FOURIER TRANSFORM APPLICATIONS

Edited by **Salih Mohammed Salih**

INTECHOPEN.COM

Fourier Transform Applications
Edited by Salih Mohammed Salih

Published by InTech
Janeza Trdine 9, 51000 Rijeka, Croatia

Notice
Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

Publishing Process Manager Vana Persen
Technical Editor Teodora Smiljanic
Cover Designer InTech Design Team

First published April, 2012
Printed in Croatia

A free online edition of this book is available at www.intechopen.com
Additional hard copies can be obtained from orders@intechopen.com

Fourier Transform Applications, Edited by Salih Mohammed Salih
p. cm.
ISBN 978-953-51-0518-3

Contents

Preface

During the preparation of this book, we found that almost all the textbooks on signal analysis have a section devoted to the Fourier transform theory. The Fourier transform is a mathematical operation with many applications in physics, and engineering that express a mathematical function of time as a function of frequency, known as its frequency spectrum; Fourier's theorem guarantees that this can always be done. The basic idea behind all those horrible looking formulas is rather simple, even fascinating: *it is possible to form any function as a summation of a series of sine and cosine terms of increasing frequency*. In other words, any space or time varying data can be transformed into a different domain called the *frequency space*. A fellow called *Joseph Fourier* first came up with the idea in the 19th century, and it was proven to be useful in various applications, mainly in signal processing. As far as we can tell, Gauss was the first to propose the techniques that we now call the Fast Fourier Transform (FFT) for calculating the coefficients in a trigonometric expansion of an asteroid's orbit in 1805. However, it was the seminal paper by Cooley and Tukey in 1965 that caught the attention of the science and engineering community and, in a way, founded the discipline of digital signal processing. While the Discrete Fourier Transform (DFT) can be applied to any complex valued series, in practice for large series it can take considerable time to compute, the time taken being proportional to the square of the number on points in the series. It is hard to overemphasis the importance of the DFT, convolution, and fast algorithms. The FFT may be the most important numerical algorithm in science, engineering, and applied mathematics. New theoretical results are still appearing, advances in computers and hardware continually restate the basic questions, and new applications open new areas for research. It is hoped that this book will provide the background, references and incentive to encourage further research and results in this area as well as provide tools for practical applications. One of the attractive features of this book is the inclusion of extensive simple, but practical, examples that expose the reader to real-life signal analysis problems, which has been made possible by the use of computers in solving practical design problems. The aim of this book is to expand the applications of Fourier transform into main three sections:

The first section deals with *Electromagnetic Field and Microwave Applications*. It consists of five chapters. The chapters are related to: the computation of transient near-field radiated by electronic devices, analysis of novel optically-inspired phenomena at microwaves, the computation of lightning electromagnetic field, beamforming and DOA estimation, and the solar microwave radiation.

The chapters of the second section discuss some advanced methods used in Fourier transform analysis which are related to the *Medical Applications*. This section consists of three chapters. The chapters are related to: spectral analysis of global behaviour of C. Elegans chromosomes, spectral analysis of heart rate variability in women, and the cortical specification of a fast Fourier transform supports a convolution model of visual perception.

The third section includes the *Fourier and Helbert Transform Applications*. This section consists of four chapters. The chapters are concerns to: the Fourier convolution theorem over finite fields (error control coding), application of the weighted energy method in the partial Fourier space to linearized viscous conservation laws with non-convex condition, Fourier transform methods for option pricing, and the Hilbert transform applications.

Finally, we would like to thank all the authors who have participated in this book for their valuable contribution. Also we would like to thank all the reviewers for their valuable notes. While there is no doubt that this book may have omitted some significant findings in the Fourier transform field, we hope the information included will be useful for electrical engineers, control engineers, communication engineers, signal processing engineers, medical researchers, and the mathematicians, in addition to the academic researchers working in the above fields.

Salih Mohammed Salih
College of Engineering
University of Anbar
Iraq

Part 1

Electromagnetic Field and Microwave Applications

1

Computation of Transient Near-Field Radiated by Electronic Devices from Frequency Data

Blaise Ravelo and Yang Liu
IRSEEM (Research Institute in Embedded Electronic System), EA 4353,
Graduate School of Engineering ESIGELEC,
76801 Saint Etienne du Rouvray Cedex,
France

1. Introduction

Facing to the increase of architecture complexity in the modern high-speed electronic equipments, the electromagnetic compatibility (EMC) characterization becomes a crucial step during the design process. This electromagnetic (EM) characterization can manifest with the unintentional conducting or radiating perturbations including, in particular, the near-field (NF) emissions. Accurate modelling method of this emission in NF zone becomes one of electronic engineer designers and researchers most concerns (Shi et al. 1989, Baudry et al. 2007, Vives-Gilabert et al. 2007, Vives-Gilabert et al. 2009, Song et al. 2010, Yang et al. 2010). This is why since the middle of 2000s; the NF modelling has been a novel speciality of the electronic design engineers. This modelling technique enables a considerable insurance of the reliability and the safety of the new electronic products. To avoid the doubtful issues related to the EM coupling, this analysis seems indispensable for the modern RF/digital electronic boards vis-à-vis the growth of the integration density and the operating numerical data-speed which achieves nowadays several Gbit/s (Barriere et al. 2009, Archambeault et al. 2010). In this scope, the influence of EM-NF-radiations in time-domain and in ultra-wide band (UWB) RF-/microwave-frequencies remains an open-question for numerous electronic researchers and engineer designers (Ravelo et al. 2011a & 2011b, Liu Y. et al. 2011a & 2011b). In the complex structures, the current and voltage commutations in the non-linear electronic devices such as diodes, MOSFETs and also the amplifiers can create critical undesired transient perturbations (Jauregui et al. 2010a, Vye 2011, Tröscher 2011, Kopp 2011). Such electrical perturbations are susceptible to generate transient EM-field radiations which need to be modelled and mastered by the electronic handset designers and manufacturers.

1.1 Overview on the NF radiations characterization occurring in the RF/microwave-device in time-domain

It is noteworthy that the frequency-investigations on the EM-radiation of electronic devices are not sufficient for the representation of certain EM-transient phenomena notably when the sources of perturbations behave as a short duration pulse-wave. In fact, it does not enable to precise the probably instant times and the intensity peak of the EM-pulse. That is why the time-domain representation is particularly essential for the infrequent and ultra-short duration wave emission analysis. In order to investigate more concretely the unwanted time-domain perturbations, different EM-NF modelling and measurement techniques were recently introduced and published in the literature (Cicchetti 1991, Adada 2007, Liu L. et al. 2009, Winter & Herbrig 2009, Ordas et al. 2009, Braun et al. 2009, Rioult et al. 2009, Xie & Lei 2009, Edwards et al. 2010, Jauregui et al. 2010b, Ravelo 2010). Furthermore, several EM-solvers are also integrated in the commercial simulation tools for the determination of the EM-field radiations by the RF/microwave devices especially in frequency domain (ANSOFT 2006, AGILENT 2008, ANSYS 2009, NESA 2010).

Currently, the computation method of the EM-field becomes systematically more and more complicated when the electronic systems operate with baseband UWB signals. Despite the recent investigations conducted on the finite-difference time-domain method (FDTD) method (Liu et al. 2009, Jauregui et al. 2010b), the accuracy of the computation results with these time-domain commercial tools remains difficult to evaluate when the perturbation sources are induced from ultra-short duration transient NF. In addition, more practical techniques (Cicchetti 1991, Braun et al. 2009, Winter & Herbrig 2009, Ordas 2009, Rioult et al. 2009) have been also introduced for the measurement of the electric- and electronic- system electromagnetic interference (EMI). But compared to the existing frequency measurement techniques, they are much better because of the limitations either in terms of space-resolution or electro-sensitivity or simply the calibration process. So, the evaluation of the accurate graphs of time-dependent EM-waves in NF is still an open challenge.

To cope with this limitation, in this chapter, an efficient computation methodology based on the transformation of wide bandwidth and baseband frequency-dependent data for the determination of the transient EM-NF mapping permitting is developed. In order to take into account the transient radiations specific to the expected use cases, an adequate excitation signal should be considered. This excitation is usually defined according to certain technical parameters (amplitude, temporal width, variation speed, time-duration…) which qualifies the undesired disturbing signal susceptible to propagate in the emitting circuits. Then, the fast Fourier transform (fft) mathematical treatment of the assumed disturbing signal synchronized with the given discrete frequency-dependent data in the adequate frequency range enables to determine the transient wave radiation mapping.

1.2 Background on the EMC application of the transient EM-NF

As aforementioned and discussed (Rammal et al. 2009, Jauregui et al. 2010a), the EM-transient analysis is actually important for the immunity predictions in the mixed or analogue-digital components constituting the high-speed electronic boards regarding the eventual radiations of high power electrical circuitry as the case of neighbouring hybrid electric vehicle propulsion systems. To assess such an EMC effect, as reported in (Adada 2007), the electronic circuit designers working on analogue/mixed signal (AMS) subsystems have preferred software tools such as SPICE, while those working on RF/microwave front-end components have tended to manipulate S-parameter frequency-domain design and simulation tools. By cons, currently, the fusion of the both approaches as AMS engineers are required to make further analysis on the critical components is needed by using the adequate EM simulation tools. In this case, we have to elaborate the context of ultra-wide band (UWB). Currently, this topic is one of improvement techniques in the area of EMC application. In this optic, the modelling of mixed component EM-NF emission becomes one of the crucial steps before the implementation process. Therefore, the undesired EMC radiations should be investigated not only in frequency-domain but also in time-domain. For this reason, we propose in this chapter an extraction method enabling to determine the time-domain EM-NF maps from the frequency-dependent data by using the Fourier transform of the 2D data.

1.3 Outline of the presented chapter

To make this chapter better to understand, it is organized in three main sections. Section 2 describes the methodology of the time-frequency computation-method proposed. It details how to extract the transient EM-NF radiation from the given time-dependent excitation sampled signal and the frequency-dependent data. Then, more concrete validation of the computation-method investigated by considering the EM-NF radiated by an arbitrary set of magnetic dipoles is devoted in Section 3. The EM-NF reference data are calculated with the theoretical formulas introduced in (Baum 1971 & 1976, Singaraju & Baum 1976). As reviewed by certain research works (Hertz 1892, Chew & Kong 1981, Lakhtakiaa et al. 1987, Song & Chen 1993, Jun-Hong et al 1997, Schantz 2001, Selin 2001, Smagin & Mazalov 2005, Sten & Hujanen 2006, Ravelo 2010), the analytical calculation performed with the EM-wave emitted by elementary dipoles allows to realize more practical and more explicit mathematical analyses of the EM-field expressions in different physic areas. We point out that the EM-field emitted by electronic devices can be modelled by the radiations of the optimized combination of elementary EM-dipoles (Fernández-López et al. 2009). To confirm the feasibility of the method proposed, an application with another proof of concept with a concrete electronic device is also offered in Section 4. This practical verification will be made toward a microwave electronic design of low-pass planar microstrip filter operating up until some GHz. Lastly; Section 5 draws the conclusion of this chapter.

2. Methodology of the time-frequency computation method investigated

The present section is divided in two different parts. First, an explicit description illustrates how to examine the transient excitation signal for the UWB applications. Afterward, the development of the routine process indicating the algorithm of the computation method proposed is elaborated.

2.1 Frequency coefficient extraction

Let us denote $i(t)$ the transient current which is considered also as the excitation of the under test electronic structure. The sampled data corresponding to this test signal is supposed discretized from the starting time t_{min} to the stop time t_{max} with time step equal to Δt. In this case, the number n of time-dependent samples is logically, equal to:

$$n = \mathrm{int}\left(\frac{t_{\max} - t_{\min}}{\Delta t}\right), \tag{1}$$

with int(x) expresses the lowest integer number greater than the real x. Accordingly, via the fast Fourier transform (fft), the equivalent frequency-dependent spectrum of $i(t_k)$ (with t_k = $k.\Delta t$ and k = {1…n}) can be determined. The frequency data emanated by this mathematical transform are generally as a complex number denoted by $\underline{I}(f_k) = fft\left[i(t_k)\right]$. Therefore, the magnitude of this signal spectrum from DC to a certain frequency in function of the initial time-step sampled data can be extracted as depicted in Fig. 2. Consequently, by denoting I_0 the sinusoidal current magnitude for generating the magnetic field spectrum $\underline{H}_0(f)$, the following complex coefficients of the input current in function of frequency as illustrated in Fig. 1:

$$\underline{c}_k = \frac{\underline{I}(k \cdot f_{step})}{\underline{I}_0}. \tag{2}$$

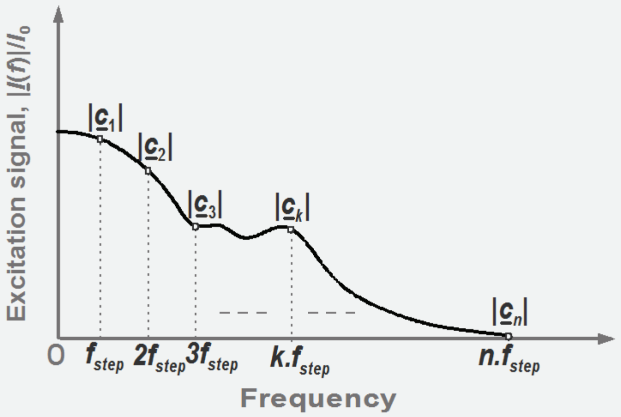


Fig. 1. Extraction of the frequency coefficients from the excitation signal spectrum

It corresponds to the discrete data at each sample of frequency $f_k = k.f_{step}$ for $k = \{1\ldots n\}$ having a step-frequency given by:

$$f_{step} = \frac{1}{t_{\max} - t_{\min}}. \tag{3}$$

We underline that this method requires a frequency range $[f_{min}, f_{max}]$ whose the lowest frequency value f_{min} of $\underline{I}(f)$-data is equal to the step frequency f_{step}. It means that the spectrum value can be extrapolated linearly to generate the excitation signal steady-state component at $f = 0$. In practice, it does not change the calculation results because according to the signal processing theory, the DC-component of transient waves with ultra-short time duration at very low frequencies is usually negligible. The upper frequency f_{max} should correspond to the frequency bandwidth containing 95-% of the excitation signal considered spectrum energy.

2.2 Routine process of the computation-method proposed

The computation-method under study is mainly consisted of two different steps. The first step is focused on the time-domain characterization of the excitation current considered $i(t_k)$ in the specific interval range varying from t_{min} to t_{max} with step Δt. In this first step, the complex frequency-coefficients $\underline{c}_k$ should be extracted through the fft-operation as explained in previous subsection. The following second step is the conversion of the frequency-dependent magnetic- or H-field expressed by $\underline{H}_0(x,y,z_0,f)$ which is recorded at the point $M(x,y,z_0)$ chosen arbitrarily, into a time-dependent data denoted by $H(x,y,z_0,t)$ by using the ifft-operation. These points $M(x,y,z_0)$ belong in the X-Y plane positioned at $z = z_0$. In this case, the frequency range considered varying from f_{min} to f_{max} and the frequency step f_{step} of $H_0(f) = H_0(x,y,z_0,f)$ must be well-synchronized with that of the excitation signal frequency coefficients c_k. Under this condition, the time-dependent data desired $H(t) = H(x,y,z_0,t)$ which is generated by the specific excitation signal $i(t)$ can be calculated with the inverse fast Fourier transform (ifft) of the convolution product between $\underline{c}(f) = \underline{I}(f) / \underline{I}_0$ and $\underline{H}_0(f)$ written as:

$$H(t) = \Re e\{ifft[\underline{c}(f).\underline{H}_0(f)]\}\ . \tag{4}$$

with $\Re e\{x\}$ represents the real part of the complex number x. Fig. 2 depicts the flow work highlighting the different operations to be fulfilled for the achievement of the transient EM-field computation-method proposed. This method enables to provide the time-dependent

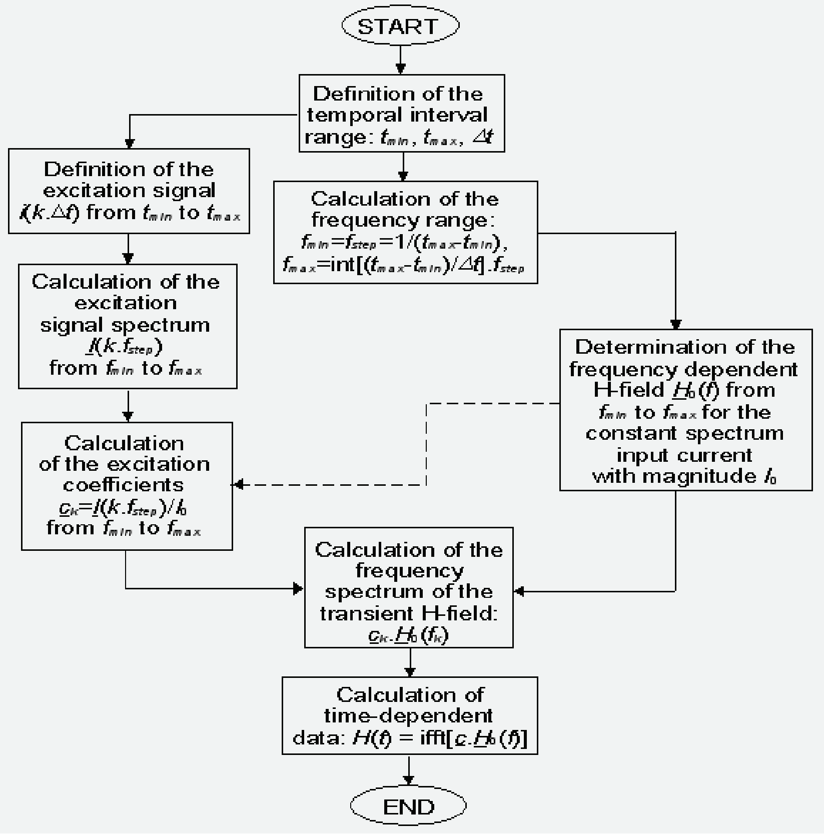

START
Definition of the temporal interval range: t_{min}, t_{max}, Δt
Definition of the excitation signal $i(k.\Delta t)$ from t_{min} to t_{max}
Calculation of the frequency range: $f_{min}=f_{step}=1/(t_{max}-t_{min})$, $f_{max}=\mathrm{int}[(t_{max}-t_{min})/\Delta t].f_{step}$
Calculation of the excitation signal spectrum $\underline{I}(k.f_{step})$ from f_{min} to f_{max}
Determination of the frequency dependent H-field $\underline{H}_0(f)$ from f_{min} to f_{max} for the constant spectrum input current with magnitude I_0
Calculation of the excitation coefficients $\underline{c}_k=\underline{I}(k.f_{step})/I_0$ from f_{min} to f_{max}
Calculation of the frequency spectrum of the transient H-field: $\underline{c}_k.\underline{H}_0(f_k)$
Calculation of time-dependent data: $H(t)$ = ifft[$\underline{c}.\underline{H}_0(f)$]
END

Fig. 2. Flow work illustrating the transient H-field radiation computation-method proposed knowing the temporal range, t_{min} and t_{max} step Δt of the excitation signal $i(t_k)$ and also the frequency-dependent H-field $\underline{H}_0(f)$ (here the under bar indicates the complex variables)

H-field here denoted as $H(t)$ according to the arbitrary form of the excitation and also knowing the frequency-dependent H-field data in the frequency range starting from the lowest value to the upper frequency limit equal to the inverse of the time-step Δt of the discrete data $x(t_k)$.

3. Validation with the radiation of a set of magnetic dipoles

This section presents the verification of the computation method established regarding the radiation of magnetic dipoles shown in Fig. 3 in time-domain. The results were realized by considering the frequency-dependent EM-NF-field directly calculated from predefined analytical formulas (Baum 1971 & 1976, Singaraju & Baum 1976, Balanis 2005). After the analytical description of the considered dipole source and also the mathematical definitions of the H-field expressions in frequency- and time-domains, we will explore the numerical computation with Matlab programming.

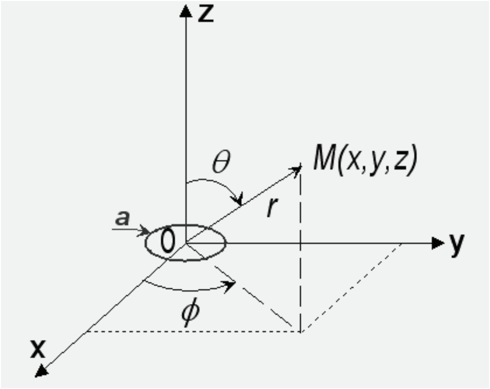


Fig. 3. Configuration of the elementary magnetic dipole formed by a circular loop with radius a placed at the centre $O(0,0,0)$ of the cartesian (x,y,z)-coordinate system

In this figure, the magnetic dipole assumed as a circular wire having radius a is positioned at the origin of the system. Then the considered point M where the EM-field will be evaluated can be referred either in cartesian coordinate (x,y,z) or in spherical coordinate (r,θ,φ). By assuming that the magnetic loop depicted in Fig. 3 is fed by transient current denoted $i(t)$, we have the H-field expressions recalled in appendix A (Baum 1971 & 1976, Singaraju & Baum 1976).

3.1 Description of the radiation source considered

Fig. 4 displays the arbitrary placement representing the set of eight magnetic dipoles in the horizontal plane Oxy considered for the validation of the calculation method under study. To generate the various behaviours of EM-field graphs, the dipole axes are randomly oriented as follows: M_1, M_4, M_5 and M_8 along Ox-axis, M_2 and M_6 along Oy, and M_3 and M_7 along Oz. These dipoles are in this case, flowed by an ultra-short duration transient current $i(t)$ which is considered as a pulse signal. These elementary dipoles are supposed formed by wire circular loops having radius a = 0.5 mm.

Note that in the present study, the elementary dipoles are supposed ideal, thus, there is no coupling between each other. By reason of linearity, the total EM-field at any point $M(x,y,z)$ is consequently the sum of each dipole contribution:

$$\vec{H}(x,y,z) = \sum_{k=1}^{8} \vec{H}_{M_k}(x,y,z) . \tag{5}$$

After implementation of the computation algorithm introduced in Fig. 2 in Matlab program, we obtain the EM-calculation results presented in the next subsection.

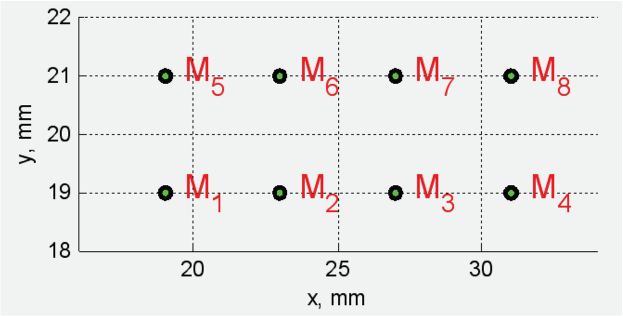


Fig. 4. Configuration of the radiating source considered which is comprised of magnetic dipoles placed in the horizontal plane *xy*

3.2 Validation results

In this subsection, comparison results between the transient EM-field maps radiated by the elementary dipoles displayed in Fig. 4 from the direct calculation and from the method proposed are presented.

3.2.1 Description of the excitation signal

In order to highlight the influence of the form and the transient variation of the disturbing currents in the electronic structure, the considered short-duration pulse excitation current $i(t)$ is assumed as a bi-exponential signal analytically defined in appendix B. One points out that in order to take into account the truncation effect between t_{min} and t_{max}, the considered sampling data from $i(t)$ should be multiplied by a specific time gate. So that accordingly, each component $\underline{I}(\omega)$ should be assumed as a sine cardinal. But here, this effect can be negligible if the assumed time step is well-accurate. The numerical application was made by taking the current amplitude I_M = 1A and the time-constants $\tau_1 = \tau_2/2 = 2$ ns. So, from the analytical relation expressed in (B-4), we have $\omega_{95\%} \approx 3.07$ Grad.s^{-1}. Fig. 5 displays the transient plot of this current excitation.

The time interval range of signal test was defined from t_{min} = 0 ns to t_{max} = 20 ns with step Δt = 0.2 ns. One can see that this baseband signal presents a frequency bandwidth f_{max} of about 2 GHz, where belongs more than 95-% of the spectrum signal energy. The data calculated $\underline{I}(\omega) = fft[i(t)]$ generates the frequency-coefficient values of $i(t)$ according to the relation expressed in (2) as described earlier in subsection 2.2.

3.2.2 Discussions on the computed results

By considering the set of eight magnetic dipoles presented in Fig. 4 which are excited by the same pulse current plotted in Fig. 5 yields the H-field component (H_x, H_y and H_z) mappings depicted in Fig. 6 at the arbitrary time t_0 = 2 ns and in the horizontal plane parallel to (Oxy) referenced z_0 = 6.5 mm above the radiating source.

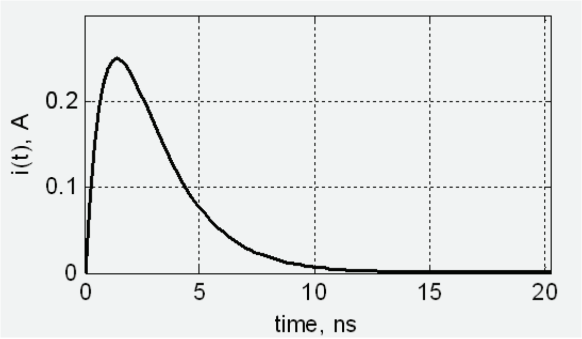


Fig. 5. Transient plot of the considered excitation current $i(t)$

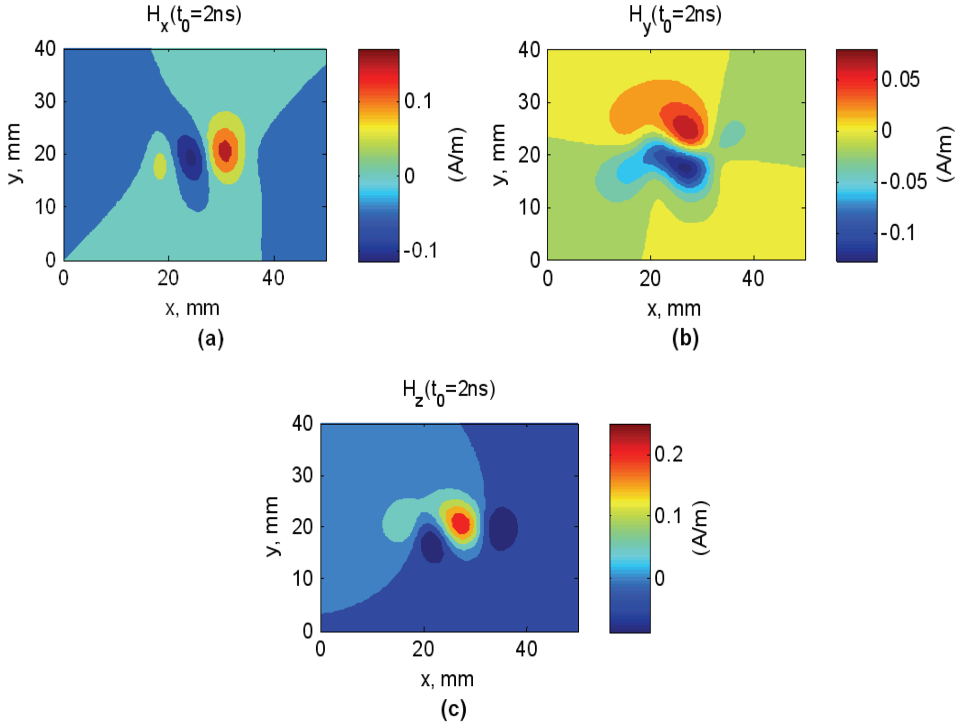


Fig. 6. Maps of H-field components detected at the height z_0 = 6.5 mm above the radiating source: (a) H_x, (b) H_y and (c) H_z obtained from the direct calculation

This height was arbitrarily chosen in order to generate a significant NF effect in the considered frequency range. The dimensions of the mapping plane were set at L_x = 110 mm and L_y = 100 mm with space-resolution equal to $\Delta x = \Delta y$ = 2 mm. First, by using the harmonic expressions of the magnetic field components, the maps of the frequency-dependent H-field are obtained from f_{min} = 0.05 GHz to f_{max} = 2.50 GHz with step Δf = 0.05 GHz. Fig. 7 represents the corresponding mappings of the H-field component magnitudes at f_0 = 2 GHz.

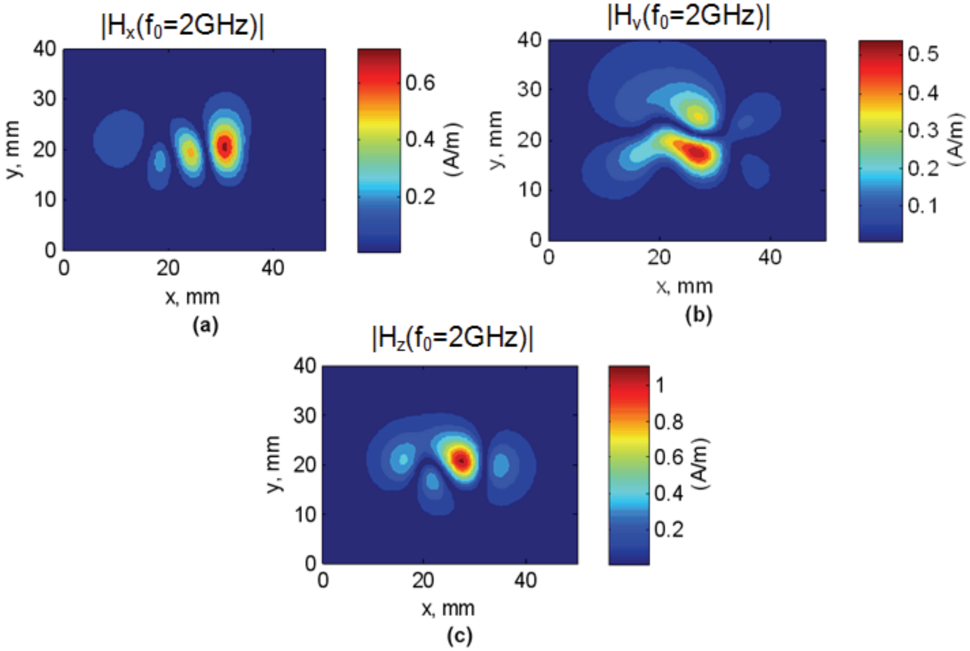


Fig. 7. Maps of H-field components magnitude obtained at the frequency f_0 = 2 GHz: (a) H_x, (b) H_y and (c) H_z directly calculated from expressions (A-8), (A-9) and (A-10)

After the Matlab program implementation of the algorithm indicated by the flow chart schematized in Fig. 2, the results shown in Figs. 8(a)-(d) are obtained via the combination of the frequency-dependent data of the H-field components associated to the frequency-coefficients of the excitation signal plotted in Fig. 5. One can see that the EM-maps presenting the same behaviors as those obtained via the direct calculations displayed in Fig. 6 were established. In addition, we compare also as illustrated in Fig. 9 the modulus of the H-fields from the method under study and from the 3D EM Field Simulator - CST (Computer Simulation Technology). Furthermore, as evidenced by Figs. 10(a)-(c), very good correlation between the profiles of the H-field components detected in the vertical cut-plane along Oy and localized at x = 23 mm was observed. To get further insight about the time-dependent representation of the H-field components, curves showing the variations of $H_x(t)$, $H_y(t)$ and $H_z(t)$ at the arbitrary point chosen of the mapping plane having coordinates (x = 19 mm, y = 35 mm) are plotted in Figs. 11(a)-(c).

As results, once again, we can find that the H-field components from the frequency data fit very well the direct calculated ones. As aforementioned, due to the truncation effects, the H_x-component presents a slight divergence at the ending time of the signal. This is particularly due to the numerical noises at the very low value of the EM field as the case of the x-component which is absolutely twenty times less than the two other components.

In order to prove in more realistic way the relevance of the investigated method, one proposes to treat the radiation of concrete electronic devices in the next section.

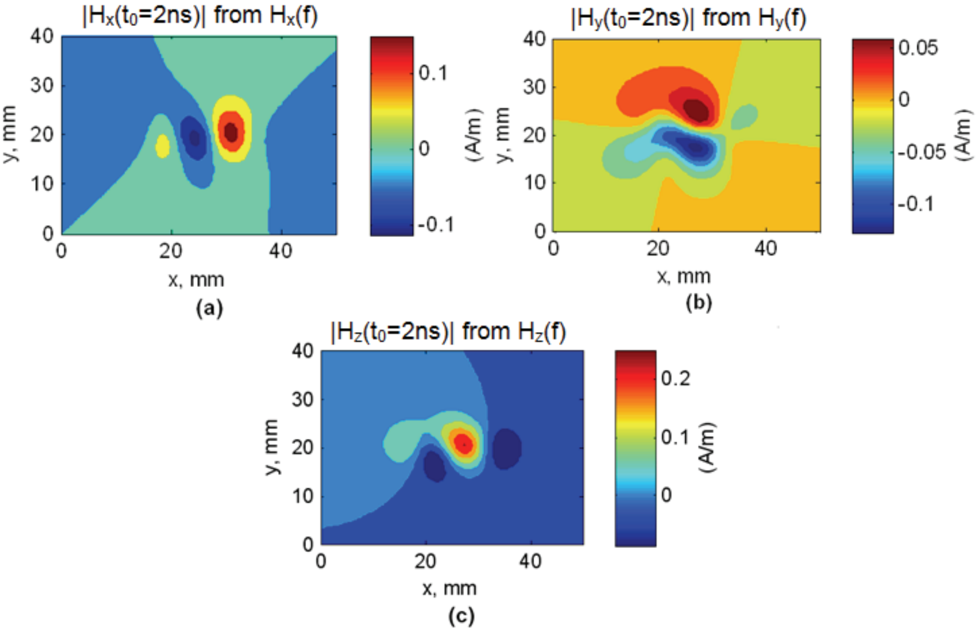


Fig. 8. Maps of H-field components calculated from the time-frequency computation method proposed for $z_0 = 6.5$ mm: (a) H_x, (b) H_y, (c) H_z

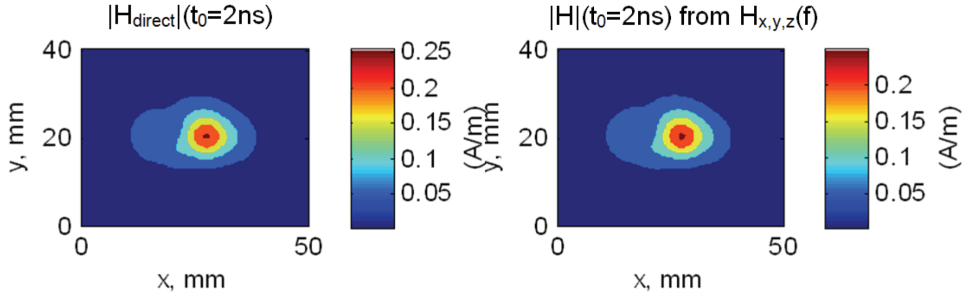


Fig. 9. Comparison of H-field maps modulus $|H|(t_0 = 2\text{ ns})$ obtained from the direct formulae (in left) and from the time-frequency computation method proposed for $z_0 = 6.5$ mm

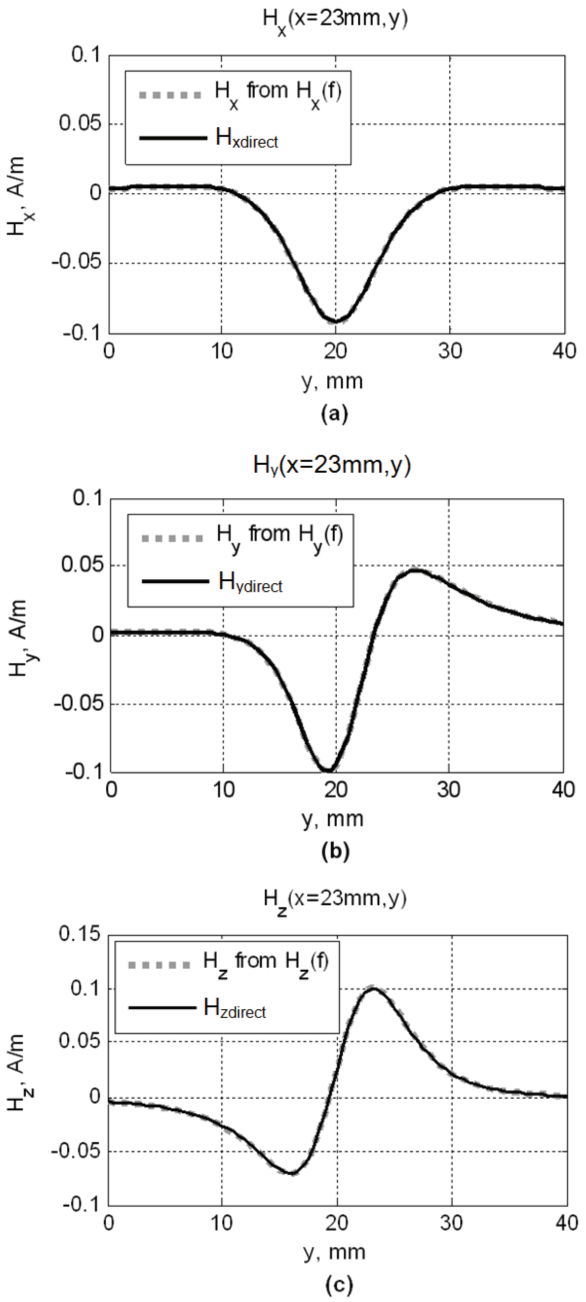


Fig. 10. Comparisons of the H-field component profiles obtained from the time-frequency computation method proposed and the direct calculation, detected in the vertical cut-plane x = 23 mm: (a) H_x, (b) H_y and (c) H_z

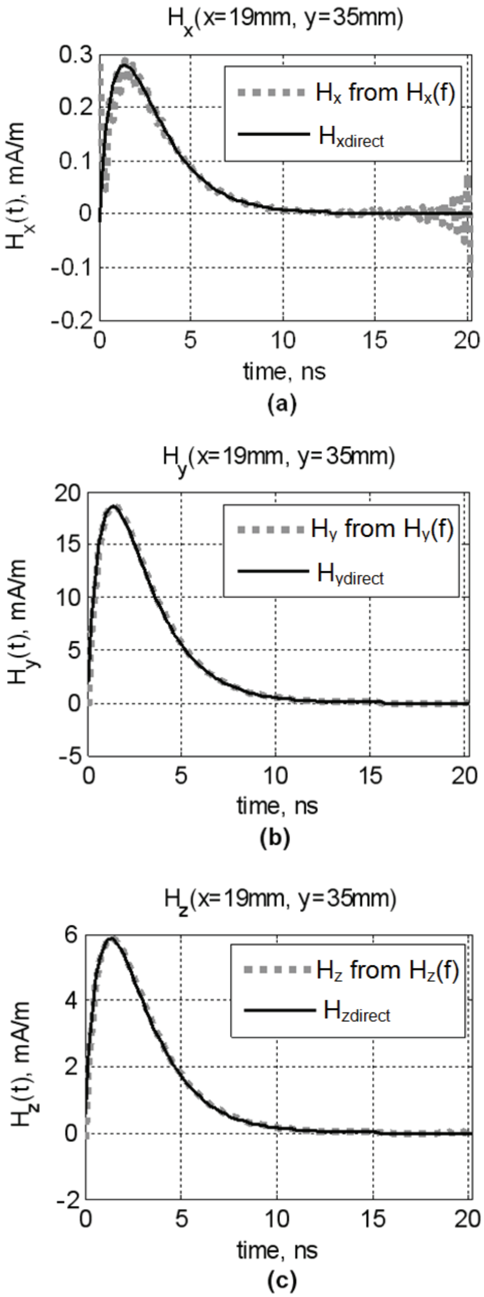
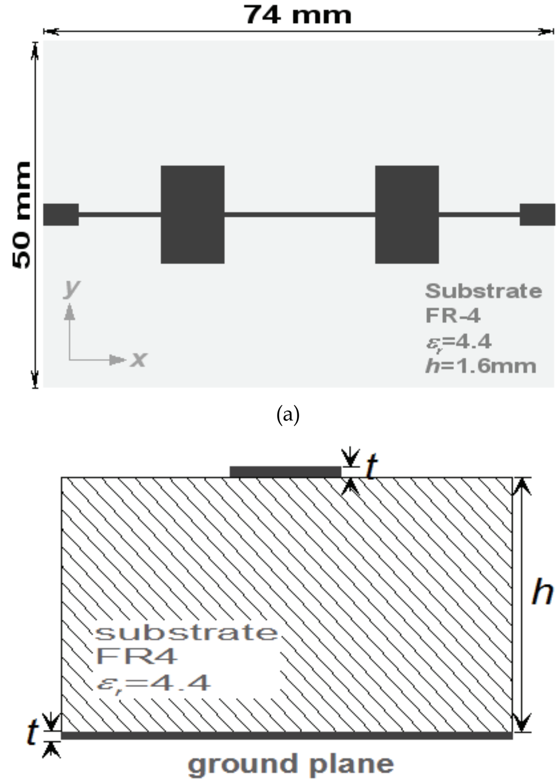


Fig. 11. Comparisons of the H-field components temporal variation obtained from the time-frequency computation method proposed and the direct calculation, detected in the arbitrary point (x = 19 mm, y = 35 mm): (a) H_x, (b) H_y and (c) H_z

4. Application with the Transient NF emitted by a microstrip device proof-of-concept

To get further insight about the feasibility of the computation method under study, let us examine the transient EM-wave emitted by an example of more realistic microwave device. This latter was designed with the standard 3-D EM-tools HFSS for generating the frequency-dependent data which is used for the determination of transient H-NF. Then, the simulation with CST microwave studio (MWS) was performed for the computation of the reference H-NF mappings in time-domain. As realistic and concrete demonstrator, a low-pass Tchebychev filter implemented in planar microstrip technology was designed. Its layout top view including the geometrical dimensions is represented in Fig. 12(a).

This device was printed on the FR4-epoxy substrate having relativity permittivity $\varepsilon_r = 4.4$, thickness $h = 1.6$ mm and etched Cu-metal thickness $t = 35$ µm. The cut cross section of this microwave circuit is pictured in Fig. 12(b).



(a)

(b)

Fig. 12. (a) Top view of the CST design of the under test low-pass microstrip filter (b) Cross-section cut of the under test microstrip filter with metallization thickness $t = 35$ µm and dielectric substrate height $h = 1.6$ mm

After simulations, we realize the results obtained and discussed in next subsections.

4.1 CST-computation results

The low-pass filter under test was simulated with CST MWS in the time interval range delimited by t_{min} = 0 ns and t_{max} = 20 ns with step Δt = 0.2 ns. Therefore, the magnetic-field maps are presented in Fig. 13. These curves are recorded at the arbitrary instant time t_0 = 2 ns. Note that this structure was excited by the input transient current plotted in Fig. 5. It results the graphs of H-field components displayed in Fig. 13.

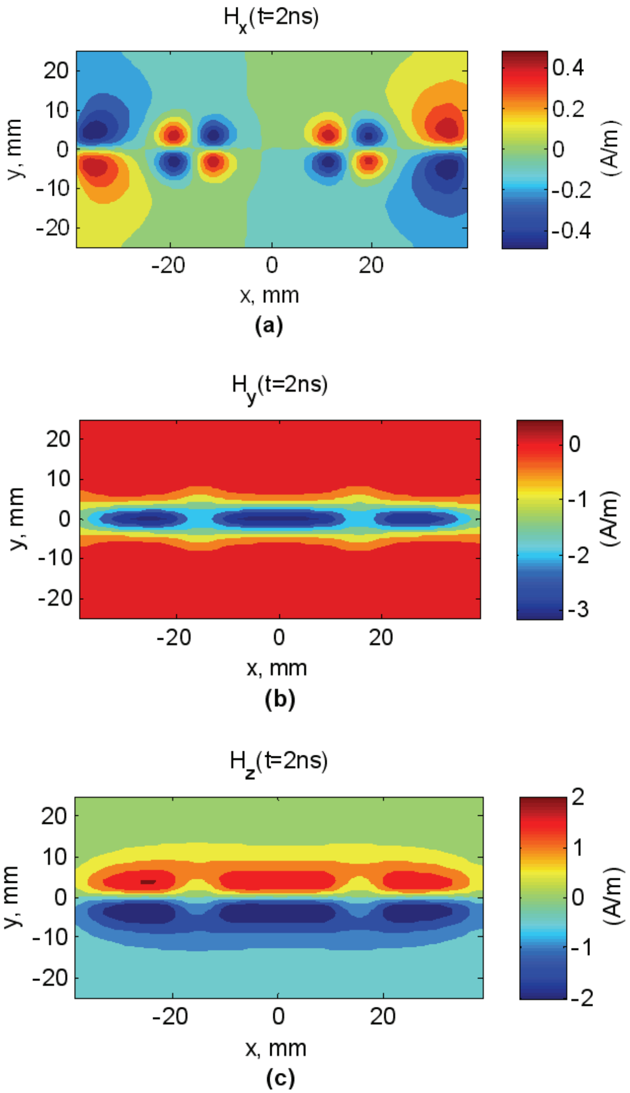


Fig. 13. Maps of transient H-field components detected at t = 2 ns: (a) H_x, (b) H_y and (c) H_z computed from the commercial tool CST

Similar to the previous section, these H-field components were mapped in the horizontal plane placed at the height z_0 = 6 mm above the bottom surface of the considered filter and delimited by -38 mm < x < 38 mm and -24 mm < y < 24 mm with space-step $\Delta x = \Delta y$ = 2 mm. By using the frequency-dependent data convoluted with the excitation test signal, we will present next that we can regenerate these transient H-field maps.

4.2 Analysis of the results obtained from the transient field computation method proposed

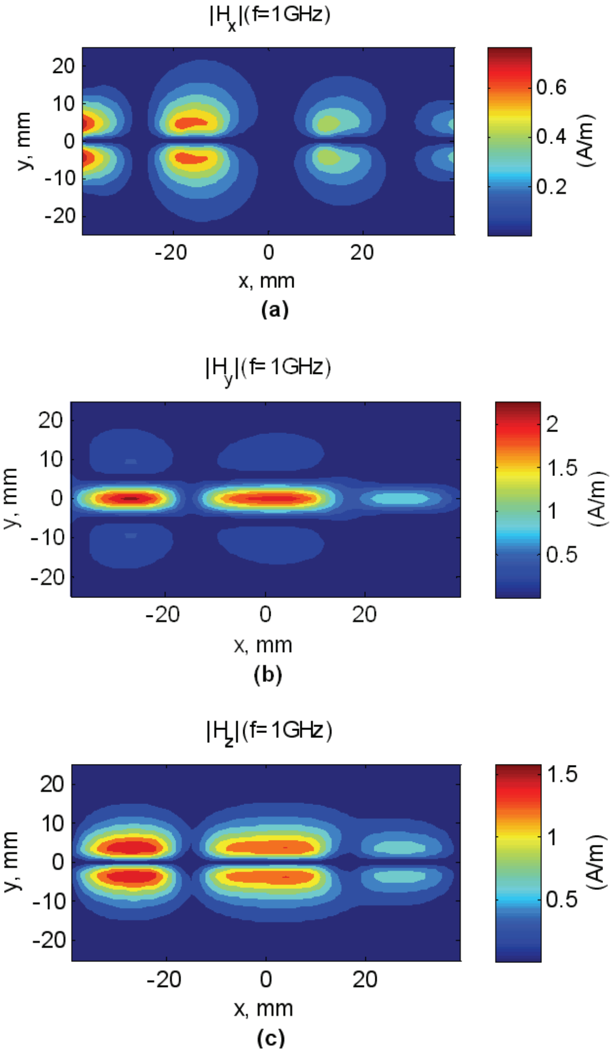
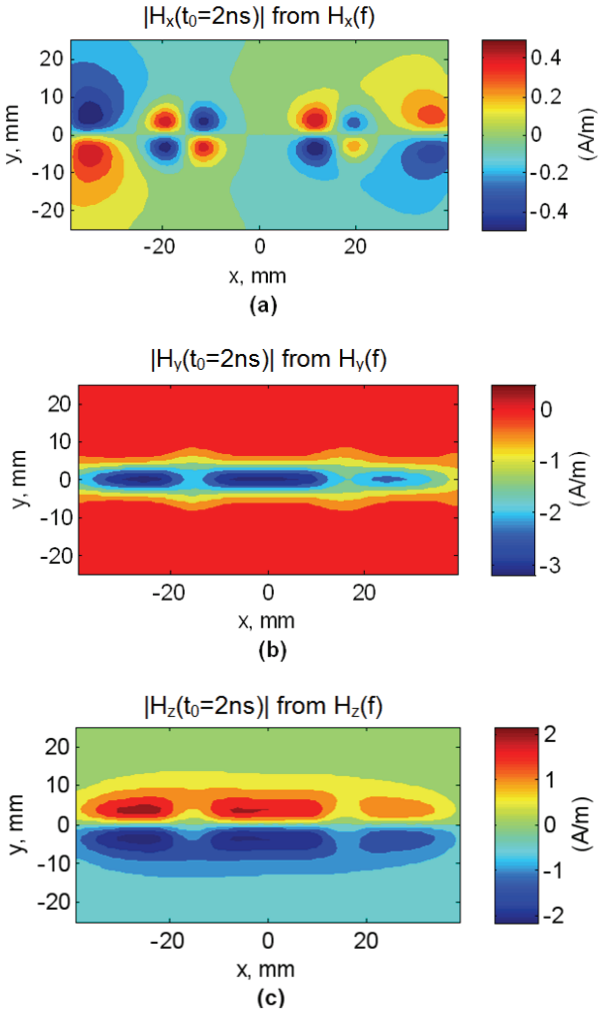


Fig. 14. Maps of frequency-dependent H-field components magnitude: (a) H_x, (b) H_y and (c) H_z computed from HFSS at f = 1 GHz

After HFSS-simulations carried out in the frequency range starting from f_{min} = 0.05 GHz to f_{max} = 2.5 GHz with frequency-step Δf = 0.05 GHz, the maps of H-field component magnitudes displayed in Fig. 14 are recorded. These field components are mapped in the same horizontal plane as in the previous subsection by taking the height z_0 = 6 mm. Among the series of the field maps obtained, we show here the data simulated at the frequency f = 1 GHz. According to the flow work depicted in Fig. 2, the frequency-dependent data $H_x(f)$, $H_y(f)$ and $H_z(f)$ are employed for the determination of the time-dependent data $H_x(t)$, $H_y(t)$ and $H_z(t)$ regarding the transient input current plotted earlier in Fig. 5. So that by application of the computation algorithm under investigation, the H-field maps calculated from Matlab are respectively, depicted in Figs. 15 at the instant time t = 2 ns.



(a)

(b)

(c)

Fig. 15. Maps of H-field components obtained from the proposed method and regarding the simulated frequency-dependent data from HFSS: (a) H_x, (b) H_y and (c) H_z

Despite the slight difference of H_x-maps, we observe that the maps are perfectly well-correlated to those introduced in Fig. 13 of subsection 4.1. For the further smart illustration of the results correlation, comparisons of H-field profiles calculated with the method proposed (grey curves) and those from CST (black curves) for x = -5 mm are plotted Fig. 16.

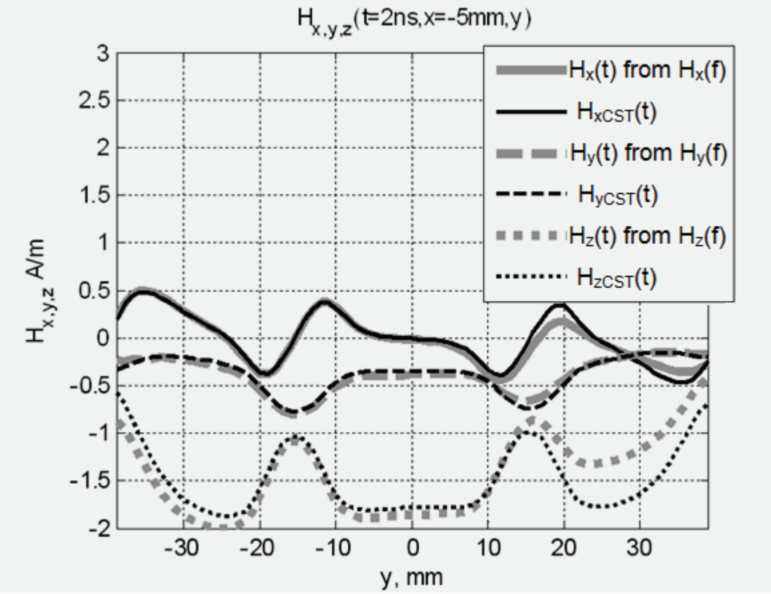


Fig. 16. Comparisons between *Oy*-profiles of the H-field components computed with CST software and those obtained from the proposed method

The imperfection of the results presented here are due to the numerical errors mainly caused by the solver and the meshing inaccuracies. Note that one evaluates relative errors of about 10 % for $|H_x|$, $|H_y|$ and $|H_z|$. Despite the apparent difference between the results from the method under investigation and the commercial EM-tool CST-computations, once again, very good correlations between the profiles of the H-field components are realized by considering the data recorded in the vertical cut-plane equated by x = -5 mm as explained in Fig. 16.

In nutshell, the computation results exposed in this paper reveal the effectiveness and the operability of the method developed for the case of elementary magnetic dipoles and also by considering the NF EM-radiation of realistic use case electronic devices.

5. Concluding remarks

A computation method of transient NF EM-field radiated by electronic devices excited by a complex wave or ultra-short duration transient signal is stated in this chapter. In addition to the evanescent wave integration, the originality of the NF calculation method developed lies on the consideration of the radiation deeming the UWB structures which is literally from DC to microwave frequency ranges. It is based on the convolution of the frequency-dependent

EM-wave data with a transient excitation pulse current. A methodological analysis was made by taking into account a complex waveform of the transient pulse signal exciting the radiation source structure considered. It was explained how the frequency-bandwidth of the frequency-dependent baseband EM-field must be chosen according to the excitation current considered.

In order to demonstrate the relevance of the method investigated, it was first, implemented into Matlab program and then, applied to the determination of the H-field radiated by a microwave circuit in UWB. As consequence, the feasibility of the method was verified with two types of structures. First, with the semi-analytical calculation implemented in Matlab by considering the frequency- and time-dependent expressions of the magnetic NF radiated by a set of magnetic dipoles, an excellent agreement with the results from the calculation method developed were found. Then, further more practical analysis was performed with the determination of transient H-NF from the frequency-dependent data computed with a standard commercial 3-D EM-tool. For this second test, the H-NF emitted by a low-pass planar microstrip filter was treated. For both cases, the excitation current injected to the structures was assumed as an ultra-short transient pulse having half-bandwidth lower than 5 ns which presents a baseband frequency spectrum with bandwidth of about 2.5 GHz from DC. With the examples of complex structures tested, very good agreement between the transient H-field component maps and profiles was realized from the method proposed and those directly calculated from the well-known standard tools and from classical mathematical EM-formulae.

It is interesting to point up that the NF computation method introduced in this chapter is advantageous in terms of:

1. Simplicity of the EM-field maps determination for any waveform of transient excitation even with ultra-short duration which is very hard to simulate with most of commercial simulation tools. The method developed can be used for the determination of the NF maps in time-domain which is practically very difficult to measure in the realistic contexts.
2. It is flexible for various types of excitation signals which can be expressed analytically and also from the realistic use case of disturbing signal generally met in EMC area (Wiles 2003, Liu, K. 2011, Hubing 2011).
3. It can be adapted also to different forms of electrical and electronic structures for low- and high-frequency applications. Globally speaking, it offers a possibility to work in UWB from DC to unlimited upper frequency limit.
4. One can achieve significant EM-field measurement in very short time-duration with base band measured data in wide bandwidth.

However, its main drawback is the limitation in term of time step which depends on the frequency range of the initial frequency-data considered and also the necessity of powerful computer for the achievement of high accurate results.

In the next step of this work, we plane to extend this method to transpose in time-domain the modelling of EM-radiation with the optimized association of elementary dipoles (Vives-Gilabert et al. 2009, Fernández-López et al. 2009). Then, we are hopeful that the method developed in this chapter is very helpful for EMC/EMI investigations of modern electrical/electronic systems as the case of hybrid vehicle embedded circuits (Vye 2011, Tröscher 2011, Kopp 2011) where the transient NF effects are susceptible to disturb the system functioning.

6. Appendix

This appendix contains two parts of theoretical parts concerning the transient NF radiated by the elementary magnetic dipoles (Baum 1971 & 1976, Singaraju & Baum 1976, Ravelo et al. 2011a & 2011b, Lui Y. et al. 2011a & 2011b) and the bi-exponential signal processing.

6.1 Appendix A: Analytical study of the magnetic dipole radiation in time-domain

By definition, the magnetic dipole moment of the elementary circular loop shown earlier in Fig. 3 (see section 3) is written as:

$$\vec{p}_M(r,t) = p(t) \cdot \delta(r) \cdot \vec{u}_z \,, \tag{A-1}$$

with *r* is the distance between the dipole centre and the point *M*(*r*,*θ*,*φ*) as shown in Fig. 3. By analogy with the definition of the time-variant vector established by Hertz in 1892 (Hertz 1892), the H-field components in the spherical coordinate system:

$$\vec{H} = H_r(r,\theta,\varphi,t)\vec{u}_r + H_\theta(r,\theta,\varphi,t)\vec{u}_\theta + H_\varphi(r,\theta,\varphi,t)\vec{u}_\varphi \,, \tag{A-2}$$

are expressed as:

$$H_r(r,\theta,\varphi,t) = \frac{\cos(\theta)}{2\pi r^2}\left[\frac{p(\tau)}{r} + \frac{1}{v}\frac{\partial p(\tau)}{\partial t}\right], \tag{A-3}$$

$$H_\theta(r,\theta,\varphi,t) = \frac{\sin(\theta)}{4\pi r^2}\left[\frac{p(\tau)}{r} + \frac{1}{v}\frac{\partial p(\tau)}{\partial t} + \frac{r}{v^2}\frac{\partial^2 p(\tau)}{\partial t^2}\right], \tag{A-4}$$

$$H_\varphi(r,\theta,\varphi,t) = 0 \,, \tag{A-5}$$

where v is the wave-velocity, and τ is the time delayed variable which is defined as $\tau = t - r/v$. One underlines that the magnetic dipole is also an Hertzian dipole so that $\partial i(r,t)/\partial r = 0$. In the frequency domain, the spherical coordinate of the H-field component formulas radiated by the magnetic dipole pictured in Fig. 3 which is supposed flowed by an harmonic current with amplitude I_M denoted:

$$\underline{I}(f) = I_M e^{-j2\pi ft} = I_M e^{-j\omega t} \,, \tag{A-6}$$

are written as (Balanis 2005):

$$H_r(r,\theta,\varphi,f) = \frac{I_M a^2}{2r^3}(1 + jkr)\cos(\theta)e^{-jkr} \,, \tag{A-8}$$

$$H_\theta(r,\theta,\varphi,f) = \frac{I_M a^2 \sin(\theta)}{4r^3}(1 + jkr - k^2r^2)e^{-jkr} \,, \tag{A-9}$$

$$H_{\varphi}(r,\theta,\varphi,f)=0\ , \tag{A-10}$$

where *j* is the complex number $\sqrt{-1}$ and the real $k=2\pi f\ /\ v$ expresses the wave number at the considered frequency *f*. Then, through the classical relationship between the spherical and cartesian coordinate systems, one can determine easily the expressions of the components H_x, H_y and H_z.

6.2 Appendix B: Spectrum analysis of bi-exponential signal

A bi-exponential form signal with parameters τ_1 and τ_2 is analytically expressed as:

$$i(t)=I_M(e^{-t/\tau_1}-e^{-t/\tau_2})\ . \tag{B-1}$$

The analytical Fourier transform expression of this current is written as:

$$\underline{I}(\omega)=I_M(\frac{\tau_1}{1+j\omega\tau_1}-\frac{\tau_2}{1+j\omega\tau_2})\ , \tag{B-2}$$

with ω is the angular frequency. This yields the signal frequency spectrum formulation expressed as:

$$\left|\underline{I}(\omega)\right|=I_M\frac{\left|\tau_1-\tau_2\right|}{\sqrt{(1+\tau_1^2\omega^2)(1+\tau_2^2\omega^2)}}\ . \tag{B-3}$$

To achieve at least 95-% of excitation signal spectrum energy, the frequency-data should be recorded in baseband frequency range with angular frequency bandwidth equal to:

$$\omega_{95\%}=\frac{\sqrt{\sqrt{(\tau_1^2-\tau_2^2)^2+1600\tau_1^2\tau_2^2}-\tau_1^2-\tau_2^2}}{\sqrt{2}\tau_1\tau_2}\ . \tag{B-4}$$

7. Acknowledgment

Acknowledgement is made to EU (European Union) and Upper Normandy region for the support of these researches. These works have been implemented within the frame of the "Time Domain Electromagnetic Characterisation and Simulation for EMC" (TECS) project No 4081 which is part-funded by the Upper Normandy Region and the ERDF via the Franco-British Interreg IVA programme.

8. References

Adada, M. (2007). High-Frequency Simulation Technologies-Focused on Specific High-Frequency Design Applications. *Microwave Engineering Europe*, (Jun. 2007), pp. 16-17

Agilent EEsof EDA. (2008). *Overview: Electromagnetic Design System (EMDS)*, (Sep. 2008) [Online]. Available from: http://www.agilent.com/find/eesof-emds

Ansoft corporation. (2006). *Simulation Software: High-performance Signal and Power Integrity*, Internal Report

ANSYS, (2009). *Unparalleled Advancements in Signal- and Power-Integrity, Electromagnetic Compatibility Testing*, (Jun. 16 2009) [Online]. Available from: http://investors.ansys.com/

Archambeault, B.; Brench, C. & Connor, S. (2010). Review of Printed-Circuit-Board Level EMI/EMC Issues and Tools. *IEEE Trans. EMC*, (May 2010), Vol. 52, No.2, pp. 455-461, ISSN 0018-9375

Balanis, C. A. (2005). *Antenna Theory: Analysis and Design*, in Wiley, (3rd Ed.), 207–208, New York, USA, ISBN: 978-0-471-66782-7

Barriere, P.-A.; Laurin, J.-J. & Goussard, Y. (2009). Mapping of Equivalent Currents on High-Speed Digital Printed Circuit Boards Based on Near-Field Measurements. *IEEE Trans. EMC*, (Aug. 2009), Vol.51, No.3, pp. 649 - 658, ISSN 0018-9375

Baudry, D.; Arcambal, C.; Louis, A.; Mazari, B. & Eudeline, P. (2007). Applications of the Near-Field Techniques in EMC Investigations. *IEEE Trans. EMC*, (Aug. 2007), Vol.49, No.3, pp. 485-493, ISSN 0018-9375

Baum, C. E. (1971). *Some Characteristics of Electric and Magnetic Dipole Antennas for Radiating Transient Pulses*, Sensor and Simulation Note 405, 23 Jan. 71

Baum, C. E. (1976). Emerging Technology for Transient and Broad-Band Analysis and Synthesis of Antennas and Scaterrers, Interaction Note 300, *Proceedings of IEEE*, (Nov. 1976), pp. 1598-1616

Braun, S.; Gülten, E.; Frech, A. & Russer, P. (2009). Automated Measurement of Intermittent Signals using a Time-Domain EMI Measurement System, *Proceedings of IEEE Int. Symp. EMC*, pp. 232-235, ISBN 978-1-4244-4266-9, Austin, Texas (USA), Aug. 17-21 2009

Chew, W. C. & Kong, J. A. (1981). Electromagnetic field of a dipole on a two-layer earth. *Geophysics*, (Mar. 1981), Vol. 46, No. 3, pp. 309-315

Cicchetti, R. (1991). Transient Analysis of Radiated Field from Electric Dipoles and Microstrip Lines. *IEEE Trans. Ant. Prop.*, (Jul. 1991), Vol.39, No.7, pp. 910-918, ISSN 0018-926X

Edwards, R. S.; Marvin, A. C. & Porter, S. J. (2010). Uncertainty Analyses in the Finite-Difference Time-Domain Method. *IEEE Trans. EMC*, (Feb. 2010), Vol.52, No.1, pp. 155-163, ISSN 0018-9375

Fernández-López, P.; Arcambal, C.; Baudry, D.; Verdeyme, S. & Mazari, B. (2009). Radiation Modeling and Electromagnetic Simulation of an Active Circuit, *Proceedings of EMC Compo 09*, Toulouse, France, Nov. 17-19 2009

Hertz, H. R. (1892). Untersuchungen ueber die Ausbreitung der Elektrischen Kraft (in German). *Johann Ambrosius Barth*, Leipzig, Germany, ISBN-10: 1142281167/ISBN-13: 978-1142281168

Hubing, T. (2011). Ensuring the Electromagnetic Compatibility of Safety Critical Automotive Systems. *Invited Plenary Speaker at the 2011 APEMC*, Jeju, South-Korea, May 2011

Jauregui, R.; Pous, M.; Fernández, M. & Silva, F. (2010). Transient Perturbation Analysis in Digital Radio, *Proceedings of IEEE Int. Symp. EMC*, pp. 263-268, ISBN 978-1-4244-6307-7, Fort Lauderdale, Florida (USA), Jul. 25-30 2010

Jauregui, R.; Riu, P. I. & Silva, F. (2010). Transient FDTD Simulation Validation, *Proceedings of IEEE Int. Symp. EMC*, pp. 257-262, ISBN 978-1-4244-6305-3, Fort Lauderdale, Florida (USA), Jul. 25-30 2010

Jun-Hong, W.; Lang, J. & Shui-Sheng, J. (1997). Optimization of the Dipole Shapes for Maximum Peak Values of the Radiating Pulse, *Proceedings of IEEE Tran. Ant. Prop. Society Int. Symp.*, Vol.1, pp. 526-529, Montreal, Que., Canada, 13-18 Jul 1997, ISBN 0-7803-4178-3

Kopp, M. (2011). Automotive EMI/EMC Simulation. *Microwave Journal*, (Jul. 2011), Vol.54, No.7, pp. 24-32

Lakhtakiaa, A.; Varadana, V. K. & Varadana, V. V. (1987). Time-Harmonic and Time-Dependent Radiation by Bifractal Dipole Arrays. *Int. J. Electronics*, (Dec. 1987), Vol.63, No.6, pp. 819-824, DOI:10.1080/00207218708939187

Liu, K. (2011). An Update on Automotive EMC Testing. *Microwave Journal*, (Jul. 2011), Vol.54, No.7, pp. 40-46

Liu, L.; Cui, X. & Qi, L.. (2009). Simulation of Electromagnetic Transients of the Bus Bar in Substation by the Time-Domain Finite-Element Method. *IEEE Trans. EMC*, (Nov. 2009), Vol.51, No.4, pp. 1017-1025, ISSN 0018-9375

Liu, Y.; Ravelo, B.; Jastrzebski, A. K. & Ben Hadj Slama, J. (2011). Calculation of the Time Domain z-Component of the EM-Near-Field from the x- and y-Components, *Accepted for communication in EuMC 2011*, Manchester, UK, Oct. 9-14 2011

Liu, Y.; Ravelo, B.; Jastrzebski, A. K. & Ben Hadj Slama, J. (2011). Computational Method of Extraction of the 3D E-Field from the 2D H-Near-Field using PWS Transform, *Accepted for communication in EMC Europe 2011*, York, UK, Sep. 26-30 2011

North East Systems Associates (NESA), (2010). *RJ45 Interconnect Signal Integrity, (2010 CST Computer Simulation Technology AG.)* [Online]. Available from: http://www.cst.com/Content/Applications/Article/Article.aspx?id=243

Ordas, T.; Lisart, M.; Sicard, E.; Maurine, P. & Torres, L. (2009). Near-Field Mapping System to Scan in Time Domain the Magnetic Emissions of Integrated Circuits, *Proceedings of PATMOS' 08: Int. Workshop on Power and Timing Modeling Optimization and Simulation*, Ver. 1-11, Lisbon, Portugal, Sep. 10-12 2008, ISBN 978-3-540-95947-2

Rammal, R.; Lalande, M.; Martinod, E.; Feix, N.; Jouvet, M.; Andrieu, J. & Jecko, B. (2009). Far Field Reconstruction from Transient Near-Field Measurement Using Cylindrical Modal Development, *Int. J. Ant. Prop.*, Hindawi, Vol. 2009, Article ID 798473, 7 pages, doi:10.1155/2009/798473

Ravelo, B. (2010). E-Field Extraction from H-Near-Field in Time-Domain by using PWS Method. *PIER B Journal*, Vol.25, pp. 171-189, doi:10.2528

Ravelo, B.; Liu, Y.; Louis, A. & Jastrzebski, A. K. (2011). Study of high-frequency electromagnetic transients radiated by electric dipoles in near-field. *IET Microw. Antennas Propag.*, (Apr. 2011), Vol. 5, No, 6, pp 692 - 698, ISSN 1751-8725

Ravelo, B.; Liu, Y. & Slama, J. B. H. (2011). Time-Domain Planar Near-Field/Near-Field Transforms with PWS Method. *Eur. Phys. J. Appl. Phys. (EPJAP)*, (Feb. 2011), Vol.53, No.1, 30701-pp. 1-8, doi: 10.1051/epjap/2011100447

Rioult, J.; Seetharamdoo, D. & Heddebaut, M. (2009). Novel Electromagnetic Field Measuring Instrument with Real-Time Visualization, *Proceedings of IEEE Int. Symp. EMC*, pp. 133-138, ISBN 978-1-4244-4266-9, Austin, Texas (USA), Aug. 17-21 2009

Schantz, H. G. (2001). Electromagnetic Energy around Hertzian Dipoles. *IEEE Tran. Ant. Prop. Magazine*, (Apr. 2001), Vol.43, No.2, pp. 50-62, ISBN 9780470688625

Selin, V. I. (2001). Asymptotics of the Electromagnetic Field Generated by a Point Source in a Layered Medium. *Computational Mathematics and Mathematical Physics*, Vol.41, No.6, pp. 915-939, ISSN 0965-5425

Shi, J.; Cracraft, M. A.; Zhang J. & DuBroff, R. E. (1989). Using Near-Field Scanning to Predict Radiated Fields, *Proceedings of IEEE Ant. Prop. Int. Symp.*, Vol.3, pp. 1477-1480, San Jose, CA (USA)

Singaraju, B. K. & Baum, C. E. (1976). *A Simple Technique for Obtaining the Near Fields of Electric Dipole Antennas from Their Far Fields*, Sensor and Simulation Note 213, Mar. 76

Smagin, S. I. & Mazalov, V. N. (2005). Calculation of the Electromagnetic Fields of Dipole Sources in Layered Media. *Doklady Physics*, (Apr. 2005), Vol.50, No.4, pp. 178-183, DOI:10.1134/1.1922556

Song, J. & Chen, K.-M. (1993). Propagation of EM Pulses Excited by an Electric Dipole in a Conducting Medium. *IEEE Tran. Ant. Prop.*, (Oct. 1993), Vol.41, No.10, pp. 1414-1421, ISSN 0018-926X

Song, Z.; Donglin, S.; Duval, F.; Louis, A. & Fei, D. (2010). A Novel Electromagnetic Radiated Emission Source Identification Methodology. *Proceedings of Asia-Pacific Symposium on EMC*, Pekin (China), Apr. 12-16 2010, ISBN 978-1-4244-5621-5

Sten, J. C.-E. & Hujanen, A. (2006). Aspects on the Phase Delay and Phase Velocity in the Electromagnetic Near-Field. *PIER Journal*, Vol. 56, pp. 67-80, doi:10.2528

Tröscher, M. (2011). 3D EMC/EMI Simulation of Automotive Multimedia Systems. *Microwave Journal*, (Jul. 2011), Vol.54, No.7, pp. 34-38

Vives-Gilabert, Y.; Arcambal, C.; Louis, A.; Daran, F.; Eudeline, P. & Mazari, B. (2007). Modeling Magnetic Radiations of Electronic Circuits using Near-Field Scanning Method. *IEEE Tran. EMC*, (May 2007), Vol.49, No.2, pp. 391-400, ISSN 0018-9375

Vives-Gilabert, Y.; Arcambal, C.; Louis, A.; Eudeline, P. & Mazari, B. (2009). Modeling Magnetic Emissions Combining Image Processing and an Optimization Algorithm. *IEEE Tran. EMC*, (Nov. 2009), Vol.51, No.4, pp. 909-918, ISSN 0018-9375

Vye, D. (2011). EMI by the Dashboard Light. *Microwave Journal*, (Jul. 2011), Vol.54, No.7, pp. 20-23

Wiles, M. (2003). An Overview of Automotive EMC Testing Facilities, *Proceedings of Automotive EMC Conf. 2003*, Milton Keynes, UK, Nov. 6 2003

Winter, W. & Herbrig, M. (2009). Time Domain Measurement in Automotive Applications, *Proceedings of IEEE Int. Symp. EMC*, pp. 109-115, ISBN 9781424442669, Austin, Texas (USA), Aug. 17-21 2009

Xie, L. & Lei, Y. (2009). Transient Response of a Multiconductor Transmission Line With Nonlinear Terminations Excited by an Electric Dipole. *IEEE Trans. EMC*, (Aug. 2009), Vol.51, No.3, pp. 805-810, ISSN 0018-9375

Yang, T.; Bayram, Y. & Volakis, J. L. (2010). Hybrid Analysis of Electromagnetic Interference Effects on Microwave Active Circuits Within Cavity Enclosures. *IEEE Trans. EMC*, (Aug. 2010), Vol.52, No.3, pp. 745-748, ISSN 0018-9375

Impulse-Regime Analysis of Novel Optically-Inspired Phenomena at Microwaves

J. Sebastian Gomez-Diaz[1], Alejandro Alvarez-Melcon[1],
Shulabh Gupta[2] and Christophe Caloz[2]
[1]*Universidad Politécnica de Cartagena*
[2]*École Polytechnique de Montréal*
[1]*Spain*
[2]*Canada*

1. Introduction

The ever increasing of needs for high data-rate wireless system is currently producing a shift from narrow-band radio towards ultra-wideband (UWB) radio operation [see Ghavami et al. (2007)]. Novel microwave tools, concepts, phenomena and direct applications must be developed to meet this demand. While the past decades have been focused on the "magnitude engineering" and filter design [see Pozar (2005)], there is a renewed interest in the "dispersion engineering". In the dispersion engineering approach, the phase is engineered to met various specifications within a given frequency range, so as to process signals in real time.

In this context, the development of electromagnetic metamaterials over the last years [see Caloz & Itoh (2006) or Marques et al. (2008)], with their intrinsically dispersive nature and subsequent impulse-regime properties, may provide novel and original solutions (see Fig. 1). Metamaterials can easily be synthesized in planar technology under the form of composite right/left-handed (CRLH) transmission lines (TLs), using non-resonant [see Caloz & Itoh (2006)] or resonant [see Duran-Sindreu et al. (2009)] approaches. These structures have provided novel and exciting applications, such as multi-band components, diplexers, couplers, phase-shifters, power-dividers or antennas with enhanced features, to mention just a few [see Caloz (2009) or Eleftheriades (2009) for a recent review]. However, CRLH structures have mostly been analyzed in the harmonic regime to date, and therefore only a few impulse-regime components and systems have been proposed so far. An example of these applications is the tunable pulse delay line presented in in Abielmona et al. (2007).

In this chapter, we present recent advances based on Fourier transformation techniques to model dispersive UWB phenomena and far-field radiation from complex CRLH structures. Section 2 first employs inverse Fourier transforms to study pulse propagation along this type of medium. Then, a Fourier transform approach is applied to the current which flows along the CRLH line, accurately retrieving the time-domain far-field radiation of the structure [which behaves as a leaky-wave antenna, (LWA)]. The main advantages of the proposed techniques are the easy treatment of complex CRLH structures, a deep insight into the physics of the phenomena, and an accurate and a fast computation, which avoids the time-consuming analysis required by completely numerical simulations.

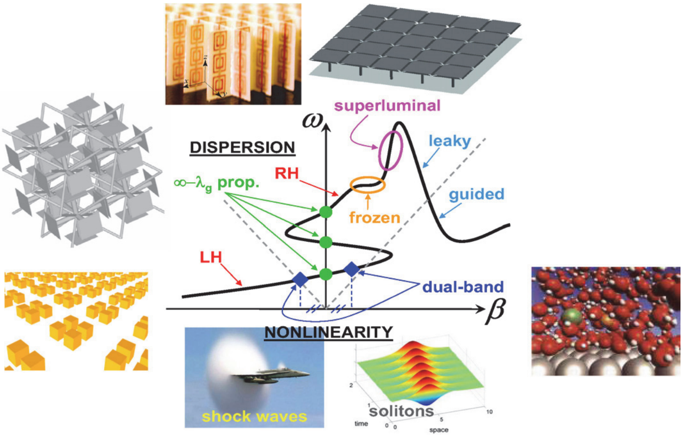


Fig. 1. Illustration of the dispersion engineering concept using metamaterial and CRLH structures. Reprinted with permission from Abielmona et al. (2008). Copyright 2008, URSI.

Section 3 applies the previously derived theory to study the impulse-regime phenomenology of CRLH structures, and subsequently demonstrates several optically-inspired phenomena and applications at microwaves in both the guided and the radiative regime. This study is divided into two main groups, according to the *guided-wave* or *radiative-wave* natures of the proposed phenomena and applications.

In the guided-regime, the dispersive properties of pulse propagation along a CRLH line and the temporal Talbot effect, which allows the period rate multiplication of an input pulse train [see Gómez-Díaz et al. (2009b)], will be described and analyzed.

In the radiative regime, the spectral-spatial decomposition property of CRLH LWAs will be exploited to characterize both, in time and in frequency, unknown fast-transient input signals. For this purpose, two different systems (namely, a real time spectrogram analyzer (RTSA) [see Gupta, Abielmona & Caloz (2009)], and a frequency resolved electrical gating (FREG) [see Gupta, Gómez-Díaz & Caloz (2009)]) will be carefully studied and simulated.

In all cases, the proposed theory and phenomena will be validated by using simulation results from full-wave commercial softwares and measurements, from fabricated prototypes. Therefore, the usefulness of the Fourier transform approach will be fully demonstrated as an essential mathematical tool for the fast and accurate modeling of very complex UWB structures and phenomena.

2. Impulse-regime analysis of CRLH structures using a Fourier transform approach

Propagation of electromagnetic short pulses in complex media, see Felsen (1969), has been a field of great interest for a long time. Most practical developments for pulse propagation in dispersive media have been carried out for optical systems, including optical fibers, couplers, switches and soliton devices [see Saleh & Teich (2007)]. At microwaves, pulse propagation has been far less studied, and the temporal analysis of dispersive metamaterial structures is usually performed with time-domain full-wave methods. However, these accurate techniques require a high computational cost, due to the meshing of the whole geometry under study.

In this section, a general time-domain Green's function approach is presented for the analysis of pulse propagation in electrically thin CRLH TLs. This method is based on the transient analysis of 1D transmission lines [see Paul (2007)] combined for the first time with CRLH TL concepts introduced in Caloz & Itoh (2006). With this equivalent transmission line simplification of the geometry, the Green's functions are available in closed-form, and directly correspond to the voltages and currents along the transmission line. The main advantages of this approach are the unconditional stability and fast computation, due to the continuous treatment of time, and the insight into the physical phenomena provided by the Green's functions. Subsequently, the method is extended to analyze impulse-regime CRLH leaky-wave antennas. The approach is based on the use of the time-domain current which flows along the structure to compute the far-field radiation of the antenna. This technique is especially appropriate to characterize complex radiated-wave UWB phenomena and devices, as it will be demonstrated in Section 3.

2.1 Composite right/left-handed structures

The introduction of metamaterials [see Caloz & Itoh (2006) or Marques et al. (2008)] in the last decade has paved the road to the development of new devices and applications based on the novel fundamental features and phenomena associated to this type of media. Among the most useful metamaterials, one can find the CRLH transmission lines [see Caloz & Itoh (2006)]. This type of transmission lines, which are inherently nonresonant and low-loss, can be easily implemented in planar technology (such as microstrip or coplanar waveguide, for instance) and provides a practical realization of electromagnetic metamaterials. As any metamaterial, CRLH TLs are generally periodic structures formed by the repetition of unit-cells (an example of this type of cells is shown in Fig. 2) whose size, p, must fulfill the condition $p \ll \lambda_g$ (where λ_g is the guided wavelength) ir order to emulate an effectively homogeneous material. A powerful method to analyze these metamaterial lines is the TL approach, presented in Caloz & Itoh (2006), which employs an ABCD-matrix technique of periodically arranged unit-cells to model the artificial transmission line (see Fig. 3) and to determine its wave propagation characteristics (such as propagation constant or Bloch impedances, see Fig. 4).

There are many examples of interesting and groundbreaking applications of planar TL metamaterials at microwaves, such as for instance multi-band components, filters and diplexers, couplers, power-dividers, phase-shifters, lenses, or backfire to endfire leaky-wave antennas. A review of these and much more applications and devices can be found in metamaterials textbooks, such as in Eleftheriades & Balmain (2005), in Caloz & Itoh (2006), or in Marques et al. (2008). Note that all previously mentioned and most of metamaterials applications operate in the harmonic regime up to now, and they have been designed for narrow-band components and systems (even though some of them may support a multi-band operation).

2.2 Impulse regime analysis of CRLH transmission lines

Many media, ranging from traditional purely right-handed materials to recent CRLH metamaterials [see Caloz & Itoh (2006)], can be advantageously analyzed by the transmission line theory described in Pozar (2005). So far, this theory has been applied mostly in the *harmonic* regime, where Green's functions for both the voltage and the current along the line are available. However, it may also be employed in the time-domain, where the Green's function approach provides an efficient tool to analyze *impulse regime* signals along

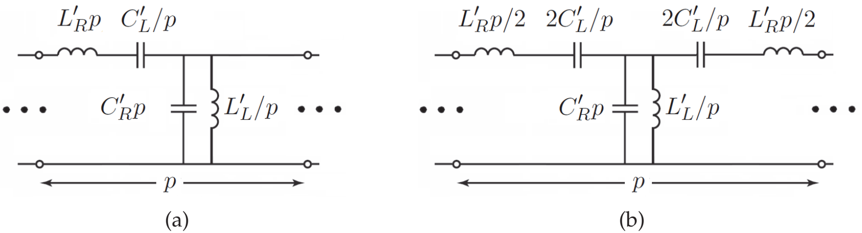


Fig. 2. Equivalent unit cell circuit model of a lossless CRLH transmission line. (a) Asymmetric configuration. (b) Symmetric configuration.

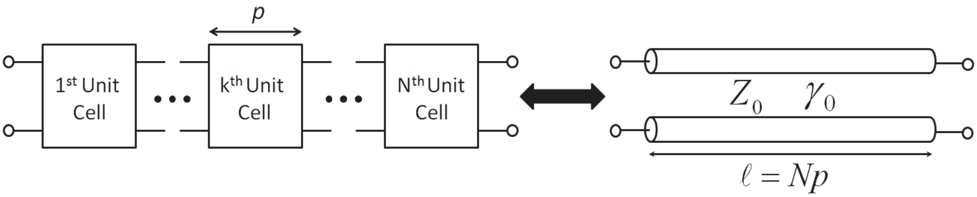


Fig. 3. Equivalence between N cascaded unit cells and a transmission line of length ℓ, characterized by an equivalent complex propagation constant γ_0 and Bloch impedance Z_0.

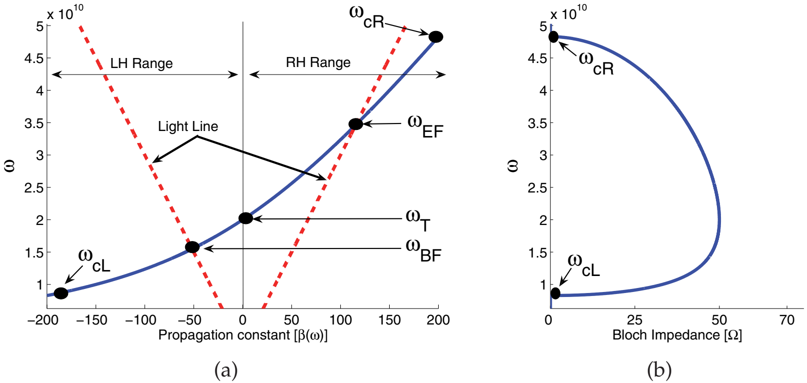


Fig. 4. Dispersion diagram (a) and frequency-dependent Bloch impedance (b) related to a *balanced* CRLH unit cell. The size of the unit cell is $p = 1$ cm and its circuital parameters are $C_R = C_L = 1.0$ pF and $L_L = L_R = 2.5$ nH. The frequencies ω_{cL} and ω_{cR} denote the bandpass frequency region of the line.

transmission lines. In this case, the point source model accurately characterizes a pulse generator, and the computed quantities are the voltages and currents along the line as a function of time.

Consider an electric source $\vec{J}(\vec{r},t)$ placed in an arbitrary homogeneous and dispersive medium. The wave equation in this case reads [see Collin (1991)]

$$\nabla \times \nabla \times \vec{E}(\vec{r},t) + \mu\varepsilon \frac{\partial^2}{\partial t^2}\vec{E}(\vec{r},t) = -\mu \frac{\partial}{\partial t}\vec{J}(\vec{r},t). \tag{1}$$

The spatial-temporal dyadic Green's function $\overline{\overline{G}}(\vec{r},\vec{r}\,';t,t')$ in this equation for a specific medium and a specific source is obtained as the response to a unitary point source $\vec{J}(\vec{r}\,',t') = \delta(\vec{r}\,';t')$. As described in Barton (1989), once this Green's function is known, the electric field may be computed using

$$\vec{E}(\vec{r},t) = \int\int \bar{\bar{G}}(\vec{r},\vec{r_g}\,';t,t') \cdot \vec{J}(\vec{r_g}\,',t')d\vec{r_g}\,'dt', \tag{2}$$

where the spatial-temporal Green's function may be expressed in terms of its inverse Fourier transform

$$\bar{\bar{G}}(\vec{r},\vec{r_g}\,';t,t') = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \tilde{\bar{\bar{G}}}(\vec{r},\vec{r_g}\,';\omega)e^{j\omega(t-t')}d\omega. \tag{3}$$

Inserting Eq. (3) into Eq. (2), yields

$$\vec{E}(\vec{r},t) = \frac{1}{2\pi}\int\int\int \tilde{\bar{\bar{G}}}(\vec{r},\vec{r_g}\,';\omega) \cdot \vec{J}(\vec{r_g}\,',t')e^{j\omega(t-t')}dr_g'dt'd\omega. \tag{4}$$

This expression provides the field radiated by an arbitrary source (in space and time) in an arbitrary dispersive homogenous medium. Since the spatial-temporal distribution of the source is generally known, only the Green's function needs to be computed to provide the field solution. An analogous formulation may naturally be obtained for the magnetic field.

In case of electrically thin 1D transmission lines (placed along the z direction, as shown in Fig. 5), the generator source may be reduced to a point source, greatly reducing the complexity of the problem. Consider a point source placed at the position $\vec{r}_g$, and with a temporal dependence

$$\vec{J}(\vec{r}_g,t') = \vec{\kappa}(\vec{r}_g)I_g(t') = \delta(\vec{r}-\vec{r}_g)I_g(t')\hat{e}_z. \tag{5}$$

In this case Eq. (4) is reduced to

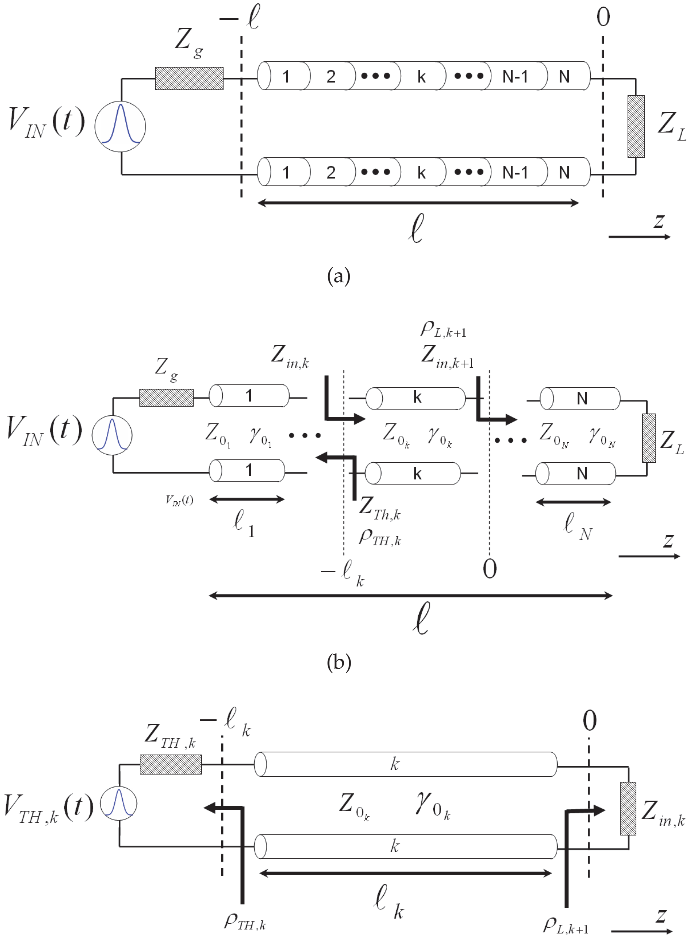
$$\vec{E}(\vec{r},t) = \frac{1}{2\pi}\int\int \tilde{\bar{\bar{G}}}(\vec{r},\vec{r_g}\,';\omega) \cdot I_g(t')\hat{e}_z e^{j\omega(t-t')}dt'd\omega. \tag{6}$$

At this point, we use the Fourier transform of the temporal source [$\tilde{I}_g(\omega) = \mathfrak{F}\{I_g(t')\}$, where the operator $\mathfrak{F}$ denotes a Fourier transform, see Pipes & Harvill (1971)]. Note that the input pulse is usually modulated at a frequency ω_0, which is included in the $\tilde{I}_g(\omega)$ notation as a $e^{j\omega_0 t}$ term. Besides, a transmission-line Green's function is used to obtain the voltage (V) or the current (I) along the 1D line, which may be expressed as

$$X(z,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{G}_X(z,z_g;\omega)\,\tilde{I}_g(\omega)\,e^{j\omega t}d\omega, \tag{7}$$

where z_g is the source position ($\vec{r}_g = z_g\hat{e}_z$), z is the observation point and $X(z,t)$ denotes the voltage or current along the line (in the z direction), as a function of the Green's functions employed [$\tilde{G}_V(z,z_g;\omega)$ or $\tilde{G}_I(z,z_g;\omega)$, related to the voltage or current, respectively]. It should be noted that the space dependence has been absorbed in the Green's function term, while the temporal information is described by Fourier and inverse-Fourier transforms.

Let us consider a simple matched transmission line, as shown in Fig. 5(a) [$Z_g = Z_L = Z_0(\omega), \forall\omega$]. In this simple case, the transmission line Green's functions for the voltages and



(a)

(b)

(c)

Fig. 5. Dispersive artificial transmission line excited by a point source generator. (a) Uniform case. The line, composed of N unit cells, is defined by its characteristic impedance $[Z_0(\omega)]$, complex propagation constant $[\gamma(\omega)]$ and length (ℓ). (b) Non-uniform case. The line is composed of N uniform transmission line sections. Each k^{th} section has its own length (ℓ_k), characteristic impedance $[Z_{0_k}(\omega)]$ and propagation constant $[\gamma_k(\omega)]$. (c) Thévenin equivalent circuit for the k^{th} uniform transmission line section. Reprinted with permission from Gómez-Díaz et al. (2010). Copyright 2010, IET.

currents may be expressed as

$$\tilde{G}_V(\vec{r},\vec{r}_g;\omega) = e^{-\gamma(\omega)R}, \tag{8}$$

$$\tilde{G}_I(\vec{r},\vec{r}_g;\omega) = \frac{e^{-\gamma(\omega)R}}{Z_0(\omega)}, \tag{9}$$

respectively, where $\gamma(\omega)$ is the complex propagation constant (or dispersion relation), $Z_0(\omega)$ is the characteristic impedance, and $R = |z - z_g|$ is the distance between the observation point z along the line and the source point z_g (generator).

Using these expressions, the voltage or the current along the line can easily be found with Eq. (7). Note that this equation applies to any type of transmission line, including metamaterial CRLH lines, provided that the propagation constant $[\gamma(\omega)]$ is known.

Consider now the more general case of a nonuniform transmission line medium composed of N uniform transmission line sections (or unit-cells), as shown in Fig. 5(b). The sections may be different from each other and may be of different type. Therefore, reflections occur due to the transition between two consecutive cells, and different propagation conditions appear at each cell. The Green's function along the k^{th} uniform transmission line section ($z \in [-\ell_k, 0]$, possibly infinitesimal), including generator and load mismatches, reads

$$G_k(z, z' = -\ell; \omega) = A(\omega)\left[e^{-\gamma_k(\omega)z} + \rho_{l,k}(\omega)e^{\gamma_k(\omega)z}\right], \tag{10}$$

where [see Pozar (2005)]

$$A(\omega) = \frac{V_{Th,k}(\omega)Z_{in,k}(\omega)}{Z_{in,k}(\omega) + Z_{Th,k}} \frac{e^{-\gamma_k(\omega)\ell_k}}{1 - \rho_{l,k}(\omega)\rho_{Th,k}(\omega)e^{-2\gamma_k(\omega)\ell_k}}, \tag{11}$$

and

$$\rho_{Th,k}(\omega) = \frac{Z_{Th,k}(\omega) - Z_{0_k}(\omega)}{Z_{Th,k}(\omega) + Z_{0_k}(\omega)}, \tag{12}$$

$$\rho_{l,k}(\omega) = \frac{Z_{in,k+1}(\omega) - Z_{0_k}(\omega)}{Z_{in,k+1}(\omega) + Z_{0_k}(\omega)}, \tag{13}$$

which was obtained by equating the Green's function evaluated at the input of the k^{th} section, to the Thévenin's voltage evaluated at the same section $V_{Th,k}$ [see Fig. 5(c)]. By repeatedly applying Eq. (10) from the load to the generator or vice-versa, the voltage as a function of time may be computed at any point along the nonuniform transmission medium.

Finally, note that the use of periodic input signals is important to produce some periodic phenomena or devices, such as the Talbot effect introduced in Azaña & Muriel (2001) or the UWB resonator presented in Gómez-Díaz et al. (2009a), among many others. For this purpose, the theory previously introduced can easily be extended to consider this type of input signals. In this case, the input signal may be represented in the time domain by

$$I_g^p(t) = \sum_{k=-\infty}^{k=+\infty} I_g(t - kT_0), \tag{14}$$

where T_0 is the period rate. The voltage or current along the transmission line may then be computed by inserting Eq. (14) in Eq. (7). Note that for many input pulses, such as modulated Gaussian pulses, the interchange between the integral and summation operations (related to Eq. (7) and Eq. (14), respectively), allowed by the assumed linearity of the system, will further contribute to reduce the computational cost required by this technique.

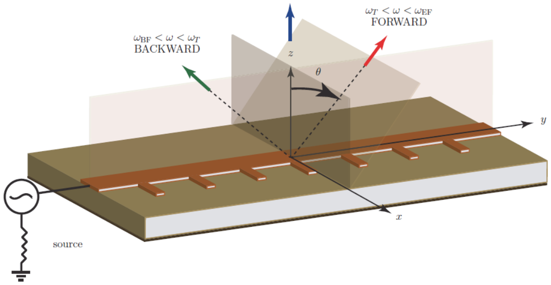

$\omega_{BF} < \omega < \omega_T$
BACKWARD
$\omega_T < \omega < \omega_{EF}$
FORWARD
z
θ
y
x
source

Fig. 6. Illustration of a CRLH LWA. The antenna can be configured to radiate at backwards [$\omega < \omega_T$ and $\beta(\omega) < 0$], forwards [$\omega > \omega_T$ and $\beta(\omega) > 0$] or broadside [$\omega = \omega_T$ and $\beta(\omega) = 0$].

2.3 Impulse regime analysis of CRLH leaky-wave antennas

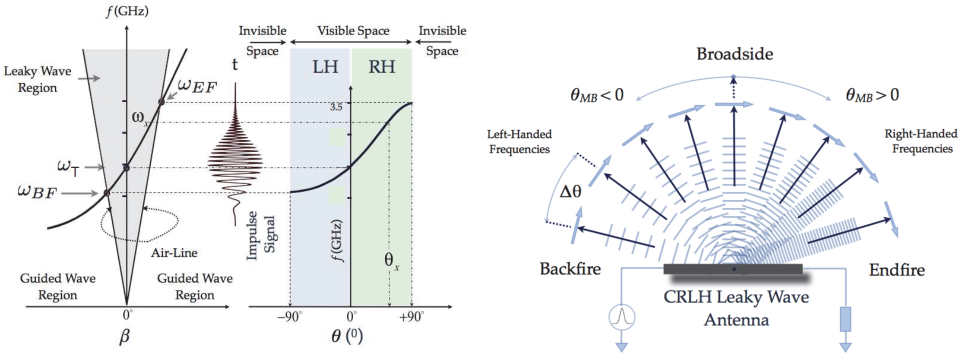
A CRLH transmission line supports a fast-wave mode [see Oliner & Jackson (2007)] which penetrates inside the fast-wave region. Therefore a CRLH structure behaves as a leaky-wave antenna when it is excited by a source with a frequency within a range of ($\omega_{BF} < \omega < \omega_{EF}$), where ω_{BF} and ω_{EF} are the fast-wave region limits [see Fig. 4(a) and Caloz & Itoh (2006)]. Since a CRLH line behaves as a LWA, the direction of the radiated main beam follows the LWA scanning law, which is given by [see Oliner & Jackson (2007)]

$$\sin(\theta) \approx \frac{\beta(\omega)}{k_0}. \tag{15}$$

In the above equation, θ is the radiation angle (measured from the perpendicular direction over the CRLH structure), $\beta(\omega)$ is the phase constant, and k_0 is the free-space wavenumber. Fig. 6 presents an illustration related to the operation principle of a CRLH LWA. As can be seen in the figure, and following Eq. (15), the antenna is able to radiate at backwards [when $\omega < \omega_T$ and $\beta(\omega) < 0$], forwards [$\omega > \omega_T$ and $\beta(\omega) > 0$] and broadside [$\omega = \omega_T$ and $\beta(\omega) = 0$]. Therefore, this type of structure is able to provide a full-space scanning, from backfire ($\theta = -90°$) to endfire ($\theta = +90°$), including the broadside ($\theta = 0°$) direction. The use of CRLH transmission lines as leaky-wave antennas has led to the development of many radiated-wave applications, most of them in the harmonic regime [see Caloz & Itoh (2006) and Eleftheriades & Balmain (2005)].

This subsection proposes an impulse-regime analysis of leaky-wave structures. Even though the proposed study is valid for all types of LWAs, we will focus on CRLH LWA structures because they are broadband in nature and they are able to radiate from backfire to endfire, including the broadside direction [see Caloz & Itoh (2006)]. As previously pointed out, a CRLH leaky-wave antenna follows the beam-scanning law of Eq. (15). Therefore, each input frequency, which lies inside the fast-wave region, is radiated towards a different direction into space. This situation is explicitly depicted in Fig. 7(a). As can be seen in the figure, there is a unique correspondence between each input frequency [ω_x, with ($\omega_{BF} \leq \omega_x \leq \omega_{EF}$)] and its associated radiation angle (θ_x). Therefore, each frequency is mapped into a different angle in space.

According to Eq. (15), if a CRLH LWA is excited by a modulated input pulse, as shown in Fig. 7(b), each spectral component of the signal is radiated towards a different direction



(a) (b)

Fig. 7. Impulse-regime behavior of CRLH LWAs. (a) Frequency-space relationship of a CRLH LWA. The dispersion curve is graphically related to its corresponding beam scanning law. (b) Spectral decomposition of a pulse obtained by the frequency-space mapping property of a CRLH leaky-wave antenna. Reprinted with permission from Gupta, Abielmona & Caloz (2009). Copyright 2009, IEEE.

in space at any particular instant. Hence, the CRLH LWA performs an *instantaneous spectral-to-spatial decomposition* of the input pulse. This decomposition allows to discriminate the various spectral components present in the input signal. In this sense, there is clear parallelism between LWAs (which usually operates at microwaves) and diffraction gratings [see Saleh & Teich (2007)], which usually operate at the optics regime and radiate each spatial input frequency towards a different direction angle in space. The main advantage of CRLH LWA over diffraction gratings is its simple point feeding system as compared to the diffraction gratings, which require plane-wave illumination.

Let us consider a single 1D (or electrically thin) LWA, located along the z axis, as shown in Fig. 8. It is important to clarify the notation employed to describe the situation under analysis. First, the line is defined by a characteristic impedance $[Z_0(\omega)]$, a complex propagation constant $[\gamma(\omega) = \alpha(\omega) + j\beta(\omega)]$, and a total length, ℓ. Second, the generator which excites the transmission line is placed at the position $\vec{r}_g$, with $\vec{r}_g = z_g\hat{e}_z$. Third, any point along the line is denoted as $\vec{r}\,'$, with $\vec{r}\,' = z'\hat{e}_z$ (note that $z_{start} \le z' \le z_{end}$, see Fig. 8). Besides, note that shall study the far-field radiation of the transmission line towards an observation point P (denoted as $\vec{r}$, with $\vec{r} = x\hat{e}_x + y\hat{e}_y + z\hat{e}_z$). Therefore, any observation point P must be located in the far-field region of the transmission line [see Balanis (2005)], fulfilling the far-field radiation condition, which is given by

$$R_{ant} = \sqrt{x^2 + y^2 + \left(z - \frac{\ell}{2}\right)^2} > \frac{2\ell^2}{\lambda_0}. \tag{16}$$

In this last equation, R_{ant} is the distance between the observation point P (placed at $\vec{r}$) and the transmission line, which can approximately be considered as a point source (placed at the center of the line) from a far-field point of view.

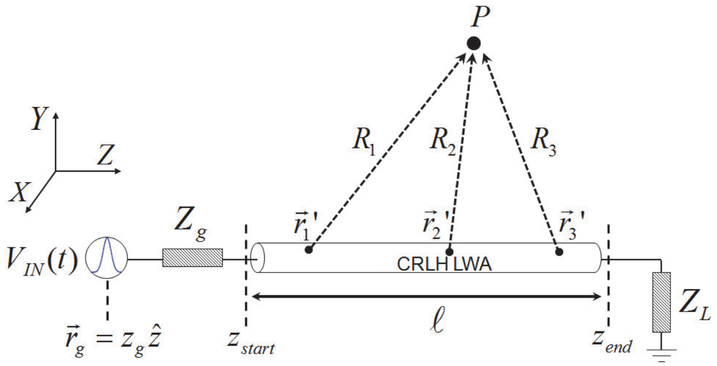


Fig. 8. Sketch of a single 1D CRLH LWA. The electrically thin antenna is considered as a linear wire from a far-field point of view. It is placed along the z-axis, it has a length of $\ell = z_{end} - z_{start}$, and it is fed by a point generator, placed at $\vec{r} = z_g \hat{e}_z$.

Let us also assume that the CRLH LWA is excited by a modulated input pulse. In this case, the time-domain theory developed in Section 2.2 for the analysis of impulse-regime CRLH TL can be employed to obtain the current which flows along the structure, which may be expressed as

$$I(z', t) = \int_{-\infty}^{\infty} \tilde{G}_I(z', z_g, \omega)\tilde{I}_g(\omega)e^{j\omega t}d\omega. \tag{17}$$

In this last equation, $\tilde{G}_I(z', z_g, \omega)$ represents the transmission line Green's function, related to the current, and $\tilde{I}_g(\omega)$ denotes the Fourier transform of the temporal input pulse.

The electric field radiated by the transmission line under study, in the far-field region, is approximately given by

$$\vec{E}(\vec{r}, \omega) \approx -j\omega\vec{A}(\vec{r}, \omega), \tag{18}$$

where $\vec{A}$ is the magnetic vector potential. It is important to keep in mind that this potential is related to the physical current which is flowing on the transmission line [see Balanis (2005)]. In the harmonic case, this potential may be expressed as

$$\vec{A}(\vec{r}, \omega) = \frac{\mu_0}{4\pi}\int_{z_{start}}^{z_{end}} I(z', \omega)\frac{e^{-jk_0 R}}{R}dz'\hat{e}_z, \tag{19}$$

where R is the distance between the pair of source-observation points.

However, in the impulse-regime case the CRLH LWA is excited by a temporal pulse. This time-domain behavior of the current induced on the structures propagates into the magnetic vector potential, which now reads

$$\vec{A}(\vec{r}, t) = \frac{\mu_0}{4\pi}\int_{\omega_{BF}}^{\omega_{EF}}\int_{z_{start}}^{z_{end}} I(z', \omega)\frac{e^{-jk_0 R}}{R}d\omega dz'\hat{e}_z. \tag{20}$$

It is important to remark that the frequency integration limits have been modified with respect to Eq. (7). Specifically, the frequency limits directly corresponds to the fast-wave region range ($\omega_{BF} \leq \omega \leq \omega_{EF}$). This is because LWAs are only able to radiate spectral components which lies inside this region. Outside of the fast-wave region, this type of structures behaves as a transmission lines as described in Caloz & Itoh (2006).

Finally, the time-domain electric field radiated by a CRLH LWA, under a far-field assumption, can easily be recovered using

$$\vec{E}(\vec{r},t) = \frac{-j\mu_0}{4\pi} \int_{\omega_{BF}}^{\omega_{EF}} \int_{z_{start}}^{z_{end}} \omega I(z',\omega) \frac{e^{-jk_0 R}}{R} d\omega dz' \hat{e}_z. \tag{21}$$

This expression provides the time-domain electric field radiation from a CRLH LWA excited by a modulated input pulse, at any observation point placed in the far-field region. The main features of this closed-form formulation are:

- All CRLH LWA radiation features at far-field are taken into account by using a *time-domain current along the structure*. This current is closely related to the complex propagation constant of the structure.
- Physical insight into the antenna radiation properties. The Green's function and the current flowing along the structure, related to the CRLH structure, completely define the antenna behavior. These parameters are obtained in closed-form.
- Extremely fast computation, because most expressions are simple and well-behaved integrals (and some of them are analytical for specific input pulses).
- Capability to deal with any type of input pulse, providing a continuous temporal output with unconditional stability.

The proposed approach can characterize complex radiated-wave UWB phenomena and devices. Specifically, this formulation will be employed in the next section to model several systems, such as a real-time spectrum analyzer (RTSA) or a frequency-resolved electrical gating system (FREG). It will be shown that the proposed formulation is able to completely characterize these systems with high efficiency and accuracy .

3. Optically-inspired phenomena at microwaves

3.1 Introduction

This section explores the impulse-regime phenomenology of CRLH structures and the subsequent theoretical and practical demonstration of several novel optically-inspired phenomena and applications at microwaves, in both, the guided and the radiative regime. The time-domain Green's function approach introduced in Section 2 has opened the door to a very fast, but still accurate, analysis of these ***novel microwave phenomena and applications, most of them transported from optics***, exploiting either the group velocity or the group velocity dispersion parameters of CRLH TL. The study can be divided into two main groups, related to the *guided-wave* or *radiative-wave* natures of the proposed phenomena and applications.

The analogy between these phenomena at microwave and their corresponding counterpart at optics [see Saleh & Teich (2007)] is deduced from the dispersive properties of CRLH structures. Specifically, *in the guided mode* there is a clear parallelism between the dispersive behavior of a CRLH line and an optical component (such as an optical fiber), which is inherently dispersive. Therefore, optical phenomena [such as the well-known temporal Talbot effect] can be reproduced at microwaves. In the *radiative mode*, the beam scanning law of the CRLH LWA is analog to a diffraction grating, where different spectral components are radiated (or diffracted) at different angles causing spatial dispersion. This may be exploited to develop spectrogram analyzers, capable to fully characterize both, in time and frequency, any unknown input signal.

3.2 Guided-regime

This subsection presents a study on the phenomenology of pulse propagation along CRLH transmission lines. As previously stated, the CRLH TL represents a general transmission medium, which is highly dispersive, especially in the left-handed frequency range. Next, the temporal Talbot effect [see Azaña & Muriel (1999)] is introduced, theoretically described and numerically confirmed for the case of CRLH media.

3.2.1 Pulse propagation along CRLH structures

The main goal of this subsection is to experimentally validate the theory presented in Section 2. For this purpose, the propagation of a modulated pulse along a CRLH TL is studied. The tunability of the pulse delay as a function of the modulation frequency is demonstrated by measuring the temporal delay from different modulated pulses, leading to a tunable delay system [see Abielmona et al. (2007)]. Then, the dispersive features of the CRLH TL are further demonstrated monitoring the effects of pulse propagation along a matched and mismatched line, in a cell-by-cell fashion.

First, we consider a CRLH transmission line composed of 30 unit cells and with the circuit parameters $C_R = 1.8$ pF, $C_L = 0.9$ pF, $L_R = 3.8$ nH and $L_L = 1.9$ nH (transition frequency $f_0 = 2.55$ GHz), excited by a modulated Gaussian pulse [$\sigma = 3.0$ ns, following the notation introduced in Saleh & Teich (2007)]. Fig. 9 shows the time-delayed waveforms obtained by the proposed theory for different carrier frequencies, and by experiments using a real-time oscilloscope (Agilent Infiniium DS0871204B). Excellent agreement is observed between theory and experiments. The group delay response is an important parameter in dispersive systems for analog signal processing applications. This parameter may be computed either in the time domain by determining the time differences between the maxima of the input and output pulses, or in the frequency domain by taking the derivative of the unwrapped phase of the transmission scattering parameter S_{21}. Fig. 10 shows the group delay along the same CRLH line and for the same pulse as in Fig. 9, computed by the discussed theory using the first approach, and validated by experiment using both approaches. Again, very good agreement is observed between theory and experiment. The small discrepancies between the two measured results may be explained by the tolerance in the localization of the pulse maxima. As it may be seen from these results and was originally presented in Abielmona et al. (2007), the CRLH transmission line acts as an impulse tunable delay system.

To better visualize the dispersion of a pulse along a CRLH dispersive medium, consider now a CRLH transmission line twice as long as before (60 unit cells), but with the same parameters. To completely validate the discussed theory, this time the line is excited by a modulated square pulse [$f_0 = 2.05$ GHz, $T = 2.2$ ns, following the notation of Saleh & Teich (2007)]. An ABCD matrix approach, described in Caloz & Itoh (2006), is employed to compute the propagation constant of the line, taking into account the finite number of unit cells in the experiment. The position-time trajectory of the pulse is presented in Fig. 11. Fig. 11(a) shows the computed

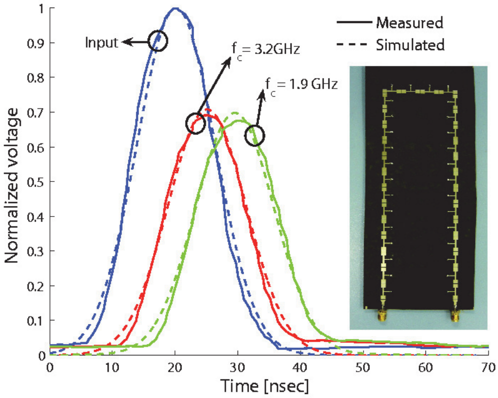


Fig. 9. Time-delayed Gaussian waveforms at the input/output of a CRLH transmission line for different carrier frequencies, obtained with the method proposed in Section 2.2. Measurement results are also shown for validation. The manufactured CRLH transmission line is shown in the inset. Reprinted with permission from Gómez-Díaz et al. (2009b). Copyright 2009, IEEE.

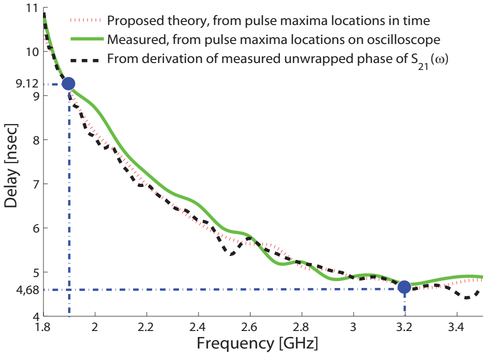
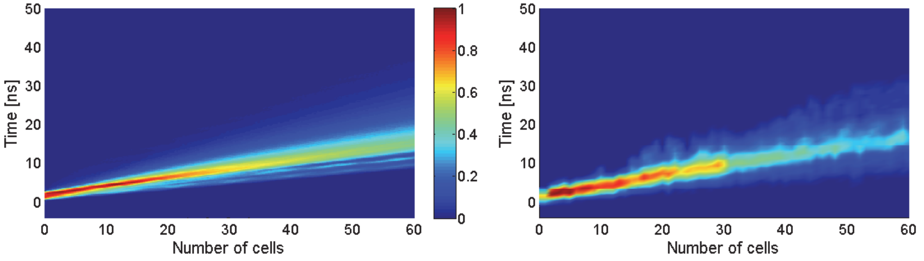


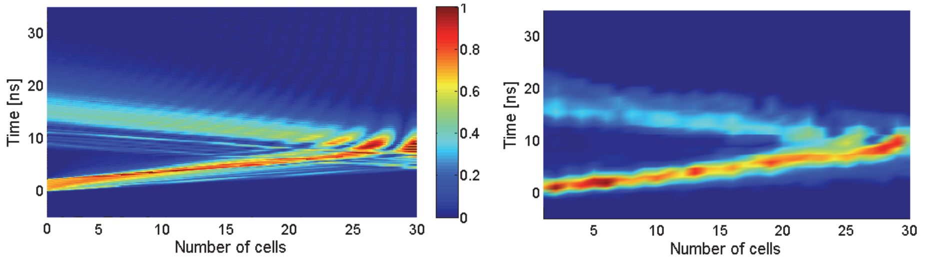
Fig. 10. Time delay versus modulation frequency, using the time difference between the maxima of the input and output pulses along the same CRLH line as in Fig. 9. Measured data using both, the same procedure as before and unwrapping the phase of $S_{21}(\omega)$, are also shown for validation. The delays obtained in Fig. 9 for $f_c = 3.2$ GHz and $f_c = 1.9$ GHz, which are 4.68 and 9.12 ns, correspond to the two highlighted points. Reprinted with permission from Gómez-Díaz et al. (2009b). Copyright 2009, IEEE.

results, from which two observations may be made: i) at the end of the line, the temporal width of the pulse has increased by a factor of 5 (at 50% of the magnitude), ii) the edges of the square envelope have been rounded off by the band-pass filtering response of the CRLH line. The small ripples near the end of the structure are explained by the Gibbs effect on the input pulse due to the finite computational interval and resolution. Fig. 11(b) shows the measured result using a high-impedance probe connected to the oscilloscope. The abrupt decrease of the voltage magnitude after the 30th cell is due to the fact that the 60-cell experimental line is



(a) (b)

Fig. 11. Propagation of a modulated square pulse ($f_0 = 2.05$ GHz, $T = 2.2$ ns) along a matched CRLH transmission line. The line includes 60 unit cells of length $p = 2.0$ cm and the circuital parameters are $C_R = 1.8$ pF, $C_L = 0.9$ pf, $L_R = 3.8$ nH and $L_L = 1.9$ nH. (a) Simulation. (b) Measurement. Reprinted with permission from Gómez-Díaz et al. (2009b). Copyright 2009, IEEE.



