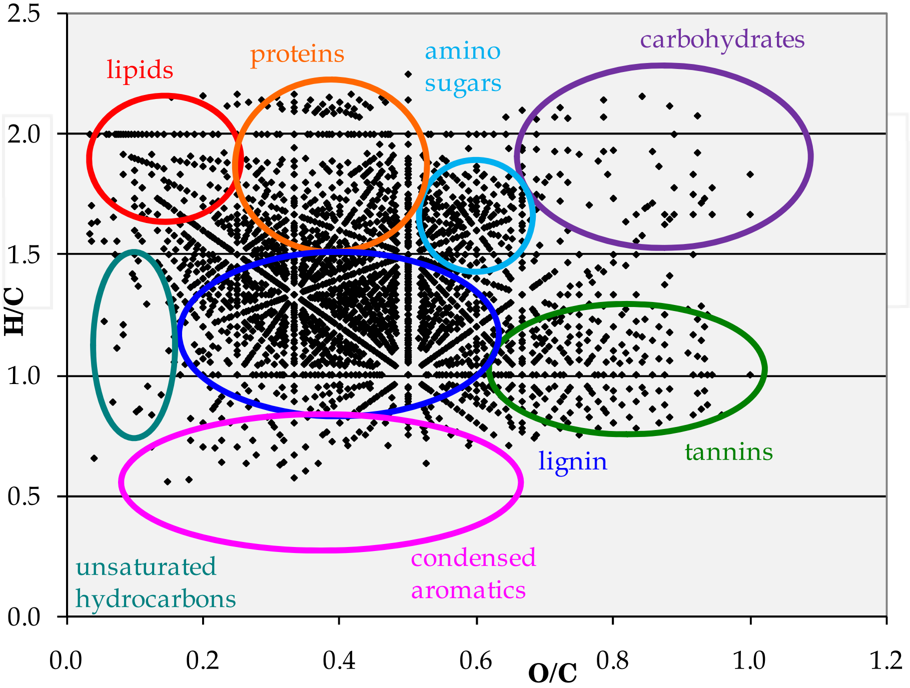


Fig. 5. van Krevelen diagram for Elizabeth River NOM isolated by small-scale electrodialysis and subsequent analysis by negative ion mode ESI-FTICR-MS. Overlain circles are used as broad indicators of where biomolecules fall on the plot (Sleighter and Hatcher, 2008; Hockaday et al., 2009; Ohno et al., 2010)

Another diagram that is often used for NOM characterization is the Kendrick mass defect (KMD) plot (Kendrick, 1963). KMD analysis converts m/z values to Kendrick mass values by multiplying the m/z value by the ratio of the nominal mass CH_2 group (14.00000) to the exact mass of a CH_2 group (14.01565), as shown below in equation 3. Then, KMD is determined by subtracting the nominal Kendrick mass (KM) from KM, as shown in equation 4. KMD values can then be plotted against their nominal KM values (as shown in Fig. 6), and formulas with the same KMD, those falling on a horizontal line, differ only by a CH_2 group (or multiple CH_2 groups). KMD values increase with the number of added H atoms, thus aliphatic compounds will have high KMD values and aromatic compounds will have lower KMD values. This coincides with mass spectral data expanded at individual nominal masses, where m/z values with low mass defects are hydrogen-poor and m/z values with higher mass defects are hydrogen-rich (as shown in Fig. 2e and Table 1).

$$\text{Kendrick Mass (KM)} = \text{m/z value} * (14.00000/14.01565) \qquad (3)$$

$$\text{Kendrick Mass Defect (KMD)} = \text{KM} - \text{nominal KM} \tag{4}$$

Originally, KMD was utilized for assigning molecular formulas, by establishing homologous CH_2 series that could be expanded from low m/z to high m/z (Stenson et al., 2003). Peaks at low m/z can more easily be assigned a molecular formula because fewer formulas exist within the selected error limit (usually 0.5 ppm). Generally, only 1 molecular formula exists within this error for peaks less than 500 m/z, but beyond this value, multiple formulas are possible. Once molecular formulas are unambiguously assigned to peaks of low m/z, peaks at high m/z values that have multiple formula choices can be related to formulas assigned at lower mass by assuming that they belong to a CH_2 homologous series. If one of the formulas belongs to a homologous series, then it is very likely the correct formula, and the others can be eliminated. This approach to formula assignment is called 'formula extension' and can be performed manually or written into software designed to assist in formula assignment (Kujawinski and Behn, 2006; Grinhut et al., 2010). While CH_2 is the most commonly used group, other functional groups can be utilized (i.e., OCH_2, COO, O, H_2O, H_2, etc.) depending upon the make-up of the sample (Sleighter and Hatcher, 2007).

Because of the recent growth in use of FTICR-MS for the analysis of NOM, articles have been published summarizing the findings of these studies and making suggestions for future work (Sleighter and Hatcher, 2007; Reemtsma, 2009). Most recently, multivariate statistical analysis in combination with visualization diagrams have been utilized to evaluate relationships among sample sets. Hierarchal cluster analysis (HCA) seeks correlations among samples displayed in a data matrix and illustrates the results in a hierarchical tree, or a dendrogram, where the branching reveals the similarity among the samples. HCA has been utilized in numerous studies, one of which shows that there are no significant differences between the formulas assigned to a depth profile of DOM from the Weddell Sea (Koch et al., 2005). In other studies, Dittmar et al. (2007) compared DOM from coastal mangrove forests (before and after photo-irradiation) to open ocean seawater. They found that photo-degraded mangrove DOM becomes similar in composition to open ocean seawater DOM. Koch et al. (2008) evaluated the various fractions of DOM collected from reversed phase HPLC separations, and Schmidt et al. (2009) used HCA statistical correlations to differentiate between pore water DOM and riverine DOM. While HCA is very useful for grouping samples based on their similarity, it does not indicate the reasons why the samples are similar or different. The variation between the samples is not explained, and another method is required for this determination.

By combining HCA with another statistical method, the variance between samples can be elucidated. Kujawinski et al. (2009) employed HCA combined with both non-metric multi-dimensional scaling (NMS) and indicator species analysis (ISA) to optimize the *a priori* grouping of samples for subsequent ISA, where specific mass spectral m/z values can be identified as an indicator species for a certain group of samples. Once all the indicator species for the samples are identified, then molecular formulas were examined more closely. Indicator species for surface ocean DOM samples were speculated to be biologically-derived and to represent a more labile component of the marine DOM pool, while indicator species for riverine/estuarine DOM were found to be similar in composition to lignin-derived species that have been linked to terrestrially-sourced DOM. Bhatia et al. (2010) also used these multivariate statistical methods recently to characterize DOM from the Greenland ice sheet and were able to link subglacial, supraglacial, and proglacial DOM to various allochthonous and autochthonous sources and processes. Another recent study utilizing ISA distinguished compounds that are specific to various fractions of NOM in soils and crop biomass (Ohno et al., 2010). It was discovered that water extractable OM from plant biomass had marker components that could be classified as lipids, proteins, carbohydrates, lignin, and unsaturated hydrocarbons, while the water extractable OM from soils contained more lignin- and carbohydrate-sourced compounds. The mobile humic acid extract of soils displayed mostly lignin-like markers and the immobile humic acid markers clustered in the condensed aromatic space. A general trend of increasing aromaticity (i.e., decreasing H/C ratio) was observed along the humification gradient, from plant biomass to water-extractable soil OM to refractory, stabilized humic acids.

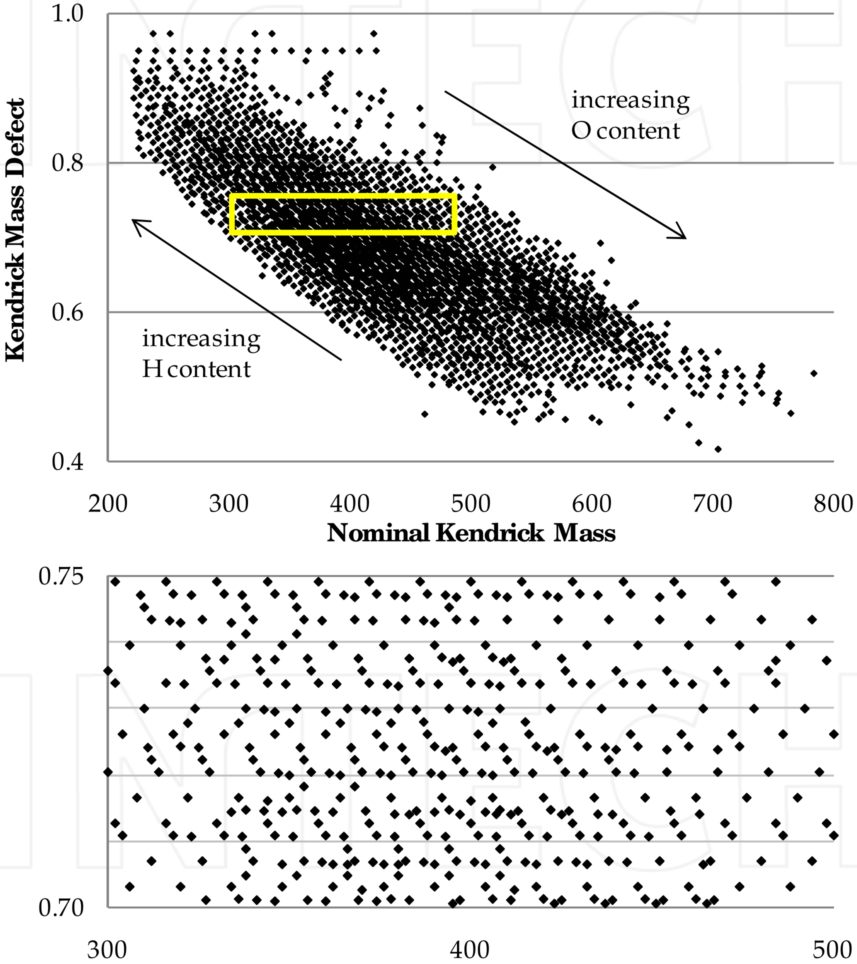


Fig. 6. Kendrick mass defect plot (using a CH_2 group) for Elizabeth River NOM isolated by small-scale electrodialysis and subsequent analysis by negative ion mode ESI-FTICR-MS. The yellow box is expanded in the lower diagram to highlight the points that fall on a horizontal line, indicating that they are part of a homologous CH_2 series

By combining HCA with PCA, Hur et al. (2010) developed a method for comparing multiple petroleum samples that were analyzed by APPI-FTICR-MS. Using these multivariate statistical tools, twenty petroleum samples could be compartimentalized into numerous compositional groups, such as those enriched in hydrocarbons, oxygen series (O_1, O_2), nitrogen series (N_1, N_1O_1), or sulfur series (S_1, S_2, O_1S_1). Sleighter et al. (2010) also employed a combined approach of HCA and PCA, but in this study the goal was to characterize 38 NOM samples along a terrestrial to marine transect of the lower Chesapeake Bay and coastal Atlantic Ocean, using samples that were prepared for mass spectral analysis by either sterile filtration only, solid phase C_{18} extraction, or small-scale electrodialysis. Not only were differences detected between NOM samples from various locations, but it was also found that the method of preparation for the same NOM sample changed its composition, as determined by ESI-FTICR-MS. Terrestrial samples contained lignin, tannin, and condensed aromatic structures in high relative magnitude, while marine DOM was composed more of aliphatic and lignin-like compounds, as well as compounds containing more heteroatom (NSP) functionalities. Samples desalted by electrodialysis retained the more polar compounds (tannins, carbohydrates, and those containing heteroatoms) that were eliminated during the C_{18} extraction procedure. Overall, these recent developments in NOM characterization, made by exploiting multivariate statistical analyses, highlight the direction in which the NOM community is progressing. It is very likely that statistical analyses will become more commonplace, especially as researchers continue to accumulate large sample sets that would be otherwise very difficult to analyze manually.

5. Limitations of FTICR-MS

A major concern regarding the use of FTICR-MS for NOM characterization in recent years has been establishing an appropriate instrumental/solvent blank, to evaluate the background peaks detected in the mass spectra. Most analytical techniques have a straightforward manner for nulling the blanks, such as a basic background subtraction of the blank analysis from the sample analysis. Unfortunately, this method cannot be applied so simply when using a competitive ionization source such as ESI. Most NOM samples are analyzed in a mixture of methanol and water, thus a clean solvent blank of methanol and water is typically evaluated before analyzing samples. However, any contaminants present within the instrument itself will not have any analyte molecules with which to compete for charge. Thus, these substances acquire most of the charge, giving them an enhanced magnitude in the mass spectra. Once a sample is introduced, the analyte molecules out-compete the contaminants for the charge in most cases, and then those contaminants are only observed at lower absolute and relative magnitudes, as shown for Mount Rainier humic acid in Fig. 7. The NOM sample is analyzed at various concentrations, to determine which concentration is sufficient to out-compete most (if not all) of the contaminants present in the FTICR-MS. For peaks that are detected in both the blank and the NOM sample, it is difficult to determine if that peak is actually in the sample or if it is due to contamination. In the case of Fig. 7, m/z 415.322 is detected at nearly a constant magnitude in the instrument/ solvent blank and at each concentration of Mount Rainier humic acid. Because the other contaminants (at 414.8–414.9 and 415.27) are nearly eliminated at 50 mg/L OM, it is likely that m/z 415.322 is present in the NOM sample, since it continues to be detected at all NOM concentrations. Nonetheless, some researchers err on the side of caution and argue that this peak should be removed from further analysis, since it is detected in the blank as well. Only 1 nominal mass region is highlighted in Fig. 7 as an example, but this trend in blank dilution/elimination occurs across the entire mass spectral range of 200-800 m/z. Thus, it is common to question the existence of several hundred peaks that are detected in each NOM sample that also overlap with the peaks in the blank. There is not currently a standard protocol among researchers for determining when to remove peaks from analysis, but typically the most conservative approach is taken and overlapping peaks in the blank and sample are removed from consideration as analyte peaks. It is important, though, to understand the overlap between solvent blank and sample analysis, so that peaks that are important to the NOM composition are not considered artifacts of the instrument itself.

Another concern, when analyzing samples as complex as NOM, is peak reproducibility. The vast majority of published studies only analyze each sample once, rather than in replicate. This is due to sample throughput and cost. However, more recently, there has been more concern for reproducibility in order to characterize NOM reliably. The sample must be analyzed, in either duplicate or triplicate, to ensure reproducibility. This facilitates comparison of solvent or instrument blanks to replicate injections, enabling decisions as to which peaks should be included for further data analysis. Kido Soule et al. (2009) discovered that broadband acquisitions offer the highest repeatability for peaks detected and peak height (along with highest throughput), when compared to SSIA (or SIM) acquisition. In a recent study, Hur et al. (2010) analyzed 20 petroleum samples by APPI in triplicate over the course of 2 consecutive days. They reported that the three mass spectra for each sample were quite consistent, showing standard deviations of less than 5% of the initial values. Sleighter et al. (2010) performed a study comparing a single sample analyzed on different days by replicate injections of Dismal Swamp (Suffolk, VA, USA) whole water over the course of 31 days. The mass spectra produced were visually very similar, and multivariate statistics was utilized to evaluate the reproducibility. Based on HCA and PCA, the different analyses were found to be virtually identical when compared to a variety of other samples. These three studies highlight that ESI-FTICR-MS has the potential to be very reproducible and reliable, when the same instrumental parameters are utilized. However, no studies have tested the repeatability for various instrument operators or for the same sample analyzed on different FTICR-MS instruments.

Another unknown factor in the analysis is the fraction of the NOM that is identified by FTICR-MS. The ion sources described here for coupling to FTICR-MS are well known to exhibit biases for certain types of molecules, depending on the ionization mechanism and the ionization efficiency of the analyte in the midst of a complex matrix. If only 5% of the NOM is 'observed' by FTICR-MS, rather than 50% (or perhaps even higher), then the implications are enormous for complete characterization. Based on this concern, Hockaday et al. (2009) suggest that an internal standard be made widely available for spiking into NOM samples prior to FTICR-MS analysis. Using this approach, the ion source performance on various instruments could be assessed and compared. Furthermore, this standard could also used for internal calibration, making datasets acquired on different instruments by various operators more comparable.

Once a high quality mass spectrum is obtained for a NOM sample, data analysis can proceed in order to obtain molecular level characterization. This characterization is performed by assigning the m/z values of peaks detected to molecular formulas, by using a molecular formula calculator. First, one must determine which atoms to consider for assignment. Typically C, H, O, N, and S are used, although some studies also include P, Na, and/or Cl. It is well known that the number of possible molecular formulas increases with increasing 1) permissible error difference between measured m/z and exact calculated m/z for the formula in question; 2) m/z value (higher m/z peaks are inherently less precisely measured); and 3) number of atoms used for formula assignment (Kim et al., 2006;

Kujawinski and Behn, 2006; Koch et al., 2007; Reemstma, 2009), as shown in Fig. 8. As one can see, increasing the allowable error and including more elements significantly increases the number of chemically possible formulas. The relationship between resolving power and the error difference between two formulas with slight differences in m/z values is reviewed in the four references given above, but Kim et al. (2006) reported that all theoretically possible elemental compositions of C,H,N,O,S up to 500 Da could be resolved at an accuracy of approximately 0.1 mDa (corresponding to a resolving power of 5,000,000 at m/z 500), allowing for a unique molecular formula to be assigned with confidence. While resolving powers of this magnitude are not currently routinely achieved during the analysis of NOM, it is likely that as technology continues to improve and FTICR-MS instruments at high magnetic fields are utilized, that these values will be achieved in the near future.

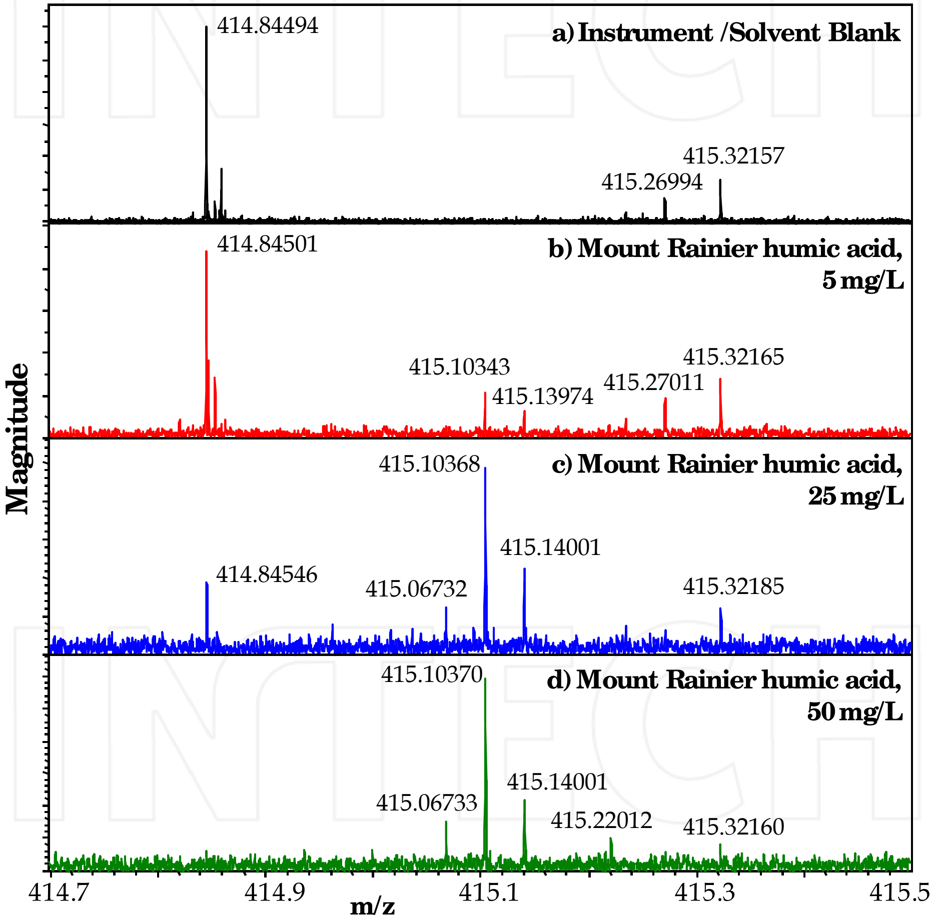


Fig. 7. Negative ion mode ESI-FTICR mass spectra of a) instrument/solvent blank (1:1 water:methanol with 0.1% ammonium hydroxide) and Mount Rainier humic acid dissolved at 5 (b), 25 (c), and 50 (d) mg/L organic matter in 1:1 water:methanol with 0.1% ammonium hydroxide. The broadband mass spectrum has been expanded at 414.7 – 415.5 m/z

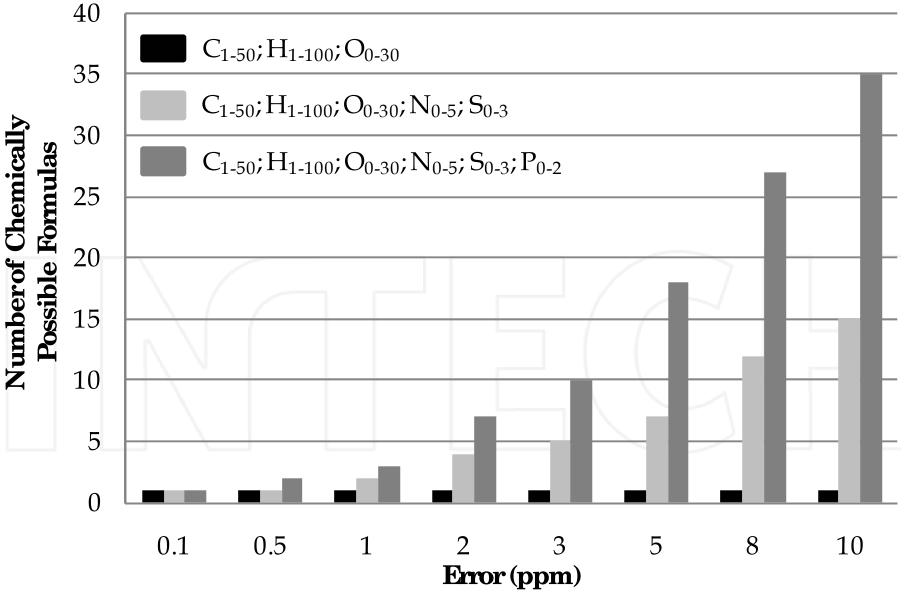
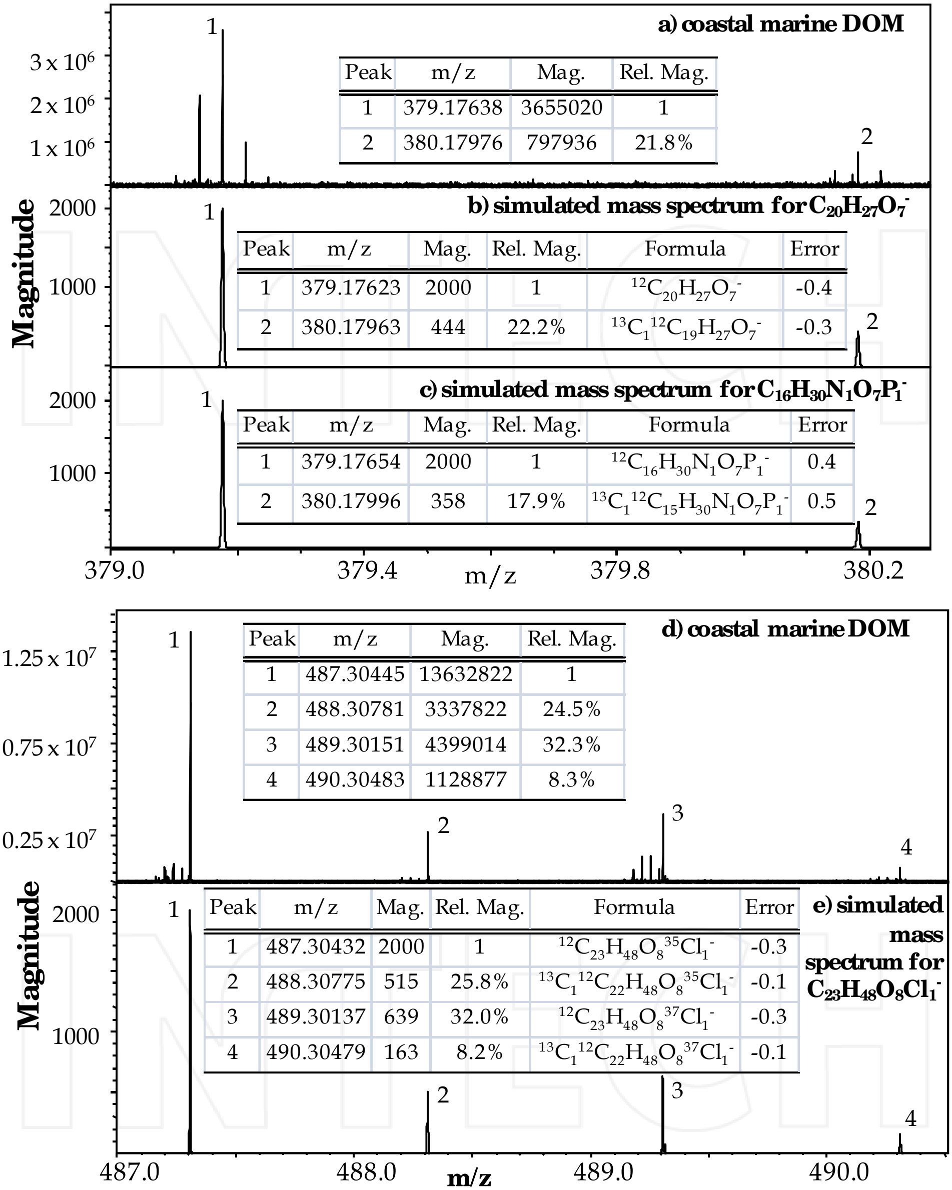


Fig. 8. The number of chemically possible molecular formulas for hypothetical m/z 499.21257 at various error values for the different elemental compositions specified in the legend

The inherent complexity incurred while assigning a unique molecular formula to an individual m/z value is highlighted in Fig. 8, and this difficulty is amplified when the task at hand is to assign thousands of m/z values to formulas for a single NOM sample. This process can be very labor-intensive. However, by carefully calibrating the mass spectral data, lower error differences can be tolerated (0.5 ppm has become quite common in the literature), to minimize the number of formulas that match the measured mass within the selected error. There are also other methods to assist in the determination of molecular formulas, such as using KMD analysis and the 'formula extension' approach, as described above. Furthermore, establishing rules that the formulas must obey (as in Stubbins et al., 2010), to be assigned to a chemically relevant molecule, also helps to reduce the number of possible formulas. Formulas that are unlikely to be acceptable can be identified by examining the isotopic peak(s) as described by Koch et al. (2007) and featured in Fig. 9. Based on the natural abundance of ^{13}C (1.1%), the number of carbons in the correct formula can be calculated from the relative magnitude of the ^{13}C peak, eliminating incorrect formulas (Fig. 9a). In the case of Fig. 9b and 9c, $C_{20}H_{27}O_7^-$ is determined to be the correct formula for the peak, because the predicted relative abundance of its ^{13}C isotopologue is much closer to what is detected in Fig. 9a than it is for the formula $C_{16}H_{30}N_1O_7P_1^-$ in Fig. 9c. For Fig. 9d, the peak at 487.30445 is not initially assigned a molecular formula, because Cl is not included in the formula assignment parameters. Upon closer inspection of the peak, the distinctive isotope pattern for Cl^- adducts is observed and matches very closely to that of a simulated theoretical pattern for $C_{23}H_{48}O_8Cl_1^-$ (Fig. 9e). Koch et al. (2007) cautions that the S/N of the isotopic peak should be at least 25, otherwise the deviation (between the number of carbons in the proposed formula and that calculated from the relative magnitude) is too great to be a reliable tool for formula elimination. In general, by ensuring that an accurate



a) coastal marine DOM

Peak	m/z	Mag.	Rel. Mag.
1	379.17638	3655020	1
2	380.17976	797936	21.8%

b) simulated mass spectrum for $C_{20}H_{27}O_7^-$

Peak	m/z	Mag.	Rel. Mag.	Formula	Error
1	379.17623	2000	1	$^{12}C_{20}H_{27}O_7^-$	-0.4
2	380.17963	444	22.2%	$^{13}C_1{}^{12}C_{19}H_{27}O_7^-$	-0.3

c) simulated mass spectrum for $C_{16}H_{30}N_1O_7P_1^-$

Peak	m/z	Mag.	Rel. Mag.	Formula	Error
1	379.17654	2000	1	$^{12}C_{16}H_{30}N_1O_7P_1^-$	0.4
2	380.17996	358	17.9%	$^{13}C_1{}^{12}C_{15}H_{30}N_1O_7P_1^-$	0.5

d) coastal marine DOM

Peak	m/z	Mag.	Rel. Mag.
1	487.30445	13632822	1
2	488.30781	3337822	24.5%
3	489.30151	4399014	32.3%
4	490.30483	1128877	8.3%

e) simulated mass spectrum for $C_{23}H_{48}O_8Cl_1^-$

Peak	m/z	Mag.	Rel. Mag.	Formula	Error
1	487.30432	2000	1	$^{12}C_{23}H_{48}O_8{}^{35}Cl_1^-$	-0.3
2	488.30775	515	25.8%	$^{13}C_1{}^{12}C_{22}H_{48}O_8{}^{35}Cl_1^-$	-0.1
3	489.30137	639	32.0%	$^{12}C_{23}H_{48}O_8{}^{37}Cl_1^-$	-0.3
4	490.30479	163	8.2%	$^{13}C_1{}^{12}C_{22}H_{48}O_8{}^{37}Cl_1^-$	-0.1

Fig. 9. Coastal marine DOM isolated by C_{18} solid phase extraction and subsequent analysis by negative ion mode ESI-FTICR-MS, expanded at 379.0-380.2 m/z (a) and 487.0-490.5 m/z (d), along with the simulated isotope patterns for the possible formulas specified (b, c, e). Error values are in ppm

calibration is achieved, setting a low error difference, utilizing KMD analysis and elemental ratio rules, and exploiting isotopic relative abundances, one can assign molecular formulas more easily and with additional confidence. Ideally, a generally accepted protocol should be established and each of these tools could be written into a standard software program, similar to that developed by Kujawinski and Behn (2006). Then data analysis consistency between samples would be more reliable, and human error introduced by different analysts could be minimized. This specific step has yet to be taken, but with emerging literature in which suggestions are made regarding formula assignment rules, some standardization will become reality in the near future.

Even though data analysis of FTICR mass spectra of NOM can provide accurate molecular formulas for the peaks detected, structural information for these peaks, which is often the key to understanding the source of the NOM, still remains elusive. Previous FTICR-MS studies focused on molecular characterization of DOM from terrestrial and marine waters have found that a significant overlap of the mass spectra is observed, as approximately 30% of the assigned formulas are shared between the two different types of DOM (Koch et al., 2005; Sleighter and Hatcher, 2008). However, mass spectrometry cannot distinguish structural isomers of an elemental composition, so the possibility exists that identifying the same elemental formula in multiple samples does not necessarily mean that their molecular structures also correspond. For these reasons, tandem mass spectrometry (MS/MS) has been employed to isolate and fragment ions in the mass spectrometer, and this can provide structural information. Because the ions detected during the mass spectral analysis of NOM are predominantly singly charged, only neutral losses (i.e., H_2O, COO, CO, and OCH_2) have been observed by MS/MS using FTICR-MS (Stenson et al., 2003; Reemtsma et al., 2008; Witt et al., 2009; Liu et al., 2011) and other lower resolution mass spectrometers (Fievre et al., 1997; Plancque et al., 2001; Leenheer et al., 2001; McIntyre et al., 2002). Furthermore, the complexity of NOM samples and the proximity of peaks detected at each nominal mass makes it particularly difficult to isolate a single peak in the mass spectrometer for fragmentation and subsequent detection of those fragments. While it is possible to do this, as shown by Witt et al. (2009), fragmenting each peak detected at every nominal mass from approximately 200-800 m/z by FTICR-MS/MS is quite a daunting task, one that will be labor intensive and require long instrument run times. The mass of data that would be acquired during such a study would also require months (if not longer) to correlate and understand. Nonetheless, the wealth of information obtained from a very thorough study could perhaps answer many questions regarding the structures of overlapping molecular formulas, but it is more likely that MS/MS will be utilized in the near future for specific target compounds that have been found to be markers for particular samples.

6. Conclusions

FTICR-MS coupled to API sources is clearly a powerful technique for the examination of the complex composition of NOM and has facilitated the overall characterization of NOM from a variety of source materials. The advent of API sources has allowed for the ionization of large, nonvolatile compounds, and its application to the polar, polyelectrolytic NOM mixtures has made a significant contribution to the understanding of the composition and reactivity of NOM. The ultrahigh resolving powers of FTICR-MS, those in the range of 10^5, can separate and resolve the numerous peaks per nominal mass detected for NOM samples, making it the mass spectrometer of choice for investigations of NOM. Furthermore, the accuracy of the instrument is capable of determining m/z values to the fifth decimal place, from which molecular formula assignments can be made fairly reliably. The ability to assign formulas to the multitude of peaks detected across the mass range of 200-1000 allows for the characterization of NOM at the molecular level.

In this review, we have described the application of FTICR-MS to the analysis of NOM, discussed some important parameters for sample preparation and the acquisition of high quality data, reviewed some of the recent publications of NOM studies utilizing FTICR-MS, assessed the emergence and importance of recent statistical methods, and evaluated the current limitations of the instrument and subsequent data processing. All things considered, the precision, sensitivity, and ultrahigh resolution offered by FTICR-MS reveals molecular level details of NOM composition, which has transformed our knowledge of NOM chemistry. We are confident that as FTICR-MS technology continues to improve, that this instrumentation will be utilized more widely and will lead to significant advancements in not only the areas of analytical and environmental chemistry, but also for the NOM, chemical oceanography, and mass spectrometry communities.

7. References

Bhatia, M.P.; Das, S.B.; Longnecker, K.; Charette, M.A. & Kujawinski, E.B. (2010). Molecular characterization of dissolved organic matter associated with the Greenland ice sheet. *Geochimica et Cosmochimica Acta*, 74, 13, 3768-3784.

Bos, S.J.; van Leeuwen, S.M. & Karst, U. (2006). From fundamentals to applications: recent developments in atmospheric pressure photoionization mass spectrometry. *Analytical and Bioanalytical Chemistry*, 384, 1, 85-99.

Bruins, A.P. (1991). Mass spectrometry with ion sources operating at atmospheric pressure. *Mass Spectrometry Reviews*, 10, 1, 53-77.

Canadell, J.G.; Le Quéré, C.; Raupach, M.R.; Field, C.B.; Buitenhuis, E.T.; Ciais, P.; Conway, T.J.; Gillett, N.P.; Houghton, R.A. & Marland, G. (2007). Contributions to accelerating atmospheric CO2 growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proceedings of the National Academy of Sciences*, 104, 47, 18866-18870.

Cech, N.B. & Enke, C.G. (2001). Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews*, 20, 6, 362-387.

Dittmar, T.; Whitehead, K.; Minor, E.C. & Koch, B.P. (2007). Tracing terrigenous dissolved organic matter and its photochemical decay in the ocean by using liquid chromatography/mass spectrometry. *Marine Chemistry*, 107, 3, 378-387.

Dittmar, T.; Koch, B.; Hertkorn, N. & Kattner, G. (2008). A simple and efficient method for the solid phase extraction of dissolved organic matter (SPE-DOM) from seawater. *Limnology and Oceanography: Methods*, 6, 230-235.

Eglinton, T.I. & Repeta, D.J. (2003). Organic matter in the contemporary ocean, In: *Treatise on Geochemistry*, *volume 6*, H.D. Holland & K.K. Turekian (Eds.), 145-180, Elsevier, Morristown.

Fievre, A.; Solouki, T.; Marshall, A.G. & Cooper, W.T. (1997). High-resolution Fourier transform ion cyclotron resonance mass spectrometry of humic and fulvic acids by laser desorption/ionization and electrospray ionization. *Energy and Fuels*, 11, 3, 554-560.

Gaskell, S.J. (1997). Electrospray: principles and practice. *Journal of Mass Spectrometry*, 32, 7, 677-688.

Grinhut, T.; Lansky, D.; Gaspar, A.; Hertkorn, N.; Schmitt Kopplin, P.; Hadar, Y. & Chen, Y. (2010). Novel software for data analysis of Fourier transform ion cyclotron resonance mass spectra applied to natural organic matter. *Rapid Communications in Mass Spectrometry*, 24, 19, 2831-2837.

Harrison, A.G. (1992). *Chemical ionization mass spectrometry*, CRC Press, Boca Raton.

Hatcher, P.G.; Dria, K.J.; Kim, S. & Frazier, S.W. (2001). Modern analytical studies of humic substances. *Soil Science*, 166, 11, 770-794.

Hedges, J.I. (1992). Global biogeochemical cycles: progress and problems. *Marine Chemistry*, 39, 1-3, 67-93.

Hockaday, W.C.; Purcell, J.M.; Marshall, A.G.; Baldock, J.A. & Hatcher, P.G. (2009). Electrospray and photoionization mass spectrometry for the characterization of organic matter in natural waters: a qualitative assessment. *Limnology and Oceanography: Methods*, 7, 81-95.

Hoffmann, E. & Stroobant, V. (2003). *Mass Spectrometry: Principles and Applications*, Wiley, Chichester.

Hur, M.; Yeo, I.; Park, E.; Kim, Y.; Yoo, J.; Kim, E.; No, M.; Koh, J. & Kim, S. (2010). Combination of statistical methods and Fourier transform ion cyclotron resonance mass spectrometry for more comprehensive, molecular-level interpretations of petroleum samples. *Analytical Chemistry*, 82, 1, 211-218.

Kendrick, E. (1963). A mass scale based on CH2= 14.0000 for high resolution mass spectrometry of organic compounds. *Analytical Chemistry*, 35, 13, 2146-2154.

Kido Soule, M.C.; Longnecker, K.; Giovannoni, S.J. & Kujawinski, E.B. (2010). Impact of instrument and experiment parameters on reproducibility of ultrahigh resolution ESI FT-ICR mass spectra of natural organic matter. *Organic Geochemistry*, 41, 8, 725-733.

Kim, S.; Kramer, R.W. & Hatcher, P.G. (2003). Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van Krevelen diagram. *Analytical Chemistry*, 75, 20, 5336-5344.

Kim, S.; Rodgers, R.P. & Marshall, A.G. (2006). Truly "exact" mass: Elemental composition can be determined uniquely from molecular mass measurement at similar to 0.1 mDa accuracy for molecules up to similar to 500 Da. *International Journal of Mass Spectrometry*, 251, 2-3, 260–265.

Koch, B.P.; Witt, M.; Engbrodt, R.; Dittmar, T. & Kattner, G. (2005). Molecular formulae of marine and terrigenous dissolved organic matter detected by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Geochimica et Cosmochimica Acta*, 69, 13, 3299-3308

Koch, B.P.; Dittmar, T.; Witt, M. & Kattner, G. (2007). Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter. *Analytical Chemistry*, 79, 4, 1758-1763.

Koch, B.P.; Ludwichowski, K.U.; Kattner, G.; Dittmar, T. & Witt, M. (2008). Advanced characterization of marine dissolved organic matter by combining reversed-phase liquid chromatography and FT-ICR-MS. *Marine Chemistry*, 111, 3-4, 233-241.

Kujawinski, E.B.; Freitas, M.A.; Zang, X.; Hatcher, P.G.; Green-Church, K.B. & Jones, R.B. (2002). The application of electrospray ionization mass spectrometry (ESI MS) to the structural characterization of natural organic matter. *Organic Geochemistry*, 33, 3, 171-180.

Kujawinski, E.B. & Behn, M.D. (2006). Automated analysis of electrospray ionization Fourier transform ion cyclotron resonance mass spectra of natural organic matter. *Analytical Chemistry*, 78, 13, 4363-4373.

Kujawinski, E.B.; Longnecker, K.; Blough, N.V.; Del Vecchio, R.; Finlay, L.; Kitner, J.B. & Giovannoni, S.J. (2009). Identification of possible source markers in marine dissolved organic matter using ultrahigh resolution mass spectrometry. *Geochimica et Cosmochimica Acta,* 73, 15, 4384-4399.

Leenheer, J.A.; Rostad, C.E.; Gates, P.M.; Furlong, E.T. & Ferrer, I. (2001). Molecular resolution and fragmentation of fulvic acid by electrospray ionization/multistage tandem mass spectrometry. *Analytical Chemistry*, 73, 7, 1461-1471.

Leenheer, J.A. & Croué, J.P. (2003). Characterizing aquatic dissolved organic matter. *Environmental Science and Technology*, 37, 1, 18A-26A.

Liu, Z.; Sleighter, R.L.; Zhong, J.; & Hatcher, P.G. (2011). The chemical changes of DOM from black waters to coastal marine waters by HPLC combined with ultrahigh resolution mass spectrometry. *Estuarine, Coastal and Shelf Science,* in press.

Marshall, A.G.; Hendrickson, C.L. & Jackson, G.S. (1998). Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews*, 17, 1, 1-35.

McIntyre, C.; Batts, B.D. & Jardine, D.R. (1997). Electrospray mass spectrometry of groundwater organic acids. *Journal of Mass Spectrometry*, 32, 3, 328–330.

McIntyre, C.; McRae, C.; Jardine, D. & Batts, B.D. (2002). Identification of compound classes in soil and peat fulvic acids as observed by electrospray ionization tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 16, 16, 1604-1609.

Mopper, K.; Stubbins, A.; Ritchie, J.D.; Bialk, H.M. & Hatcher, P.G. (2007). Advanced instrumental approaches for characterization of marine dissolved organic matter: extraction techniques, mass spectrometry, and nuclear magnetic resonance spectroscopy. *Chemical Reviews*, 107, 2, 419-442.

Ohno, T.; He, Z.; Sleighter, R.L.; Honeycutt, C. & Hatcher, P.G. (2010). Ultrahigh resolution mass spectrometry and indicator species analysis to identify marker components of soil- and plant biomass- derived organic matter fractions. *Environmental Science and Technology*, 44, 22, 8594-8600.

Perdue, E.M. & Ritchie, J.D. (2003). Dissolved organic matter in freshwaters, In: *In: Treatise on Geochemistry*, *volume 5*, 273-318, Elsevier.

Plancque, G.; Amekraz, B.; Moulin, V.; Toulhoat, P. & Moulin, C. (2001). Molecular structure of fulvic acids by electrospray with quadrupole time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 15, 10, 827-835.

Purcell, J.M.; Hendrickson, C.L.; Rodgers, R.P. & Marshall, A.G. (2007). Atmospheric pressure photoionization proton transfer for complex organic mixtures investigated by Fourier transform ion cyclotron resonance mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 18, 9, 1682-1689.

Raffaelli, A. & Saba, A. (2003). Atmospheric pressure photoionization mass spectrometry. *Mass Spectrometry Reviews*, 22, 5, 318-331.

Reemtsma, T.; These, A.; Linscheid, M.; Leenheer, J. & Spitzy, A. (2008). Molecular and structural characterization of dissolved organic matter from the deep ocean by

FTICR-MS, including hydrophilic nitrogenous organic molecules. *Environmental Science and Technology*, 42, 5, 1430-1437.

Reemtsma, T. (2009). Determination of molecular formulas of natural organic matter molecules by (ultra-) high-resolution mass spectrometry: Status and needs. *Journal of Chromatography A*, 1216, 18, 3687-3701.

Sabine, C.L. & Feely, R.A. (2007). The oceanic sink for carbon dioxide, In: *Greenhouse gas sinks*, D. Reay; C. Hewitt; K. Smith & J. Grace (Eds.), 31-49, CAB International, Oxfordshire.

Schmidt, F.; Elvert, M.; Koch, B.P.; Witt, M. & Hinrichs, K.U. (2009). Molecular characterization of dissolved organic matter in pore water of continental shelf sediments. *Geochimica et Cosmochimica Acta,* 73, 11, 3337-3358.

Schnitzer, M. & Khan, S.U. (1978). *Soil Organic Matter*, Elsevier, Amsterdam.

Simjouw, J.P.; Minor, E.C. & Mopper, K. (2005). Isolation and characterization of estuarine dissolved organic matter: Comparison of ultrafiltration and C18 solid-phase extraction techniques. *Marine Chemistry*, 96, 3-4, 219-235.

Sleighter, R.L. & Hatcher, P.G. (2007). The application of electrospray ionization coupled to ultrahigh resolution mass spectrometry for the molecular characterization of natural organic matter. *Journal of Mass Spectrometry*, 42, 5, 559-574.

Sleighter, R.L. & Hatcher, P.G. (2008). Molecular characterization of dissolved organic matter (DOM) along a river to ocean transect of the lower Chesapeake Bay by ultrahigh resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Marine Chemistry*, 110, 3-4, 140-152.

Sleighter, R.L.; McKee, G.A.; Liu, Z. & Hatcher, P.G. (2008). Naturally present fatty acids as internal calibrants for Fourier transform mass spectra of dissolved organic matter. *Limnology and Oceanography: Methods*, 6, 246-253.

Sleighter, R.L.; McKee, G.A. & Hatcher, P.G. (2009). Direct Fourier transform mass spectral analysis of natural waters with low dissolved organic matter. *Organic Geochemistry*, 40, 1, 119-125.

Sleighter, R.L.; Liu, Z.; Xue, J. & Hatcher, P.G. (2010). Multivariate statistical approaches for the characterization of dissolved organic matter analyzed by ultrahigh resolution mass spectrometry. *Environmental Science and Technology*, 44, 19, 7576-7582.

Stenson, A.C.; Landing, W.M.; Marshall, A.G. & Cooper, W.T. (2002). Ionization and fragmentation of humic substances in electrospray ionization Fourier transform-ion cyclotron resonance mass spectrometry. *Analytical Chemistry*, 74, 17, 4397-4409.

Stenson, A.C.; Marshall, A.G. & Cooper, W.T. (2003). Exact masses and chemical formulas of individual Suwannee River fulvic acids from ultrahigh resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectra. *Analytical Chemistry*, 75, 6, 1275-1284.

Stevenson, F.J. (1994). *Humus Chemistry: Genesis, Composition, Reactions,* John Wiley and Sons, New York.

Stubbins, A.; Spencer, R.G.M.; Chen, H.; Hatcher, P.G.; Mopper, K.; Hernes, P.J.; Mwamba, V.L.; Mangangu, A.M.; Wabakanghanzi, J.N. & Six, J. (2010). Illuminated darkness: Molecular signatures of Congo River dissolved organic matter and its photochemical alteration as revealed by ultrahigh precision mass spectrometry. *Limnology and Oceanography*, 55, 4, 1467-1477.

Thurman, E.M. & Malcolm, R.L. (1981). Preparative isolation of aquatic humic substances. *Environmental Science and Technology*, 15, 4, 463-466.

Thurman, E.M. (1985). *Organic Geochemistry of Natural Waters*, Kluwer Academics, Boston.

van Krevelen, D. (1950). Graphical – statistical method for the study of structure and reaction process of coal. *Fuel*, 29, 269-284.

Witt, M.; Fuchser, J. & Koch, B.P. (2009). Fragmentation studies of fulvic acids using collision induced dissociation Fourier transform ion cyclotron resonance mass spectrometry. *Analytical Chemistry*, 81, 7, 2688-2694.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Rachel L. Sleighter and Patrick G. Hatcher (2011). Fourier Transform Mass Spectrometry for the Molecular Level Characterization of Natural Organic Matter: Instrument Capabilities, Applications, and Limitations, Fourier Transforms - Approach to Scientific Principles, Prof. Goran Nikolic (Ed.), ISBN: 978-953-307-231-9, InTech, Available from: http://www.intechopen.com/books/fourier-transforms-approach-to-scientific-principles/fourier-transform-mass-spectrometry-for-the-molecular-level-characterization-of-natural-organic-matt

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Enhanced Fourier Transforms for X-Ray Scattering Applications

Benjamin Poust and Mark Goorsky
University of California, Los Angeles, Department of Materials Science and Engineering,
Los Angeles, CA 90095
USA

1. Introduction

A new method for enhancing the Fourier transforms of x-ray reflectivity data is presented. This enhanced Fourier analysis, which employs differentiation of the scattered intensity signal, is shown to be extremely effective in extracting layer thicknesses from specular x-ray reflectivity scans from single and multi-layer structures. This is a powerful technique that complements simulations of x-ray scattering patterns that employ dynamical diffraction models. Examples of the procedure, data analysis, and comparison of the results with methods that have been described previously will be presented.

A Fourier Transform (FT) power spectrum peak represents the frequency or period length of an interference oscillation. X-ray scattering measurements provide information in the reciprocal space domain. Therefore, the FT of an x-ray scattering measurement would be expected to provide information concerning layer properties, especially the layer thicknesses which establish the interference fringes in scattering measurements including reflectivity measurements and higher angle diffraction measurements. Indeed, the intensity modulations that are observed in specular x-ray reflectivity measurements are related to the layer thicknesses and to the difference in refractive index between one layer and the next. At x-ray wavelengths, the refractive index is determined by the material density. Discrete Fourier transforms and their application to x-ray reflectivity data will be discussed subsequently in terms of the mathematics, challenges inherent to x-ray scatter FTs, and enhancement techniques that have already been discussed in the literature.

The key to the enhancement method described here is based around **differentiating** the specular intensity with respect to vertical reciprocal space coordinate Q_Z. This differentiation retains the important and useful components of the x-ray reflectivity measurements while minimizing the impact of features of the measurements that obscure the transformation of the interference pattern. The reflectivity data is transformed according to

$$I_j^T = \frac{dI_j}{dQ_{Z,j}} \approx \frac{1}{N}\sum_{i=1}^{N} \frac{I(Q_{Z,j+i}) - I(Q_{Z,j-i})}{Q_{Z,j+i} - Q_{Z,j-i}} \tag{1}$$

The summation on the right side of the equation is over the N nearest data points on either side of the jth data point. The number of neighboring data points used to calculate the average slope at the *j*th data point is set just high enough to average out noise fluctuations, but kept well below the period lengths of any possible thickness signals. In general, this differentiation approach is far more effective at removing the sloping background than logarithmic compression alone, average subtraction alone, or Q_z^4 leveling (Durand's method (Durand 2004)) methods in the literature which have been previously employed to enhance FT of the reflectivity data and these methods are described in more detail below. When combined with any of the other enhancement techniques, however, differentiation yields readily distinguishable FT peaks for even the weakest and most truncated of sloping oscillations. It is not proposed here that differentiation should replace the other enhancement techniques, but rather that it should be used with them to achieve the best possible FT enhancement. The background into the development of this approach is presented below with illustrations and comparison with the other techniques.

2. Background

Electromagnetic radiation that travels through one medium and passes into another will be partially reflected at that interface if the media have different indices of refraction. Radiation reflecting from interfaces leads to interference in the scattered wave and it is this effect that makes x-ray scatter based film thickness measurements useful. X-ray reflectivity can be especially amenable to extraction of layer thicknesses as the interference pattern includes only information on changes in refractive index (which is the electron density at x-ray wavelengths) as a function of depth. The crystallinity, strain state, or other such crystallographic factors do not play roles in determining the reflectivity curve for both substrates and multi-layer structures deposited on the substrates. In fact, specular x-ray reflectivity measurements have proven to be extremely valuable for determining the properties of multi-layer structures. In a typical case, the fringes that are introduced from a single layer are used to determine the layer thickness; for more complicated multi-layer structures, simulation programs are used to help extract layer thicknesses.

To demonstrate the information gained from a specular x-ray reflectivity scan, consider that the reflectivity scan is simply a specular scan from the origin (000) of reciprocal space along Q_z, i.e., perpendicular to the surface. This is depicted in Figure 1 which shows a two dimensional section of reciprocal space that includes the co-planar scattering conditions; the so-called Ewald construction. An 'off-specular' scan, which will be described later as an important scan used in the literature to help extract thickness information via FT, is also included in the figure.

The axes are Q_x, which is along the surface, and Q_z, which is perpendicular to the surface. K_o represents the incident wavevector (with a magnitude 1/λ where λ is the radiation wavelength) and K_H represents the scattered wavevector (also 1/λ). For specular reflectivity scans, the angle between the incident beam and the surface is ω (= Θ) and the angle between the incident and scattered beams is 2Θ. Changing ω (=Θ) by a small increment and the angle between the incident and scattered beams by twice that increment traces a vertical line in reciprocal space. The relationship between Q_z and ω and 2Θ is simply

$$Q_z = \frac{\left(\sin\omega - \sin\left(\omega - 2\Theta\right)\right)}{\lambda} \tag{2}$$

and, for the specific case for specular reflectivity where ω = Θ,

$$Q_z = \frac{2(\sin\Theta)}{\lambda} \tag{3}$$

It should be noted that, at x-ray wavelengths, the refractive index is important. Including this effect leads to

$$Q_z = \frac{2}{\lambda}\sqrt{\cos^2\Theta_c - \cos^2\Theta} \tag{4}$$

where $\cos(\Theta_c)$ = the refractive index and Θ_c represents the critical angle below which the incident x-ray beam is totally reflected at the surface. (Durand 2004)

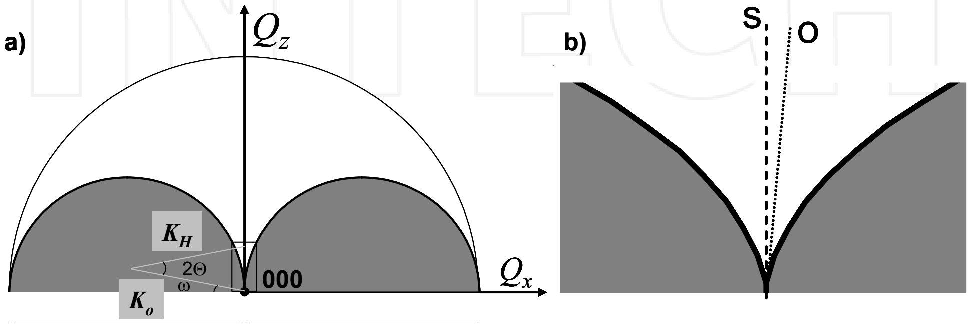


Fig. 1. Reciprocal Space and the Ewald Construction. (a) The two gray half circles represent regions for which scattering involves the transmission mode. The region near the origin of reciprocal space (000) is expanded in (b). The vertical dashed line (labeled '**S**') represents a specular reflectivity scan. This is achieved by rotating the sample by an angle ω = Θ with respect to the incident beam and the angle between the incident and scattered beam by 2Θ. The dotted line (labeled '**O**') represents an off-specular scan in which the incident beam angle with the surface ω ≠ Θ but both the angle ω and 2Θ are moved by increments of Θ and 2Θ, respectively

The development of x-ray reflectivity expanded significantly with the work of L.G. Parratt (Parratt 1954), who introduced a theory that related the layer thicknesses and electron densities to the reflectivity curve. Following the approach of von Laue, Parratt used Maxwell's equations and appropriate boundary conditions to solve for the reflected to transmitted amplitude ratio at the bottom of layer *j*, X_j:

$$X_j = \frac{r_j + X_{j+1}\varphi_{j+1}^2}{1 + r_j X_{j+1}\varphi_{j+1}^2} \tag{5}$$

where layer *j* + 1 is below layer *j*. The phase offset of the wave scattered from the bottom of and traversing the thickness of layer *j* + 1, φ_{j+1}, is:

$$\varphi_{j+1} = \exp(ik_{z,j+1}t_{j+1}) \tag{6}$$

The component of the wave vector perpendicular to the surface in layer *j*, $k_{z,j}$, is

$$k_{z,j} = \frac{1}{\lambda\sqrt{n_j^2 - \cos^2\Theta}} \tag{7}$$

where ω (= Θ) is the incidence angle, as noted above, and n_j is the index of refraction of layer j at wavelength λ.

The Fresnel coefficient of reflection from the interface between layers j and $j + 1$, r_j, is given by

$$r_j = \frac{k_{z,j} - k_{z,j+1}}{k_{z,j} + k_{z,j+1}} \tag{8}$$

for a sharp interface.

Overall, the x-ray reflectivity $R(Q_z)$ for a layer on a substrate can be described as

$$R(q_z) = \frac{I(Q_z)}{I_o} = R_F(Q_z)\left|\frac{1}{\rho_\infty}\int\frac{d\rho(z)}{dz}\exp(iQ_z z)\,dz\right|^2 \tag{9}$$

where I_o is the incident intensity, $I(Q_z)$ is the measured reflectivity intensity, and $R_F(Q_z)$ is the Fresnel reflectivity (which is effectively the Fresnel coefficient [8] squared). Equation [9] demonstrates that the reflectivity intensity as a function of Q_z (or Θ) depends only upon the density change at the interfaces and surface.

This relationship later modified the Fresnel coefficient to account for rough or graded interfaces:

$$r_j = \frac{k_{z,j} - k_{z,j+1}}{k_{z,j} + k_{z,j+1}}\exp\left[-2\sigma_{j+1}\sqrt{k_{z,j}k_{z,j+1}}\right] \tag{10}$$

where σ_{j+1} is the effective width associated with interface roughness and/or compositional grading. Further development of the theory of x-ray reflectivity scattering is addressed elsewhere. (Bowen and Tanner 1998; Daillant and Gibaud 1999; Wormington, Panaccione et al. 1999)

Examples of reflectivity curves based on this formalism are shown below to illustrate the application and challenges associated with using Fourier Transforms of x-ray reflectivity data to extract layer thickness information. Figure 2 shows reflectivity from a silicon surface, a silicon surface with a layer of SiO_2, a silicon surface with a layer of germanium, and a silicon surface with an r.m.s. roughness of 5 Å.

For such specular reflectivity scans, it is customary to plot the intensity on a logarithmic scale against the angle ω (which, for a specular scan is equal to Θ) with the understanding that the angle between the incident beam and the detector is also changing at twice the angle, hence 2Θ. In some cases, the x-axis will indicate this, e.g., with a label 'ω - 2Θ' or 'Θ - 2Θ'. In other cases, the reflectivity scan will be plotted as a function of Q_z (transformed using equation [2] or [4]).

Starting with the simulated reflectivity curve from a smooth silicon surface, there are two notable characteristics of the specular reflectivity scans that are important for the subsequent Fourier transform:

i. the reflected intensity is unity when the incident angle is below a certain value – the critical angle for total external reflection which is identified as $Q_{z,c}$ or Θ_c as noted above.

ii. the reflected intensity drops off strongly for $Q \geq Q_{z,c}$ ($\omega \geq \Theta_c$) with a with a Q_z^4 dependence at higher angles.

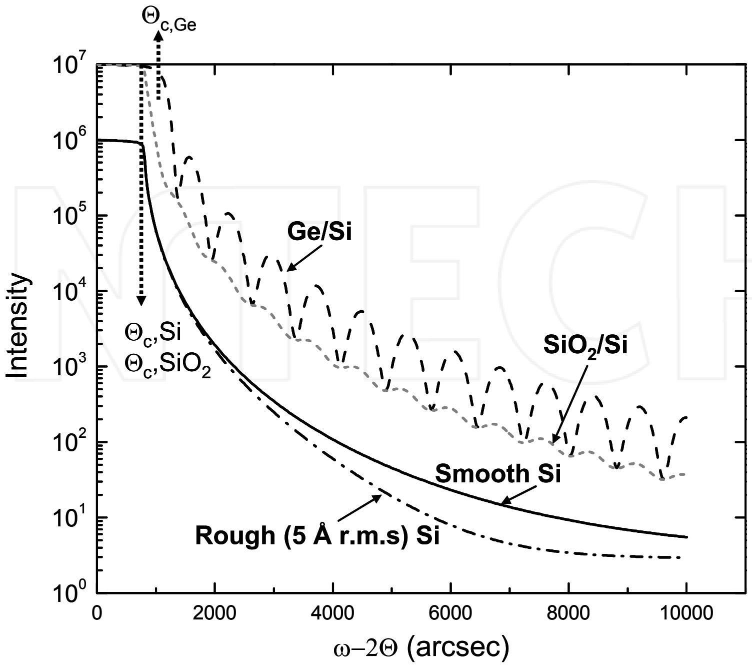


Fig. 2. Simulated X-ray reflectivity scans for a bare surface (Si) comparing \smooth and rough surfaces as well as the scans for 200 Å layers (either SiO$_2$ or Ge) on smooth Si in which the thickness fringes are clearly visible. The scans from the structures with the SiO$_2$ and Ge layers are vertically offset (10X) for clarity

For the case of a layer deposited on the silicon surface, a series of fringes is observed. The fringe spacing provides information on the layer thickness and represents interference associated with the difference in electron density between the layer and the substrate. This effect is clearly depicted for the examples of a 200 Å germanium layer on the silicon when compared to a 200 Å layer of SiO$_2$ layer on silicon. The difference in electron density between silicon and germanium is significantly greater than that between SiO$_2$ and silicon. The fringe spacing remains the same, whereas the fringe amplitude is correspondingly greater for the combination with the greater electron density difference, in this case, for the germanium on silicon. In addition, the critical angle, Θ_c for the Ge layer on the surface is greater than that for the SiO$_2$ layer on the surface. Silicon and SiO$_2$ have very similar densities, so the critical angles for those surfaces are also nearly the same value, as shown in the figure.

For the silicon surface with the rougher interface, the intensity drops off at a different Q dependence (and this decay depends on the extent of roughness). Figure 3 plots the silicon reflectivity multiplied by Q_z^4 (ω^4 for these scans to compare directly to the data in Figure 2) for the specular scans from both the smooth and the rough surfaces. In these case, it is clear that a horizontal line is generated for the smooth surface for $\omega > 3\Theta_c$ (i.e., the decay exhibits a Q_z^{-4} dependence) and a decreasing slope is observed for the rougher surface.

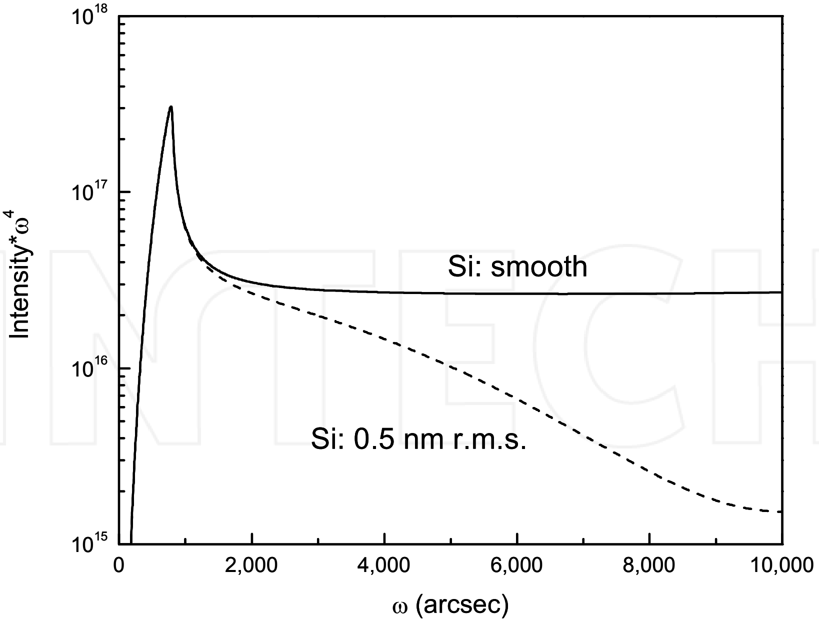


Fig. 3. The intensity at each position is multiplied by the incident angle to the fourth power. For a smooth surface, this product is a constant value at higher angles. For a rough surface, the product decreases at higher angles, but not as strongly as the intensity alone decreases

Based on the x-ray reflectivity curves and as is deduced from Equation [9], x-ray reflectivity scans along Q_z measure the Fourier transform of the derivative of the electron density with respect to z. Indeed, Equation [9] can be rewritten as (Russel 1990) (Li, Muller et al. 1996) (Daillant and Gibaud 1999) (Durand 2004)

$$I(Q_Z) \approx \frac{1}{Q_Z^4}\left[F\left(\frac{d\rho(z)}{dz}\right)\right]^2 \approx \frac{1}{Q_Z^4}P\left(\frac{d\rho(z)}{dz}\right) \tag{10}$$

where, in this case, $F(\mathrm{d}\rho(z)/\mathrm{d}z)$ is the Fourier Transform amplitude, and $P(\mathrm{d}\rho(z)/\mathrm{d}z)$ is the Fourier transform power of the derivative electron density at a depth z from the surface. This formulation of the equation more clearly brings out the general $Q_z{}^4$ dependence and also suggests that the extraction of the density distribution as a function of depth using Fourier Transforms should be straightforward.

In other words, the Fourier transform of a specular reflectivity measurement should yield the autocorrelation function of the derivative of the electron density. Thus, x-ray reflectivity Fourier transforms are expected to produce peaks corresponding to distances between interfaces, which can be the thickness of individual layers or the sum of the thicknesses of multiple layers, etc. While the order of the interface stack sequence and information about roughness are not extracted directly from a Fourier Transform of the data, the FT extraction is nonetheless anticipated to be powerful techniques for automatically extracting layer thickness from specular reflectivity measurements.

3. Experimental background

A Bede Scientific, Inc. model D1 high resolution x-ray diffractometer was used for all x-ray measurements in this work. The diffractometer is comprised of three primary components, often referred to as 'axes.' The first 'axis' is the beam conditioning component of the diffractometer, the second is the sample alignment and the third includes the scattered beam conditioning.

The first axis collimates and monochromates the x-ray radiation produced using a copper anode ($\lambda_{CuK\alpha1}$ = 1.540562 Å) vacuum tube x-ray. A Maxflux™ specular mirror is used to redirect divergent x-rays toward the collimator crystal in a parallel path, resulting in an approximately tenfold increase in usable x-rays. In the reflectivity measurements discussed here, the beam is diffracted twice by one channel-cut 220 Si collimator crystal and then passed through a slit before it is incident upon the sample crystal. Together, the specular mirror, collimator crystal, and monochromator slit comprise the first axis.

The second axis supports and manipulates the sample crystal, or specimen. The specimen can also be rotated along orthogonal axes in the plane of the sample surface: χ and ω. The χ axis of rotation allows for adjustment of sample tilt and is located in the dispersion plane of the diffractometer. ω is orthogonal to χ and describes the angle of incidence between the x-ray beam and the sample surface, as noted above.

The third axis assembly is placed on a cantilever that revolves around the 2Θ axis of rotation, which coincides with the ω axis of rotation. A scintillating x-ray detector tuned for optimum response to the copper K□ line, a dual-channel analyzer crystal (DCA) ((220) silicon reflections), and a pair of x-ray acceptance limiting detector slits constitutes the scattered beam conditioning axis. The configuration of the third axis determines the angular resolution of the instrument. Angular resolution is determined by the width of the detector slits used in double axis measurements which are employed here. In triple axis measurements, the DCA essentially serves as an extremely narrow detector slit and determines the angular resolution of the measurement.

A commercial simulation program (Bede REFS) employs a distorted wave Born approximation to the dynamical scattering conditions. Following the general approach described by Parratt, (Parratt 1954) (Wormington, Panaccione et al. 1999), this program can be used to generate scans for illustrative purposes as well as to help understand the Fourier transforms from multi-layer structures. The examples that are illustrated here employ two different sets of samples. The first includes a thin AlN layer deposited on a sapphire substrate. The second is a multi-layer structure based on an AlSb / InAs structure deposited on a GaAs substrate. The techniques and procedures described here, however, are not material dependent and the structures should be considered as illustrating the general principles of the technique.

4. Prior studies: Fourier Transform of specular x-ray reflectivity

The first published use of FT with x-ray scattering data is that of Sakurai and Iida. (Sakurai and Iida 1992) As it happens, a few properties of x-ray reflectivity data that are shown in Figures 2 and 3 severely limit the effectiveness of Fourier transforms of the raw data. Consider that thickness fringes in reflectivity measurements typically follow an intensity distribution illustrated by the curve of the Ge (or SiO_2) layer on silicon (as in Figure 2); this presents a challenge in that the oscillations appear on a sloping background that ranges over several orders of magnitude. In the case of a specular x-ray reflectivity scan, the intense peak adjacent to the oscillations is the intense highly reflected component near the critical angle, Θ_c. The part of the curve containing the oscillations represents a short non-background interval. Any short non-background interval is essentially a signal consisting of a single pulse. A non-square pulse yields a Fourier transform with a truncation peak centered at zero frequency similar to the squared sync function associated with a square pulse FT. When the oscillations are weak or the number of oscillations measured is small, the truncation peak can obscure the oscillation peaks. This effect is illustrated in Figure 4. Here, a simple sinusoidal oscillation is compared to the same sinusoidal oscillation with the addition of an intense peak at the left that corresponds to the overall decay of the reflectivity signal.

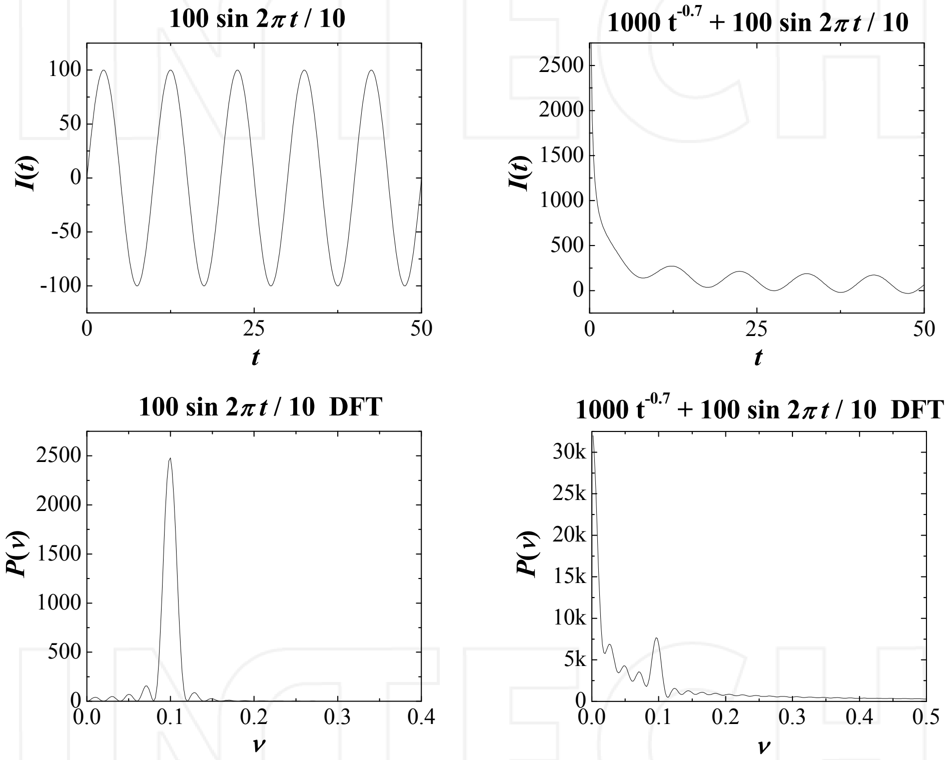


Fig. 4. A sinusoidal oscillation (left, upper) and FT (left, lower), the same oscillation added to half of an adjacent intense peak (right, upper) with the FT showing the peak due to the sinusoidal oscillation partially buried by the truncation FT peak

The FT that includes the high intensity signal at low angles superimposes a truncation FT that significantly distorts the peak which originates from the sinusoidal function. Therefore, the FT provides only vague information about the sinusoidal function peak. This effect is often so severe for experimental data that it renders Fourier transforms on raw data totally useless. For example, see Figure 5. The specular x-ray reflectivity data is from a 320 Å AlN layer deposited on a sapphire substrate and the expected AlN thickness fringe pattern is clearly exhibited. The transform shown on the right, however, only exhibits a diffuse feature where a sharp peak associated with the 320 Å AlN film would be expected. Clearly, the FT of this data does not provide practical information about the layer thickness and the culprit is the high intensity at low angles that introduces the background slope.

In large part, these types of constraints have limited the popularity of FT analysis with x-ray scatter data. However, a few approaches for enhancing specular x-ray reflectivity FTs exist in the literature. They are based around (i) flattening the overall shape of the x-ray scatter data prior to the FT and (ii) maintaining as much of the interference pattern as possible without artificially modifying the data.

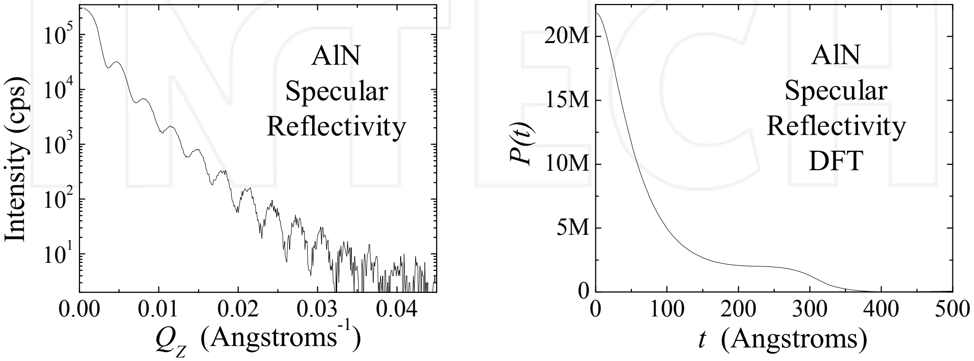


Fig. 5. (Left) X-ray reflectivity scan from a 320 Å AlN film on a sapphire substrate and (right) Discrete Fourier Transform

For example, Banerjee et al., (Banerjee, Raghavan et al. 1999) attempted to flatten the scatter data by subtracting an off-specular scan from the specular reflectivity scan, as is depicted in Figure 1(b). The second technique proposed by Grave de Peralta and Temkin (Peralta and Temkin 2003) involves logarithmic compression of the intensity and subtraction of a heavily smoothed version of the x-ray reflectivity curve. The third technique proposed by Durand (Durand 2004) involved multiplication of the intensity by the reciprocal space coordinate value, Q_z^4 as was already shown in Figure 3. The analysis of each of these concepts is included below.

Because off-specular scans diverge from the interference direction, they often do not show interference fringes. They do, however, have a sloping background shape similar to that of a specular scan. The idea of leveling the specular reflectivity scan for FT enhancement was first demonstrated through subtraction of an off-specular scan for just this reason. (Banerjee, Raghavan et al. 1999) However, the effectiveness of off-specular scan subtraction depends heavily upon the sample measured. A film that is uniform in thickness but conforms to a rough substrate, for example, may have the interference fringes broadened horizontally in reciprocal space. This is because areas of different surface height may scatter incoherently with respect to one another. Such broadened interference fringes are typically referred to as Bragg sheets. They are illustrated in Figure 6.

This effect is demonstrated in Figure 7 which shows the specular scan from the 200 Å Ge layer on Si but with 5 Å roughness at both the Si-Ge interface and the Ge surface. In the latter case, subtraction of the fringes in the off-specular (otherwise known as a longitudinal scan) case will introduce an additional, artificial modulation.

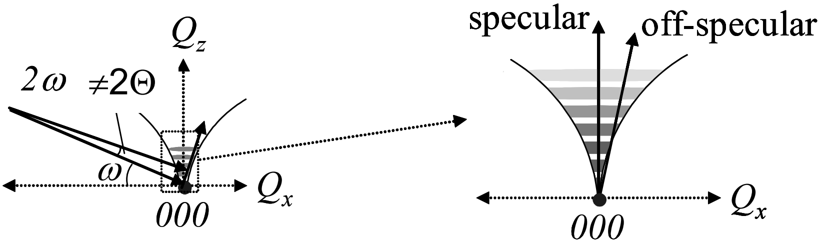


Fig. 6. Specular and off-specular scans in the case where interference fringes are horizontally broadened into Bragg sheets

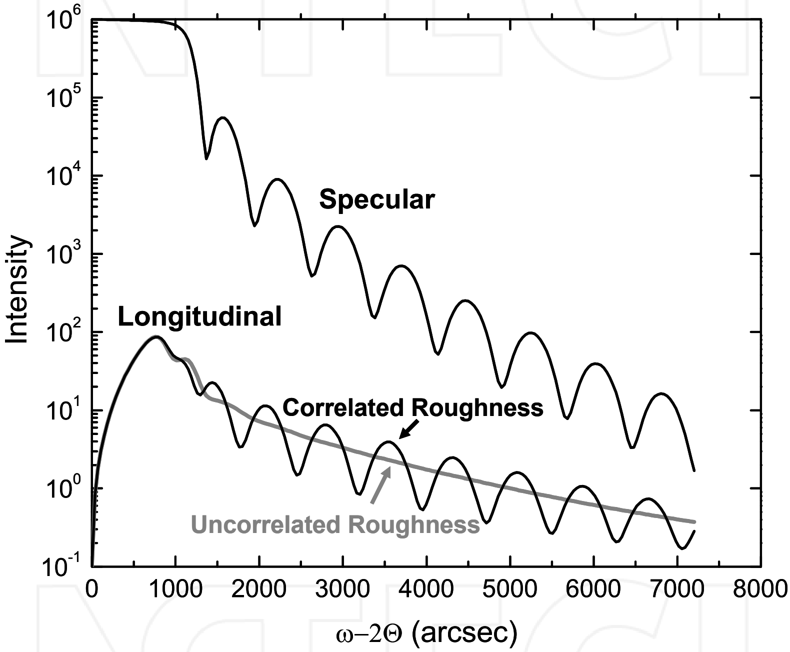


Fig. 7. X-ray reflectivity scans from a 200 Å Ge layer on Si with 5 Å roughness at the interface and 5 Å roughness at the Ge surface. The top scan is the specular scan; the lower scans represent the longitudinal scans for the cases of uncorrelated and correlated interfaces

Still, the use of the off-specular subtraction introduced the idea of leveling the scan by some type of subtraction to enhance the FTs and thus paved the way for more effective techniques.

One improvement is logarithmic compression of the experimental data followed by subtraction of a semi-local average intensity.(Peralta and Temkin 2003) Here logarithmic compression means transforming the intensity of the jth datapoint, I_j, according to:

$$I_j^T = \log I_j \tag{11}$$

Equation [11] transforms the data such that its shape becomes numerically what it appears to be when plotted on a logarithmic intensity scale. When logarithmic compression is first applied, and the semi-local average is then calculated and subtracted, a dramatic reduction of the pulse transform peak is observed upon the Fourier Transform and the oscillation peak is clearly observed, as shown in Figure 8 for the same AlN on sapphire sample discussed earlier. The combination of logarithmic compression and average subtraction is effective because logarithmic compression aids in calculation of a more useful average curve. In this work, the local average is calculated at the j^{th} data point as the average over the N nearest data points in each direction. The average at each point is referred to here as a 2N point local average, where 2N is the number of neighboring points contributing to the average. Within N data points of each edge of the scan, the number of points used to calculate the average must be reduced. For example, at the first data point (e.g. j = 1) the average can only be calculated over data points 1 to 1 + N. Logarithmic compression prior to calculation of the average reduces this effect through simple compression of the intensity dynamic range.

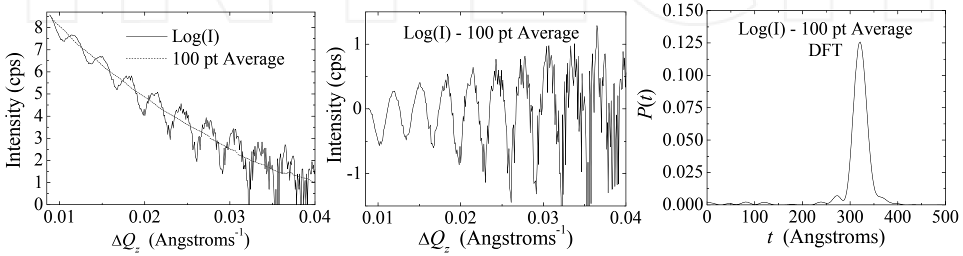


Fig. 8. (Left) logarithmically compressed specular scan showing 100 point local average, (center) specular scatter intensity following subtraction of 100 point local average, and (right) FT of the center curve

Clearly logarithmic compression followed by subtraction of a semi-local average can be very useful. When extremely thin films are measured, however, the approach loses its effectiveness because the wavelength of the oscillation begins to approach the total length of the scan. To prevent the average from itself including oscillations, the term N must at least be larger than a period length. When that period length is long, N must be so large that the average curve no longer matches the sloping background shape. This particular effect is demonstrated in Figure 9 using part of the AlN on sapphire scan. Under these conditions, two extra peaks are artificially introduced into the FT and two thinner layers would appear to be present in the original data.

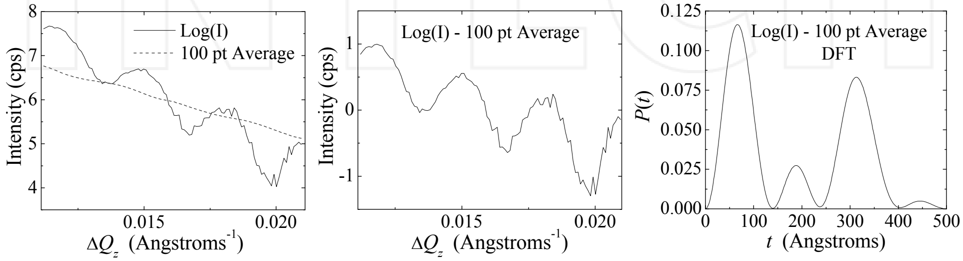


Fig. 9. Log compression and average subtraction with FT for a sample with few fringes

Additional challenges arise when intense peaks exist in the specular scan. This occurs, for example, in the case of a superlattice structure. The peak elevates the average curve above the background slope and thus limits the effectiveness of average subtraction in leveling the data. Improper leveling leads to pulse transform artifacts similar to those observed in Figure 9. Multiplication of the intensity by Q_Z^4 (or ω^4) (Durand 2004) should thus be an effective method for removing the sloping background of a reflectivity scan, i.e., leveling the curve, as depicted in Figure 3. The FT of the leveled curve is calculated and peaks are taken to represent layer thicknesses and sums of layer thicknesses. The effectiveness of this approach is demonstrated using the AlN on sapphire reflectivity data and is shown below in Figure 10.

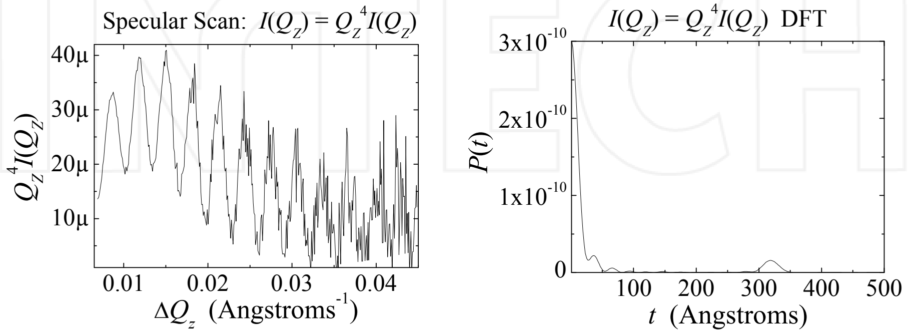


Fig. 10. Multiplication of intensity by Q_z^4 at each data point and the FT showing a distinct peak corresponding to the 320 Å thickness of the layer

As seen in Figure 10, this approach is clearly less effective at leveling the specular scan than are the averaging techniques and thus yields a stronger pulse transform peak. Challenges arise when roughness or graded interfaces cause the specular scatter intensity to drop-off faster with increasing Q_Z (for example, see Figure 2) than the assumed fourth order dependence so the resulting curve is not necessarily flat. The main advantage to Durand's approach, however, lies in the fact that it is independent of oscillation period length. Recall that when subtracting a semi-local average, the optimum local average size must be selected very carefully. Furthermore, when the oscillation period length approaches the length of the measurement, the average subtraction technique becomes ineffective. These disadvantages do not apply to Durand's method. Each of the above techniques has merits, but each suffers from some aspect of the nature of the x-ray reflectivity curve.

5. Presentation of a new enhancement approach

5.1 Differentiation and application to a single layer structure

An approach to better flatten the reflectivity data stems from considering that the x-ray reflectivity data effectively consists of a well-behaved fringe pattern (sine curves or the combination of several sine curves) with the addition of a sloped background (i.e., the Q_z^4 decay with Q_z). These extrinsic influences may be better separated using differentiation than the techniques described in the literature and this concept forms the basis for the approach described here. For example, the differentiation of the fringe pattern will produce a curve with an identical periodicity as the original data while the differentiation of the sloping component will provide a much reduced contribution to the entire curve. The data is transformed according to

$$I_j^T = \frac{dI_j}{dQ_{Z,j}} \approx \frac{1}{N}\sum_{i=1}^{N}\frac{I(Q_{Z,j+i}) - I(Q_{Z,j-i})}{Q_{Z,j+i} - Q_{Z,j-i}} \tag{12}$$

The summation on the right side of equation [12] is over the *N* nearest data points on either side of the *j*th data point. The number of neighboring data points used to calculate the average slope at the *j*th data point is set just high enough to average out noise fluctuations, but kept well below the period lengths of any possible thickness signals.

Differentiation alone is extremely effective at leveling the data for FT enhancement. As a comparison to logarithmic compression, semi-local average subtraction, and multiplication of the intensity by Q_Z^4, the same specular reflectivity scan of an approximately 320 Å AlN film deposited on a sapphire substrate is shown following the differentiation transformation according to Equation [12] with the FT in Figure 11. The surface truncation peak is significantly reduced and does not overlap with the FT peak due to the layer. Also, the layer peak is relatively sharp and provides the correct layer thickness of 320 Å. Thus, for this example, differentiation is clearly one of the most effective single enhancement techniques. Multiple differentiation processes improve the result further.

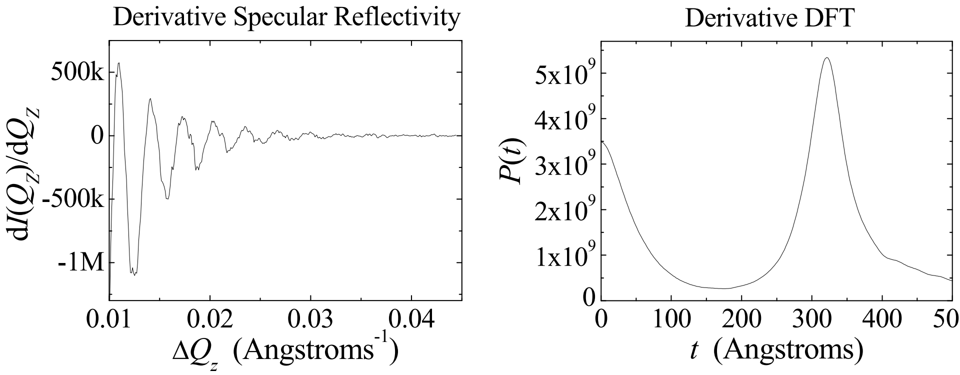


Fig. 11. 320 Å AlN film on sapphire derivative specular reflectivity scan and the FT

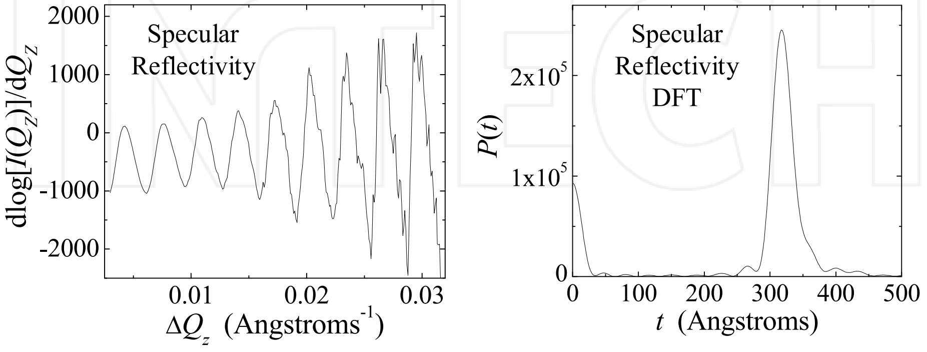


Fig. 12. Specular reflectivity scan with logarithmic compression followed by differentiation (left) and the FT (right)

In general, differentiation is far more effective at removing the sloping background than logarithmic compression alone, average subtraction alone, or the Q_z^4 leveling methods. However, these techniques can be combined to produce even better results. For example, when combined with any of the other enhancement techniques, differentiation yields distinguishable FT peaks for even the weakest and most truncated of sloping oscillations. Therefore, it is not proposed here that differentiation should replace the other enhancement techniques, but rather that it should be used synergistically with one or more of them to achieve the best possible FT enhancement. To illustrate this point, as shown in Figure 12, logarithmic compression followed by differentiation is a very effective transform for FT enhancement of specular reflectivity scans. The peak near 320 Å corresponding to the thickness of the AlN film is clearly observed in the FT and the layer FT peak is sharper than with either technique alone.

Multiplication by Q_Z^4 followed by differentiation is also an extremely effective enhancement for specular reflectivity FTs. Using data from the same AlN thin film discussed in the examples above, the FT result is improved through multiplication of the intensity by Q_Z^4 followed by differentiation. The curve and FT are shown below in Figure 13.

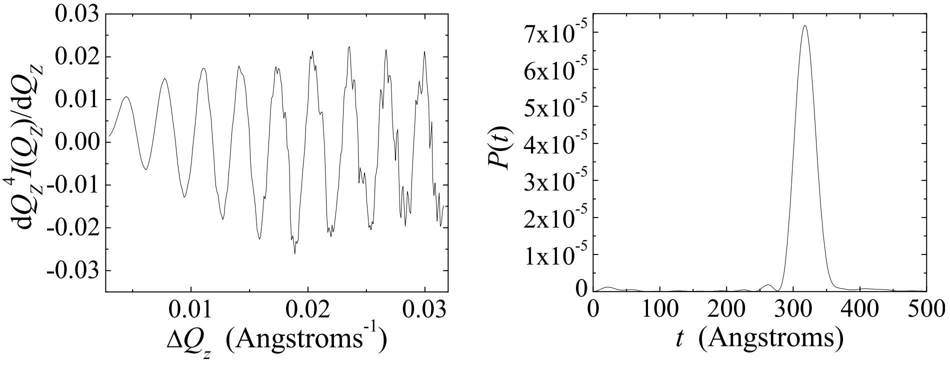


Fig. 13. 320 Å AlN specular x-ray reflectivity curve leveled using Q_z^4 leveling followed by differentiation (left) with the FT (right)

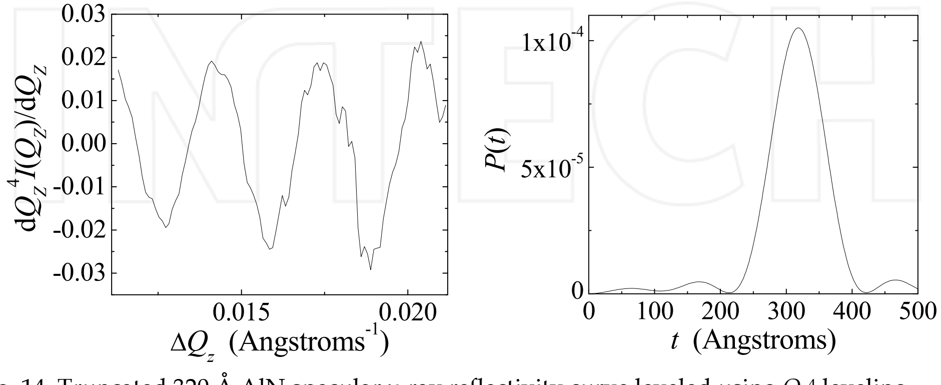


Fig. 14. Truncated 320 Å AlN specular x-ray reflectivity curve leveled using Q_z^4 leveling followed by differentiation (left) with the FT (right)

For comparison to logarithmic compression followed by local average subtraction, the same truncated segment of the AlN reflectivity scan (compare to Figure 9) is processed using Q_z^4 leveling followed by differentiation in Figure 14. Despite the severe wave truncation, the enhanced FT still shows an easily distinguishable single layer thickness peak and greatly reduced artifacts when compared to the results in Figure 9 in which the truncation artificially introduces additional peaks of comparable intensity to the layer FT peak.

5.2 Determination of layer thicknesses in multi-layer structures

Multiple layer structures present a more substantial challenge to the FT approaches in general, so the effectiveness of including the differentiation method in the transformation should be assessed in terms of analyzing such a multi-layer structure. The strategy for including the differentiation method is illustrated using a multiple layer (>5) InAs / AlSb heterostructure grown on GaAs. These structures are used for high speed electronic devices. Reflectivity data from such a multi-layer structure is presented next to assess how well the differentiation process combined – in this case – with the Q_z^4 multiplication can extract important layer thickness information.

As noted above, Fourier analysis affords a reasonably straightforward method for determining layer thicknesses, though initially it may be difficult to interpret which layer thicknesses are represented by which FT peaks in these more complex, multi-layer structures. One method for identifying FT peaks is to Fourier transform simulated x-ray reflectivity scans of the structure adding one layer at a time. By tracking the changes to FT peaks with the addition of layers, it is possible to identify the relationship between the FT peaks and the particular layer thicknesses. This process was carried out for the AlSb/InAs structure that is illustrated in Figure 15.

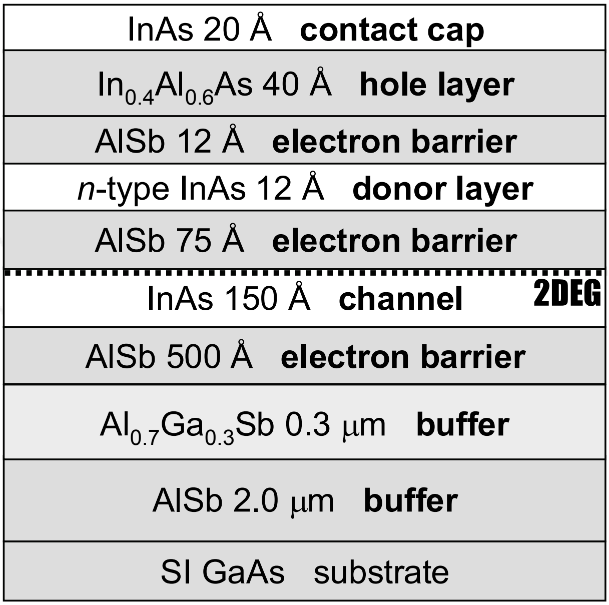

InAs 20 Å **contact cap**
$In_{0.4}Al_{0.6}As$ 40 Å **hole layer**
AlSb 12 Å **electron barrier**
n-type InAs 12 Å **donor layer**
AlSb 75 Å **electron barrier**
InAs 150 Å **channel**
2DEG
AlSb 500 Å **electron barrier**
$Al_{0.7}Ga_{0.3}Sb$ 0.3 μm **buffer**
AlSb 2.0 μm **buffer**
SI GaAs substrate

Fig. 15. Schematic of InAs / AlSb multi-layer structure

Starting with the lower AlSb electron barrier, this process is shown in Figure 16 through Figure 22.

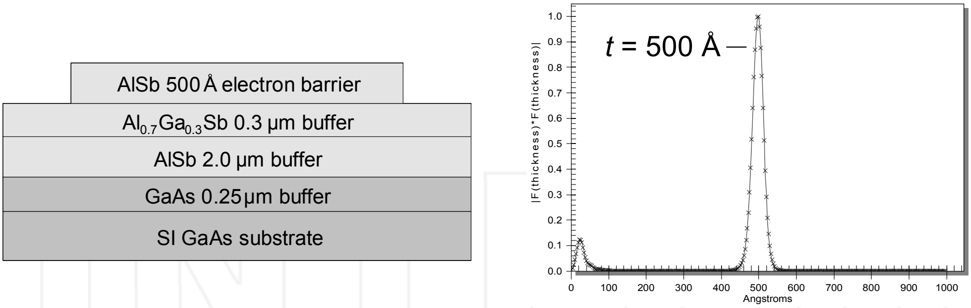


Fig. 16. Simulated reflectivity scan FT of the structure up to the lower AlSb electron barrier

Figure 16 shows the Fourier transform of the reflectivity data; peaks for the buffer layers are not present (beyond the scale of the graph in Figure 16) which indicates that the buffer layers are too thick to yield measurable thickness fringes in the reflectivity scan and therefore do not yield FT peaks. In effect, the fringe spacing for these thick layers is very small (~ 8 arcsec for a 2.0 μm thick AlSb layer on GaAs using Cu kα_1 radiation) and is readily washed out with even small magnitude beam divergence or curvature that would be typical for modern x-ray sources, equipment, and substrates. The lower AlSb electron barrier, on the other hand, yields a FT peak corresponding to its thickness (500 Å). As shown in Figure 17, addition of the InAs electron channel to the simulation yields a new peak at the channel thickness (150 Å), and shifts the original lower AlSb barrier FT peak by approximately the channel thickness. The peak labeled t_2 in Figure 17 thus represents the sum of those two layer thicknesses.

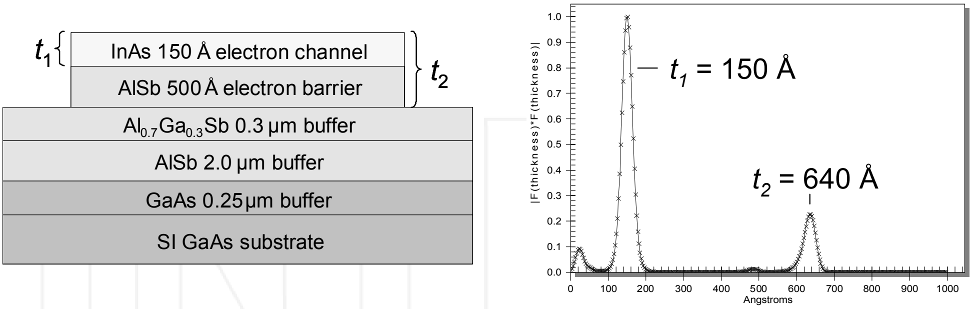


Fig. 17. Simulated reflectivity scan FT of the structure up to the InAs channel

Next, as shown in Figure 18, adding the 75 Å AlSb electron barrier and doped 12 Å InAs layer to the simulation introduces a new FT peak (t_1) representing the sum of the two new layers. The InAs channel peak, which is now t_2 in Figure 18, is still present but is very weak. Peak t_3 represents the sum of the InAs channel, middle AlSb barrier, and doped InAs layer thicknesses. FT peak t_4 in Figure 18 represents the sum of the four layer thicknesses above the buffer layers.

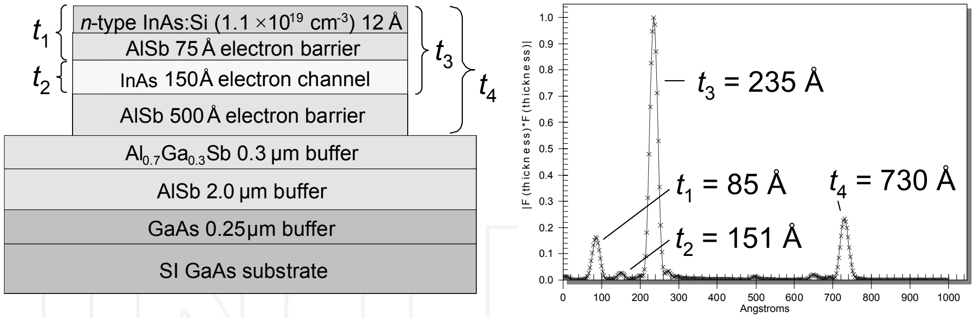


Fig. 18. FT of simulated reflectivity scan of the structure up to the middle AlSb electron barrier

Adding the upper 12 Å AlSb electron barrier, we see in Figure 19 a shift of peaks t_1, t_3, and t_4, by approximately the thickness of the new layer.

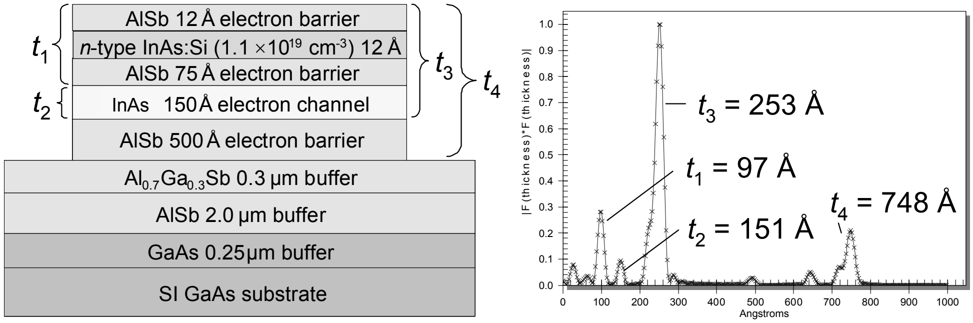


Fig. 19. FT from simulated reflectivity scan of the structure up to the top AlSb electron barrier

Carrying this process through to the top of the structure in Figure 20 and 21, we arrive at the relationships between the FT peaks and layer thicknesses summarized in Figure 22.

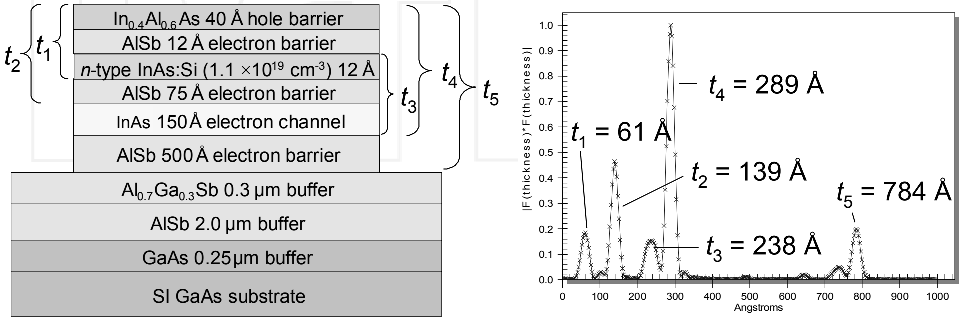


Fig. 20. FT of simulated reflectivity scan of the multi-layer structure up to the InAlAs hole barrier

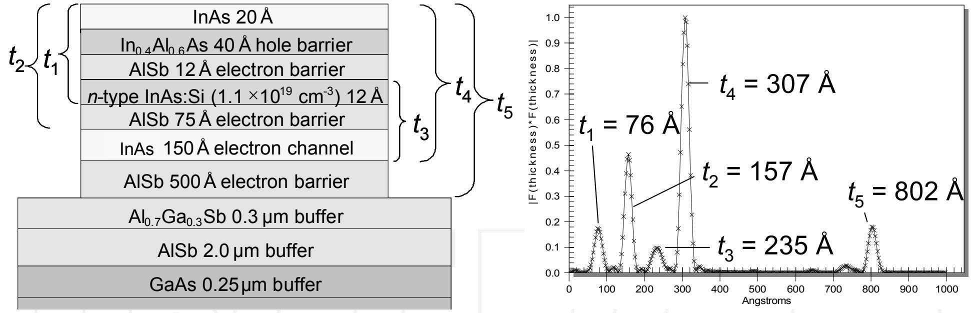


Fig. 21. FT of simulated reflectivity scan of the structure up to the InAs cap layer

$$t_{\text{channel}} = t_4 - t_2$$
$$t_{\text{lower barrier}} = t_5 - t_4$$
$$t_{\text{upper barrier}} = t_2 - t_1$$
$$t_{\text{cap}} + t_{\text{hole barrier}} + t_{\text{thin e- barrier}} = t_4 - t_3$$
$$t_{\text{donor layer}} = t_1 - (t_4 - t_3)$$

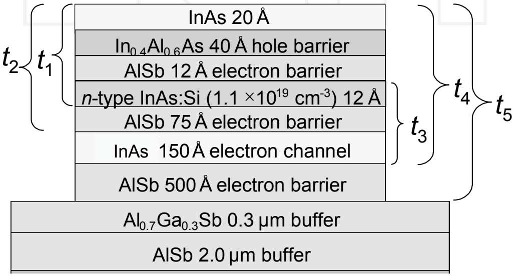

InAs 20 Å
In0.4Al0.6As 40 Å hole barrier
AlSb 12 Å electron barrier
n-type InAs:Si (1.1 ×10^19 cm^-3) 12 Å
AlSb 75 Å electron barrier
InAs 150 Å electron channel
AlSb 500 Å electron barrier
Al0.7Ga0.3Sb 0.3 μm buffer
AlSb 2.0 μm buffer
t2
t1
t3
t4
t5

Fig. 22. Relationships between various layer thicknesses and the XRR FT peaks

Through a series of simulated reflectivity scans and the application of the Q_z^4 multiplication and the differentiation described above to the simulation scans, four clear peaks are expected to emerge from the FT. These peaks represent different combinations of the layer thicknesses as identified in Fig. 23. For example, layer "t_1" is a combination of the thicknesses of the top four layers, t_2 is a combination of the top five layers, t_3 is from the bottom three layers and t_4 is a sum of thicknesses from the entire stack. Figure 24 shows the experimental specular reflectivity scan from this structure, and the FT of the data after multiplication by Q_z^4 and differentiation as described above for the AlN layer. A simulation based on the FT values is included with the experimental reflectivity data as a dashed line.

The FT transform of the processed data clearly show the same four peaks. With these values, the thicknesses of several of the individual layer peaks can be determined. Of prime importance is the thickness of the InAs channel layer. This is $t_4 - t_2 = 150$ Å which is the intended thickness and matches that from the best fit XRR simulation (154 Å) as well as transmission electron microscopy measurements (148-152 Å). The other layer thicknesses obtained by our new method also match those from the simulated reflectivity scan and from TEM. In fact, even the thin InAs donor layer (This can be obtained, for example, as $t_1 + t_3 - t_4$) is extracted. Using the values in the transformation, the layer thickness is 10 ± 3 Å, also comparable to the simulated value (10 Å) and the TEM measurement (10 ± 3 Å). The benefit with the differentiation approach is that the pulse truncation peak near the origin would overwhelm the peaks due to the thinner layers using any of the previously published processes. The identification of these peaks is confirmed by FT transforms of simulated scans which also shows that errors related to ignoring refractive index differences from layer to layer are small (thickness differences of less than ± 2 Å).

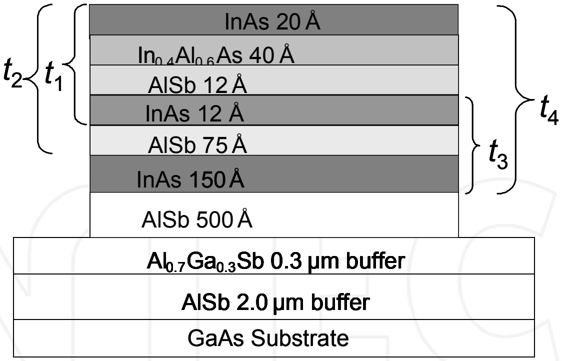


Fig. 23. Schematic of AlSb / InAs heterostructures. t_1, t_2, t_3, and t_4 represent the major thickness values from the FT

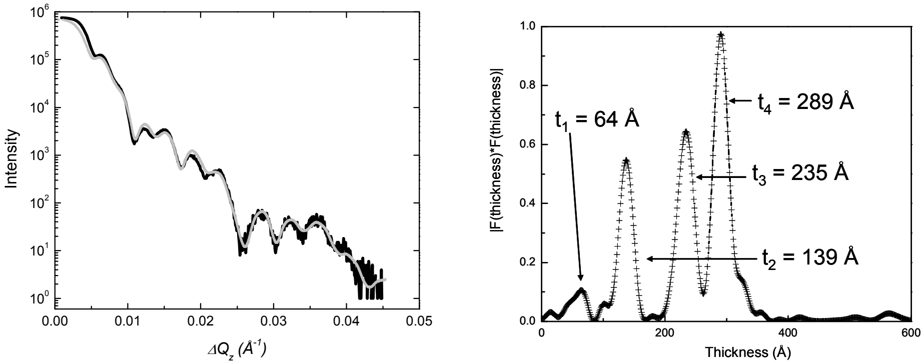


Fig. 24. The specular reflectivity scan and simulation (left, simulation is gray line superimposed on the experimental data, black line) and the FT (right) - after Q_z^4 leveling and differentiation - with the extracted film thicknesses

6. Summary

Fourier transforms of X-ray reflectivity scans from multi-layer structures provide useful layer thickness information. A challenge in the transform process is to address the Q_z^{-4} sloping background which is present with the layer thickness fringe oscillations. Previous studies included several different methods to address this sloping background issue with limited success. Our approach combines one of these previous methods – logarithmic compression or Q_z^4 multiplication – with differentiation of the x-ray reflectivity data. The differentiation more effectively separates the contribution from the sloping background from the thickness fringe oscillations, making the subsequent Fourier transform more closely related to only the layer thicknesses. This approach is also less susceptible to the presence of a limited number of thickness oscillations or a truncated data set. Data from a single layer (320 Å AlN on an Al_2O_3 substrate) was utilized to compare the results from each of the previous techniques with the differentiation step. The combination of the differentiation step with either the logarithmic compression or the Q_z^4 leveling produces clear layer thickness peaks after the Fourier transform.

The differentiation approach described here also proved to be very effective at extracting layer thickness information from multi-layer structures that produce complicated specular x-ray reflectivity scans. The Fourier transform produces a series of sharp peaks that originate from different layer thicknesses. The peaks represent the combination of different layer thicknesses and the thickness of each layer can be determined with the assistance of the use of specular x-ray reflectivity simulation scans. This process was demonstrated for an InAs / AlSb multi-layer structure. This technique has applications to any multi-layer system with thin layers including complex optical coatings, multi-layer metallic coatings for giant magnetoresistance measurements and for any other multi-layer structure that can be measured using specular x-ray reflectivity scans.

7. Acknowledgments

The authors acknowledge funding for this work from the National Science Foundation and Northrop Grumman, Inc.

8. References

Banerjee, S., G. Raghavan, et al. (1999). *Journal of Applied Physics* 85: 7135.

Bowen, D. K. and B. K. Tanner (1998). *High Resolution X-ray Diffractometry and Topography*. Bristol, PA, Taylor & Francis Inc.

Daillant, J. and A. Gibaud (1999). *X-ray and Neutron Reflectivity: Principles and Applications*. Berlin, Heidelberg, Springer Verlag.

Durand, O. (2004). *Thin Solid Films* 450: 51-59.

Li, M., M. O. Muller, et al. (1996). *Journal of Applied Physics* 80: 2788.

Parratt, L. G. (1954). *Physical Review* 95: 359-369.

Peralta, L. G. d. and H. Temkin (2003). *Journal of Applied Physics* 93(4): 1974-1977.

Russel, T. P. (1990). *Mater. Sci. Rep*. 5: 171.

Sakurai, K. and A. Iida (1992). *Adv. X-ray Anal.* 35: 813.

Wormington, M., C. Panaccione, et al. (1999). *Philosophical Transactions of the Royal Society of London* A 357: 2827-2848.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

18

Fourier Transform on Group-Like Structures and Applications

Massoud Amini[1], Mehrdad Kalantar[2], Hassan Myrnouri[3] and Mahmood M. Roozbahani[4]

[1,4]*Department of Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran 14115-134,*

[2]*School of Mathematics and Statistics, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6,*

[3]*Department of Mathematics, Islamic Azad University, Lahijan Branch, Lahijan*

[1,3,4]*Iran*

[2]*Canada*

1. Introduction

The Fourier analysis was invented by the French mathematician, Joseph Fourier, in the early nineteenth century (28). The Fourier transform (FT) is an operation that transforms a function of a real variable in a given domain into another function in another domain (there are generalizations to complex or several real variables). The domains differ from one application to another. In signal processing, typically the original domain is time and the target domain is frequency. The transform captures those frequencies present in the original function. The Fourier transform and its generalizations are part of the Fourier analysis (23). The Fourier transform on discrete structures such as finite groups (DFT), opens up the issue of the time complexity of the algorithm which computes FT. Particularly, efficient computation of a fast Fourier transform (FFT) is essential for high-speed computing (8). There is a vast literature on the theory of Fourier transform on groups, including Fourier transform on locally compact abelian groups (27), on compact groups (34), and finite groups (73). In this chapter, we review the generalizations of the Fourier transform on group-like structures, including inverse semigroups (43), hypergroups (10), and groupoids (70). This is a vast subject with an extensive literature, but here a personal view based on the authors' research is presented. References to the existing literature is given, as needed. The Fourier transform on inverse semigroups are introduced in (47). The basics of the theory of Fourier transform on commutative hypergroups are discussed in (10). The Fourier transform on arbitrary compact hypergroups is first studied in (74). The section on the Fourier transform of unbounded measures on commutative hypergroups is taken from (2). An application of the quantum Fourier transform (QFT) on finite commutative hypergroups in quantum computation is given in (4). Finally, the Fourier transform on compact groupoids is discussed in (1), (3). The case of abelian groupoids (57) is considered in (6).

2. Outline

3. Classical Fourier transform

Fourier transform has a long history. Some variants of the discrete (cosine) Fourier transform were used by Alexis Clairaut in 1754 to do some astronomical calculations, and the discrete (sine) Fourier transform in 1759 by Joseph Louis Lagrange to compute the coefficients of a trigonometric series for a vibrating string. The latter used a discrete Fourier transform of order 3 to study the solution of a cubic. Finally Carl Friedrich Gauss used a full discrete Fourier transform in 1805 to find trigonometric interpolation of asteroid orbits.

Although d'Alembert and Gauss had already used trigonometric series to study the heat equation, it was first in the 1807 seminal paper of Joseph Fourier that the idea of expanding all (continuous) functions by trigonometric series was explored.

Broadly speaking, the Fourier transform is a systematic way to decompose generic functions into a superposition of symmetric functions (72). In this broad sense, even the decomposition of a real function into its odd and even parts is an instance of a Fourier series. Similarly (72) if complex function $w = f(z)$ is a harmonic of order j, that is $f(e^{2\pi i/n}z) = e^{2\pi ij/n}f(z)$ for all $z \in \mathbb{C}$ then

$$f(z) = \sum_{j=0}^{n-1} \frac{1}{n} \sum_{k=0}^{n-1} f(e^{2\pi ik/n}z)e^{-2\pi ijk/n}.$$

In general, the monomial $f(z) = z^n$ has rotational symmetry of order n and each continuous function $f : \mathbb{T} \to \mathbb{C}$ could be expanded (in an appropriate norm) as $f(z) = \sum_{n=-\infty}^{\infty} \hat{f}(n)z^n$ where

$$\hat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta})e^{-in\theta}d\theta.$$

One important feature of this expansion is that it could be considered as a generalization of the case where a complex analytic function f on the closed unit disk $\mathbb{D}$ is expanded in its Taylor series.

For a function $f : \mathbb{R}^d \to \mathbb{C}$, under some very natural conditions we have the dual formulas

$$f(x) = \int_{\mathbb{R}^d} \hat{f}(\xi)e^{2\pi ix\xi}d\xi, \quad \hat{f}(\xi) = \int_{\mathbb{R}^d} f(x)e^{-2\pi ix\xi}dx.$$

Consider an integrable (real or complex) function on the interval $[0, 2\pi]$ with Fourier coefficients $\hat{f}(n)$ as above. Then an important classical problem is the problem of convergence of the corresponding Fourier series. More precisely, if

$$S_N(f)(t) = \sum_{n=-N}^{N} \hat{f}(n)e^{int}$$

then it is desirable to have (necessary and) sufficient conditions on f so that the sequence $\{S_N(f)\}$ converges to f in a given function topology.

For norm convergence, we know that if $f \in L^p[0, 2\pi]$ for $1 < p < \infty$, then $\{S_N(f)\}$ converges to f in L^p-norm (for $p = 2$, this is Riesz-Fisher theorem). The pointwise convergence is more delicate. There are many known sufficient conditions for $\{S_N(f)\}$ to converge to f at a given point x, for example if the function is differentiable at x; or even if the function has a discontinuity of the first kind at x and the left and right derivatives at x exist and are finite, then $\{S_N(f)(x)\}$ will converge to $\frac{1}{2}(f(x^+) + f(x^-))$ (but by the Gibbs phenomenon, it has large oscillations near the jump). By a more general sufficient condition (called the Dirichlet-Dini Criterion) if f is 2π-periodic and locally integrable and

$$\int_0^{\pi} |\frac{f(x+t) - f(x-t)}{2} - \ell| \frac{dt}{t} < \infty$$

then $\{S_N(f)(x)\}$ converges to ℓ. In particular, if f is of Holder class $\alpha > 0$, then $\{S_N(f)(x)\}$ converges everywhere to $f(x)$. Indeed, in the latter case, the convergence is uniform (by Dini-Lipschitz test). If f is only continuous, the sequence of n-th averages of $\{S_N(f)\}$ converge uniformly to f, that is the Fourier series is uniformly Cesaro summable (also the Gibbs phenomenon disappears in the pointwise convergence of the Cesaro sum).

In general, for a given continuous function f the Fourier series converges almost anywhere to f (this holds even for a square integrable function: Carleson theorem, or L^p function: Hunt theorem) and when f is of bounded variation, the Fourier series converges everywhere to f (by Dini test). If a function is of bounded variation and Holder of class $\alpha > 0$, then the Fourier series is absolutely convergent.

However, the family of continuous functions whose Fourier series converges at a given x is of first Baire category in the Banach space of continuous functions on the circle. This means that for most continuous functions the Fourier series does not converge at a given point. Kolmogorov constructed a concrete example of a function in L^1 whose Fourier series diverges almost everywhere (later examples showed that this may happen everywhere). More generally, for any given set E of measure zero, there exists a continuous function f whose Fourier series fails to converge at any point of E (Kahane-Katznelson theorem).

4. Fourier transform on groups

The Fourier transform could be defined in general on any locally compact Hausdorff group G. This is done using the (left) Haar measure, a (left) translation invariant positive Borel measure λ on G whose existence and uniqueness (up to positive scalars) is proved in abstract harmonic analysis (27, 2.10). If $f \in L^1(G) := L^1(G, \lambda)$ and $\pi : G \to \mathcal{B}(\mathcal{H}_\pi)$ is an irreducible (unitary) representation of G (we write $\pi \in \hat{G}$) then we may define the Fourier transform of f

$$\hat{f}(\pi) = \int_G f(x)\pi(x^{-1})d\lambda(x).$$

Similarly the Fourier-Stieljes transform of a bounded Borel measure $\mu \in M(G)$ is defined by

$$\hat{\mu}(\pi) = \int_G \pi(x^{-1})d\mu(x),$$

for $\pi \in \hat{G}$.

The Fourier transform is well studied in the two special cases where G is abelian or compact (finite). We give more details in the next two sections.

4.1 Abelian groups

When G is abelian, each irreducible representation is one dimensional (27, 3.6) and could be identified with a (continuous) character $\chi \in \hat{G}$. A character is a continuous group homomorphism $\chi : G \to \mathbb{T}$. In this case, the dual object $\hat{G}$ is itself a locally compact abelian group and one could employ the Fourier transform to prove the Pontrjagin duality, that is $(\hat{G})\hat{} \simeq G$ as topological groups.

If $f \in L^1(G)$ then the Fourier transform of f is defined on $\hat{G}$ by

$$\hat{f}(\chi) = \int_G f(x)\overline{\chi(x)}d\lambda(x)$$

and $\hat{f} \in C_0(\hat{G})$ (Riemann-Lebesgue lemma) and $\|\hat{f}\|_\infty \leq \|f\|_1$, but the Fourier transform is not an isometry from $L^1(G)$ to $C_0(\hat{G})$ (even for $G = \mathbb{R}$). However we have the inversion formula: if $B(G)$ is the linear span of the positive definite functions on G (27, p. 84) and $f \in B(G) \cap L^1(G)$ then $\hat{f} \in L^1(\hat{G})$ and with a suitable normalization of the Haar measure ϖ of $\hat{G}$

$$f(x) = \int_{\hat{G}} \hat{f}(\chi)\chi(x)d\varpi(\chi),$$

for almost all $x \in G$. By the Pontrjagin duality, this means that $f(x) = (\hat{f})\hat{}(x^{-1})$ for λ-a.e. $x \in G$.

When G is compact and abelian and $f \in L^2(G)$ then the Fourier-Plancherel transform of f is defined on the discrete abelian group $\hat{G}$ by

$$\hat{f}(\chi) = \int_G f(x)\overline{\chi(x)}d\lambda(x),$$

which converges by Cauchy-Schwartz inequality (similar definition works for $f \in L^p(G)$, $1 < p < \infty$, by Holder inequality). In this case $\hat{f} \in \ell^2(\hat{G})$ ($f \in \ell^q(\hat{G})$, $\frac{1}{p} + \frac{1}{q} = 1$, respectively) and $\|\hat{f}\|_2 = \|f\|_2$, hence the Fourier-Plancherel transform is an isometry from $L^2(G)$ to $\ell^2(\hat{G})$.

4.2 Compact and finite groups

If G is a compact Hausdorff group, each irreducible representation $\pi \in \hat{G}$ is finite dimensional (27, 5.2), say with dimension d_π, and the linear span of the matrix elements of π (coefficients of π) is a finite dimensional space $\mathcal{E}_\pi$ of dimension at most d_π^2. Moreover the linear span $\mathcal{E}(G)$ of the union $\cup_{\pi\in\hat{G}}\mathcal{E}_\pi$ (trigonometric functions on G) is uniformly dense in $C(G)$ and $L^2(G) = \bigoplus_{\pi\in\hat{G}} \mathcal{E}_\pi$ (Peter-Weyl theorem). If $f \in L^1(G)$, the Fourier transform of f

$$\hat{f}(\pi) = \int_G f(x)\pi(x^{-1})d\lambda(x).$$

defines an element $\hat{f} \in \bigoplus_{\pi\in\hat{G}} \mathbb{M}_{d_\pi}(\mathbb{C})$ with coefficients $\hat{f}(\pi)_{ij} = \int_G f(x)\langle\pi(x^{-1})e_i, e_j\rangle d\lambda(x)$, hence if $f \in L^1(G) \cap L^2(G)$,

$$f = \sum_{\pi\in\hat{G}} d_\pi tr(\hat{f}(\pi)\pi(.))$$

in L^2-norm with $\|f\|_2^2 = \sum_{\pi\in\hat{G}} d_\pi tr(\hat{f}(\pi)^*\hat{f}(\pi))$. If χ_π is the character associated to the irreducible representation $\pi \in \hat{G}$ by $\chi_\pi = tr(\pi(.))$, then $\{\chi_\pi : \pi \in \hat{G}\}$ is an orthonormal basis of $ZL^2(G)$ consisting of central functions in $L^2(G)$ (27, 5.23).

Given a complex-valued function f on a finite group G, we may present f as

$$f = \sum_{g\in G} f(g)g,$$

viewing f as an element of the group algebra $\mathbb{C}G$. The Fourier basis of $\mathbb{C}G$ comes from the decomposition of the semisimple algebra $\mathbb{C}G$ as the direct sum of its minimal left ideals

$$\mathbb{C}G = \oplus_{i=1}^n M_i,$$

and taking a basis for each M_i. When $G = \mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$ is the cyclic group of order n, an element f of the group algebra $\mathbb{C}\mathbb{Z}_n$ is a signal, sampled at n evenly spaced points in time. The minimal left ideals of $\mathbb{C}\mathbb{Z}_n$ are one dimensional, and the Fourier basis is unique (up to scaling factors), and is indeed the usual basis of exponential functions given by the classical discrete Fourier transform. Hence the expansion of f in a Fourier basis corresponds to a re-expression of f in terms of the frequencies that comprise f (45).

The efficiency of computing the Fourier transform of an arbitrary function on G depends on the choice of basis in which f is expanded. A naive computation requires $|G|^2$ operations, where an operation is a complex multiplication followed by a complex addition. Much better results are obtained for different groups (21), (48), (49), (50), and (51) (see also (68) and references therein). For example, computing the Fourier transform requires no more than $O(|G|log^k|G|)$ operations on the cyclic group $G = \mathbb{Z}_n$ (13; 19), symmetric group $G = S_n$ (47), and hyper-octahedral group $G = B_n$ (67), with $k = 1, 2, 4$, respectively. On the other hand, there is no known $O(|G|log^c|G|)$ algorithms for matrix groups over a finite field (48).

4.2.1 Discrete Fourier transform

The classical discrete Fourier transform of a sequence $\{f(j)\}_{0\leq j\leq n-1}$ is defined

$$\hat{f}(k) = \sum_{j=0}^{n-1} f(j) e^{-2ikj/n},$$

for $0 \leq k \leq n-1$, and

$$f(j) = \frac{1}{n}\sum_{k=0}^{n-1} \hat{f}(k) e^{2ikj/n},$$

for $0 \leq j \leq n-1$.

For $G = \mathbb{Z}_n$, the complex irreducible representations χ_k of $\mathbb{Z}/n\mathbb{Z}$ are one dimensional,

$$\chi_k(j) = e^{-2ikj/n},$$

and the group algebra $\mathbb{C}\mathbb{Z}_n$ has only one natural basis,

$$b_k = \frac{1}{n}\sum_{j=0}^{n-1}(e^{2ikj/n})j,$$

and $\hat{f}(k) = \hat{f}(\chi_k)$.

4.2.2 Fast Fourier transform

The classical FFT has revolutionized signal processing. Applications include fast waveform smoothing, fast multiplication of large numbers, and efficient waveform compression, to name just a few (14). FFTs for more general groups have applications in statistical processing. For example, the FFT on Z_2^k (76) allows for efficient 2^k-factorial analysis. That is, it allows for the efficient statistical analysis of an experiment in which each of k variables may take on one of two states. The FFT on S_n allows for an efficient statistical analysis of votes cast in an election involving n candidates (20).

A direct calculation of the Fourier transform of an arbitrary function on $\mathbb{Z}_n$ requires $n^2 = |\mathbb{Z}_n|^2$ operations. The classical fast discrete Fourier transform (of Cooley-Tukey (19)) expressed in the group language (49) is as follows: Suppose $n = pq$ where p is a prime. Then $\mathbb{Z}_n$ has a subgroup H isomorphic to $\mathbb{Z}_q$. By the reversal decomposition $\mathbb{Z}_n = \cup_{0\leq j\leq p-1}(j+H)$. For $0 \leq k \leq n-1$,

$$\hat{f}(k) = \hat{f}(\chi_k) = \sum_{j=0}^{n-1} f(j)\chi_k(j) = \sum_{j=0}^{p-1}\chi_k(j)\sum_{h\in H} f(j+h)\chi_k(h).$$

Now the inner sums in the last term are Fourier coefficients on H. This reduces the problem of computing the Fourier transform on G to the same computation on H. When $n = 2^k$ (a k-bit number in digital terms), the Fourier transform on $\mathbb{Z}_{2^k}$ could be computed in $2^k + k2^{k+1} = O(nlogn)$ operations.

The reduction argument involved in the above FFT algorithm could basically be applied to any finite group (or semigroup). However, since the irreducible representations are not one dimensional for non abelian case, more is needed to be explored in the general case. In particular, one needs to understand the way (finite dimensional) irreducible representations behave when restricted to subgroups.

The basic idea of Cooley-Tukey is adapted by Clausen (18) to create an FFT for the symmetric group S_n. A standard reference for representation theory of the symmetric group is (38). The Clausen algorithm for FFT on the symmetric group S_n requires a set of inequivalent, irreducible, chain-adapted representations relative to the chain $S_n > S_{n-1} > \cdots > S_1 = \{e\}$, where e is the identity of S_n and S_k is the subgroup of S_n consisting of permutations which fix $k+1$ through n. Two such sets of representations are provided by the Young seminormal (62) and orthogonal forms (77). (for generalizations to seminormal representations of Iwahori-Hecke algebras see (35) and (62)).

A partition of a nonnegative integer k is a weakly decreasing sequence λ of nonnegative integers whose sum is k. Two partitions are equal if they only differ in number of zeros they contain. We write $\lambda \vdash k$. A complete set of (equivalence classes of) irreducible representations for S_n is indexed by the partitions of n (38). A partition corresponds to its Young diagram, consisting of a table whose i-th row has $\lambda(i)$ boxes. If we fill these boxes with numbers from $\{1, 2, ...n\}$ such that each number appears exactly once and column (row) entries increase from top to bottom (left to right), we get a standard tableau of shape λ. Now S_n acts on tableaux by permuting their entries. If L is a standard tableau of shape λ, and $L(i)$ is the box in L in the position (k, ℓ), we put $|L(i)| = \ell - k$ and define the action of the basic permutation $\sigma = \sigma_i = (i-1, i)$ on the vector space V^λ generated by a basis $\{v_L\}$ indexed by all standard tableaux L of shape λ by

$$\sigma v_L = (|L(i)| - |L(i-1)|)^{-1} v_L + (1 + (|L(i)| - |L(i-1)|)^{-1}) v_{\sigma L},$$

with the convention that $v_{\sigma L} = 0$, if σL is not standard. Young showed that these actions exhaust $\hat{S}_n = \{\rho^\lambda : \lambda \vdash n\}$. Moreover if λ^- is the set of all partitions of $n-1$ obtained by removing a corner (a box with no box to the right or below) of λ then $V^\lambda = \bigoplus_{\mu \in \lambda^-} V^\mu$ and if we order the basis of V^λ with the last letter ordering of standard tableaux,

$$\rho^\lambda|_{S_{n-1}} = \bigoplus_{\mu \in \lambda^-} \rho^\mu.$$

Now Clausen algorithm uses the above Young seminormal representations to find FFT for S_n (18). If $S_n = \cup_{1 \le i \le n} \tau_i S_{n-1}$ is the transversal decomposition of S_n into left cosets of the subgroup S_{n-1}, where $\tau_i = \sigma_{i+1}\sigma_{i+2}\ldots\sigma_n$, for $i < n$ and $\tau_n = e$, the identity of S_n, then for $\lambda \vdash n$,

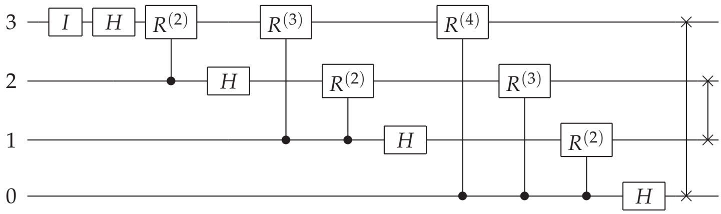
$$\hat{f}(\rho^\lambda) = \sum_{\sigma \in S_n} f(\sigma)\rho^\lambda(\sigma) = \sum_{i=1}^{n} \rho^\lambda(\tau_i) \sum_{\sigma \in S_{n-1}} f(\tau_i \sigma)\rho^\lambda(\sigma).$$

Therefore the minimum number of needed operations is at most $\frac{2}{3}n(n+1)^2 n! = O(n! log^3 n!)$.

4.2.3 Quantum Fourier transform

In quantum computing, the quantum Fourier transform (QFT) is the quantum analogue of the discrete Fourier transform applied to quantum bits (qubits). Mathematically, QFT as a unitary operator is nothing but the Fourier-Plancherel transform on the Hilbert space $\ell^2(G)$, for some appropriate (abelian) group G. For n-qubits, this group could be considered to be $G = \mathbb{Z}_{2^n} = \mathbb{Z}/2^n\mathbb{Z}$, but in practice it could be any finite group.

QFT is an integral part of many famous quantum algorithms, including factoring and discrete logarithm, the quantum phase estimation (estimating the eigenvalues of a unitary matrix) and the hidden subgroup problem (HSP).



Scheme 1. *Quantum Circuit of 4-qubit Quantum Fourier Transform*

The quantum Fourier transform can be performed efficiently on a quantum computer, with a particular decomposition into a product of simpler unitary matrices. Using a canonical decomposition, the discrete Fourier transform on n-qubits can be implemented as a quantum circuit consisting of only $O(n^2)$ Hadamard gates and controlled phase shift gates (58). There are also QFT algorithms requiring only $O(nlogn)$ gates (32). This is exponentially faster than the classical discrete Fourier transform (DFT) on n bits, which takes $O(n2^n)$ gates. However, QFT can not give a generic exponential speedup for any task which requires DFT.