(a) (b)

Fig. 12. Propagation of a modulated square pulse ($f_0 = 2.05$ GHz, $T = 2.2$ ns) along an open-ended CRLH transmission line. The line is identical to that of Fig. 11 except that it includes only 30 unit cells. (a) Simulation. (b) Measurement. Reprinted with permission from Gómez-Díaz et al. (2009b). Copyright 2009, IEEE.

in fact constituted of two cascaded 30-cell lines with a small loss in the interconnection. The agreement with theory is reasonable, considering the tolerances of the measurement setup. Finally, let us investigate the effects of the reflection of a pulse at a discontinuity of a CRLH transmission line. For this purpose, the line is open-ended and it is excited by the same modulated square pulse as in the previous case. The propagation and the reflection of the pulse as a function of space and time is presented in Fig. 12. Again, good agreement is observed between the theory [Fig. 12(a)] and the experiment [Fig. 12(b)]. It is worth noticing that an interesting interference pattern, between propagating and reflected pulses, occurs near the discontinuity.

3.2.2 Temporal Talbot effect

The Talbot effect is a periodic constructive interference pattern produced by a dispersive transmission medium with second-order dispersion for a periodic input signal. It was first reported by H. F. Talbot in 1836 for the case of a source with periodic spatial variation [see Talbot (1836)]. The temporal counterpart of this effect [see Azaña & Muriel (1999) or Azaña & Muriel (2001)] occurs when a time-periodic signal is propagating along the same kind of medium. An input pulse train with temporal period T_0 and pulse width ΔT is replicated at the position nz_T ($n \in \mathbb{N}$), where z_T is called the Talbot distance (or self-imaging distance). Also, an increased repetition rate of m pulses per T_0 is obtained at the fractional distances $z_f = (s/m)z_T$ (where $s, m \in \mathbb{N}$), provided that (s/m) is an irreducible fraction, and under the condition that: $m < T_0/\Delta T$ [see Azaña & Muriel (2001)]. The temporal Talbot effect has been employed mainly in the optical regime, for applications such as the generation of signals with ultrahigh repetition rate (THz) from slower ranges (GHz), as shown in Azaña & Muriel (2001), or pulse compression, as presented in detail in Berger et al. (2004).

At microwaves, this phenomenon is more difficult to reproduce because it is difficult to obtain a second-order dispersive medium at this frequency region. However, the recent introduction of CRLH TLs [see Caloz & Itoh (2006)], which provides a dispersive broadband behavior, may lead to novel scenarios where the temporal Talbot effect and their subsequent applications can be reproduced. This section provides a theoretical demonstration of the temporal Talbot effect in CRLH media, including a full-wave validation of the phenomena at microwaves. Furthermore, the conditions for the phenomena existence in CRLH media are stated, in connection with potential device applications.

Intuitively, the temporal Talbot effect occurs when a periodic pulse signal is transmitted through a second order dispersion medium where the neighboring pulses interfere as a result of dispersion so as to produce new temporal components. For the analysis, let us consider a single modulated pulse, denoted by $\Psi(t) = \Psi_0(t)e^{j\omega_0 t}$, where $\Psi_0(t)$ is a slowly varying envelope. The periodic signal is then represented in the time domain, at the position $z = 0$, as

$$A(z=0,t) = \sum_{n=-\infty}^{n=+\infty} \Psi(t - nT_0), \tag{22}$$

where T_0 is the repetition rate of the signal. In the spectral domain, the periodic signal becomes discrete, and it may be expressed as

$$\tilde{A}(z=0,\omega) = \omega_r \sum_{n=-\infty}^{n=+\infty} \tilde{\Psi}(\omega = n\omega_r)\delta(\omega - n\omega_r), \tag{23}$$

where $\omega_r = 2\pi/T_0$ is the spectral repetition frequency.

On the other hand, the transfer function of a lossless CRLH TL is given by

$$\tilde{H}(z,\omega) = e^{-j\beta(\omega)z}, \tag{24}$$

where $\beta(\omega)$ is the CRLH TL propagation constant. This propagation constant, around the CRLH TL transition frequency, may be approximated as

$$\beta(\omega) = \frac{\omega}{\omega'_R} - \frac{\omega'_L}{\omega}, \tag{25}$$

where $\omega'_R = p/\sqrt{L_R C_R}$ and $\omega'_L = 1/\sqrt{p^2 L_L C_L}$, as described in Caloz & Itoh (2006). In the right handed side of Eq. (25), the first term provides a simple time delay (or linear frequency phase), whereas the second order term is responsible for the line dispersion. Eq. (25) can be further expanded, employing Taylor series, around a modulation frequency (ω_0) as

$$\beta(\omega) = \beta_0 + \beta_1(\omega - \omega_0) + \frac{1}{2}\beta_2(\omega - \omega_0)^2 + O(\omega^3), \tag{26}$$

where the term $O(\omega^3)$ is related to the order of the error committed in the approximation, and the term β_n is defined as

$$\beta_n = \left[\frac{\beta^n(\omega)}{\partial\omega^n}\right]\bigg|_{\omega=\omega_0}. \tag{27}$$

Note that Eq. (26) is only valid for the case of narrowband pulses (centered at the frequency ω_0 and with bandwidth $\Delta\omega$).
Employing Eq. (26), the transfer function of the CRLH TL $\left[\tilde{H}(\omega') = \tilde{H}(\omega = \omega_0 + \omega')\right.$, where a change of variable has been introduced from ω to ω'] takes the form

$$\tilde{H}(z,\omega') = \exp\left\{-j\left[\beta_0 + \beta_1\omega' + \frac{1}{2}\beta_2\omega'^2\right]z\right\}. \tag{28}$$

In order to derive the Talbot distance, only the third expansion term of the exponential is considered. This is because the first two terms do not provide any information related to the Talbot distance [which is only due to second order dispersion, see Azaña & Muriel (2001)]. Specifically, the first term is related to the modulation frequency of the pulse and does not carry any information about the envelope, and the second term represent the group delay (or retarded frame) of the signal.

Using Eq. (23) and Eq. (28), the spectrum of the signal at the distance z can be written as

$$\begin{aligned}\tilde{A}(z,\omega') =& \omega_r \sum_{n=-\infty}^{n=+\infty} \exp\left\{-j\frac{\beta_2\omega'^2 z}{2}\right\}\tilde{\Psi}(n\omega_r)\delta(\omega' - n\omega_r) = \\ & \omega_r \sum_{n=-\infty}^{n=+\infty} \exp\left\{-j\frac{\beta_2 z}{2}\left(\frac{2\pi n}{T_0}\right)^2\right\}\tilde{\Psi}(n\omega_r)\delta(\omega' - n\omega_r).\end{aligned} \tag{29}$$

Note that the appearance of ω' squared term in this equation is due to the quadratic phase factor in the spectral response of the CRLH TL dispersive medium.

Eq. (29) reveals that the Talbot effect $\left[\text{i.e., } \tilde{A}_z(z,\omega) = \tilde{A}(z=0,\omega)\right]$ occurs under the condition

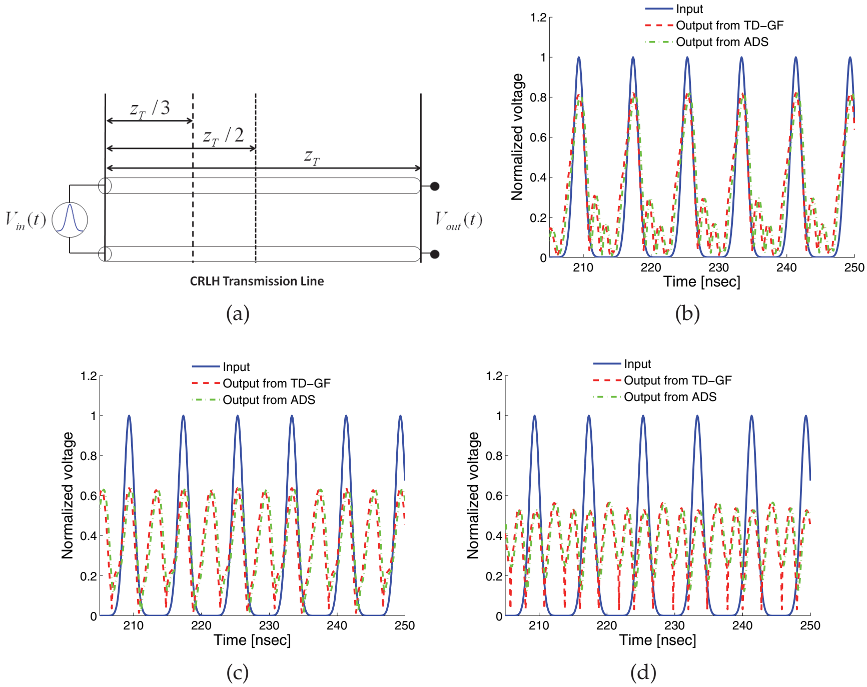
$$\frac{\beta_2 z}{2}\left(\frac{2\pi n}{T_0}\right)^2 = \pi p, \tag{30}$$

where $p \in \mathbb{N}$. The case of the first integer Talbot distance ($p = 1$) is given by

$$z_T = \frac{T_0^2}{2\pi|\beta_2|} = \frac{T_0^2\omega_0^3}{4\pi\omega_L'^2}, \tag{31}$$

where the identity $\beta_2 = -2\omega_L'^2/\omega_0^3$ [see Eq. (27)] has been employed.

As demonstrated in Azaña & Muriel (2001), the Talbot distance can also be obtained at fractionary distances given by $z_f = (s/m)zT$, where s and m are irreducible integers. At this fractionary distance, the periodic input signal is also self-imaged but with an increase repetition rate by a factor of m. This phenomenon corresponds to the fractionary Talbot effect [see Azaña & Muriel (2001)]. It is important to note that in the case of a CRLH TL, the Talbot distance can be tuned externally by modifying the parameters T_0 and ω_0, without changing the intrinsic parameters of the line.



(a) (b) (c) (d)

Fig. 13. Talbot repetition rate multiplication effect. a) CRLH TL with length corresponding to the basic Talbot distance z_T. b) Reconstruction of the original pulse train at the Talbot distance z_T. c) Repetition rate doubling at the distance $z_T/2$. d) Repetition rate tripling at the distance $z_T/3$. Results obtained from the proposed approach, and validated using the commercial software ADS©. Reprinted with permission from Gómez-Díaz et al. (2009b). Copyright 2009, IEEE.

After theoretically deriving the Talbot distance related to CRLH TLs, the phenomena will be verified using the numerical technique proposed in Section 2.2 and commercial full-wave simulations. Consider a balanced lossless CRLH transmission line with circuital parameters $C_R = C_L = 1.0$ pF and $L_R = L_L = 2.5$ nH. A modulated train of Gaussian pulses, with temporal width of $\sigma = 0.75$ ns and period rate of $T_0 = 8$ ns is used to excite the line shown in Fig. 13(a). Fig. 13(b) presents the input pulse train and the output pulse train at the Talbot distance z_T. The reconstruction of the initial train of pulses is confirmed, although a small disagreement, due to higher order terms (greater than 2) of the CRLH dispersion relation, is observed between two consecutive pulses. Fig. 13(c) and Fig. 13(d) show the input and output voltages at the fractional Talbot distances of $z_T/2$ and $z_T/3$, respectively. The effect of pulse multiplication is thus clearly confirmed, while the distortion is smaller because less higher-order dispersion effects occur over a shorter distance of propagation.

The practical implementation of Talbot devices based on CRLH transmission lines depends on the technology employed. For instance, with the CRLH parameters used in Fig. 13, a microstrip implementation with a typical unit cell size of 1 cm would lead to a Talbot distance z_T of around 17 meters. This is unpractical, specially due to the losses. However, using multilayer technology [see Horii et al. (2005)], this distance may be dramatically reduced to sizes in the order of several centimeters, while even lower sizes apply for pulse rate repetition multiplication. This decrease in length also decreases the total amount of losses, allowing the Talbot effect to be applied in practical situations, such as the generation of signals with ultrahigh repetition rate or pulse compression [see Berger et al. (2004)].

3.3 Radiative-regime

The full-space scanning capabilities of CRLH LWAs are exploited in this Section to achieve the *spectral-spatial* decomposition of an input broadband signal. This property is then applied to the development of a real-time spectrum analyzer [RTSA, see Gupta, Abielmona & Caloz (2009)] and a frequency resolved electrical gating system [FREG, see Gupta, Gómez-Díaz & Caloz (2009)], which are UWB applications able to fully characterize, in both time and frequency, an unknown signal. These systems are fully modeled by the technique presented in Section 2.3, which constitutes an ideal tool to provide not only a fast numerical system characterization, but a deep physical insight into the electromagnetic properties of the devices.

3.3.1 Real time spectrogram analyzer (RTSA)

A real time spectrogram analyzer (RTSA) is a device able to provide a joint time-frequency representation of an unknown input signal. This device is specially useful in modern UWB systems, as described in Ghavami et al. (2007), where ultra-fast transient signals are involved. The common output of a RTSA system is an *spectrogram* [see Cohen (1989)], which is a 2-D plot of a signal where the energy distribution is related to an image in a time-frequency plane. The joint time-frequency representation provides information related to the temporal evolution of each spectral component, and the exact amplitude and location in time of each frequency. It is important to mention that spectrograms suffer from the fundamental "uncertainty principle" limitation, which states

$$\Delta t \Delta f \geq \frac{1}{2}, \tag{32}$$

where Δf is the bandwidth of the gated signal and Δt is the gate duration. This implies an inherent trade off between time and frequency resolution for all spectrograms.

There are two main approaches to obtain the spectrogram of an unknown input test signal at microwave frequencies. The first and more usual technique is based on the use of *digital components*, which performs the short-time Fourier transform (STFT) [see Oppenheim (1996)] of the unknown input signal. However, the use of this approach at microwaves has the important drawback of requiring fast processors and large memories, which limits its use to input signals of just a few hundreds megahertz and temporal resolutions of only a few microseconds. The second option, purely analog, is based on the use of a *bank of filters*, as described in Amin & Feng (1995). The main disadvantage of this approach is that it requires a very large number of channels and extremely narrow-band filters, which is specially difficult to achieve at high frequencies. Consequently, this option is very complex and expensive.

In this context, a novel RTSA based on the spectral-spatial decomposition of CRLH LWAs was proposed in Gupta, Abielmona & Caloz (2009). In this approach, the CRLH LWA provides an analog implementation of the STFT, in a similar way as it has been implemented in the optical regime. A schematic of the CRLH LWA RTSA system is shown in Fig. 14. The description of the system behavior is as follows. First, the spectral-spatial decomposition of the CRLH LWA

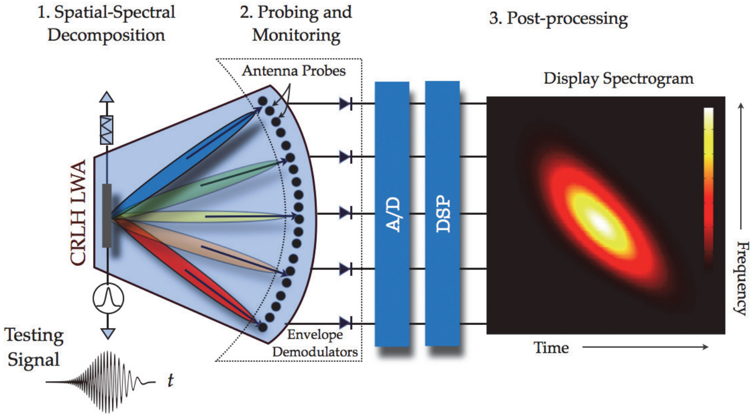


Fig. 14. Analog real-time spectrogram analyzer showing the CRLH LWA, the antenna probes, the envelope detectors, the A/D converters, the DSP block, and the display with the spectrogram. Reprinted with permission from Gupta, Abielmona & Caloz (2009). Copyright 2009, IEEE.

is employed to discriminate the frequency components of the input test signal (see Section 2.3). Second, a set of probes (antenna receivers) monitor the time variation of each frequency component. Finally, a postprocessing step performs the analog/digital (A/D) conversion, the data processing and the display of the spectrogram.

The analog CRLH LWA RTSA system provides several advantages and benefits as compared with other RTSAs at microwave frequencies. First, this approach is completely analog and real time. Therefore, there is not requirement of large memories and fast processors, just a light postprocessing stage. Second, the same RTSA system can use different CRLH LWAs, in order to be flexible and to cover several frequency ranges. The use of new technologies to fabricate CRLH LWAs allows the use of a wide variety of input signals, from microwave up to potentially millimeter-wave frequencies. Third, the CRLH LWA RTSA system is inherently broadband, and it can be designed to process the 100% of an input signal bandwidth.

On the other hand, the CRLH LWA RTSA system has also to deal with some drawbacks. First, it requires a far-field probe configuration, which makes the system relatively large. However, the system can possibly be compacted employing near-field to far-field transformations. Second, the physical length of the LWA represents a space-gating mechanism which controls both, the temporal and the frequency resolution of the resulting spectrograms. This is clearly shown in Fig. 15, where a slice of the signal's spatial waveform is presented on and radiating from the antenna. And third, the time and frequency resolution also depends on the number of detectors, and their associated response time and sampling frequency. Therefore, there is a fundamental tradeoff between time and frequency resolutions.

In order to characterize this system, full-wave commercial software may be employed. However, the generation of the output results are extremely time consuming due to the complexity of the system and to the temporal nature of the analysis. An interesting and efficient alternative is to employ the theory developed for the modeling of impulse-regime radiation (see Section 2.3), to fully characterize the CRLH LWA RTSA system. For the sake of

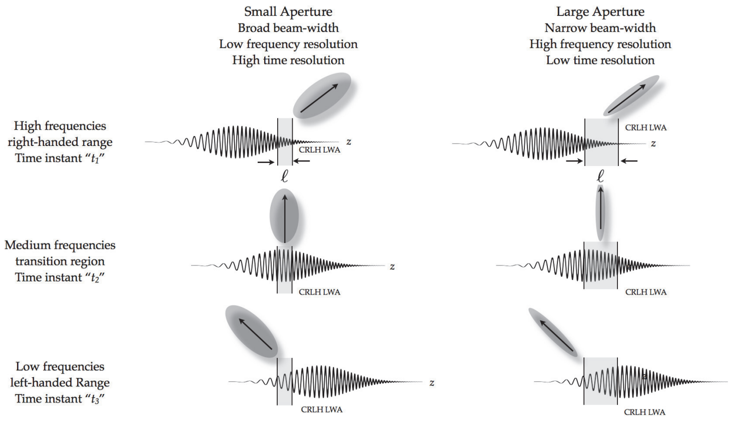
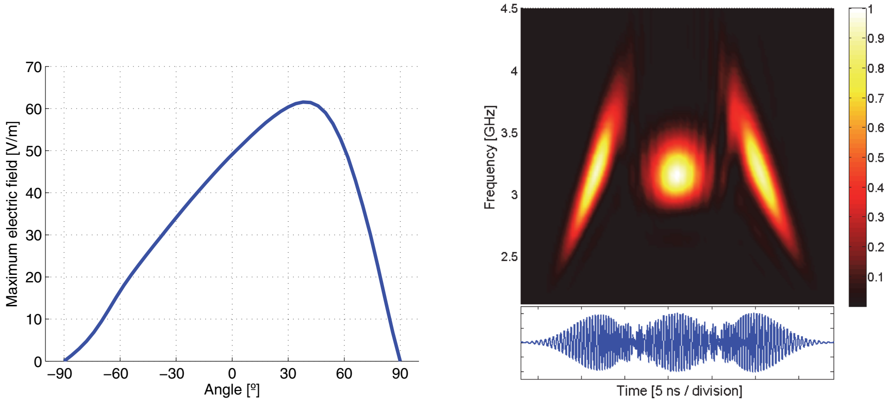


Fig. 15. Impact of the LWA size ℓ on the time-frequency resolution of the spectrograms generated by the an analog CRLH LWA RTSA.

validation, let us consider a complete RTSA system based on a CRLH LWA which is composed of 32 unit cells of length $p = 1.0$ cm, with circuital parameters of $C_R = C_L = 1.0$ pF and $L_R = L_L = 2.5$ nH. In order to complete the analog RTSA system, a total of number of 181 observation probes are placed in a semi-circular configuration, following the configuration of Fig. 14. The first step to model the RTSA is to perform the calibration of the system. This is necessary to compensate for the different power levels received at each probe [as described in Gómez-Díaz et al. (2010)], due to the directivity variation with frequency of this type of antennas. For the calibration, a narrow-band signal is modulated to the different fast-wave frequencies [following Eq. (15) (scanning law)] and subsequently radiated by the LWA. Then, the maximum power received at each probe is stored, obtaining a normalization rule for this particular system configuration. In our example, the calibration data is shown in Fig. 16(a). After the RTSA system has been calibrated, it can be efficiently used to obtain spectrograms of an unknown input signal. For the method validation, the actual temporal and frequency information of an input test signal, previously known, is employed.

Let us consider now that the CRLH LWA is fed by a signal composed of three modulated Gaussian pulses. The first pulse has a positive-chirp modulation (which means that the modulation frequency is increasing with time), and the third pulse has a negative-chirp modulation. The spectrogram obtained with the proposed method, after calibration, is depicted in Fig. 16(b), including an additional graph showing the analytical temporal representation of the signal. As can be observed in the figure, the spectrogram follows the signal variations in time (the three pulses are clearly observable) and also, simultaneously, in frequency. It is especially interesting to observe the transition between two consecutive pulses, where frequencies corresponding to different pulses appear at the same instant.

The method proposed here is able to perform a quick (about 30 seconds) and accurate modeling of the RTSA system, with deep insight into the CRLH LWA time-radiation properties, and avoiding the extremely time-consuming analysis required in full-wave simulations (usually between $8 - 12$ hours). Therefore, it provides a fast tool to configure an RTSA system and to determine a priory the range of input signals which can accurately be characterized for a given CRLH LWA.

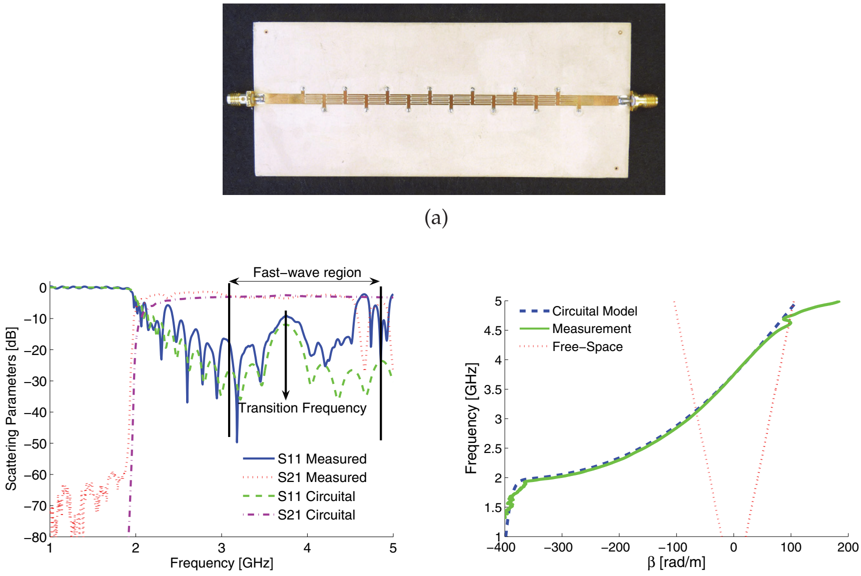


(a) (b)

Fig. 16. Results obtained from an RTSA system based on a CRLH LWA composed of 32 unit cells of length $p = 1.0$ cm, with circuital parameters of $C_R = C_L = 1.0$pF and $L_R = L_L = 2.5$nH. (a) Maximum electric field obtained at the different positions of the probes, used for the calibration of the RTSA system. (b) Normalized spectrogram of a three chirp-modulated Gaussian pulses signal, with chirp parameters $C = -[10, 0, 10]$, modulation frequency $f_0 = 3.19$ GHz and temporal width $\sigma = 1.0$ ns, computed with the proposed technique. The inset shows the analytical time response of the signal. Reprinted with permission from Gómez-Díaz et al. (2010). Copyright 2010, IET.

Finally, experimental results from a RTSA prototype are included for a complete system validation and to confirm the accuracy of the proposed modeling technique. For this purpose, a CRLH LWA fabricated in microstrip technology and composed of 14 -0.8 cm-long unit cells, with circuital parameters $C_R = 1.29$ pF, $C_L = 0.602$ pF, $L_R = 3.0$ nH and $L_L = 1.4$ nH, is employed. A photo of the fabricated antenna is depicted in Fig. 17(a), whereas a comparison of the measured and simulated scattering parameters is shown in Fig. 17(b). Furthermore, Fig. 17(c) presents a simulation-measurements comparison of the antenna dispersion relationship. As can be observed, the CRLH LWA presents its transition frequency at 3.745 GHz. The antenna fast-wave frequency region starts at about 3.1 GHz (related to backfire radiation), and it is extended until 4.7 GHz (related to endfire radiation). Note that several resonances occur within the fast-wave frequency region (close to endfire), degrading the antenna performance. These resonances are due to internal resonances of the interdigital capacitors employed in the antenna prototype, and they are not considered in the circuital model of the LWA.

In order to measure the spectrogram of the test signals, a single receiver was used and rotated around the circular far-field trajectory of the system (specifically, between $\theta = -80°$ and $\theta = 80°$ with increments of 5°). The RTSA system is calibrated employing a linear frequency ramp. After calibration, the CRLH LWA is excited by a modulated Gaussian pulse, with FWHM of 3.5 ns. Fig. 18 presents a comparison between simulations (obtained by the proposed time-domain Green's function approach) and measurements, as a function of the pulse modulation frequency. It can be observed that a very good agreement is achieved in all cases. First, the pulse modulation frequency is set to 3.3 GHz, which corresponds to backwards radiation. This is clearly shown in the spectrograms of Fig. 18(a) and Fig. 18(b). Then the



(a)

(b)

(c)

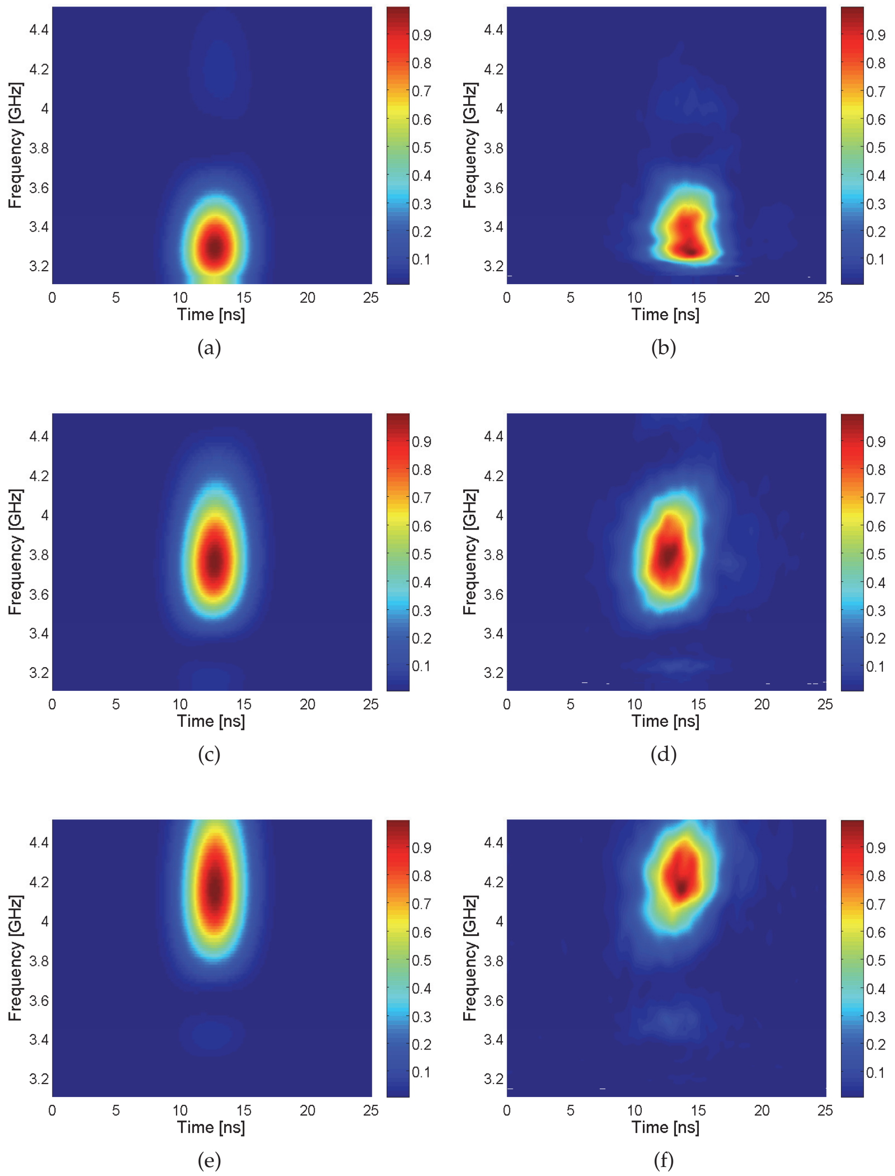
Fig. 17. 1D CRLH LW antenna composed by 14 -0.8 cm- long cells, with circuital parameters $C_R = 1.29$ pF, $C_L = 0.602$ pF, $L_R = 3.0$ nH and $L_L = 1.4$ nH. a) Photo of a microstrip CRLH LWA prototype. b) Scattering parameters. c) Dispersion relation. Reprinted with permission from Gómez-Díaz et al. (2009c). Copyright 2009, American Institute of Physics.

modulation frequency is set to the CRLH LWA transition frequency, which corresponds to broadside radiation. As expected, the obtained spectrograms [see Fig. 18(c) and Fig. 18(d)] confirm the change in frequency. Finally, pulse modulation frequency is set to 4.2 GHz, which corresponds to a forward direction. Again the spectrograms show the changes in the pulse frequency, experimentally verifying the RTSA system and confirming the usefulness of the proposed theoretical approach to model this type of analog systems.

3.3.2 Frequency resolved electrical gating system (FREG)

Analog CRLH LWA RTSAs generate the spectrogram of an unknown input signal in real-time, using the spectral-spatial decomposition property of the leaky-wave antenna, with minimal requirements on computational resources. This system provides important benefits over any digital RTSAs at microwave frequencies. However, as explained in the previous section, the time and frequency resolution of the generated spectrograms directly depends on the physical length of the CRLH LWA, which is fixed in a given system. Therefore, a particular antenna can only handle a limited range of input signals, which must fulfil specific time and frequency constrains. This imposes an important limitation to the CRLH LWA RTSA systems.

This section proposes a novel analog approach to obtain spectrograms, where the hardware dependence is suppressed, at the cost of the requirement of periodicity of the input signals. This approach is inspired from a similar system known in optics as frequency resolved optical gating (FROG) [see Trebino (2002)], where a self-gating principle is applied to provide close to ideal spectrograms for arbitrary test signals. Here, we propose a microwave counterpart of



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 18. Spectrograms obtained by the proposed RTSA model (figures on the left) and by experiments (figures on the right), employing the CRLH LWA of Fig. 17. A modulated Gaussian pulse with $FWHM = 3.5$ ns feeds the antenna. The pulse modulation frequency is set to 3.3, 3.745 and 4.2 GHz, corresponding to backward [(a) and (b)], broadside [(c) and (d)] and forward [(e) and (f)] radiation, respectively.

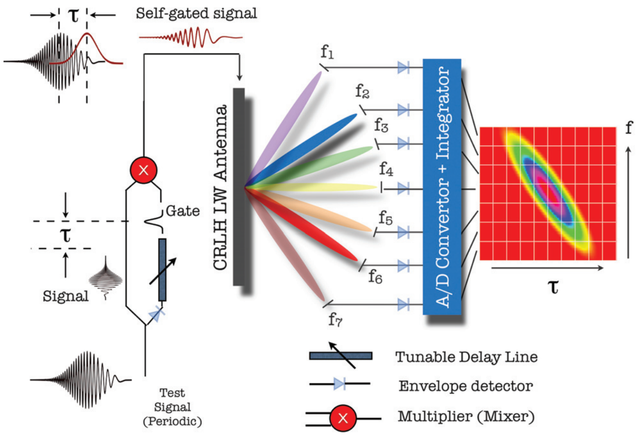


Fig. 19. Proposed frequency resolved electrical gating (FREG) system. Reprinted with permission from Gupta, Gómez-Díaz & Caloz (2009). Copyright 2009, EuMC.

the FROG system, which is termed frequency resolved electrical gating (FREG). This system is very useful for the measurement and characterization of fast-varying non-stationary UWB signals and ultrashort pulses.

In order to compute a spectrogram of a signal $x(t)$, a temporal gating function $g(t)$ is required [see Cohen (1989)]. Using a self-gating approach, instead of using a separate time signal as the gate function, the *envelope of the testing signal* itself is used as the gating function, i.e. $g(t) = |x(t)|$. The spectrogram of a signal $x(t)$ is then obtained as

$$S(\tau,\omega) = \left| \int_{-\infty}^{\infty} x(t)|x(t-\tau)|e^{-j\omega t}dt \right|^2. \tag{33}$$

The proposed FREG system, based on this self-gating principle and on the spectral-spatial decomposition property of the CRLH LWA, is depicted in Fig. 19. The testing signal, whose spectrogram is to be generated, is split into two channels. One of the channels is envelope detected and passed through a tunable delay line. The two channels are then mixed together. The mixer thus performs the self-gating process at a given time delay instant τ. This self-gated signal is then injected into a CRLH LWA which spectrally resolves it in space. Once the frequency components are separated in space, antennas circularly placed in the far-field of the LWA receive the different frequency components corresponding their angular position. All the received signals are then digitized and summed, before being stored for spectrogram display. This process is repeated for different values of the time delay τ so that the entire test signal is scanned, according to Eq. (33), until the spectrogram is fully constructed. Since, the beam scanning law of CRLH LWA is nonlinear in nature, a final post-processing step is required to linearize the spectrogram.

The proposed system exhibits significant advantages over the analog RTSA system and purely digital systems. Due to the self-gating process, neither the time nor the frequency resolutions of the generated spectrogram depend on the physical length of the antenna. The time and frequency resolutions are thus dependent only on the time signal itself and the hardware

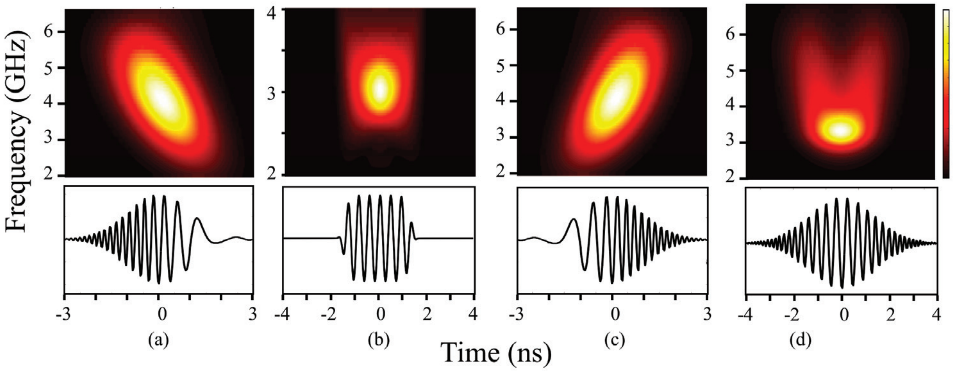


Fig. 20. Simulated spectrograms. a) Down-chirped gaussian pulse ($C_1 = -10$, $C_2 = 0$, $f_0 = 4$ GHz). b) Non-chirped super-gaussian pulse ($C_1 = C_2 = 0$, $f_0 = 3$ GHz). c) Up-chirped gaussian pulse ($C_1 = +10$, $C_2 = 0$, $f_0 = 4$ GHz). d) Cubically chirped gaussian pulse ($C_1 = 0$, $C_2 = 0.25 \times 10^{28}$). All pulse have a FWHM duration of 1 ns with a initial pulse offset of $t_0 = 6.5$ ns, and are described using Eq. (34). Reprinted with permission from Gupta, Gómez-Díaz & Caloz (2009). Copyright 2009, EuMC.

dependence spectrogram is suppressed. The LWA simply plays a role of spectral decomposer which, when longer (higher directivity), provides better separation of frequencies in space.

The choice of the gate duration is an important parameter to achieve an optimal time-frequency resolution in the spectrogram. An optimal gate duration for pulses with dominating phase variations is given by $T_g \approx 1/\sqrt{2|\phi''(t)|}$, where $\phi''(t)$ is the second time derivative of phase [see Cohen (1989)]. This duration permits the resolution of the fastest phase variations. For general pulse measurement, a gate duration as short as the testing signal itself or slightly shorter is thus desirable. Since the FREG system is based on self-gating, the gate duration is close to optimal and the corresponding spectrograms are ideal, as demonstrated in Trebino (2002).

Moreover, the proposed system being analog in nature, neither require fast processors nor huge memory buffers, which avoid placing a heavy computational burden on the system. Furthermore, the system is frequency scalable and sufficiently broadband to handle a wide variety of UWB signals. As mentioned above, the length of the LWA controls the spectral decomposition of the gated signal, which is improved as the physical length of the antenna is increased. Finally, since it uses a multi-shot measurement procedure, where the testing signal is gated several times with different time delays τ, the proposed FREG system requires a periodic input signal. This is the main limitation of the FREG system.

In order to simulate the proposed FREG system, the time-domain Green's functions approach presented in Section 2.3 is employed. The role of this theory is to model the transient CRLH LWA behavior, which provides the spectral-spatial decomposition property and is a key component of the proposed FREG system. Then, the other components of the system are implemented as follows. The envelope of the testing signals are numerically obtained and used as a gating function. The tunable time delay between the replica of the test signal and the gate signal is applied, and the mixer, which performs the self-gating operation, is modeled by a simple mathematical product. The resulting signal is then fed into the CRLH LWA. As a final stage, the temporal radiation computed at the probe locations are integrated as a function of the gate delay (τ), in order to recompose the desired spectrogram.

Once the numerical model of the system is complete, spectrograms obtained from various test signals are computed. For this purpose, consider a 16-cell CRLH LWA with the circuital parameters $C_L = C_R = 1$ pF, $L_L = L_R = 2.5$ nH and a unit cell size of $p = 2$ cm, easily implemented in metal-insulator-metal (MIM) technology [see Abielmona et al. (2007)]. The various modulated testing pulses are gaussian and super gaussian-type signals which follows the equation

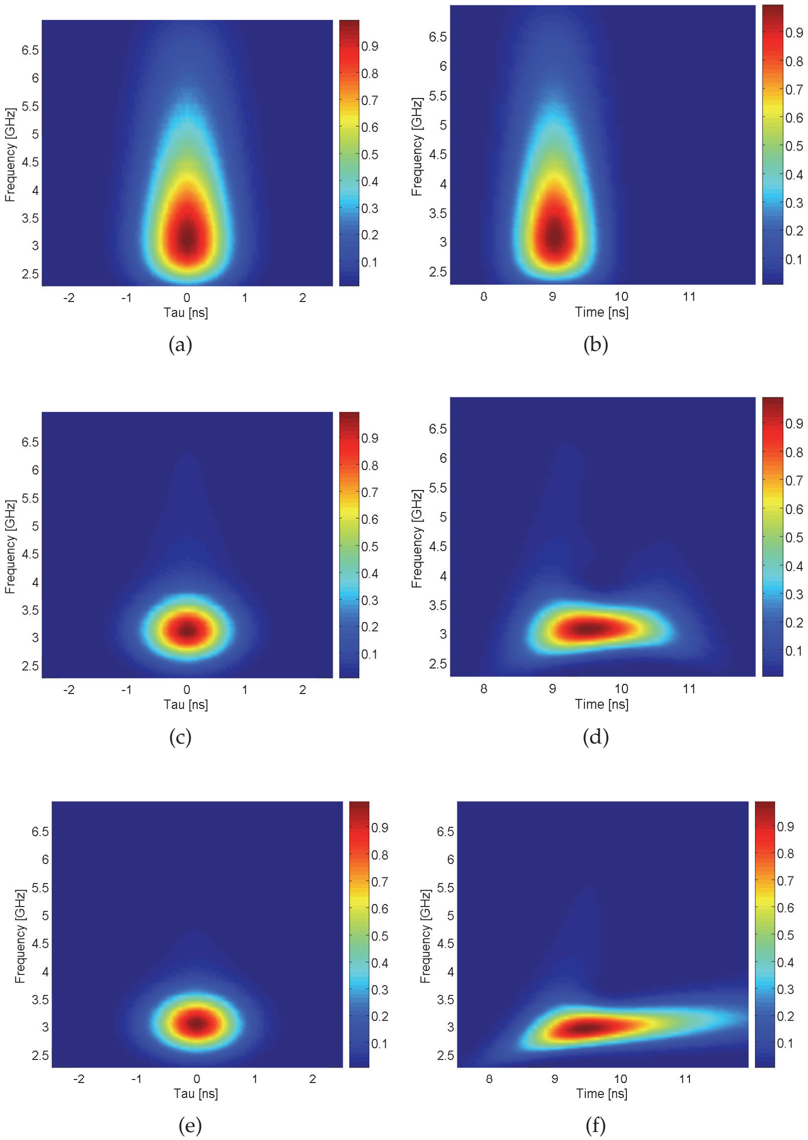
$$v(t) = C_0 \mathbf{Re} \left\{ \exp \left[j2\pi f_0 t - \frac{1}{2}(1 + jC_1) \left(\frac{t - t_0}{\sigma} \right)^{2m} + jC_2(t - t_0)^3 \right] \right\}, \tag{34}$$

where C_0 controls the pulse amplitude, t_0 is the time offset, σ is related to the pulse duration, C_1 and C_2 control the linear and quadric chirp modulation, f_0 is the pulse modulation frequency and m is an integer.

Fig. 20 shows FREG-generated spectrograms. Figs. 20(a) and (c) show the spectrograms of a down-chirped and up-chirped gaussian pulses, respectively. A faithful representation of a linear instantaneous frequency variation is obtained. The spectrogram of a modulated un-chirped super-gaussian pulse is shown in Fig 20(b), where the occurrence of all the frequency components of the signal at the same time instant are clearly seen. Finally, Fig. 20(d) shows the spectrogram of a cubically chirped (down and up) gaussian pulse. The high frequency components occurring at two different times, characteristic of cubically chirped pulses, can be clearly identified. These few examples demonstrate the capability of the proposed FREG system to analyze a wide variety of non-stationary signals.

It is important to point out that a full-wave simulation of the FREG system is extremely time-consuming. Specifically, the FREG system requires multiple analysis of the impulse-regime response of a CRLH LWA, fed by different input signals. Since each of these analysis lasts between $8 - 10$ hours, the simulation of a complete FREG spectrogram may easily lasts few days, which is completely prohibitive. On the other hand, the use of the time-domain Green's functions approach reduces this time to a few ($5 - 8$) minutes. Furthermore, the use of this numerical tool provides a deep insight into the physics of the system, including an electromagnetic modeling of the antenna and a clear understanding of each step of the proposed FREG system.

Finally, a comparison between the FREG and the RTSA systems is given in Fig. 21. The goal of this analysis is to point out the advantages and disadvantages of each method. For the comparison, the same input pulse feeds the RTSA and FREG systems, which are based on identical CRLH LWAs. In the comparison, the number of unit cells N (with size $p = 1.56$ cm and circuital parameters of $C_L = C_R = 1$ pF and $L_L = L_R = 2.5$ nH) of the antenna is modified to perform several tests. The spectrogram results are then given as a function N, i.e. as a function of the total length of the antenna ($\ell = N \cdot p$). For the test, a modulated Gaussian pulse is employed (with $f_0 = 3.0$ GHz and $\sigma = 0.5$ ns). In the figure, the results from the FREG system are placed on the left, whereas the spectrograms computed by the RTSA system are located on the right. First, we set in both systems an antenna with $N = 5$ unit cells, obtaining the spectrograms shown in Fig. 21(a) and Fig. 21(b). This antenna is physically very short, which turns out into a very low directivity. This leads to a very bad frequency resolution



(a) (b)

(c) (d)

(e) (f)

Fig. 21. Spectrograms obtained by the proposed FREG (figures on the left) and RTSA (figures on the right) systems, based on identical CRLH LWAs for the different tests. The antennas are composed of different numbers of N cells, with length $p = 1.56$ cm and circuital parameters of $C_L = C_R = 1$ pF and $L_L = L_R = 2.5$ nH. A modulated Gaussian pulse feeds the systems ($f_0 = 3.0$ GHz, $\sigma = 0.5$ ns). The resulting spectrograms are given for the case of $N = 5$ [(a) and (b)], $N = 20$ [(c) and (d)] and $N = 40$ [(e) and (f)] unit cells.

in both spectrograms. On the other hand, this antenna provides an excellent time-gating trade-off, because the energy is instantaneously radiated, almost without propagation along the structure, leading to a large time resolution. Thereby, the use of a very short antenna leads to generally erroneous spectrograms, due to the wide detection of frequencies which are not part of the input pulse. Second, we modify the CRLH LWA antenna, including now a total of $N = 20$ unit cells. This configuration provides a good frequency resolution in both systems, while the temporal resolution is deteriorated in the RTSA system (due to the use of a longer antenna). The resulting spectrograms are depicted on Fig. 21(c) (FREG) and Fig. 21(d) (RTSA). As it can be observed, the FREG system provides a completely realistic spectrogram, which faithfully reproduces the input signal in terms of frequency and time (location and spreading). On the other hand, the spectrogram obtained by the RTSA has a good frequency resolution, but has some problems dealing with the temporal duration of the pulse. As previously commented, this problem is due to the propagation of the input pulse as it is being radiated, as graphically illustrated in Fig. 15. And third, we simulate the FREG and RTSA systems based on the same CRLH LWA, but composed now of $N = 40$ unit cells. The results are shown in Fig. 21(e) (FREG) and Fig. 21(f) (RTSA). The spectrogram obtained using the FREG system is quite similar to the previous FREG spectrogram ($N = 20$ unit cells), keeping the temporal characteristics but improving the frequency resolution (because a longer antenna provides higher directivity). All relevant features of the input modulated Gaussian pulse, in terms of frequency and time, can easily be extracted from this spectrogram. However, the RTSA system provides a completely wrong result. This is because of the excessive length of the CRLH LWA, which completely destroy the temporal resolution of the system.

The above comparison demonstrates that the proposed FREG system presents important advantages over the RTSA system, specially in terms on temporal resolution, being able to characterize any unknown UWB input signal. Furthermore, this comparison has shown that the RTSA system can only deal with signals whose frequency and temporal characteristics are -at least overall- previously known. On the other hand, the main constrains of the FREG system are the complex equipment required, the requirement of a periodic input signal, and the fact that it is not a completely real-time system.

4. Conclusions

This chapter has introduced an impulse-regime analysis of metamaterial-type transmission lines and antennas. Specifically, a novel formulation, based on Fourier transformations, has been proposed to describe pulse propagation along dispersive linear CRLH lines. The proposed theory is capable to model complex impulse-regime phenomena, such as dispersion, in a simple, accurate and fast way. Then, the method has been extended to consider CRLH leaky-wave antennas, allowing a fast and accurate analysis of the far-field radiation of these structures in time-domain. The proposed formulation has then been applied to *the development of novel phenomena and applications in the microwave domain, most of them transported from optics*. Instead of the usual magnitude engineering and filter design, a dispersion or phase engineering has been applied. In this approach, the dispersive nature and subsequent impulse-regime properties of CRLH structures have been exploited to obtain novel phenomena/applications. Each phenomenon or application proposed has theoretically been described, numerically verified, and in most of the cases, experimentally demonstrated. The shift from narrow band systems (mostly used in the past) to ultra wide band systems, required by current high date rate wireless communication systems, suggests that the forthcoming decades will experience a major interest on this dispersive engineering approach, providing new, novel and more exciting effects and devices at microwaves.

5. References

Abielmona, S., Gupta, S. & Caloz, C. (2007). Experimental demonstration and characterization of a tunable CRLH delay line system for impulse/continuous wave, *IEEE Microwave and Wireless Components Letters* 17(12): 864–866.

Abielmona, S., Gupta, S., Nguyen, H. V. & Caloz, C. (2008). Dispersion engineered impulse regime memataterial devices, *Proc. XXIXth Assembly of Union Radio Science International (URSI)*, Chicago, IL, USA.

Amin, G. & Feng, K. D. (1995). Short-time fourier transforms using cascade filter structures, *IEEE Transaction on Circuits and Systems II, Analog Digital Signal Processing* 20(3): 631–641–408.

Azaña, J. & Muriel, M. A. (1999). Technique for multiplying the repetition rates of periodic trains of pulses by means of a temporal self-imaging effect in chirped fiber gratings, *Optics Lett.* 24: 1672–1674.

Azaña, J. & Muriel, M. A. (2001). Temporal self-imaging effects: theory and application for multiplying pulse repetition rates, *IEEE J. Sel. Top. Quantum Electron* 7: 728–744.

Balanis, C. A. (2005). *Antenna Theory: Analysis and Design. 3rd Edition.*, John Wiley and Sons.

Barton, G. (1989). *Elements of Green's Functions and Propagation*, Oxford Science Publications.

Berger, N. K., Levit, B. S., Bekker, A. & Fischer, B. (2004). Compression of periodic optical pulses using temporal fractional talbot effect, *IEEE Photonics Technology Letters* 16(8): 1855–1857.

Caloz, C. (2009). Perspectives on EM metamaterials, *Materials Today* 12(3).

Caloz, C. & Itoh, T. (2006). *Electromagnetic Metamaterials: Transmission Line Theory and Microwave Applications.*, Wiley and IEEE Press.

Cohen, L. (1989). Time-frequency distributions-a review, *Proceedings of the IEEE* 77(2): 941–981.

Collin, R. E. (1991). *Field Theory of Guided Waves*, IEEE Press, Piscataway, N.J.

Duran-Sindreu, M., Velez, A., Aznar, F., Bonache, J. & Martin, F. (2009). Application of open split ring resonators and open complementary split ring resonators to the synthesis of artificial transmission lines and microwave passive components, *IEEE Transactions on Microwave Theory and Techniques* 57(2): 3395–3403.

Eleftheriades, G. (2009). EM transmission-line metamaterials, *Materials Today* 12(3).

Eleftheriades, G. V. & Balmain, K. G. (eds) (2005). *Negative-Refraction Metamaterials: Fundamental Principles and Applications*, Wiley & IEEE Press, Hoboken, NJ.

Felsen, L. B. (1969). Transients in dispersive media, part I: Theory, *IEEE Transactions on Antennas and Propagation* 17: 191–200.

Ghavami, M., Michael, L. B. & Kohno, R. (2007). *UWB Signals and Systems in Communication Engineering*, J. Wiley & Sons.

Gómez-Díaz, J. S., Gupta, S., Álvarez-Melcón, A. & Caloz, C. (2009a). Impulse-regime CRLH resonator for tunable pulse rate multiplication, *Radio Science* 44(doi:10.1029/2008RS003991): 1–9.

Gómez-Díaz, J. S., Gupta, S., Álvarez-Melcón, A. & Caloz, C. (2009b). Investigation on the phenomenology of impulse-regime metamaterial transmission lines, *IEEE Transactions on Antennas and Propagation* 57(12): 4010–4014.

Gómez-Díaz, J. S., Gupta, S., Álvarez-Melcón, A. & Caloz, C. (2009c). Tunable talbot imaging distance using an array of beam-steered metamaterial leaky-wave antennas, *Journal of Applied Physics* 106: 084908–9.

Gómez-Díaz, J. S., Gupta, S., Álvarez-Melcón, A. & Caloz, C. (2010). Effcient time-domain análisis of highly-dispersive linear and non-linear metamaterial waveguide and antenna structures, *IET Microwaves, Antennas and Propagation* 4(10): 1617–1625.

Gupta, S., Abielmona, S. & Caloz, C. (2009). Microwave analog real-time spectrum analyzer (rtsa) based on the spatial-spectral decomposition property of leaky-wave structures, *IEEE Transactions on Microwave Theory and Techniques* 57(12): 4010–4014.

Gupta, S., Gómez-Díaz, J. S. & Caloz, C. (2009). Frequency resolved electrical gating principle for UWB signal characterization using leaky-wave structures, *39th European Microwave Conference*, Rome, Italy.

Horii, Y., Caloz, C. & Itoh, T. (2005). Super-compact multilayered left-handed transmission line and diplexer application, *IEEE Transactions on Microwave Theory and Techniques* 53(4).

Marques, R., Martín, F. & Sorolla, M. (eds) (2008). *Metamaterials with Negative Parameters: Theory, Design and Microwave Applications*, Wiley, Hoboken, NJ.

Oliner, A. A. & Jackson, D. R. (2007). Leaky-wave antennas, *in* J. L. Volakis (ed.), *Antenna Engineering Handbook*, 4 edn, McGraw-Hill, New York.

Oppenheim, A. V. (1996). *Signals and Systems*, Prentice Hall.

Paul, C. R. (2007). *Analysis of Multiconductor Transmission Lines*, 2^{nd} edition edn, Wiley-IEEE Press.

Pipes, L. A. & Harvill, L. R. (1971). *Applied Mathematics for Engineers and Physicist*, 3^{rd} edn, McGraw Hill.

Pozar, D. (2005). *Microwave Engineering*, 3^{rd} edn, John Wiley and Sons.

Saleh, B. E. A. & Teich, M. C. (2007). *Fundamentals of Photonics*, 2nd edition edn, Wiley-Interscience.

Talbot, H. F. (1836). Facts relating to optical science no. IV, *Philos. Mag.* 9: 401–407.

Trebino, R. (ed.) (2002). *Frequency-Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses*, Springer.

Fourier Transform Application in the Computation of Lightning Electromagnetic Field

Vesna Javor
University of Nis / Faculty of Electronic Engineering
Serbia

1. Introduction

Atmospheric discharge is one of the most interesting and powerful natural phenomenon hiding from men its undiscovered features and secrets for centuries. Lightning discharges have been studied in many theoretical and experimental ways. However, application of Fourier transform has been introduced in this research only in recent few decades. It proved very useful in solving many lightning research problems.

There are two main groups of problems in lightning studies that involve using Fourier transform. One deals with determining how the energy is distributed over a continuous frequency spectrum for the quantity of interest. Channel-base currents, induced voltages and currents, electric and magnetic field components, so as integrals and derivatives of the same functions, distribute components over the entire frequency range in different ways. These are non-periodic functions scattering their energy throughout the frequency spectra.

Another important group of problems to be solved by Fourier transform deals with the calculation of lightning induced effects at different distances from the lightning discharge and risk assessment for buildings, various structures, people and property in an external impulse electromagnetic field. These calculations are based on experimental results for the measured lightning electromagnetic field (LEMF) and channel-base currents, which are used in different types of lightning stroke models including lossy ground effects. Lightning discharge channel is usually modeled by a thin vertical antenna at a lossy ground of known electrical parameters. Both ground and air are treated as linear, isotropic and homogeneous half-spaces. Even for such a simple approximation of the lightning channel - calculations can not be easily done in time domain, but transformation to frequency domain is used instead. Once the calculations are done in frequency domain, way back to time domain is made by Inverse Fourier transform applied to the obtained results.

In both groups of problems an impulse function which has the analytical derivative, integral and integral transformations is very useful. New functions proposed by the author for representing lightning currents are presented in this Chapter, Section 2. These can be also used in other high voltage technique calculations. The main problem for a user of the impulse functions already given in literature to approximate some quantity is the choice of parameters so to obtain desired waveshape characteristics or values adequate to experimentally measured. Parameters of the new channel-base current (NCBC) function (Javor, 2008; Javor & Rancic, 2011) are calculated according to IEC 62305 standard lightning currents (International Electrotechnical Commission, Technical Committee 81 [IEC, TC 81], 2006), and the procedure to choose function parameters is also explained in (Javor & Rancic, 2011). Functions to represent other typical lightning currents are proposed in (Javor, 2011b, 2011c), such as long stroke current (LSC), and two-rise function (TRF) as a multi-peaked current. These functions can be used to obtain the desired peak value, rise-time, decaying time to half of the peak value, current steepness, integral of the function (representing also impulse charge), or integral of the square of the function (representing also specific energy), etc. Fourier transform is obtained analytically and the results are presented in Section 3. Application of the Fourier transform in LEMF computation is shown in Section 4. Based on these results, some conclusions are given in Section 5.

2. NCBC function

Any mathematical function capable to approximate the impulse lightning channel-base current in electromagnetic field calculations can be also used to represent far lightning electromagnetic field as an external excitation inducing currents and voltages in objects and structures inside such a field. Such functions are necessary in lightning stroke modeling.

There is an overview of lightning return-stroke models in (Rakov & Uman, 1998, 2006) where they are classified into four classes according to the type of governing equations. An adequate return-stroke model would be the one that provides simultaneously an approximation to the experimentally measured channel-base current, to the lightning electric and magnetic field waveshapes and intensities at various distances, and to the observed return-stroke speeds. Functions implied in these models are based either on the double-exponential (DEXP) function (Bewley, 1929; Bruce & Golde, 1941) or Heidler's function (Heidler, 1985). DEXP first order derivative at t=0^+ is too large, thus causing numerical problems in LEMF calculations. It also has physically non-realistic convex waveshape in the rising part. Heidler's function reproduces concave rising part and its derivative is equal to zero at t=0+. It is also used for representing lightning currents in the International standard IEC 62305 (IEC/TC 81, 2006). In order to obtain better agreement with experimental results (Berger et al., 1975) one function was proposed (Nucci et al., 1990) as a linear combination of the DEXP and Heidler's function. Another pulse function was also proposed in (Feizhou & Shange, 2004). All these functions need peak correction factors, and their parameters cannot be easily chosen according to the desired waveshape. However, NCBC function parameters in the rising part can be chosen independently from parameters in the decaying part. The exact rising time to the peak value can be chosen in advance, and the decaying time to half of the peak value can be selected in the approximation procedure. The maximum value of the function can be chosen without the peak correction factor. Maximum current steepness can be adjusted (Javor & Rancic, 2011) using analytical expression for the first derivative. Experimentally measured impulse charge and specific energy values (Berger et al., 1975; Anderson & Eriksson, 1980) can be achieved using analytically obtained integral of NCBC function and integral of the square of the function (Javor, 2011a).

NCBC function (Fig. 1) is given with the following expression:

$$i(t)=\begin{cases} I_m\left(t/t_m\right)^a \exp\left[a(1-t/t_m)\right], & 0\le t\le t_m, \\ I_m\sum_{i=1}^{n} c_i\left(t/t_m\right)^{b_i}\exp\left[b_i(1-t/t_m)\right], & t_m\le t\le\infty, \end{cases} \tag{1}$$

where a and b_i are parameters, c_i coefficients, n the chosen number of expressions in the decaying part, so that the total sum of n weighting coefficients c_i is equal to unit, and t_m is the rise-time to the maximum current value I_m. For n=1, c_1=1 and b_1=b, NCBC function reduces to CBC function (Javor & Rancic, 2006) with four parameters (I_m, t_m, a and b). In the special case, for n=1, a=4 and b=0.0312596735, CBC function reduces to High-Voltage Pulse (HVP) function 1.2/50μs (Velickovic & Aleksic, 1986). Impulse duration time is defined as t_i=t_k-$t_{a'}$, for t_k the time in which the current decreased to half of its peak value (Fig. 1). The rising part of the function and its front rise-time are given in Fig. 2. The front rise-time is defined as t_c=$t_{b'}$-$t_{c'}$, for $t_{b'}$ and $t_{c'}$ determined as the time values corresponding to the points B′ and A′, obtained from intersecting horizontal lines for the maximum I_m and the zero function value (time axis) with the line drawn through the points A, for $i(t)$=0.3I_m (Fig. 2) or for $i(t)$=0.1I_m in some other definitions, and B, for $i(t)$=0.9I_m.

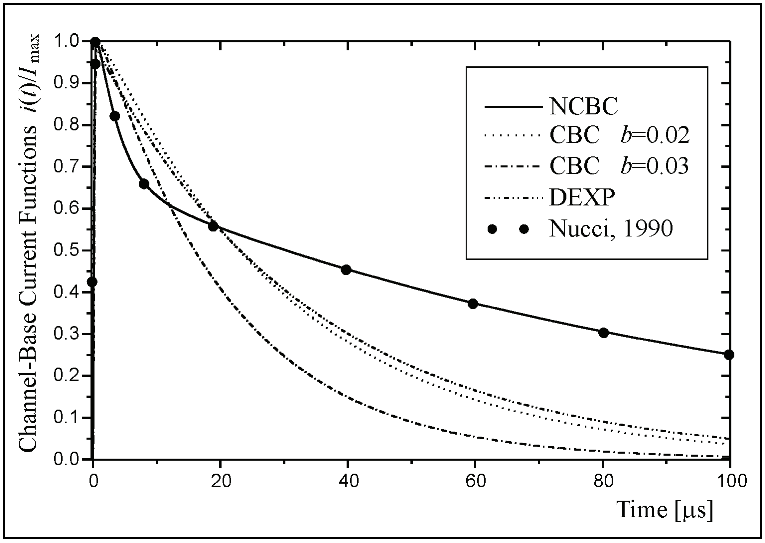


Fig. 1. Normalized NCBC function

NCBC function is an analytically prolonged mathematical function (but still continuous, so as its first derivative, whereas higher order derivatives are not continuous at the point of function maximum I_m), the parameter a in the rising part can be chosen to approximate the front of the waveshape independently from parameters b_i and weighting coefficients c_i in the decaying part, which facilitates the approximation procedure. NCBC function belongs to C^1 differentiability class. Parameters of NCBC function can be chosen so that it represents waveshape of the often used DEXP function with parameters given in (Bruce & Golde, 1941). DEXP function $i(t)$=I_m[exp(-αt)-exp(-βt)] for I_m=11kA, t_m=0.5826μs, α=3·10^{4}s^{-1}, and β=10^{7}s^{-1}, has the decreasing time to half of the peak value of approximately 23μs. It can be approximated with NCBC function for n=1 and I_m=11kA, t_m=0.5826μs, a=0.5 and b=0.02, having also the decreasing time to half of the peak value of about 23μs. If using lightning stroke models and electromagnetic theory relations, lightning electric and magnetic field components above perfectly conducting ground in general have three terms, depending on the integral of the channel-base current function, on the function itself, and on the function derivative. Consequently, the DEXP function having just a convex rising part and great values of the first derivative at $t=0^+$ makes numerical problems in LEMF calculations.

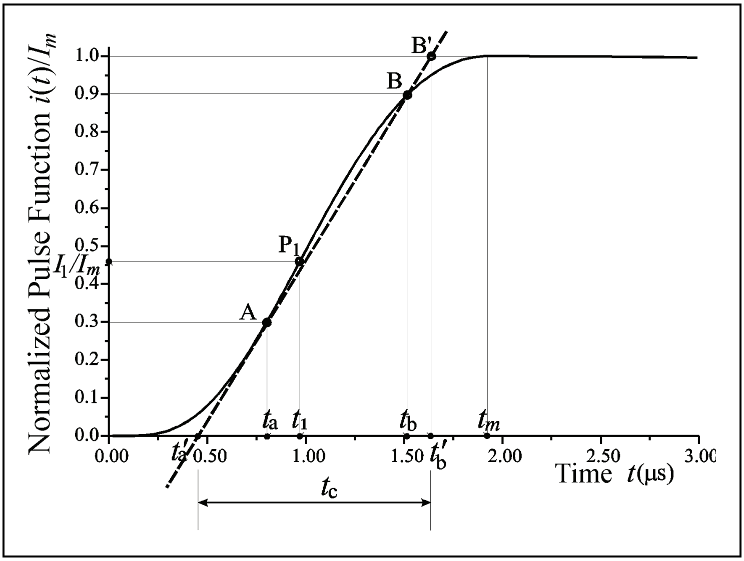


Fig. 2. The rising part of the normalized CBC function representing current t_c/t_i=1.2/50μs

NCBC function, for a > 1, satisfies the demand of having the first derivative equal to zero at $t=0^+$ and the concave to convex rising part. The value of parameter a has to be chosen greater than 1 in order to obtain one saddle point in the rising part (Javor & Rancic, 2006). Heidler's function also has the first derivative equal to zero at $t=0^+$, but it needs peak correction factor, and its rise-time to the maximum current value cannot be chosen in advance. Heidler's function doesn't have analytical integral and its Fourier transform cannot be obtained analytically, but just numerically (Heidler & Cvetic, 2002).

NCBC function can approximate also other channel-base currents used in lightning return-stroke modeling, as the one proposed in (Nucci et al., 1990). This function is used in many papers and approximates well experimental results for subsequent negative strokes. In order to represent the same current two terms are needed in the decaying part of NCBC function (n=2) and its parameters are calculated as I_m=11kA, t_m=0.472μs, a=1.1, b_1=0.16, c_1=0.34, b_2=0.0047, and c_2=0.66.

2.1 Parameters of NCBC function

The normalized NCBC function for different values of parameter a, in the first 1μs is presented in Fig. 3, for t_m=0.5826μs. The parameter a determines the rising part of the function. NCBC function is presented in Fig. 4 for different values of parameter *b*, for t_m=1.9μs, but in a longer time period (300μs), as parameter *b* determines the decaying part of the function. If the rising time to the maximum value t_m has values e.g. 0.5, 1, 2, 5, or 10μs, and other parameters are *a*=1.5 and *b*=0.01, the changes in the function waveshape are presented in Fig. 5.

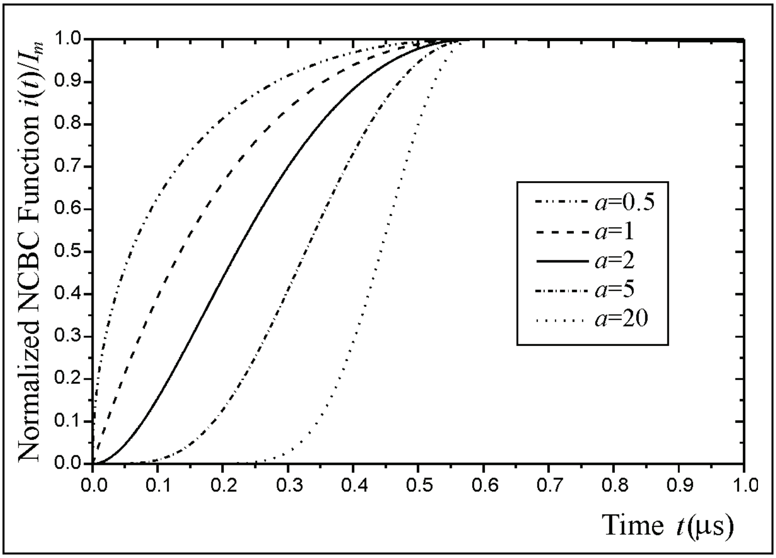


Fig. 3. Normalized NCBC function in the first 1μs, for *b* = 0.03, t_m=0.5826μs, and different values of *a* as parameter

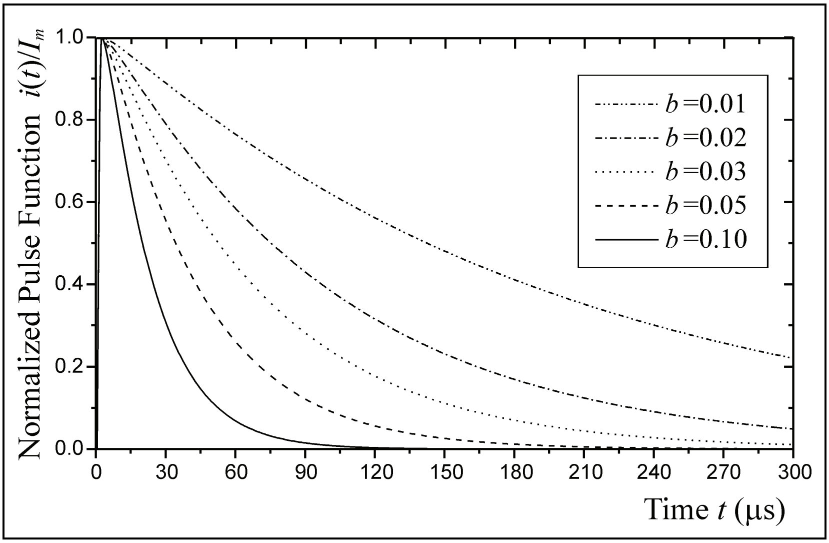


Fig. 4. Normalized CBC function for *a* = 1.5, t_m =1.9μs, and different values of *b* as parameter

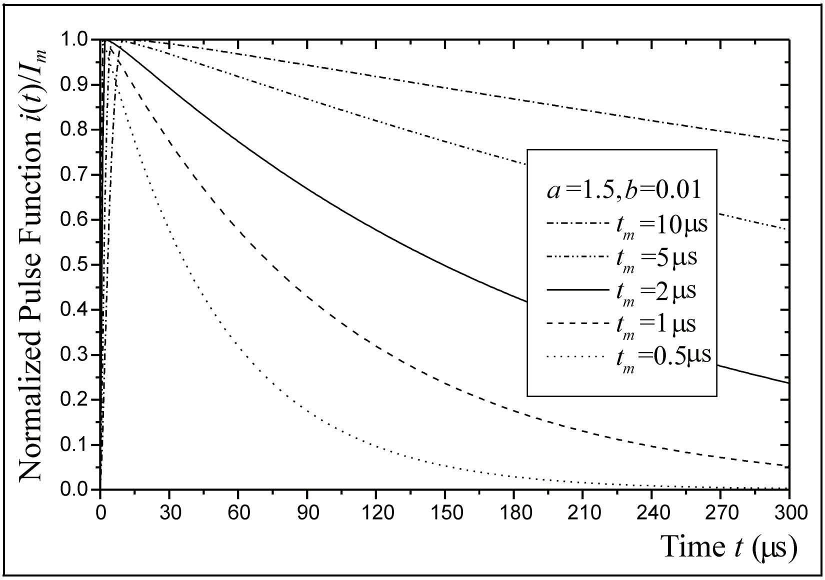


Fig. 5. Normalized CBC function for a = 1.5, b = 0.01, and different values of t_m as parameter

2.2 Derivative of NCBC function

NCBC function first order derivative is:

$$\frac{\mathrm{d}i(t)}{\mathrm{d}t} = \begin{cases} aI_m(t/t_m)^{a-1}(1-t/t_m)\exp[a(1-t/t_m)], & 0 \le t \le t_m, \\ I_m\sum_{i=1}^{n} c_i b_i (t/t_m)^{b_i-1}(1-t/t_m)\exp[b_i(1-t/t_m)], & t_m \le t \le \infty, \end{cases} \tag{2}$$

The first derivative of CBC function (NCBC for n=1) is presented in Fig. 6 up to 1μs, for b = 0.03 (but note that parameter b is irrelevant for the rising part), and different values of a, t_m and I_m.

2.3 Integral of NCBC function

Integral of NCBC function is calculated as:

$$\int_0^t i(t)\mathrm{d}t = \begin{cases} I_m t_m a^{-(a+1)}\exp(a)\gamma(a+1, at/t_m), \ 0 \le t \le t_m \\ I_m t_m \begin{Bmatrix} a^{-(a+1)}\exp(a)\gamma(a+1,a) + \\ \sum_{i=1}^{n} c_i b_i^{-(b_i+1)}\exp(b_i)[\gamma(b_i+1, b_i t/t_m) - \gamma(b_i+1, b_i)] \end{Bmatrix}, \ t_m \le t \le \infty \end{cases} \tag{3}$$

Impulse charge is defined in IEC 62305 standard (IEC/TC 81, 2006) as the integral of the channel-base current function. For standard lightning currents impulse charge is calculated in (Javor, 2011a), so as specific energy as the integral of the square of NCBC function.

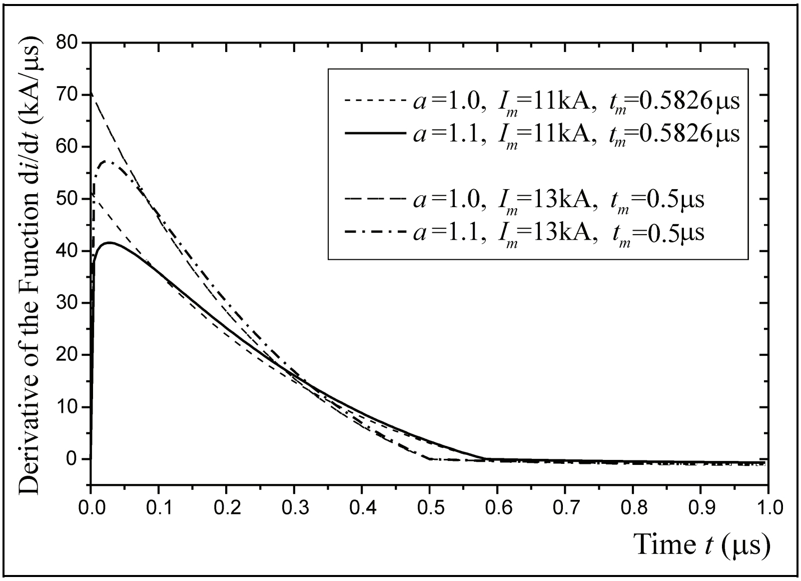


Fig. 6. The first derivative of CBC function in the first 1μs

3. Fourier transform of NCBC function

Whether a function is given analytically, graphically or numerically, its Fourier transform can be obtained analytically, numerically or using some of the commercial programs. Faster or slower rising/decaying of the function and its waveshape at the end of the covered impulse duration time determine the number N of needed points for Fast Fourier Transform.

For NCBC function Fourier and Laplace transforms are calculated analytically. It should be noted that we obtain Fourier transform for any Fourier-transformable function which is zero for $t<0$ if we substitute variable s with $j2\pi f$ in its Laplace transform.

The analytical expression for the unilateral Laplace transform of NCBC function is:

$$I(s) = I_m t_m \left[\frac{\exp(a)}{\left(a + st_m\right)^{a+1}} \gamma(a+1, a+st_m) + \sum_{i=1}^{n} \frac{c_i \exp(b_i)}{\left(b_i + st_m\right)^{b_i+1}} \Gamma(b_i+1, b_i+st_m) \right] \tag{4}$$

for $\gamma(a+1,x) = \int_0^x t^a \exp(-t)\mathrm{d}t$ and $\Gamma(a+1,x) = \int_x^\infty t^a \exp(-t)\mathrm{d}t$ the incomplete Gamma functions, as defined in (Abramowitz, Stegun, 1970).

Fourier transform of NCBC function is obtained from (4) for $s= j2\pi f$, so:

$$I(f)=I_m t_m\left[\frac{\exp(a)}{\left(a+\mathrm{j}2\pi f\,t_m\right)^{a+1}}\gamma(a+1,a+\mathrm{j}2\pi f\,t_m)+\sum_{i=1}^{n}\frac{c_i\exp(b_i)}{\left(b_i+\mathrm{j}2\pi f\,t_m\right)^{b_i+1}}\Gamma(b_i+1,b_i+\mathrm{j}2\pi f\,t_m)\right]. \quad (5)$$

It can be obtained also numerically, or by using some computer program such as FAS (Walker, 1996) with the application of FFT (Fast Fourier Transform). Any frequency domain calculations can be done just for the positive frequencies, as for FFT of real functions the following relations are valid:

$$\mathrm{Re}\{\underline{I}_k(f)\}=\mathrm{Re}\{\underline{I}_{N-k}(f)\} \quad (6)$$

and

$$\mathrm{Im}\{\underline{I}_k(f)\}=-\mathrm{Im}\{\underline{I}_{N-k}(f)\}\,. \quad (7)$$

This feature makes the time for computing FFT twice shorter for real functions, which is very useful if the number of FFT points is great so that calculations are time-consuming. For smaller number of FFT points the computing time is shortened. FFT results given for NCBC function representing lightning current indicate that the major part of its power is in the lower frequency range. The modulus of FFT decays fast as the function of frequency, so it is better to use some other than linear scaling of frequencies in order to cover the lower part of the frequency range. For very high frequencies, the values of real and imaginary parts of FFT of the pulse functions near the end of the frequency interval can be taken as zeros, as they are relatively very small comparing to the values at lower frequencies. That doesn't make much influence on computation results if IFFT (Inverse Fast Fourier Transform) is done for obtaining time domain results based on frequency domain calculations. The sufficient number of FFT points for lightning research studies should be 1024, but better results are obtained with 2048 or 4096 points. This depends on the time interval to cover and the current waveshape.

Fourier transform of Heidler's function is calculated approximately (Andreotti et al., 2005). Some FFT results for Heidler's function are also given in (Vujevic & Lovric, 2010). NCBC/CBC have analytical solutions for Fourier transform which enables obtaining analytical solution for LEMF results in the case of perfectly conducting ground.

The modulus of Fourier transform of HVP function 1.2/50μs is presented in Fig. 7, for $\underline{I}(f)$ normalized to $\underline{I}(0)$, as the function of $\log_{10}f$. For f_2≈6.4kHz≈$10^{3.8}$Hz the normalized modulus of FFT is $|\underline{I}(f_2)/\underline{I}(0)|$≈0.3657. The results are obtained for FFT calculated in N=8192 points, for the time step ΔT≈19.064ns, frequency step Δf≈6.4kHz, and limit frequency f_g≈52.455 MHz. From Fig. 7 is obvious that using frequencies higher than a few MHz is not necessary.

For some frequencies up to the chosen limit frequency the results are presented in Table 1 for FFT of HVP function 1.2/50μs calculated in N=8192 points, for ΔT≈19.064ns, Δf≈6.4kHz, and f_g≈52.455MHz. For the chosen number of points for FFT, the corresponding time interval is $T=N\Delta T$≈156.17μs and the frequency step is $\Delta f=(N\Delta T)^{-1}=T^{-1}$. For such values there are 8192 frequencies in the chosen frequency interval [$-0.5f_g$,$+0.5\,f_g$], and the highest positive frequency is f_{4097}≈26.2275MHz. FFT results for $|\underline{I}(f)|\approx 10^{-n}|\underline{I}(0)|$ for n=1, 2, 3, …, 8, are also given in Table 1, corresponding approximately to frequencies f_5, f_{35}, f_{108}, f_{352}, f_{2478}, f_{3912}, f_{4078}, f_{4095}, and they also point out to the fast decaying of FFT modulus. As (6) and (7) are valid for the real functions, this makes the time for computing twice shorter. This is very useful if the number of FFT points is great and relevant calculations in frequency domain are time-consuming. For smaller number of FFT points the total computing time is shortened, but the sampling interval in time domain is critical in that case.

f_k	f(Hz)	Re{$\underline{I}(f)$}	Im{$\underline{I}(f)$}	$\lvert\underline{I}(f)/\underline{I}(0)\rvert$
f_1	0.0000E+00	0.6232E-04	0.0000E+00	0.1000E+01
f_2	0.6403E+04	0.6892E-05	-0.2172E-04	0.3657E+00
f_3	0.1281E+05	0.8713E-06	-0.1179E-04	0.1896E+00
f_4	0.1921E+05	-0.2839E-06	-0.7867E-05	0.1263E+00
f_5	0.2561E+05	-0.6610E-06	-0.5842E-05	0.9433E-01
f_{35}	0.2177E+06	-0.6334E-06	-0.6164E-07	0.1021E-01
f_{108}	0.6851E+06	0.4510E-07	0.4597E-07	0.1033E-02
f_{352}	0.2248E+07	0.4931E-09	0.6259E-08	0.1007E-03
f_{2478}	0.1586E+08	0.8158E-12	0.6237E-09	0.1001E-04
f_{3912}	0.2503E+08	-0.2014E-12	0.6261E-10	0.1005E-05
f_{4078}	0.2611E+08	-0.1417E-12	0.6350E-11	0.1019E-06
f_{4095}	0.2621E+08	-0.1377E-12	0.6689E-12	0.1096E-07
f_{4097}	0.2623E+08	-1.3765E-13	0.0000E+00	2.2088E-09

Table 1. FFT of the HVP function in 8196 points

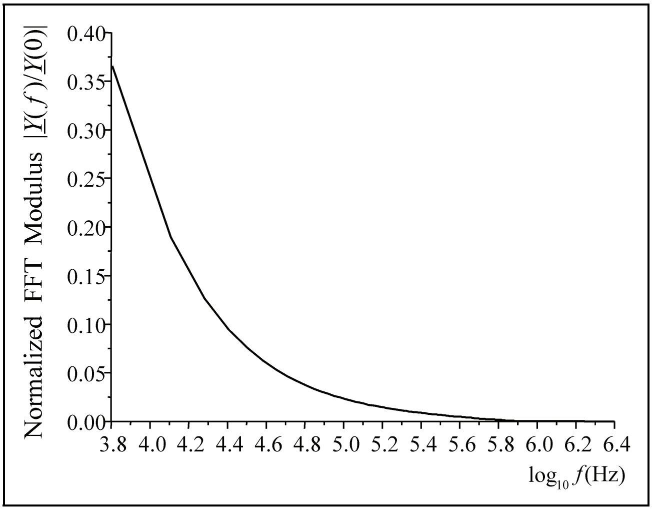


Fig. 7. Normalized FFT modulus of HVP as the function of frequency

For example, for N=2048 and the same time interval $T=N\Delta T\approx 156.172\mu s$ the sampling frequency is $f_g=1/\Delta T\approx 13.114MHz$, which gives the same frequency step $\Delta f\approx 6.4kHz$, but four times greater step in time domain $\Delta T\approx 76.256ns$ than if N=8092.

Fig. 8 shows real part of FFT for HVP function in the considered frequency range [-4096Δ*f*, +4096Δ*f*], and enlarged rectangle area for the range of frequencies around zero *f*∈[-200Δ*f*, +200Δ*f*]=[-1280.638kHz,+1280.638kHz]. It can be seen that Eq. (6) is valid.

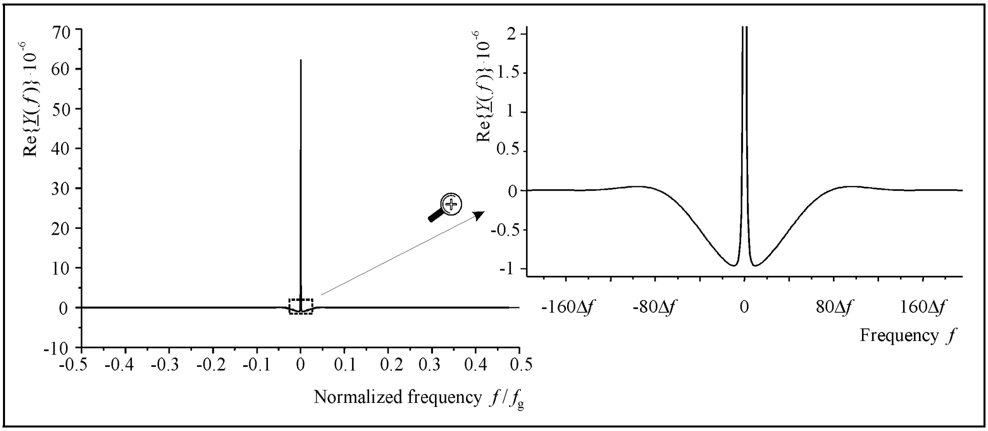


Fig. 8. Real part of FFT for HVP in the frequency range [-0.5f_g,+0.5f_g]=[-4096Δ*f*,+4096Δ*f*] and enlarged rectangle area for frequencies *f*∈[-200Δ*f* ,+200Δ*f*]

Fig. 9 shows imaginary part of FFT for HVP function in the same considered frequency range [-0.5f_g,+0.5f_g], and enlarged rectangle area for *f*∈[-200Δ*f*, +200Δ*f*]. It can be seen that Eq. (7) is valid.

FFT results given in Table 1 for this pulse function indicate that the major part of its power is in the lower frequency range. The modulus of FFT decays rapid as the function of frequency which can be noticed both in Table 1 and Fig. 7.

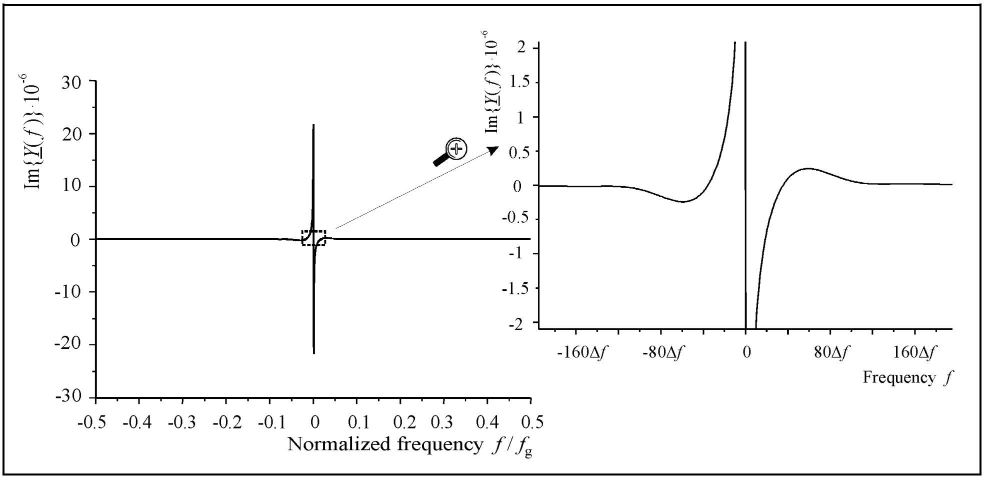


Fig. 9. Imaginary part of FFT for HVP in the range [-0.5f_g,+0.5f_g]=[-4096Δ*f*,+4096Δ*f*] and enlarged rectangle area for frequencies *f*∈[-200Δ*f* ,+200Δ*f*]

As the value $|\underline{I}(f_2)/\underline{I}(0)|\approx0.3657$ is obtained already for the frequency $f_2\approx6.4$kHz (Fig. 7), it is better to use some other then linear scaling of frequencies in order to cover better the lower part of the frequency range. For very high frequencies, the values of real and imaginary part of FFT near the end of the frequency interval can be taken as zeros, as they are relatively very small comparing to their values at low frequencies. That doesn't make much influence on the computation results if IFFT (Inverse Fast Fourier Transform) is done for obtaining time domain results based on frequency domain calculations. Both FFT and IFFT can be done using program FAS (Walker, 1996) in 128, 256, 512, 1024, 2048, 4096 or 8192 points. The input data can be given as analytical functions (as in the case of NCBC or other channel-base current functions) or discrete values in the corresponding set of values for that quantity (as in the case of experimentally measured data).

In order to analyze lightning electromagnetic field in a far zone, as an external excitation of some receiving antenna structure, vertical electric field and azimuthal magnetic field can be represented with the waveshape of NCBC function according to the measured data and its Fourier transform should be known. Fourier transform of the channel-base current is also necessary if calculating equivalent voltage source as a product of the input impedance and the channel-base current in frequency domain (Moini et al. 2000, Shoory et al. 2005), instead of the current source itself at the channel-base. A few computer programs for antenna analysis are used for calculations (Richmond, 1974, 1992; Bewensee, 1978; Burke & Poggio, 1980, 1981; Djordjevic et al., 2002), and most of these use the voltage generator as excitation (Grcev et al., 2003).

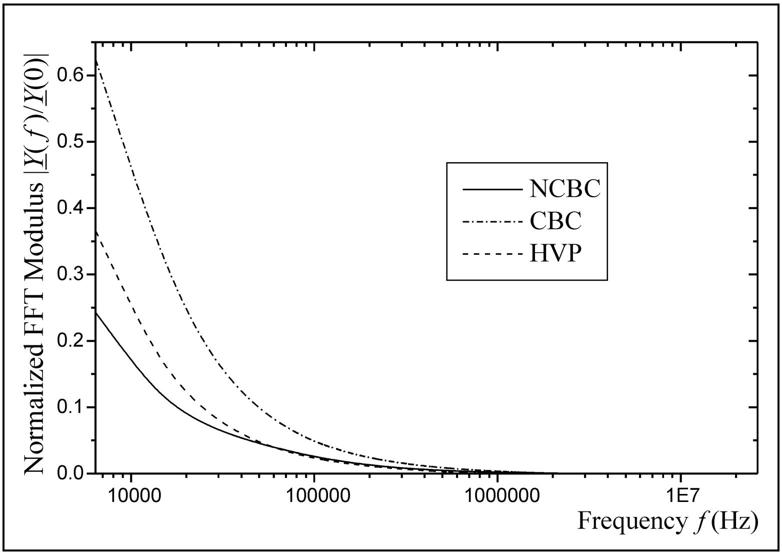


Fig. 10. Normalized FFT modulus for NCBC, CBC and HVP as a function of frequency

For CBC function with parameters I_m=11kA, t_m=0.5826μs, a=1.5 and b=0.02 the results of FFT are presented in Figs. 10-12. This function is chosen adequate to the DEXP, except for the parameter a = 1.5 instead of a = 0.5 in the rising part, in order to satisfy the mentioned condition for lightning discharge channel-base current functions (having the first derivative equal to zero at t=0$^+$ and the concave to convex shape in the rising part). In the same figures there are the results for the new channel base current function (NCBC) with parameters I_m=11kA, t_m=0.472μs, a=1.1, b_1=0.16, c_1=0.34, b_2=0.0047, and c_2=0.66. Just for the comparison,

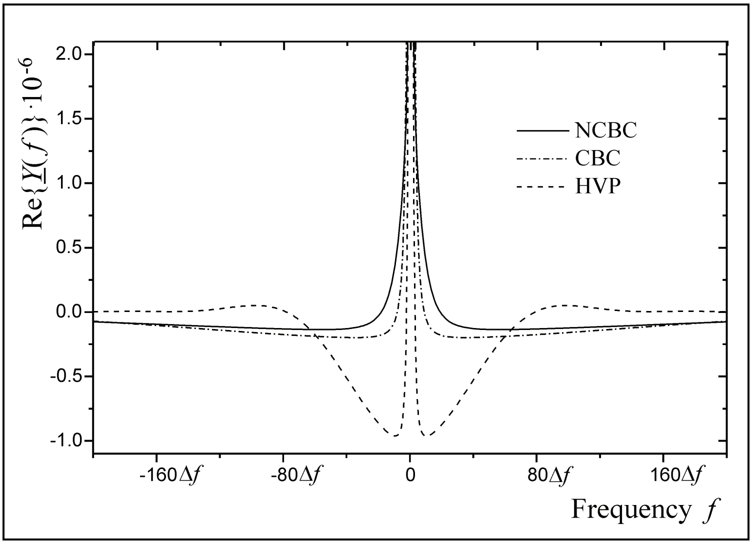


Fig. 11. Real part of the normalized FFT for NCBC, CBC and HVP functions in the frequency range [-200Δf,+200Δf]

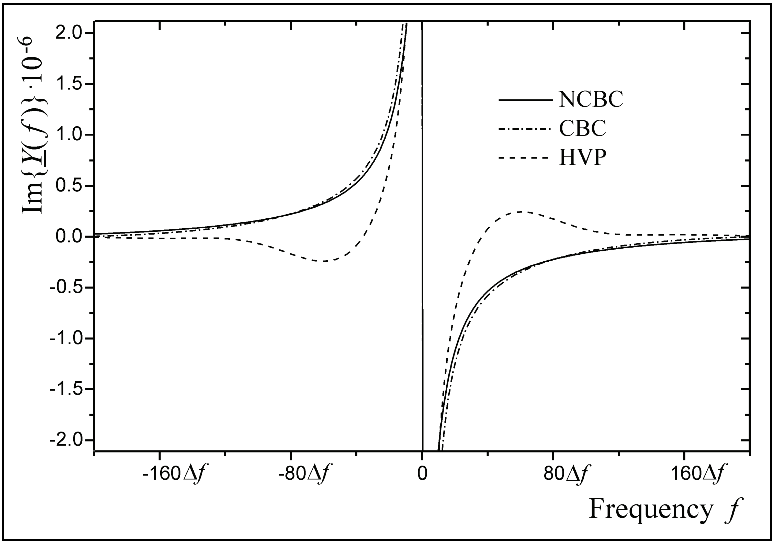


Fig. 12. Imaginary part of the normalized FFT for NCBC, CBC and HVP functions in the frequency range [-200Δf,+200Δf]

the results are given in these figures also for the high voltage pulse (HVP) function 1.2/50μs, although the function with those parameters is not adequate for lightning channel-base currents modeling. All the results are presented for the functions normalized to the maximum values. The results for moduli of FFT for $\underline{I}(f)$ normalized to $\underline{I}(0)$, as the function of frequency, are presented in Fig. 10. For f_2≈6.4kHz the normalized modulus of FFT for NCBC function is $|\underline{I}(f_2)/\underline{I}(0)|$≈0.2426, for CBC function $|\underline{I}(f_2)/\underline{I}(0)|$≈0.6242, and for HVP function $|\underline{I}(f_2)/\underline{I}(0)|$≈0.3657.

Real and imaginary parts of FFT in the frequency range fє[-200Δf ,+200Δf] are presented in Figs. 11 and 12.

Log-log dependence of FFT modulus on frequency is presented in Fig. 13 for CBC function with parameters I_m=11kA, t_m=0.5826μs, a=1.5 and b=0.02 and for NCBC function with parameters I_m=11kA, t_m=0.472μs, a=1.1, b_1=0.16, c_1=0.34, b_2=0.0047, and c_2=0.66.

Normalized FFT modulus as the function of frequency for NCBC with parameters I_m=25.176kA, t_m=0.65μs, a=20, b=0.00467, and for Heidler's function with parameters I_0=25kA, η=0.993, τ_1=0.454μs, τ_2=143μs, both representing IEC 62305 standard negative first stroke lightning current 0.25/100μs for lightning protection level (LPL) III-IV, is given in Fig. 14.

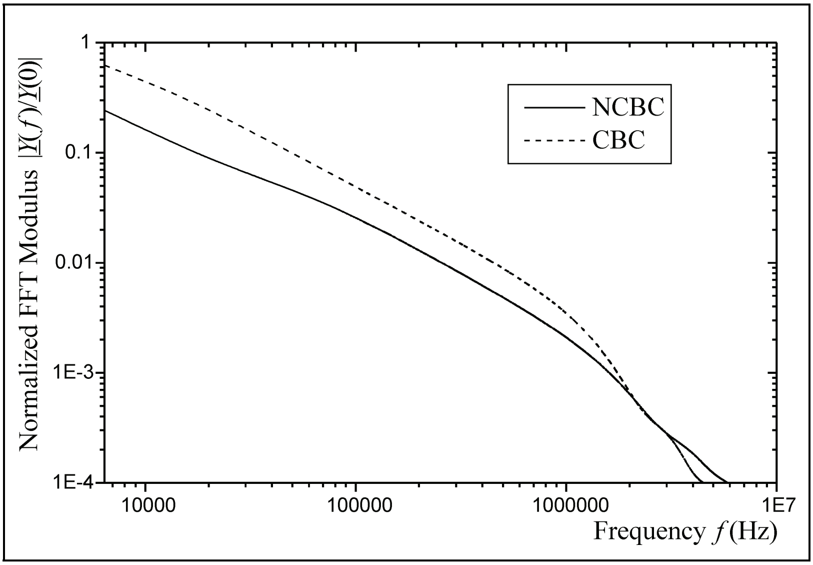


Fig. 13. Normalized FFT moduli for NCBC and CBC functions as the function of frequency

The results show that for frequencies higher than 10MHz FFT modulus of these channel-base current functions is less than 1‰ of its value at 10kHz, so for frequencies higher than 10MHz calculations of LEMF are not needed in frequency domain.

FFT results for NCBC function given in Fig. 14 are also in good agreement with the results given in (Vujevic & Lovric, 2010) for the corresponding Heidler's function with parameters I_0=25kA, η=0.993, τ_1=0.454μs, τ_2=143μs, and for LPL III-IV.

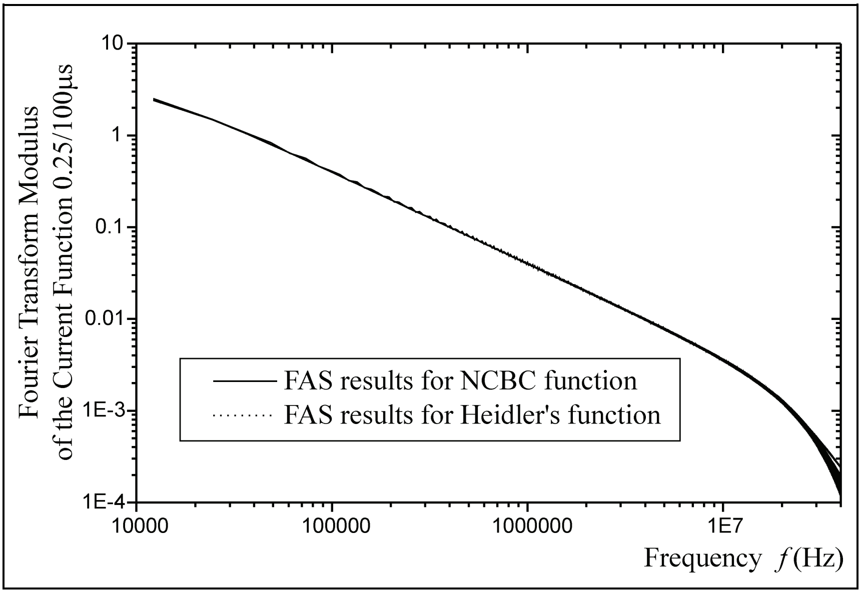


Fig. 14. Normalized FFT moduli for NCBC and Heidler's functions representing standard 0.25/100μs as the function of frequency

4. Electromagnetic field calculation using antenna model of the lightning channel at a lossy ground

Based on experimentally measured characteristics of natural lightning (Berger et al., 1975; Lin et al. 1979; Anderson & Eriksson 1980) and also artificially initiated lightning discharges, LEMF can be estimated using some of the models from literature (Rakov & Uman, 1998, 2006), and lightning currents are assumed to propagate with attenuation and distortion while distributing charge along the channel. In the case of perfectly conducting ground calculations are simple in time domain, but full-wave approach in the case of a lossy ground is complex to implement, so Fourier transform is applied. After calculations are done in frequency domain, the conversion to time domain is performed by Inverse Fast Fourier transform (IFFT). A solution of the Sommerfeld's problem is required for the lossy ground. A few alternatives are proposed in literature.

4.1 Alternatives to full-wave time domain computation

The problem of LEMF calculation can be easily solved directly in time domain if the ground is treated as perfectly conductive. In such a case there exist vertical electric and azimuthal magnetic field components in the observed point at the ground surface. Horizontal component of electric field is zero at the perfectly conducting ground surface, but non-zero above the ground surface. However, it exists above, under, and at the surface of the lossy ground. Vertical component of the lightning electric and azimuthal component of the magnetic field can be easily (but in that case approximately) determined at the distances greater than a kilometer under the assumption of perfectly conducting ground. For smaller distances, propagation above ground of finite conductivity results in the noticeable attenuation of the high-frequency components of electric and magnetic field, and thus in appearance of the horizontal electric field at the surface. Finite conductivity has greater impact on horizontal than on vertical electric field, so calculation of horizontal component requires rigorous computation or, at least, acceptable approximations.

Approximate formulas in frequency domain are often used for calculation of the horizontal electric field in air, up to heights of tens of meters above the ground surface. These formulas can be integrated in the calculation of LEMF in time domain, but the obtained expressions are much more complex. There are simple approximations: the assumption of perfectly conducting ground, "wavetilt" formula, Cooray's approach and Rubinstein's approach. Cooray, 1992, proposed the calculation of horizontal electric field at the surface of finitely conductive ground using azimuthal magnetic induction and the expression for ground surface impedance. He showed that this simple formula provides very accurate results at the distances of about 200m. Rubinstein, 1996, proposed expression for the horizontal electric field with two terms: 1) horizontal electric field calculated under the assumption of perfectly conducting ground, and 2) the correction factor, given as a product of the magnetic field calculated under the same assumption and the function similar as in "wavetilt" formula which represents the effect of finite conductivity. The basic assumption of Rubinstein's approximation is $\sigma_1 >> \varepsilon_0\varepsilon_{r1}$, and that finite ground conductivity does not affect the horizontal magnetic field at the surface. If this is not the case, then more general formula can be written, known in literature as the Cooray-Rubinstein's formula (Cooray, 2002). Wait gave generalization of Cooray-Rubinstein's formula and the exact evaluation of horizontal electric field, showing under which circumstances this general expression reduces to Cooray-Rubinstein's formula (Wait, 1997). Cooray and Lindquist, 1983, and Cooray, 1987, using the attenuation function in time domain proposed by Wait, 1956, included effects of the finite conductivity, and obtained results for the electric field that are in better agreement with experiments. Terms for approximate formulas in time domain are complex, so the approach in frequency domain is preferred.

4.2 Method of images

Method of images gives an approximate solution to the Sommerfeld's integral. Complex image technique often uses one or more terms of exponential series to approximate plane wave reflection coefficient. Thus, multiple discrete and/or continuous image sources are introduced, and this technique also proved not to be limited to a quasi-static range alone. Using this technique, different authors (Shubair & Chow, 1993; Yang & Zhou, 2004; Popovic & Petrovic, 1993; and Petrovic, 2005) obtained results for Sommerfeld's integral kernel for vertical dipoles above a lossy half-space which were used for the comparison with the new Two-image approximation (TIA). Approximate formulas are often valid for a limited range of ground electrical parameters, field point distances, or heights of dipoles above ground, but TIA approximation of Sommerfeld's integral kernel proposed in (Rancic & Javor, 2006 & 2007) has the advantage of being valid in a wide range of lossy ground electrical parameters, for various heights of vertical electric dipoles above the ground, and for possibility to calculate electromagnetic field in both near and far zone.

LEMF can be calculated using thin wire antenna modeling of a lightning channel assumed as vertical, without branches and reflections from the end. If using electromagnetic model, boundary conditions are fulfilled at the wire antenna surface, and a voltage or current source is assumed at the channel-base. Unknown current distribution along the antenna is determined by solving some system of integro-differential equations of EFIE or MFIE type satisfying boundary condition for electric or magnetic field components, respectively. One of those is e.g. System of integral equations of two potentials (SIE-TP), (Rancic, 1995), which is of EFIE type. Current distribution along the antenna can be approximated e.g. by polynomials (Popovic, 1970) with unknown complex coefficients. These can be determined by using Method of Moments (MoM) (Harrington, 1968) i.e. by matching in a sufficient number of points along the antenna. Based on that current distribution LEMF is calculated using electromagnetic theory relations. Thus, the problem is solved in full (without mentioned approximations of some LEMF components using others) on the basis of approximate solution of Sommerfeld's problem of a dipole radiating at the arbitrary height above the lossy half-space, which is a classical problem in electromagnetics.

The problem of a dipole radiation in the presence of a lossy medium was noticed yet in 1909, by Sommerfeld, who determined the solution in the form of superposition of inhomogeneous plane waves. His work had such a great influence on later theoretical research in this area that the solution of more complex problems which shows the influence of a lossy medium on the properties of linear antennas is marked as Sommerfeld's solution. Sommerfeld was the first who treated all four cases of elementary Hertz's dipoles in the air applying his formulation through the cylindrical waves. Van der Pol, 1931, proposed approximate method for solving Sommerfeld's integrals. Class of integrals obtained by Sommerfeld's formulation which strictly defines boundary conditions on the surface of discontinuity is known in literature as "integrals of Sommerfeld's type". These integrals have limits of integration 0 and infinity, and the integrand is a product of Bessel's function (J_0 or J_1), exponential, and one more function, so it is of very complex shape. Depending on the complexity of the model, this function has different forms so that different problems appear in its numerical integration. Since these integrals are highly oscillatory and slowly decreasing functions they don't have solutions in the closed form. In the analysis of antennas above lossy media, there is an inevitable and major problem to determine Sommerfeld's integral as accurate as possible in a wide range of values of electrical parameters of the medium, for different positions of current elements, different frequencies and positions of the observed point in the field. There are many different ways to approximate Sommerfeld's integral which can be found in literature. The model that includes complex mathematical functions in the case of lossy ground gives result equivalent to the field of infinite number of current images. Simpler approximations have limitations as for the values of electrical parameters of the medium, the position of the point in the field (near or far zone) and positions of matching points near or at the interface of two half-spaces. In order to include the influence of electrical parameters of a real ground on antenna characteristics, direct Sommerfeld's approach to the problem can be also used, but it is very complex for solving from both the analytical and numerical point of view. In addition to the direct approach to this problem, it is possible to modify the path of integration of Sommerfeld's integral, or to use interpolation and speed up the calculation, or to use tables of Sommerfeld's integrals, but there were also some attempts to solve the problem by obtaining approximations in closed form for some special cases of Sommerfeld's integrals. A suitable solution would be the one to determine efficiently Sommerfeld's integrals for arbitrary positions and orientations of dipoles, arbitrary positions of the points in LEMF and a wide range of frequencies.

One approximate solution is obtained by use of approximation of the spectral reflection coefficient of a plane wave with exponential series, having one or more terms which is named in literature as the method of images. For a horizontal electric dipole one complex image is the simplest approximation, which was first introduced by Wait, 1955, and later justified in (Wait & Spies, 1969). An important contribution was given by Bannister, (Bannister, 1966), who also showed that techniques of the theory of images in the case of finitely conductive ground are not restricted just to a quasi-static extent of the problem (Bannister, 1978). He later published a review paper "Summary of image theory expressions for the quasi-static fields of antennas at or above the earth's surface" (Bannister, 1979). Bannister extended the validity of the approximation to high-frequency problems by introducing exponential function exp($-u_0\underline{d}$) where $u_0=(\alpha^2+\underline{\gamma}_0^2)^{1/2}$ and $\underline{\gamma}_0$ is the complex propagation constant in the air. Thus the effects of propagation are included, the same solution obtained, and for both approximations there is the assumption that the refraction coefficient of ground-to-air is much greater than 1 ($\underline{n}_{10}=\underline{\gamma}_1/\underline{\gamma}_0>>1$). Mahmoud & Metwally, 1981, used discrete, so as the combination of discrete and continuous images. Mahmoud, 1984, extended the theory to multiple images and to a layered ground. Distributed images were also used and named “the exact theory of continuous images” in (Lindell & Alanen, 1984). By using Prony’s method and nonlinear optimization technique Chow et al., 1991, analytically expressed Sommerfeld's integrals in closed form of spatial complex images. Shubair & Chow, 1993, discussed the impact of complex images on the problem of vertical antennas in the presence of lossy media and took advantage of the spatial Green's function in closed form in order to obtain a superposition of the sources influence, quasi-dynamic image and three complex images. Arand et al., 2003, used the method of discrete complex images and Generalized pencil of function technique to obtain the locations and impact of current sources. A procedure to approximate Sommerfeld’s integrals was carried out by Popovic & Petrovic, 1993, named "a simple near-exact solution" which uses a few images determined so to approximately satisfy the boundary conditions in a limited area of the interface and to yield accurate field values in the domain of the antenna. Electromagnetic models including influence of the lossy ground on LEMF calculation often use approximations of Sommerfeld’s integral kernel on the basis of theory of images in frequency domain. Takashima et al., 1980, proposed a modified theory of images for smaller distances from the point of observation to the vertical electric dipole. As in lightning research is necessary to treat also larger distances from the point of observation to the vertical electric dipole, this approximation can not satisfy the needs of computation in the range of distances of interest. For larger distances from the point of observation to the vertical electric dipole complete expressions, derived by King, 1990, were used for EM field of a vertical electric dipole over the lossy medium, if the condition $|\underline{\gamma}_2|^2>>|\underline{\gamma}_1|^2$ or $|\underline{\gamma}_2|>>3|\underline{\gamma}_1|$ is satisfied. King & Sandler, 1994, confirmed with their results the validity of these expressions over certain types of ground.

The new approximation TIA can be classified into two-image approximations. The basic idea for obtaining a new approximation of the spectral reflection coefficient is matching the approximation of reflection coefficient as a two-terms function (with three unknown constants) at two points, but also matching its first derivative at one point, which resulted in a very efficient approximation of Sommerfeld's integral. TIA gives good results for modeling in both near and far zone, as well as physically conceivable arrangement of images and representation of the problem (as the distances of the images are real values and not complex as in other approximations). TIA is similar to two-image approximations of Sommerfeld's integral, and therefore is their abbreviation retained, but the way of deriving the corresponding expressions is different. Sommerfeld's integral kernel results are compared to the results from literature for different values of lossy ground electrical parameters and various heights of vertical dipoles above the ground (Javor & Rancic, 2009). For the spectral reflection coefficient the following approximation is used

$$\tilde{R}_{z10}(u_0) \cong B + A\,\mathrm{e}^{-(u_0 - u_{0c})d_0}\,, \tag{8}$$

where A and B are unknown complex coefficients, d_0 is the distance from the source to the second image, and u_{0c} is the characteristic value chosen so that $u_{0c}=\underline{\gamma}_0$. Unknown constants are determined from equations obtained by matching the value of spectral reflection coefficient in two points ($u_0\to\infty$ and $u_0=\underline{\gamma}_0$), and its first derivative at $u_0=\underline{\gamma}_0$. This results in $B=R_\infty$, $A=R_0\text{-}R_\infty$ and $\underline{d}_0=\underline{\gamma}_0^{-1}(1+\underline{n}_{10}^{-2})$, for R_∞ and R_0 the quasi-stationary reflection coefficients

$$R_\infty = ({n_{10}}^2 - 1)/({n_{10}}^2 + 1) \tag{9}$$

and

$$R_0 = (n_{10} - 1)\,/\,(n_{10} + 1) \tag{10}$$

Instead of a complex distance of the second image the value is selected as $d_{im}=|\,\underline{\gamma}_0^{-1}(1+\underline{n}_{10}^{-2})\,|$, based on numerical experiments.

4.3 Antenna model of a lightning discharge at a lossy ground

The simplest approximation of a LD channel is a vertical transmitting antenna, with an excitation at its base, positioned at a lossy ground surface as in Fig. 15. For the application of an electromagnetic model and equations of antenna theory, it is necessary to divide this vertical antenna into segments of the length not greater than approximately half of the wavelength corresponding to the frequency for which the analysis is done. Vertical rod antenna with an excitation by an ideal Dirac's voltage source in the channel-base, having voltage $u(t)=U\delta(t)$ and frequency f, is presented in Fig. 15. The total height h of the antenna models the height of a LD channel, which may be as high as several thousands of meters in natural conditions. In such a model is assumed that the antenna is of a circular cross section with the radius a, so that $a<<\lambda_0$, where λ_0 is the wave-length in the air corresponding to the frequency f. The corresponding angular frequency is $\omega=2\pi f$. The antenna is divided into N segments, so that $h=l_1+l_2+...+l_N$, and the length of each segment is $l_k>>a_k$, whereas radius $a_k<<\lambda_0$. It can be simply taken that $a_k=a$ for all segments, for k=1, 2,..., N. The division into segments is necessary because of the wide range of frequencies of interest and for some of those the antenna should be divided into hundreds of segments. The lossy ground is treated as homogeneous, linear and isotropic half-space of electrical parameters: specific conductivity σ_1, electric permittivity $\varepsilon_1=\varepsilon_0\varepsilon_{r1}$, and magnetic permeability $\mu_1=\mu_0$. Other parameters of two half-spaces are defined as: the complex conductivity $\underline{\sigma}_i=\sigma_i+\mathrm{j}\omega\varepsilon_i$, complex propagation constant $\underline{\gamma}_i=(\mathrm{j}\omega\mu_i\,\underline{\sigma}_i)^{1/2}$, for i=0 denoting air and i=1 denoting ground of the relative complex permittivity $\underline{\varepsilon}_{r1}=\varepsilon_{r1}\text{-j}\varepsilon_{i1}=\varepsilon_{r1}\text{-j}60\sigma_1\lambda_0$, and ground to the air refraction index $\underline{n}_{10}=\underline{\gamma}_1/\underline{\gamma}_0=(\underline{\varepsilon}_{r1})^{1/2}$.