The QFT on $\mathbb{Z}_N$ maps a quantum state $\sum_{j=0}^{N-1} x_j|j\rangle$ to the quantum state $\sum_{k=0}^{N-1} y_k|k\rangle$ with $y_k = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} x_j\omega^{jk}$, where $\omega = e^{2\pi i/N}$ is a primitive N-th root of unity. This is a surjective isometry on the finite dimensional Hilbert space $\ell^2(\mathbb{Z}_N)$ whose corresponding unitary matrix is $\mathfrak{F}_N = \frac{1}{\sqrt{N}}[\omega^{jk}]_{0\leq j,k\leq N-1}$. The case $N = 2^n$ handles the transformation of n-qubits.

Using the Hadamard and phase gates

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad R^{(k)} = \begin{pmatrix} 1 & 0 \\ 0 & 2\pi i/2^k \end{pmatrix}$$

one could implement the QFT over $n = 4$ qubits efficiently as in the above circuit (46) (see also (15)).

5. Fourier transform on monoids and inverse semigroups

A semigroup is a nonempty set S together with an associative binary operation. If S has an identity element, it is called a monoid. A (finite dimensional) representation of S (of dimension d) is a homomorphism from S to the semigroup $\mathbb{M}_d(\mathbb{C})$ with matrix multiplication. An inverse semigroup is a semigroup S such that for each $x \in S$ there is a unique $y = x^* \in S$ such that $xyx = x$ and $yxy = y$. The set $E = \{xx^* : x \in S\}$ is a commutative subsemigroup of S.

For a finite inverse semigroup S, the semigroup algebra $\mathbb{C}S$ with convolution

$$f * g(s) = \sum_{r,t\in S, rt=s} f(r)g(t)$$

is semisimple (53), hence representations of S are (equivalent to) a direct sum of irreducible and null representations of S. By Wedderburn theorem, the set $\hat{S}$ of (equivalence classes of) irreducible representations of S is finite and the map

$$\bigoplus_{\rho\in\hat{S}} \rho : \mathbb{C}S \to \bigoplus_{\rho\in\hat{S}} \mathbb{M}_{d_\rho}(\mathbb{C})$$

is an isomorphism of algebras sending f to $\bigoplus_{\rho\in\hat{S}}\sum_{s\in S} f(s)\rho(s)$. This isomorphism is indeed the Fourier transform on S, that is $\hat{f}(\rho)=\sum_{s\in S} f(s)\rho(s)$, for $f\in\mathbb{C}S$.

A standard example of finite inverse semigroups is the rook monoid R_n consisting of all injective partial functions on $\{1,\dots,n\}$ under the operation of partial function composition. R_n is isomorphic to the semigroup of all $n\times n$ matrices with all 0 entries except at most one 1 in each row or column (corresponding to the set of all possible placements of non-attacking rooks on an $n\times n$ chessboard). The rook monoid plays the role of the symmetric group for finite groups in the category of finite inverse semigroups. Each finite inverse semigroup S is isomorphic to a subsemigroup of R_n, for $n=|S|$ (43). The (fast) Fourier transform on R_n is studied in details in (45), which is applied to the analysis of partially ranked data. We give a brief account of FFT on rook monoids in the next section, and refer the interested reader to (45) for details.

5.1 Fast Fourier transform on rook monoids

In this section we give a fast algorithm to compute $\hat{f}$ for $f\in\mathbb{C}R_n$ mimicing the Clausen algorithm for FFT on S_n (45). There is a much faster algorithm based on groupoid bases (45, 7.2.3), but the present approach has the advantage that explicitly employs the reversal decomposition.

As in S_n, let $\sigma_i=(i-1,i)\in R_n$ and put $\tau_i=\sigma_{i+1}\sigma_{i+2}\dots\sigma_n$, for $i<n$ and $\tau_n=e$, the identity of R_n. Moreover, put $\tau^i=\sigma_n\sigma_{n-1}\dots\sigma_{i+1}$, for each i, and let R_k be the subsemigroup of R_n consisting of those partial permutations σ with $\sigma(j)=j$, for $j>k$. Then Halverson has characterized $\hat{R}_n$ using the semigroup chain $R_n>R_{n-1}>\cdots>R_1$, similar to the Young seminormal representations of S_n (35). Considering the transversal decomposition of R_n into cosets of R_{n-1}, one should note that two distinct left cosets of R_{n-1} do not necessarily have the same cardinality. Indeed, we have (45, 2.4.4)

$$|R_n|=2n|R_{n-1}|-(n-1)^2|R_{n-2}|.$$

This is because R_n consists of those rook matrices having all zeros in column and row 1, those having 1 in position $(\alpha,1)$ for $1\leq\alpha\leq n$, or in position $(1,\alpha)$ for $2\leq\alpha\leq n$, counting twice those with ones in positions $(\alpha,1)$ and $(1,\beta)$ for $2\leq\alpha,\beta\leq n$. This suggests the following decomposition for $n\geq 3$ and $\rho\in\hat{R}_n$,

$$\begin{aligned}\hat{f}(\rho)=&\sum_{i=1}^{n}\rho(\tau_i)\sum_{\sigma\in R_{n-1}} f(\tau_i\sigma)\rho(\sigma)+\rho([n])\sum_{\sigma\in R_{n-1}} f([n]\sigma)\rho(\sigma)\\&+\sum_{i=1}^{n-1}\rho(\tau^i)\sum_{\sigma\in R_{n-2}} f(\sigma\tau^i)\rho(\sigma),\end{aligned}$$

where $[n]$ is the identity of R_{n-1} considered as an element of R_n (not defined at n). This decomposition follows from dividing elements $\sigma\in R_n$ into three parts: those with $\sigma(n)=i$ for some $1\leq i\leq n$, those not defined at n with $\sigma(i)=n$ for some $1\leq i\leq n-1$, and those not defined at n with $\sigma(i)\neq n$ for all $1\leq i\leq n$ (45, 9.1.1). This leads the upper bound

$$2\sum_{\rho\in\hat{R}_n}\sum_{i=1}^{n}2(n-i)d_\rho^2+\sum_{\rho\in\hat{R}_n} d_\rho^2+\sum_{\rho\in\hat{R}_n}(2n-1)d_\rho^2,$$

where twice the first sum calculates the maximum number of operations involved in calculation of the matrix products in the first and third terms of the above decomposition, the second sum calculates the maximum number of operations involved in calculation of the matrix products in the second term, and the last sum calculates the maximum number of operations involved in calculation of all matrix sums involved in the whole decomposition. This upper bound is at most

$$2n(n-1)|R_n| + |R_n| + (2n-1)|R_n|.$$

This leads to the upper bound $n2^n|R_n| = O(|R_n|^{1+\varepsilon})$ for the minimum number of operations needed to calculate the FFT on R_n (45, 9.2.3), which could be improved as $(\frac{3}{4}n^2 + \frac{2}{3}n^3)|R_n| = O(|R_n|log^3|R_n|)$ using more sophisticated decompositions (45, 7.2.3).

6. Fourier transform on hypergroups

Fourier and Fourier-Stieltjes transforms play a central role in the theory of absolutely integrable functions and bounded Borel measures on a locally compact abelian group G (70). They are particularly important because they map the group algebra $L^1(G)$ and measure algebra $M(G)$ onto the Fourier algebra $A(G)$ and Fourier-Stieltjes algebra $B(G)$, respectively. There are important Borel measures on a locally compact, non-compact, abelian group G which are unbounded. A typical example is a left Haar measure. L. Argabright and J. de Lamadrid in (7) explored a generalized Fourier transform of unbounded measures on locally compact abelian groups. This theory has recently been successfully applied to the study of quasi-crystals (8).

A version of generalized Fourier transform is defined for a class of commutative hypergroups (10). Some of the main results (7) are stated and proved in (10) for hypergroups, but the important connection with Fourier and Fourier-Stieltjes spaces are not investigated in (10) (in contrast with the group case, these may fail to be closed under pointwise multiplication for general hypergroups.) Recently these spaces are studied in (5), and under some conditions, they are shown to be Banach algebras. Section 6.2 investigates the relation between transformablity of unbounded measures on strong commutative hypergroups and these spaces. In the next section we study unbounded measures on locally compact (not necessarily commutative) hypergroups. The main objective of this section is the study of translation bounded measures. These are studied in (10) in commutative case. The main results (Theorem 6.18 and Corollary 6.20) are stated and proved in section 6.2.1. The former states that an unbounded measure on a strong commutative hypergroup is transformable if and only if its convolution with any positive definite function of compact support is positive definite. The latter gives the condition for the transform of this measure to be a function.

A hypergroup is a triple $(K, ^-, *)$ where K is a locally compact space with an involution $^-$and a convolution $*$ on $M(K)$ such that $(M(K), *)$ is an algebra and for $x, y \in K$,

(i) $\delta_x * \delta_y$ is a probability measure on K with compact support,

(ii) the map $(x, y) \in K^2 \mapsto \delta_x * \delta_y \in M(K)$ is continuous,

(iii) the map $(x, y) \in K^2 \mapsto \text{supp}\,(\delta_x * \delta_y) \in \mathcal{C}(K)$ is continuous with respect to the Michael topology on the space $\mathcal{C}(K)$ of nonvoid compact sets in K,

(iv) K admits an *identity* e satisfying $\delta_e * \delta_x = \delta_x * \delta_e = \delta_x$,

(v) $(\delta_x * \delta_y)^- = \delta_{\bar{y}} * \delta_{\bar{x}}$,

(vi) $e \in \text{supp}\,(\delta_x * \delta_y)$ if and only if $x = \bar{y}$.

For $\mu \in M(K)$ and Borel subset $E \subseteq K$ put $\bar{\mu}(E) = \mu(\bar{E})$, where $\bar{E} = \{\bar{x} : x \in E\}$. This is an involution on $M(K)$ making it a Banach $*$-algebra. A representation of K is a $*$-representation π of $M(K)$ such that $\pi(\delta_e) = id$ and $\pi \mapsto \pi(\mu)$ is weak-weak operator-continuous from $M_+(K)$ to $\mathcal{B}(\mathcal{H}_\pi)$. When π is irreducible, we write $\pi \in \hat{K}$. Define the conjugation operator D_π on $\mathcal{H}_\pi$ by

$$D_\pi(\sum_{i=1}^{d_\pi} \alpha_i \xi_i^\pi) = \sum_{i=1}^{d_\pi} \bar{\alpha}_i \xi_i^\pi,$$

and put $\bar{\pi} = D_\pi \pi D_\pi$. For $\mu \in M(K)$, the Fourier-Stieltjes transform of μ is defined by $\hat{\mu}(\pi) = \bar{\pi}(\mu)$. Then $\hat{\mu} \in \mathcal{E}(K) := \bigoplus_{\pi\in\hat{K}} \mathcal{B}(\mathcal{H}_\pi)$ and the Fourier-Stieltjes transform from $M(K)$ to $\mathcal{E}(K)$ is one-one.

6.1 Compact hypergroups

Let K be a compact hypergroup and $\hat{K}$ denote the set of equivalence classes of all continuous irreducible representations of K. For $\pi \in \hat{K}$, let $\{\xi_i^\pi\}_{i=1}^{d_\pi}$ be an orthonormal basis for the corresponding (finite dimensional) Hilbert space $\mathcal{H}_\pi$ and put

$$\pi_{i,j}(x) = \langle \pi(x)\xi_i^\pi, \xi_j^\pi \rangle \quad (1 \le i,j \le d_\pi).$$

Let $Trig_\pi(K) = span\{\pi_{i,j} : 1 \le i,j \le d_\pi\}$ and $Trig(K) = span\{\pi_{i,j} : \pi \in \hat{K}, 1 \le i,j \le d_\pi\}$. Then $dimTrig_\pi(K) = d_\pi^2$ and there is $k_\pi \ge d_\pi$ such that

$$\int_K \pi_{i,j}\bar{\sigma}_{r,s} dm = k_\pi^{-1}\delta_{\pi,\sigma}\delta_{i,r}\delta_{j,s} \quad (\pi,\sigma \in \hat{K}) \text{ (74, 2.6).}$$

Also $\{k_\pi^{\frac{1}{2}}\pi_{i,j} : \pi \in \hat{K}, 1 \le i,j \le d_\pi\}$ is an orthonormal basis of $L^2(K)$ and

$$Trig(K) = \oplus_{\pi\in\hat{K}} Trig_\pi(K) \text{ (74, 2.7).}$$

In particular, $Trig(K)$ is norm dense in both $C(K)$ and $L^2(K)$ (74, 2.13, 2.9).
For each $f \in L^2(K)$ we have the Fourier series expansion

$$f = \sum_{\pi\in\hat{K}} \sum_{i,j=1}^{d_\pi} k_\pi \langle f, \pi_{i,j} \rangle \pi_{i,j}$$

where the series converges in L^2-norm.
Consider the $*$-algebra

$$\mathcal{E}(\hat{K}) := \prod_{\pi\in\hat{K}} B(\mathcal{H}_\pi)$$

with coordinatewise operations. For $f = (f_\pi) \in \mathcal{E}(\hat{K})$ and $1 \le p < \infty$ put

$$\|f\|_p := \Big(\sum_{\pi\in\hat{K}} k_\pi \|f_\pi\|_p^p\Big)^{\frac{1}{p}}, \quad \|f\|_\infty := sup_{\pi\in\hat{K}} \|f_\pi\|_\infty,$$

where the right hand side norms are operator norms as in (34, D.37, 36(e)). Define $\mathcal{E}_p(\hat{K})$, $\mathcal{E}_\infty(\hat{K})$, and $\mathcal{E}_0(\hat{K})$ as in (34, 28.24). These are Banach spaces with isometric involution (34, 28.25), (10). Also $\mathcal{E}_0(\hat{K})$ is a C^*-algebra (34, 28.26). For each $\mu \in M(K)$, define $\hat{\mu} \in \mathcal{E}_\infty(\hat{K})$ by $\hat{\mu}(\pi) = \bar{\pi}(\mu)$, then $\mu \mapsto \hat{\mu}$ is a norm-decreasing $*$-isomorphism of $M(K)$ into $\mathcal{E}_\infty(\hat{K})$. Similarly one can define a norm-decreasing $*$-isomorphism $f \mapsto \hat{f}$ of $L^1(K)$ onto a dense subalgebra of $\mathcal{E}_0(\hat{K})$ (74, 3.2, 3.3). Also there is an isometric isomorphism $g \mapsto \hat{g}$ of $L^2(K)$ onto $\mathcal{E}_2(\hat{K})$. Each $g \in L^2(K)$ has a Fourier expansion

$$g = \sum_{\pi \in \hat{K}} \sum_{i,j=1}^{d_\pi} k_\pi \langle \hat{g}(\pi)\xi_i^\pi, \xi_j^\pi \rangle \pi_{i,j},$$

where the series converges in L^2-norm (74, 3.4).
For $\mu \in M(K)$ and $\pi \in \hat{K}$, we set $a_\pi = \bar{\pi}(\mu)^*$, and write

$$\mu \approx \sum_{\pi \in \hat{K}} k_\pi tr(a_\pi \pi).$$

If $\mu = fdm$, where $f \in L^1(K)$, then we write

$$f \approx \sum_{\pi \in \hat{K}} k_\pi tr(a_\pi \pi).$$

If moreover $\sum_{\pi \in \hat{K}} k_\pi \|a_\pi\|_1 < \infty$, we write $f \in A(K)$ and put

$$\|f\|_A = \sum_{\pi \in \hat{K}} k_\pi \|\hat{f}(\pi)\|_1.$$

$A(K)$ is a Banach space with respect to this norm, and $f \mapsto \hat{f}$ is an isometric isomorphism of $A(K)$ onto $\mathcal{E}_1(\hat{K})$. Also for each $f \in A(K)$ with $f \simeq \sum_{\pi \in \hat{K}} k_\pi tr(a_\pi \pi)$ we have

$$f(x) = \sum_{\pi \in \hat{K}} k_\pi tr(a_\pi \pi(x)),$$

m-a.e. (74, 4.2). If moreover f is positive definite, we have

$$f(e) = \|f\|_u := \sum_{\pi \in \hat{K}} k_\pi tr(\hat{f}(\pi)),$$

where the series converges absolutely (74, 4.4). If we denote the set of all continuous positive definite functions on K by $P(K)$, then $f \in P(K)$ if and only if $f \in A(K)$ and each operator $\hat{f}(\pi)$ is positive definite (74, 4.6) and $A(K) = span(P(K)) = L^2(K) * L^2(K)$ (74, 4.8, 4.9).

6.2 Commutative hypergroups

6.2.1 Fourier transform of unbounded measures

There is a fairly successful theory of Fourier transform on commutative hypergroups (10) which goes quite parallel to its group counterpart (except that there is no Pontrjagin duality for commutative hypergroups in general). The dual object $\hat{K}$ is not a hypergroup for some commutative hypergroups K, and even when $\hat{K}$ is a commutative hypergroup, its dual object is not necessarily K (10, Section 2.4). We do not review the standard results on the Fourier transform on commutative hypergroups and refer the interested reader to (10). Instead, here we develop a theory of Fourier transform for unbounded measures on commutative hypergroups. This is analogous to the Argabright-Lamadrid theory on abelian groups (7).
Let K be a locally compact hypergroup (a convo in the sense of (36)). We denote the spaces of bounded continuous functions and continuous functions of compact support on K by $C_b(K)$

and $C_c(K)$, respectively. The latter is an inductive limit of Banach spaces $C_A(K)$ consisting of functions with support in A, where A runs over all compact subsets of K. This in particular implies that if X is a Banach space, a linear transformation $T : C_c(K) \to X$ is continuous if and only if it is locally bounded, that is for each compact subset $A \subseteq K$ there is $\beta = \beta_A > 0$ such that

$$\|T(f)\| \leq \beta \|f\|_\infty \quad (f \in C_A(K)).$$

Also a version of closed graph theorem is valid for $C_c(K)$, namely T is locally bounded if and only if it has a closed graph (12). Throughout $C_c(K)$ is considered with the inductive limit topology (*ind*). Following (12) we have the following definitions.

Definition 6.1. A measure on K is an element of $C_c(K)^*$. The space of all measures on K is denoted by $M(K)$. The subspaces of bounded and compactly supported measures are denoted by $M_b(K)$ and $M_c(K)$, respectively.

Definition 6.2. For $\mu, \nu \in M(K)$, we say that μ is convolvable with ν if for each $f \in C_c(K)$, the map $(x, y) \mapsto f(x * y)$ is integrable over $K \times K$ with respect to the product measure $|\mu| \times |\nu|$. In this case $\mu * \nu$ is defined by

$$\int_K f d(\mu * \nu) = \int_K \int_K f(x * y) d\mu(x) d\nu(y) = \int_K \int_K \int_K f(t) d(\delta_x * \delta_y)(t) d\mu(x) d\nu(y).$$

Let $C(\nu)$ denote the set of all measures convolvable with ν. When K is a measured hypergroup with a left Haar measure m, a locally integrable function f is called convolvable with ν if $fm \in C(\nu)$. In this case, we put $f * \nu = fm * \nu$. Similarly if $\nu \in C(fm)$ then $\nu * f = \nu * fm$. The next two lemmas are a straightforward calculation and we omit the proof. Here Δ denotes the modular function of K. Also for a function f on K, we define $\bar{f}$ and $\tilde{f}$ by

$$\bar{f}(x) = f(\bar{x}), \; \tilde{f}(x) = \overline{f(\bar{x})} \quad (x \in K).$$

We denote the complex conjugate of f by f^-. For $\mu \in M(K)$, μ^-, $\bar{\mu}$ and $\tilde{\mu}$ are defined similarly.

Lemma 6.3. *If K is a measured hypergroup, $\mu \in M(K)$, and f is locally integrable on K, then*
(i) $(\mu * f)(x) = \int_K f(\bar{y} * x) d\mu(y) = \int_K f d(\bar{\mu} * \delta_x)$,
(ii) $(f * \mu)(x) = \int_K f(x * \bar{y}) \bar{\Delta}(y) d\mu(y) = \int_K f d(\delta_x * \Delta\bar{\mu})$,
locally almost everywhere. If $f \in C_c(K)$, the above formulas are valid everywhere and define continuous (not necessarily bounded) functions on K.

If we want to avoid the modular function in the second formula, we should define the convolution of functions and measures differently, but then this definition does not completely match with the formula for convolution of measures (see (10)).

Lemma 6.4. *If K is a measured hypergroup, $\mu, \nu \in M(K)$, $f \in C_c(K)$, and $\mu \in C(\nu)$, then $\bar{\mu} \in C(f)$, $\bar{\mu} * f \in L^1(\nu)$, and*

$$\int_K f d\mu * \nu = \int_K \bar{\mu} * f d\nu.$$

Definition 6.5. A measure $\mu \in M(K)$ is called left translation bounded if for each compact subset $A \subseteq K$

$$\ell_\mu(A) := \sup_{x \in K} |\bar{\mu} * \delta_x|(A) < \infty.$$

Similarly μ is called right translation bounded if for each compact subset $A \subseteq K$

$$r_\mu(A) := \sup_{x \in K} |\delta_x * \Delta\bar{\mu}|(A) < \infty.$$

We denote the set of left and right translation bounded measures on K by $M_{\ell b}(K)$ and $M_{rb}(K)$, respectively.

Proposition 6.6. *Let $\mu \in M(K)$, then*
(i) μ is left translation bounded if and only if $\mu * f \in C_b(K)$, for each $f \in C_c(K)$. In this case for each compact subset $A \subseteq K$,

$$\|\mu * f\|_\infty \leq \ell_\mu(A)\|f\|_\infty \quad (f \in C_A(K)).$$

(ii) μ is right translation bounded if and only if $f * \mu \in C_b(K)$, for each $f \in C_c(K)$. In this case for each compact subset $A \subseteq K$,

$$\|f * \mu\|_\infty \leq r_\mu(A)\|f\|_\infty \quad (f \in C_A(K)).$$

Proof. We prove (i), (ii) is proved similarly. If $f \in C_A(K)$, then by Lemma 6.3,

$$|(\mu * f)(x)| \leq \|f\|_\infty |\bar{\mu} * \delta_x|(A) \leq \|f\|_\infty \ell_\mu(A),$$

for each $x \in K$. Conversely, if $\mu * f$ is bounded for each $f \in C_c(K)$, then $f \mapsto \mu * f$ defines a linear map from $C_c(K)$ into $C_b(K)$. We claim that it has a closed graph. If $f_\alpha \to 0$ in (ilt), then there is a compact subset $A \subseteq K$ such that eventually $supp(f_\alpha) \subseteq A$ and $f_\alpha \to 0$, uniformly on A. If $\mu * f_\alpha \to g$, uniformly on K, then $|(\mu * f_\alpha)(x)| \leq \|f\|_\infty |\bar{\mu} * \delta_x|(A) \to 0$, for each $x \in K$. Hence $g = 0$. By closed graph theorem, for each compact subset $B \subseteq K$, there is $k_B > 0$ such that

$$\|\mu * f\|_\infty \leq k_B \|f\|_\infty \quad (f \in C_B(K)).$$

Now let $A \subseteq K$ be compact and choose $B \subseteq K$ compact with $B^\circ \supseteq A$. By Lemmas 6.3 and 6.4, for each $x \in K$,

$$\begin{aligned} |\bar{\mu} * \delta_x|(A) &= \sup\{|\int_K g d(\bar{\mu} * \delta_x)| : \|g\|_\infty \leq 1, supp(g) \subseteq A\} \\ &\leq \sup\{|\int_K g d(\bar{\mu} * \delta_x)| : \|g\|_\infty \leq 1, supp(g) \subseteq B\} \\ &= \sup\{|\mu * g(x)| : \|g\|_\infty \leq 1, supp(g) \subseteq B\} \leq k_B. \quad \square \end{aligned}$$

Example 6.7. All elements of $M_b(K)$ and $L^p(K, m)$ are translation bounded. Also each hypergroup has a left-translation invariant measure n (36, 4.3C). We have

$$(n * f)(x) = \int_K (\delta_x * \bar{f}) dn \leq \int_K \bar{f} dn,$$

for each $x \in K$ and $f \in C_c^+(K)$. Hence $n \in M_{\ell b}(K)$. Similarly $\Delta n \in M_{rb}(K)$.

Lemma 6.8. *(i) If $\theta \in M_c(K)$ then $\theta \in C(\mu)$ and $\mu \in C(\theta)$, for each $\mu \in M(K)$.*
(ii) If $\nu \in M(K)$, then
$a)\nu \in M_{\ell b}(K)$ if and only if $\mu \in C(\nu)$, for each $\mu \in M_b(K)$.
$b)\nu \in M_{rb}(K)$ if and only if $\nu \in C(\mu)$, for each $\mu \in M_b(K)$.

Proof. (i) follows from the fact that $|\bar{\theta}| * |f| \in C_c(K)$ for each $f \in C_c(K)$ (36, 4.2F). If $f \in C_c(K)$, $\nu \in M_{\ell b}(K)$ and $\mu \in M_b(K)$, then $|\nu| * |\bar{f}| \in C_b(K)$, hence

$$\int_K \int_K |f(x * y)| d|\mu|(x) d|\nu|(y) = \int_K (|\nu| * |\bar{f}|) d|\mu| < \infty,$$

so $\mu \in C(\nu)$. Conversely, if $\nu \notin M_{\ell b}(K)$ then there is $f \in C_c^+(K)$ such that $\nu * f$ is unbounded. Hence there is $\mu \in M_b^+(K)$ with $\nu * f \notin L^1(K)$, i.e.

$$\int_K \bar{f}(x * y) d\bar{m}(x) d\nu(y) = \int_K (\nu * f) d\mu = \infty,$$

that is $\bar{\mu} \notin C(\nu)$. This proves (iia). (iib) is similar. □

Corollary 6.9. *If $\nu \in M_{\ell b}(K)$ then for each $\mu \in M_b(K)$ and $f \in C_c(K)$ we have $\mu * f \in L^1(K, \nu)$ and*

$$\int_K (\mu * f) d\nu = \int_K (\nu * \bar{f}) d\mu.$$

Following (12, chap. 8), we have the following associativity result which follows from the above lemma and a straightforward application of Fubini's Theorem.

Theorem 6.10. *(i) If $\mu, \nu \in M(K)$, $\mu \in C(\nu)$, and $\theta \in M_c(K)$, then $\theta \in C(\mu) \cap C(\mu * \nu)$ and $\theta * \mu \in C(\nu)$, and*

$$(\theta * \mu) * \nu = \theta * (\mu * \nu).$$

(ii) If $\nu \in M_{\ell b}(K)$, $\mu_1, \mu_2 \in M_b(K)$ then $\mu_1 * (\mu_2 * \nu) = (\mu_1 * \mu_2) * \nu$.

6.2.2 Transformable measures

In this section we assume that K is a commutative hypergroup such that $\hat{K}$ is a hypergroup, namely K is strong (10, 2.4.1). We don't assume that $(\hat{K})\hat{} = K$, unless otherwise specified. We denote the Haar measure on K and $\hat{K}$ by m_K and $m_{\hat{K}}$, respectively. For $\mu \in M_b(K)$, $\hat{\mu} \in C_b(\hat{K})$ is defined by

$$\hat{\mu}(\gamma) = \int_K \overline{\gamma(x)} d\mu(x) \quad (\gamma \in \hat{K}).$$

We usually identify $\hat{\mu}$ with $\hat{\mu} m_{\hat{K}}$. Put $\check{\mu} = (\hat{\mu})\check{}$.

Definition 6.11. (10, 2.3.10) A measure $\mu \in M(K)$ is called transformable if there is $\hat{\mu} \in M^+(\hat{K})$ such that

$$\int_K (f * \tilde{f}) d\mu = \int_{\hat{K}} |\check{f}| d\hat{\mu} \quad (f \in C_c(K)).$$

We denote the space of transformable measures on K by $M_t(K)$ and put $\hat{M}_t(\hat{K}) = \{\hat{\mu} : \mu \in M_t(K)\}$. Also we put $C_2(K) = span\{f * \tilde{f} : f \in C_c(K)\}$. Then above relation could be rewritten as

$$\int_K g d\mu = \int_{\hat{K}} \check{g} d\hat{\mu} \quad (g \in C_2(K)),$$

or

$$\int_K f * g d\mu = \int_{\hat{K}} \check{f} \check{g} d\hat{\mu} \quad (f, g \in C_c(K)).$$

Example 6.12. The Haar measure m_K is transformable and $\hat{m}_K = \pi_K$ is the Levitan-Plancherel measure on $\hat{K}$ (10, 2.2.13). Also all bounded measures are transformable. If $\mu \in M_b(K)$ and $\hat{\mu} \geq 0$ on $supp(\pi_K)$, then the transform of μ in the above sense is $\hat{\mu}\pi_K$, where $\hat{\mu}$ is the Fourier-Stieltjes transform of μ. In particular, $\hat{\delta}_e = \pi_K$ (10). Finally, if f is a bounded positive definite function on K, then by Bochner's Theorem (10, 4.1.16), there is a unique $\sigma \in M_b^+(K)$ such that $f = \check{\sigma}$. Then it is easy to see that $f m_K \in M_t(K)$ and $(f m_K)\hat{} = \tilde{\sigma}$.

Lemma 6.13. *$C_2(K)$ is dense in $C_c(K)$.*

Proof. Let $\{u_\alpha\}$ be a bounded approximate identity of $L^1(K, m_K)$ consisting of elements of $C_c(K)$ with $u_\alpha \geq 0$, $u_\alpha = \bar{u}_\alpha$, and $\int_K u_\alpha dm_K = 1$ such that $supp(u_\alpha)$ is contained in a compact neighborhood V of e for each α (29). For $f \in C_c(K)$ with $supp(f) = W$, we have $supp(f), supp(f * u_\alpha) \subseteq W * V =: U$. By uniform continuity of f, given $\varepsilon > 0$, there is α_0 such that $|f(x * y) - f(x)| < \varepsilon$, for each $x \in U$ and $y \in supp(u_\alpha)$ with $\alpha \geq \alpha_0$.

$$|(f * u_\alpha - f)(x)| \leq \int_K |f(x*y) - f(x)||u_\alpha(y)|dm_K(y) = \int_V \varepsilon u_\alpha dm_K = \varepsilon,$$

for $\alpha \geq \alpha_0$. □

By the above lemma and an argument similar to (7, Thm. 2.1) we have:

Theorem 6.14. *(Uniqueness Theorem) If $\mu \in M_t(K)$ then μ and $\hat{\mu}$ determine each other uniquely and the map $\mu \mapsto \hat{\mu}$ is an isomorphism of $M_t(K)$ onto $\hat{M}_t(\hat{K})$.*

The proof of the next result is straightforward.

Lemma 6.15. *For each $\mu \in M_t(K)$, the following measure are transformable with the given transform.*
(i) $\hat{\bar{\mu}} = \bar{\hat{\mu}}$,
(ii) $(\tilde{\mu})\hat{} = (\hat{\mu})^-$,
(iii) $(\mu^-)\hat{} = (\hat{\mu})\tilde{}$,
(iv) $(\delta_x * \mu)\hat{} = \bar{x}\hat{\mu} \quad (x \in K)$,
(v) $(\gamma\mu)\hat{} = \delta_{\tilde{\gamma}} * \hat{\mu} \quad (\gamma \in \hat{K})$.

If $\mu \in M_t(K)$, then by definition we have

$$\hat{C}_2(\hat{K}) := \{\hat{g} : g \in C_2(K)\} \subseteq L^2(\hat{K}, \hat{\mu}).$$

Next lemma is proved with the same argument as in (7, Prop. 2.2). We bring the proof for the sake of completeness.

Lemma 6.16. *$\hat{C}_2(\hat{K}) \subseteq L^2(\hat{K}, \hat{\mu})$ is dense.*

Proof. Since $C_c(\hat{K}) \subseteq L^2(\hat{K}, \hat{\mu})$ is dense, we need to show that given $\varphi \in C_c(\hat{K})$ and $\varepsilon > 0$, there is $g \in C_2(K)$ such that $\int_{\hat{K}} |\varphi - \hat{g}|^2 d\hat{\mu} < \varepsilon^2$.
Put $A = supp(\varphi)$. If $|\hat{\mu}|(A) = 0$, we take $g = 0$. Assume that $|\hat{\mu}|(A) > 0$. By (10, 2.2.4(iv)), $\hat{C}_c(\hat{K})$ is dense in $C_0(\hat{K})$, so there is $f \in C_c(K)$ such that

$$|\hat{f}(\gamma) - 1| < \varepsilon \|\varphi\|_\infty^{-1} (|\hat{\mu}|(A))^{\frac{-1}{2}} =: \delta \quad (\gamma \in A).$$

We may assume that $\delta < \frac{1}{2}$, so $|\hat{f}| > \frac{1}{2}$, and so $\int_{\hat{K}} |\hat{f}|^2 d\hat{\mu} \neq 0$. Choose $h \in C_c(K)$ such that $\|\hat{h} - \varphi\|_\infty < \varepsilon (\int_{\hat{K}} |\hat{f}|^2 d\hat{\mu})^{\frac{-1}{2}}$. Put $g = h * f \in C_2(K)$, then

$$\begin{aligned}(\int_{\hat{K}} |\hat{g} - \varphi|^2 d\hat{\mu})^{\frac{1}{2}} &= (\int_{\hat{K}} |\hat{h}\hat{f} - \varphi|^2 d\hat{\mu})^{\frac{1}{2}} \\ &\leq (\int_{\hat{K}} |\hat{h}\hat{f} - \varphi\hat{f}|^2 d\hat{\mu})^{\frac{1}{2}} + (\int_{\hat{K}} |\varphi\hat{f} - \varphi|^2 d\hat{\mu})^{\frac{1}{2}} < 2\varepsilon. \quad \square\end{aligned}$$

Lemma 6.17. *If $\mu \in M_t(K)$ and $g \in C_2(K)$, then $\hat{g} \in L^1(\hat{K}, \hat{\mu})$ and $g * \mu = (\hat{g}\hat{\mu})\check{}$.*

Proof. The first assertion follows from definition of $\hat{\mu}$. For the second, since K is unimodular, given $x \in K$,

$$(g * \mu)(x) = \int_K g d(\delta_x * \bar{\mu}) = \int_{\hat{K}} \check{g}(\gamma)\gamma(\bar{x}) d\hat{\mu}$$
$$= \int_{\hat{K}} \hat{g}(\gamma)\gamma(x) d\hat{\mu} = (\hat{g}\hat{\mu})\check{}. \quad \square$$

Now we are ready to prove the main result of this section.

Theorem 6.18. *For $\mu \in M(K)$, the following are equivalent:*
(i) $\mu \in M_t(K)$,
(ii) $g * \mu \in B(K)$, for each $g \in C_2(K)$.

Proof. If $g \in C_2(K)$, then $\hat{g} \in L^1(\hat{K}, \hat{\mu})$, hence $\hat{g}\hat{\mu} \in M_b(\hat{K})$. Therefore, by the above lemma and (10, 4.1.15), $g * \mu = (\hat{g}\hat{\mu})\check{} \in B(K)$. Conversely if $g * \mu \in B(K)$, for each $g \in C_2(K)$, then by Bochner's Theorem, there is a unique $\nu_g \in M_b(\hat{K})$ such that

$$g * \mu(x) = \int_{\hat{K}} \gamma(x) d\nu_g(\gamma) \quad (x \in K).$$

For each $f \in L^1(K, m)$,

$$\int_K f(g * \mu) dm = \int_K \int_{\hat{K}} f(x)\gamma(x) d\nu_g(\gamma) dm(x) = \int_{\hat{K}} \check{f} d\nu_g.$$

Fix $g, h \in C_2(K)$, then for each $f \in C_c(K)$,

$$\int_K (f * \bar{g})(h * \mu) dm = \int_K (f * \bar{g} * \bar{h}) d\mu = \int_K (f * \bar{h} * \bar{g}) d\mu = \int_K (f * \bar{h})(g * \mu) dm.$$

Hence

$$\int_{\hat{K}} \check{f}\hat{g} d\nu_h = \int_{\hat{K}} \check{f}\hat{h} d\nu_g.$$

By density of $\hat{C}_c(\hat{K})$ in $C_0(\hat{K})$, we get $\hat{g} d\nu_h = \hat{h} d\nu_g$. Define $\hat{\mu}$ on $\hat{K}$ by

$$\int_{\hat{K}} \psi d\hat{\mu} = \int_{\hat{K}} \frac{\psi}{\hat{h}} d\hat{\nu}_h,$$

where $h \in C_2(K)$ is such that $\hat{h} > 0$ on $supp(\psi)$. This is a well defined linear functional by what we just observed. It is easy to see that this is locally bounded on $C_c(\hat{K})$. Also

$$\int_{\hat{K}} \psi \hat{g} d\hat{\mu} = \int_{\hat{K}} \frac{\psi \hat{g}}{\hat{h}} d\nu_h = \int_{\hat{K}} \psi d\nu_g,$$

that is $\hat{g} d\hat{\mu} = d\nu_g$, and so $\hat{g} \in L^1(\hat{K}, \hat{\mu})$ and

$$\int_K g d\mu = (\bar{g} * \mu)(e) = \int_{\hat{K}} d\nu_{\bar{g}} = \int_{\hat{K}} \check{g} d\hat{\mu}.$$

Finally, since $\hat{h} > 0$ on $supp(\psi)$, we have $\int_{\hat{K}} \psi d\hat{\mu} \geq 0$, for $\psi \geq 0$. These all together show that $\mu \in M^+(\hat{K})$ is the transform of μ and we are done. $\square$

In (5) the authors introduced the concept of tensor hypergroups and showed that for a tensor hypergroup K, the Fourier space $A(K)$ is a Banach algebra. The following lemma follows from Plancherel Theorem exactly as in the group case (26, 3.6.2°).

Lemma 6.19. *If K is a commutative, strong, tensor hypergroup, then $A(K)$ is isometrically isomorphic (through Fourier transform) to $L^1(\hat{K})$.*

The last result of this paper is a direct consequence of the above lemma and Theorem 6.18 (see the proof of (7, Theorem 2.4)).

Corollary 6.20. *If K is a commutative, strong, tensor hypergroup, then for each $\mu \in M(K)$, the following are equivalent:*
(i) $\mu \in M_t(K)$ and $\hat{\mu}$ is a function,
(ii) $g * \mu \in A(K)$, for each $g \in C_2(K)$.

6.3 Finite hypergroups

Peter Shor in his seminal paper presented efficient quantum algorithms for computing integer factorizations and discrete logarithms. These algorithms are based on an efficient solution to the hidden subgroup problem (HSP) for certain abelian groups. HSP was already appeared in Simon's algorithm implicitly in form of distinguishing the trivial subgroup from a subgroup of order 2 of $\mathbb{Z}_{2^n}$.
The efficient algorithm for the abelian HSP uses the Fourier transform. Other methods have been applied by Mosca and Ekert (52). The fastest currently known (quantum) algorithm for computing the Fourier transform over abelian groups was given by Hales and Hallgren (32). Kitaev (39) has shown us how to efficiently compute the Fourier transform over any abelian group (see also (37)).
For general groups, Ettinger, Hoyer and Knill (25) have shown that the HSP has polynomial query complexity, giving an algorithm that makes an exponential number of measurements. Several specific non-abelian HSP have been studied by Ettinger and Hoyer (24), Rotteler and Beth (69), and Puschel, Rotteler, and Beth (61). Ivanyos, Mangniez, and Santha (37) have shown how to reduce certain non-abelian HSP's to an abelian HSP. The non-abelian HSP for normal subgroups is solved by Hallgren, Russell, and Ta-Shma (33).
As for the Graph Isomorphism Problem (GIP), which is a special case of HSP for the symmetric group S_n, Grigni, Schulman, Vazirani and Vazirani (31) have shown that measuring representations is not enough for solving GIP. However, they show that the problem can be solved when the intersection of the normalizers of all subgroups of G is large. Similar negative results are obtained by Ettinger and Hoyer (24). At the positive side, Beals (9) showed how to efficiently compute the Fourier transform over the symmetric group S_n (see also (40)).

6.3.1 Hidden subgroup problem

Definition 6.21. (Hidden Subgroup Problem (HSP)). Given an efficiently computable function $f : G \to S$, from a finite group G to a finite set S, that is constant on (left) cosets of some subgroup H and takes distinct values on distinct cosets, determine the subgroup H.

An efficient quantum algorithms for abelian groups is given in the next page. Note that the resulting distribution over ρ is independent of the coset cH arising after the first stage. Thus, repetitions of this experiment result in the same distribution over $\hat{G}$. Also by the principle of delayed measurement, measuring the second register in the first step can in fact be delayed until the end of the experiment.
In the next algorithm, one wishes the resulting distribution to be independent of the actual coset cH and depend only on the subgroup H. This is guaranteed by measuring only the name of the representation ρ and leaving the matrix indices unobserved. The fact that $O(log(|G|))$ samples of this distribution are enough to determine H with high probability is proved in (33).

1. Prepare the state

$$\frac{1}{\sqrt{|G|}}\sum_{g\in G}|g\rangle|f(g)\rangle$$

and measure the second register, the resulting state is

$$\frac{1}{\sqrt{|H|}}\sum_{h\in H}|ch\rangle|f(ch)\rangle$$

where c is an element of G selected uniformly at random.
2. Compute the Fourier transform of the "coset" state above, resulting in

$$\frac{1}{\sqrt{|H|.|G|}}\sum_{\rho\in\hat{G}}\sum_{h\in H}\rho(ch)|\rho\rangle|f(ch)\rangle$$

where $\hat{G}$ denotes the Pontryagin dual of G, namely the set of homomorphisms $\rho : G \rightarrow \mathbb{T}$.
3. Measure the first register and observe a homomorphism ρ.

Algorithm 1. *Algorithm for Abelian Hidden Subgroup Problem*

1. Prepare the state $\sum_{g\in G}|g\rangle|f(g)\rangle$ and measure the second register $|f(g)\rangle$. The resulting state is $\sum_{h\in H}|ch\rangle|f(ch)\rangle$ where c is an element of G selected uniformly at random. As above, this state is supported on a left coset cH of H.
2. Let $\hat{G}$ denote the set of irreducible representations of G and, for each $\rho\in\hat{G}$, fix a basis for the space on which ρ acts. Let d_ρ denote the dimension of ρ. Compute the Fourier transform of the coset state, resulting in

$$\sum_{\rho\in\hat{G}}\sum_{1\leq i,j\leq d_\rho}\frac{\sqrt{d_\rho}}{\sqrt{|H|.|G|}}\sum_{h\in H}\rho(ch)|\rho,i,j\rangle|f(ch)\rangle$$

3. Measure the first register and observe a representation ρ.

Algorithm 2. *Algorithm for Non-abelian Hidden Normal Subgroup Problem*

A finite hypergroup is a set $K = \{c_0, c_1, \ldots, c_n\}$ together with an $*$-algebra structure on the complex vector space $\mathbb{C}K$ spanned by K which satisfies the following axioms. The product of elements is given by the structure equations

$$c_i * c_j = \sum_k n^k_{i,j} c_k,$$

with the convention that summations always range over $\{0, 1, \ldots, n\}$. The axioms are

1. $n^k_{i,j}\in\mathbb{R}$ and $n^k_{i,j}\geq 0$,
2. $\sum_k n^k_{i,j} = 1$,
3. $c_0 * c_i = c_i * c_0 = c_i$,
4. $K^* = K$, $n^0_{i,j}\neq 0$ if and only if $c^*_i = c_j$,

for each $0 \leq i, j, k \leq n$.

If $c_i^* = c_i$, for each i, then the hypergroup is called hermitian. If $c_i * c_j = c_j * c_i$, for each i, j, then the hypergroup is called commutative. Hermitian hypergroups are automatically commutative.

In harmonic analysis terminology, we have a convolution structure on the measure algebra $M(K)$. This means that we can convolve finitely additive measures on K and, for $x, y \in K$, the convolution $\delta_x * \delta_y$ is a probability measure. Indeed $\delta_{c_i} * \delta_{c_j}\{c_k\} = n_{i,j}^k$. We follow the convention of harmonic analysis texts and denote the involution by $x \mapsto \bar{x}$ (instead of x^*), and the identity element by e (instead of c_0). For a function $f : K \to \mathbb{C}$, and sets $A, B \subseteq K$ we put

$$f(x * y) = \sum_{z \in K} f(z)(\delta_x * \delta_y)\{z\}, \quad (x, y \in K),$$

and

$$A * B = \cup\{supp(\delta_x * \delta_y) : x \in A, y \in B\}.$$

A finite hypergroup K always has a left Haar measure (positive, left translation invariant, finitely additive measure) $\omega = \omega_K$ given by

$$\omega\{x\} = \big((\delta_{\bar{x}} * \delta_x)\{e\}\big)^{-1} \quad (x \in K).$$

A function $\rho : K \to \mathbb{C}$ is called a character if $\rho(e) = 1, \rho(x * y) = \rho(x)\rho(y)$, and $\rho(\bar{x}) = \overline{\rho(x)}$. In contrast with the group case, characters are not necessarily constant on conjugacy classes. Let K be a finite commutative hypergroup, then $\hat{K}$ denotes the set of characters on K. In this case, for $\mu \in M(K)$ and $f \in \ell^2(K)$, we put

$$\hat{\mu}(\rho) = \sum_{x \in K} \rho(x)\mu\{x\}, \quad \hat{f}(\rho) = \sum_{x \in K} f(x)\rho(x)\omega\{x\} \quad (\rho \in \hat{K}).$$

Hence $\hat{f} = (f\omega)\hat{}$.

6.3.2 Subhypergroups

If $H \subseteq K$ is a subhypergroup (i.e. $\bar{H} = H$ and $H * H \subseteq H$), then $\hat{\omega}_H = \chi_{H^\perp}$ (10, 2.1.8), where the right hand side is the indicator (characteristic) function of

$$H^\perp = \{\rho \in \hat{K} : \rho(x) = 1 \ (x \in H)\}.$$

If K/H is the coset hypergroup (which is the same as the double coset hypergroup $K//H$ in finite case (10, 1.5.7)) with hypergroup epimorphism (quotient map) $q : K \to H/K$ (10, 1.5.22), then $(K/H)\hat{} \simeq H^\perp$ (with isomorphism map $\chi \mapsto \chi \circ q$) (10, 2.2.26, 2.4.8). Moreover, for each $\mu \in M(K)$, $q(\mu * \omega_H) = q(\mu)$ (10, 1.5.12). We say that K is strong if $\hat{K}$ is a hypergroup with respect to some convolution satisfying

$$(\rho * \sigma)\check{} = \rho\check{}\sigma\check{} \quad (\rho, \sigma \in \hat{K}),$$

where

$$\check{k}(x) = \sum_{\rho \in \hat{K}} k(\rho)\rho(x)\pi\{\rho\} \quad (x \in K, k \in \ell^2(\hat{K}, \pi))$$

is the inverse Fourier transform. In this case, for $\rho, \sigma \in \hat{K}$, we have $\rho \in \sigma * H^\perp$ if and only if $Res_H\rho = Res_H\sigma$, where $Res_H : \hat{K} \to \hat{H}$ is the restriction map (10, 2.4.15). Also H is strong and $\hat{K}/H^\perp \simeq \hat{H}$ (10, 2.4.16). Moreover $(\hat{K})\hat{} \simeq K$ (10, 2.4.18).

6.3.3 Fourier transform

Let us quote the following theorem from (10, 2.2.13) which is the cornerstone of the Fourier analysis on commutative hypergroups.

Theorem 6.22. *(Levitan) If K is a finite commutative hypergroup with Haar measure ω, there is a positive measure π on $\hat{K}$ (called the Plancherel measure) such that*

$$\sum_{x\in K} |f(x)|^2 \omega\{x\} = \sum_{\rho\in\hat{K}} |\hat{f}(\rho)|^2 \pi\{\rho\} \quad (f \in \ell^2(K,\omega)).$$

Moreover $supp(\pi) = \hat{K}$ and $\pi\{\rho\} = \pi\{\bar{\rho}\}$. In particular the Fourier transform $\mathfrak{F}$ is a unitary map from $\ell^2(K,\omega)$ onto $\ell^2(\hat{K},\pi)$.

In quantum computation notation,

$$\mathfrak{F} : |x\rangle \mapsto \frac{1}{\tau(x)} \sum_{\rho\in\hat{K}} \rho(x)\pi\{\rho\}|\rho\rangle,$$

where

$$\tau(x) = \big(\sum_{\rho\in\hat{K}} |\rho(x)|^2 \pi^2\{\rho\}\big)^{\frac{1}{2}} \quad (x \in K).$$

When K is a group, $\tau(x) = |\hat{K}|^{\frac{1}{2}}$, for each $x \in K$. It is essential for quantum computation purposes to associate a unitary matrix to each quantum gate. however, if we write the matrix of $\mathfrak{F}$ naively using the above formula we don't get a unitary matrix. The reason is that, in contrast with the group case, the discrete measures on ℓ^2 spaces are not counting measure. More specifically, when K is a group, $\ell^2(K) = \bigoplus_{x\in K}\mathbb{C}$, where as here $\ell^2(K,\omega) = \bigoplus_{x\in K}\omega\{x\}^{\frac{1}{2}}\mathbb{C}$ and $\ell^2(K) = \bigoplus_{\rho\in\hat{K}}\pi\{\rho\}^{\frac{1}{2}}\mathbb{C}$. The exponent $\frac{1}{2}$ is needed to get the same inner product on both sides. If we use change of bases $|x\rangle' = \omega\{x\}^{\frac{1}{2}}|x\rangle$ and $|\rho\rangle' = \pi\{\rho\}^{\frac{1}{2}}|\rho\rangle$, the Fourier transform can be written as

$$\mathfrak{F} : |x\rangle' \mapsto \omega\{x\}^{\frac{1}{2}} \sum_{\rho\in\hat{K}} \rho(\bar{x})\pi\{\rho\}^{\frac{1}{2}}|\rho\rangle',$$

and the corresponding matrix turns out to be unitary.

6.3.4 Examples

There ar a variety of examples of (commutative hypergroups) whose dual object is known. One might hope to relate the HSP on a (non-abelian) group G to the HSHP on a corresponding commutative hypergroup like $\hat{G}$ (see next example). The main difficulty is to go from a function f which is constant on cosets of some subgroup $H \leq G$ to a function which is constant on cosets of a subhypergroup of $\hat{G}$. The canonical candidate $\hat{f}$ fails to be constant on costs of $H^{\perp} \leq \hat{G}$.

We list some of the examples of commutative hypergroups and their duals, hoping that one can get such a relation in future.

Example 6.23. If G is a finite group, then $\hat{G} := (G^G)\hat{}$ is a commutative strong (and so Pontryagin (10, 2.4.18)) hypergroup (10, 8.1.43). The dual hypergroups of the Dihedral group D_n and the (generalized) Quaternion group Q_n are calculated in (10, 8.1.46,47).

Example 6.24. If G is a finite group and H is a (not necessarily normal) subgroup of G then the double coset space $G//H$ (which is basically the same as the homogeneous space G/H in the finite case) is a hypergroup whose dual object is $A(\hat{G}, H)$ (10, 2.2.46). It is easy to put conditions on H so that $G//H$ is commutative.

There are also a vast class of special hypergroups (see chapter 3 of (10) for details) which are mainly infinite hypergroups, but one might mimic the same constructions to get similar finite hypergroups in some cases.
There are not many finite hypergroups whose character table is known (75). Here we give two classical examples (of order two and three) and compute the corresponding Fourier matrix.

Example 6.25 (Ross)**.** The general form of an hypergroup of order 2 is known. It is denoted by $K = \mathbb{Z}_2(\theta)$ and consists of two elements 0 and 1 with multiplication table

$*$	δ_0	δ_1
δ_0	δ_0	δ_1
δ_1	δ_1	$\theta\delta_0 + (1-\theta)\delta_1$

and Haar measure and character table

	0	1
ω	1	$\frac{1}{\theta}$
χ_0	1	1
χ_1	1	$-\theta$

When $\theta = 1$ we get $K = \mathbb{Z}_2$. The dual hypergroup is again $\mathbb{Z}_2(\theta)$ with the Plancherel measure

	χ_0	χ_1
π	$\frac{\theta}{1+\theta}$	$\frac{1}{1+\theta}$

The unitary matrix of the corresponding Fourier transform is given by

$$\mathfrak{F}_2 = \frac{1}{\sqrt{1+\theta^2}} \begin{pmatrix} \theta & 1 \\ 1 & -\theta \end{pmatrix}$$

Example 6.26 (Wildberger)**.** The general form of hypergroups of order 3 is also known. We know that it is always commutative, but in this case, the Hermitian and non Hermitian case should be treated separately. Let $K = \{0, 1, 2\}$ be a Hermitian hypergroup of order three and put $\omega_i = \omega\{i\}$, for $i = 0, 1, 2$. Then the multiplication table of K is

$*$	δ_0	δ_1	δ_2
δ_0	δ_0	δ_1	δ_2
δ_1	δ_1	$\frac{1}{\omega_1}\delta_0 + \alpha_1\delta_1 + \beta_1\delta_2$	$\gamma_1\delta_1 + \gamma_2\delta_2$
δ_2	δ_2	$\gamma_1\delta_1 + \gamma_2\delta_2$	$\frac{1}{\omega_2}\delta_0 + \beta_2\delta_1 + \alpha_2\delta_2$

where

$$\beta_1 = \frac{\gamma_1\omega_2}{\omega_1},\ \beta_2 = \frac{\gamma_2\omega_1}{\omega_2},\ \alpha_1 = 1 - \frac{1+\gamma_1\omega_2}{\omega_1},\ \alpha_2 = 1 - \frac{1+\gamma_2\omega_1}{\omega_2}\ \gamma_2 = 1-\gamma_1,$$

and γ_1, ω_1 and ω_2 are arbitrary parameters subject to conditions $0 \leq \gamma_1 \leq 1, \omega_1 \geq 1, \omega_2 \geq 1$, and

$$1+\gamma_1\omega_2 \leq \omega_1$$
$$1+(1-\gamma_1)\omega_1 \leq \omega_2.$$

The Plancherel measure and character table are given by

	π	0	1	2
χ_0	$\frac{s_1}{t}$	1	1	1
χ_1	$\frac{s_2}{t}$	1	x	z
χ_2	$\frac{s_3}{t}$	1	y	v

where

$$x = \frac{\alpha_1-\gamma_1}{2} + \frac{D}{2\omega_2}, \quad y = \frac{\alpha_1-\gamma_1}{2} - \frac{D}{2\omega_2}$$
$$z = \frac{\alpha_2-\gamma_2}{2} - \frac{D}{2\omega_2}, \quad v = \frac{\alpha_2-\gamma_2}{2} + \frac{D}{2\omega_2}$$
$$D = \sqrt{(1+\gamma_1\omega_2-\gamma_2\omega_1)^2 + 4\gamma_2\omega_1}$$

and

$$s_1 = x^2v^2 + \frac{y^2}{\omega_2} + \frac{z^2}{\omega_1} - (y^2z^2 + \frac{x^2}{\omega_2} + \frac{v^2}{\omega_1})$$
$$s_2 = y^2 + \frac{v^2}{\omega_1} + \frac{1}{\omega_2} - (v^2 + \frac{y^2}{\omega_2} + \frac{1}{\omega_1})$$
$$s_3 = z^2 + \frac{x^2}{\omega_2} + \frac{1}{\omega_1} - (x^2 + \frac{z^2}{\omega_1} + \frac{1}{\omega_1})$$
$$t = x^2v^2 + y^2 + z^2 - (x^2 + y^2z^2 + v^2).$$

Let $\pi_i = \pi\{\chi_i\} = \frac{s_i}{t}$ and $w_{ij} = \sqrt{\omega_i\pi_j}$, for $i,j = 0,1,2$, then the Fourier transform is given by the unitary matrix

$$\mathfrak{F}_3 = \begin{pmatrix} w_{00} & w_{10} & w_{20} \\ w_{01} & xw_{11} & zw_{21} \\ w_{02} & yw_{12} & vw_{22} \end{pmatrix}$$

One concrete example is the normalized Bose Mesner algebra of the square. In this case, $\omega_1 = 1, \omega_2 = 2, \gamma_1 = \beta_1 = \alpha_1 = \alpha_2 = 0, \gamma_2 = 1$, and $\beta_2 = \frac{1}{2}$. A simple calculation gives $D = 2, x = 1, y = z = -1, v = 0$, and if we put $\pi_1 = \frac{1}{4}$, we get $\pi_2 = \frac{1}{4}$ and $\pi_3 = \frac{1}{2}$. In this case, the Fourier transform matrix is

$$\mathfrak{F}_3 = \frac{1}{2}\begin{pmatrix} 1 & 1 & \sqrt{2} \\ 1 & 1 & -\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & 0 \end{pmatrix}$$

In the non-Hermitian case, the multiplication table of K is

$*$	δ_0	δ_1	δ_2
δ_0	δ_0	δ_1	δ_2
δ_1	δ_1	$\gamma\delta_1 + (1-\gamma)\delta_2$	$\alpha\delta_0 + \gamma\delta_1 + \gamma\delta_2$
δ_2	δ_2	$\alpha\delta_0 + \gamma\delta_1 + \gamma\delta_2$	$(1-\gamma)\delta_1 + \gamma\delta_2$

where $\gamma = \frac{1-\alpha}{2}$, and α is an arbitrary parameter with $0 < \alpha \leq 1$. When $\alpha = 1$, we get $K = \mathbb{Z}_3$. The dual hypergroup is again K and the Plancherel measure and character table are given by

	π	0	1	2
χ_0	$\frac{s_1}{t}$	1	1	1
χ_1	$\frac{s_2}{t}$	1	z	$\bar{z}$
χ_2	$\frac{s_2}{t}$	1	$\bar{z}$	z

where

$$z = \frac{-\alpha \pm i\sqrt{\alpha^2 + 2\alpha}}{2}.$$

$$s_1 = 2 - \omega_1(\alpha^2 + \alpha), \quad s_2 = \omega_1 - 1, \quad t = \omega_1(2 - \alpha^2 - \alpha).$$

Put $\pi_i = \pi\{\chi_i\}$ and $w_{ij} = \sqrt{\omega_i \pi_j}$, for $i, j = 0, 1, 2$, then the Fourier transform is given by the unitary matrix

$$\mathfrak{F}_3 = \begin{pmatrix} w_{00} & w_{10} & w_{20} \\ w_{01} & zw_{11} & \bar{z}w_{21} \\ w_{02} & \bar{z}w_{12} & zw_{22} \end{pmatrix}$$

As a concrete example, let us put $\omega_1 = \omega_2 = 2, \gamma = \frac{1}{4}$ and $\alpha = \frac{1}{2}$ to get $z = \frac{-1+i\sqrt{5}}{4}$ and $\pi_1 = \frac{1}{5}, \pi_2 = \pi_3 = \frac{2}{5}$. In this case, the Fourier transform matrix is

$$\mathfrak{F}_3 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & \frac{-1+i\sqrt{5}}{4} & \frac{-1-i\sqrt{5}}{4} \\ \sqrt{2} & \frac{-1-i\sqrt{5}}{4} & \frac{-1+i\sqrt{5}}{4} \end{pmatrix}$$

6.3.5 Hidden subhypergroup problem

In this section we give an algorithm for solving hidden sub-hypergroup problem (HSHP) for abelian (strong) hypergroups. This algorithm is efficient for those finite commutative hypergroups whose Fourier transform is efficiently calculated. It is desirable that, following Kitaev (39), one shows that the Fourier transform could be efficiently calculated on each finite commutative hypergroup. This could be difficult, as there is yet no complete structure theory for finite commutative hypergroups (see chapter 8 of (10)).

Definition 6.27. (Hidden Sub-hypergroup Problem (HSHP)). Given an efficiently computable function $f : K \to S$, from a finite hypergroup K to a finite set S, that is constant on (left) cosets of some subhypergroup H and takes distinct values λ_c on distinct cosets $c * H$, for $c \in K$. Determine the subhypergroup H.

Lemma 6.28. *Let K be commutative and H be a sub-hypergroup of K and $\rho \in \hat{K}$, then the following are equivalent.*
(i) $\rho \in H^{\perp}$,
(ii) $\sum_{m \in c*H} \omega\{m\}\rho(\bar{m}) \neq 0$, *for each* $c \in K$,
(iii) $\sum_{m \in c*H} \omega\{m\}\rho(m) \neq 0$, *for some* $c \in K$.

Proof. $(i) \Rightarrow (ii)$ If $\rho \in H^{\perp}$ and $q : K \to K/H$ is the quotient map, then given $c \in K$, $q(\mu * \omega_H) = q(\mu)$ for $\mu = \delta_c\omega \in M(K)$. But clearly

$$q(\delta_c\omega) = \delta_{c*H}\omega = \sum_{m \in c*H} \delta_m\omega.$$

Hence $\rho(\delta_{c*H})\omega = \rho \circ q(\delta_c\omega) \neq 0$, where the last equality is because $\rho \circ q \in (K/H)\hat{}$ and a character is never zero.
$(iii) \Rightarrow (i)$ If $\rho \notin H^{\perp}$ then the multiplicative map $\rho \circ q$ should be identically zero on K/H (otherwise it is a character and $\rho \in H^{\perp}$). Hence $\sum_{m \in c*H} \rho(m)\omega = \rho(\delta_{c*H})\omega = 0$, for each $c \in K$. □

1. Prepare the state $|\chi_0\rangle'|0\rangle$.
2. Apply $\mathfrak{F}^{-1}$ to the first register to get

$$\sum_{x \in K} \omega\{x\}^{\frac{1}{2}}|x\rangle'|0\rangle.$$

3. Apply the black box to get

$$\sum_{x \in K} \omega\{x\}^{\frac{1}{2}}|x\rangle'|f(x)\rangle,$$

and measure the second register, to get

$$\frac{\sqrt{|K|}}{\sqrt{|c * H|}} \sum_{m \in c*H} \omega\{m\}^{\frac{1}{2}}|m\rangle'|\lambda_c\rangle,$$

where c is an element of K selected uniformly at random, and λ_c is the value of f on the coset $c * H$.
4. Apply $\mathfrak{F}$ to the first register to get

$$\frac{\sqrt{|K|}}{\sqrt{|c * H|}} \sum_{m \in c*H} \sum_{\rho \in \hat{K}} \omega\{m\}\pi\{\rho\}^{\frac{1}{2}}\rho(m)|\rho\rangle'|\lambda_c\rangle = \frac{\sqrt{|K|}}{\sqrt{|c * H|}} \sum_{\rho \in \hat{K}} \pi\{\rho\}^{\frac{1}{2}} \sum_{m \in c*H} \omega\{m\}\rho(m)|\rho\rangle'|\lambda_c\rangle$$

5. Measure the first register and observe a character ρ.

Algorithm 3. *Algorithm for Abelian Hidden Subhypergroup Problem*

Theorem 6.29. *If the Fourier transform could be efficiently calculated on a finite commutative hypergroup K, then the above algorithm solves HSHP for K in polynomial time.*

Note that in the above algorithm, the resulting distribution over ρ is independent of the coset $c * H$ arising after the first step. Also note that by Lemma 6.28, the character observed in step 3 is in $H^{\perp}$.

7. Fourier transform on groupoids

7.1 Abelian groupoids

The structure of abelian groupoids is recently studied by the third author in (57). Our basic reference for groupoids is (70).

A groupoid G is a small category in which each morphism is invertible. The unit space $X = G^{(0)}$ of G is is the subset of elements $\gamma\gamma^{-1}$ where γ ranges over G. The range map $r : G \to G^{(0)}$ and source map $s : G \to G^{(0)}$ are defined by $r(\gamma) = \gamma\gamma^{-1}$, and $s(\gamma) = \gamma^{-1}\gamma$, for $\gamma \in G$. We set $G^u = r^{-1}(u)$ and $G_u = s^{-1}(u)$. The loop space $G_u^u = \{\gamma \in G | r(\gamma) = s(\gamma)\}$ is a called the isotropy group of G at u.

Definition 7.1. An abelian groupoid is a groupoid whose isotropy groups are abelian.

Definition 7.2. (57) An equivalence relation R on X is r-discrete proper if its graph R is closed in $X \times X$ and the quotient map $q : X \to X/R$ is a local homeomorphism.

If the equivalence relation R on X is r-discrete proper, then the groupoid R is proper in the sense of (30, page 14), that is, the inverse image of every compact subset of $X \times X$ is compact in R. For a groupoid Γ with unit space $X = \Gamma^{(0)}$, the equivalence relation $R(\Gamma)$ and isotropy groupoid $\Gamma(X)$ are defined by $R(\Gamma) = \{(s(\gamma), r(\gamma)) : \gamma \in \Gamma\}$ and $\Gamma(X) = \{(\gamma, \sigma) : \gamma, \sigma \in \Gamma, r(\gamma) = s(\sigma)\}$.

Definition 7.3. An abelian groupoid Γ is called decomposable if $R = R(\Gamma)$ is r-discrete proper and covered with compact open R-sets.

Lemma 7.4. *The isotropy subgroupoid* $\Gamma(X)$ *is open in an decomposable abelian groupoid* Γ.

Proof. Since X is open in Γ and $r \times s$ is a continuous map when R has quotient topology, $\Gamma(X) = (r \times s)^{-1}(X)$ is open in Γ, because X is open in $R(\Gamma)$. □

Example 7.5. If Γ is an r-discrete abelian groupoid with finite unit space, then it is decomposable abelian groupoid.

7.1.1 Eigenfunctionals

A pair $(\mathcal{C}, \mathcal{D})$ is called a *regular C^*-inclusion* if $\mathcal{D}$ is a maximal abelian C*-subalgebra of the unital C^*-algebra $\mathcal{C}$ such that $1 \in \mathcal{D}$ and (*i*) $\mathcal{D}$ has the extension property in $\mathcal{C}$; (*ii*) $\mathcal{C}$ is regular (as a $\mathcal{D}$-bimodule). Let P denote the (unique) conditional expectation of $\mathcal{C}$ onto $\mathcal{D}$. We call $(\mathcal{C}, \mathcal{D})$ a C^*-diagonal if in addition, (*iii*) P is faithful.