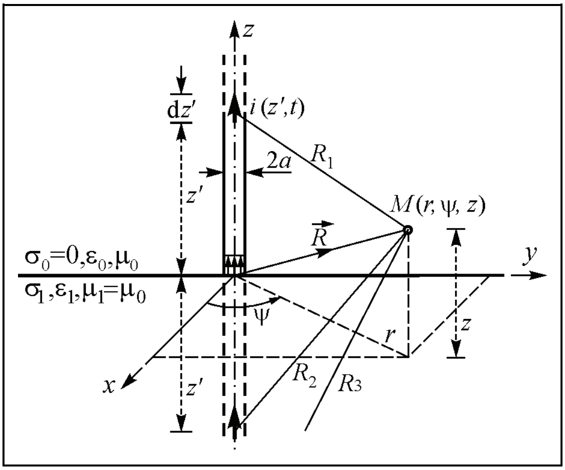

z
dz'
i (z',t)
2a
R_1
$M(r, \psi, z)$
$\vec{R}$
z'
σ_0=0,ε_0,μ_0
σ_1,ε_1,μ_1=μ_0
y
ψ
r
z
x
z'
R_2
R_3

Fig. 15. Lightning channel model at a lossy ground

Electric field is determined based on the current along the antenna from the expression

$$\vec{E} = -\operatorname{grad}\varphi - {\gamma_0}^2\vec{\Pi} = E_\rho\hat{\rho} + E_z\hat{z}\ . \tag{11}$$

Polynomial approximation used to represent the current distribution along *k*-th segment with axis s_k' is given with:

$$I_k(s_k') = \sum_{m=0}^{n_k} B_{mk}(s_k' / l_k)^m , \tag{12}$$

where n_k is the polynomial degree and B_{mk} are unknown complex coefficients to be determined from the system of equations SIE-TP. By satisfying boundary condition for the tangential component of electric field, SIE-TP is obtained as:

$$\gamma_0 \int_{s=0}^{s_n} \varphi_0(s)\operatorname{ch}\left[\underline{\gamma}_0(s_n - s)\right]\mathrm{d}s + {\gamma_0}^2 \int_{s=0}^{s_n} \mathit{\Pi}_{s_n}(s)\operatorname{sh}[\gamma_0(s_n - s)]\mathrm{d}s$$

$$+ \int_{s=0}^{s_n} Z_n'(s)I_n(s)\operatorname{sh}[\gamma_0(s_n - s)]\mathrm{d}s = C_{2n}\operatorname{sh}(\gamma_0 s_n), \tag{13}$$

$$\varphi_0(s_n) + \gamma_0 \int_{s=0}^{s_n} \varphi_0(s)\operatorname{sh}[\gamma_0(s_n - s)]\mathrm{d}s + {\gamma_0}^2 \int_{s=0}^{s_n} \Pi_{s_n}(s)\operatorname{ch}[\gamma_0(s_n - s)]\mathrm{d}s$$

$$+ \int_{s=0}^{s_n} Z_n'(s)I_n(s)\operatorname{ch}[\gamma_0(s_n - s)]\mathrm{d}s = C_{2n}\operatorname{ch}(\gamma_0 s_n)\,. \tag{14}$$

Polynomial approximation is used to represent the current distribution along *k*-th segment with axis s_k', so that in (13) and (14), $\varphi(s)$=-div $\vec{\Pi}_0$=-$\partial\Pi/\partial z_0$ represents the scalar potential at the antenna surface, Π_{sn}(s)= Π_{z0}(s) Hertz vector tangential component, whereas $Z_n'(s)$ is the impedance per unit length, s_n the matching point, and s the local coordinate along the n-th conductor, so that 0≤s≤ l_n. Π_{z0}(s) and φ_0(s) are potentials in the upper half-space to be calculated from:

$$\Pi_{z0}(s)=\frac{1}{4\pi\underline{\sigma}_0}\sum_{k=1}^{N}\int_{s'_k=0}^{l_k} I_k(s'_k)[K_0(r_{1k})+S_{00}^{v}(r_{2k})]\mathrm{d}s'_k\,, \tag{15}$$

$$\varphi_0(s)=\frac{1}{4\pi\underline{\sigma}_0}\sum_{k=1}^{N}\int_{s'_k=0}^{l_k} I_k(s'_k)\frac{\partial}{\partial s'_k}[K_0(r_{1k})-S_{00}^{v}(r_{2k})]\mathrm{d}s'_k\,, \tag{16}$$

for $K_0(r_{1k})=\exp(-\underline{\gamma}_0 r_{1k})/r_{1k}$ the standard potential kernel, and S_{00}^{v} (r_{2k}) the Sommerfeld's integral kernel defined as

$$S_{00}^{v}(r_{2k})=\int_{\alpha=0}^{\infty}\tilde{R}_{z10}(\alpha)\tilde{K}_0(\alpha,r_{2k})\mathrm{d}\alpha. \tag{17}$$

$\tilde{R}_{z10}(\alpha)$ is the reflection coefficient in the spectral domain, for the variable 0≤α<∞, so that

$$\tilde{R}_{z10}(\alpha)=\tilde{R}_{z10}(u_0)=\frac{n_{10}{}^2u_0-u_1}{n_{10}{}^2u_0+u_1}, \tag{18}$$

for $u_1=\sqrt{\alpha^2+\gamma_1{}^2}=\sqrt{u_0{}^2+\gamma_1{}^2-\gamma_0{}^2}$, and $u_0=\sqrt{\alpha^2+\gamma_0{}^2}$, whereas $\tilde{K}_0(\alpha,r_{2k})$ is the spectral form of the standard potential kernel

$$\tilde{K}_0(\alpha,r_{2k})=\frac{\mathrm{e}^{-u_0(z+s_k')}}{u_0}\alpha J_0(\alpha\rho) \tag{19}$$

and

$$K_0(r_{2k})=\frac{e^{-\gamma_0 r_{2k}}}{r_{2k}}=\int_{\alpha=0}^{\infty}\tilde{K}_0(\alpha,r_{2k})\mathrm{d}\alpha, \tag{20}$$

for $r_{2k}=\sqrt{\rho^2+(z+s_k')^2}$ the distance from the field point M (ρ, ψ, z) to the first image of the k-th antenna segment for $\rho=\sqrt{x^2+y^2}$, and $J_0(\alpha\rho)$ the Bessel function of the first kind and order zero.

Thus, for the Sommerfeld's integral kernel (7) is obtained:

$$S_{00}^{v}(r_{2k})\cong BK_0(r_{2k})+A\exp(\gamma_0 d_{im})K_0(r_{3k}), \tag{21}$$

for $r_{3k}=\sqrt{\rho^2+(z+s_k'+d_{im})^2}$ the distance from the second image to the field point, and $K_0(r_{3k})=\exp(-\gamma_0 r_{3k})/r_{3k}$.

LEMF calculations were also done for the antenna models of cage structures (Javor, 2003) and in the case of lightning protection rods at a lossy ground (Javor & Rancic, 2009).

4.4 Current distribution along the vertical antenna at the lossy ground

The results obtained with TIA approximation are shown in Fig. 16 for real and imaginary part of the current along the vertical antenna having height h=300m, radius a=0.05m, and resistivity per unit length R'=0.1Ω/m, at the lossy ground of relative dielectric constant ε_{r1}=2 and ε_{r1}=10, specific conductivity σ_1=10^{-1}S/m or 10^{-5}S/m, for frequency f=3MHz, the number of antenna segments N=30, and the degree of polynomial approximation n_k =3, for k=1,...,30.

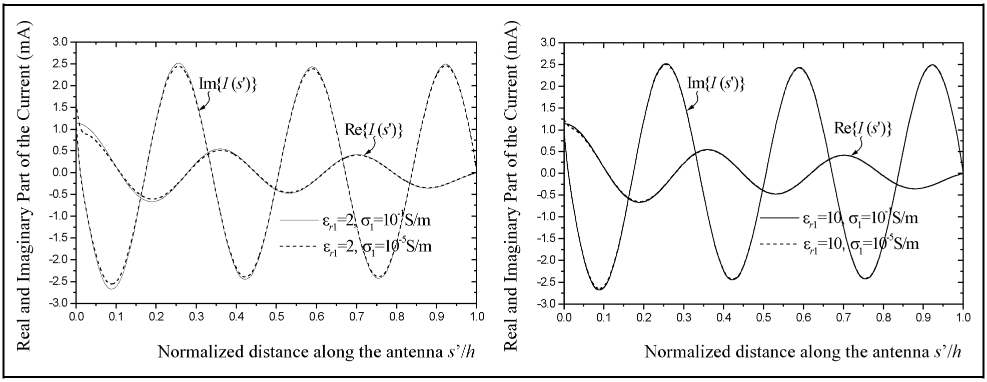


Fig. 16. Real and imaginary part of the current along the antenna for h=300m and radius a=0.05m, for different lossy ground parameters and f=3MHz

4.5 Input impedance of a vertical antenna modeling the lightning discharge channel

Input impedance, Z_{ul}, is important for the antenna analysis in frequency domain and presents integral characteristic of the antenna structure, which also allows checking the accuracy of TIA for Sommerfeld's integral kernel. Satisfactory agreement for the polynomial degree n>2 was confirmed. It is enough to choose a polynomial degree n=2 if the length of the antenna is not greater than 0.6λ_0, for N=1 segment of the antenna. The polynomial degree should not take values n>8 as the polynomial approximation of the antenna current is not appropriate for those. For $\sigma_1\lambda_0<10^{-1}$ the obtained values for the input impedance/admittance are not dependent on the normalized conductivity, but approximately equal to the values of the input impedance/admittance in the case of a perfect dielectric of relative dielectric constant ε_{r1}.

For the antenna of length h=2600m the results for input impedance are obtained for the frequency step Δf =6.425kHz and the selected maximum frequency $f_{\max}/2$=3.2896MHz for FFT transforming time interval [0,t] into the frequency interval [-$f_{\max}/2$, $f_{\max}/2$]. For an arbitrary overall height h the segmentation should be done depending on the frequency i.e. wavelength into the segments of length $l_k \leq \lambda_0/2$ for a selected polynomial degree n_k=3. For frequencies f <500kHz the antenna can be treated as one segment, so that N=1 is enough for calculations, whereas for higher frequencies is necessary to divide the antenna into segments. E.g. about 20 segments are required for frequencies around 1MHz, and about 200 segments for frequencies around 10MHz, if the chosen polynomial degree of the current approximation is n_k=3 along each of the segments. Fig. 17 shows the results for input resistance and input reactance of the antenna modeling lightning discharge channel, for

h=2600m, a=5cm, Z'=0.1 Ω/m, ε_{r1}=10 and σ_1=0.01S/m, N_{FFT}=1024 points for FFT and Δf=6.425kHz. Fig. 18 shows the results for input conductance and input susceptance. For different frequencies, different number of segments along the VMA in the range 1≤N≤200 was chosen, depending on the observed frequency.

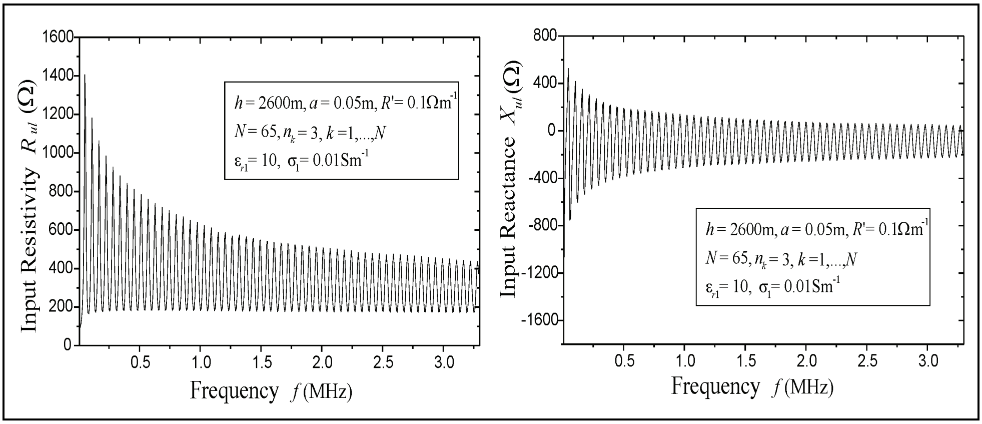


Fig. 17. Input resistance and reactance of the vertical antenna, for h=2600m, a=0.05m, and Z'=0.1Ω/m, at the lossy ground of parameters ε_{r1}=10 and σ_1=0.01S/m, versus frequency

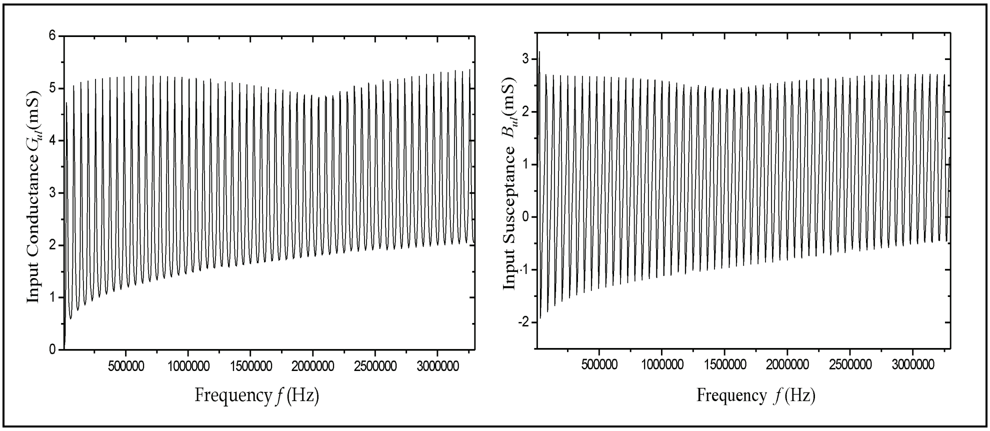


Fig. 18. Input conductance and susceptance of the vertical antenna, for h=2600m, a=0.05m, and Z'=0.1Ω/m, at the lossy ground of parameters ε_{r1}=10 and σ_1=0.01S/m, versus frequency

4.6 Frequency domain results for lightning electric field

Results for vertical and radial electric field as functions of the normalized radial distance $\beta_0\rho$ for the normalized antenna height h/λ_0=0.25, normalized radius of the antenna a/λ_0=0.0005 at the ground of specific conductivity σ_1=0.01S/m and for different values of electric permittivity ε_{r1} as parameter, are presented in Fig. 19, and for ε_{r1}=10 and for different values of specific conductivity σ_1 as parameter in Fig. 20. Vertical and radial component of the electric field at radial distances r from 50 to 250m (0.5λ_0≤r≤2.5λ_0) from

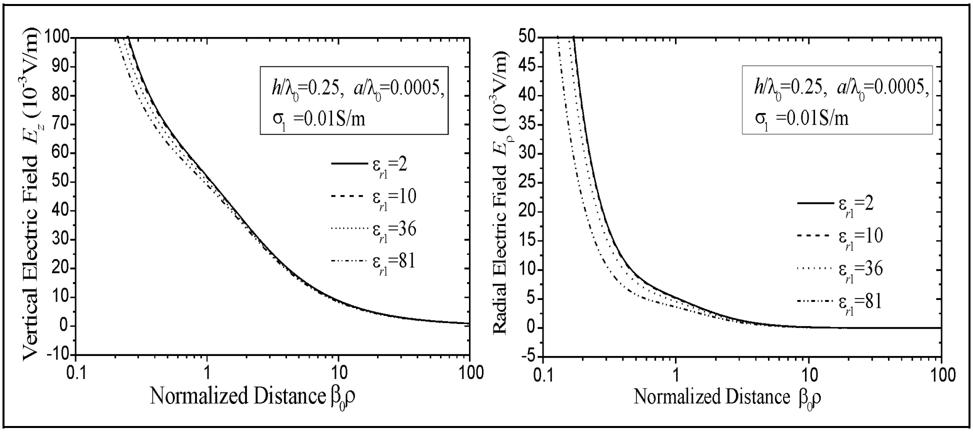


Fig. 19. Vertical and radial electric field of the quarter-wavelength monopole antenna at the lossy ground, for σ_1=0.01S/m and ε_{r1} as parameter, as functions of normalized distance

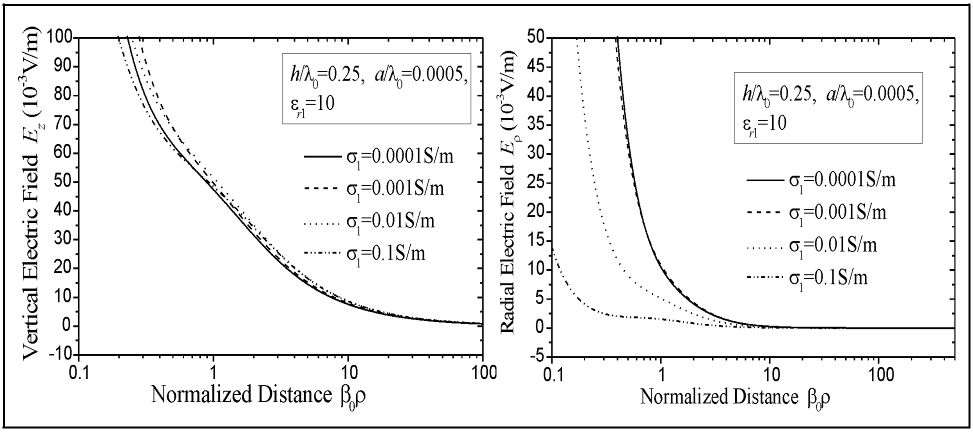


Fig. 20. Vertical and radial electric field of the quarter-wavelength monopole antenna at the lossy ground, for ε_{r1}=10 and σ_1 as parameter, as the functions of normalized distance

the base of the antenna having height h=300m, circular cross section of radius a=5cm at the ground surface of parameters ε_{r1}=10 and σ_1=0.01S/m, for the frequency f=3MHz, the polynomial degree of the current distribution approximation along each of the segments n_k=3, and the number of segments 20, 30 and 50, is presented in Fig. 21. The results obtained for the electric field in the points at a height z=1.5m above the ground surface, for the radial distances $0.5\lambda_0 \le r \le 2.5\lambda_0$, differ a little from results for the field at the ground surface (z=0).

4.7 Time domain results for lightning electromagnetic field

Results for vertical electric and azimuthal magnetic field components are presented in Figs. 22-24 for different distances from the channel-base: r=500m, 5km and 100km. For CBC

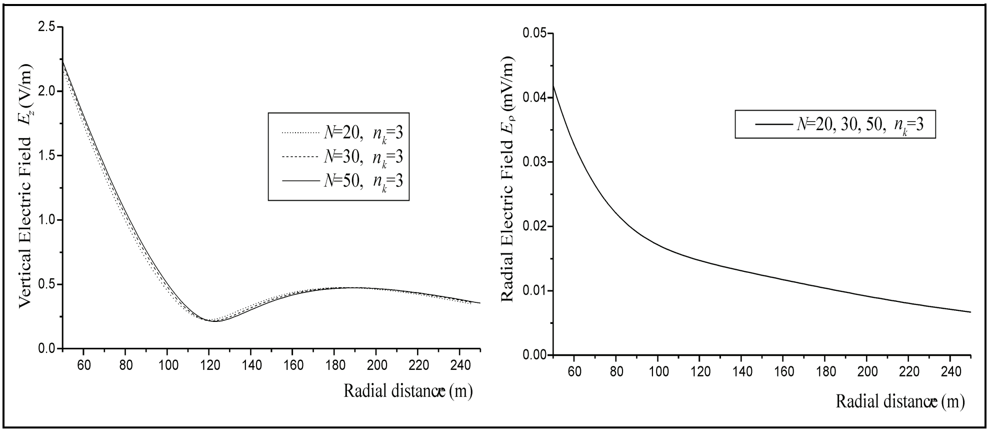


Fig. 21. Vertical and radial electric field of the antenna *h*=300m at the lossy ground, for ε_{r1}=10 and σ_1=0.01S/m for *f*=3MHz, as the functions of distance

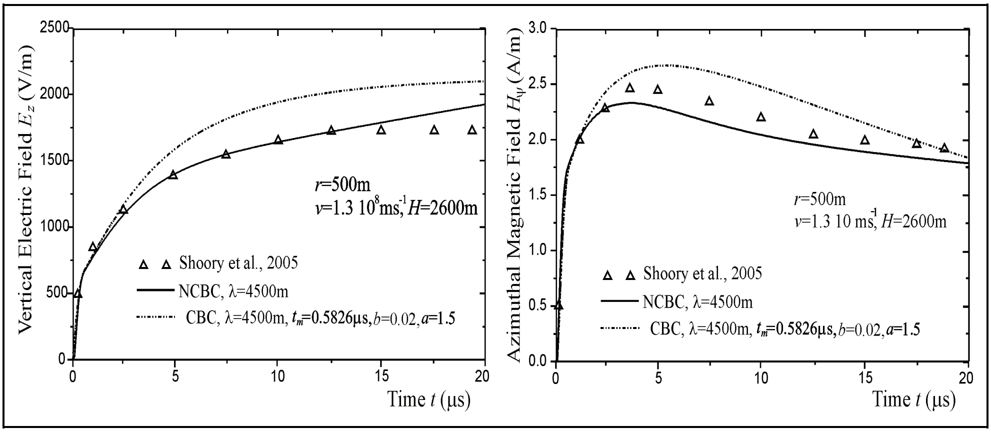


Fig. 22. Vertical electric and azimuthal magnetic field at the ground surface (*z*=0) for *r*=500m

function with parameters I_m=11kA, t_m=0.5826μs, *a*=1.5 and *b*=0.02, and for NCBC function with parameters I_m=11kA, t_m=0.472μs, *a*=1.1, b_1=0.16, c_1=0.34, b_2=0.0047, and c_2=0.66, and two different decaying constants λ=2000m and λ=4500m, the results are compared to the results from (Nucci, 1990) calculated for perfectly conducting ground using the same Modified Transmission Line Model with Exponential Decay (MTLE) with the decaying constant λ=2000m (Nucci et al., 1990), the same return-stroke speed v=1.3·10^{8}m/s and the channel having height *H*=2600m and radius *a*=0.05m. For the same *v*, *H* and *a*, for the distributed resistance *R'*=0.1Ω/m along the antenna, driven by a Dirac delta current source connected across a 3.25m gap, the results are presented also for the perfectly conducting ground (Shoory et al., 2005). Shoory et al., 2005, used the electromagnetic model and a similar procedure to here presented: EFIE type equation, Method of Moments (MoM), method of images for the approximate solution of Sommerfeld's integrals (another approximation), concept of current source, NEC2 computer program (Burke & Poggio, 1981) and FFT/IFFT in 8192 points, i.e. calculations for 4092 positive frequencies. It is interesting that results very similar to (Shoory et al., 2005) would be obtained with NCBC & MTLE, but for λ=6000m.

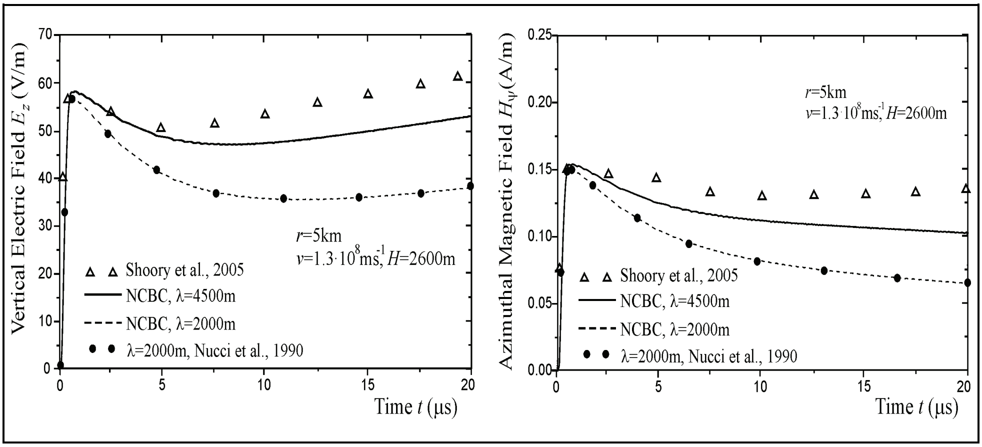


Fig. 23. Vertical electric and azimuthal magnetic field at the ground surface (z=0) for r=5km

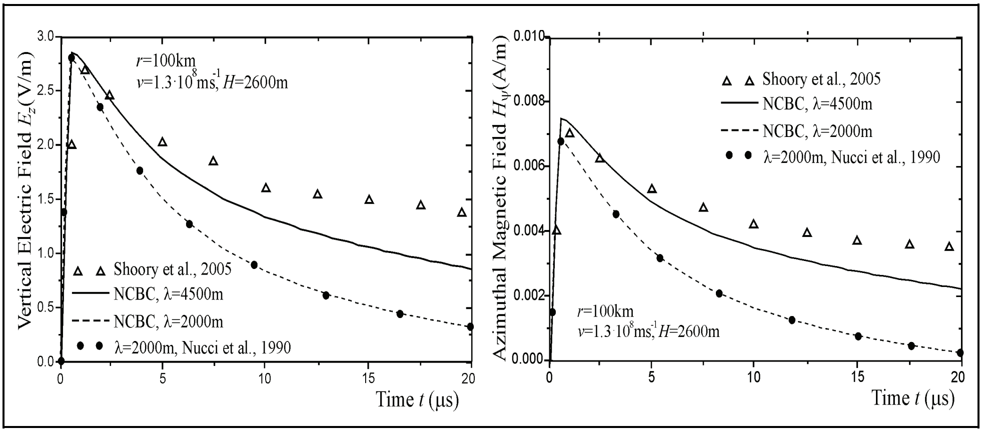


Fig. 24. Vertical electric and azimuthal magnetic field at the ground surface (z=0) for r=100km

5. Conclusion

Fourier transform proved to be very successful in lightning research. It enables calculations in frequency domain which are more suitable for including lossy ground effects than in time domain. It also provides information about frequency spectra of the quantities of interest in lightning research. For antenna modeling of a lightning stroke in frequency domain, Sommerfeld's integral is calculated efficiently using Two-image approximation.

Results are in good agreement with the results from literature for various ground electrical parameters, heights of vertical dipoles above ground, in the near and far field. Based on these results, it can be concluded that the effects of lossy ground are greater on horizontal than on vertical electric field, and that specific conductivity influences more than electrical permittivity. Fourier transform application has to be further investigated in terms of optimal choice of FFT parameters in order to reduce computing time which can be important in antenna modeling of lightning discharge channels and in analysis of lightning electromagnetic fields.

6. Acknowledgment

This work is supported by Project Grant III44004 (2011-2014) financed by Ministry of Education and Science, Republic of Serbia.

7. References

Abramowitz, M. & Stegun, I. (1970). *Handbook of Mathematical Functions,* pp. 253-256, Dover Publications, ISBN 486-61272-4, New York, USA

Anderson, R. B. & Eriksson, A. J. (1980). Lightning parameters for engineering application, *Electra,* Vol.69, pp. 65-102

Andreotti, A.; Falco, S. & Verolino L. (2005). Some integrals involving Heidler's lightning return stroke current expression, *Electrical Engineering (Archiv für Elektrotechnik)*, No.87, pp. 121-128, ISSN 0948-7921

Arand, B. A.; Hakkak M.; Forooraghi K. & Mohassel, J. R. (2003). Analysis of vertical wire antenna above lossy ground using discrete complex image method, *Int. Journal of Electronics and Communications (AEÜ)*, Vol.57, No.5, 2003, pp. 333-337

Baba, Y. & Ishii, M. (2003). Characteristics of electromagnetic return-stroke models, *IEEE Trans. on Electromagn. Compat.*, Vol.45, No.1, (Feb. 2003), pp.129-135

Bannister, P. R. (1966). Quasi-static Fields of Dipole Antennas at the Earth's Surface, *Radio Science*, Vol.1, No.11, (November 1966), pp. 1321-1330

Bannister, P. R. (1978). Extension of quasi-static range finitely conducting earth-image theory techniques to other ranges, *IEEE Trans. on Antennas and Propagation*, Vol.AP-26, No.3, (May 1978), pp. 507-508

Berger, K.; Anderson, R. B. & Kröninger, H. (1975). Parameters of lightning flashes, *Electra,* Vol.41, pp. 22-37

Bewensee, R. M. (1978). WF-SYR/LLL1: A thin-wire computer code for antennas and scatterers with pulse expansion functions for currents, Lawrence Livermore Laboratory, CA, Rep. UCRL 52 028

Bewley, L. V. (1929) Traveling waves due to lightning, *AIEEE Transactions*, Vol. 48, pp. 1050-1064.

Bruce, C. E. R. & Golde, R. H. (1941). The lightning discharge, *J. Inst. Electr. Eng.,* Vol.88, pp. 487-520, ISSN 0360-6449

Burke, G. J. & Poggio, A. J. (1980, 1981). Numerical Electromagnetics Code (NEC) - Method of Moments, *Technical Document 116, Naval Ocean Systems Center*, San Diego, California

Chow, Y. L.; Yang, J. J.; Fang, D. G. & Howard G. E. (1991). A closed-form spatial Green's function for the thick microstrip substrate, *IEEE Trans. on Microwave theory and techniques*, Vol.39, No.3, (Mar. 1991), pp. 588-592

Cooray, V. (1987). Effects of propagation on the return stroke radiation fields, *Radio Sci.*, Vol.22, pp. 757-768

Cooray, V. (1992). Horizontal fields generated by return strokes, *Radio Sci.*, Vol.27, pp. 529-537

Cooray, V. (2002). Some considerations on Cooray-Rubinstein formulation used in deriving the horizontal electric field of lightning return strokes over finitely conducting ground, *IEEE Trans. on Electromagn. Compat.*, Vol.44, No.4, (Nov. 2002), pp. 560-566

Djordjevic, A. R.; Bazdar, M. B.; Petrovic, V. V.; Olcan, D. I.; Sarkar, T. K. & Harrington, R. F. (2002). AWAS for Windows, Version 2.0, Analysis of wire antennas and scatterers - Software and User's manual, Artech House

Feizhou, Z. & Shange L. (2002). A new function to represent the lightning return-stroke currents, *IEEE Trans. on Electromagn. Compat.*, Vol.44, No.4, (Nov. 2002), pp. 595-597

Golde, R. H. (1977). Lightning currents and related parameters, In : *Lightning, Vol. 1, Physics of Lightning*, R. H. Golde, (Ed.), pp. 309-350, Academic Press, ISBN 0122878027, London, UK

Grcev, L.; Rachidi, F. & Rakov, V. A. (2003). Comparison of electromagnetic models of lightning return strokes using current and voltage sources, *International Conference on Atmospheric Electricity, ICAE'03*, (June 2003), Versailles, France

Harrington, R. F. (1968). *Field Computation by Moment Methods*, New York: Macmillan, New York, Sec. 6.2.

Heidler, F. & Cvetic, J. (2002). A class of analytical functions to study the lightning effects associated with the current front, *ETEP,* Vol.12, No.2, (March/April 2002), pp. 141-150, ISSN 1546-3109

Heidler, F. (1985). Travelling current source model for LEMP calculation, *Proc. 6th Int. Zurich Symp. EMC,* pp. 157-162, Zurich, Switzerland, March 1985

IEC 62305-1 Ed. 1.0 (2006). *Protection against lightning* - Part 1: General principles, International Standard, IEC TC/SC 81

Javor, V. & Rancic, P. D. (2006). Application of One Suitable Lightning Return Stroke Current Model, *Proceedings of the International Symposium on Electromagnetic Compatibility EMC Europe 2006*, pp. 941-946, Barcelona, Spain, September 4-8, 2006

Javor, V. & Rancic, P. D. (2007). Vertical Monopole Antenna Model of the Lightning Discharge Current: Two-Image Approximation of Sommerfeld's Integral Kernel, *CD Proc. of the 16th Conf. COMPUMAG 2007,* Aachen, June 2007

Javor, V. & Rancic, P. D. (2009). Electromagnetic field in the vicinity of lightning protection rods at a lossy ground, *IEEE Trans. on Electromagn. Compat.*, Vol.51, No.2, (May 2009), pp.320-330, ISSN 0018-9375

Javor, V. & Rancic, P. D. (2011). A Channel-Base Current function for lightning return-stroke modeling, *IEEE Transactions on Electromagn. Compat.,* Vol.53, No.1, (February 2011), pp. 254-259, ISSN 0018-9375

Javor, V. (2003). Electromagnetic Field Distribution of a nearby Lightning Discharge inside an Overground Conductive Wire Structure, *Facta Universitatis, Series Electronics and Energetics,* Nis, Vol.16, (Apr. 2003), pp. 41-53, ISSN 0353-3670

Javor, V. (2008). Approximating Decaying Part of the Lightning Return Stroke Channel-Base Current, *Proceedings of the 3rd International Symposium on Lightning Physics and Effects*, pp. 26, Vienna, Austria, April 14-15, 2008

Javor, V. (2011a). Impulse charge and specific energy of lightning discharge currents, *CD Proceedings of the 10th International Conference on Applied Electromagnetics, PES 2011*, Paper O1_1, (September 2011), Nis, Serbia

Javor, V. (2011b). Approximation of a lightning channel-base current with a two-rise front, *The Int. Journal for computation and mathematics in electric and electronic engineering COMPEL*, ISSN 0332-1649, (accepted for publication), 2011

Javor, V. (2011c). New functions for representing IEC 62305 standard and other typical lightning currents, *Journal of lightning research*, ISSN 1652-8034, Ref. BSP-JLR-2011-4, (accepted for publication), 2011

Lin, Y. T.; Uman, M. A.; Tiller, J. A.; Brantley R. D.; Beasley, W. H.; Krieder, E. P. & Weidman, C. D. (1979). Characterization of lightning return stroke electric and magnetic fields from simultaneous two-station measurements, *J. Geophys. Res.*, Vol.84, pp. 6307-6314, ISSN 0148-0227

Lindell, I. V. & Alanen, E. (1984). Exact Image Theory for the Sommerfeld Half-Space Problem, Part III: General Formulation, *IEEE Trans. on Ant. and Prop.*, Vol.32, No.10, pp. 1027-1032

Mahmoud, S. F. & Metwally, A. D. (1981). New image representation for dipoles near a dissipative earth: Part 1. Discrete images; Part 2. Discrete plus continuous images, *Radio Science*, Vol.16, No. 6, (Nov-Dec. 1981), pp. 1271-1283

Mahmoud, S. F. (1984). Image theory for electric dipoles above a conducting anisotropic earth, *IEEE Trans. on Antennas and Propagation*, Vol.AP-32, No.7, (July 1984), pp. 679-683

Moini, R.; Kordi, B.; Rafi, G. Z. & Rakov, V. A. (2000). A new lightning return stroke model based on antenna theory, *Journal of Geophys. Research,* Vol.105, No.D24, pp. 29693-29702, ISSN 0148-0227

Nucci, C. A.; Diendorfer, G.; Uman, M. A.; Rachidi, F.; Ianoz, M. & Mazzetti, C. (1990). Lightning return stroke current models with specified channel-base current: a review and comparison, *Journal of Geophys. Research,* Vol.95, No.D12, (November 1990), pp. 20395-20408, ISSN 0148-0227

Petrovic, V. V. (2005). Private communication, Department of Theoretical Electrical Eng., ETF-School of EE, University of Belgrade

Popovic, B. D. & Petrovic, V. V. (1993). Vertical wire antenna above ground: simple near-exact image solution, *IEE Proceedings-H*, Vol.140, No.6, (Dec. 1993), pp. 501-507

Popovic, B. D. (1970). Polynomial approximation of current along thin symmetrical cylindrical dipoles, *Proceedings of IEEE*, Vol.117, No.5, (May 1970), pp. 873-878

Rakov, V. A. & Uman, M. A. (1998). Review and evaluation of lightning return stroke models including some aspects of their application, *IEEE Trans. on Electromagn. Compat.,* Vol.40, No.4, (November 1998), pp. 403-426, ISSN 0018-9375

Rakov, V. A. & Uman, M. A. (2006). *Lightning Physics and Effects*, pp. 394-431, Cambridge University Press, ISBN 978-0-521-58327-5, Cambridge, UK

Rancic, M. P. & Rancic, P. D. (2006). Vertical dipole antenna above a lossy half space: Efficient and accurate two-image approximation for the Sommerfeld's integral, *CD Proc. of EuCAP 2006*, paper No.121 (Ref. No. 362128), Nice, Nov. 2006

Rancic, P. D. & Javor, V. (2007). New Two-Image Approximation of Sommerfeld's Integral Kernel in Calculation of Electromagnetic Field of Vertical Mast Antenna in Frequency Domain, *2nd Int. Symp. on Lightning Physics and Effects,* (April 2007), European COST Action P-18, Vienna, Austria

Rancic, P. D. (1995). Contribution to linear antennas analysis by new forms of integral equations of two potentials, *Proc. of 10th Conf. COMPUMAG'95*, Berlin, (July 1995), pp. 328-329

Richmond, J. H. (1974). Computer program for thin-wire structures in a homogeneous conducting medium, *NASA Report CR-2399*, National Technical Information Service, Springfield, Virginia

Richmond, J. H. (1992). Radiation and scattering by thin-wire structures in the complex frequency domain, In: *Computational Electromagnetics*, E. K. Miller (Ed.), IEEE Press, New York, USA

Rubinstein, M. (1996). An approximate formula for the calculation of the horizontal electric field from lightning at close, intermediate, and long range, *IEEE Trans. on Electromagn. Compat.*, Vol.38, No.3, (August 1996), pp. 531-535

Shoory, A.; Moini, R.; Sadeghi, H. & Rakov, V. A. (2005). Analysis of lightning-radiated electromagnetic fields in the vicinity of lossy ground, *IEEE Trans. on Electromagn. Compat.*, Vol.47, No.1, (February 2005), pp. 131-145, ISSN 0018-9375

Shubair, R. M. & Chow, Y. L. (1993). A simple and accurate complex image interpretation of vertical antennas present in contiguous dielectric half-spaces, *IEEE Trans. on Antennas and Propagation*, Vol.41, No.6, (June 1993), pp. 806-812

Sommerfeld, A. N. (1909). Über die Ausbreitung der Wellen in der Drahtlosen Telegraphie, *Ann. der Physik*, Vol.28, pp. 665-736

Takashima, T., Nakae, T. & Ishibashi R. (1980). Calculation of complex fields in conducting media, *IEEE Trans. on Electr. Insul.*, Vol.EI-15, pp. 1-7

Van der Pol, B. (1931). Über die Ausbreitung elektromagnetischer Wellen, *Z. Hochfrequenctechnik 87*, pp. 152-157

Velickovic, D. M. & Aleksic, S. R. (1986). A new approximation of pulse phenomena *Proc. II Serbian Symp. Applied Electrostatics PES'86,* Nis, Serbia, (Nov. 1986), pp. 6.1-6.9, (*in Serbian*)

Vujevic, S. & Lovric, D. (2010). Exponential approximation of the Heidler function for the reproduction of lightning current waveshapes, *Electric Power Systems Research*, Vol.80, No.10, (October 2010), pp. 1293-1298, ISSN 0378-7796

Wait, J. R. & Spies, K. P. (1969). On the Image Theory Representation of the Quasi-static Fields of a Line Source Above the Ground, *Can. J. Phys.*, No.47, pp. 2731-2733

Wait, J. R. (1997). Concerning the horizontal electric field of lightning, *IEEE Trans. on Electromagn. Compat.*, Vol.39, No.2, (1997), pp. 186

Walker, J. S. (1996). *Fast Fourier Transforms,* pp. 149-169, Boca Raton: CRC Press, ISBN 0-8493-7163-5

Yang, C. & Zhou, B. (2004). Calculation methods of electromagnetic fields very close to lightning, *IEEE Trans. on Electromagn. Compat.*, Vol.46, No.1, (Feb. 2004), pp. 133-141

4

Robust Beamforming and DOA Estimation

Liu Congfeng
Electronic Countermeasure Research Institute, Xidian University, Xian Shaanxi, China

1. Introduction

1.1 Robust beamforming overview

Beamforming is a ubiquitous task in array signal processing with applications, among others, in radar, sonar, acoustics, astronomy, seismology, communications, and medical imaging. Without loss of generality, we consider herein beamforming in array processing applications. The introduction to beamforming can be found in [1]-[9] and the references herein.

The traditional approach to the design of adaptive beamformers assumes that the desired signal components are not present in training data, and the robustness of beamformer is known to depend essentially on the availability of signal-free training data. However, in many important applications such as mobile communications, passive location, microphone array speech processing, medical imaging, and radio astronomy, the signal-free training data cells are unavailable. In such scenarios, the desired signal is always present in the training snapshots, and the adaptive beamforming methods become very sensitive to any violation of underlying assumptions on the environment, sources, or sensor array. In fact, the performances of the existing adaptive array algorithms are known to degrade substantially in the presence of even slight mismatches between the actual and presumed array responses to the desired signal [10]-[12]. Similar types of degradation can take place when the array response is known precisely but the training sample size is small, namely the mismatch between the actual and the estimated covariance matrix [13]-[15]. Therefore, robust approaches to adaptive beamforming appear to be of primary importance in these cases [16][17].

Many approaches have been proposed to improve the robustness of the adaptive beamformer during the past three decades. Indeed, the literatures on the robust adaptive beamformer are quite extensive. We provide a brief review as fellows. For more recent detailed critical reviews, see [18] and [19]-[25].

1.1.1 Robust approaches for signal direction mismatch

For the specific case of the signal direction mismatch, there are several efficient methods have been developed. Representative examples of such techniques are the linearly constrained minimum variance (LCMV) beamformer [26], which is also denoted as the linearly constrained minimum power (LCMP) beamformer in other references [27] and this chapter, signal blocking-based algorithms [10][28], and Bayesian beamformer [29]. Although all these methods provide excellent robustness against the signal direction mismatch, they are not robust against other types of mismatches caused by poor array calibration, unknown sensor mutual coupling, near-far wavefront mismodeling, signal wavefront distortions, source spreading, and coherent/incoherent local scattering, as well as other effects [17].

Chun-Yang Chen and P.P.Vaidyanathan consider a simplified uncertainty set which contains only the steering vectors with a desired uncertainty range of direction of arrival (DOA) [25], although the closed-form solution is given, and the diagonal loading level can be computed by the iteration method systematically, but how to determine the DOA uncertainty range is the critical problem.

1.1.2 Robust approaches for general mismatch

Several other approaches are known to provide the improved robustness against more general types of mismatches, for example, the algorithms that use the diagonal loading of the sample covariance matrix [14][16], the eigenspace-based beamformer [11][30][31], and the covariance matrix taper (CMT) approach [32]-[34]. For the diagonal loading method, a serious drawback is that there is no reliable way to choose the diagonal loading level, F.Vincent and O.Besson propose the method to select the optimal loading level with a view to maximizing the signal-to-noise ratio (SNR) in the presence of steering vector errors and it is shown that the loading is negative, but they can't give the exact solution, instead of the approximate solution, moreover, they can't give the expression of steering vector errors [35]. The eigenspace-based approach is essentially restricted in its performance at low SNR and when the dimension of the signal-plus-interference subspace is high, and the dimension must be known in the latter technique [31]. The CMT approach is known to provide an excellent robustness in scenarios with nonstationary interferers, however, its robustness against mismatches of the desired signal array response may be unsatisfactory, furthermore, it can also be explained as the diagonal loading [33].

1.1.3 Uncertainty set constraint approaches for general mismatch

Very recently, many approaches have been proposed for improving the robustness of the standard minimum variance distortionless response (MVDR) beamformer. Their main ideas are based on the definition of the uncertainty set and the worst-case performance optimization, but these algorithms are all classified to the diagonal loading technique.

Jian Li et al propose the robust Capon beamformer under the constraint of steering vector uncertainty set [20], then the constraint of steering vector norm is imposed and the doubly constraint robust Capon beamformer is proposed [22]. For the two beamformers, although they give the exact weight vectors, and the methods of finding the optimal loading level, but their performance improvements are not obvious. Actually, the constraint of uncertainty set is the essence of the two robust beamformers, and the two beamfomer have the same robustness characteristic. F.Vincent and O.Besson also analyze the performance of the beamformer under the uncertainty set constraint approximatively, but they can't give the exact loading level [36].

S. A. Vorobyov et al propose a robust beamformer in the presence of an arbitrary unknown signal steering vector mismatch [19], although they prove the proposed approach equivalent to the loading sample matrix inversion (LSMI) algorithm, but they can't give the direct method to compute the optimal weight vector, and the second-order cone (SOC) programming-based approach is used to solve the original problem. Ayman Elnashar et al make use of the diagonal loading technique to implement the robust beamformer [24], but the optimal value of diagonal loading level is not solved exactly, alternatively, the diagonal loading technique is integrated into the adaptive update schemes by means of optimum variable loading technique. R. G. Lorenz and S. P. Boyd also solve the similar beamformer by the Lagrange multiplier techniques [23], but they express the weight vector and the array manifold as the direct sum of the corresponding real and imaginary components. Almir Mutapcic et al show that worst-case robust beamforming with multiplicative uncertainty in the weights can be cast as a tractable convex optimization problem [37], but they can't give the solving method, In fact, the proposed robust beamformer with uncertain weights can be converted to that in [19] equivalently.

S. Shahbazpanahi et al consider the general-rank signal model, and the robust beamformer is proposed for the distributed sources [21], therein, an elegant closed-form solution is given, but its performance improvement depends on the constraint parameter severely, and is not up to optimal.

1.1.4 Weight norm constraint approaches for general mismatch

Jian Li et al propose a Capon beamforming approach with the norm inequality constraint (NIC) to improve the robustness against array steering vector errors and noise [22], although the exact solution is given, and optimal loading level can be computed via the proposed method, but by analysis and simulation, its efficiency is not as good as expectation. Since the constraint parameter determines its robustness, but how to select the constraint parameter is not discussed.

Quadratic inequality constraints (QIC) on the weight vector of LCMP beamformer can improve robustness to pointing errors and random perturbations in sensor parameter [27]. The weights that minimize the output power subject to linear constraints and an inequality constraint on the norm of the weight vector have the same form as that of the optimum LCMP beamformer with diagonal loading of the data covariance matrix. But the optimal loading level cannot be directly expressed as a function of the constraint in a closed form, and cannot be solved exactly. Hence, its application is restricted by the optimal weight vector finding. So that some numerically algorithms are proposed to implement the QICLCMP, such as Least Mean Squares (LMS) or Recursive Least Squares (RLS) [27], but the application effect isn't good as the expectation.

This chapter is organized as follows [38]. First, the norm inequality constraint Capon beamformer (NICCB) is introduced and analyzed particularly. Second, the choice of the norm constraint parameter and the selecting bound is discussed. Third, the norm equality constraint Capon beamformer (NECCB) is proposed and is solved effectively. Finally, the simulation analyses and the conclusion are given.

1.2 Capon beamformer under norm inequality constraint (NICCB)

The Capon beamformer can experience significant performance degradation when there is a mismatch between the presumed and actual characteristics of the source or array. The goal of NICCB is to impose an additional inequality constraint on the Euclidean norm of $\mathbf{w}$ for the purpose of improving the robustness to pointing errors and random perturbations in sensor parameters, here $\mathbf{w}$ denotes the array weight vector. This requires incorporating a norm inequality constraint on $\mathbf{w}$ of the form:

$$\|\mathbf{w}\|^2 \leq \varsigma \tag{1.1}$$

where ς is the norm constraint parameter. Consequently, the NICCB problem is formulated as follows:

$$\begin{cases} \min\limits_{\mathbf{w}} \mathbf{w}^H \mathbf{R} \mathbf{w} \\ s.t. \quad \mathbf{w}^H \overline{\mathbf{s}} = 1 \\ \qquad\ \|\mathbf{w}\|^2 \leq \varsigma \end{cases} \tag{1.2}$$

where $\mathbf{R}$ is the data covariance matrix, $\overline{\mathbf{s}}$ is the presumed signal steering vector, and $(\cdot)^H$ denotes the conjugate transposition, $\|\cdot\|$ denotes the vector l_2 norm. For the convenience of analysis, and analyzing the choice of the norm constraint parameter, the solution to NICCB [22] is introduced as follows.

1.2.1 Solution to NICCB

Let S be the set defined by the constraints in above optimization problem, namely:

$$S = \left\{ \mathbf{w} \middle| \mathbf{w}^H \overline{\mathbf{s}} = 1, \|\mathbf{w}\|^2 \leq \varsigma \right\} \tag{1.3}$$

Define a function:

$$f_1(\mathbf{w}, \lambda, \mu) = \mathbf{w}^H \mathbf{R} \mathbf{w} + \lambda\left(\|\mathbf{w}\|^2 - \varsigma\right) + \mu\left(-\mathbf{w}^H \overline{\mathbf{s}} - \overline{\mathbf{s}}^H \mathbf{w} + 2\right) \tag{1.4}$$

where λ is the real-valued Lagrange multiplier, and $\lambda \geq 0$ satisfying $\mathbf{R} + \lambda \mathbf{I} > 0$ so that $f_1(\mathbf{w}, \lambda, \mu)$ can be minimized with respect to $\mathbf{w}$. μ is the arbitrary Lagrange multiplier. Then:

$$f_1(\mathbf{w}, \lambda, \mu) \leq \mathbf{w}^H \mathbf{R} \mathbf{w}, \qquad \mathbf{w} \in S \tag{1.5}$$

with equality on the boundary of S.

For the standard Capon beamformer

$$\begin{cases} \min\limits_{\mathbf{w}} \mathbf{w}^H \mathbf{R} \mathbf{w} \\ s.t. \quad \mathbf{w}^H \overline{\mathbf{s}} = 1 \end{cases} \tag{1.6}$$

The optimal solution is:

$$\mathbf{w} = \frac{\mathbf{R}^{-1} \overline{\mathbf{s}}}{\overline{\mathbf{s}}^H \mathbf{R}^{-1} \overline{\mathbf{s}}} \tag{1.7}$$

where $\mathbf{R}^{-1}$ is the inversion of $\mathbf{R}$, namely $(\cdot)^{-1}$ denotes the matrix inversion. Here, we can have:

$$\|\mathbf{w}\|^2 = \mathbf{w}^H\mathbf{w} = \left(\frac{\mathbf{R}^{-1}\overline{\mathbf{s}}}{\overline{\mathbf{s}}^H\mathbf{R}^{-1}\overline{\mathbf{s}}}\right)^H \frac{\mathbf{R}^{-1}\overline{\mathbf{s}}}{\overline{\mathbf{s}}^H\mathbf{R}^{-1}\overline{\mathbf{s}}} = \frac{\overline{\mathbf{s}}^H\mathbf{R}^{-2}\overline{\mathbf{s}}}{\left(\overline{\mathbf{s}}^H\mathbf{R}^{-1}\overline{\mathbf{s}}\right)^2} \tag{1.8}$$

where $\mathbf{R}^{-2} = \left(\mathbf{R}^{-1}\right)^2 = \mathbf{R}^{-1}\cdot\mathbf{R}^{-1}$, the above result using the Hermitian property of $\mathbf{R}$.

Consider the condition:

$$\frac{\overline{\mathbf{s}}^H\mathbf{R}^{-2}\overline{\mathbf{s}}}{\left(\overline{\mathbf{s}}^H\mathbf{R}^{-1}\overline{\mathbf{s}}\right)^2} \le \varsigma \tag{1.9}$$

when the above condition is satisfied, the standard Capon beamformer solution (1.7) satisfies the norm constraint of NICCB, hence, is also the solution to NICCB. For this case, $\lambda = 0$ and the norm constraint in NICCB is inactive.

Otherwise, we have the condition:

$$\varsigma < \frac{\overline{\mathbf{s}}^H\mathbf{R}^{-2}\overline{\mathbf{s}}}{\left(\overline{\mathbf{s}}^H\mathbf{R}^{-1}\overline{\mathbf{s}}\right)^2} \tag{1.10}$$

which is an upper bound on ς so that NICCB is different from the standard Capon beamformer. To deal with this case, we can rewrite the $f_1(\mathbf{w},\lambda,\mu)$ as follows:

$$\begin{aligned} f_1(\mathbf{w},\lambda,\mu) = &\left[\mathbf{w} - \mu(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}}\right]^H (\mathbf{R}+\lambda\mathbf{I})\left[\mathbf{w} - \mu(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}}\right] - \\ &-\mu^2\overline{\mathbf{s}}^H(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}} - \lambda\varsigma + 2\mu \end{aligned} \tag{1.11}$$

Hence, the unconstrained minimizer of $f_1(\mathbf{w},\lambda,\mu)$, for fixed λ and μ, is given by:

$$\hat{\mathbf{w}}_{\lambda,\mu} = \mu(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}} \tag{1.12}$$

Clearly, we have:

$$f_2(\lambda,\mu) \overset{\Delta}{=} f_1\left(\hat{\mathbf{w}}_{\lambda,\mu},\lambda,\mu\right) = -\mu^2\overline{\mathbf{s}}^H(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}} - \lambda\varsigma + 2\mu \le \mathbf{w}^H\mathbf{R}\mathbf{w}, \qquad \mathbf{w}\in S \tag{1.13}$$

The maximization of $f_2(\lambda,\mu)$ with respect to μ. Hence, μ is given by:

$$\hat{\mu} = \frac{1}{\overline{\mathbf{s}}^H(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}}} \tag{1.14}$$

Insert $\hat{\mu}$ into $f_2(\lambda,\mu)$, and let:

$$f_3(\lambda) \overset{\Delta}{=} f_2(\lambda,\hat{\mu}) = -\lambda\varsigma + \frac{1}{\overline{\mathbf{s}}^H(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}}} \tag{1.15}$$

The maximization of the above function $f_3(\lambda)$ with respect to λ gives:

$$\frac{\overline{\mathbf{s}}^H(\mathbf{R}+\lambda\mathbf{I})^{-2}\overline{\mathbf{s}}}{\left[\overline{\mathbf{s}}^H(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}}\right]^2}=\varsigma \tag{1.16}$$

Hence, the optimal Lagrange multiplier $\hat{\lambda}$ can be obtained efficiently via, for example, a Newton's method from the above equation of λ.

Note that using $\hat{\mu}$ in $\hat{\mathbf{w}}_{\lambda,\mu}$ yields:

$$\hat{\mathbf{w}}=\frac{(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}}}{\overline{\mathbf{s}}^H(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}}} \tag{1.17}$$

which satisfies the constraints of NICCB, namely:

$$\hat{\mathbf{w}}^H\overline{\mathbf{s}}=1 \tag{1.18}$$

and

$$\|\hat{\mathbf{w}}\|^2=\varsigma \tag{1.19}$$

Hence, $\hat{\mathbf{w}}$ belongs to the boundary of S. Therefore, $\hat{\mathbf{w}}$ is our sought solution to the NICCB optimization problem, which has the same form as the Capon beamformer with a diagonal loading term $\lambda\mathbf{I}$ added to $\mathbf{R}$, namely, NICCB also belongs to the class of diagonal loading approaches.

From the above analysis, we can see that if the Lagrange multiplier λ is obtained, the optimal weight vector for NICCB will be solved. In order to obtain the Lagrange multiplier λ, we must solve the following equation via Newton's method, and let:

$$h(\lambda)=\frac{\overline{\mathbf{s}}^H(\mathbf{R}+\lambda\mathbf{I})^{-2}\overline{\mathbf{s}}}{\left[\overline{\mathbf{s}}^H(\mathbf{R}+\lambda\mathbf{I})^{-1}\overline{\mathbf{s}}\right]^2} \tag{1.20}$$

Hence, the key problem of NICCB is finding the optimal Lagrange multiplier by above equation (1.20). In this chapter, we will give the complete investigation on NICCB, and the existence of its solution is analyzed as follows.

1.2.2 Solution to the optimal Lagrange multiplier

In order to solve the equation (1.20), we perform an eigenvalue decomposition (EVD) of the sample covariance matrix as follows:

$$\mathbf{R}=\mathbf{U}\cdot\mathbf{\Gamma}\cdot\mathbf{U}^H=\sum_{i=1}^{M}\lambda_i\mathbf{u}_i\mathbf{u}_i^H \tag{1.21}$$

where $\mathbf{\Gamma}=diag(\lambda_1,\lambda_2,\cdots,\lambda_M)$ is diagonal matrix, $\mathbf{U}=(\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_M)$ is an Hermitian matrix, λ_i ($i=1,2\cdots M$) and $\mathbf{u}_i$ ($i=1,2\cdots M$) are the eigenvalues and eigenvectors of $\mathbf{R}$, respectively, M is the total number of degrees-of-freedom. For the convenience of analysis, we assume that the eigenvalues / eigenvectors of $\mathbf{R}$ are sorted in descending order, i.e. ,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \tag{1.22}$$

We can have:

$$h(\lambda) = \frac{\sum_{i=1}^{M} \frac{\bar{\mathbf{s}}^H \mathbf{u}_i \mathbf{u}_i^H \bar{\mathbf{s}}}{(\lambda_i + \lambda)^2}}{\left[\sum_{i=1}^{M} \frac{\bar{\mathbf{s}}^H \mathbf{u}_i \mathbf{u}_i^H \bar{\mathbf{s}}}{\lambda_i + \lambda}\right]^2} = \frac{\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{(\lambda_i + \lambda)^2}}{\left[\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{\lambda_i + \lambda}\right]^2} \tag{1.23}$$

Therefore, $h(\lambda)$ is monotonically increasing function of $\lambda \geq 0$ [22] , then:

$$h(\lambda) = \frac{\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{(\lambda_i + \lambda)^2}}{\left[\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{\lambda_i + \lambda}\right]^2} \leq \frac{\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{(\lambda_M + \lambda)^2}}{\left[\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{\lambda_1 + \lambda}\right]^2} = \left(\frac{\lambda_1 + \lambda}{\lambda_M + \lambda}\right)^2 \frac{1}{\sum_{i=1}^{M} \left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2} \tag{1.24}$$

and

$$h(\lambda) = \frac{\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{(\lambda_i + \lambda)^2}}{\left[\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{\lambda_i + \lambda}\right]^2} \geq \frac{\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{(\lambda_1 + \lambda)^2}}{\left[\sum_{i=1}^{M} \frac{\left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2}{\lambda_M + \lambda}\right]^2} = \left(\frac{\lambda_M + \lambda}{\lambda_1 + \lambda}\right)^2 \frac{1}{\sum_{i=1}^{M} \left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2} \tag{1.25}$$

and let:

$$\gamma = \sum_{i=1}^{M} \left\|\bar{\mathbf{s}}^H \mathbf{u}_i\right\|^2 \tag{1.26}$$

Alternately, the above inequality relationship can be expressed as:

$$\sqrt{\gamma\varsigma} \leq \frac{\lambda_1 + \lambda}{\lambda_M + \lambda} \tag{1.27}$$

and

$$\sqrt{\gamma\varsigma} \geq \frac{\lambda_M + \lambda}{\lambda_1 + \lambda} \tag{1.28}$$

Next, we analyze the bound of the Lagrange multiplier λ and its existence.

1. If $\sqrt{\gamma\varsigma} > 1$, using (1.27) and (1.28), we can have:

$$\begin{cases} \sqrt{\gamma\varsigma}\left(\lambda_M + \lambda\right) \le \lambda_1 + \lambda \\ \sqrt{\gamma\varsigma}\left(\lambda_1 + \lambda\right) \ge \lambda_M + \lambda \end{cases} \Rightarrow \begin{cases} \lambda \le \dfrac{\lambda_1 - \sqrt{\gamma\varsigma}\lambda_M}{\sqrt{\gamma\varsigma} - 1} \\ \lambda \ge \dfrac{\lambda_M - \sqrt{\gamma\varsigma}\lambda_1}{\sqrt{\gamma\varsigma} - 1} \end{cases} \tag{1.29}$$

Since $\sqrt{\gamma\varsigma} > 1$, but $\lambda_M - \sqrt{\gamma\varsigma}\lambda_1 < 0$, and $\lambda \ge 0$, so that $\lambda_1 - \sqrt{\gamma\varsigma}\lambda_M > 0 \Leftrightarrow \sqrt{\gamma\varsigma} < \lambda_1 / \lambda_M$. Therefore, the bound of the Lagrange multiplier λ under $1 < \sqrt{\gamma\varsigma} < \lambda_1 / \lambda_M$ is given as follows:

$$\lambda_{\min}^{(1)} \overset{\Delta}{=} 0 \le \lambda \le \frac{\lambda_1 - \sqrt{\gamma\varsigma}\lambda_M}{\sqrt{\gamma\varsigma} - 1} \overset{\Delta}{=} \lambda_{\max}^{(1)} \tag{1.30}$$

Then, we have:

$$h\left(\lambda_{\min}^{(1)}\right) = h(0) = \frac{\overline{\mathbf{s}}^H \mathbf{R}^{-2} \overline{\mathbf{s}}}{\left[\overline{\mathbf{s}}^H \mathbf{R}^{-1} \overline{\mathbf{s}}\right]^2} > \varsigma \tag{1.31}$$

and

$$h\left(\lambda_{\max}^{(1)}\right) = h(\lambda)\Big|_{\lambda_{\max}^{(1)}} \le \left(\frac{\lambda_1 + \lambda}{\lambda_M + \lambda}\right)^2 \frac{1}{\gamma}\Bigg|_{\lambda_{\max}^{(1)}} = \varsigma \tag{1.32}$$

Hence, when $1 < \sqrt{\gamma\varsigma} < \lambda_1 / \lambda_M$, there is unique solution $\lambda \in \left[\lambda_{\min}^{(1)},\ \lambda_{\max}^{(1)}\right]$ satisfies $h(\lambda) = \varsigma$.

2. If $\sqrt{\gamma\varsigma} < 1$, using (1.27) and (1.28), we can have:

$$\begin{cases} \sqrt{\gamma\varsigma}\left(\lambda_M + \lambda\right) \le \lambda_1 + \lambda \\ \sqrt{\gamma\varsigma}\left(\lambda_1 + \lambda\right) \ge \lambda_M + \lambda \end{cases} \Rightarrow \begin{cases} \lambda \ge \dfrac{\sqrt{\gamma\varsigma}\lambda_M - \lambda_1}{1 - \sqrt{\gamma\varsigma}} \\ \lambda \le \dfrac{\sqrt{\gamma\varsigma}\lambda_1 - \lambda_M}{1 - \sqrt{\gamma\varsigma}} \end{cases} \tag{1.33}$$

Since $\sqrt{\gamma\varsigma} < 1$, but $\sqrt{\gamma\varsigma}\lambda_M - \lambda_1 < 0$, and $\lambda \ge 0$, so that $\sqrt{\gamma\varsigma}\lambda_1 - \lambda_M > 0 \Leftrightarrow \sqrt{\gamma\varsigma} > \lambda_M / \lambda_1$. Therefore, the bound of the Lagrange multiplier λ under $\lambda_M / \lambda_1 < \sqrt{\gamma\varsigma} < 1$ is given as follows:

$$\lambda_{\min}^{(2)} \overset{\Delta}{=} 0 \le \lambda \le \frac{\sqrt{\gamma\varsigma}\lambda_1 - \lambda_M}{1 - \sqrt{\gamma\varsigma}} \overset{\Delta}{=} \lambda_{\max}^{(2)} \tag{1.34}$$

Then, we have:

$$h\left(\lambda_{\min}^{(2)}\right)=h(0)=\frac{\overline{\mathbf{s}}^H\mathbf{R}^{-2}\overline{\mathbf{s}}}{\left[\overline{\mathbf{s}}^H\mathbf{R}^{-1}\overline{\mathbf{s}}\right]^2}>\varsigma \tag{1.35}$$

and

$$h\left(\lambda_{\max}^{(2)}\right)=h(\lambda)\Big|_{\lambda_{\max}^{(2)}}\geq\left(\frac{\lambda_M+\lambda}{\lambda_1+\lambda}\right)^2\frac{1}{\gamma}\Bigg|_{\lambda_{\max}^{(2)}}=\varsigma \tag{1.36}$$

Hence, when $\lambda_M/\lambda_1<\sqrt{\gamma\varsigma}<1$, there isn't a solution $\lambda\in\left[\lambda_{\min}^{(2)},\lambda_{\max}^{(2)}\right]$ satisfies $h(\lambda)=\varsigma$.

In a word, we can conclude that when $1<\sqrt{\gamma\varsigma}<\lambda_1/\lambda_M$, there is a unique solution $\lambda\in\left[\lambda_{\min}^{(1)},\lambda_{\max}^{(1)}\right]$ satisfies $h(\lambda)=\varsigma$.

1.3 Norm inequality constraint parameter selection

From above analysis, we can see that it is important to select the norm inequality constraint parameter ς for NICCB. If the norm inequality constraint parameter ς is large, it is inactive. On the contrary, if the norm inequality constraint parameter ς is small, there isn't a solution to satisfy NICCB.

We have analyzed that when $1<\sqrt{\gamma\varsigma}<\lambda_1/\lambda_M$, there is a unique solution $\lambda\in\left[\lambda_{\min}^{(1)},\lambda_{\max}^{(1)}\right]$ satisfies $h(\lambda)=\varsigma$. Hence, we can have the selecting bound of the norm inequality constraint parameter ς as follows:

$$1<\sqrt{\gamma\varsigma}<\frac{\lambda_1}{\lambda_M} \tag{1.37}$$

Namely:

$$\frac{1}{\gamma}<\varsigma<\frac{1}{\gamma}\cdot\left(\frac{\lambda_1}{\lambda_M}\right)^2 \tag{1.38}$$

Add the condition of $\varsigma<\dfrac{\overline{\mathbf{s}}^H\mathbf{R}^{-2}\overline{\mathbf{s}}}{\left(\overline{\mathbf{s}}^H\mathbf{R}^{-1}\overline{\mathbf{s}}\right)^2}\overset{\Delta}{=}\varsigma_0$, we can obtain:

$$\varsigma_{\min}\overset{\Delta}{=}\frac{1}{\gamma}<\varsigma<\min\left\{\varsigma_0,\ \frac{1}{\gamma}\left(\frac{\lambda_1}{\lambda_M}\right)^2\right\}\overset{\Delta}{=}\varsigma_{\max} \tag{1.39}$$

If the norm inequality constraint parameter ς is out of the above bound, there is no solution to NICCB. Hence, the norm inequality constraint parameter ς should be chosen in the interval defined by the above inequalities.

1.4 Capon beamformer under norm equality constraint (NECCB)

From above analyses, we can see that the norm inequality constraint can enhance the robustness of NICCB. Since the inequality relationship has a wide range, the norm of the weight vector will vary in the relevant wide range. If the fluctuation of weight vector norm is acutely, the performance improvement will be weakened greatly. Because the norm equality constraint (NEC) is stronger than the norm inequality constraint (QIC), NECCB will have more ascendant robust performance than NICCB. Hence, NECCB is proposed and is solved effectively in this chapter.

NECCB is to impose an additional equality constraint on the Euclidean norm of $\mathbf{w}$. The NECCB problem is formulated as follows:

$$\begin{cases} \min\limits_{\mathbf{w}} \mathbf{w}^H \mathbf{R} \mathbf{w} \\ s.t. \quad \mathbf{w}^H \overline{\mathbf{s}} = 1 \\ \qquad \|\mathbf{w}\|^2 = \varsigma \end{cases} \tag{1.40}$$

Compare NECCB with NICCB, we can educe the conclusion as follows: (1) The solution to NICCB is obtained on the boundary of its constraint, similarly, for NECCB, the solution is also obtained on its constraint boundary. (2) The solving methods of the two beamformers (or the optimization problem) is different, such as the forenamed solution to NICCB, the Lagrange multiplier of NICCB is taken as positive real-value only, but for NECCB, the Lagrange multiplier is taken as arbitrary real-value, namely, it will be not only the positive real-value, but also the negative real-value. Hence, if we analyze from the point of view of the solving optimization problem, NECCB has two solutions to the optimal Lagrange multiplier, one is positive, and another is negative. Actually, the positive one is the solution to NICCB. For the sake of distinguishing the otherness, the negative solution is interested to NECCB. In order to solve NECCB, we must make use of the discussed results of NICCB, since the manipulation of some inequality, such as the inequality lessening and enlarging is only right for the positive real-value when we solve NICCB.

Similar to NICCB, the solution to NECCB can also be solved by the Lagrange multiplier methodology. And the optimal weight vector of NECCB has the same form as NICCB. The difference between NECCB and NICCB is only the Lagrange multiplier $\breve{\lambda}$, for NICCB, $\lambda \geq 0$, here $\breve{\lambda}$ is arbitrary real-value.

Although the solution to NECCB has the same form as NICCB, but the bound of the Lagrange multiplier is different. In order to use the analyzed results of NICCB for NECCB, replace the Lagrange multiplier by its absolute value, namely the bound of the Lagrange multiplier $\breve{\lambda}$ for NECCB is given by:

$$\sqrt{\gamma\varsigma} \leq \frac{\lambda_1 + |\breve{\lambda}|}{\lambda_M + |\breve{\lambda}|} \tag{1.41}$$

and

$$\sqrt{\gamma\varsigma} \geq \frac{\lambda_M + |\breve{\lambda}|}{\lambda_1 + |\breve{\lambda}|} \tag{1.42}$$

1. If $\sqrt{\gamma\varsigma} > 1$, then:

$$\frac{\lambda_M - \sqrt{\gamma\varsigma}\lambda_1}{\sqrt{\gamma\varsigma}-1} \le \left|\breve{\lambda}\right| \le \frac{\lambda_1 - \sqrt{\gamma\varsigma}\lambda_M}{\sqrt{\gamma\varsigma}-1} \tag{1.43}$$

If $\lambda_1 - \sqrt{\gamma\varsigma}\lambda_M > 0$, then $\sqrt{\gamma\varsigma} < \lambda_1/\lambda_M$, since $\lambda_M - \sqrt{\gamma\varsigma}\lambda_1 < 0$, but $\left|\breve{\lambda}\right| > 0$. Therefore, if $1 < \sqrt{\gamma\varsigma} < \lambda_1/\lambda_M$, we can have:

$$\breve{\lambda}_{\min}^{(1)} \overset{\Delta}{=} -\frac{\lambda_1 - \sqrt{\gamma\varsigma}\lambda_M}{\sqrt{\gamma\varsigma}-1} \le \breve{\lambda} \le \frac{\lambda_1 - \sqrt{\gamma\varsigma}\lambda_M}{\sqrt{\gamma\varsigma}-1} \overset{\Delta}{=} \breve{\lambda}_{\max}^{(1)} \tag{1.44}$$

Since $\breve{\lambda}_{\max}^{(1)} > 0$, and $\breve{\lambda}_{\min}^{(1)} = -\breve{\lambda}_{\max}^{(1)} < 0$. Hence, when $1 < \sqrt{\gamma\varsigma} < \lambda_1/\lambda_M$, the solution to NECCB in the bound of $\left[0,\ \breve{\lambda}_{\max}^{(1)}\right]$ is the same as NICCB, but the solution in the bound of $\left[\breve{\lambda}_{\min}^{(1)},\ 0\right]$ is the true solution to NECCB.

2. If $\sqrt{\gamma\varsigma} < 1$, then:

$$\frac{\sqrt{\gamma\varsigma}\lambda_M - \lambda_1}{1-\sqrt{\gamma\varsigma}} \le \left|\breve{\lambda}\right| \le \frac{\sqrt{\gamma\varsigma}\lambda_1 - \lambda_M}{1-\sqrt{\gamma\varsigma}} \tag{1.45}$$

If $\sqrt{\gamma\varsigma}\lambda_1 - \lambda_M > 0$, then $\sqrt{\gamma\varsigma} > \lambda_M/\lambda_1$, since $\sqrt{\gamma\varsigma}\lambda_M - \lambda_1 < 0$, but $\left|\breve{\lambda}\right| > 0$. Therefore, if $\lambda_M/\lambda_1 < \sqrt{\gamma\varsigma} < 1$, we can have:

$$\lambda_{\min}^{(2)} \overset{\Delta}{=} -\frac{\sqrt{\gamma\varsigma}\lambda_1 - \lambda_M}{1-\sqrt{\gamma\varsigma}} \le \breve{\lambda} \le \frac{\sqrt{\gamma\varsigma}\lambda_1 - \lambda_M}{1-\sqrt{\gamma\varsigma}} \overset{\Delta}{=} \lambda_{\max}^{(2)} \tag{1.46}$$

Since $\breve{\lambda}_{\max}^{(2)} > 0$, and $\breve{\lambda}_{\min}^{(2)} = -\breve{\lambda}_{\max}^{(2)} < 0$, with the above analysis of NICCB, we can obtain that when $\lambda_M/\lambda_1 < \sqrt{\gamma\varsigma} < 1$ there isn't a solution in the bound of $\left[0,\ \breve{\lambda}_{\max}^{(2)}\right]$ to NECCB, but the solution in the bound of $\left[\breve{\lambda}_{\min}^{(2)},\ 0\right]$ is the true solution to NECCB.

From above analysis, we can conclude as follows:

1. When $1 < \sqrt{\gamma\varsigma} < \lambda_1/\lambda_M$, the solution in the bound of $\left[\breve{\lambda}_{\min}^{(1)},\ 0\right]$ is the true solution to NECCN, and the norm equality constraint parameter ς should be chosen in the interval defined by $1/\gamma < \varsigma < \min\left\{\left(\lambda_1/\lambda_M\right)^2/\gamma,\ \ \varsigma_0\right\}$.
2. When $\lambda_M/\lambda_1 < \sqrt{\gamma\varsigma} < 1$, the solution in the bound of $\left[\breve{\lambda}_{\min}^{(2)},\ 0\right]$ is the true solution to NECCB, and the norm equality constraint parameter ς should be chosen in the bound of $\left(\lambda_M/\lambda_1\right)^2/\gamma < \varsigma < \min\{1/\gamma,\ \ \varsigma_0\}$.
3. NECCB has the form of diagonal loading with negative loading level, but NICCB has the form of diagonal loading with positive loading level.

1.5 Simulation analysis

In order to validate the correctness and the efficiency of the proposed algorithms, we analyze as follows. In our simulations, we assume a uniform linear array with N=10 omnidirectional sensors spaced half a wavelength apart. Through all examples, we assume that there is one desired source, namely, there is a signal from direction 0^o, the Signal Noise Ratio (SNR) is -5dB. Therein, the presumed signal direction is equal to 5^o (i.e., there is a 5^o direction mismatch).

For the comparison, the benchmark standard Capon beamforming algorithm that corresponds to the ideal case when the covariance matrix is estimated by the maximun likelihood estimator (MLE) and the actual steering vector is used, this algorithm does not correspond to any real situation but is included in our simulations for the sake of comparison only, and is denoted by Ideal-SCB in the figure. The other algorithms include: standard Capon beamformer (SCB), NICCB, NECCB. For NICCB and NECCB, the constraint parameter is selected as the median of the allowable bound.

1.5.1 Effectivity analyzing

In order to show the effectivity of the proposed algorithms, we first compare the pattern of the mentioned Capon beamforming algorithms. The Capon beamformer pattern is given in Fig. 1. Since the signal direction mismatch, the mainlobe of SCB departs from the signal direction. The performance of NICCB is slightly better than SCB, and NECCB is the best of all, the direction mismatch is overcame commendably and NECCB also has lower sidelobe level. Here, NICCB uses the positive optimal loading level, NECCB uses the negative optimal loading level. From the comparison, we can see that NECCB has the better performance than NICCB.

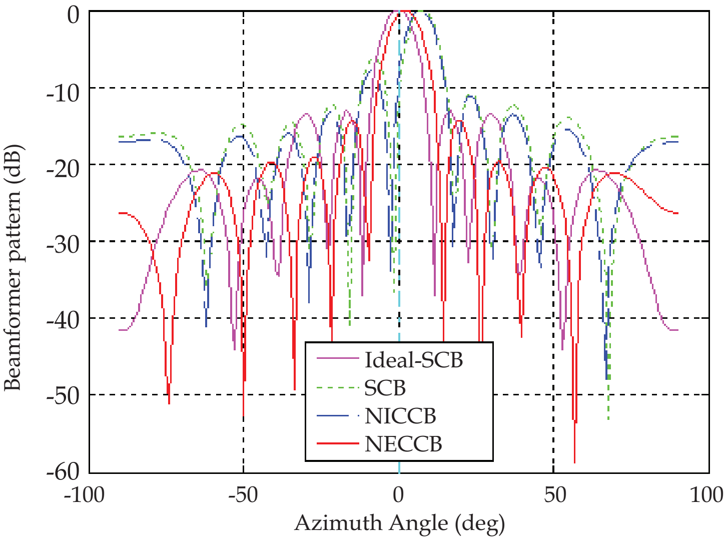


Fig. 1. Capon beamformer pattern comparison

The variation of the beamformer SNR versus samples number is given in Fig. 2. We can see that with the change of the samples number, the SNRs varies accordingly. The SNR of NICCB is almost closed to the SNR of SCB, and is lower than the SNR of Ideal-SCB, but NECCB is the best of all, especially for the small number, it has preferable performance. Hence, the norm constraint can improve the SNR, and NECCB has the highest SNR among the listed algorithms.

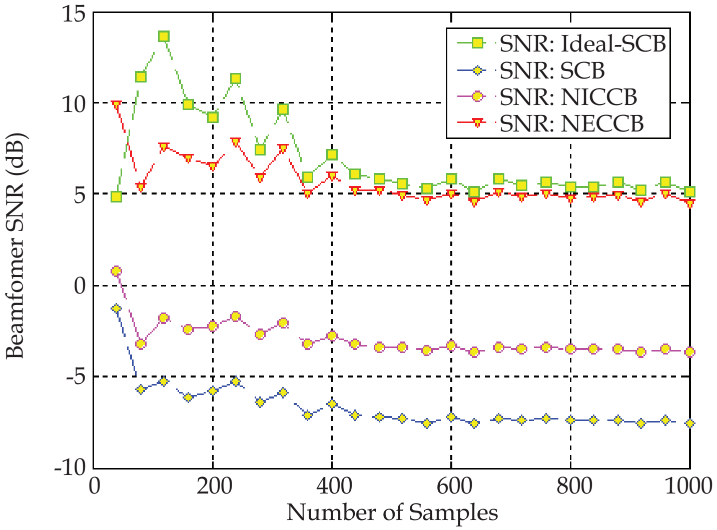


Fig. 2. Output SNR versus samples number

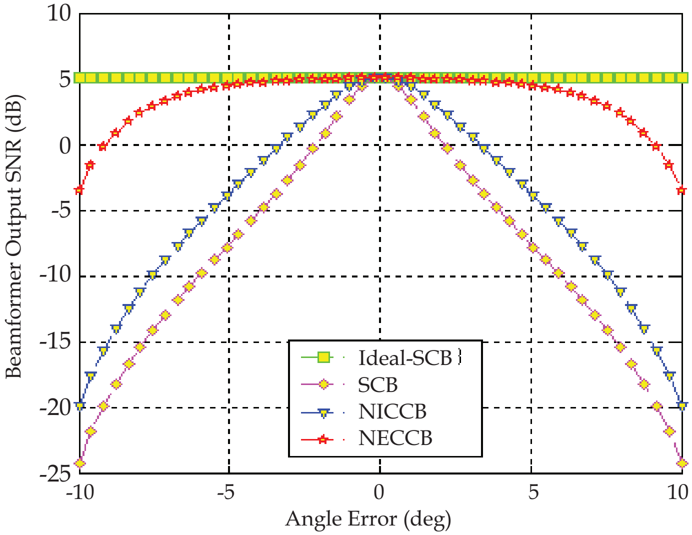


Fig. 3. Output SNR versus angle mismatch

The variation of the Capon beamformer output signal-to-noise-ratio (SNR) versus signal direction mismatch is given in Fig. 3. We can see that with the change of the signal direction mismatch, the SNR varies accordingly, when the angle error is in the range of [-7°, 7°], NECCB will has higher SNR than SCB, NICCB. The NECCB has the higher SNR can be explained by the Fig. 1 of the beam pattern comparison, NECCB not only has the good pointing performance, but also has the lower sidelobe level. Namely, for the same desired signal output, the output noise of NECCB is lower. The simulation results can also be explained as follows, for the used scene, the Signal Noise Ratio is -5dB, and for NECCB, the optimal Lagrange is negative, namely the optimal loading level is negative, but for others, the loading level is zero or positive. Therefore, for the NECCB beamformer, the output noise power is decreased, but for other beamformers, the output noise power is increased. Hence, the NECCB has higher output SNR than others. For the sake of saving space, the corresponding beam pattern comparison isn't given, but in the simulation, NECCB pattern also points to the actual signal direction exactly. Hence, NECCB has the better robustness in the signal direction mismatch case.

From above analysis, we can see that NECCB has the best robustness against the signal direction mismatch.

1.5.2 Correctness analyzing

Since NICCB and NECCB have the same form as that of SCB with diagonal loading. But their key problems are how to find their own optimal loading level or Lagrange multiplier, In order to show the impact of loading level on the Capon beamformer under norm constraint (NCCB) and attest the correctness of the proposed algorithms, the simulation results are given as follows.

The variation of the output SNR versus diagonal loading level is given in Fig. 4. We can see that with the change of the loading level in the bound of $\left[\breve{\lambda}_{\min}^{(1)}, \breve{\lambda}_{\max}^{(1)}\right]$, the SNR of NCCB varies accordingly. When the loading level is positive, NCCB is NICCB, whereas, when the loading level is negative, NCCB is NECCB. By comparison, we can see that NECCB has higher SNR than NICCB, but for the optimal loading, namely when the loading level is equal to -6.09, NECCB has the best pointing performance, and its SNR is the highest one, where the optimal loading level -6.09 is calculated using the equation $h(\lambda)=\varsigma$ with $1<\sqrt{\gamma\varsigma}<\lambda_1/\lambda_M$ in the bound of $\left[\breve{\lambda}_{\min}^{(1)},\ \ 0\right]$. Hence, the loading level has a great impact on the SNR of the Capon beamformer, and determines the performance improvement.

The variation of the weight vector norm versus diagonal loading level is given in Fig. 5. We can see that with the change of the loading level in the bound of $\left[\breve{\lambda}_{\min}^{(1)}, \breve{\lambda}_{\max}^{(1)}\right]$, the weight vector norm of NCCB varies accordingly. For most loading levels, the weight vector norm varies slightly, but when the loading level is small in negative domain, the weight vector norm is a little high, and the highest point is corresponding to the lowest point in Fig. 4. Therefore, the loading level has a great impact on the weight vector norm, especially for the negative loading.

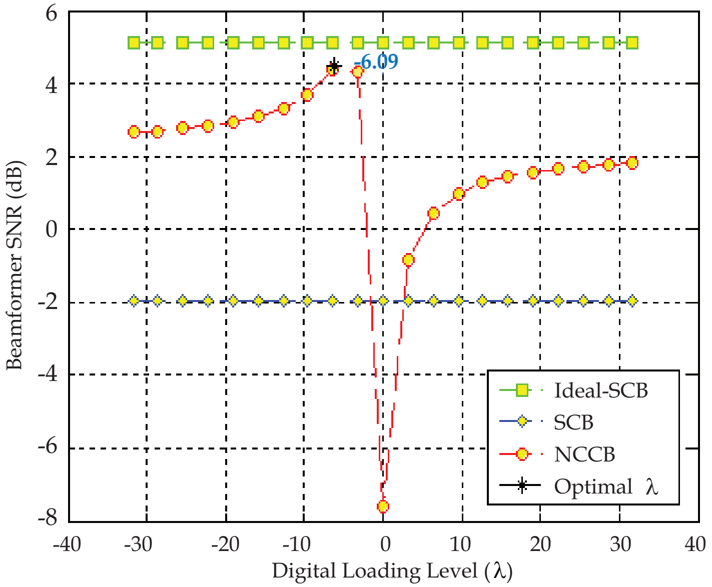


Fig. 4. Output SNR versus loading level

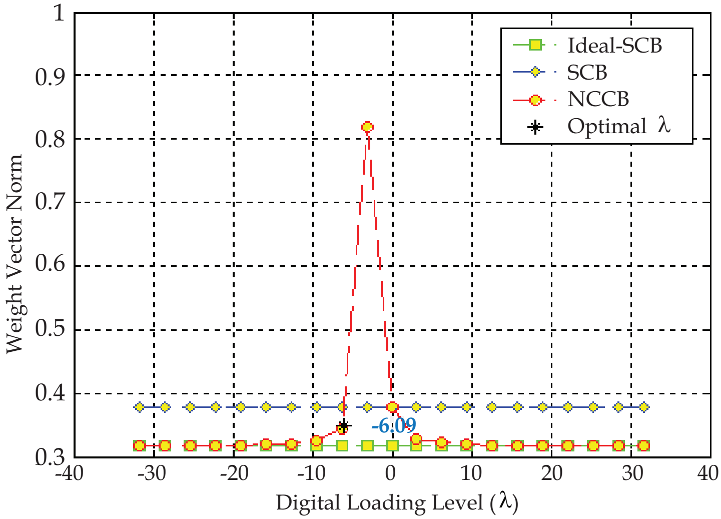


Fig. 5. Weight vector norm versus loading level

From the above simulation results, we can see that the loading level has a great impact on the performance of the Capon beamformer, and NECCB has the best pointing performance, namely, the optimal negative loading is the best. This is also consistent to the theory analysis, for the robust beamformer with diagonal loading, the improvement is determined by the optimal loading level, when the loading level is optimal, the performance improvement will be the optimal, but for other values, the improvement will be little, or even worse.

1.5.3 Constraint parameter selection analyzing

For NCCB, there are two key problems, one is how to find the optimal loading level, and the other is how to select the norm constraint parameter. Although we have solved the two problems in theory, but there is another key problem, namely, how to select the optimal norm constraint parameter. Therefore, the impact of norm constraint parameter on NCCB is analyzed here particularly.

The variation of the output SNR versus norm constraint parameter is given in Fig. 6. We can see that with the change of the norm constraint parameter in the allowable bound of $(\varsigma_{\min}, \ \varsigma_{\max})$, the SNR of the Capon beamformer varies accordingly. NICCB has a little higher SNR than that of SCB, NECCB has the highest SNR. And with the norm constraint parameter increasing, the SNR of NECCB increases correspondingly, but the SNR of NICCB is inclined to the SNR of SCB. When the norm constraint parameter is equal to the maximum, the constraint is inactive, and the three SNRs tend to the same value. Hence, the SNR is determined by the choice of the norm constraint parameter, especially for NECCB.

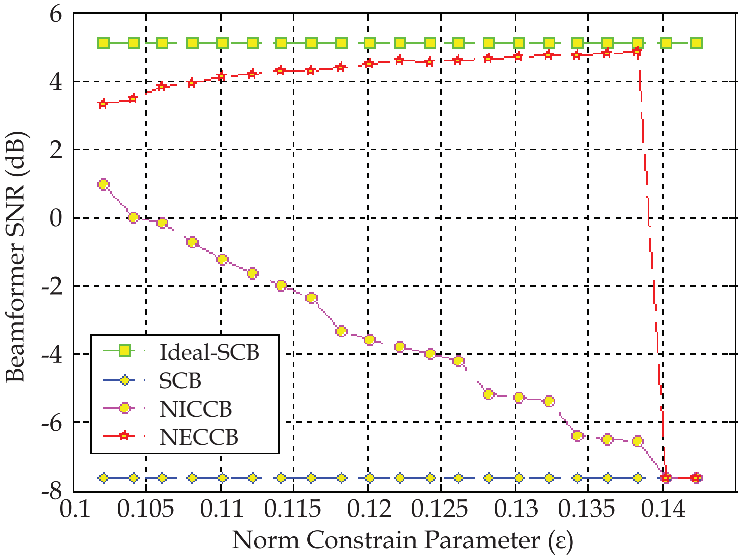


Fig. 6. Output SNR versus constraint parameter

The variation of the weight vector norm versus norm constraint parameter is given in Fig. 7. When the norm constraint parameter is selected in the allowable bound of $(\varsigma_{\min}, \ \varsigma_{\max})$, the weight vector norms of NICCB and NECCB vary adaptively, and are equal to the square root of the constraint parameter approximatively, this is consistent with the theory, namely the solution is obtained on the constraint boundary. The slight difference is caused by the approximative computation.

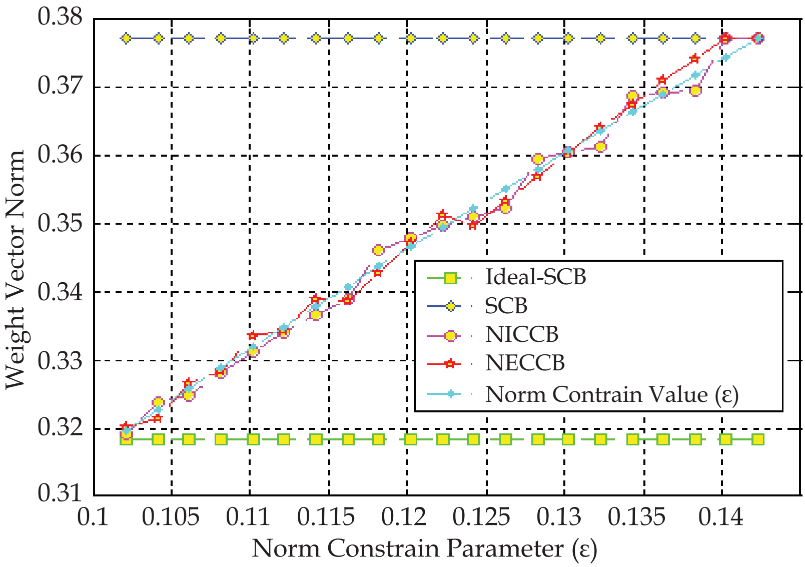


Fig. 7. Weight norm versus constraint parameter

From above simulation results, we can see that if the norm constraint parameter is selected in the allowable bound, the norm constraint parameter has a great impact on the performance of NICCB and NECCB, especially for NECCB. But NECCB with the larger constraint parameter has the better pointing performance, namely, when the constraint parameter is selected as more larger in its allowable bound, the optimal negative loading has the optimal improvement.

1.6 Conclusion

From the above analysis, we can conclude as follows. (I) The proposed algorithm is correct and effective. (II) The norm constraint can improve the robustness of the Capon beamformer. Especially, the equality constraint has the preferable improvement to overcome the steering vector mismatch, and also has good robustness for the samples number. (III) When the norm constraint parameter is selected in the allowable bound, NECCB has the best performance, namely the optimal negative loading has the optimal improvement, this is because that the norm equality constraint is stronger than the norm inequality constraint.

2. Improved pattern synthesis method with linearly constraint minimum variance criterion

Antenna pattern synthesis becomes the fundamental research contents with the wide application of the array antenna in communication, radar and other areas, and catches the attentions widely. The array antenna pattern synthesis is the task which solves the weight values of the every element to force the antenna pattern inclining to the anticipant shape. Dolph has first given the method of getting the weight function for uniform linear array to achieve the Chebychev pattern [39], therefore the optimal solution can be achieved in the sense of giving the mainlobe width and the maximum lowest sidelobe level. However, how to implement the pattern synthesis for the arbitrary array antenna efficiently is a challenging research task in array signal processing society.

Currently, the methods of pattern synthesis can be classified as the two types, one is the traditional vector weight methods [40-42], the other one is the matrix weight methods [43], therein, the intelligent computer methods are used to improve the calculating efficiency of the optimal weight vector, such as the genetic evolution algorithms [44] and the particle swarm optimization algorithms [45]. However, for any pattern synthesis method, the iterative operation can't be avoided, and the iterative number determines the operation load directly, the operation load, or titled as the compute efficiency is the key metric to evaluate the validity of the pattern synthesis.

Guo Q et al propose the pattern synthesis method for the arbitrary array antenna with the linearly constraint minimun variance criterion (LCMV-PS) [45], compared with the traditional vector weight methods, this algorithm has the small iterative number and the preferable convergence. However, by analysis and simulation, it is found that the iterative coefficient determines its performance, namely, the iterative coefficient not only determines the pattern shape, but also determines the iterative number, or titled as the compute load. Therefore, how to select the iterative formula and its iterative coefficient is the key problem to reduce the compute load and enhanced the applicability.

In this chapter, for the LCMV-PS method proposed in [45], by analyzing its implementation and jammer power iterative formula, the improved fast robust LCMV-PS method is proposed [46]. This algorithm takes into account the effect of the relative amplitude between synthesis pattern and its reference upon the pattern synthesis adequately, via adding a proportion constant to the iterative formula, the effect of their relative amplitude upon the changing ratio of the jammer power is strengthened, not only the iterative efficiency of the jammer power is improved, namely the iterative number is reduced, and the pattern synthesis efficiency is improved, but also the selecting bound of the iterative coefficient is extended, namely the effect of the iterative coefficient upon the pattern synthesis is weakened, and the application area and applicability of the pattern synthesis method is enhanced greatly. The last simulation attests its correctness and effectiveness.

2.1 Pattern synthesis method with LCMV criterion

The problem of array pattern synthesis can be simplified as follows, namely for the given element number M and element position $\{\mathbf{x}_i\}_{i=1}^{M}$, solving the complex weight vector **w**, and force the array pattern $P(\theta)$ with the definite width and maximum value in the desired direction, at the same time, make the sidelobe level according to the requirement.

The target of the pattern synthesis method for arbitrary arrays based on LCMV criterion (LCMV-PS) is making all the sidelobe peak level equal to the minimum that the array can achieve as possible. Furthermore, this method constructs many illusive jammers in the sidelobe region, and the jammer power will be justed by the synthesis pattern amplitude in its relative direction, namely, if the synthesis pattern amplitude is high in this direction, the illusive jammer power will be increased. Therefore, the LCMV-PS method can be simple described as follows:

1. Specify the mainlobe region $\left[\theta_{ML1},\theta_{ML2}\right]$ and sidelobe envelope $D(\theta_i)$.
   Set the initial value of jammer power $f_0(\theta_i)$, if θ_i is in the sidelobe region, $f_0(\theta_i)=1$, otherwise if in the mainlobe region, $f_0(\theta_i)=0$, $i=1,\cdots N$, where N is the number of the uniformly distributed jammers with one degree spacing, namely, $\theta_1,\theta_2,\cdots,\theta_N$ are the degree values with the one degree spacing in the array pattern overlay region. And $D(\theta_i)$ is the given reference sidelobe envelope of the synthesis pattern.
2. Calculate jammer powers for the k-th iteration $f_k(\theta_1),f_k(\theta_2),\cdots,f_k(\theta_N)$.
   If *k*=0, then the jammer powers are the initial values $f_0(\theta_1),f_0(\theta_2),\cdots,f_0(\theta_N)$.
   If *k*≥1, there is the iterative formula as follows:

$$f_k(\theta)=\begin{cases} 0 & \theta\in\left[\theta_{ML1},\theta_{ML2}\right] \\ \max\left\{f_{k-1}(\theta)+Kf_{k-1}(\theta)\dfrac{P_{k-1}(\theta)-Pr_{k-1}}{Pr_{k-1}},\ \ 0\right\} & \theta\notin\left[\theta_{ML1},\theta_{ML2}\right]\end{cases} \tag{2.1}$$

where $f_{k-1}(\theta)$ is the jammer powers of the *k*-1-th iteration, K is the iterative coefficient. $P_{k-1}(\theta)=\left|\mathbf{w}^H\mathbf{a}(\theta)\right|$ is the pattern of the *k*-1-th iteration, therein $\mathbf{w}$ is the relative weight vector, $\mathbf{a}(\cdot)$ is the steering vector, and $(\cdot)^H$ denotes the conjugate transposition. Pr_{k-1} is the sidelobe reference amplitude, if the arbitrary sidelobe shape is required in the pattern synthesis, it is only to substitute $Pr_{k-1}(\theta)=Pr_{k-1}\cdot D(\theta)$ for Pr_{k-1} in the above formula, and $D(\theta)$ is the given sidelobe envelope.

3. Calculate the data covariance matrix $\mathbf{R}_x$, namely:

$$\mathbf{R}_x=\mathbf{A}\cdot diag\left[f_k(\theta_1),f_k(\theta_2),\cdots,f_k(\theta_N)\right]\cdot\mathbf{A}^H+\sigma\mathbf{I} \tag{2.2}$$

where $\mathbf{A}=\left[\mathbf{a}(\theta_1),\mathbf{a}(\theta_2),\cdots,\mathbf{a}(\theta_N)\right]$ is the array manifold matrix. σ is a given small quantity, and $\mathbf{I}$ is the identity matrix, $\sigma\mathbf{I}$ is added to prevent the covariance matrix from being ill-conditioned.

4. Calculate the weight vector $\mathbf{w}$ according with the following LCMV beamforming algorithm, then synthesize the pattern. If it is satisfactory, stop; otherwise, go to step (2) and continue. Therein, $\mathbf{w}$ is solved by the below LCMV optimization problem, namely:

$$\begin{cases}\min\limits_{\mathbf{w}}\mathbf{w}^H\mathbf{R}_\mathbf{x}\mathbf{w} \\ s.t.\quad \mathbf{C}^H\mathbf{w}=\mathbf{f}\end{cases} \tag{2.3}$$

where $\mathbf{C}$ is the M×m constraint matrix, and $\mathbf{f}$ is the m×1 constraint value vector. Its optimal solution is:

$$\mathbf{w}=\mathbf{R}_\mathbf{x}^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_\mathbf{x}^{-1}\mathbf{C}\right)^{-1}\mathbf{f} \tag{2.4}$$

In the constraint condition of the optimization problem, the constraint of the mainlobe can be imposed, the constraint of the sidelobe can also be added, in other words, the constraint condition and parameter can be selected according to the pattern synthesis requirement.

2.2 Improvement of the jammer power iteration formula

From the step of the LCMV-PS method, we can see that the key is the jammer power iteration in step (2), since it not only determines the synthesis pattern shape, but also determines the final iterative number.

By the particular analysis of the LCMV-PS implementing steps, it is not difficulty to find that although the relative difference of the synthesis pattern and the reference pattern $\left(P_{k-1}(\theta) - Pr_{k-1}\right)/Pr_{k-1}$ is used as the ratio factor for the gain change, and to control the change direction and quality of the jammer powers, actually, the expression of the iterative formula $f_{k-1}(\theta) + Kf_{k-1}(\theta)\left(\left(P_{k-1}(\theta) - Pr_{k-1}\right)/Pr_{k-1}\right)$ can be transformed as:

$$f_{k-1}(\theta) + Kf_{k-1}(\theta)\frac{P_{k-1}(\theta) - Pr_{k-1}}{Pr_{k-1}} = f_{k-1}(\theta)\left(K \cdot \frac{P_{k-1}(\theta)}{Pr_{k-1}} + (1-K)\right) \tag{2.5}$$

This expression indicates that the jammer powers between the adjacent iterations are different by a proportional factor, when the iterative coefficient K is given, the jammer power ratio of the adjacent iterations is determined by the relative amplitude of the synthesis pattern and the reference pattern $P_{k-1}(\theta)/Pr_{k-1}$. Therefore, for the given K, the change of the jammer powers in the iteration process is determined by $P_{k-1}(\theta)/Pr_{k-1}$, and the relationship is a linear function.

From the pattern synthesis process of the LCMV-PS method, when the synthesis pattern is higher than the reference pattern, the jammer powers should increase, and is in direct proportion to the difference of the two patterns. When the synthesis pattern is more higher than the reference pattern, the jammer powers should increase more larger. But when the synthesis pattern is close to the reference pattern, the change of the jammer power should be small, namely the adjustment should be precise at this time. Although the original iterative formula is consistent to the analyzing idea, and the change ratio of the iterative jammer powers is $K \cdot \left(P_{k-1}(\theta)/Pr_{k-1}\right) + (1-K)$, namely is in direct proportion to $P_{k-1}(\theta)/Pr_{k-1}$. Therefore, for the original method, K is the main parameter to determine the iterative effect and efficiency, and by the simulation, it is found that the synthesis pattern will be good when the parameter K is selected in a small region, such as the reference value K=0.1 in [42]. Actually, for the difference element number or array parameter, the optimal value of K will vary correspondingly. Hence, for the original method, how to select the optimal parameter K is the key matter, it not only determines the effect of the synthesis pattern, but also determines the efficiency of the jammer power iteration, namely the iterative number.

Since in the iterative process, it is the factor $K \cdot \left(P_{k-1}(\theta)/Pr_{k-1}\right) + (1-K)$ determining the change quantity and direction of the jammer powers iteration, for the given K, the second item is constant, but the first item is the linear function of $P_{k-1}(\theta)/Pr_{k-1}$, and its proportional coefficient is K, namely the slope K determines the change quantity of the jammer power with $P_{k-1}(\theta)/Pr_{k-1}$. With the slope of the linear function increasing, namely for the given parameter K_p>1, the change ratio of $K \cdot K_p \cdot \left(P_{k-1}(\theta)/Pr_{k-1}\right) + (1-K)$ with $P_{k-1}(\theta)/Pr_{k-1}$ will be larger, namely the efficiency of the jammer power iteration will be improved. At the same time, for the given parameter K_p, when the effect of the jammer power iteration is better, the selection of K will be loosened, namely K can be selected in a wider region. Therefore, if the constant factor K_p (K_p>1) can be added as this method, the efficiency of the jammer power iteration will be improved, and the bound for selecting the iterative coefficient *K* will be enlarged, namely the selection of *K* will be simplified greatly.

Hence, in order to improve the iterative efficiency of the LCMV-PS method and simplify the selection of the iterative coefficient *K*, the iterative formula of the jammer power can be improved as follows, namely

$$f_k(\theta) = \begin{cases} 0 & \theta \in [\theta_{ML1}, \theta_{ML2}] \\ \max\left\{ f_{k-1}(\theta) + K \cdot f_{k-1}(\theta) \cdot \dfrac{K_p \cdot P_{k-1}(\theta) - Pr_{k-1}}{Pr_{k-1}},\ 0 \right\} & \theta \notin [\theta_{ML1}, \theta_{ML2}] \end{cases} \tag{2.6}$$

where K_p is used to adjust the effect of the relative amplitude $P_{k-1}(\theta)/Pr_{k-1}$ of the synthesis pattern and reference pattern upon the change ratio of the jammer power, namely is used to adjust the iterative efficiency of the pattern synthesis method, and other parameters have the same sense as forenamed. If K_p>1, the iterative efficiency will be advanced, whereas the iterative efficiency will be reduced. It is important that the effect of the iterative coefficient *K* upon the pattern synthesis is reduced greatly by adding the parameter K_p, therefore, the selection of parameter *K* will be simplified greatly.

Compared with the LCMV-PS method proposed in [42], the iterative formula of the jammer power is added by a constant K_p to adjust the iterative efficiency in this chapter, if K_p>>1, the efficiency of the proposed method will be improved greatly, therefore, the iterative number will be reduced, so that the operation load will be reduced greatly by the proposed method. At the same time, the bound for selecting *K* will also be enlarged greatly, and the application area and applicability of the pattern synthesis method is enhanced.

2.3 Simulations

Since the proposed method has the higher iterative efficiency and stronger applicability as compared with the LCMV-PS method proposed in [42], the simulation keystone is to compare the iterative efficiency of the two methods, and the effect of the iterative coefficient upon the two methods. The simulations are as follows.

2.3.1 Efficiency analyzing

In order to validate the efficiency of the proposed method, the single beam and multi-beam pattern synthesis examples of the uniform and non-uniform linear array are given respectively, and the single beam pattern synthesis examples of uniform and non-uniform planar array are also given respectively. For the convenience of comparison, the proposed improved LCMV-PS method is denoted as I-LCMV-PS, the LCMV-PS in [42] is denoted as LCMV-PS, the reference pattern is denoted as Reference.

The single beam synthesis pattern of the uniform linear array is given in Fig. 8. Therein, the element number is 32, the elements inter-space is half wavelength (λ/2), the mainlobe is point to 0°, the mainlobe width is 22°, the iterative parameter of jammer power *K*=0.5, and K_p=100. When the optimal weight vector is solved under the LCMV criterion, the mainlobe direction constraint is used only. Therein the iterative number of I-LCMV-PS is 5, but the iterative number of LCMV-PS is 20.

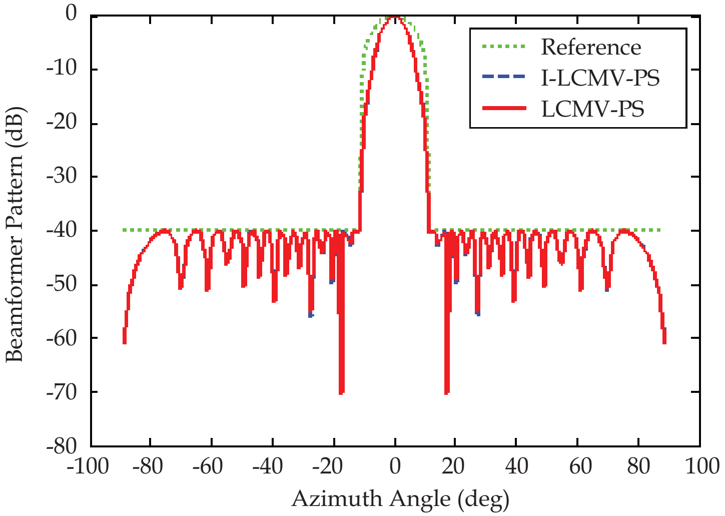


Fig. 8. Single beam synthesis pattern of uniform linear array

The single beam synthesis pattern of the non-uniform linear array is given in Fig. 9. Therein, the element number is 32, the element space vector is (λ/2)×[0.29595, 1.5655, 2.7845, 3.9334, 4.999, 5.9753, 6.8645, 7.6764, 8.428, 9.1413, 9.842, 10.557, 11.311, 12.127, 13.021, 14.002, 15.073, 16.226, 17.449, 18.72, 20.017, 21.312, 22.579, 23.795, 24.939, 26, 26.971, 27.856, 28.664, 29.413, 30.125, 30.825], the mainlobe is also point to 0°, the mainlobe width is also 22°, and K=0.5, K_p=100. When the optimal weight vector is solved under the LCMV criterion, the mainlobe direction constraint is used only. The iterative number of the two methods are 5 and 20, respectively.

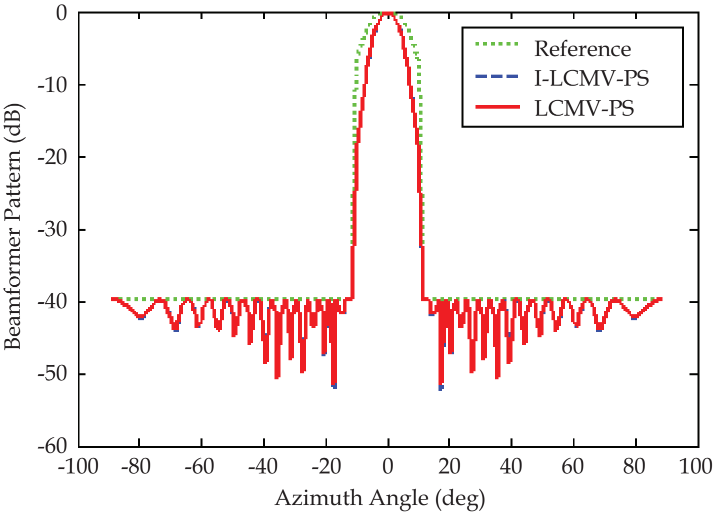


Fig. 9. Single beam synthesis pattern of non-uniform linear array

In order to attest the applicability of the improved algorithm to the planar array, and its effectivity of multi-beam synthesis, namely the arbitrary array and arbitrary pattern synthesis, the particular simulation examples are given as Fig. 11.~Fig. 16.

The multi-beam synthesis patterns of the uniform and non-uniform linear array are given in Fig. 11. and Fig. 12. The parameters are same as Fig. 8. and Fig. 9., the two beams point to 45^o and -45^o respectively. Therein the iterative number of I-LCMV-PS is 6, but the iterative number of LCMV-PS is 25.

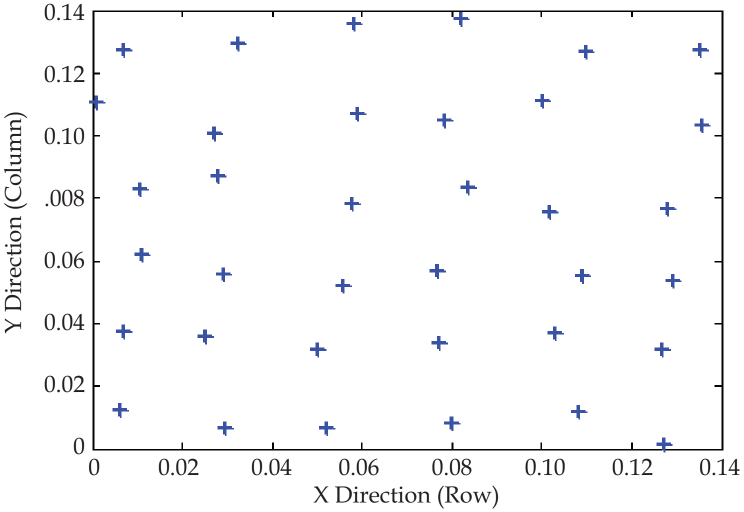


Fig. 10. Element position of the non-uniform planar array

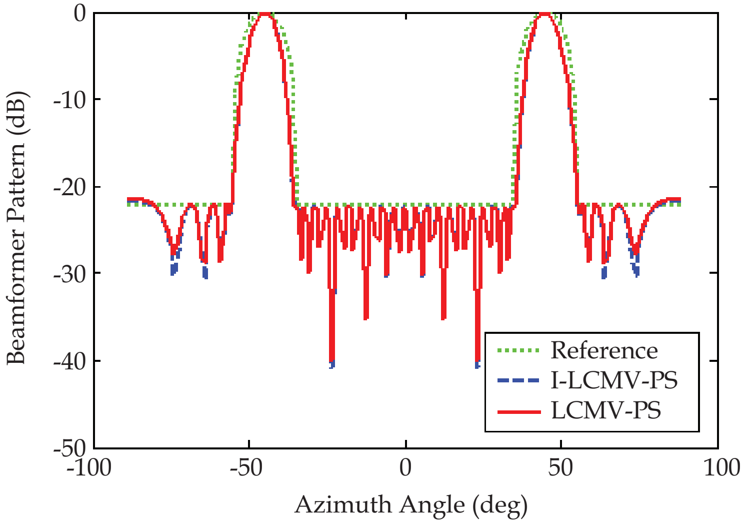


Fig. 11. Multi-beam synthesis pattern of uniform linear array

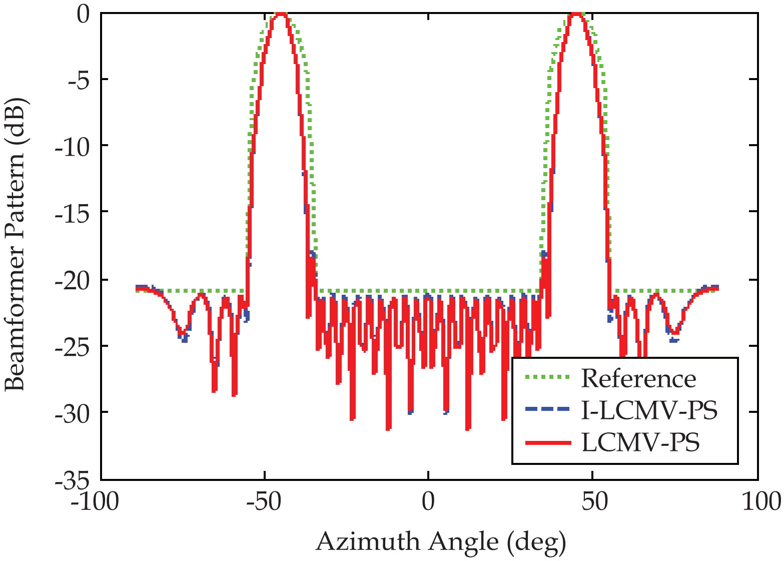


Fig. 12. Multi-beam synthesis pattern of non-uniform linear array

The single beam synthesis patterns of the uniform and non-uniform planar array with I-LCMV-PS are given in Fig. 13.~ Fig. 16. Therein, they have the same element number 36, the uniform planar array is phalanx, and element space is half wavelength, but the element position of the non-uniform planar array is given as Fig. 10. The mainlobe of the two array point to (0º,0º), and the beam-widths in the azimuth and elevation direction are all 30º. In the simulation, the iterative number of I-LCMV-PS is 8, and the Fig. 14. and Fig. 16. is the side view figure of Fig. 13. and Fig. 15. respectively.

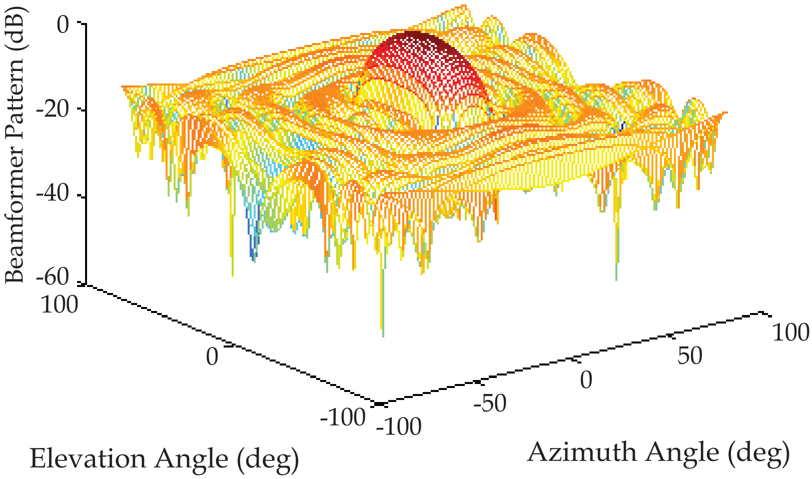


Fig. 13. Single beam synthesis pattern of uniform planar array (1)

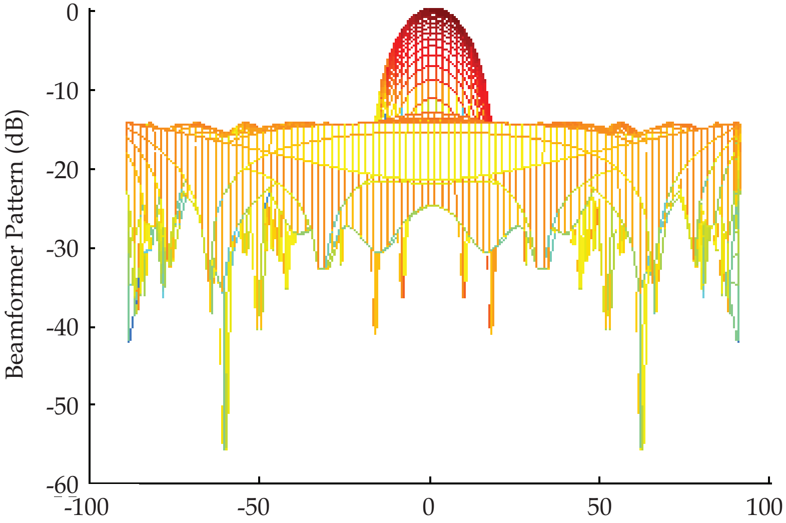


Fig. 14. Single beam synthesis pattern of uniform planar array (2)

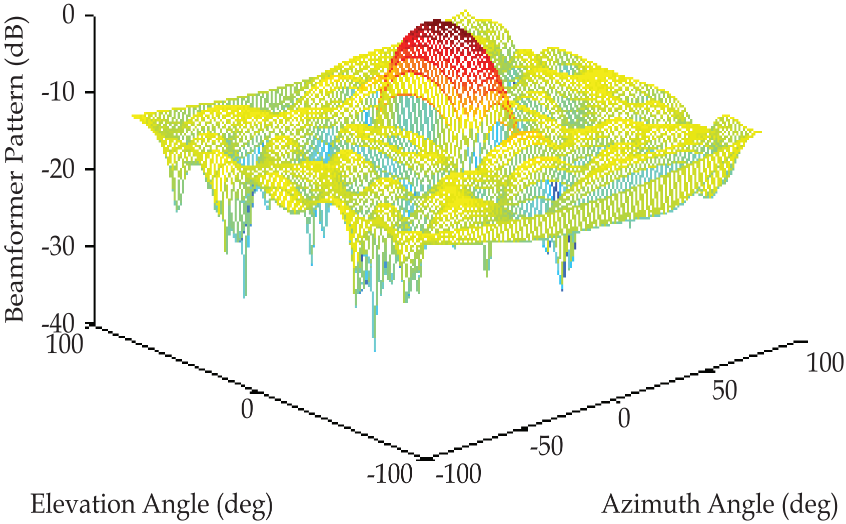


Fig. 15. Single beam synthesis pattern of non-uniform planar array (1)

From the above simulations, we can see that the two methods have the preferable synthesis pattern, but I-LCMV-PS has small iterative number, namely it has the higher pattern synthesis efficiency.

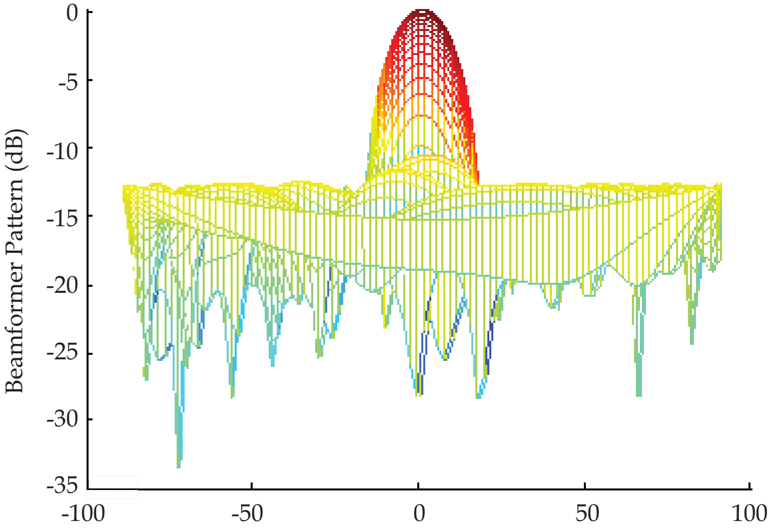


Fig. 16. Single beam synthesis pattern of non-uniform planar array (2)

2.3.2 Convergence analyzing

In order to compare the convergence characteristic, limit by the chapter length, here the example of the uniform linear array is given, about the examples of the non-uniform linear array and the planar array are similar to the uniform linear array.

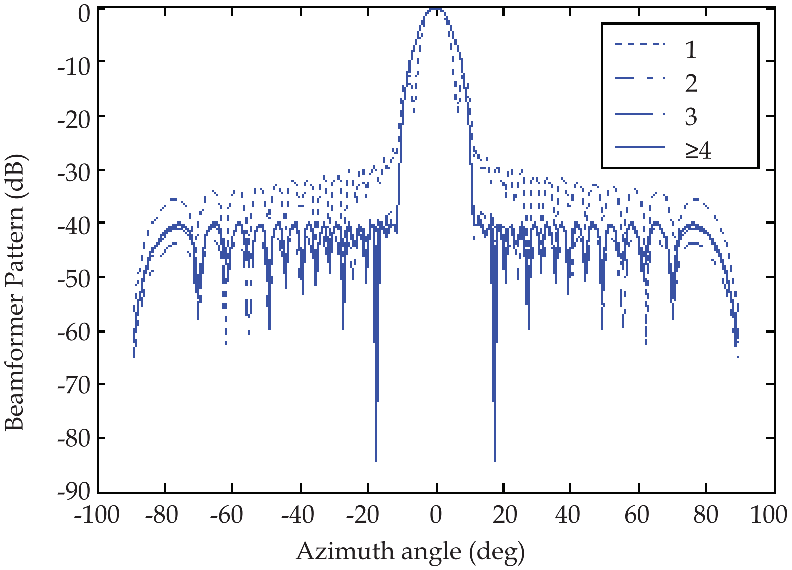


Fig. 17. Synthesis pattern versus iterative number of I-LCMV-PS

Fig. 10. and Fig. 11. give the convergence of the synthesis pattern for the uniform linear array with I-LCMV-PS and LCMV-PS respectively. Therein, the parameters are as same as 2.3.1. From Fig. 10., we can see that when the iterative number is larger than 4, I-LCMV-PS can achieve the preferable pattern synthesis effect, but for LCMV-PS, the iterative number must be larger than 20, because I-LCMV-PS has the higher efficiency of the jammer power iteration.

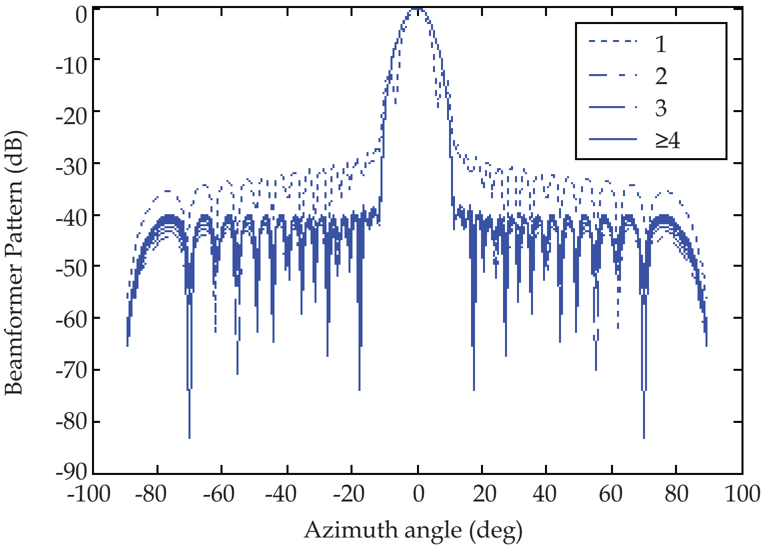


Fig. 18. Synthesis pattern versus iterative number of LCMV-PS

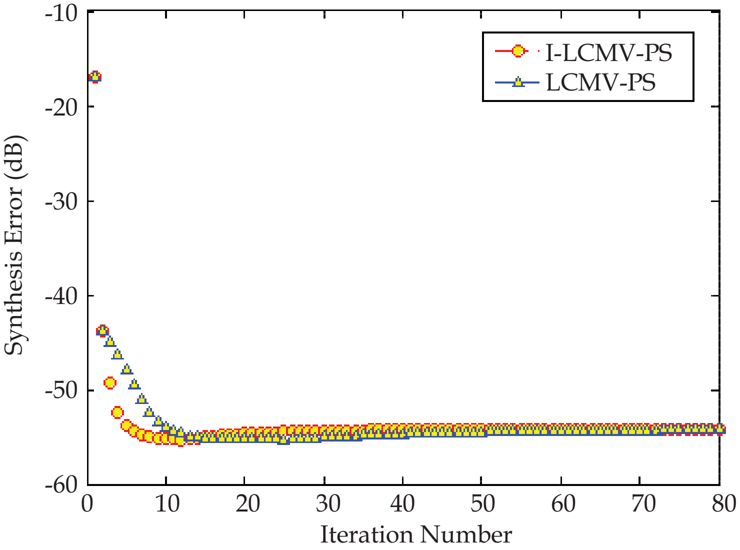


Fig. 19. Pattern synthesis error versus iteration number

In order to attest the convergence and synthesis precision of the improved algorithm, Fig. 19. gives the pattern synthesis error versus the iterative number. Therein, the pattern synthesis error is calculated as follows:

$$E_{sum} = \sum_{i=1}^{N} \left| P_k(\theta_i) - Pr(\theta_i) \right| \tag{2.7}$$

where $Pr(\theta)$ is the reference pattern, $P_k(\theta)$ is the synthesis pattern after k-th iteration. From the error curve, we can see that the improved algorithm has the fast convergence performance, and fall rapidly from beginning, at last, the two curves converge at the same value, namely the convergence is consistent with the Fig. 17. and Fig. 18. Therefore, the proposed algorithm has the good convergence and synthesis precision.

In order to analyze the effect of the iterative coefficient upon the pattern synthesis, Fig. 20. and Fig. 21. give the synthesis pattern versus iterative coefficient for the uniform linear array with I-LCMV-PS and LCMV-PS respectively. Therein, the parameters are as same as 2.3.1. From Fig. 20., we can see that when K_p=100, the selection of iterative coefficient K has very little effect upon the pattern synthesis, in simulation, when 0.005<K<2000, the preferable performance can be achieved, the 2000 is the upper bound in simulation, if the parameter is larger than this value, the good performance can also be achieved. Actually, if K_p larger, the selecting bound for K wll be wider, it is consistent with the theory analysis. But from Fig. 21., we can also see that the efficiency of the pattern synthesis is determined by K, in the simulation, we find that when 0.1<K<1.6, the preferable effect is achieved. By the comparison, we can see that the improved LCMV-PS method has the lower dependence upon the iterative coefficient, and enables the selecting bound for the iterative coefficient K enlarged greatly, and enhanced its application area and applicability.

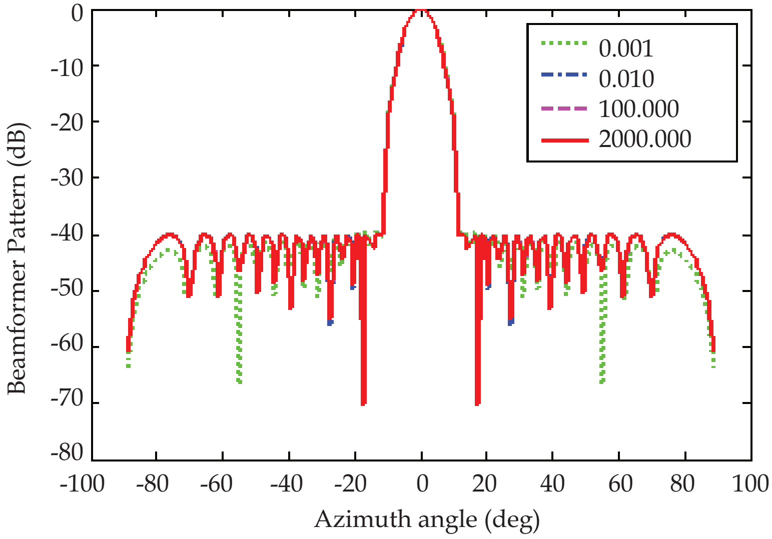


Fig. 20. Synthesis pattern versus iterative coefficient of I-LCMV-PS

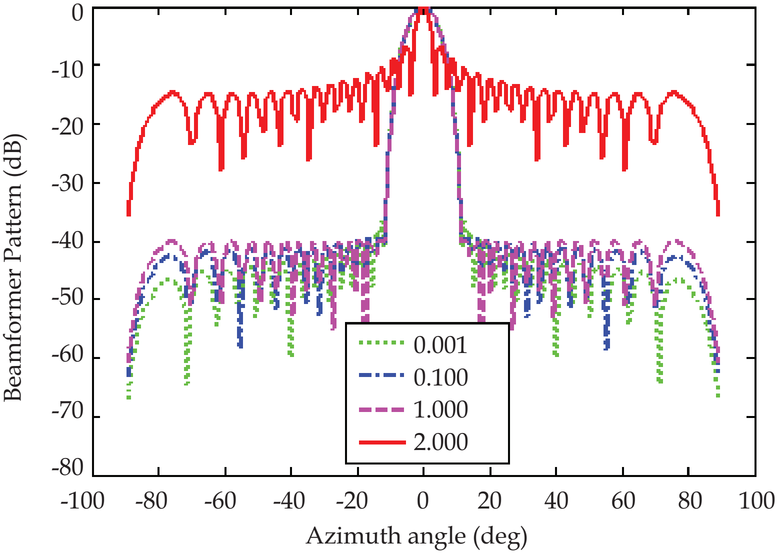


Fig. 21. Synthesis pattern versus iterative coefficient of LCMV-PS

2.4 Conclusion

From the above analysis and simulation, we can conclude as follows: (I) The proposed jammer power iterative formula is correct and effective. (II) By the improvement for the iterative formula of the original method, the iterative efficiency is increased greatly, and the iterative number is reduced greatly, therefore the operation load of the pattern synthesis is reduced efficiently. (III)The improved jammer power iterative formula enlarges the selecting bound for the iterative coefficient, and reduces the effect of the parameter upon the pattern synthesis, and enhances the application area and applicability of the proposed pattern synthesis method.

3. Unitary Root-MUSIC

The problem of estimating the direction-of-arrival (DOA) of narrowband sources from sensor array data has received considerable attention. The eigen-based methods for DOA estimation represent a class techniques that offer a much better resolution performance than that of conventional beamformers. In eigen-based methods, signal and noise subspaces are identified first via a $M \times M$ generalized EVD (GEVD) of the array data/noise correlation matrix pencil, where M equals the number of array elements. A search is then conducted over a null spectrum associated with the noise subspace, to locate the minima, from which the source DOA’s can be determined. In the case where a uniform linear array (ULA) is employed, the null-spectrum searching can be converted into a polynomial rooting problem. Two well known examples are the Root-MUSIC[47] and Root-Minimum-Norm[48] method. They belong to the so-called weighted root-form eigen-based methods. Compared to their spectrum-searching or spectral-form counterparts, root-form methods exhibit a higher resolution capability in dealing with closely spaced sources.Rao and Hari[49] argue that a zero of the null spectrum, having a large radial error, will cause the corresponding spectral minima to be less defined,but does not affect the DOA estimates. As for the mean-squared errors of the DOA estimates,Li and Vaccaro[50] show that both spectral and root-form methods yield the same expression. It should be borne in mind, however, that the result holds only when each of the sources has a minimum corresponding to it in the null spectrum.

A major issue regarding eigen-based methods is the heavy computational load associated with the GEVD. This is more significant when M is large. To remedy this the concept of beamspace transformation was proposed[51] as a means of reducing the dimension of the array data. Ta.S.L proposed a novel iterative implementation of beamspace root-form methods without the need for large-order polynomial rooting[52]. Marius.P, Alex.B.G and Martin.H proposed the unitary root MUSIC algorithm reduces the computational complexity because it exploits the eigendecomposition of a real-valued covariance matrix[53].

In this chapter [54], combining the algorithms of Root-MUSIC and Unitary-root MUSIC, the Root-MUSIC algorithm with real-valued eigendecomposition is given.

3.1 Array signal model

Assume a uniform linear array (ULA) is composed of M sensors, and let it receive q ($q < M$) narrowband signals impinging with unknown directions of arrival ((DOA) $\theta_1, \theta_2, \dots, \theta_q$. Assume that there are N snapshots $\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(N)$ available. The tth measured snapshot of the array is generally modeled as:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{3.1}$$

where $\mathbf{A} = \left[\mathbf{a}(\theta_1), \ \cdots, \ \mathbf{a}(\theta_q)\right]$ is the $M \times q$ composite steering matrix, the columns of which represent a basis for the signal subspace, $\mathbf{a}(\theta)$ represents the array's $M \times 1$ complex manifold:

$$\mathbf{a}(\theta) = \left[1, \ e^{j(2\pi/\lambda)d\sin\theta}, \ \cdots, \ e^{j(2\pi/\lambda)d(M-1)\sin\theta}\right]^T \tag{3.2}$$

In addition

$\mathbf{s}(t)$ denotes the $q \times 1$ vector of source waveforms;
$\mathbf{n}(t)$ denotes the $M \times 1$ vector of white sensor noise;
λ is the wavelength;
d is the interelement space;
$(\cdot)^T$ denotes transpose.

It is generally assumed that signals are uncorrelated with the noise $\mathbf{n}(t)$. The sensor noise is assumed to be a zero-mean spatially and temporally uncorrelated Gaussian process with the uunknown diagonal covariance matrix given by

$$\mathbf{R}_n = E\left\{\mathbf{n}(t)\mathbf{n}(t)^H\right\} = diag\left\{\sigma^2, \ \sigma^2, \ \cdots, \ \sigma^2,\right\} = \sigma^2\mathbf{I} \tag{3.3}$$

where $E\{\cdot\}$ is the expectation operator,and $(\cdot)^H$ stands for the Hermitian transpose, $\mathbf{I}$ is the identity matrix, σ^2 is the noise variance.

Data model (1) allows us to write the covariance matrix of the array measurements as:

$$\mathbf{R} = E\left\{\mathbf{x}(t)\mathbf{x}(t)^H\right\} = \mathbf{ASA}^H + \sigma^2\mathbf{I} \tag{3.4}$$

where $\mathbf{S} = E\left\{\mathbf{s}(t)\mathbf{s}(t)^H\right\}$ is the $q \times q$ source waveform covariance matrix.

3.2 Root-MUSIC

Root-MUSIC is the polynomial rooting form of MUSIC, namely, the spectrum peak searching is replaced by polynomial rooting in MUSIC implementation.

In Root-MUSIC, the polynomial should be defined as follows

$$f_i(z) = \mathbf{u}_i^H\mathbf{a}(z) \qquad i = q+1,\cdots,M \tag{3.5}$$

where, $\mathbf{u}_i$ is the eigenvector corresponding to the *M*-*q* minimum eigen-value of the data covariance matrix, and

$$\mathbf{a}(z) = \begin{bmatrix} 1 & z & \cdots, & z^{M-1} \end{bmatrix}^T \tag{3.6}$$

From the above definition, we can include that the polynomial roots lie on the unit circle properly when $z = \exp(j\omega)$, and $\mathbf{a}(e^{j\omega})$ is the steering vector of space frequency ω. From the eigen-space algorithms, $\mathbf{a}(e^{j\omega_m}) = \mathbf{a}_m$ is the signal steering vector, and it is orthogonal to the space of the noise. Therefore, the polynomial definition can be modified as the following form

$$f(z) = \mathbf{a}^H(z)\mathbf{U}_N\mathbf{U}_N^H\mathbf{a}(z) \tag{3.7}$$

where $\mathbf{U}_N$ is the noise space, namely, let the eigendecomposition of the matrix $\mathbf{R}$ be defined in a standard way

$$\mathbf{R} = \mathbf{U}\ \boldsymbol{\Lambda}\ \mathbf{U}^H = \mathbf{U}_S\boldsymbol{\Lambda}_S\mathbf{U}_S^H + \sigma^2\mathbf{U}_N\mathbf{U}_N^H \tag{3.8}$$

where

$\mathbf{U}_S = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_q\end{bmatrix}$, $\mathbf{U}_N = \begin{bmatrix}\mathbf{u}_{q+1}, & \cdots & \mathbf{u}_M\end{bmatrix}$, $\boldsymbol{\Lambda}_S = \mathrm{diag}\left\{\lambda_1, \ \cdots, \ \lambda_q\right\}$
and the subscripts S and N stand for signal- and noise-subspace, respectively.

Therefore, the source DOA information can be obtained when the above roots are solved. At the same time, we found the item of z^* in the polynomial, and the solving process will become complex and difficult. In order to simplify problem, the above polynomial can be modified as

$$f(z) = z^{M-1}\mathbf{a}^T\left(z^{-1}\right)\mathbf{U}_N\mathbf{U}_N^H\mathbf{a}(z) \tag{3.9}$$

Here, the polynomial order is 2(*M*-1), and has (*M*-1) pairs roots, the every pair roots have the mutual conjugate relationship. In the (M-1) pairs roots, there are q roots $z_1,\cdots,z_q$ are distributed on the unit circle, and

$$z_i = \exp\left(j\omega_i\right), \qquad 1 \le i \le q \tag{3.10}$$

For the ULA, the corresponding DOA of signal can be calculated as

$$\theta_i = \arcsin\left(\frac{\lambda}{2\pi d}\arg\left(z_i\right)\right), \qquad i = 1,\cdots,q \tag{3.11}$$

where λ is the signal wavelength, d is the array space.

A simple alternative method is proposed in Ref.[55]. From above analysis, we can see that the signal space is orthogonal to the noise space, therefore

$$z_i = \exp\left(j2\pi\frac{d}{\lambda}\sin\theta_i\right), \qquad i = 1,\cdots,q \tag{3.12}$$

should be q roots of all M-q polynomials in Eq. (3.9), namely

$$f_i\left(z_j\right) = 0, \qquad j = 1,\cdots,q, \qquad i = q+1,\cdots,M \tag{3.13}$$

Eq.(3.9) represents M-q polynomials of M-1 order. From Eq.(3-13), they should have a q-order maximum common factor, which can be denoted as $f\left(z\right)$. The DOAs of all the sources can be obtained by rooting $f\left(z\right)$. From the eigenvectors of the noise space, $f\left(z\right)$ can be obtained as follows

There exists a vector $\mathbf{b} = \begin{bmatrix} b_1 & \cdots & b_{M-q} \end{bmatrix}^T$ which satisfies

$$\mathbf{U}_N^H\mathbf{b} = \begin{bmatrix} \mathbf{U}_{N1}^H \\ \mathbf{U}_{N2}^H \end{bmatrix}\mathbf{b} = \begin{bmatrix} c_1 & \cdots & c_q & 1 & 0 & \cdots & 0 \end{bmatrix}_{1\times M}^T \tag{3.14}$$

where $\mathbf{U}_{N1}$ is q×(M-q) sub-matrix and $\mathbf{U}_{N2}$ is (M-q)×(M-q) sub-matrix of $\mathbf{U}_N$. This can be understood by noticing that the product of $\mathbf{U}_N$ and $\mathbf{b}$ represents a linear combination of noise vectors represented in the M-q dimensional noise space. The product $\mathbf{U}_{N2}$ and $\mathbf{b}$ defines a system of M-q equations with M-q unknowns. $\mathbf{b}$ can be fixed to be the solution that results in a product $\begin{bmatrix} 1 & 0 & \cdots \end{bmatrix}^T$. The product of $\mathbf{U}_{N1}$ and $\mathbf{b}$ is then a set of coefficients that are determined.

Adopting this approach $\mathbf{b}$ is obtained by

$$\mathbf{b} = \mathbf{U}_{N2}^{-1}\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}_{1\times(M-q)}^T \tag{3.15}$$

And $\mathbf{c} = \begin{bmatrix} c_1 & \cdots & c_q \end{bmatrix}$ is determined easily as

$$\mathbf{c} = \mathbf{U}_{N1}\mathbf{b} = \mathbf{U}_{N1}\mathbf{U}_{N2}^{-1}\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}_{1\times(M-q)}^T \tag{3.16}$$

Now that $\mathbf{c}$ has been determined, the polynomial $f\left(z\right)$ is formed by

$$f\left(z\right) = \sum_{i=1}^{q+1} c_i z^{i-1}, \qquad c_{M+1} = 1 \tag{3.17}$$

Eq.(3.17) has q roots which are correspond to the DOAs of q sources. After obtaining q roots of Eq.(3.13), $\{z_i\}_{i=1}^{q}$, the DOAs of the sources are obtained by Eq.(3.12).

From the well known conventional Root-MUSIC polynomial of Eq.(3.7), we can conclude that it is a function of z, namely

$$f_{\text{Root-MUSIC}}(z) = z^{M-1}\mathbf{a}^T(1/z)\mathbf{U}_N\mathbf{U}_N^H\mathbf{a}(z) = z^{M-1}\mathbf{a}^T(1/z)\left\{1-\mathbf{U}_S\mathbf{U}_S^H\right\}\mathbf{a}(z) \tag{3.18}$$

Here the orthogonal property of the signal and noise subspace is used.

3.3 Root-MUSIC with real-valued eigendecomposition

The matrix $\mathbf{R}$ is the centro-Hermitian if

$$\mathbf{R} = \mathbf{J}\mathbf{R}^*\mathbf{J} \tag{3.19}$$

where $\mathbf{J}$ is the exchange matrix with ones on its antidiagonal and zeros elsewhere, and $(\cdot)^*$ stands for complex conjugate. The matrix (3.3)is known to be centro-Hermitian if and only if $\mathbf{S}$ is a diagonal matrix, i.e., when the signal source are uncorrelated. However, to 'double' the number of snapshots and decorrelate possibly correlated source pairs in the case of an arbitrary matrix $\mathbf{S}$, the centro-Hermitian property is sometimes forced by means of the so-called forward-backward (FB) averaging:

$$\mathbf{R}_{FB} = \frac{1}{2}\left(\mathbf{R} + \mathbf{J}\mathbf{R}^*\mathbf{J}\right) = \mathbf{A}\tilde{\mathbf{S}}\mathbf{A}^H + \sigma^2\mathbf{I} \tag{3.20}$$

where

$$\tilde{\mathbf{S}} = \frac{1}{2}\left(\mathbf{S} + \mathbf{D}\mathbf{S}^*\mathbf{D}\right) \tag{3.21}$$

$$\mathbf{D} = diag\left\{e^{-j(2\pi/\lambda)d(M-1)\sin\theta_1}, \quad \cdots, \quad e^{-j(2\pi/\lambda)d(M-1)\sin\theta_q},\right\} \tag{3.22}$$

Let us define the matrix as:

$$\mathbf{C} = \mathbf{Q}^{\mathbf{H}}\mathbf{R}_{\mathbf{FB}}\mathbf{Q} \tag{3.23}$$

therefore, the $\mathbf{C}$ is a real-valued covariance matrix, where $\mathbf{Q}$ is any unitary,column conjugate symmetric $M \times M$ matrix. Any matrix $\mathbf{Q}$ is column conjugate symmetric if

$$\mathbf{J}\mathbf{Q}^* = \mathbf{Q} \tag{3.24}$$

For example, the following sparse matrices

$$\mathbf{Q}_{2n} = \frac{1}{\sqrt{2}}\begin{bmatrix} \mathbf{I} & j\mathbf{I} \\ \mathbf{J} & -j\mathbf{J} \end{bmatrix} \tag{3.25}$$

$$\mathbf{Q}_{2n+1} = \frac{1}{\sqrt{2}}\begin{bmatrix} \mathbf{I} & \mathbf{0} & j\mathbf{I} \\ \mathbf{0}^T & \sqrt{2} & \mathbf{0}^T \\ \mathbf{J} & \mathbf{0} & -j\mathbf{J} \end{bmatrix} \tag{3.26}$$

can be chosen for arrays with an even and odd number of sensors,respectively, where the vector $\mathbf{0} = (0,\ \ 0,\ \ \cdots,\ \ 0)^T$

From (3.23), and insert (3.20) to it, it follows that

$$\mathbf{C} = \mathbf{Q^H R_{FB} Q} = \mathbf{Q^H}\left[\frac{1}{2}\left(\mathbf{R} + \mathbf{JR^*J}\right)\right]\mathbf{Q} = \frac{1}{2}\left[\mathbf{Q^H RQ} + \mathbf{Q^H}\left(\mathbf{JR^*J}\right)\mathbf{Q}\right] \tag{3.27}$$

using $\mathbf{JQ}^* = \mathbf{Q}$, $\mathbf{Q}^* = \mathbf{JQ}$ and $\mathbf{J}^H = \mathbf{J}$, we obtain that

$$\begin{aligned}\mathbf{C} &= \frac{1}{2}\left[\mathbf{Q^H RQ} + \mathbf{Q^H JR^*JQ}\right] = \frac{1}{2}\left[\mathbf{Q^H RQ} + (\mathbf{JQ})^{\mathbf{H}}\mathbf{R}^*(\mathbf{JQ})\right] \\ &= \frac{1}{2}\left[\mathbf{Q^H RQ} + \left(\mathbf{Q}^*\right)^{\mathbf{H}}\mathbf{R}^*\left(\mathbf{Q}^*\right)\right] = \frac{1}{2}\left[\mathbf{Q^H RQ} + \left(\mathbf{Q^H RQ}\right)^*\right] \\ &= \mathrm{Re}\left[\mathbf{Q^H RQ}\right]\end{aligned} \tag{3.28}$$

therefore, we prove that $\mathbf{C}$ is a real-valued covariance matrix.

Let the eigendecompositions of the matrices $\mathbf{R}$, $\mathbf{R_{FB}}$ and $\mathbf{C}$ be defined in a standard way

$$\mathbf{R} = \mathbf{V}\ \mathbf{\Pi}\ \mathbf{V}^H = \mathbf{V}_S\mathbf{\Pi}_S\mathbf{V}_S^H + \sigma^2\mathbf{V}_N\mathbf{V}_N^H \tag{3.29}$$

$$\mathbf{R}_{FB} = \mathbf{U}\ \mathbf{\Lambda}\ \mathbf{U}^H = \mathbf{U}_S\mathbf{\Lambda}_S\mathbf{U}_S^H + \sigma^2\mathbf{U}_N\mathbf{U}_N^H \tag{3.30}$$

$$C = \mathbf{E}\ \mathbf{\Gamma}\ \mathbf{E}^H = \mathbf{E}_S\mathbf{\Gamma}_S\mathbf{E}_S^H + \sigma^2\mathbf{E}_N\mathbf{E}_N^H \tag{3.31}$$

where

$\mathbf{V}_S = \left[\mathbf{v}_1,\ \cdots,\ \mathbf{v}_q\right]$, $\mathbf{V}_N = \left[\mathbf{v}_{q+1},\ \cdots,\ \mathbf{v}_M\right]$, $\mathbf{\Pi}_S = diag\left\{\pi_1,\ \cdots,\ \pi_q\right\}$
$\mathbf{U}_S = \left[\mathbf{u}_1,\ \cdots,\ \mathbf{u}_q\right]$, $\mathbf{U}_N = \left[\mathbf{u}_{q+1},\ \cdots,\ \mathbf{u}_M\right]$, $\mathbf{\Lambda}_S = diag\left\{\lambda_1,\ \cdots,\ \lambda_q\right\}$
$\mathbf{E}_S = \left[\mathbf{e}_1,\ \cdots,\ \mathbf{e}_q\right]$, $\mathbf{E}_N = \left[\mathbf{e}_{q+1},\ \cdots,\ \mathbf{e}_M\right]$, $\mathbf{\Gamma}_S = diag\left\{\gamma_1,\ \cdots,\ \gamma_q\right\}$
and the subscripts S and N stand for signal- and noise-subspace,respectively.

Assume the Characteristic equation for the matrix $\mathbf{R_{FB}}$ as

$$\mathbf{R_{FB}} \cdot \mathbf{u} = \lambda \cdot \mathbf{u} \tag{3.32}$$

we can obtain that

$$\mathbf{Q}^H \cdot \mathbf{R_{FB}u} = \mathbf{Q}^H \cdot \lambda\mathbf{u} = \lambda \cdot \mathbf{Q}^H\mathbf{u} \tag{3.33}$$

with the use of equation: $\mathbf{QQ^H} = \mathbf{I}$ and the definition of $\mathbf{C}$,we obtain that

$$\mathbf{Q}^H \cdot \mathbf{R_{FB}u} = \mathbf{Q}^H \cdot \mathbf{R_{FB}} \cdot \mathbf{QQ^H} \cdot \mathbf{u} = \mathbf{C} \cdot \mathbf{Q^H u} = \lambda \cdot \mathbf{Q}^H\mathbf{u} \tag{3.34}$$

Equation $\mathbf{C} \cdot \mathbf{Q^H u} = \lambda \cdot \mathbf{Q}^H\mathbf{u}$ can be identified as the characteristic one for the real-valued covariance matrix $\mathbf{C}$.

Hence, using (3.30), (3.31), (3.32) and (3.34), the eigenvectors and eigenvalues of the matrices $\mathbf{R_{FB}}$ and $\mathbf{C}$ are related as

$$\mathbf{E} = \mathbf{Q}^H \mathbf{U} \tag{3.35}$$

$$\mathbf{\Gamma} = \mathbf{\Lambda} \tag{3.36}$$

It is well known that the conventional Root-MUSIC polynomial is given by

$$f_{MUSIC}(z) = \mathbf{a}^T(1/z)\mathbf{V}_N\mathbf{V}_N^H\mathbf{a}(z) \tag{3.37}$$

$$= \mathbf{a}^T(1/z)\left\{1 - \mathbf{V}_S\mathbf{V}_S^H\right\}\mathbf{a}(z) \tag{3.38}$$

where

$$\mathbf{a}(z) = \begin{bmatrix}1, & z, & \cdots, & z^{M-1}\end{bmatrix}^T \tag{3.39}$$

$z = e^{j\omega}$, and $\omega = (2\pi/\lambda)d\sin\theta$. Similarly to(3.37) and (3.38), the FB root-MUSIC polynomial can be used:

$$f_{FB-MUSIC}(z) = z^{M-1}\mathbf{a}^T(1/z)\mathbf{U}_N\mathbf{U}_N^H\mathbf{a}(z) \tag{3.40}$$

$$= z^{M-1}\mathbf{a}^T(1/z)\left\{1 - \mathbf{U}_S\mathbf{U}_S^H\right\}\mathbf{a}(z) \tag{3.41}$$

A simple manipulation with the use of (3.35) and $\mathbf{QQ^H} = \mathbf{I}$, we can obtain that:

$$f_{FB-MUSIC}(z) = z^{M-1}\mathbf{a}^T(1/z)\cdot\mathbf{QQ^H}\cdot\mathbf{U}_N\mathbf{U}_N^H\cdot\mathbf{QQ^H}\cdot\mathbf{a}(z) \tag{3.42}$$

$$= z^{M-1}\mathbf{a}^T(1/z)\cdot\mathbf{Q}\cdot\left(\mathbf{Q^H}\mathbf{U}_N\right)\cdot\left(\mathbf{Q^H}\mathbf{U}_N\right)^{\mathbf{H}}\mathbf{Q^H}\cdot\mathbf{a}(z) \tag{3.43}$$

$$= z^{M-1}\mathbf{a}^T(1/z)\cdot\mathbf{Q}\cdot\mathbf{E}_N\cdot\mathbf{E}_N^H\cdot\mathbf{Q^H}\cdot\mathbf{a}(z) \tag{3.44}$$

$$= z^{M-1}\tilde{\mathbf{a}}^T(1/z)\cdot\mathbf{E}_N\cdot\mathbf{E}_N^H\cdot\tilde{\mathbf{a}}(z) \tag{3.45}$$

$$= f_{C-MUSIC}(z) \tag{3.46}$$

where the manifold

$$\tilde{\mathbf{a}}(z) = \mathbf{Q^H}\cdot\mathbf{a}(z) \tag{3.47}$$

should be exploited for the polynomial rooting in (3.45). The relationship between the former and the new manifolds follows from the expression for the real-valued covariance matrix (3.23). From (3.23)and (3.20), we have

$$\mathbf{C}=\mathbf{Q^H R_{FB} Q}=\mathbf{Q^H}\left(\mathbf{A}\tilde{\mathbf{S}}\mathbf{A}^H+\sigma^2\mathbf{I}\right)\mathbf{Q}=\mathbf{Q^H A}\tilde{\mathbf{S}}\mathbf{A}^H\mathbf{Q}+\sigma^2\mathbf{Q^H Q} \tag{3.48}$$

$$=\tilde{\mathbf{A}}\tilde{\mathbf{S}}\tilde{\mathbf{A}}^H+\sigma^2\mathbf{Q^H Q} \tag{3.49}$$

where

$$\tilde{\mathbf{A}}=\mathbf{A}^H\mathbf{Q} \tag{3.50}$$

Let us term the polynomial (3.46) as the polynomial of Root-MUSIC with real-valued eigendecomposition (RVED-Root-MUSIC), since it exploits the eigendecomposition of the real-valued matrix (3.24) instead of that of the complex matrices (3.18) or (3.20). But from (3.42) to (3.44), it is clear that the FB and RVED-Root-MUSIC polynomials are identical. Hence, the performance of RVED-ROOT-MUSIC does not depend on a particular choice of the unitary column conjugate symmetric matrix $\mathbf{Q}$.

3.4 Polynomial coefficient finding

From (3.44) and (3.45), we obtain the polynomial of RVED-Root-MUSIC, which is a function of z. The next thing is finding the coefficient of the polynomial[56]..

Using (3.44), we have:

$$\begin{aligned} f_{C-MUSIC}(z)&=z^{M-1}\mathbf{a}^T(1/z)\cdot\mathbf{Q}\cdot\mathbf{E}_N\cdot\mathbf{E}_N^H\cdot\mathbf{Q^H}\cdot\mathbf{a}(z)\\ &=z^{M-1}\mathbf{a}^T(1/z)\cdot\mathbf{G}\cdot\mathbf{a}(z)\end{aligned} \tag{3.51}$$

where

$$\mathbf{G}=\mathbf{Q}\cdot\mathbf{E}_N\cdot\mathbf{E}_N^H\cdot\mathbf{Q^H}=\left(g_{i,j}\right)_{M\times M} \tag{3.52}$$

Inserting (3.39) into (3.52), and with simple manipulation, we obtain that

$$\begin{aligned} f_{C-MUSIC}(z)&=z^{M-1}\left[1,\ \ z^{-1},\ \ \cdots,\ \ z^{-(M-1)}\right]\cdot\mathbf{G}\cdot\left[1,\ \ z^1,\ \ \cdots,\ \ z^{(M-1)}\right]^T\\ &=\left[z^{M-1},\ \ z^{M-2},\ \ \cdots,\ \ 1\right]\cdot\begin{bmatrix} g_{1,1} & \cdots & g_{1,M}\\ \vdots & \cdots & \vdots\\ g_{M,1} & \cdots & g_{M,M}\end{bmatrix}\cdot\left[1,\ \ z^1,\ \ \cdots,\ \ z^{(M-1)}\right]^T\\ &=\left[\sum_{i=1}^M g_{i,1}z^{M-i},\ \ \sum_{i=1}^M g_{i,2}z^{M-i},\ \ \cdots,\ \ \sum_{i=1}^M g_{i,M}z^{M-i}\right]\cdot\begin{bmatrix}1\\ z^1\\ \vdots\\ z^{(M-1)}\end{bmatrix}\\ &=1\cdot\left(\sum_{i=1}^M g_{,i1}z^{M-i}\right)+z^1\cdot\left(\sum_{i=1}^M g_{i,2}z^{M-i}\right)+\cdots+z^{(M-1)}\cdot\left(\sum_{i=1}^M g_{i,M}z^{M-i}\right)\end{aligned}$$

the polynomial of RVED-Root-MUSIC is given by

$$\begin{aligned}
f_{C-MUSIC}(z) = & \left(g_{1,M}\right)\cdot z^{2M-2-0} \\
& +\left(g_{2,M}+g_{2,M-1}\right)\cdot z^{2M-2-1} \\
& +\left(g_{3,M}+g_{2,M-1}+g_{1,M-2}\right)\cdot z^{2M-2-2} \\
& +\cdots \\
& +\left(g_{M,M}+g_{M-1,M-1}+\cdots+g_{2,2}+g_{1,1}\right)\cdot z^{2M-2-(M-1)} \\
& +\left(g_{M,M-1}+g_{M-1,M-2}+\cdots+g_{3,2}+g_{2,1}\right)\cdot z^{2M-2-M} \\
& +\left(g_{M,M-1}+g_{M-1,M-2}+\cdots+g_{3,2}+g_{2,1}\right)\cdot z^{2M-2-(M+1)} \\
& +\cdots \\
& +\left(g_{M,2}+g_{M-1,1}\right)\cdot z^{2M-2-(M+M-3)} \\
& +\left(g_{M,1}\right)\cdot z^{2M-2-(M+M-2)}
\end{aligned}$$

So the number of coefficient of the polynomial is $2M-1$, and the computation of the coefficient is given as follows

$$a_k = \begin{cases} \sum_{i=1}^{k} g_{i,M-k+i} & k=1,\ 2,\ \cdots,\ L \\ \sum_{i=1}^{(2M-1)-k+1} g_{k-M+i,i} & k=L+1,\ L+2,\ \cdots,\ 2L-1 \end{cases} \tag{3.53}$$

where a_k denotes the kth coefficient of the polynomial.

Based upon our analysis, using (3.4), (3.20), (3.23), (3.35), (3.26), (3.31), (3.35), (3.44), (3.51), (3.52) and (3.53), the fast algorithm for RVED-Root-MUSIC can be formulated as the following seven-step procedure:

Step 1. Compute $\mathbf{R}$ and $\mathbf{R_{FB}}$ with the use of (3.4) and (3.6).and the estimate is given by $\hat{\mathbf{R}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}(k)\mathbf{x}^H(k)$ then $\hat{\mathbf{R}}_{FB} = \frac{1}{2}\left(\hat{\mathbf{R}}+\mathbf{J}\hat{\mathbf{R}}^*\mathbf{J}\right)$.

Step 2. Compute $\mathbf{C}$, and the $\mathbf{Q}$ is dependent on the number of array sensors. The estimate of the real-valued covariance matrix is given by $\hat{\mathbf{C}} = \mathbf{Q^H}\hat{\mathbf{R}}_{\mathbf{FB}}\mathbf{Q}$

Step 3. Obtain $\mathbf{E}_N$ from the eigendecomposition of $\mathbf{C}$ .and the estimate of $\mathbf{E}_N$, $\hat{\mathbf{E}}$ is given by the eigendecomposition of $\hat{\mathbf{C}}$

Step 4. Compute $\mathbf{G}$ with the use of (3.52). And the estimate of $\mathbf{G}$, $\hat{\mathbf{G}}$ is given by $\mathbf{G} = \mathbf{Q}\cdot\hat{\mathbf{E}}_N\cdot\hat{\mathbf{E}}_N^H\cdot\mathbf{Q^H}$

Step 5. Compute the coefficient of the polynomial by (3.53).

Step 6. Find the root of the polynomial (3.51), and select the q roots that are nearest to the unit circle as being the roots corresponding to the DOA estimates.

Step 7. DOA estimate, using:

$$\theta_k = \arcsin\left(\frac{\lambda}{2\pi d}\arg(z_k)\right) \qquad k = 1, \quad \cdots, \quad q$$

where z_k represents one of the q roots selected for DOA estimation.

From the above analysis, we can conclude that the RVED-Root-MUSIC has a lower computational complexity than the conventional root-MUSIC technique thanks to the eigendecomposition of the real-valued matrix instead of that of the complex matrices, and the asymptotic performance of it is better than of conventional root-MUSIC due to the FB averaging effect..

3.5 Simulations

In this section, we present some simulation results to illustrate the performance of RVED-Root-MUSIC. We consider a ULA with M=8 elements and the inter-element space is equal to a half of wavelength. There are three signals with SNRs of 30 dB impinges on the array from $\theta_1 = -80$, $\theta_2 = -20$, $\theta_3 = 40$. The detailed simulation results are shown as Fig. 22. ~ Fig. 25.

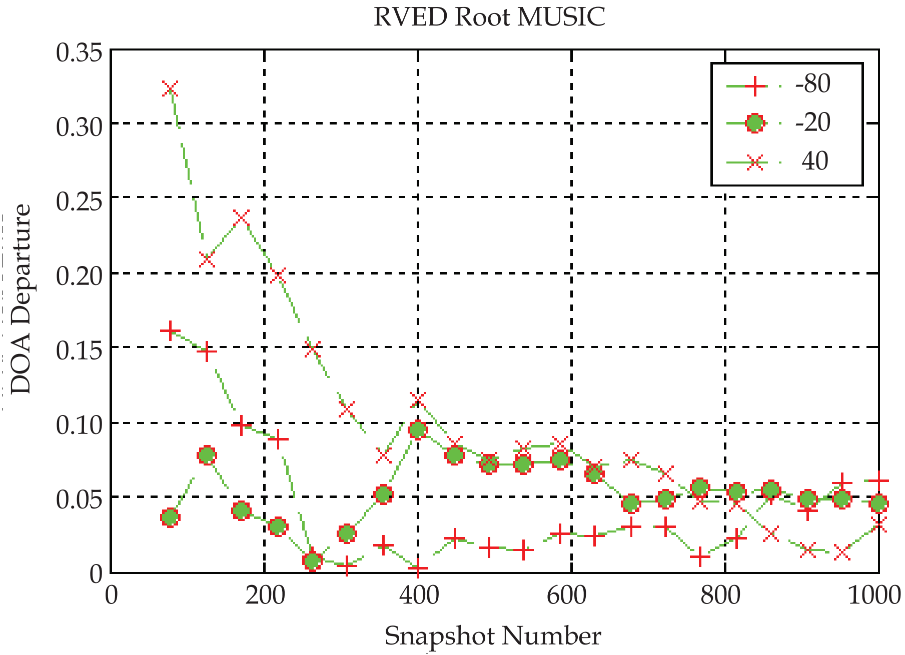


Fig. 22. DOA departure vs dnapshot number. Signal DOA=[-80 -20 40], SNR=5dB

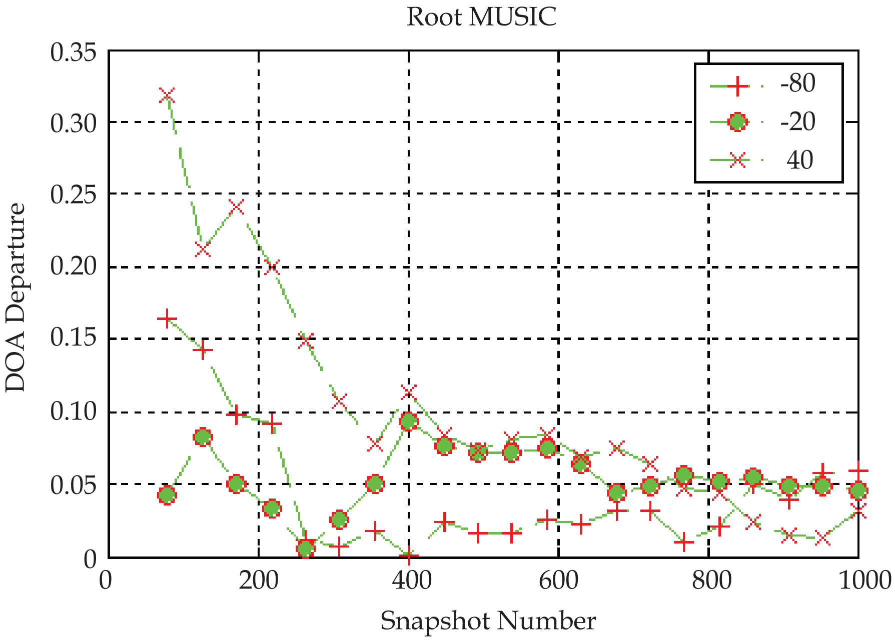


Fig. 23. DOA departure vs snapshot number. Signal DOA=[-80 -20 40], SNR=5dB

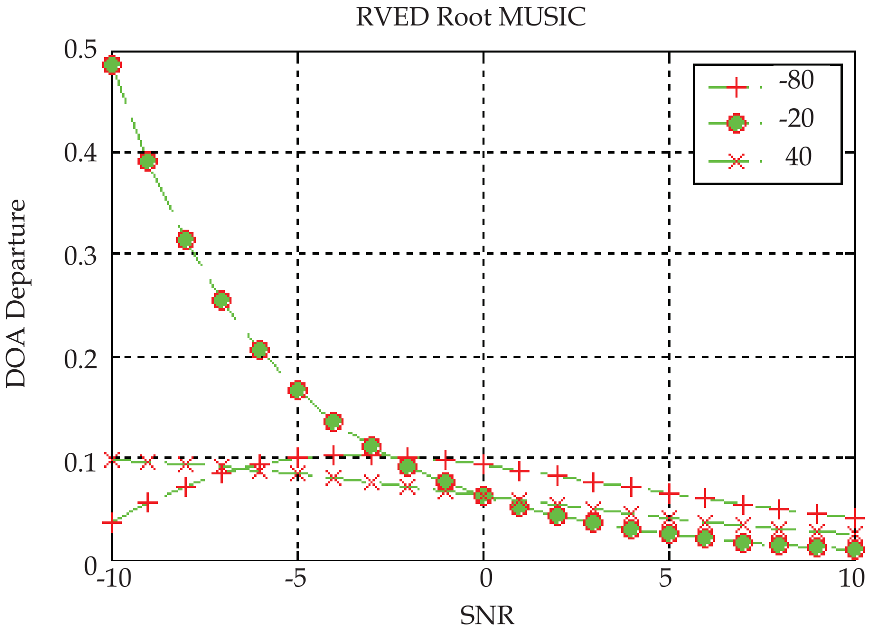


Fig. 24. DOA departure vs SNR. Signal DOA=[-80 -20 40], Snapshot number =1000

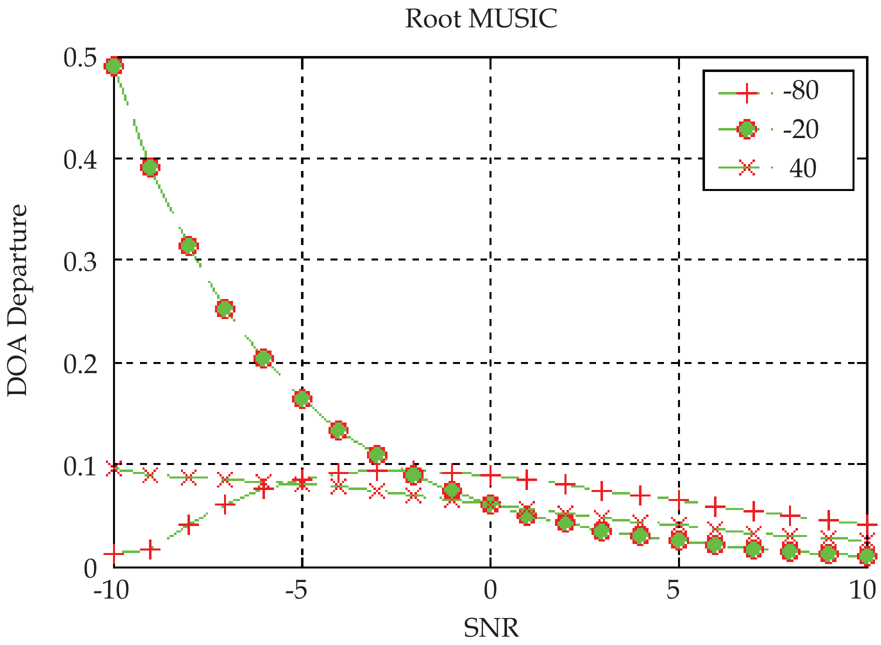


Fig. 25. DOA departure vs SNR. Signal DOA=[-80 -20 40], Snapshot number =1000

Fig. 22. and Fig. 23. depict DOA departure versus snapshot number results of RVED-Root-MUSIC and Root-MUSIC respectively, where the SNR=5dB. In figure 22. and 23., the x-axis denotes the snapshot number, and y-axis denotes the departure of signal DOA.

Fig. 24. and Fig. 25. depict DOA departure versus SNR results of RVED-Root-MUSIC and Root-MUSIC respectively, where the snapshot number =1000. In figure 24. and 25., the x-axis denotes the SNR, and y-axis denotes the departure of signal DOA .

From the detecting results and the comparison between RVED-Root-MUSIC and Root-MUSIC, we can conclude that RVED-Root-MUSIC can detect DOA of signal quickly and effectively. At the same time, the results validate the correctness and effective of this algorithm.

3.6 Conclusion

An improved version of the Root-MUSIC algorithm, called Root-MUSIC with real-valued eigendecomposition (RVED-Root-MUSIC), has been presented in this chapter. The computational complexity is reduced significantly by exploiting the one-to-one correspondence between centro-Hermitian and real matrices, allowing a transformation to real matrices, which can be maintained for all steps of the algorithm. Due to the inherent forward-backward averaging effect, RVED-Root-MUSIC can separate two completely coherent sources and provide improved estimates for correlated signals.

4. Real-value space ESPRIT algorithm and its implement

The recovery of signal parameters from noisy observations is a fundamental problem in (real-time) array signal processing. Due to their simplicity and high-resolution capability,the subspace estimation schemes have been attracting considerable attention. Among them the most representative are MUSIC and ESPRIT methods. MUSIC utilizes the orthogonal characteristic of noisy subspace of data covariance matrix,but ESPRIT exploits the rotational invariance structure of the signal subspace[57,58]. The virtue of ESPRIT is the low computational burden,and not requiring spectrum peak searching by contrast with MUSIC. Comparing with Root-MUSIC, ESPRIT obtains the information of signal direction of arriving (DOA) via exploiting the rotational invariance of every subarray (every subarray 's signal subspace), but Root-MUSIC estimates the signal DOA by solving the polynomial, which is constructed by using the orthogonal between the steering vector and noise subspace.

Unitary ESPRIT achieves even more accurate results than previous ESPRIT techniques by taking advantage of the unit magnitude property of the phase factors that represent the phase delays between the two subarrays [59]. It has been shown in [63] that constraining the phase factors to the unit circle can also give some improvement for correlated sources. For centro-symmetric sensor arrays with a translational invariance structure, Unitary ESPRIT provides a very simple and efficient solution to this task.

Although Unitary ESPRIT effectively doubles the number of data samples, the computational complexity is reduced by transforming the required rank-revealing factorizations of complex matrices into decompositions of real-valued matrices of the same size. Thus, we obtain increased estimation accuracy with a reduced computational load. This reduction can be achieved by constructing invertible transformations that map centro-Hermitian matrices to real matrices.

The real-value ESPRIT algorithm is proposed by [62] and [63], which is on the foundation of the Unitary ESPRIT, by constructing a transformation matrix, transforms the complex data of original array into real-value data. Thus lowered the computational burden. Moreover this algorithm is also applicable to centro-symmetric sensor arrays.

This chapter bases on the foundation of the algorithm that above references proposes and reference [64], analyzes the rotational invariance principle of RVS-ESPRIT algorithm, and the relationship of RVS-ESPRIT and complex space ESPRIT(CS-ESPRIT), definitely give:

1. The rotational invariance relationship of the real-value space array steering,
2. The rotational invariance relationship of the real-value space signal subspace,
3. The rotational invariance relationship between the array steering and the signal subspace of the real-value space,
4. The rotational invariance relationship between the real-value space signal subspace and the complex value space signal space,
5. The rotational invariance relationship between the real-value space array steering and the complex value space array steering.

And give the implementing algorithm of REV-ESPRIT. At last compares its performance with other algorithm by simulation.

This chapter is organized as follows[65]. It starts with a review of the signal model and the rotational invariance subspace principle. Next the RVS-ESPRIT algorithm is analyzed, among which includes the transformation from the complex space to real-value space, the rotational invariance principle of real-value space, and its implementing algorithm. Finally, the computer simulations with the comparison the performance of RVS-ESPRIT and the well-known LS- ESPRIT algorithm are given.

4.1 Signal model

Assume that there are two completely same subarray, their space $\varDelta$ is already known, and every subarray consists of m elements. Consider N $(N < m)$ narrowband plane waves from far-field of the array, these plane waves are assumed to be impinging on the array from directions $\theta_1,\ \theta_2,\ \cdots,\ \theta_N$, among them, θ_i, $i = 1,\ 2,\ \cdots,\ N$ is angle between the array normal and the direction of the ith signal of the N narrowband planes waves imping. Because the structure of two arrays is completely same, therefore, for a signal, the difference of the two subarray outputs is only one phase difference φ_i, $i = 1,\ 2,\ \cdots,\ N$. Suppose the first subarray receives the data for $\mathbf{X}_1$, the second receives the data for $\mathbf{X}_2$, then:

$$\mathbf{X}_1 = \left[\mathbf{a}(\theta_1) \quad \cdots \quad \mathbf{a}(\theta_N)\right]\mathbf{S} + \mathbf{N}_1 = \mathbf{A}\cdot\mathbf{S} + \mathbf{N}_1 \tag{4.1}$$

$$\mathbf{X}_2 = \left[\mathbf{a}(\theta_1)e^{j\varphi_1} \quad \cdots \quad \mathbf{a}(\theta_N)e^{j\varphi_N}\right]\mathbf{S} + \mathbf{N}_2 = \mathbf{A\Phi}\cdot\mathbf{S} + \mathbf{N}_2 \tag{4.2}$$

Where, the direction matrix of subarray 1 is $\mathbf{A}_1 = \mathbf{A} = \left[\mathbf{a}(\theta_1) \quad \cdots \quad \mathbf{a}(\theta_N)\right]$, the direction matrix of subarray 2 is $\mathbf{A}_2 = \mathbf{A\Phi}$, $\mathbf{S}$ is the space signal vector, $\mathbf{N}_1$ and $\mathbf{N}_2$ are the noise vectors of the subarray 1 and 2, respectively,and are assumed to be white Gaussian, and among the formula:

$$\mathbf{\Phi} = diag\left[e^{j\varphi_1} \quad \cdots \quad e^{j\varphi_N}\right] \tag{4.3}$$

4.2 The rotational invariance subspace principle

From the above mathematics model, we can know that the signal direction information is included in $\mathbf{A}$ and $\mathbf{\Phi}$, because $\mathbf{\Phi}$ is a diagonal matrix, so that we can obtain the DOA of signal through solving $\mathbf{\Phi}$, that is:

$$\varphi_k = \frac{2\cdot\pi|\varDelta|\sin\theta_k}{\lambda} \tag{4.4}$$

where λ is the center wave-length of Arriving the wave. So if we obtain the rotational invariance relationship $\mathbf{\Phi}$ of the two subarray, we can get the signal DOA information. First uniting the two subarray models, namely:

$$\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix} = \begin{bmatrix}\mathbf{A} \\ \mathbf{A}\cdot\mathbf{\Phi}\end{bmatrix}\mathbf{S} + \begin{bmatrix}\mathbf{N}_1 \\ \mathbf{N}_2\end{bmatrix} = \overline{\mathbf{A}}\cdot\mathbf{S} + \mathbf{N} \tag{4.5}$$

Under the ideal condition, the covariance matrix is estimated as fellows:

$$\mathbf{R} = E\left\{\mathbf{X}\cdot\mathbf{X}^H\right\} = \overline{\mathbf{A}}\cdot\mathbf{R}_S\cdot\overline{\mathbf{A}}^H + \mathbf{R}_N \tag{4.6}$$

where $\mathbf{R}_S = E\left\{\mathbf{S}\cdot\mathbf{S}^H\right\}$, $\mathbf{R}_N = E\left\{\mathbf{N}\cdot\mathbf{N}^H\right\}$.

Let the eigendecompositions of the covariance matrix, there is:

$$\mathbf{R} = \sum_{i=1}^{2m}\lambda_i e_i e_i^{\ H} = \mathbf{U}_S\cdot\mathbf{\Sigma}_S\cdot\mathbf{U}_S^{\ H} + \mathbf{U}_N\cdot\mathbf{\Sigma}_N\cdot\mathbf{U}_N^{\ H} \tag{4.7}$$

Very obviously, the eigenvalue that gets from the top have the relationship as follows: $\lambda_1 \geq \cdots \geq \lambda_N > \lambda_{N+1} = \cdots = \lambda_{2m}$. where $\mathbf{U}_S$ is signal subspace that spanned by eigenvectors which are corresponding to large eigenvalues, $\mathbf{U}_N$ is noise subspace that spanned by eigenvectors which are corresponding to small eigenvalues.

We know that the signal subspace is spanned by large eigenvector is equal to that is spanned by array direction matrix in the above eigendecomposition, that is:

$$span\left\{\mathbf{U}_S\right\} = span\left\{\overline{\mathbf{A}}(\theta)\right\} \tag{4.8}$$

At this time, existing a nonsingular matrix $\mathbf{T}$, which can make:

$$\mathbf{U}_S = \overline{\mathbf{A}}(\theta)\cdot\mathbf{T} \tag{4.9}$$

Obviously above-mentioned structure is coming into existence to the two subarrays, so have:

$$\mathbf{U}_S = \begin{bmatrix}\mathbf{U}_{S1}\\ \mathbf{U}_{S2}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\cdot\mathbf{T}\\ \mathbf{A\Phi}\cdot\mathbf{T}\end{bmatrix} \tag{4.10}$$

Very obvious, the subspace spanned by array direction matrix $\mathbf{A}$ is equal to $\mathbf{U}_{S1}$ and $\mathbf{U}_{S2}$ which are spanned by the large eigenvectors of subarray 1 and 2 respectively.

$$span\left\{\mathbf{U}_{S1}\right\} = span\left\{\mathbf{A}(\theta)\right\} = span\left\{\mathbf{U}_{S2}\right\} \tag{4.11}$$

Moreover, from the relationship of the two subarrays with regard to signal direction matrix, we can know:

$$\mathbf{A}_2 = \mathbf{A}_1\mathbf{\Phi} \tag{4.12}$$

Again from (4.10),we can know:

$$\begin{cases}\mathbf{U}_{S1} = \mathbf{A}\cdot\mathbf{T}\\ \mathbf{U}_{S2} = \mathbf{A\Phi}\cdot\mathbf{T}\end{cases} \Rightarrow \begin{cases}\mathbf{A} = \mathbf{U}_{S1}\cdot\mathbf{T}^{-1}\\ \mathbf{U}_{S2} = \mathbf{A\Phi}\cdot\mathbf{T} = \mathbf{U}_{S1}\cdot\mathbf{T}^{-1}\cdot\mathbf{\Phi}\cdot\mathbf{T}\end{cases} \tag{4.13}$$

$$\mathbf{U}_{S2} = \mathbf{U}_{S1}\cdot\mathbf{T}^{-1}\cdot\mathbf{\Phi}\cdot\mathbf{T} = \mathbf{U}_{S1}\cdot\mathbf{\Psi} \tag{4.14}$$

where $\mathbf{\Psi} = \mathbf{T}^{-1}\cdot\mathbf{\Phi}\cdot\mathbf{T}$. (4.12) reflects the rotational invariance characteristic of the signal direction matrix of the two subarrays, but (4.14) reflects the rotational invariance characteristic of the received signal data subspace of the two subarrays.

If the signal direction matrix $\mathbf{A}$ is full rank, we can obtain form (4.14) as fellows:

$$\mathbf{\Phi} = \mathbf{T} \cdot \mathbf{\Psi} \cdot \mathbf{T}^{-1} \tag{4.15}$$

So that,the diagonal matrix which is consisted of the eigenvalues of $\mathbf{\Psi}$ certainly be equal to $\mathbf{\Phi}$, but the every column of $\mathbf{T}$ is the eigenvectors of $\mathbf{\Psi}$. Therefore, once we get the rotational invariance matrix $\mathbf{\Psi}$, we can obtain the signal DOA from (4.4) directly.

4.3 Real-value space ESPRIT algorithm

4.3.1 The transformation from complex space into realvalue space

We know that the uniform linear array is centro-symmetric, and its signal direction matrix satisfy the nether formula:

$$\mathbf{J}_M \cdot \mathbf{A}^* = \mathbf{A} \cdot \mathbf{\Delta} \tag{4.16}$$

where, $\mathbf{J}_M$ is the $M \times M$ exchange matrix with ones on its antidiagonal and zeros elsewhere, and the signal direction matrix makes reference to the first element of the array, the diagonal matrix $\mathbf{\Delta} = \mathbf{\Phi}^{-(M-1)}$, and the $\mathbf{\Phi}$ is expressed as (4.3). If the reference point is selected as the central point of the array, so we have:

$$\mathbf{A}_C = \mathbf{A} \cdot \mathbf{\Delta}^{1/2} = \left[\mathbf{a}_C(\beta_1) \quad \cdots \quad \mathbf{a}_C(\beta_N) \right] \tag{4.17}$$

where

$$\mathbf{a}_C(\beta_i) = e^{-j\left(\frac{M-1}{2}\right)\beta_i} \left[1 \quad e^{-j\beta_i} \quad \cdots \quad e^{-j(M-1)\beta_i} \right]^T = e^{-j\left(\frac{M-1}{2}\right)\beta_i} \mathbf{a}(\beta_i) \tag{4.18}$$

If matrix $\mathbf{Q}$ satisfying:

$$\mathbf{J}_M \cdot \mathbf{Q}^* = \mathbf{Q} \tag{4.19}$$

we call it as the left real transformation matrix.

For example, $\mathbf{Q}$ can be chosen for arrays with an even and odd number of sensors respectively as the following sparse matrices:

$$\mathbf{Q}_{2n} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_n & j\mathbf{I}_n \\ \mathbf{J}_n & -j\mathbf{J}_n \end{bmatrix} \tag{4.20}$$

$$\mathbf{Q}_{2n+1} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_n & \mathbf{0} & j\mathbf{I}_n \\ \mathbf{0}^T & \sqrt{2} & \mathbf{0}^T \\ \mathbf{J}_n & \mathbf{0} & -j\mathbf{J}_n \end{bmatrix} \tag{4.21}$$

Moreover, from the bidirectional averaging algorithm, we can process the array data by once bidirectional averaging, and insert (4.16) into it, we can obtain:

$$\mathbf{R}_{FB} = \frac{1}{2}\left(\mathbf{R} + \mathbf{J}_M \cdot \mathbf{R}^* \cdot \mathbf{J}_M \right) \tag{4.22}$$

Insert $\mathbf{R} = \mathbf{A} \cdot \mathbf{R}_S \cdot \mathbf{A}^H + \mathbf{R}_N$ into(4.13), we can obtain:

$$\begin{aligned}\mathbf{R}_{FB} &= \frac{1}{2}\left(\mathbf{A} \cdot \mathbf{R}_S \cdot \mathbf{A}^H + \mathbf{R}_N + \mathbf{J}_M \cdot \left(\mathbf{A} \cdot \mathbf{R}_S \cdot \mathbf{A}^H + \mathbf{R}_N \right)^* \cdot \mathbf{J}_M \right) \\ &= \frac{1}{2}\left(\mathbf{A} \cdot \mathbf{R}_S \cdot \mathbf{A}^H + \mathbf{R}_N + \mathbf{J}_M \cdot \mathbf{A}^* \cdot \mathbf{R}_S^* \cdot \mathbf{A}^T \cdot \mathbf{J}_M + \mathbf{J}_M \cdot \mathbf{R}_N^* \cdot \mathbf{J}_M \right)\end{aligned} \tag{4.23}$$

because of $\mathbf{J}_M \cdot \mathbf{A}^* = \mathbf{A} \cdot \mathbf{\Delta} \Rightarrow \left(\mathbf{J}_M \cdot \mathbf{A}^* \right)^H = \left(\mathbf{A} \cdot \mathbf{\Delta} \right)^H \Rightarrow \mathbf{A}^T \cdot \mathbf{J}_M = \mathbf{\Delta}^H \cdot \mathbf{A}^H$, and insert it into (4.23), get the result:

$$\begin{aligned}\mathbf{R}_{FB} &= \mathbf{A} \cdot \frac{1}{2}\left(\mathbf{R}_S + \mathbf{\Delta} \cdot \mathbf{R}_S^* \cdot \mathbf{\Delta}^H \right) \cdot \mathbf{A}^H + \frac{1}{2}\left(\mathbf{R}_N + \mathbf{J}_M \cdot \mathbf{R}_N^* \cdot \mathbf{J}_M \right) \\ &= \mathbf{A} \cdot \frac{1}{2}\left(\mathbf{R}_S + \mathbf{\Delta} \cdot \mathbf{R}_S^* \cdot \mathbf{\Delta}^H \right) \cdot \mathbf{A}^H + \mathbf{R}_N^{'} && (4.24) \\ &= \frac{1}{2L} \mathbf{Z} \cdot \mathbf{Z}^H && (4.25)\end{aligned}$$

where

$$\mathbf{Z} = \begin{bmatrix} \mathbf{X} & \mathbf{J}_M \cdot \mathbf{X}^* \cdot \mathbf{J}_L \end{bmatrix} \tag{4.26}$$

Since:

$$\begin{aligned}\frac{1}{2L}\mathbf{Z} \cdot \mathbf{Z}^H &= \frac{1}{2L}\begin{bmatrix} \mathbf{X} & \mathbf{J}_M \cdot \mathbf{X}^* \cdot \mathbf{J}_L \end{bmatrix} \cdot \begin{bmatrix} \mathbf{X} & \mathbf{J}_M \cdot \mathbf{X}^* \cdot \mathbf{J}_L \end{bmatrix}^H \\ &= \frac{1}{2L}\left(\mathbf{X}\mathbf{X}^H + \mathbf{J}_M \cdot \mathbf{X}^* \cdot \mathbf{J}_L \cdot \mathbf{J}_L^H \cdot \mathbf{X}^T \cdot \mathbf{J}_M^H \right) \\ &= \frac{1}{2L}\left(\mathbf{X}\mathbf{X}^H + \mathbf{J}_M \cdot \mathbf{X}^* \cdot \mathbf{X}^T \cdot \mathbf{J}_M^H \right) \\ &= \frac{1}{2L}\left(\mathbf{X}\mathbf{X}^H + \mathbf{J}_M \cdot \left(\mathbf{X}\mathbf{X}^H \right)^* \cdot \mathbf{J}_M^H \right) \\ &= \frac{1}{2}\left[\frac{1}{L}\left(\mathbf{X} \cdot \mathbf{X}^H \right) + \mathbf{J}_M \cdot \left(\frac{1}{L}\left(\mathbf{X} \cdot \mathbf{X}^H \right) \right)^* \cdot \mathbf{J}_M^H \right]\end{aligned} \tag{4.27}$$

Because $\hat{\mathbf{R}} = \frac{1}{L}\left(\mathbf{X} \cdot \mathbf{X}^H \right)$ is the estimating formula of $\mathbf{R}$. Thus (4.25) is established. When the row number of data vector $\mathbf{X}$ is odd, we can definite:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{x}^T \\ \mathbf{X}_2 \end{bmatrix}_{M \times L} \tag{4.28}$$

If we process $\mathbf{Z}$ which is defined by (4.26) By means of matrix $\mathbf{Q}$ which is defined by (4.20) or (4.21) as fellows:

$$\mathbf{T}(\mathbf{X}) = \mathbf{Q}_M^H \cdot \mathbf{Z} \cdot \mathbf{Q}_{2L}$$

$$\mathbf{T}(\mathbf{X}) = \mathbf{Q}_M^H \cdot \mathbf{Z} \cdot \mathbf{Q}_{2L} = \begin{bmatrix} \mathrm{Re}\{\mathbf{X}_1 + \mathbf{J}\mathbf{X}_2^*\} & -\mathrm{Im}\{\mathbf{X}_1 - \mathbf{J}\mathbf{X}_2^*\} \\ \sqrt{2}\,\mathrm{Re}\{\mathbf{x}^T\} & -\sqrt{2}\,\mathrm{Im}\{\mathbf{x}^T\} \\ \mathrm{Im}\{\mathbf{X}_1 + \mathbf{J}\mathbf{X}_2^*\} & \mathrm{Re}\{\mathbf{X}_1 - \mathbf{J}\mathbf{X}_2^*\} \end{bmatrix} \tag{4.29}$$

If the row dimension of the data vector is even, the transformation matrix is:

$$\mathbf{T}(\mathbf{X}) = \mathbf{Q}_M^H \cdot \mathbf{Z} \cdot \mathbf{Q}_{2L} = \begin{bmatrix} \mathrm{Re}\{\mathbf{X}_1 + \mathbf{J}\mathbf{X}_2^*\} & -\mathrm{Im}\{\mathbf{X}_1 - \mathbf{J}\mathbf{X}_2^*\} \\ \mathrm{Im}\{\mathbf{X}_1 + \mathbf{J}\mathbf{X}_2^*\} & \mathrm{Re}\{\mathbf{X}_1 - \mathbf{J}\mathbf{X}_2^*\} \end{bmatrix} \tag{4.30}$$

What to need to be noticed here is, the matrix $\mathbf{Q}$ which defined by (4.20) and (4.21) satisfies

$$\mathbf{Q} \cdot \mathbf{Q}^H = \mathbf{I} \tag{4.31}$$

From the transformation relationship of (4.28) and (4.29), we can see that $\mathbf{T}(\mathbf{X})$ transforms complex data into real data, so that the computational burden is lowered greatly, and we can obtain:

$$\begin{aligned} \mathbf{R}_T &= \frac{1}{2L}\mathbf{T}(\mathbf{X}) \cdot \mathbf{T}^H(\mathbf{X}) \\ &= \frac{1}{2L}\mathbf{Q}_M^H \cdot \mathbf{Z} \cdot \mathbf{Q}_{2L} \cdot \left(\mathbf{Q}_M^H \cdot \mathbf{Z} \cdot \mathbf{Q}_{2L}\right)^H = \frac{1}{2L}\mathbf{Q}_M^H \cdot \mathbf{Z} \cdot \mathbf{Q}_{2L} \cdot \mathbf{Q}_{2L}^H \cdot \mathbf{Z}^H \cdot \mathbf{Q}_M \\ &= \frac{1}{2L}\mathbf{Q}_M^H \cdot \mathbf{Z} \cdot \mathbf{Z}^H \cdot \mathbf{Q}_M = \mathbf{Q}_M^H \cdot \left[\frac{1}{2L}\left(\mathbf{Z} \cdot \mathbf{Z}^H\right)\right] \cdot \mathbf{Q}_M \\ &= \mathbf{Q}_M^H \cdot \mathbf{R}_{FB} \cdot \mathbf{Q}_M \end{aligned} \tag{4.32}$$

If the eigendecompositions of $\mathbf{R}_{FB}$ as follows:

$$\mathbf{R}_{FB} = \begin{bmatrix} \mathbf{U}_S & \mathbf{U}_N \end{bmatrix} \cdot \mathbf{\Sigma} \cdot \begin{bmatrix} \mathbf{U}_S^H \\ \mathbf{U}_N^H \end{bmatrix} \tag{4.33}$$

Insert (4.33) into (4.32), we can obtain:

$$\mathbf{R}_T = \mathbf{Q}_M^H \cdot \begin{bmatrix} \mathbf{U}_S & \mathbf{U}_N \end{bmatrix} \cdot \mathbf{\Sigma} \cdot \begin{bmatrix} \mathbf{U}_S^H \\ \mathbf{U}_N^H \end{bmatrix} \cdot \mathbf{Q}_M \tag{4.34}$$

(4.34) shows that the signal subspace of the transformation matrix $\mathbf{R}_T$ is:

$$\mathbf{E}_S = \mathbf{Q}_M^H \cdot \mathbf{U}_S \tag{4.35}$$

Insert (4.24) into (4.32), we can obtain:

$$\begin{aligned}\mathbf{R}_T &= \mathbf{Q}_M^H \cdot \mathbf{R}_{FB} \cdot \mathbf{Q}_M = \mathbf{Q}_M^H \cdot \left[\mathbf{A} \cdot \frac{1}{2}\left(\mathbf{R}_S + \mathbf{\Delta} \cdot \mathbf{R}_S^* \cdot \mathbf{\Delta}^H\right) \cdot \mathbf{A}^H + \mathbf{R}_N^{'} \right] \cdot \mathbf{Q}_M \\ &= \mathbf{Q}_M^H \cdot \mathbf{A} \cdot \frac{1}{2}\left(\mathbf{R}_S + \mathbf{\Delta} \cdot \mathbf{R}_S^* \cdot \mathbf{\Delta}^H\right) \cdot \mathbf{A}^H \cdot \mathbf{Q}_M + \mathbf{Q}_M^H \cdot \mathbf{R}_N^{'} \cdot \mathbf{Q}_M \\ &= \left(\mathbf{Q}_M^H \cdot \mathbf{A}\right) \cdot \frac{1}{2}\left(\mathbf{R}_S + \mathbf{\Delta} \cdot \mathbf{R}_S^* \cdot \mathbf{\Delta}^H\right) \cdot \left(\mathbf{Q}_M^H \cdot \mathbf{A}\right)^H + \mathbf{Q}_M^H \cdot \mathbf{R}_N^{'} \cdot \mathbf{Q}_M \\ &= \mathbf{A}_T \cdot \frac{1}{2}\left(\mathbf{R}_S + \mathbf{\Delta} \cdot \mathbf{R}_S^* \cdot \mathbf{\Delta}^H\right) \cdot \mathbf{A}_T^H + \mathbf{Q}_M^H \cdot \mathbf{R}_N^{'} \cdot \mathbf{Q}_M\end{aligned} \quad (4.36)$$

Therefore, the relationship between the real-value transformed signal direction matrix $\mathbf{A}_T$ and the original complex signal direction matrix $\mathbf{A}$ is given by:

$$\mathbf{A}_T = \mathbf{Q}_M^H \cdot \mathbf{A} \quad (4.37)$$

4.3.2 The real-value space rotational invariance principle

We analyze the signal subspace relationship of the two subarray data in the rotational invariance subspace algorithm theory, which is given by (4.14) $\mathbf{U}_{S2} = \mathbf{U}_{S1} \cdot \mathbf{\Psi}$. If the array is uniform linear array, and the overlap element of the two subarrays is maximum, namely, $m = M - 1$, so the signal subspace rotational invariance of the two subarray data can be expressed as:

$$\mathbf{K}_2 \cdot \mathbf{U}_S = \mathbf{K}_1 \cdot \mathbf{U}_S \cdot \mathbf{\Psi} \quad (4.38)$$

where $\mathbf{U}_S$ is the signal subspace of the received data of the whole uniform linear array, and:

$$\mathbf{K}_1 = \left[\mathbf{I}_{M-1} \quad 0\right]_{(M-1)\times M} \quad (4.39)$$

$$\mathbf{K}_2 = \left[0 \quad \mathbf{I}_{M-1}\right]_{(M-1)\times M} \quad (4.40)$$

In the same way, the rotational invariance of the two subarray signal direction matrix can be given as follows:

$$\mathbf{K}_2 \cdot \mathbf{A} = \mathbf{K}_1 \cdot \mathbf{A} \cdot \mathbf{\Phi} \quad (4.41)$$

where $\mathbf{A}$ is the signal direction matrix of the whole array.

From the definition of (4.39) and (4.40), we can see that $\mathbf{K}_1$ and $\mathbf{K}_2$ satisfies:

$$\mathbf{K}_1 = \mathbf{J}_m \cdot \mathbf{K}_2 \cdot \mathbf{J}_M \quad (4.42)$$

Utilize the relationship of the definition (4.19): $\mathbf{J}_M \cdot \mathbf{Q}^* = \mathbf{Q} \Rightarrow \mathbf{J}_M \cdot \mathbf{Q} = \mathbf{Q}^*$ again, we can obtain:

$$\begin{aligned}\mathbf{Q}_m^H \cdot \mathbf{K}_2 \cdot \mathbf{Q}_M &= \mathbf{Q}_m^H \cdot \mathbf{J}_m \cdot \mathbf{J}_m \cdot \mathbf{K}_2 \cdot \mathbf{J}_M \cdot \mathbf{J}_M \cdot \mathbf{Q}_M = \left(\mathbf{J}_m^H \cdot \mathbf{Q}_m\right)^H \cdot \mathbf{J}_m \cdot \mathbf{K}_2 \cdot \mathbf{J}_M \cdot (\mathbf{J}_M \cdot \mathbf{Q}_M) \\ &= (\mathbf{J}_m \cdot \mathbf{Q}_m)^H \cdot \mathbf{K}_1 \cdot (\mathbf{J}_M \cdot \mathbf{Q}_M) = \left(\mathbf{Q}_m^*\right)^H \cdot \mathbf{K}_1 \cdot \mathbf{Q}_M^* = \left(\mathbf{Q}_m^H\right)^* \cdot \mathbf{K}_1 \cdot \mathbf{Q}_M^* \\ &= \left(\mathbf{Q}_m^H \cdot \mathbf{K}_1 \cdot \mathbf{Q}_M\right)^*\end{aligned} \quad (4.43)$$

therefore, define:

$$\mathbf{H}_1 \overset{\Delta}{=} \mathbf{Q}_m^H \cdot \mathbf{K}_1 \cdot \mathbf{Q}_M + \mathbf{Q}_m^H \cdot \mathbf{K}_2 \cdot \mathbf{Q}_M = \mathbf{Q}_m^H \cdot (\mathbf{K}_1 + \mathbf{K}_2) \cdot \mathbf{Q}_M = 2\,\mathrm{Re}\left\{\mathbf{Q}_m^H \cdot \mathbf{K}_2 \cdot \mathbf{Q}_M\right\} \quad (4.44a)$$

$$\mathbf{H}_2 \overset{\Delta}{=} j \cdot \mathbf{Q}_m^H \cdot \mathbf{K}_1 \cdot \mathbf{Q}_M - j \cdot \mathbf{Q}_m^H \cdot \mathbf{K}_1 \cdot \mathbf{Q}_M = \mathbf{Q}_m^H \cdot j \cdot (\mathbf{K}_1 - \mathbf{K}_2) \cdot \mathbf{Q}_M = 2\,\mathrm{Im}\left\{\mathbf{Q}_m^H \cdot \mathbf{K}_2 \cdot \mathbf{Q}_M\right\} \quad (4.44b)$$

so that:

$$\mathbf{Q}_m^H \cdot \mathbf{K}_1 \cdot \mathbf{Q}_M = \frac{1}{2}(\mathbf{H}_1 - j\mathbf{H}_2) \quad (4.45a)$$

$$\mathbf{Q}_m^H \cdot \mathbf{K}_2 \cdot \mathbf{Q}_M = \frac{1}{2}(\mathbf{H}_1 + j\mathbf{H}_2) \quad (4.45b)$$

From the result given by (4.37): $\mathbf{A}_T = \mathbf{Q}_M^H \cdot \mathbf{A} \Rightarrow \mathbf{A} = \mathbf{Q}_M \cdot \mathbf{A}_T$, and insert it into the formula defined by (4.41): $\mathbf{K}_2 \cdot \mathbf{A} = \mathbf{K}_1 \cdot \mathbf{A} \cdot \mathbf{\Phi}$,we can obtain the results as follows:

$$\mathbf{K}_2 \cdot \mathbf{Q}_M \cdot \mathbf{A}_T = \mathbf{K}_1 \cdot \mathbf{Q}_M \cdot \mathbf{A}_T \cdot \mathbf{\Phi} \quad (4.46)$$

The both side of the upper formula multiplies by the $\mathbf{Q}_m^H$ together, we can obtain:

$$\mathbf{Q}_m^H \cdot \mathbf{K}_2 \cdot \mathbf{Q}_M \cdot \mathbf{A}_T = \mathbf{Q}_m^H \cdot \mathbf{K}_1 \cdot \mathbf{Q}_M \cdot \mathbf{A}_T \cdot \mathbf{\Phi} \quad (4.47)$$

Using (4.45), and removing the constant factor $1/2$, we can obtain that:

$$(\mathbf{H}_1 + j\mathbf{H}_2) \cdot \mathbf{A}_T = (\mathbf{H}_1 - j\mathbf{H}_2) \cdot \mathbf{A}_T \cdot \mathbf{\Phi} \quad (4.48)$$

Via moving item, combination and so on simplifications, we will have:

$$\mathbf{H}_1 \cdot \mathbf{A}_T \cdot (\mathbf{\Phi} - \mathbf{I}) = \mathbf{H}_2 \cdot \mathbf{A}_T \cdot j \cdot (\mathbf{\Phi} + \mathbf{I}) \quad (4.49)$$

From the definition of (4.3) $\mathbf{\Phi} = diag\left[e^{j\varphi_1} \quad \cdots \quad e^{j\varphi_N}\right]$ again, (4.49) can be simplified as：

$$\mathbf{H}_2 \cdot \mathbf{A}_T = \mathbf{H}_1 \cdot \mathbf{A}_T \cdot \frac{1}{j}(\mathbf{\Phi} - \mathbf{I}) \cdot (\mathbf{\Phi} + \mathbf{I})^{-1} = \mathbf{H}_1 \cdot \mathbf{A}_T \cdot \mathbf{\Phi}_T \quad (4.50)$$

where

$$\begin{aligned} \mathbf{\Phi}_T &= \frac{1}{j}(\mathbf{\Phi} - \mathbf{I}) \cdot (\mathbf{\Phi} + \mathbf{I})^{-1} & (4.51) \\ &= \frac{1}{j} \cdot diag\left\{e^{j\varphi_1} - 1 \quad \cdots \quad e^{j\varphi_N} - 1\right\} \cdot diag\left\{\frac{1}{e^{j\varphi_1} + 1} \quad \cdots \quad \frac{1}{e^{j\varphi_N} + 1}\right\} \\ &= \frac{1}{j} \cdot diag\left\{\frac{e^{j\varphi_1} - 1}{e^{j\varphi_1} + 1} \quad \cdots \quad \frac{e^{j\varphi_N} - 1}{e^{j\varphi_N} + 1}\right\} \\ &= diag\left\{\tan\left(\frac{\varphi_1}{2}\right) \quad \cdots \quad \tan\left(\frac{\varphi_N}{2}\right)\right\} & (4.52) \end{aligned}$$

So that, (4.50) reflects the rotational invariance relationship of the real-value space array steering, but (4.51) reflects the rotational invariance relationship between the real-value space array steering and the complex value space array steering.

Resembling the derivation of (4.50), from $\mathbf{E}_S = \mathbf{Q}_M^H \cdot \mathbf{U}_S \Rightarrow \mathbf{U}_S = \mathbf{Q}_M \cdot \mathbf{E}_S$, and insert it into the formula given by (4.38): $\mathbf{K}_2 \cdot \mathbf{U}_S = \mathbf{K}_1 \cdot \mathbf{U}_S \cdot \mathbf{\Psi}$, we can obtain:

$$\mathbf{K}_2 \cdot \mathbf{Q}_M \cdot \mathbf{E}_S = \mathbf{K}_1 \cdot \mathbf{Q}_M \cdot \mathbf{E}_S \cdot \mathbf{\Psi} \tag{4.53}$$

The both side of the upper formula multiplies by the $\mathbf{Q}_m^H$ together, we can obtain:

$$\mathbf{Q}_M^H \cdot \mathbf{K}_2 \cdot \mathbf{Q}_M \cdot \mathbf{E}_S = \mathbf{Q}_M^H \cdot \mathbf{K}_1 \cdot \mathbf{Q}_M \cdot \mathbf{E}_S \cdot \mathbf{\Psi} \tag{4.54}$$

Using (4.45), and removing the constant factor $1/2$, we can obtain that:

$$\left(\mathbf{H}_1 + j\mathbf{H}_2\right) \cdot \mathbf{E}_S = \left(\mathbf{H}_1 - j\mathbf{H}_2\right) \cdot \mathbf{E}_S \cdot \mathbf{\Psi} \tag{4.55}$$

Via moving item, combination and so on simplifications, we will have:

$$\mathbf{H}_2 \cdot \mathbf{E}_S \cdot \left(j\mathbf{\Psi} + \mathbf{I}\right) = \mathbf{H}_1 \cdot \mathbf{E}_S \cdot \left(\mathbf{\Psi} - \mathbf{I}\right) \tag{4.56}$$

$$\mathbf{H}_2 \cdot \mathbf{E}_S = \mathbf{H}_1 \cdot \mathbf{E}_S \cdot \left(\mathbf{\Psi} - \mathbf{I}\right) \cdot \left(j\mathbf{\Psi} + \mathbf{I}\right)^{-1} = \mathbf{H}_1 \cdot \mathbf{E}_S \cdot \mathbf{\Psi}_T \tag{4.57}$$

where

$$\mathbf{\Psi}_T = \left(\mathbf{\Psi} - \mathbf{I}\right) \cdot \left(j\mathbf{\Psi} + \mathbf{I}\right)^{-1} \tag{4.58}$$

So that, (4.57) reflects the rotational invariance relationship of the real-value space signal subspace, but (4.58) reflects the rotational invariance relationship between the real-value space signal subspace and the complex value space signal space.

Utilizing the character that the space spanned by array direction matrix is equal to which is spanned by the signal subspace, so a nonsingular matrix $\mathbf{T}_T$ exists, and satisfying $\mathbf{A}_T = \mathbf{E}_S \cdot \mathbf{T}_T$, thus using (4.50): $\mathbf{H}_2 \cdot \mathbf{A}_T = \mathbf{H}_1 \cdot \mathbf{A}_T \cdot \mathbf{\Phi}_T$, we can obtain that:

$$\mathbf{H}_2 \cdot \mathbf{E}_S \cdot \mathbf{T}_T = \mathbf{H}_1 \cdot \mathbf{E}_S \cdot \mathbf{T}_T \cdot \mathbf{\Phi}_T \Rightarrow \mathbf{H}_2 \cdot \mathbf{E}_S = \mathbf{H}_1 \cdot \mathbf{E}_S \cdot \mathbf{T}_T \cdot \mathbf{\Phi}_T \cdot \mathbf{T}_T^{-1} \tag{4.59}$$

Comparing with (4.57), we can obtain that:

$$\mathbf{\Psi}_T = \mathbf{T}_T \cdot \mathbf{\Phi}_T \cdot \mathbf{T}_T^{-1} \tag{4.60}$$

This formula reflects the rotational invariance relationship between the array steering and the signal subspace of the real-value space.

4.3.3 The real-value space ESPRIT algorithm

The observational data of M elements are given as:

$$x_1(t),\ \cdots,\ x_M(t),\ t = 1,\ \cdots\ L$$

Step 1. Construct the $M \times L$ observational data matrix $\mathbf{X} = \left[\mathbf{x}(1), \ \cdots, \ \mathbf{x}(L)\right]$, where $\mathbf{x}(t) = \left[x_1(t), \ \cdots, \ x_M(t)\right]^T$ is the observational data vector which is consists of M elements observational signals.

Step 2. Get the estimating formula of $\mathbf{R}$ by $\overset{\wedge}{\mathbf{R}} = \frac{1}{L}\left(\mathbf{X} \cdot \mathbf{X}^H\right)$, and transform the received array data into real-value space $\overset{\wedge}{\mathbf{R}}_T$ via (4.32).

Step 3. Compute the eigendecompositions of the real-value space $\overset{\wedge}{\mathbf{R}}_T$, and get the signal subspace $\overset{\wedge}{\mathbf{E}}_S$, and the source number $\hat{N}$.

Step 4. Solve the rotational invariance of (4.57) by least square method (or total least square method), and gain $\overset{\wedge}{\mathbf{\Psi}}_T$.

Step 5. Compute the eigendecompositions of $\overset{\wedge}{\mathbf{\Psi}}_T$, where $\overset{\wedge}{\mathbf{\Psi}}_T = \overset{\wedge}{\mathbf{T}}_T \cdot \overset{\wedge}{\mathbf{\Phi}}_T \cdot \overset{\wedge}{\mathbf{T}}_T^{-1}$, get $\overset{\wedge}{\mathbf{\Phi}}_T = diag\left\{\Omega_1, \ \cdots, \ \Omega_{\hat{N}}\right\}$.

Step 6. If $\overset{\wedge}{\mathbf{\Phi}}_T$ is the real diagonal matrix, according as (4.3) and (4.52), compute the DOA of imping signal as fellows:

$$\begin{cases} \varphi_k = 2 \cdot \arctan\left(\Omega_k\right) \\ \theta_k = \arcsin\left(\dfrac{\lambda}{2 \cdot \pi |\Delta|} \cdot \varphi_k\right) \end{cases} \qquad \left(k = 1, \ \cdots, \ \hat{N}\right) \tag{4.61}$$

If $\Omega_k\left(k = 1, \ \cdots, \ \hat{N}\right)$ is complex, compute the DOA by (4.61) with the real part of Ω_k.

4.4 Simulations

In order to validating the correctness and the effective of the proposed algorithm, we present some simulation results to illustrate the performance of RVS-ESPRIT. We consider a ULA with M=8 element and the interelement space is equal to a half of wavelength. There are three signals impinge on the array from $\theta_1 = -80$, $\theta_2 = -20$, $\theta_3 = 40$. The detailed simulation results are shown as Fig. 26. ~ Fig. 29.

Fig. 26. and Fig. 27. depicts DOA departure versus snapshot number results of RVS-ESPRIT and TLS-ESPRIT respectively, where the SNR=5dB. In figure 26. and 27., the x-axis denotes the snapshot number, and y-axis denotes the departure of signal DOA.

Fig. 28. and Fig. 29. depicts DOA departure versus SNR results of RVS-ESPRIT and TLS-ESPRIT respectively, where the snapshot number =1000. In figure 28. and 29., the x-axis denotes the SNR, and y-axis denotes the departure of signal DOA.

From the detecting results and comparison between RVS-ESPRIT and TLS-ESPRIT, we can conclude that RVS-ESPRIT can detect DOA of signal quickly and effectively. At the same time, the results validate the correctness and effective of this algorithm.

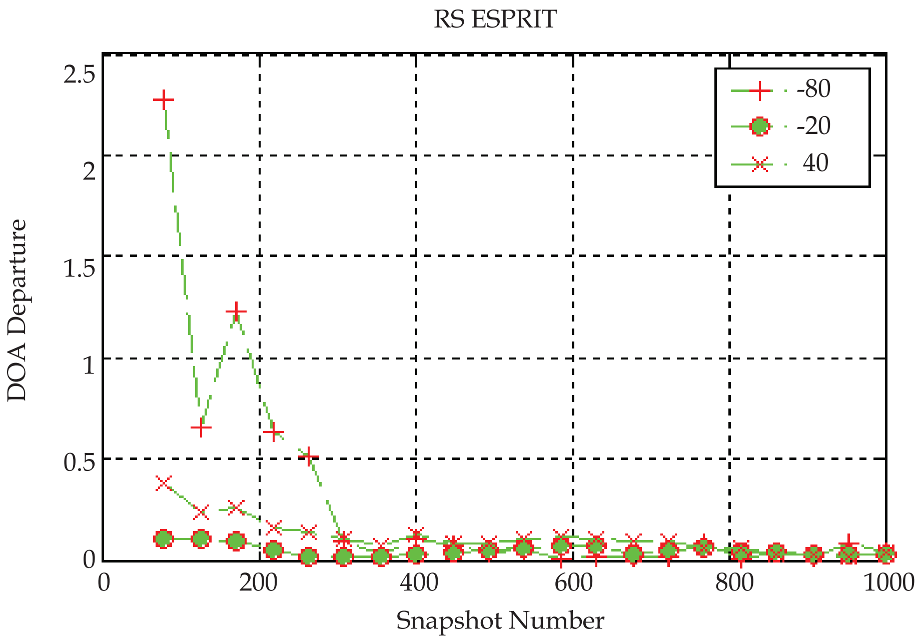


Fig. 26. DOA departure vs snapshot number. Signal DOA=[-80 -20 40], SNR=5dB

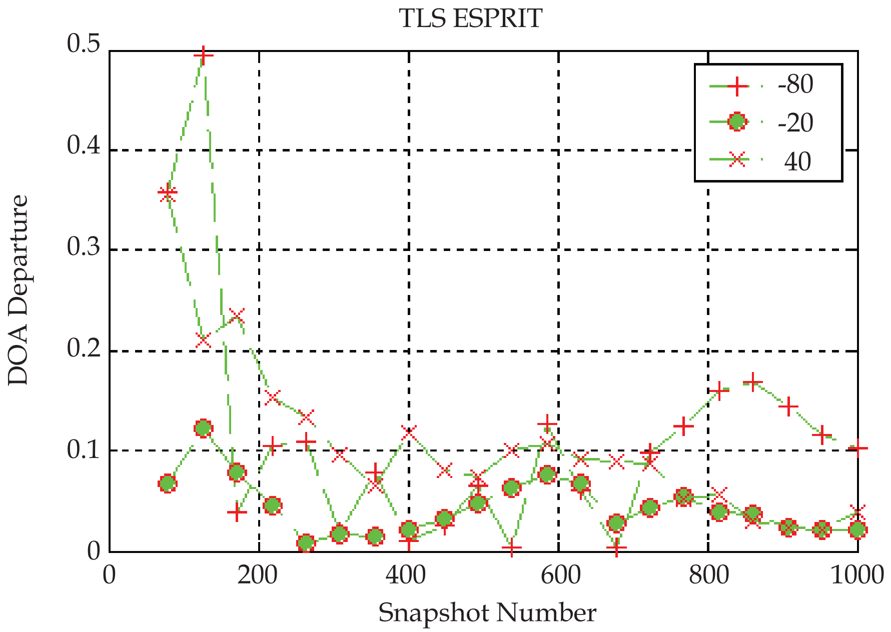


Fig. 27. DOA departure vs snapshot number. Signal DOA=[-80 -20 40], SNR=5dB

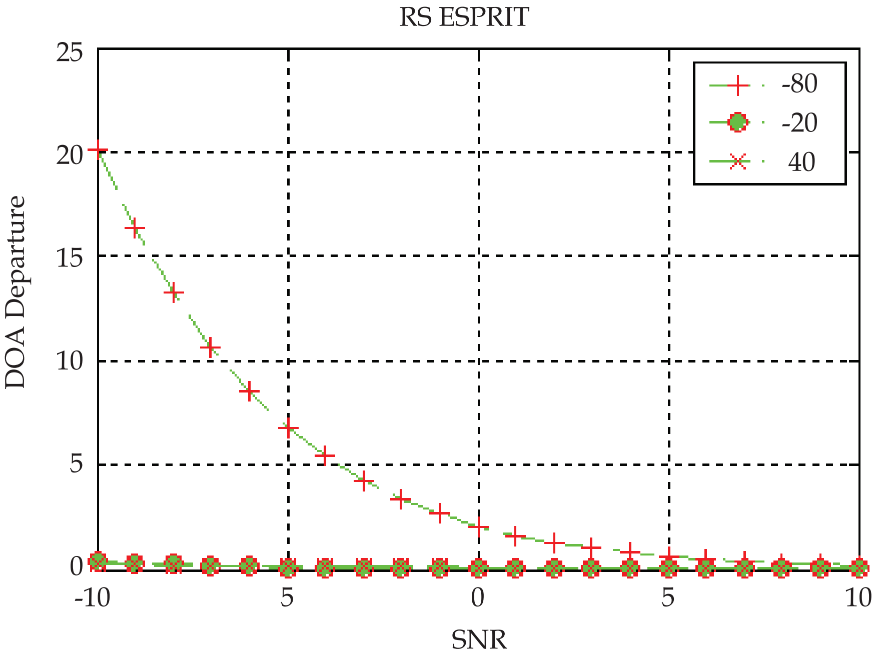


Fig. 28. DOA departure versus SNR. Signal DOA=[-80 -20 40], Snapshot number =1000

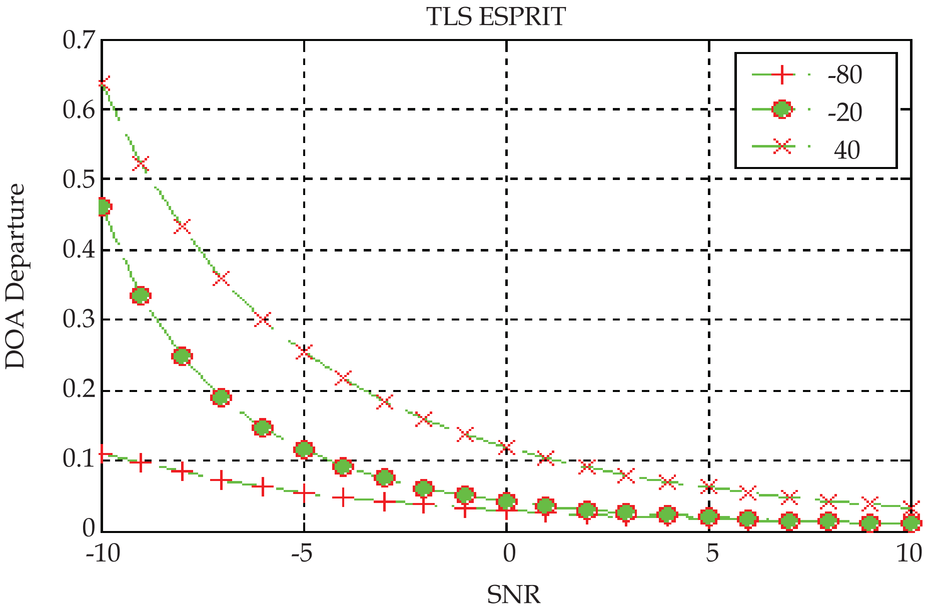


Fig. 29. DOA departure versus SNR. Signal DOA=[-80 -20 40], Snapshot number =1000

4.5 Conclusion

This chapter carrys on the detailed theories analysis of RVS-ESPRIT based on the theory of CS-ESPRIT, and gives the concrete implementing algorithm. Because the eigendecompositions of RVS-ESPRIT is in real domain, so the calculation speed is raised consumedly, then the speed of DOA estimating is improved largely also. Due to the inherent forward-backward averaging effect, RVS-ESPRIT can separate two completely coherent sources and provides improved estimates for correlated signals.

5. References

[1] J. Capon. High resolution frequency-wavenumber spectrum analysis. Proc. IEEE, Vol.57, pp: 1408-1418, Aug.1969.

[2] J. S. Reed, J. D. Mallet, L. E. Brennan. Rapid convergence rate in adaptive arrays. IEEE Trans. Aerosp. Electron. Syst., Vol. AES-10, No.6, pp: 853-863, Nov.1974.

[3] D. H. Johnson, D. E. Dudgeon. Array signal processing: Concepts and Techniques. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[4] H. Krim, M. Viberg. Two decades of array signal processing research. IEEE Signal Process. Mag., Vol.13, No.4, pp: 67-94, Jul.1996.

[5] L. C. Godara. Application of antenna arrays to mobile communications, Part II: Beam-forming and direction-of-arrival considerations. Proc. IEEE, Vol.85, No.8, pp: 1195-1245, Aug.1997.

[6] H. L. Van Trees. Detection, estimation, and modulation theory, Part IV, Optimum array processing. New York: Wiley, 2002.

[7] P. S. Naidu. Sensor array signal processing. Boca Raton, FL: CRC, 2001.

[8] J. R. Guerci. Space-time adaptive processing. Norwood, MA: Artech House, 2003.

[9] Jian Li, Petre Stotica. Robust adaptive beamforming. New York: Wiley, 2006.

[10] H. Cox. Resolving power and sensitivity to mismatch of optimum array processors. J. Acoust. Soc. Amer., Vol.54, No.3, pp: 771-785, Mar.1973.

[11] D. D. Feldman, L. J. Griffiths. A projection approach to robust adaptive beamforming. IEEE Trans. Signal Processing, Vol.42, pp: 867-876, Apr.1994.

[12] M. Wax, Y. Anu. Performance analysis of the minimum variance beamformer in the presence of steering vector errors. IEEE Trans. Signal Processing, Vol.44, pp: 938-947, Apr.1996.

[13] E. K. Hung, R. M. Turner. A fast beamforming algorithm for large arrays. IEEE Trans. Aerosp. Electron. Syst., Vol.AES-19, pp: 598-607, Jul.1983.

[14] B. D. Carlson. Covariance matrix estimation errors and diagonal loading in adaptive arrays. IEEE Trans. Aerosp. Electron. Syst., Vol.24, pp: 397-401, Jul.1988.

[15] M. Wax, Y. Anu. Performance analysis of the minimum variance beamformer. IEEE Trans. Signal Processing, Vol.44, pp: 928-937, Apr.1996.

[16] H. Cox, R. M. Zeskind, M. H. Owen. Robust adaptive beamforming. IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-35, pp:1365-1376, Oct.1987.

[17] A. B. Gershman. Robust adaptive beamforming in sensor arrays. Int. J. Electron. Commun., Vol.53, pp: 305-314, Dec.1999.

[18] A. B. Gershman. Robust adaptive beamforming: an overview of recent trends and advances in the field. International Conference on Antenna Theory and Technique, 9-12 September, 2003, Sewstopol, Ukraine, pp: 30-35.

[19] Sergiy A. Vorobyov, A. B. Gershman, Zhi-Quan Luo. Robust adaptive beamforming using worst-case performance optimization: a solution to the signal mismatch problem. IEEE Trans. Signal Processing. Vol.51, No.2, pp: 313-324, Feb.2003.

[20] Jian Li, Petre Stotica, Zhisong Wang.On robust capon beamformer and diagonal loading. IEEE Trans. Signal Processing. Vol.51, No.7, pp: 1702-1715, Jul.2003.

[21] S. Shahram, A. B. Gershman, Zhiquan Luo, et al. Robust adaptive beamforming for general-rank signal models. IEEE Trans. Signal Processing, Vol.51, No.9, pp: 2257-2269, Sep.2003.

[22] Jian Li, Petre Stotica, Zhisong Wang, Doubly constrained robust capon beamformer. IEEE Trans. Signal Processing. Vol.52, No.9, pp: 2407-2423, Sep.2004.

[23] G.L. Robert, P.B. Stephen. Robust minimum variance beamforming. IEEE Trans. Signal Processing, Vol.53, No.5, pp: 1684-1696, May.2005.

[24] Ayman Elnashar, Said M. Elnoubi, Hamdi A. El-Mikati. Further study on robust adaptive beamforming with optimum diagonal loading. IEEE Trans. Antennas Propagation, Vol.AP-54, No.12, pp: 3647-3658, Dec.2006.

[25] C.Y.Chen, P.P.Vaidyanathan. Quadratically constrained beamforming robust against direction-of- arrival mismatch. IEEE Trans. Signal Processing, Vol.55, No.8, pp: 4139-4150, Aug.2007.

[26] R. A. Monzingo, T. W. Miller. Introduction to adaptive arrays. New York: Wiley, 1980.

[27] Z. Tian, K. L. Bell, H. L. Van Trees. A recursive least squares implementation for LCMP beamforming under quadratic constraint. IEEE Trans. Signal Processing, Vol.49, No.6, pp: 1365-1376, Jun.2001.

[28] L. C. Godara. Error analysis of the optimal antenna array processors. IEEE Trans. Aerosp. Electron. Syst., Vol. AES-22, pp: 395-409, Jul.1986.

[29] K. L. Bell, Y. Ephraim, H. L. Van Trees, A Bayesian approach to robust adaptive beamforming. IEEE Trans. Signal Processing, Vol.48, pp: 386-398, Feb.2000.

[30] L. Chang, C. C. Yeh. Performance of DMI and eigenspace-based beamformers. IEEE Trans. Antennas Propagation, Vol.40, pp: 1336-1347, Nov.1992.

[31] Cheng-Chou Lee, Ju-Hong Lee. Eigenspace-based adaptive array beamforming with robust capabilities. IEEE Trans. Antennas Propagation, Vol.45, pp: 1711-1716, Dec.1997.

[32] J. Riba, J. Goldberg, G. Vazquez. Robust beamforming for interference rejection in mobile communications. IEEE Trans. Signal Processing, Vol.45, pp: 271-275, Jan.1997.

[33] J. R. Guerci. Theory and application of covariance matrix tapers for robust adaptive beamforming. IEEE Trans. Signal Processing, Vol.47, pp: 997-985, Apr.1999.

[34] J. R. Guerci, J.S.Bergin. Principal components,covariance matrix tapers,and the subspace leakage problem. IEEE Trans. Aerosp. Electron. Syst, Vol.38, pp: 152-162, Jan.2002.

[35] F. Vincent, O. Besson. Steering vector errors and diagonal loading, IEE Proceedings.- Radar Sonar Navig., Vol.151, No.6, pp: 337-343, Dec.2004.

[36] O. Besson, F. Vincent. Performance analysis of beamformers using generalized loading of the covariance matrix in the presence of random steering vector errors. IEEE Trans. Signal Processing, Vol.53, pp: 452-459, Feb.2005.

[37] Almir Mutapcic, Seung-Jean Kim, Stephen Boyd. Beamforming with uncertain weights. IEEE Signal Processing Letter, Vol.14, No.5, pp: 348-351, May.2007.

[38] Liu Congfeng, Liao Guisheng, Robust capon beamformer under norm constraint, Signal Processing, May.2010, Vol.90, Issue.5, pp: 1573-1581.

[39] Dolph C L. A current distribution for broadside arrays which optimizes the relationship between beam width and sidelobe level. Proc IRE, June.1946, Vol.34, pp:335–348.

[40] Zhou P, Ingram M. Pattern synthesis for arbitrary arrays using an adaptive array method. IEEE Transactions on Antennas And Propagation. May.1999, Vol.47, No.5, pp:862-869.

[41] Wang F, Balakrishnan V, Zhou P Y, et al. Optimal array pattern synthesis using semidefinite programming. IEEE Transactions on Signal Processing, May.2003, Vol.51, No.5, pp:1172-1183.

[42] Guo Q, Liao G., Wu Y, et al. Pattern synthesis method for arbitrary arrays based on LCMV criterion. Electronics Letters. May.2003, Vol.39, No.23, pp:1628-1630.

[43] Xie Yao, Jian Li, Zheng Xiayu, et al. Optimal array pattern synthesis via matrix weighting, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, Apr 15-20 2007, Honolulu, HI, United States, IEEE Inc., Piscataway, NJ 08855-1331, United States, Vol.2, pp: 885-888.

[44] Kurup D G., Himdi M, Rydberg A. Synthesis of uniform amplitude unequally spaced antenna arrays using the differential evolution algorithm, IEEE Transactions on Antennas And Propagation. Sep.2003, Vol.51, No.9, pp:2210-2217.

[45] Boeringer D W, Werner D H. Particle swarm optimization versus genetic algorithms for phased array synthesis, IEEE Transactions on Antennas And Propagation. Mar.2004, Vol.52, No.3, pp:771-779.

[46] Liu Xiaojun, Liu Congfeng, Liao Guisheng, Improved pattern synthesis method with linearly constraint minimum variance criterion, 2010 IEEE Internation Conference on Wireless Communications, Networking and Information Security (WCNIS2010), 2010.6,25-27, Vol.2, Beijing China。

[47] Barabell A.J, Improving the resolution performance of eigenstructure-based direction-finding algorithms.Proc,ICASSP 83,1983,pp.336-339.

[48] Kumaresan R. and Tufts D.W., Estimating the angles of arrival of multiple plane wave,IEEE Trans,1983,AES-19,pp.134-138.

[49] Bao B and Hari.K, Performance analysis of root-MUSIC, IEEE Trans, 1989, ASSP-37,pp.1939-1949.

[50] Li F and Vaccaro.R.J,Analytical performance prediction of subspace-based algorithms for DOA estimation: SVD and Signal Processing II Algorithms,Analysis and Applications (Elsevier,New York,1991)

[51] Xu X.L and Buckley K.M, Reduced dimension beamspace broad-band localization, preprocessor design and evaluation.Proc.IEEE Forth Workshop on Spectrum Estimation and Modeling, 1988,pp22-26.

[52] Ta sung Lee, Fast implementation of root-form eigen-based methods for detecting closely spaced sources IEE Proceedings-F Vol.139.No.4,August 1992.

[53] Marius P, Alex B.G and Martin H, Unitary root-MUSIC with a real-valued eigendecomposition: a theoretical and experimental performacce study, IEEE Transactions on SP,Vol.48,No.5,May 2000,pp:1306-1314.

[54] Liu Congfeng, Liao Guisheng, Fast algorithm for root-MUSIC with real-valued eigendecomposition, Radar-2006: 2006 CIE International Conference on Radar 2006.10 ShangHai China.

[55] Q. S Ren and A. J. Willis, Extending MUSIC to single snapshot and on line direction finding applications, International Radar Conference, IEE, Stevenage, United Kingdom, Oct.1997: 783-787.

[56] Liu Xiaojun, Liu Congfeng, Liao Guisheng, Polynomial coefficient finding for root-MUSIC, Journal of Electronic (China), 2009.8, Vol.26(4).

[57] R. Roy and T. Kailath, "ESPRIT-Estimation of signal parameters via rotational invariance techniques," in Signal Processing Part 11: Control Theory and Applications (L. Auslander, F. A. Griinbaum, J. W. Helton, T. Kailath, P. Khargonekar, and S. Mitter, Eds.). Berlin, Vienna, New York Springer-Verlag, 1990, pp. 36941 1.

[58] R. H. Roy, "ESPRIT-estimation of signal parameters via rRotational invariance techniques," Ph.D. thesis, Stanford Univ., Stanford, CA, Aug. 1987.

[59] M. Haardt and M. E. Ali-Hackl, "Unitary ESPRIT: How to exploit additional information inherent in the rotational invariance structure," in Proc. IEEE Int. ConJ Acoust., Speech, Signal Processing, Adelaide, Australia, Apr. 1994, pp. 229-232, vol. IV.

[60] S. U. Pillai and B. H. Kwon, "Forwardhackward spatial smoothing techniques for coherent signal identification," IEEE Trans. Acoust., Speech, Signal Processing, vol. 37, pp. 8-15, Jan. 1989.

[61] B. D. Rao and K. V. S. Hari, "Weighted subspace methods and spatial smoothing: Analysis and comparison," IEEE Trans. Signal Processing, vol. 41, pp. 788-803, Feb. 1993.

[62] Martin Haardt, Josef A. Nossek Unitary ESPRIT: how to obtain increased estimation accuracy with a reduced computational burden IEEE Trans. Signal Processing, Vol. 43, No. 5, pp.1232-1242. MAY 1995

[63] B. Ottersten, M. Viberg, and T. Kailath, "Performance analysis of the total least squares ESPRIT algorithm," IEEE Trans. Signal P mcessing, vol. 39, pp. 1122-1135, May 1991.

[64] Wang Yongliang, Chen Hui, Peng Yingning, Wan qun Spatial spectrum estimation techniques and algorithm Tsinghua university press Beijing China 2004.

[65] Liu Congfeng, Liao Guisheng, Real-value space ESPRIT algorithm and Its implement, WiCOM 2006: 2006 International Conference on Wireless Communications, Networking, and Mobile Computing 2006.09 WuHan China.

5

Analysis of Long-Periodic Fluctuations of Solar Microwave Radiation, as a Way for Diagnostics of Coronal Magnetic Loops Dynamics

Maxim L. Khodachenko[1], Albert G. Kislyakov[2] and Eugeny I. Shkelev[2]
[1]*Space Research Institute, Austrian Academy of Sciences, Graz,*
[2]*Lobachevsky State University, Nizhny Novgorod,*
[1]*Austria*
[2]*Russia*

1. Introduction

The solar corona has a very complex and highly dynamic structure. It consists of a large number of constantly evolving, loops and filaments, which interact with each other and are closely associated with the local magnetic field. The non-stationary character of solar plasma-magnetic structures manifests itself in various forms of the coronal magnetic loops dynamics as rising motions, oscillations, meandering, twisting (Aschwanden et al., 1999; Schrijver et al., 1999), as well as in formation, sudden activation and eruption of filaments and prominences. Energetic phenomena, related to these types of magnetic activity, range from tiny transient brightenings (micro-flares) and jets to large, active-region-sized flares and coronal mass ejections (CMEs). They are naturally accompanied by different kinds of electromagnetic (EM) emission, covering a wide frequency band from radio waves to gamma-rays. Radiation, produced within a given plasma environment, carries an information on physical and dynamic conditions in a radiating source. This causes an exceptional importance of the EM radiation, as a diagnostic tool, for understanding the nature and physics of various solar dynamic phenomena. As a relatively new, in that context, direction of study in the traditional branch of the solar microwave radio astronomy appears the analysis of the slow, long-periodic (e.g., > 1 s) fluctuations of the radiation intensity (Khodachenko et al., 2005; Zaitsev et al., 2003).