If $\mathcal{C}$ is non-unital, $\mathcal{D}$ is called a diagonal in $\mathcal{C}$ if its unitization $\widetilde{\mathcal{D}}$ is a diagonal in $\widetilde{\mathcal{C}}$. A twist $\mathcal{G}$ is a proper $\mathbb{T}$-groupoid such that $\mathcal{G}/\mathbb{T}$ is an r-discrete equivalence relation $\mathcal{R}$. There is a one-to-one correspondence between twists and diagonal pairs of C*-algebras: Define

$$C_c(\mathcal{R}, \mathcal{G}) = \{f \in C_c(\mathcal{G}) : f(t\gamma) = tf(\gamma)\,(t \in \mathbb{T}, \gamma \in \mathcal{G})\}.$$

Then $\mathcal{G}$ is a $\mathbb{T}$-groupoid over $\mathcal{R}$ and

$$D(\mathcal{G}) = \{f \in C_c(\mathcal{R}, \mathcal{G}) : suppf \subset \mathbb{T}\mathcal{G}^0\}$$

where $\mathbb{T}\mathcal{G}^0$ denotes the isotropy group bundle of $\mathcal{G}$.

$C_c(\mathcal{R}, \mathcal{G})$ is a C^*-algebra with a distinguished abelian subalgebra $D(\mathcal{G})$. Furthermore, $D(\mathcal{G}) \cong C_0(\mathcal{G}^0)$. Now $C_c(\mathcal{R}, \mathcal{G})$ becomes a pre-Hilbert $D(\mathcal{G})$-module, whose completion $\mathbf{H}(E)$ is a

Hilbert $C_0(E^0)$-module. One may construct a $*$-homomorphism $\pi : C_c(\mathcal{R}, \mathcal{G}) \to \mathcal{B}(\mathbf{H}(\mathcal{G}))$ such that $\pi(f)g = fg$ (the convolution product) for all $f, g \in C_c(\mathcal{R}, \mathcal{G})$ and define $\mathcal{C}(\mathcal{G})$ to be the closure of $\pi(C_c(\mathcal{R}, \mathcal{G}))$ and $\mathcal{D}(\mathcal{G})$ to be the closure of $\pi(D(\mathcal{G}))$. It turns out that $\mathcal{D}(\mathcal{G})$ is a diagonal in $\mathcal{C}(\mathcal{G})$, hence every twist gives rise to a diagonal pair of C^*-algebras. Conversely, given C*-algebras $\mathcal{C}$ and $\mathcal{D}$ such that $\mathcal{D}$ is a diagonal in $\mathcal{C}$, one may construct a twist $\mathcal{G}$ such that $\mathcal{C} = \mathcal{C}(\mathcal{G})$ and $\mathcal{D} = \mathcal{D}(\mathcal{G})$, giving a bijective correspondence between twists and diagonal pairs (see (41) for more details).

Theorem 7.6. *If Γ is an decomposable abelian groupoid, then $(C^*(\Gamma), C^*(\Gamma(X)))$ is a C^*-diagonal pair.*

Proof. Since $\mathcal{R}^{(0)} = \widehat{\Gamma(X)}$, where $\mathcal{R} := \widehat{\Gamma(X)} \rtimes_c R$, and $\widehat{\Gamma(X)} \rtimes_c R$ is a principal r-discrete groupoid, $(C^*(\widehat{\Gamma(X)} \rtimes_c R, \mathcal{G}), C_0(\widehat{\Gamma(X)}))$ is a diagonal pair (55, Theorem VIII.6), hence $(C^*(\Gamma), C^*(\Gamma(X)))$ is a C^*-diagonal, because the isomorphism between groupoid C^*-algebras preserves their diagonals. □

Another piece of structure of the pair $(C^*(\mathcal{R}, \mathcal{G}) = C^*(\Gamma), C_0(\mathcal{R}^{(0)}) = C^*(\Gamma(X)))$ is the restriction map $P : f \mapsto f|_{C^*(\Gamma(X))}$ from $C^*(\Gamma)$ to $C^*(\Gamma(X))$ (65). In (54), the authors show that there is generalized conditional expectation $C^*(\Gamma) \to C^*(\Gamma(u))$.

Corollary 7.7. *If Γ is an decomposable abelian groupoid, then the restriction map $P : C^*(\Gamma) \to C^*(\Gamma(X))$ is the unique faithful conditional expectation onto $C^*(\Gamma(X))$.*

Corollary 7.8. *If Γ is an decomposable abelian groupoid, then $C^*(\Gamma(X))$ (resp. $C^*_\mu(\Gamma(X))$) is a maximal abelian sub-algebra (masa) in $C^*(\Gamma)$ (resp. $C^*_\mu(\Gamma)$).*

If Γ is a nontrivial decomposable abelian groupoid, X can not be the interior of $\Gamma(X)$, therefore $C^*(X)$ is not a maximal subalgebra of $C^*(\Gamma)$ (70, prop. II.4.7(ii)). By theorem 7.7, $C^*(\Gamma(X))$ is commutative and we can conclude that there is no $f \in C_c(\Gamma)$ with support outside $\Gamma(X)$ which is in the commutative C^*-algebra $C^*(\Gamma(X))$.

Corollary 7.9. *If Γ is an decomposable abelian groupoid, $C^*(\Gamma)$ is regular as a $C^*(\Gamma(X))$-bimodule.*

For a Banach $C^*(\Gamma(X))$-bimodule $\mathcal{M}$, An element $m \in \mathcal{M}$ is called an *intertwiner* if

$$m.C^*(\Gamma(X)) = C^*(\Gamma(X)).m.$$

If $m \in \mathcal{M}$ is an intertwiner such that for every $f \in C^*(\Gamma(X))$, $f.m \in \mathbb{C}m$, we call m a minimal intertwiner (compare intertwiners to normalizers (22, prop 3.3)). Regularity of a bimodule $\mathcal{M}$ is equivalent to norm-density of the $C^*(\Gamma(X))$-intertwiners (22, remarks 4.2).
A $C^*(\Gamma(X))$-eigenfunctional is a nonzero linear functional $\phi : \mathcal{M} \to \mathbb{C}$ such that, for all $f \in C^*(\Gamma(X))$, $g \to \phi(f * g)$, $g \to \phi(g * f)$ are multiples of ϕ. A minimal intertwiner of $\mathcal{M}^*$ is an eigenfunctional. We equip the set $\mathcal{E}_{C^*(\Gamma(X))}(\mathcal{M})$ of all $C^*(\Gamma(X))$- eigenfunctionals with the relative weak* topology $\sigma(\mathcal{M}^*, \mathcal{M})$. Its normalizer is the set $N(C^*(\Gamma(X))) = \{v \in C^*(\Gamma) : vC^*(\Gamma(X))v^* \subset C^*(\Gamma(X))$ and $v^*C^*(\Gamma(X))v \subset C^*(\Gamma(X))\}$. For $v \in N(C^*(\Gamma(X)))$, let $dom(v) := \{\phi \in \mathcal{Z} : \phi(v^*v) > 0\}$, note this is an open set in $\mathcal{Z} = \widehat{C^*(\Gamma(X))}$. There is a homeomorphism $\beta_v : dom(v) \to dom(v^*) := ran(v)$ given by

$$\beta_v(\phi)(f) = \frac{\phi(v^* * f * v)}{\phi(v^* * v)}.$$

It is easy to show that $\beta_v^{-1} = \beta_{v^*}$ (41).

Remark 7.10. Intertwiners and normalizers are closely related, at least when $C^*(\Gamma(X))$ is a masa in the unital C^*-algebra $C^*(\Gamma)$ containing $\mathcal{M}$ (22). Indeed, if $v \in C^*(\Gamma)$ is an intertwiner for $C^*(\Gamma(X))$, then $v^*v, vv^* \in C^*(\Gamma(X))' \bigcap C^*(\Gamma)$. If $C^*(\Gamma(X))$ is maximal abelian in $C^*(\Gamma)$, then v is a normalizer of $C^*(\Gamma(X))$ (22, Proposition 3.3). Conversely, for $v \in N(C^*(\Gamma(X)))$, if β_{v^*} extends to a homeomorphism of $S(v^*)$ onto $S(v)$, then v is an intertwiner. Moreover, if I is the set of intertwiners, then $N(C^*(\Gamma(X)))$ is contained in the norm-closure of I, and when $C^*(\Gamma(X))$ is a masa in $C^*(\Gamma)$, $N(C^*(\Gamma(X))) = I$ (22, proposition 3.4).

Since $C^*(\Gamma(X))$ is a C^*-subalgebra of the C^*-algebra $C^*(\Gamma)$, and it contains an approximate unit of $C^*(\Gamma)$, for each $v \in N(C^*(\Gamma(X)))$, we have $vv^*, v^*v \in C^*(\Gamma(X))$ (65, lemma 4.6).
Given an eigenfunctional $\phi \in \mathcal{E}_{C^*(\Gamma(X))}(\mathcal{M})$, the associativity of the maps $f \in C^*(\Gamma(X)) \mapsto f.\phi$ and $f \in C^*(\Gamma(X)) \mapsto \phi.f$ yields the existence of unique multiplicative linear functionals $s(\phi)$ and $r(\phi)$ on $C^*(\Gamma(X))$ satisfying $s(\phi)(f)\phi = f.\phi$ and $r(\phi)(f)\phi = \phi.f$, that is

$$\phi(g * f) = \phi(g)[s(\phi)(f)], \ \phi(f * g) = [r(\phi)(f)]\phi(g).$$

We call $s(\phi)$ and $r(\phi)$ the source and range of ϕ respectively (22, page 6). There is a natural action of the nonzero complex numbers z on $\mathcal{E}(\mathcal{M})$, sending (z, ϕ) to the functional $m \to z\phi(m)$; clearly $s(z\phi) = s(\phi)$ and $r(z\phi) = r(\phi)$. Also $\mathcal{E}(\mathcal{M}) \cup \{0\}$ is closed in the weak*-topology. Furthermore, $r : \mathcal{E}(\mathcal{M}) \to \mathcal{Z}$ and $s : \mathcal{E}(\mathcal{M}) \to \mathcal{Z}$ are continuous.

Notation 7.11. We define $\mathcal{G} = \mathcal{E}^1(C^*(\Gamma))$, where $\mathcal{E}^1(C^*(\Gamma))$ is the collection of norm-one eigenvectors for the dual action of $C^*(\Gamma(X))$ on the Banach space dual $C^*(\Gamma)^*$. For a bimodule $\mathcal{M} \subset C^*(\Gamma)$, $\mathcal{G}|\mathcal{M}$ can be defined directly in terms of the bimodule structure of $\mathcal{M}$, without explicit reference to $C^*(\Gamma)$ as in (22, remark 4.16).

The groupoid $\mathcal{G}$, with a suitable operation and the relative weak*- topology admits a natural $\mathbb{T}$-action. If $\phi, \psi \in \mathcal{E}(\mathcal{M})$ satisfy $r(\phi) = r(\psi)$ and $s(\phi) = s(\psi)$, then there exists $z \in \mathbb{C}$ such that $z \neq 0$ and $\phi = z\psi$ (22, corollary 4.10). Also $\mathcal{E}^1(\mathcal{M}) \cup \{0\}$ is weak*-compact (22, prop. 4.17). Thus, $\mathcal{E}^1(\mathcal{M})$ is a locally compact Hausdorff space. We may regard an element $m \in \mathcal{M}$ as a function on $\mathcal{E}^1(\mathcal{M})$ via $\hat{m}(\phi) = \phi(m)$. When $\mathcal{A}$ is both a norm-closed algebra and a $C^*(\Gamma(X))$-bimodule, the coordinate system $\mathcal{E}^1(\mathcal{A})$ has the additional structure of a continuous partially defined product as described in (22, remark4.14).

Notation 7.12. Let $\mathcal{R}(\mathcal{M}) := \{|\phi| : \phi \in \mathcal{E}^1(\mathcal{M})\}$. Then $\mathcal{R}(\mathcal{M})$ may be identified with the quotient $\mathcal{E}^1(\mathcal{M})\backslash\mathbb{T}$ of $\mathcal{E}^1(\mathcal{M})$ by the natural action of $\mathbb{T}$. A twist is a proper $\mathbb{T}$-groupoid $\mathcal{G}$ so that $\mathcal{G}\backslash\mathbb{T}$ is an principal r-discrete groupoid. The topology on $\mathcal{R}(C^*(\Gamma))$ is compatible with the groupoid operations, so $\mathcal{R}(C^*(\Gamma))$ is a topological equivalence relation (22).

We know that $(C^*(\Gamma), C^*(\Gamma(X)))$ is a C^*-diagonal, let $\mathcal{M} \subset C^*(\Gamma)$ be a norm-closed $C^*(\Gamma(X))$-bimodule. Then the span of $\mathcal{E}^1(\mathcal{M})$ is $\sigma(\mathcal{M}^*, \mathcal{M})$-dense in $\mathcal{M}^*$. Suppose $\mathcal{A}$ is a norm closed algebra satisfying $C^*(\Gamma(X)) \subset \mathcal{A} \subset C^*(\Gamma)$. If $\mathcal{B}$ is the C^*-subalgebra of $C^*(\Gamma)$ generated by $\mathcal{A}$, then $\mathcal{B}$ is the C^*-envelope of $\mathcal{A}$. If in addition, $\mathcal{B} = C^*(\Gamma)$, then $\mathcal{R}(C^*(\Gamma))$ is the topological equivalence relation generated by $\mathcal{R}(\mathcal{A})$ (22).
Eigenfunctionals can be viewed as normal linear functional on $\mathcal{B}^{**}$ and one may use the polar decomposition to obtain a minimal partial isometry for each eigenfunctional (22). Indeed, by the polar decomposition for linear functionals, there is a partial isometry $u^* \in C^*(\Gamma)^{**}$ and positive linear functionals $|\phi|, |\phi^*| \in C^*(\Gamma)^*$ so that $\phi = u^*.|\phi| = |\phi^*|.u^*$. Then $r(\phi) = |\phi|$ and $s(\phi) = |\phi^*|$. Moreover, uu^* and u^*u are the smallest projections in $C^*(\Gamma)^{**}$ which satisfy $u^*u.s(\phi) = s(\phi).u^*u = s(\phi)$ and $uu^*.r(\phi) = r(\phi).uu^* = r(\phi)$. For $\phi \in \mathcal{E}^1(C^*(\Gamma))$, we call the

above partial isometry u the partial isometry associated to ϕ, and denote it by v_ϕ. If $\phi \in \mathcal{Z}$, then u is a projection and we denote it by p_ϕ. The above equations show that $v_\phi^* v_\phi = p_{s(\phi)}$ and $v_\phi v_\phi^* = p_{r(\phi)}$. Moreover, given $\phi \in \mathcal{E}^1(C^*(\Gamma))$, v_ϕ may be characterized as the unique minimal partial isometry $w \in C^*(\Gamma)^{**}$ such that $\phi(w) > 0$. Recall that $\chi, \xi \in C^*(\hat{\Gamma}(X))$ satisfy $(\chi, \xi) \in \mathcal{R}(C^*(\Gamma))$ if and only if there is $\phi \in \mathcal{E}^1(C^*(\Gamma))$ with $r(\phi) = \chi$ and $s(\phi) = \xi$. For brevity, we write $\chi \sim \xi$ in this case (22).

Remark 7.13. For $\chi \in \mathcal{Z}$, we denote the GNS-representation of $C^*(\Gamma)$ associated to the unique extension of χ by $(\mathbf{H}_\chi, \pi_\chi)$. Let $\chi, \xi \in \mathcal{Z}$, then $\xi \sim \chi$ if and only if the GNS-representations π_χ and π_ξ are unitarily equivalent (22, lemma 5.8). Therefore if $\chi, \xi \in \mathcal{Z}(x)$, then $\xi \sim \chi$ if and only if $\xi = \chi$ (54, lemma 2.11).

If we set $\mathcal{M} = \{f \in C^*(\Gamma) : \chi(f^* f) = 0\}$, then $C^*(\Gamma)/\mathcal{M}$ is complete relative to the norm induced by the inner product $\langle f + \mathcal{M}, g + \mathcal{M}\rangle = \chi(g^* f)$, and thus $\mathbf{H}_\chi = C^*(\Gamma)/\mathcal{M}$.

Proposition 7.14. *Suppose $\chi \in \mathcal{Z}$ and $\phi \in \mathcal{E}^1(C^*(\Gamma))$ satisfy $\chi \sim s(\phi)$. Then there exist unit orthogonal unit vectors $\omega_1, \omega_2 \in \mathbf{H}_\chi$ such that for every $f \in C^*(\Gamma)$, $\phi(f) = \langle \pi_\chi(f)\omega_1, \omega_2\rangle$ (22).*

Theorem 7.15. *We have $\mathcal{R}(C^*(\Gamma)) \cong \mathcal{Z} \rtimes_c R(\Gamma)$, algebraically and topologically.*

Proof. Ionescu and Williams showed that every representation of Γ induced from an irreducible representation of a stability group is irreducible (30, page 296). We can extend a character $\chi \in \mathcal{Z}$ to $\chi \in \widehat{C^*(\Gamma)}$ such that $\chi|C^*(\Gamma|_{[y]\neq[x]}) = 1$, but since the extension is unique in C^*-diagonals, it will be equal to $Ind(x, \Gamma(X)_x\ , \chi)$. Let $\chi, \xi \in \widehat{\Gamma(X)}$, then $\chi \sim \xi$ if and only if the GNS-representations π_χ and π_ξ are unitarily equivalent. By (54, lemma 2.5), $Ind(x, \Gamma(X)_x, \chi)$ is unitarily equivalent to $Ind(x.k,\ \Gamma(X)_{s(k)}, \chi.k)$ in $\widehat{C^*(\Gamma)}$, hence they are in the same class in $\widehat{C^*(\Gamma)}$, that is $\chi \sim \chi.k$, where $k \in R$. But if two stability groups of Γ are not in the same orbit, none of their irreducible representation can be equivalent. This shows that the two sets are the same algebraically. Since the topology is r-discrete, they are also topologically isomorphic. □

Remark 7.16. We know that $C^*(\Gamma|_{[x]}) \cong K(L^2(X|_{[x]}, \mu)) \otimes C^*(\Gamma(x))$ (16, page 107). Also we have $L^2(X|_{[x]}, \mu) \cong L^2(R_x, \alpha_x)$ (56, page 135), therefore $\mathbf{H}_\chi = C_c(\Gamma_x) \otimes \mathbf{H}_{\pi_\chi}$ and

$$Ind(x, \Gamma(X)_x, \chi) = \pi_\chi,$$

where $\mathbf{H}_{\pi_\chi}$ is the Hilbert space constructed in $C^*(\Gamma(X)_x)$ by χ (54), but $\mathbf{H}_\chi$ is the Hilbert space constructed in $C^*(\Gamma)$ by the unique extension of χ (22). Hence

$$\frac{C^*(\Gamma|_{[x]})}{\{f \in C^*(\Gamma) : \chi(f^* * f) = 0\}} \cong C_c(\Gamma_x) \otimes \mathbf{H}_{\pi_\chi}$$

and $\pi_\chi(f)\mathbf{(}g \otimes \omega\mathbf{)} = Ind(x, \Gamma(X)_x, \chi)(f)\mathbf{(}g \otimes \omega\mathbf{)} = \mathbf{(}f * g \otimes \omega\mathbf{)}$.

Corollary 7.17. *The groupoid $\mathcal{G}$ above is the $\mathbb{T}$-groupoid of $\mathcal{Z} \rtimes_c R$. In other words, we have the exact sequence*

$$\mathcal{Z} \to \mathcal{Z} \times \mathbb{T} \to \mathcal{G} \to \mathcal{G}\backslash\mathbb{T} \cong \mathcal{Z} \rtimes_c R.$$

7.1.2 Fourier transform

We know that $\mathcal{Z} \subset \mathcal{E}^1(\mathcal{M})$. If Γ is an abelian group, we have $C^*(\Gamma(X)) = \mathcal{M} = C^*(\Gamma)$, hence $\mathcal{E}_{C^*(\Gamma(X))}(\mathcal{M}) \supset \widehat{C^*(\Gamma)}$, hence eigenfunctionals are generalizations of multiplicative linear functionals. However, in general, $\mathcal{E}^1(\mathcal{M})$ need not separate points of $\mathcal{M}$.
We define the open support of $f \in C^*(\mathcal{R})$ as

$$supp'(f) = \{\gamma \in \mathcal{R} : f(\gamma) \neq 0\}.$$

Let

$$C_0(\mathcal{Z}) = \{f \in C^*(\mathcal{R}, \mathcal{G}) : supp' f \subset \mathcal{Z}\}$$

(65). Since $\mathcal{R}$ is (topologically) principal, and the normalizer $N(C^*(\Gamma(X)))$ consists exactly of the elements of $C^*(\Gamma)$ whose open support is a bisection (65, Proposition 4.8), we may define $\alpha_v = \alpha_{supp' v}$, where α_S for a bisection S of an r-discrete groupoid $\mathcal{R}$ is defined in (65). We define the Weyl groupoid $\mathcal{K}$ of $(C^*(\Gamma), C^*(\Gamma(X)))$ as the groupoid of germs of

$$K(C^*(\Gamma(X))) = \{\alpha_v, v \in N(C^*(\Gamma(X)))\}.$$

By proposition 4.13 in (65), the Weyl groupoid $\mathcal{K}$ of $(C^*(\Gamma), C^*(\Gamma(X)))$ is canonically isomorphic to $\mathcal{R}$.
Following Renault, let us define $D = \{(z_1, v, z_2) \in \mathcal{Z} \times N(C^*(\Gamma(X))) \times \mathcal{Z} : v^*v(z_2) > 0$ and $z_1 = \alpha_v(z_2)\}$. Consider the quotient $\mathcal{G}(C^*(\Gamma(X))) = D/ \sim$ by the equivalence relation: $(z_1, v, z_2) \sim (z_1', v', z_2')$ if and only if $z_2 = z_2'$ and there exist $f, f' \in C^*(\Gamma(X))$ with $f(z_2), f'(z_2) > 0$ such that $v * f = v' * f'$. The class of (z_1, v, z_2) is denoted by $[z_1, v, z_2]$. Then $\mathcal{G}(C^*(\Gamma(X)))$ has a natural groupoid structure over $\mathcal{Z}$, defined exactly in the same fashion as the groupoid of germs: the range and source maps are defined by $r[z_1, v, z_2] = z_1$, $s[z_1, v, z_2] = z_2$, the product by $[z_1, v, z_2][z_2, v', z] = [z_1, v * v', z]$ and the inverse by $[z_1, v, y]^{-1} = [y, v^*, z_1]$.
The map $(z_1, v, z_2) \to [z_1, \alpha_v, z_2]$ from D to $\mathcal{K}$ factors through the quotient and defines a groupoid homomorphism from $\mathcal{G}(C^*(\Gamma(X)))$ onto $W(C^*(\Gamma(X)))$. Moreover the subset $\mathcal{H} = \{[z, f, z] : f \in C^*(\Gamma(X)), f(z) \neq 0\} \subset \mathcal{G}(C^*(\Gamma(X)))\}$ can be identified with the trivial group bundle $\mathbb{T} \times \mathcal{Z}$. Since $C^*(\Gamma(X))$ is maximal abelian and contains an approximate unit of $C^*(\Gamma)$, the sequence

$$\mathcal{H} \to \mathcal{G}(C^*(\Gamma(X))) \to \mathcal{K}$$

is (algebraically) an extension (65, Proposition 4.14). Also we have a canonical isomorphism of extensions:

$$\begin{tikzcd}
\mathcal{H} \arrow[r] \arrow[d] & \mathcal{G}(C^*(\Gamma(X))) \arrow[r] \arrow[d] & \mathcal{K} \arrow[d] \\
\mathbb{T} \times \mathcal{Z} \arrow[r] & \mathcal{G} \arrow[r] & \mathcal{R}
\end{tikzcd}$$

It is easy to recover the topology of $\mathcal{G}(C^*(\Gamma(X)))$. Indeed, every $v \in N(C^*(\Gamma(X)))$ defines a trivialization of the restriction of $\mathcal{G}(C^*(\Gamma(X)))$ to the open bisection $S = supp'(v)$. Therefore $\mathcal{G}(C^*(\Gamma(X)))$ is a *locally trivial* topological twist over $\mathcal{K}$ (65, lemma 4.16).

Definition 7.18 (Fourier transform)**.** The Fourier transform is defined in (54, equaion (4.5)) as

$$\hat{f}(\chi, \gamma) = \frac{\int_{\Gamma(r(\gamma))} \overline{\chi(g)} f(g\gamma) d\beta^{r(\gamma)}(g)}{\sqrt{\omega(\gamma)}},$$

where ω is a continuous $\Gamma(X)$-invariant homomorphism from Γ to $\mathbb{R}^+$ such that for all $f \in C_c(\Gamma(X))$,

$$\int_{\Gamma(X)} f(g)d\beta^{r(\gamma)}(g) = \omega(\gamma)\int_{\Gamma(X)} f(\gamma g\gamma^{-1})d\beta^{s(\gamma)}(g).$$

Definition 7.19 (Gelfand transform). Given $f \in C^*(\Gamma)$ and $(z_1, v, z_2) \in D$, define the Gelfand transform of f by

$$\hat{f}(z_1, v, z_2) = \frac{P(v^* * f)(z_2)}{\sqrt{v^* * v(z_2)}}.$$

Then $\hat{f}(z_1, v, z_2)$ depends only on its class in $\mathcal{G}(C^*(\Gamma(X)))$, $\hat{f}$ defines a continuous section of the *line bundle* $L(C^*(\Gamma(X))) := (\mathbb{C} \times \mathcal{G}(C^*(\Gamma(X))))/\mathbb{T}$, and the map $f \mapsto \hat{f}$ is linear and injective (65, Lemma 5.3). When f belongs to $C^*(\Gamma(X))$, then $\hat{f}$ vanishes off $\mathcal{Z}$ and its restriction to $\mathcal{Z}$ is the classical Gelfand transform. If v belongs to $N(C^*(\Gamma(X)))$, then the open support of $\hat{v}$ is the open bisection of $\mathcal{K}$ defined by the partial homeomorphism α_v (65, corollary 5.6). The Weyl groupoid $\mathcal{K}$ is a Hausdorff r-discrete groupoid (65, proposition 5.7). Let $N_c(C^*(\Gamma(X)))$ be the set of elements v in $N(C^*(\Gamma(X)))$ such $\hat{v}$ has compact support and let $C^*(\Gamma)_c$ be its linear span. Then $N_c(C^*(\Gamma(X)))$ is dense in $N(C^*(\Gamma(X)))$ and $C^*(\Gamma)_c$ is dense in $C^*(\Gamma)$, and the Gelfand map $\Psi : f \to \hat{f}$ defined above sends $C^*(\Gamma)_c$ bijectively onto $C_c(\mathcal{K}, \mathcal{G})$ and $C^*(\Gamma(X))_c = C^*(\Gamma(X)) \bigcap C^*(\Gamma)_c$ onto $C_c(\mathcal{Z})$. Hence the Gelfand map $\Psi : C^*(\Gamma)_c \to C_c(\mathcal{K}, \mathcal{G})$ is an $*$-algebra isomorphism (65, lemma 5.8).

7.2 Compact groupoids

All over this section, we assume that $\mathcal{G}$ is compact and the Haar system on $\mathcal{G}$ is normalized so that $\lambda_u(\mathcal{G}_u) = 1$, for each $u \in X$, where $X = \mathcal{G}^{(0)}$ is the unit space of $\mathcal{G}$. In general $\mathcal{G}$ may be non-Hausdorff, but as usual we assume that $\mathcal{G}^{(0)}$ and $\mathcal{G}^u$, $\mathcal{G}_v$, for each $u, v \in \mathcal{G}^{(0)}$ are Hausdorff.

7.2.1 Fourier transform

Let $\mathcal{G}$ be a groupoid (70, 1.1). The unit space of $\mathcal{G}$ and the range and source maps are denoted by $X = \mathcal{G}^{(0)}$, r and s, respectively. For $u, v \in \mathcal{G}^{(0)}$, we put $\mathcal{G}^u = r^{-1}\{u\}$, $\mathcal{G}_v = s^{-1}\{v\}$, and $\mathcal{G}^u_v = \mathcal{G}^u \cap \mathcal{G}_v$. Also we put $\mathcal{G}^{(2)} = \{(x, y) \in \mathcal{G} \times \mathcal{G} : r(y) = s(x)\}$.

We say that $\mathcal{G}$ is a topological groupoid (70, 2.1) if the inverse map $x \mapsto x^{-1}$ on $\mathcal{G}$ and the multiplication map $(x, y) \mapsto xy$ from $\mathcal{G}^{(2)}$ to $\mathcal{G}$ are continuous. This implies that the range and source maps r and s are continuous and the subsets $\mathcal{G}^u$, $\mathcal{G}_v$, and $\mathcal{G}^u_v$ are closed, and so compact, for each $u, v \in X$. We fix a left Haar system $\lambda = \{\lambda^u\}_{u \in X}$ and put $\lambda_u(E) = \lambda^u(E^{-1})$, for Borel sets $E \subseteq \mathcal{G}_u$, and let λ^v_u be the restriction of λ_u to the Borel σ-algebra of $\mathcal{G}^v_u$. The integrals against λ^u and λ^v_u are understood to be on $\mathcal{G}^u$ and $\mathcal{G}^v_u$, respectively. The functions on $\mathcal{G}^v_u$ are extended by zero, if considered as functions on $\mathcal{G}$. All over this section, we assume that $\lambda_u(\mathcal{G}^v_u) \neq 0$, for each $u, v \in X$. This holds in transitive groupoids. In this case, we say that $\mathcal{G}$ is *locally non-trivial* and we denote the restriction of λ_u to the σ-algebra of Borel subsets of $\mathcal{G}^v_u$ by λ^v_u.

The convolution product of two measurable functions f and g on $\mathcal{G}$ is defined by

$$f * g(x) = \int f(y)g(y^{-1}x)d\lambda^{r(x)}(y) = \int f(xy^{-1})g(y)d\lambda_{s(x)}(y).$$

A (continuous) *representation* of G is a double $(\pi, \mathcal{H}_\pi)$, where $\mathcal{H}_\pi = \{\mathcal{H}^\pi_u\}_{u \in X}$ is a continuous bundle of Hilbert spaces over X such that:

(i) $\pi(x) \in \mathcal{B}(\mathcal{H}^{\pi}_{s(x)}, \mathcal{H}^{\pi}_{r(x)})$ is a unitary operator, for each $x \in \mathcal{G}$,
(ii) $\pi(u) = id_u : \mathcal{H}^{\pi}_u \to \mathcal{H}^{\pi}_u$, for each $u \in X$,
(iii) $\pi(xy) = \pi(x)\pi(y)$, for each $(x, y) \in \mathcal{G}^{(2)}$,
(iv) $\pi(x^{-1}) = \pi(x)^{-1}$, for each $x \in \mathcal{G}$,
(v) $x \mapsto \langle \pi(x)\xi(s(x)), \eta(r(x))\rangle$ is continuous on $\mathcal{G}$, for each $\xi, \eta \in C_0(G^{(0)}, \mathcal{H}_{\pi})$.

Two representations π_1, π_2 of $\mathcal{G}$ are called (unitarily) *equivalent* if there is a (continuous) bundle $U = \{U_u\}_{u\in X}$ of unitary operators $U_u \in \mathcal{B}(\mathcal{H}^{\pi}_u, \mathcal{H}^{\pi}_u)$ such that

$$U_{r(x)}\pi_1(x) = \pi_2(x)U_{s(x)} \quad (x \in \mathcal{G}).$$

We use $\mathcal{R}ep(\mathcal{G})$ to denote the category consisting of (equivalence classes of continuous) representations of $\mathcal{G}$ as objects and intertwining operators as morphisms (3, Notation 2.5). Let $\pi \in \mathcal{R}ep(\mathcal{G})$, the mappings

$$x \mapsto \langle \pi(x)\xi_{s(x)}, \eta_{r(x)}\rangle,$$

where ξ, η are continuous sections of $\mathcal{H}_{\pi}$ are called *matrix elements* of π. This terminology is based on the fact that if $\{e^i_u\}$ is a basis for $\mathcal{H}^{\pi}_u$, then $\pi_{ij}(x) = \langle \pi(x)e^j_{s(x)}, e^i_{r(x)}\rangle$ is the (i, j)-th entry of the (possibly infinite) matrix of $\pi(x)$. We denote the linear span of matrix elements of π by $\mathcal{E}_{\pi}$. By continuity of representations, $\mathcal{E}_{\pi}$ is a subspace of $C(\mathcal{G})$. It is clear that $\mathcal{E}_{\pi}$ depends only on the unitary equivalence class of π. For $u, v \in X$, $\mathcal{E}^{\pi}_{u,v}$ consists of restrictions of elements of $\mathcal{E}_{\pi}$ to $\mathcal{G}^v_u$. Also we put $\mathcal{E}_{u,v} = span\big(\cup_{\pi\in\hat{\mathcal{G}}} \mathcal{E}^{\pi}_{u,v}\big)$ and $\mathcal{E} = span\big(\cup_{\pi\in\hat{\mathcal{G}}} \mathcal{E}_{\pi}\big)$.
It follows from the Peter-Weyl theorem (1, Theorem 3.10) (note that there is a typo in (3, Theorem 3.10), and the orthonormal basis elements $\sqrt{d^{\pi}_u/\lambda_u(\mathcal{G}^v_u)}\pi^{ij}_{u,v}$ is wrongly inscribed as $\sqrt{d^{\pi}_u\lambda_u(\mathcal{G}^v_u)}\pi^{ij}_{u,v}$) that, for $u, v \in X$, if $\lambda_u(\mathcal{G}^v_u) \neq 0$, then for each $f \in L^2(\mathcal{G}^v_u, \lambda^v_u)$,

$$f = \sum_{\pi\in\hat{\mathcal{G}}} \sum_{i=1}^{d^{\pi}_v} \sum_{j=1}^{d^{\pi}_u} c^{ij}_{u,v,\pi}\pi^{ij}_{u,v} \ ,$$

where

$$c^{ij}_{u,v,\pi} = \frac{d^{\pi}_u}{\lambda_u(\mathcal{G}^v_u)} \int_{\mathcal{G}^v_u} f(x)\overline{\pi^{ij}_{u,v}(x)}d\lambda^v_u(x) \quad (1 \leq i \leq d^{\pi}_v, 1 \leq j \leq d^{\pi}_u).$$

This is a local version of the classical non commutative Fourier transform. As in the classical case, the main drawback is that it depends on the choice of the basis (which in turn gives the choice of the coefficient functions). The trick is similar to the classical case, that's to use the continuous decomposition using integrals. This is the content of the next definition. As usual, all the integrals are supposed to be on the support of the measure against which they are taken.

Definition 7.20. Let $u, v \in X$ and $f \in L^1(\mathcal{G}^v_u, \lambda^v_u)$, then the Fourier transform of f is $\mathfrak{F}_{u,v}(f) : \mathcal{R}ep(\mathcal{G}) \to \mathcal{B}(\mathcal{H}^{\pi}_v, \mathcal{H}^{\pi}_u)$ defined by

$$\mathfrak{F}_{u,v}(f)(\pi) = \int f(x)\pi(x^{-1})d\lambda^v_u(x).$$

To better understand this definition, let us go back to the group case for a moment. Let's start with a locally compact abelian group G. Then the Pontryagin dual $\hat{G}$ of G is a locally compact abelian group and for each $f \in L^1(G)$, its Fourier transform $\hat{f} \in C_0(\hat{G})$ is defined by

$$\hat{f}(\chi) = \int_G f(x)\overline{\chi(x)}dx \quad (\chi \in \hat{G}).$$

The continuity of $\hat{f}$ is immediate and the fact that it vanishes at infinity is the so called *Riemann-Lebesgue lemma*. For non abelian compact groups, a similar construction exists, namely, with an slight abuse of notation, for each $f \in L^1(G)$ one has $\hat{f} \in C_0(\hat{G}, \mathcal{B}(\mathcal{H}))$, where $\hat{G}$ is the set of (unitary equivalence classes of) irreducible representations of G endowed with the Fell topology. In the groupoid case, one has a similar local interpretation. Each $f \in L^1(\mathcal{G}_u^v, \lambda_u^v)$ has its Fourier transform $\mathfrak{F}_{u,v}(f)$ in $C_0(\hat{\mathcal{G}}, \mathcal{B}_{u,v}(\mathcal{H}))$, where $\hat{\mathcal{G}}$ is the set of (unitary equivalence classes of) irreducible representations of $\mathcal{G}$ endowed again with the Fell topology, and $\mathcal{B}_{u,v}(\mathcal{H})$ is a bundle of operator spaces over $\hat{\mathcal{G}}$ whose fiber at π is $\mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$, and $C_0(\hat{\mathcal{G}}, \mathcal{B}_{u,v}(\mathcal{H}))$ is the set of all continuous sections vanishing at infinity.
Now let us discuss the properties of the Fourier transform. If we choose (possibly infinite) orthonormal bases for $\mathcal{H}_u^\pi$ and $\mathcal{H}_v^\pi$ and let each $\pi(x)$ be represented by the (possibly infinite) matrix with components $\pi_{u,v}^{ij}(x)$, then $\mathfrak{F}_{u,v}(f)$ is represented by the matrix with components $\mathfrak{F}_{u,v}(f)(\pi)^{ij} = \frac{\lambda_u(\mathcal{G}_u^v)}{d_u^\pi} c_{u,v,\pi}^{ji}$. When $f \in L^2(\mathcal{G}_u^v, \lambda_u^v)$, summing up over all indices i, j, we get the following.

Proposition 7.21. (Fourier inversion formula) *For each $u, v \in X$ and $f \in L^2(\mathcal{G}_u^v, \lambda_u^v)$,*

$$f = \sum_{\pi \in \hat{\mathcal{G}}} \frac{d_u^\pi}{\lambda_u(\mathcal{G}_u^v)} Tr\big(\mathfrak{F}_{u,v}(f)(\pi)\pi(\cdot)\big),$$

where the sum converges in the L^2 norm and

$$\|f\|_2^2 = \sum_{\pi \in \hat{\mathcal{G}}} \frac{d_u^\pi}{\lambda_u(\mathcal{G}_u^v)} Tr\big(\mathfrak{F}_{u,v}(f)(\pi)\mathfrak{F}_{u,v}(f)(\pi)^*\big).$$

We collect the properties of the Fourier transform in the following lemma. Note that in part (iii), $f^*(x) = \overline{f(x^{-1})}$, for $x \in \mathcal{G}_u^v$ and $f \in L^1(\mathcal{G}_u^v, \lambda_u^v)$.

Lemma 7.22. *Let $u, v, w \in X$, $a, b \in \mathbb{C}$, and $f, f_1, f_2 \in L^1(\mathcal{G}_u^v, \lambda_u^v)$, $g \in L^1(\mathcal{G}_v^w, \lambda_v^w)$, then for each $\pi \in \mathcal{R}ep(\mathcal{G})$,*
$(i)\mathfrak{F}_{u,v}(af_1 + bf_2) = a\mathfrak{F}_{u,v}(f_1) + b\mathfrak{F}_{u,v}(f_2)$,
$(ii)\mathfrak{F}_{u,w}(f * g)(\pi) = \mathfrak{F}_{u,v}(f)(\pi)\mathfrak{F}_{v,w}(g)(\pi)$,
$(iii)\mathfrak{F}_{v,u}(f^*)(\pi) = \mathfrak{F}_{u,v}(f)(\pi)^*$,
(iv) $\mathfrak{F}_{u,w}(\ell_x(f))(\pi) = \mathfrak{F}_{u,v}(f)(\pi)\pi(x^{-1})$ and $\mathfrak{F}_{w,v}(r_y(f))(\pi) = \pi(y)\ \mathfrak{F}_{u,v}(f)(\pi)$, whenever $x \in \mathcal{G}_v^w, y \in \mathcal{G}_u^w$.

As in the group case, there is yet another way of introducing the Fourier transform. For each finite dimensional continuous representation π of $\mathcal{G}$, let the *character* χ_π of π be the bundle of functions χ_π whose fiber χ_u^π at $u \in X$ is defined by $\chi_u^\pi(x) = \mathrm{Tr}(\pi(x))$, for $x \in \mathcal{G}_u^u$, where Tr is the trace of matrices. Note that one can not have these as functions defined on $\mathcal{G}_u^v$, since when $x \in \mathcal{G}_u^v$, $\pi(x)$ is not a square matrix in general. Also note that the values of the above character functions depend only on the unitary equivalence class of π, as similar matrices have the same trace. Now if $\pi \in \hat{\mathcal{G}}$, $x \in \mathcal{G}_u^v$, and $f \in L^1(\mathcal{G}_u^v, \lambda_u^v)$, then

$$\mathrm{Tr}\big(\mathfrak{F}_{u,v}(f)(\pi)\pi(x)\big) = \int f(y)\mathrm{Tr}(\pi(y^{-1}x))d\lambda_u^v(y) = f * \chi_u^\pi(x),$$

where in the last equality f is understood to be extended by zero to $\mathcal{G}_u$.

Corollary 7.23. *The map $P_{u,v}^\pi : L^2(\mathcal{G}_u^v, \lambda_u^v) \to \mathcal{E}_{u,v}^\pi$, $f \mapsto d_u^\pi f * \chi_u^\pi$ is a surjective orthogonal projection and for each $f \in L^2(\mathcal{G}_u^v, \lambda_u^v)$, we have the decomposition*

$$f = \sum_{\pi \in \hat{\mathcal{G}}} d_u^\pi f * \chi_u^\pi,$$

which converges in the L^2 norm.

Applying the above decomposition to the case where $u = v$ and $f = \chi_u^\pi$, we get

Corollary 7.24. *For each $u \in X$ and $\pi, \pi' \in \hat{\mathcal{G}}$,*

$$\chi_u^\pi * \chi_u^{\pi'} = \begin{cases} d_u^{\pi\,-1} & \text{if } \pi \sim \pi', \\ 0 & \text{otherwise.} \end{cases}$$

7.2.2 Inverse Fourier and Fourier-Plancherel transforms

Next we are aiming at the construction of the inverse Fourier transform. This is best understood if we start with yet another interpretation of the local Fourier transform. It is clear from the definition of $\mathfrak{F}_{u,v}$ that if $u, v \in X$, $\pi_1, \pi_2 \in \mathcal{R}ep(\mathcal{G})$, and $f \in L^1(\mathcal{G}_u^v, \lambda_u^v)$, then

$$\mathfrak{F}_{u,v}(f)(\pi_1 \oplus \pi_2) = \mathfrak{F}_{u,v}(f)(\pi_1) \oplus \mathfrak{F}_{u,v}(f)(\pi_2),$$

and the same is true for any number (even infinite) of continuous representations, so it follows from (1, Theorem 2.16) that $\mathfrak{F}_{u,v}(f)$ is uniquely characterized by its values on $\hat{\mathcal{G}}$, namely we can regard

$$\mathfrak{F}_{u,v} : L^1(\mathcal{G}_u^v, \lambda_u^v) \to \prod_{\pi \in \hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi),$$

where the Cartesian product is the set of all choice functions $g : \mathcal{G} \to \bigcup_{\pi \in \hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$ with $g(\pi) \in \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$, for each $\pi \in \hat{\mathcal{G}}$. Consider the ℓ^∞-direct sum $\sum_{\pi \in \hat{\mathcal{G}}} \bigoplus \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$. The domain of our inverse Fourier transform then would be the algebraic sum $\sum_{\pi \in \hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$, consisting of those elements of the direct sum with only finitely many nonzero components. An element $g \in \mathcal{D}(\mathfrak{F}_{u,v}^{-1})$ is a choice function such that $g(\pi) \in \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$, for each $\pi \in \hat{\mathcal{G}}$ is zero, except for finitely many π's.

Definition 7.25. Let $u, v \in X$. The inverse Fourier transform

$$\mathfrak{F}_{u,v}^{-1} : \sum_{\pi \in \hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi) \to C(\mathcal{G}_u^v)$$

is defined by

$$\mathfrak{F}_{u,v}^{-1}(g)(x) = \sum_{\pi \in \hat{\mathcal{G}}} \frac{d_u^\pi}{\lambda_u(\mathcal{G}_u^v)} \mathrm{Tr}\big(g(\pi)\pi(x)\big) \quad (x \in \mathcal{G}_u^v).$$

To show that this is indeed the inverse map of the (local) Fourier transform we need a version of the Schur's orthogonality relations (1, Theorem 3.6).

Proposition 7.26. (Orthogonality relations) *Let $\tau, \rho \in \hat{\mathcal{G}}$, $u, v \in X$, $T \in \mathcal{B}(\mathcal{H}_\tau)$, $S \in \mathcal{B}(\mathcal{H}_\rho)$, $A \in B(\mathcal{H}_\rho, \mathcal{H}_\tau)$, and $\xi \in \mathcal{H}_\tau$, $\eta \in \mathcal{H}_\rho$, then*

$$(i) \quad \int \tau(x^{-1})A_{r(x)}\rho(x)d\lambda_u^v(x) = \begin{cases} \frac{\lambda_u(\mathcal{G}_u^v)}{d_u^\tau} Tr(A_u) id_{\mathcal{H}_u^\tau} & \textit{if } \tau = \rho, \\ 0 & \textit{otherwise,} \end{cases}$$

$$(ii) \quad \int \tau(x^{-1})\xi_{r(x)} \otimes \rho(x)\eta_{s(x)} d\lambda_u^v(x) = \begin{cases} \frac{\lambda_u(\mathcal{G}_u^v)}{d_u^\tau} \eta_u \otimes \xi_u & \textit{if } \tau = \rho, \\ 0 & \textit{otherwise,} \end{cases}$$

$$(iii) \quad \int Tr(T_{s(x)}\tau(x^{-1}))Tr(S_{r(x)}\rho(x))d\lambda_u^v(x) = \begin{cases} \frac{\lambda_u(\mathcal{G}_u^v)}{d_u^\tau} Tr(T_u S_u) & \textit{if } \tau = \rho, \\ 0 & \textit{otherwise,} \end{cases}$$

$$(iv) \quad \int Tr(T_{s(x)}\tau(x^{-1}))\rho(x)d\lambda_u^v(x) = \begin{cases} \frac{\lambda_u(\mathcal{G}_u^v)}{d_u^\tau} T_u & \textit{if } \tau = \rho, \\ 0 & \textit{otherwise.} \end{cases}$$

In some applications we need to use the orthogonality relations over $\mathcal{G}_u$ (not $\mathcal{G}_u^v$). In this case, using the normalization $\lambda_u(\mathcal{G}_u) = 1$, and essentially by the same argument we get the following result (3, Proposition 3.3).

Proposition 7.27. (Orthogonality relations) *Let $\tau, \rho \in \hat{\mathcal{G}}$, $u \in X$, $T \in \mathcal{B}(\mathcal{H}_\tau)$, $S \in \mathcal{B}(\mathcal{H}_\rho)$, $A \in B(\mathcal{H}_\rho, \mathcal{H}_\tau)$, and $\xi \in \mathcal{H}_\tau$, $\eta \in \mathcal{H}_\rho$, then*

$$(i) \quad \int \tau(x^{-1})A_{r(x)}\rho(x)d\lambda_u(x) = \begin{cases} \frac{Tr(A_u)}{d_u^\tau} id_{\mathcal{H}_u^\tau} & \textit{if } \tau = \rho, \\ 0 & \textit{otherwise,} \end{cases}$$

$$(ii) \quad \int \tau(x^{-1})\xi_{r(x)} \otimes \rho(x)\eta_{s(x)} d\lambda_u(x) = \begin{cases} \frac{1}{d_u^\tau} \eta_u \otimes \xi_u & \textit{if } \tau = \rho, \\ 0 & \textit{otherwise,} \end{cases}$$

$$(iii) \quad \int Tr(T_{s(x)}\tau(x^{-1}))Tr(S_{r(x)}\rho(x))d\lambda_u(x) = \begin{cases} \frac{1}{d_u^\tau} Tr(T_u S_u) & \textit{if } \tau = \rho, \\ 0 & \textit{otherwise,} \end{cases}$$

$$(iv) \quad \int Tr(T_{s(x)}\tau(x^{-1}))\rho(x)d\lambda_u(x) = \begin{cases} \frac{1}{d_u^\tau} T_u & \textit{if } \tau = \rho, \\ 0 & \textit{otherwise.} \end{cases}$$

Now we are ready to state the properties of the local inverse Fourier transform (3, Proposition 3.4). But let us first introduce the natural inner products on its domain and range. For $f, g \in C(\mathcal{G}_u^v)$ and $h, k \in \sum_{\pi \in \hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$ put $\langle f, g \rangle = \int \bar{f}.g d\lambda_u^v$, and

$$\langle h, k \rangle = \sum_{\pi \in \hat{\mathcal{G}}} \frac{d_u^\pi}{\lambda_u(\mathcal{G}_u^v)} \mathrm{Tr}(k(\pi)h^*(\pi)),$$

where the right hand side is a finite sum as h and k are of finite support. Also note that if $\epsilon_u : C(\mathcal{G}_u^u) \to \mathbb{C}$ is defined by $\epsilon_u(f) = f(u)$, then for each $f, g \in C(\mathcal{G}_u^v)$, we have $g * f^* \in C(\mathcal{G}_u^u)$ and $\langle f, g\rangle = \epsilon_u(g * f^*)$, where $f^* \in C(\mathcal{G}_v^u)$ is defined by $f^*(x) = \overline{f(x^{-1})}$, for $x \in \mathcal{G}_v^u$. Similarly, $h^* \in \sum_{\pi\in\hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$ is defined by $h^*(\pi) = \bar{h}(\check{\pi})$, where $\bar{h}(\pi) = h(\pi)^*$, $\bar{\pi}(x) = \pi(x)^*$, and $\check{\pi}(x) = \pi(x^{-1})^*$, for each $\pi \in \hat{\mathcal{G}}$ and $x \in \mathcal{G}$. The star superscript denotes the conjugation of Hilbert space operators.

Proposition 7.28. *For each $u, v \in X$ and $h, k \in \sum_{\pi\in\hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$, $\ell \in \sum_{\pi\in\hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_w^\pi, \mathcal{H}_v^\pi)$ we have*
$(i)\mathfrak{F}_{u,v}\mathfrak{F}_{u,v}^{-1}(h) = h,$
$(ii)\lambda_u(\mathcal{G}_u^v)\mathfrak{F}_{u,w}^{-1}(hk) = \mathfrak{F}_{u,v}^{-1}(h) * \mathfrak{F}_{v,w}^{-1}(k),$
$(iii)\mathfrak{F}_{v,u}^{-1}(h^*) = (\mathfrak{F}_{u,v}^{-1}(h))^*,$
$(iv)\langle\mathfrak{F}_{u,v}^{-1}(h), \mathfrak{F}_{u,v}^{-1}(k)\rangle = \langle h, k\rangle.$

Next we define a norm on the domain of the inverse Fourier transform in order to get a Plancherel type theorem. Let $u, v \in X$, for $h \in \sum_{\pi\in\hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$ we put $\|h\|_2 = \langle h, h\rangle^{\frac{1}{2}}$. This is the natural norm on the algebraic direct sum, when one endows each component $\mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$ with the Hilbert space structure given by $\langle T, S\rangle = \frac{d_u^\pi}{\lambda_u(\mathcal{G}_u^v)}\mathrm{Tr}(ST^*)$. We denote the completion of $\sum_{\pi\in\hat{\mathcal{G}}} \mathcal{B}(\mathcal{H}_v^\pi, \mathcal{H}_u^\pi)$ with respect to this norm by $\mathcal{L}_{u,v}^2(\mathcal{G})$. The above map is called the (local) *Fourier-Plancherel transform*. The next result is a direct consequence of (3, Proposition 3.2).

Theorem 7.29. (Plancherel Theorem) *For each $u, v \in X$ such that $\lambda_u(\mathcal{G}_u^v) \neq 0$, $\mathfrak{F}_{u,v}$ extends to a unitary $\mathfrak{F}_{u,v} : L^2(\mathcal{G}_u^v, \lambda_u^v) \to \mathcal{L}_{u,v}^2(\mathcal{G})$.*

8. Conclusion

The classical Fourier transform extends to most of the group-like structures. The fast Fourier transform could be computed on finite abelian groups and monoids. There are more sophisticated versions of the Fourier transform on compact groups and hypergroups. The theory of abelian groupoids should be developed with care, but one could safely define the Fourier transform in this case. The theory of Fourier transform on compact groups extends to compact groupoids and a local (bundle-wise) as well as a global Fourier transform is defined in this case.

9. References

[1] Amini, M. (2007). Tannaka-Krein duality for compact groupoids I, Representation theory. *Advances in Mathematics*, 214, 1, 78-91.

[2] Amini, M. (2007). Fourier transform of unbounded measures on hypergroups. *Bull. Ital. Union Math.*, 8, 819-828.

[3] Amini, M. (2010). Tannaka-Krein duality for compact groupoids II, duality. *Operators and Matrices*, 4, 573-592.

[4] Amini, M.; Kalantar, M.; Roozbehani, M.M. (2005). Hidden Sub-hypergroup Problem, Technical Report, Institute for Researches in Theoretical Physics and Mathematics, Tehran.

[5] Amini, M.; Medghalchi, A. (2004). Fourier algebras on tensor hypergrups. *Contemporary Math.*, 363, 1-14.

[6] Amini, M.; Myrnouri, H. (2010). Gelfand transform in abelian groupoid C^*-algebras. Preprint.

[7] Argabright, L.; de Lamadrid, J. (1974). *Fourier Analysis of Unbounded Measures on Locally Compact Abelian Groups*. Memoirs of the American Mathematical Society, No. 145, American Mathematical Society, Providence, R.I.

[8] Baake, M. (2002). Diffraction of weighted lattice subsets. *Canad. Math. Bull.*, 45, 4, 483-498.

[9] Beals, R. (1997). Quantum computation of Fourier transforms over symmetric groups, In: *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, 48-53, El Paso, Texas.

[10] Bloom, R.; Heyer, H. (1995). *Harmonic Analysis of Probability Measures on Hypergroups*. de Gruyter Stud. Math., vol. 20, Walter de Gruyter, Berlin and Hawthorne.

[11] Bluestein, L. I. (1970). A linear filtering approach to the computation of the discrete Fourier transform.*IEEE Trans. Electroacoustics*, 18, 451-455.

[12] Bourbaki, N. (1965). *Elements de Mathematiques, Integration*. Second edition, Herman, Paris.

[13] Bracewell, R. N. (2000). *The Fourier Transform and Its Applications*. 3rd ed., McGraw-Hill, Boston.

[14] Brigham, E. O. (1988). *The Fast Fourier Transform and Applications*. Prentice Hall, Englewood Cliffs.

[15] Brown, D.E. (2007). Efficient classical simulation of the quantum Fourier transform. *New J. Phys.,* 9, 146, 1-7.

[16] Clark, L. O. (2007). Classifying the types of groupoid C^*-algebras, *J. Operator Theory*, 57, 2, 101-116.

[17] Cleve, R. (2004). A note on computing Fourier transforms by quantum programs, http://pages.cpsc.ucalgary.ca/ cleve.

[18] Clausen, M.; Baum, U. (1993). Fast Fourier transforms for symmetric groups: Theory and implementation. *Math. Comput.,* 61, 204, 833-847.

[19] Cooley, J. W.; Tukey, J. W. (1965). An algorithm for machine calculation of complex Fourier series. *Math. Comput.,* 19, 297-301.

[20] Diaconis, P. (1989). A generalization of spectral analysis with application to ranked data. *Ann. Statist.,* 17, 3, 949-979.

[21] Diaconis P.; Rockmore, D. (1990). Efficient computation of the Fourier transform on finite groups. *J. Amer. Math. Soc.,* 3, 2, 297-332.

[22] Donsig, A. P.; Pitts, D.R. (2008). Coordinate systems and bounded isomorphisms, *J. Operator Theory*, 59, 2, 359-416.

[23] Duoandikoetxea, J. (2001). *Fourier Analysis*, American Mathematical Society, Providence.

[24] Ettinger, M.; Hoyer, P. (2000). On quantum algorithms for noncommutative hidden subgroups. *Advances in Applied Mathematics,* 25, 239-251.

[25] Ettinger, M.; Hoyer, P.; Knill, R. (1999). Hidden subgroup states are almost orthogonal, Technical report, quant-ph/9901034.

[26] [E]] Eymard, P. (1964). L'algèbre de Fourier d'un groupe localement compact. *Bull. Soc. Math. France*, 92, 181-236.

[27] Folland, G. (1995). *A Course in Abstract Harmonic Analysis*. Studies in Advanced Mathematics, CRC Press, Boca Raton.

[28] Fourier, J. B. (1822). *Theorie Analytique de la Chaleur*, Paris.

[29] Ghahramani, F.; Medghalchi, A.R. (1985-86). Compact multipliers on weighted hypergroup algebras I,II. *Math. Proc. Camb. Phil. Soc.*, 98, 493-500; 100, 145-149.

[30] Goehle, G. (2009). *Groupoid Crossed Products*. Ph.D dissertation, Dartmouth College.

[31] Grigni, M.; Schulman, L.; Vazirani, M.; Vazirani, U. (2001). Quantum mechanical algorithms for the nonAbelian hidden subgroup problem, In: *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, Crete, Greece, 6-8 July.

[32] Hales, L.; Hallgren, S. (2000). An Improved Quantum Fourier Transform Algorithm and Applications, In: *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 515-525, Redondo Beach, California, 12-14 November.

[33] Hallgren, S.; Russell, A.; Ta-Shma, A. (2000). Normal subgroup reconstruction and quantum computation using group representations, In: *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, 627-635, Portland, Oregon, 21-23 May.

[34] Hewitt, E.; Ross, K. A. (1970). *Abstract Harmonic Analysis*, Vol. II, Die Grundlehren der mathematischen Wissenschaften, Band 152, Springer-Verlag, Berlin.

[35] Halverson, T. (2004). Representations of the q-rook monoid. *J. Algebra*, 273, 227-251.

[36] Jewett, R.I. (1975). Spaces with an abstract convolution of measures. *Advances in Math.*, 18, 1-110.

[37] Ivanyos, G.; Magniez, F.; Santha, M. (2001). Efficient quantum algorithms for some instances of the non-abelian hidden subgroup problem, In: *Proceedings of the Thirteenth Annual ACM Symposium on Parallel Algorithms and Architectures*, 263-270, Heraklion, Crete Island, Greece, 4-6 July.

[38] James, G.; Kerber, A. (1984). *The Representation Theory of the Symmetric Group*, Encyclopedia of Mathematics and its Applications, vol. 16, Cambridge University Press, Cambridge.

[39] Kitaev, A. Y. (1997). Quantum computations: algorithms and error correction. *Russian Mathematical Surveys*, 52, 6, 1191-1249.

[40] Kobler, J.; Schoning, U.; Toran, J. (1993). *The Graph Isomorphism Problem: Its Structural Complexity*. Birkhauser, Boston, MA.

[41] Kumjian, A. (1986). On C^*-diagonals. *Can. J. Math.*, 38, 4, 969-1008.

[42] Kumjian, A.; Muhly, P.S.; Renault J.; Williams, D.P. (1998). The Brauer group of a locally compact groupoid. *J. Amer. Math.*, 120, 901-954.

[43] Lawson, M. V. (1998). *Inverse Semigroups: The Theory of Partial Symmetries*. World Scientific, Singapore.

[44] Lomont, C. (2004). The hidden subgroup problem: review and open problems, quant-ph/0411037.

[45] Malandro, M. (2008). *Fast Fourier Transforms for Inverse Semigroups*, PhD. Thesis, Dartmouth College.

[46] Marquezino, F.L.; Portugal, R.; Sasse, F.D. (2010). Obtaining the Quantum Fourier Transform from the Classical FFT with QR Decomposition, quant-ph/1005.3730v1.

[47] Maslen, D.K. (1998). The efficient computation of Fourier transforms on the symmetric group. *Math. Comput.*, 67, 223, 1121-1147.

[48] Maslen, D.K.; Rockmore, D. N. (1995). Adapted Diameters and FFTs on Groups, Proc. 6th ACM-SIAM SODA, 1995, 253-62.

[49] Maslen, D.K.; Rockmore, D. N. (1997). Generalized FFTs-A survey of some recent results. *Proc. 1995 DIMACS Workshop on Groups and Computation*, 28, 183-238.

[50] Maslen, D.K.; Rockmore, D. N. (1997). Separation of variables and the computation of Fourier transforms on finite groups I. *J. Amer. Math. Soc.*, 10, 1, 169-214.

[51] Maslen, D.K.; Rockmore, D. N. (2001). The Cooley-Tukey FFT and group theory. *Notices of the AMS*, 48, 10, 1151-1161.

[52] Mosca, M.; Ekert, A. (1999). The hidden subgroup problem and eigenvalue estimation on a quantum computer, In: Proceedings if the 1st NASA International Conference on Quantum Computing and Quantum Communications, C.P. Williams (Ed.), Lecture Notes in Computer Science, 1509, 174-188, Springer-Verlag, Berlin.

[53] Munn, W. D. (1957). Matrix representations of semigroups. *Proc. Cambridge Philos. Soc.*, 53, 5-12.

[54] Muhly, P. S.; Renault, J. N.; Williams, D. P. (1996). Continuous-trace groupoid C^*-algebras III. *Trans. Amer. Math. Soc.*, 348, 9, 3621-3641.

[55] Muhly, P. S. (2003). Private communication with M. Amini.

[56] Muhly, P. S.; Williams, D. P. (1992). Continuous-trace groupoid C^*-algebras.II, *Math.Scand.*, 70, 127-145.

[57] Myrnouri, H. (2010). *Abelian Groupoids*. Ph.D. Thesis, Islamic Azad University.

[58] Nielsen, M. A.; Chuang, I. L. (2000). *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge.

[59] Parthasarathy, K. R. (2001). *Lectures on Quantum Computation and Quantum Error Correcting Codes*. Indian Statistical Institute, Delhi Center, Delhi.

[60] Preskill, J. (1998). *Lecture Notes for Physics 229: Quantum Information and Computation*, CIT.

[61] Puschel, M.; Rotteler, M. Beth, T. (1999). Fast quantum Fourier transforms for a class of non-Abelian groups, In: *Proc. 13th AAECC*, LNCS, 1719, 148-159.

[62] Ram, A. (1997). Seminormal representations of Weyl groups and Iwahori-Hecke algebras. *Proc. London Math. Soc.*, 75, 1, 99-133.

[63] Renault, J. (1980). *A Groupoid Approach to C^*-Algebras*. Lecture Notes in Mathematics 793, Springer Verlag, Berlin.

[64] Renault, J. (1983). Two applications of the dual groupoid of a C^*-algebra, In: *Lecture Notes in Mathematics*, 434-445, Springer-Verlag, New York.

[65] Renault, J. (2008). Cartan subalgebras in C^*-algebras. *Irish Math. Soc. Bulletin*. 61, 29-63.

[66] Rhodes, J.; Zalcstein, Y. (1991). Elementary representation and character theory of finite semigroups and its application, In: *Monoids and Semigroups with Applications*, 334-367, World Scitific, River Edge.

[67] Rockmore, D. N. (1995). Fast Fourier transforms for wreath products, *Appl. Comput. Harmon. Anal.*, 2, 279-292.

[68] Rockmore, D. N. (2005). Recent progress and applications in group FFTs, In: *Computational Noncommutative Algebra and Applications*, NATO Science Series, 136, 227-254, Springer, Netherlands.

[69] Rotteler, M.; Beth, T. (1998). Polynomial-time solution to the hidden subgroup problem for a class of non-abelian groups, quant-ph/9812070.

[70] Rudin, W. (1990). *Fourier Analysis on Groups*. Wiley Classics Library, John Wiley & Sons, New York.

[71] Shor, P. W. (1994). Algorithms for quantum computation: Discrete logarithms and factoring, In: *Proceedings of the 35th Annual Symposium on the Foundations of Computer Science*, S. Goldwasser (Ed.), 124-134, IEEE Computer Society, Los Alamitos, CA.

[72] Terence Tao, Fourier transform, UCLA preprints.

[73] Terras, A. (1999). *Fourier Analysis on Finite Groups and Applications.* Cambridge University Press, Cambridge.

[74] Vrem, R.C. (1979). Harmonic analysis on compact hypergroups. *Pacific J. Math.*, 85, 239-251.

[75] Wildberger, N. J. (1993). Finite commutative hypergroups and applications from group theory to conformal field, In: *Applications of Hypergroups and Related Measure Algebras*, Contemp. Math. 183, 413-434, Proceedings Seattle Conference.

[76] Yates, F. (1937). The design and analysis of factorial experiments. *Imp. Bur. Soil Sci. Tech. Comm.*, 35.

[77] Young, A. (1977). The Collected Papers of Alfred Young, 1873-1940, University of Toronto Press, Toronto.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

19

Reduced Logic and Low-Power FFT Architectures for Embedded Systems

Erdal Oruklu, Jafar Saniie and Xin Xiao
Illinois Institute of Technology
USA

1. Introduction

Discrete Fourier Transform (DFT) is one of the core operations in digital signal processing and communication systems. Many fundamental algorithms can be realized by DFT, such as convolution, spectrum estimation, and correlation. Furthermore, DFT is widely used in standard embedded system applications such as wireless communication protocols requiring Orthogonal Frequency Division Multiplexing (Wey et al., 2007), and radar image processing using Synthetic Aperture Radar (Fanucci et al., 1999). In practice, DFT is difficult to implement directly due to its computational complexity. To reduce the degree of computation, Cooley and Tukey proposed the well-known Fast Fourier Transform (FFT) algorithm, which reduces the calculation of N-point DFT from O(N^2) to O($N/2log_2N$). (Proakis & Manolakis, 2006). Nevertheless, for embedded systems, in particular portable devices; efficient hardware realization of FFT with small area, low-power dissipation and real-time computation is a significant challenge. The challenge is even more pronounced when FFTs with large transform lengths (>1024 points) need to be realized in embedded hardware. Therefore, the objective of this research is to investigate hardware efficient FFT architectures, emphasizing compact, low-power embedded realizations.