Microwave radiation from the magnetic loops in solar active regions (e.g., during solar flares) is usually interpreted as a gyro-synchrotron radiation, produced by fast electrons on harmonics of the gyro-frequency ν_B in the magnetic field B of the loop. In the case of a power-law distribution of electrons in energy as $f(\mathcal{E}) \propto \mathcal{E}^{-\delta}$, the intensity of gyro-synchrotron radiation I_ν from an optically thin loop (Dulk, 1985; Dulk & Marsh, 1982) is

$$I_\nu \propto B^{-0.22+0.9\delta}(\sin\theta)^{-0.43+0.65\delta}, \qquad (1)$$

where θ is the angle between magnetic field and the direction of electromagnetic wave propagation. For the observed typical values of the electron energy spectrum index $2 \leq \delta \leq 7$ this implies the proportionality of intensity to a moderately high power of the background magnetic field and essential anisotropy of the radiation: $I_\nu \propto B^{1.58 \div 6.08}(\sin\theta)^{0.87 \div 4.12}$. Equation (1) is obtained within an assumption of an optically thin source, when the radiation intensity is proportional to the emissivity η_ν (Dulk, 1985). According to the estimations in Urpo et al. (1994), a coronal loop with diameter of about 4×10^8 cm is optically thin in the considered frequency range for the gyro-synchrotron absorption if the density of fast electrons is $< 2 \times 10^9$ cm^{-3}. The typical density of > 10 keV electrons in the microwave burst events is usually $10^6 \div 10^7$ cm^{-3} and therefore, it stays well within the above indicated limit.

It follows from the equation (1) that variations of the loop magnetic field, associated with disturbances of the electric current in a radiating source, should modulate the intensity of the microwave radiation of the loop (Khodachenko et al., 2005; Zaitsev et al., 2003). Another origin for the modulation of intensity of the observed microwave radiation can be due to the quasi-periodic motion (oscillation) of a coronal magnetic loop, containing the radiation source. This mechanism is connected with the anisotropy of the gyro-synchrotron emission, as well as with the variation of the magnetic field value during the oscillatory motion of the loop, which both, according to the equation (1), can result in a quasi-periodic modulation of the received signal (Khodachenko et al., 2006; 2011). Therefore, the analysis of slow modulations of solar microwave radiation may be used for the diagnostics of oscillating electric currents in the coronal loops, as well as for the investigation of large-scale motion of the loops (including loop oscillations) in solar active regions. By this, it is natural to expect that structural complexity of solar active regions will manifest itself in peculiarities of the emitted radiation.

The dynamic spectra of the long-periodic oscillations, modulating the intensity of microwave radiation from solar active regions, have been found to contain quite often several spectral tracks, demonstrating a specific temporal behaviour (Zaitsev et al., 1998; 2001a;b; 2003). Khodachenko et al. (2005) considered these multi-track features as an indication that the detected microwave radiation is produced within a system of a few closely located, magnetic loops, having slightly different parameters and involved in a kind of common global dynamic process. In several cases such slow modulations of solar microwave radiation (with multi-track spectra) were interpreted as the signatures of oscillating electric currents, running within the circuits of moving relative each other inductively connected coronal magnetic loops (Khodachenko et al., 2006). The dynamics of these electric currents has been described by means of the equivalent electric circuit (LCR-circuit) models of the coronal loops (Khodachenko et al., 2003; Zaitsev et al., 1998) characterized by time-dependent inductance L, capacitance C, resistance R, as well as mutual inductance coefficients M_j (Khodachenko et al., 2003; 2009). The L, C, R and M_j parameters of electric circuit of a current-carrying loop depend on shape, scale, position of the loop with respect to other loops, as well as on the plasma parameters and value of the total longitudinal current in the magnetic tube. In that respect it is worth to mention that the LCR-circuit model ignores the fact that changes of the magnetic field and related electric current propagate in plasma at the Alfvén speed. It ignores any short-time variations of plasma parameters, which appear to be averaged in course of derivation of the LCR model equation (Khodachenko et al., 2009; Zaitsev et al., 2001a). The LCR approach assumes instant changes of the electric current over the whole electric circuit according to the varying potential and ignores all the "propagation effects" related to the system MHD modes. The LCR equations correctly describe temporal evolution of electric currents in a system of solar magnetic current-carrying loops only at a time scale longer than the Alfvén wave propagation time. More generally, the equivalent electric circuit model of a coronal loop tends to emphasize the global electric circuit, obscuring the effects of the ambient plasma and details of the magnetic structure. The LCR model approach was applied in particular for interpretation of the solar microwave burst long-periodic modulations with drifting modulation frequencies $\nu(t)$ in the interval $0.03 \div 1$ Hz. Based on the analysis of the frequency drift of the modulations Zaitsev et al. (1998; 2001a; 2003) and Khodachenko et al. (2005) estimated also the values of the electric current in flaring loops ($10^8 \div 10^{11}$ A), which appeared to be close to the values obtained by other methods (Hardy et al., 1998; Leka et al., 1996; Moreton & Severny, 1968; Spangler, 2007; Tan et al., 2006).

However, not all long-periodic modulations of solar microwave radiation demonstrate frequency drift and occupy a similar frequency range. Khodachenko et al. (2009) pointed out that the frequency of LCR-oscillations of the electric current, which depends on specific parameters of a coronal loop, usually stays within the interval $\nu_{LCR} \approx (0.03 \div 1)$ Hz. Therefore, the modulations of the solar microwave radiation intensity with 0.03Hz$< \nu <$ 1Hz are very likely to be due to the electric currents, oscillating in LCR-circuits of coronal loops. At the same time, modulations caused by the oscillatory motions of loops that contain the radiation sources, because of their direct connection with the large-scale dynamics of loops, should have typical frequencies < 0.01 Hz and exhibit no drift. Thus, it has been proposed in (Khodachenko et al., 2009) that one should distinguish, when speaking about different kinds of long-periodic modulations of the solar microwave radiation, between the *low-frequency (LF)* ($\approx 0.03 \div 1$ Hz) and *very-low-frequency (VLF)* (< 0.01 Hz) modulations, assuming the first to be connected with the LCR-oscillations of electric currents in the coronal loops and the second to be caused by large-scale motions of the radiation sources confined within the oscillating loops. LF modulations (e.g., $0.03 \div 1$ Hz) have been studied in details and interpreted in terms of the equivalent electric circuit models of coronal loops in Khodachenko et al. (2005; 2006); Zaitsev et al. (1998; 2001a;b; 2003). They are not considered in this chapter, whereas we addresses here the VLF modulations (< 0.01 Hz) of solar microwave radiation and their possible relation to the large-scale dynamics of coronal loops in solar active regions. Some preliminary results on that subject were reported recently in a short publication Khodachenko et al. (2011), and the present paper addresses this topic in more details.

Transverse oscillations of the coronal magnetic loops considered here are triggered by flares and filament eruptions, i.e. by phenomena in which the significant Lorentz forces are likely acting in association with the magnetic field adjustments. By this, all considered oscillating loops have at least one footpoint in the immediate vicinity of a separatrix surface or of a flare ribbon. Using the potential-field extrapolations Schrijver & Brown (2000) demonstrated that the field lines close to a separatrix surface exhibit strongly amplified displacements in response to small displacements in the photospheric "roots". This means that the magnetic field lines in the proximity of separatrix are much more sensitive to changes in the field sources than are the field lines that lie well within domains of connectivity. Speaking about a nature of the observed transversal oscillations of coronal loops Schrijver et al. (2002) address two models: (a) transverse waves in coronal loops that act as wave guides and (b) mentioned above, strong sensitivity of the shape of magnetic field lines near separatrix to changes in the bottom field sources. By this, the authors outline several observational features that favor the model (b). Based on the extensive study of properties of transverse loop oscillations triggered by flares, Aschwanden et al. (2002) also concluded that most of the loops do not fit the simple model of a kink eigen-mode oscillation. Therefore, the present paper is not dedicated to the study of MHD oscillations in solar coronal loops. Our goal consists in demonstration of the fact that quasi-periodic transverse motions of a coronal magnetic loop, which contains a source of microwave emission, may be connected with a specific modulation of radiation intensity received by a remote observer.

Speaking about other possible mechanisms (besides of the microwave radiation source large-scale oscillatory motion), which may cause a quasi-periodic modulation of the non-thermal electron gyro-synchrotron radiation, it is necessary to mention that a quasi-periodically varying flow of the non-thermal electrons may also result in oscillations of intensity of microwave radiation. Generation of energetic electrons usually is believed to be associated with the processes of magnetic reconnection during solar flares (Miller et al., 1997). There are also theories which suggest acceleration of particles by the inductive and charge separation electric fields, build in course of the continuous motion of solar large-scale coronal magnetic structures (Khodachenko et al., 2003; Zaitsev & Stepanov, 1992). Besides of that, particle acceleration in a collapsing magnetic trap (Karlický & Kosugi, 2004), in the MHD turbulence (LaRosa & Moore, 1993; Miller et al., 1996), and in shocks (Cargill et al., 1988; Holman & Pesses, 1983) are addressed as secondary possible mechanisms for energetic particle production. An extended review of particle acceleration processes in solar flares was recently published by Aschwanden (2002). In most of these cases the typical periods are shorter than those of the VLF transverse oscillations of coronal loops. On the other hand, there are also models in which VLF large-scale oscillations of coronal loops control the process of generation of energetic particles after the impulsive phase of a flare (Nakariakov et al., 2006). This case, however, deserves a special study, which appears beyond the scope of the present paper. Our analysis here is based on the traditional scenario, according to which the non-thermal particles, produced during a flare in particle acceleration regions (e.g., sites of magnetic reconnection or the area of separatrix currents), are injected into oscillating loops.

2. Peculiarities of VLF modulations of microwave radiation related to large-scale transverse motions of the radiating sources in the oscillating coronal loops

Even taking into account the typical ranges of the modulation frequencies of microwave radiation (mentioned above), it is usually difficult to identify the modulation mechanism acting in each particular case. Indeed, having in mind only Equation (1) one cannot say for sure if the observed modulation is due to *a)* the electric current oscillations in the radiation source or *b)* the large-scale oscillatory motion of the loop. However, regarding the last modulation mechanism, an attention should be paid to the fact that large-scale transverse oscillatory motion of a coronal loop is accompanied by the periodic stress of magnetic field created in the loop, especially near its footpoints, during each inclination, i.e. two times per oscillation cycle. This means that the magnetic field strength fluctuates during the oscillatory motion of the loop with a half-period $P_{\rm osc}/2$ of the loop oscillation. Therefore, according to Equation (1), for a transverse oscillating loop, a properly located observer, in addition to the modulation caused by the emission diagram motion at the main oscillation frequency $\nu_0 = 1/P_{\rm osc}$, may see in some cases the modulation at the double frequency of the loop oscillation $2\nu_0$, as well as weak higher order harmonics caused by the non-linearity of Equation (1). However, as it will be shown below the relative amplitude of the higher-order harmonics, i.e., with numbers > 2 is rather low, and in the most cases only the first two harmonic frequencies can be detected. The domination of the main and double-frequency harmonics in the spectrum is caused by implicit presence of these frequencies in the signal, according to the above described character of the radiation modulating factors. Therefore, the presence of the "modulation pairs" in the low-frequency spectra, i.e., the lines which can be associated with the main and double frequency of the loop oscillation (ν_0 and $2\nu_0$) may indicate about a transverse oscillatory dynamics of the loop.

Formation of the "modulation pairs" and their higher-order harmonic companions in multi-line dynamic spectrum of the VLF modulation of microwave radiation emitted from a transverse oscillating coronal loop may be illustrated with a simple model. Let's suppose that the loop undergoes oscillations in the direction transverse to the loop plane as shown in Figure 1. The loop inclination relative to the vertical direction varies as $\alpha(t) = \alpha_0 \sin(2\pi\nu_0 t)$, where α_0 and ν_0 are the angular amplitude and frequency of the loop oscillations, respectively. Assuming that this loop, when oriented vertically, is seen by a remote observer at the angle Θ_0, we get that in course of the loop oscillation the viewing angle changes as $\Theta(t) = \Theta_0 - \alpha(t)$.

Irrespectively of the nature of a coronal loop oscillation, the important feature of the large-scale transverse motion of the loop, consists in an oscillating magnetic stress, created in the loop during its quasi-periodic inclinations. Assuming the local transverse disturbance of the magnetic field relative its initial vertical direction to be δB, we find that the total disturbed magnetic field is $\delta B/\cos\alpha(t)$. For sufficiently small $\alpha(t)$ the following approximation can be used: $1/(\cos\alpha(t)) \approx 1/(1-\alpha(t)^2)^{1/2} \approx 1 + (1/2)\alpha(t)^2$. This means that the disturbed magnetic field in the loop varies in time as $B(t) \approx \delta B(1 + (1/2)\alpha(t)^2)$. Therefore, for the assumed above sinusoidal character of $\alpha(t)$, we finally obtain that the local magnetic field in a transverse oscillating magnetic loop may be approximated as $B(t) \propto (1 + 0.5\alpha_0^2 \sin^2(2\pi\nu_0 t))$. Substitution of the expressions for $\Theta(t)$ and $B(t)$ into (1) enables to construct a modeling signal for the varying intensity of microwave radiation emitted from a transverse oscillating magnetic loop. The examples of dynamic spectra of this signal obtained with $\alpha_0 = \pi/6$ and $\delta = 5$ for different viewing angles $\Theta_0 = \pi/2; \pi/3; \pi/4; \pi/6$ are shown in Figure 2.

Dynamical spectra of the modeling signal in Figure 2 demonstrate several important features, typical for the radiation emitted from a microwave source located in a transverse large-scale oscillating magnetic loop, which may be observed in the solar microwave emission. In particular, for the most of the viewing angles (except of $\Theta_0 = \pi/2$) the dynamic spectra contain well pronounced "modulation pairs", e.g. the lines at the main ν_0 and double $2\nu_0$ frequency of the oscillation. Besides of that, sometimes also a weak third harmonic at $3\nu_0$ may be observed, which appears due to essentially non-sinusoidal (non-harmonic) character of the signal resulted from the joint action of two modulating factors: quasi-periodic magnetic stress and emission diagram motion. In a special case of $\Theta_0 = \pi/2$, the absence of the main frequency component is caused by a "symmetrizing" (in this case) of the varying angular part of the emission intensity. This results in a situation when the diagram motion and magnetic stress factors work synchronously.

As it can be seen in Figure 2, only the first two harmonics have high enough amplitudes. In particular, the spectral amplitude of third harmonic in the cases with $\Theta_0 = \pi/3; \pi/4; \pi/6$, never exceeds 25% of the main frequency component, whereas the second harmonic constitutes usually about 65% of the last. Therefore, the detection of harmonics with numbers higher than 2 in a natural signal, will be in the most cases difficult due to the noise contamination. The presence of modulation pairs in the VLF spectra of solar microwave radiation may be considered as an imprint of a transverse kink-type motion of a loop containing the radiation source. This feature may be used for the indirect identification of candidates for transverse oscillating coronal loops by finding specific modulation lines in the VLF dynamic spectra of microwave radiation. However, the exact detection of transverse

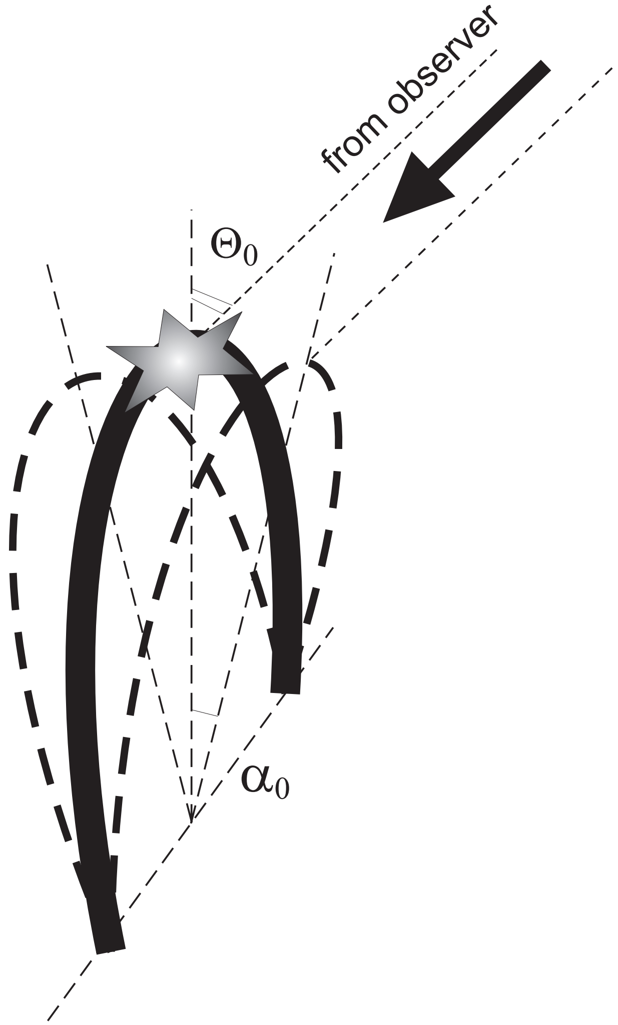


Fig. 1. Schematic view of an oscillating loop which contains a microwave radiation source. Θ_0 is the direction to a remote observer and α_0 is the angular amplitude of the loop oscillations.

motion of the radiating loops by the dynamical spectra of microwave emission (with the exclusion of other mechanisms which may also generate higher spectral harmonics) requires quantitative study of the measured radio signal and superimposing these results with the precise calculation of the radiation from the loop, taking into account the loop position relative observer.

Several real observational examples are considered below, for which VLF modulations of microwave radiation could be associated with the observed in EUV post-flare oscillating coronal loops.

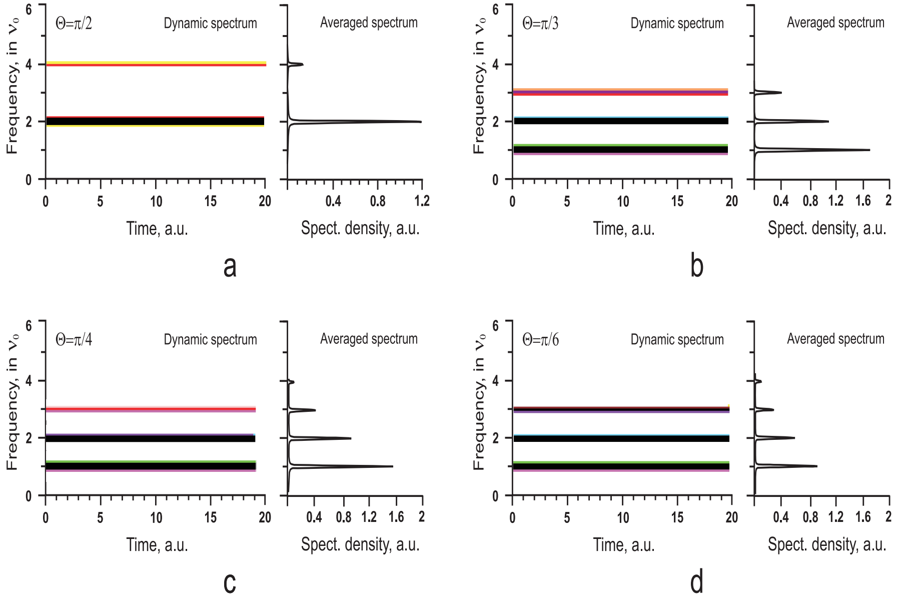


Fig. 2. Dynamical spectra of the modeling signal with $\alpha_0 = \pi/6$ and $\delta = 5$ demonstrating the peculiarities of the microwave radiation VLF modulations produced due to large-scale transverse oscillations of a coronal loop containing the radiating source. Different directions to an observer Θ_0 (viewing angles) are considered: (a) $\pi/2$; (b) $\pi/3$; (c) $\pi/4$; (d) $\pi/6$.

3. Data preparation and analysis methods

Differently to the idealized infinite in time analytical modelling signal considered in section 2, the natural radio emission received from a solar active region with oscillating coronal loop(s) is essentially time-dependent. It usually begins with an impulsive phase of a solar flare and has duration of only several periods of decaying oscillations of the loop(s). Besides of that, a signal from a particular oscillating loop is quite often strongly contaminated by interfering signals from neighboring loops in the active region of interest, as well as by the radiations emitted from other solar active regions. This complicates the task of detection and diagnostics of coronal magnetic loop oscillations in microwaves and requires special data preparation procedures with consequent application of high spectral and time resolution data analysis techniques.

Since the analyzed data appear in a form of discrete counts of the signal intensity, it is natural that digital methods are applied for their processing. A basic specifics of the digital methods consists in certain limitation of dynamical range of the resulting spectra which may lead to the loss of relatively weak and short-time, but important parts of the whole spectra-temporal picture of the studied phenomenon. To avoid of that, the analyzed data pass certain pre-processing preparation which (depending on particular case and task) may include the following procedures: 1) Subtraction of a constant component of a signal, or the signal average; 2) subtraction of a slow (as compared to the analyzed oscillations) major trend of the signal; 3) slow polynomial approximation of the analyzed data with the consequent subtraction of the approximating signal; 4) signal "normalization" (will be described below);

and 5) digital filtration of contaminating components. Altogether, the "subtraction" methods 1-3 enhance visibility of weaker fluctuations of the radiation, enabling better analysis of their spectral and temporal characteristics. Note, that the digital filtration (e.g., the method 5) with an appropriate filter(s) may efficiently substitute all the "subtraction" methods. However, the inertness of a filter results in some smoothing of the radio emission fluctuations of interest. Besides of that, information about the intensity of the radiation fluctuations is lost during the frequency filtration.

The filtration algorithm consists in the application of a Gaussian window to the analyzed signal spectrum, obtained with the discrete Fourier transform (DFT), with the consequent performance of the reverse DFT. The Gaussian window for the low frequency (LF), high frequency (HF), band-pass (BP), and band-lock (BL) filters, respectively, is determined by the following expressions:

$$\begin{aligned}
W_{LF}(k) &= \exp\left\{-\frac{k^2}{2}\cdot\left(\frac{f_1}{f_s}\cdot d\cdot N\right)^{-2}\right\}, && k=0,...,N/2;\\
W_{HF}(k) &= 1-\exp\left\{-\frac{k^2}{2}\cdot\left(\frac{f_1}{f_s}\cdot d\cdot N\right)^{-2}\right\}, && k=0,...,N/2;\\
W_{BP}(k) &= \exp\left\{-\frac{1}{2}\cdot\left(\frac{k}{N}-\frac{f_0}{f_s}\right)^{2}\cdot\left(\frac{\Delta f}{f_s}\cdot d\right)^{-2}\right\}, && k=0,...,N/2;\\
W_{BL}(k) &= 1-\exp\left\{-\frac{1}{2}\cdot\left(\frac{k}{N}-\frac{f_0}{f_s}\right)^{2}\cdot\left(\frac{\Delta f}{f_s}\cdot d\right)^{-2}\right\}, && k=0,...,N/2.
\end{aligned}\tag{2}$$

Here N is the number of counts in the analyzed signal, f_s is the frequency of discretization, d is decimation coefficient (Marple, 1986), f_1 is the cut-off frequency for the LF and HF filters, f_0 and Δf, are the central frequency and the band, respectively, for the BP and BL filters.

Let's consider now the "normalization" method. It is based on a treatment of an analytical signal $z(n)$ (Marple, 1986):

$$z(n) = s(n) + is_H(n) = M(n)\exp(i\Psi(n)), \tag{3}$$

where $s(n)$ and $s_H(n)$ are the analyzed digital signal and its Hilbert conjugate, respectively, and n is the count number. The parameters $M(n)$ and $\Psi(n)$ are the module and phase of the analytical signal, which are defined as the following:

$$\begin{aligned}
M(n) &= \sqrt{s(n)^2 + s_H(n)^2},\\
\Psi(n) &= \arctan\left[\frac{s_H(n)}{s(n)}\right].
\end{aligned}\tag{4}$$

The "normalization" procedure consists in division of the analytical signal $z(n)$ on $M(n)$. By this, the amplitudes of all spectral components of the analyzed process $s(n)$ become to be equalized. Note, that due to the orthogonality of functions $s(n)$ and $s_H(n)$, the value of function $M(n)$ never becomes zero. The "normalization" method is especially efficient for the analysis of non-stationary modulation processes. It enables to detect and to follow the variations of an "instantaneous" frequency of an oscillatory component of radiation.

An important role in the present work belongs also to the method, used for the detection of quasi-periodic features in the time records of the solar microwave radiation intensity. It consists in application of an original data analysis algorithm (Shkelev et al., 2002; Zaitsev et al., 2001b) made as a combination of the "sliding window" Fourier (SWF) transform technique and the nonlinear Wigner-Ville (WV) method (Cohen, 1989; Ville, 1948; Wigner, 1932). Below we outline the idea of this data analysis algorithm and its main features.

The classical Fourier transform enables the analysis of a given signal in terms of separate spectral frequency components. It is applied for the study of relative distribution of energy between the spectral components in the case of sufficiently long (ideally infinite) duration of the analyzed signal. However such energetic spectrum does not provide an information on a time when each particular spectral component appears. Possible improvement of the classical Fourier transform method in that respect consists in its application within a certain interval of time Δt (so called "window") and in a consequent shift of this "window" along the time axis. This approach became a standard method for the analysis of non-stationary signals. Further generalization of SWF transform method leads to the wavelet analysis, where effect of the "window" is produced by means of a certain mother-wavelet function. Wavelet transform enables judging about energy distribution over the time and frequency of an analyzed signal. Nowadays this method is also widely used for the analysis of non-stationary and impulsive signals. Its efficiency is however strongly dependent on parameters of the applied mother-wavelet function, which needs to be specially selected and adjusted to the type of particular analyzed signal.

Recently one more spectral analysis method has been applied in astrophysics. This method is based on the Wigner-Ville (WV) transform

$$P(f,t) = \int_{-\infty}^{\infty} z\left(t+\frac{\tau}{2}\right) z^{*}\left(t-\frac{\tau}{2}\right) e^{-i2\pi f\tau} d\tau, \tag{5}$$

where $z(t) = s(t) + is_H(t)$ is an analytic signal, made of the analyzed sample of real signal $s(t)$, and its Hilbert conjugate $s_H(t)$. Function $P(f,t)$ gives distribution of the signal energy over frequency f and time, and may be visualized in the form of a dynamical spectrum of the signal. According to its definition (5), WV transform may be also interpreted as Fourier image (relative the shifted time) of the local autocorrelation function for the analytical signal $z(t)$.

Since the time t appears explicitly among the arguments of the WV spectrum $P(f,t)$, this method is most efficient for high-resolution spectra-temporal analysis of non-stationary signals with varying spectra, such as quasi-harmonic signals with a changing frequency, or varying impulsive signals. In these cases SWF transform and wavelet methods are less efficient, because the averaging over the analysis window (or over the wavelet) results in a decrease of spectral density of the signal components, varying within the corresponding time intervals. At the same time, the non-linearity and non-locality of WV method cause appearance of artificial inter-modulation spectral components at combination frequencies (artifacts) and may result also in suppression of weak spectral components of the signal by its more intense or noisy parts (Cohen, 1989; Shkelev et al., 2002). To compensate the drawbacks of SWF and WV data analysis methods, when they are used separately, and to keep their strong features, the methods were combined in a proper way in the SWF-WV algorithm, which uses various types of signal processing and filtration (with variable shape and size of the analysis windows) in order to eliminate possible artifacts and to provide high spectral and temporal resolution (Kislyakov et al., 2011; Shkelev et al., 2002).

To avoid the appearance of spurious spectra caused by the signal edge effects in the case of analysis of real finite in time signal samples, the so called "weighting" functions with smoothed edges are used. In the present study, the SWF spectrum S_k of a discrete signal

s_n consisting of N counts is calculated by a discrete Fourier transform (DFT) (Allen & Mills, 2004; Marple, 1986):

$$S_k = \sum_{n=0}^{N-1} \mathrm{wnd}(n) s_n \exp\left\{-i\frac{2\pi nk}{N}\right\}, \quad k = 0, 1, ..., N-1 \tag{6}$$

with a sliding window $\mathrm{wnd}(n)$. The following window functions are used (see also Pollock (1999)):

$$\mathrm{wnd}_1(n) = \begin{cases} 1, & 0 \le n < N \\ 0, & n \ge N \end{cases} \qquad \text{– a rectangular window}$$

$$\mathrm{wnd}_2(n) = \begin{cases} \cos^2\left(\frac{\pi \cdot n}{2N}\right), & 0 \le n < N/2 \\ 0, & n \ge N/2 \end{cases} \qquad \text{– Henning's window}$$

$$\mathrm{wnd}_3(n) = \begin{cases} 1 - 6 \cdot \left(\frac{2 \cdot n}{N}\right)^2 + 6 \cdot \left(\frac{2 \cdot n}{N}\right)^3, & 0 \le n < N/4 \\ 2 \cdot \left(1 - \frac{2 \cdot n}{N}\right)^3, & N/4 \le n < N/2 \\ 0, & n \ge N/2 \end{cases} \qquad \text{– Parsen's window.} \tag{7}$$

The advantage of this method consists in its high performance speed, especially if the standard algorithms of fast Fourier transform (FFT) are used in the calculations (Allen & Mills, 2004; Marple, 1986). On the other hand, frequency resolution of SWF is reverse proportional to the number of signal counts in the applied window. Therefore the size of window should be sufficiently large. This in its turn decreases the temporal resolution of the method. In practice, the choice of particular type and width of the window is determined by dynamical features of the analyzed signal.

An algorithm of the discrete WV transform is determined by the following expression:

$$P_{mk} = P(m\Delta t, k\Delta f) = 2\Delta t \sum_{n=0}^{2N-1} \left[z_{m+n} z_{m-n}^* \exp\left\{-i\frac{\pi nk}{N}\right\}\right], \ \{m, k\} = 0, 1, 2, ..., 2N, \tag{8}$$

where z_n and z_n^* are discrete values of the analyzed complex analytical signal made, as determined above, of the real discrete signal and its Hilbert conjugate; Δt is a period of discretization, and $\Delta f = 1/(4N\Delta t)$ is the frequency step. Note, that WV transform results in real values only in the case of continuous functions integrated in infinite limits (like in (5)). The discrete WV transform (8) yields a complex function P_{mk}, which is called as "pseudo-WV transform" (Cohen, 1989). In that respect in practice only a module of P_{mk} or its real part are considered.

In course of the comparative study performed in Kislyakov et al. (2011); Shkelev et al. (2002) it has been shown that WV data analysis technique enables higher spectral and temporal resolution than that of the SWF. In Shkelev et al. (2002) the efficiency of WV and SWF methods was checked with various test signals, made as combinations of impulsive and quasi-harmonic processes. By this, along with the meaningful spectrum, WV transform generated specific artificial spectral features (due to the non-linearity of the method). It has been shown in that respect that superimposing of the higher resolution WV spectra with those of lower resolution provided by SWF, may help to identify and to exclude these artificial spectral features form the consideration. Altogether, combined with each other the described above SWF and WV methods provide an efficient data analysis algorithm characterized by high sensitivity, high spectral and temporal resolution, and ability to detect complex multi-signal modulations in the analyzed data records, enabling the dynamical spectra of these modulations. For successful operation of the SWF-WV algorithm, the sampling cadence of analyzed data series should provide sufficient number of the data points, e.g. $\geq 10,000$ points per realization. The length of the analyzed data series should be consistent with the time scales of considered dynamic phenomena, i.e. the duration of an analyzed data set should include at least several periods of the modulating oscillatory component. The SWF-WV method appears the most efficient for the study of signals with non-stationary complex modulations. In such cases the traditional Fourier transform and wavelet methods are less efficient. This feature of the algorithm has been, in particular, used to distinguish between the modulations, possibly caused by the large-scale transverse quasi-periodic motion of the loops, which are the subject of the present study, and the modulations with frequency drifts related with the electric current LCR-oscillations in the loops.

For the visualization of the whole variety of the detected modulations, so called averaged spectral density plots are produced along with the dynamic spectra by the SWF-WV algorithm. These plots are obtained by averaging of multiple instantaneous cuts of the dynamic spectrum taken at given moments of time, so that short-living modulation features also become clearly seen among the longer lasting modulation lines. These both types of spectra (dynamic and averaged) enable the detection of the large-scale transverse oscillatory dynamics of the radiating coronal magnetic loops.

The universality of SWF-WV algorithm resulted in its successful application in different branches of space physics. The algorithm was used for the diagnostics of intrinsic physical and dynamical conditions in the stellar and planetary systems, solar/stellar winds, as well as in solar and planetary radiation sources and planetary environments (Khodachenko et al., 2006; Kislyakov et al., 2006; Panchenko et al., 2009; Zaitsev et al., 2003; 2004). Nowadays, the link to SWF-WV data analysis algorithm is available for the scientific community via the on-line catalogue of models and data analysis tools (http://europlanet-jra3.oeaw.ac.at/catalogue/), developed within the JRA3-EMDAF (European modelling and data analysis facilities) activity (http://europlanet-jra3.oeaw.ac.at/) of the European FP7 research infrastructure project Europlanet-RI.

4. Diagnostics of large-scale oscillations of coronal loops by the analysis of VLF modulations of microwave radiation

5. Instrumentation and modulations detection capabilities

We analyze the VLF ($<$ 0.01 Hz) modulations of solar microwave bursts recorded in Metsähovi Radio Observatory (Finland) with the 14-m and 1.8 m radio telescope antenna at 37 GHz and 11.7 GHz, respectively. The key selection criterion for the analyzed microwave data was their synchronism with the oscillating loops observed in extreme ultraviolet (EUV) by TRACE (Aschwanden et al., 2002). The width of the antenna beam pattern of the Metsähovi radio telescope at 37 GHz is 2.4′, the sensitivity of the receiver is about 0.1 sfu (10^{-23} W m^{-2} Hz^{-1}), and time resolution, $0.05 \div 0.1$ s. Therefore, at 37 GHz the spatial resolution of the radio telescope is sufficient for identification of an active region that contains a radiating source. This enables to analyze microwave radiation emitted directly from the region, imaged in EUV (e.g., observed by TRACE), and to perform the comparison of the radiation features and dynamics of coronal loops. At 11.7 GHz the radiation is collected from the whole solar disk and the position of the radiating source cannot be resolved. However, even in this case, by comparison with observations in other wavelengths and timing of the events, it is usually possible to identify the microwave radiation features related to the energy release and dynamic phenomena in particular active regions.

Variations of the background magnetic field in a radiating source may cause not only the amplitude modulation of microwave radio emission from solar active regions (according (1)), but also could result in a certain frequency modulation (due to the dependence of electron gyro-frequency ν_B on the magnetic field). The estimated width of the gyro-frequency variation interval due to this effect is from several tens to several hundreds MHz. At the same time, the bandwidth of the receiver at Metsähovi is much larger than this interval and a possible frequency modulation of the microwave radio emission cannot be resolved. Thus, one can detect only the effects of varying magnetic field, manifested in the intensity modulation of the microwave signal, due to (1).

5.1 Observations and interpretation

In this subsection, we show how the analysis of long-periodic modulations of solar microwave radiation accompanying the explosive events on the Sun may be used to obtain information on the details of large-scale dynamics of the coronal loops associated with flares and the overall structure of solar active regions. By this, the primary focus here is made on VLF modulations of the radiation, emitted from active regions where TRACE observed in EUV at the same time the large-scale oscillations of coronal loops. Careful check of the solar microwave radio emission records available at Metsähovi revealed several events which coincide in time with the EUV observations of oscillating coronal loops. Below these events are considered in details.

Figure 3c shows the intensity profile and dynamic spectrum of VLF modulations of microwave emission from the active region AR8910 on the limb (see Figure 3a,b) where a group of oscillating loops (Figure 4a) was observed by TRACE after M2.0 flare on 2000-Mar-23, at 11:30-12:00 UT (Aschwanden et al., 2002). These observations were performed at 37 GHz, and the spatial resolution of the Metsähovi radio telescope was sufficient to resolve the radiating source in the active region AR8910.

A remarkable feature of the VLF modulation dynamic spectrum in Fig. 3c,d and the averaged spectral density plot in Fig. 3d consists in the presence of several "modulation pairs". These are the modulations (a) at 1.7 mHz ($\sim$ 10 min) and 3.4 mHz ($\sim$ 4.9 min); (b) at 6.0 mHz ($\sim$ 2.8 min) and 12.0 mHz ($\sim$ 1.4 min), as well as (c) at 7.8 mHz ($\sim$ 2.1 min) and 15.6 mHz ($\sim$ 64 s). According to the considerations in Section 2, these "modulation pairs" could indicate the transverse oscillating loops with the periods, corresponding to the main frequencies of the pairs, i.e. $\sim$ 10 min; $\sim$ 2.8 min, and $\sim$ 2.1 min for the cases (a), (b), and (c), respectively. By this, the first "modulation pair" (case (a)) as a signature of the loop transverse oscillation with a period $\sim$ 10, fits quite well the results of TRACE observations, which found the oscillating loop with approximately the same period (615 s) (Aschwanden et al., 2002; Schrijver et al., 2002). As it can be seen from the dynamical spectrum in Figure 3c, the spectral resolution of the analysis performed in this particular case was about 0.3 mHz. By this, the frequency 1.62 mHz corresponding to the detected TRACE period of 615 s is definitely within the frequency

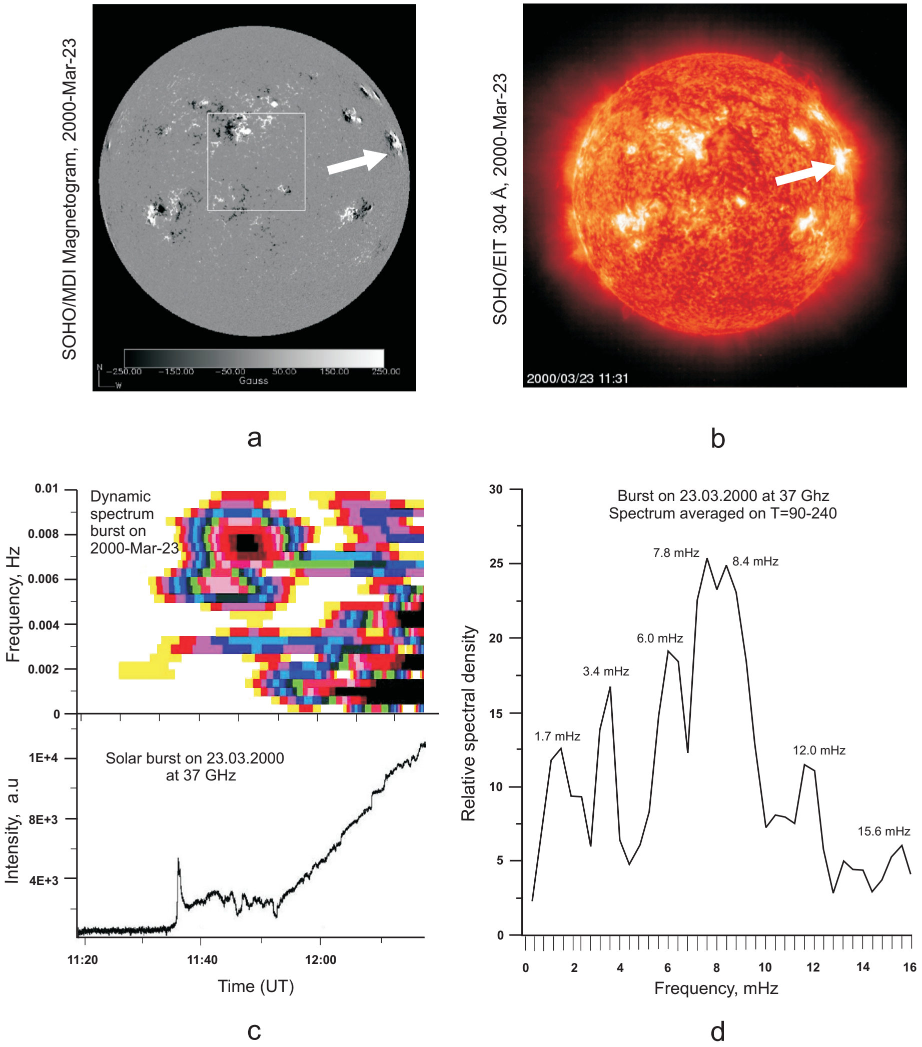


Fig. 3. (a) SOHO/MDI Magnetogram of the Sun on 2000-Mar-23, white arrow points at the active region AR8910; (b) The Sun image in 304 Å on 2000-Mar-23 from SOHO/EIT, white arrow points at the active region AR8910; (c) Intensity profile and corresponding VLF modulation dynamic spectrum of the microwave radiation, recorded from the active region AR8910 on 2000-Mar-23, at 11:30-12:00; Color codes the dynamic spectral relative intensity (arbitrary units), more dark features correspond to stronger (better pronounced) modulations; (d) averaged spectral density of the VLF modulation.

interval 1.7 ± 0.3 mHz of the modulation feature revealed by the analysis of the microwave radiation.

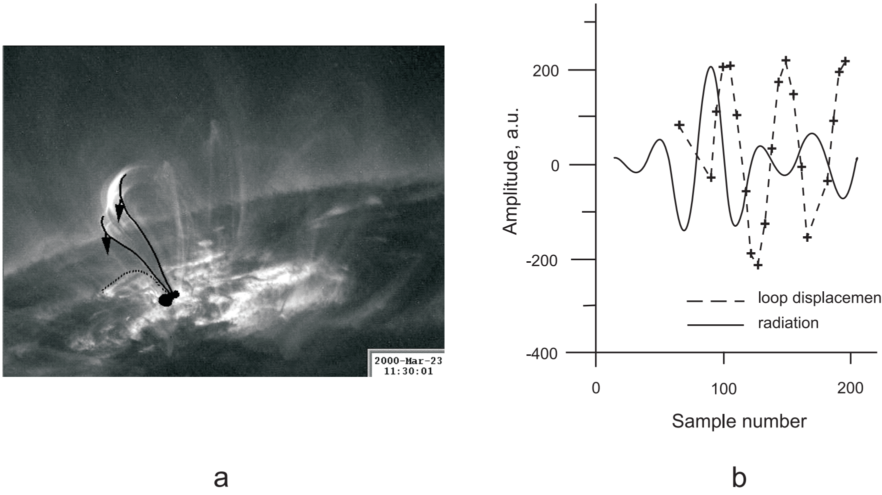


Fig. 4. (a) Transverse oscillating coronal loops observed by TRACE in the active region AR8910 after an M2.0 flare on 2000-Mar-23 at 11:30-12:00 UT (Aschwanden et al., 2002); (b) Phase comparison of the $\sim$ 10 min modulation component of the microwave emission on 2000-Mar-23 and the amplitude of the corresponding 615 s oscillation of the TRACE loop.

A higher level of the second harmonic in the modulation pair 1.7 mHz ($\sim$ 10 min) and 3.4 mHz ($\sim$ 4.9 min) is very likely due to the fact that in this particular case two different mechanisms, modulating the microwave emission of the loop, are by chance manifested simultaneously. The first mechanism is that considered in this paper, which is connected with a large-scale transverse oscillation of the loop. The second mechanism is due to the parametric resonance between 5-min velocity oscillations in the solar photosphere and acoustic oscillations of coronal magnetic loop modulating the microwave emission (Zaitsev et al., 2008). The effect consists in simultaneous excitation in the loop, which occasionally appeared to have a resonant frequency close to 10 min, of oscillations with periods $\sim$5 min, $\sim$10 min, and $\sim$3 min, which correspond to the 5-min pumping frequency of the photospheric convection velocity oscillations, subharmonic, and the first upper frequency of the parametric resonance, respectively (Zaitsev et al., 2008; Zaitsev & Kislyakov, 2006).

It makes no sense to search in TRACE data for the signatures of other oscillating loops (cases (b) and (c)), indicated by the VLF modulations of the microwave radiation during the 2000-Mar-23 event, since with the usual 40 s image sampling cadence of TRACE and the 4-point resolution limit of the instrument (Aschwanden et al., 2002) the fastest detectable by TRACE period is about 3 min. The remaining short-periodic "non-paired" modulation feature at 8.4 mHz ($\sim$ 1.9 min) may also be a part of a "modulation pair", of which the second harmonic counterpart could not be resolved in the VLF spectrum due to the strong contamination of the analyzed microwave signal. Such weak higher harmonic components may be as well a signature of another oscillatory process, which is unrelated to the large-scale transverse motion of loops, e.g., a sausage-type MHD wave excited in a loop. Detailed analysis of this special case remains however beyond the scope of the present study.

An additional confirmation of the fact that $\sim$ 10 min modulation of the microwave radiation emitted from the active region AR8910 on 2000-Mar-23, is connected with the transverse oscillatory motion of a coronal loop, shown in Fig. 4a, and that it is in very likely related to the motion of the emission diagram pattern, comes from the graphs in Fig. 4b. This figure enables phase comparison for the transverse motion of the loop (observed with TRACE Aschwanden et al. (2002)) and the filtered 1.7 mHz ($\sim$ 10 min) component of the radiation. The last characterizes temporal behaviour of the radio emission received from the oscillating loop. The shifted phase ($\sim \pi$) means that the maxima and minima of the radiation part controlled by the loop motion correspond to the specific orientations of the loop and are connected with a certain direction of the emission diagram relative to the observer.

The microwave burst on 2001-Sep-07 represents another example of manifestation of the coronal loop transverse oscillations in modulation of solar radio emissions. The burst was produced during M-flare activity at 15:30 UT in the active region AR9601, close to the solar disc center (see Figure 5a,b), where TRACE observed a group of oscillating loops, immediately after the flare (Aschwanden et al., 2002). The corresponding microwave radiation record was made at 11.7 GHz with the Metsähovi radio telescope. At this frequency, the Metsähovi antenna cannot resolve the position of a radiating source, and the emission from the whole solar disk contributed to the analyzed microwave intensity profile. At the same time, as can be seen in Fig. 5c, which presents the analyzed microwave radiation record with the burst and its VLF modulation dynamic spectrum, the spectral features related to the processes in the flaring active region AR9601 can easily be identified by timing of the event. In particular, the dynamic spectrum of VLF modulations of the microwave radiation exhibits several lines, which start simultaneously with the impulsive phase of the flare (at 15:30 UT). These may be the signatures of different oscillating loops excited by the flare. By this, most of the oscillations (e.g. the dynamic spectrum lines) decay at the time intervals $>$ 20 min. Unfortunately for this particular event it is impossible to determine exact duration of each of these decaying modulations because the available microwave radiation record does not cover the end of the event. As it can be seen in Fig. 5c, some of the dynamic spectrum lines continue beyond the analyzed record time frame.

Three "modulation pairs" can be identified in the dynamic and averaged spectra in Fig. 5c,d: (a) 1.8 mHz ($\sim$ 9.2 min) and 3.6 mHz ($\sim$ 4.6 min); (b) 2.7 mHz ($\sim$ 6.2 min) and 5.4 mHz ($\sim$ 3.1 min); as well as (c) 4.3 mHz ($\sim$ 3.8 min) and 8.6 mHz ($\sim$ 1.9 min), which may be the signatures of transverse oscillating loops with periods $\sim$ 9.2 min, $\sim$ 6.2 min, and $\sim$ 3.8 min, respectively. We note that the loop periods in the cases (a) and (b) are consistent with the 6-10 min oscillating loops observed with TRACE (Aschwanden et al., 2002), whereas the shorter period oscillation (case (c)) cannot be resolved by TRACE because of the relatively long image sampling cadence. The modulation at 5.4 mHz ($\sim$ 3.1 min) may also be a weak third harmonic produced by the 9.2 min oscillating loop. If this is true, then the line at 2.7 mHz ($\sim$ 6.2 min) will have no a pair-companion, and one should exclude the possibility of the $\sim$ 6.2 min transverse oscillating loop. A "non-paired" weak modulation feature at 6.4 mHz ($\sim$ 2.6 min) may be a signature of an oscillating loop with not resolved second harmonic. At the same time, as for the 2000-Mar-23 event, weak short-period harmonics may be the signatures of oscillatory processes that are unrelated to the transverse motion of a loop, but caused only by a changing magnetic field in the radiating source.

The strong modulation line at 0.6 mHz ($\sim$ 27.7 min) should be considered separately from all other modulations mentioned above. The dynamical spectrum in Fig. 5c, as well as a separate study of VLF modulations of the microwave radiation recorded before the faring burst at 15:30 UT, reveal the presence of the $\sim$ 27.7 min component also before the flare. In view of the fact

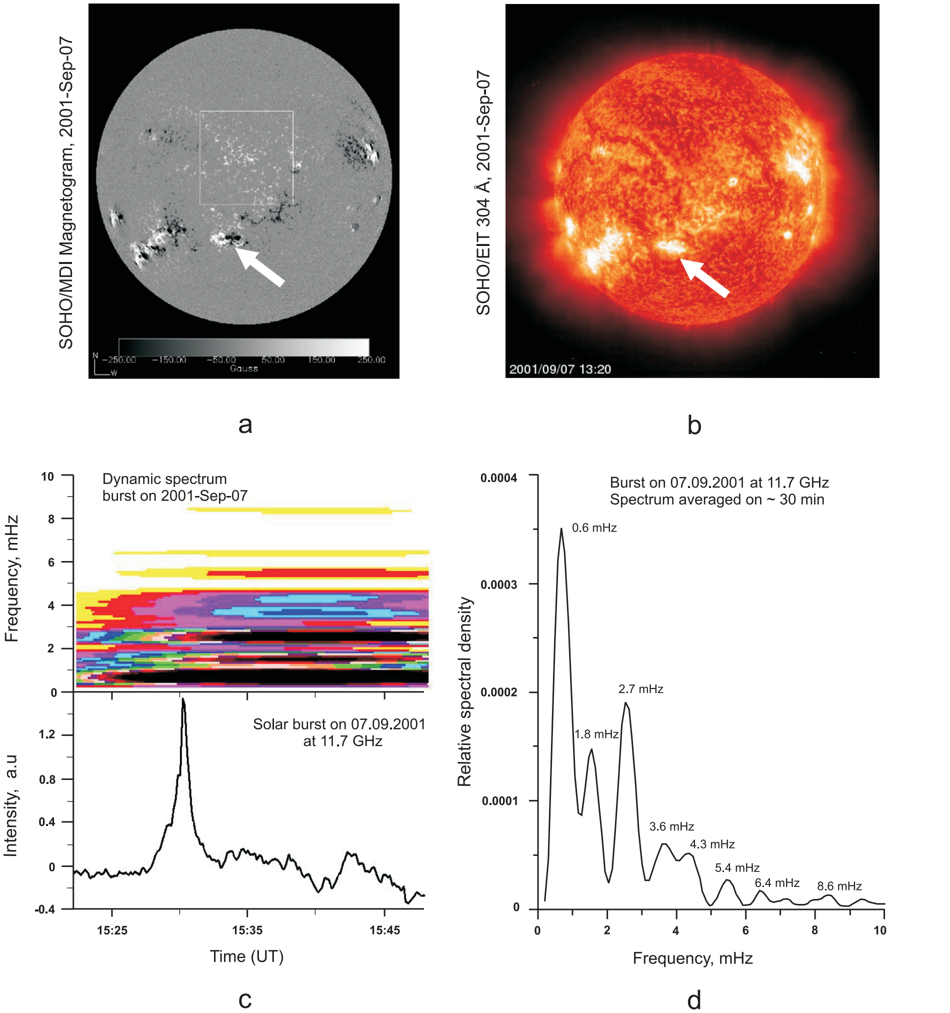


Fig. 5. (a) SOHO/MDI Magnetogram of the Sun on 2001-Sep-07, white arrow points at the active region AR9601; (b) The Sun image in 304 Å on 2001-Sep-07 from SOHO/EIT, white arrow points at the active region AR9601; (c) Intensity profile and corresponding VLF modulation dynamic spectrum of the microwave burst on 2001-Sep-07, at 15:30-15:50 associated with an M-flare in the active region AR9601; Color codes the dynamic spectral relative intensity (arbitrary units), more dark features correspond to stronger (better pronounced) modulations; (d) averaged spectral density of the VLF modulation.

that the analyzed microwave emission (at 11.7 GHz) was received from the whole solar disk, the $\sim$ 27.7 min modulated part of radiation very likely originates in another active region. It may also be connected with a kind of global solar seismology process.

The last example of possible manifestation of transverse oscillations of coronal loops in microwaves which we present here, is the event on 2001-Sep-15, when TRACE observed oscillating loops, associated with M flare at 11:23 UT in the active region AR9608, close to the limb (see Figure 6a,b). Similar to the case of the microwave burst on 2001-Sep-07, the event on 2001-Sep-15 was observed at 11.7 GHz, and the microwave radiation source in AR9608 was not resolved by the Metsähovi antenna. Thus, the emission from the whole solar disk contributed to the analyzed microwave intensity profile. However, as for the burst on 2001-Sep-07, all the spectral features related to the flaring active region AR9608 can be identified by event timing. As it can be seen in Fig. 6c, the dynamic spectrum of VLF modulations of the microwave radiation consists of several lines, most of which begin simultaneously with the impulsive phase of the flare at 11:23 UT. These lines may be associated with oscillatory processes in the active region loops triggered by the flare. All the post-flare oscillations decay at the time intervals from $\sim$ 20 min up to $\sim$ 1 hour.

Fig. 6d shows the averaged spectral density of the VLF modulations of the microwave burst on 2001-Sep-15. At least one "modulation pair": 1.3 mHz ($\sim$ 12.8 min) and 2.6 mHz ($\sim$ 6.4 min) can be identified among the detected modulation lines. It is very likely connected with a transverse oscillating loop having the period $\sim$ 12.8 min. This result agrees with the reported TRACE observations of the oscillating loop in the active region AR9608 with a period $\sim 12 - 15$ min. Other detected in the microwave record on 2001-Sep-15 (see Fig. 6c,d), more short-periodic modulations at 3.8 mHz ($\sim$ 4.4 min) and 5.2 mHz ($\sim$ 3.2 min) could be higher-order harmonics produced by the 12.8 min oscillating loop, or the modulations associated with oscillatory processes not connected with the transverse motion of loops. They may also be the signatures of oscillatory processes in small loops, which cannot be seen in the TRACE EUV movies due to the limitations of the operated observing mode. Additionally should be mentioned a relatively weak line at 0.7 mHz ($\sim$ 23.8 min). Its possible second harmonic could contribute to the broad line near 1.3 mHz, which is identified as a main frequency line in another modulation pair. If that is the case, then this may be a signature of another transverse oscillating loop with the period $\sim$ 23.8 min. Verwichte et al. (2010) reported recently detection of such an oscillating loop related with the considered flaring event in the the active region AR9608.

A special remark deserves also the long lasting ultra-low-frequency (ULF) modulation at 0.3 mHz ($\sim$ 56 min), clearly visible in both, the dynamic and averaged, spectra of the long-periodic modulations of solar microwave radiation on 2001-Sep-15 in Figs. 6c,d. Similarly to the 27 min line in the case of the 2001-Sep-07 burst, this ULF modulation appears before the flaring burst and lasts much longer than all other modulation lines in the spectrum (Fig. 6c). Therefore, it cannot be related to the flare in the active region AR9608 and consequent post-flare dynamics of coronal loops. This modulation feature is probably connected with the solar seismology processes, or a slow dynamics of another active region. As an additional argument in support of the solar global i.e., helioseismic nature of the ULF ($\nu_0 < 0.6$ mHz, i.e. $\tau > 30$ min) modulations may be the fact that these modulations are usually detected in the radiation records made at 11.7 GHz when the radio emission from the whole solar disk contributes to the analyzed microwave intensity profile. The radiation emitted from the spatially resolved separate active regions, for example, at 37 GHz with Metsähovi radio telescope, does not exhibit any ULF modulation features. In more details the ULF modulations of solar microwave radiation are considered in Kislyakova et al. (2011).

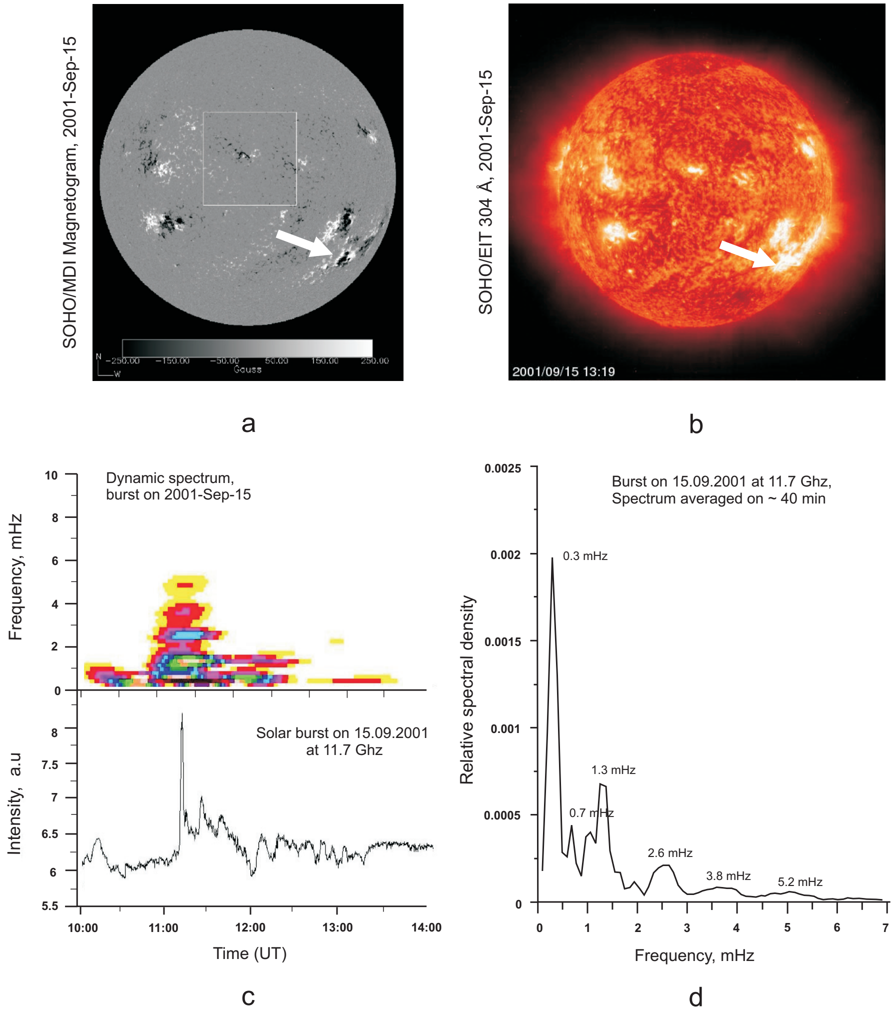


Fig. 6. (a) SOHO/MDI Magnetogram of the Sun on 2001-Sep-15, white arrow points at the active region AR9608; (b) The Sun image in 304 Å on 2001-Sep-15 from SOHO/EIT, white arrow points at the active region AR9608; (c) Intensity profile and corresponding VLF modulation dynamic spectrum of the microwave burst on 2001-Sep-15, at 11:23-12:15 associated with an M-flare in the active region AR9608; Color codes the dynamic spectral relative intensity (arbitrary units), more dark features correspond to stronger (better pronounced) modulations; (d) averaged spectral density of the VLF modulation.

5.2 On the magnetic field variations, estimated from VLF spectra

It is impossible, using the available data, to perform an exact calculation for the amplitude of magnetic field variations. That is because the analyzed microwave signals were recorded in relative units without calibration to the radiation intensity scale. At the same time, taking into account known values of the maximal intensity measured during the radio bursts and the lower limit of sensitivity of the Metsähovi receiver, the intensity modulation amplitude ΔI_ν can be estimated roughly from the obtained VLF spectra, by their comparison with the spectra of specially created modelling signals (Khodachenko et al., 2005). For the considered in the paper events this estimation gives the value of the relative variation of intensity $\Delta I_\nu / I_\nu^0 \sim 10^{-2} \div 10^{-1}$. Assuming, as an upper rough limit, that this variation of intensity is connected only with the varying magnetic field, we come to a relation $\Delta I_\nu / I_\nu^0 = (B/B_0)^\gamma - 1 = ((\Delta B/B_0 + 1)^\gamma - 1)$, where $\Delta B/B_0$ is the relative variation of magnetic field, and $\gamma = -0.22 + 0.9\delta$ is the power index of magnetic field in the equation (1). For $2 < \delta < 7$ one gets that $\gamma = 1.58 \div 6.08$. For the case of small relative variations of the magnetic field we finally obtain $\Delta I_\nu / I_\nu^0 \approx \gamma \Delta B/B_0 = 10^{-2} \div 10^{-1}$, which for $B_0 = 100$G gives the amplitude $\Delta B = (0.16 \div 6.3)$ G.

6. Discussion, conclusions and perspectives

The analysis of VLF modulations of solar microwave bursts presented in this work shows good temporal coincidence of the modulations and their oscillatory parameters with the observed decaying large-scale transverse oscillations of the coronal loops triggered by flares. This indicates about a physical link between the oscillatory motion of the loops and variations of the observed radio emission. As a working hypothesis to take this link into account, a loop with the propagating beams of non-thermal particles which produce microwave emission due to the electron gyro-synchrotron mechanism, has been considered. As pointed out by Schrijver et al. (2002), who considered several cases of transverse oscillations of coronal loops observed with TRACE (including the event on 2000-Mar-23 addressed in the present paper), in almost all cases the oscillating loops lie at, or near, the large-scale separatrices, or near the sites involved in reconnection. These regions may be the sources of non-thermal particles injected into the loops and generating microwave emission there. Moreover, in the case on 2000-Mar-23 the loop oscillation happened in response to a flaring event located at the loop base Schrijver et al. (2002). That could provide a direct input of energetic particles into the loop.

In the most general case, thermal bremsstrahlung mechanism of the radiation should be also considered, besides of the gyro-synchrotron, for the analyzed frequency range of the solar microwave emission. If the last mechanism assumes that there are high energy electrons passing through the magnetic loop, the first one is connected with the radiation of hot plasma heated by the electron beams in the chromospheric footpoints of the loop. A comparative study of contribution of the bremsstrahlung and gyro-synchrotron radiation to the microwave emission of a flare, performed in Urpo et al. (1994), shows that thermal bremsstrahlung is more important for the microwave events that have an intensity of the order of or less than 100 SFU, with the exception of cases when the electron spectrum is sufficiently hard. Therefore, correct interpretation of the microwave radiation source requires consideration of both mechanisms, for example by involving of the hybrid thermal/nonthermal model of the solar flare emission (Holman & Benka, 1992). However, looking at a possibility of an oscillatory behaviour of the microwave radiation source which constitutes the primary subject of the present study, we notice that in the case of the bremsstrahlung mechanism it is possible only for a varying energy deposition into the system, i.e. a varying flow of non-thermal particles heating the loop footpoints. In view of the unclearness of how the post-flare transverse oscillating loop may modulate the source of accelerated electrons, we built our analysis with the assumption that the non-thermal particle population remains more-or-less stable, and all the variations of the observed microwave radiation are connected with the non-thermal gyro-synchrotron part of the radiation, modulated by the large-scale oscillatory motion of the loop. The bremsstrahlung component, even being present in the microwave radiation, does not contribute to the analyzed oscillating part of the emission provided by the gyro-synchrotron mechanism.

Our analysis here is based on the assumption of an optically thin microwave source. In this case the radiation intensity is proportional to emissivity η_ν, given by equation (Dulk, 1985)

$$\eta_\nu \approx 3.3 \times 10^{-24} 10^{\,-0.52\delta} NB(\sin\theta)^{\,-0.43+0.65\delta} \left(\frac{\nu}{\nu_B}\right)^{1.22-0.9\delta}, \tag{9}$$

where ν is radiation frequency, N is the number of electrons per cubic centimeter with energy higher than 10 keV, B is magnetic field, $\nu_B = eB/(2\pi m_e c)$, is the electron-cyclotron frequency, and δ is the electron energy spectrum index. This fact has been used for obtaining our basic equation (1) in Section 1. In the opposite case of an optically thick source, the intensity of microwave radiation is proportional to the effective temperature T_{eff}, i.e., to the ratio of emissivity η_ν and absorption coefficient κ_ν. For the last, Dulk (1985) provides the following expression:

$$\kappa_\nu \approx 1.4 \times 10^{-9} 10^{\,-0.22\delta} (N/B)(\sin\theta)^{\,-0.09+0.72\delta} \left(\frac{\nu}{\nu_B}\right)^{-1.3-0.98\delta}. \tag{10}$$

Therefore, in the case of an optically thick source the dependence of radiation intensity on the varying magnetic field $B(t)$ and direction to the observer $\Theta(t)$ will be different than that given by the equation (1). However, the character of this dependence, i.e. $I_\nu \propto (\sin(\Theta(t)))^k B(t)^l$, where k and l are the numbers depending on the non-thermal electron spectral index δ, will remain the same. Therefore, one may expect similar manifestation of the varying $B(t)$ and $\Theta(t)$ in the modulation of the received microwave emission also in the case of an optically thick source.

Analysis of LF and VLF modulations of solar microwave radiation is a relatively new direction in solar radio astronomy which appears nowadays a subject of a certain interest. VLF variations of solar microwave radiation intensity may be related to slow variations of magnetic field in a radiating source, as well as to large-scale motions of coronal structures containing the radiating source. Joint action of two radiation modulating factors: (i) the quasi-periodic fluctuation of magnetic field and (ii) motion of the radio emission diagram, in the case of a transverse oscillating coronal loop results in essentially non-sinusoidal (non-harmonic) character of the signal received by a remote observer, with strongly pronounced two first harmonics (at the main and double frequency of the oscillation), called here as "modulation pairs". Such specifics of the VLF spectrum has been used for the identification of transverse oscillating loops triggered by flares. The analysis of solar microwave records has been performed with an algorithm based on Sliding Window Fourier transform and Wigner-Ville techniques. This high sensitive algorithm provides, significant spectral and temporal resolution which enable clear detection of VLF "modulation pairs" in the solar flaring microwave radiation (microwave bursts). Comparison of parameters of these "modulation pairs" with the simultaneous TRACE observations in EUV of the corresponding solar active regions enabled to associate some of the paired VLF modulation features with the large-scale transverse oscillations of coronal loops. The "non-paired" features also detected in the VLF modulation dynamic spectra, may be either parts of "modulation pairs" in which the second harmonic cannot be resolved because of the strong contamination of the analyzed signal, or be the signatures of other oscillatory processes (MHD modes) in the loops, unrelated to their large-scale transverse motion, e.g., sausage-type MHD waves.

The presence of "modulation pairs" in the VLF spectra of solar microwave radiation is considered here as an indication of the transverse oscillating coronal loops. However, the exact characterization of the transverse motion of radiating loops by the dynamical spectra of microwave emission needs a quantitative analysis of the measured radio signal and superimposing these results with a precise calculation of the radiation from the loop, taking into account the loop position relative observer. This topic requires a dedicated study. It outlines the general direction for further development of the ideas expressed in the paper.

Besides of that, of certain interest appear the long lasting ULF modulations of solar microwave radiation detected at < 0.6 mHz (> 30 min) in the absence of bursts, i.e. during the periods of quiet Sun. It is remarkable that these modulations are not visible in the emissions from separate active regions, recorded in particular at 37 GHz with Metsähovi radio telescope. But they appear in the integrated radiation received from the whole solar disk at 11.7 GHz. This fact may be considered as an argument in support of the global helioseismic nature of these ULF modulations, which also require further more detailed study.

7. Acknowledgements

This work was supported by the Austrian Fond zur Förderung der wissenschaftlichen Forschung (project P21197-N16). The authors are thankful to the JRA3/EMDAF (European Modelling and Data Analysis Facility; http://europlanet-jra3.oeaw.ac.at) – a subdivision of the European research infrastructure EUROPLANET-RI and EU FP7 project IMPEx (http://impex-fp7.oeaw.ac.at) for the support of their research communication and collaboration visits. The authors would like to thank M.Aschwanden for useful discussions and the data regarding the transverse large-scale dynamics of coronal loops.

8. References

Allen, R.L. & Mills, D.W. (2004). *Signal analysis: time, frequency, scale, and structure*. IEEE Press, Wiley-Interscience, ISBN:0-471-23441-9.

Aschwanden, M.J.; Fletcher, L.; Schrijver, C.J.; Alexander, D. (1999). Coronal Loop Oscillations Observed with the Transition Region and Coronal Explorer. *The Astrophysical Journal*, Vol. 520, No. 2, 880–894, (DOI:10.1086/307502).

Aschwanden, M.J. (2002). Particle acceleration and kinematics in solar flares - A Synthesis of Recent Observations and Theoretical Concepts (Invited Review). *Space Science Review*, Vol. 101, No. 1, 1–227, (DOI:10.1023/A:1019712124366).

Aschwanden, M.J.; DePontieu, B.; Schrijver, C.J.; Title, A. (2002). Transverse Oscillations in Coronal Loops Observed with TRACE II. Measurements of Geometric and Physical Parameters. *Solar Physics*, Vol. 206, No. 1, 99–132, (DOI:10.1023/A:1014916701283).

Cargill, P.J.; Goodrich, C.C.; Vlahos, L. (1988). Collisionless shock formation and the prompt acceleration of solar flare ions. *Astronomy and Astrophysics*, Vol. 189, No. 1-2, Jan., 254–262, ISSN 0004-6361.

Cohen, L. (1989). Time-frequency distributions - A review. *IEEE Proc.*, Vol. 77, July-1989, 941–981, ISSN 0018-9219.

Dulk, G.A. (1985). Radio emission from the sun and stars. IN: *Annual review of astronomy and astrophysics*, Vol. 23(A86-14507 04-90), 169–224, Palo Alto, CA, Annual Reviews, Inc., (DOI:10.1146/annurev.aa.23.090185.001125).

Dulk, G.A. & Marsh, K.A. (1982). Simplified expressions for the gyrosynchrotron radiation from mildly relativistic, nonthermal and thermal electrons. *The Astrophysical Journal*, Vol. 259, Aug-1982, 350–358, (DOI:10.1086/160171).

Hardy, S.J.; Melrose, D.B.; Hudson, H.S. (1998). Observational tests of a double loop model for solar flares. *Publications Astronomical Society of Australia*, Vol. 15, No. 3, 318–324.

Holman, G.D. & Benka, S.G. (1992). A hybrid thermal/nonthermal model for the energetic emissions from solar flares. *The Astrophysical Journal*, Vol. 400, No. 2, L79–L82, ISSN 0004-637X, (DOI:10.1086/186654).

Holman, G.D. & Pesses, M.E. (1983). Solar type II radio emission and the shock drift acceleration of electrons. *The Astrophysical Journal*, Vol. 267, No. 15, 837–843, ISSN 0004-637X, (DOI:10.1086/160918).

Karlický, M. & Kosugi, T. (2004). Acceleration and heating processes in a collapsing magnetic trap. *Astronomy and Astrophysics*, V. 419, 1159–1168, (DOI:10.1051/0004-6361:20034323).

Khodachenko, M.L.; Haerendel, G.; Rucker, H.O. (2003). Inductive electromagnetic effects in solar current-carrying magnetic loops. *Astronomy and Astrophysics*, Vol. 401, 721–732, (DOI:10.1051/0004-6361:20030146).

Khodachenko, M.L.; Zaitsev, V.V.; Kislaykov, A.G.; Rucker, H.O.; Urpo, S. (2005). Low-frequency modulations in the solar microwave radiation as a possible indicator of inductive interaction of coronal magnetic loops. *Astronomy and Astrophysics*, Vol. 433, No. 2, 691–699, (DOI:10.1051/0004-6361:20041988).

Khodachenko, M.L.; Rucker, H.O.; Kislaykov, A.G.; Zaitsev, V.V.; Urpo, S. (2006). Microwave Diagnostics of Dynamic Processes and Oscillations in Groups of Solar Coronal Magnetic Loops. *Space Science Reviews*, Vol. 122, No. 1-4, 137–148, (DOI:10.1007/s11214-006-7767-0).

Khodachenko, M.L.; Zaitsev, V.V.; Kislyakov, A.G.; Stepanov, A.V. Equivalent Electric Circuit Models of Coronal Magnetic Loops and Related Oscillatory Phenomena on the Sun. (2009). *Space Science Reviews*, Vol. 149, No. 1-4, 83–117, (DOI:10.1007/s11214-009-9538-1).

Khodachenko, M.L.; Kislyakova, K.; Zaqarashvili, T.V.; Kislyakov, A.G.; Panchenko, M.; Zaitsev, V.V.; Arkhypov, O.V.; Rucker, H.O. (2011). Possible manifestation of large-scale transverse oscillations of coronal loops in solar microwave emission. *Astronomy and Astrophysics*, Vol. 525, CiteID A105, (DOI:10.1051/0004-6361/201014860).

Kislyakov, A. G.; Zaitsev, V. V.; Stepanov, A. V.; Urpo, S. (2006). On the Possible Connection between Photospheric 5-Min Oscillation and Solar Flare Microwave Emission. *Solar Physics*, Vol. 233, No. 1, 89–106, (DOI:10.1007/s11207-006-2850-y).

Kislyakova, K.G.; Zaitsev, V.V.; Urpo, S.; Riehokainen, A. (2011). Long-period oscillations of the solar microwave emission. *Astronomy Reports*, Vol. 55, No. 3, 275–283, (DOI:10.1134/S1063772911030036).

Kislyakov, A.G.; Shkelev, E.I.; Lupov, S.Y.; Kislyakova, K.G. (2011). Parameters of astrophysical objects according to their electromagnetic emission intensity modulation data. I. Observational data processing algorithms. *Bulletin of the Nizhny Novgorod State University*, No.2(1), 46–54, (in russian).

LaRosa, T.N. & Moore, R.L. (1993). A Mechanism for Bulk Energization in the Impulsive Phase of Solar Flares: MHD Turbulent Cascade. *The Astrophysical Journal*, Vol. 418, 912–918, (DOI:10.1086/173448).

Leka, K.D.; Canfield, R.C.; McClymont, A.N.; van Driel-Gesztelyi, L. (1996). Evidence for Current-carrying Emerging Flux. *The Astrophysical Journal*, Vol. 462, 547–560, (DOI:10.1086/177171).

Marple, S. L., Jr. (1986). *Digital spectral analysis with applications*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, ISBN:0-132-14149-3.

Miller, J.A.; LaRosa, T.N.; Moore, R.L. (1996). Stochastic Electron Acceleration by Cascading Fast Mode Waves in Impulsive Solar Flares. *The Astrophysical Journal*, Vol. 461, 445–464, (DOI:10.1086/177072).

Miller, J.A.; Cargill, P.J.; Emslie, A.G.; Holman, G.D.; Dennis, B.R.; LaRosa, T.N.; Winglee, R.M.; Benka, S.G.; Tsuneta, S. (1997). Critical issues for understanding particle acceleration in impulsive solar flares. *Journal of Geophysical Research*, Vol. 102, No. A7, 14631–14660, (DOI:10.1029/97JA00976).

Moreton, G.E. & Severny, A.B. (1968). Magnetic Fields and Flares in the Region CMP 20 September 1963. *Solar Physics*, Vol. 3, No. 2, 282–297, (DOI:10.1007/BF00155163).

Nakariakov, V. M.; Foullon, C.; Verwichte, E.; Young, N. P. (2006). Quasi-periodic modulation of solar and stellar flaring emission by magnetohydrodynamic oscillations in a nearby loop. *Astronomy and Astrophysics*, Vol. 452, No. 1, 343–346, (DOI:10.1051/0004-6361:20054608).

Panchenko, M.; Khodachenko, M. L.; Kislyakov, A. G.; Rucker, H. O.; Hanasz, J.; Kaiser, M. L.; Bale, S. D.; Lamy, L.; Cecconi, B.; Zarka, P.; Goetz, K. (2009). Daily variations of auroral kilometric radiation observed by STEREO, *Geophysical Research Letters*, Vol. 36, No. 6, CiteID L06102, (DOI:10.1029/2008GL037042).

Pollock, D.S.G. (1999). *A handbook of time-series analysis, signal processing and dynamics*. London: Academic Press, ISBN:0-12-560990-6.

Schrijver, C. J.; Title, A. M.; Berger, T. E.; Fletcher, L.; Hurlburt, N. E.; Nightingale, R. W.; Shine, R. A.; Tarbell, T. D.; Wolfson, J.; Golub, L.; Bookbinder, J. A.; Deluca, E. E.; McMullen, R. A.; Warren, H. P.; Kankelborg, C. C.; Handy, B. N.; de Pontieu, B. (1999). A new view of the solar outer atmosphere by the Transition Region and Coronal Explorer. *Solar Physics*, Vol. 187, No. 2, 261–302, (DOI:10.1023/A:1005194519642).

Schrijver, C.J.; Aschwanden, M.J.; Title, A.M. (2002). Transverse oscillations in coronal loops observed with TRACE I. An Overview of Events, Movies, and a Discussion of Common Properties and Required Conditions. *Solar Physics*, Vol. 206, No. 1, 69–98, (DOI:10.1023/A:1014957715396).

Schrijver, C.J. & Brown, D.S. (2000). Oscillations in the Magnetic Field of the Solar Corona in Response to Flares near the Photosphere. *The Astrophysical Journal*, Vol. 537, No. 1, L69–L72, (DOI:10.1086/312753).

Shkelev, E.I.; Kislyakov, A.G.; Lupov, S.Y. (2002). Methods of decreasing of a cross-modulation effects in the Wigner-Ville distribution. *Izv. Vyss. Uchebn. Zaved., Ser. Radiofiz.* (transl. as *Radiophys. Quantum Electron.*), Vol. 45, No. 5, 433-442.

Spangler, S.R. (2007). A Technique for Measuring Electrical Currents in the Solar Corona. *The Astrophysical Journal*, Vol. 670, No. 1, 841–848, (DOI:10.1086/521995).

Tan, B.; Ji, H.; Huang, G.; Zhou, T.; Song, Q.; Huang, Y. (2006). Evolution of Electric Currents Associated with Two M-Class Flares. *Solar Physics*, Vol. 239, No. 1-2, 137–148, (DOI:10.1007/s11207-006-0120-7).

Urpo, S.; Bakhareva, N.M.; Zaitsev, V.V.; Stepanov, A.V. (1994). Comparison of mm-wave and X-ray diagnostics of flare plasma. *Solar Physics*, Vol. 154, No. 2, 317–334, ISSN 0038-0938, (DOI:10.1007/BF00681102).

Verwichte, E.; Foullon, C.; Van Doorsselaere, T. (2010). Spatial Seismology of a Large Coronal Loop Arcade from TRACE and EIT Observations of its Transverse Oscillations. *The Astrophysical Journal*, Vol. 717, No. 1, 458–467, (DOI:10.1088/0004-637X/717/1/458).

Ville, J. (1948). *Câbles et Transm.*, Vol. 2A, 61.

Wigner, E. (1932). On the Quantum Correction For Thermodynamic Equilibrium. *Phys. Rev.*, Vol. 40, No. 5, 749–759, (DOI:10.1103/PhysRev.40.749).

Zaitsev, V.V. & Stepanov, A.V. (1992). Towards the circuit theory of solar flares. *Solar Physics*, Vol. 139, No. 2, 343–356, ISSN 0038-0938, (DOI:10.1007/BF00159158).

Zaitsev, V.V.; Stepanov, A.V.; Urpo, S.; Pohjolainen, S. (1998). LRC-circuit analog of current-carrying magnetic loop: diagnostics of electric parameters. *Astronomy and Astrophysics*, Vol. 337, 887–896.

Zaitsev, V.V.; Kislyakov, A.G.; Urpo, S.; Shkelev E.I. (2001a). Observational evidence for energy accumulation and dissipation in coronal magnetic loops. *Izv. Vyssh. Uchebn. Zaved., Ser. Radiofiz.* (transl. as *Radiophys. & Quant. Electronics*), Vol. 44, No. 9, 697–709, (DOI:0033-8443/01/4409-0697$25.00).

Zaitsev, V.V.; Kislyakov, A.G.; Stepanov, A.V.; Urpo, S.; Shkelev E.I. (2001b). Low-frequency pulsations of coronal magnetic loops. *Izv. Vyssh. Uchebn. Zaved., Ser. Radiofiz.* (transl. as *Radiophys. & Quant. Electronics*), Vol. 44, No. 1-2, 36–52, (DOI:0033-8443/01/441-2-0036$25.00).

Zaitsev, V.V.; Kislyakov, A.G.; Urpo, S.; Stepanov, A.V.; Shkelev E.I. (2003). Spectral-Temporal Evolution of Low-Frequency Pulsations in the Microwave Radiation of Solar Flares. *Astronomy Reports*, Vol. 47, No. 10, 873–882, (DOI:10.1134/1.1618999).

Zaitsev, V.V.; Kislyakov, A.G.; Stepanov, A.V.; Kliem, B.; Fürst, E. (2004). Pulsating Microwave Emission from the Star AD Leo. *Astronomy Letters*, Vol. 30, 319–324, (DOI:10.1134/1.1738154).

Zaitsev, V. V.; Kislyakov, A. G.; Kislyakova, K. G. (2008). Parametric resonance in the solar corona. *Cosmic Research*, Vol. 46, No. 4, 301–308, (DOI:10.1134/S0010952508040035).

Zaitsev, V. V. & Kislyakov, A. G. (2006). Parametric excitation of acoustic oscillations in closed coronal magnetic loops. *Astronomy Reports*, Vol. 50, No. 10, 823–833, (DOI:10.1134/S1063772906100076).

Part 2

Medical Applications

Spectral Analysis of Heart Rate Variability in Women

Ester da Silva[1,2], Ana Cristina S. Rebelo[1], Nayara Y. Tamburús[2], Mariana R. Salviati[2], Marcio Clementino S. Santos[3] and Roberta S. Zuttin[2]

[1]*Federal University of São Carlos, São Carlos, SP,*
[2]*College of Health Sciences, Methodist University of Piracicaba, Piracicaba, SP,*
[3]*Pará State University, Belém, PA,*
Brazil

1. Introduction

This chapter discusses heart rate variability (HRV) to understand autonomic mechanisms and the use of linear analysis tools for frequency domain measures of HRV and spectral analysis by fast Fourier transform (FFT), and describes some results found in women.

Heart activity is largely modulated by the autonomic nervous system (ANS), which promotes rapid adjustments in the cardiovascular system during different stimuli (i.e., physical exercise, mental stress and postural change) (Hainsworth, 1998). HRV is a non-invasive measure used to analyze the influence of the autonomic nervous system on the heart, providing information about both sympathetic and parasympathetic contributions to consecutive heart rate (HR) oscillations. It has been proposed that a decrease in HRV is a powerful predictor of morbidity and mortality resulting from arrhythmic complications. HRV decreases with age (Catai et al., 2002; Melo et al., 2005) as a consequence of parasympathetic reduction and predominance of sympathetic modulation (Lipsitz et al., 1990; Longo & Correia, 1995; Akselrod, 1995).

The tool most commonly used in the frequency domain is spectral analysis, which consists of decomposing the HR variation in a given period into its fundamental oscillatory components, defining them by their frequency and amplitude. One of the mathematical algorithms most commonly used to determine the number, frequency and amplitude of these components is the FFT. The sum of all the components constitutes the so-called total power spectral density. Spectral analysis involves three distinct spectral components: 1) very low-frequency (VLF) fluctuations related to the renin-angiotensin system and thermoregulation; 2) low-frequency (LF) fluctuations related to the sympathetic and parasympathetic nervous systems and to baroreflex activity; and 3) high-frequency (HF) fluctuations associated with vagal activity (Longo & Correia, 1995; Task Force, 1996). The sympathovagal balance can also be expressed by the LF/HF ratio. Based on this analysis, it is possible to observe the predominance of one component over the other and the relationship between them, reflecting the autonomic modulation of the heart in the control of HR.

Given the importance of the autonomic nervous system to cardiovascular health, several analytical measures, grouped into linear and non-linear methods, can be used to assess HRV. The ECG is recorded with the subject in a steady state (when rhythms are stationary) for a sufficiently long period to determine events occurring within the frequencies of interest. R-R interval spectral power is calculated from this series of intervals using an autoregressive algorithm, which yields center frequencies and absolute power of component fluctuations (Task Force, 1996).

Sympathovagal balance (in dimensionless units) is simply the ratio of absolute LF to absolute HF power, or the LF/HF ratio. The literature on sympathovagal balance is replete with disclaimers that spectral power reflects fluctuations, not absolute levels of autonomic nerve traffic (Akselrod 1995). If mathematical manipulation of R-R interval spectral power is to inspire confidence as a robust, reliable metric, it must be grounded solidly on physiological principles. It must stand on its own and calculations of sympathovagal balance may obscure rather than illuminate human physiology and pathophysiology (Eckberg 1997).

This chapter discusses the measurement and analysis of HRV, as well as results of data for women and the relationship between aging and hormonal changes (oral contraceptives and hormone replacement therapy), which contribute to modifications of the autonomic control of the heart. Each item will be discussed in a separate subitem of this chapter.

2. Measurement of heart rate variability

An electrocardiogram and HR data were obtained using a one-channel heart monitor (MINISCOPE II Instramed, Porto Alegre, RS, Brazil) and processed using a Lab. PC+ analog-to-digital converter (Lab PC + / National Instruments, Co., Austin, TX, USA) acting as an interface between the heart rate monitor and a microcomputer. The ECG signal was recorded in real time after analog-to-digital conversion at a sampling rate of 500 Hz and the R-R intervals (ms) were calculated on a beat-to-beat basis using specific software (Silva et al., 1994). To evaluate the effect of body position on the HR response and its variability, R-R intervals were recorded over a 15-min period under resting conditions with the subjects in the supine and sitting positions, respectively.

HR and R-R intervals (RRI) can be obtained in real time, beat-by-beat, using the ECG and specific software (Silva et al., 1994). First, a visual inspection of RRI (ms) distribution obtained during 900s of collection at rest in the supine condition was carried out in order to eliminate the fragments containing spikes, which resulted in an interval with higher stability of ECG RRI tracing (Task Force, 1996).

3. Spectral analysis

Linear HRV can be assessed by frequency domains. For the frequency domain, a spectral analysis was performed by FFT applied to a single window after the subtraction of a linear trend, at the R-R intervals previously chosen. The power spectral components were obtained at low (LF: 0.04 to 0.15 Hz) and high (HF: 0.15 to 0.4 Hz) frequencies, in absolute units (ms^2), and the normalized units (nu) were computed by dividing the absolute power of a given LF or HF component (ms^2) by the total power minus very low frequency (0.003-0.04 Hz) power and then multiplying this ratio by 100. Since the LF band is modulated by both sympathetic and parasympathetic activity and the HF band is correlated with vagal cardiac control, the LF/HF ratio was calculated to determine the sympathovagal balance (Task Force, 1996). Sympathovagal balance is the ratio between LF and respiratory-frequency powers. Based on this analysis, it is possible to determine the predominance of one component over the other and the relationship between them, reflecting the autonomic modulation of the heart in the control of heart rate.

Figure 1, which is based on an autoregressive model, illustrates the HRV power spectra at rest in the supine and sitting positions of a representative subject in different conditions.

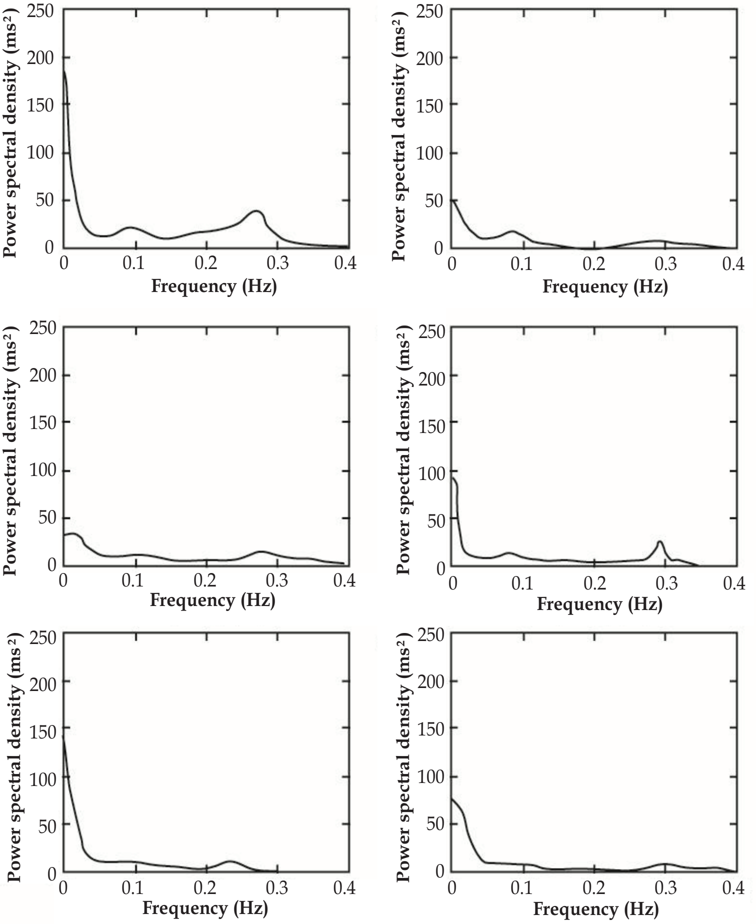


Fig. 1. Power spectral density of heart rate variability of a representative subject from the groups of young women (A and B), and postmenopausal women undergoing (C and D) and not undergoing (E and F) estrogen therapy, obtained at rest in the supine and sitting positions, respectively. Spectral components are shown as LF (0.04 to 0.15 Hz), HF (0.15 to 0.4 Hz) and VLF (below 0.04 Hz). (adapted by Neves et al., 2007)

Figure 2 illustrates the analysis of the RRI (ms) of a volunteer at rest in the supine position, using the power spectrum of the autoregressive model for a better view of the spectral components. Three spectral frequency bands were obtained: 1) VLF, corresponding to frequencies varying from 0 to 0.04 Hz; LF, corresponding to the interval of 0.04 Hz to 0.15 Hz; and AF, corresponding to the interval of 0.15 Hz to 0.40Hz. The LF and HF components are expressed in normalized units (UN) which correspond to the percentage of the total power spectrum subtracted from the VLF component. These components were also expressed as the ratio between the absolute areas of low and high frequency (LF/HF ratio), which is indicative of the vagosympathetic equilibrium. Figure 2 illustrates the temporal series of the RRI corresponding to the 256 values of analysis selected previously.

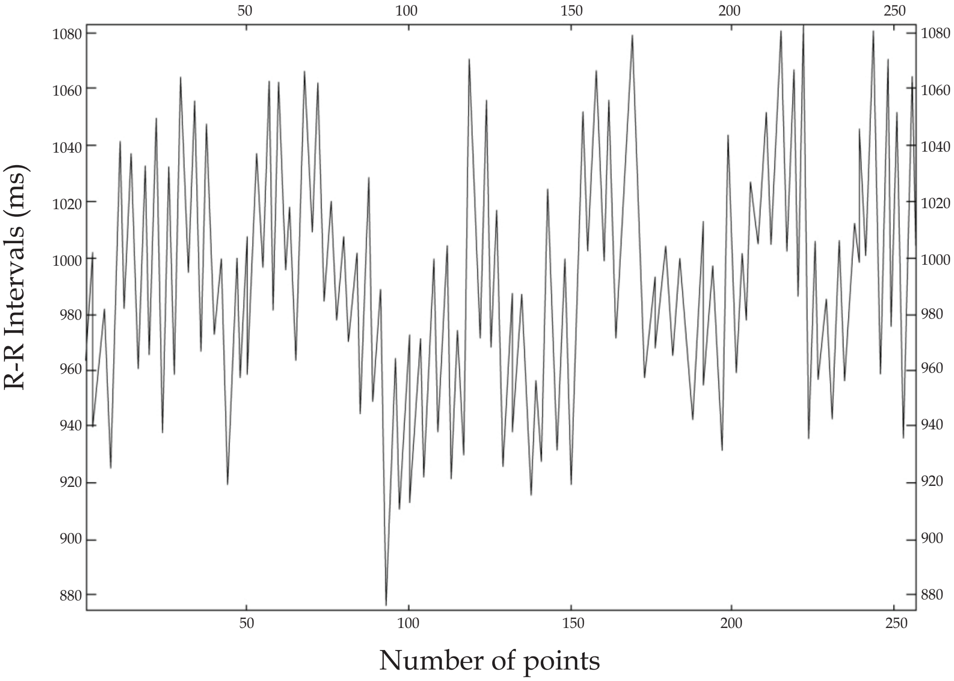


Fig. 2. Temporal series of 256 values of R-R intervals (ms) of a volunteer in the supine position

Because the HR presents fluctuations that are, in large part, periodic, a continuous electrocardiographic record over short or long periods (24 h) and a subsequent graphical representation of the normal R-R intervals over time (tachogram) produce a complex undulatory phenomenon that can be decomposed into simpler waves through mathematical algorithms, such as the FFT or the autoregressive model. This process, called spectral analysis, enables the electrocardiographic signal from the temporal series (tachogram) to be decomposed into its different frequency components, i.e., into so-called frequency bands. It should be noted that frequency refers to the number of times a given phenomenon (e.g., a sound wave, electric current or any form of cyclic wave) occurs over time. Normally, the frequency unit employed is Hertz (Hz), which is equivalent to one cycle per second. Figure 3 shows the application of an autoregressive model to view the power spectrum of the analysis of heart rate variability corresponding to these values of a volunteer of this study.

In long records (24 h), the total power is decomposed into four distinct bands: 1) high frequency band (HF), oscillating at a frequency of 0.15 a 0.40 Hz, i.e., 9-24 cycles/min, corresponding to the heart rate variations related to the respiratory cycle (respiratory sinus arrhythmia), which are typically modulated by parasympathetic activity; 2) low frequency or LF band (0.04 to 0.15 Hz or 2.4 to 9 cycles/min), modulated by both sympathetic and parasympathetic activities, with a predominance of sympathetic in some specific situations, and which reflects the oscillations of the baroreceptors system; 3) very low frequency or VLF band (0.003 to 0.04 Hz or 0.2 to 2.4 cycles/min), depending on the thermoregulatory mechanisms and the renin-angiotensin system, which is also regulated by sympathetic and parasympathetic activities; and 4) ultra low frequency or ULF band (< 0.003 Hz or < 0.2 cycles/min), which corresponds to most of the total variance, but whose physiological significance is not yet well defined. This band is influenced by the parasympathetic and sympathetic systems and is obviously absent from short duration records. It appears to be related with the neuroendocrine system, circadian rhythm, and other systems (Task Force, 1996).

A high frequency component equivalent to 0.25 Hz (15 cycles/min = 15 cycles/60 s = 0.25 cycles/s = 0.25 Hz), a low frequency component equivalent to 0.1 Hz (6 cycles/min) and a very low frequency component of 0.016 Hz (1 cycle/min). The combination of these three sine waves generates a complex wave signal that can be compared to the signal obtained when the heart rate is expressed on a temporal graph (tachogram). Moreover, the calculation of the area covered by each frequency band (which is proportional to the square of the amplitude of the original signal and hence, in this case, is expressed in ms^2) enables one to separate the amount of variance (power) ascribed to each frequency. This allows for a more detailed study of the individual participation of each of the divisions of the ANS (sympathetic and parasympathetic) in different physiological and pathological situations, as well as its relationship with the main systems that interfere with HRV (respiratory, vasomotor, thermoregulatory, renin-angiotensin and central nervous systems). In fact, this is the main difference between spectral analysis and time domain analysis, since the latter generally fails to distinguish the dominant rhythms or oscillations that give the heart rate its variability (Task Force, 1996).

Spectral components are usually measured in absolute values of power (ms^2). However, the values of LF and HF can also be expressed in normalized units (nu), which represent the value of each of these components in relation to the total power (TP) minus the VLF component. These values are calculated by means of the following formulas: HF (nu) = HF/(TP – VLF) x 100 and LF (nu) = LF/(TP – VLF) x 100. This minimizes the effects of changes in the VLF range on the other two components with faster frequencies (LF and HF). Another frequently used measure is the LF/HF ratio, which can provide useful information about the balance between the sympathetic and parasympathetic systems. It should also be noted that, because absolute values in ms^2 are highly variable and distributed asymmetrically, they usually require logarithmic transformation (Task Force, 1996).

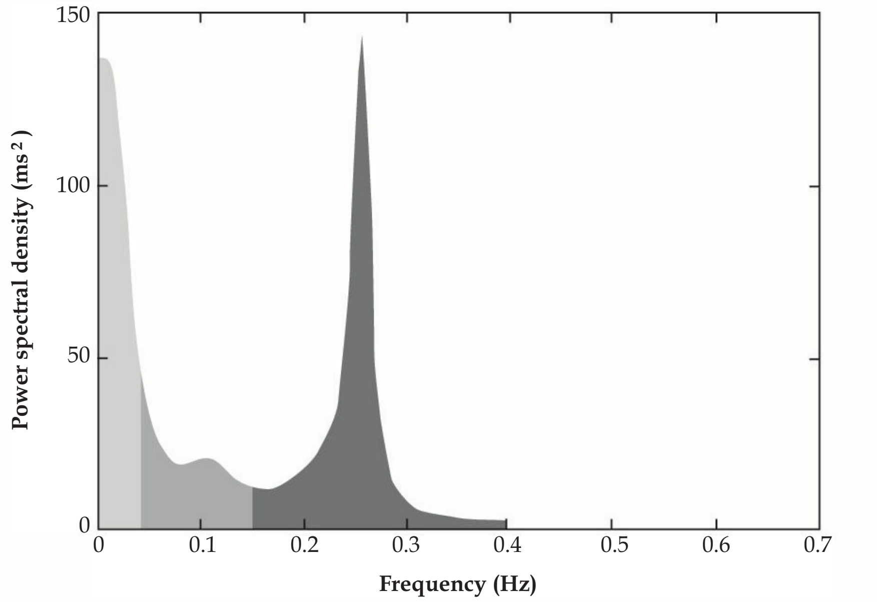


Fig. 3. Power spectrum of the analysis of HRV obtained by applying an autoregressive model to a dataset of 256 values of R-R intervals in the supine position from one of the volunteers of this study, showing the VLF (light gray), LF (medium gray) and HF (dark gray) bands

4. Heart rate variability and oral contraceptives

Third-generation combined oral contraceptives (COCs) containing desogestrel and gestodene (GEST) were introduced to reduce adverse effects such as fluid retention, nausea, headaches, and weight changes (Arangino et al., 1998; Read, 2010). The balance of risks and benefits of COC use varies, depending on patterns of usage and background risk of disease (Hannaford et al., 2010). The repercussions of COCs on cardiac autonomic modulation have not yet been thoroughly investigated. Studies reveal that female sex hormones influence cardiovascular autonomic function (Minson et al., 2000; Neves et al., 2007; Carter et al., 2009). Leicht et al. (2003) reported a positive correlation between circulating estrogen levels and HRV.

Furthermore, the cardioprotective effects of endogenous estrogen through vasodilation and inhibition of blood vessel injuries have been reported (Mendelsohn & Karas, 1999). Low levels of estrogen are associated with a reduction of cardiac autonomic modulation (Moodithaya 2009). Large clinical trials have shown that the long-term use of estrogen in combination with a progestogen may not be beneficial, and could even compromise the efficiency of autonomic HR modulation. Minson et al. (2000) confirmed that COC use can modify baroreflex sensitivity and sympathetic activity. However, Santos et al. (2008) and Schueller et al. (2006) found that COC users and non-users showed similar HRV indices.

Carter et al. (2009) observed no effects of OC use on the sympathetic modulation of the heart during orthostatic stress, nor differences in that regard between the phase of intake of active pills and that of intake of inactive pills. Women with greater physical activity, both users and non-users of OCs, showed a predominance of parasympathetic modulation and presented a greater complexity of pattern distribution and less regularity and predictability of sequential patterns than sedentary groups. Wenner et al. (2006) evaluated amenorrheic and eumenorrheic athletes who were users and non-users of OCs, and observed no influence on cardiac autonomic function. However, other studies suggest that there is a relationship between OC use and autonomic HR modulation, which the authors attribute to changes in vagal peripheral modulation caused by high levels of circulating estrogen (Minson, 2000; Leicht et al., 2003).

Santos et al. (2008) analyzed the autonomic modulation of HR based on frequency domain (LF, HF and LF/HF) indices and found that the use of contraceptives did not affect the results, since they detected no difference among the groups under study. This finding may be attributed to the pharmacological properties of low estrogen/progesterone dosages, as well as to the maintenance of the integrity of the autonomic modulation of HR, since the values found here fall within the range of normality. The results of this study suggest that low estrogen/progesterone dosages do not impair autonomic modulation in the age group under study.

5. Heart rate variability and hormonal therapy

The aging process causes changes in the autonomic modulation of the cardiovascular system, and particularly in HR. The literature reports that parasympathetic activity in the sinus node decreases with age, leading to a reduction in HRV and a greater risk for cardiovascular events (Lipsitz et al., 1990). Structural and functional changes in the blood vessels, in the cardiac conduction system and in the sensitivity of baroreceptors, as well as increased myocardial stiffness, leading to greater force of contraction and reduced ventricular filling, contribute to reduce the functional capacity of the cardiovascular and hemodynamic system (Walsh, 1987). In addition, with increasing age, submaximal physical activity and decline in functional capacity lead to increased physiological stress (Perini et al., 2002).

The incidence of cardiovascular diseases among premenopausal women is low when compared to that of men in the same age group, but increases significantly after this period (Gensini et al., 1996). In several countries, cardiovascular diseases are the major cause of morbidity and mortality among postmenopausal women, representing an important public health problem (Mosca et al., 1997). The increase in the incidence of cardiovascular events among middle-aged women has been associated with the hypoestrogenism typical of this period of women's lives (Greendale et al., 1999).

With regard to autonomic heart function, some studies have demonstrated the harmful effects of hypoestrogenism on HRV. Mercuro et al. (2000) found a reduction in HRV indices, analyzed in the time and frequency domains, after bilateral oophorectomy, i.e., through the interruption of estrogen production, as occurs in menopause. Liu et al. (2003) demonstrated higher values of HRV, analyzed in the time domain, in premenopausal women than in postmenopausal women and men in the same age group, illustrating the importance of estrogens in the autonomic differences brought about by menopause. Similar findings, also analyzed in the time domain, were reported by Brockbank et al. (2000) for premenopausal women compared to women after more than one year of menopause. Davy et al. (1998) reported that young women have higher HRV than menopausal women, and that HRV in both active and sedentary women tends to decline with advancing age. In earlier studies conducted in our laboratory (Ribeiro et al., 2001; Neves et al., 2007), lower levels of HRV in menopausal women compared to young women were also recorded.

To ascertain if a physical training program could promote physiological adaptations and improved sympathovagal balance of the heart, attenuating the deleterious effects of menopause on the cardiovascular system, Sakabe (2007) evaluated 18 sedentary women divided into two groups: Control Group – 10 postmenopausal women (50 to 60 years old) without hormone therapy (HT); and HT Group – 8 postmenopausal women (50 to 60 years old) undergoing HT (estradiol plus levonorgestrel). Both groups were assessed at two different times: before (assessment) and after (reassessment) a 3-month physical training program (PTP). Protocol 1 – to evaluate the autonomic modulation of the HR, the HR was recorded under resting conditions, supine and sitting positions, for 15 minutes in each position. The indices evaluated in Protocol 1 were: mean HR and R-R intervals (RRI), RMSSD index of the RRI, low (LF) and high (HF) frequency bands of the spectral analysis, in normalized units, and LF/HF ratio. It was concluded that hormone replacement therapy did not have a significant effect on HRV.

6. Heart rate variability and menopause

Postmenopausal women have greater sympathetic and less parasympathetic activity than premenopausal women (Brockbank et al., 2000; Earnest et al., 2010). Moreover, Mercuro et al.'s study (2000) reveals the harmful effects of hypoestrogenism on the autonomic modulation of the HR, while other studies have demonstrated numerous evidences that endogenous hormones (estrogen and progesterone) contribute to a cardioprotective phenotype in women (Vitale et al., 2009).

Parasympathetic modulation shifts to a lower range with normal aging. Although parasympathetic modulation is generally higher in women than men, aging reduces the difference between genders, with changes in HRV beginning approximately at menopause (Earnest et al., 2010). Boettger (2010) examined changes in cardiovascular autonomic parameters obtained from short-term recordings over time. The data he collected indicated a lifelong shift in autonomic balance toward sympathetic predominance, starting at the age of 30 years.

Zuttin (2009) evaluated and compared autonomic modulation of the HR at rest in healthy young, premenopausal and postmenopausal women leading a sedentary lifestyle, to verify cardiovascular adjustment in response to postural changes. This investigation involved 113 healthy sedentary women, who were divided into a young group (YG) with an average age of 23±3.4 years (n=40), a premenopausal group (PreMG) aged 36±3.1 years (n= 39), and a postmenopausal group (PostMG) with an average age of 55±4.5 years (n=34).

In the supine position, it was found that the YG presented significantly higher values of the HF index in absolute units (ms^2) and lower LF values (ms^2) and ratio than the PostMG. In addition, the YG and PostMG showed a statistical difference in all the evaluated indices ($p<0.05$), while no difference was found between the PreMG and PostMG groups ($p>0.05$). In a comparison of the YG vs. PreMG and YG vs. PostMG groups in the sitting position, the YG presented significantly higher values for the ratio ($p<0.05$).

With regard to the effect of postural adjustment on the autonomic HR modulation, a comparison of the indices obtained in the supine and sitting positions revealed significant differences ($p<0.05$) in all the indices. On the other hand, the PreMG groups showed a difference in the LF/HF index ($p<0.05$), while the PostMG group showed no significant difference ($p<0.05$).

Having calculated the regression coefficients, it was found that the straight line of the adjusted regression indicates that, as the age of the subjects increases, it is possible to estimate the reduction of the HF index (ms^2). The parameters indicate mainly a reduction of the postural change in parasympathetic modulation. With aging, the adjustment capacity diminishes, as indicated by the delta between the supine and sitting positions.

7. Conclusions

This chapter discussed the measurement and analysis of HRV, as well as results of data for women and the relationship between aging, hormonal changes (oral contraceptives and hormone replacement therapy) which contribute to modifications in the autonomic control of heart rate.

8. Acknowledgments

This chapter was financed by the Brazilian research funding agencies CNPq (National Council for Scientific and Technological Development – Process no. 370448/2007-3) and FAPESP (São Paulo Research Foundation – Process no. 2006/56788-1).

9. References

Akselrod, S. (1995). Components of heart rate variability: Basic Studies. In: *Heart Rate Variability* Malik, M & Camm, AJ, (Ed.), pp. 147-163, Futura Publishing Company, New York.

Arangino, S. et al. (1998). Effect of desogestrel-containing oral contraceptives on vascular reactivity and catecholamine levels. *Contraception*, Vol. 58, pp. 289 – 93, ISSN 1879-0518 (Electronic).

Boettger, M. K. et al. (2010). Influence of Age on Linear and Nonlinear Measures of Autonomic Cardiovascular Modulation. *Annals Noninvasive Electrocardiology*, Vol. 15, No. 2, pp. 165–174., ISSN 1542-474X (Electronic).

Brockbank, C. L. et al. (2000). Heart rate and its variability change after the menopause. *Experimental Physiology*, Vol. 85, No. 3, pp. 327-330, ISSN 1469-445X (Electronic).

Carter, J. B., Banister, E. W. & Blaber, A. P. (2003). The effect of age and gender on heart rate variability after endurance training. *Medicine and Science in Sports and Exercise*, Vol. 35, No. 8, pp. 1333-1340, ISSN 1530-0315 (Electronic).

Catai, A. M. et al. (2002). Effects of aerobic exercise training on heart rate variability during wakefulness and sleep and cardiorespiratory responses of young and middle-aged healthy men. *Brazilian Journal of Medical and Biology Research*, Vol. 35, pp. 741–752, ISSN 1414-431X (Electronic).

Collier, S. R. (2008). Sex differences in the effects of aerobic and anaerobic exercise on blood pressure and arterial stiffness. *Gender Medicine*, Vol.5, No. 2, pp. 115-123, ISSN 1878-7398 (Electronic).

Davy, K. P. et al. (1998). Elevated heart rate variability in physically active young and older adult women. *Clinical Science*, Vol. 94, No. 6, pp. 579-584, ISSN 1470-8736 (Electronic).

Earnest, C. et al. (2010). Autonomic function and change in insulin for exercising postmenopausal women. *Maturitas*, Vol. 65, No. 3, pp. 284–291, ISSN 1873-4111 (Electronic).

Eckberg, M. D. (1997). Sympathovagal Balance: a critical appraisal. *Circulation*, Vol. 96, No. 9, pp. 3224-32, ISSN 1524-4539 (Electronic).

Gensini, G. F. et al. (1996). Menopause and risk of cardiovascular disease. *Thrombosis Research*, Vol. 84, No. 1, pp. 1-19, ISSN 1879-2472 (Electronic).

Greendale, G. A., Lee, N. P. & Arriola, E. R. (1999). The menopause. *Lancet*, Vol. 353, No. 9152, pp. 571-580, ISSN 1474-547X (Electronic).

Hainsworth (1998). Physiology of the cardiac autonomic system. In: *Clinical guide to cardiac autonomic tests,* Malik, M., (Ed.), pp. 51-65. Kluwer Academic Publishers, Dordrecht, Boston, London.

Hannaford, P. C. et al. (2010). Mortality among contraceptive pill users: Cohort evidence from Royal College of General Practitioners' Oral Contraception Study. *British medical journal*, Vol. 340, pp. c927, ISSN 1468-5833 (Electronic).

Leicht, A.S., Hirning, D. A. & Allen, G. D. (2003). Heart rate variability and endogenous sex hormones during the menstrual cycle in young women. *Experimental Physiology*, Vol. 3, pp. 441-446, ISSN 1469-445X (Electronic).

Lipsitz, L. A. et al. (1990). Spectral characteristics of heart rate variability before and during postural tilt. Relations to aging and risk of syncope. *Circulation*, Vol. 81, pp. 1803–1810, ISSN 1524-4539 (Electronic).

Liu, C. C., Kuo, T. B., Yang, C. C. (2003). Effects of estrogen on gender-related autonomic differences in humans. *American Journal of Physiology-Heart and Circulation Physiology*, Vol. 285, No. 5, pp. 2188-2193, ISSN 1522-1539 (Electronic).

Longo, D. F. & Correia, M. J. (1995). Variabilidade da frequência cardíaca. *Revista Portuguesa de Cardiologia,* Vol. 14, pp. 241-262, ISSN 0870-2551.

Melo, et al. (2005). Effects of age and physical activity on the autonomic control of heart rate in healthy men. *Brazilian Journal of Medical and Biological Research*, Vol. 38, n. 9, p. 1331-1338, ISSN 1414-431X (Electronic).

Mendelsohn, M. E. & Karas, R. H. (1999). The protective effects of estrogen on the cardiovascular system. *New England Journal of Medi*cine, Vol. 340, pp. 1801-1811, ISSN 1533-4406 (Electronic).

Mercuro, G. et al. (2000). Evidence of a role of endogenous estrogen in the modulation of autonomic nervous system. *American Journal of Cardiology,* Vol. 85, pp. 787-789, ISSN 1879-1913 (Electronic).

Minson et al. (2000). Sympathetic activity and baroreflex sensitivity in young women taking oral contraceptives. *Circulation*, Vol. 102, pp. 1473 – 1476, ISSN 1524-4539 (Electronic).

Moodithaya, S. S. & Avadhany, S. T. (2009). Comparison of cardiac autonomic activity between pre and post menopausal women using heart rate variability. *Indian Journal of Physiology and Pharmacology*, Vol. 53, pp. 227 – 234, ISSN 0019-5499 (Print).

Mosca L, et al. (1997). Cardiovascular disease in women: a statement for healthcare professionals from the American Heart Association. Writing Group. *Circulation,* Vol. 96, No. 7, pp. 2468-2482, ISSN 1524-4539 (Electronic).

Neves, V. F. C, et al. (2007). Autonomic modulation of heart rate of young and postmenopausal women undergoing estrogen therapy. *Brazilian Journal of Medical and Biological Research*, Vol. 40, pp. 491-499, ISSN 1414-431X (Electronic).

Perini, R. et al. (2002). Aerobic training and cardiovascular responses at rest and during exercise in older men and women. Medicine and Science in Sports and Exercise, Vol. 34, No. 4, pp. 700-708, ISSN 1530-0315 (Electronic).

Read, C. M. (2010). New regimens with combined oral contraceptive pills-moving away from traditional 21/7 cycles. *European Journal of Contraception and Reproductive Health Care*, Vol. 15, No. 2, pp. 32-41, ISSN 1473-0782 (Electronic).

Ribeiro, T. F. et al. (2001). Heart rate variability under resting conditions in postmenopausal and young women. *Brazilian Journal of Medical and Biological Research*, Vol. 34, No. 7, pp. 871-877, ISSN 1414-431X (Electronic).

Sakabe, D. W. (2007). Efeitos do treinamento físico sobre a modulação autonômica da frequência cardíaca e a capacidade aeróbia de mulheres pós-menopausa sem o uso de terapia hormonal, pp. 1-170, Avariable from:
http://www.teses.usp.br/teses/disponiveis/17/17145/tde-21052008-135005/pt-br.php

Santos, M. C. S. et al. (2008). Influence of oral contraceptive use on lipid levels and cardiorespiratory responses among healthy sedentary women. *Brazilian Journal of Physical Therapy*, Vol. 12, pp. 188 – 94, ISSN 1413-3555.

Schueller, P. O. et al. (2006). Effects of synthetic progestagens on autonomic tone, neurohormones and C-reactive protein levels in young healthy females of reproductive age. *International Journal of Cardiology*, Vol. 111, pp. 42 – 48, ISSN 1874-1754 (Electronic).

Silva, E et al. (1994). Design of a computerized system to evaluate the cardiac function during dynamic exercise. *Physics in Medicine & Biology*, Vol. 33, p. 409 abstract.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996). Heart rate variability: standards of measurements, physiological interpretation, and clinical use. *Circulation*, Vol. 93, pp. 1043-1065, ISSN 1524-4539 (Electronic).

Vitale, C., Mendelsohn, M. E. & Rosano, G. M. C. (2009). Gender differences in the cardiovascular effect of sex hormones. *Nature Reviews. Cardiology*, Vol. 6, pp. 532–542.

Walsh, R. A. (1987). Cardiovascular effects of the aging process. *The American Journal of Medicine*, Vol. 82, No,1B, pp. 34-40, ISSN 1555-7162 (Electronic).

Wenner, M. M. et al. (2006). Preserved autonomic function in amenorrheic athletes. *Journal of Applied Physiology*, Vol. 101, pp. 590 –597, ISSN 0021-8987 (Print).

Zuttin, R. Z. (2009). Influência da idade sobre a modulação autonômica da frequência cardíaca e a capacidade aeróbia em mulheres. pp. 1-90, Piracicaba, Brazil, Available from https://www.unimep.br/phpg/bibdig/aluno/visualiza.php?cod=528.

7

Cortical Specification of a Fast Fourier Transform Supports a Convolution Model of Visual Perception

Phillip Sheridan
Griffith University
Australia

1. Introduction

Currently, the full extent of the role Fourier analysis plays in biological vision is unclear. Although we have examples of sensory organs that perform Fourier transforms, e.g. the lens of the eye and the cochlear, to date there is no direct empirical evidence for its implementation in cortical architecture. However, there does exist intriguing theoretical evidence that suggests a role for the Fourier transform in a primate's primary visual cortex (area V1) which emerges from recent developments in our knowledge of contextual modulation. This paper proposes a new Fourier transform and a specification of how this transform has a natural implementation in cortical architecture. The significance of this new Fourier transform and its specification in neural circuitry is that it provides a plausible explanation for previously unexplained observable properties of the primate vision system.

1.0.0.1

The spatial response properties, such as orientation tuning and spatial frequency tuning, of neurons in area V1 have been known for some time (Schiller et al., 1976). For a while, it was generally accepted that these tuning functions of receptive fields are largely context-independent (De Valois et al., 1979). However, later research has demonstrated contextual influences from the region close to the receptive field (Sceniak et al., 2001); (Cavanaugh et al., 2002); (Bair & Movshen, 2004). Moreover, it has been found that this near surround region of a receptive field can modify receptive field responses through suppression (Blakemore & Tobin, 1972) and by cross-orientation facilitation effects (Sillito & Jones, 1996); (Cavanaugh et al., 2002); (Kimura & Ohzawa, 2009). It has also been demonstrated that long-range contextual modulation is as robust a feature of neural function in area V1 as the extensively studied receptive field properties of this area (Lamme, 1995). Since that time, the evidence for long-range contextual modulation continues to grow, e.g. (Zipser et al., 1996); (Lamme et al., 1998); (Lee et al., 1998).

1.0.0.2

Concurrent with this research establishing the empirical evidence for contextual modulation has been research aimed at developing functional models of V1 that are consistent with the empirical evidence. In the early 1980's the concept of convolution was employed by David Marr (Marr & Hildreth, 1980) as a model that accounted for considerable observable properties of the human vision system. Since that time further theoretical and empirical evidence has been mounting that supports such a model. In particular, it has been shown that response properties of neurons in area V1 are modeled by convolution of the input image with a family of Gabor functions (Sanger, 1988). Further research has demonstrated that the upper layers of area V1 are modeled well by a bank of Gabor filters (Grigorescu et al., 2003); (Huang et al., 2008); (Lee & Choe, 2003); (Ursine et al., 2004); (Tang et al., 2007). A related, but alternative, approach to the Gabor response functions to model simple and complex cells of V1 is the use of Gaussian derivatives (Huang et al., 2009). The common denominator of these contextual modulation models is long-range convolution. However, the issue of accepting these state of the art computational models of contextual modulation as plausible functional models of Layer 2/3 of V1 thus becomes one of addressing the *cortical convolution conundrum*, more specifically: how are the large scale convolutions required by such models accounted for in cortical architecture?

1.0.0.3

This paper's goal is to address the cortical convolution conundrum. In the process, we will propose a new fast Fourier transform, named Generalised Overarching SHIA Fast Fourier Transform (GOSH-FFT) and argue:

- GOSH-FFT has a natural implementation in the cortical architecture of visual area V1, and
- Its implementation provides a plausible cortical mechanism to account for the convolutions implied by long-range contextual modulation.

The rest of this paper is organised as follows: Section 2 provides a description of key neurophysiological and mathematical concepts underpinning the main thrust of this paper. Section 3 describes the Generalised Overarching SHIA Fast Fourier Transform (GOSH-FFT). Section 4 proposes a new interpretation of the physiology of long-range intrinsic connections and reinterprets previously introduced physiological concepts to propose a plausible cortical implementation of GOSH-FFT. Section 5 discusses various implications of the novel material of this paper. Section 6 summarises and concludes the paper. Section 7 is an appendix that contains a MatLab-like pseudo-code description of GOSH-FFT and a mathematical proof of GOSH-FFT.

2. Background

This section reviews neurophysiological and mathematical concepts pertinent to the understanding of GOSH-FFT and its cortical implementation.

2.1 Physiological background

The primary visual cortex (V1) has a distinctive layer structure. Inputs from the lateral geniculate nucleus (LGN) of the thalamus arrive chiefly to layer 4 in the monkey (Fitzpatrick et al., 1985). Layers 2/3 of monkey V1 contain the first neurons in the feedforward hierarchy that are strongly responsive to orientated stimuli (Schiller et al., 1976); (De Valois et al., 1979); (Parker & Hawken, 1988); (Leventhal et al., 1995). Central to understanding response properties of neurons in V1 is the notion of the receptive field (Hubel & Wiesel, 1962). The receptive field of a neuron is that region of its visual field to which it responds most strongly. Neurons within layers 2/3 V1 have receptive fields that are oriented. That is, they prefer stimuli that are lines of a certain orientation, or oriented texture elements (Hubel & Wiesel, 1974). The spatial and temporal frequency tuning preferences of neurons in V1 can also be measured. The neuron's response properties measured via the receptive fields resemble spatially localized filters with a preferred orientation and spatial frequency (Schiller et al., 1976); (Foster et al., 1985); (Mikami et al., 1986); (Edwards et al., 1995) or spatio-temporal energy (Basole et al., 2003); (Basole et al., 2006).

2.1.0.4

The orientation preference of neurons can be mapped using optical imaging techniques and neurological studies, which show good agreement with single cell measurements (Blasdel, 1992); and groups of neurons which act as a single unit. It has been experimentally shown that this single unit activity of large groups of single cells are composed of 10^4 (first order approximation) interconnected cells even in one local V1 column (Siegel, 1990). The advantages of modeling large scale neuron activity which exhibit cohort macroscopic organisation was shown by (Sirovitch et al., 1996). No model was presented but organising principles for analyzing and viewing data were presented. These techniques have revealed an intricate structure to the orientation preference map in layers 2/3. A critical feature of these structures is the orientation pinwheel (local map), in which the orientation preference of the neuronal population changes through the entire range of 180 degrees of orientations over the 360 degrees of polar range of the circular pinwheel. At the centre of the pinwheel is the singularity, which is the point at which lines of iso-orientation preference meet (Obermayer & Blasdel, 1993).

2.1.0.5

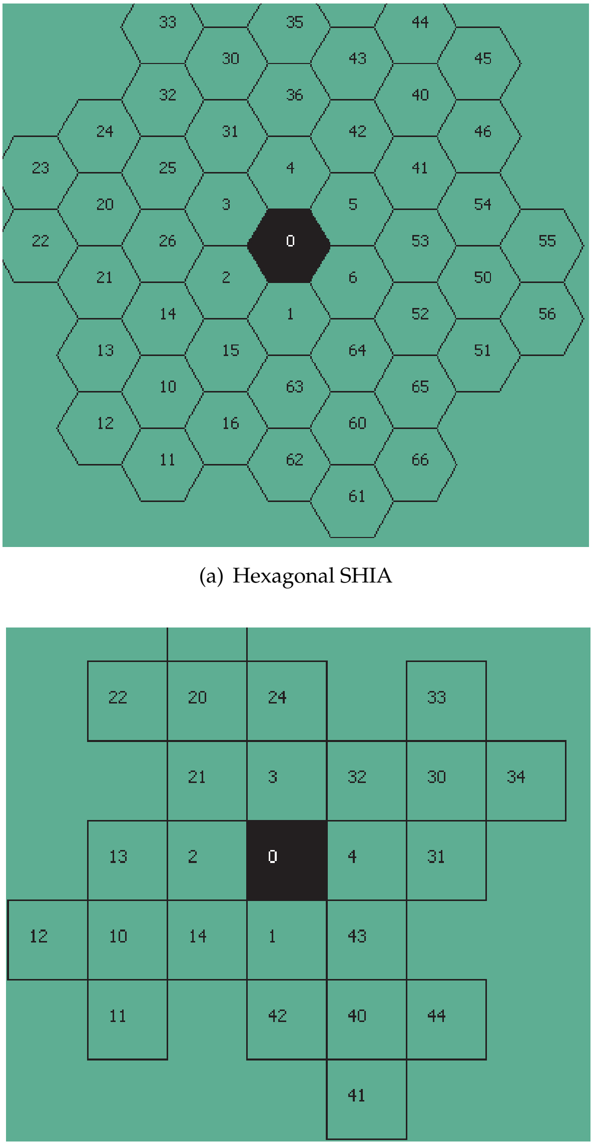
The cortex is often called the iso-cortex because of the repeated structures of which it is comprised (Douglas & Martin, 1991). The smallest scale of structure is the minicolumn which, in the monkey, consists of 30 adjacent pyramidal cell shafts in layers 2/3 packed within a diameter of 23 μm (Peters & Sethares, 1996). There are approximately 20 cell bodies within a minicolumn in layers 2/3. The next largest physical scale in V1 at which repeated structures occur is the cortical column (Lund et al., 2003). The cortical column is 200 μm in diameter and is the scale at which long-range patchy connections terminate. A number of anatomical and functional markers repeat at a larger scale of 400 μm. These include the distance between CO blobs, the approximate periodicity of the orientation preference map, and the spatial scale of a single ocular dominance band (Lund et al., 2003). Orientation pinwheels are also of approximately this spatial scale. Each of these functional markers has been shown to be closely related to the system of patchy connections, in which like response preference connects to like, and the inter-patch distance in V1 has this same periodicity of 400 μm (Bartfeld & Grinvald, 1992); (Malach et al., 1993); (Bosking et al., 1997). The largest spatial scale is V1 itself, which is some 4 cm wide in the monkey. There are of the order of 10,000 CO blobs in layers 2/3 of V1 (Murphy et al., 1998), and ocular dominance bands of 120 in number (Horton & Hocking, 1998), suggesting that the multiple response property maps with periodicity of 400 μm repeat around 10,000 times over layers 2/3 of V1. The input connections from the LGN arborize at a range of scales within layer 4C of V1. These inputs are arranged in block-like structures at the approximate scale of an ocular dominance band in layer 4C, but at a finer scale of approximately one column in layer 4C (Fitzpatrick et al., 1985). Further fine scale arborizations occur at approximately the scale of one minicolumn in layer 4A. At the global scale of the cortex, inputs from the LGN are organized into a retinotopic mapping of the visual field (Rolls and Cowey 1970; (Tootell et al., 1988). Connectivity into the layers 2/3 of V1 occurs via a number of anatomical routes, apart from the well described feedforward connections from layer 4 (Fitzpatrick et al., 1985). Other routes of information transfer include extra-striate feedback (Rockland et al., 1994); (Rockland & Vanhoesen, 1994); (Angelucci et al., 2002), long-range intrinsic fibres within V1 (Blasdel et al., 1985), as well as feedback from V1 to the lateral geniculate nucleus (Marrocco et al., 1982); (Briggs & Usrey, 2007), and diffusion of visual signal in the retina (Kruger et al., 1975); (Berry et al., 1999).

2.1.0.6

The finest scale of axonal projections within V1 are the short-range intrinsic connections that provide connectivity between neurons up to the range approximated by an ocular dominance column width, or 400 μm. Within V1, long-range patchy connections extend for 3 mm within the supra-granular layers (Stettler et al., 2002) and long-range connections within the infra-granular layers extend for up to 6 mm (Rockland & Knutson, 2001). V1 also receives feedback from at least nine extra-striate areas (Rockland & Vanhoesen, 1994). Extra-striate feedback is considered by most researchers to be the primary source of long-range horizontal interactions measured in V1 (Alexander & Wright, 2006). These feedback connections are fast conducting myelinated cortico-cortical fibres, and while they traverse distances of up to 10 cm in the monkey, the transmission delays are of the same order as intrinsic short and long-range axons within V1 (Bringuier et al., 1999); (Girard et al., 2001). These feedback connections are often in register with the intrinsic patchy system within V1, depending on the area of origin (Angelucci et al., 2002); (Lund et al., 2003). The middle temporal (MT) visual area will serve here as brief illustration of the role of extra-striate feedback in V1. Receptive field sizes in MT are about 10 times larger than in V1 at all eccentricities (Albright & Desimone, 1987). Small focal injections of tracer into V1 indicate that the sizes of the feedback fields from MT to V1 are 21-fold larger than the aggregate receptive size of the V1 injection sites (Angelucci et al., 2002). These feedback connections are an obvious substrate for the integration of global signals into V1 (Bullier, 2001). The local-global map hypothesis (Alexander et al., 2004) of V1 posits a non-local influence on the structure of local maps in V1. This hypothesis states that the global visual map in V1 is remapped to the local map scale in V1 in the form of a map of response properties, e.g. orientation, and in the case of the monkey, spatial frequency preference and colour selectivity. These local maps tile the surface of V1 and each receive inputs from a large extent of the visual field. *So rather than the local map being simply a map of primitive visual features that apply to a point in visual space, the local map is a map of primitive visual features as they arise in the organisation of the visual field and become relevant to a location in visual space.* As the maximum range of contextual modulation in V1 approaches the size of the visual field (Alexander & Wright, 2006), the local organisation of response properties can be influenced by the functional properties of the global visual field.

2.2 Mathematical background

Fundamental to the Fourier transform proposed in this paper is the Spiral Honeycomb Image Algebra (SHIA). This is a data structure that embodies important properties of the natural visual constraints imposed by the primate eye (Sheridan et al., 2000). In particular, SHIA has a discrete, finite and bounded domain which mimics the distribution of photo receptors on the retinal field. The underlying geometry of the SHIA is a hexagonal or rectangular lattice. In the former case, each hexagon has a designated positive integer address expressed in base seven. The numbered hexagons form clusters of super-hexagons of size 7^n. These self-similar super-hexagons tile the plane in a recursively modular manner. As an example, a super-hexagon of size $7^2 = 49$ and its concomitant addressing scheme is displayed in Fig. 1 (a).



(a) Hexagonal SHIA

(b) Rectangular SHIA

Fig. 1. Displays the two-level addressing scheme of SHIA: (a) Hexagonal and (b) Rectangular.

In the latter case, each rectangle has a designated positive integer address expressed in base five. An example of this addressing scheme is displayed in Fig. 1 (b).

2.2.0.7

The importance of the SHIA addressing scheme is that it facilitates primitive image transformations of translation, rotation and scaling. One of these transformations that has proven to be of particular relevance to the Fourier transform is one that provides rotation and scaling. It is referred to as mapping $M10$ in the notation of SHIA. Fig. 2 (a) displays an image represented in a four level SHIA, size is $7^4 = 2401$. Fig. 2 (b) represents the effect of applying $M10^2 = M100$ to this image.

The critical observation to make in regard to the effect of $M10$ is that it produces multiple 'near' copies at reduced resolution of the input image. This transform will play a critical role in the proposed FFT.

2.2.0.8

The origin of what is now called a Fourier transform dates back to 1807 when Jean Baptiste Joseph Fourier defined the notion of representing a function as a trigonometric series. The discrete version of a Fourier transform (DFT) for a one-dimensional signal is defined as:

$$F(u) = \frac{1}{N} \sum_{x=0}^{N-1} f(x) e^{-j2\pi ux/N} \tag{1}$$

for $u = 0, ...N - 1$, where $f(x)$ is a real valued function, N represents the number of elements in the signal and $j^2 = -1$.

The effect of this transform is to capture the spatial relationships inherent in the signal f(x) and express these relationships as the sum of sinusoidal function (frequency components). Similarly, the discrete version of an inverse Fourier transform (IDFT) for a one-dimensional signal is defined as:

$$f(x) = \sum_{u=0}^{N-1} F(x) e^{j2\pi ux/N} \tag{2}$$

for $x = 0, ...N - 1$, where $F(u)$ is the Fourier transform of the real valued function $f(x)$, N represents the number of elements in the signal and $j^2 = -1$.

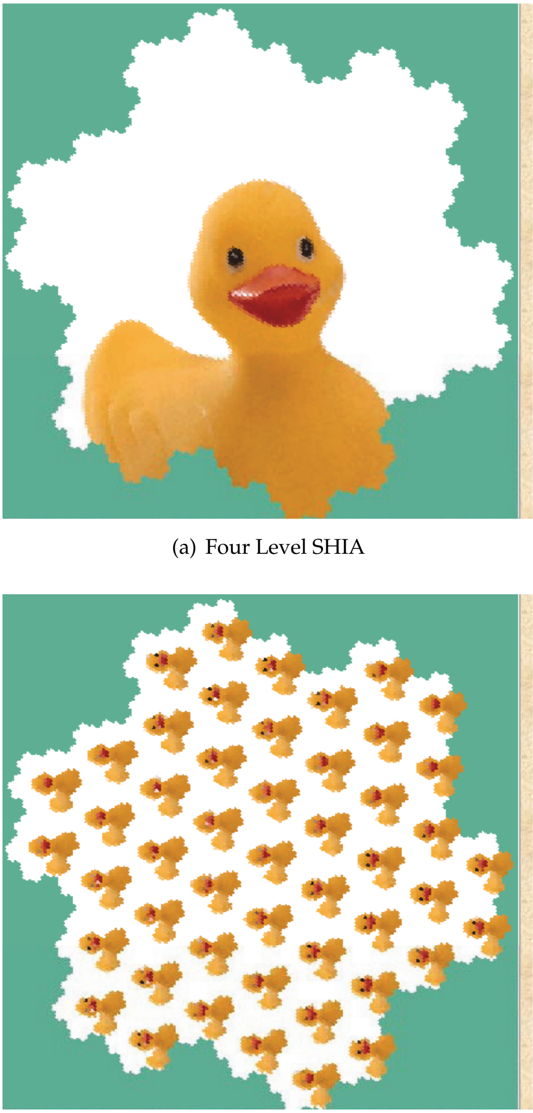
The effect of this inverse Fourier transform is to take a signal in frequency domain back to the spatial domain.

2.2.0.9

Prior to the invention of the digital computer, the Fourier series was employed as a purely analytic tool. However, since that time, the development of a class of computationally efficient algorithms, known as fast Fourier transforms (FFT), has meant the notion has become a useful computational tool (VanLoan, 1992). One of the most attractive computational properties of the FFT is its ability to process signals at higher resolution with a minimal increase in cost to complexity. Today, most of us benefit from fast Fourier transforms every day without even knowing it as these algorithms power a vast range of electronic technology such as digital cameras and cell phones.

2.2.0.10

The relevance of a fast Fourier transform to this paper is its relationship to the notion of convolution. The convolution of two functions $f(x)$ and $g(x)$ is denoted by $f(x) * g(x)$ and



(a) Four Level SHIA

(b) M10

Fig. 2. Displays (a) an image of a duck represented on a four-level SHIA; (b) the result of applying SHIA transform M10 twice to the image displayed in (a). There are four observable effects: 1) multiple near copies of the input image (a), 2) each copy is rotated by the same angle, 3) each copy is scaled by the same amount, 4) applying *M*10 twice to the image displayed in (b) results in the image displayed in (a).

its discrete definition is

$$f(x) * g(x) = \sum_{a \in g} f(x) g(x-a) \tag{3}$$

A well known result to researchers in the field of signal processing is the Convolution Theorem, which relates convolution in the spatial domain to convolution in the frequency domain. For two functions, $f(x)$ and $g(x)$, let $F(x)$ and $G(x)$ represent the Fourier transform of $f(x)$ and $g(x)$ respectively. The Convolution Theorem states that,

$$f(x) * g(x) \rightleftharpoons F(x)G(x) \tag{4}$$

In other words, the convolution of two functions in the spatial domain can be achieved by the multiplication of the functions in the frequency domain.

3. Generalised Overarching SHIA fast Fourier transform (GOSH-FFT)

In this section we propose a new fast Fourier transform that, as we will see later, possesses the potential to be implemented in cortical architecture and thereby address the cortical convolution conundrum. Associated with SHIA, as described in Section 2.2, is a Cooley-Tucky type fast Fourier transform, named Generalised Overarching SHIA Fast Fourier Transform (GOSH-FFT). This novel fast Fourier transform employs the transform $M10$, as described in Section 2.2, as the critical mechanism that turns a Fourier transform into a fast Fourier transform.

3.0.0.11

Suppose an image is represented on a SHIA of size 7^n, where $n \geq 0$. Let $m \geq 0$ such that $nmodulom = 0$. Let $k = n \div m$ and M denote $M10^m$, the composite of $M10 \circ M10 \circ ...$ m times. For(i:0:k)

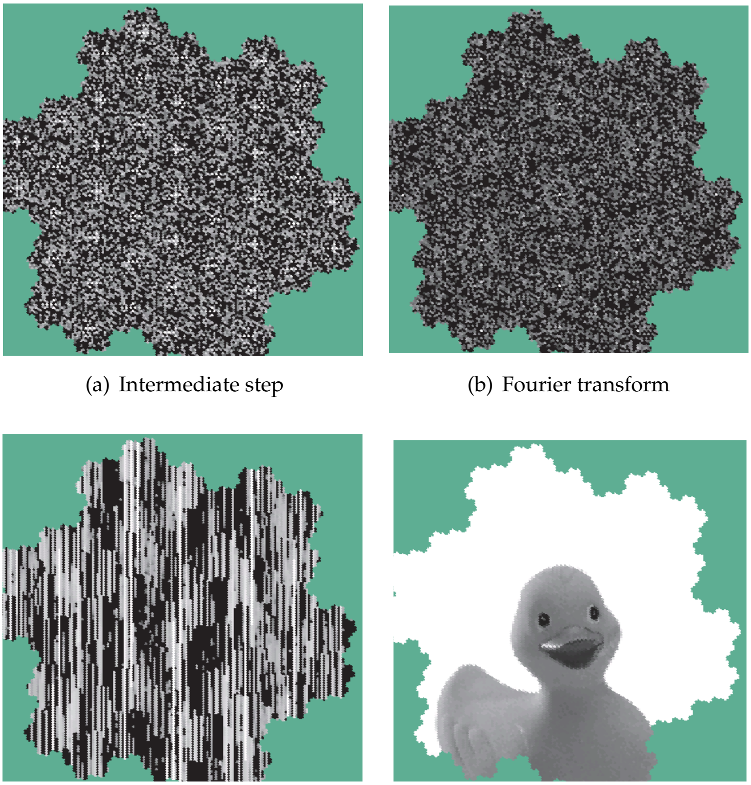
1. Apply M to the input;
2. Perform a discrete Fourier transform over a sequence of sub images of size 7^m;
3. Apply the inverse of M^i locally.

A special case of GOSH-FFT was initially described in (Sheridan, 2007), with $m = 1$. The significance of the initial work was that it demonstrated the intrinsic connection between the Fourier transform and primitive image transformations of translation, rotation and scaling. It also turns out that another special case of GOSH-FFT, when $n = 2m$, will play a critical role in the core hypothesis of this paper. This special case, named Particular SHIA FFT (PaSH-FFT), is illustrated in Fig. 3.

A complete statement of Algorithm 3.0.0.11 is written in MatLab-like pseudo-code and can be found in Section 7 along with a mathematical proof that GOSH-FFT delivers a Fourier transform.

4. Cortical implementation of contextual modulation

In Section 1, we reviewed state of the art models of contextual modulation and concluded that these models implied the cortical convolution conundrum. We further motivate this conundrum by observing that as a consequence of Equation 3, a convolution of the entire visual field requires every minicolumn in Layer 2/3 of area V1 to receive an input from every other minicolumn of that layer. As there just are not enough connections to convolve the visual field in one cortical step, the convolution problem reduces to one of determining, given the known connectivity, what is the sequence of intermediate neurons an initial input must pass



(a) Intermediate step

(b) Fourier transform

(c) Intermediate step

(d) Inverse Fourier transform

Fig. 3. Displays the results of applying the special case of GOSH-FFT, that is PaSH-FFT, to image of Fig. 2 (a), with n=4 and m=2. The four sub figures display intermediate results of PaSH-FFT: (a) on completion of first iteration of PaSH-FFT to Fig. 2; (b) Fourier transform on completion of second iteration; (c) on completion of first iteration of inverse PaSH-FFT; (d) Inverse PaSH-FFT on completion of second iteration.

through before being output as a convolved value. With the cortical convolution conundrum thus fully formulated, in this section we will establish a specification of a sufficient sequence of steps to address the issue. This specification will unfold in three steps. First, we will discuss how the SHIA transform *M*10 manifests in cortical architecture. We will then employ this manifestation to demonstrate how neural circuitry accommodates PaSH-FFT. Lastly, we will show how the cortical manifestation of PaSH-FFT supports long-range convolution.

4.1 Cortical manifestation of M10

A critical component of the fast Fourier transform, PaSH-FFT, is the transform *M*10. Consequently, it is an imperative of our argument that the redistribution properties of *M*10 be accounted for in the neural circuitry of the visual system. To this end we now argue that the required effects of *M*10 are accounted for by the long-range properties of patchy connections between columns of Layer 2/3 and similarly patchy extra-striate feedback connections to area V1.

4.1.0.12

It has been argued that the orientation pinwheel comprises a unitary organisational structure or local map in layer 2/3 of area V1 (Hubel & Wiesel, 1974); (Bartfeld & Grinvald, 1992); (Blasdel, 1992). When four pinwheels are reflected about their common borders, a saddle point arises at the centre of the four pinwheels. See Fig. 4.

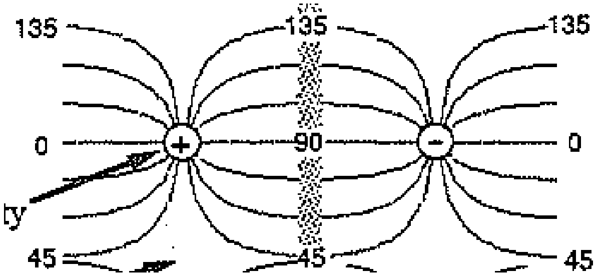


Fig. 4. Displays a schematic diagram of the pinwheel like structures of visual area V1, extracted from Figure 10 page 43 of Bruce et al. (2003).

4.1.0.13

In the macaque, the preferred response properties of V1 neurons can be influenced by activity from a wide extent of the visual field. A review of contextual modulation in the monkey demonstrated contextual modulation in V1 from long-ranges in the visual field (Alexander & Wright, 2006). The review was compiled from a number of experimental paradigms, including visual stimulation with long lines while the neuron's receptive field is occluded (Fiorani et al., 1992), surround only textures (Rossi et al., 2001) and colour patches placed distally to the neuron's receptive field (Wachtler et al., 2003). It was shown that the maximum range of contextual modulation measurable in V1 approaches a large extent of the visual field relative to a neuron's receptive field size or the local cortical magnification factor. Some experimental paradigms, such as the curve tracing effect (Roelfsema & Lamme, 1998); (Khayat et al., 2004), relative luminance (Kinoshita & Komatsu, 2001), and texture defined boundaries (Lee et al., 1998) show excitatory contextual modulation with 'tuning curves' that are flat out to the maximum distance tested. The functional connectivity that underlies this long-range contextual modulation in the monkey is likely to involve cortico-cortical feedback from higher visual areas working in concert with long-range intrinsic patchy connectivity. In the monkey, the feedback connections to V1 from higher visual areas incorporate inputs from a very large extent of the visual field (Angelucci et al., 2002); (Lund et al., 2003).

4.1.0.14

In the analysis that follows, the combination of patchy intrinsic connections and patchy feedback connections are therefore assumed to enable transfer of visual information at ranges approaching the global scale of the visual field. Moreover, we assume that the quantity and distribution of these connections are adequate to deliver the effects of transform $M10$ at the scale of the visual field.

4.2 Cortical manifestation of PaSH-FFT

The next step in accounting for global convolution in cortical circuitry is to explore how PaSH-FFT manifests itself in cortical architecture. The raw data, at the lowest level of PaSH-FFT, are complex numbers that must be multiplied and added. The first issue to address is to justify our assumption that the operations being performed by a neuron could be represented as complex arithmetical operations on complex numbers. Specifically, PaSH-FFT requires that a neuron can be regarded as a mechanism capable of representing and manipulating complex numbers in accordance with the arithmetical operations of addition and multiplication. There are many ways in which to interpret neuronal function in terms of complex addition and multiplication. The model presented by (MacLennan, 1999) is adequate for the purposes of this paper, where it is shown how the representation of complex numbers can be encoded as the rate and relative phase of axonal impulse. From this encoding, complex multiplication is associated with the strength of a synaptic connection as the signal passes through it and complex addition is associated with the summing of the neuronal inputs. Thus at the lowest level of computation in our model, we assume that the operation being performed by a neuron can be represented as complex addition and multiplication.

4.2.0.15

In area V1, each neuron makes use of information available to it in real time. There is evidence that contextual information is projected to widespread regions in V1 in an anticipatory manner. Since the spatial changes in the visual field tend to be predictable from previous visual inputs, anticipatory contextual inputs can arrive in time to be integrated in an adaptive manner with ongoing feedforward input. In order to express the properties of widespread contextual integration in a more formal manner, however, we will use the mathematical convenience of assuming that each of the distinct mathematical processes to be described occurs in a step-wise fashion. This more constrained approach allows not only each distinct part of the process to be formulated, but also formulates the inter-relationships between the various sub-processes. Although it is claimed that this approach is appropriate for the purposes of this paper, it must be acknowledged that the question of how such "contextual integration" actually occurs in the neuronal system remains open.

4.2.0.16

At the finest scale of connectivity via short-range intrinsic connections, each neuron of a local map is treated as if it were connected to every other neuron minicolumn of that local map. While this is not literally true, considerations of poly-synaptic interactions at this local scale, and the real-time, anticipatory nature of visual processing means that it is a reasonable approximation of the functional connectivity. Consequently, we can assume that each neuron in a local map can sum the outputs of all other neurons in that local map which have been multiplied by unique complex numbers. We call such a collection of parallel computations a local computation. See Fig. 5, which is a schematic diagram of a local computation.

4.2.0.17

This paper now needs to discuss three types of local computations, each of which is determined by the interpretation of the input signal and the collection of synaptic strengths which weight the input signal. If the input signal is the spatial domain and the weights are associated with a set of primitive roots of unity, then the resulting local computation is a Fourier transform, denoted F. (See Equation 1 for a definition of a Fourier transform.) If the

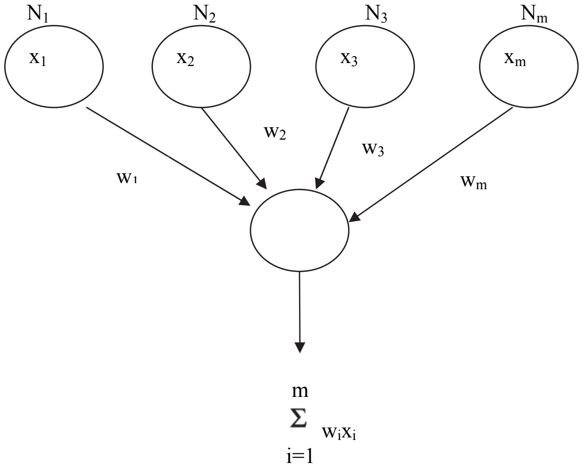


Fig. 5. Displays a schematic diagram of a computational unit. The circles represent neurons and the straight lines connecting the circles represent cortical connections. Each neuron depicted at the top of the figure outputs a value x_i. The neuron depicted at the bottom of the figure inputs the sum of each x_i multiplied by weight w_i.

input signal is the frequency domain and the weights are associated with a set of inverse primitive roots of unity, then the resulting local computation is an inverse Fourier transform, denoted *I*. (See Equation 2 for a definition of an inverse Fourier transform.) If the input signal is a frequency domain and the weights represent Fourier components, then the resulting local computation is a convolution in the frequency domain, denoted *C*. (See Equations 3 and 4.) Table 1 provides a summary of this notation.

Symbol	**Local Computation**	**Input Signal**	**Weights**
F	Fourier transform	spatial domain	primitive roots of unity
I	Inverse Fourier	frequency domain	primitive roots of unity
C	Convolution	frequency domain	Fourier components

Table 1. Summary of notation and definitions of local computations

4.2.0.18

At the next higher scale of connectivity each local map is assumed to have access to ongoing activity of every other local map via long-range patchy connections and striate-extrastriate interactions. These provide the means by which the results of local computation can be transported to another local map as input for a further local computation. We denote such a projection as *P* to represent the class of transformations (as described in Section 4.1).

4.2.0.19

All of the components of PaSH-FFT can be considered as being built of sequences of *P* projections followed by a local computation. Likewise, when we employ the term local computation in conjunction with the symbols *F*, *I* or *C*, it is implied that the signal within a local map has been sent to every other neuron in that local map through a synapse whose strength is associated with the appropriate weight of F, I or C respectively.

4.2.0.20

Although it is commonly accepted that the cortex has a massively parallel architecture, currently there exists no comprehensive model to describe these dynamics. The absence of such a model means that in any particular cortical process, we cannot be sure which aspects of the process are parallel and which are intrinsically sequential. We will employ the following notation to show how the inherently sequential steps of PaSH-FFT can be mapped into neural circuitry. Let the symbol $\odot$ denote the composition of two local computations as follows: given arbitrary local computations A and B to operate on a signal in sequence let $(s)A \odot B = ((s)A)B$.

Note that the operator is to the right of the input signal it operates on, which is enclosed in left and right parenthesis ().

4.2.0.21

With these concepts in hand, we can now identify the sequential steps of PaSH-FFT. In this special case, the size of the input signal is the square of the size of the local computation and represents two iterations of GOSH-FFT, as described in Section 3. The identification of the sequential steps also suggests the sequence of connections that the input signal must traverse. We now illustrate this with PaSH-FFT, given an input signal s, then the application of PaSH-FFT would be expressed as follows:

$$(s)PaSH - FFT = (s)P \odot F \odot P \odot F \odot P \tag{5}$$

Given the assumed neural parallelism, a count of the number of components on the right hand side of the equals sign in Equation 5, reveals that a Fourier transform of the entire visual field can be completed by the signal traversing a sequential path connecting five neurons. Likewise, an inverse Fourier transform can be delivered in cortical circuitry as follows:

$$(s)inversePaSH - FFT = (s)P \odot I \odot P \odot I \odot P \tag{6}$$

4.3 Cortical manifestation of convolution

We now progress to the issue of how convolution could be implemented in cortical architecture. To this end, we describe the various computational constraints imposed by the computational requirements of convolution and argue that the known cortical architecture satisfies these constraints.

4.3.0.22

The key to the solution of the convolution problem in the neurological domain is provided by the Convolution Theorem, the same one employed by numerous digital signal processing applications. This theorem was discussed in Section 2.2. The importance of the theorem is that the convolution of two functions in the spatial domain can be achieved by the multiplication of the functions in the frequency domain. The implications of this theorem to the cortical convolution conundrum are significant. In our model, the components of the Fourier transform of the function the input signal is to be convolved with are represented by connection weights. Then, once the input signal has been transformed to the frequency domain the required convolutions can be performed by mere multiplications. In cortical terms, each component of the signal, in the frequency domain, must traverse a connection to one more neuron to achieve the desired multiplication. However, the resulting convolution, in the frequency domain, must be transformed back to the spatial domain to complete the convolution. This is achieved with an inverse Fourier transform. Accordingly, the sequence of connections along the path that terminates in the output of a convolved value in the spatial domain is thus given by:

$$(s)Convolution = (s)P \odot F \odot P \odot F \odot C \odot P \odot I \odot P \odot I \odot P \tag{7}$$

4.3.0.23

It is assumed that each component of the input signal traverses parallel paths along the network. Thus, the net time cost to complete a convolution is equivalent to the time required for a component of the input signal to traverse a path connecting 10 neurons. This path is composed of five short-range intrinsic connections and five long-range connections.

4.4 Analysis

The plausibility of the cortical model of convolution proposed in this paper is fundamentally predicated on the assumptions made in its formulation. Consequently, we summarise these assumptions along with the arguments offered to justify them before we provide an analysis of the model's parameterisation:

1. The number of long-range patchy connections is adequate to achieve a redistribution of the global signal via transform $M10$. This was argued in Section 4.1 and heavily relied on a conclusion based on a review article reported in (Alexander & Wright, 2006).
2. A first order approximation of the number of minicolumns in a local map as 10,000. This was discussed in Section 2.1 and relied on the work reported in (Siegel, 1990).
3. The number of short-range intrinsic connections is adequate to consider each local map as being fully connected. This was discussed in Section 4.2 and relied on the work reported in (Siegel, 1990).
4. A first order approximation of the number of minicolumns in the global map is $10,000^2 =$ 100 million. This was discussed in Section 2.1 and was based on the work reported in (Murphy et al., 1998) and Assumption 2.

4.4.0.24

The first assumption is possibly the most critical as it establishes the fundamental architectural relationship between the local and global maps and is essential to PaSH-FFT. The second two assumptions implied that a Fourier transform of the portion of the signal represented in a local map would be completed by each component of the signal traversing one cortical connection and that on completion of the first iteration of PaSH-FFT, the global signal consists of 10,000 local discrete Fourier transforms each of which is at the scale of a local map. Then, on completion of the second iteration of PaSH-FFT, the 10,000 local Fourier transforms would be transformed into a global Fourier transform of size $10,000^2 = 100$ million, which by the fourth assumption represents the size of the global signal. From this we are able to assert that the input spatial signal would be transformed to frequency space at a cost of the signal traversing a path connecting four neurons. With the signal in frequency space, we employed the Convolution Theorem to assert that with each component of the global signal traversing one additional connection, the state of the signal would represent a convolved signal in frequency space. This assertion was predicated on the assumption that the weight of each of these last connections represented the Fourier weight of the appropriate gaussian. The final step was to transform the convolved signal back from frequency space to the spatial domain. This was achieved with the inverse PaSH-FFT, which would be completed at the additional cost of the signal traversing a path connecting a further five neurons. Putting these three steps together, we arrived at a total path length of 10 connections for the global input signal to be transformed into a representation of global convolution of the visual field. We also note that this analysis accounted for a single global convolution of the input signal. However, there will be many global convolutions required, possibly up to one for every orientation preference and spatial frequency preference represented in a local map. Although the input spatial signal needs only to be transformed into the frequency domain once, each distinct convolution would require a distinct set of parallel paths to transform the signal back into the spatial domain. Consequently, the multiple convolutions would not necessarily result in a longer path. Accordingly, we assert that the transform, PaSH-FFT, with appropriate parameterisation, would deliver a global convolution of the visual field. Moreover, this output signal is generated within the required time constraints imposed by observed contextual modulation. Given our assumptions, the lowest number of iterations required to complete a Fourier transform is two. Consequently, 10 represents the length of the shortest path (see equation 7) possible to deliver a global convolution via PaSH-FFT.

5. Discussion

The signal processing literature describes many different types of fast Fourier transforms (FFT). Although any one of them represents an alternative candidate to PaSH-FFT, the problem to address is accounting for how they might be implemented within the known constraints of cortical architecture. All fast Fourier transforms need to rearrange components between their intermediate steps of multiply and add. PaSH-FFT derives its rearrangements of components with the transform $M10$ that, as argued, is compatible with the distribution and quantity of long-range cortical connections. If any other FFT could be substituted for PaSH-FFT in the model, one would need to account for the rearrangement phase of that FFT within the known connectivity of area V1.

5.0.0.25

Another issue worthy of some discussion pertains to the Fourier transform and the absence of empirical evidence that would irrefutably demonstrate its cortical implementation. Part of the explanation for this lack of evidence may be provided by the role the Fourier transform plays in the vision process as suggested by this paper. That is, PaSH-FFT was shown to be a means to an end (convolution), not the end itself. Consequently, the question of finding neurons through empirical experimentation that measures response properties of neurons that closely model the profile of a Fourier transform may remain unanswered for some time to come.

5.0.0.26

Underpinning the proposed implementation of PaSH-FFT in cortical architecture is a highly simplistic model of the parallelism inherent in the cortex. The model employed did not take into account at least two well accepted features of this parallel architecture. First, the system itself somehow synchronises the flow of the signal. Second, the cortex does not need computation-like synchronisation or state update. Synchronisation can be provided by considering "Small World" relationships. (Gao et al., 2001) have shown that a "Small World" network needs only a small fraction of long-range couplings to obtain a great improvement in both stochastic resonance and synchronisation in network connectivity of bistable oscillators. We suggest that the known topology of the visual cortex (Zeki, 1993) if considered as a "Small World" network can provide the foregoing benefits. They would be consistent with the long-range and short-range connectivities of V1 to retinal neurons which have the required bistable oscillator condition provided by on-centre or off-centre neurons responses to light and dark and including those with colour opponency properties. The long and short-range selectivity for connections can be dynamic based on the neuron threshold levels and spatial frequency channels (Dudkin, 1992). The system updates a neuronal state only when new information indicates a change in the input signal.

5.0.0.27

The cortical implementation of PaSH-FFT was discussed in Section 4.2 where it was argued that the known connectivity of area V1 was sufficient to support its cortical implementation. It was then argued that this implementation could deliver the required convolution in a 'small' number of sequential steps. However, the argument did not rule out the possibility of an alternative mechanism that would deliver the required convolution in fewer steps than PaSH-FFT. It would appear that without a sufficiently developed model of the brain's parallelism, it is unlikely that a mathematical proof of a lower bound for the minimum number of sequential steps could be produced. Currently, the only bound that we can be sure of is that the required convolution could not be completed in one step. The question of determining the minimum lower bound remains an open question.

5.0.0.28

The role of the frequency domain was at the heart of the solution to the cortical convolution conundrum proposed in this paper. However, the possibility of performing the convolution in the spatial domain without resorting to the frequency domain cannot be ruled out by any argument presented in this paper. Although it is unclear how this could be accomplished without resorting to a highly asymmetric model of the distribution of the connectivity of long-range connections. In any case, the search for an explanation of how the dynamic reconfiguration implied by the analysis of this paper is actually accomplished is likely to provide many different conjectures along the way. One possible avenue in this endevour might be provided by further tracer experiments such as those reported in (Angelucci et al., 2002).

6. Summary and conclusion

This paper reviewed the evidence for long-range contextual modulation and concluded that it implied cortical convolution at the scale of the visual field. This resulted in the need to address the problem of how such long-range convolution could be accounted for with known cortical connectivity and within known time constraints. The paper proposed a solution to the problem that emerged from a mathematical analysis of cortical connectivity to account for the implied constraints of long-range convolution. In particular, it was argued that the known distribution of the long-range patchy connections and extrastriate connections is adequate to provide the means by which the global visual signal can be transformed into frequency space where the convolution can be performed. The main thrust of the argument was that these long-range connections facilitated the transformation of the signal into and out of the frequency domain via a new fast Fourier transform named PaSH-FFT. A mathematical proof of the most general form of this FFT, GOSH-FFT, was provided in the appendix along with MatLab-like pseudo-code to facilitate the implementation of GOSH-FFT in computer software.

6.0.0.29

It was shown that, to a first order approximation, a cortical implementation of PaSH-FFT could account for the large scale convolution implied by known models of contextual modulation. The significance of PaSH-FFT is that it:

- represents a plausible cortical mechanism to account for long-range contextual modulation;
- suggests a theoretical explanation of how the brain might be wired to achieve large scale Fourier analysis;
- opens up the possibility of explaining other cortical processes via frequency space computations.

It is the conclusion of this paper that the processing of the visual signal in the frequency domain via a fast Fourier transform plays a fundamental role in primate vision.

7. Appendix

In this appendix, we present the pseudo-code for GOSH-FFT and a mathematical proof of GOSH-FFT.

7.1 Appendix A

This section presents a formal statement of GOSH-FFT in MatLab-like pseudo-code.

Notation:

$x =$ Complex array specifying the input signal
$base = 7^{\alpha}$, where α is an integer greater than zero
$n = 7^{\beta}$, where β/α , is an integer greater than zero.
$\otimes$ denotes scalar multiplication in SHIA

```
function y = $GOSH - FFT(x, n, base, \alpha)$
  $y \leftarrow x$
  $forlevel = 1 : log_{base}(n$
    $f \leftarrow m10(y, \alpha)$
    $r \leftarrow rootsOfUnity(base^{level})$
    $w \leftarrow m10(r, \alpha)$
    $F \leftarrow localDFFT(f, w, level, n, base)$
    $y \leftarrow localM10Inverse(F, level, n, base, \alpha)$
    $r \leftarrow w$
  end
end
```

```
function T = $m10(t, \alpha)$
    $fori = 0 : n - 1$
        $a \leftarrow convertToSHIA(i)$
        $b \leftarrow 10^{\alpha} \otimes a$
        $c \leftarrow convertFromSHIA(b)$
        $T(c) \leftarrow t(i)$
    end
end
```

```
function F = localDFT( f, w, level, n, base )
    $subSignalSize \leftarrow base^{level}$
    $numSubSignals \leftarrow n/subSignalSize$
    $fors = 0 : numSubSignals - 1$
        $tab \leftarrow s \times subSignalSize$
        $fors = 0 : subSignalSize - 1$
            $wLB \leftarrow i \times base$
            $lB \leftarrow i \times base + tab$
            $uB \leftarrow lB + base - 1$
            $foru = lB : uB$
                $F(u) \leftarrow \sum_{x=0}^{base-1} f(x + lB) \times w(x + wLB)$
            end
        end
    end
end
```

```
function T = localM10Inverse( t, level, n, base )
    $subSignalSize \leftarrow base^{level}$
    $numSubSignals \leftarrow n/subSignalSize$
    for s = 0: numSubSignals-1
        for i =0: subSignalSize-1
            $a \leftarrow convertToSHIA(i)$
            $b \leftarrow 10^{-\alpha} \otimes a$
            $c \leftarrow convertFromSHIA(b)$
            $T(c) \leftarrow t(i)$
        end
    end
end
```

7.2 Appendix B

This section presents the mathematical proof of GOSH-FFT.

Notation: The symbol % will be employed to mean modular arithmetic. Let $B = 7^m$, where m is a positive integer. $N = B^p$, where p is a positive integer.

Let M denote the compound transform $M10^m$ from SHIA.

Let $1, e, e^2, e^3, ..., e^{N-1}$ represent N roots of unity, where multiplication of two arbitrary roots of unity denoted e^i and e^j is defined as: $e^i e^j = e^{(i+j)\%N}$.

$e(u) = e^{(u\%B)+(u/B)n} = 1, e^1, ..., e^{B-1}, ..., e^{(B-1)n}, ..., e^{(B-1)n+(B-1)}$, for $p > 1$ and $=$ $1, e^1, ..., e^{B-1}$, for p = 1.

$F(u) = F^{u\%B}_{u/B} = F^0_0, ..., F^{B-1}_0, ..., F^0_{n-1}, ...F^{B-1}_{n-1}$, denote a sequance of N Fourier components.

$f(x) = f^{x\%B}_{x/B} = f^0_0, ...f^{B-1}_0, ..., f^0_{n-1}, ..., f^{B-1}_{n-1}$, denote a sequence of N points in the input signal.

The proof is by induction on p. When p=1, GOSH-FFT is simply a DFT. Assume the GOSH-FFT computes a Fourier transform for all levels less than p.

$M^{-1}(f^{x\%B}_{x/B}) = f^0_0, \ldots, f^0_{n-1}, \ldots, f^{B-1}_0, \ldots, f^{B-1}_{n-1}$.

We now have B sub-signals, each of which is composed of n points. Then, by the induction hypothesis, we can apply GOSH-FFT to obtain B individual transforms of the B sub-signals; which yields:

Let

$$\begin{aligned}
\hat{f} &= \hat{F}^0_0, \ldots, \hat{F}^0_{n-1}, \ldots, \hat{F}^{B-1}_0, \ldots, \hat{F}^{B-1}_{n-1}.\\
M(\hat{f}) &= \hat{F}^0_0, \ldots, \hat{F}^{B-1}_0, \ldots, \hat{F}^0_{n-1}, \ldots, \hat{F}^{B-1}_{n-1}\\
M(e(u)) &= e((u/B) + (u\%B)n)\\
&= 1, e^n, \ldots, e^{(B-1)n}, e^1, e^{n+1}, \ldots, e^{(B-1)n+1}, \ldots, e^{n-1}, e^{2n-1}, \ldots, e^{Bn-1}
\end{aligned}$$

Perform a local DFT on each of the n groups of B points. The general term is:

$$\sum_{q=0}^{B-1} \hat{F}^q_{u/B} e^{q((u/B)+((u\%B)n))\%N} \tag{8}$$

$$= \sum_{q=0}^{B-1} \left[\sum_{r=0}^{n-1} f^q_r e^{Br(u/B)\%N} \right] e^{q(u/B)+((u\%B)n))\%N} \tag{9}$$

$$= \sum_{q=0}^{B-1} \left[\sum_{r=0}^{n-1} f^q_r e^{Bn(u/B)\%N} e^{(u\%B)rBn\%N} \right] e^{q((u/B)+((u\%B))n))\%N} \tag{10}$$

$$= \sum_{q=0}^{B-1} \left[\sum_{r=0}^{n-1} f^q_r e^{Br((u/B)+((u\%B)n))\%N} \right] e^{q((u/B)+((u\%B)n))\%N} \tag{11}$$

$$= \sum_{q=0}^{B-1} \left[\sum_{r=0}^{n-1} f^q_r e^{(Br+q)((u/B)+((u\%B)n))\%N} \right] \tag{12}$$

$$= \sum_{s=0}^{N-1} f^{s\%B}_{s/B} e^{s((u/B+((u\%B)n))\%N} \tag{13}$$

$$= F((u/B) + (u\%B)n)) \tag{14}$$

$= F(u)$

Although this proof is for signals of size 7^n represented on a hexagonal lattice, it applies to the case of a signal of size 5^n on a rectangular lattice with a few modifications, i.e. all appearances of the number 7 needs to be replaced with the number 5.

8. Acknowlegements

The author thanks Valerie Christian McDougall for her work on improving the presentation of this manuscript.

The source code that implements GOSH-FFT is available in a software package named VisioLab from the following website: http://www.ict.griffith.edu.au/sheridan/

9. References

Albright, T. D. & Desimone, R. (1987). Local precision of visuotopic organization in the middle temporal area (Mt) of the macaque. *Experimental Brain Research*, Vol. 65, No. 3, 1987, 582-592

Alexander, D. M. & Bourke, P. D. (2004). Intrinsic connections in tree shrew V1 imply a global to local mapping. *Vision Research*, Vol. 44, No. 9, 2004, 857-876

Alexander, D. M. & Wright, J. J. (2006). The maximum range and timing of excitatory contextual modulation in monkey primary visual cortex. *Visual Neuroscience*, Vol. 23, No. 5, 2006, 721-728

Angelucci, A. & Levitt, J. B. (2002). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, Vol. 22, No. 19, 2002, 8633-8646

Bair, W. & Movshon, J. A. (2004) Adaptive temporal integration of motion in direction-selective neurons in macaque visual cortex. *J Neurosci.*, Vol. 24, No. 33, 2004, 7305-7323

Bartfeld, E. & A. Grinvald (1992). Relationships between Orientation-Preference Pinwheels, Cytochrome-Oxidase Blobs, and Ocular-Dominance Columns in Primate Striate Cortex. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, No. 24, pp. 11905-11909, 1992

Basole, A. & Kreft-Kerekes, V. (2006). Cortical cartography revisited: a frequency perspective on the functional architecture of visual cortex. *Visual Perception, Part 1, Fundamentals of Vision: Low and Mid-Level Processes in Perception* Vol. 154, 2006, 121-134

Basole, A. & E. White, L. (2003). Mapping multiple features in the population response of visual cortex. *Nature* Vol. 23, No. 6943, 2003, 986-990

Berry, M. J. & Brivanlou, I. H. (1999). Anticipation of moving stimuli by the retina. *Nature*, Vol. 398, No. 6725, 1999, 334-338

Blakemore, C. & Tobin, E. A. (1972). Lateral inhibition between orientation detectors in cats visual-cortex. *Exp Brain Res.*, Vol. 15, No. 4), 1972, 439-441

Blasdel, G. G. & Lund, J. S. (1985). Intrinsic Connections of Macaque Striate Cortex - Axonal Projections of Cells Outside Lamina 4C. *Journal of Neuroscience*, Vol. 5, No. 12, 1985, 3350-3369

Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, Vol. 12, No. 8, 1992, 3139-61

Bosking, W. H. & Zhang, Y. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of Neuroscience*, Vol. 17, No. 6, 1997, 2112-2127

Briggs, F. & Usrey, W. M. (2007). A fast, reciprocal pathway between the lateral geniculate nucleus and visual cortex in the macaque monkey. *Journal of Neuroscience*, Vol. 27, No. 20, 2007, 5431-5436s

Bringuier, V. & Chavane, F. (1999). Horizontal propagation of visual activity in the synaptic integration field of area 17 neurons. *Science*, Vol. 283, No. 5402, 1999, 695-699

Bruce, V.; Green, P.R.; Georgeson, M. A. (2003). *Visual perception physiology,psychology, and ecology fourth edition*, Psychology Press Taylor and Francis Group

Bullier, J. (2001). Feedback connections and conscious vision. *Trends in Cognitive Sciences*, Vol. 5, No. 9, 2001, 369-370

Cavanaugh, J. R.; Bair, W.; Movshon, J. (2002) A. Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *J Neurophysiol.*, Vol. 88, No. 5, 2002, 2547-2556

De Valois, K. K. & De Valois, R. L. (1979) Responses of Striate Cortex Cells to Grating and Checkerboard Patterns. *Journal of Physiology-London*, Vol. 291, Jun 1979, 483-505

Douglas, R. J. & Martin, K. A. C. (1991). A Functional Microcircuit for Cat Visual-Cortex. *Journal of Physiology-London*, Vol. 440, 1991, 735-769

Dudkin, K. (1992). Cooperative neural networks underlying image description in visual cortex. *NeuROInformatics and Neurocomputers*, Vol. 1, 1992, 387 - 398

Edwards, D. P. & Purpura, K. P. (1995). Contrast Sensitivity and Spatial-Frequency Response of Primate Cortical-Neurons in and around the Cytochrome-Oxidase Blobs. *Vision Research*, Vol. 35, No. 11, 1995, 1501-1523

Fiorani, M. & Rosa, M. G. (1992). Dynamic surrounds of receptive fields in primate striate cortex: a physiological basis for perceptual completion? *Proceedings of the National Academy of Sciences U S A*, Vol. 89, No. 18, 1992, 8547-51

Fitzpatrick, D. & Lund, J. S. (1985). Intrinsic Connections of Macaque Striate Cortex - Afferent and Efferent Connections of Lamina 4C. *Journal of Neuroscience*, Vol. 5, No. 12, 1985, 3329-3349

Foster, K. H. & Gaska, J. P. (1985). Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. *Journal of Physiology*, Vol. 365, 1985, 331-63

Gao, Z. & Hu, B. (2001). Stochastic resonance of Small World Networks. *Physicas Review*, Vol. 65, No. 16209, 1-4

Girard, P. & Hupe, J. M. (2001). Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *Journal of Neurophysiology*, Vol. 85, No. 3, 2001, 1328-1331

Grigorescu, C.; Petkov N.; Westenberg, M. A. (2003) Contour detection based on nonclassical receptive field inhibition. *IEEE Trans. Image Process.*, Vol. 12, No. 7, 2003, 729-739

Horton, J. C. & Hocking, D. R. (1998). Effect of early monocular enucleation upon ocular dominance columns and cytochrome oxidase activity in monkey and human visual cortex. *Visual Neuroscience*, Vol. 15, No. 2, 1998, 289-303

Huang, W.; Jia, J.; Jia, J. ((2008) Modeling contextual modulation in the primary visual cortex. *Neural Networks*, Vol. 21, No. 8, 2008, 1182-1196

Huang, W.; Jiao, L.; Jia, J.; Yu, H. (2009) A neural contextual model for detecting perceptually salient contours. *Pattern Recognition Letters*, Vol. 30, No. 11, 2009, 985-993

Hubel, D. H. & Wiesel, T. N. (1962). Receptive Fields, Binocular Interaction and Functional Architecture in Cats Visual Cortex. *Journal of Physiology-London*, Vol. 160, No. 1, 1962, 106

Hubel, D. H. & Wiesel, T. N. (1974). Sequence Regularity and Geometry of Orientation Columns in Monkey Striate Cortex. *Journal of Comparative Neurology*, Vol. 158, No. 3, 1974, 267-294

Khayat, P. S. & Spekreijse, H. (2004). Correlates of transsaccadic integration in the primary visual cortex of the monkey. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 34, 2004, 12712-12717

Kimura, R. & Ohzawa, I. (2009) Time course of cross-orientation suppression in the early visual cortex. *J Neurophysiol.*, Vol. 101, No. 3, 2009, 1463-1479

Kinoshita, M. & Komatsu, H. (2001). Neural representation of the luminance and brightness of a uniform surface in the macaque primary visual cortex. *Journal of Neurophysiology*, Vol. 86, No. 5, 2001, 2559-2570

Kruger, J. & Fischer, B. (1975). Shift-Effect in Retinal Ganglion-Cells of Rhesus-Monkey. *Experimental Brain Research*, Vol. 23, No. 4, 1975, 443-446

Lamme, v. A. F. (1995). The neurophysiology of figure-ground segmentation in primary visual cortex. *J. Neuroscience*, Vol. 15, 1995, 1605-1615

Lamme, V. A. F.; Zipser, K.; Spekreijse, H. (1998). Figure-ground activity in primary visual cortex is suppressed by anesthesia. *Proc Natl Acad Sci USA.*, Vol. 95, No. 6, 1998, 3263-3268

Lee, T. S.; Mumford, D.; Romero, R.; Lamme, V. A. F. (1998). The role of the primary visual cortex in higher level vision. *Vision Res.*, Vol. 38, No. 15, 1998, 2429-2454

Lee, H. C. & Choe, Y. (2003). Detecting salient contours using orientation energy distribution. *Proc.Internat. Joint. Conf. on Neural Networks*, Vol. 1, pp. 206-211

Leventhal, A. G. & Thompson, K. G. (1995). Concomitant Sensitivity to Orientation, Direction, and Color of Cells in Layer-2, Layer-3, and Layer-4 of Monkey Striate Cortex. *Journal of Neuroscience*, Vol. 15, No. 3, 1995, 1808-1818

Lund, J. S. & Angelucci, A. (2003). Anatomical substrates for functional columns in macaque monkey primary visual cortex. *Cerebral Cortex*, Vol. 13, No. 1, 2003, 15-24

MacLennan, B. J. (1999). Field computation in natural and artificial intelligence. *Information Sciences International Journal*, Vol. 119 1999, 73-79

Malach, R. & Amir, Y. (1993). Relationship between Intrinsic Connections and Functional Architecture Revealed by Optical Imaging and in-Vivo Targeted Biocytin Injections in Primate Striate Cortex. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 90, No. 22, 1993, 10469-10473

Marr, H. & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society*, Vol. B-207, 1980, 187-217

Marrocco, R. T. & McClurkin, J. W. (1982). Modulation of Lateral Geniculate-Nucleus Cell Responsiveness by Visual Activation of the Cortico Geniculate Pathway. *Journal of Neuroscience*, Vol. 2, No. 2, 1982, 256-263

Mikami, A. & Newsome, W. T. (1986). Motion selectivity in macaque visual cortex. II. Spatiotemporal range of directional interactions in MT and V1. *Journal of Neurophysiology*, Vol. 55, No. 6, 1986, 1328-39

Murphy, K. M. & Jones, D. G. (1998). Spacing of cytochrome oxidase blobs in visual cortex of normal and strabismic monkeys. *Cerebral Cortex*, Vol. 8, No. 3, 1998, 237-244

Obermayer, K. & Blasdel, B. G. (1993). Geometry of Orientation and Ocular Dominance Columns in Monkey Striate Cortex. *Journal of Neuroscience*, Vol. 13, No. 10, 1993, 4114-4129

Parker, A. J. & Hawken, M. J. (1988). Two-Dimensional Spatial Structure of Receptive-Fields in Monkey Striate Cortex. *Journal of the Optical Society of America a-Optics Image Science and Vision*, Vol. 5, No. 4, 1988, 598-605

Peters, A. & Sethares, C. (1996). Myelinated axons and the pyramidal cell modules in monkey primary visual cortex. *Journal of Comparative Neurology*, Vol. 365, No. 2, 1996, 232-255

Rockland, K. S. & Knutson, T. (2001). Axon collaterals of Meynert cells diverge over large portions of area V1 in the macaque monkey. *Journal of Comparative Neurology*, Vol. 441, No. 2, 2001, 134-147

Rockland, K. S. & Lund, J. S. (1983). Intrinsic Laminar Lattice Connections in Primate Visual-Cortex. *Journal of Comparative Neurology*, Vol. 216, No. 3, 1983, 303-318

Rockland, K. S. & Saleem, K. S. (1994). Divergent Feedback Connections from Areas V4 and Teo in the Macaque. *Visual Neuroscience*, Vol. 11, No. 3, 1994, 579-600

Rockland, K. S. & Vanhoesen, G. W. (1994). Direct Temporal-Occipital Feedback Connections to Striate Cortex (V1) in the Macaque Monkey. *Cerebral Cortex*, Vol. 4, No. 3, 1994, 300-313

Roelfsema, P. R. & Lamme, V. A. F. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, Vol. 395, No. 6700 376-381

Rolls, E. T. & Cowey, A. (1970). Topography of the retina and striate cortex and its relationship to visual acuity in rhesus monkeys and squirrel monkeys. *Experimental Brain Research*, Vol. 10, No. 3, 1970, 298-310

Rossi, A. F. & Desimone, R. (2001). Contextual modulation in primary visual cortex of macaques. *Journal of Neuroscience*, Vol. 21, No. 5, 2001, 1698-1709

Sanger, T. D. (1988). Stereo disparity computation using Gabor filters. *Biological Cybernetics*, Vol. 59, 1988, 405-418

Sceniak, M. P.; Hawken, M. J.; Shapley, R. (2001). Visual spatial characterization of macaque V1 neurons. *J Neurophysiol.*, Vol. 85, No. 5, 2001, 1873-1887

Schiller, P. H. & Finlay, B. L. (1976). Quantitative Studies of Single-Cell Properties in Monkey Striate Cortex .2. Orientation Specificity and Ocular Dominance. *Journal of Neurophysiology*, Vol. 39, No. 6, 1976, 1320-1333

Schiller, P. H. & Finlay, B. L. (1976). Quantitative studies of single-cell properties in monkey striate cortex. III. Spatial frequency. *J Neurophysiol*, Vol. 39, No. 6, 1976, 1334-51

Sheridan, P.; Hintz, T.; Alexander, D. (2000). Pseudo invariant transformations on a hexagonal lattice. *Image Vision Comput.*, Vol. 18, No. 11, 2000, 907-917

Sheridan, P. E. (2007). A Method to Perform a Fast Fourier Transform with Primitive Image Transformations. *IEEE Transactions on Image Processing*, Vol. 16, No. 5, 2007, 1355-1369

Siegel, R. M. (1990). Nonlinear dynamic system-theory primary visual cortical processing. *Physica*, Vol. 42, No. 1-3, 1990, 385-395

Sillito, A. M. & Jones, H. E. (1996). Context-dependent interactions and visual processing in V1. *J Physiol Paris.*, Vol. 90, NO. 3-4, 1996, 205-209

Sirovitch, L.; Everson, R.; Kaplan, E.; Knight, B.; O'Brien, E.; Orbach, D. (1996). Modelling the functional organisation of the visual cortex. *Physica*, Vol. D96, 1996, 355-366

Stettler, D. D. & Das, A. (2002). Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron*, Vol. 36, No. 4, 2002, 739-750

Tang, Q., Sang, N. & Zheng, P. (2007). Extraction of salient contours from cluttered scenes. *Pattern Reccognition*, Vol. 40, No. 11, 2007, 3100-3109

Tootell, R. B. H. & Hamilton, S. L. (1988). Functional-Anatomy of Macaque Striate Cortex. *Journal of Neuroscience*, Vol. 8, No. 5, 1988, 1500-1530

Ursine, M.; La, G. E.; Cara (2004). A model of contextual interactions and contour detection in primary visual cortex. *Neural Networks*, Vol. 17, No. 5-6, 2004, 719-735

Van Loan, C. (1992). *Computational Frameworks for the Fast Fourier Transform*, SIAM, Philadelphia, PA.

Wachtler, T. &. Sejnowski, T. J. (2003). Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, Vol. 37, No. 4, 2003, 681-691

Zeki, S. (1993). *A Vision of the Brain*, Blackwell Scientific Publishing, Oxford.

Zipser, K.; Lamme, V.; Schiller, P. (1996). Contextual modulation in primary visual cortex. *J. of Neuroscience*, Vol. 16, No. 22, 1996, 736-738

8

Spectral Analysis of Global Behaviour of C. Elegans Chromosomes

Afef Elloumi Oueslati[1], Imen Messaoudi[1],
Zied Lachiri[2] and Noureddine Ellouze[1]
Unité Signal, Image et Reconnaissance de Formes, Département de Génie Electrique,
[1]Ecole Nationale d'Ingénieurs de Tunis, BP 37, Campus Universitaire,
Le Belvédère, 1002, Tunis,
[2]Département de Génie Physique et Instrumentation
Institut National des Sciences Appliquées et de Technologie, BP 676,
Centre Urbain Cedex, 1080, Tunis,
Tunisie

1. Introduction

Fourier analysis is one of the most useful decomposition into frequency bands to provide a signal's variations and irregularities measure. DNA spectral analysis based on Fourier Transform contributes in the systematic search of special DNA patterns which may correspond to biological important markers. For example, the Fourier harmonic analysis of the occurrence of a base "A" can give us the corresponding frequency with amplitude and a phase without being able to locate it in time. However it is interesting to also detect the moments of "silence" of base "A" i.e. the moments when this base does not exist. Such a representation of Fourier is thus limited with signals which contain transitory elements or evolutions in their spectral contents. For these non stationary signals, the DNA sequences, to highlight the frequency behavior, it becomes necessary to give the frequency the possibility of changes over time. It's the time frequency analysis aim assured by the Short Time Fourier Transform. In fact, the punctual aspect is very important to localize particular regions in chromosomes, to characterize the beginning of a protein coding regions or a nucleosome or its end. By depicting the frequencies by a smoothed STFT, a 2D or 3D spectrogram representation, specific regions appear distinctly. In this paper, we are concerned with the periodicities 3, 6, 9 and 10.5. The periodicity 3 discussed in (Anastassiou, 2001; Berger et al, 2003; Cohanim et al, 2005; Kornberg, 1977; Segal et al, 2006; Susillo et al 2003; Trifonov & Sussman, 1980; Trifonov, 1998; Vaidyanathan & Yoon, 2004) is related with protein coding regions (called exons) in the gene. The periodicity 10.5 is related with nucleosome's positions in the DNA sequence and the degree of deformability of the sequence in the DNA helix (Hayes et al, 1990; Trifonov & Sussman, 1980; Widom, 1996; Worcel et al 1981). The periodicity 6 and 9 are specific to C. Elegans organism.

This chapter is divided in five parts. First, we expose an introduction for relevant regions on chromosomes. In part three, we detailed the DNA's sequence analysis approach, related to sequence global behavior problem. It exposes the spectral analysis, which follows a certain methodology that generates results to highlight the periodicities studied. This analysis is based on organisms translated into signals by three coding techniques. The algorithm steps of this technique are detailed to mention the generation method of spectrums and spectrograms. The fourth part deals with the study of the frequencies' evolution. It present results for smoothed STFT, as a 1D, 2D or 3D spectrogram representation. Part five concludes this chapter.

2. The relevant regions in chromosomes

The specific succession in the bases (A, G, C, and T) constitutes the hereditary message. Each DNA fragment involves a specific protein synthesis process. Proteins are synthesized from a set composed of 20 different amino acids, which are determined by three bases occurring in subsequent order. A group of three consecutive nucleotides with desoxyribose and phosphoric group is called a codon and a total of 64 different combinations specify 20 amino acids and three stop codons, namely TAA, TAG, and TGA. The protein synthesis (Fig.1) is realized in two steps: (1) the transcription within which the hereditary information is copied into the messenger RNA and, (2) the translation in which the messenger RNA is exploited by the ribosome to form the amino acid chain. To obtain numerical data from this succession of symbolic bases of a DNA sequence, we use binary indicator coding techniques.

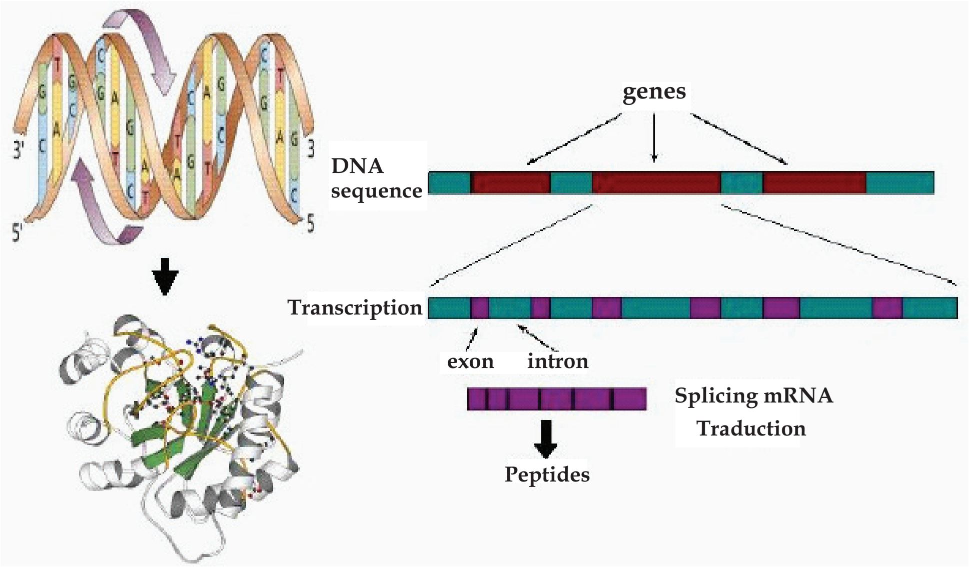


Fig. 1. The protein's synthesis steps

In a DNA sequence, electron microscopy and biochemical studies have established that the bulk of the chromatin DNA is compacting into repeating structural units, named nucleosomes. A model of this DNA structure in such regions is proposed by Kornberg in (Kornberg, 1974, 1977). The chromatin is a dynamic structure, oscillating between the nucleosome and open structures depending on the environmental conditions (Kornberg, 1974, 1977; Oudet et al, 1978). And each nucleosome is formed by two molecules of each histone (protein) H2A, H2B, H3 and H4. Each nucleosome has a diameter of 12.5±1 nm and contains about 200 base pairs of DNA. This number is varying according to the chromatin’s origin (Hayes et al 1990; Kornberg, 1977; Oudet et al, 1978; Worcel et al 1981). In contrast a particle named ‘nucleosome core’ is invariant in its DNA content about 146 base pairs. Interesting electron microscopic evidence elaborated in (Oudet et al, 1978) suggests that under appropriate conditions a nucleosome could open up into two separate half nucleosomes of diameter 9.3±1 nm. The finding of each type of histones in the nucleosome has suggested that a nucleosome could be made up of two symmetrical halves (Altenburger, 1976).

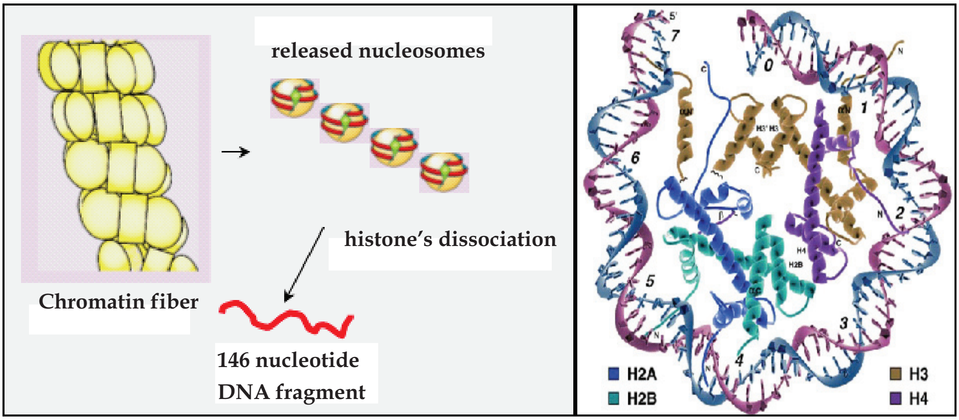


Fig. 2. Chromatine’s and nucleosome’s structure

In order to study the protein coding regions signals and the nucleosome regions ones, the DNA symbolic data must be converted to DNA signals.

3. Genomic sequence analysis based on Short Fourier Transform

In order to give frequencies more precise location in time, Gabor proposes to use a Fourier local analyze with windows. The technique consists in segmenting signal by multiplication by sliding window of fixed length (Mallat, 1999). Each part is analyzed independently with a classic Fourier transform to enhance frequencies behavior. The totality of these transforms forms the short Fourier transform and precise the frequencies location in time.

Applying coding process, the numerical signals are obtained by base’s succession description as follows:

$$x[n] = \{x(i), i \in [1,..,N]\} \tag{1}$$

The classic discrete Fourier transform related to numerical sequence is expressed as:

$$X[k] = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}nk} \tag{2}$$

In order to locate the signal frequencies in time, the analysis is applied to sequence’s parts generated by multiplication with a sliding analysis window.

For this purpose, the numerical signal x[n] is divided into frames of N length. The expression become

$$x_w[n,i] = x[n].\omega[i - \Delta n] \tag{3}$$

When based on the binary indicator 'A', the equation becomes:

$$x_{Aw}[n,i] = U_A[n].\omega[i - \Delta n] \tag{4}$$

With i is the window's order and the Δn is the adopted sliding value. The window's length must be chosen to have an appropriate number of samples to guarantee the best frequency resolution. On each block x_w [n], is applied a Fourier transform to determine Xw [k], k∈[0:N-1], k represents the frequency index. The FT expression associated with each frame is as follows:

$$X_\omega^i[k] = \sum_{n=0}^{N-1} x_\omega[n,i] e^{-j\frac{2\pi}{N}nk} \tag{5}$$

With binary indicator 'A' coding, the equation is:

$$X_A[k] = \sum_{n=0}^{N-1} x_{A\omega}[n,i] e^{-j\frac{2\pi}{N}nk} \tag{6}$$

On the basis of this expression, many representations can be obtained. The sequence is associated to chromosome, the first analyze consists in studying the frequency global behavior. To enhance the frequencies, we used a mean smoothed spectrum. The principle consists in calculating the mean of the obtained spectrum of equation.

$$\bar{X}_\omega^j[k] = \frac{1}{N} \sum_{i=0}^{N-1} X_\omega^i[k] \tag{7}$$

The chromosomes are generally constituted by more than 10 Mbp, so the obtained spectrum needs to be smoothed. A second mean of the mean spectrums is applied. The converted DNA sequence x[n] is divided into frames of M length with an overlap Δm. Each of these frames is also divided into N frames by multiplication with a sliding analysis window w[n]. On each part, a mean smoothed spectrum is generated. Finally, the mean of the spectrum for all the parts is calculated. The final expression of the spectrum is:

$$\bar{X}[k] = \frac{1}{M} \sum_{j=0}^{M-1} \bar{X}_\omega^j[k] \tag{8}$$

3.1 The chromosomes coding techniques

This analysis aims to study the chromosome's frequency global behaviour. For this purpose, it is important to enhance particularly the signals generated by the protein coding regions and the nucleosome regions. That's why, three types of coding techniques are considered:

- A linear coding based on Binary indicators which is related to the base succession,
- A Structural coding with Pnuc, which is an experimental coding based on the helix deformability
- A two-dimensional coding based on Frequency Chaos Game Representation which has submerged from the field of physics known as 'chaotic dynamical systems'

3.1.1 Binary indicator's techniques

The linear coding consists in attributing a binary value for each unit of the all indicators. Which are included in {’A’,’T’,’C’,’G’, ’TT’, ’TA’, ’GC’', ‘AAA’… ‘GGG’}. The marker associated takes the value of either 1 or 0 at location n for the first character, depending on whether or not the corresponding character group exists from the location n.

Base’s binary indicator:

$$S[n] = \sum_{b \in B} U_b[n] \tag{9}$$

Where:

$$U_b[n] = \begin{cases} 1 \; if \;\; base\; b\; is\; in\; position\; n \\ \quad 0\; else \end{cases} \tag{10}$$

is the binary indicator of the base B={A,T,C,G}

Considering sequence S_{DNA} and $U_A[n]$ the associated base’s binary indicator
S_{DNA} = ’**AA**TCGCG**A**C**A**CTC**A**TTCGG’
$U_A[n] = 1\,1\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0$

Two Base’s binary indicator:

$$S[n] = \sum_{bb \in BB} U_{bb}[n] \tag{11}$$

where

$$U_{bb}[n] = \begin{cases} 1 \; if \;\; base\; bb\; is\; in\; position\; n \\ \quad 0\; else \end{cases} \tag{12}$$

is the binary indicator of the base B

B={AA,AT,AC,AG,TA,TT,TC,TG,CA,CT,CC,CG,GA,GT,GC,GG}

Considering sequence S_{DNA} and $U_{CG}[n]$ the associated base’s binary indicator
S_{DNA} = ’AAT**CGCG**ACACTCATT**CG**G’
$U_{CG}[n] = 0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0$

Some dinucleotides as ‘AA’, ’TT’, ‘TA’ are enhancing the ADN flexibility around histones to constitute nucleosomes (Fig. 3).

Codon’s binary indicator: the three bases association called triplet or codon have a fundamental role in the process of amino acids fabrication. For these reasons, a coding technique based on these base's association is used. We adopt binary indicators to each of the 64 codons (Table 1)

$$S[n] = \{U_{cod}[i], i = 1\ldots N_s\} \tag{13}$$

where:

$$U_{cod}[i] = \begin{cases} 1 \; \textit{if the codon cod starts at position n} \\ 0 \; \textit{else} \end{cases} \tag{14}$$

is the binary indicator of the codon cod and Ns is the sequence's length. This marker takes the value of either 1 or 0 at location n for the first character depending on whether the corresponding character exists from the location n. Let's consider the codon binary indicator $U_{TCG}[n]$.

Considering sequence S_{DNA} and the associated codon's binary indicator
S_{DNA} = 'AAT**CGCG**ACACTCATT**CG**G'
$U_{TCG}[n]$ = 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0

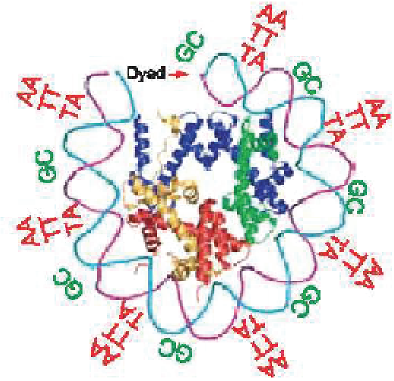


Fig. 3. DNA flexibility around histones is enhanced by dinucleotide as 'AA', 'TT', 'TA'

BASE	Codon associated
A	AAA, AAT, AAC, AAG, ATA, ATT, ATC, ATG ACA, ACT, ACC, ACG, AGA, AGT, AGC, AGG
T	TAA, TAT, TAC, TAG,TTA, TTT, TTC, TTG TCA, TCT, TCC, TCG, TGA, TGT, TGC, TGG
C	CAA, CAT, CAC, CAG, CTA, CTT, CTC, CTG CGT, CGC, CGG, CCA, CCT, CCC, CCG, CGA
G	GAA, GAT, GAC, GAG, GTA, GTT, GTC, GTG GCA, GCT, GCC, GCG, GGA, GGT, GGC, GGG

Table 1. Codon associated to each base

3.1.2 Pnuc: the structural coding techniques

The second coding technique is the Pnuc which is based on local bending and flexibility properties of the double helix; it is deduced experimentally from nucleosome positioning (Pnuc). By considering the matching of both stalks (A-T and C-G) along the helix, one base's pair defines a plane and a direction in this plane. A description of the double helix shows the overlapping of the plans (Fig. 4). When considering that the planes are parallel, passing between planes needs translation and rotation of 34,3° of the orientation of the connection of the plan.

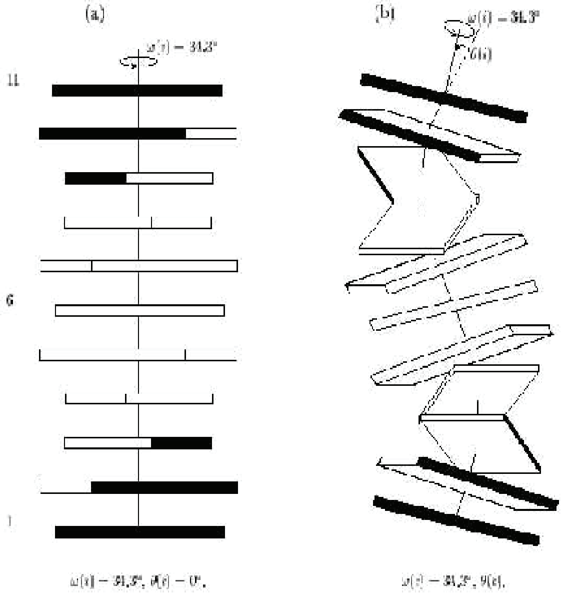

(a)
(b)
$\omega(i) = 34.3°$
$\omega(i) = 34.3°$
$\theta(i)$
11
6
1
$\omega(i) = 34.3°, \theta(i) = 0°,$
$\omega(i) = 34.3°, \theta(i),$

Fig. 4. A description of the double helix shows the overlapping of the plans

Now the plans are not parallel and the axis of the double helix presents curvature. By considering the interaction between a protein, a histone and a DNA's sequence, this interaction is stronger when the contact area between both objects is the biggest. To increase this surface, it is necessary to roll up as much as possible the segment of DNA around the protein, in this way, we have two properties:

If the segment of DNA is not rolled up around the protein, it is in position of equilibrium, the curvature is static

The stalk must be flexible to allow the additional curvature around the protein. These two properties generate the nucleosome which generates an excessive curvature of the stalk.

Each trinucleotide is replaced by its numerical value given by the Pnuc table. The S_{DNA} is then replaced by the numerical sequence C_{PNUC} .

S_{DNA} = 'AAT**CGCG**ACACTCATT**CG**G'
C_{PNUC} = 0.7 5.3 8.3 7.5 7.5 6.0 5.4 5.2 6.5 5.8 5.4 5.4 6.7 0.7 3.0 8.3 4.7

Trinucleotide	PNUC	Trinucleotide	PNUC
AAA/TTT	0.0	CAG/CTG	0.042
AAC/GTT	0.037	CCA/TGG	0.054
AAG/CTT	0.052	CCC/GGG	0.060
AAT/ATT	0.07	CCG/CGG	0.047
ACA/TGT	0.052	CGA/TCG	0.083
ACC/GGT	0.054	CGC/GCG	0.075
ACG/CGT	0.054	CTA/TAG	0.022
ACT/AGT	0.058	CTC/GAG	0.054
AGA/TCT	0.033	GAA/TTC	0.030
AGC/GCT	0.075	GAC/CTG	0.054
AGG/CCT	0.054	GCA/TGC	0.060
ATA/TAT	0.028	GCC/GGC	0.0100
ATC/GAT	0.053	GGA/TCC	0.038
ATG/CAT	0.067	GTA/TAC	0.037
CAA/TTG	0.033	TAA/TTA	0.020
CAC/GTG	0.065	TCA/TGA	0.054

Table 2. The PNuc table

The signal generated from this coding for a part of chromosome is given by Fig. 5. For clarity purpose the signal is multiplied by 10. Fig. 6 illustrate the stft method applied on this resulting signal. First, subfigure a shows a mean spectrum for distinct window of length $5*10^5$.. The spectrum obtained needs smoothing so for the second figure (subfigure b) a blackman smoothing window is applied on each signal part before calculating the mean spectrum of equation 7. In the third and last figures (subfigure c and d) the equation 8 is used and the parameters chosen are: Blackman window, $M=5*10^5$, $N=5*10^4$ and overlap 50% for subfigure c and $N=5*10^3$ with overlap 10% for subfigure d. The figure shows that meaning and smoothing are very efficient to have the best signal (subfigure d). In this signal, the periodicity 10 is enhanced to prove that this is a characteristic of helix flexibility.

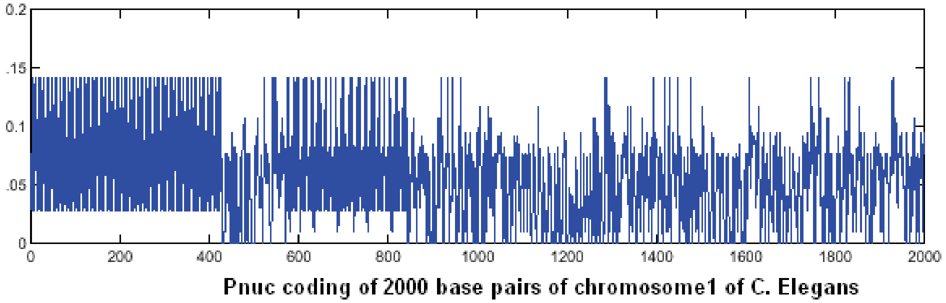
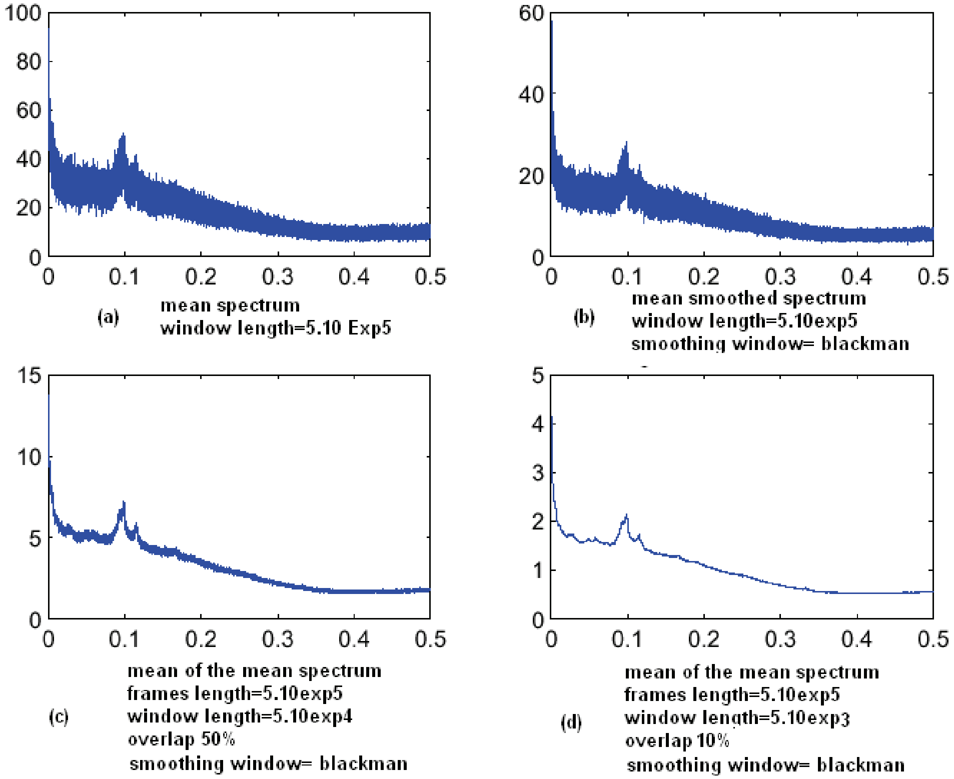


Fig. 5. Pnuc signal of 2000 base pairs of chromosome 1 of C. Elegans genome



(a) mean spectrum
window length=5.10 Exp5

(b) mean smoothed spectrum
window length=5.10exp5
smoothing window= blackman

(c) mean of the mean spectrum
frames length=5.10exp5
window length=5.10exp4
overlap 50%
smoothing window= blackman

(d) mean of the mean spectrum
frames length=5.10exp5
window length=5.10exp3
overlap 10%
smoothing window= blackman

Fig. 6. Illustration of the smoothed mean spectrum applied on Pnuc signal of 2000 base pairs of chromosome 1 of C. Elegans genome

3.1.3 Fcgr: the two dimensionnal coding techniques

The third technique is submerged from the Chaos Game Representation (CGR) images which can forms a global signature of bio-sequences (Almeida et al, 2001; Cenac et al, 2004; Deshavanne et al, 1999, 2000; Joseph & Sasikumar, 2006; Oliver et al 1993; Fiser et al, 1994). The CGR paradigm is a holistic way of DNA representation. It provides a unique scatter pictures. In 1999, H. Joel Jeffrey uses for the first time this representation for studying the "non-randomness" of genomic sequences (Jeffrey, 1990). The CGR is an iterative algorithm for drawing fractal images to any desired scale. It maps nucleotide sequences in the [0,1]x[0,1] square. The four letters A, C, G and T are placed at the corners. The binary CGR vertices are assigned to the four nucleotides as:

$$l_A = (0,0), l_C = (0,1), l_G = (1,1), l_T = (1,0) \tag{15}$$

Deriving scatter pictures, the CGR's construction algorithm consists of three steps. First, the four letters A, C, G and T are placed at the corners of a rectangular unit square. Second, the first point is plotted halfway between the center of the square, and the corner corresponding to the first nucleotide of the sequence. Third, the new point are marked successively half way between the previous point and the corner corresponding to the base of each nucleotide read from the sequence (Almeida et al, 2001; Joseph 2006). A generated CGR image can be viewed as an image of distributed dots. Subdividing the unit square into a set of square entries of equal size n, the number of square entries obtained is equal to 2n ×2n. The number of points counted in each sub-square represents the number of occurrence of a particular n-lengthen pattern.

For illustration, let's consider a DNA sequence $S=\{S_1,S_2,....,S_N\}$ of N nucleotides, the CGR value along this sequence is defined by equation 16. The result will be a square uniformly and randomly filled with dots.

$$X_{n+1}=\frac{1}{2}\left(X_n+l_{s_{n+1}}\right) \tag{16}$$

The first point X_0 is usually placed at the center of the square having thus the coordinates (0.5, 0.5). Then, the next point X_{n+1} is repeatedly placed halfway between the previous plotted point X_n and the segment joining the vertex corresponding to the letter s_{n+1} of the sequence. Fig. 7 illustrates the construction process of CGR trajectory for sequence "ATCGG".

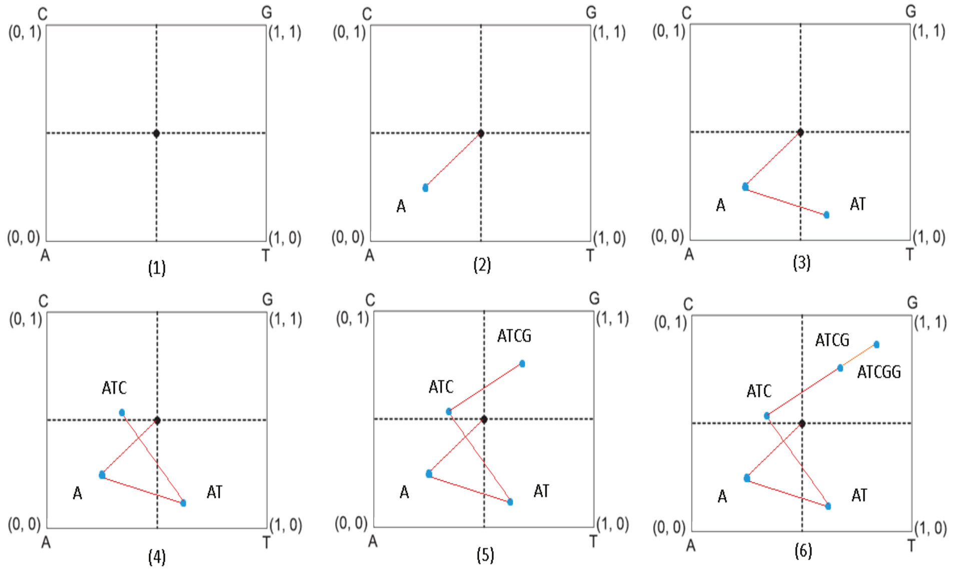


Fig. 7. An illustration of CGR trajectory for sequence "ATCGG"

To derive the CGR plot, the following steps are taken: First place X_0 at the square's center and the four letters at the corners as described before (subfigure 1). From center to vertex A, mark midpoint 1 (address A) (subfigure 2). From 1 to T, mark midpoint 2 (address AT) (subfigure 3). From 2 to C, mark midpoint 3 (address ATC) (subfigure 4). From 3 to G, mark midpoint 4 (address ATCG) (subfigure 5). From 4 to G, mark midpoint 5 (address ATCGG) (subfigure 6).

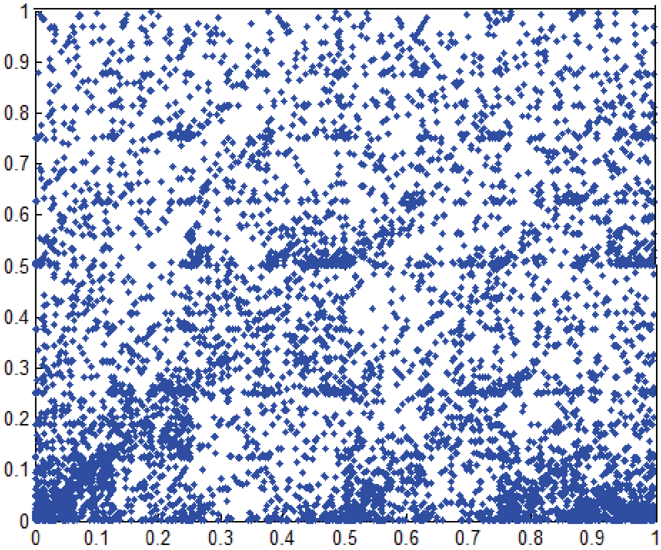


Fig. 8. Chaos Game Representation of the C. Elegans's gene F56F11.4

By identifying local patterns displayed in the CGR square, it is possible to identify correspondent features of DNA sequences (Yu et al, 2008). The fractal nature of this kind of DNA representation can be observed Fig. 8. The clustering dots in the lower corners indicate a slightly high concentration in A and T. It is known that CGR patterns depict base composition. In fact, we divide the CGR space with a grid of size k (i.e ($2^k \times 2^k$) pixels) and we count occurrence in each quadrant, the frequency of k-lengthen words occurrence can be estimated and the frequency matrix then extracted is called FCGR (Frequency Chaos Game Representation) (Almeida et al, 2001; Deshavanne et al, 2000; Jeffrey, 1990).

The FCGR was first investigated by Deschavanne in (Deshavanne et al, 1999) and later by Almeida in (Almeida et al, 2001). To show the frequencies of the K-tuples, a color scheme normalized to the distribution of frequency of occurrence of associated patterns is used (Joseph & Sasikumar, 2006; Oliver et al, 1993; Tavassoly, 2007a; Tavassoly, 2007b; Makula, 2009; Goldman, 1993; Cénac, 2006; Tino, 1999; reference 44). A grayscale color mapping may also be used. In Fig. 9, the dinucleotide and trinucleotide frequency matrices (k ={2,3}) are obtained for the gene F56F11.4 of C.elegans. Thus, 2^2x2^2=16 cells are needed for motifs of length two and 2^3x2^3=64 regions to count motifs of length 3. The darker pixels represent the most frequently used words; when the clearest ones represent the fewer used words. CGRs were used for displaying the behavior of sub-patterns within the same input sequence and depicting oligo_mer composition. It forms the basis for similarity and self-similarity algorithms in a different way from traditional alignment of nucleotides.

This FCGR cannot follow the evolution of frequencies from the beginning to the end of a given sequence. So, we propose to generate signals from FCGR. We Generate the nth-order FCGR for the hole sequence, and we replace the reading the first n-lengthen word in the sequence, by the correspondent frequency of the same sub-pattern in the FCGRn matrix.

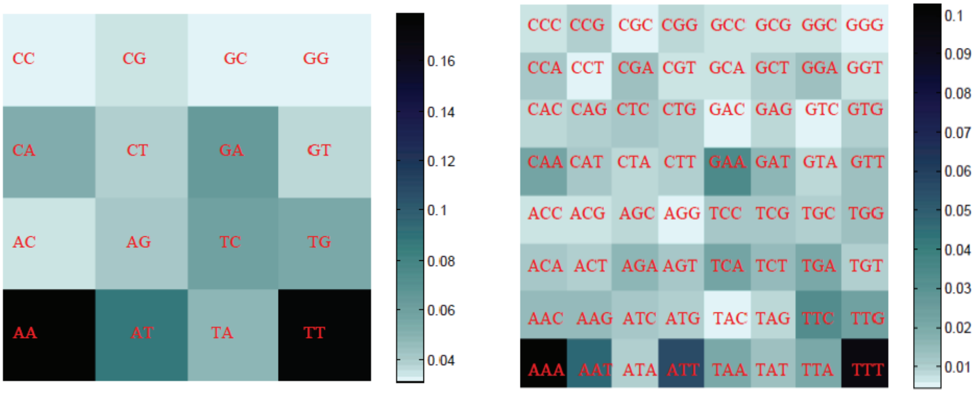


Fig. 9. The FCGR2 (k=2) and the FCGR3 (k=3) for the gene F56F11.4 of C. Elegans

Generating signals from FCGRs was a good way to capture such variability. For this fact, a new 1D graphical representation of DNA sequences is introduced, which provide useful insights into local and global characteristics of genomic sequences. This novel algorithm of DNA coding consists of computing the kth-order FCGR for the whole sequence and assigning then the value of the correspondent frequency to each k-lengthen word in the sequence. Thus allows us to follow the frequencies' evolution along a given sequence. The the obtained plot set is called FCGR$_k$-signal.

Let's We consider the given sequence S_{DNA}
S_{DNA} = ' TTTAAAAGCTCGCGCTAAAA'

The given sequence is divided with a k-length sliding window. A set of K-frames are obtained which are denoted by K-mers. For example when k= {2, 3, 6}, we have 2-mers (S_{DNA}), 3-mers (S_{DNA}) and 6-mers (S_{DNA}).

F_K(s) is defined to be the frequencies' set of the k-substrings that appear in the sequence S. Obviously; these frequencies derive from the appropriate FCGR$_k$ matrices. It follows that:

F_2(S_{DNA})= {0.1579, 0.1579, 0.1579, 0.3684, 0.3684, 0.3684, 0.1053, 0.2105, 0.1579, 0.1053, 0.1579, 0.2105, 0.1579, 0.2105, 0.1579, 0.1579, 0.3684, 0.3684, 0.3684}

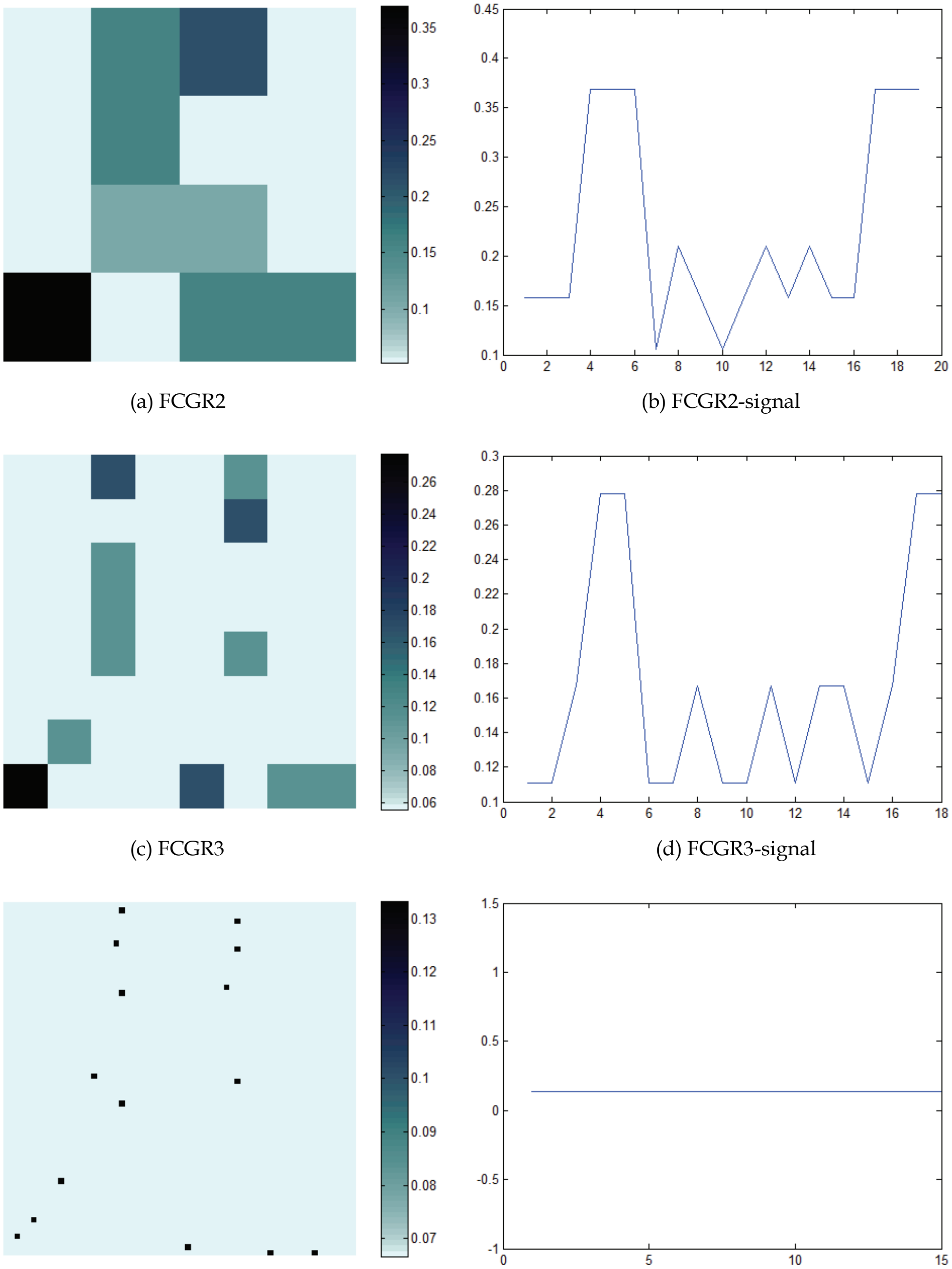
F_3(S_{DNA})= {0.1111, 0.1111, 0.1667, 0.2778, 0.2778, 0.1111, 0.1111, 0.1667, 0.1111, 0.1111, 0.1667, 0.1111, 0.1667, 0.1667, 0.1111, 0.1667, 0.2778, 0.2778}

F_6(S_{DNA})= {0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333}

Fig. 10. illustrates the FCGR drawings for the case of k = {2,3and 6} and the corresponding plot sets for the considered sequence.

Fig. 10. Also shows the slightly high concentration in AA and AAA motifs in FCR2 and FCGR3 which are expressed by the high-rise blocks in the correspondent signals.

On the signals obtained, a spectral analysis is applied to detect the frequency global behaviour in the spectrum for each C. Elegans chromosome.



(a) FCGR2

(b) FCGR2-signal

(c) FCGR3

(d) FCGR3-signal

(e) FCGR6

(f) FCGR6-signal

Fig. 10. FCG representation for k=2, 3 and 6 and the FCGR_signals associated

3.2 The Fourier analysis method steps

The short time analysis is the technique used in order to locate specific regions in a DNA sequence. In this purpose, a mean values of Smoothed Discrete Fourier Transform is applied on sliding window along the DNA sequence to follow the peak's evolution for specific frequencies points. The Fourier analysis algorithm steps are:

The converted DNA sequence x[n] is divided into frames of M length with an overlap Δm. Each of these frames is also divided into N frames by multiplication with a sliding analysis window w[n]:

$$x_w[n,i] = x[n]w[n - i\Delta n] \tag{17}$$

Where i is the window index, and Δn the overlap. The weighting w[n] is assumed to be non zero in the interval [0, N-1]. The frame length value N is chosen in such a way that, on the one hand, the parameters to be measured remain constant and, on the other hand, that there are enough samples of x[n] within the frame to guarantee reliable frequency parameter determination. The choice of the windowing function influences the values of the short term parameters, the shorter the window the greater his influence (Mallat, 1999). We select N and M frame length as power of two to apply the Fast Fourier Transform algorithm.

Each weighted block $x_w[n]$, of the frame is transformed in the spectral domain using Discrete Fourier Transform (DFT), in order to extract the spectral parameters $X_w[k]$, where k represents the index of the frequency ([0, N-1]). The DFT of each frame (in one of M sequence parts) is expressed as follows:

$$X_w^i[k] = \sum_{n=0}^{N-1} x_w[n,i] e^{-j\frac{2\pi}{N}nk} \tag{18}$$

Using the mean values, we calculate a DFT mean value for each frame (1: M). The expression of mean DFT is expressed as:

$$Xm_w^j[k] = \frac{1}{N}\sum_{i=0}^{N-1} X_w^i[k] \tag{19}$$

Where *i* correspond to the index frame of N frames ([1...N]), *k* is the index of the frequency and *j* correspond to the index frame of M frames ([1: M]).
We constitute the matrix

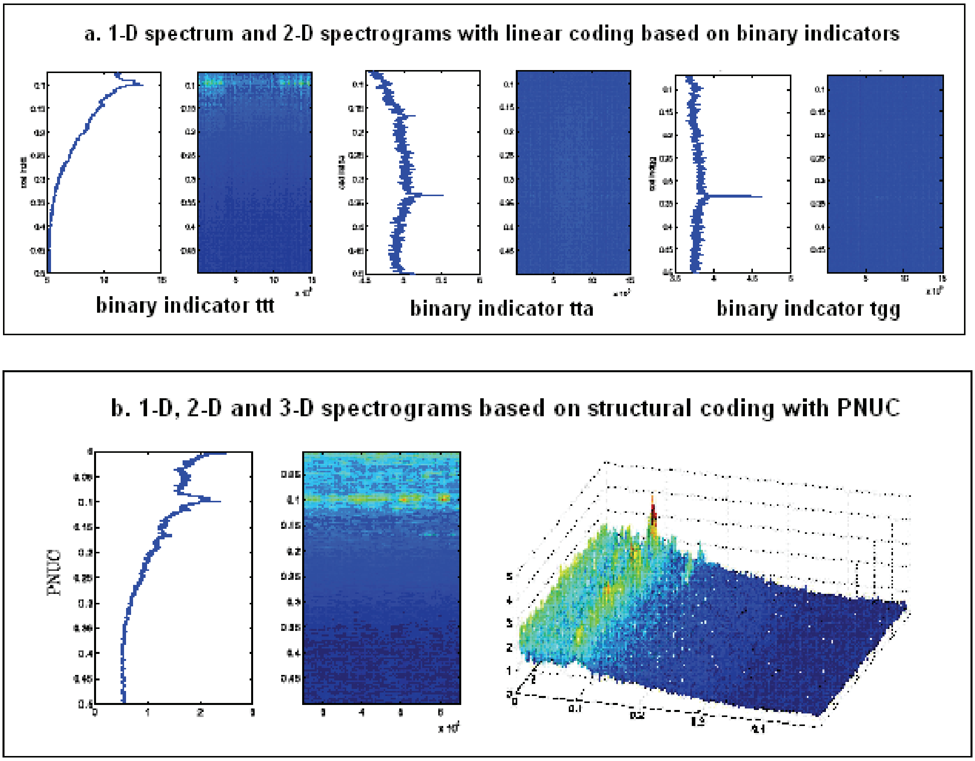
$$MAT(j,k) = Xm_w^j[k] \tag{20}$$

With these obtained values, we can constitute the matrix to represent restricted join time frequency information, known as 2D or 3D DNA spectrograms. This 2D or 3D representation consists of the spectrogram amplitude for a specific index periodicity in a specific nucleotide position in the chromosome.

4. Results

The method has been applied on C. Elegans genome. The chromosomes have been divided on 1- million's parts. The M frames have a length of 1024 bp and an overlap Δm=256, the

N frames of each M frames have length of 256 with Δn =128. The fig. 11 presents some examples for the spectrum related to each of the three coding technique used. In this figure, we show particularly the periodicities 3 and 10 which are closely depending on coding.



a. 1-D spectrum and 2-D spectrograms with linear coding based on binary indicators

b. 1-D, 2-D and 3-D spectrograms based on structural coding with PNUC

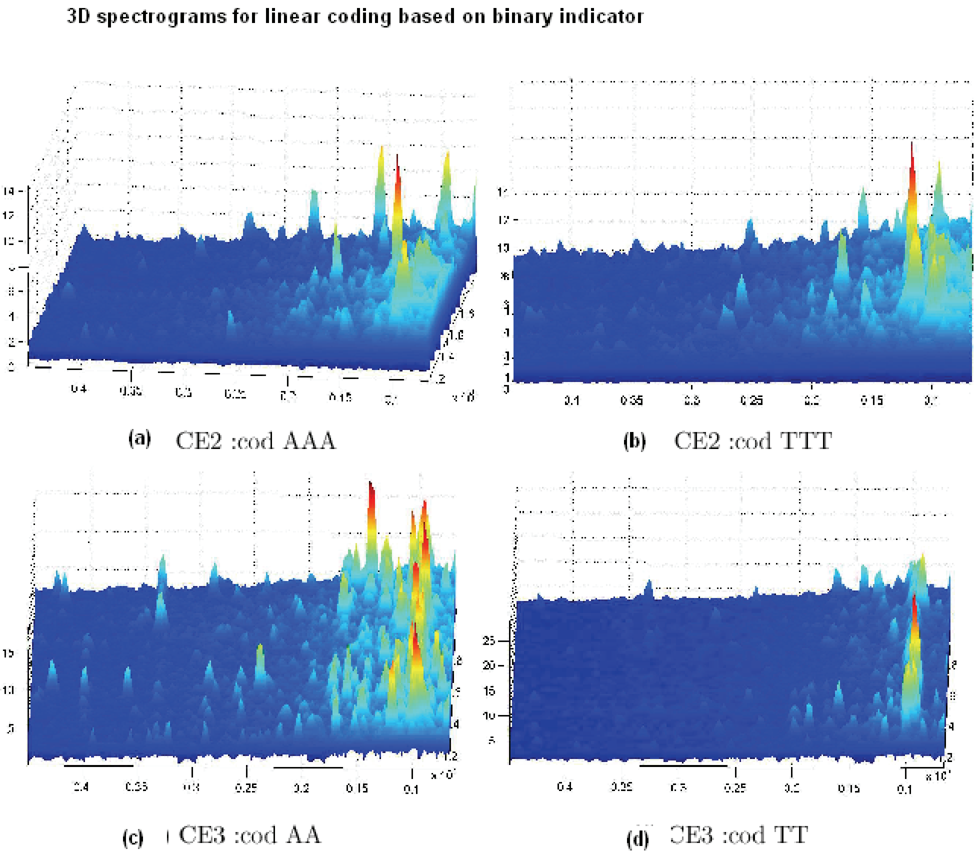
Fig. 11. Examples of spectrums and spectrograms generated with a mean valued technique based on smoothed Discrete Fourier Transform applied on sliding window along the DNA sequence parts of C. elegans genome. Two coding methods are used: a- linear coding technique (binary indicator) (subfigure a), b- structural coding technique (PNUC) (subfigure b)

In order to highlight the various frequencies characteristic of an organism, the tests were carried out with various coding over various sizes of segments and various widths. The example presented in the Table 3 presents the percentage of contribution of the trinucleotides in the highlighting of the various characteristic frequencies at the frequencies 1/3, 1/6.5, 1/9 and 1/10. The table shows that the organism C. Elegans is rich in periodicities and that these periodicities are raised by more than the 3/4 of these coding technique. We notice clearly that for periodicity 3, the rate has raised more 97 %, followed by periodicity 10 which has 90 % and periodicity 9 with 85 %. Periodicity 6.5 is a periodicity which is very marked for this organism 70 % of code contributes to its raising. It translates the existence with a very high rate of 6 bases groups at the periodicity 6. The majority of these groups represent polyA, generally associated for gene purposes.

period	chromosomes				
	Ch1	Ch2	Ch3	Ch4	Ch10
P=3	96.8%	96.8%	96.8%	96.8%	96.8%
P=6.5	64%	60.9%	81.25%	73.43%	71.9%
P=9	85.9%	89%	89.1%	82.81%	79.7%
P=10	70.3%	90.6%	90.5%	90.6%	84.3%

Table 3. The proportion of contribution of the trinucleotides in the highlighting of the various characteristic frequencies

The Fig. 11 presents some spectrum with linear coding based on binary indicator. Each indicator contributes on a specific periodicity enhancement. The ttt binary indicator enhances the periodicity 10 when for the indicators tta et tgg the periodicity 3 is picked up. The 3D spectrograms give more precision on the power's spread around these periodicities. In fact, the peaks in these frequency locations have different power values (Fig. 12).



(a) CE2 :cod AAA

(b) CE2 :cod TTT

(c) CE3 :cod AA

(d) CE3 :cod TT

Fig. 12. 3-D spectrograms for binary indicators coding

The spectrogram 3D adds a third element to the representation 2D. In addition to the localization of the periodicities in the segment, we visualize power associated with each peak. We can distinguish between the peaks which we can find in all the segments for a given periodicity: 10 and 3 and the peaks which are present in certain segments and which were eliminated by carrying out the average

The Fig. 12 is divided on 4 subfigures. Each one add to the 2D spectrograms the power values and locations of the periodicities Enhanced: it represents the 3-D spectrograms. Subfigures (a) and (b) are related to chromosomes2 of C. Elegans when the subfigures (c) and (d) concern chromosome 3.