As VLSI technology evolves, different architectures have been proposed for improving the performance and efficiency of the FFT hardware. Pipelined architectures are widely used in FFT realization (Li & Wanhammar, 1999; He & Torkelson, 1996; Hopkinson & Butler, 1992; Yang et al., 2006) due to their speed advantages. Higher radix (Hopkinson & Butler, 1992; Yang et al., 2006) and multi-butterfly (Bouguezel et al., 2004; X. Li et al., 2007) structures can also improve the performance of the FFT processor significantly, but these structures require substantially more hardware resources. Alternatively, shared memory based schemes with a single butterfly calculation unit (Cohen, 1976; Ma, 1994, 1999; Ma & Wanhammar, 2000; Wang et al., 2007) are preferred in many embedded FFT processors since they require least amount of hardware resources. Furthermore, "in-place" addressing strategy is a practical choice to minimize the amount of data memory. With "in-place" strategy, the two outputs of the butterfly unit can be written back to the same memory locations of the two inputs, and replace the old data. For in-place FFT processing, two data read and two data write operations occur at every clock cycle. Multiple memory banks and conflict-free addressing logic are required to realize four data accesses in one clock cycle. Consequently, a typical FFT processor is composed of three major components: i) butterfly calculation units, ii) conflict free address generators for both data and coefficient accesses and iii) multi-bank memory units.

In this study, several techniques are developed for reducing the hardware logic and power requirements for these three components:

1. In order to optimize the conflict free addressing logic, a modified butterfly structure with input/output exchange circuits is presented in Section 2.
2. CORDIC based FFT algorithms are presented for multiplier-less and coefficient memory-less implementation of the butterfly unit in Section 3.
3. Memory bank partitioning and bitline segmentation techniques are presented for dynamic power reduction of data memory accesses. Furthermore, a special coefficient memory addressing logic which reduces the switching activity is proposed in Section 4.

Case studies with ASIC and FPGA synthesis results demonstrate the performance gains and feasibility of these FFT implementations on embedded systems.

2. Hardware efficient realization of fast Fourier transform

There is an ongoing interest in hardware efficient FFT architectures. Cohen (Cohen, 1976) introduced a simplified control logic for FFT address generation, which is composed of parity checks, barrel shifters and counters based on the fact that two data addresses of every butterfly operations differ in their parity. Ma (Ma, 1999) proposed a method to realize the radix-2 addressing logic which reduces the address generation delay by avoiding parity check (XOR operations), but barrel shifters are still needed. Furthermore, Ma's approach is not "in-place", so more registers and related control logic are needed to buffer the interim data to avoid the memory conflict. Yang (Yang et al., 2006) proposed a locally pipelined radix-16 FFT realized by two radix-2 deep feedback (R2SD^{2}F) butterflies. This architecture can improve the throughput of the FFT processing and reduce the complex multipliers and adders compared to other pipelined methods, but it needs extra memory and there is significantly more coefficient access due to radix-16 implementation. Li (X. Li et al., 2007) proposed a mixed radix FFT architecture, which contains one radix-2 butterfly and one radix-4 butterfly. The two butterflies share the multipliers, which reduce the hardware consumption, but the address generation is based on XOR logic, and similar to Cohen's design. Next section describes in detail addressing schemes that emphasize reduced hardware.

2.1 Conflict-free addressing for FFT

The N-point discrete Fourier transform is defined by

$$X(k)=\sum_{n=0}^{N-1} x(n)W_N^{nk} \qquad k=0,1,...,N-1, \quad W_N^{nk}=e^{-j\frac{2\pi}{N}nk} \tag{1}$$

Fig. 1 shows the signal flow graph of 16-point decimation-in-frequency (DIF) radix-2 FFT (Proakis & Manolakis, 2006). FFT algorithm is composed of butterfly calculation units:

$$x_{m+1}(p)=x_m(p)+x_m(q) \tag{2}$$

$$x_{m+1}(q)=[x_m(p)-x_m(q)]W_N^r \tag{3}$$

Equations (2), (3) describe the radix-*2* butterfly calculation at Stage m as shown in Fig. 2. Parallel and "in-place" butterfly operation using two memory banks of two-port memory units requires that the two inputs of any butterfly are read from different banks of memory and the two outputs are written to the same address locations as the inputs. As shown in Fig. 1, in the conventional FFT addressing scheme, only the butterflies in the first stage satisfy this requirement. Two inputs and two outputs of butterfly operations in all other stages are originating from and sinking to the same memory bank. Therefore, a special addressing scheme is required to prevent the conflicting addresses.

Cohen (Cohen, 1976) used parity check to separate the data into two memory banks. Fig. 3 is the signal flow graph of Cohen's approach and it shows that inputs and outputs of any butterfly stage utilize separate memory banks. The addresses of butterfly operations are "in-place" located. The drawback of Cohen's method is the address generation delay. In order to reduce the delay of the address generation, Ma (Ma, 1999) proposed an alternative addressing scheme which avoids using parity check. The signal flow graph of Ma's scheme is shown in Fig. 4. In Ma's scheme, two inputs of a butterfly unit originate from two separate memory banks but two outputs of the butterfly unit utilize the same memory bank. The inputs and outputs of a butterfly unit are not "in-place". Therefore, extra registers and related control logic are needed to buffer the outputs of the butterfly until next butterfly calculation is finished in order to realize the "in place" operation. Compared to Cohen's approach which uses both parity check and barrel shifters, Ma's method needs only barrel shifters and avoids parity check, resulting in a reduced address generation delay. However, Ma's approach consumes more hardware resources to realize the "in-place" operation.

In the following section, a hardware efficient FFT engine with reduced critical path delay is proposed. Addressing logic is reduced by using a butterfly structure which modifies the conventional one by adding exchange circuits at the input and output of the butterfly (Xiao, et al., 2008]. With this butterfly structure, the two inputs and two outputs of any butterfly can be exchanged; hence all data addresses in FFT processing can be reordered. Using this flexible input and output ordering, addressing logic is designed to be "in-place" and it does not need barrel shifters.

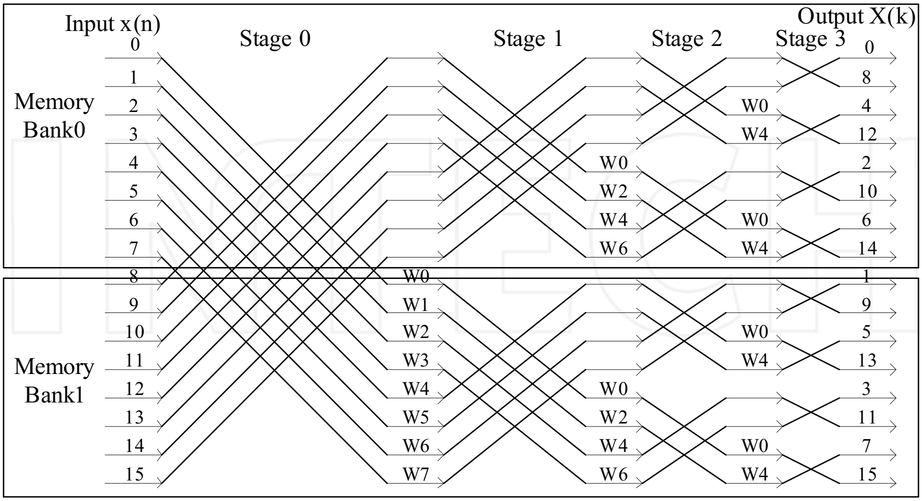


Fig. 1. Signal flow graph of 16-point FFT

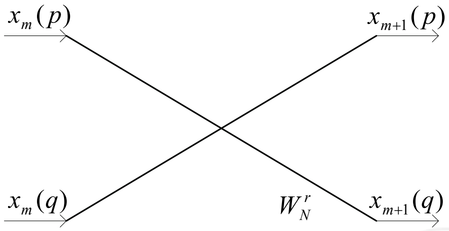


Fig. 2. Butterfly unit at stage m

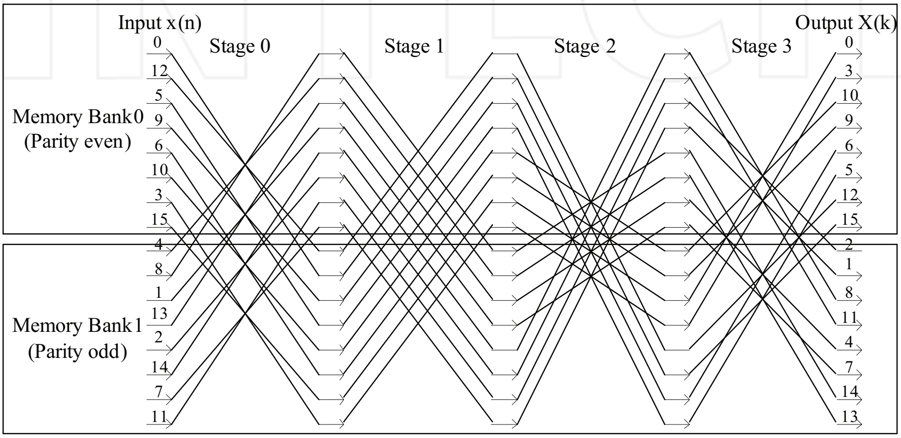


Fig. 3. Signal flow graph of 16-point FFT using Cohen's method (Cohen, 1976)

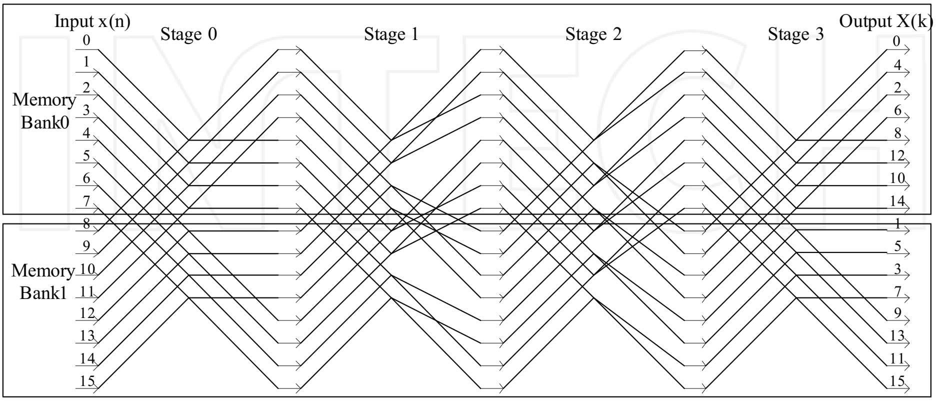


Fig. 4. Signal flow graph of 16-point FFT using Ma's method (Ma, 1999)

2.2 Reduced address generation logic with the modified butterfly FFT (mbFFT)

This addressing scheme is based on a **m**odified **b**utterfly **FFT** (mbFFT) structure, which is shown in Fig. 5. The main difference between the modified butterfly structure and the conventional one is the addition of two exchange circuits that are placed at both the input and the output of the butterfly unit. Each exchange circuit is composed of two (2:1) multiplexers; when the exchange control signal *C1* or *C2* is *1*, the data will be exchanged, otherwise they keep their locations.

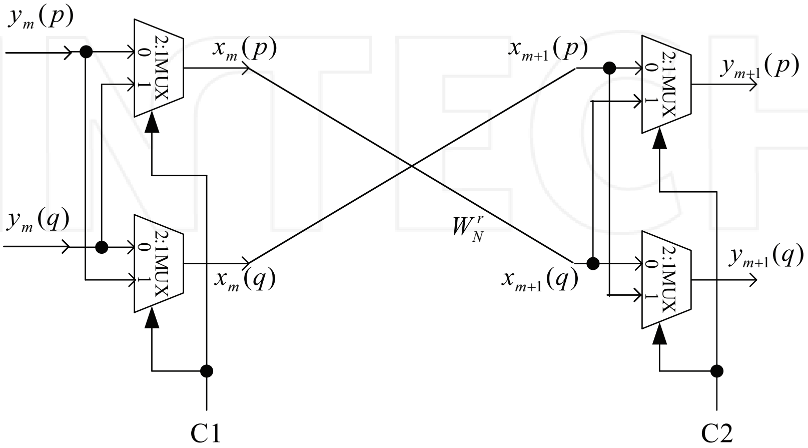


Fig. 5. Modified butterfly structure

Equation (4) shows the function:

If C1=1: $x_m(p) = y_m(q), \quad x_m(q) = y_m(p);$

Else: $x_m(p) = y_m(p), \quad x_m(q) = y_m(q);$

If C2=1: $y_{m+1}(p) = x_{m+1}(q), \quad y_{m+1}(q) = x_{m+1}(p);$

else: $y_{m+1}(p) = x_{m+1}(p), \quad y_{m+1}(q) = x_{m+1}(q);$ (4)

Based on this butterfly structure, all data within the FFT processing can be reordered by setting the different values of the exchange control signals *C1* and *C2*. The control signals are chosen such that the input data always originate from two separate memory banks and output data are written to the same memory location in order to achieve in-place operation.

2.2.1 *16*-point mbFFT implementation

For 16-point mbFFT, the signal flow graph is shown in Fig. 6. In the figure, the butterfly inputs or outputs indicated by broken lines denote that the data have been exchanged. Fig. 7 shows the complete address generation architecture and components for 16-point FFT implementation. The address generation logic is composed of a 5-bit counter *D*, three inverters, a 3-bit shifter, three (2:1) multiplexers, two (4:1) multiplexers, four multi-bit (2:1) multiplexers and delay elements. *Stage Counter* S indicates which stage of FFT is currently in progress and controls the two (4:1) multiplexers to generate the correct exchange control signals *C1* and *C2* for the butterfly operation. The 3-bit shifter shifts one bit at each stage and it controls three (2:1) multiplexers to generate the correct *M1* address. Since this technique is "in-place", the addresses for read and write are same with the exception of a delay introduced for compensating the butterfly computation time. Table I presents the counter values (control logic) which are used to generate the addresses for *M0* and *M1* memory banks.

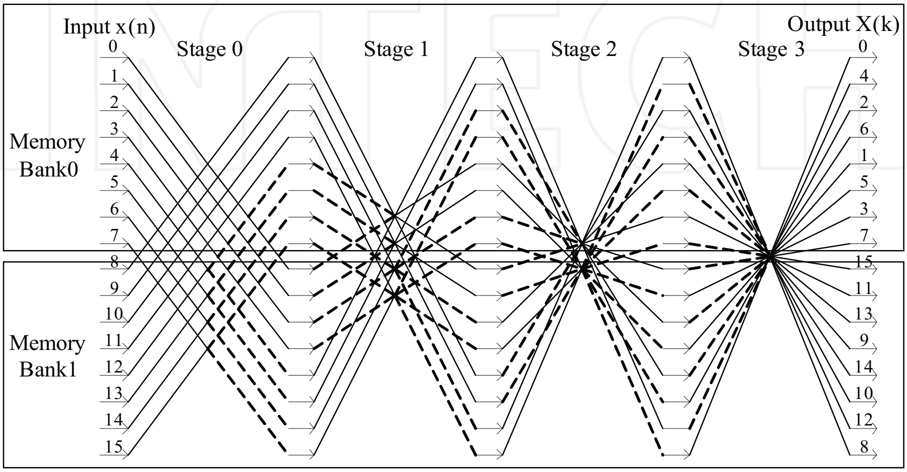


Fig. 6. Signal flow graph of 16-point mbFFT

Counter $B(b_2b_1b_0)$	Counter $\overline{B}(\overline{b}_2\overline{b}_1\overline{b}_0)$	Stage 0 (exchange control signal: C1=0,C2=b_2)		Stage 1 (exchange control signal: C1= b_2,C2= b_1)		Stage 2 (exchange control signal: C1= b_1,C2= b_0)		Stage 3 (exchange control signal: C1= b_0,C2=0)	
		Bank0 address $b_2b_1b_0$	Bank1 address $b_2b_1b_0$	Bank0 address $b_2b_1b_0$	Bank1 address $\overline{b}_2b_1b_0$	Bank0 address $b_2b_1b_0$	Bank1 address $\overline{b}_2\overline{b}_1b_0$	Bank0 address $b_2b_1b_0$	Bank1 address $\overline{b}_2\overline{b}_1\overline{b}_0$
000	111	000	000	000	100	000	110	000	111
001	110	001	001	001	101	001	111	001	110
010	101	010	010	010	110	010	100	010	101
011	100	011	011	011	111	011	101	011	100
100	011	100	100	100	000	100	010	100	011
101	010	101	101	101	001	101	011	101	010
110	001	110	110	110	010	110	000	110	001
111	000	111	111	111	011	111	001	111	000

Table 1. Address generation table for the 16-point mbFFT

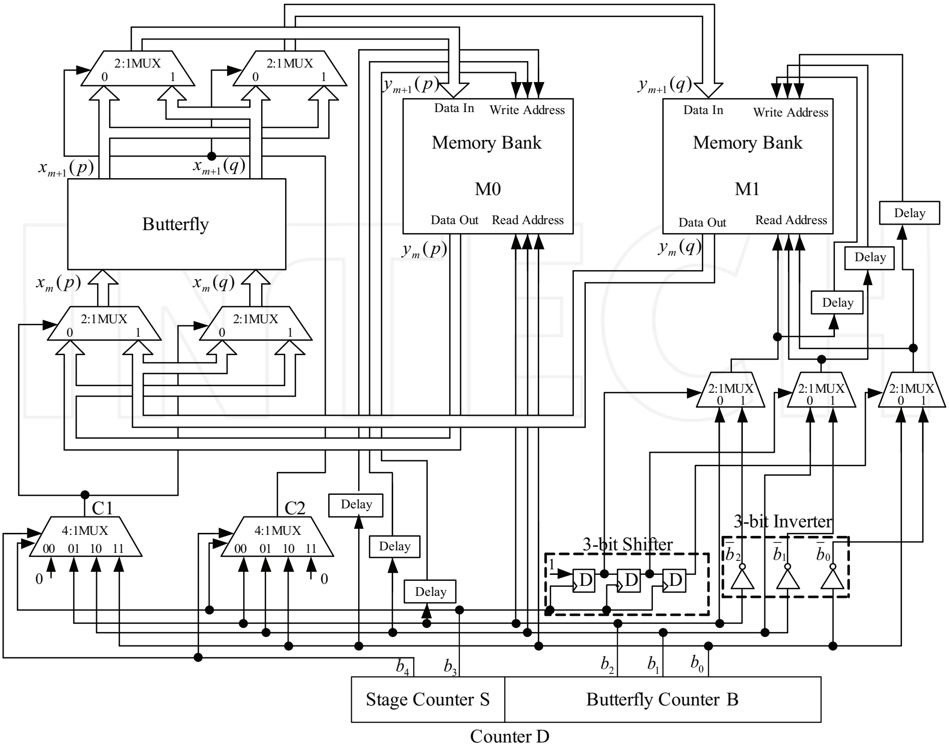


Fig. 7. Address generation circuits for 16-point mbFFT

2.2.2 *N*-point mbFFT implementation

In order to generalize the addressing scheme for $N = 2^n$ - point FFT, the necessary circuit components of the addressing and control logic can be listed as follows:

- (n-1)-bit *Butterfly Counter* $B = b_{n-2}b_{n-3}...b_1b_0$,
- (n-1) inverters which generate the complement of the *Butterfly Counter* $\overline{B} = \overline{b}_{n-2}\overline{b}_{n-3}...\overline{b}_1\overline{b}_0$ from counter B ,
- $\lceil \log_2 n \rceil$ - bit *Stage Counter* $S = (n-1),...,2,1,0$.
- *Two memory banks, Bank 0 (M0) and Bank 1 (M1).*

In practice, *Stage Counter S* and *Butterfly Counter B* can be combined to a single *counter D*, where *B* is the least significant (n-1) bits of *counter D,* and *S* is the most significant $\lceil \log_2 n \rceil$ bits of *counter D*. At any time, the read and write addresses of *M0* is exactly same as the value of *Butterfly Counter B*. For *M1*, the read and write address at Stage s is $\overline{b}_{n-2}\overline{b}_{n-3}...\overline{b}_{n-s-1}b_{n-s-2}...b_1b_0$, which is a combination of counters B and $\overline{B}$. The exchange control signal *C1* is equal to b_{n-s-1} (assume $b_{n-1} \equiv 0$), and *C2* is equal to b_{n-s-2} (assume $b_{-1} \equiv 0$). The address of twiddle factors at stage s is given by $b_{n-s-2}b_{n-s-3}...b_0 0...0$ (s '0's).

2.3 VLSI synthesis results

The mbFFT architecture is synthesized using TSMC CMOS 0.18µm technology. Synthesis is performed with Cadence Build Gates and Encounter tools. The synthesis results for 16-point FFT with 32-bit complex number input show a maximum clock frequency of 280MHz with 0.665mm^2 area and 0.645mW total power consumption for the complete FFT operation including butterfly unit, address generation unit, and memory circuits.

In order to compare different FFT addressing methods, the logic complexity can be evaluated similar to (Ma, 1999), based on gate counts. The sizes of some basic circuits and gates are listed in Table 2. Estimated gate count comparison for 1024-point FFT of 32-bit complex data (16-bit each for the real part and imaginary part) is shown in the Table 3. In terms of area, mbFFT scheme requires 24% fewer number of transistors. This reduction is mainly due to the difference in logic complexity of the multiplexers and barrel shifters. Based on the gate counts in Table 2 (and confirmed by synthesis results), *r*-input (r:1) multiplexer is approximately 4 times smaller than *(r-1)* barrel shifter in terms of area.

The delay of address generation for both read and write operations in the mbFFT addressing scheme is determined by two stages of multiplexers, where the first stage uses an *r*-input (r:1) multiplexer and the second stage uses a *2*-input (2:1) multiplexer for a 2^r-point FFT operation (see Fig 7). In (Ma, 1999), worst-case address generation delay is dominated by an *(r-1)*-bit barrel shifter and a (2:1)-multiplexer. An *(r-1)*-bit barrel shifter requires $\lceil \log_2(r-1) \rceil$ stages of (2:1) multiplexers in the critical path. Cohen's address generation method (Cohen, 1976) uses an *r*-bit parity check unit, an *(r-1)*-bit barrel shifter, and two (2:1) multiplexers in the critical path. Standard cell synthesis results in Table 4 show that the proposed mbFFT address generation scheme is faster compared to (Cohen, 1976) and (Ma, 1999) for large FFTs, due to the complex wiring and parasitic capacitances in barrel shifters and elimination of the parity-check operation.

Compared to a pipelined FFT architecture such as R2SD^{2}F given in (Yang et al., 2006), the shared memory architectures such as mbFFT offer significantly reduced hardware cost and power consumption at the expense of (slower) throughput. R2SD^{2}F requires log_4N-1 multipliers, $2log_4N$ adders and $10log_4N$ multiplexers for the butterfly operations in an *N*-point FFT. In contrast, only one multiplier, two adders and four multiplexers are used in the mbFFT architecture datapath. The latency (total clock cycles) of a pipelined FFT architecture is faster by a factor of $\frac{1}{2}\log_2 N$. However, the maximum achievable clock frequency would be less than the mbFFT design due the increased complexity of the R2SD^{2}F datapath and address generation. Hence, for embedded applications, the proposed reduced logic, shared memory FFT approach with modified butterfly units presents a more viable solution.

Types of Gates and Circuits	No. of. Transistors
2-Input XOR	10
2-1 Multiplexer	6
10-1 Multiplexer	42
1-bit Register/Latch	10
9-bit Counter	182
13-bit Counter	270
9-bit Barrel Shifter	152
10-bit Barrel Shifter	168

Table 2. Transistor counts for CMOS cells (Ma, 1999)

Design Schemes	Components		Transistor Counts
	Quantity	Type	
Proposed mbFFT Design	1	13-bit Counter	1562
	9	Inverters	
	1	9-bit Shifter	
	9	1-bit 2:1 Multiplexer	
	2	1-bit 10:1 Multiplexer	
	4	32-bit 2:1 Multiplexer	
	2	9-bit Latches	
(Ma, 1999)	1	13-bit Counter	2066
	2	9-bit Barrel Shifters	
	4	9-bit Latches	
	2	32-bit Latches	
	2	9-bit 2:1 Multiplexers	
	2	32-bit 2:1 Multiplexers	
(Cohen, 1976)	1	13-bit Counter	1924
	1	9-bit Counter	
	2	9-bit Latch	
	2	10-bit Barrel Shifter	
	2	9-bit 2:1 Multiplexer	
	4	32-bit 2:1Multiplexer	
	1	9-bit Address Parity Generator	

Table 3. Address generation logic comparison for 1024-point FFT with 32-bit complex data

FFT size =2^n	Proposed mbFFT	(Ma, 1999)	(Cohen,1976)
n=4	1.28 ns	1.28 ns	1.82 ns
n=8	1.40 ns	1.53 ns	2.50 ns
n=10	1.47 ns	1.71 ns	2.61 ns
n=16	1.59 ns	1.85 ns	2.87 ns

Table 4. Delay comparison of address generation circuits

3. Multiplierless FFT architectures using CORDIC algorithm

In FFT processors, butterfly operation is the most computationally demanding stage. Traditionally, a butterfly unit is composed of complex adders and multipliers. A complex multiplier can be very large and it is usually the speed bottleneck in the pipeline of the FFT processor. The Coordinate Rotation Digital Computer (CORDIC) (Volder, 1959) algorithm is an alternative method to realize the butterfly operation without using any dedicated multiplier hardware. CORDIC algorithm is versatile and hardware efficient since it requires only add and shift operations, making it suitable for the butterfly operations in FFT (Despain, 1974). Instead of storing actual twiddle factors in a ROM, the CORDIC-based FFT processor needs to store only the twiddle factor *angles* in a ROM for the butterfly operation. In recent years, several CORDIC-based FFT designs have been proposed for different applications (Abdullah et al., 2009; Lin & Wu, 2005; Jiang, 2007; Garrido & Grajal, 2007). In (Abdullah et al., 2009), non-recursive CORDIC-based FFT was proposed by replacing the twiddle factors in FFT architecture by non-iterative CORDIC micro-rotations. It reduces the ROM size, however, it does not eliminate it completely. (Lin & Wu, 2005) proposed a "mixed-scaling-rotation" CORDIC algorithm to reduce the total iterations, but it increases the hardware complexity. (Jiang, 2007) introduced Distributed Arithmetic (DA) to the CORDIC-based FFT algorithms, but the DA look-up tables are costly in implementation. (Garrido & Grajal, 2007) proposed a memory-less CORDIC algorithm to reduce the memory requirements for a CORDIC-based FFT processor by using only shift operations for multiplication.

Conventionally, a CORDIC-based FFT processor needs a dedicated memory bank to store the necessary twiddle factor angles for the rotation. In our earlier work (Xiao et al., 2010), a modified CORDIC algorithm for FFT processors is proposed which eliminates the need for storing the twiddle factor angles. The algorithm generates the twiddle factor angles successively by an accumulator. With this approach, memory requirements of an FFT processor can be reduced by more than 20%. Memory reduction improves with the increasing radix size. Furthermore, the angle generation circuit consumes less power consumption than angle memory accesses. Hence, the dynamic power consumption of the FFT processor can be reduced by as much as 15%. Since the critical path is not modified with the CORDIC angle calculation, system throughput does not change.

In the following sections, CORDIC algorithm fundamentals and the design of the proposed memory efficient CORDIC-based FFT processor are described.

3.1 CORDIC algorithm

CORDIC algorithm was proposed by J.E. Volder (Volder, 1959). It is an iterative algorithm to calculate the rotation of a vector by using only additions and shifts. Fig. 8 shows an example for rotation of a vector V_i.

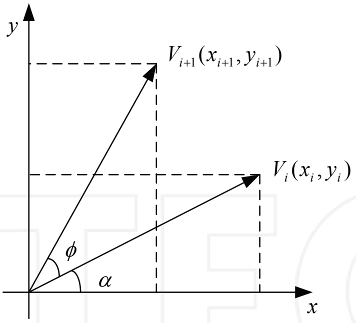


Fig. 8. Rotate vector $V_i(x_i, y_i)$ to $V_{i+1}(x_{i+1}, y_{i+1})$

The following equations illustrate the steps for calculating the rotation:

$$\begin{aligned} x_{i+1} &= r\cos(\alpha+\phi) = r(\cos\alpha\cos\phi - \sin\alpha\sin\phi) \\ &= x_i\cos\phi - y_i\sin\phi \end{aligned} \tag{5}$$

$$\begin{aligned} y_{i+1} &= r\sin(\alpha+\phi) = r(\sin\alpha\cos\phi + \cos\alpha\sin\phi) \\ &= y_i\cos\phi + x_i\sin\phi \end{aligned} \tag{6}$$

If each rotate angle ϕ is equal to $\arctan 2^{-i}$, then:

$$x_{i+1} = \cos\phi(x_i - y_i \cdot 2^{-i}) \quad (7)$$

$$y_{i+1} = \cos\phi(y_i + x_i \cdot 2^{-i}) \quad (8)$$

Since $\phi = \arctan 2^{-i}$, $\cos\phi$ can be simplified to a constant with fixed number of iterations:

$$x_{i+1} = K_i(x_i - y_i \cdot d_i \cdot 2^{-i}) \quad (9)$$

$$y_{i+1} = K_i(y_i + x_i \cdot d_i \cdot 2^{-i}) \quad (10)$$

where $K_i = \cos(\arctan(2^{-i}))$ and $d_i = \pm 1$. Product of K_i's can be represented by the K factor which can be applied as a single constant multiplication either at the beginning or end of the iterations. Then, (9) and (10) can be simplified to:

$$x_{i+1} = x_i - y_i \cdot d_i \cdot 2^{-i} \quad (10)$$

$$y_{i+1} = y_i + x_i \cdot d_i \cdot 2^{-i} \quad (11)$$

The direction of each rotation is defined by d_i and the sequence of all d_i 's determines the final vector. d_i is given as*:*

$$d_i = \begin{Bmatrix} -1 & \text{if } \mathrm{z_i} < 0 \\ +1 & \text{if } \mathrm{z_i} \geq 0 \end{Bmatrix} \quad (12)$$

where $\mathrm{z_i}$ is called angle accumulator and given by

$$z_{i+1} = (z_i - d_i \cdot \arctan 2^{-i}) \quad (13)$$

All operations described through equations (10)-(13) can be realized with only additions and shifts; therefore, CORDIC algorithm does not require dedicated multipliers. CORDIC algorithm is often realized by pipeline structures, leading to high processing speed. Fig. 9 shows the basic structure of a pipelined CORDIC unit.

As shown in equation (1), the key operation of FFT is $x(n) \cdot W_N^{nk}$, ($W_N^{nk} = e^{-j\frac{2\pi}{N}nk}$). This is equivalent to "Rotate $x(n)$ by angle $-\frac{2\pi}{N}nk$" operation which can be realized easily by the CORDIC algorithm. Without any complex multiplications, CORDIC-based butterfly can be fast. An FFT processor needs to store the twiddle factors in memory. CORDIC-based FFT doesn't have twiddle factors but needs a memory bank to store the rotation angles. For radix-2, N-point, m-bit FFT, $\frac{mN}{2}$ bits memory needed to store $\frac{N}{2}$ angles. In the next section, a new CORDIC based FFT design which does not require any twiddle factor or angle memory units is presented. This design uses a single accumulator for generating all the necessary angles instantly and does not have any precision loss.

3.2 Reduced memory CORDIC based FFT

Although several multi-bank addressing schemes have been used to realize parallel and pipelined FFT processing (Ma, 1999; Xiao et al., 2008), these methods are not suitable for the reduced memory CORDIC FFT. In these schemes, the twiddle factor angles are not in regular increasing order (see Table 5), resulting in a more complex design for angle generators. As shown in Table 6, using a special addressing scheme first proposed in (Xiao et al., 2009), the twiddle factor angles follow a regular, increasing order, which can be generated by a simple accumulator. Table 6 shows the address generation table of the 16-point radix-2 FFT. It can be seen that twiddle factor angles are sequentially increasing, and every angle is a multiple of the basic angle $2\pi/N$, which is $\pi/8$ for 16-point FFT. For different FFT stages, the angles increase always one step per clock cycle. Hence, an angle

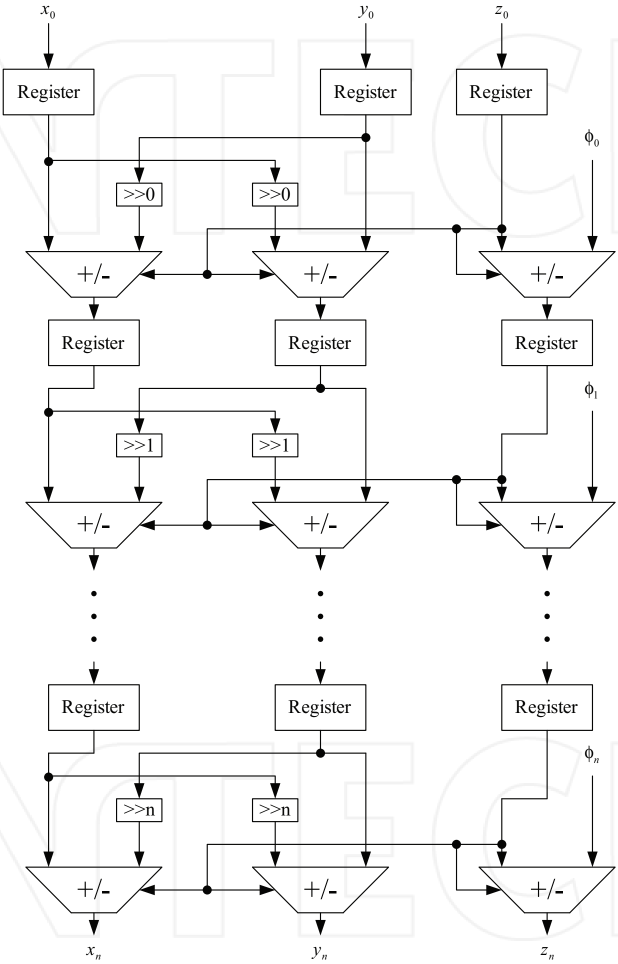


Fig. 9. Basic structure of a pipelined CORDIC unit

generator circuit composed of an accumulator, and an output latch can realize this function, as shown in Fig. 10. Control signal for the latch that enables or disables the accumulator output is simple and it is based on the current FFT butterfly stage and RAM address bits *$b_2b_1b_0$* (see Table 6).

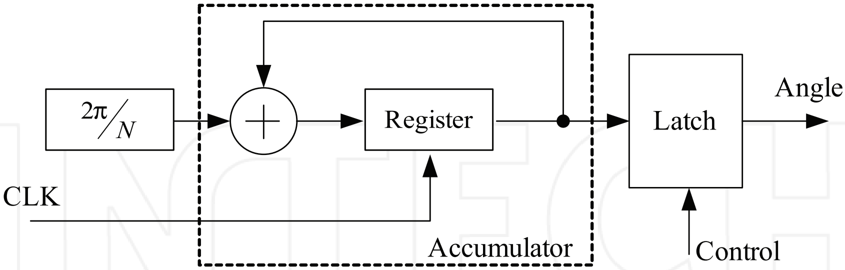


Fig. 10. Angle generator for the CORDIC based FFT

Butterfly Counter B(b2b1b0)	**Stage 0**		**Stage 1**		**Stage 2**		**Stage 3**	
	RAM address b0b2b1	**Twiddle factor angle**	**RAM address b1b0b2**	**Twiddle factor angle**	**RAM address b2b1b0**	**Twiddle factor angle**	**RAM address b0b2b1**	**Twiddle factor angle**
000	000	0	000	0	000	0	000	0
001	100	$4\pi/8$	010	$4\pi/8$	001	$4\pi/8$	100	0
010	001	$\pi/8$	100	0	010	0	001	0
011	101	$5\pi/8$	110	$4\pi/8$	011	$4\pi/8$	101	0
100	010	$2\pi/8$	001	$2\pi/8$	100	0	010	0
101	110	$6\pi/8$	011	$6\pi/8$	101	$4\pi/8$	110	0
110	011	$3\pi/8$	101	$2\pi/8$	110	0	011	0
111	111	$7\pi/8$	111	$6\pi/8$	111	$4\pi/8$	111	0

Table 5. Address generation table of Ma's (Ma, 1999) design for 16-point radix-2 FFT

Fig. 11 shows the architecture of the proposed *no-twiddle-factor-memory* design for radix-2 FFT. Four registers and eight 2-to-1 multiplexers are used. Registers are needed before and after the butterfly unit to buffer the intermediate data in order to group two sequential butterfly operations together. Therefore, the conflict-free "in-place" data accessing can be realized. This register-buffer design can be extended to any radix FFTs. For radix-2, the structure can be simplified by using just 4 registers, but for radix-r FFT, $2 \times r^2$ registers are needed. Fig. 12 shows the structure for radix-*r* FFT.

Butterfly Counter B(b2b1b0)	Stage 0		Stage 1		Stage 2		Stage 3	
	RAM address b2b1b0	Twiddle factor angle	RAM address b0b2b1	Twiddle factor angle	RAM address b1b0b2	Twiddle factor angle	RAM address b2b1b0	Twiddle factor angle
000	000	0	000	0	000	0	000	0
001	001	$\pi/8$	100	0	010	0	001	0
010	010	$2\pi/8$	001	$2\pi/8$	100	0	010	0
011	011	$3\pi/8$	101	$2\pi/8$	110	0	011	0
100	100	$4\pi/8$	010	$4\pi/8$	001	$4\pi/8$	100	0
101	101	$5\pi/8$	110	$4\pi/8$	011	$4\pi/8$	101	0
110	110	$6\pi/8$	011	$6\pi/8$	101	$4\pi/8$	110	0
111	111	$7\pi/8$	111	$6\pi/8$	111	$4\pi/8$	111	0

Table 6. Address generation table for 16-point radix-2 FFT with the proposed angle generator

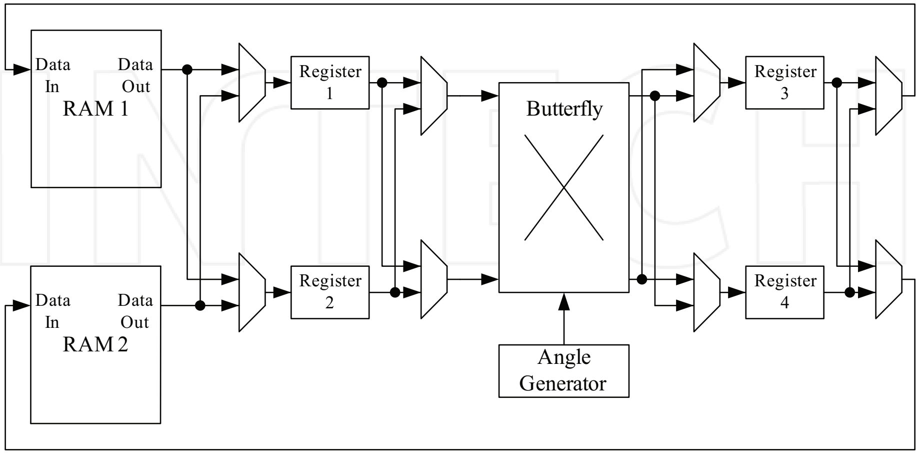


Fig. 11. Radix-2 FFT processor with no-twiddle-factor-memory

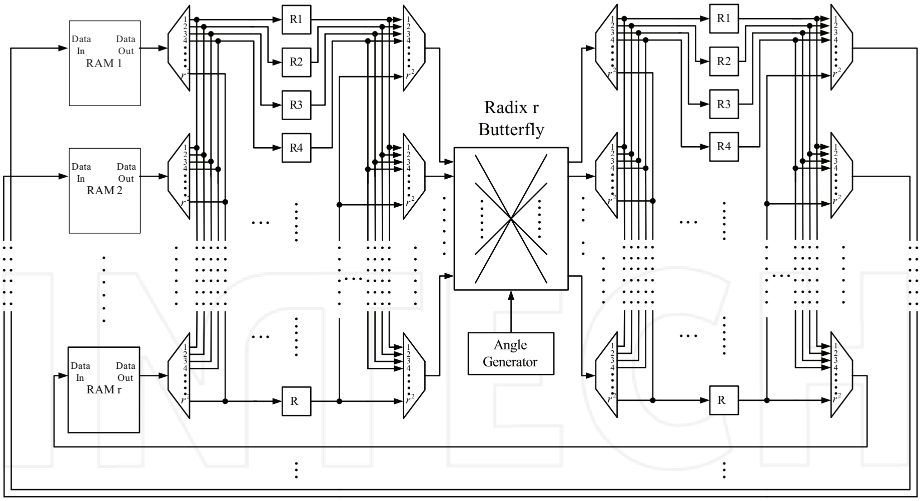


Fig. 12. Proposed radix-r CORDIC-based FFT

For an $N = 2^n$-point FFT, the addressing and control logic are composed of several components: An $(n-1)$-bit butterfly counter $B = b_{n-2}b_{n-3}...b_1b_0$ will provide the address sequences and the control logic of the angle generator. In stage S, the memory address is given by $b_{s-1}b_{s-2}...b_1b_0b_{n-2}b_{n-3}...b_s$, which is rotate right S bits of butterfly counter B. Meanwhile, the control logic of the latch of the angle generator is determined by the sequence of the pattern; $b_{n-2}b_{n-3}...b_s0...0$ (S "0"s).

For radix-2, $N = 2^n$-point, *m*-bit FFT, (each data is *2m*-bit complex number; *m*-bit each for the real part and imaginary part) by using the proposed angle generator, $\frac{5mN}{2}$ bits memory required by the conventional CORDIC can be reduced to $\frac{4mN}{2}$ which corresponds to 20% reduction. For higher radix FFT, the reduction is even more significant. For radix-r FFT, the saving is $\frac{(r-1)mN}{r}$ bits out of $\frac{(3r-1)mN}{r}$, which converges to 33.3% reduction.

Due to finite wordlength, as the accumulator operates, the precision loss will accumulate as well. In order to address this issue, more bits (wider wordlength) can be used for the fundamental angle *2π/N* and the accumulator logic. For example, for 1024-point FFT, the accumulator is extended from 16 bits to 21 bits and no precision loss is observed compared to a conventional angle-stored CORDIC FFT processor.

3.3 FPGA synthesis results

The proposed reduced memory CORDIC based FFT designs for both radix-2 and radix-4 FFT algorithms have been realized by Verilog-HDL and implemented on an FPGA chip (STRATIX-III EP3SE50C2). Synthesis results shown in Table 7 show that these designs can reduce memory usage for FFT processors without any tangible increase in the number of logic elements used when compared against the conventional CORDIC implementation (i.e., angles are stored in memory). Furthermore, dynamic power consumption is reduced (up to 15%) with no delay penalties. The synthesis results match with the theoretical analysis.

		Radix-2		Radix-4	
		Proposed CORDIC FFT (angle generator)	Conventional CORDIC FFT (angles stored)	Proposed CORDIC FFT (angle generator)	Conventional CORDIC FFT (angles stored)
256-point FFT	Total logic elements	1,427 (19-bit accum.)	1,386	5,892 (20-bit accum.)	5,763
	Total memory	8,672	10,720	8,728	11,800
	Dynamic Power	136.87 mW	156.22mW	437.53 mW	495.06 mW
1024-point FFT	Total logic elements	1,773 (21-bit accum.)	1,718	5,991 (22-bit accum.)	5,797
	Total memory	33,248	41,440	33,304	45,592
	Dynamic Power	135.07 mW	175.98 mW	439.40 mW	496.64 mW
4096-point FFT	Total logic elements	1,809 (23-bit accum.)	1,757	5,993 (24-bit accum.)	5,863
	Total memory bits	131,552	164,320	131,608	180,760
	Dynamic Power	212.78 mW	242.85 mW	501.11 mW	571.72 mW

Table 7. FPGA implementation results for Radix-2 and Radix-4 FFT

4. Low-power FFT addressing schemes

For embedded applications, power dissipation is often a crucial design goal. (Ma & Wanhammar, 1999) proposed a new addressing logic to improve the memory accessing speed and to reduce the power consumption. (Hasan et al., 2003) designed a new coefficient ordering method to reduce the power consumption of radix-4 short-length FFTs. Gate-level algorithms have also been proposed (Zainal at al., 2009; Saponara, 2003) to reduce the FFT processor's power consumption by lower supply voltage techniques and/or voltage scaling. Power consumption of FFT processors can be significantly reduced by optimizing both data and coefficient memory accesses. Dynamic power consumption in CMOS circuits can be characterized by the following equation:

$$P_{dynamic} = \alpha \cdot C_{total} \cdot V_{DD}^{2} \cdot f \tag{14}$$

where *a* is the switching activity, V_{DD} is the supply voltage, *f* is the frequency and C_{total} is the total switching capacitance charging and discharging in the circuit. In particular, architectural techniques can reduce two parameters in (14), C_{total} and α. These techniques are discussed next: First, a multi-bank memory structure is proposed for data memory accesses, resulting in reduced overall capacitance load on the SRAM bit-lines. Second, a new butterfly calculation order reduces the memory access frequency for twiddle factors and minimizes the switching activity.

4.1 Memory bank partitioning

Since FFT operation largely consists of data and twiddle factor memory accesses, it is desirable to reduce the power dissipation caused by memory accesses. Memory bank partitioning and bitline segmentation is an important technique to reduce the power dissipation in SRAMs. The bitlines (each read and write port is associated with one bitline) in the SRAM logic are a significant source of energy dissipation due to the large capacitive load. This capacitance has two components, wire capacitance of the bitlines and the diffusion capacitance of each pass transistor connecting bitline to bitcells. Hence, the capacitive load increases linearly with the components attached to the bitline i.e., the number of words or size of the memory. In order to reduce this large capacitive load, the data memory can be partitioned into four memory banks instead of two. As a result, the capacitive loading in each memory bank is lowered since the bitline wire length and the number of pass transistors connected to the bitline is now only one fourth of the original bitline. The first two memory banks, *bank0* and *bank1* are accessed by the upper leg of the butterfly structure, and *bank2* and *bank3* are accessed by the lower leg of the butterfly (see Fig. 13). The most significant bit (MSB) of the addresses determine which two memory banks will be accessed; the remaining two memory banks will be inactive. Multi-bank memory structure has been proposed before (Ma & Wanhammar, 2000), but a major advantage of the proposed addressing scheme is that the memory bank switching occurs only once in the middle of a stage. In the first half of the stage, same two memory banks are used and in the second half of the stage, the other two memory banks are accessed. There is no precharging and discharging of bitlines in the inactive memory banks.

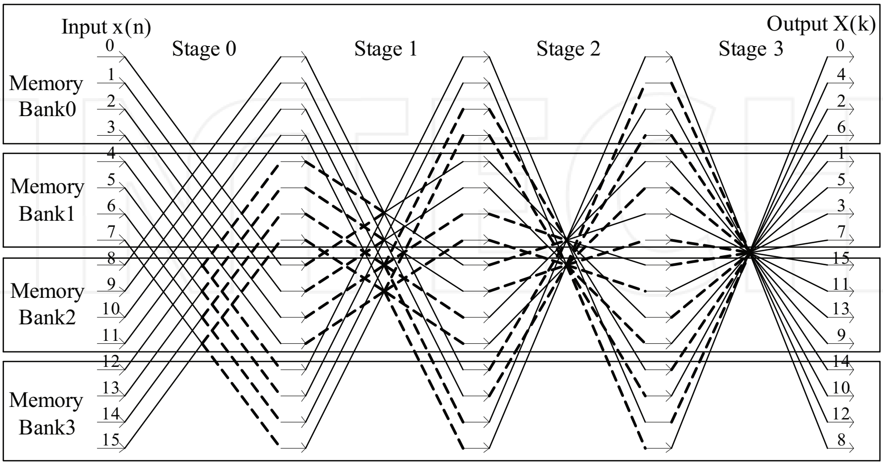


Fig. 13. Signal flow graph of 16-point FFT using memory partitioning

4.2 Reordering coefficient access sequence

The mbFFT architecture (see Section 2.2) can be used to generate the addressing scheme for reducing twiddle factor memory accesses and switching activity power. The twiddle factor access sequence is optimized for minimizing data bus changes. For all butterfly stages, the twiddle factor addresses are ordered in such a way that the twiddle factors at the same address are grouped together and accessed sequentially. This way, the twiddle factor ROM is not accessed every clock cycle. Reordering of the coefficient access sequences is shown in Table 8 and Table 9. For example, in *stage 1* in Table 9, only 8 accesses are needed instead of 16, and in *stage 2*, only 4 accesses instead of 8 and so on.

Counter $B(b_2b_1b_0)$	Stage 0			Stage 1		
	Bank 0,1 address $b_2b_1b_0$	Twiddle factor address b_1b_0	Bank 2,3 address $b_2b_1b_0$	Bank 0,1 address $b_2b_0b_1$	Twiddle Factor address b_10	Bank 2,3 address $\bar{b}_2b_0b_1$
000	000	00	000	000	00	100
001	001	01	001	010	00	110
010	010	10	010	001	10	101
011	011	11	011	011	10	111
100	100	00	100	100	00	000
101	101	01	101	110	00	010
110	110	10	110	101	10	001
111	111	11	111	111	10	011

Stage 2			Stage 3		
Bank0,1 address $b_2b_1b_0$	Twiddle factor address 00	Bank2,3 address $\bar{b}_2\bar{b}_1b_0$	Bank0,1 address $b_2b_1b_0$	Twiddle factor address b_00	Bank2,3 address $\bar{b}_2\bar{b}_1\bar{b}_0$
000	00	110	000	00	111
001	00	111	001	00	110
010	00	100	010	00	101
011	00	101	011	00	100
100	00	010	100	00	011
101	00	011	101	00	010
110	00	000	110	00	001
111	00	001	111	00	000

Table 8. Address generation table for the 16-point, reduced memory access FFT

Counter $B(b_3b_2b_1b_0)$	Stage 0			Stage 1		
	Bank 0,1 address $b_3b_2b_1b_0$	Twiddle factor address $b_2b_1b_0$	Bank 2,3 address $b_3b_2b_1b_0$	Bank 0,1 address $b_3b_0b_2b_1$	Twiddle factor Address $b_2b_1 0$	Bank 2,3 address $\bar{b}_3b_0b_2b_1$
0000	0000	000	0000	0000	000	1000
0001	0001	001	0001	0100	000	1100
0010	0010	010	0010	0001	010	1001
0011	0011	011	0011	0101	010	1101
0100	0100	100	0100	0010	100	1010
0101	0101	101	0101	0110	100	1110
0110	0110	110	0110	0011	110	1011
0111	0111	111	0111	0111	110	1111
1000	1000	000	1000	1000	000	0000
1001	1001	001	1001	1100	000	0100
1010	1010	010	1010	1001	010	0001
1011	1011	011	1011	1101	010	0101
1100	1100	100	1100	1010	100	0010
1101	1101	101	1101	1110	100	0110
1110	1110	110	1110	1011	110	0011
1111	1111	111	1111	1111	110	0111

Stage 2			Stage 3			Stage 4		
Bank0,1 address $b_3b_1b_0b_2$	Twiddle factor address $b_2 00$	Bank2,3 address $\bar{b}_3\bar{b}_1b_0b_2$	Bank0,1 address $b_3b_2b_1b_0$	Twiddle factor address 000	Bank2,3 address $\bar{b}_3\bar{b}_2\bar{b}_1b_0$	Bank0,1 address $b_3b_0b_2b_1$	Twiddle factor Address 000	Bank2,3 address $\bar{b}_3\bar{b}_0\bar{b}_2\bar{b}_1$
0000	000	1100	0000	000	1110	0000	000	1111
0010	000	1110	0001	000	1111	0100	000	1011
0100	000	1000	0010	000	1100	0001	000	1110
0110	000	1010	0011	000	1101	0101	000	1010
0001	100	1101	0100	000	1010	0010	000	1101
0011	100	1111	0101	000	1011	0110	000	1001
0101	100	1001	0110	000	1000	0011	000	1100
0111	100	1011	0111	000	1001	0111	000	1000
1000	000	0100	1000	000	0110	1000	000	0111
1010	000	0110	1001	000	0111	1100	000	0011
1100	000	0000	1010	000	0100	1001	000	0110
1110	000	0010	1011	000	0101	1101	000	0010
1001	100	0101	1100	000	0010	1010	000	0101
1011	100	0111	1101	000	0011	1110	000	0001
1101	100	0001	1110	000	0000	1011	000	0100
1111	100	0011	1111	000	0001	1111	000	0000

Table 9. Address generation table for the 32-point, reduced memory access FFT

Equations (15) and (16) show the twiddle factor memory access frequency for shared memory methods (Xiao et al., 2008) and the proposed reduced memory access method for $N = 2^n$ point FFT.

Conventional method: $$\frac{N}{2} \times (n-2) + 2 = \frac{N}{2}\left(\left(\log_2 N\right) - 2\right) + 2 \quad (15)$$

Reduced memory access method: $$\sum_{i=2}^{n-1} 2^i + 2 = 2^n - 2 = N - 2 \quad (16)$$

Table 10 shows the twiddle factor memory access frequency for different FFT lengths. As FFT length increases, the power saving also scales up.

4.3 Implementation

To implement an $N = 2^n$-point FFT with reduced coefficient memory accesses, an *(n-1)*-bit *Butterfly Counter* $B = b_{n-2}b_{n-3}...b_1b_0$, and a $\lceil \log_2 n \rceil$-bit *Stage Counter* $S = (n-1), \ldots ,2,1,0$ is needed. In addition, one *(n-2)*-bit barrel shifter is used: Assume $RR(x_u x_{u-1} x_{u-2} \ldots x_1 x_0, v)$ indicates rotate-right counter $x_u x_{u-1} x_{u-2} \ldots x_1 x_0$ by v bit. At *stage s*, the read and write addresses of the upper legs of the butterfly is $A_u = RR(b_{n-3} \ldots b_1 b_0, s) = a_{n-3}a_{n-4} \ldots a_1 a_0$, and b_{n-2} decides if *bank0* or *bank1* will be accessed.

	16-point FFT	32-point FFT	64-point FFT	128-point FFT	256-point FFT	512-point FFT	1024-point FFT	2048-point FFT	4096-point FFT	8192-point FFT
Conventional FFT design	18	50	130	322	770	1794	4098	9218	20482	45058
Reduced memory access FFT design	14	30	62	126	254	510	1022	2046	4094	8190
Reduction	22%	40%	52%	61%	67%	72%	75%	78%	80%	82%

Table 10. Reduction in twiddle factor memory access frequency

For example, for the 32-point FFT shown in Table 9, at *stage 2*, the address of the upper legs of the butterfly is $RR(b_2 b_1 b_0, 2) = b_1 b_0 b_2$, and when b_3=0, memory *bank0* will be accessed, when b_3=*1*, memory *bank1* will be accessed. For the read and write addresses of the lower legs of the butterfly, *(n-2)* inverters are needed. The address is given by $\bar{a}_{n-3}\bar{a}_{n-4} \ldots \bar{a}_{n-s-1}a_{n-s-2} \ldots a_1 a_0$, and b_{r-2} decides if *bank2* or *bank3* will be accessed at *stage 0*. At stage 0, when $b_{n-2} = 0$, *bank2* will be accessed. When $b_{n-2} = 1$, *bank3* will be accessed. For other stages $b_{n-2} = 0$ means *bank3* will be accessed, $b_{n-2} = 1$ means *bank4* will be accessed. The address of twiddle factors is given by $a_{n-s-3} \ldots a_0 0 \ldots 0$ (S '0's). Fig 14 shows the components of the address generation logic using mbFFT and four memory banks.

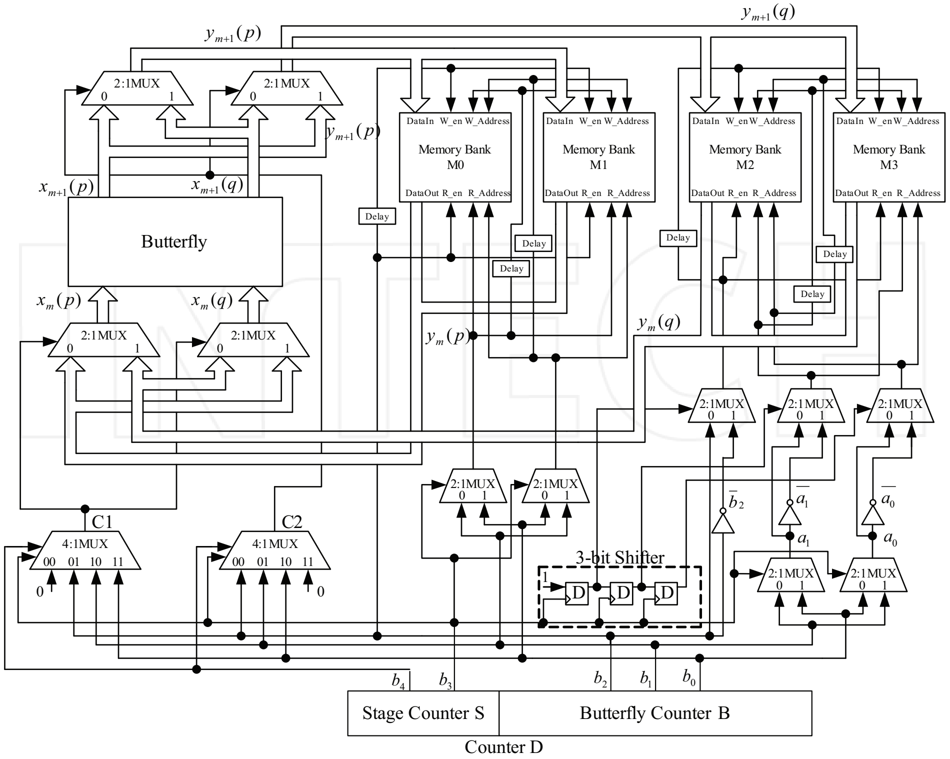


Fig. 14. Address generation circuits for low-power 16-point FFT using mbFFT and four memory banks

	Shared memory design (Xiao et al., 2008)			Power optimized design		
	Total power	Dynamic power	Static power	Total power	Dynamic power	Static power
512 point FFT	653.14mw	203.13mw	450.00mw	635.47mw	185.47mw	450.00mw
1024 point FFT	715.79mw	265.79mw	450.00mw	676.79mw	226.78mw	450.00mw
2048point FFT	840.49mw	390.49mw	450.00mw	764.31mw	314.31mw	450.00mw
4096 point FFT	1089.33mw	639.33mw	450.00mw	939.25mw	489.24mw	450.00mw
8192 point FFT	1595.13mw	1145.13mw	450.00mw	1289.17mw	839.17mw	450.00mw

Table 11. FPGA synthesis results – Reduction in dynamic power

4.4 FPGA synthesis results

The low-power FFT algorithm is implemented on an FPGA chip (ALTERA STRATIX EP1S25F780C5) with FFT length up to 8192 points as shown in Table 11. The synthesis results demonstrate that dynamic power reduction grows with the transform size, making this architecture ideal for applications requiring long FFT operations.

5. Conclusion

This study focused on hardware efficient and low-power realization of FFT algorithms. Recent novel techniques have been discussed and presented to realize conflict-free memory addressing of FFT. Proposed methods reorder the data and coefficient address sequences in order to achieve significant logic reduction (24% less transistors) and delay improvements within FFT processors. Multiplierless implementation of FFT is shown using a CORDIC algorithm that does not need any coefficient angle memory, resulting in 33% memory and 15% power reduction. Finally, optimization of FFT dynamic power consumption is presented through memory partitioning and reducing coefficient memory access frequency (26% power reduction for 8192 point-FFT).

6. References

Abdullah, S. S.; Nam, H.; McDermot, M. & Abraham, J. A. (2009). A High Throughput FFT Processor with No Multipliers. *IEEE International Conference on Computer Design*, pp. 485-490, 2009.

Bouguezel, S.; Ahmad, M. O. & Swamy, M. N. S. (2004). A New Radix-2/8 FFT Algorithm for Length-*Q* X 2^m DFTs. *IEEE Transactions on Circuits and Systems I,* vol. 51, no. 9, pp. 1723-1732, September 2004.

Cohen, D. (1976). Simplified Control of FFT Hardware. *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 24, pp. 577–579, December 1976.

Despain, A. M. (1974). Fourier Transform Computers Using CORDIC Iterations. *IEEE Transactions on Computers*, vol. c-23, no.10, pp. 993-1001, October 1974.

Fanucci, L.; Forliti, M. & Gronchi, F. (1999). Single-Chip Mixed-Radix FFT Processor for Real-Time On-Board SAR Processing. *6th IEEE International Conference on Electronics, Circuits and Systems, ICECS '99*, vol. 2, pp. 1135-1138, September 1999.

Garrido, M. & Grajal, J. (2007). Efficient Memory-Less CORDIC for FFT Computation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, no. 2, pp. 113-116, April 2007.

Hasan, M.; Arslan, T. & Thompson, J. S. (2003). A Novel Coefficient Ordering Based Low Power Pipelined Radix-4 FFT Processor for Wireless LAN Applications, *IEEE Transactions on Consumer Electronics*, vol. 49, no.1, pp. 128-134, February 2003.

He, S. S. & Torkelson, M. (1996). A New Approach to Pipeline FFT Processor. *Proceedings of 10th International Parallel Processing Symposium*, pp. 766-770, April 1996.

Hopkinson, T. M. & Butler, G. M. (1992). A Pipelined, High-Precision FFT Architecture. *Proceedings of the 35th Midwest Symposium Circuits and Systems*, vol. 2, pp. 835-838, August 1992.

Jiang, R. M. (2007). An Area-Efficient FFT Architecture for OFDM Digital Video Broadcasting. *IEEE Transactions on Consumer Electronics*, vol. 53, no. 4, pp. 1322-1326, 2007.

Li, W. D. & Wanhammar, L. (1999). A Pipeline FFT Processor. *Proceedings of IEEE Workshop on Signal Processing Syst*ems, pp. 654-662, October 1999.

Li, X.; Lai, Z. & Cui, J. (2007). A Low Power and Small Area FFT Processor for OFDM Demodulator. *IEEE Transactions on Consumer Electronics,* vol. 53, no. 2, pp. 274-277, May 2007.

Lin, C. & Wu, A. (2005). Mixed-Scaling-Rotation CORDIC (MSR-CORDIC) Algorithm and Architecture for High-Performance Vector Rotational DSP Applications. *IEEE Transactions on Circuits and Systems I*, vol. 52, no. 11, pp. 2385-2396, 2005.

Ma, Y. (1994). A Fast Address Generation Scheme for FFT Processors, *Chinese Journal Computers*, vol. 17, no. 7, pp. 505-512, July 1994.

Ma, Y. (1999). An Effective Memory Addressing Scheme for FFT Processors. *IEEE Transactions on Signal Processing*, vol. 47, no. 3, pp. 907–911, March 1999.

Ma, Y. & Wanhammar, L. (1999). A Coefficient Access Control for Low Power FFT Processors. *IEEE 42nd Midwest Symposium on Circuits and Systems*, vol.1, pp. 512-514, Aug. 1999.

Ma, Y. & Wanhammar, L. (2000). A Hardware Efficient Control of Memory Addressing for High-Performance FFT Processors. *IEEE Transactions on Signal Processing,* vol. 48, no. 3, pp. 917-921, March 2000.

Proakis, J. G.; & Manolakis, D. G. (2006). *Digital Signal Processing Principles, Algorithms, and Applications,* Prentice Hall, ISBN 978-0131873742.

Saponara, S.; Serafini, L. & Fanucci, L. (2003). Low-Power FFT/IFFT VLSI Macro Cell for Scalable Broadband VDSL Modem. *The 3rd IEEE International Workshop on System-on-Chip for Real-Time Applications,* pp.161-166, June 2003.

Volder, J. (1959). The CORDIC Trigonometric Computing Technique. *IEEE Transactions on Electronic Computers,* vol. EC-8, no. 8, pp. 330-334, September 1959.

Wang, Y.; Tang, Y.; Jiang, Y.; Chung, J.; Song, S. & Lim, M. (2007). Novel Memory Reference Reduction Methods for FFT Implementations on DSP Processors. *IEEE Transactions on Signal Processing,* vol. 55, no. 5, pp. 2338-2349, May 2007.

Wey, C.; Lin, S. & Tang, W. (2007). Efficient Memory-Based FFT Processors For OFDM Applications. *IEEE International Conference on Electro- Information Technology,* pp.345 - 350, May 2007.

Xiao, X.; Oruklu, E. & Saniie, J. (2008). An Efficient FFT Engine with Reduced Addressing Logic. *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 55, no. 11, pp.1149-1153, November 2008.

Xiao, X.; Oruklu, E. & Saniie, J. (2009). Fast Memory Addressing Scheme for Radix-4 FFT Implementation. *IEEE International Conference on Electro/Information Technology, EIT 2009,* pp. 437-440, June 2009.

Xiao, X.; Oruklu, E. & Saniie, J. (2010). Reduced Memory Architecture for CORDIC-based FFT. *IEEE International Symposium on Circuits and Systems (ISCAS),* June 2010.

Yang, L.; Zhang, K.; Liu, H.; Huang, J. & Huang, S. (2006). An Efficient Locally Pipelined FFT Processor. *IEEE Transactions on Circuits and Systems II, Exp. Briefs,* vol. 53, issue 7, pp. 585-589, July 2006.

Zainal, M. S.; Yoshizawa, S. & Miyanaga, Y. (2009). Low Power FFT Design for Wireless Communication Systems. *International Symposium on Intelligent Signal Processing and Communications Systems ISPACS 2008*, pp. 1-4, February 2009.

20

The Effect of Local Field Dispersion on the Spectral Characteristics of Nanosized Particles and their Composites

T.S. Perova[1], I.I. Shaganov[2] and K. Berwick[2]
[1]*Trinity College Dublin,*
[2]*Vavilov State Optical Institute, St.-Petersburg,*
[3]*Dublin Institute of Technology, Dublin*
[1,3]*Ireland*
[2]*Russia*

1. Introduction

Infrared (IR) spectroscopy of micro- and nanosized particles and their composites is currently one of the most important enabling technologies in the development of micro- and nanostructures and their application to various areas of science and technology. Decreasing the characteristic size of metallic, dielectric and semiconductor materials results in a dramatic alteration to their optical, electrical and mechanical properties, allowing the fabrication of new materials with unique physical properties (Lamberti, 2008; Cao, 2004). These alterations in the optical properties are related to a quantum confinement effect, as well as to a dielectric, or electrostatic, confinement effect (Cahay et al., 2001; Chemla & Miller, 1986). The effect of quantum confinement is most pronounced in semiconductor materials, where the transition from the bulk state to the microcrystalline state causes a substantial change in the band structure and an enhancement of the non-linear electro-optical properties. Dielectric, or polarisation, confinement has a wider impact, since it influences the frequencies and intensities of absorption bands in the spectra of any condensed matter, including crystalline and amorphous solids, as well as liquids. This is because considerable changes in the polarisation of micro/nanoparticles occur, depending on their form and orientation with respect to the external electromagnetic field and the details of the spatial restriction.

So, the dielectric confinement effect is due to abrupt changes in the intensity of the internal (E_{in}(v)), local electric field E_{loc}(v), causing significant changes in the spectroscopic characteristics, depending on the direction of the external field E(v), and the size and shape of the submicron sized particles, or micro-objects. Dielectric confinement occurs when the absorbing material consists of micro-particles with characteristic sizes significantly smaller than the wavelength of the probe beam. These particles are generally embedded in a transparent dielectric matrix, or deposited on a transparent substrate as an ultra-thin film (Fig. 1). A good analogy to these systems is that of an aerosol suspended in air or stained glass, that is, glass doped with small metal particles (Gehr & Boyd, 1996). In the long wavelength limit, $d << \lambda$, for the determination of the spectroscopic characteristics of micro-particles with size *d* in the direction of dielectric confinement, one can use an effective medium theory model, while taking into account the dispersive local field (DLF) (Chemla & Miller, 1986; Schmitt-Rink, 1987; Cohen, 1973; Spanier & Herman, 2000). The important role of the local-field effect in the derivation of the equations of the effective medium theory of composites was considered in the paper by Aspnes, 1982.

The local-field approach is widely used for the analysis of the spectral characteristics of condensed matter under dielectric confinement. In particular, in Ref. (Liu, 1994), a description of the distribution of the *p*-component of the local electric field within the quantum wells in multi-quantum well GaAs-Al$_x$Ga$_{1-x}$As structures and the absorption band for intersubband transitions has been obtained, using a self-consistent integral equation for the local field. The development of the approach suggested for the analysis of the spectral features observed from materials based on porous structures is of particular importance (Spanier & Herman, 2000; Timoshenko et al., 2003; Golovan et al., 2007). These investigations largely involve extending the models used in effective medium theory.

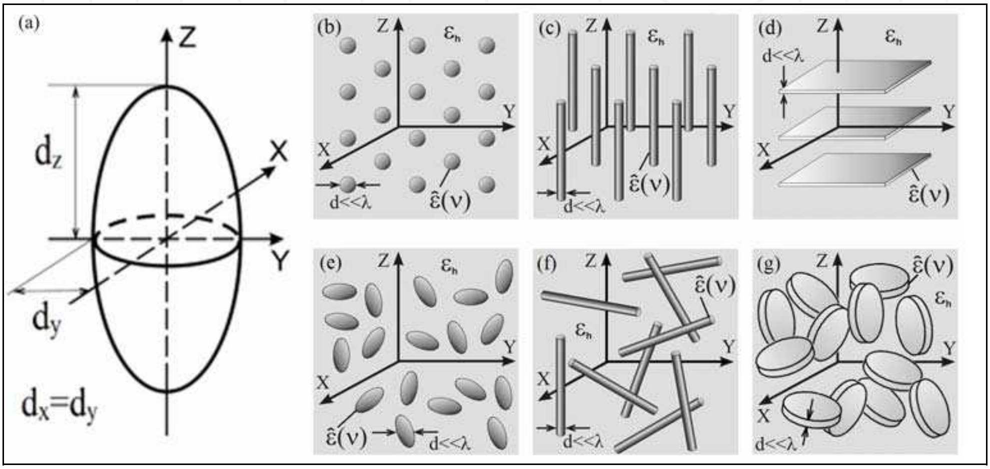


Fig. 1. a) The modeled spheroidal shape of the absorbing mesoparticles. Schematic depicting different types of size confinement for ordered (in (b) 3D confinement, c) 2D confinement ($\vec{E}\perp z$) and d) 1D confinement ($\vec{E}\parallel z$)) and disordered (in (e) 3D confinement, *f*) 2D confinement and g) 1D confinement) mesoparticles

The effective medium theory (EMT) approach is widely used for modelling the optical and spectroscopic properties of a variety of composite media. The most extensively used EMT models are the Maxwell-Garnett (MG) and Bruggeman models, however other models are also used in some specific cases (Aspnes, 1982; Cohen et al., 1973; Spanier & Herman, 2000; Maxwell-Garnett, 1906; Bruggeman, 1935). For example in Ref. (Spanier & Herman, 2000), hybrid models, containing both phenomenological features and statistical theories of the dielectric function of dielectric media were used for modelling the infrared spectra from porous silicon carbide films. In Ref. (Mallet et al., 2005), an analysis of the accuracy of the modified Maxwell-Garnett equation, taking into account multiple scattering of light by the composite medium with spherical inclusions, has also been performed. In Ref. (Gehr & Boyd, 1996) the authors reviewed the theories and models developed for relating the linear and non-linear optical properties of composite materials to those of the constituent materials, and to the morphology of the composite structure. The authors of Ref. (Hornyak et al., 1997) experimentally determined the size of gold nanoparticles, satisfying the quasi-static limit of applicability of the Maxwell-Garnett equation. As shown in this paper, inaccuracies in the MG expression, related to the scattering of light on large particles which do not satisfy the limit discussed earlier, can be eliminated by a dynamic modification of this expression (Foss et al., 1994). In the work (Ung et al., 2001), it was shown that the MG expression adequately describes the influence of inter-particle interactions on the position of the plasmon resonance band in colloidal solutions of gold. The influence of the local field on the enhancement of the light emission from various composite materials is described by (Dolgaleva et al., 2009).

In this Chapter, an overview of our recent work developing the effective medium approach and dispersive local field theory is presented. We also discuss the application of these models to nanocomposite materials, based on liquids and amorphous solids, for simulation of the experimentally obtained infrared spectra. We focus on a consideration of dielectric confinement only within the linear optical response. The influence of the dielectric confinement effect on the infrared absorption spectra of composite media will be demonstrated experimentally. We also present a theoretical analysis of this effect on the value of the frequency shift. The influence of the integrated intensity of the IR bands under consideration and the dielectric constant of the surrounding matrix will also be explored. The results obtained will assist in improving the reliability of IR spectral analysis.

2. Theoretical considerations

In the long-wave limit, when the absorbing material consists of micro-particles, with characteristic sizes significantly smaller than the wavelength of the probe beam, i.e. satisfying the condition $d \ll \lambda$, the spectroscopy of intermolecular interactions (IMI) can be used for analysis of their spectroscopic characteristics. The influence of the dielectric effect on the absorption spectra of molecular condensed systems was described for the first time in the work of Backshiev, Girin and Libov (BGL) (Backshiev et al., 1962; 1963), based on accounting for the spectral difference in the intensity of the effective, internal, field in the vicinity of the optical resonance and the average macroscopic field in condensed matter. A similar approach for calculating the spectral dependence of the microscopic susceptibility in the wavelength range of the intermolecular vibrations of organic liquids was used by Clifford & Crawford, 1966. In accordance with the BGL approach, the relationship between the micro- and macro-characteristics of condensed matter can be presented as

$$B(\nu) = \frac{2\pi \operatorname{Im}\hat{\varepsilon}(\nu)\theta(\nu)}{Nh} \tag{1a}$$

or

$$\frac{B(\nu)Nh}{2\pi} = \operatorname{Im}\hat{\varepsilon}(\nu)\theta(\nu) \tag{1b}$$

Here $B(\nu)$ is the spectral density of the quantum intramolecular transition probability (Heitler, 1975), $\operatorname{Im}\hat{\varepsilon}(\nu)$ is the imaginary part of the dielectric function in the vicinity of this transition and $\theta(v) = \left|E_o(\nu) / E_{in}(\nu)\right|^2$ is the correction factor accounting for the spectral difference between the internal, local, micro $E_{in}(\nu)$ and the average macro $E_0(\nu)$ fields of the electromagnetic wave in condensed matter. We note that the average electromagnetic field, $E_o(v)$, is considered here as a small perturbation and, therefore, the approach presented is still valid within the framework of linear molecular optics.

The $B(\nu)$ spectrum in Eqn. (1) is considered to be characteristic of intramolecular quantum transitions with absorption. In case of the lattice vibrations, this spectrum is related to the dipole moment of the quantum transition, localized in a physically small volume of the crystal. The size of this elemental volume is significantly smaller than the wavelength of the probe beam, but is substantially larger than the size of the elemental crystal cell (Tolstykh et al., 1973). This conclusion can also be generalised to non-crystalline media. Indeed, it can be easily shown that expression (1b) corresponds to the spectrum of the imaginary part of the complex microscopic susceptibility of the medium $\operatorname{Im}\chi^{micro}(\nu)=\chi_2(\nu)$, which, in accordance with the Lorentz local field model, is related to the macroscopic susceptibility of the isotropic medium by this expression

$$\hat{\chi}^{micro}=\frac{3\hat{\chi}}{\hat{\chi}+3} \tag{2}$$

Where $\hat{\chi}(\nu)=\chi_1(\nu)-i\chi_2(\nu)$ is the macroscopic dielectric susceptibility of the medium under consideration. Solving Eqn. (2) with respect to $\operatorname{Im}\chi^{micro}$ we obtain $\chi_2^{micro}=\chi_2\theta(\nu)$. Since $\chi_2(\nu)=\operatorname{Im}\hat{\varepsilon}(v)$ we can consider $\chi_2^{micro}(\nu)$ as the spectrum of $\varepsilon_2^{micro}(\nu)$. This allows us to conclude that Eqn. (1a) corresponds to the spectral characteristics of a spherical micro-volume, or microparticle, of the condensed medium under consideration, with the particle size satisfying the condition $\lambda >> d >> a_{molec}$ and represented by the following expression

$$\varepsilon_2^{micro}(\nu)=\varepsilon_2(\nu)\theta(\nu) \tag{3}$$

Using a continuum model of the local field allows us to use this expression with $\theta(\nu)=9/\left|\varepsilon(\nu)+2\right|^2$ in order to establish the relationship between the dielectric loss spectrum of the bulk sample and that of the material under three dimensional (3D) size confinement, that is, for an isolated spherical particle. In general, the relationship between the local and average field in a condensed medium under 1D, 2D and 3D confinement can be written as (Ghiner & Surdotovich, 1994)

$$E_{in}(\nu)/E_0(\nu)=1+(\varepsilon(\nu)-1)/m\,, \tag{4}$$

where $m=$ 1, 2 and 3 respectively for the case of 1D, 2D and 3D confinement. The general equation describing the spectra of micro-objects, $\varepsilon_2^{micro}(\nu)$, satisfying the conditions above can be expressed as

$$\varepsilon_2^{micro}(\nu)=\varepsilon_2(\nu)\theta(\nu)=\varepsilon_2(\nu)\left|1+(\varepsilon(\nu)-1)/m\right|^{-2} \tag{4a}$$

The use of molecular spectroscopy approaches when considering the spectral features characteristic of microparticles is justified. As shown by (Ghiner & Surdotovich, 1994), micro-particles, satisfying the conditions for dielectric confinement, can be considered as meso-oscillator molecules or meso-molecules, possessing their own spectroscopic characteristic, i.e. $\varepsilon_2^{micro}(\nu)$ (or $\varepsilon_2^{meso}(\nu)$), spectrum. A similar conclusion follows from the work of (Chemla & Miller, 1986) where an expression similar to Eqn. (3) here was used to describe the spectral properties of semiconductor particles. It is worth noting that the basic mechanism responsible for the blue shift of the absorption spectra of nanoparticles with respect to their bulk counterpart is the decrease in the intermolecular interaction potential due to the reduction in, or elimination of, resonant dipole-dipole interactions of the molecules both inside and outside the particles. This decrease occurs as a result of the decrease in particle size from d ≤ λ to d << λ, as well as the increase in the distance between the particles. The decrease in the resonant dipole-dipole interactions and, consequently, the intermolecular interaction potential can be taken into account by considering the dispersion of the effective field, from which the expressions (1a) and (4) are derived. We note that expression (4) describes only limited cases of dielectric confinement. In accordance with the expression for the local field inside a spherically shaped particle (Böttcher, 1952), the correction factor in Eqn. (4) can be written as

$$\varepsilon_2^{micro}(\nu) = \varepsilon_2(\nu)\theta(\nu) = \varepsilon_2(\nu)\left|1 + L(\varepsilon(\nu) - 1)\right|^{-2} \tag{5}$$

L is the form factor, the ratio of the semi-axes for an ellipsoidal particle shown in Fig. 1e. For an ellipsoid of revolution, the corresponding components of the form factor for two orientations of the electric field vector *E*, parallel, L_z, or perpendicular, $L_{x,y}$, to the rotation axis of the spheroid, are determined by the following expressions (Osborn, 1945; Golovan et al., 2003):

$$L_z = \frac{1}{1-P^2}\left[1 - P\frac{\arcsin\left(\sqrt{1-P^2}\right)}{\sqrt{1-P^2}}\right]; \quad L_{x,y} = \frac{1-L_z}{2} \tag{6}$$

where $P = d_z/d_x = d_z/d_y$ and d_z and $d_x = d_y$ are the sizes of the corresponding polar and equatorial semi-axes of the spheroid (Fig. 1a). We note that for a spherical particle, the form-factor is *L*=1/3 (Fig. 1b), while for a rod, along the short axis, *L* = 1/2, and along the long axis, *L*=0 (Figs. 1c and 1f). For a strongly oblate spheroid stretched in the perpendicular direction, *L*=1 (Figs. 1d and 1g).

Equation (5) shows that a particle with a dielectric function, ε, corresponding to the bulk material, can be considered as a particle with an effective microscopic spectrum $\varepsilon_2^{micro}(\nu)$. If this particle is embedded in a dielectric host matrix with $\varepsilon_h > 1$, then expression (5) can be written as

$$\varepsilon_2^{micro}(\nu) = \varepsilon_2(\nu)\theta(\nu) = \varepsilon_2(\nu)\varepsilon_h^2\left|1 + L(\varepsilon(\nu) - \varepsilon_h)\right|^{-2} \tag{7}$$

The dielectric loss spectrum for a diluted composite medium would be determined by spectrum $\varepsilon_2^{micro}(\nu)$ and the volume concentration of particles *f* in the composite

$$\varepsilon_2^{comp}(\nu) = f\varepsilon_2(\nu)\theta(\nu) = f\varepsilon_2(\nu)\varepsilon_h^2\left|1 + L(\varepsilon(\nu) - \varepsilon_h)\right|^{-2} \tag{8}$$

Obviously, we ignore the resonant dipole-dipole interactions of the particles, which are practically insignificant when the filling factor, *f*, is smaller than 1%. This does not generate significant errors in calculations until *f* is over 10%. Eqn. (8) was obtained earlier in paper (Shaganov et al., 2005). A more accurate equation can be obtained by modifying the Maxwell-Garnett expression using an effective media approach (Aspnes, 1982; Cohen et al., 1973; Spanier & Herman, 2000). For a composite medium containing absorbing particles of spheroidal shape, the corresponding expression can be written as

$$\frac{\hat{\varepsilon}_i(\nu)-\varepsilon_h}{L_i\hat{\varepsilon}_i(\nu)+(1-L_i)\varepsilon_h}=\frac{f\cdot(\hat{\varepsilon}(\nu)-\varepsilon_h)}{L_i\hat{\varepsilon}(\nu)+(1-L_i)\varepsilon_h} \tag{9}$$

where L_i is the corresponding component of the form factor, $\hat{\varepsilon}_i(\nu)$ is the component of the tensor of the effective complex dielectric permittivity of the media and $\hat{\varepsilon}(\nu)$ is the complex dielectric permittivity of the bulk material of the embedded particles. For $L_i = 1/3$ expression (9) can be converted to the typical form of the Maxwell-Garnett equation.

$$\frac{\hat{\varepsilon}_i(\nu)-\varepsilon_h}{\hat{\varepsilon}_i(\nu)+2\varepsilon_h}=\frac{f\cdot(\hat{\varepsilon}(\nu)-\varepsilon_h)}{\hat{\varepsilon}(\nu)+2\varepsilon_h} \tag{10}$$

These expressions have been widely used in the past for modelling the spectral properties of metal-dielectric composites (Cohen et al., 1973; Foss et al. 1994; Hornyak et al., 1997; Ung et al., 2001). We note that the limits of applicability of this approximation are defined by the applicability of the electrostatic model of the effective medium, because this approximation does not take into account the size of the particles under consideration. A more precise approach is required to consider so-called dynamic polarisation, which takes into consideration the size of the particle, and its interaction time, with the field of the electromagnetic wave (Golovan et al., 2003; 2007). It is reasonable to assume that dynamic polarisation is significant only in the visible range, playing a minor role in the mid-infrared range, to a first approximation. Solving expression (9) for the desired value, we obtain the following expression for the dielectric permittivity spectrum of the composite media

$$\hat{\varepsilon}_i=\frac{f(\hat{\varepsilon}(\nu)-\varepsilon_h)\,\varepsilon_h\,(1-L_i)+\varepsilon_h\left[L_i\hat{\varepsilon}(\nu)+(1-L_i)\varepsilon_h\right]}{L_i\hat{\varepsilon}(\nu)+(1-L_i)\varepsilon_h-f\cdot(\hat{\varepsilon}(\nu)-\varepsilon_h)L_i} \tag{11}$$

From expression (10), the effective dielectric loss spectrum of the ordered composite medium, $\operatorname{Im}(\hat{\varepsilon}_i(\nu))$, in general, can be presented in the following form.

$$\operatorname{Im}(\hat{\varepsilon}_i(\nu))=\frac{A_2B_1-A_1B_2}{B_1^2+B_2^2} \tag{11a}$$

where

$$A_{1i}(\nu)=\{(1-f)(1-L_i)\varepsilon_h+\left[f(1-L_i)+L_i\right]\varepsilon_1(\nu)\}\varepsilon_h$$
$$A_{2i}(\nu)=[f(1-L_i)+L_i]\varepsilon_h\varepsilon_2(\nu)$$
$$B_{1i}(\nu)=(1-L_i)\varepsilon_h+f\varepsilon_hL_i+L_i(1-f)\varepsilon_1(\nu)$$
$$B_{2i}(\nu)=L_i(1-f)\varepsilon_2(\nu)$$

Here $\varepsilon_1(\nu)$ and $\varepsilon_2(\nu)$ are the real and imaginary parts of the dielectric permittivity spectrum of the particle material in the bulk state $\hat{\varepsilon}(\nu)=\varepsilon_1(\nu)-i\varepsilon_2(\nu)$. For a random particle orientation, the effective dielectric loss spectrum of the isotropic composite medium can be presented as

$$\varepsilon_2^{eff}(\nu) = \frac{1}{3}\sum_{L_i}\varepsilon_{2i}(\nu) \tag{12}$$

where the addition of the index L_i takes into account the difference in form factor of the particles in the x, y, z directions. For the specific case of 1D, 2D and 3D confinement, and at $f << 1$, Eqn.(11) can be transformed to the more simple form given in Ref. (Shaganov et al., 2010)

$$\mathrm{Im}(\hat{\varepsilon}_i(\nu)) = f\varepsilon_2(\nu)\theta_{iD}(\nu) \tag{13}$$

where θ_{iD} (ν) is the correction factor for the internal, local, field, acting on the particles under 1D, 2D and 3D dielectric confinement, in agreement with Expression (4) obtained previously (Shaganov et al., 2005).

$$\theta_{iD}(\nu) = \left(\left|1 + \frac{\hat{\varepsilon}(\nu) - \varepsilon_h}{m_i\varepsilon_h}\right|\right)^{-2} \tag{14}$$

where $i = m_i = 1, 2, 3$ for 1D, 2D and 3D confinement, respectively. For randomly oriented particles, expressions (13) and (14) can be transformed to the following form (Shaganov et al., 2010)

$$\varepsilon_2^{eff}(\nu) = \frac{1}{3}f\varepsilon_2(\nu)\left[3 - m_i + m_i^3\varepsilon_h^2 \cdot \left|\hat{\varepsilon}(\nu) + (m_i - 1)\varepsilon_h\right|^{-2}\right] \tag{15}$$

It is worth noting that the expressions above are valid only for diluted composites, where the resonant dipole-dipole interaction between the particles can be neglected. Depending on the intensity of the absorption band, or oscillator strength, resonant interactions between the particles become significant when the volume fraction of the particles is in the range $f = 0.1 - 0.2$. In this case, the local field factor, $\theta_{iD}(\nu)$, becomes dependent on the particle concentration (Shaganov et al., 2005) and expression (15) is transformed to the following (Shaganov et al., 2010)

$$\varepsilon_2^{eff}(\nu) = \frac{1}{3}f\varepsilon_2(\nu)\left[3 - m_i + m_i^3\varepsilon_h^2 \cdot \left|(\hat{\varepsilon}(\nu) - \varepsilon_h)(1 - f) + m_i\varepsilon_h\right|^{-2}\right] \tag{15a}$$

We note that Eqn. (15a) can be used not only in the limited cases of 1D, 2D and 3D confinement, but also for composites of spheroidal particles where the ratio of the semi-axes $P \geq 10$ or $P \leq 0.1$. For intermediate values of P: $0.2 < P < 9$, account must be taken of the specific values of the form factors for the three axes of the spheroidal particles, assuming that $m_i = 1/L_i$ (see Shaganov et al., 2010 for details).
As shown in Ref. (Shaganov et al., 2005), the difference between the spectral characteristics of the bulk materials and those from a composite of micro-particles can be substantial. The shift in the peak position of the intense absorption bands due to dielectric confinement can be far greater than the linewidth of the absorption band observed in the bulk material. The peak position, or maximum frequency, for isolated particles in the case of 3D confinement (ν_{3D}) is close to Fröhlich's frequency (ν_F) (Fröhlich, 1949), corresponding to the condition $\varepsilon_1(\nu_F) = -2\varepsilon_h$. The maximum shift in the peak position of the absorption spectrum occurs for 1D confinement, where the peak position is observed at a frequency ν_l, satisfying the minimum of the function $\varepsilon_1(\nu_l) = 0$. Thus, it is not surprising that the values for ν_l, obtained

from calculations for polar crystals, coincide with the frequency of the corresponding band of longitudinal optical (LO) phonon vibrations. The absorption spectra of the amorphous media at frequencies ν_l has already been discussed in numerous papers (Berreman, 1963; Röseler, 2005; Tolstoy et al., 2003; Iglesias et al., 1990; DeLeeuw & Thorpe, 1985). Conclusions on the size dependent nature of this effect have been made earlier in the theoretical work of (Lehmann, 1988). We would like to emphasise that we obviously cannot discuss LO-phonons in amorphous solids and, in particular, in liquids, since the new bands observed arise as a result of the interaction of the transverse electromagnetic wave with a condensed medium under dielectric confinement, when the contribution from surface vibrations becomes greater than that from the bulk. The maximum frequency of the spectrum from a composite medium, for 2D confinement, lies between the frequencies for 1D and 3D dielectric confinement, i.e. $\nu_{3D} < \nu_{2D} < \nu_{1D}$. In practice, the microparticles will not all be spheroidal, particularly in microcrystalline powders, for which the shape of particles often depends on the crystalline structure of the material. For a more detailed discussion see Shaganov et al., 2010.

3. Results and discussion

The objective of this section is to demonstrate, both theoretically and experimentally, the role of various types of dielectric confinement on the absorption spectra of organic liquids and amorphous solids. Amorphous SiO_2 and three organic liquids of spectroscopic grade viz. benzene (C_6H_6), chloroform ($CHCl_3$), and carbon disulphide (CS_2), have been chosen for the experiments, because of their well characterized, strong absorption in the infrared range (Zolotarev et al., 1984; Barnes & Schatz, 1963).

3.1 Calculations

The method in which dielectric mesoparticles are embedded in the host medium is important in the engineering of the optical properties of a composite. For example, depending on the alignment and distribution of the mesoparticles in the host medium, the composite medium can possess optical anisotropy, which is apparent in phenomena such as birefringence, anisotropy in the real part of the refractive index, and dichroism, anisotropy in the imaginary part of the refractive index (Golovan et al., 2007). In this study, we discuss the influence of dielectric confinement on the resonant part of the dielectric permittivity, leading to phenomena such as a spectral shift in the resonant absorption band and its anisotropy. We consider two extreme cases only, viz. completely ordered and completely disordered (randomly oriented) dielectric mesoparticles, uniformly distributed in a host medium (Fig. 1). It is worth noting that deliberately varying the degree of mesoparticle disorder in a composite medium can be used in order to tune its optical properties. Eqns. (8), (11a) and (12), (15) describe dielectric loss spectra for completely ordered and disordered composites, respectively. Note that in the case of a disordered composite, as described by Eqn. (12), (15) and (15a), the solution consists of two bands for all mesoparticles, with the exception of those with a spherical shape. The splitting apparent in the dielectric loss spectrum and, therefore, in the absorption spectrum of the composite, is most pronounced for 1D confinement.

In the calculations presented in Figs. 2-4, and summarized in Tables 1 and 2, we used Eqns. (11a) and (12) for 1D, 2D and 3D confinement. These situations can also be described using the simplified Eqns. (8) and (15a). In Table 1, experimental data described in Section 3.2 are also shown for comparison. Calculations have been performed for liquid benzene, chloroform and carbon disulphide at f = 0.1 (for ε_h = 11.56 (Si) and ε_h = 13.6) and additionally for carbon disulphide at f = 0.1 and ε_h = 3, and for benzene at f = 0.1 and ε_h = 2.2, ε_h = 5, and ε_h = 16 (Ge). The optical constants in the infrared range for benzene, chloroform and carbon disulphide were taken from references (Zolotarev et al., 1984; Barnes & Schatz, 1963). The value of ε_h = 13.6 was an average, calculated by taking the square root of the product of ε_{Si} =11.56 and ε_{Ge} = 16. This allowed us to model liquids films between a Ge ATR prism and a Si slide in a GATR attachment as described in the experimental Section. Additional calculations for carbon disulphide at ε_h = 3 and for benzene at ε_h = 2.2 were performed to illustrate absorption band splitting in a disordered composite, apparent on the bottom panels of Fig.2 (c and d).

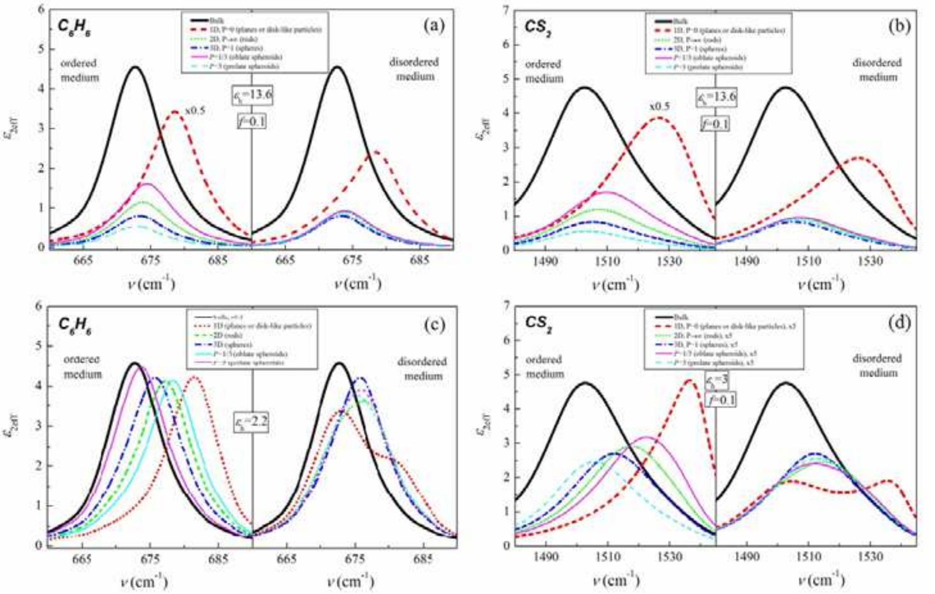


Fig. 2. Dielectric loss spectra calculated for ordered and disordered mesoparticles with filling factor f=0.1 under different confinement conditions for (a) benzene and (b) carbon disulfide in host media with ε_h=13.6 (Si/Ge) and for (c) benzene at ε_h=2.2 and (d) carbon disulfide at ε_h=3

As can be seen from Fig. 2, by changing the particle shape for ordered mesoparticles, we can gradually change the peak position of the absorption spectrum of the composite media in the range of 15 cm^{-1} for benzene and 30 cm^{-1} for CS_2. The peak position of the dielectric loss spectrum for oblate spheroids (P = 1/3) is close to the peak position corresponding to 1D confinement in planes or disks (Figs. 1d and 1g), while the peak position for prolate spheroids (P = 3) is close to that observed from bulk benzene and carbon disulphide, since the amount of dielectric confinement is reduced in the direction of the field, that is, along the rotation axis of the particles.

The situation for disordered mesoparticles is quite different. In both cases, namely P = 1/3 and P = 3, the dielectric loss spectra are close to the spectrum from spherical particles. It is interesting that, in this case, the dielectric loss spectrum from oblate spheroids is closer to the spectrum of the bulk, while the spectrum for prolate spheroids is closer to the spectrum characteristic of 1D confinement. The similarity of the dielectric loss spectra for P = 1/3 and P = 3 to the spectrum from spherical particles under 3D confinement can be explained as being due to averaging of the disordered spheroids in every direction, resulting in an isotropic medium, the properties of which will be close to those in spherical particles. This is true despite despite the strong anisotropy of the particles themselves. From a comparison of Figs. 2(a) and 2(c) for C_6H_6 and Figs. 2(b) and 2(d) for CS_2 it can also be seen that splitting of the dielectric loss spectrum depends strongly on the value of ε_h.

Liquid	Calculations		Experiment	
	Ordered medium	Disordered medium	GATR I (Si window)	GATR II (Al window)
I. $CHCl_3$			752	
Bulk	756.0			753
3D	759.1	758.1		
2D	760.1	759.8		759.2
1D	771.1	771	772	
P=1/3	761.1	673.2		
P=3	756.9	659.5		
II. C_6H_6			671	
Bulk	672.7			671
3D	673.3	673.3		
2D	673.8	673.6		675
1D	678.6	678.4	678.6	
P=1/3	674.4	673.8		
P=3	672.8	673.4		
III. CS_2			1501	
Bulk	1502.7			1500
3D	1505.4	1505.4		
2D	1507.5	1506.7		1513
1D	1527.7	1527.2	1531	
P=1/3	1509.7	1507.4		
P=3	1503.4	1506.0		

Table 1. Calculated and experimental peak positions, ν (cm^{-1}), of the most intense IR absorption band observed for liquid $CHCl_3$, C_6H_6 and CS_2 under various dielectric confinement conditions (ε_h=13.6, Si/Ge)

Host Matrix, ε_h	Bulk benzene, ν_{bulk}	Ordered			Disordered		
		1D	2D	3D	1D	2D	3D
2.2	672.7 (4.56)	681.4 (0.42)	677.3 (0.41)	675.7 (0.42)	673 (0.34) 681 (2.2)	676.2 (0.36)	675.7 (0.42)
5		680.4 (1.69)	675.3 (0.76)	674.3 (0.63)	680 (0.65) 673 (1.4)	674.8 (0.63)	674.2 (0.63)
11.56		678.9 (5.63)	674 (1.09)	673.5 (0.78)	678.8 (1.99)	673.8 (0.87)	673.5 (0.78)
16		678.2 (8.26)	673.7 (1.19)	673.3 (0.82)	678.1 (2.19)	673.6 (0.94)	673.3 (0.82)

Table 2. Peak position (in cm^{-1}) and intensity (in brackets) for dielectric loss spectra of two component composite medium consisting of ordered and disordered benzene mesoparticles under 1D, 2D and 3D dielectric confinement in various host matrices (f=0.1)

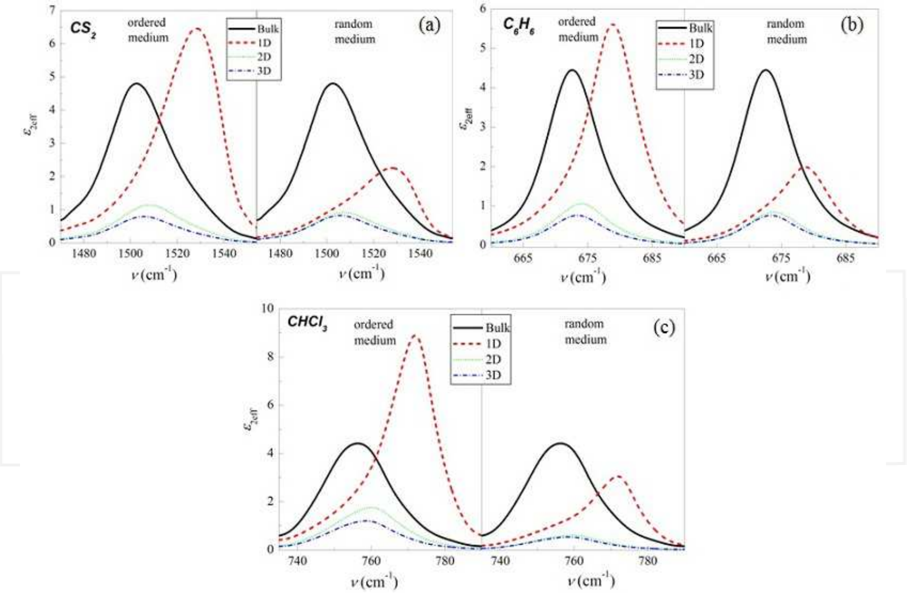


Fig. 3. Calculated dielectric loss spectra $\varepsilon_{2\text{eff}}(\nu)$ of liquids (a) CS_2, (b) C_6H_6 and (c) $CHCl_3$ under the conditions of different size confinement for ordered and disordered (random) media. The calculations were performed using equations (8) and (15a) at f=0.1, ε_h=11.56

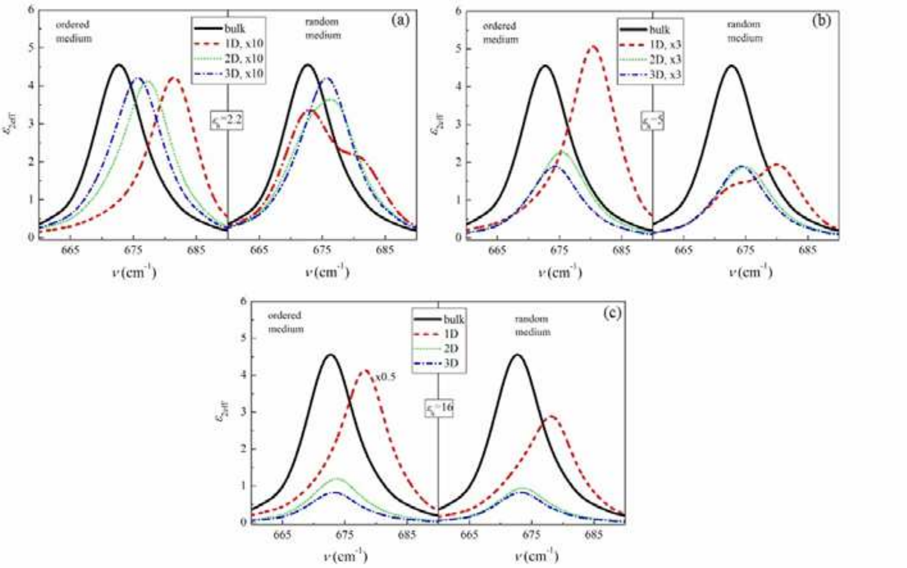


Fig. 4. Dielectric loss spectra of benzene mesoparticles calculated for ordered and random composite media at different confinement conditions for various matrixes (a) ε_h=2.2 glass, (b) ε_h=5, (c) ε_h=11.56 silicon, and (d) ε_h=16 germanium

Calculations of the dielectric loss spectra of mesoparticles of amorphous SiO_2 under various types of dielectric confinement are presented in Fig. 5 and summarised in Table 3. The calculations were performed at f = 0.2 and ε_h = 2.34 for KBr. The optical properties of amorphous SiO_2 were obtained from Ref. (Efimov, 1995). Amorphous SiO_2 has several absorption bands, with peaks at 468 cm^{-1} (Si-O-Si rocking vibrational mode), 808 cm^{-1} (O-Si-O bending mode) and 1082 cm^{-1} (Si-O asymmetric stretching mode). In our analysis, we focus mainly on the most intense band at 1082 cm^{-1}.

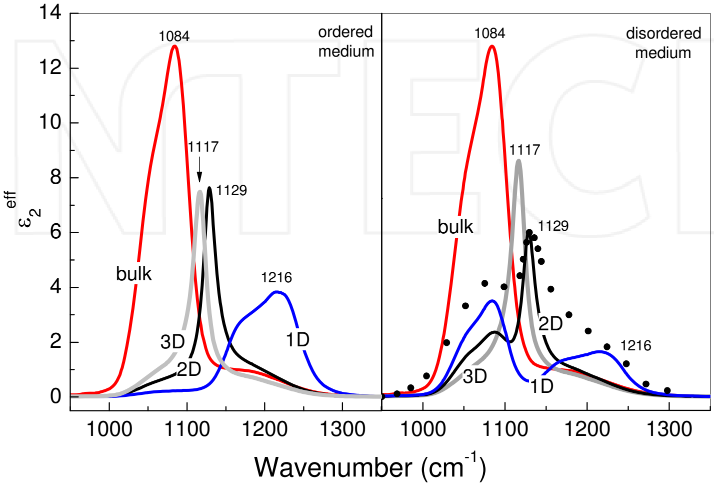


Fig. 5. Calculated dielectric loss spectra of bulk SiO_2 and its composites under dielectric confinement in a host medium with ε_h=2.34 (KBr) and filling factor f=0.2 for ordered (left panel) and disordered (right panel) mesoparticles. The circles on the right panel correspond to experimental data for SiO_2/Si rods in a KBr matrix from Noda et al., 2005

The principal features of the calculated spectra are shown in Figs. 2 - 5 and the results of our calculations are summarized in Tables 1 - 3. For all the calculated model composites viz. benzene, chloroform, carbon disulphide and SiO_2, the position of the dielectric loss spectral maximum, and its intensity, depends on the dielectric permittivity of the host medium. For larger ε_h, the peak position is shifted to smaller wavenumbers towards the peak of the bulk medium. The peak intensity increases significantly for larger values of ε_h, for example, by a factor of 2 for C_6H_6 in Figs. 2(a) and 4(c), for $CHCl_3$ in Fig. 3c and for CS_2 in Fig. 2(b).

In all cases, the maximal spectral shift of the dielectric loss spectrum is observed under 1D dielectric confinement. The peak positions for 2D and 3D confinement are closer to the peak position observed from bulk benzene and carbon disulphide. The difference in peak position for 2D and 3D confinement is very small and is more apparent for small ε_h. For benzene mesoparticles embedded in the host matrix, with ε_h = 2.2 or ε_h = 5, the appearance of the second peak is clearly seen under 1D confinement in disordered media (see Fig. 4 (a and b) and Table 2). Similar results are also apparent in Fig. 2(d) for CS_2 at ε_h = 3. Note that at smaller ε_h, the peak related to the bulk mode is more intense, while for larger ε_h, the peak corresponding to 1D confinement has a higher intensity. For both C_6H_6 and CS_2, the peak related to the bulk mode significantly reduces in intensity, indeed, it practically disappears at ε_h = 13.6 as seen in Fig. 2 (a and b). It is also worth noting that the peak intensity of the dielectric loss spectrum for an ordered composite medium at f = 0.1 and ε_h = 16, in a germanium host matrix, is approximately two times higher than that for bulk benzene (Table 2), while it is 10 times lower at ε_h = 2.2.

Sample	Dielectric matrix	Bulk, peak position	Calculations, DLF method, peak position, (cm^{-1})			Experiment, peak position, (cm^{-1})		
	ε_h	ν_{bulk}, cm^{-1}	ν_{3D}	ν_{2D}	ν_{1D}	ν_{3D}	ν_{2D}	ν_{1D}
	1, air		1143	1164	1252			1257[d]
					1253[a]			1253[e]
SiO$_2$					1250[b]			
	2.34, KBr	1084	1117	1129	1216		1130[c]	
						1109		
	1.77, water		1109	1118	1142			

[a]From transmission spectra calculated at 65° of incident light using TMM; [b]from minimum of the reflection spectrum calculated for p-polarized light at incidence angle of 70° using expressions of multilayer stack optics; [c]experimental data from Noda et al., 2005; [d]experimental data from Shaganov et al., 2001; [e]experimental data from Röseler, 2005.

Table 3. Experimental and calculated peak positions, ν (cm^{-1}), of IR absorption bands observed for SiO$_2$ under various dielectric confinement conditions

3.2 Experimental

Infrared absorption spectra were measured on an FTS 6000 Fourier Transform Infrared (FTIR) spectrometer using a commercially available Grazing angle Attenuated Total Reflection (ATR), attachment from the Harrick Scientific Corporation. Absorption measurements were made on both thick and thin layers of liquid, as well as on thin solid films. For measurements of absorption for the bulk, thick layer, a drop of liquid approximately 1 mm thick was placed in the middle of the Ge ATR element. In order to achieve dielectric confinement in the liquids studied, three methods were used. In the first technique, absorption spectra were measured using the Grazing angle Attenuated Total Reflection (GATR) attachment. A thin film of liquid was obtained by confining the liquid between the Ge ATR prism and the 4 mm thick silicon top window, see Fig. 6(a). In the second method, an Al coated glass substrate was used instead of the Si top window Fig. 6(b). The strength of window compression was changed using the GATR pressure applicator control. Measurements were performed in p-polarized light at a 60° angle of incidence. The third method for exploring dielectric confinement effects is based on the use of a macro-porous silicon matrix, with liquid infiltrated into the pores (Perova et al., 2009). In our study, porous Si samples were fabricated by electrochemical etching of single-crystalline (100) n-type Si wafers in a HF (48%) : H$_2$O = 2:3 solution. Etching was performed for 30 mins at a current density of 16 mA/cm^2. The resulting pore diameter was about 0.8 μm. All three liquids studied evaporated completely from the pores approximately 30-40 minutes after infiltration. Therefore, in-situ FTIR measurements were carried out immediately after liquid infiltration using a registration time of ~ 20 sec and a dwell time of ~ 5 sec.

All the liquids investigated were of high purity, purchased from Sigma-Aldrich. The Ge ATR prism and Si windows were new and of excellent optical quality. The Ge ATR element was carefully cleaned before the drop of liquid was placed on it. The glass substrate with the

Al layer was freshly prepared; a new element was used for each experiment. At least five separate experiments were performed for each liquid. These precautions enabled us to avoid the influence of any unwanted interactions.

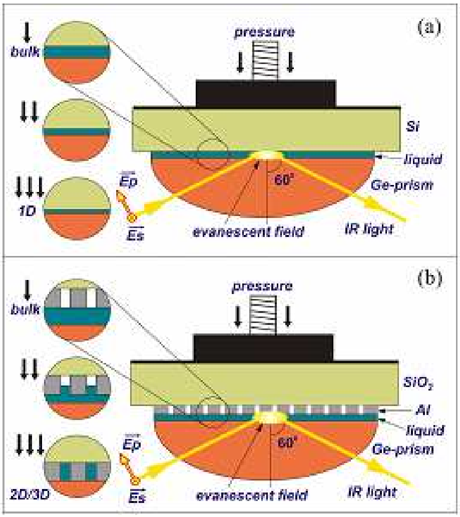


Fig. 6. Schematic of FTIR experiments using GATR attachment for (a) Si and (b) Al top window

Thin films of SiO_2 were deposited onto an Al coated glass substrate using an electron-gun evaporator. In order to register the LO-phonons, or the absorption spectrum of these materials under 1D confinement, we used an oblique incidence of light in *p*- and *s*-polarisations, using the Reflection-Absorption (RA) and GATR attachments, see Ref. (Shaganov et al., 2003) for details. For registration of spectra under 3D confinement, we used SiO_2 spherical particles of different diameters dissolved in water. Spherical particles of SiO_2, with a diameter of 193 nm, coated with an ultra-thin layer of surfactant to prevent particle conglomeration and dissolved in water, were purchased from Sigma-Aldrich. The distribution of particles size in the solution is ± 5 - 10 nm, as guaranteed by the manufacturer.

3.3 Comparison of experimental and calculated data

a) Liquid systems

Absorption spectra obtained for three of the liquids under investigation are shown in Figs. 7(a) - 7(c). From Fig. 7(a), it is apparent that, for a thick chloroform layer, an absorption band with a peak position at ν = 752 cm^{-1}, corresponding to the bulk mode, is observed. At the maximum possible confinement, when the layer is only ~ 100-200 nm thick, the absorption peak shifts to ν = 771 cm^{-1}. This value agrees very well with calculations of the dielectric loss spectrum of liquid $CHCl_3$ under 1D dielectric confinement (see Table 1). We believe that the line width increase observed for the absorption band at 760 cm^{-1} is due to the fact that the absorption band for this intermediate case is a superposition of the absorption bands obtained in the presence and absence of the dielectric confinement effect. We would like to emphasise that the position and shape of the weaker absorption band, observed for $CHCl_3$ at ν = 1215 cm^{-1}, remained unchanged as expected (see Fig. 7(a)).

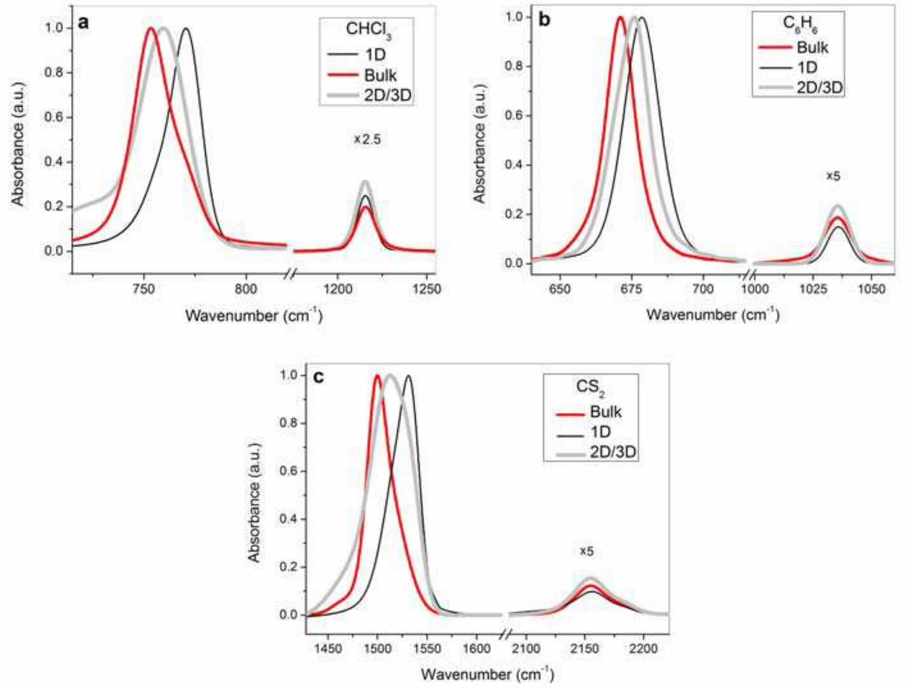


Fig. 7. Normalized infrared spectra of liquids (*a*) $CHCl_3$, (*b*) C_6H_6 and (*c*) CS_2 registered with GATR attachment. Note that the absorbance of the vibrational bands with small intensities was multiplied by slightly different factors, shown beside the bands, to demonstrate clearly that they have the same peak position

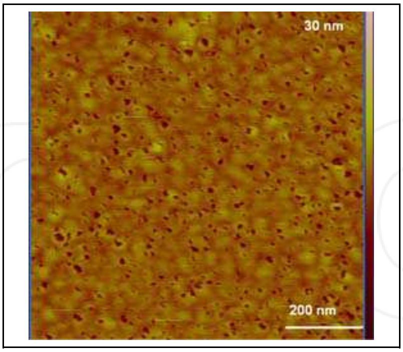


Fig. 8. Top view AFM image of the Al coated glass substrate

Similar behaviour was observed for liquid benzene (Fig. 7(b)), where the frequency of the spectral maximum for the bulk liquid, initially observed at ν = 671 cm^{-1}, was shifted under strong confinement to ν = 679 cm^{-1}, corresponding to 1D dielectric confinement of a very thin layer of C_6H_6. The same effect was observed in liquid CS_2 (Fig. 7(c)) with a frequency shift from v = 1501 cm^{-1}, observed for the bulk material, to v = 1532 cm^{-1}, measured under 1D confinement. As in the case of chloroform, the layer thickness for the benzene and carbon disulphide was estimated to be 100 – 200 nm. The position and shape of the weak absorption band observed at 1036 cm^{-1} for C_6H_6, and at 2155 cm^{-1} for CS_2, remain unchanged (see Figs. 7(b) and 7(c)). We also note that the largest peak shift due to dielectric confinement was observed for CS_2 with the largest integrated intensity of infrared absorption band of all the liquids investigated.

In order to measure the vibrational spectra of these liquids under 2D/3D dielectric confinement, we modified the experimental setup as follows (see Fig. 6(b)). The top silicon window was replaced with a 5 mm thick glass plate, coated with a thin, ~ 0.1 μm, Al layer. The coating was applied by evaporation of Al wire in a bell jar evaporator. Under these evaporation conditions, the Al film contains pores, with diameters ranging from a few microns to tens of nanometers. A small drop of liquid was placed on top of the ATR Ge prism, then the Al coated glass plate was placed on top and the experiment was immediately run as the level of compression of the top glass window was increased. The effects of confinement on the liquid spectra were practically identical to those described earlier, with the exception of the last stage. When the thin layer of liquid evaporated completely, the maximum frequency in the spectrum shifted to ~ 760 cm^{-1} for $CHCl_3$, 676 cm^{-1} for C_6H_6 and to 1513 cm^{-1} for CS_2 (see Fig. 7 and Table 1). These frequencies are in good agreement with data calculated for 2D or 3D dielectric confinement. This can be seen in Table 1 and from Figs. 2, 3 and 7. Note that the infiltration of the liquid into the voids, or pores, in the Al layer was confirmed by the fact that the spectra related to 2D/3D confinement were still observed several hours after initial sample preparation, when the thin layer of liquid between the Ge prism and the Al coated substrate had definitely evaporated. As the deposited layer of Al is too thin to consider the "porous" Al layer obtained as a matrix for the fabrication of liquid wires, the diameter/length ratio of the pores obtained suggests that we are dealing with liquid spheres embedded in a porous Al matrix situated at the top of the Ge prism. The results of surface analysis of the Al coated glass substrate using an Atomic-Force Microscope (AFM) confirms the existence of the void structure (with width and depth of voids at around ~20-40 nm and ~10-15 nm, accordingly) of the substrates used for these experiments (see Fig. 8). From Table 1, the peak positions observed under 2D and 3D confinement are close, making it difficult to draw firm conclusions. Nevertheless, we believe that with this experiment, it is possible to obtain information on the absorbance spectra of the liquids investigated under 3D confinement.

Fig. 9 shows the behaviour of the absorption spectra of benzene infiltrated into a macro-porous silicon matrix registered at various times after infiltration. The position of the absorption band for benzene immediately after infiltration was close to the frequency of the bulk mode of C_6H_6 (ν = 673 cm^{-1}). During the course of evaporation, the peak position shifted to higher frequencies and ν = 682 cm^{-1} at the end of the registration process. We believe that at the beginning of the registration process, the pores were totally filled with liquid benzene. Since the pore diameter is larger than that necessary to satisfy the criteria for dielectric confinement, the absorption spectrum observed is that from the bulk liquid. In the course of drying out, the liquid bulk phase of the C_6H_6 evaporates, leaving a thin layer of adsorbed liquid on the pore surface. When the electric field of the incident light is oriented parallel to the sample surface, the conditions for the registration of 1D dielectric confinement are met, as shown in the insert in Fig. 9. These results are in good agreement with our calculations shown in Fig. 3(b). Similar results were obtained for $CHCl_3$ and for CS_2, these results are summarized in Table 1. It should be noted that, due to the faster evaporation of $CHCl_3$ and the CS_2 liquids from the pores, it was not possible to register the bulk mode at the beginning of the registration process. In conclusion, we note that since exact values of layer thickness and the sphere diameter distribution were not known, we were unable to calculate the imaginary part of the dielectric function from the experiment, in order to compare this with the calculated values $\varepsilon_2^{eff}(\nu)$. Therefore, the position of the absorbance spectra, A(v), was used for this comparison. However, it is well known that for strong and narrow isolated absorption bands, the peak positions of $\varepsilon_2(\nu)$ and A(v) are close to each other. Our estimates have shown that, in this case, the deviation does not exceed 1 - 2 cm^{-1}.

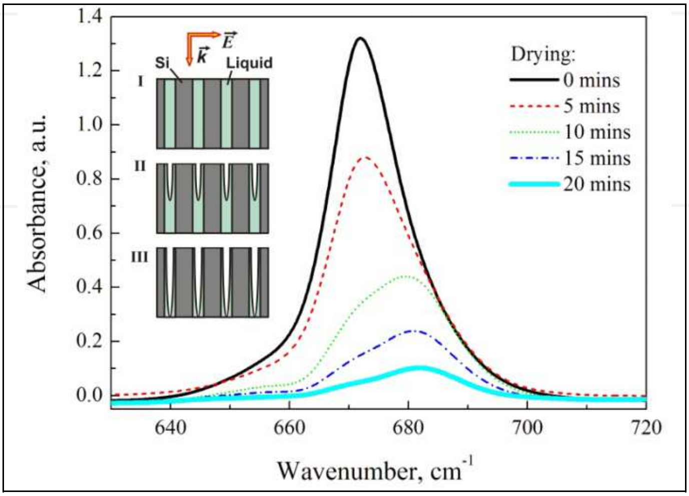


Fig. 9. Absorbance spectra of benzene infiltrated into silicon pores. Insert: Schematic diagram of the conversion of liquid infiltrated into the macro-porous silicon matrix from a bulk liquid phase to a liquid under 1D dielectric confinement as a result of the drying process. Reproduced with permission of journal Chemical Physics Letters (Perova et al., 2009)

(b) Amorphous solids (SiO_2)

Calculations of the dielectric loss spectra of mesoparticles of amorphous SiO_2 under various types of dielectric confinement are presented in Fig. 4 and summarised in Table 3. The peak position for 1D dielectric confinement is confirmed experimentally in our earlier paper (Shaganov et al., 2003) for 70 nm thick thermally grown oxide, as well as by a number of other papers where the IR spectra of thin (Shaganov et al., 2001; Almeida, 1992; Olsen & Schimura, 1989) and ultra-thin (with a thickness of 5nm) (Tolstoy et al., 2003) films of amorphous SiO_2 were measured under oblique incidence of IR light. It is worth noting that the shift in the peak position of the Si-O-Si band at ~ 1100 cm^{-1} to higher frequencies (~ 1253 cm^{-1}) was also observed in an SiO_2 thin film under oblique incidence of light using infrared spectroscopic ellipsometry (see Ref. (Röseler , 2005) for details).

In order to lend further support to the model suggested, we performed calculations of the transmission spectra for SiO_2 thin films under oblique incidence of light, corresponding to 1D confinement, using a 2 x 2 Transfer Matrix Method (TMM) (Azzam & Bashara, 1977). The peak positions of the transmission spectra obtained are included in Table 3. In addition, the peak position of the transmission spectra for thin SiO_2 films, calculated at oblique incidence of light, is also shown in Table 3. These calculations are performed using expressions from paper (Shaganov et al., 2001). The results obtained for 1D confinement in an SiO_2 thin film demonstrate very good agreement between the theory developed for the calculation of the optical properties of a multilayer stack and the approach suggested in this paper. These results are also in agreement with both spectroscopic ellipsometry and infrared spectroscopy experiments at oblique incidences of light.

The spectra calculated for the disordered composite, for 2D confinement, are confirmed experimentally using results published recently in Ref. (Noda et al., 2005), where the infrared spectra of SiO_2/Si disordered nanowires embedded in KBr pellets were investigated. We believe that the peak observed in paper (Noda et al., 2005) at ~ 1130 cm^{-1} and assigned by the authors to a highly disordered structure of thin SiO_2/Si nanowires can be reinterpreted, in the light of the results presented in this paper, as being due to 2D dielectric confinement of amorphous SiO_2.

Finally, the infrared spectra of spherical SiO_2 particles in an aqueous solution have been measured using a GATR attachment. Spherical particles, with diameters of 193 nm, coated with an ultra-thin layer of surfactant to prevent particle conglomeration and dissolved in water, were supplied by Sigma-Aldrich. As noted earlier, the particle size distribution guaranteed by the manufacturer is ± 5 - 10 nm. The solution was shaken intensely before placing a drop of the liquid onto the Ge ATR prism. The infrared spectra of the SiO_2 mesoparticles obtained in this experiment are shown for spherical particles of diameter 193 nm in Fig. 10, along with calculations for 3D confinement at ε_h = 1.77. Good agreement between the experimental and calculated spectra can be seen for this composite. The minor discrepancies between the experimental and calculated spectra of the spherical SiO_2 particles observed in the spectrum wing regions is due to the fact that silicon dioxide can exist in various forms such as amorphous quartz, fused quartz or quartz doped with impurities. The exact structure of the Sigma-Aldrich silicon dioxide spherical particles is not known. For our calculations, the optical constants of amorphous quartz were taken from the literature, which can result in differences between the calculated spectra from the experimental data in the wing regions.

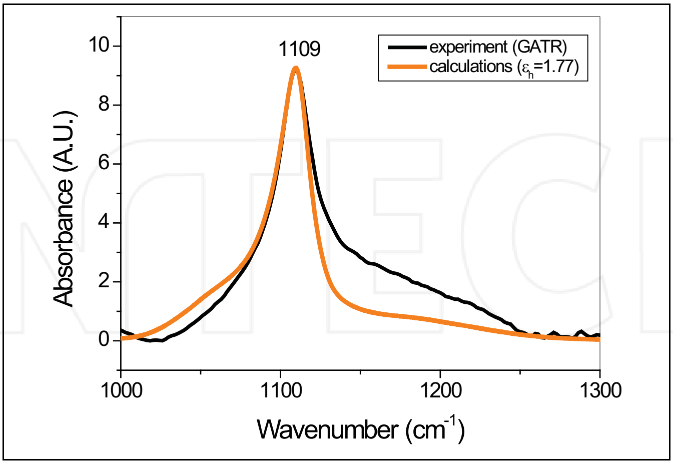


Fig. 10. Absorbance spectra of SiO_2 spherical particles (thin line) diluted in water with a diameter of 193 nm shown together with the calculated spectra (thick line) based on Eqn. (15a) at ε_h= 1.77

4. Conclusions

The experimental results presented here demonstrate good agreement with calculations made using the model suggested for estimating the effect of 1D, 2D and 3D dielectric confinement on the IR spectra of condensed matter. The results obtained allow us to conclude that the physical mechanism responsible for the shift of the absorption peak of small particles experiencing different types of dielectric confinement is the same, regardless of the nature of the condensed medium, whether crystalline or amorphous, solid or liquid. The shift is due to the local field effect acting on the size confined particles, surrounded by the dielectric matrix.

The expression obtained for particle absorption under 1D confinement is the same as that for the dielectric loss spectrum of the crystal at the frequency of the longitudinal-optical phonons (Berreman, 1963). This indicates that similar absorption bands to those seen under 1D confinement will be observed near the minimum of the real part of the dielectric function ($\mathrm{Re}\varepsilon(\nu)$) function in any condensed medium. This has been confirmed already in other studies on thin films of amorphous solids (Payne & Inkson, 1984; Röseler, 2005; Tolstoy et al., 2003; Röseler, 2005; Shaganov et al., 2005), polymer monolayers (see (Yamamoto & Masui, 1996) and references therein) as well as for the thin liquid films investigated in this work. We conclude that the absorption bands, observed earlier and ascribed to the Berreman effect (Berreman, 1963), are a particular case of the manifestation of 1D confinement. This conclusion is supported by a study by (Lehmann, 1988), where it was shown that the appearance of the absorption band at the frequency of the LO-phonons in an amorphous dielectric is a consequence of the boundary conditions in a dielectric film at an oblique incidence of the probe beam.

The numerical and experimental results described above indicate that relatively large spectral effects can be expected as a result of dielectric confinement. Our results convincingly demonstrate that the blue shift of the absorption bands under dielectric confinement can be significant, and must be taken into account when interpreting experimental spectroscopic data from composite systems. Of course, we ignored the resonance dipole-dipole interactions, which are negligible at particle volume concentrations of less than 1% and will not impact the accuracy of the calculations for filling factors of less than 10%.

We note that the expressions obtained only deal with isolated particles of spheroidal shape and are valid when the spheroidal semi-axes in either one, two or three directions are satisfied by the conditions $d_z << \lambda$ or $d_x = d_z << \lambda$ while remaining larger than atomic dimensions. Therefore, the approach suggested here can be used for a general description of the spectral characteristics of arbitrary micro-objects, or more specifically, sub-micron microcrystalline particles, under dielectric confinement. One of the disadvantages of our approach is the absence of the size parameter in the model. Obviously, the response of the dielectric medium will change with a decrease in the particle size, approaching frequencies characteristic of the limited cases of 1D, 2D and 3D confinement considered in this work. There is evidence to indicate that this assumption is justified, assuming the optical properties are linear. We believe that particle size will play a significant role only when quantum confinement effects influences their non-linear optical properties. Other physical phenomena need to be taken into account in order to calculate the absolute value of the imaginary part of dielectric function. These phenomena include sample surface roughness, polaritons, diffraction and scattering. In addition, the specificity of the molecular orientation in, for example, Langmuir-Blodgett films, or the film structure, that is, anisotropy, of the oxide or island-like surface structure for ultra-thin films, or monolayers, may also influence the shape and position of the IR spectra. Therefore, further development of this theory and its experimental verification is required.

We conclude that dielectric confinement offers considerable promise as a method for tuning the absorption properties of composite media. The approach allows control of both the position and intensity of the dielectric loss spectrum of the absorbing medium embedded in a composite. Furthermore, the absorption efficiency can be increased significantly due to local field effects. Clearly, further development of simple models for the description of the spectral properties of composite media, including meso-composites based on porous semiconductors, as well as other porous media with absorbent inclusions, is still necessary. The most important applications of these studies are to the analysis of the absorption spectra of industrial smokes, toxic aerosols and liquid droplets, (see Ref. (Carlton et al., 1977; Carlton, 1980) for example) as well as for colloidal optofluidic systems (Psaltis, 2006).

5. Acknowledgments

The authors would like to acknowledge V. Melnikov, S. Dyakov and Alex Gunko for help with the experiments and calculations. V. Tolmachev is acknowledged for useful discussions.

6. References

Aspnes, D.E. (1982). Local field effects and effective-medium theory: A macroscopic perspective. *Amer.J.Phys.*, 50, 8, 704-708, ISSN: 00029505.

Almeida, R.M. (1992). Detection of LO modes in glass by infrared reflection spectroscopy at oblique incidence, *Phys. Rev. B*, 45, 1, 161-170, 1 January. ISSN: 1098-0121.

Azzam, R.M.A. & Bashara, N.M. (1997). *Ellipsometry and polarized light*, Elsevier B.V., ISBN: 0-444-87016-4, Amsterdam, The Netherlands.

Bakhshiev, N. G.; Girin, O. P. & Libov, V. S. (1962). The relationship between the measured and intrinsic absorption spectra in condensed medium.I. *Sov.Phys.Dokl.* 145, 3, 1025-1027.

Bakhshiev, N. G.; Girin, O. P. & Libov, V. S. (1963). The relationship between the measured and intrinsic absorption spectra in condensed medium.II. *Sov. Opt.* & Spectr. 1963, 14, 745-750 (English. Translation: *Opt.Spectry.* 14, 2, 395-400).

Barnes, D. W. & and Schatz, P. N. (1963). Optical Constants and Absolute Intensities from Infrared Reflection Measurements. The 6.6-μ Band of Liquid CS_2 and 13-μ Doublet of liquid CCL_4. *J. Chem. Phys.*, 38, 11, 2662-2667, ISSN: 0021-9606.

Berreman, D.W. (1963). Infrared Absorption at Longitudinal Optic Frequency in Cubic Crystal Films, *Phys. Rev.*, 130, 6, 2193-2198, 15 June.

Bruggeman, D.A.G. (1935). Berechnung verschiedener physikalischer Konstanten von heterogenen Substanzen, *Ann. Phys.* (Leipzig) 24, 636-679.

Bőttcher C.I.F. (1952). *Theory of Electric Polarisation*, Elsevier, Amsterdam.

Cahay, M.; Leburton, J.-P.; Lockwood, D.J.; Bandyopadhyay, S. & Harris, J.S. (2001). *Quantum Confinement VI: Nanostructured materials and Devices*, Electrochemical Society, Inc., ISBN: 1-56677-352-0, New Jersey, USA.

Cao, G. (2004). *Nanostructures and Nanomaterials: Synthesis, Properties and Applications*, Imperial College Press, ISBN: 1-86094-415-9, London, UK.

Carlon, H. R.; Anderson, D. H.; Milham, M. E.; Tarnove, T. L.; Frickel, R. H. & Sindoni, I. (1977). Infrared extinction spectra of some common liquid aerosols, *Appl. Opt.*, 16, 6, 1598-1605, ISSN: 1559-128X.

Carlon, H. R. (1980). Aerosol spectrometry in the infrared, *Appl. Opt.*, 19, 13, 2210-2218. ISSN: 1559-128X

Chemla, D. S. & Miller, D.A.B. (1986). Mechanism for enhanced optical nonlinearities and bistability by combined dielectric–electronic confinement in semiconductor microcrystallites, *Opt. Letters*, 11, 8, 522-524, ISSN: 0146-9592.

Clifford, A. A. & Crawford, B. (1966). Vibrational Intensities. XIV. The Relation of Optical Constants to Molecular Parameters. *J. Phys. Chem.*, 70, 5, 1536-1543.

Cohen, R.W.; Cody, G. D.; Coutts, M. D. & Abeles, B. (1973). Optical Properties of Granular Silver and Gold Films. *Phys. Rev. B,* 8, 8, 3689-3701, ISSN: 1098-0121.

DeLeeuw, S. W. & Thorpe, M. F. (1985). Coulomb splittings in glasses. *Phys. Rev. Lett.*, 55, 26, 2879-2882, ISSN: 0031-9007.