This figure shows that for the binary indicator 'AA', 'TT', 'AAA' and 'TTT', the peaks around the frequency 1/10.5 are very pronounced. The variation of the degree view angle demonstrates that the peaks are locally spread in the chromosome part. In the literrature, it has been demonstrated both with the biochemical and signal processing studies, that the periodicity 10.5 related to the nucleosomes is varying. That's why, these figures shows in one hand that there is peaks around this periodicity and in the other hand the peaks are spread in specific regions in the chromosome.

The Fig. 13 represents the spectrograms recovered after PNUC coding. The analysis breaks up the chromosome made up of 15,2Mbp into 15 parts of 1Mbp. We find the localization of the periodicity in the ends. In reality, the periodicity peaks are missed or have of very weak power in the sequence going of 6 Mbp with 12 Mbp, it is not localized on the centromer but it is around ends of the helix. We find it towards the position 13 Mbp until the end. In the parts where it exists it is not continuous, it is localized in specific time's lapses.

In Fig. 14 mean valued technique based on smoothed Discrete Fourier Transform was applied along the parts 6, 9 and 13 of the chromosome 1 of C.elegans. From the 1D, 2D and 3D plots, it is observed that coding with FCGR$_2$ reveal the presence of both 10.5 and 3 periodicities. The peaks are spread with different values according to parts around each of these periodicities. Each part has each own specificity. In fact, in part 9 (subfigure a) , periodicities 3 and 10 just submerge from the frequency behavior with peaks of modest values. For the part 6 these periodicities have the same behavior, the specificity is the presence of horizontal peaks around the location 750 in this part. When the part 13 is rich in periodicities 10 and 12 and poor in periodicity 3.

For coding with FCGR-3 (Fig.15), the very pronounced peaks correspond to the 10.5 periodicity; just in the left side other peaks appear around the frequency 0.11 which corresponds to the 9 periodicity; in the right side a few peaks occur around the frequency 1/12. The 3 periodicity disappears in the majority of the parts and when it appears, it is present only on a few areas with very low amplitudes. In Fig 15, we can distinguish between frequency behavior in the three parts represented. The periodicity 10 is more pronounced for the part 16 (subfigure b) when comparing with part 9 (subfigure a) and 10 (subfigure c).

As for the hexamers coding (FCGR$_6$), we find that it enhances the frequency 1/10.5; upon rare zones the frequency 1/12 is observed (Fig.16). We clearly notice that this coding technique enhances the periodicity 10 and his neighbor in opposition to periodicity 3. The three parts shows different aspect of the repartition of the periodicities. In part 9 (subfigure a), the peaks are spread in a "large" frequency band around periodicity. The band is reduced for part 16 (subfigure b) to be located in two frequencies then the power is grouped in one frequency for part 12 (subfigure c).

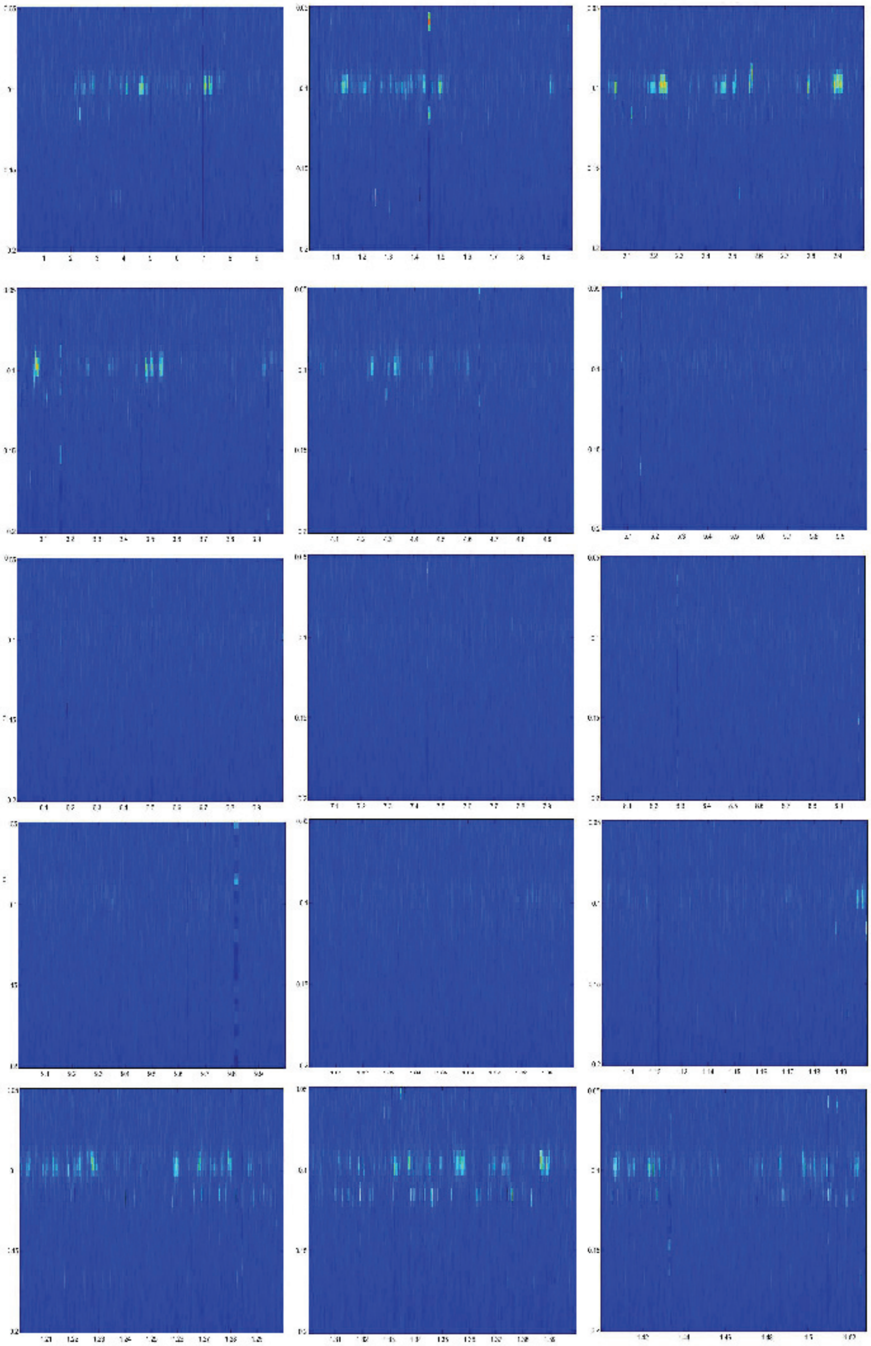
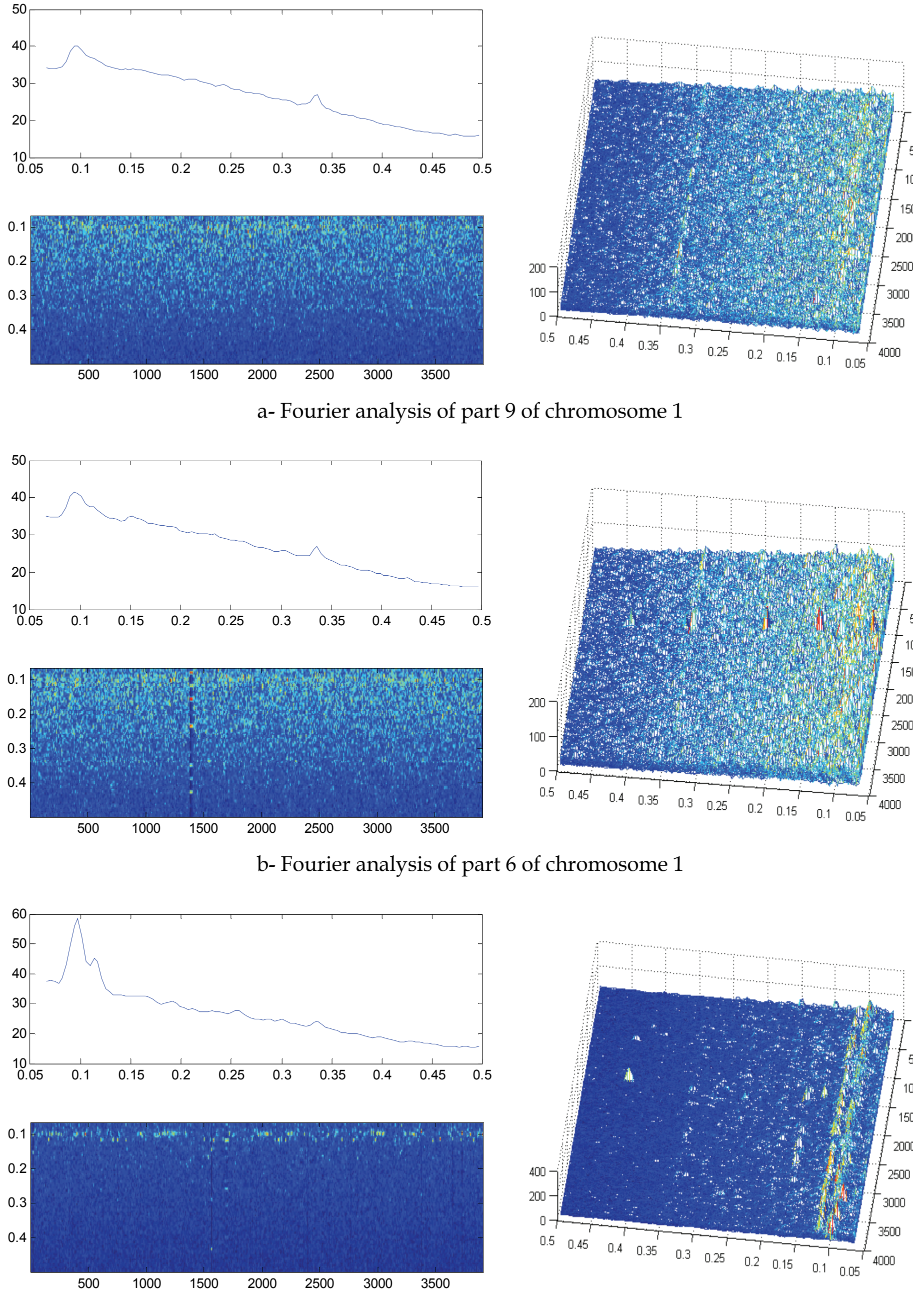


Fig. 13. Distribution of periodicity 10 for coding pnuc along chromosome 2

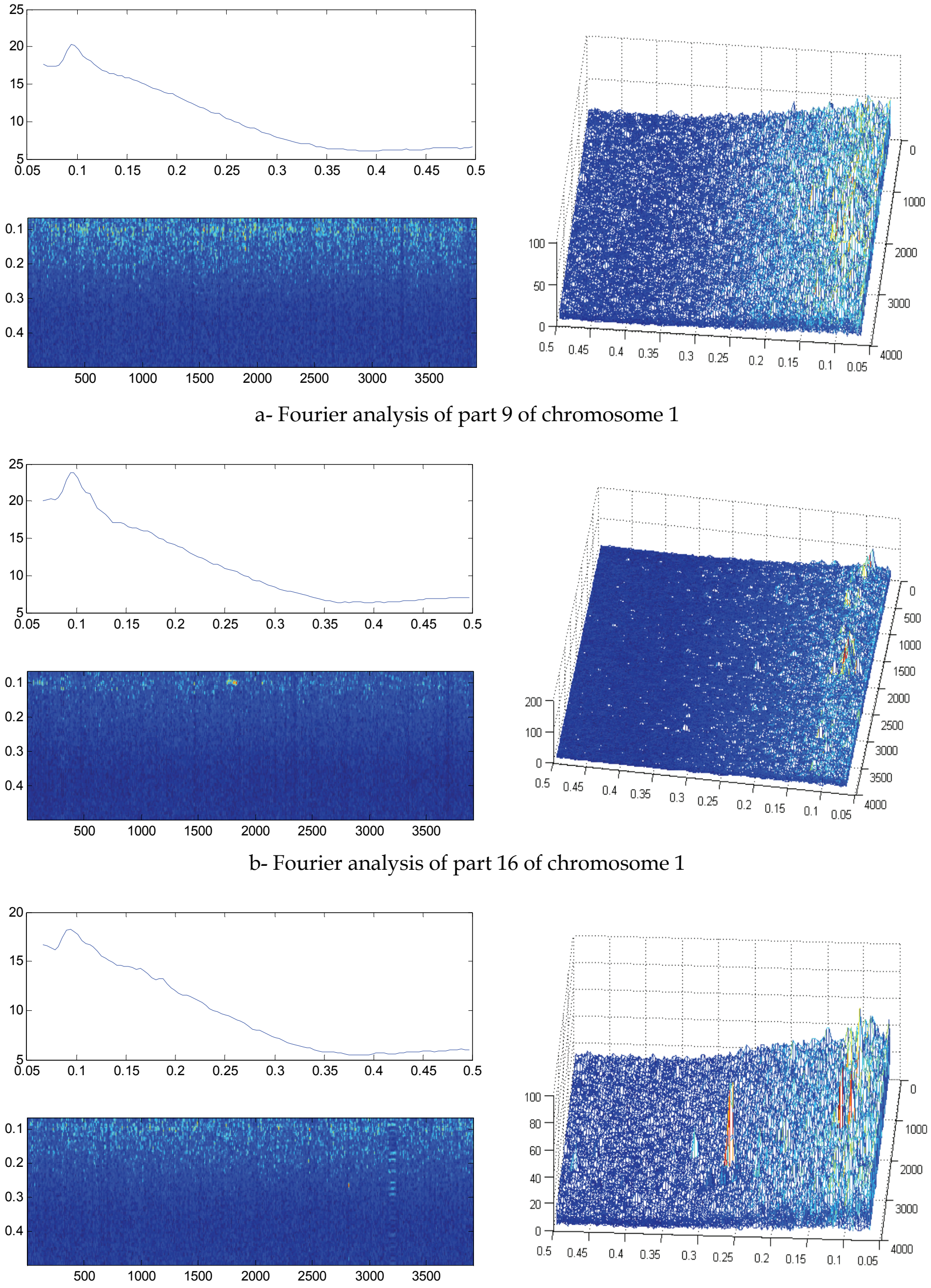


a- Fourier analysis of part 9 of chromosome 1

b- Fourier analysis of part 6 of chromosome 1

c- Fourier analysis of part 13 of chromosome 1

Fig. 14. Examples of spectrums and spectrograms of chromosome's parts with FCGR-2 signal coding

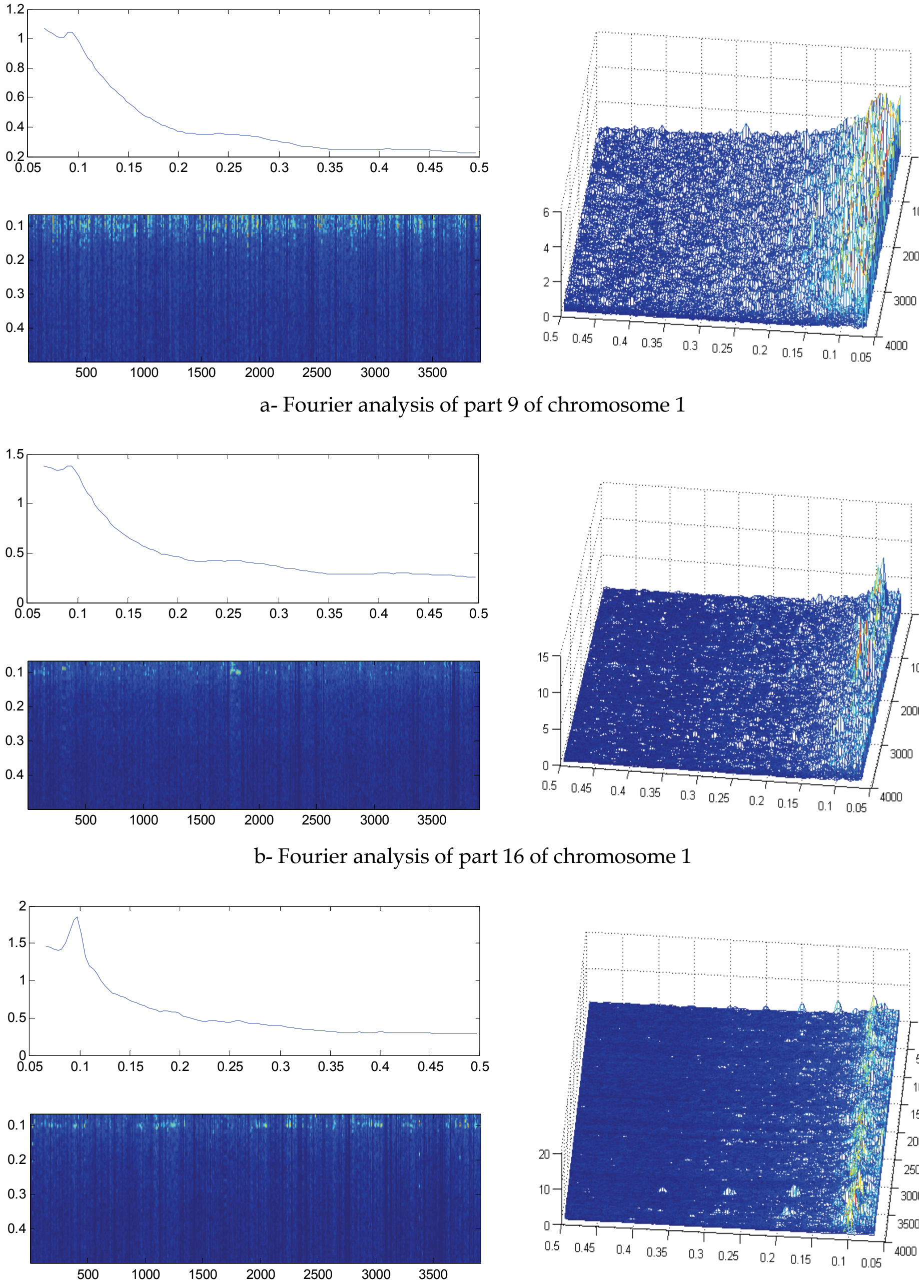


a- Fourier analysis of part 9 of chromosome 1

b- Fourier analysis of part 16 of chromosome 1

c- Fourier analysis of part 10 of chromosome 1

Fig. 15. Examples of spectrums and spectrograms of chromosome's parts with FCGR-3 signal coding



a- Fourier analysis of part 9 of chromosome 1

b- Fourier analysis of part 16 of chromosome 1

c- Fourier analysis of part 12 of chromosome 1

Fig. 16. Examples of spectrums and spectrograms of chromosome's parts with FCGR-6 signal coding

A peak around the frequency 1 / 4 nearby at position 2500 corresponds to a satellite (Fig. 16 subfigure a). This frequency derives from repetitions of certain dinucleotides in the area. The spectrogram reveals the presence of a satellite with multiple frequencies; this is manifested clearly in the 3D graph in the form of horizontally aligned peaks colored in red, the higher frequencies.

5. Conclusion

In This chapter, we investigate the contribution of each coding technique: the linear, the two-dimensional and the structural one in the enhancement of the peaks related to the C. elegans genome periodicities. For this purpose, we use a mean values of smoothed Discrete Fourier Transform applied on sliding window along the DNA sequence to follow the peak evolution for specific frequency points around the frequencies. We detect periodicities around 3, 6, 9 and 10 and found periodicities 3 and 10 related respectively to genes and the positions of the nucleosomes. First we evaluate the frequencies spread through the chromosomes with a 1-D spectrum. Second, we consider the 2-D and 3-D DNA spectrograms to visually detect the specific parts of chromosomes related with protein coding regions, nucleosomes positioning regions, and other particular regions.

The time frequency analysis made it possible to follow the periodicities' evolution. We studied the contribution of a range of binary indicators for the raising of exons' peak frequency. We also studied the localization of the areas being able to form nucleosomes. Thanks to the spectrogram with two dimensions, we visualized the localization of the areas corresponding to periodicity 10 in the limits and not in the center of the helix. The three-dimensional spectrogram showed that the raised peaks do not correspond to the periodicity 10 but we see clearly in certain sequences and for some indicators two lines of peaks of variable powers around this periodicity. This result can explain the variation between 10 and 10.7 of the periodicities associated with the nucleosomes presented in the literature. It is also observable that these peaks are alternated around two periodicities; this result could be associated with the phenomena of chromatin compaction.

6. References

Almeida, J.S., Carrico, J.A., Maretzek, A., Noble P.A. & Fletcher M. (2001) "*Analysis of genomic sequences by Chaos GameRepresentation*", Bioinformatics Vol. 17, n°5, pp 429–437.

Anastassiou D. (2001), "*Genomic Signal processing*", IEEE Signal Processing Magazine, 18 (4), pp: 8-20.

Berger J. A., Mitra S. K.& Astola J. (2003), "*Power spectrum analysis for DNA sequences*", Proc. of ISSPA 2003, pp 29-32, France, 1-4 July.

Cénac, P. (2006) "Étude statistique de séquences biologiques et convergence de martingales", PhD thesis on Applied Mathematics, Paul Sabatier University, Toulouse III, pp 17−25.

Cénac P., Fayolle G., Lasgouttes J.M., (2004) " *Dynamical Systems in the Analysis of Biological Sequences*", research report n° 5351,pp 3–50.

Cohanim, A.B., Kashi Y. & Trifinov E.N. (2005), "*Yeast Nucleosome DNA Pattern: Deconvolution from Genome Squences of S. cerevisiae*, Journal of Biomolecular Structure & Dynamics ISSN 0739-1102, volume 22, Issue Number 6, Adenine Press, pp: 687-693.

Deschavanne, P., Giron, A., Vilain, J. Dufraigne, CH., & Fertil, B. (2000), "*Genomic Signature Is Preserved in Short DNA Fragment*", International Symposium on Bio-Informatics and Biomedical Engineering, IEEE, pp 161–167.

Deschavanne, P., Giron, A., Vilain, J., Fagot, G. & Fertil, B. (1999) "*Genomic signature: characterization and classification of species assessed by chaos game representation of sequences*", Mol Biol E, Vol 16, n°10, pp 1391–1399.

Fiser, A., Tusnady, G.E.& Simon, I.(1994) "*Chaos game representation of protein structures*", J.Mol Graphics, Vol 12, pp 295,302–304.

Fukushima, A., Ikemurab, T., Kinouchie, M., Oshima, T., Kudod, Y., Morig, H. & Kanaya, S. (2002) "*Periodicity in prokaryotic and eukaryotic genomes identified by powerspectrum analysis*", Elsevier, Gene 300, pp 203–211.

Goldman, N. (1993) "*Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences*", Nucleic Acids Research Vol.21, n°10, pp 2487–2491.

Godsell, D.S. & Dickerson, R.E. (1994) "*Bending and curvature calculations in b-dna*", Nucl. Acids Res, vol 22, pp 5497-5503.

Hayes, J.J., Tullius, T.D. & wolffe, A. P. (1990) "*The structure of DNA in a nucleosome*", Proceedings of the National Academy of sciences of the United States of America, vol 87 No 19, pp 7405-7409, October.

Jeffrey, H.J., (1990) "*Chaos game visualization of sequences*", Computers & Graphics, Elsevier, Vol.16, n°1, pp 25–33.

Joseph, J. & Sasikumar, R. (2006) "*Chaos game representation for comparison of whole genomes*", BMC Bioinformatics, Vol 7, n°1, pp 1-10..

Kornberg, R.D. (1974), "*Chromatin structure: a repeating unit of histones and DNA.*" Science 184, pp:868-871,

Kornberg, R.D. (1977), "*Structure of Chromatin*", Annu Rev Biochem. 46 , pp 931-954.

Makula, M., (2009) "*Interactive visualization of oligomer frequency in DNA*" Computing and Informatics, Vol. 28, pp 1001–1016.

Nicorici, D., Berger, J. A. , Astola, J. & Mitra, S. K. ,(2003), "*Finding borders between coding and non coding DNA regions using recursive segmentation and statistics of stop codons", Finnish Signal Processing Symposium (FINSIG'03)*, Tampere, Finland, pp. 231-235.

Oliver, J. L., Bernaola-Galvan, P., Guerrero, G. & Foman-Roldan, R. (1993), "*Entropic profiles of DNA sequences through chaos-game-derived images,*" J. Theor. Biol.,Vol 160, n°4, pp 457–470.

Oppenheim, A. V., Schafer ,R. W. & Buck, J. R., (1999) "*Discrete Time Signal Processing*", 2nd Edition, Prentice Hall.

Oudet, P., Germond, J.E., Bellard, M., Spadafora, C. & Chambon, P. (1978) "*Structure of Eucaryotic Chromosomes and Chromatin",* Phylosophical Transactions of the Royal Society of London. Series B, Biological Sciences, vol 283, No 997, pp: 241-258,

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thamstrom, A., Field, Y., Moore, I. K., Wang, J.P.Z. & widom, J. (2006) " *a genomic code for nucleosome positioning, nature* vol 442, pp: 772-778, August

Sussillo, D., Kundaje, A. & Anastassiou, D., 2003 "*Spectrogram analysis of genomes,*" *Eurasip* Journal of Applied Signal Processing, vol. 2003, no. 4, .

Tavassoly, I., Tavassoly, O., Rad, M.S.R & Dastjerdi, N.M. (2007a) "*Multifractal Analysis of Chaos Game Representation Images of Mitochondrial DNA*", Frontiers in the Convergence of Bioscience and Information Technologies FBIT, IEEE, pp 224–229.

Tavassoly, I., Tavassoly, O., Rad, M.S.R & Dastjerdi, N.M. (2007b) "*Three dimensional Chaos Game Representation of genomic sequences*", Frontiers in the Convergence of Bioscience and Information Technologies FBIT, IEEE, pp 219–223.

Tino, P. (1999) "*Spatial representation of symbolic sequences through Iterative Function System*", IEEE, Vol 29, n°4, pp 386–393.

Trifonov, E. N. & Sussman, J.L. (1980) " *The Pitch of chromatin DNA is Reflected in its Nucleotide Sequence*", Proceedings of the National Academy of Sciences of the United States of America, Vol 77, No 7, part 2: Biological Sciences, pp:3816-3820.

Trifonov, E. N. (1998), "3-*, 10.5-, 200- and 400-base periodicities in genome sequences*", Elsevier Physica A 249, pp :511-516,.

Vaidyanathan, P. P. & Yoon, B. J. (2004) "*The role of signal processing concepts in genomics and proteomics",* Journal of the *Franklin Institute (Special Issue on Genomics*), vol. 341, pp. 111-135.

Widom, J. (1996) "*Short-range Order in Two Eucaryotic Genomes: Relation to chromosome Structure*" J. Mol. Biol. 259 pp 579-588

Worcel, A., Strogatz, S. & Riley, D. "*Structure of chromatin and the linking number of DNA*", Proceedings of the National Academy of Sciences of the United States of America, Vol 78, No 3, part 2: Biological Sciences, pp:1461-1465, 1981

Yu, Z.G., Shi, L., Xiao, Q.J. & Anh, V., (2008) "*Chaos game representation of genomes and their simulation by recurrent iterated function systems*", Bioinformatics and Biomedical Engineering ICBBE, IEEE , pp 41–46.

Part 3

Fourier and Helbert Transform Applications

9

The Fourier Convolution Theorem over Finite Fields: Extensions of Its Application to Error Control Coding

Eric Sakk and Schinnel Small
*Department of Computer Science, Morgan State University,
Baltimore, MD,
USA*

1. Introduction

Linear spectral transform techniques such as the discrete Fourier transform and wavelet analysis over real and complex fields have been routinely applied in the literature (Burrus et al. (1998); Strang & Nuygen (1996)). Furthermore, extensions of these techniques over finite fields (Blahut & Burrus (1991); Caire et al. (1993)) have led to applications in the areas of information theory and error control coding (Blahut (2003); Dodd (2003); Sakk (2002); Wicker (1994)). The goal of this chapter is to review the Galois Field Fourier Transform, the associated convolution theorem and its application in the field of error control coding. In doing so, an interesting connection will be established relating the convolution theorem over finite fields to error control codes designed using finite geometries (Blahut (2003); Lin & Costello (1983); Wicker (1994)).

While a complete exposition of the field of error control would be out of context for this chapter, we refer the interested reader to the recent characterizations of Low-Density Parity Check (LDPC) codes (Pusane et al. (2011); Smarandache et al. (2009); Xia & Fu (2008)). Such formulations have led to a resurgence of interest in the design (Kou et al. (2001); O.Vontobel et al. (2005); Tang et al. (2005); Vandendriesscher (2010)) and decoding (Kou et al. (2001); Li et al. (2010); Liu & Pados (2005); Ngatched et al. (2009); Tang et al. (2005); Zhang et al. (2010)) of finite geometry codes. The formulation in this chapter is meant to serve as a guiding principle relating finite geometric properties to algebraic ones. The vehicle we have chosen to demonstrate these relationships is an example from the field of error control. In particular, we show how a generalized Fourier-like convolution theorem can be applied as a decoding methodology for finite geometry codes.

We begin in Section 2 by reviewing the Galois Field Fourier Transform (GFFT) followed by an overview of error control coding in Section 3. In addition, in Section 3.1 it is demonstrated how the GFFT can be applied within the context of error control coding. Section 4 then goes on to generalize these results to linear transformations using Pascal's triangle as an example. The combinatorics of such a transformation naturally lead to the design of codes derivable from finite geometries. Finally, Sections 5 and 6 conclude this chapter by deriving and applying the generalized convolution theorem.

2. The Galois Field Fourier Transform

We are particulary interested in the case of finite fields where p is a prime number and $\alpha \in GF(p^m)$ is an element of order n. The Galois Field Fourier Transform (GFFT) and its inverse of a vector $v = \{v_0, v_1, ..., v_{n-1}\}$ over $GF(p)$ of length n can be related via the equations:

$$V_j = \sum_{i=0}^{n-1} \alpha^{ij} v_i \qquad j = 0, ..., n-1$$

and

$$v_i = (n)^{-1} \sum_{j=0}^{n-1} \alpha^{-ij} V_j \qquad i = 0, ..., n-1.$$

For any vector f over $GF(p)$ where the above equations hold true, we define

$$\mathcal{F}(v) \equiv V = \{V_0, V_1, ..., V_{n-1}\} \tag{1}$$

as the GFFT of v and

$$\mathcal{F}^{-1}(V) = v = \{v_0, v_1, ..., v_{n-1}\} \tag{2}$$

as the inverse GFFT of F.

Using this formulation, given two vectors

$$\begin{aligned} v &= \{v_0, v_1, ..., v_{n-1}\} \\ w &= \{w_0, w_1, ..., w_{n-1}\} \end{aligned} \tag{3}$$

over $GF(p)$ and their associated transforms

$$\begin{aligned} \mathcal{F}(v) &= V = \{V_0, V_1, ..., V_{n-1}\} \\ \mathcal{F}(w) &= W = \{W_0, W_1, ..., W_{n-1}\}, \end{aligned} \tag{4}$$

the familiar convolution theorem can be demonstrated to hold true for the finite field case. Specifically, computing

$$x_j = \sum_{k=0}^{n-1} v_k w_{(j-k)} \tag{5}$$

is equivalent to computing

$$x_j = \mathcal{F}^{-1}(V_j W_j). \tag{6}$$

3. Error control coding

Given a message encoded as a vector μ of length k over $GF(p)$, the goal of error control coding (ECC) is to transform the message vector into a code vector C of length $n > k$ in a way that causes C to be robust to errors arising over a communication channel (such as a wireless link, fiber optic cable, etc). Rather than the message vector μ, it is the code vector C that is transmitted over a channel where the receiver is only able to observe a received vector $\hat{C}$. Ideally, in the absence of any noise, it should be the case that $\hat{C} = C$. On the other hand, if noise is present on the channel, the method used to transform (i.e. 'encode') the message μ into the code vector C provides a way to recover μ from $\hat{C}$. The basic strategy behind ECC is, given a message,

a. Embed a k dimensional message vector μ in a larger vector space of dimension n to create the code vector C.

b. The addition of channel noise converts C into the received vector $\hat{C}$.

c. If the channel noise does not cause $\hat{C}$ to be confused with other possible encodings, the original code vector C can be recovered using some predetermined decoding scheme. Conceptually speaking, the $\hat{C}$ that lies within a predefined noise 'sphere' with respect to the original C will be decoded as the (ideally) unique C; hence, μ can be recovered as well. The size of the noise sphere (which is designed as part of the code) determines how many errors can be corrected.

The general idea behind ECC then is to find a C that minimizes $||C - \hat{C}||$; however, numerically determining the minimum distance solution is wrought with dimensionality issues that can lead to computational intractability. Hence, classes of codes have been devised that relate the message encoding method to the decoding algorithm. Such algorithms are often iterative (Blahut (2003); Lin & Costello (1983); Wicker & Kim (2003)) and converge upon the optimal solution by exploiting the mathematical structure designed into the code.

Two important quantities in the field of ECC are the Hamming weight and the Hamming distance. Consider two vectors v and w of length n over $GF(p)$.

Definition 3.1. *The Hamming weight $w_H(v)$ of a vector v is defined as the number of non-zero components in v.*

Definition 3.2. *The Hamming distance between v and w is defined as the number of components that differ between v and w.*

For example, over $GF(3)$, assuming $n = 5$, $v = \{0\ 2\ 1\ 0\ 2\}$ and $w = \{0\ 2\ 2\ 1\ 2\}$, according to the above definitions we have that $w_H(v) = 3$, $w_H(w) = 4$ and $d_H(v, w) = 2$.

An important quantity for defining the noise sphere is referred to as d_{min} which is the minimum Hamming distance between all code vectors defined in the code class. To correct up to t errors in any code vector, it turns out that $d_{min} = 2t + 1$. Furthermore, when the ECC is a linear code, a major simplification arises where d_{min} is simply the minimum Hamming weight computed over all non-zero code vectors in the code class.

3.1 Application of the GFFT to Reed-Solomon codes

The GFFT and the convolution theorem have been applied in the field of error control coding for the construction of a class of linear codes known as Reed-Solomon codes (Blahut (2003); Wicker (1994)). The algorithm for encoding a message vector μ over $GF(p^m)$ of length k is quite straightforward. To be able to correct up to t errors, create a vector of length n by appending μ with $2t$ consecutive zeros. The code vector C is then derived by computing the inverse GFFT of the appended construction. One approach to proving that this construction is capable of correcting up to t errors involves applying the GFFT convolution theorem. Specifically, given a code vector C, a *locator* vector Λ must be defined such that $C_j \Lambda_j = 0$ for all $j = 0, \cdots, n$. Letting c and λ denote the GFFT of C and Λ, the convolution theorem implies $c * \lambda = 0$. Based upon the convolution approach, the conclusion can be reached that the inverse GFFT construction leads to Reed-Solomon codes capable of correcting up to t errors in the code vector (Blahut (2003); Wicker (1994)).

The key feature of the GFFT approach to constructing Reed-Solomon codes described above is that restrictions are placed on the position and the number of zeros appended to the message vector. To summarize:

i. Addition of zeros to the message vector μ of length k is performed at prescribed locations.

ii. The resulting vector is then inverse transformed in order to compute the code vector C.

iii. The error correcting properties of this code can be demonstrated by applying the convolution theorem.

In this work, one our of goals is to demonstrate that, given other linear transformations inducing a convolution theorem, the above steps can be generalized to other classes of codes. As we shall see, the key is to define the transform and the structure of how zeros are introduced into the message vector.

4. Pascal codes

4.1 The Pascal matrix over finite fields

Let us now focus our attention on the case of $GF(p)$ where p is prime. Our starting point will be:

Definition 4.1. *Let p be a prime number, then the ij^{th} entry of a $p^m \times p^m$* ***m^{th} order Pascal matrix P_{p^m} over $GF(p)$*** *is defined as*

$$\begin{aligned} p_{ij} &= (j!)((j-i)!i!)^{-1} \mod p \\ &= \binom{j}{i} \mod p \end{aligned} \tag{7}$$

for $i, j = 0,\ 1,\ ...,\ p^m - 1$ and, by convention, if $i > j$, then $p_{ij} = 0$.

In other words, P_{p^m} is an upper triangular matrix whose non-zero entries are the elements of Pascal's triangle taken *mod p*. For the purposes of this work, it is useful to observe that P_{p^m} also has a Kronecker product description (Sakk & Wicker (2003)):

$$P_{p^m} = P_p \otimes P_{p^{m-1}} \mod p \tag{8}$$

where P_p is a 1^{st} order Pascal matrix.

Example 4.2. *Consider the binary case where $p = 2$ and $m = 3$. Equation (8) gives*

$$P_{2^3} = \begin{bmatrix} 1&1&1&1&1&1&1&1 \\ 0&1&0&1&0&1&0&1 \\ 0&0&1&1&0&0&1&1 \\ 0&0&0&1&0&0&0&1 \\ 0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&1 \end{bmatrix}.$$

Example 4.3. *Consider the ternary case where $p = 3$ and $m = 2$. Equation (8) gives*

$$P_{3^2} = \begin{bmatrix} 1&1&1&1&1&1&1&1&1 \\ 0&1&2&0&1&2&0&1&2 \\ 0&0&1&0&0&1&0&0&1 \\ 0&0&0&1&1&1&2&2&2 \\ 0&0&0&0&1&2&0&2&1 \\ 0&0&0&0&0&1&0&0&2 \\ 0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&1&2 \\ 0&0&0&0&0&0&0&0&1 \end{bmatrix}.$$

4.2 The inverse of the Pascal matrix

As Section 5 will require understanding $Q_{p^m} \equiv P_{p^m}^{-1}$, we introduce the following.

Observation 4.4. *Let p be prime and let Q_p be the $p \times p$ matrix defined by*

$$q_{ij} = \begin{cases} (-1)^{j-i}\binom{j}{i} \mod p & \text{if } j \geq i, \\ 0 & \text{otherwise} \end{cases} \quad \text{for i,j=0, 1, ..., p-1.} \tag{9}$$

Then $Q_p = P_p^{-1} \mod p$.

This result easily follows from the integer case (Call & Velleman (1993); Heller (1963)). Furthermore, it has been demonstrated that (Sakk (2002)):

Observation 4.5. *If p is prime and P_p is a 1^{st} order Pascal matrix over $GF(p)$, then*

$$P_p^p \; mod \; p = I_p \tag{10}$$

where $I_p = p \times p$ identity matrix.

Hence, it easily follows that

Corollary 4.6. *If p is prime and P_p is a Pascal matrix over $GF(p)$, then*

$$Q_p = P_p^{-1} \; mod \; p = P_p^{p-1} \; mod \; p. \tag{11}$$

Example 4.7. *A Pascal matrix over $GF(5)$ and its inverse:*

$$P_p = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad Q_p = P_p^4 = \begin{bmatrix} 1 & 4 & 1 & 4 & 1 \\ 0 & 1 & 3 & 3 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Based upon Equation (8), it should be clear that

$$Q_{p^m} = Q_p \otimes Q_{p^{m-1}} \; mod \; p. \tag{12}$$

Finally, based upon Equation (10), it also follows that, for the m^{th} order case,

$$P^p_{p^m} \; mod \; p = I_{p^m} \tag{13}$$

where $I_{p^m} = p^m \times p^m$ identity matrix. In a manner similar to the $m = 1$ case, this characterization provides a path to computing the m^{th} order inverse

$$Q_{p^m} \equiv P^{-1}_{p^m} \; mod \; p = P^{p-1}_{p^m} \; mod \; p. \tag{14}$$

4.3 Error control codes designed from Pascal matrices

In a manner similar to the GFFT approach to Reed-Solomon codes summarized in Section 3.1, it has been pointed out that P_{p^m} can also be used to transform message vectors with the appropriate coordinates set equal to zero (Sakk & Wicker (2003)). More precisely, we have the following:

Definition 4.8. *Consider an m^{th} order Pascal matrix over $GF(p)$ and let r be an integer such that $0 \le r \le m(p-1)$. Also, consider the p-ary expansion of an index*

$$i = i_0 p^0 + i_1 p^1 + \cdots + i_{m-1} p^{m-1}$$

*where $0 \le i_j \le p-1$ for $0 \le j \le m-1$. A codeword c for an **r^{th} order Pascal code of length p^m**, denoted by $P_p(r,m)$, is generated by*

$$C = \mu P_{p^m} \tag{15}$$

where

$$\mu = \begin{pmatrix} \mu_0 \; \mu_1 \; \cdots \; \mu_{p^m-1} \end{pmatrix}$$

is a message vector of length $p^m - 1$ such that $\mu_i \in GF(p)$,

$$\begin{cases} \mu_i = 0 & if \quad w_p(i) > r \\ \mu_i \neq 0 & if \quad w_p(i) \le r \end{cases} \tag{16}$$

and

$$w_p(i) \equiv \sum_{j=0}^{m-1} i_j.$$

Error control codes derived from the m^{th} order Pascal matrix over $GF(2)$ (i.e. binary data) have been related (Forney (1988); Massey et al. (1973)) to a class of codes known as r^{th} order binary Reed-Muller codes $RM(r,m)$ of length 2^m (MacWilliams & Sloane (1977); Wicker (1994)). In addition, it has been further demonstrated (Sakk (2002)) that $P_2(r,m)$ codes over $GF(2)$ are equivalent to $RM(r,m)$ codes with minimum distance $d_{min} = 2^{m-r}$. These observations have been extended where it has been demonstrated that $P_p(r,m)$ codes over $GF(p)$ are equivalent to generalized Reed-Muller codes (GRM) codes (Sakk (2002)).

To place this class of codes in the same context as that outlined in Section 3.1, we must show how to introduce zeros into the message vector, apply the Pascal matrix as the linear transformation and, based upon this transformation, introduce a convolution theorem. From the definition above, a given code is specified by choosing p, m and a value of $0 \leq r \leq m(p-1)$. The code vector length then becomes $n = p^m$; and, for this class of codes, a given value of r defines the length k of the message. The rest of the $n-k$ components of μ must be set to zero in a systematic way that leads to the minimum distance property of the code.

Example 4.9. *Consider P_{2^3} in Example 4.2 (hence, $n = 2^3 = 8$) and a message vector $\mu = (\mu_0,\ \mu_1,\ ...,\ \mu_7)$ and let s be the number of consecutive zeros in the vector μ for a given value of r:*

$$\begin{array}{lll}
r = 0\ (d_{min} = 8): & s = 7 & \mu = (\mu_0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0) & (k = 1) \\
r = 1\ (d_{min} = 4): & s = 3 & \mu = (\mu_0,\ \mu_1,\ \mu_2,\ 0,\ \mu_4,\ 0,\ 0,\ 0) & (k = 4) \\
r = 2\ (d_{min} = 2): & s = 1 & \mu = (\mu_0,\ \mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \mu_5,\ \mu_6,\ 0) & (k = 7) \\
r = 3\ (d_{min} = 1): & s = 0 & \mu = (\mu_0,\ \mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \mu_5,\ \mu_6,\ \mu_7) & (k = 8)
\end{array}$$

Example 4.10. *Consider P_{3^2} in Example 4.3 (hence, $n = 3^2 = 9$) and a message vector $\mu = (\mu_0,\ \mu_1,\ ...,\ \mu_8)$ and let s be the number of consecutive zeros in the vector μ for a given value of r:*

$$\begin{array}{llll}
r = 0\ (d_{min} = 9): & s = 8 & \mu = (\mu_0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0) & (k = 1) \\
r = 1\ (d_{min} = 6): & s = 5 & \mu = (\mu_0,\ \mu_1,\ 0,\ \mu_3,\ 0,\ 0,\ 0,\ 0,\ 0) & (k = 3) \\
r = 2\ (d_{min} = 3): & s = 2 & \mu = (\mu_0,\ \mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ 0,\ \mu_6,\ 0,\ 0) & (k = 6) \\
r = 3\ (d_{min} = 2): & s = 1 & \mu = (\mu_0,\ \mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \mu_5,\ \mu_6,\ \mu_7,\ 0) & (k = 8) \\
r = 4\ (d_{min} = 1): & s = 0 & \mu = (\mu_0,\ \mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \mu_5,\ \mu_6,\ \mu_7,\ \mu_8) & (k = 9)
\end{array}$$

In the above examples, d_{min} is shown in parentheses for each value of r; furthermore, observe that $d_{min} = s + 1$. Recalling for a moment the GFFT approach to Reed-Solomon code design, the minimum distance of a code where the message vector has $n-k$ consecutive zeros can be shown to be $d_{min} = n - k + 1$ (Blahut (2003); Wicker (1994)). It is apparent that, by using a Pascal matrix as the transform, a result similar to that of the GFFT can be ascertained. The major difference is that, for Reed-Solomon codes, the string of zeros must occur at the end of the message vector before applying the GFFT to create C. For $P(r,m)$, in addition to the string of consecutive zeros, based upon the structure of P_{p^m}, zeros must also be dispersed in other positions within μ to form code vectors $C = \mu P_{p^m}$.

5. Extensions of the Fourier convolution theorem over finite fields

The convolution operation involves relating the componentwise product of two vectors in one domain to the convolution of their transforms (Blahut & Burrus (1991)). Many linear transforms have well-defined convolution operations. For instance, the Hadamard transform yields the so-called logical or 'dyadic' convolution operation (Ahmed et al. (1973); Dodd (2003); Robinson (1972)). In this chapter, we develop extensions of the convolution theorem that can be used to reveal useful properties of other classes of codes. As an example, we demonstrate how the GFFT approach can be applied to describe generalized Reed-Muller codes (Blahut (2003)).

To begin the formulation, we consider the componentwise product $\gamma_j = \mu_j \lambda_j$ of two vectors $\mu = (\mu_0 \; ... \; \mu_{n-1})$ and $\lambda = (\lambda_0 \; ... \; \lambda_{n-1})$. Furthermore, we consider matrix transforms such that $C \equiv \mu P_{p^m}$ and $\Lambda \equiv \lambda P_{p^m}$ or, equivalently, $\mu = C Q_{p^m}$ and $\lambda = \Lambda Q_{p^m}$ where $(P_{p^m})^{-1} \equiv Q_{p^m}$. (here, '$\mu$' denotes the message vector and 'C' denotes the code vector). We demonstrate a formulation analogous to the convolution operation that describes $\gamma = \Gamma Q_{p^m}$:

$$\begin{aligned}
\Gamma_i &= \sum_{j=0}^{n-1} \gamma_j p_{ji} \mod p \qquad i = 0, \, 1, \, ..., \, n-1 \\
&= \sum_{j=0}^{n-1} (\mu_j \lambda_j) p_{ji} \mod p \\
&= \sum_{j=0}^{n-1} \mu_j (\sum_{k=0}^{n-1} \Lambda_k q_{kj}) p_{ji} \mod p \\
&= \sum_{k=0}^{n-1} \Lambda_k (\sum_{j=0}^{n-1} \mu_j q_{kj} p_{ji}) \mod p \\
&\equiv \sum_{k=0}^{n-1} \Lambda_k T_{i,k} \mod p \qquad i = 0, \, 1, \, ..., \, n-1
\end{aligned} \tag{17}$$

where $n = p^m$.

Notice that if we are dealing with familiar spectral transforms such as the Fourier or the Hadamard transform (where P denotes the forward transform and Q denotes the inverse transform), $T_{i,k}$ takes on a simple form. This is because the product $q_{kj} p_{ji}$ in $\sum_{j=0}^{n-1} \mu_j q_{kj} p_{ji}$ reduces to a term that enables us to take the transform of μ as $C_{f(i,k)} = \frac{1}{n}(\sum_{j=0}^{n-1} \mu_j p_{j,f(i,k)})$. For the case of the Fourier transform $f(i,k) = i - k$ and $T_{i,k} = C_{(i-k)}$; as expected, one ends up with the convolution theorem (Blahut (2003); Wicker (1994)). In the case of a Hadamard transform, $f(i,k) = i \oplus k$ (where $\oplus$ denotes bit-by-bit addition of the binary expansions of i and k) and $T_{i,k} = C_{(i \oplus k)}$. Here, the bit-by-bit addition $\oplus$ of the binary expansions of i and k over GF(2) would result in the dyadic convolution (Ahmed et al. (1973); Robinson (1972)).

For the codes in this presentation, the $q_{kj} p_{ji}$ term in the above summation leads to a convolution theorem that depends on the matrix P_{p^m}. Furthermore, this theorem can also be applied to demonstrate how to decode C to recover the message vector μ. In Equation (17)

$q_{kj} = (-1)^{j-k}\binom{j}{k} \mod p$ and $p_{ji} = \binom{i}{j} \mod p$; therefore, the product $q_{kj}p_{ji}$ will not lead to an expression that readily reduces the inner summation to a single term. To see why, let's write out $T_{i,k}$ as follows:

$$\begin{aligned} T_i &= (T_{i,0}\ T_{i,1}\ ...\ T_{i,n-1}) \\ &= (\mu_0\ \mu_1\ ...\ \mu_{n-1}) \begin{bmatrix} q_{00}p_{0i} & q_{10}p_{0i} & \cdots & q_{(n-1)0}p_{0i} \\ q_{01}p_{1i} & q_{11}p_{1i} & \cdots & q_{(n-1)1}p_{1i} \\ \vdots & \vdots & \cdots & \vdots \\ q_{0(n-1)}p_{(n-1)i} & q_{1(n-1)}p_{(n-1)i} & \cdots & q_{(n-1)(n-1)}p_{(n-1)i} \end{bmatrix} \\ &= (\mu_0\ \mu_1\ ...\ \mu_{n-1}) \begin{bmatrix} p_{0i} & & & \\ & p_{1i} & & \\ & & \ddots & \\ & & & p_{(n-1)i} \end{bmatrix} \begin{bmatrix} q_{00} & q_{10} & \cdots & q_{(n-1)0} \\ q_{01} & q_{11} & \cdots & q_{(n-1)1} \\ \vdots & \vdots & \cdots & \vdots \\ q_{0(n-1)} & q_{1(n-1)} & \cdots & q_{(n-1)(n-1)} \end{bmatrix} \\ &\equiv \mu D_i Q_{p^m}^T \end{aligned} \tag{18}$$

where T denotes the matrix transpose.

Observation 5.1. *The components of the vector $T_i = (T_{i,0}\ T_{i,1}\ ...\ T_{i,n-1})$ can be written as a linear combination of the components of $C = (C_0\ ...\ C_{n-1})$.*

Proof: Let

$$M_i \equiv D_i Q_{p^m}^T \tag{19}$$

where D_i is defined in Equation (18) and

$$\begin{aligned} A_i &\equiv Q_{p^m} M_i = Q_{p^m} D_i Q_{p^m}^T \\ &\Rightarrow M_i = P_{p^m} A_i. \end{aligned} \tag{20}$$

Then,

$$T_i = \mu M_i = \mu P_{p^m} A_i = C A_i. \tag{21}$$

Combining this result with Equation (17) we conclude

$$\begin{aligned} \Gamma_i &= \sum_{k=0}^{n-1} \Lambda_k T_{i,k} \mod p \qquad i = 0,\ 1,\ ...,\ n-1 \\ &= \sum_{k=0}^{n-1} \Lambda_k (C A_i)_k \mod p \qquad i = 0,\ 1,\ ...,\ n-1 \end{aligned} \tag{22}$$

So, instead of $T_{i,k}$ reducing to one single component of the vector C (as one might expect from a typical convolution operation), the Pascal convolution requires a linear combination of the components of C. Although this operation is slightly more complicated than the Fourier approach, the identity in Equation (8) does induce a simplification.

Observation 5.2. *(Symbolic Computation of Pascal Convolution)*
For the 1st order case where $n = p$ and $i = 0,\ ...,\ p-1$, using Equation (19) let $\hat{M}_i \equiv M_i$,

using Equation (18) let $\hat{D}_i \equiv D_i$ and let $\hat{A}_i \equiv Q_p \hat{M}_i$. Then, for any $0 \leq j \leq p^m - 1$ where $j = j_0 p^0 + j_1 p^1 + ... + j_{m-1} p^{m-1}$ and $A_j = Q_{p^m} M_j$,

$$A_j = \hat{A}_{j_{m-1}} \otimes ... \otimes \hat{A}_{j_1} \otimes \hat{A}_{j_0} \tag{23}$$

where $M_j \equiv D_j Q_{p^m}^T$.

Proof: The statement is clearly true for the first order case $m = 1$ since $j = j_0$. By induction let $j = j_0 p^0 + j_1 p^1 + ... + j_{m-1} p^{m-1}$ and assume that

$$D_j = \hat{D}_{j_{m-1}} \otimes ... \otimes \hat{D}_{j_1} \otimes \hat{D}_{j_0}$$

where $0 \leq j_k \leq p - 1$ for all $k = 0, \; ..., \; m - 1$. Consider any $j' = j_0 p^0 + ... + j_{m-1} p^{m-1} + j_m p^m$ and apply Equation (18) along with Lucas' theorem to obtain the following intermediate result:

$$\begin{aligned}
\hat{D}_{j_m} \otimes \hat{D}_{j_{m-1}} \otimes ... \otimes \hat{D}_{j_0} &= \hat{D}_{j_m} \otimes D_j \\
&= \begin{bmatrix} \binom{j_m}{0} & & & \\ & \binom{j_m}{1} & & \\ & & \ddots & \\ & & & \binom{j_m}{p-1} \end{bmatrix} \otimes \begin{bmatrix} \binom{j}{0} & & & \\ & \binom{j}{1} & & \\ & & \ddots & \\ & & & \binom{j}{p^m-1} \end{bmatrix} \\
&= \begin{bmatrix} \binom{j'}{0} & & & \\ & \binom{j'}{1} & & \\ & & \ddots & \\ & & & \binom{j'}{p^{m+1}-1} \end{bmatrix} \\
&= D_{j'}
\end{aligned} \tag{24}$$

Therefore, $D_j = \hat{D}_{j_{m-1}} \otimes ... \otimes \hat{D}_{j_1} \otimes \hat{D}_{j_0}$ is true. Next, successively apply the identity $(AC) \otimes (BD) = (A \otimes B)(C \otimes D)$ to obtain:

$$\begin{aligned}
\hat{M}_{j_{m-1}} \otimes ... \otimes \hat{M}_{j_1} \otimes \hat{M}_{j_0} &= (\hat{D}_{j_{m-1}} Q_p^T) \otimes ... \otimes (\hat{D}_{j_1} Q_p^T) \otimes (\hat{D}_{j_0} Q_p^T) \\
&= (\hat{D}_{j_{m-1}} \otimes ... \otimes \hat{D}_{j_1} \otimes \hat{D}_{j_0})(Q_p^T \otimes Q_p^T \otimes ... \otimes Q_p^T) \\
&= D_j Q_{p^m}^T \\
&= M_j
\end{aligned}$$

Finally, we arrive at the desired conclusion

$$\begin{aligned}
(\hat{A}_{j_{m-1}} \otimes ... \otimes \hat{A}_{j_1} \otimes \hat{A}_{j_0}) &= (Q_p \hat{M}_{j_{m-1}}) \otimes ... \otimes (Q_p \hat{M}_{j_1}) \otimes (Q_p \hat{M}_{j_0}) \\
&= (Q_p \otimes Q_p \otimes ...Q_p)(\hat{M}_{j_{m-1}} \otimes ... \otimes \hat{M}_{j_1} \otimes \hat{M}_{j_0}) \\
&= Q_{p^m} M_j \\
&= A_j.
\end{aligned}$$

Observation 5.2 tells us that, in order to calculate $T_j = CA_j$ for arbitrary $n = p^m$, one need only calculate $\hat{A}_i$ for $i = 0, \ ..., \ p-1$ and then take successive Kronecker products. The initial set of $\hat{A}_i$ for $i = 0, \ ..., \ p-1$ can easily be calculated by referring back to Equation (20) where $\hat{A}_i = Q_p\hat{M}_i = Q_p\hat{D}_iQ_p^T$.

An interesting property concerning the A_i is that the sum

$$\sum_{i=0}^{p^m-1} A_i = \sum_{i=0}^{p^m-1} Q_p\hat{D}_iQ_p^T$$

(where the sum is taken mod p) is a matrix of ones. This follows from two observations. First, from the definition of D_i in Equation (18), $\sum_{i=0}^{p^m-1} D_i$ is a matrix whose (p^m-1, p^m-1) entry is one and all other entries are zero. Second, it can also be demonstrated that the last column of Q_{p^m} must be a column of ones. Therefore, $Q_p\sum_{i=0}^{p^m-1} \hat{D}_iQ_p^T = \sum_{i=0}^{p^m-1} A_i$ is a matrix of ones.

Example 5.3. *For $p = 2$, the 1st order case $n = p$ gives $i = 0, 1$; hence, over $GF(2)$,*

$$P_p = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad Q_p = P_p = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

we calculate

$$\hat{A}_0 = Q_p\hat{D}_0Q_p^T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\hat{A}_1 = Q_p\hat{D}_1Q_p^T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

From Observation 5.2, to obtain the A_j for $n = p^2$ and $j = 0, 1, 2, 3$, one need only take successive Kronecker products as:

$$A_0 = \hat{A}_0 \otimes \hat{A}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_1 = \hat{A}_0 \otimes \hat{A}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$A_2 = \hat{A}_1 \otimes \hat{A}_0 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_3 = \hat{A}_1 \otimes \hat{A}_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

As expected, the A_i are symmetric matrices. Also, notice, as mentioned above, that $\sum_{i=0}^{p^m-1} A_i$ is a matrix of ones. For the case where $n = p^2$, let us now apply Observation 5.1 to calculate the Pascal convolution of the vectors $C = (C_0, C_1, C_2, C_3)$ and $\Lambda = (\Lambda_0, \Lambda_1, \Lambda_2, \Lambda_3)$. Using Equation (22), we have:

$$\begin{aligned}
\Gamma_0 &= \Lambda_0C_0 + \Lambda_1(0) + \Lambda_2(0) + \Lambda_3(0) \\
\Gamma_1 &= \Lambda_0C_1 + \Lambda_1(C_0 + C_1) + \Lambda_2(0) + \Lambda_3(0) \\
\Gamma_2 &= \Lambda_0C_2 + \Lambda_1(0) + \Lambda_2(C_0 + C_2) + \Lambda_3(0) \\
\Gamma_3 &= \Lambda_0C_3 + \Lambda_1(C_2 + C_3) + \Lambda_2(C_1 + C_3) + \Lambda_3(C_0 + C_1 + C_2 + C_3).
\end{aligned} \tag{25}$$

To close this section, we draw some immediate conclusions from Equation (25):

- Because of the Kronecker product, a good deal of self-similar structure can be observed in the resulting vector Γ. For instance, the coefficients of the Λ_i can be computed by iteration starting with the initial 'seed' generated by $\hat{A}_0$ and $\hat{A}_1$. As an example, the coefficient of Λ_1 in Γ_1 can be computed by adding the coefficient of Λ_0 in Γ_0 to the coefficient of Λ_0 in Γ_1. The coefficients of Λ_2 and Λ_3 in Γ_2 and Γ_3 can be computed by adding the coefficients of Λ_0 and Λ_1 in Γ_0 and Γ_1 to the coefficients of Λ_0 and Λ_1 in Γ_2 and Γ_3, and so on.
- Looking at the result columnwise, the set of coefficients associated with a given Λ_i appear to be the checksums for an $R(r, 2)$ binary Reed-Muller code ((MacWilliams & Sloane, 1977, p.385-388), (Wicker, 1994, p.155-165)). As pointed out in the next section, although this observation is true for the binary case, an orthogonal set of checksums for $p > 2$ will not come about by this method. It is the dual of the Pascal convolution that will lead to the decoding of GRM codes.

6. Majority logic decoding using Pascal convolution

GRM codes fall into a larger category of codes known as Euclidean geometry codes (Blahut (2003); Lin & Costello (1983); MacWilliams & Sloane (1977); Wicker (1994)) where it is well-known that a technique known as 'majority logic decoding' (MLD) can be used to recover the message vector. Based upon statements made in Section 4, it should be clear that Pascal codes are also MLD. However, the role played by the Pascal convolution in the decoding strategy is worthy of mention. As pointed out in the conclusions of Example 5.3, the checksums of a majority logic decoding (MLD) scheme for GRM codes can be derived using the dual of the convolution relation derived above. We now demonstrate this observation more clearly.

Because of the similar forms of P_{p^m} and Q_{p^m} the dual convolution relation is easily derived from the inverse transform. Consider the componentwise product $\Gamma_j = C_j\Lambda_j$ of two vectors where $C = \mu P_{p^m}$ and $\Lambda = \lambda P_{p^m}$:

$$\begin{aligned}
\gamma_i &= \sum_{j=0}^{n-1} \Gamma_j q_{ji} \mod p \qquad i = 0,\ 1,\ ...,\ n-1 \\
&= \sum_{j=0}^{n-1} (C_j\Lambda_j) q_{ji} \mod p \\
&= \sum_{j=0}^{n-1} C_j \Big(\sum_{k=0}^{n-1} \lambda_k p_{kj}\Big) q_{ji} \mod p \\
&= \sum_{k=0}^{n-1} \lambda_k \Big(\sum_{j=0}^{n-1} C_j p_{kj} q_{ji}\Big) \mod p \\
&\equiv \sum_{k=0}^{n-1} \lambda_k s_{i,k} \mod p \qquad i = 0,\ 1,\ ...,\ n-1
\end{aligned} \tag{26}$$

where $n = p^m$. Similar to Equation (18), one can also show that

$$\begin{aligned} s_i &= (s_{i,0}\ s_{i,1}\ ...\ s_{i,n-1}) \\ &= C\Delta_i P_{p^m}^T \end{aligned} \tag{27}$$

which can also be written as

$$s_i = \mu P_{p^m} \Delta_i P_{p^m}^T \tag{28}$$

where Δ_i is a diagonal matrix with elements $(q_{0i}\quad q_{1i}\quad ...\quad q_{(n-1)i})$ along its diagonal. Furthermore, if we define

$$B_i \equiv P_{p^m} \Delta_i P_{p^m}^T \tag{29}$$

then results similar to Observations 5.1 and 5.2 can also be demonstrated. However, in proving the dual of Observation 5.2 there is one difference be aware of. Since $q_{ji} = (-1)^{i-j}\binom{i}{j}$, the Kronecker product in the dual of Equation (24) will contain extra factors of $(-1)^{i-j}$. To achieve the equality $\Delta_j = \hat{\Delta}_{j_{m-1}} \otimes ... \otimes \hat{\Delta}_{j_1} \otimes \hat{\Delta}_{j_0}$ where $j = j_0 p^0 + j_1 p^1 + ... + j_{m-1} p^{m-1}$ the following identity will be required:

$$\begin{aligned} (-1)^k &= (-1)^{k_0 p^0 + k_1 p^1 + ... + k_{m-1} p^{m-1}} \\ &= (-1)^{k_0}((-1)^p)^{k_1}((-1)^{p^2})^{k_2} ... ((-1)^{p^{m-1}})^{k_{m-1}} \\ &= (-1)^{\sum_{l=0}^{m-1} k_l} \end{aligned}$$

for any $0 \le k \le p^m - 1$ where we have applied $a^p = a$ for any $a \in GF(p)$. Then, following the proof of Observation 5.2, it is straightforward to show that for any $0 \le j \le p^m - 1$ where $j = j_0 p^0 + j_1 p^1 + ... + j_{m-1} p^{m-1}$,

$$B_j = \hat{B}_{j_{m-1}} \otimes ... \otimes \hat{B}_{j_1} \otimes \hat{B}_{j_0} \tag{30}$$

where

$$\hat{B}_{j_k} = P_p \hat{\Delta}_{j_k} P_p^T.$$

In Section 4, we explained that the form of message vectors when applying P_{p^m} as the transformation where the message vector $\mu = (\mu_0,\ ...,\ \mu_{p^m-1})$ should have all components $\mu_j = 0$ if $w_p(j) > r$ (see Examples 4.9 and 4.10). To see how this formulation can lead to a decoding scheme, let us examine the case where $p = 2$, $m = 2$ and $r = 1$ (i.e. - a 1^{st} order binary Reed-Muller code of length 4). Consider first using Equations (26) and (27) to calculate Pascal convolution of the vectors $\mu = (\mu_0, \mu_1, \mu_2, \mu_3)$ and $\lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_3)$:

$$\begin{array}{lrcccccc} 00: & \gamma_0 = & \lambda_0 C_0 & + & \lambda_1(0) & + & \lambda_2(0) & + \lambda_3(0) \\ 01: & \gamma_1 = & \lambda_0(C_0 + C_1) & + & \lambda_1 C_1 & + & \lambda_2(0) & + \lambda_3(0) \\ 10: & \gamma_2 = & \lambda_0(C_0 + C_2) & + & \lambda_1(0) & + & \lambda_2 C_2 & + \lambda_3(0) \\ 11: & \gamma_3 = & \lambda_0(\sum_{i=0}^{3} C_i) & + & \lambda_1(C_1 + C_3) & + & \lambda_2(C_2 + C_3) & + \lambda_3 C_3 \end{array}. \tag{31}$$

where the binary expansion of the γ index has been explicitly written out at the beginning of each row. Next, consider Equations (26) and (28) to calculate the same convolution:

$$\begin{array}{llllll} 00: & \gamma_0 = \lambda_0\mu_0 + & \lambda_1(0) & + \quad \lambda_2(0) & + \quad \lambda_3(0) \\ 01: & \gamma_1 = \lambda_0\mu_1 + & \lambda_1(\mu_0+\mu_1) & + \quad \lambda_2(0) & + \quad \lambda_3(0) \\ 10: & \gamma_2 = \lambda_0\mu_2 + & \lambda_1(0) & + \lambda_2(\mu_0+\mu_2) & + \quad \lambda_3(0) \\ 11: & \gamma_3 = \lambda_0\mu_3 + & \lambda_1(\mu_2+\mu_3) & + \lambda_2(\mu_1+\mu_3) & + \lambda_3(\sum_{i=0}^{3}\mu_i) \end{array}.$$

Since, for $P_2(1,2)$, $\mu = (\mu_0,\ \mu_1,\ \mu_2,\ 0)$, this set of equations can be simplified as

$$\begin{array}{llllll} 00: & \gamma_0 = \lambda_0\mu_0 + & \lambda_1(0) & + \quad \lambda_2(0) & + \quad \lambda_3(0) \\ 01: & \gamma_1 = \boldsymbol{\lambda_0\mu_1} + & \lambda_1(\mu_0+\mu_1) & + \quad \lambda_2(0) & + \quad \lambda_3(0) \\ 10: & \gamma_2 = \lambda_0\mu_2 + & \boldsymbol{\lambda_1(0)} & + \lambda_2(\mu_0+\mu_2) & + \quad \lambda_3(0) \\ 11: & \gamma_3 = \lambda_0(0) + & \lambda_1\mu_2 & + \quad \boldsymbol{\lambda_2\mu_1} & + \lambda_3(\mu_0+\mu_1+\mu_2) \end{array}. \qquad (32)$$

Equations (31) and (32) must hold for *any* vector λ. Therefore, for a specific γ_j, we can equate the coefficients of the λ_i in Equation (31) with those in Equation (32). So, for example, we end with the result that

$$\mu_2 = C_0 + C_2$$
$$\mu_2 = C_1 + C_3$$

and

$$\mu_1 = C_0 + C_1$$
$$\mu_1 = C_2 + C_3.$$

For this first order $r = 1$ code, we can generate a set of checksums using a simple algorithm. Start at an index i of γ such that $w_2(i) = 1$ and equate Equations (31) and (32) along a *diagonal path* in order to 'collect' all checksum equations associated associated with μ_i. For example, the bold symbols in Equation (32) generate the checksums for μ_1. It turns out that these diagonal equations actually generate what are known as the 'incidence vectors' of the MLD strategy (Blahut (2003); MacWilliams & Sloane (1977); Wicker (1994)).

We now provide an algorithm for $GF(p)$ to show how the Pascal convolution approach is equivalent to a typical MLD using finite Euclidean geometry ((Wicker, 1994, p.155-165)). The interesting aspect of this algorithm is that the Pascal convolution generates the correct checksums for *any* $GF(p)$. Consider a $P_p(r,m)$ code where $C = \mu P_{p^m}$ such that $\mu_j = 0$ if $w_p(j) > r$:

(0) Let $j = r$.

(1) Let S_j be the set of indices i such that $w_p(i) = j$.

(2) Apply Equation (27) to calculate γ.

(3) Apply Equation (28) to calculate γ (these equations will simplify based upon which of the μ_i are zero).

(4) For each $i \in S_j$, start at λ_0 associated with γ_i and construct checksum equations by equating the result in Step (2) with that of Step (3) along a *diagonal path* (i.e. - starting at k=0, choose the coefficient of λ_k associated with γ_{i+k}).

(5) For $i \in S_j$, create estimates $\bar{\mu}_i$ by a majority logic decision on the checksums.

(6) $j = j - 1$. If $j < 0$, stop.

(7) Remove the estimated components as:

$$\bar{C} = \bar{\mu} P_{p^m}$$
$$\hat{C} \equiv C - \bar{C}\ (= (\mu - \bar{\mu}) P_{p^m}).$$

(8) Adjust μ to reflect the change in step (7) as follows. Construct a new vector $\tilde{\mu}$ where $\tilde{\mu}_i = \mu_i$ if $i \in S_j$ and $\tilde{\mu}_i = 0$ otherwise. Then let

$$\hat{\mu} \equiv \mu - \tilde{\mu}.$$

(9) Let $C = \hat{C}$ and $\mu = \hat{\mu}$ and go to Step (1).

As with typical MLD schemes, this algorithm starts with the highest order r to obtain estimates of the code vector components and then successively estimates the lower order components.

Example 6.1. *Let $p = 3$, $m = 2$ and $r = 2$. Consider decoding a $P_3(2, 2)$ code. From Example 4.10,*

$$\mu = (\mu_0,\ \mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ 0,\ \mu_6,\ 0,\ 0).$$

Also, we know that $P_3(2, 2)$ has $d_{min} = 3$ implying that we can correct a single error. Therefore, we expect that the MLD equations should have at least three checksums.

(0) Start with $j = 2$.

(1) Let $S_2 = \{2, 4, 6\}$ (i.e. - $i = i_0 + i_1 p$ such that $w_3(i) = 2$).

(2,3,4) Rather than write out the equations for γ_i, we summarize by equating the results of step (2) and step (3):

$$i = 2: \quad \begin{aligned} \mu_2 &= c_0 + c_1 + c_2 \\ \mu_2 &= c_3 + c_4 + c_5 \\ \mu_2 &= c_6 + c_7 + c_8 \end{aligned}$$

$$i = 4: \quad \begin{aligned} \mu_4 &= c_0 + 2c_1 + 2c_3 + c_4 \\ 2\mu_4 &= 2c_1 + c_2 + c_4 + 2c_5 \\ 2\mu_4 &= 2c_3 + c_4 + c_6 + 2c_7 \\ \mu_4 &= c_4 + 2c_5 + 2c_7 + c_8 \end{aligned}$$

$$i = 6: \quad \begin{aligned} \mu_6 &= c_0 + c_3 + c_6 \\ \mu_6 &= c_1 + c_4 + c_7 \\ \mu_6 &= c_2 + c_5 + c_8 \end{aligned}$$

After estimating the message components dictated by S_2 (step (5)), remove the code estimates from C (step (7)) and begin work on S_1 where now (step(8)) $\mu_i = 0$ if $w_p(i) > 1$. For S_1, we have the checksums:

$$i = 1: \quad \begin{aligned} \mu_1 &= 2c_0 + c_1 \\ 2\mu_1 &= c_1 + 2c_2 \\ \mu_1 &= 2c_3 + c_4 \\ 2\mu_1 &= c_4 + 2c_5 \\ \mu_1 &= 2c_6 + c_7 \\ 2\mu_1 &= c_7 + 2c_8 \end{aligned}$$

$$i = 3: \quad \begin{aligned} \mu_3 &= 2c_0 + c_3 \\ \mu_3 &= 2c_1 + c_4 \\ \mu_3 &= 2c_2 + c_5 \\ 2\mu_3 &= c_3 + 2c_6 \\ 2\mu_3 &= c_4 + 2c_7 \\ 2\mu_3 &= c_5 + 2c_8 \end{aligned}$$

After estimating the message components dictated by S_1, once again, remove the code estimates from C and begin work on S_0 where now $\mu_i = 0$ if $w_p(i) > 0$. At this stage, with all other components of $\mu = 0$ except μ_0, we are left with $\mu = C$ (i.e. - nine estimate of the check on μ_0).

7. Conclusions

When considering the design of error control codes, it is interesting to look for guiding principles that can account for whole classes of codes. In this presentation, we have shown how the GFFT convolution approach to Reed-Solomon codes can be extended to other classes of codes such as generalized Reed-Muller codes.

Code	Convolution Principle	Decoding Strategy
Reed-Solomon	GFFT-based	iterative
GRM	generalized	iterative

Table 1. Comparison of Fourier and generalized convolution techniques.

Instead of applying a Fourier matrix to encode the message, we have applied a Pascal matrix and extended the convolution theorem over finite fields. In doing so, we have observed that this formulation leads to the well-known majority logic decoding algorithm. Additional investigations have also considered codes in the context of the wavelet transform (Sakk & Wicker (2003)). The block codes addressed in this chapter have been shown to lend themselves to graph-based iterative decoding strategies (see Table 1). The results derived above suggest that the generalized convolution approach is useful for understanding the systematic introduction of redundancy for the sake of error control.

8. References

Ahmed, N., Rao, K. & Abdussattar, A. (1973). On cyclic autocorrelation and the Walsh-Hadamard transform, *IEEE Transactions on Electromagnetic Compatibility* 18: 141–146.

Blahut, R. (2003). *Algebraic Codes for Data Transmission*, Cambridge University Press.

Blahut, R. & Burrus, C. (1991). *Algebraic Methods for Signal Processing and Communications Coding*, Springer.

Burrus, C., Gopinath, R. & Guo, H. (1998). *Introduction to Wavelets and Wavelet Transforms*, Prentice-Hall, NJ.

Caire, G., Grossman, R. & Poor, H. (1993). Wavelet transforms and associated finite cyclic groups, *IEEE Transactions on Information Theory* 39: 1157–1166.

Call, G. & Velleman, D. (1993). Pascal's matrices, *American Mathematical Monthly* 100: 372–376.

Dodd, M. (2003). *Applications of the Discrete Fourier Transform in Information Theory and Cryptology*, PhD thesis, Royal Holloway and Bedford New College, University of London.

Forney, G. D. (1988). Coset codes - Part II: Binary lattices and related codes, *IEEE Transactions on Information Theory* 34: 1152–1187.

Heller, S. (1963). Inverse of triangular matrix, *American Mathematical Monthly* 70: 334.

Kou, Y., Lin, S. & Fossorier, M. P. C. (2001). Low-density parity-check codes based on finite geometries: A rediscovery and new results, *IEEE Transactions on Information Theory* 47: 2711–2736.

Li, G., Li, D., Wang, Y. & Sun, W. (2010). Hybrid decoding of finite geometry low-density parity-check codes, *IET Communications* 4(10): 1238–1246.

Lin, S. & Costello, D. (1983). *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, New York.

Liu, Z. & Pados, D. A. (2005). Decoding algorithm for finite-geometry LDPC codes, *IEEE Transactions on Communications* 53: 415–421.

MacWilliams, F. J. & Sloane, N. J. A. (1977). *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam.

Massey, J. L., Costello, D. J. & Justesen, J. (1973). Polynomial weights and code constructions, *IEEE Transactions on Information Theory* 19: 101–110.

Ngatched, T. M. N., F, T. & Bossert, M. (2009). An improved decoding algorithm for finite-geometry LDPC codes, *IEEE Transactions on Communications* 57: 302–306.

O.Vontobel, P., Smarandache, R., Kiyavash, N., Teutsch, J. & D.Vukobratovic (2005). On the minimal pseudocodewords of codes from finite geometries, *Proc. IEEE Int. Symp. Inf. Theory*, Adelaide, Australia.

Pusane, A. E., Smarandache, R., Vontobel, P. O. & Costello, D. J. (2011). Deriving good LDPC convolutional codes from LDPC block codes, *IEEE Transactions on Information Theory* 57: 835–857.

Robinson, G. (1972). Logical convolution and discrete Walsh and Fourier power spectra, *IEEE Transactions on Audio and Electroacoustics* 20: 271–280.

Sakk, E. (2002). *Wavelet Packet Formulation of Generalized Reed Muller Codes*, PhD thesis, Cornell University, Ithaca, NY.

Sakk, E. & Wicker, S. (2003). Wavelet packets for error control coding, *Proceedings of the SPIE Volume 5207 (Wavelets X)*, San Diego, CA.

Smarandache, R., Pusane, A. E., Vontobel, P. O. & Costello, D. J. (2009). Pseudocodeword performance analysis for LDPC convolutional codes, *IEEE Transactions on Information Theory* 55: 2577–2598.

Strang, G. & Nuygen, T. (1996). *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA.

Tang, H., Xu, J., Lin, S. & Abdel-Ghaffar, K. A. S. (2005). Codes on finite geometries, *IEEE Transactions on Information Theory* 51: 572–596.

Vandendriesscher, P. (2010). Some low-density parity-check codes derived from finite geometries, *Designs, Codes and Cryptography* 54: 287–297.

Wicker, S. (1994). *Error Control Systems for Digital Communication and Storage*, Prentice Hall.

Wicker, S. & Kim, S. (2003). *Fundamentals of codes, graphs, and iterative decoding*, Kluwer.

Xia, S.-T. & Fu, F.-W. (2008). Minimum pseudoweight and minimum pseudocodewords of LDPC codes, *IEEE Transactions on Information Theory* 54: 480–485.

Zhang, L., Huang, Q. & Lin, S. (2010). Iterative decoding of a class of cyclic codess, *Information Theory and Applications Workshop (ITA)*, San Diego, CA.

Application of the Weighted Energy Method in the Partial Fourier Space to Linearized Viscous Conservation Laws with Non-Convex Condition

Yoshihiro Ueda
Faculty of Maritime Sciences, Kobe University
Japan

1. Introduction

As you know, the energy method in the Fourier space is useful in deriving the decay estimates for problems in the whole space $\mathbb{R}^n$. Recently, the author studied half space problems in $\mathbb{R}^n_+ = \mathbb{R}_+ \times \mathbb{R}^{n-1}$ and developed the energy method in the partial Fourier space obtained by taking the Fourier transform with respect to the tangential variable $\mathbb{R}^{n-1}$. Then the author applied this energy method to the half space problem for linearized viscous conservation laws with convex condition and proved the asymptotic stability of planar stationary waves by showing a sharp convergence rate for $t \to \infty$ (see, [14]).

In this chapter, we consider the half space problem for linearized viscous conservation laws with non-convex condition, and derive the asymptotic stability of planar stationary waves and the corresponding convergence rate. Our proof is based on the energy method in the partial Fourier space with the anti-derivative method.

In this present chapter, we are concerned with the half space problem for the viscous conservation laws:

$$u_t - \Delta u + \nabla \cdot f(u) = 0, \tag{1.1}$$

$$u(0, x', t) = u_b, \tag{1.2}$$

$$u(x, 0) = u_0(x). \tag{1.3}$$

Here $x = (x_1, \cdots, x_n)$ is the space variable in the half space $\mathbb{R}^n_+ = \mathbb{R}_+ \times \mathbb{R}^{n-1}$ with $n \geq 2$; we sometimes write as $x = (x_1, x')$ with $x_1 \in \mathbb{R}_+$ and $x' = (x_2, \cdots, x_n) \in \mathbb{R}^{n-1}$; $u(x, t)$ is the unknown function, $u_0(x)$ is the initial data satisfying

$$u_0(x) \to 0 \quad \text{as} \quad x_1 \to \infty,$$

and u_b is the boundary data (assumed to be a constant) with $u_b < 0$; $f(u) = (f_1(u), \cdots, f_n(u))$ is a smooth function of $u \in \mathbb{R}$ with values in $\mathbb{R}^n$ and satisfies

$$f_1(0) = 0, \qquad f_1(u) > f_1(0)\,(= 0) \tag{1.4}$$

for $u \in [u_b, 0)$. Here we note that the condition (1.4) is the necessary condition for the existence of the planar stationary waves (for the detail, see Section 2.2). We emphasize that the assumption (1.4) is weaker than the convex condition

$$f_1''(u) > 0 \tag{1.5}$$

for $u \in [u_b, 0]$ and $f_1(0) = 0$. Namely, we do not assume the convex condition for our problem (1.1)–(1.3).

For viscous conservation laws (1.1) with the convex condition (1.5), there are many results on the asymptotic stability of nonlinear waves. First, Il'in and Oleinik in [3] studied the stability of nonlinear waves in the one-dimensional whole space. Liu, Matsumura and Nishihara in the paper [8] discussed the stability of stationary waves in one-dimensional half space. More precisely, they proved the asymptotic stability of several kind of nonlinear waves such as rarefaction waves, stationary waves, and the superposition of stationary waves and rarefaction waves. Later, in a series of papers [5–7], their stability result of stationary waves in one-space dimension was generalized to the multi-dimensional case. Kawashima, Nishibata and Nishikawa [5] first considered the stability of non-degenerate planar stationary waves in two-dimensional half space and obtained the convergence rate $t^{-1/4-\alpha/2}$ in L^∞ norm by assuming that the initial perturbation is in $L^2_\alpha(\mathbb{R}_+; L^2(\mathbb{R}))$. Furthermore, the papers [6, 7] studied the n-dimensional problem in the L^p framework. In particular, the paper [7] showed the stability of non-degenerate planar stationary waves and obtained the convergence rate $t^{-(n/2)(1/2-1/p)-\alpha/2}$ in L^p norm under the assumption that the initial perturbation belongs to $L^2_\alpha(\mathbb{R}_+; L^2_{x'})$.

Next, we refer to viscous conservation laws with non-convex condition. Liu and Nishihara in [9] and Nishikawa in [10] investigated the asymptotic stability of travelling waves in the one-dimensional and multi-dimensional whole space, respectively. On the other hand, Hashimoto and Matsumura in [1] studied the asymptotic stability of stationary waves in the one-dimensional half space. Especially, in order to relax the convex condition, Liu and Nishihara in [9] and Nishikawa in [10] employed the anti-derivative method and achieved the desired result. Moreover, Hashimoto and the author in [2] used the same method to derive the asymptotic stability of stationary waves for damped wave equations with non-convex convection term in one-dimensional half space. Inspired by these arguments, we try to relax the convex condition (1.5) and get the asymptotic stability of planar stationary wave for the multi-dimensional problem (1.1)–(1.3). Unfortunately, Nishikawa in the paper [10] considered some special situation for the nonlinear term to make a good combination of the energy method and the anti-derivative method. For the same reason, we will treat the special situation (for the detail, see Section 3).

All these stability results mentioned above are obtained by employing the energy method in the physical space. On the other hand, it is useful to apply the energy method in the partial Fourier space to show sharper convergence rate. Indeed the author's paper [14] considered our problem (1.1)–(1.3) with the convex condition (1.5) and obtained the sharper convergence rate of the planar stationary waves. We shall show the result of the paper [14] in detail.

We are interested in the asymptotic stability of one-dimensional stationary solution $\phi(x_1)$ (called planar stationary wave) for the problem (1.1)–(1.3): $\phi(x_1)$ is a solution to the problem

$$-\phi_{x_1x_1} + f_1(\phi)_{x_1} = 0, \tag{1.6}$$

$$\phi(0) = u_b, \qquad \phi(x_1) \to 0 \quad \text{as} \quad x_1 \to \infty. \tag{1.7}$$

To show the stability, it is convenient to introduce the perturbation v and write the solution u in the form

$$u(x,t) = \phi(x_1) + v(x,t).$$

The original problem (1.1)–(1.3) is then reduced to

$$v_t - \Delta v + \nabla \cdot (f(\phi + v) - f(\phi)) = 0, \tag{1.8}$$

$$v(0, x', t) = 0, \tag{1.9}$$

$$v(x, 0) = v_0(x), \tag{1.10}$$

where $v_0(x) = u_0(x) - \phi(x_1)$; notice that $v_0(x) \to 0$ as $x_1 \to \infty$.

Under the convex condition (1.5), the author in [14] showed the asymptotic stability of the planar stationary wave $\phi(x_1)$ by proving a sharp decay estimate for the perturbation $v(x,t)$. To this end we employed the energy method in the partial Fourier space $\hat{\mathbb{R}}^n_+ = \mathbb{R}_+ \times \mathbb{R}^{n-1}_\xi$ which is obtained by taking the Fourier transform with respect to the tangential variable $x' = (x_2, \cdots, x_n) \in \mathbb{R}^{n-1}$; $\xi = (\xi_2, \cdots, \xi_n) \in \mathbb{R}^{n-1}_\xi$ is the Fourier variable corresponding to $x' \in \mathbb{R}^{n-1}$. For the variable $x_1 \in \mathbb{R}_+$ in the normal direction, we use L^2 space (or weighted L^2 space). As the result, for the corresponding linearized problem with $f(\phi + v) - f(\phi)$ replaced by $f'(\phi)v$ in (1.8), we showed the following pointwise estimate with respect to $\xi \in \mathbb{R}^{n-1}_\xi$:

$$|\mathcal{F}v(\cdot, \xi, t)|_{L^2} \le Ce^{-\kappa|\xi|^2 t}|\mathcal{F}v_0(\cdot, \xi)|_{L^2}, \tag{1.11}$$

where C and κ are positive constants. Here $\mathcal{F}$ denotes the Fourier transform with respect to $x' \in \mathbb{R}^{n-1}$ and $|\cdot|_{L^2}$ is the L^2 norm with respect to $x_1 \in \mathbb{R}_+$. This pointwise estimate (1.11) enables us to get the following sharp decay estimate:

$$\|v(t)\|_{\mathcal{L}^2} \le Ct^{-(n-1)/4}\|v_0\|_{L^2(L^1)}, \tag{1.12}$$

where $\|\cdot\|_{\mathcal{L}^2}$ denotes the L^2 norm with respect to $x = (x_1, x') \in \mathbb{R}^n_+$, $\|\cdot\|_{L^2(L^1)}$ is the norm in $L^2(\mathbb{R}_+; L^2_{x'} \cap L^1_{x'})$, and C is a positive constant.

Furthermore, when the planar stationary wave $\phi(x_1)$ is non-degenerate, the author applied the weighted energy method in the partial Fourier space $\hat{\mathbb{R}}^n_+ = \mathbb{R}_+ \times \mathbb{R}^{n-1}_\xi$. Namely, we used the weighted space L^2_α with respect to $x_1 \in \mathbb{R}_+$. In this case, the pointwise estimate (1.11) is improved to

$$|\mathcal{F}v(\cdot, \xi, t)|_{L^2} \le C(1+t)^{-\alpha/2}e^{-\kappa|\xi|^2 t}|\mathcal{F}v_0(\cdot, \xi)|_{L^2_\alpha}, \tag{1.13}$$

where $|\cdot|_{L^2_\alpha}$ denotes the L^2_α norm with respect to $x_1 \in \mathbb{R}_+$. Consequently, we had the decay estimate

$$\|v(t)\|_{\mathcal{L}^2} \le C(1+t)^{-\alpha/2}t^{-(n-1)/4}\|v_0\|_{L^2_\alpha(L^1)}, \tag{1.14}$$

where $\|\cdot\|_{L^2_\alpha(L^1)}$ is the norm in $L^2_\alpha(\mathbb{R}_+; L^2_{x'} \cap L^1_{x'})$. For the above results, we refer to the reeder [14] in detail.

The main purpose of this chapter is to derive the sharp decay estimate (1.11)–(1.14) for the linearized problem of (1.8)–(1.10) with non-convex condition (1.4), i.e.,

$$v_t - \Delta v + \nabla \cdot (f'(\phi)v) = 0 \tag{1.15}$$

with (1.9), (1.10). To overcome the difficulty occured by the non-convex condition, we make a good combination of the weighted energy method in partial Fourier space employed in [14] and the anti-derivative method employed in [2, 9], and get the desired results. Once we obtain the linear stability results for the problem (1.15), (1.9), (1.10), we may apply this results to the asymptotic stability for the nonlinear problem (1.8)–(1.10).

The remainder of this chapter is organized as follows. In Section 2, we introduce function spaces and some preliminaries used in this chapter. Especially, we reformulate our problem (1.15), (1.9), (1.10) by using the anti-derivative method in Section 2.3. In the final section, we treat the half space problem for the reformulated viscous conservation laws and (1.15), (1.9), (1.10), and develop the weighted energy method in the partial Fourier space with the anti-derivative method. In this section, we derive pointwise estimates of solutions and prove the corresponding decay estimates.

2. Preliminaries

2.1 Notations and function spaces

Let us consider functions defined in the half space $\mathbb{R}^n_+ = \mathbb{R}_+ \times \mathbb{R}^{n-1}$. We sometimes write the space variable $x = (x_1, \cdots, x_n) \in \mathbb{R}^n_+$ as $x = (x_1, x')$ with $x_1 \in \mathbb{R}_+$ and $x' = (x_2, \cdots, x_n) \in \mathbb{R}^{n-1}$. The symbols $\nabla = (\partial_{x_1}, \cdots, \partial_{x_n})$ and $\Delta = \sum_{j=1}^n \partial_{x_j}^2$ denote the standard gradient and Laplacian with respect to $x = (x_1, \ldots, x_n)$, respectively. The symbol $\nabla_{x'} = (\partial_{x_2}, \cdots, \partial_{x_n})$ denotes the gradient of the tangential direction with respect to $x = (x_2, \ldots, x_n)$. Thus we have $\nabla \cdot g = \sum_{j=1}^n \partial_{x_j} g_j$ for $g = (g_1, \cdots, g_n)$, and $\nabla_{x'} \cdot g_* = \sum_{j=2}^n \partial_{x_j} g_j$ for $g_* = (g_2, \cdots, g_n)$. Let $\hat{v}(x_1, \xi)$ be the Fourier transform of $v(x_1, x')$ with respect to $x' \in \mathbb{R}^{n-1}$:

$$\hat{v}(x_1, \xi) = \mathcal{F}[v(x_1, \cdot)](\xi) = (2\pi)^{-(n-1)/2} \int_{\mathbb{R}^{n-1}} v(x_1, x') e^{-ix' \cdot \xi} dx', \tag{2.1}$$

where $\xi = (\xi_2, \cdots, \xi_n) \in \mathbb{R}^{n-1}_\xi$ is the Fourier variable corresponding to $x' = (x_2, \cdots, x_n) \in \mathbb{R}^{n-1}$ and $x' \cdot \xi = \sum_{j=2}^n x_j \xi_j$.

Let $1 \le p \le \infty$. We denote by $L^p_{x'} = L^p(\mathbb{R}^{n-1})$ the L^p space with respect to $x' \in \mathbb{R}^{n-1}$, with the norm $\| \cdot \|_{L^p_{x'}}$. For a nonnegative integer s, we denote by $H^s_{x'} = H^s(\mathbb{R}^{n-1})$ the Sobolev space over $\mathbb{R}^{n-1}$ with the norm

$$\|v\|_{H^s_{x'}} = \Big(\sum_{k=0}^{s} \|\partial_{x'}^k v\|_{L^2_{x'}}^2 \Big)^{1/2},$$

where $\partial_{x'}^k$ denotes the totality of all the k-th order derivatives with respect to $x' \in \mathbb{R}^{n-1}$. Also, we denote by $L^p(\mathbb{R}_+)$ the L^p space with respect to $x_1 \in \mathbb{R}_+$, with the norm $| \cdot |_{L^p}$. For a nonnegative integer s, we denote by $H^s(\mathbb{R}_+)$ the Sobolev space over $\mathbb{R}_+$, with the norm $| \cdot |_{H^s}$. For $\alpha \in \mathbb{R}$, we denote by $L^2_\alpha(\mathbb{R}_+)$ the weighted L^2 space over $\mathbb{R}_+$ with the norm

$$|v|_{L^2_\alpha} = \Big(\int_0^\infty (1 + x_1)^\alpha |v(x_1)|^2 dx_1 \Big)^{1/2}.$$

Now we introduce function spaces over the half space $\mathbb{R}^n_+ = \mathbb{R}_+ \times \mathbb{R}^{n-1}$. Let $1 \le p,\ q \le \infty$, s be a nonnegative integer, and $\alpha \in \mathbb{R}$. The space $L^q(L^p) = L^q(\mathbb{R}_+; L^p_{x'})$ consists of L^q functions of $x_1 \in \mathbb{R}_+$ with values in $L^p_{x'}$ with respect to $x' \in \mathbb{R}^{n-1}$. The norm is denoted by $\|\cdot\|_{L^q(L^p)}$. When $q = p$, we simply write as

$$\mathcal{L}^p = L^p(L^p), \qquad \|\cdot\|_{\mathcal{L}^p} = \|\cdot\|_{L^p(L^p)}.$$

The space $H^s(L^p) = H^s(\mathbb{R}_+; L^p_{x'})$ consists of H^s functions of $x_1 \in \mathbb{R}_+$ with values in $L^p_{x'}$ with respect to $x' \in \mathbb{R}^{n-1}$. The norm is denoted by $\|\cdot\|_{H^s(L^p)}$. Also, $L^2_\alpha(L^p) = L^2_\alpha(\mathbb{R}_+; L^p_{x'})$ denotes the space of L^2_α functions of $x_1 \in \mathbb{R}_+$ with values in $L^p_{x'}$ with respect to $x' \in \mathbb{R}^{n-1}$. The norm is denoted by

$$\|v\|_{L^2_\alpha(L^p)} = \Big(\int_0^\infty (1+x_1)^\alpha \|v(x_1,\cdot)\|^2_{L^p_{x'}} dx_1\Big)^{1/2}.$$

We sometimes use

$$\mathcal{L}^2_\alpha = L^2_\alpha(L^2), \qquad \|\cdot\|_{\mathcal{L}^2_\alpha} = \|\cdot\|_{L^2_\alpha(L^2)}.$$

$L^2(H^s) = L^2(\mathbb{R}_+; H^s_{x'})$ denotes the space of L^2 functions of $x_1 \in \mathbb{R}_+$ with values in $H^s_{x'}$ with respect to $x' \in \mathbb{R}^{n-1}$, whose norm is given by

$$\begin{aligned}\|v\|_{L^2(H^s)} &= \Big(\int_0^\infty \|v(x_1,\cdot)\|^2_{H^s_{x'}} dx_1\Big)^{1/2}\\ &= \Big(\sum_{k=0}^{s}\int_0^\infty \|\partial^k_{x'} v(x_1,\cdot)\|^2_{L^2_{x'}} dx_1\Big)^{1/2} = \Big(\sum_{k=0}^{s} \|\partial^k_{x'} v\|^2_{\mathcal{L}^2}\Big)^{1/2}.\end{aligned}$$

By the definition (2.1) of the Fourier transform, we see that

$$\sup_{\xi \in \mathbb{R}^{n-1}_\xi} |\hat{v}(\cdot,\xi)|_{L^2} \le C\|v\|_{L^2(L^1)} \tag{2.2}$$

with $C = (2\pi)^{-(n-1)/2}$. Also, it follows from the Plancherel theorem that

$$\|\partial^k_{x'} v\|_{\mathcal{L}^2_\alpha} = \Big(\int_{\mathbb{R}^{n-1}_\xi} |\xi|^{2k} |\hat{v}(\cdot,\xi)|^2_{L^2_\alpha} d\xi\Big)^{1/2}. \tag{2.3}$$

Let $T > 0$ and let X be a Banach space defined on the half space $\mathbb{R}^n_+$. Then $C([0,T]; X)$ denotes the space of continuous functions of $t \in [0,T]$ with values in X.

In this paper, positive constants will be denoted by C or c.

2.2 Stationary solution

We review the results on the stationary problem (1.6)–(1.7). For the details, we refer the reader to [2, 8, 11–13].

Proposition 2.1 ([8]). *Assume the condition* (1.4). *Then $f_1'(0) \le 0$ is necessary for the existence of solutions to the stationary problem* (1.6)–(1.7). *Conversely, under the condition $f_1'(0) \le 0$, we have the following existence result:*

(i) *Non-degenerate case where $f_1'(0) < 0$: In this case the stationary problem* (1.6)–(1.7) *admits a unique smooth solution $\phi(x_1)$ with $\phi_{x_1} > 0$ (resp. $\phi_{x_1} < 0$), provided that $u_b < 0$ (resp. $0 < u_b$ and $f_1(u) < f_1(0)$ for $0 < u < u_b$). The solution verifies*

$$|\phi(x_1)| \le Ce^{-cx_1}, \qquad x_1 > 0,$$

where C and c are positive constants.

(ii) *Degenerate case where* $f_1'(0) = 0$*: In this case the problem* (1.6)–(1.7) *admits a unique smooth solution* $\phi(x_1)$ *if and only if* $u_b < 0$*. The solution verifies* $\phi_{x_1} > 0$ *and*

$$|\phi(x_1)| \le C(1 + x_1)^{-1/q}, \qquad x_1 > 0,$$

where q *is the degeneracy exponent of* f_1 *and* C *is a positive constant.*

In this chapter we only treat the stationary solutions $\phi(x_1)$ with $\phi_{x_1} > 0$ and discuss their stability; however, we must get the similar stability result of the monotone decreasing stationary solutions by using the same argument introduced in this chapter. (We refer the reader to [2].)

2.3 Reformulated problem

In this subsection we reformulate our problem by the anti-derivative method. To this end we introduce a new function $z(x,t)$ as

$$z(x,t) = -\int_{x_1}^{\infty} v(y, x', t)\, dy. \tag{2.4}$$

Here, we assume the integrability of $v(x,t)$ over $\mathbb{R}_+$. This transformation is motivated by the argument in Liu-Nishihara [9]. By using (2.4), we can reformulate (1.8)–(1.10) in terms of $z(x,t)$ as

$$z_t - \Delta z + f'(\phi) \cdot \nabla z + \int_{x_1}^{\infty} \phi_{x_1} f_*''(\phi) \cdot \nabla_{x'} z\, dy = -g_1 + \nabla_{x'} \cdot h_*, \tag{2.5}$$

$$z_{x_1}(0, x', t) = 0, \tag{2.6}$$

$$z(x,0) = z_0(x), \tag{2.7}$$

where $z_0(x) = -\int_{x_1}^{\infty}(u_0(y,x') - \phi(y))dy$, $f_*''(\phi) = (f_2''(\phi), \cdots, f_n''(\phi))$, and g_1, $\nabla_{x'} \cdot h_*$ are nonlinear terms defined by $h_* = (h_2, \cdots, h_n)$ and

$$g_j = f_j(\phi + z_{x_1}) - f_j(\phi) - f_j'(\phi) z_{x_1}, \qquad h_j = \int_{x_1}^{\infty} g_j\, dy.$$

Once we obtain the solution for the problem (2.5)–(2.7), the differentiation $v = z_{x_1}$ is the solution for (1.8)–(1.10). Namely, we will apply the weighted energy method in the partial Fourier space and try ot derive the global solution in time to the reformulated problem (2.5)–(2.7). We will discuss this reformulated problem in Section 3 to prove our main theorems.

2.4 Weight function

We introduce the weight function employed in the weighted energy method. Our weight function is defined as

$$w(u) = (-e^{Au} + 1)/f_1(u) \qquad \text{for} \qquad u \in [u_b, 0], \tag{2.8}$$

where A is a positive constant determined in Lemma 2.2. This weight function is very important to derive a priori estimate in the latter section. For this weight function, we obtain the following lemma.

Lemma 2.2 ([2]). *Suppose that $f_1(u)$ satisfies* (1.4). *Let $w(u)$ be the weight function defined in* (2.8). *Then there exists a positive constant δ such that if $A \geq \delta$, then $w(u)$ satisfies the following conditions:*

$$\begin{aligned} &\text{(i) } (wf_1)'(u) < 0 \qquad \textit{for} \quad u \in [u_b, 0], \\ &\text{(ii) } (wf_1)''(u) < 0 \qquad \textit{for} \quad u \in [u_b, 0]. \end{aligned} \tag{2.9}$$

Moreover, let ϕ be the stationary solution constructed in Proposition 2.1. Then the weight function satisfies the following properties.

(i) *Non-degenerate case where $f_1'(0) < 0$: The weight function $w(\phi)$ satisfies*

$$c < w(\phi) < C \qquad \textit{for} \qquad \phi \in [u_b, 0].$$

(ii) *Degenerate case where $f_1'(0) = 0$: The weight function $w(\phi)$ satisfies*

$$c(1 + x_1) < w(\phi) < C(1 + x_1) \qquad \textit{for} \qquad \phi \in [u_b, 0].$$

Here, C and c are some positive constants which independent of x_1.

The detail of the proof is omitted here. For the details, we refer the reader to [2].

3. Asymptotic stability with convergence rates

In the final section, we apply our weighted energy method in the partial Fourier space to the linearized problem. We consider the linearized problem corresponding to the half space problem (2.5)–(2.7). Namely, we consider (2.5) with $g_j = 0$ for $j = 1, \cdots, n$. For this linearized equation, we treat the special situation that

"$f_j(u)$ are linear in $u \in [u_b, 0]$ for $j = 2, \cdots, n$."

Then our initial value problem of the linearized equation is written as

$$z_t - \Delta z + f'(\phi) \cdot \nabla z = 0 \tag{3.1}$$

together with (2.6) and (2.7). Taking the Fourier transform with respect to $x' \in \mathbb{R}^{n-1}$ for the linearized problem (3.1), (2.6), (2.7), we obtain

$$\begin{aligned} &\hat{z}_t - \hat{z}_{x_1x_1} + |\xi|^2\hat{z} + f_1'(\phi)\hat{z}_{x_1} + i\xi \cdot f_*'(\phi)\hat{z} = 0, \\ &\hat{z}_{x_1}(0, \xi, t) = 0, \\ &\hat{z}(x_1, \xi, 0) = \hat{z}_0(x_1, \xi), \end{aligned} \tag{3.2}$$

where $\xi = (\xi_2, \cdots, \xi_n) \in \mathbb{R}^{n-1}_\xi$ is the Fourier variable corresponding to $x' = (x_2, \cdots, x_n) \in \mathbb{R}^{n-1}$, $f_*'(\phi) = (f_2'(\phi), \cdots, f_n'(\phi))$, and $\xi \cdot f_*'(\phi) = \sum_{j=2}^n \xi_j f_j'(\phi)$. This is the formulation of our linearized problem in the partial Fourier space $\hat{\mathbb{R}}^n_+ = \mathbb{R}_+ \times \mathbb{R}^{n-1}_\xi$.

Furthermore, we sometimes use the differentiated problem. We differentiate the problem (3.1), (2.6), (2.7) with respect to x_1. Then this yields our problem (1.15) together with (1.9) and (1.10), and the corresponding problem in the partial Fourier space:

$$\begin{aligned} &\hat{v}_t - \hat{v}_{x_1x_1} + |\xi|^2\hat{v} + (f_1'(\phi)\hat{v})_{x_1} + i\xi \cdot f_*'(\phi)\hat{v} = 0, \\ &\hat{v}(0,\xi,t) = 0, \\ &\hat{v}(x_1,\xi,0) = \hat{v}_0(x_1,\xi). \end{aligned} \tag{3.3}$$

Here we note that $v = z_{x_1}$. By applying the weighted energy method to the above problems, we obtain the pointwise estimate of solutions.

3.1 Energy method

We apply the energy method to the problems (3.2) and (3.3) formulated in the partial Fourier space and derive pointwise estimates of solutions to (3.2). We use L^2 space for the variable $x_1 \in \mathbb{R}_+$ in the normal direction. The result is given as follows.

Theorem 3.1 (Pointwise estimate). *Let $\phi(x_1)$ be a stationary solution with $\phi_{x_1} > 0$. Then the solution to the problem* (3.2) *verifies the following pointwise estimate.*

(i) *Non-degenerate case where $f_1'(0) < 0$: Suppose that $\hat{z}_0(\cdot,\xi) \in H^2(\mathbb{R}_+)$ for each $\xi \in \mathbb{R}^{n-1}_\xi$. Then it holds*

$$|\hat{z}(\cdot,\xi,t)|_{L^2_\alpha} \leq Ce^{-\kappa|\xi|^2t}|\hat{z}_0(\cdot,\xi)|_{L^2_\alpha}, \tag{3.4}$$

$$|\hat{z}_{x_1}(\cdot,\xi,t)|_{L^2} \leq Ce^{-\kappa|\xi|^2t}(|\hat{z}_0(\cdot,\xi)|_{L^2_\alpha} + |(\hat{z}_0)_{x_1}(\cdot,\xi)|_{L^2}), \tag{3.5}$$

$$|\hat{z}_{x_1x_1}(\cdot,\xi,t)|_{L^2} \leq Ce^{-\kappa|\xi|^2t}\big(|\hat{z}_0(\cdot,\xi)|_{L^2_\alpha} + |(\hat{z}_0)_{x_1}(\cdot,\xi)|_{H^1}\big) \tag{3.6}$$

with $\alpha = 0$, for $\xi \in \mathbb{R}^{n-1}_\xi$ and $t \geq 0$, where $|\cdot|_{L^2_\alpha}$ denotes the L^2_α norm with respect to $x_1 \in \mathbb{R}_+$, and C and κ are positive constants.

(ii) *Degenerate case where $f_1'(0) = 0$: Suppose that $\hat{z}_0(\cdot,\xi) \in L^2_1(\mathbb{R}_+)$ and $(\hat{z}_0)_{x_1}(\cdot,\xi) \in H^1(\mathbb{R}_+)$ for each $\xi \in \mathbb{R}^{n-1}_\xi$. Then it holds that* (3.4)–(3.6) *with $\alpha = 1$.*

As a simple corollary we have the following decay estimate.

Corollary 3.2 (Decay estimate). *Assume the same conditions of Proposition 3.1. Then the solution to the problem* (3.1), (2.6), (2.7) *satisfies the following decay estimate.*

(i) *Non-degenerate case where $f_1'(0) < 0$: Suppose that $z_0 \in H^2(L^1)$. Then this yields*

$$\|\partial^k_{x'}z(t)\|_{\mathcal{L}^2_\alpha} \leq Ct^{-(n-1)/4-k/2}\|z_0\|_{L^2_\alpha(L^1)}, \tag{3.7}$$

$$\|\partial^k_{x'}z_{x_1}(t)\|_{\mathcal{L}^2} \leq Ct^{-(n-1)/4-k/2}\big(\|z_0\|_{L^2_\alpha(L^1)} + \|(z_0)_{x_1}\|_{L^2(L^1)}\big), \tag{3.8}$$

$$\|\partial^k_{x'}z_{x_1x_1}(t)\|_{\mathcal{L}^2} \leq Ct^{-(n-1)/4-k/2}\big(\|z_0\|_{L^2_\alpha(L^1)} + \|(z_0)_{x_1}\|_{H^1(L^1)}\big) \tag{3.9}$$

with $\alpha = 0$, for $t > 0$, where $k \geq 0$ is an integer and C is a positive constant.

(ii) *Degenerate case where $f_1'(0) = 0$: Suppose that $z_0 \in L^2_1(L^1)$ and $(z_0)_{x_1} \in H^1(L^1)$. Then this yields that* (3.7)–(3.9) *with $\alpha = 1$.*

Proof of Theorem 3.1. Throughout this proof, we use the weighted L^2 norm:

$$|a|_{L^2_w} = \Big(\int_0^\infty w(\phi(x_1))|a(x_1)|^2 dx_1\Big)^{1/2},$$

where w is the weight function defined by (2.8). For this weighted norm, by using Lemma 2.2, we see the following properties.

$$\begin{aligned} &c|\cdot|_{L^2} \le |\cdot|_{L^2_w} \le C|\cdot|_{L^2} \quad \text{for the non-degenerate case} : f'(0) < 0, \\ &c|\cdot|_{L^2_1} \le |\cdot|_{L^2_w} \le C|\cdot|_{L^2_1} \quad \text{for the degenerate case} : f'(0) = 0. \end{aligned} \tag{3.10}$$

We prove (i) and (ii) in Theorem 3.1 in parallel. We first derive (3.4). We multiply $(3.2)_1$ by $w(\phi)\bar{\hat{z}}$ and take the real part, obtaining

$$\frac{1}{2}w(\phi)\frac{\partial}{\partial t}|\hat{z}|^2 + \frac{\partial}{\partial x_1}\mathcal{F}_1 + \mathcal{D}_1 = 0, \tag{3.11}$$

where

$$\begin{aligned} \mathcal{D}_1 &= w(\phi)(|\hat{z}_{x_1}|^2 + |\xi|^2|\hat{z}|^2) - \frac{1}{2}(wf_1)''(\phi)\phi_{x_1}|\hat{z}|^2, \\ \mathcal{F}_1 &= \frac{1}{2}(wf_1)'(\phi)|\hat{z}|^2 - w(\phi)\mathrm{Re}(\bar{\hat{z}}\hat{z}_{x_1}), \end{aligned}$$

and w is a weight function defined by (2.8). By virtue of $(2.9)_2$ in Lemma 2.2, we have

$$\mathcal{D}_1 \ge cw(\phi)(|\hat{z}_{x_1}|^2 + |\xi|^2|\hat{z}|^2) + c\phi_{x_1}|\hat{z}|^2, \tag{3.12}$$

where c is a some positive constant. Therefore, integrating (3.11) in $x_1 \in \mathbb{R}_+$, we get

$$\frac{\partial}{\partial t}|\hat{z}|^2_{L^2_w} + c_1\hat{D}_1 - (wf_1)'(u_b)|\hat{z}(0,\xi,t)|^2 \le 0 \tag{3.13}$$

with a positive constant c_1, where

$$\hat{D}_1 = |\hat{z}_{x_1}|^2_{L^2_w} + |\xi|^2|\hat{z}|^2_{L^2_w} + |\sqrt{\phi_{x_1}}\hat{z}|^2_{L^2}. \tag{3.14}$$

Here, by virtue of $(2.9)_1$, the last term of the left-hand side of (3.13) is positive. We multiply (3.13) by $e^{\kappa|\xi|^2 t}$ ($\kappa > 0$) to get

$$\frac{\partial}{\partial t}(e^{\kappa|\xi|^2 t}|\hat{z}|^2_{L^2_w}) + e^{\kappa|\xi|^2 t}(c_1\hat{D}_1 - \kappa|\xi|^2|\hat{z}|^2_{L^2_w}) \le 0. \tag{3.15}$$

Noting that $\hat{D}_1 \ge |\xi|^2|\hat{z}|^2_{L^2_w}$, we choose $\kappa > 0$ such that $\kappa < c_1$ and integrate (3.15) over $[0,t]$. This yields

$$e^{\kappa|\xi|^2 t}|\hat{z}(\cdot,\xi,t)|^2_{L^2_w} + \int_0^t e^{\kappa|\xi|^2\tau}\hat{D}_1(\xi,\tau)d\tau \le C|\hat{z}_0(\cdot,\xi)|^2_{L^2_w}, \tag{3.16}$$

where C is a positive constant.