Dolgaleva, K.; Boyd, R.W. & Millionni, P.W. (2009). The effects of local fields on laser gain for layered and Maxwell Garnett composite materials. *J.Opt. A: Pure Appl. Opt.*, 11, 2, ISSN:14644258.

Efimov, A.M. (1995). *Optical Constants of Inorganic Glasses*, CRC Press, Inc., ISBN: 0-8493-3783-6, New York, USA.

Foss, C.A.; Hornyak, G.I., Stockert, J.A. & Martin C.R. (1994). Template-Synthesized Nanoscopic Gold Particles: Optical Spectra and the Effects of Particle Size and Shape. *J.Phys.Chem. B,* 98, 11, 2963-2971, ISSN: 1520-6106.

Fröhlich, H. (1949). *Theory of Dielectrics*, Clarendon Press, Oxford.

Ghiner, A.V. & Surdutovich, G. I. (1994). Method of integral equations and an extinction theorem for two-dimensional problems in nonlinear optics. *Phys. Rev. A*, 50, 1, 714-723, ISSN: 1050-2947; (994). Beyond the Lorentz-Lorenz Formula. *Optics & Photonics News*, 5, 12, December, 34-35, ISSN: 10476938.

Golovan, L. A.; Kuznetsova, L. P.; Fedotov, A. B.; Konorov, S. O.; Sidorov-Biryukov, D. A.; Timoshenko, V. Yu.; Zheltikov, A. M.; and Kashkarov, P. K. (2003). Nanocrystal-size-sensitive third-harmonic generation in nanostructured silicon. *Appl. Phys. B,* 76, 4, 429-433, ISSN: 0946-2171.

Golovan, L. A.; Timoshenko, V. Yu. & Kashkarov, P. K. (2007). Optical properties of porous-system-based nanocomposites, *Physics-Uspekhi*, 50, 6, 595-612, ISSN: 1063-7869; Golovan', L.; Kashkarov, P. & Timoshenko, V. (2007). Form birefringence in porous semiconductors and dielectrics: A review. *Crystal. Reports,* 52, 4, 672-685, ISSN: 1063-7745.

Heitler, W. (1975). *Quantum Theory of Radiation*, 3rd ed., Wiley, ISBN: 0486645584, New York.

Hornyak, G.L. ; Patrissi, C.J. & Martin, Charles R. (1997). Fabrication, Characterization, and Optical Properties of Gold Nanoparticle/Porous Alumina Composites: The Nonscattering Maxwell-Garnett Limit. *J.Phys.Chem B*, 101, 9, 1548-1550, ISSN: 1520-6106.

Iglesias, J. E.; Ocana, M. & Serma, C.J. (1990). Aggregation and Matrix Effects on the Infrared Spectrum of Microcrystalline Powders. *Appl. Spectr.*, 44, 3, 418, ISSN: 0021-9037.

Lamberti, C. (2008). *Characterization of Semiconductor Heterostructures and Nanostructures*, Elsevier, ISBN: 0-44453-099-1, Amsterdam, The Netherland; Oxford, UK.

Lehmann, A. (1988). Theory of Infrared Transmission Spectra of Thin Insulating Films, *Phys. Stat. Sol. B,* 148, 1, 401-405.

Liu, A. (1994). Local-field effect on the linear optical intersubband absorption in multiple quantum wells, *Phys. Rev. B,* 50, 12, 8569-8576, ISSN: 1098-0121.

Mallet, P.; Guerin, C. A. & Sentenac, A. (2005). Maxwell-Garnett mixing rule in the presence of multiple scattering: Derivation and accuracy, *Phys. Rev. B,* 72, 1, 14205/1-9, ISSN: 1098-0121.

Maxwell-Garnett, J.C. (1904). Colours in metal glasses and metal films. *Philos. Trans. R. Soc. London, Sect. A*, Vol. 203, 385-420; (1906). *Philos. Trans. A,* 205, 237-288.

Noda, T.; Suzuki, H.; Araki, H.; Yang, W.; Ying, S. & Tosa, M. (2005). Microstructures and IR spectra of long amorphous SiO2/Si nanowires. *Appl.Sur.Sci*., 241, 1-2, 231-235. ISSN: 0169-4332.

Olsen, J.E. & Schimura, F. (1989). Infrared reflection spectroscopy of the SiO2-silicon interface. *J. Appl. Phys*., 66, 3, 1353-*1358*, ISSN: 0021-8979.

Osborn, J. A. (1945). Demagnetizing Factors of the General Ellipsoid. *Phys. Rev*., 67, 11-12, 351-357.

Palik, D. (1978). *Optical Constants of Solids*, Delta Academic Press, New York, *ISBN* 0-521-46829-9.

Perova, T.S.; Shaganov, I. I.; Melnikov, V.A. & Berwick, K. (2009). Direct evidence of the dielectric confinement effect in the infrared spectra of organic liquids, *Chem.Phys.Lett*., 479, 1-3, 81-85, ISSN: 0009-2614.

Psaltis, D.; Quake, S.R. & Yang, C. (2006). Developing optofluidic technology through the fusion of microfluidics and optics, *Nature* 442, 7101, 381-386. ISSN: 0028-0836.

Röseler, A. (2005). Spectroscopic Infrared Ellipsometry. In: *Handbook of Ellipsometry,* H.G. Tompkins, E.I. Irene, (Ed.), 789-797, Springer-Verlag GmbH & Co. KG, ISBN: 3-540-22293-6, Heidelberg, Germany.

Schmitt-Rink, S., Miller, D. A. B. & Chemla, D. S. (1987). Theory of the linear and nonlinear optical properties of semiconductor microcrystallites. *Phys.Rev. B* 35, 15, 8113-8125, ISSN: 1098-0121.

Shaganov, I. I.; Perova, T. S.; Moore, R. A. & Berwick, K. (2001). Spectroscopic characteristics of SiO and SiO_2 solid films: Assignment and local field effect influence. *J. Mater. Science: Mater. Electron.,* 12, 4-6, 351-355, ISSN: 09574522.

Shaganov, I. I.; Perova, T.S.; Moore, R.A. & Berwick, K. (2003). Local field effect on infrared phonon frequencies of thin dielectric films. *Proceed. SPIE*, 4876, 1, 1158-1167, ISSN: 0277-786X.

Shaganov, I. I.; Perova, T. S.; Moore A. R. & Berwick, K. (2005). Effect of the Internal Field on the IR Absorption Spectra of Small Particles in the Case of 3D, 2D, and 1D Size Confinement. *J. Phys. Chem. B,* 109, 20, 9885-9891, ISSN: 1520-6106.

Shaganov, I.; Perova, T.; Melnikov, V.; Dyakov, S. & Berwick, K. (2010). The Size Effect on the Infrared Spectra of Condensed Media under Conditions of 1D, 2D and 3D Dielectric Confinement. *J. Phys.Chem. C*, 114, 39, 16071-16081, ISSN: 19327447.

Spanier, J. E. & Herman, I. P. (2000). Use of hybrid phenomenological and statistical effective-medium theories of dielectric functions to model the infrared reflectance of porous SiC films, *Phys. Rev. B*, 61, 15, 10437-10450, ISSN: 1098-0121.

Timoshenko, V. Yu.; Osminkina, L. A.; Efimova, A. I.; Golovan, L. A. & Kashkarov, P. K., (2003). Anisotropy of optical absorption in birefringent porous silicon. *Phys. Rev. B*, 67, 11, 113405/1-4, ISSN: 1098-0121.

Tolstoy, V.P.; Chernyshova, I.V. & Skryshevsky, V.A. (2003). *Handbook of Infrared spectroscopy of ultrathin films*, John Wiley & Sons, Inc., ISBN: 9780471234326, Hoboken, New Jersey.

Tolstykh, T.S.; Shaganov, I.I. & Libov, V.S. (1974) Spectroscopic properties of optical transitions in the lattice vibration region for ionic crystals, *Sov. Phys. Solid State*, 16, 3, 431-434.

Ung, T.; Liz-Marzán, L.M. & Mulvaney, P. (2001). Optical Properties of Thin Films of Au@SiO_2 Particles. *J.Phys.Chem B,* 105, 17, 3441-3452, ISSN: 1520-6106.

Yamamoto K. & Masui A. (1996). TO-LO Splitting in Infrared Spectra of Thin Films. *Appl. Spectr.,* 50, 6, 759-763. ISSN: 0021-9037

Zolotarev, V. M.; Morozov, V. N. & Smirnova, E. V. (1984). *Optical constants of natural and technical media*, Chemistry, Leningrad.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Fourier Transform Based Hyperspectral Imaging

Marco Q. Pisani and Massimo E. Zucco
National Institute of Metrological Research
Italy

1. Introduction

A hyperspectral imaging system (HSIS) is a combination of an imaging device and a spectrometer. The result is a 2D image combined with the third dimension containing the spectral composition of each pixel of the image. Spectrometers normally implemented in hyperspectral imaging systems are made by integrating a dispersive means (a prism or a grating) in an optical system, with the drawback of having the image analyzed per lines and some mechanics integrated in the optical system, cfr. Fig 1, (Sellar & Boreman, 2005).

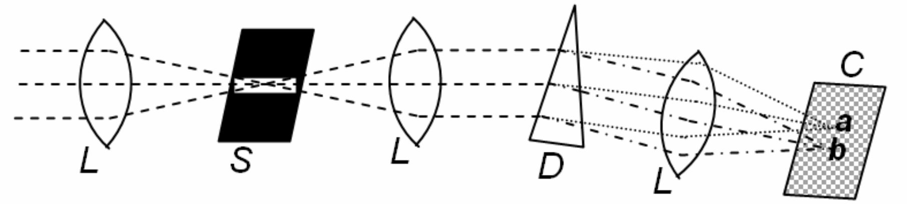


Fig. 1. Classical "pushbroom" hyperspectral imaging camera. L: collimating lenses, S: entrance slit selecting a row of the image in the focal plane; D: dispersive means (prism or grating) dispersing light in the direction orthogonal to the entrance slit; C: camera sensor where the combined image is focused. Points *a* and *b*, representing a pixel of the row selected at the entrance, are imaged at different vertical coordinates

Alternatively, HSIS devices are based on optical band-pass filters either tuneable or fixed and the spectrum has to be scanned in steps. In Fig 2 an example of the spectral transmissivity of a tunable band pass filter. Since the spectral transmissivity depends on the wavelength, HSIS systems have to be calibrated in advance and some mathematical manipulations are required to obtain the final hyperspectral image.
A third kind of spectrometer implemented in HSIS is based on interferometers (Alcock & Coupland, 2006), where the spectrum for each pixel is obtained by applying Fourier transform based algorithm to the signal (called interferogram) obtained by scanning the optical path difference *OPD*. The same technique has been used for decades by spectroscopists to obtain high resolution absorption spectra by using Michelson (or two-beam) or Fabry-Perot F-P (multi-beam) interferometer. There are many features that make

interferometer based spectrometers superior to conventional spectrometers. First, the Felgett or multiplex advantage arises from the fact that there is no spectral scanning and all the spectral components are acquired at the same time. Second, the Jacquinot or throughput advantage originates from the fact that the aperture used in FTIR spectrometers has a larger area than the slits used in dispersive spectrometers, thus enabling higher throughput of radiation. These two effects combined together make the interferometer based spectrometer a faster (or equivalently having a higher luminosity) instrument with respect to the other spectrometers at the same resolution. We have realized a HSIS based on a F-P spectrometer that will be discussed in details in section 3. In section 4 the application of our HSIS will be presented. In section 2 the mathematical manipulation to obtain spectra with the Michelson spectrometer will be discussed.

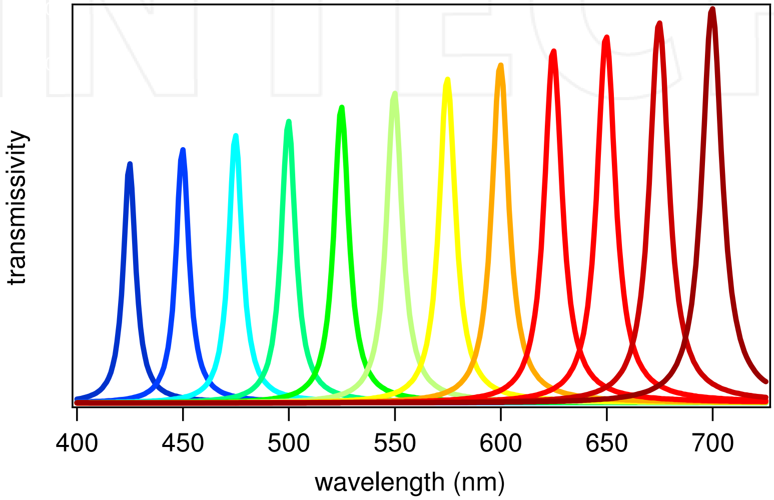


Fig. 2. The spectral response of a tunable filter to be used for HSIS

2. Michelson spectrometer

This section is devoted to the presentation of the Michelson spectrometer and to the mathematical manipulation to calculate spectra. In a Michelson spectrometer (Fig 3) the incoming light is divided in two beams by the beam splitter, after the two beams have travelled different paths, they are finally recombined on the detector where interference is measured. The intensity on the detector varies with the optical path difference *OPD* or retardation δ, double of the mirror displacement x.
When the incoming light is emitted by a monochromatic source and the two beams have the same intensity on the detector, the interferogram signal is represented by the equation

$$\overline{I}(\delta)=0.5B(\tilde{\nu}_o)(1+\cos(2\pi\tilde{\nu}_o\delta)) \tag{1}$$

where $\tilde{\nu}_o$ is the wavenumber $\tilde{\nu}_o=1/\lambda_o$ and $B(\tilde{\nu}_o)$ represents the intensity of the source at $\tilde{\nu}_o$. By using the frequency $\nu_o=c\tilde{\nu}_o=c/\lambda_o$, equation (1) is transformed in:

$$\overline{I}(\delta) = 0.5B(\nu_o)\big(1 + \cos(2\pi\nu_o\delta/c)\big) \tag{2}$$

The interferogram in (2) has two components, a constant component equal to $0.5B(\tilde{\nu}_o)$ and a modulated component equal to

$$\overline{I}(\delta) = 0.5B(\nu_o)\cos(2\pi\nu_o\delta/c) \tag{3}$$

Equation (3) is incomplete, there are several factors that reduce the spectral response of the system and the resulting signal: the optical elements (beam splitter, mirrors, lenses) and detectors normally have a non uniform responsivity. Moreover, the electronic devices used to condition the signal have a non uniform frequency dependence. All these different contributions are counted in the term $S(\nu_o)$, giving

$$\overline{I}(\delta) = S(\nu_o)\cos(2\pi\nu_o\delta/c) \tag{4}$$

When the source is broadband and continuum, the interferogram can be represented by the cosine Fourier Transform integral

$$I(\delta) = \frac{1}{c}\int_{-\infty}^{\infty} S(\nu)\cos(2\pi\nu\delta/c)d\nu \tag{5}$$

and the spectrum by

$$S(\nu) = 2\int_{0}^{\infty} I(\delta)\cos(2\pi\nu\delta/c)d\delta \tag{6}$$

I(δ) in (6) is based on an infinite and continuum retardation δ. In practice, the signal is sampled at finite sampling interval Δs and consists of N discrete, equidistant points and equation (6) transforms in (7), where all the constants have been discarded: the discrete version of the cosine FT, discrete cosine transform DCT. The maximum retardation is N Δs.

$$S[k\cdot\Delta\nu] = \sum_{n=0}^{N-1} I[n\cdot\Delta s]\cos(2\pi nk/N) \tag{7}$$

DCT means that in the time domain the Fourier series decomposes the periodic function into a sum of cosine and in the frequency domain it could be seen as if the signal intensity is divided in multiple adjacent frequency bins. For N equi-spaced points in the retardation at interval N Δs, we have N equi-spaced bins in the spectrum with spacing $\Delta\nu$ related to Δs by the formula

$$\Delta\nu = \frac{c}{N\Delta s} = \frac{c}{\delta_{\max}} \tag{8}$$

Therefore the resolution $\Delta\nu$ is inversely proportional to total retardation $\delta_{\max}$. Considering the total mirror displacement L from zero retardation, the resolution will be $\Delta\nu = \frac{c}{2L}$. As an example, for L = 20 μm, we obtain $\Delta\nu$ = 7.5 THz.

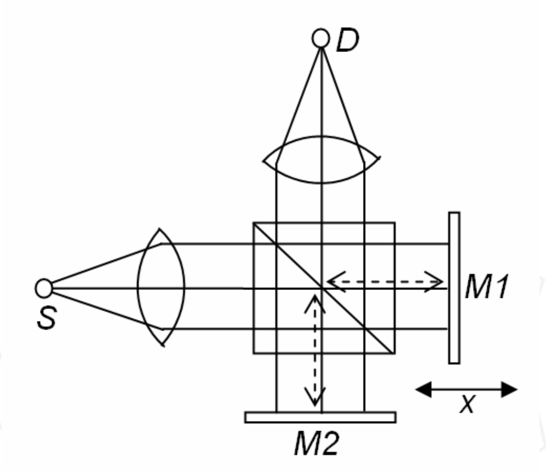


Fig. 3. Set-up of the Michelson spectrometer. S is the light source, D is the detector, M1 and M2 are two plane mirrors (or corner cube retro-reflectors). M1 moves back and forth unbalancing the optical path difference (*OPD*) of the interferometer

A laser at 633 nm is normally incorporated in Michelson spectrometers to calibrate the mirror displacement and to acquire the retardation at equal intervals of retardation. There are two ways to perform the interferogram acquisition synchronized with the mirror displacement, one is by moving the mirror at constant velocity in the retardation, this means that for a mirror velocity of 1 cm/s the acquired signal is at tens of kHz. The other method is by equispaced steps, for each single step one point of the interferogram is acquired.

As will be seen in Fig 6 in the next section, DCT generates a spectrum formed by the fundamental spectrum plus its mirror image, only the first N/2 points of equation (7) are useful, the second set of N/2 are redundant and discarded. A spectrum is meaningful if there is no overlap between the fundamental spectrum and the symmetrical replicas, therefore if the fundamental spectrum is completely contained in the first N/2 bins and is zero in the remaining N/2 bins. This is called Nyquist criterion, in order to sample a signal the sampling device should include a low pass filter that cuts the frequencies higher than half the sampling rate.

DCT in (7) is based on symmetrical interferograms around the zero retardation. When optics is dispersive and/or conditioning electronics have a phase dependence on frequency the interferogram becomes asymmetrical and DCT cannot be used directly. There are some techniques (Griffiths & de Haseth, 2007) to calculate the phase correction from the complex DFT.

The instrumental lineshape ILS function represents the resolution of the spectrometer and corresponds to the spectrum measured by the spectrometer when the radiation is monochromatic. ILS is the filter shape of each frequency bin. When the interferogram acquisition is abruptly truncated at the extremes, the rectangular or boxcar cutoff creates an ILS having the shape of a “sinc” function centered on the frequency bin, having a narrow peak but with important sidelobes that would hide possible neighbor lines. There is a palette of ILS function available to trade between the resolution (related to the width of the peak) and to the amplitude accuracy (related to the amplitude of the tails). ILS function can varied by multiplying the acquired interferogram by an appropriate tapering function (Smith, 1999), (Weisstein).

As was described before, the frequency bin interval is inversely proportional to the total retardation. If a monochromatic component has a period that is not exactly a submultiple of the total retardation, the frequency falls between adjacent bins and spectral components would be spread in several adjacent bins. Adding a series of zeros at the end of the interferogram has the important effect that new bins are created and the spectrum is interpolated in correspondence of the new bins and more frequencies could be represented without being spread.

Apparently, as stated in eq (8) the resolution *Δν* is only limited by the maximum retardation of the interferometer and therefore by increasing the retardation there is no physical limit to the attainable resolution. In practice the detector has a finite dimension and the considered rays pass through the interferometer with a divergence half-angle *a*. For a certain frequency *νo* and divergence *a*, it is always possible to find a retardation $\delta_{\max}$ such that the rays interfere destructively on the detector. This retardation $\delta_{\max}$ inserted in eq. (8) limits the minimum attainable resolution and there is no advantage to use a retardation longer than $\delta_{\max}$. The formula that relates the divergence half-angle *a* to the resolution *Δν* is the following:

$$\Delta\nu = \alpha^2 \nu_o \qquad (9)$$

3. F-P spectrometer

After having considered how it is possible to calculate a spectrum with the Michelson spectrometer by using the mathematical manipulations based on DCT, we now describe F-P spectrometers, core of the HSIS we have developed, and the associated algorithm to calculate the spectrum (Pisani & Zucco, 2009).

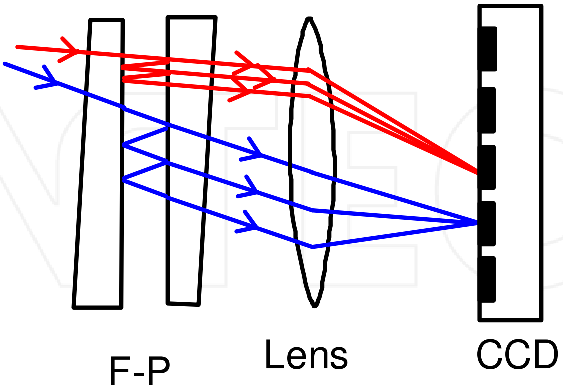


Fig. 4. Scheme of the F-P interferometer

The F-P spectrometer set-up is presented in Fig 4, and it is formed by two semireflective mirrors having reflectivity *R* at a distance *d*. For simplicity we consider mirrors with metallic coating with the peculiarity of having a negligible dispersion in the visible and in the IR region or equivalently a constant penetration depth versus the frequency. The incoming beam is reflected many times by the reflective surfaces and the different refracted beams are finally combined on the detector. The resulting interferogram for a monochromatic source at frequency ν_o is the Airy function

$$\bar{I}(\delta)=S(\nu)\frac{1}{1+\left(\frac{4R}{(1-R)^2}\right)\sin^2(2\pi\nu d/c)} \qquad (10)$$

Comparing the F-P interferogram (10) with the Michelson interferogram in (4) for a monochromatic source, it is evident that the interferogram is formed by fringes having the same periodicity, but with fringes more pronounced at the increasing of *R*, as in Fig 5.

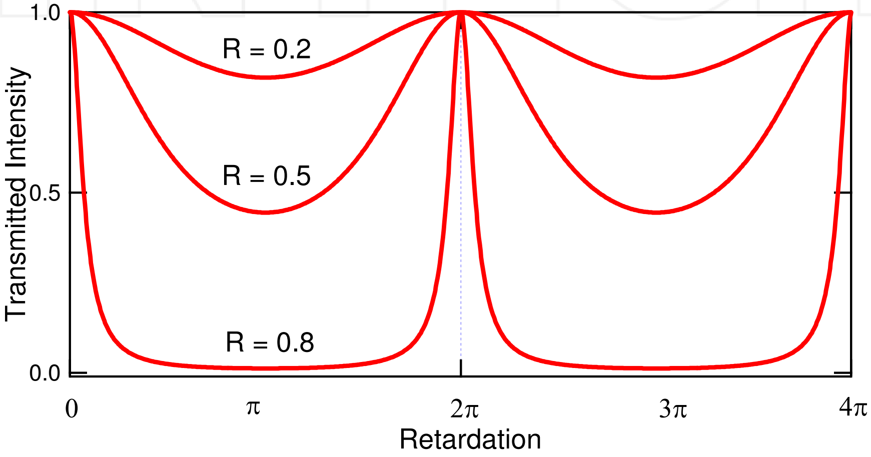


Fig. 5. Interferogram from the F-P spectrometer for different *R*

When the source is broadband and continuum, the resulting interferogram can be obtained by the integration of the different monochromatic contributions giving:

$$I(\delta)=\int_{-\infty}^{\infty}S(\nu)\frac{1}{1+\left(\frac{4R}{(1-R)^2}\right)\sin^2(2\pi\nu d/c)}d\nu \qquad (11)$$

In the approximation that the reflectivity *R* of the mirrors is very low *R*<<1, the Airy fringes in (10) could be approximated with cosine function and eq (10) becomes

$$\bar{I}(\delta)=S(\nu)\frac{1}{1+\left(\frac{4R}{(1-R)^2}\right)\sin^2(2\pi\nu d/c)}\approx S(\nu)(1-2R)+S(\nu)2R\cos(\pi\nu d/c) \qquad (12)$$

Taking into account only the modulated part of the interferogram (12), and including all the responsivity contributions in intensity of the radiation at the detector in $S(\nu_0)$, we obtain a equation similar to (3) and therefore the interferogram from the F-P could be solved using the DCT as in (6), provided that Nyquist criterion is respected, i.e. at least two points per fringe are acquired.

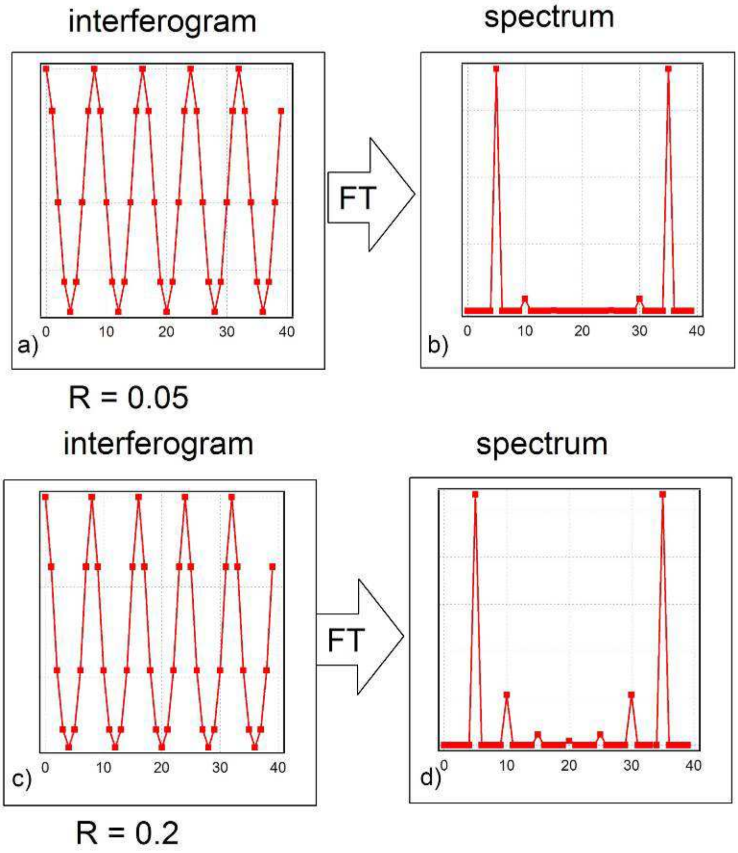


Fig. 6. Spectra obtained with DCT from F-P interferogram with different reflectivity. a) and b) interferogram and spectrum for reflectivity *R* = 0.05. c) and d) interferogram and spectrum for reflectivity *R* = 0.2

When the approximation presented in (12) is not possible because *R* is close to 1, we could still apply DCT to the Airy fringes, since the interferogram is periodic and can be decomposed by the Fourier series into a sum of cosines. Since the Airy fringes are far from being a cosine, also the higher harmonic components are presents. As an example in Fig 6. we have represented a F-P interferogram of a monochromatic source when *R* = 0.05, cfr. Fig 6(a), the associated DCT is presented in Fig 6(b) and the second harmonic is about 1/20th of the fundamental whereas the third harmonic is negligible. In this example, the fringe is well oversampled, 8 points per fringe are acquired and Nyquist criterion is respected since it is possible to see that the harmonics fade away in the higher part of the spectrum, and the aliases do not "superpose" with the fundamental spectrum. On the contrary, in a situation similar to the practical case that we will describe in the next section, when R = 0.2 the Airy fringes are more pronounced and by applying the DCT we see in Fig 6(d) that the 3rd and 4th harmonics are present and the 7th harmonic of the alias would sum up with the fundamental. Since the 6th harmonic is about $3 \cdot 10^{-5}$ of the fundamental, we state that is effect is negligible and by acquiring 8 points per fringe for R = 0.2 is enough to respect Nyquist criterion.

If now we consider a source with a broad spectrum larger than an octave in the electromagnetic spectrum, the second harmonic of the low frequency side would sum up with the fundamental spectrum in the high frequency side as is depicted in Fig 7(a). In this figure the fundamental spectrum is represented in red and is larger than an octave, the second harmonic in blue is superposed with the high frequency side of the spectrum. The sum of all the harmonics would give the resulting spectrum in black, and from its analysis is nearly impossible to obtain information about the fundamental spectrum in red. When the fundamental spectrum is smaller than an octave, as in Fig 7(b), from the resulting spectrum it is possible to discriminate from the fundamental and the second harmonic spectra.

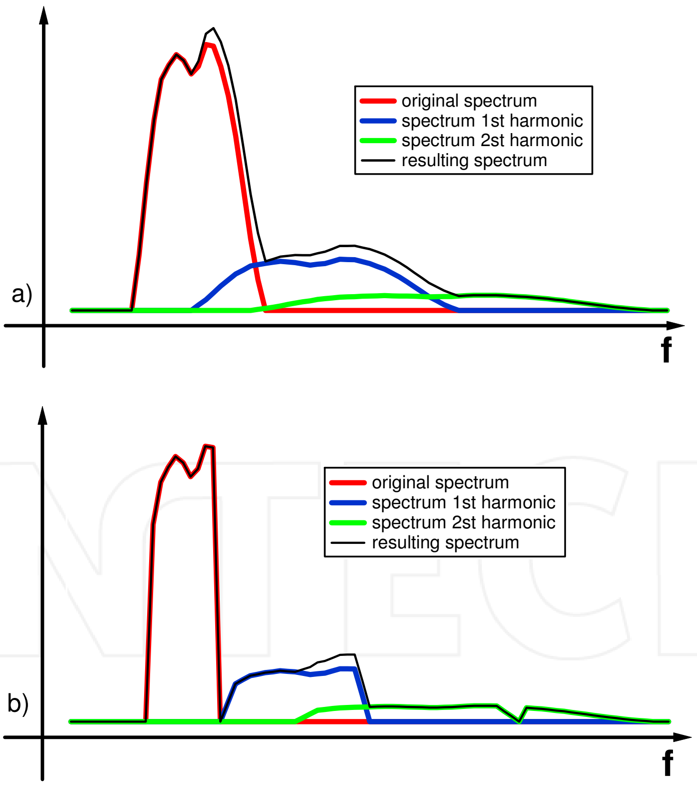


Fig. 7. Effect of superposition between harmonics of the fundamental spectrum. a) the spectrum is larger than an octave and it is not possible to discriminate from the fundamental and second harmonic b) the spectrum is smaller than an octave

While Michelson spectrometers have a double sided interferogram and this fact is useful in order to obtain information about the phase correction, the F-P interferogram is evidently single-sided and it does not start from the central or zero fringe when the mirrors come in contact because of the penetration depth of the radiation in the metallic coating. The latter implies that the interferogram is incomplete (i.e. it does not contain the data corresponding to the zero retardation condition), therefore it is not possible to apply directly the DCT to the interferogram as in two beam interferometers. As an example, in Fig 8 is presented the measured interferogram obtained from a F-P having a metallic coating of a laser radiation at λ = 410 nm. It is possible to see that the first half of the fringe is missing, corresponding to a retardation of 205 nm or equivalently to a mirror distance of 102.5 nm, we can estimate that the penetration depth of the metallic layer is smaller than 50 nm.

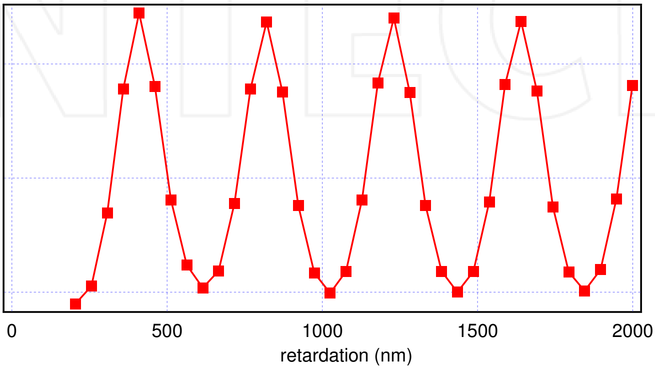


Fig. 8. The measured interferogram of a laser radiation at λ = 410 nm. The estimated penetration depth of the metallic layer is smaller than 50 nm

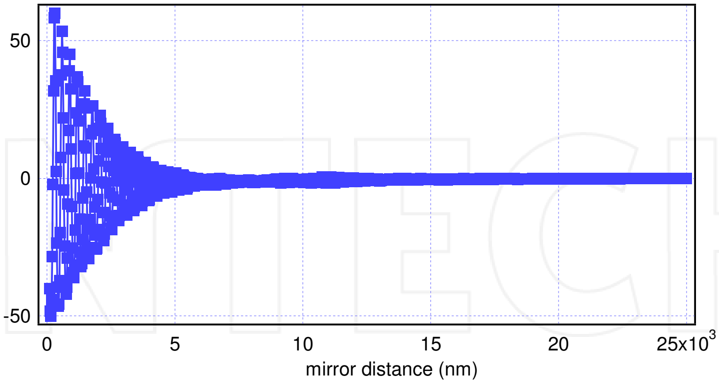


Fig. 9(a). The interferogram of a yellow LED with a maximum mirror distance δ = 25 µm

The two aforementioned problems (presence of harmonics and missing points) are solved in our prototype by introducing an optical bandpass filter in the optical system transmitting slightly less than one octave of the electromagnetic spectrum, as the spectrum reported in Fig 7(b). This solution reduces the region of measurable spectrum but has the important consequence that by using the information that certain region of the spectrum have zero intensity it is possible to estimate the value of the missing points in the interferogram and reconstruct the spectrum. As an example, consider the interferogram in Fig 9 of a yellow LED with a maximum mirror scan distance of δ = 25 µm corresponding to a resolution of about 6 THz, according to equation (8). To calibrate the mirror distance a blue laser at 410 nm is used and the points are acquired at interval of 25.5 nm in mirror distance. In Fig 9(a) it is presented the plot of the acquired interferogram. Due to the penetration depth, the first four points are missing. In Fig 9(b) the zoom of Fig 9(a) near the contact zone, zero retardation.

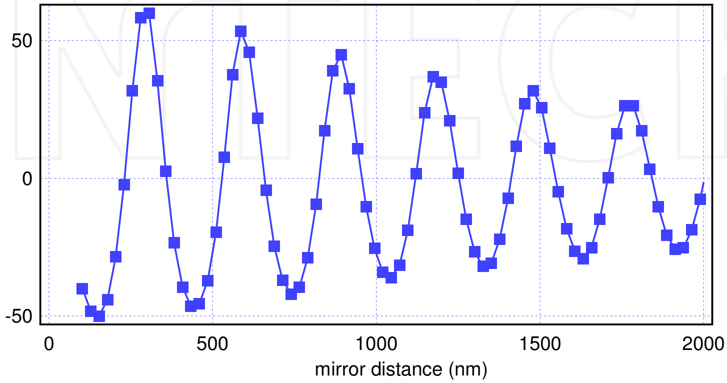


Fig. 9(b). The zoom of the interferogram in Fig 9(a) with the 4 missing points

Since the interferogram of Fig 9 is incomplete, four points are missing, it cannot be elaborated by the DCT in eq (7). In order to apply DCT the interferogram is completed with four points set at zero, not having any information *a priori*, and the result is presented in Fig 10.

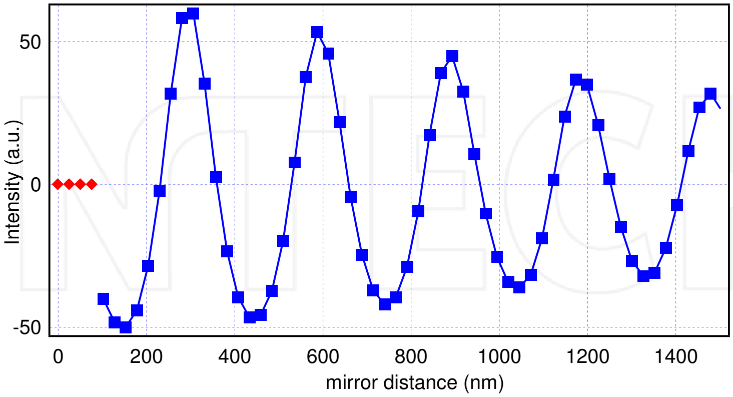


Fig. 10. The interferogram in Fig 9 with the 4 missing points (red diamonds) set at zero

The DCT of the interferogram gives the spectrum presented in Fig 11 where it is visible the spectrum of the yellow LED at about 500 THz and the second harmonic at about 1000 THz.

The spectra are sitting on a curved background due to the 4 missing points. In fact, according to the DCT equation where each point in the interferogram corresponds to a cosine contribution in the spectrum: the first missing point corresponds to the DC level in the spectrum, the second missing point corresponds to the cosine component having a period equal to the full spectrum span, and so on.

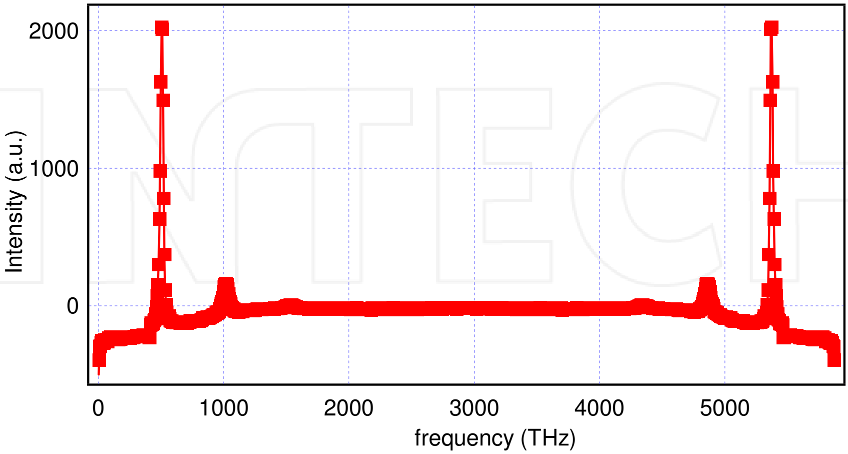


Fig. 11. The "biased" spectrum obtained by applying the DCT to the incomplete interferogram in Fig 10

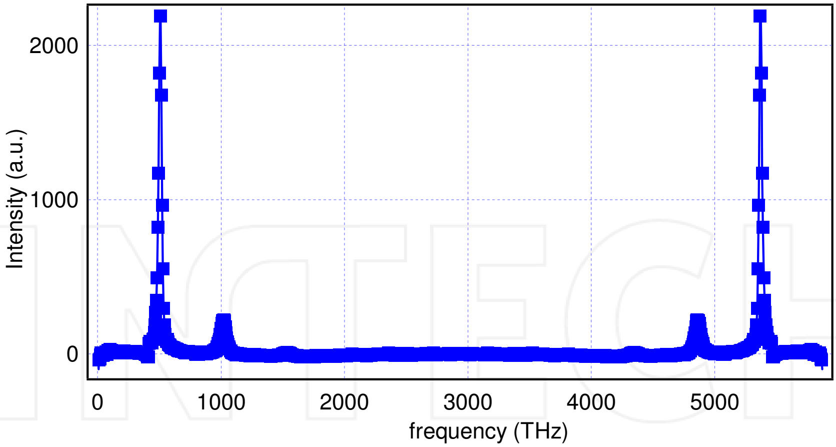


Fig. 12. The spectrum versus frequency of the yellow LED obtained by reconstructing the four missing points

By using the information that the spectrum has to be zero in certain region (the optical filters stops optical frequencies lower than 380 THz and higher than 720 THz), it is possible to find the amplitude of the four missing cosines and the value of the four missing points. In Fig 12 the reconstructed spectrum of the yellow LED with the fundamental spectrum at about 500

THz, the second harmonic at about 1000 THz and the third harmonic is nearly visible. In Fig 13 the spectrum in the pass band region of the optical filter represented in wavelength. In Fig 14 the interferogram with the reconstructed four missing points.

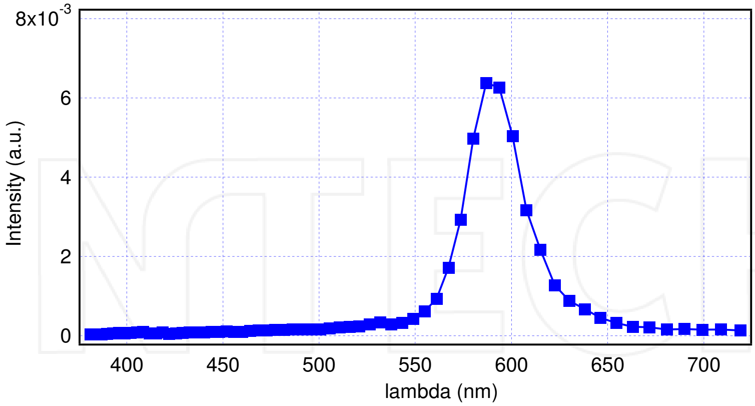


Fig. 13. The spectrum versus wavelength of the yellow LED in Fig 12

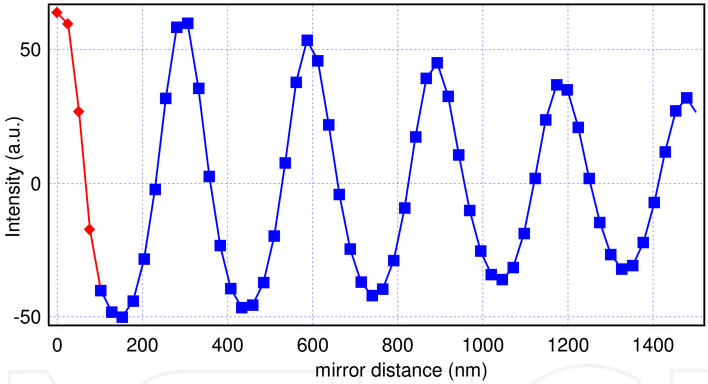


Fig. 14. Corrected interferogram to obtain the spectrum in Fig 12 and Fig 13. Blue squares: the original incomplete interferogram, red diamonds the reconstructed four points

In this section we have described the mathematical manipulations to calculate the spectrum from the incomplete F-P interferogram, in the next section we will explain how the F-P spectrometer could be integrated in an imaging system and obtain a HSIS.

4. Hyperspectral imaging prototype

In this section we describe how the F-P spectrometer is integrated in the HSIS, how the acquisition system is done and some applications in spectroscopy, colorimetry and thermal imaging.

The core of the device is the scanning F-P spectrometer described in section 3 and whose rendering is presented in Fig 15. The two mirrors are coated with a thin aluminum layer and have a reflectivity of about 20%. The dispersion of the metallic coating has been demonstrated to be negligible for our applications in the spectral interval (0.4 – 1.7 um). The mirrors are mounted in aluminum frames and the distance is scanned by means of three piezo actuators allowing a maximum displacement of 60 μm at 100 V. A system made by three elastic hinges and three screws allows the optimal alignment and working distance of the mirrors to be found, so that, when the actuators are completely retracted (maximum voltage applied), the mirrors are in contact and the contact area is sufficiently large and centered.

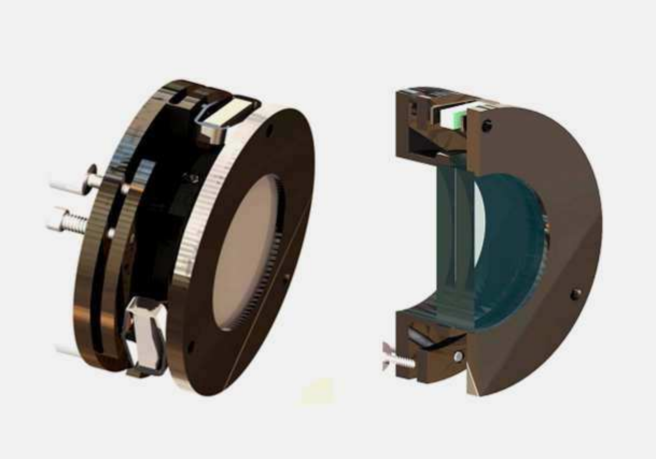


Fig. 15. Rendering of the F-P device and its section. The piezo actuator and the trimming screws are visible

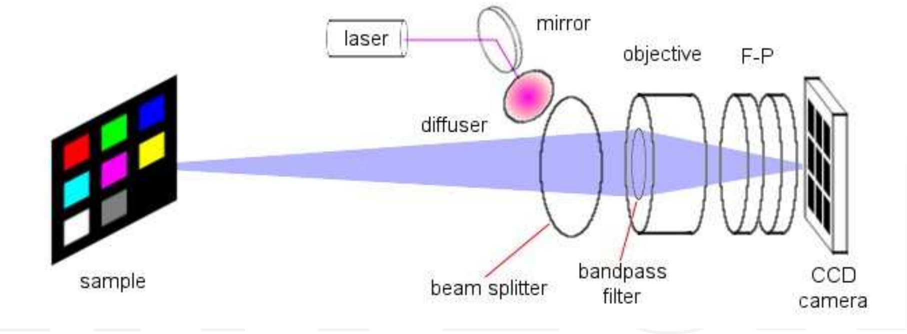


Fig. 16. Optical set-up formed by a beam splitter to couple the laser radiation to calibrate the retardation, the bandpass filter, the photographic objective and the CCD camera

The optical part of the HSIS is made of a photographic objective coupled with a CCD camera. The F-P is placed as close as possible to the camera sensor. Between the objective and the F-P is placed the optical band pass filter needed to select the wanted portion of the spectrum as explained before in order to apply the algorithm to find the missing points. The layout of the experiment used to obtain the data described in next section is schematized in

Fig 16. Other setups used in this work are variations of this one. The laser radiation to calibrate the mirror displacement is sent either directly on the sample or directly to the CCD through the F-P by inserting a beam splitter on the optical axis. We have used the HSIS in two different regions of the spectrum, in the visible by means of an optical band pass filter (380 – 720 nm), a reference blue laser at 410 nm and a 12-bit Si–CCD camera, in the NIR using a calibration laser at 980 nm and a 14-bit InGaAs CCD camera that has a response in the optical region 900 – 1700 nm.

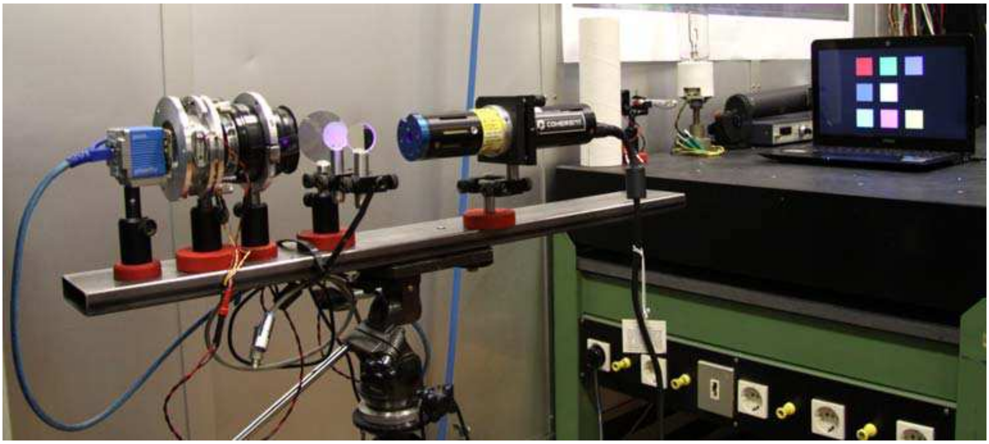


Fig. 17. Picture of the experiment where it has been acquired the spectrum of a laptop monitor

The HSIS is controlled by two boards on a PC, one used to drive the F-P and the other to drive the camera. The first, equipped with a 16-bit digital-to-analog converter, generates a triangular voltage ramp sent to a HV amplifier which generates the 0-100 V signal to drive the three piezo actuators (connected in parallel). Through the same board a trigger signal is generated synchronously with the ramp. The second board acquires a video starting from the trigger signal (corresponding to the maximum mirror distance condition) and ending with complete contact. In order to have a sufficient sampling rate to respect Nyquist criterion, about 1000 frames each acquisition are required for a 20 μm scan. This figure, in combination with the maximum frame rate of the camera sets the maximum ramp speed. The video is saved in TIFF format.

5. Hyperspectral imaging applications

This section shows some application of our HSIS prototype. We start with a calibrated reflective target to test the accuracy of the system, on a selectively absorptive target to test the potentialities as a spectroscopic analytic instrument, with laser sources to test the spectral resolution and on a heated tungsten plate to test thermal imaging.

5.1 Reflective target

The first application of the HSIS is a ColorChecker®. The image covers a rectangle of about 670x460 pixels on the CCD area. Each colour tab is illuminated by a laser spot in order to calibrate the retardation for each frame in the video. In Fig 18 a video frame is presented.

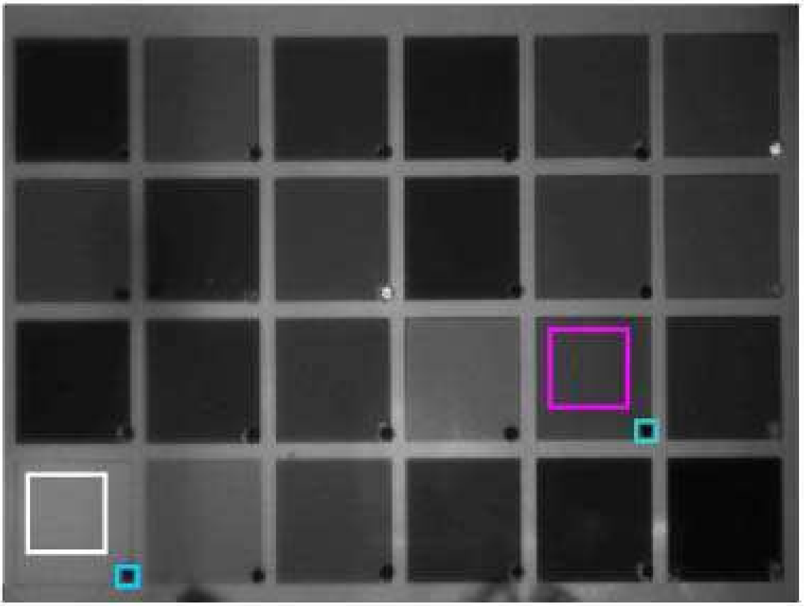


Fig. 18. Video frame of the colour checker target. In magenta and white are indicated the areas used for the calculation of the spectrum of the magenta and white tabs. In blue are indicated the pixels used to calibrate the retardation for both tabs

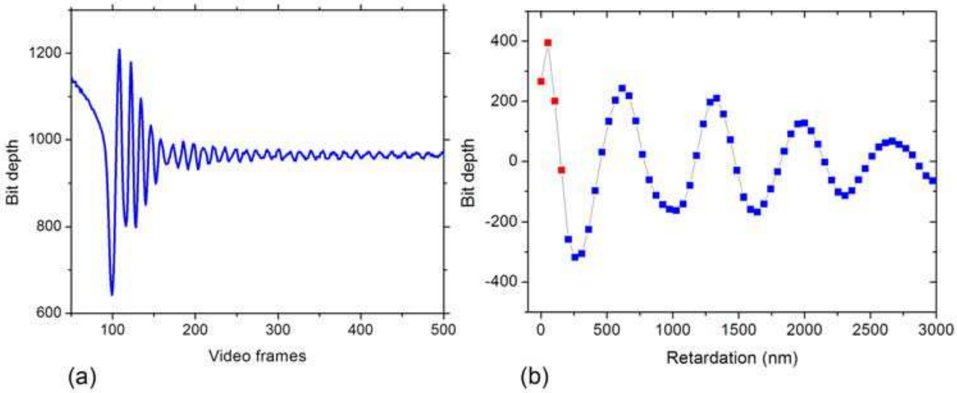


Fig. 19(a). the interferogram of the magenta set of pixels in Fig 18; Fig 19(b) the same interferogram after having subtracted the mean and calibrated the retardation using the calibration of the blue laser. The blue squares are the recorded data, the red squares are the values found with the reconstruction algorithm

The magenta square in Fig 18 contains the pixels used for the calculation of the spectrum of the magenta coloured tab. The blue square on the bottom right hand side contains the pixels illuminated by the blue laser used to calibrate the retardation for the magenta. The same applies for the white tab (bottom left) used for normalization purposes. Fig 19(a) reports the interferogram of the pixel set in the magenta square in Fig 18 with the x-axis represented in video frames. In Fig 19(b) the blue squares are the first part of the same interferogram after the re-sampling using the reference scale from Fig 6(b) (as explained in section 4). The retardation sampling interval is 51.25 nm and the first record corresponds to a retardation of 205 nm (half wavelength of the blue laser).

The spectra obtained with the reconstructing algorithm are presented in Fig 20(b) (in black and magenta respectively the spectra of the white and magenta tabs). The resolution in frequency is about 14 THz corresponding to a resolution in wavelength of about 12 nm at a wavelength of 500 nm, (by applying the zero padding method the number of points is artificially increased in order to have one point each nanometer for practical computational reasons). The absolute reflectivity spectrum is given by the ratio of the coloured tab spectrum and the white tab spectrum used as a reference. In this way the effect of the non uniform spectral responses of the optical system, of the camera and of the light source is cancelled. In Fig 20(a) the reflectivity spectrum is shown and compared with the same spectrum measured with a commercial spectrometer (thin black line); relatively large differences are evident at the extremes of the spectrum mainly due to the reduced intensity of the reference spectrum in Fig 20(b).

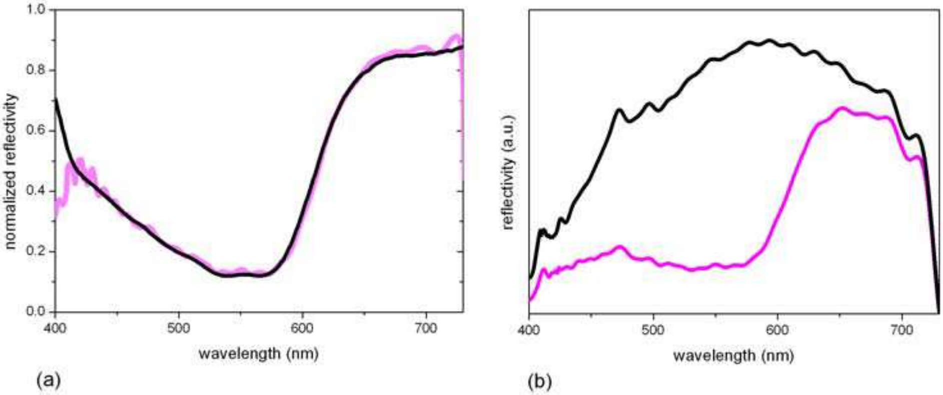


Fig. 20. (a) Magenta spectrum normalized with respect to the white to obtain the absolute reflectivity spectrum, the black trace is the reflectivity spectrum obtained with a spectrometer. (b) In black and magenta respectively the spectra of the white and magenta tabs

Similarly we can obtain a hyperspectral image from an emitting surface. Fig 21 shows the emission spectra from a target made of LED and lamp sources. Even in this case a normalization that allows to take in account the responsivity of the whole system is necessary. The normalization in this case has been done comparing the spectrum obtained from a broadband source (a tungsten lamp or a white LED) with the same spectrum measured with a calibrated spectrophotometer. This normalization function is a constant of the system.

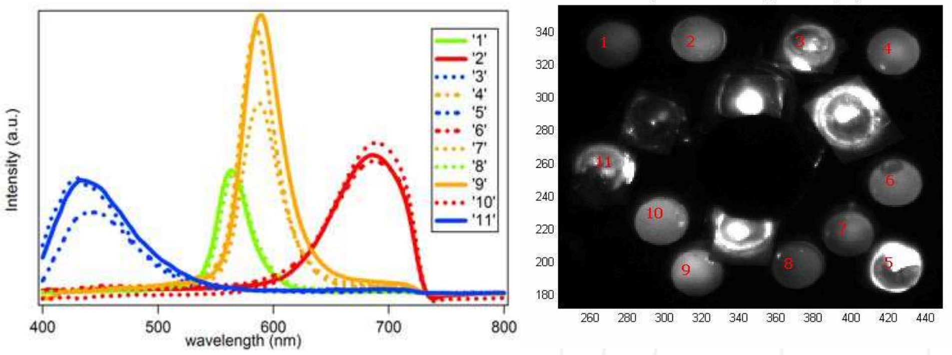


Fig. 21. The emission spectra on the left correspond to the pixel area (10×10) indicated in the image on the right. The target is composed of LED of four different colors (blue, green, yellow and red)

5.2 Spectroscopy applications

As a second application we have measured the transmission of a didymium oxide optical filter, and the results are presented in Fig 22. A white screen placed in front of the camera is illuminated with the xenon discharge lamp. A portion of the field of view of the camera is covered with the filter so that the light reflected by the screen passes through it before entering the objective. Another portion is illuminated with the blue laser again used for the retardation calibration. The transmittance spectrum is obtained by the ratio between the spectrum of the filtered portion and the spectrum of the white portion of the same image. Again the result is compared with the measurement done with a spectrometer (black line). This was done to demonstrate the potentialities of the instrument to detect complex absorption spectra, thus to be used in spectroscopy based chemical analysis.

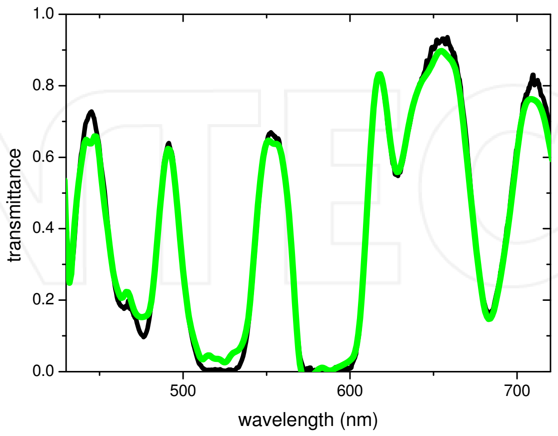


Fig. 22. The complex transmission spectrum of a glass doped with didymium oxide. The black line is the same spectrum measured with a classical spectrometer

5.3 Thermal imaging

The device can be used in different regions of the electromagnetic spectrum provided the transmissivity of the mirrors glass and the reflectivity of the metallic layer is adequate. We have implemented the system with a InGaAs camera capable of detecting infrared radiation in the 900-1700 nm band. One interesting application in this region is thermal imaging. By exploiting the change in shape of the blackbody radiation curve with temperature it is possible to infer the temperature of the emitting body. Fig 23 shows thermal imaging of a heated tungsten plate. With respect to classical thermal cameras which infer the temperature of a body by measuring the amount of radiation emitted in a given band, with an hyperspectral system the temperature is inferred by fitting the blackbody curve. The difference is that in the latter we do not need an *a priori* knowledge of the emissivity of the body.

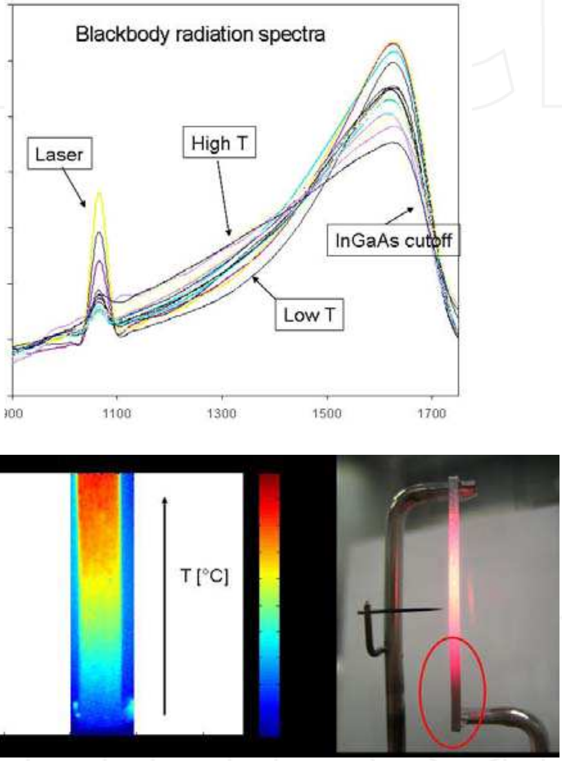


Fig. 23. Up: spectra obtained at different temperatures. The curves represent the blackbody emission spectra truncated by the responsivity of the camera for wavelengths larger than 1700 nm

5.4 Spectral resolution

In order to test the spectral resolution of the device a white target has been illuminated with lasers at five different wavelengths: a blue diode laser at 410 nm (used as a reference), a green duplicated Nd:YAG laser at 532.4 nm, a red He-Ne laser at 633 nm and two red diode lasers (637.5 and 674 nm). The scan applied to the mirrors is 50 µm corresponding to about

240 entire fringes in the blue interferogram. Fig 24 shows the obtained spectrum using Welch windowing. The experimental FWHM is about 3 THz corresponding to the theoretical resolution in eq. (8). The maximum difference of the measured wavelengths with respect to the nominal values is 1 nm. The resolution can be appreciated in the pair 633 and 638 nm whose peaks are well distinguishable.

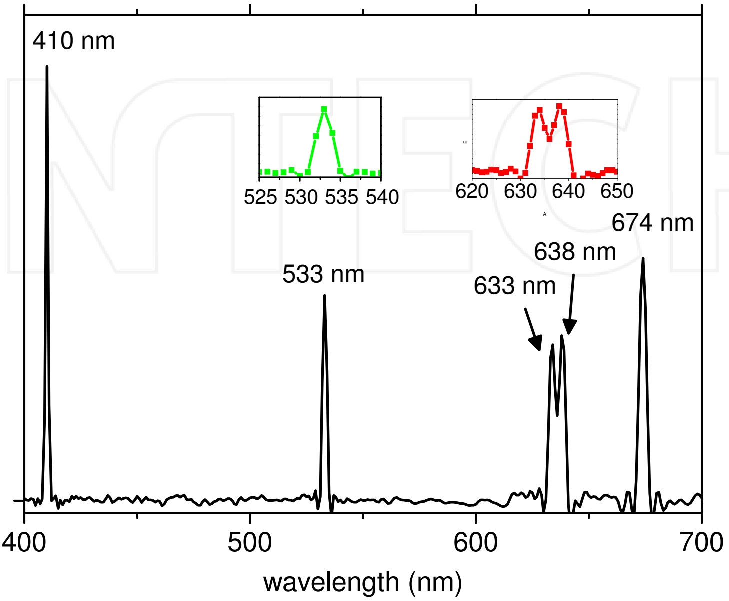


Fig. 24. Spectrum of a target illuminated by five laser beams having wavelength 410 nm (used as a reference), 532.4 nm, 633 nm, 637.5 nm and 674 nm. In the box a zoom of the green line showing the resolution

6. Conclusion

Hyperspectral imaging is a technique that consists in associating to each pixel of an image the spectral composition of the light hitting the same. The simplest example of Hyperspectral imaging is the RGB standard where color content of digital images is represented by giving the amount of red, green and blue. In a similar way we have three different cone cells in our retina, sensible to the three colors. By increasing the number of components of the spectral content would permit to have more information about the radiation emitted or reflect by an object. There are many applications to this technique in different fields, to mention but a few: fluorescence microscopy, thermal imaging, chemistry (infrared spectroscopic analysis), space missions (Earth survey for environmental or security), colorimetry. In this work we have realized a hyperspectral imaging device based on a Fabry-Perot interferometer. We have introduced the algorithm based on the Fourier transform to obtain the spectral content of each pixel from the measured interferogram. Moreover we have discussed the accuracy and resolution that characterize this system and some applications.

7. Acknowledgements

We would like to thank Maria Luisa Rastello for helpful discussions, Marco Santiano for the design and realization of the F-P cavity and Paolo Bianco that has contributed to the most recent results.

8. References

Alcock, R. D. & Coupland, J. M. (2006). A compact, high numerical aperture imaging Fourier transform spectrometer and its application. *Meas. Sci. Technol.,* Vol. 17, No. 11, (November 2006) page numbers (2861-2868), ISSN 0957-0233

Griffiths P. R. & de Haseth, J. A. (2007). *Fourier Transform Infrared Spectroscopy,* Wiley Interscience, ISBN 978-0-471-19404-0, Hoboken (U.S.A)

Pisani, M. & Zucco, M. (2009). Compact imaging spectrometer combining Fourier transform spectroscopy with a Fabry-Perot interferometer. *Optics Express,* Vol. 17, No. 7, (May 2009) page numbers (8319-8331), ISSN 1094-4087

Sellar, R. G. & Boreman, G. D. (2005). Classification of imaging spectrometers for remote sensing applications. *Opt Eng.,* Vol. 44, No. 1, (December 2004) page numbers (2861-2868), ISSN 0091-3286

Smith, S. W. (1999). *The Scientist and Engineer's Guide to Digital Signal Processing,* California Technical Publishing, ISBN 0-9660176-4-1, San Diego (U.S.A)

Weisstein, E. W. Apodization Function. From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/ApodizationFunction.html

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

This book aims to provide information about Fourier transform to those needing to use infrared spectroscopy, by explaining the fundamental aspects of the Fourier transform, and techniques for analyzing infrared data obtained for a wide number of materials. It summarizes the theory, instrumentation, methodology, techniques and application of FTIR spectroscopy, and improves the performance and quality of FTIR spectrophotometers.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

22

Application of Fast Fourier Transform for Accuracy Evaluation of Thermal-Hydraulic Code Calculations

Andrej Prošek and Matjaž Leskovar
Jožef Stefan Institute
Slovenia

1. Introduction

To study the behaviour of nuclear power plants, sophisticated and complex computer codes are needed. Before the computer codes are used for safety evaluations they have first to be validated. The assessment process of system codes involves the comparison of code results against experimental data and measured plant data. The accuracy of the code is the capability of the code to correctly predict the physical behaviour. Therefore the evaluation of accuracy coincides with code validation. In the past a few methods to quantify the code accuracy of thermal-hydraulic system codes have been proposed. Among the proposed methods, the approach using the fast Fourier transform has been proposed as one of the most effective approaches in 1990s (Ambrosini et al., 1990; D'Auria et al., 1994). The fast Fourier transform based method (FFTBM) shows the measurement-prediction discrepancies, i.e. the accuracy quantification, in the frequency domain. From the amplitudes of the component frequencies, the average amplitude (AA) is calculated. AA sums the difference between experimental and calculated signal discrete Fourier transform amplitudes at each frequency. To get a dimensionless accuracy measure, the sum of the amplitudes of the experimental signal is used for normalization. The closer the non-dimensional AA value is to zero, the better the agreement between the calculated results and the experimental measurements is judged. However, some problems involved in FFTBM, such as proper selection of time windows, weighting factors, number of discrete data used, consistency of the method in all cases and time dependent accuracy, still remain open and partly limit its application, especially those requiring a consistent accuracy judgement. For example, in early applications of FFTBM problems in evaluating signals, where experimental or error signal has the shape similar to triangle (i.e. first increases and then decreases), the accuracy value regularly overshoots at triangle peak, stabilising at lower value when discrepancy decease (Mavko et al. 1997). Not being aware about the reasons of such or some other strange behaviour, the FFTBM method has been also criticized. In general it is required that at any time into the transient the accuracy measure should remember the previous history. But on the other hand, the original FFTBM method has been effectively applied in obtaining information on code accuracy by several researchers in the literature.

More recently, an automated code assessment program (ACAP) has been developed to provide a quantitative comparison between nuclear reactor system code results and experimental measurements (Kunz et al., 2002). For the time record data the original FFTBM accuracy measure was modified and a new continuous wavelet transformation accuracy measure was included among several other accuracy measures developed for timing of events tables, scatter plots and steady state data. Unfortunately, the ACAP tool was developed for single variable comparison only. Besides, not many measures were effective in evaluating time record data. This means that the original FFTBM remained in use. In 2002, the review of important applications was done (Prošek et al., 2002). Much of the work was performed mostly in the application domain. The comparisons between the experimental data and calculated results were done for different transients and accidents on different experimental facilities. The first large FFTBM application was to the international standard problem no. 27 (ISP-27) in which primary system thermal-hydraulic system codes were used (D'Auria et al., 1994) to the BETHSY facility simulating the French pressurized water reactor. The maturity was shown in that the method was sensible in highlighting the differences between pre- and post-test calculations for the same user, normally originating by an ad-hoc code tuning operated in post-test analyses and by the code use at the international level. The first application of FFTBM to containment code calculations was to ISP-35 performed on the NUPEC facility (D'Auria et al., 1995). The need for potential further efforts to refine the weighting factors was expressed. The application to ISP-39 performed on the FARO facility (D'Auria & Galassi, 1997) was the first application of FFTBM to severe accidents. The application confirmed the capabilities of the FFTBM method only in ranking generic calculation results. The application to ISP-42 performed on the PANDA facility (Aksan et al., 2001) showed that ten variables were not enough to completely characterize the transient. Finally, the application of FFTBM to the ISP-13 exercise, post-test calculations of the LOFT L2-5 test was performed in the frame of the BEMUSE program (OECD/NEA, 2006). The original FFTBM approach was used in this application. The numerous applications showed that there are some deficiencies of the original FFTBM, which were resolved in the proposed improved FFTBM.

In this Chapter, we first describe the original FFTBM approach. Then the time dependent accuracy measures are introduced. By calculating the time dependent accuracy it can be answered, which discrepancy contributes and how much is its contribution to the inaccuracy. Then, the index for time shift indication is proposed. The application of the time dependent accuracy showed that the original FFTBM gave an unrealistic judgment of the accuracy for monotonically increasing or decreasing functions, causing problems in FFTBM results interpretation. This problem was hidden in the past when FFTBM was applied only to a few time windows and/or intervals. It was found out that the reason for such an unrealistic calculated accuracy of increasing/decreasing signals is the edge between the first and last data point of the investigated signal, when the signal is periodically extended. Namely, if the values of the first and last data point of the investigated signal differ, then there are discontinuities present in the periodically extended signal seen by the discrete Fourier transform, which views the finite domain signal as an infinite periodic signal. The discontinuities give several harmonic components in the frequency domain, thus increasing the sum of the amplitudes, on which FFTBM is based, and by this influencing the accuracy. The influence of the edge due to the periodically extended signal is for clarity reasons called edge effect. It should be noted that the signal may include several other rising and falling edges, which influences are not considered as edge effect in this chapter. The quantitative contribution of the edge effect on the accuracy may be very unpredictable and can overshadow the contribution of the discrepancies of the compared functions on the accuracy. Therefore it is proposed how to resolve the problem of the edge effect on a unique way by signal mirroring.

In order to demonstrate its application, the proposed improved FFTBM by signal mirroring is tested to show that it gives a realistic and consistent judgment of the accuracy also for monotonically increasing or decreasing functions, and for all other signals influenced by the edge effect. The results obtained with FFTBM results were compared to results obtained with ACAP. At the end general recommendations for applying FFTBM are given.

2. Original FFTBM description

The methodology of the code-accuracy assessment consists of three steps: a) selection of the test case (experimental or plant measured data to compare), b) qualitative analysis, and c) quantitative analysis. The qualitative analysis is a prerequisite to perform the quantitative analysis. The qualitative analysis, including visual observation of plots, is done by evaluating and ranking the discrepancies between the measured and calculated variable trends. The quantitative analysis (applying FFTBM) is meaningless unless all the important phenomena are predicted.

The original FFTBM is a method (Ambrosini et al., 1990), which shows the measurement-prediction discrepancies in the frequency domain. The method purpose is to quantify the accuracy of code calculations based on the amplitudes of the discrete experimental and error signal calculated by the fast Fourier transform (FFT). On the other hand, the digital computers can only work with information that is discrete and finite in length and there is no version of the Fourier transform that uses finite length signals (Smith, 1999). The way around this is to make the finite data look like an infinite length signal. This is done by imagining that the signal has an infinite number of samples on the left and right of the actual points. The imagined samples can be a duplication of the actual data points. In this case, the signal looks discrete and periodic. This calls for the discrete Fourier transform (DFT) to be used. There are several ways to calculate DFT. One method is FFT. While it produces the same results as the other approaches, it is incredibly more efficient. The key point to understand the FFTBM is that the periodicity is invoked in order to be able to use a mathematical tool, i.e., the DFT. Therefore, the periodic nature of DFT is explained first before the accuracy measures used in FFTBM are described.

2.1 Periodic nature of discrete Fourier transform

The discrete Fourier transform views both, the time domain and the frequency domain, as periodic (Smith, 1999). The signals used for comparison are not periodic. Nevertheless, the user must conform to the DFT’s view of the world. Figure 1 shows two different interpretations of the time domain signal. In the upper part of Fig. 1 the time domain is viewed as N points. This represents how digital signals are typically acquired in experiments and code calculations. For instance, these 64 samples might have been acquired by sampling some parameters at regular intervals of time. Sample 0 is distinct and separate from sample 63 because they were acquired at different times. The samples on the left side are not related to the samples on the right.

As shown in the lower part of Fig. 1, the DFT views these 64 points to be a single period of an infinitely long periodic signal. This means that the left side of the signal is connected to the right side of a duplicate signal, and vice versa. The most serious consequence of time domain periodicity is the occurrence of the edge. When the signal spectrum is calculated with DFT, the edge is taken into account, despite the fact that the edge has no physical meaning for comparison, since it was introduced artificially by the applied numerical method. It is known that the edge produces a variegated spectrum of frequencies due to the discontinuity of the edge. These frequencies originating from the artificially introduced edge may overshadow the frequency spectrum of the investigated signal.

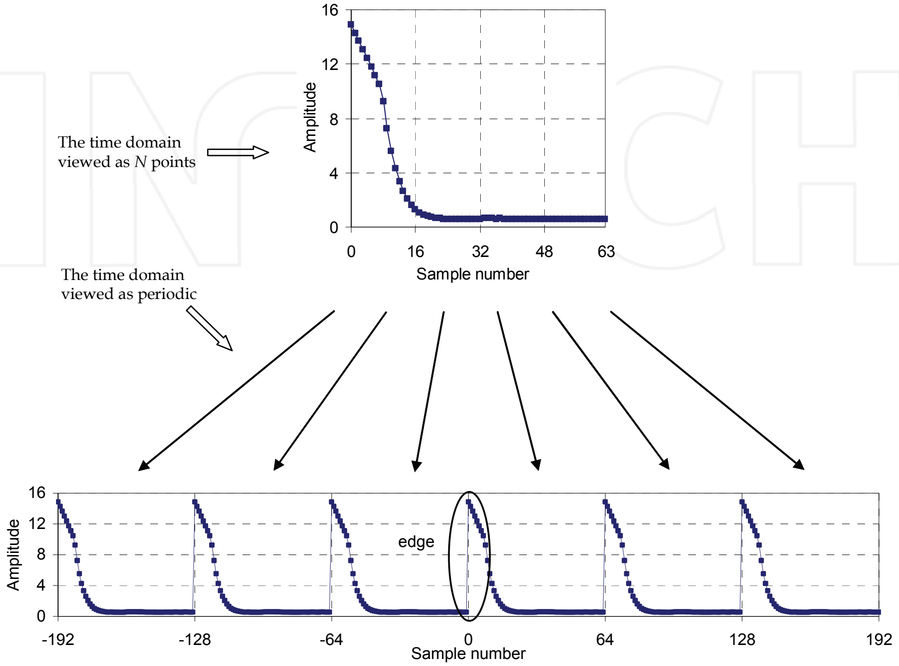


Fig. 1. Periodicity of the DFT's time domain original signal. The time domain can be viewed as N samples in length, shown in the upper part of the figure, or as an infinitely long periodic signal, shown in the lower part of the figure

2.2 Average amplitude

For the calculation of measurement-prediction discrepancies the experimental signal $F_{\exp}(t)$ and the error signal $\Delta F(t)$ are needed. The error signal in the time domain is defined as $\Delta F(t) = F_{cal}(t) - F_{\exp}(t)$ where $F_{cal}(t)$ is the calculated signal. The code accuracy quantification for an individual calculated variable is based on the amplitudes of the discrete experimental and error signal obtained by FFT at frequencies f_n, where $n = 0,1,...,2^m$ and m is the exponent defining the number of points $N = 2^{m+1}$. The average amplitude AA is defined:

$$\mathrm{AA} = \frac{\sum_{n=0}^{2^m} \left| \tilde{\Delta} F(f_n) \right|}{\sum_{n=0}^{2^m} \left| \tilde{F}_{\exp}(f_n) \right|}, \qquad (1)$$

where $\left|\tilde{\Delta}F(f_n)\right|$ is the error signal amplitude at frequency f_n and $\left|\tilde{F}_{\exp}(f_n)\right|$ is the experimental signal amplitude at frequency f_n. The AA factor can be considered a sort of average fractional error and the closer the AA value is to zero, the more accurate is the result. Typical values of AA are from 0 to 1.

2.3 Total average amplitude

The overall picture of the accuracy for a given code calculation is obtained by defining average performance index, that is the AA_{tot} (total average amplitude or total accuracy)

$$\mathrm{AA}_{\mathrm{tot}} = \sum_{i=1}^{N_{\mathrm{var}}} \left(\mathrm{AA}\right)_{\mathrm{i}} \cdot \left(w_{\mathrm{f}}\right)_{\mathrm{i}} \,, \tag{2}$$

with

$$\sum_{i=1}^{N_{\mathrm{var}}} \left(w_{\mathrm{f}}\right)_{\mathrm{i}} = 1 \,, \tag{3}$$

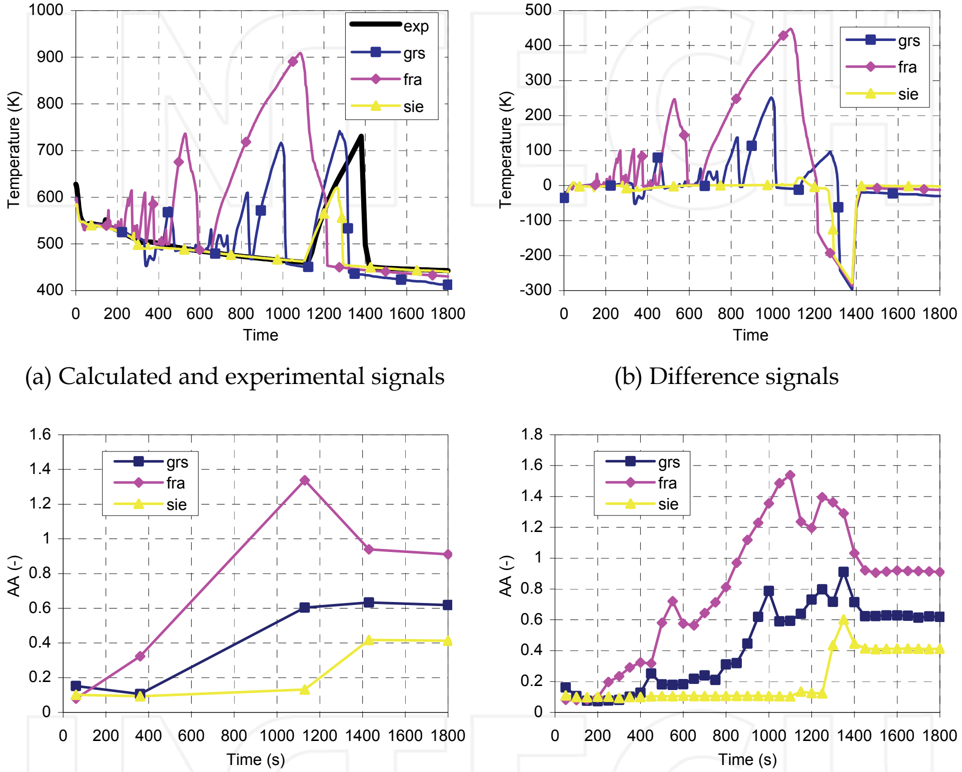
where N_{var} is the number of the variables analyzed (typically from 20 to 25), and $\left(\mathrm{AA}\right)_i$ and $\left(w_f\right)_i$ are the average amplitude and the weighting factor for the i-th analyzed variable, respectively. Each $\left(w_f\right)_i$ accounts for the experimental accuracy, the safety relevance of particular variables and its relevance with respect to the primary pressure. The weights must remain unchanged during each comparison between code results and experimental data concerning the same class of a transient (for more information on weighting factors see D'Auria et al., 1994). The acceptability factor for AA_{tot} was set to 0.4 and for primary system pressure to 0.1.

3. Time dependent accuracy

As mentioned in Section 2, the FFTBM methodology requires the qualitative assessment and the subdivision of the transient into phenomenological windows. Normally, the accuracy analysis is performed for time windows and time intervals, where each phenomenological window represents one time window, while time intervals start at the beginning of the transient and end at each phenomenological window end time.

Instead of a few phenomenological windows a series of narrow windows (phases) is proposed (around 40 windows / intervals for a transient). This gives the possibility to check the accuracy of each part of the transient and to get time dependency of accuracy measures. In the quantitative assessment with 3 to 5 phenomenological windows only global trends were available. In the present analysis by the term moving time window a set of equidistant narrow time windows as we progress into the transient is meant (like a moving chart strip). By the term increasing time interval a set of time intervals each increased for the duration of one narrow time window is meant, where the last time interval is equal to the whole transient duration time. The moving time window shows instantaneous details of $\Delta F(t)$ and consequently cannot draw an overall judgement about the accuracy, but focuses the analysis only on instantaneous discrepancies. An integral approach is needed to draw an overall judgement about accuracy and this is achieved by increasing the time interval, what also shows how the accuracy changes with time progression. From these time dependant accuracy measures it can be easily seen when the largest total discrepancy occurs and what is its influence on the total accuracy. They also show how the transient duration selected for the analysis influences the results.

In Fig. 2 are shown the results for the three different participants using different computer codes for the standard problem exercise no. 4 (SPE-4) (Szabados et al., 2009), simulating the small-break loss of coolant accident on the PMK-2 facility.



(a) Calculated and experimental signals

(b) Difference signals

(c) Total accuracy trend

(d) Time dependent total accuracy

Fig. 2. Results for SPE-4 test calculations of rod surface temperature

The PMK-2 facility is the first integral type facility for VVER-440/213 plants. In the study by Szabados et al., 2009 the results for the signals are plotted (see Fig. 2(a) and the accuracy for five selected time windows is given (see Fig. 2(c)). Figure 2(b) shows the difference signals, which are used for the AA calculation beside the experimental signal used for the normalization. Finally, Fig. 2(d) shows the AA calculated for increasing time intervals (each time 10 s), thus obtaining time dependent AA. When comparing Figs. 2(c) and 2(d) the contribution of discrepancies is better evident from Fig. 2(d). Also it can be seen how the accuracy overshoots due to edge effects, later stabilizing at some lower value.

4. Index for time shift detection

It should be noted that the AA accuracy measure (Equation 1) is not obtained by comparing the experimental and calculated magnitude spectra, but by calculating the magnitude spectrum of the difference signal. Nevertheless, due to the Fourier transform properties the magnitude spectrum of the difference signal can also be obtained by adding the experimental and calculated signal magnitude spectra (actually subtraction); they must be converted into a rectangular notation, added, and then reconverted back to a polar form. When the spectra are in a polar form, they cannot be added by simply adding the magnitudes and phases. The error function amplitude spectrum $\left|\tilde{\Delta}F(f_n)\right|$ can be expressed as:

$$\begin{aligned}&\left|\tilde{\Delta}F(f_n)\right| = \left|\tilde{F}_{exp}(f_n) - \tilde{F}_{cal}(f_n)\right| = \\ &= \sqrt{\left(\mathrm{Re}(\tilde{F}_{exp}(f_n) - \tilde{F}_{cal}(f_n))\right)^2 + \left(\mathrm{Im}(\tilde{F}_{exp}(f_n) - \tilde{F}_{cal}(f_n))\right)^2} = \\ &= \sqrt{M_1^2 + M_2^2 - 2M_1M_2\left(\cos\varphi_1\cos\varphi_2 + \sin\varphi_1\sin\varphi_2\right)},\end{aligned} \tag{4}$$

where $\tilde{F}_{exp} = M_1\cos\varphi_1 + iM_1\sin\varphi_1$ and $\tilde{F}_{cal} = M_2\cos\varphi_2 + iM_2\sin\varphi_2$ (rectangular form).

This example shows that to calculate the difference magnitude spectrum we need both the magnitude and the phase of the experimental and calculated spectra. Information about the shape of the time domain signal is contained in the magnitude and in the phase. In other words, comparing the shapes of the time domain signals is done through calculating the difference signal magnitude spectrum. At the time of the development of the original FFTBM (Ambrosini et al., 1990) it was mentioned that a possible improvement of the method could involve "the development of the procedure taking into account the information represented by the phase spectrum of the Fast Fourier Transform in the evaluation of accuracy". As we can see from Equation 4, the difference signal magnitude inherently includes the magnitude and the phase of the experimental and calculated signal. The finding that both, the magnitude and the phase of the experimental and calculated frequency spectra are contained in AA by making the Fourier transform of the difference signals is very important as this gives the possibility to compare the shapes of the signals. The authors agree with Smith et al., 1999 that it is difficult to imagine which information is contained in the phase spectrum of the difference signal, since the experimental and calculated phase cannot be simply added. Therefore, to the authors' opinion the difference signal phase information is not applicable for the comparison of two signals. In the following, AA will be referred to as $AA^{M\varphi}$, since it contains the magnitude M and phase φ information.

The original FFTBM package allows time shifting of data trends to analyze separately the effects of delayed or anticipated code predictions concerning some particular phenomena or systems interventions. It is a Fourier transform property that a shift in the time domain corresponds to a change in the phase. This property was used to identify the signals which differ in the time shift. Namely, the magnitudes of such signals are the same and only their phases are different. Therefore, the following expression, not taking into account the phase, is proposed (Prošek & Leskovar, 2009):

$$AA^{M}=\frac{\sum_{n=0}^{2^m}\left\|\tilde{F}_{exp}(f_n)\right|-\left|\tilde{F}_{cal}(f_n)\right\|}{\sum_{n=0}^{2^m}\left|\tilde{F}_{exp}(f_n)\right|}, \tag{5}$$

where:

$$\begin{aligned}&\left\|\tilde{F}_{exp}(f_n)\right|-\left|\tilde{F}_{cal}(f_n)\right\|=\\&=\left|\left(\left(\mathrm{Re}(\tilde{F}_{exp}(f_n))\right)^2+\left(\mathrm{Im}(\tilde{F}_{exp}(f_n))\right)^2\right)^{1/2}-\left(\left(\mathrm{Re}(\tilde{F}_{cal}(f_n))\right)^2+\left(\mathrm{Im}(\tilde{F}_{cal}(f_n))\right)^2\right)^{1/2}\right|=\\&=\left|M_1-M_2\right|=\sqrt{\left(M_1-M_2\right)^2}.\end{aligned} \tag{6}$$

When $\varphi_1=\varphi_2$, Equation 4 is equal to Equation 6. This means that expression AA^M is a measure containing information from magnitudes M only. It is known that when two signals are only time shifted, the magnitude spectra are the same and the value of AA^M is consequently zero. It is very unlikely that a calculated signal, which is not shifted, would have a shape giving the same magnitudes as the experimental signal, as the predictions are required to be qualitatively correct. Therefore, AA^M can be used to establish the value by which $AA^{M\varphi}\equiv AA$ is increased due to the time shift contribution. In $AA^{M\varphi}$, the information from both, the shape of the time domain signal and the time shift, is provided, while in AA^M, only the time invariant information of the time domain signal is provided, what can be regarded to a certain degree as the shape of the time domain signal. Therefore, the difference $AA^{\varphi}=AA^{M\varphi}-AA^M$ gives the information about the time shift contribution. This difference is further normalized to:

$$I=\frac{AA^{M\varphi}-AA^M}{AA^M}=\frac{AA^{\varphi}}{AA^M}, \tag{7}$$

where the indicator *I* tells how the compared time signals are shifted, and is therefore called the time shift indicator. The larger the value of the time shift indicator *I*, the larger is the contribution of the time shift to $AA^{M\varphi}$ of the difference signal. A large value of *I* (*I* > 1) indicates that the compared signals are maybe shifted in time. When *I* > 2, we can be quite confident into time shift.

5. Signal mirroring

Since the original FFTBM is based on the sum of the amplitudes of the frequency spectrum of the investigated signal, the frequencies originating from the artificially introduced edge may significantly contribute to the sum of the frequency spectrum amplitudes of the investigated signal. Consequently the accuracy measure based on the original FFTBM is significantly influenced by the edge and therefore does not present a consistent accuracy measure of the signals being compared. This inconvenient drawback of the original FFTBM may be completely cured by eliminating the artificially numerically introduced edge. This may be efficiently done by signal mirroring, where the investigated signal is mirrored before the FFTBM is applied.

5.1 Symmetrised signal

If we have a function $F(t)$, where $0 \le t \le T_d$ and T_d is the transient time duration, its mirrored function is defined as $F_{mir}(t) = F(-t)$, where $-T_d \le t \le 0$. From these two functions a new function is composed which is symmetrical around T_d: $F_m(t)$, where $0 \le t \le 2T_d$. By composing the original signal and its mirrored signal, a signal without an edge between the first and the last data sample is obtained when periodically extended, and is called symmetrised signal. The symmetrised signal is shown in Fig. 3. The upper figure shows the finite length signal and the lower figure shows the infinite length periodic signal.

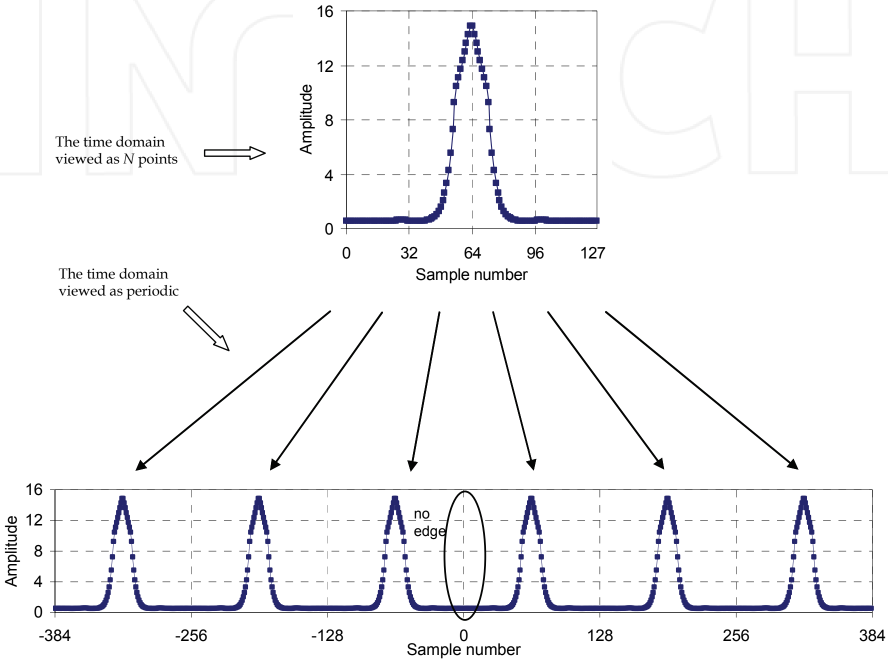


Fig. 3. Periodicity of the DFT's time domain symmetrised signal. The time domain can be viewed as N samples in length, shown in the upper part of the figure, or as an infinitely long periodic signal, shown in the lower part of the figure. We see that the symmetrised signal has no edge also when viewed as a periodic signal. Therefore in the sum of the frequency spectrum amplitudes only the amplitudes of the investigated signal are considered, as it should be. The FFTBM using the symmetrised signal is called "improved FFTBM by signal mirroring"

5.2 Deficiency of original FFTBM

When the original accuracy measures were proposed (Ambrosini et al., 1990) it seems that the impact of the edge effect was not considered. It is evident that this is a deficiency if the accuracy measure depends on the unphysical edge resulting from the intrinsic property of the DFT mathematical method, which treats the investigated finite length signal as an infinite length periodic signal. Namely, for the comparison the shape of the finite discrete signal is important and not its edge characteristics. Also the visual comparison of signals is done in such a way.

It was already mentioned that the periodicity is invoked in order to use a mathematical tool. Therefore this influence should be eliminated. This was done by signal mirroring. When DFT is applied to the finite length symmetrised time domain signal the edge effect is obviously not introduced anymore.

5.3 Calculation of AA_m

For the calculation of the average amplitude by signal mirroring (AA_m) Equation 1 is used as for the calculation of AA, except that, instead of the original signal, the symmetrised signal is used. The reason to symmetrise the signal was to exclude the artificial edge from the signal without influencing the characteristics of the investigated signal. The signal is automatically symmetrised in the computer program for the improved FFTBM by signal mirroring (updated version of software described in Prošek & Mavko, 2003).

As already mentioned, the edge has no physical meaning for comparison, since it was introduced artificially by the applied numerical method, but FFT produces harmonic components because of it. By mirroring, the shapes of the experimental and error signal are symmetric and their spectra are different from the original signals spectra, mainly because they are without unphysical edge frequency components. Due to different spectra the sum of the amplitudes changes in both, the numerator and the denominator of Equation 1. To further demonstrate this in Sections 6.2 and 6.3, two new definitions are introduced for the average amplitude of the error signal (AA_{err}) and the average amplitude of the experimental signal (AA_{exp}), related with the numerator and denominator of Equation 1:

$$\mathrm{AA}_{\mathrm{err}} = \frac{1}{2^m + 1} \sum_{n=0}^{2^m} \left| \tilde{\Delta} F(f_n) \right|, \tag{8}$$

$$\mathrm{AA}_{\mathrm{exp}} = \frac{1}{2^m + 1} \sum_{n=0}^{2^m} \left| \tilde{F}_{\mathrm{exp}}(f_n) \right|. \tag{9}$$

It should be noted that also when both, the original and error signal are without the artificial edge, in principal different AA_{err} and AA_{exp} may be obtained with the original FFTBM and the improved FFTBM by signal mirroring. Indeed AA and AA_m are slightly different measures also if the signals are without an artificial edge. The values obtained with the original FFTBM and the improved FFTBM by signal mirroring are the same only for symmetrical original signals. But this is not really a deficiency of the proposed improved FFTBM by signal mirroring, since it is important only that the method judges the accuracy on a realistic and unbiased way and that it is consistent within itself. In Section 6.4 it is presented how the accuracy calculated with the improved FFTBM by signal mirroring can be directly compared to the accuracy calculated with the original FFTBM.

6. Demonstration application

In this section some results are shown to see the advantages of the improved FFTBM by signal mirroring compared to the original FFTBM. First the test and the calculations used in the demonstration application of the improved FFTBM by signal mirroring are briefly described. Then two case studies are presented. The case 1 study is presented to show how the artificial edge (when present) always changes the accuracy even if this is logically not expected. In the case 2 study the accuracy of one variable is calculated for two time windows. Average amplitudes of the error and experimental signal are calculated to show the impact of the edge effect. Finally, the improved FFTBM by signal mirroring is applied to LOFT L2-5 calculations performed in the frame of the Best-Estimate Methods Uncertainty and Sensitivity Evaluation (BEMUSE) Phase II to further show that the improved FFTBM by signal mirroring is more consistent in the quantitative assessment than the original FFTBM.

6.1 Test description

The LOFT L2-5 test was selected for this demonstration because the huge amount of data was available and the assessment of these test results with the original FFTBM was already published in the literature (OECD/NEA, 2006). The calculations of the LOFT L2-5 test, which is the re-analysis of the ISP-13 exercise, were performed in the phase II of the BEMUSE research program. The nuclear LOFT integral test facility is a scale model of a pressurized water reactor (OECD/NEA, 2006). The objective of the ISP-13 test was to simulate a loss of coolant accident (LOCA) caused by a double-ended, off-shear guillotine cold leg rupture coupled with a loss of off-site power. The experiment was initiated by opening the quick opening blowdown valves in the broken loop hot and cold legs. The reactor scrammed and emergency core cooling systems started their injection. After initial heatup the core was quenched at 65 s, following the core reflood. The LPIS injection was stopped at 107.1 s, after the experiment was considered complete. In total 14 calculations from 13 organizations were performed using 6 different codes (9 different code versions). The code most frequently used was RELAP5/MOD3.3. For more detailed information on the calculations the reader is referred to (OECD/NEA, 2006, Prošek et al., 2008).

6.2 Case 1 study by signal mirroring

To demonstrate how signal mirroring works, FFT was applied to the signals shown in Figs. 1 and 3. The AA_{exp} values of signals were calculated per Equation 9. Imagine that you would quantitatively assess two variables, with the shape of the experimental signals as shown in Figs. 1 and 3. Most probably you would judge that the judgment based on FFTBM should be the same for both signals as the area below the curve is the same when normalized with the number of data samples (the area below the symmetrised curve is the double area of the original signal). Nevertheless, in the case of the original FFTBM different values of AA_{exp} are obtained (25.87 for the original signal and 16.29 for the symmetrised signal), while in the case of the improved FFTBM by signal mirroring the same results are obtained for AA_{exp} of the original and symmetrised signal (16.29). This means that both, the original FFTBM and the improved FFTBM by signal mirroring produce the same results when no artificial edge is present in the signal and when the signal is symmetrical. This example clearly shows that the original FFTBM is not consistent when an artificial edge is present in the signal.

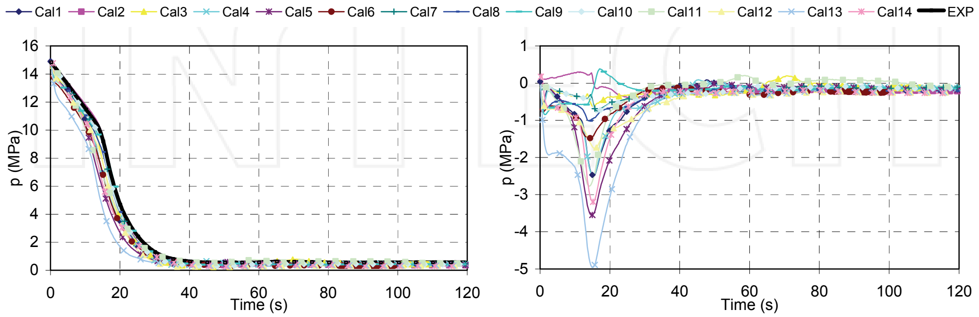
The difference between the original FFTBM and the improved FFTBM by signal mirroring results mainly due to the unphysical edge introduced by the applied numerical method, and can be directly quantified. The AA_{exp} of the symmetrised signal has to be extracted from the AA_{exp} of the original signal. In our example this contribution is 9.58 ((25.87 – 16.29) = 9.58). This means that the AA_{exp} of the experimental signal (used in the denominator of Equation 1) is 37% smaller when the edge effect is not considered, what would increase the value of accuracy measure AA for 59% in this example. This means that all integral variables (e.g. integrated break flow, ECCS injected mass) and variables dropping from the nominal to a low value (e.g. power, primary pressure during LOCA) exhibit lower AA values just because the artificially introduced unphysical edge is present in the experimental signal. This basically explains the, in general, very high accuracy of these integral variables (Prošek et al., 2002) in comparison to other variables and why the acceptability factor for primary pressure (D'Auria et al., 1994) (dropping during small break LOCAs) had to be set to the very low value 0.1 (for other parameters there is no need for a special criterion).

The improved FFTBM by signal mirroring provides a realistic, unbiased and consistent judgment, since it eliminates the effect of the unphysical edge, which sometimes is present and sometimes not. For example, when comparing primary pressures, during blowdown the pressure is decreasing and so a huge edge is present (it significantly decreases AA calculated by the original FFTBM), while during a very small break the pressure may recover to normal pressure after the initial drop due to emergency core cooling injection and consequently there is no edge (the original FFTBM then calculates similar values of AA as the improved FFTBM by signal mirroring).

6.3 Case 2 study by signal mirroring

To further demonstrate how signal mirroring works, in the second example the pressurizer pressure accuracy of LOFT L2-5 test calculations (see Fig. 4) is calculated for two time intervals, the blowdown phase time interval (0-20 s) and the whole transient time interval (0 – 119.5 s), as shown in Tables 1 and 2, respectively. Both, the original FFTBM and the improved FFTBM by signal mirroring were used. For each calculation the values of the average amplitude of the error signal (see Equation 8) and the average amplitude per Equation 1 are shown with the corresponding average amplitude of the experimental signal (see Equation 9). It should be noted that two calculations (Cal3, Cal6) did not provide data for the whole transient time interval; therefore for them the quantitative assessment was not applicable.

To see the influence of the edge elimination, the ratios of average amplitudes of the error signal obtained by the original FFTBM and the improved FFTBM by signal mirroring are shown in Tables 1 and 2. Besides the ratios, average amplitudes of the error signal, average amplitudes of the experimental signal, average amplitudes and rank of average amplitudes



(a) Calculated and experimental signals (b) Difference signals

Fig. 4. Results of the LOFT L2-5 test for pressurizer pressure

ID	AA_{err}	AA	$AA_{err\ m}$	AA_m	AA_{err}/ $AA_{err\ m}$	Rank AA	Rank AA_m
Cal1	4.28	0.136	3.63	0.209	1.2	11	9
Cal2	0.86	0.027	0.88	0.050	1.0	1	1
Cal3	1.34	0.043	1.14	0.066	1.2	2	2
Cal4	4.13	0.131	3.64	0.209	1.1	10	10
Cal5	6.80	0.216	4.76	0.274	1.4	13	13
Cal6	2.91	0.092	2.40	0.138	1.2	7	7
Cal7	1.80	0.057	1.35	0.078	1.3	5	4
Cal8	2.06	0.065	1.67	0.096	1.2	6	5
Cal9	1.73	0.055	2.16	0.124	0.8	4	6
Cal10	1.70	0.054	1.24	0.071	1.4	3	3
Cal11	4.01	0.128	3.86	0.222	1.0	9	11
Cal12	3.73	0.118	2.55	0.147	1.5	8	8
Cal13	9.86	0.314	6.92	0.399	1.4	14	14
Cal14	5.77	0.183	4.34	0.250	1.3	12	12
ID		AA_{exp}		$AA_{exp\ m}$		AA_{exp}/ $AA_{exp\ m}$	
EXP		31.45		17.37		1.8	

Table 1. Calculation of AA and AA_m for pressurizer pressures in time interval (0–20 s)

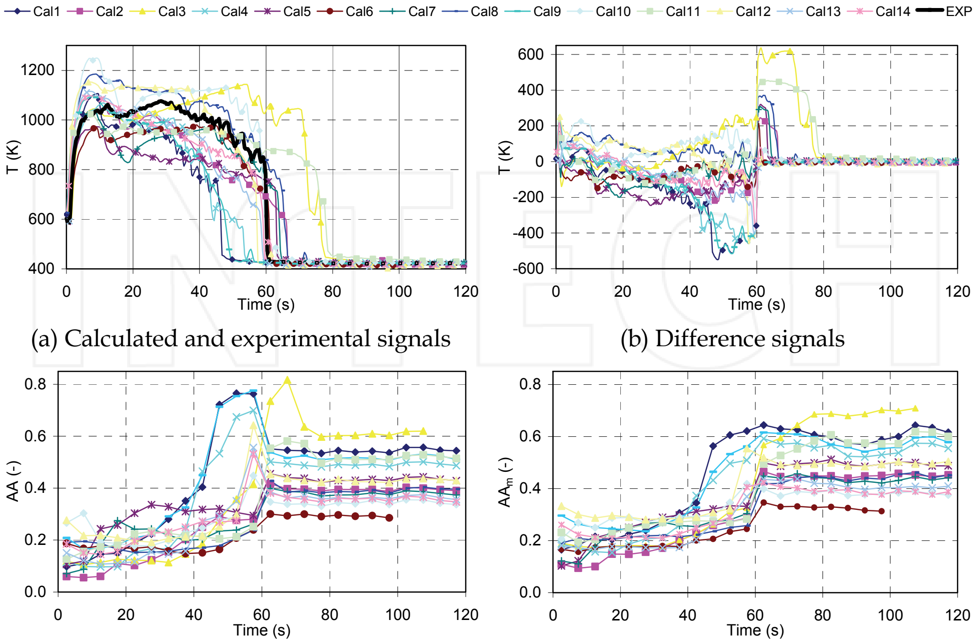
ID	AA_{err}	AA	$AA_{err\ m}$	AA_m	AA_{err}/ $AA_{err\ m}$	Rank AA	Rank AA_m
Cal1	3.06	0.096	3.85	0.237	0.8	8	8
Cal2	1.08	0.034	1.23	0.076	0.9	1	1
Cal3	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Cal4	3.10	0.097	3.97	0.244	0.8	9	9
Cal5	3.93	0.123	4.96	0.305	0.8	11	11
Cal6	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Cal7	1.17	0.036	1.38	0.085	0.8	3	3
Cal8	1.49	0.047	1.80	0.111	0.8	4	4
Cal9	1.64	0.051	2.09	0.129	0.8	5	5
Cal10	1.10	0.034	1.34	0.082	0.8	2	2
Cal11	2.97	0.093	3.74	0.230	0.8	7	7
Cal12	2.23	0.070	2.71	0.167	0.8	6	6
Cal13	5.96	0.186	7.44	0.458	0.8	12	12
Cal14	3.80	0.119	4.47	0.275	0.8	10	10
ID		AA_{exp}		$AA_{exp\ m}$		AA_{exp}/ $AA_{exp\ m}$	
EXP		32.00		16.26		2.0	

Table 2. Calculation of AA and AA_m for pressurizer pressures in time interval (0–119.5 s)

for both, the original FFTBM and the improved FFTBM by signal mirroring are shown. It can be seen that the average amplitude of the experimental signal is similar for both time intervals. The reason is that after 20 s the pressure signal (see Fig. 4(a)) is not changing very much. As the pressure at 20 s significantly dropped, the edge effect at 20 s is rather similar to the edge effect at 119.5 s. The average amplitudes of the experimental signal obtained by the original FFTBM are 1.8 and 2.0 times larger than by the improved FFTBM by signal mirroring for the first and second time interval, respectively. The conclusion for the error signals (see Fig. 4(b)) is different. The ratio of the average amplitudes of the error signal varies between 0.8 and 1.5 in the first time interval, while in the second time interval this ratio is around 0.8. The reason for the varying ratio in the first time interval is that the edges between calculations are quite different, while in the second time interval the edges are rather similar between calculations. Ranking of the AA values may change only in the case when the ratio of AA_{err} varies, i.e. in the first time interval, as it can be seen from Table 1. In the second time interval (in the whole calculation) the rank of AA remains unchanged, as shown in Table 2. Nevertheless, the absolute value of AA changes when the edge is not considered in the experimental signal and this influences the total accuracy.

6.4 Application to single variable

Fig. 5 shows the comparison between experimental and calculated data, the error signals between the calculation and the experiment, AA calculated with the original FFTBM, and AA_m calculated with the improved FFTBM by signal mirroring. For the rod surface temperature shown in Fig. 5(a) the calculated maximum values of the rod surface temperature were in rather good agreement with the experimental value. However, the trends were in general under or over predicted, with some calculations that predicted quench too early. From the error signals shown in Fig. 5(b) it can be seen that discrepancies were present until core quench. From Fig. 5(c) it can be seen that the edge effect is present in the calculation of AA. When looking AA_m in Fig. 5(d) it can be seen that the value of AA_m monotonically increases as long as the discrepancy is present. Nevertheless, when considering the whole transient time interval, there is only a slight difference between AA and AA_m. The reason is the small edge in the whole transient time interval. This is the reason why the original FFTBM in several cases produced reasonable results (but not for monotonically increasing and decreasing signals like the pressurizer pressure). Nevertheless, for investigating the influence of discrepancies as we progress into the transient the edge effect needs to be eliminated to make the right conclusions. Only the improved FFTBM by signal mirroring gives consistent results. Consistent judgment of the time dependent accuracy is very important as the analyst in this way gets an objective picture how each discrepancy decreases the accuracy. On the other hand, from Fig. 5(c) it can be very easily verified that the requirements for accuracy measures (Ambrosini et al., 1990) that at any time into the transient the previous history should be remembered and that the measure should be independent upon the transient duration are not well fulfilled in the case of the original FFTBM when the edge influences the results. Through performing the time dependent accuracy, tens of calculations for different time intervals were performed demonstrating the consistency of the improved FFTBM by signal mirroring comparing to the original FFTBM and therefore there is no need to use further experiments for the validation of the improved FFTBM. On the other hand, very frequently at the end of the transient the edges are rather small and in such cases also the original FFTBM produces consistent results. Luckily, this was the case in several studies performed with the original FFTBM (Prošek et al., 2002). Nevertheless, before the analyst is confident to the results obtained by the original FFTBM he should always verify that the edge is not present in the signal. In the opposite, the results are doubtful. It should be also noted that the methodology using the FFTBM requires qualitative analysis with visual observation and only for discrepancies which are reasonable and understood the quantitative assessment using FFTBM could be done.



(a) Calculated and experimental signals

(b) Difference signals

(c) AA accuracy measure

(d) AA_m accuracy measure

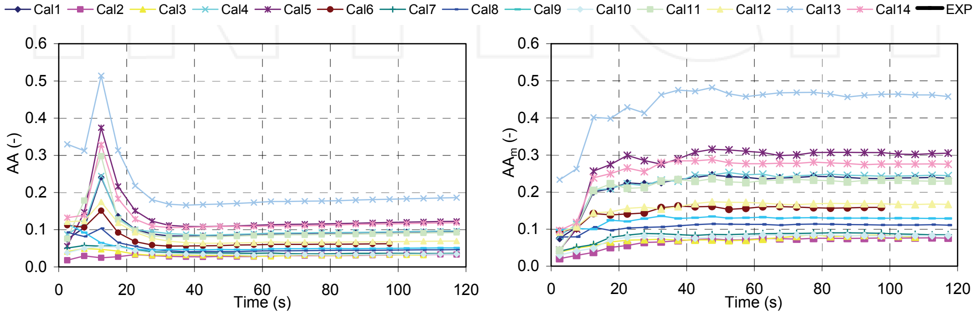
Fig. 5. Results for the LOFT L2-5 test calculations of rod surface temperature

6.5 Accuracy criterion for primary pressure

The differences between AA and AA_m as a function of time were clearly shown to be due to the edge contribution. On the other hand, for the whole transient time interval with stabilized conditions resulting in small edges also the judgment by the original FFTBM is qualitatively correct. However, this is never the case for monotonic trends where the edge increases with increasing the transient time. This can be seen from Fig. 6 showing AA and AA_m for pressurizer pressure shown in Fig. 4. After 20 s the AA value is low due to the large edge present in the experimental signal, which is used for normalization.

Based on the results in Fig. 6 it seems that the restrictive pressure criterion (AA below 0.1) in the original FFTBM was set, because it was based on pressure trends during small break LOCAs in facilities simulating typical PWRs (high initial pressure and large pressure drop after break occurrence, therefore high edge). When tests on different facilities were simulated, there were difficulties in satisfying the primary pressure criterion. The first example is the accuracy quantification of four standard problem exercises (SPEs) organized by IAEA (D'Auria et al., 1996). In this study only the primary system pressure has been considered. Among other things it was also concluded that in the case of SPE-3 the calculation is clearly unacceptable (AA was 0.31) and that more complex transients lead to worse results than simple one's. As no plots are shown in the paper by D'Auria et al. (1996)

no further conclusion can be done except that the pressure drop (edge) is smaller than in a typical PWR. Namely, the initial pressure in this test is lower than in the typical PWR test. By lowering the pressure edge the values of AA are increased. This can be still better illustrated in the recent application of FFTBM to heavy water reactors. In the study (Prošek et al., 2006) all participants fulfilled the acceptance criterion for the total accuracy K< 0.4 while the primary pressure criterion was not fulfilled. In the blind accuracy calculation the AA value for the primary pressure was 0.117 for the best calculation. The header 7 pressure with initial pressure around 10 MPa was selected as a variable representing the primary pressure. In the open accuracy analysis header 6 pressure was proposed by a representative from Italy. The initial value of this pressure was around 12 MPa. Now the value of AA was below 0.1 for most of participants mostly due to the increased pressure edge effect (the best AA was 0.074) due to the higher pressure.



(a) AA for LOFT L2-5 (b) Difference signals

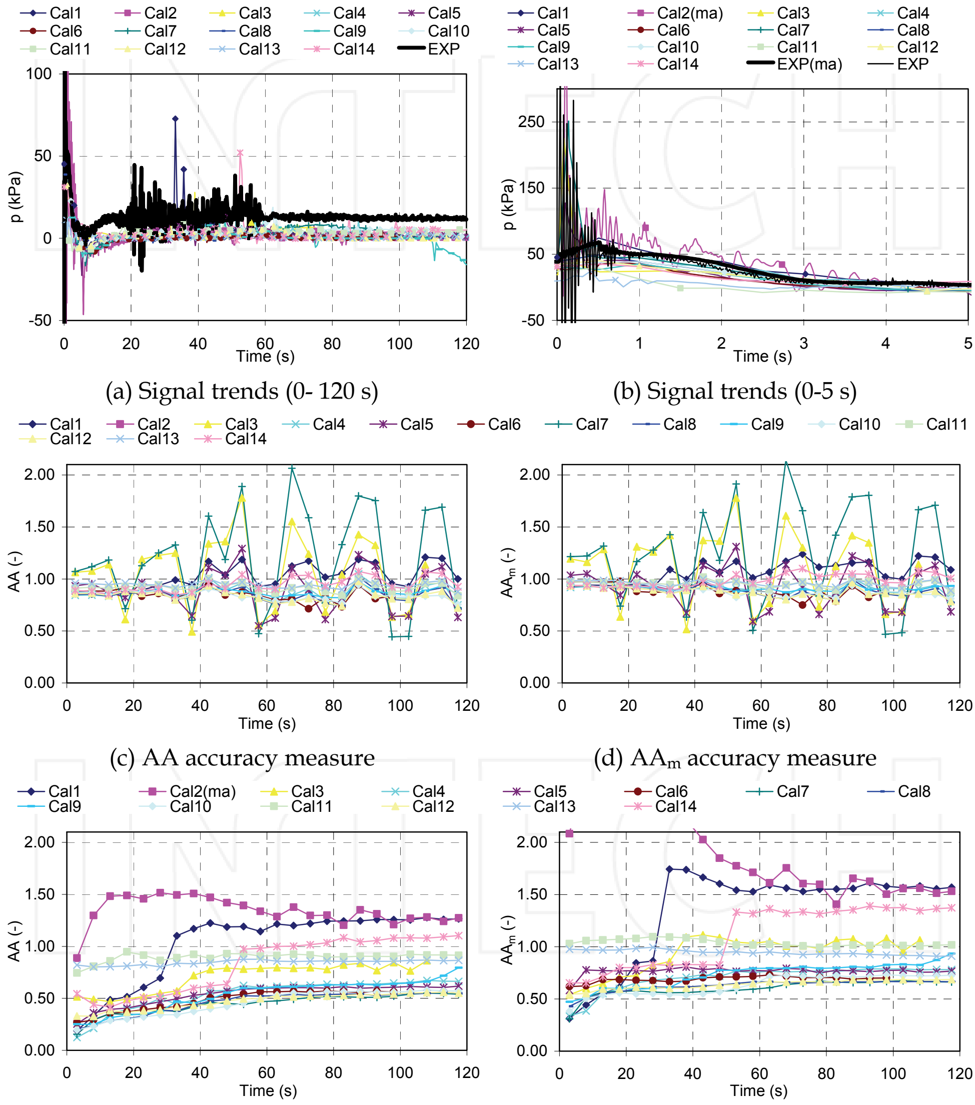
Fig. 6. Accuracy measures for the LOFT L2-5 test calculations of pressurizer pressure

The last example of the AA calculation for the primary pressure is for ISP-22 calculations (loss of feedwater test). From paper by Prošek et al. (2002) it may be seen that the AA value for the primary pressure in the best posttest calculation is 0.21, the worst (as judged by the original FFTBM) among summarized ISPs. From the original report (Ambrosini et al., 1992) showing plots it can be easily concluded that the edge contribution in the experimental signal is smaller than typically for small break LOCAs due to lower pressure drop and the complex shape of the experimental signal, resulting in larger AA.

All these examples demonstrate that due to the unpredictable edge contribution a consistent criterion for the primary pressure cannot be defined for the original FFTBM, while for the improved FFTBM with signal mirroring this can be done.

6.6 Moving average

When trends oscillate greatly (e.g., the steam generator pressure drops shown in Fig. 7), special treatment is needed (Prošek & Mavko, 2009). To correctly reproduce the experimental signal by linear interpolation, many points are needed. This is achieved by increasing the maximum frequency component of the signal. However, it makes no sense to increase the number of points, as some cases have a sampling frequency 30 times larger than the calculated data. When many points are used, the main contribution to the amplitude spectrum comes from the oscillations (very often noise) in the experimental signal for which the calculated data have no information. The correct procedure is therefore to smooth the data. Smoothing data removes random variations and shows trends and cyclic components. The simplest way to smooth the data is by taking the averages. This is done by use of the moving average of the experimental signal. Mathematically, the moving average is an example of a convolution of the input signal with a rectangular pulse having an area of 1.



(a) Signal trends (0- 120 s)

(b) Signal trends (0-5 s)

(c) AA accuracy measure

(d) AA$_m$ accuracy measure

(e) AA accuracy measure (moving average)

(f) AA$_m$ accuracy measure (moving average)

Fig. 7. Results for the LOFT L2-5 test calculations of steam generator pressure drop

Without using the moving average, AA varies around a certain value. In the presented case for steam generator pressure drop, the values of AA and AA_m are close to 1 (see Figs. 7(c) and 7(d)), because the calculated values are much smaller than the experimental values. The exception is Cal2 which values vary around 4 and is not visible in Figs. 7(c) and 7(d). Variations in AA are the consequence of inappropriately prepared experimental data for the FFTBM analysis. The problem of the oscillatory signal was less significant in the past, because the original FFTBM limited the number of data points to 1,000, and data reduction was needed when this value was exceeded. Thus, data reduction is another possibility to use for partially smoothing the signal and thereby increasing the accuracy by eliminating some noise. However, as shown by Figure 4(e) in Prošek et al., 2006, the AA still varies because the moving average was not used. The reason is that, by increasing the time interval and not increasing the number of points, the amplitude spectrum changes as the signal between two consecutive data points is not a monotonic function (it oscillates). This gives a different amplitude spectrum of the experimental and difference signal. When moving average was used in the case of the steam generator pressure drop experimental signal, the AA values no longer oscillate in phase because of AA_{exp}, as shown in Figs. 7(e) and 7(f). This suggests that the observation of oscillations being in phase in the calculated AAs indicates that moving average should be used. Figures 7(e) and 7(f) show a sudden increase in AA in the Cal1 and Cal5 calculations. The reason for this increase are the pressure spikes clearly shown in Fig. 7(a). Each spike significantly deteriorates the results. Finally, FFTBM was able to detect the deviation in the Cal9 calculation at the end of the transient.

Another important finding is that the mismatch between the experimental data and the calculations for the steam generator pressure drop variable is present from the very beginning of the transient, as shown in Fig. 7(b). Only the Cal2 calculation reproduced the frequency of oscillations in the first second. However, because the peaks were too high, the calculation was not very accurate. Use of moving average removes the large oscillations from the experimental signal (EXP(ma)), while in the Cal2(ma) calculation, the oscillations still remain in the beginning of the transient. Later (at approximately 15 seconds), the pressure drop stabilizes and the values oscillate around their mean values. This means that the transient related to the pressure drop has more or less ended.

6.7 Comparison of results obtained by FFTBM and ACAP

Tables 3 and 4 shows the comparison of FFTBM and Automated Code Assessment Program (ACAP) (Kunz et al., 2002) accuracy measures for the calculated pressurizer pressure and rod surface temperature shown in Figs. 4 and 5, respectively. This comparison was made for the independent assessment that FFTBM provides for consistent accuracy measures. The calculations are sorted according to AA_m in ascending manner. For the pressurizer pressure it can be seen that AA_m, AA, mean square error (MSE), and cross-correlation coefficient (XCC) accuracy measures agree well. The only difference is that MSE and XCC indicate that all calculations of pressurizer pressure are very good, while FFTBM shows that some are not so accurate and some do not even fulfil the original FFTBM primary pressure criterion. As the pressure criterion was developed without consideration of the edge effect, care must be taken in its use, as indicated by the ACAP results. Finally, D'Auria fast Fourier transform (DFFT) and continuous wavelet transform (CWT) accuracy measures do not help much in this case.

Method	FFTBM		ACAP			
Calculation	AA_m	AA	DFFT	MSE	XCC	CWT
Cal6 (100 s)	0.076	0.034	0.194	1.000	1.000	0.154
Cal10	0.079	0.032	0.132	1.000	0.999	0.008
Cal14	0.082	0.034	0.173	1.000	0.999	0.116
Cal13	0.085	0.036	0.223	1.000	0.999	0.059
Cal7	0.111	0.047	0.173	0.999	0.999	0.148
Cal8	0.129	0.051	0.194	1.000	0.999	0.179
Cal2	0.159	0.062	0.168	0.999	0.999	0.008
Cal5	0.167	0.070	0.134	0.998	0.997	0.126
Cal12	0.230	0.093	0.082	0.998	0.993	0.006
Cal4	0.237	0.096	0.129	0.998	0.994	0.140
Cal9	0.244	0.097	0.128	0.998	0.995	0.220
Cal11	0.275	0.119	0.110	0.997	0.992	0.069
Cal1	0.305	0.123	0.089	0.996	0.983	0.091
Cal3 (110 s)	0.458	0.186	0.096	0.991	0.972	0.053

Table 3. Comparison of FFTBM and ACAP accuracy measures for pressurizer pressure in time interval (0–119.5 s)

Method	FFTBM		ACAP			
Calculation	AA_m	AA	DFFT	MSE	XCC	CWT
Cal6 (100 s)	0.313	0.285	0.245	0.989	0.992	0.020
Cal10	0.375	0.337	0.197	0.984	0.987	0.193
Cal14	0.388	0.347	0.228	0.988	0.973	0.171
Cal13	0.409	0.356	0.206	0.982	0.960	0.106
Cal7	0.442	0.374	0.203	0.983	0.968	0.006
Cal8	0.451	0.396	0.182	0.980	0.972	0.010
Cal2	0.452	0.391	0.208	0.982	0.962	0.055
Cal5	0.488	0.429	0.181	0.968	0.962	0.055
Cal12	0.504	0.429	0.207	0.981	0.967	0.004
Cal4	0.555	0.487	0.158	0.948	0.883	0.043
Cal9	0.578	0.511	0.150	0.938	0.853	0.049
Cal11	0.600	0.515	0.152	0.940	0.832	0.000
Cal1	0.616	0.544	0.151	0.929	0.841	0.055
Cal3 (110 s)	0.708	0.620	0.149	0.901	0.780	0.000

Table 4. Comparison of FFTBM and ACAP accuracy measures for rod surface temperature (i.e. rod cladding temperature) in time interval (0–119.5 s)

For rod surface temperature (see Table 4), AA_m, AA, MSE, and XCC accuracy measures agree well. The XCC figure of merit is in especially good agreement with AA_m. When comparing the Cal12 and Cal13 calculations, FFTBM slightly favours the Cal13 calculation, while ACAP gives comparable values. The qualitative analysis of dryout occurrence reported in Table 13 of the BEMUSE Phase II Report (OECD/NEA, 2006) showed, that the Cal13 calculation receives three excellent and one minimal mark, while the Cal12 calculation receives two excellent, one reasonable, and one minimal mark. One parameter representing the dryout occurrence is the peak cladding temperature and for it the Cal13 calculation is qualitatively judged better than the Cal12 calculation. These BEMUSE results support the FFTBM judgments for cladding temperature. Examination of AA_m in Fig. 5(d) shows that, in the initial period of 40 seconds, the Cal13 calculation is significantly better because of the Cal12 calculation's large overprediction of cladding temperature.

6.8 Discussion

A demonstration application of the improved FFTBM by signal mirroring was done for a design basis accident. In the case of the LOFT L2-5 test calculation it was shown that only the improved FFTBM by signal mirroring gives a realistic judgment for the time dependent accuracy. The differences between AA and AA_m as a function of time were clearly shown to be due to the edge contribution. On the other hand, for the whole transient time interval with stabilized conditions resulting in small edges also the judgment by the original FFTBM is qualitatively correct. However, this is never the case for monotonic trends where the edge increases with increasing the transient time.

In general there is a need to make comparisons for any time window and the transient may not be terminated at stable conditions resulting in small edges. For the proposed improved FFTBM by signal mirroring the acceptability criteria need to be defined in the same way as this was done for the original FFTBM. The easiest way would be to use the same set of calculations as for the original FFTBM. The obtained results for LOFT L2-5 suggest slightly higher acceptability limits for the improved FFTBM by signal mirroring than for the original FFTBM, including the restrictive pressure criterion.

7. Conclusions

In the past the most widely used method for code accuracy quantification of primary system thermal-hydraulic codes was the original FFTBM. Recently, in the original FFTBM an important deficiency was discovered. It turned out that the accuracy measure depends on the difference between the first and last data point of the investigated signal. Namely, the DFT mathematical method, on which the FFTBM is based, treats the investigated finite length signal as an infinite length periodic signal, introducing discontinuities if the first and last data point of the finite signal differ. These discontinuities produce a variegated spectrum of frequencies when applying DFT, which may overshadow the frequency spectrum of the investigated signal. This so called edge effect is a significant deficiency of the original FFTBM since for the comparison the shape of the investigated signal is important and not the artificially introduced unphysical edge.

Therefore the authors proposed to resolve the edge effect problem on a unique way by signal mirroring, where the investigated signal is mirrored before FFTBM is applied. By composing the original signal and its mirrored signal a symmetric signal with the same characteristics is obtained, but without introducing artificial discontinuities when viewed as a periodically extended infinite signal. With the so improved FFTBM by signal mirroring a consistent and unbiased tool for quantitative assessment is obtained. An additional good property of the improved FFTBM is that the same FFTBM procedure (numerical tools etc.) may be applied as with the original FFTBM.

The benefits of the improved FFTBM by signal mirroring were demonstrated on the large break LOCA test LOFT L2-5. The results show that the so improved FFTBM judges the accuracy of variables in a reliable, unbiased and consistent way. Nevertheless, the new measure for indication of the time shift between the experimental and the calculated signal can be used only by the original FFTBM. It is also suggested to make all operations in the time domain for both, the original and the improved FFTBM, as it is very difficult to make adjustments in the frequency domain (e.g. logarithmic scale, moving average). There is no way to make such adjustments automatically. Finally, it should not be forgotten, that the qualitative analysis is a prerequisite to perform the quantitative analysis. This means that thermal hydraulic code calculations must be analyzed by experts first, and only then FFTBM can assist in conducting an objective comparison and answering if improvements to the input model are needed.

8. References

Aksan, S.N., D’Auria, F. & Bonato, S. (2001). Application of Fast Fourier Transform Based Method (FFTBM) to the results of ISP-42 PANDA test calculations: Phase A. *Proceedings of the ICONE-9*, April 8-12, 2001, Nice, France.

Ambrosini, W., Bovalini, R. & D’Auria, F. (1990). Evaluation of accuracy of thermalhydraulic code calculations. *Energia Nucleare*, 7, 2 (May-September 1990), 5–16.

Ambrosini, W., Breghi, M.P., D’Auria, F. & Galassi, G.M. (1992). Evaluation of post-test analyses of OECD-CSNI International Standard Problem 22. Report, University of Pisa, NT 184 (91) Rev.1.

D’Auria, F. & Galassi, M. (1997). Accuracy Quantification by the FFT method in FARO L-14 (ISP 39) open calculations, University of Pisa, NT 309(97).

D’Auria, F., Eramo, A., Frogheri, M. & Galassi, G.M. (1996). Accuracy quantification in SPE-1 to SPE-4 organised by IAEA. *Proc. Int. Conf. on Nucl. Engineering (ICONE-4)*, Vol. 3, pp. 461-469, ISBN 0-7918-1226-X, New Orleans, Louisiana, March 10-14, 1996.

D’Auria, F., Oriolo, F., Leonardi, M. & Paci, S. (1995). Code Accuracy Evaluation of ISP35 Calculations Based on NUPEC M-7-1 Test. *Proc. Second Regional Meeting: Nuclear Energy in Central Europe*, pp. 516–523. ISBN 961-900004-9-8, Portorož, Slovenia, September 2005, Nuclear Society of Slovenia, Ljubljana.

D'Auria, F., Leonardi, M. & Pochard, R. (1994). Methodology for the evaluation of thermalhydraulic codes accuracy. *Proc. Int. Conf. on New trends in Nuclear System Thermohydraulics*, pp. 467-477, Pisa, Italy, May 2004, Edizioni ETS Pisa.

Kunz, R.F., Kasamala, G.F., Mahaffy, J.H. & Murray, C.J. (2002). On the automated assessment of nuclear reactor systems code accuracy. *Nuclear Engineering and Design*, 211, 2-3 (February 2002), 245-272, ISSN 0029-5493.

Mavko, B., Prošek, A. & D'Auria, F. (1997). Determination of code accuracy in predicting small-break LOCA experiment. *Nuclear Technology*, 120, 1, (October 1997), 1-19, ISSN 0029-5450.

OECD/NEA (2006). BEMUSE Phase 2 Report: Re-Analysis of the ISP-13 Exercise, Post Test Analysis of the LOFT L2-5 Test Calculation. OECD/NEA Report, Committee on the Safety of Nuclear Installations (CSNI), NEA/CSNI/R(2006)2.

Prošek, A. & Leskovar, M. (2009). Extensions of the fast Fourier transform based method for quantitative assessment of code calculations. *Electrotechnical Review*, 76, 5 (December 2009), 251–256, ISSN 0013-5852, (http://ev.fe.uni-lj.si/5-2009/Prosek.pdf).

Prošek, A. & Mavko, B. (2003). A tool for quantitative assessment of code calculations with an improved fast Fourier transform based method. *Electrotechnical Review*, 70, 5 (December 2003), 291–296, ISSN 0013-5852, (http://ev.fe.uni-lj.si/5-2003/prosek.pdf).

Prošek, A. & Mavko, B. (2009). Quantitative code assessment with fast Fourier transform based method improved by signal mirroring, (International agreement report, NUREG/IA-0220). Washington: U. S. NRC (http://www.nrc.gov/reading-rm/doc-collections/nuregs/agreement/ia0220/ia0220.pdf).

Prošek, A., D'Auria, F., Richards, D.J. & Mavko, B. (2006). Quantitative assessment of thermal–hydraulic codes used for heavy water reactor calculations. *Nuclear Engineering and Design,* 236, 3 (February 2006), 295–308, ISSN 0029-5493.

Prošek, A., D'Auria, F. & Mavko, B. (2002). Review of quantitative accuracy assessments with fast Fourier transform based method (FFTBM). *Nuclear Engineering and Design*, 217, 1-2 (August 2002), 179-206, ISSN 0029-5493.

Prošek, A., Leskovar, M. & Mavko, B. (2008). Quantitative assessment with improved fast Fourier transform based method by signal mirroring. Nuclear Engineering and Design, 238, 10 (October 2008), 2668-2677, ISSN 0029-5493. (Figs. 1, 3, 4, 5, 6 and Tables 1, 2 Reprinted from Nuclear Engineering and Design, Volume 238, Issue 10, A. Prošek, M. Leskovar, B. Mavko, Pages 2668-2677, Copyright (2010), with permission from Elsevier.)

Smith S. W. (1999). *The Scientist and Engineer's Guide to Digital Signal Processing*, Second Edition, California Technical Publishing, ISBN 0-9660176-7-6, San Diego, California.

Szabados, L., Ézsöl, Gy., Perneczky, L., Tóth, I., Guba, A., Takács, A. & Trosztel, I. (2009). *Volume II. Major findings of PMK-2 Test Results and Validation of thermohydraulic System Codes for VVER Safety Studies*, Akadémiai Kiadó, ISBN 978-963-05-8810-2, Budapest, Hungary.

Fourier Transforms - Approach to Scientific Principles
Edited by Prof. Goran Nikolic

ISBN 978-953-307-231-9
Hard cover, 468 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821