We next prove (3.5). Multiplying $(3.3)_1$ by $\bar{\hat{v}}$ and taking the real part, then we have

$$\frac{1}{2}\frac{\partial}{\partial t}|\hat{v}|^2 + \frac{\partial}{\partial x_1}\mathcal{F}_2 + \mathcal{D}_2 = 0, \tag{3.17}$$

where

$$\mathcal{D}_2 = |\hat{v}_{x_1}|^2 + |\xi|^2|\hat{v}|^2 + \frac{1}{2}f_1''(\phi)\phi_{x_1}|\hat{v}|^2,$$

$$\mathcal{F}_2 = \frac{1}{2}f_1'(\phi)|\hat{v}|^2 - \mathrm{Re}(\bar{\hat{v}}\hat{v}_{x_1}).$$

We integrate (3.17) in $x_1 \in \mathbb{R}_+$ to obtain

$$\frac{\partial}{\partial t}|\hat{v}|_{L^2}^2 + 2\hat{D}_2 \leq C|\hat{v}|_{L^2}^2, \tag{3.18}$$

where $\hat{D}_2 = |\hat{v}_{x_1}|_{L^2}^2 + |\xi|^2|\hat{v}|_{L^2}^2$ and C is a positive constant. We multiply (3.18) by $e^{\kappa|\xi|^2 t}$ $(\kappa > 0)$ to get

$$\frac{\partial}{\partial t}(e^{\kappa|\xi|^2 t}|\hat{v}|_{L^2}^2) + e^{\kappa|\xi|^2 t}(2\hat{D}_2 - \kappa|\xi|^2|\hat{v}|_{L^2}^2) \leq Ce^{\kappa|\xi|^2 t}|\hat{v}|_{L^2}^2. \tag{3.19}$$

Then we choose $\kappa > 0$ such that $\kappa < 2$ and integrate (3.19) over $[0, t]$. This yields

$$e^{\kappa|\xi|^2 t}|\hat{v}(\cdot,\xi,t)|_{L^2}^2 + \int_0^t e^{\kappa|\xi|^2\tau}\hat{D}_2(\xi,\tau)d\tau \leq C|\hat{v}_0(\cdot,\xi)|_{L^2}^2 + C\int_0^t e^{\kappa|\xi|^2\tau}|\hat{v}(\cdot,\xi,\tau)|_{L^2}^2 d\tau.$$

Noting that $v = z_{x_1}$ and $|\hat{v}|_{L^2} \leq C|\hat{v}|_{L^2_w}$, we apply (3.16) to the above inequality. Then we get

$$e^{\kappa|\xi|^2 t}|\hat{v}(\cdot,\xi,t)|_{L^2}^2 + \int_0^t e^{\kappa|\xi|^2\tau}\hat{D}_2(\xi,\tau)d\tau \leq C(|\hat{z}_0(\cdot,\xi)|_{L^2_w}^2 + |\hat{v}_0(\cdot,\xi)|_{L^2}^2). \tag{3.20}$$

We shall show (3.6). Multiplying $(3.3)_1$ by $-\bar{\hat{v}}_{x_1x_1}$ and taking the real part, then we have

$$\frac{1}{2}\frac{\partial}{\partial t}|\hat{v}_{x_1}|^2 + \frac{\partial}{\partial x_1}\mathcal{F}_3 + \mathcal{D}_3 = 0, \tag{3.21}$$

where

$$\begin{aligned}\mathcal{D}_3 = {} & |\hat{v}_{x_1x_1}|^2 + |\xi|^2|\hat{v}_{x_1}|^2 + \frac{3}{2}f_1''(\phi)\phi_{x_1}|\hat{v}_{x_1}|^2 \\ & - \frac{1}{2}\big((f_1f_1'')'f_1\big)'(\phi)\phi_{x_1}|\hat{v}|^2 - \frac{1}{2}i\xi(f_1f_*'')'(\phi)\phi_{x_1}|\hat{v}|^2, \\ \mathcal{F}_3 = {} & \frac{1}{2}f_1'(\phi)|\hat{v}_{x_1}|^2 + \frac{1}{2}\{(f_1f_1'')'(\phi) + i\xi f_*''(\phi)\}\phi_{x_1}|\hat{v}|^2 \\ & - \mathrm{Re}(\hat{v}_t\bar{\hat{v}}_{x_1}) - (|\xi|^2 + f_1''(\phi)\phi_{x_1} + i\xi f_*'(\phi))\mathrm{Re}(\hat{v}\bar{\hat{v}}_{x_1}).\end{aligned}$$

Integrating (3.21) in $x_1 \in \mathbb{R}_+$, we have

$$\frac{\partial}{\partial t}|\hat{v}_{x_1}|_{L^2}^2 + 2\hat{D}_3 \le C(|\hat{v}|_{H^1}^2 + |\xi||\hat{v}|_{L^2}^2), \tag{3.22}$$

where $\hat{D}_3 = |\hat{v}_{x_1x_1}|_{L^2}^2 + |\xi|^2|\hat{v}_{x_1}|_{L^2}^2$ and C is a positive constant. We multiply (3.22) by $e^{\kappa|\xi|^2 t}$ ($\kappa > 0$) to get

$$\frac{\partial}{\partial t}(e^{\kappa|\xi|^2 t}|\hat{v}_{x_1}|_{L^2}^2) + e^{\kappa|\xi|^2 t}(2\hat{D}_3 - \kappa|\xi|^2|\hat{v}_{x_1}|_{L^2}^2) \le Ce^{\kappa|\xi|^2 t}(|\hat{v}|_{H^1}^2 + |\xi||\hat{v}|_{L^2}^2). \tag{3.23}$$

Then we choose $\kappa > 0$ such that $\kappa < 2$ and integrate (3.23) over $[0,t]$. This yields

$$\begin{aligned} &e^{\kappa|\xi|^2 t}|\hat{v}_{x_1}(\cdot,\xi,t)|_{L^2}^2 + \int_0^t e^{\kappa|\xi|^2\tau}\hat{D}_3(\xi,\tau)d\tau \\ &\le C|(\hat{v}_0)_{x_1}(\cdot,\xi)|_{L^2}^2 + C\int_0^t e^{\kappa|\xi|^2\tau}(|\hat{v}(\cdot,\xi,\tau)|_{H^1}^2 + |\xi|^2|\hat{v}(\cdot,\xi,\tau)|_{L^2}^2)d\tau. \end{aligned}$$

Thus, employing (3.16) and (3.20) to the above inequality, we obtain

$$e^{\kappa|\xi|^2 t}|\hat{v}_{x_1}(\cdot,\xi,t)|_{L^2}^2 + \int_0^t e^{\kappa|\xi|^2\tau}\hat{D}_3(\xi,\tau)d\tau \le C(|\hat{z}_0(\cdot,\xi)|_{L^2_w}^2 + |\hat{v}_0(\cdot,\xi)|_{H^1}^2). \tag{3.24}$$

Finally, we apply (3.10) to the estimates (3.16), (3.20) and (3.24). Then this gives the desired estimates (3.4)–(3.6) with $\alpha = 0, 1$. Hence the proof of Theorem 3.1 is completed. □

Here, for later use, we derive the corresponding time weighted estimate. We multiply (3.15) (or (3.19), (3.23)) with $0 < \kappa \le c_1$ (or $0 < \kappa < 2$) by $(1+t)^\gamma$ ($\gamma \ge 0$) and integrate over $[0,t]$. Then this yields the desired estimate:

$$\begin{aligned} &(1+t)^\gamma e^{\kappa|\xi|^2 t}|\hat{z}(\cdot,\xi,t)|_{L^2}^2 + \int_0^t (1+\tau)^\gamma e^{\kappa|\xi|^2\tau}\hat{D}_1(\xi,\tau)d\tau \\ &\qquad \le C|\hat{z}_0(\cdot,\xi)|_{L^2}^2 + \gamma C\int_0^t (1+\tau)^{\gamma-1}e^{\kappa|\xi|^2\tau}|\hat{z}(\cdot,\xi,\tau)|_{L^2}^2 d\tau, \end{aligned} \tag{3.25}$$

$$\begin{aligned} &(1+t)^\gamma e^{\kappa|\xi|^2 t}|\hat{v}(\cdot,\xi,t)|_{L^2}^2 + \int_0^t (1+\tau)^\gamma e^{\kappa|\xi|^2\tau}\hat{D}_2(\xi,\tau)d\tau \\ &\qquad \le C|\hat{v}_0(\cdot,\xi)|_{L^2}^2 + C\int_0^t (1+\tau)^\gamma e^{\kappa|\xi|^2\tau}|\hat{v}(\cdot,\xi,\tau)|_{L^2}^2 d\tau, \end{aligned} \tag{3.26}$$

$$\begin{aligned} &(1+t)^\gamma e^{\kappa|\xi|^2 t}|\hat{v}_{x_1}(\cdot,\xi,t)|_{L^2}^2 + \int_0^t (1+\tau)^\gamma e^{\kappa|\xi|^2\tau}\hat{D}_3(\xi,\tau)d\tau \\ &\le C|(\hat{v}_0)_{x_1}(\cdot,\xi)|_{L^2}^2 + C\int_0^t (1+\tau)^\gamma e^{\kappa|\xi|^2\tau}(|\hat{v}(\cdot,\xi,\tau)|_{H^1}^2 + |\xi|^2|\hat{v}(\cdot,\xi,\tau)|_{L^2}^2)d\tau, \end{aligned} \tag{3.27}$$

where $\gamma \ge 0$, and C is a positive constant. These will be used in the next subsection.

Proof of Corollary 3.2. By virtue of the Plancherel theorem, we have (2.3). Substituting (3.4) into (2.3), we obtain

$$\begin{aligned}\|\partial_{x'}^k z(t)\|_{\mathcal{L}^2_\alpha}^2 &\le C\int_{\mathbb{R}^{n-1}_\xi} |\xi|^{2k}e^{-\kappa|\xi|^2t}|\hat{z}_0(\cdot,\xi)|_{L^2_\alpha}^2 d\xi\\ &\le C\sup_{\xi\in\mathbb{R}^{n-1}_\xi}|\hat{z}_0(\cdot,\xi)|_{L^2_\alpha}^2\int_{\mathbb{R}^{n-1}_\xi}|\xi|^{2k}e^{-\kappa|\xi|^2t}d\xi\\ &\le Ct^{-(n-1)/2-k}\|z_0\|_{L^2_\alpha(L^1)}^2,\end{aligned}$$

where C is a positive constant. Here we used (2.2) and the simple inequality

$$\int_{\mathbb{R}^{n-1}_\xi}|\xi|^{2k}e^{-\kappa|\xi|^2t}d\xi \le Ct^{-(n-1)/2-k}$$

with a constant C. By applying the same argument to the pointwise estimates (3.5) and (3.6) we can derive the decay estimate (3.8) and (3.9), respectively. Thus this completes the proof. □

3.2 Weighted energy method

In the last subsection, we restrict to the non-degenerate case $f_1'(0) < 0$ and apply the weighted energy method to the problems (3.2) and (3.3). This yields sharp pointwise estimates of solutions to (3.2). We use the weighted space $L^2_\alpha(\mathbb{R}_+)$ ($\alpha \ge 0$) for $x_1 \in \mathbb{R}_+$ in the normal direction and this gives the additional decay $(1+t)^{-\alpha/2}$. The result is stated as follows.

Theorem 3.3 (Pointwise estimate)**.** *Let $f_1'(0) < 0$ and let $\phi(x_1)$ be a stationary solution with $\phi_{x_1} > 0$. Let $\alpha \ge 0$ and suppose that $\hat{z}_0(\cdot,\xi) \in L^2_\alpha(\mathbb{R}_+)$ and $(\hat{z}_0)_{x_1}(\cdot,\xi) \in H^1(\mathbb{R}_+)$ for each $\xi \in \mathbb{R}^{n-1}_\xi$. Then the solution to the problem* (3.2) *verifies the pointwise estimate*

$$|\hat{z}(\cdot,\xi,t)|_{L^2} \le C(1+t)^{-\alpha/2}e^{-\kappa|\xi|^2t}|\hat{z}_0(\cdot,\xi)|_{L^2_\alpha}, \tag{3.28}$$

$$|\hat{z}_{x_1}(\cdot,\xi,t)|_{L^2} \le C(1+t)^{-\alpha/2}e^{-\kappa|\xi|^2t}(|\hat{z}_0(\cdot,\xi)|_{L^2_\alpha} + |(\hat{z}_0)_{x_1}(\cdot,\xi)|_{L^2}), \tag{3.29}$$

$$|\hat{z}_{x_1x_1}(\cdot,\xi,t)|_{L^2} \le C(1+t)^{-\alpha/2}e^{-\kappa|\xi|^2t}\big(|\hat{z}_0(\cdot,\xi)|_{L^2_\alpha} + |(\hat{z}_0)_{x_1}(\cdot,\xi)|_{H^1}\big) \tag{3.30}$$

for $\xi \in \mathbb{R}^{n-1}_\xi$ and $t \ge 0$, where the norms $|\cdot|_{L^2}$, $|\cdot|_{H^1}$ and $|\cdot|_{L^2_\alpha}$ are with respect to $x_1 \in \mathbb{R}_+$, and C and κ are positive constants.

As an easy consequence, we have the following decay estimate.

Corollary 3.4 (Decay estimate)**.** *Assume the same conditions of Theorem 3.3. Let $z_0 \in L^2_\alpha(L^1)$ and $(z_0)_{x_1} \in H^1(L^1)$ for $\alpha \ge 0$. Then the solution to the problem* (3.1), (2.6), (2.7) *satisfies the decay estimate*

$$\begin{aligned}&\|\partial_{x'}^k z(t)\|_{\mathcal{L}^2} \le C(1+t)^{-\alpha/2}t^{-(n-1)/4-k/2}\|z_0\|_{L^2_\alpha(L^1)},\\ &\|\partial_{x'}^k z_{x_1}(t)\|_{\mathcal{L}^2} \le C(1+t)^{-\alpha/2}t^{-(n-1)/4-k/2}\big(\|z_0\|_{L^2_\alpha(L^1)} + \|(z_0)_{x_1}\|_{L^2(L^1)}\big),\\ &\|\partial_{x'}^k z_{x_1x_1}(t)\|_{\mathcal{L}^2} \le C(1+t)^{-\alpha/2}t^{-(n-1)/4-k/2}\big(\|z_0\|_{L^2_\alpha(L^1)} + \|(z_0)_{x_1}\|_{H^1(L^1)}\big)\end{aligned}$$

for $t > 0$, where $k \geq 0$ is an integer and C is a positive constant.

The proof of Corollary 3.4 is completely same as the proof of Corollary 3.2 and omitted here.

Proof of Theorem 3.3. We use the weighted energy method to the problems (3.2) and (3.3) formulated in the partial Fourier space. Our computation is similar to the one used in [4, 10, 12, 14] and divide into four steps.

Step 1. First, we show the following space-time weighted energy inequality:

$$\begin{aligned}&(1+t)^{\gamma}e^{\kappa|\xi|^2 t}|\hat{z}(\cdot,\xi,t)|^2_{L^2_\beta}+\int_0^t(1+\tau)^{\gamma}e^{\kappa|\xi|^2\tau}\big(\hat{\mathbf{D}}_\beta(\xi,\tau)+\beta|\hat{z}(\cdot,\xi,\tau)|^2_{L^2_{\beta-1}}\big)d\tau\\&\leq C|\hat{z}_0(\cdot,\xi)|^2_{L^2_\beta}+\gamma C\int_0^t(1+\tau)^{\gamma-1}e^{\kappa|\xi|^2\tau}|\hat{z}(\cdot,\xi,\tau)|^2_{L^2_\beta}d\tau\end{aligned}\tag{3.31}$$

for $\gamma \geq 0$ and $0 \leq \beta \leq \alpha$, where

$$\hat{\mathbf{D}}_\beta = |\hat{z}_{x_1}|^2_{L^2_\beta} + |\xi|^2|\hat{z}|^2_{L^2_\beta} + |\sqrt{\phi_{x_1}}\hat{z}|^2_{L^2_\beta},$$

and C and κ are positive constants. Notice that $\hat{\mathbf{D}}_0$ coincides with $\hat{D}_1$ in (3.14).

To prove (3.31), we use the equality (3.11). Notice that, by virtue of $(2.9)_1$ in Lemma 2.2, we have

$$-\mathcal{F}_1 \geq c_2|\hat{z}|^2 - C|\hat{z}_{x_1}|^2 \tag{3.32}$$

with positive constants c_2 and C. Now we multiply (3.11) by $(1+x_1)^\beta$ $(0 \leq \beta \leq \alpha)$ to get

$$\begin{aligned}&\frac{1}{2}\frac{\partial}{\partial t}\{(1+x_1)^\beta|\hat{z}|^2\} + \{(1+x_1)^\beta\mathcal{F}_1\}_{x_1}\\&\qquad + (1+x_1)^\beta\mathcal{D}_1 + \beta(1+x_1)^{\beta-1}(-\mathcal{F}_1) = 0.\end{aligned}$$

We integrate this equality over $x_1 \in \mathbb{R}_+$ and use (3.12) and (3.32), obtaining

$$\frac{\partial}{\partial t}|\hat{z}|^2_{L^2_{\beta,w}} + c_1\hat{\mathbf{D}}_\beta + 2\beta c_2|\hat{z}|^2_{L^2_{\beta-1}} - (wf_1)'(u_b)|\hat{z}(0,\xi,t)|^2 \leq \beta C|\hat{z}_{x_1}|^2_{L^2_{\beta-1}} \tag{3.33}$$

for $0 \leq \beta \leq \alpha$, where we define

$$|v|_{L^2_{\beta,w}} = \Big(\int_0^\infty(1+x_1)^\beta w\big(\phi(x_1)\big)|v(x_1)|^2dx_1\Big)^{1/2},$$

and C is a positive constant. Here, by virtue of $(2.9)_1$, the last term of the left-hand side of (3.33) is positive. We now observe that

$$|a|^2_{L^2_{\beta-1}} \leq \epsilon|a|^2_{L^2_\beta} + C_\epsilon|a|^2_{L^2}$$

for any $\epsilon > 0$, where C_ϵ is a constant depending on ϵ. We apply this inequality to the term on the right-hand side of (3.33) by taking $a = \hat{z}_{x_1}$. Noting that $\hat{\mathbf{D}}_\beta \geq |\hat{z}_{x_1}|^2_{L^2_\beta}$, we choose $\epsilon > 0$ so small that $\alpha C\epsilon \leq c_1$. This yields

$$\frac{\partial}{\partial t}|\hat{z}|^2_{L^2_{\beta,w}} + c_3\hat{\mathbf{D}}_\beta + 2\beta c_2|\hat{z}|^2_{L^2_{\beta-1}} \leq \beta C|\hat{z}_{x_1}|^2_{L^2} \tag{3.34}$$

for constants c_3 and C. Multiplying (3.34) by $e^{\kappa|\xi|^2 t}$ $(\kappa > 0)$, we obtain

$$\frac{\partial}{\partial t}\{e^{\kappa|\xi|^2 t}|\hat{z}|^2_{L^2_{\beta,w}}\} + e^{\kappa|\xi|^2 t}\big(c_3\hat{\mathbf{D}}_\beta - \kappa|\xi|^2|\hat{z}|^2_{L^2_\beta}\big) + 2\beta c_2 e^{\kappa|\xi|^2 t}|\hat{z}|^2_{L^2_{\beta-1}} \le \beta C e^{\kappa|\xi|^2 t}|\hat{z}_{x_1}|^2_{L^2}.$$

As in (3.15), we choose $\kappa > 0$ such that $\kappa \le c_3$. Then we multiply the resulting inequality by $(1+t)^\gamma$ $(\gamma \ge 0)$ and integrate over $[0,t]$. This yields

$$\begin{aligned}(1+t)^\gamma e^{\kappa|\xi|^2 t}|\hat{z}(\cdot,\xi,t)|^2_{L^2_\beta} &+ \int_0^t (1+\tau)^\gamma e^{\kappa|\xi|^2\tau}\big(\hat{\mathbf{D}}_\beta(\xi,\tau) + \beta|\hat{z}(\cdot,\xi,\tau)|^2_{L^2_{\beta-1}}\big)d\tau \\ &\le C|\hat{z}_0(\cdot,\xi)|^2_{L^2_\beta} + \gamma C\int_0^t (1+\tau)^{\gamma-1}e^{\kappa|\xi|^2\tau}|\hat{z}(\cdot,\xi,\tau)|^2_{L^2_\beta}d\tau \\ &\qquad + \beta C\int_0^t (1+\tau)^\gamma e^{\kappa|\xi|^2\tau}|\hat{z}_{x_1}(\cdot,\xi,\tau)|^2_{L^2}d\tau,\end{aligned} \tag{3.35}$$

where C is a positive constant. Here the last term on the right-hand side of (3.35) is already estimated in (3.25) because $|\hat{z}_{x_1}|^2_{L^2} \le \hat{\mathbf{D}}_0 = \hat{D}_1$. Therefore the proof of (3.31) is complete.

Step 2. Next we show the following estimate for $\alpha \ge 0$:

$$\begin{aligned}&(1+t)^l e^{\kappa|\xi|^2 t}|\hat{z}(\cdot,\xi,t)|^2_{L^2_{\alpha-l}} \\ &+ \int_0^t (1+\tau)^l e^{\kappa|\xi|^2\tau}\big(\hat{\mathbf{D}}_{\alpha-l}(\xi,\tau) + (\alpha-l)|\hat{z}(\cdot,\xi,\tau)|^2_{L^2_{\alpha-l-1}}\big)d\tau \le C|\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha}\end{aligned} \tag{3.36}$$

for each integer l with $0 \le l \le [\alpha]$, where C and κ are positive constants. Note that if $\alpha \ge 0$ is an integer, then (3.36) with $l = \alpha$ gives the desired estimate (3.28).

We prove (3.36) by induction with respect to the integer l with $0 \le l \le [\alpha]$. First we put $\gamma = 0$ and $\beta = \alpha$ in (3.31). This shows that (3.36) holds true for $l = 0$. Now, let $1 \le j \le [\alpha]$ (for $\alpha \ge 1$) and suppose that (3.36) holds true for $l = j-1$. In particular, we suppose that

$$\int_0^t (1+\tau)^{j-1}e^{\kappa|\xi|^2\tau}|\hat{z}(\cdot,\xi,\tau)|^2_{L^2_{\alpha-j}}d\tau \le C|\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha}. \tag{3.37}$$

Then we prove (3.36) for $l = j$. To this end, we put $\gamma = j$ and $\beta = \alpha - j$ in (3.31). This gives

$$\begin{aligned}&(1+t)^j e^{\kappa|\xi|^2 t}|\hat{z}(\cdot,\xi,t)|^2_{L^2_{\alpha-j}} \\ &+ \int_0^t (1+\tau)^j e^{\kappa|\xi|^2\tau}\big(\hat{\mathbf{D}}_{\alpha-j}(\xi,\tau) + (\alpha-j)|\hat{z}(\cdot,\xi,\tau)|^2_{L^2_{\alpha-j-1}}\big)d\tau \\ &\le C|\hat{z}_0(\cdot,\xi)|^2_{L^2_{\alpha-j}} + jC\int_0^t (1+\tau)^{j-1}e^{\kappa|\xi|^2\tau}|\hat{z}(\cdot,\xi,\tau)|^2_{L^2_{\alpha-j}}d\tau \le C|\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha},\end{aligned}$$

where we used (3.37) in the last estimate. This shows that (3.36) holds true also for $l = j$ and therefore the proof of (3.36) is complete.

Step 3. Next, when $\alpha > 0$ is not an integer, we show that

$$\begin{aligned}&(1+t)^\gamma e^{\kappa|\xi|^2 t}|\hat{z}(\cdot,\xi,t)|^2_{L^2} + \int_0^t (1+\tau)^\gamma e^{\kappa|\xi|^2\tau}\hat{D}_1(\xi,\tau)d\tau \\ &\le C(1+t)^{\gamma-\alpha}|\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha}\end{aligned} \tag{3.38}$$

for $\gamma > \alpha$, where $\hat{D}_1$ is defined in (3.12), and C and κ are positive constants. Notice that (3.38) gives the desired estimate (3.28) even if $\alpha > 0$ is not an integer.

To prove (3.38), we recall the inequality (3.25) which is the same as (3.31) with $\beta = 0$. We need to estimate the second term on the right-hand side of (3.25). This can be done by applying the technique due to Nishikawa in [10]. When $\alpha > 0$ is not an integer, we have from (3.36) with $l = [\alpha]$ that

$$\begin{aligned} (1+t)^{[\alpha]} e^{\kappa|\xi|^2 t} |\hat{z}(\cdot,\xi,t)|^2_{L^2_{\alpha-[\alpha]}} &\le C|\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha}, \\ \int_0^t (1+\tau)^{[\alpha]} e^{\kappa|\xi|^2 \tau} |\hat{z}(\cdot,\xi,\tau)|^2_{L^2_{\alpha-[\alpha]-1}} &\le C|\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha}, \end{aligned} \tag{3.39}$$

where C is a positive constant. Now, using a simple interpolation inequality $|a|_{L^2} \le |a|^{\theta}_{L^2_{\theta-1}} |a|^{1-\theta}_{L^2_\theta}$ $(0 \le \theta \le 1)$ and the Hölder inequality, we see that

$$\begin{aligned} &\int_0^t (1+\tau)^{\lambda} |a(\tau)|^2_{L^2} d\tau \\ &\le \Big(\int_0^t (1+\tau)^{\mu} |a(\tau)|^2_{L^2_{\theta-1}} d\tau \Big)^{\theta} \Big(\int_0^t (1+\tau)^{\nu} |a(\tau)|^2_{L^2_\theta} d\tau \Big)^{1-\theta}, \end{aligned} \tag{3.40}$$

provided that $\lambda = \mu\theta + \nu(1-\theta)$ with $0 \le \theta \le 1$. We use (3.40) for $a = e^{\kappa|\xi|^2 t/2} \hat{z}(x_1, \xi, t)$, $\theta = \alpha - [\alpha]$, $\lambda = \gamma - 1$, $\mu = [\alpha]$ and the corresponding ν determined by $\lambda = \mu\theta + \nu(1-\theta)$. Then, using (3.39), we arrive at the estimate

$$\begin{aligned} &\int_0^t (1+\tau)^{\gamma-1} e^{\kappa|\xi|^2 \tau} |\hat{z}(\cdot,\xi,\tau)|^2_{L^2} d\tau \\ &\le C|\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha} \Big(\int_0^t (1+\tau)^{\nu-[\alpha]} d\tau \Big)^{1-\theta} \le C(1+t)^{\gamma-\alpha} |\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha}, \end{aligned}$$

where we have used the fact that $(\nu - [\alpha] + 1)(1-\theta) = \gamma - \alpha$. Substituting this estimate into (3.25), we get the desired estimate (3.38).

Step 4. Finally, we prove (3.29) and (3.30). Employing (3.38), we can estimate the last term of the right-hand side of (3.26). Namely, we obtain

$$\begin{aligned} &(1+t)^{\gamma} e^{\kappa|\xi|^2 t} |\hat{v}(\cdot,\xi,t)|^2_{L^2} + \int_0^t (1+\tau)^{\gamma} e^{\kappa|\xi|^2 \tau} \hat{D}_2(\xi,\tau) d\tau \\ &\le C|\hat{v}_0(\cdot,\xi)|^2_{L^2} + C(1+t)^{\gamma-\alpha} |\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha} \le C(1+t)^{\gamma-\alpha} (|\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha} + |\hat{v}_0(\cdot,\xi)|^2_{L^2}) \end{aligned} \tag{3.41}$$

for $\gamma > \alpha$. Thus this yields (3.29).

On the other hand, by applying (3.38) and (3.41) to (3.27), we get

$$\begin{aligned} &(1+t)^{\gamma} e^{\kappa|\xi|^2 t} |\hat{v}_{x_1}(\cdot,\xi,t)|^2_{L^2} + \int_0^t (1+\tau)^{\gamma} e^{\kappa|\xi|^2 \tau} \hat{D}_3(\xi,\tau) d\tau \\ &\le C(1+t)^{\gamma-\alpha} (|\hat{z}_0(\cdot,\xi)|^2_{L^2_\alpha} + |\hat{v}_0(\cdot,\xi)|^2_{H^1}) \end{aligned}$$

for $\gamma > \alpha$. This means that (3.30). Hence this completes the proof of Theorem 3.3.

□

4. References

[1] Hashimoto, I. & Matsumura, A. (2007). Large time behavior of solutions to an initial boundary value problem on the half line for scalar viscous conservation law, *Methods Appl. Anal.*, Vol. 14, 45–60.

[2] Hashimoto, I. & Ueda, Y. (2011). Anti-derivative method in the half space and application to damped wave equations with non-convex convection, to appear in *Kyushu Journal of Mathematics*.

[3] Il'in, A. M. & Oleinik, O. A. (1964). Behavior of the solution of the Cauchy problem for certain quasilinear equations for unbounded increase of the time, *Amer. Math. Soc. Transl.*, Vol. 42, 19–23.

[4] Kawashima, S. & Matsumura, A. (1985). Asymptotic stability of traveling wave solutions of systems for one-dimensional gas motion, *Commun. Math. Phys.*, Vol. 101, 97–127.

[5] Kawashima, S.; Nishibata, S. & Nishikawa, M. (2003). Asymptotic stability of stationary waves for two-dimensional viscous conservation laws in half plane, *Discrete Contin. Dyn. Syst., Suppl.*, 469–476.

[6] Kawashima, S.; Nishibata, S. & Nishikawa, M. (2004). L^p energy method for multi-dimensional viscous conservation laws and application to the stability of planar waves, *J. Hyperbolic Differential Equations*, Vol. 1, 581–603.

[7] Kawashima, S.; Nishibata, S. & Nishikawa, M. (2004). Asymptotic stability of stationary waves for multi-dimensional viscous conservation laws in half space, preprint.

[8] Liu T.-P.; Matsumura, A. & Nishihara, K. (1998). Behaviors of solutions for the Burgers equation with boundary corresponding to rarefaction waves, *SIAM J. Math. Anal.*, Vol. 29, 293–308.

[9] Liu, T.-P. & Nishihara, K. (1997). Asymptotic behavior for scalar viscous conservation laws with boundary effect, *J. Differential Equations*, Vol. 133, 296–320.

[10] Nishikawa, M. (1998). Convergence rate to the traveling wave for viscous conservation laws, *Funkcial. Ekvac.*, Vol. 41, 107–132.

[11] Ueda, Y. (2008). Asymptotic stability of stationary waves for damped wave equations with a nonlinear convection term, *Adv. Math. Sci. Appl.*, Vol. 18, No. 1, 329–343.

[12] Ueda, Y.; Nakamura, T. & Kawashima, S. (2008). Stability of planar waves for damped wave equations with nonlinear convection in multi-dimensional half space, *Kinetic and Related Models*, Vol. 1, 49–64.

[13] Ueda, Y.; Nakamura, T. & Kawashima, S. (2010). Stability of degenerate stationary waves for viscous gases, *Arch. Rational Mech. Anal.*, Vol. 198, 735–762.

[14] Ueda, Y.; Nakamura, T. & Kawashima, S. (2011). Energy method in the partial Fourier space and application to stability problems in the half space, *J. Differential Equations*, Vol. 250, 1169-1199.

11

Fourier Transform Methods for Option Pricing[*]

Deng Ding
Department of Mathematics, University of Macau, Macao, China

1. Introduction

Since Bachelier, a French mathematician, first tried to give a mathematical definition for Brownian motion and used it to model the dynamics of stock process in 1900, financial mathematics has developed a lot. Black & Scholes (1973) and Merton (1973) respectively used the geometric Brownian motion (GBM) to model the underlying asset's price process so that they opened the gate of easy ways to compute option prices, which led one of the major breakthroughs of modern finance. Many good ideas have been proposed to model the stock pricing processes since then. Merton (1976) first introduced the jumps into the asset price processes in his seminar paper. More recently, a lot of exponential Lévy models, including Kou's model, Variance Gamma (VG) model, Inverse Gaussian (IG) model, Normal Inverse Gaussian (NIG) model, and CGMY model, etc., were proposed to add jumps in the financial models so that they can describe the statistical properties of financial time series better [e.g. see Cont & Tankov (2003) and references there in]. Also, serval stochastic volatility modes were presented [e.g. see Heston (1993), Bates (1996), and Duffie *et al* (2000), etc.]. Empirical financial data indicate that these models are usually more consist with financial markets.

Under assuming that the price of an underlying asset follows a GBM, Black & Scholes (1973) showed that the value of a European option satisfies a boundary problem of heat equation (Black-Scholes equation) so that they derived an explicit formula (Black-Scholes formula) for the value. However, this nice analytic tractability in option pricing can not be carried over to the most exponential Lévy models or the stochastic volatility modes for asset returns. Thus, many new approaches, including efficient numerical methods, for option pricing have been proposed. These approaches and methods can be classified into four major groups: The partial integro-differential equation (PIDE) and various numerical methods for such equations; The Monte Carlo methods via the stochastic simulation techniques for underlying asset price processes; Directly numerical integration and various numerical methods via differential integral transforms; Backward stochastic differential equation and its numerical methods. Each of them has its advantages and disadvantages for different financial models and specific applications.

Since Stein & Stein (1991) first used Fourier inversion method to find the distribution of the underlying asset in a stochastic volatility model, the Fourier transform methods have become a very active field of financial mathematics. Heston (1993) applied the characteristic function

[*] The work was supported by the research grant MYRG136(Y1-L2)-FST11-DD from the Research Committee of University of Macau.

approach to obtain an analytic representation for the valuation of European options in the Fourier domain. Duffie *et al* (2000) offered a comprehensive survey that Fourier transform methods are applicable to a wide range of stochastic processes, the class of exponential affine diffusions [e.g. see Kwok *et al* (2010) and Schmelzle (2010)].

Carr & Madan (1999) pioneered the use of the fast Fourier transform (FFT) technique by mapping the Fourier transform directly to option prices via the characteristic function. Since then, many efficient numerical methods by using FFT techniques have been proposed, and many authors have discussed these methods in rigorous detail. Lee (2004) extended their method significantly and proved an error analysis for these FFT methods. Lewis (2001) generalized this approach to several general payoff functions via the convolution of generalized Fourier transforms.

Recently, some new ideas to improve the Fourier transform methods have been raised. Fang & Oosterlee (2008) proposed the COS method which is based on the Fourier and Fourier-cosine expansion. Feng & Linetsky (2008) presented a new method which involves the relation between Fourier transform and the Hilbert transform, and the Sinc expansion in Hardy spaces. Lord *et al* (2008) extended the FFT-based methods to the CONV method, which is based on the a quadrature technique and relies heavily on Fourier transform.

In this chapter, we demonstrate the application of Fourier transform as a very effective tool in option pricing theory. Together with the fast Fourier transform technique and other important properties of Fourier transform, we survey serval different methods for pricing European options and some path dependent options under different financial models.

This chapter is organized as follows. In the next section, the mathematical formulations that will be used in this chapter are reviewed. They include a brief discussion on Fourier transform with its important properties; an introduction of discrete Fourier transform and the FFT idea; a definition of characteristic functions; and a brief review on the exponential Lévy models and stochastic volatility models in financial mathematics. In Section 3, several Fourier transform methods for pricing European options are introduced. They include the Black-Scholes type formulas, the FFT methods for signal underlying asset and for multi underlying assets, and the Fourier expansion methods. In Section 4, three new Fourier transform methods for pricing path dependent options are considered. They are the CONV method for pricing Bermudan barrier option, the COS method for pricing Bermudan barrier option, and the fast Hilbert transform method for pricing barrier option.

2. Fourier transforms and characteristic functions

2.1 Fourier transforms

Let $g(x)$ be a piecewise continuous real function over $\mathbb{R}$ which satisfies the integrability condition:

$$\int_{-\infty}^{\infty} |g(x)|dx < \infty.$$

The *Fourier transform* of $g(x)$ is defined by

$$\mathcal{F}g(u) = \int_{-\infty}^{\infty} e^{iux} g(x)dx, \quad u \in \mathbb{R}, \tag{1}$$

where $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. Given the function $\mathcal{F}g(u)$, the function $g(x)$ can be recovered by the *Fourier inversion formula*:

$$g(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathrm{i}ux} \mathcal{F}g(u) du, \quad x \in \mathbb{R}. \tag{2}$$

Sometime it may be more convenient to consider the *generalized Fourier transform* on the complex plane. Let $a, b \in \mathbb{R}$ with $a < b$. Assume

$$\int_{-\infty}^{\infty} e^{-ax} |g(x)| dx < \infty$$

and

$$\int_{-\infty}^{\infty} e^{-bx} |g(x)| dx < \infty.$$

Then, the generalized Fourier transform:

$$\mathcal{F}g(z) = \int_{-\infty}^{\infty} e^{\mathrm{i}zx} g(x) dx, \quad z \in \mathbb{C},$$

exists and is analytic for all z in the strip $\mathbb{S} = \{z \in \mathbb{C} : a < \mathrm{Im}\{z\} < b\}$. Moreover, within this strip the generalized Fourier transform may be inverted by integrating along a straight line paralleled to the real axis:

$$g(x) = \frac{1}{2\pi} \int_{\mathrm{i}w-\infty}^{\mathrm{i}w+\infty} e^{-\mathrm{i}xz} \mathcal{F}g(z) dz, \quad x \in \mathbb{R}, \tag{3}$$

with $a < w < b$. Here, $\mathrm{Im}\{\cdot\}$ denotes taking the imaginary part of argument.

The following properties of Fourier transform are useful in this chapter.

P1. Differentiation: $\mathcal{F}\big(g'(x)\big)(u) = -\mathrm{i}u\mathcal{F}g(u)$, where $g'(x)$ is the derivative of $g(x)$.

P2. Modulation: $\mathcal{F}\big(e^{\lambda x} g(x)\big)(u) = \mathcal{F}g(u - \mathrm{i}\lambda)$, where λ is a constant.

P3. Convolution: $\mathcal{F}(f * g)(u) = \mathcal{F}f(u)\mathcal{F}g(u)$, where $f * g$ is the convolution of $f(x)$ and $g(x)$, which is defined by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(y) g(x - y) dy.$$

P4. Relation with Hilbert transform: $\mathcal{F}(\mathrm{sgn} \cdot g)(u) = \mathrm{i}\mathcal{H}(\mathcal{F}g)(u)$, where $\mathrm{sgn}(x)$ is the signum function, and $\mathcal{H}$ is the Hilbert transform, which is defined by the Cauchy principal value integral:

$$\mathcal{H}f(u) = \frac{1}{\pi} P.V. \int_{-\infty}^{\infty} \frac{f(x)}{u - x} dx.$$

P5. Parseval relation: $\langle f, g \rangle = \frac{1}{2\pi} \langle \mathcal{F}f(u), \mathcal{F}g(u) \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product of two square-integrable function defined by

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x) g(x) dx.$$

2.2 Fast Fourier transforms

Let $N = 2^L$ be a power of 2, and let $x = (x_0, x_1, \ldots, x_{N-1})^T$ be a given N-dimensional vector, where v^T presents the transpose of a vector v. The *discrete Fourier transform* of x is another N-dimensional vector $\mathcal{D}x = (y_0, y_1, \ldots, y_{N-1})^T$ is defined by

$$y_p = \sum_{j=0}^{N-1} e^{\mathrm{i}\frac{2\pi}{N}jp} x_j, \quad p = 0, 1, \ldots, N-1. \tag{4}$$

Denote A_N an $N \times N$ matrix whose $(p, j)^{\text{th}}$ entry is given by

$$a_{p\,j} = e^{\mathrm{i}\frac{2\pi}{N}jp}, \quad j, p = 0, 1, \ldots, N-1.$$

Then, the *discrete Fourier transform* of x is given by

$$\mathcal{D}x = A_N x.$$

It is clear that the computation to find $\mathcal{D}x$ requires N^2 steps. However, the *fast Fourier technique* (FFT) would require only $\frac{1}{2}NL = \frac{1}{2}N\log_2 N$ steps. The idea behind the FFT algorithm is to take advantage of the periodicity property of N roots of unity. For the given N-dimensional vector x, denote $x' = (x_0, x_2, \ldots, x_{N-2})^T$ and $x'' = (x_1, x_3, \ldots, x_{N-1})^T$ two $N/2$-dimensional vectors. Then, we find

$$y' = A_M x' \quad \text{and} \quad y'' = A_M x'',$$

where $M = N/2$, $y' = (y'_0, y'_1, \ldots, y'_{M-1})^T$ and $y'' = (y''_0, y''_1, \ldots, y''_{M-1})^T$, and the matrix $A_M = [a_{j\,k}]_{M\times M}$ given by

$$a_{p\,j} = e^{\mathrm{i}\frac{2\pi}{N}jp}, \quad j, p = 0, 1, \ldots, M-1.$$

It is easy to verify that the vector $\mathcal{D}x$, which is defined in (4), now is given by

$$y_p = y'_p + e^{\mathrm{i}\frac{2\pi}{N}p} y''_p, \quad p = 0, 1, \ldots, M-1,$$

$$y_{M+p} = y'_p - e^{\mathrm{i}\frac{2\pi}{N}p} y''_p, \quad p = 0, 1, \ldots, M-1.$$

Instead of performing the matrix-vector multiplication $A_N x$, we now only need to perform two matrix-vector multiplications $A_M x'$ and $A_M x''$ so that the number of operations is reduced from N^2 to $2(N/2)^2 = N^2/2$. The same procedure of reducing the length of the vector by half can be applied repeatedly. Using this FFT algorithm, the total number of operations is reduced from $O(N^2)$ to $O(N\log_2 N)$. The similar FFT algorithm also can be used to calculate the *discrete Fourier inversion transform*:

$$x_p = \sum_{j=0}^{N-1} e^{-\mathrm{i}\frac{2\pi}{N}jp} y_j, \quad p = 0, 1, \ldots, N-1.$$

2.3 Characteristic functions

Let X be a random variable having the density function $f(x)$. Then, the *characteristic function* of X is defined by

$$\varphi_X(u) = \mathbb{E}\left[e^{\mathrm{i}uX}\right] = \int_{-\infty}^{\infty} e^{\mathrm{i}ux} f(x)dx, \quad u \in \mathbb{R},$$

i.e. it is the Fourier transform of the density $f(X)$. Here $\mathbb{E}[\cdot]$ denotes the mathematical expectation. The characteristic functions are uniformly continuous and non-negative definite. Also they possess the following properties:

P6. If X and Y are independent, then $\varphi_{X+Y}(u) = \varphi_X(u)\varphi_Y(u)$.

P7. If $a, b \in \mathbb{R}$ and $Y = aX + b$, then $\varphi_Y(u) = e^{ibu}\varphi_X(au)$.

P8. X and Y have the same distribution function if and only if they have the same characteristic function.

2.4 Exponential Lévy models

An adapted stochastic process X_t with $X_0 = 0$ is called a *Lévy process* if it has independent and stationary increments, and it is continuous in probability. Moreover, every Lévy process has a right continuous with left limits (càdlàd) version which is also a Lévy process. We always work with this càdlàg version.

Let X_t be a Lévy process, and let $\mathcal{B}_0$ be the set family of all Borel sets $U \subset \mathbb{R}$ whose closure $\bar{U}$ does not contain 0. Set

$$\mu\big((0,t], U\big) = \sum_{0<s\leq t} 1_U\big(\Delta X_s\big), \quad t > 0,\ U \in \mathcal{B}_0,$$

and denote $\mu\big(\{0\}, U\big) \equiv 0$. Then $\mu(dt, dx)$ is a σ-finite random measure on $\mathcal{B}(\mathbb{R}_+) \times \mathcal{B}_0$, which is called the *jump measure* of X_t. Set $\nu(U) = \mathbb{E}\big[\mu\big([0,1], U\big)\big]$ for any $U \in \mathcal{B}_0$. It is clear that $\nu(dx)$ is a σ-finite measure on $\mathcal{B}_0$, which is called the *Lévy measure* of X_t. Then, the characteristic function of Lévy process X_t has the *Lévy-Khintchine representation*:

$$\varphi_t(u) = e^{t\psi(u)}, \quad u \in \mathbb{R}, \tag{5}$$

where the characteristic exponent function:

$$\psi(u) = \mathrm{i}\alpha u - \frac{1}{2}\sigma^2 u^2 + \int_{\mathbb{R}} \big(e^{\mathrm{i}ux} - 1 - \mathrm{i}ux 1_{\{|x|\leq 1\}}(x)\big)\nu(dx). \tag{6}$$

Here the triple $(\alpha, \sigma^2, \nu(dx))$ is called the Lévy triple of X_t. The proof of this Lévy-Khintchine decomposition is based on the famous *Lévy-Itô decomposition*, and the detail can be found in Cont & Tankov (2003). Also a Lévy process is a strong Markov process, and a semimartingale. Furthermore, under the condition: $\int_{\{|x|\geq 1\}} e^x \nu(dx) < \infty$, the exponential e^{X_t} is a martingale if and only if

$$\alpha + \frac{1}{2}\sigma^2 + \int_{\mathbb{R}} \big(e^x - 1 - x 1_{\{|x|\leq 1\}}(x)\big)\nu(dx) = 0.$$

To ensure positivity as well as the independence and stationarity of log-returns, in financial mathematics, the price process of the underlying asset is usually modeled as the exponential of a Lévy process X_t:

$$S_t = S_0 e^{X_t}, \quad t \geq 0. \tag{7}$$

There are many models in the financing modeling literature simply correspond to different choices for σ^2 and $\nu(dx)$ to the Lévy process X_t.

The cases that $\sigma^2 \neq 0$ and $\nu(dx)$ is a finite measure, i.e. X_t is a *Lévy jump diffusion*:

$$X_t = \mu t + \sigma B_t + \sum_{k=1}^{N_t} \xi_k, \quad t \geq 0,$$

where B_t is a standard Brownian motion, $\{\xi_k\}$ is a sequence of independent and identically distributed random variables with common density $f_\xi(x)$, and N_t is a Poisson process of intensity λ, such that B_t, $\{\xi_k\}$ and N_t are mutually independent. In these models, the Lévy triple is given by

$$\alpha = \mu + \lambda \int_{\{|x|<1\}} x f_\xi(x) dx, \quad \sigma^2 \quad \text{and} \quad \nu(dx) = \lambda f_\xi(x) dx.$$

These models explain the jump part as the market responses to outside news: good news and bad news. The news arrives according to the Poisson process N_t, and the price changes in response according to the jump size ξ_k.

- The earliest of these models is due to Merton (1976). In this model, ξ_ks are normally distributed, with mean μ_J and standard deviation σ_J, so that the characteristic function of X_t is given by

$$\varphi_t(u) = \exp\Big\{ \mathrm{i}\alpha t u - \frac{1}{2}\sigma^2 t u^2 + \lambda t \big(e^{\mathrm{i}\mu_J u - \frac{1}{2}\sigma_J^2 u^2} - 1\big) \Big\}.$$

- The second one belongs to Kou (2002). In this model, ξ_ks are non-symmetric double exponentially distributed, and the characteristic function of X_t is given by

$$\varphi_t(u) = \exp\Big\{ \mathrm{i}\alpha t u - \frac{1}{2}\sigma^2 t u^2 + \lambda t \Big(\frac{1-\eta^2}{1+\eta^2 u^2} e^{\mathrm{i}\kappa u} - 1 \Big) \Big\},$$

  where η and κ are parameters.

The cases that $\sigma^2 = 0$ and $\nu(\mathbb{R} \setminus \{0\}) = \infty$, i.e., the models have infinite activity in jumps. We only list some examples here.

- The VG model, which is proposed by Madan *et al* (1998): X_t is a variance gamma process with parameters σ, κ and θ, and its characteristic function is given by:

$$\varphi_t(u) = \big(1 - \mathrm{i}\theta\kappa u + \frac{1}{2}\sigma^2 \kappa u^2\big)^{-t/\kappa}.$$

- The NIG model, which is presented by Barndorff-Nielsen (1997): X_t is a normal inverse Gaussian process with parameters α, β and δ, and its characteristic function is given by:

$$\varphi_t(u) = \exp\Big\{ -\delta t \Big(\sqrt{\alpha^2 - (\beta + \mathrm{i}u)^2} - \sqrt{\alpha^2 - \beta^2} \Big) \Big\}.$$

- The CGMY model, which is defined by Carr *et al* (2002): X_t is a CGMY process whose characteristic function is given by

$$\varphi_t(u) = \exp\Big\{ tC\Gamma(-Y) \big[(M - \mathrm{i}u)^Y - M^Y + (G + \mathrm{i}u)^Y - G^Y \big] \Big\},$$

  where the parameters C, G and M are nonnegative, $Y < 2$, and $\Gamma(x)$ is the gamma function.

- The Finite Moment Log Stable model, which is considered by Carr & Wu (2003): X_t is a finite moment log stable process whose characteristic function is given by

$$\varphi_t(u) = \exp\left\{ \mathrm{i}\omega t u - (\mathrm{i}\sigma u)^{\alpha} t \sec \frac{\pi\alpha}{2} \right\},$$

where ω, α and σ are parameters.

2.5 Stochastic volatility models

Although the class of Lévy processes is quite rich, it is sometime insufficient for multi-period financial modeling. Several models combining jumps and stochastic volatility have been appeared in the literature. We only list two such models here.

- The Heston's stochastic volatility model [Heston (1993)]: In this model, the price process S_t is defined by the system of stochastic differential equations (SDEs):

$$\begin{cases} \mathrm{d}S_t = \mu S_t \mathrm{d}t + \sqrt{V_t}\, S_t \mathrm{d}B_t^S, \\ \mathrm{d}V_t = \kappa(\theta - V_t)\mathrm{d}t + \sigma\sqrt{V_t}\, \mathrm{d}B_t^V, \end{cases}$$

with the initial values $S_0 = s_0$ and $V_0 = v_0$, where B_t^S and B_t^V are two standard Brownian motions with correlation ρ (i.e. $\langle B^S, B^V \rangle_t = \rho t$), κ is the mean reversion, θ is the long-run variance level, and σ is the volatility-of-volatility parameter. Although S_t is no longer a Lévy process, under the risk neutral probability the characteristic function of the log-price process $X_t = \log S_t$ is known in closed form:

$$\varphi_t(u) = \frac{\exp\left\{ \frac{\kappa\theta t(\kappa - \mathrm{i}\rho\sigma u)}{\sigma^2} + \mathrm{i}rtu + \mathrm{i}x_0 u \right\}}{\left(\cosh\frac{\gamma t}{2} + \frac{\kappa - \mathrm{i}\rho\sigma u}{\gamma}\sinh\frac{\gamma t}{2}\right)^{2\kappa\theta/\sigma^2}} \exp\left\{ -\frac{(u^2 + \mathrm{i}u)v_0}{\gamma\coth\frac{\gamma t}{2} + \kappa - \mathrm{i}\rho\sigma u} \right\},$$

where $x_0 = \log s_0$, r is the interest rate, and $\gamma = \sqrt{\sigma^2(u^2 + \mathrm{i}u) + (\kappa - \mathrm{i}\rho\sigma u)^2}$.

- The Baste's stochastic volatility model [Bates (1996)]: In this model, the price process S_t and the variance process V_t are given by the system of SDEs:

$$\begin{cases} \mathrm{d}S_t = \mu S_{t-}\mathrm{d}t + \sqrt{V_t}\, S_{t-}\mathrm{d}B_t^S + S_{t-}\mathrm{d}Z_t, \\ \mathrm{d}V_t = \kappa(\theta - V_t)\mathrm{d}t + \sigma\sqrt{V_t}\, \mathrm{d}B_t^V, \end{cases}$$

where B_t^S and B_t^V are two standard Brownian motions with correlation ρ, and Z_t is a compound Poisson process with intensity λ and log-normal distribution of jump sizes such that if k is its jump size then

$$\log(1+k) \sim N\left(\log(1+\bar{k}) - \frac{1}{2}\delta^2,\, \delta^2\right).$$

Similarly to the Heston's model, the characteristic function of $X_t = \log S_t$ is known in closed form:

$$\varphi_t(u) = \varphi_t^D(u) \cdot \varphi_t^J(u),$$

where the characteristic function of the diffusion part:

$$\varphi_t^D(u) = \frac{\exp\left\{ \frac{\kappa\theta t(\kappa - \mathrm{i}\rho\sigma u)}{\sigma^2} + \mathrm{i}(r - \lambda\bar{k})tu + \mathrm{i}x_0 u \right\}}{\left(\cosh\frac{\gamma t}{2} + \frac{\kappa - \mathrm{i}\rho\sigma u}{\gamma}\sinh\frac{\gamma t}{2}\right)^{2\kappa\theta/\sigma^2}} \exp\left\{ -\frac{(u^2 + \mathrm{i}u)v_0}{\gamma\coth\frac{\gamma t}{2} + \kappa - \mathrm{i}\rho\sigma u} \right\},$$

and the characteristic function of the jump part:

$$\varphi_t^J(u) = \exp\left\{t\lambda\left(e^{-\frac{1}{2}\delta^2 u^2 + \mathrm{i}\left(\log(1+\bar{k}) - \frac{1}{2}\delta^2\right)u} - 1\right)\right\}.$$

3. Fourier transform methods for pricing European options

3.1 Black-Scholes type formulas

Gurland (1948) used the Convolution property P3 to derive an formula for calculation of the distribution function $F(x)$ from the characteristic function $\varphi(u)$:

$$F(x) = \frac{1}{2} + \frac{1}{2\pi}\int_0^\infty \frac{e^{\mathrm{i}ux}\varphi(-u) - e^{-\mathrm{i}ux}\varphi(u)}{\mathrm{i}u}du.$$

Shephard (1991) gave a simple proof of this formula. From this formula by a simple calculation we can get an important formula for Fourier inversion method:

$$\mathbb{P}\big(X > x\big) = 1 - F(x) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathrm{Re}\Big(\frac{e^{-\mathrm{i}ux}\varphi(u)}{\mathrm{i}u}\Big)du. \tag{8}$$

Here $\mathrm{Re}\{\cdot\}$ denotes taking the real part of argument. Inside the field of finance, this Fourier inversion method was first considered by Stein & Stein (1991) to find the distribution of the underlying asset price in a stochastic volatility model. Heston (1993) applied the formula (8), to obtain a Black-Scholes type formula for pricing a European call option in the stochastic volatility model.

In fact, with help of the risk-neutral valuation, the price of a European call with spot price S_0 and strike price K is given by

$$\begin{aligned} V(S_0, T; K) &= e^{-rT}\mathbb{E}\big[(S_T - K)^+\big] \ = \ e^{-rT}\int_{-\infty}^\infty (e^x - e^k)^+ f_T(x)dx \\ &= \int_k^\infty e^{-rT+x} f_T(x)dx - e^{-rT}K\int_k^\infty f_T(x)dx \ = \ I_1 - e^{-rT}KI_2, \end{aligned}$$

where r is the interest rate, T is the maturity, $k = \log K$, and $f_T(x)$ is the density function of $X_T = \log S_T$. It is clear that the second integral is the probability: $\Pi_2 = \mathbb{P}(X_t > k)$. By the Fourier inversion formula (8) we have

$$\Pi_2 = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathrm{Re}\Big(\frac{e^{-\mathrm{i}uk}\varphi_T(u)}{\mathrm{i}u}\Big)du,$$

where $\varphi_T(u)$ is the characteristic function of X_T. By a change of probability measure, we can verify that the second integral is give by $I_1 = S_0\Pi_1$, where

$$\Pi_1 = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathrm{Re}\Big(\frac{e^{-\mathrm{i}uk}\varphi_T(u-\mathrm{i})}{\mathrm{i}u\varphi_T(-\mathrm{i})}\Big)du,$$

Thus, the price of the European call now is given by a Black-Scholes type formula:

$$V(S_0, T; K) = S_0\Pi_1 - Ke^{-r(T-t)}\Pi_2. \tag{9}$$

Beginning with Heston's work, many authors used the Fourier inversion methods to solve advanced valuation problems [e.g. see Bates (1996), Schmelzle (2010), and references therein]. As mentioned by Carr & Madan (1999), the numerical integrals based on these methods are generally much faster than finite difference solutions to PDEs or PIDEs. Unfortunately, the FFT cannot be used to evaluate these integrals, since the integrands are singular at the required evaluation point $u = 0$.

Recently, several authors applied the Property P5 (Parseval relation) to derive the Fourier inversion formulas in option pricing [e.g. see Dufresne *et al* (2009)]. We have well known that, under the risk-neutral probability, the option price $V(S_0, T)$ with the terminal payoff $H(S_T)$ is given by

$$V(S_0, T) = e^{-rT}\mathbb{E}\big[H(S_T)\big] = e^{-rT}\int_{-\infty}^{\infty} H(x)p_T(x)dx = e^{-rT}\langle H(x), p_T(x)\rangle,$$

where $p_T(x)$ is the density function of S_T. By the Parseval relation we have

$$V(S_0, T) = \frac{e^{-rT}}{2\pi}\langle \mathcal{F}H(u), \mathcal{F}p_T(u)\rangle.$$

3.2 Fast Fourier transform methods - single asset

1. The method of Carr and Madan. Carr & Madan (1999) developed a different method designed to use the fast Fourier transform to price options. They introduced a new technique to calculate the Fourier transform of a modified call option price with respect to the logarithmic strike price so that the fast Fourier transform can be applied to calculate the integrals.

Consider a European call with the maturity T and the strike price K, which is written on a stock whose price process is $S_t = e^{rt+X_t}$, under a risk-neutral probability. Let $f_T(x)$ be the density of X_T. Consider the price of a call option:

$$V_T(k) = e^{-rT}\mathbb{E}\big[\big(e^{X_T} - e^k\big)^+\big] = e^{-rT}\int_k^{\infty}(e^x - e^k)f_T(x)dx, \tag{10}$$

where $k = \ln K$ is the log strike price. Note that $V_T(k) \to S_0 = 1$ as $k \to -\infty$, and the function $V_T(k)$ is not square-integrable. Thus, we cannot express the Fourier transform in strike in terms of the characteristic function $\varphi_T(u)$ of X_T and then find a range of strikes by Fourier inversion.

To obtain a square-integrable function we consider the modified call price:

$$\tilde{V}_T(k : \alpha) = e^{\alpha k}V_T(k)$$

for some $\alpha > 0$, which is chosen to improve the integrability. Carr & Madan (1999) showed that a sufficient condition for square-integrability of $\tilde{V}(k)$ is given by

$$\int_{-\infty}^{\infty} f_T(x)e^{(1+\alpha)x}dx < \infty.$$

Consider the Fourier transform of $\tilde{V}(k : \alpha)$:

$$\psi_T(v : \alpha) = \mathcal{F}\tilde{V}_T(k : \alpha)(v) = \int_{-\infty}^{\infty} e^{ivk}\tilde{V}_T(k : \alpha)dk.$$

By (10) and the definition of characteristic functions we have

$$\psi_T(v:\alpha) = e^{-rT}\int_{-\infty}^{\infty} f_T(x)dx \int_{-\infty}^{x} \left(e^{x+\alpha k} - e^{(1+\alpha)k}\right)e^{\mathrm{i}vk}dk$$
$$= \frac{e^{-rT}\varphi_T\big(v-\mathrm{i}(\alpha+1)\big)}{\alpha^2+\alpha-v^2+\mathrm{i}(2\alpha+1)v}, \tag{11}$$

where $\varphi_T(u)$ is the characteristic function of X_T. Using the Fourier inverse transform we get

$$V_T(k) = e^{-\alpha k}\tilde{V}(k:\alpha) = \frac{e^{-\alpha k}}{2\pi}\int_{-\infty}^{\infty} e^{-\mathrm{i}vk}\psi_T(v:\alpha)dv.$$

Note that $\psi_T(v:\alpha)$ is odd in its imaginary part and even in its real part. Since $V(k)$ is real we have

$$V_T(k) = \frac{e^{-\alpha k}}{\pi}\int_0^{\infty} e^{-\mathrm{i}vk}\psi_T(v:\alpha)dv. \tag{12}$$

Now this integral can be evaluated by the numerical approximation using the trapezoidal rule:

$$V_T(k_p) \approx \frac{e^{-\alpha k_p}}{2\pi}\sum_{j=0}^{N-1} e^{-\mathrm{i}v_j k_p}\psi_T(v_j:\alpha)\Delta v, \quad p=0,1,\ldots,N-1.$$

to apply the FFT algorithm, we set N as a power of 2, and define the grid points:

$$\begin{cases} k_p = -\frac{1}{2}N\Delta k + p\Delta k, & p=0,1,\ldots,N-1, \\ v_j = -\frac{1}{2}N\Delta u + j\Delta v, & j=0,1,\ldots,N-1, \end{cases}$$

with the step sizes Δk and Δv, which satisfy the Nyquist relation: $\Delta k\Delta v = 2\pi/N$. Then, the numerical approximation can expressed as

$$V_T(k_p) \approx \frac{e^{-\alpha k_p}}{2\pi}\sum_{j=0}^{N-1}(-1)^{j+p}e^{-\mathrm{i}\frac{2\pi}{N}jp}\psi_T(v_j:\alpha)\Delta v, \quad p=0,1,\ldots,N-1,$$

which can be efficiently computed by using FFT algorithm.

Many authors have discussed this Carr and Madan's pricing method in rigorous detail. Lee (2004) extended this method significantly and proved an error analysis for this FFT method. On the other hand, numerical experiments show that this FFT method has a quite large error when the strike price K is small, and its stability is very dependent on the choice of the damping exponential factor α. Ding & U (2010) presented a modified FFT-based method to overcome these disadvantages.

2. The method of Lewis. Lewis (2001) introduced an option pricing method which generalizes previous work on Fourier transform methods. The main idea of his method is to express the option price via the convolution of generalized Fourier transforms, and then, to apply the property P5 (Parseval relation) to obtain the generalized Fourier transform of option price.

Consider a European type option whose payoff is $H(S_T)$ at maturity T, where $S_t = S_0 e^{rt+X_t}$ is the stock price process under the risk neutral probability. To proceed, we assume that X_t is a Lévy process having the analytic characteristic function $\varphi_T(z)$, which is regular in the strip

$\mathbb{S}_X = \{z : a < \text{Im}\{z\} < b\}$, and $h(x) = H(e^{x+rT})$ is Fourier integrable in some strip $\mathbb{S}_h$ such that

$$\mathbb{S}_V = \bar{\mathbb{S}}_X \cap \mathbb{S}_h \neq \emptyset,$$

where $\bar{\mathbb{S}}_X$ is the complex conjugate set of $\mathbb{S}_X$.

Let $s = \log S_0$ denote the logarithm of current stock value. Then, the option price is given by

$$V(s) = e^{-rT}\mathbb{E}\Big[H\big(e^{s+rT+X_T}\big)\Big] = e^{-rT}\int_{-\infty}^{\infty} H(e^{s+rT+x})f_T(x)dx,$$

where $f_T(x)$ is density function of X_T. Under the assumption we can compute the generalized Fourier transform of $V(s)$ by the Parseval relation:

$$\begin{aligned}\int_{-\infty}^{\infty} e^{\mathrm{i}zs}V(s)ds &= e^{-rT}\int_{-\infty}^{\infty} f_T(x)dx\int_{-\infty}^{\infty} e^{\mathrm{i}zs}H(e^{s+rT+x})ds\\ &= e^{-rT}\int_{-\infty}^{\infty} e^{-\mathrm{i}zx}f_T(x)dx\int_{-\infty}^{\infty} e^{\mathrm{i}zy}H(e^{y+rT})dy,\end{aligned}$$

for all $z \in \mathbb{S}$. Finally we obtain

$$\mathcal{F}V(z) = e^{-rT}\varphi_T(-z)\mathcal{F}h(z), \quad z \in \mathbb{S}_V.$$

Option prices can now be given by the generalized Fourier inversion formula (3):

$$V(s) = \frac{e^{-rT}}{2\pi}\int_{\mathrm{i}w-\infty}^{\mathrm{i}w+\infty} \varphi_T(-z)\mathcal{F}h(z)dz,$$

with $a < w < b$, for a range of initial values s.

For a European call option, the function $h(x) = (e^{rT+x} - e^k)^+$, is Fourier integrable in the region $\text{Im}\{z\} > 1$, where its generalized Fourier transform can be computed explicitly:

$$\mathcal{F}h(z) = \int_{-\infty}^{\infty} e^{\mathrm{i}xz}\big(e^{x+rT} - e^k\big)^+dx = \frac{e^{k+\mathrm{i}z(k-rT)}}{\mathrm{i}z(\mathrm{i}z+1)}.$$

Hence, the generalized Fourier transform of call option price takes the form:

$$\mathcal{F}V(z) = \varphi_T(-z)\frac{e^{(1+\mathrm{i}z)(k-rT)}}{\mathrm{i}z(\mathrm{i}z+1)}, \quad 1 < \text{Im}\{z\} < 1+\alpha.$$

and hence, the option price is given by

$$V(s) = \frac{e^{ws+(1-w)(k-rT)}}{2\pi}\int_{-\infty}^{\infty}\frac{\varphi_T(-z)e^{\mathrm{i}u(k-rT-s)}}{(\mathrm{i}u-w)(1+\mathrm{i}u-w)}du,$$

for some $1 < w < 1+\alpha$. The integral in this formula can be approximated by using the FFT algorithm as the formula (12). However, the choice of w is a delicate issue because choosing big w leads to slower decay rates at infinity and bigger truncation errors, and while w is close to one the denominator diverges and the discretization error becomes to large (see Chapter 11 in Cont & Tankov (2003)). On the other hand, Lewis (2001) also listed the generalized Fourier transforms $\mathcal{F}h(z)$ and the strip $\mathbb{S}_h$ for various claims, for instance, the put option, the covered call or the cash-secured put, etc.

3.3 Fast Fourier transform methods - multi assets

1. The most direct extension of the Carr and Madan method to the multi assets model is to price the correlation option, whose payoff at the maturity T is defined by

$$H(S_{1T}, S_{2T}) = (S_{1T} - K_1)^+ (S_{2T} - K_2)^+,$$

where K_1 and K_2 are strike prices [see, e.g. Kwok *et al* (2010)]. Define $X_{it} = \log S_{it}$ and $k_i = \log K_i, i = 1, 2,$ and let $f_T(x_1, x_2)$ be the joint density of X_{1T} and X_{2T} under the risk neutral probability. Then, the characteristic function of X_{1T} and X_{2T} is defined by the following two-dimensional Fourier transform:

$$\varphi_T(u_1, u_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\mathrm{i}(u_1x_1+u_2x_2)} f_T(x_1, x_2)dx_1dx_2.$$

Consider the price of the correlation option:

$$V_T(k_1, k_2) = e^{-rT}\int_{k_2}^{\infty}\int_{k_1}^{\infty}(e^{x_1} - e^{k_1})(e^{x_2} - e^{k_2})f_T(x_1, x_2)dx_1dx_2.$$

Following the Carr and Madam method we consider the modified call price:

$$\tilde{V}_T(k_1, k_2 : \alpha_1, \alpha_2) = e^{\alpha_1k_2+\alpha_2k_2}V_T(k_1, k_2),$$

for some parameters $\alpha_1 > 0$ and $\alpha_2 > 0$, which are chosen such that this modified price is square integrable for negative value of k_1 and k_2. Then, by a direct integral the Fourier transform of $\tilde{V}(k_1, k_2)$ is given by

$$\begin{aligned}\psi_T(v_1, v_2) &= \mathcal{F}\tilde{V}_T(k_1, k_2 : \alpha_1, \alpha_2)\\ &= \frac{e^{-rT}\varphi_T(v_1 - \mathrm{i}(\alpha_1+1), v_2 - \mathrm{i}(\alpha_2+1))}{(\alpha_1+\mathrm{i}v_1)(\alpha_1+1+\mathrm{i}v_1)(\alpha_2+\mathrm{i}v_2)(\alpha_2+1+\mathrm{i}v_2)}.\end{aligned}$$

Applying the Fourier inversion on $\psi_T(v_1, v_2)$ and using the numerical approximation of two-dimensional Fourier inversion integral we have

$$\begin{aligned}V_T(k_p^1, k_q^2) &= \frac{e^{-\alpha_1k_p^1-\alpha_2k_q^2}}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\mathrm{i}(v_1k_p^1+v_2k_q^2)}\psi_T(v_1, v_2)dv_1dv_2\\ &\approx \frac{e^{-\alpha_1k_p^1-\alpha_2k_q^2}}{(2\pi)^2}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} e^{-\mathrm{i}(v_m^1k_p^1+v_n^2k_q^2)}\psi_T(v_m^1, v_n^2)\Delta v^1\Delta v^2,\end{aligned}$$

for $p, q = 0, 1, \ldots, N-1$. To apply the two-dimensional of the FFT algorithm we set

$$v_m^1 = (m - \frac{N}{2})\Delta v^1, \quad v_n^2 = (n - \frac{N}{2})\Delta v^2, \quad m, n = 0, 1, \ldots, N-1,$$

and

$$k_p^1 = (p - \frac{N}{2})\Delta k^1, \quad k_q^2 = (q - \frac{N}{2})\Delta k^2, \quad p, q = 0, 1, \ldots, N-1,$$

where N is a power of 2, and the step sizes Δv^1, Δv^2, Δk^1, and Δk^2 observe the Nyquist relations: $\Delta v^1\Delta k^1 = \Delta v^2\Delta k^2 = 2\pi/N$. Dempster & Hong (2002) shown that the numerical approximation now is given by

$$V_T(k_p^1, k_q^2) \approx \frac{e^{-\alpha_1k_p^1-\alpha_2k_q^2}}{(2\pi)^2}\Psi_T(k_p^1, k_q^2),$$

for $k_p^1 \neq k_q^2$, where

$$\Psi_T(k_p^1, k_q^2) = (-1)^{p+q} \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} e^{-\mathrm{i}\frac{2\pi}{N}(mp+nq)} \big((-1)^{m+n} \psi_T(v_m^1, v_n^2) \Delta v^1 \Delta v^2\big),$$

which can be computed via the FFT algorithm.

There are some other types of terminal payoff functions that admit analytic representation of the Fourier transform of the damped option price [see e.g. Eberlein *et al* (2010), Lee (2004), and references therein]. However, to derive the FFT pricing algorithm for the spread option:

$$H(S_{1T}, S_{2T}) = (S_{1T} - S_{2T} - K)^+, \tag{13}$$

Dempster & Hong (2002) approximated the exercise region of $H(S_{1T}, S_{2T})$ by a combination of rectangular strips.

2. Hurd & Zhou (2010) proposed an alternative approach to pricing the European spread option (13) under the three-factor SV model and exponential Lévy models. Let $h(x_1, x_2)$ be the terminal spread option payoff with the strike price $K = 1$, i.e.

$$h(x_1, x_2) = (e^{x_1} - e^{x_2} - 1)^+. \tag{14}$$

And denote $\Gamma(z)$ the complex gamma function defined by

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt, \quad \mathrm{Re}\{z\} > 0.$$

Then, the method of Hurd and Zhou mainly relies on the following Fourier representation of the payoff function $h(x_1, x_2)$:

$$h(x_1, x_2) = \frac{1}{(2\pi)^2} \int_{\mathrm{i}\epsilon_2-\infty}^{\mathrm{i}\epsilon_2+\infty} \int_{\mathrm{i}\epsilon_1-\infty}^{\mathrm{i}\epsilon_1+\infty} e^{\mathrm{i}(u_1x_1+u_2x_2)} \hat{h}(u_1, u_2) du_1 du_2, \tag{15}$$

where ϵ_1 and ϵ_2 are any two given real numbers with $\epsilon_2 > 0$ and $\epsilon_1 + \epsilon_2 < -1$, and

$$\hat{h}(u_1, u_2) = \frac{\Gamma\big(\mathrm{i}(u_1 + u_2) - 1\big)\Gamma(-\mathrm{i}u_2)}{\Gamma(\mathrm{i}u_1 + 1)}.$$

The proof of this formula can be found in the appendix of Hurd & Zhou (2010). Let $X_{1t} = \log S_{1t}$ and $X_{2t} = \log S_{2t}$ with $X_{10} = x_1$ and $X_{20} = x_2$, and let $\varphi_T(u_1, u_2)$ be the characteristic function of $X_{1T} - x_1$ and $X_{2T} - x_2$. Then, we have

$$\mathbb{E}\big[e^{\mathrm{i}(u_1X_{1T}+u_2X_{2T})}\big] = e^{\mathrm{i}(u_1x_1+u_2x_2)} \varphi_T(u_1, u_2).$$

Now, using the formula (16), the price of the spread option (14) is expressed as an explicit two-dimensional Fourier inversion transform:

$$\begin{aligned}
V_T(x_1, x_2) &= e^{-rT}\mathbb{E}\big[\big(e^{X_{1T}} - e^{X_{2T}} - 1\big)^+\big] \\
&= e^{-rT}\mathbb{E}\Big[\frac{1}{(2\pi)^2} \int_{\mathrm{i}\epsilon_2-\infty}^{\mathrm{i}\epsilon_2+\infty} \int_{\mathrm{i}\epsilon_1-\infty}^{\mathrm{i}\epsilon_1+\infty} e^{\mathrm{i}(u_1X_{1T}+u_2X_{2T})} \hat{h}(u_1, u_2) du_1 du_2\Big] \\
&= \frac{e^{-rT}}{(2\pi)^2} \int_{\mathrm{i}\epsilon_2-\infty}^{\mathrm{i}\epsilon_2+\infty} \int_{\mathrm{i}\epsilon_1-\infty}^{\mathrm{i}\epsilon_1+\infty} \mathbb{E}\big[e^{\mathrm{i}(u_1X_{1T}+u_2X_{2T})}\big] \hat{h}(u_1, u_2) du_1 du_2 \\
&= \frac{e^{-rT}}{(2\pi)^2} \int_{\mathrm{i}\epsilon_2-\infty}^{\mathrm{i}\epsilon_2+\infty} \int_{\mathrm{i}\epsilon_1-\infty}^{\mathrm{i}\epsilon_1+\infty} e^{\mathrm{i}(u_1x_1+u_2x_2)} \varphi_T(u_1, u_2) \hat{h}(u_1, u_2) du_1 du_2.
\end{aligned}$$

This two-dimensional Fourier inversion integral can be numerically computed by the usual FFT calculations. Since this approach does not require an approximation of the exercise region, it is considered to be more computationally efficient.

3.4 Fourier expansion methods

Fang & Oosterlee (2008) proposed a new numerical method for pricing European options, which is called the COS method. This method is based on the Fourier transform and the Fourier-cosine expansion, and is shown to have the exponential convergence rate and linear computational complexity. And now, this novel method has been considered by many authors to price various options [see e.g. Ding *et al* (2011b),Fang & Oosterlee (2009), and the references therein].

We recall that the price S_t of underlying asset follows an exponential Lévy model. Let $X_t = \log(S_t/K)$, where K is the strike price. Consider the price of European call in the form:

$$V_0(K) = e^{-rT}\mathbb{E}\big[(S_T - K)^+\big] = e^{-rT}K\int_0^\infty (e^x - 1)f_T(x)dx, \tag{16}$$

where $f_T(x)$ is the probability density of X_T under the risk-neutral probability. Let $\varphi_T(u)$ be the characteristic function of $f_T(x)$. The main ideas of the COS method are to choose two numbers a and b such that the truncated integral approximates the infinite integral very well, i.e.,

$$\tilde{\varphi}_T(u) = \int_a^b e^{\mathrm{i}ux}f_T(x)dx \approx \int_{-\infty}^\infty e^{\mathrm{i}ux}f_T(x)dx = \varphi_T(u), \tag{17}$$

and to consider the Fourier-cosine expansion of $f_T(x)$ in $[a,b]$:

$$f_T(x) = \frac{1}{2}A_0 + \sum_{n=1}^\infty A_n\cos\Big(n\pi\frac{x-a}{b-a}\Big), \quad x \in [a,\ b], \tag{18}$$

where

$$A_n = \frac{2}{b-a}\int_a^b \tilde{p}_T(x)\cos\Big(n\pi\frac{x-a}{b-a}\Big)\mathrm{d}x, \quad n = 0,1,2,\ldots.$$

From (17) we have

$$A_n = \frac{2}{b-a}\mathrm{Re}\Big\{\tilde{\varphi}_T\Big(\frac{n\pi}{b-a}\Big)\exp\Big(-\mathrm{i}\frac{na\pi}{b-a}\Big)\Big\}, \quad n = 0,1,2,\ldots.$$

Then, we get an approximation of $f_T(x)$ in (18) by

$$f_T(x) \approx \frac{1}{2}\tilde{A}_0 + \sum_{n=1}^{N-1}\tilde{A}_n\cos\Big(n\pi\frac{x-a}{b-a}\Big), \quad x \in [a,\ b], \tag{19}$$

where

$$\tilde{A}_n = \frac{2}{b-a}\mathrm{Re}\Big\{\varphi_T\Big(\frac{n\pi}{b-a}\Big)\exp\Big(-\mathrm{i}\frac{na\pi}{b-a}\Big)\Big\} \approx A_n, \quad n = 0,1,\ldots,N{-}1$$

Now, substituting (19) into (16), we obtain an approximation of the option price:

$$V_0(K) \approx Ke^{-rT}\Big\{\frac{1}{2}\tilde{A}_0\big(\Phi_0(0,b) - \Psi_0(0,b)\big) + \sum_{n=1}^{N-1}\tilde{A}_n\big(\Phi_n(0,b) - \Psi_n(0,b)\big)\Big\}, \tag{20}$$

where $\Phi_n(0,b)$ and $\Psi_n(0,b)$ are two integrals given by:

$$\Phi_n(c,d) = \int_c^d e^x \cos\Big(n\pi\frac{x-a}{b-a}\Big)dx \quad \text{and} \quad \Psi_n(c,d) = \int_c^d \cos\Big(n\pi\frac{x-a}{b-a}\Big)dx,$$

for any $[c,d] \subset [a,b]$, which are analytically given by

$$\begin{aligned}
\Phi_n(c,d) &= \frac{1}{1+(\frac{n\pi}{b-a})^2}\Big[\cos\Big(n\pi\frac{d-a}{b-a}\Big)e^d - \cos\Big(n\pi\frac{c-a}{b-a}\Big)e^c \\
&\quad + \frac{n\pi}{b-a}\sin\Big(n\pi\frac{d-a}{b-a}\Big)e^d - \frac{n\pi}{b-a}\sin\Big(n\pi\frac{c-a}{b-a}\Big)e^c\Big], \\
\Psi_n(c,d) &= \begin{cases} \Big[\sin\Big(n\pi\frac{d-a}{b-a}\Big) - \sin\Big(n\pi\frac{c-a}{b-a}\Big)\Big]\frac{b-a}{n\pi} & n \neq 0, \\ (d-c) & n = 0 \end{cases}.
\end{aligned}$$

Fang & Oosterlee (2008) also showed that, in most cases, the convergence rate of the COS formula (20) is exponential and the computational complexity is linear. They also discussed the truncation range for COS method, and gave a general formula to determine the interval of integration $[a,b]$ in that paper, which is given by

$$[a,b] = \Big[c_1 - \delta\sqrt{c_2 + \sqrt{c_4}},\ c_1 + \delta\sqrt{c_2 + \sqrt{c_4}}\ \Big], \tag{21}$$

where c_1, c_2, and c_4 are the first, second, and fourth cumulates of $X_t = \log(S_t/K)$. Also, the constant δ depends on the tolerance level in the approximation (17), and usually we choose $\delta = 10$. Meanwhile, Ding & U (2011) respectively applied the Fourier-sine and Fourier expansions to substitute the Fourier-cosine expansion in (18). A comprehensive analysis with numerical comparisons for these different methods is also given in their paper.

4. Fourier transform methods for pricing path dependent options

4.1 The CONV method for pricing Bermudan barrier options

Pricing Bermudan or barrier options is much harder than pricing European options. Because these options are depended on paths of the price process for the underlying assets. Recently, some new numerical integration methods based on Fourier transforms are proposed. Lord *et al* (2008) proposed an efficient and accurate FFT-based method, called the CONV method, to price Bermudan options under exponential Lévy models.

In the following we apply the CONV method to price a Bermudan barrier option in which the monitored dates may be many times more than the exercise dates. Denote

$$G(S) = \begin{cases} (S-K)^+, \text{ for call option,} \\ (K-S)^+, \text{ for put option,} \end{cases}$$

where K is the strike price, S is the spot price of the underlying asset. Let

$$\begin{cases} \mathbb{T} \ = \{t_{mL+l} : m = 0,1,\ldots,M,\ l = 0,1,\ldots,L{-}1\}, \\ \mathbb{T}_e = \{t_{mL} : m = 1,\ldots,M{-}1\} \subset \mathbb{T}, \end{cases} \tag{22}$$

be the set of pre-specified monitored dates and the set of pre-specified exercise dates, respectively, before maturity T, where

$$0 = t_0 < t_1 < \cdots < t_{ML} = T \quad \text{with} \quad \Delta t = t_k - t_{k-1} = T/(ML).$$

Consider a discrete American barrier option, which is monitored at every $t_k \in \mathbb{T}$ and can be exercised at each $t_k \in \mathbb{T}_e$, namely Bermudan barrier option, whose payoff is given by

$$G(S_{t_k})1_{\{S_{t_k}<H\}} + R_0 1_{\{S_{t_k}\geq H\}}, \quad t_k \in \mathbb{T}.$$

Here S_{t_k} is the price of the underlying asset at time $t_k \in \mathbb{T}$, $H > K$ is the constant barrier and R_0 is the contractual rebate. That is, this Bermudan barrier option is an up-and-out barrier option that cease to exist if the asset price S_{t_k} hits the barrier level H at one time $t_k \in \mathbb{T}$, and it can also be exercised at any time $t_k \in \mathbb{T}_e$ before maturity T.

Denote $V(S, t_k)$ the value of this Bermudan barrier option at time t_k and the spot price $S_{t_k} = S$. With help of the risk-neutral valuation formula, this price process can be computed recursively by the following backward induction:

$$\begin{cases} V(S,t_{ML}) = G(S)1_{\{S<H\}} + R_0 1_{\{S\geq H\}}, \\ C(S,t_k) \quad = \mathbb{E}\big[e^{-r\Delta t} V(S_{t_{k+1}}, t_{k+1}) \mid S_{t_k} = S\big], \quad t_k \in \mathbb{T}, \\ V(S,t_k) \quad = C(S,t_k)1_{\{S<H\}} + e^{-r(T-t_k)} R_0 1_{\{S\geq H\}}, \quad t_k \in \mathbb{T} \setminus \mathbb{T}_e, \\ V(S,t_k) \quad = \max\big\{G(S), C(S,t_k)\big\}1_{\{S<H\}} + e^{-r(T-t_k)} R_0 1_{\{S\geq H\}}, \quad t_k \in \mathbb{T}_e, \end{cases} \tag{23}$$

in specialty, the initial price is given

$$V(S,t_0) = C(S,t_0) = \mathbb{E}\big[e^{-r\Delta t} V(S,t_1) \mid S_{t_0} = S\big].$$

Here $r > 0$ is the interest rate, and $\mathbb{E}[\cdot \mid \cdot]$ is the conditional expectation under the risk-neutral probability.

Assume that the price process of the underlying asset is given by

$$S_t = S_0 e^{X_t}, \quad t \geq 0,$$

where X_t is a Lévy process and S_0 is the initial price. Let $f(\cdot \mid x)$ be the condition density of $X_{t_{k+1}}$ given $X_{t_k} = x$ for $t_k \in \mathbb{T}$. Set

$$g(x) = \begin{cases} (S_0 e^x - K)^+, & \text{for a call option,} \\ (K - S_0 e^x)^+, & \text{for a put option.} \end{cases}$$

Then the backward induction (23) can be rewritten by

$$\begin{cases} v(x,t_{ML}) = g(x)1_{\{x<h\}} + R_0 1_{\{x\geq h\}}, \\ c(x,t_k) \quad = e^{-r\Delta t}\int_{-\infty}^{\infty} v(y,t_{k+1}) f(y \mid x)\mathrm{d}y, \quad t_k \in \mathbb{T}, \\ v(x,t_k) \quad = c(x,t_k)1_{\{x<h\}} + R_0 1_{\{x\geq h\}}, \quad t_k \in \mathbb{T} \setminus \mathbb{T}_e, \\ v(x,t_k) \quad = \max\big\{g(x), c(x,t_k)\big\}1_{\{x<h\}} + e^{-r(T-t_k)} R_0 1_{\{x\geq h\}}, \quad t_k \in \mathbb{T}_e, \end{cases} \tag{24}$$

and

$$v(x,t_0) = c(x,t_0) = e^{-r\Delta t}\int_{-\infty}^{\infty} v(y,t_1) f(y \mid x)\mathrm{d}y,$$

where $v(x,t_k) = V(S_0e^x, t_k)$ for any $t_k \in \mathbb{T}$, and $h = \log(H/S_0)$.

Since each Lévy process is stationary and has independent increments, the condition density $f(\cdot \mid x)$ possesses the property:

$$f(y|x) = f(y-x), \quad x, y \in \mathbb{R},$$

where $f(y)$ is the density of X_{t_1} under the initial condition $X_{t_0} = x$. Applying this property to infinite integrals $c(x,t_k)$ in (24) it becomes to

$$c(x,t_k) = e^{-r\Delta t}\int_{-\infty}^{\infty} v(y,t_{k+1})f(y-x)dy = e^{-r\Delta t}\int_{-\infty}^{\infty} v(x+z,t_{k+1})f(z)dz$$

for any $t_k \in \mathbb{T}$. Then, this integral can be rewritten as a convolution of $v(x+z,t_{k+1})$ and the function $f(-x)$, i.e.

$$c(x,t_k) = e^{-r\Delta t}f(-x) * v(x,t_{k+1}).$$

Thus we have

$$e^{\alpha x}c(x,t_k) = e^{-r\Delta t}\big[e^{\alpha x}f(-x)\big] * \big[e^{\alpha x}v(x,t_{k+1})\big], \tag{25}$$

where $\alpha > 0$ is chosen to improve the integrability. Now, the main idea of CONV method is that, taking the Fourier transform in the both sides of (25) and applying the Property P3 (Convolution), the integral becomes

$$\begin{aligned} e^{r\Delta t}\mathcal{F}\big[e^{\alpha x}c(x,t_k)\big](u) &= \mathcal{F}\big\{\big[e^{\alpha x}f(-x)\big] * \big[e^{\alpha x}v(x,t_{k+1})\big]\big\}(u) \\ &= \mathcal{F}\big[e^{\alpha x}f(-x)\big](u)\cdot\mathcal{F}\big[e^{\alpha x}v(x,t_{k+1})\big](u). \end{aligned}$$

Denote $\varphi(z)$ is the characteristic function of the density $f(x)$ on the complex plan $\mathbb{C}$. Then, by a simple calculation, we have

$$e^{r\Delta t}\mathcal{F}\big[e^{\alpha x}c(x,t_k)\big](u) = \varphi\big(-(u-\mathrm{i}\alpha)\big)\cdot\mathcal{F}\big[e^{\alpha x}v(x,t_{k+1})\big](u).$$

Thus, taking the inverse Fourier transform we obtain

$$e^{\alpha x}c(x,t_k) = e^{-r\Delta t}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\mathrm{i}ux}\varphi\big(-(u-\mathrm{i}\alpha)\big)\int_{-\infty}^{\infty} e^{\mathrm{i}uy+\alpha y}v(y,t_{k+1})dydu. \tag{26}$$

Denote

$$\kappa = \delta\sqrt{-\frac{\partial^2\varphi_T(u)}{\partial u^2}\Big|_{u=0} + \Big(\frac{\partial\varphi_T(u)}{\partial u}\Big|_{u=0}\Big)^2},$$

where $\varphi_T(u)$ is the characteristic function of X_T, and δ is a proportionality constant. According to the suggestion from Lord *et al* (2008), we can take $\delta = 20$ for the GBM model and $\delta = 40$ for other exponential Lévy models. Let N be a power of 2. We consider the grid points on x-axes:

$$x_j = (j - \frac{1}{2}N)\Delta x, \quad j = 0, 1, \ldots, N-1,$$

where $\Delta x = \kappa/N$. Furthermore, we also consider the grid points for the numerical integrals in (26):

$$u_j = (j - \frac{1}{2}N)\Delta u, \quad y_j = x_j, \quad j = 0, 1, \ldots, N-1,$$

where $\Delta u = 2\pi/\kappa$. It is clear that these grids satisfy the Nyquist relation: $\Delta u \Delta y = 2\pi/N$. Now, for each $t_k \in \mathbb{T}$ and each $p = 0, 1, \ldots, N-1$, approximating the integral in (26) with composite trapezoidal rule and the second integral with left rectangle rule yields

$$c(x_p, t_k) \approx e^{-\alpha x_p - r\Delta t} \frac{\Delta u \Delta y}{2\pi} \sum_{j=0}^{N-1} \Big(e^{-\mathrm{i} u_j x_p} \varphi\big(-(u_j - \mathrm{i}\alpha)\big) \sum_{n=0}^{N-1} \omega_n e^{\mathrm{i} u_j y_n + \alpha y_n} v(y_n, t_{k+1}) \Big)$$

where the weights ω_n are chosen as $\omega_0 = \omega_{N-1} = \frac{1}{2}$ and $\omega_n = 1$ for $n = 1, \ldots, N-2$. Noting that $u_0 = -\frac{1}{2} N \Delta u$ and $\Delta x = \Delta y = 2\pi/(N\Delta u)$, we have

$$\begin{aligned} c(x_p, t_k) \approx\ & e^{-\alpha x_p - r\Delta t} e^{\mathrm{i} u_0 (y_0 - x_0)} (-1)^p \sum_{j=0}^{N-1} \Big(e^{-\mathrm{i} j p \frac{2\pi}{N}} e^{\mathrm{i} j (y_0 - x_0) \Delta u} \varphi\big(-(u_j - \mathrm{i}\alpha)\big) \\ & \cdot \frac{1}{N} \sum_{n=0}^{N-1} e^{\mathrm{i} j n \frac{2\pi}{N}} (-1)^n \omega_n e^{\alpha y_n} v(y_n, t_{k+1}) \Big). \end{aligned} \tag{27}$$

Now, we can employ the FFT algorithm to calculate the summations in the right side of (27). Once the integral $c(x_p, t_k)$ is computed, we can determine the early-exercise price $S^*_{t_k}$, $t_k \in \mathbb{T}_e$, by the procedure: for each $t_k \in \mathbb{T}_e$, locate j_k such that

$$c(x_{j_k}, t_k) - g(x_{j_k}) = 0, \tag{28}$$

or,

$$\big(c(x_{j_k}, t_k) - g(x_{j_k})\big)\big(c(x_{j_k+1}, t_k) - g(x_{j_k+1})\big) \leq 0. \tag{29}$$

In the case (28) set $x^*(t_k) = x_{j_k}$, and in the case (29) set $x^*(t_k) = \frac{1}{2}(x_{j_k} + x_{j_{k+1}})$. Then the early-exercise price at every $t_k \in \mathbb{T}_e$ is given by $S^*_{t_k} = S_0 e^{x^*(t_k)}$. Ding *et al* (2011b) gave a detail algorithm, which summarizes the above procedure, for pricing an up-and-out Bermudan barrier option, and the corresponding numerical experiments.

4.2 The COS method for pricing Bermudan barrier options

Recently, Fang & Oosterlee (2009) extended their COS method to price discrete early-exercise options under exponential Lévy models, and Fang & Oosterlee (2011) further considered such pricing problems under Heston's model.

Assume that the price process of the underlying asset S_t follows an exponential Lévy model, and $\mathbb{T}$ and $\mathbb{T}_e$ are the set of pre-specified monitored dates and the set of pre-specified exercise dates, respectively, before the maturity T, which are defined by (22). In the following, we apply the COS method to price the Bermudan barrier option which defined in preceding subsection, whose payoff is given by

$$G(S_{t_k}) 1_{\{S_{t_k} < H\}} + R_0 1_{\{S_{t_k} \geq H\}}, \quad t_k \in \mathbb{T},$$

where $H > K$ is the constant barrier and R_0 is the contractual rebate. Denote $V(S, t_k)$, $t_k \in \mathbb{T}_e$, the value of this Bermudan barrier option at time t_k and the spot price $S_{t_k} = S$. As in the preceding subsection, with help of the risk-neutral valuation formula, this price process can be computed recursively by the backward induction (23). In specialty, the initial price is given by

$$V(S, t_0) = C(S, t_0) = \mathbb{E}\big[e^{-r\Delta t} V(S, t_1) \mid S_{t_0} = S\big]. \tag{30}$$

Here S is the spot price of underlying asset, $r > 0$ is the interest rate. Let $X_t = \log(S_t/K)$ be the logarithm of the underlying asset price S_t over the strike price K, and denote $x = \log(S/K)$ and $h = \log(H/K)$. Let $f(\cdot \mid x)$ be the condition density of $X_{t_{k+1}}$ given $X_{t_k} = x$ for $t_k \in \mathbb{T}$. Set

$$g(x) = \begin{cases} K(e^x - 1)^+, & \text{for a call option,} \\ K(1 - e^x)^+, & \text{for a put option.} \end{cases}$$

Then the backward induction (23) and the price formula (30) can be rewritten by

$$\begin{cases} v(x, t_{ML}) = g(x)1_{\{x<h\}} + R_0 1_{\{x \geq h\}}, \\ c(x, t_k) \quad = e^{-r\Delta t} \int_{-\infty}^{\infty} v(y, t_{k+1}) f(y \mid x) dy, \quad t_k \in \mathbb{T}, \\ v(x, t_k) \quad = c(x, t_k))1_{\{x<h\}} + R_0 1_{\{x \geq h\}}, \quad t_k \in \mathbb{T} \setminus \mathbb{T}_e, \\ v(x, t_k) \quad = \max\{g(x), c(x, t_k)\}1_{\{x<h\}} + e^{-r(T-t_k)} R_0 1_{\{x \geq h\}}, \quad t_k \in \mathbb{T}_e, \end{cases} \tag{31}$$

and

$$v(x, t_0) = c(x, t_0) = e^{-r\Delta t} \int_{-\infty}^{\infty} v(y, t_1) f(y \mid x) dy, \tag{32}$$

where $v(x, t_k) = V(Ke^x, t_k)$ for any $t_k \in \mathbb{T}$.

We consider the infinite integrals $c(x, t_k)$ in (31). Since $f(y|x)$ decays to zero very quickly as $y \to \pm\infty$ we may choose two bounds a and b, which can be selected by (21), such that

$$\bar{c}(x, t_k) = e^{-r\Delta t} \int_a^b v(y, t_{m+1}) f(y \mid x) dy \approx c(x, t_k), \quad t_k \in \mathbb{T}. \tag{33}$$

without losing some significant accuracy. Note that the density $f(y \mid x)$ has the following Fourier-cosine expansion on $[a, b]$:

$$f(y \mid x) = \frac{2}{b-a} \sum_{j=0}^{\infty} \left[w_j \cos\left(j\pi \frac{y-a}{b-a}\right) \int_a^b f(u \mid x) \cos\left(j\pi \frac{u-a}{b-a}\right) du \right],$$

where $w_0 = \frac{1}{2}$ and $w_j = 1$ for all $j = 1, 2, 3, \ldots$. We substitute this expansion into the integral (33) and then we can rewrite it by

$$\bar{c}(x, t_k) = \frac{(b-a)}{2} e^{-r\Delta t} \sum_{j=0}^{\infty} w_j F_j(x) V_j(t_{k+1}), \tag{34}$$

where for each $j = 1, 2, 3, \ldots$,

$$V_j(t_{k+1}) = \frac{2}{b-a} \int_a^b v(y, t_{k+1}) \cos\left(j\pi \frac{y-a}{b-a}\right) dy, \tag{35}$$

and

$$F_j(x) = \frac{2}{b-a} \int_a^b f(y \mid x) \cos\left(j\pi \frac{y-a}{b-a}\right) dy. \tag{36}$$

Since the Fourier-cosine expansion has a high accuracy with a few terms, we can truncate the infinity series (34) and approximate $\bar{c}(x, t_k)$ by leaving the first N terms, i.e.,

$$\hat{c}(x, t_k) = \frac{(b-a)}{2} e^{-r\Delta t} \sum_{j=0}^{N-1} w_j F_j(x) V_j(t_{k+1}) \approx \bar{c}(x, t_k), \tag{37}$$

for any $t_k \in \mathbb{T}$, where $w_0 = w_{N-1} = \frac{1}{2}$ and $w_j = 1$ for $j = 1, \ldots, N-2$. On the other hand, we can represent each integral (36) by approximation as the following

$$F_j(x) = \frac{2}{b-a}\mathrm{Re}\Big\{ \exp\Big(-\mathrm{i}\frac{ja\pi}{b-a}\Big) \int_a^b f(u \mid x) \exp\Big(\mathrm{i}\frac{ju\pi}{b-a}\Big) du \Big\}$$
$$\approx \frac{2}{b-a}\mathrm{Re}\Big\{ \exp\Big(-\mathrm{i}\frac{ja\pi}{b-a}\Big) \int_{-\infty}^{\infty} f(u \mid x) \exp\Big(\mathrm{i}\frac{j\pi}{b-a}u\Big) du \Big\}.$$

Let $\varphi(u; x)$ be the characteristic function of $f(\cdot \mid x)$. Then, we can approximate each $F_j(x)$ by

$$F_j(x) \approx \frac{2}{b-a}\mathrm{Re}\Big\{ \exp\Big(-\mathrm{i}\frac{ja\pi}{b-a}\Big) \cdot \varphi\Big(\frac{j\pi}{b-a}; x\Big) \Big\}.$$

And so, we get the further numerical approximations of the integrals $\hat{c}(x, t_k)$ in (37) by

$$\tilde{c}(x, t_k) = e^{-r\Delta t} \sum_{j=0}^{N-1} w_j \mathrm{Re}\Big\{ \exp\Big(-\mathrm{i}\frac{ja\pi}{b-a}\Big) \cdot \varphi\Big(\frac{j\pi}{b-a}; x\Big) \Big\} V_j(t_{k+1}) \approx \hat{c}(x, t_j). \tag{38}$$

In special, the approximation of initial price $v(x, t_0)$ in (32) is given by

$$\tilde{v}(x, t_0) = e^{-r\Delta t} \sum_{j=0}^{N-1} w_j \mathrm{Re}\Big\{ \exp\Big(-\mathrm{i}\frac{ja\pi}{b-a}\Big) \cdot \varphi\Big(\frac{j\pi}{b-a}; x\Big) \Big\} V_j(t_1). \tag{39}$$

Meanwhile, since the characteristic function $\phi(u; x)$ possesses the property:

$$\varphi(u; x) = \varphi(u) \cdot e^{\mathrm{i}ux}, \quad u \in \mathbb{R},$$

where $\varphi(u) = \varphi(u; 0)$, the approximations (38) and (39) can be simplified to

$$\tilde{c}(x, t_k) = e^{-r\Delta t} \sum_{j=0}^{N-1} w_j \mathrm{Re}\Big\{ \exp\Big(\mathrm{i}j\pi\frac{x-a}{b-a}\Big) \cdot \varphi\Big(\frac{j\pi}{b-a}\Big) \Big\} V_j(t_{k+1}), \tag{40}$$

and the initial price of option

$$\tilde{v}(x, t_0) = e^{-r\Delta t} \sum_{j=0}^{N-1} w_j \mathrm{Re}\Big\{ \exp\Big(\mathrm{i}j\pi\frac{x-a}{b-a}\Big) \cdot \varphi\Big(\frac{j\pi}{b-a}\Big) \Big\} V_j(t_1). \tag{41}$$

In order to use this approximate formulation we still need to compute the integrals $V_j(t_k)$. For convenience we introduce the following notions: for any $a \le x_1 \le x_2 \le b$,

$$C_j(x_1, x_2; t_k) = \frac{2}{b-a} \int_{x_1}^{x_2} \tilde{c}(x, t_k) \cos\Big(j\pi\frac{x-a}{b-a}\Big) dx, \quad j = 0, 1, \ldots, N-1,$$

and

$$D_j(x_1, x_2) = \frac{2R_0}{b-a} \int_{x_1}^{x_2} \cos\Big(j\pi\frac{x-a}{b-a}\Big) dx, \quad j = 0, 1, \ldots, N-1,$$

where $\tilde{c}(x, t_{ML}) = g(x)$, and $\tilde{c}(x, t_k)$, $t_k \in \mathbb{T}$, are given in (40).

We first calculate $V_j(t_{ML}), j = 0, 1, \dots, N-1$. We have

$$V_j(t_{ML}) = \begin{cases} C_j(0, h; t_{ML}) + D_j(h, b), & \text{for a call option,} \\ C_j(a, 0; t_{ML}) + D_j(h, b), & \text{for a put option,} \end{cases}$$

for all $j = 0, 1, \dots, N-1$. Denote

$$\Phi_j(x_1, x_2) = \int_{x_1}^{x_2} e^x \cos\Big(j\pi \frac{x-a}{b-a}\Big) dx \quad \text{and} \quad \Psi_k(x_1, x_2) = \int_{x_1}^{x_2} \cos\Big(j\pi \frac{x-a}{b-a}\Big) dx,$$

for any $a \leq x_1 \leq x_2 \leq b$ and $j = 0, \dots, N-1$. Then, by simple integration, these integrals admit the following analytic solutions:

$$\begin{aligned} \Phi_j(x_1, x_2) = \; & \frac{1}{1 + (\frac{j\pi}{b-a})^2} \Big[\cos\Big(j\pi \frac{x_2 - a}{b-a}\Big) e^{x_2} - \cos\Big(j\pi \frac{x_1 - a}{b-a}\Big) e^{x_1} \\ & + \frac{j\pi}{b-a} \sin\Big(j\pi \frac{x_2 - a}{b-a}\Big) e^{x_2} - \frac{j\pi}{b-a} \sin\Big(j\pi \frac{x_1 - a}{b-a}\Big) e^{x_1} \Big], \end{aligned}$$

and

$$\Psi_j(x_1, x_2) = \Big[\sin\Big(j\pi \frac{x_2 - a}{b-a}\Big) - \sin\Big(j\pi \frac{x_1 - a}{b-a}\Big) \Big] \frac{b-a}{j\pi},$$

for $j = 0, \dots, N-1$, with $\Psi_0(x_1, x_2) = x_2 - x_1$. Moreover, by a simple calculation, $j = 0, \dots, N-1$, we have the following result:

$$D_j(x_1, x_2) = \frac{2R_0}{b-a} \Psi_j(x_1, x_2), \tag{42}$$

$$C_j(x_1, x_2; t_{ML}) = \frac{2}{b-a} \alpha K \big(\Phi_j(x_1, x_2) - \Psi_j(x_1, x_2)\big), \tag{43}$$

where α is a parameter such that $\alpha = 1$ for a call option and $\alpha = -1$ for a put option.

And next, we consider to calculate the integrals $V_j(t_k)$ for $t_k \in \mathbb{T} \setminus \mathbb{T}_e$. We have

$$V_j(t_k) = C_j(a, h; t_k) + e^{-r(T - t_{k-1})} D_j(h, b), \quad j = 0, 1, \dots, N-1.$$

Since the integral $D_j(x_1, x_2)$ has the analytic representation (45), we only need to calculate the integral $C_j(x_1, x_2; t_k)$. Fang & Oosterlee (2009) gave an efficient numerical algorithm which approximates $C_j(x_1, x_2; t_k)$ by using FFT method with the operation cost $O(N \log_2(N))$.

Finally, we consider to calculate the integrals $V_k(t_k)$ for $t_k \in \mathbb{T}_e$. It is clear that we should find the value $\tilde{v}(x, t_k)$ in the last equation in (31), or equivalently, to determine the early-exercise point x_k^* at each time t_k, which is the point where the continuation value is equal to the payoff, i.e., $\tilde{c}(x_k^*, t_k) = g(x_k^*)$. Let

$$h_k(y) = \tilde{c}(y, t_k) - g(y), \quad t_k \in \mathbb{T}_e.$$

Then, the problem becomes to find the root x_k^* of each equation $h_k(y) = 0$. Note that the function $\tilde{c}(y, t_k)$, which is given in (40), is bounded and smooth, and the function $g(y)$ is smooth except for $y = 0$ and bounded in $[a, b]$. We can use the Newton's method or the secant method to find the root x_k^*. Here if x_k^* is not in the interval $[a, b]$, we set it in the nearest boundary point a or b. Once we find the early-exercise point x_k^*, $t_k \in \mathbb{T}_e$, we have two different cases for an up-and-out barrier option:

Case 1: $x_k^* < h$, which means the early-exercise point doesn't hit the up barrier. Thus, We have the authority to decide to execute the option now or reserve it to the next time point. So we can split the integral that defines $V_j(t_k)$ into three parts: $[a, x_k^*]$, (x_k^*, h) and $[h, b]$. We have

$$V_j(t_k) = \begin{cases} C_j(a, x_k^*; t_k) + C_j(x_k^*, h; t_{ML}) + e^{-r(T-t_{k-1})} D_j(h, b), \text{ for a call option,} \\ C_j(a, x_k^*; t_{ML}) + C_j(x_k^*, h; t_k) + e^{-r(T-t_{k-1})} D_j(h, b), \text{ for a put option.} \end{cases}$$

Case 2: $x_k^* \geq h$, which means the early-exercise point hits the up barrier. Thus, the option integral can be split into two parts: $[a, h)$ and $[h, b]$:

$$V_j(t_k) = \begin{cases} C_j(a, h; t_k) + e^{-r(T-t_{k-1})} D_j(h, b), \quad \text{for a call option,} \\ C_j(a, h; t_{ML}) + e^{-r(T-t_{k-1})} D_j(h, b), \text{ for a put option,} \end{cases}$$

Ding *et al* (2011a) gave a detail algorithm, which summarizes the above procedure, for pricing an up-and-out Bermudan barrier option, and the corresponding numerical experiments.

4.3 The fast Hilbert transform approach for pricing barrier options

Feng & Linetsky (2008) presented a new numerical method to price discretely monitored barrier options under exponential Lévy models. Their method involves the relation with Hilbert transform (Property P4) and the Sine expansion in Hardy spaces. They also gave an efficient computational algorithm based on the fast Hilbert transform.

Let $\mathbb{T} = \{t_k : k = 1, \ldots, M\}$ be the set of pre-specified monitored dates, where

$$0 = t_0 < t_1 < \cdots < t_M = T \quad \text{with} \quad \Delta t = t_k - t_{k-1} = T/M.$$

We consider a European-type barrier put option whose payoff at maturity T is given by

$$G = 1_{\{S_{t_1} > L\}} 1_{\{S_{t_2} > L\}} \cdots 1_{\{S_{t_M} > L\}} (K - S_T)^+,$$

where K is the strike price and $0 < L < K$ is the lower barrier. We also assume that the underlying asset price process is given by $S_t = Ke^{X_t}$ with $S_0 = S$, where X_t is a Lévy process started at $x = \log(S/K)$. Denoting $l = \log(L/K)$, then, with help of the risk-neutral valuation formula, the price of this option is given by

$$V(x, t_0) = e^{-rT} \mathbb{E}\big[1_{\{X_{t_1} > l\}} 1_{\{X_{t_2} > l\}} \cdots 1_{\{X_{t_M} > l\}} K(1 - e^{X_{t_M}})^+ \mid X_0 = x\big],$$

which can be computed recursively by the following backward induction:

$$\begin{cases} V(x, t_M) = 1_{\{x > l\}} K(1 - e^x)^+, \\ V(x, t_k) \ = e^{-r\Delta t} 1_{\{x > l\}} \mathbb{E}\big[V(X_{t_{k+1}}, t_{k+1}) \mid X_{t_k} = x\big], \quad k = M-1, \ldots, 1, \\ V(x, t_0) \ = e^{-rT} \mathbb{E}\big[e^{-r\Delta t} V(X_{t_1}, t_1) \mid X_{t_0} = x\big]. \end{cases} \tag{44}$$

Since each Lévy process is stationary and has independent increments, for each $k = 1, \ldots, M-1$, we have

$$\mathbb{E}\big[V(X_{t_{k+1}}, t_{k+1}) \mid X_{t_k} = x\big] = \mathbb{E}\big[V(X_{t_1}, t_{k+1}) \mid X_{t_0} = x\big].$$

Thus, letting $P_{t_1}v(x) = \mathbb{E}\big[v(X_{t_1}) \mid X_{t_0} = x\big]$ be the expectation operator, from the backward induction (44) we get

$$\begin{cases} v_M(x) = 1_{\{x>l\}} \cdot K(1-e^x)^+, \\ v_k(x) = 1_{\{x>l\}} P_{t_1} v_{k+1}(x), \quad j = M{-}1,\dots,1, \\ v_0(x) = P_{t_1} v_1(x), \end{cases} \tag{45}$$

where $v_k(x) = e^{r\Delta t}V(x,t_k)$. And hence, the problem now becomes to find the function $v_0(x)$ from the backward induction (45).

Note that $1_{\{x>0\}} = \frac{1}{2}\big(1+\mathrm{sgn}(x)\big)$ for all $x \in \mathbb{R}$. Using the Property P4 (Relation with Hilbert transform) we obtain

$$\mathcal{F}(1_{\{x>0\}} \cdot v)(u) = \frac{1}{2}\big(\mathcal{F}v(u) + \mathrm{i}\mathcal{H}(\mathcal{F}v)(u)\big). \tag{46}$$

Let $a \in \mathbb{R}$ and $\mathcal{T}_a$ be the transform operator: $\mathcal{T}_a v(x) = v(x-a)$. Then, we have

$$1_{\{x>a\}} = \mathcal{T}_a 1_{\{x>0\}} = \frac{1}{2}\big(1 + \mathcal{T}_a \mathrm{sgn}(x)\big),$$

and hence,

$$1_{\{x>a\}} \cdot v(x) = \frac{1}{2}\big(v(x) + v(x)\cdot \mathcal{T}_a \mathrm{sgn}(x)\big) = \frac{1}{2}\Big(v(x) + \mathcal{T}_a\big(\mathrm{sgn}\cdot \mathcal{T}_{-a}v\big)(x)\Big).$$

Taking the Fourier transform to both sides of this equation, we have

$$\mathcal{F}\big(1_{\{x>a\}} \cdot v\big)(u) = \frac{1}{2}\mathcal{F}v(u) + \frac{1}{2}\mathcal{F}\Big(\mathcal{T}_a\big(\mathrm{sgn}\cdot \mathcal{T}_{-a}v\big)\Big)(u).$$

and using equation (46) we obtain

$$\mathcal{F}\big(1_{\{x>a\}} \cdot v\big)(u) = \frac{1}{2}\mathcal{F}v(u) + \frac{1}{2}\mathrm{i}e^{\mathrm{i}au}\mathcal{H}\big(e^{-\mathrm{i}ay}\mathcal{F}v(y)\big)(u). \tag{47}$$

On other hand, noting that, for each $k = 1,\dots,M{-}1$, the condition density of $X_{t_{k+1}}$ given $X_{t_k} = x$ possesses the property:

$$f(y|x) = f(y{-}x), \quad x,y \in \mathbb{R},$$

where $f(y)$ is the density of X_{t_1} under the initial condition $X_{t_0} = x$. The infinite integrals $P_{t_1}v_{k+1}(x)$ in (45) becomes to

$$P_{t_1}v_{k+1}(x) = \int_{-\infty}^{+\infty} v_{k+1}(y) f(y{-}x)\mathrm{d}y = \int_{-\infty}^{+\infty} v_{k+1}(x{+}z) f(z) dz.$$

Then, this integral can be rewritten as a convolution of $v_{k+1}(x{+}z)$ and the function $f(-x)$, i.e.

$$P_{t_1}v_{k+1}(x) = f(-x) * v_{k+1}(x).$$

Note that $\mathrm{supp} v_k(x) \subset (l,0)$ for each $k = M,\dots,1$ from the backward induction (45). We can take the Fourier transform in the both sides of (45). Applying the Property P3 (Convolution) we have

$$\mathcal{F}\big[P_{t_1}v_k(x)\big](u) = \mathcal{F}\big[f(-x) * v_{k+1}(x)\big](u) = \mathcal{F}\big[f(-x)\big](u)\cdot \mathcal{F}\big[v_{k+1}(x)\big](u).$$

Denote $\varphi(u)$ is the characteristic function of $f(x)$ on the complex plane $\mathbb{C}$. Then, by a simple calculation, we have

$$\mathcal{F}\big[P_{t_1} v_k(x)\big](u) = \varphi(-u) \cdot \mathcal{F}v_{k+1}(u). \tag{48}$$

Thus, using the formulas (47) and (48) we obtain

$$\begin{cases} \hat{v}_M(u) = \frac{K(1-e^{\mathrm{i}ul})}{\mathrm{i}u} - \frac{K(1-e^{(1+\mathrm{i}u)l})}{1+\mathrm{i}u}, \\ \hat{v}_k(u) = \frac{1}{2}\varphi(-u)\hat{v}_{k+1}(u) + \frac{1}{2}\mathrm{i}e^{\mathrm{i}ul}\mathcal{H}\big(e^{-\mathrm{i}yl}\varphi(-y)\hat{v}_{k+1}(y)\big)(u), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = M-1, \ldots, 1, \\ \hat{v}_0(u) = \varphi(-u)\hat{v}_1(u). \end{cases} \tag{49}$$

Here $\hat{v}_k(u) = \mathcal{F}v_k(u)$ for each k. Applying the truncated Sinc approximation, Feng & Linetsky (2008) obtained the discretization of $\hat{v}_k(u)$:

$$\begin{aligned} \hat{v}_k(nh) &= \frac{1}{2}\varphi(-nh)\hat{v}_{k+1}(nh) \\ &\quad + \frac{\mathrm{i}e^{\mathrm{i}nhl}}{2\pi} \sum_{j=-N, j\neq n}^{N} e^{-\mathrm{i}jhl}\varphi(-jh)\hat{v}_{k+1}(jh)\frac{1-(-1)^{n-j}}{n-j}, \end{aligned} \tag{50}$$

for $n = -N, \ldots, N$ and $k = M-1, \ldots, 1$, where N is a positive integer and h is the discretization step size. Then, the function $v_0(x)$ can be computed by the discretised inversion Fourier transform:

$$v_0(x) = \frac{1}{2\pi}\sum_{j=-N}^{N} e^{-\mathrm{i}jhx}\varphi(-jh)\hat{v}_1(jh)h.$$

Furthermore, Feng & Linetsky (2008) showed that the computation (50) involves a Toeplitz matrix-vector multiplication, which can be accomplished in $O(N \log_2 N)$ operations using the FFT technique. They also referred the corresponding algorithm of computing the discrete Hilbert transform via the FFT as the *Fast Hilbert Transform*.

5. References

Barndorff-Nielsen, O. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling, *Scandinavian Journal of Statistics*, Vol. 24, pp. 1–13.

Bates, D. S. (1996). Jumps and stochastic volatility: exchange rate processes implicit in Deutsche mark options, *Rev. Financ. Stud.*, Vol. 9, pp. 69–107.

Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy*, Vol. 81 (No. 3), pp. 637–654.

Borak, S., Detlefsen, K. & Härdle, W. (2005). FFT based option pricing, SFB 649 Discussion paper 2005-011. http://sfb649.wiwi.hu-berlin.de

Carr, P., Geman, H., Madan, D. & Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business*, Vol. 75, pp. 305–332.

Carr, P. & Madan, D. (1999). Option valuation using the fast Fourier transform, *Journal of Computational Finance*, Vol. 2 pp. 61–73.

Carr, P. & Wu, L. (2003). The finite moment log stable process and option pricing, *Journal of Finance*, Vol. 58, pp. 753–777.

Cont, R. & Tankov, P. (2003). *Financial Modelling with Jump Processes*, Chapman & Hall/ CRC Press.

Dempster, M. A. & Hong, S. S. G. (2002). Spread option valuation and the fast Fourier transform, *Mathematical Finance–Bachelier Congress 2000*, Spring, pp. 203–220.

Ding, D & U, S. C. (2010). An accurate and stable FFT-based method for pricing options under exp-Levy processes, *The Proceedings of ISCM II - EPMSC XII*, American Institute of Physics, pp. 741–746.

Ding, D & U, S. C. (2011). Efficient option pricing methods based on Fourier series expansions, *Journal of Mathematical Research and Exposition*, Vol. 31, pp. 12–22.

Ding, D., Huang, H. & Zhao, J. (2011a). An efficient algorithm for Bermudan barrier option pricing, Working paper, University of Macau.

Ding, D., Weng, Z. & Zhao, J. (2011b). Pricing Bermudan barrier options via a fast and accurate FFT-based Method, to appear in the Proceedings of CiSE 2011, IEEE.

Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions, *Econometrica*, Vol. 68, pp. 1343–1376.

Dufresne, D., Garrido, J. & Morales, M. (2009). Fourier inversion formulas in option pricing and insurance, *Methodol. Comput. Appl. Probab.*, Vol. 11, pp. 359–383.

Eberlein, E., Glau, K., & Papapantoleon, A. (2010). Analysis of Fourier transform valuation formulas and application, *Applied Mathematical Finance*, Vol. 17, pp. 211–240.

Fang, F. & Oosterlee, C. W. (2008). A novel pricing method for European option based on Fourier-cosine series expansions, *SIAM J. Sci. Comput.*, Vol. 31, pp. 826–848.

Fang, F. & Oosterlee, C. W. (2009). Pricing early-exercise and discrete barrier options by Fourier-cosine series expansions, *Numer. Math.*, Vol. 114, pp. 27–62.

Fang, F. & Oosterlee, C. W. (2011). A Fourier-based valuation method for Bermudan and barrier options under Heston's model, *SIAM J. Financial Math.*, Vol. 2, pp. 439–463.

Feng, L. & Linetsky, V. (2008). Pricing discretely monitored barrier options and defaultable bonds in Lévy process models: a fast Hilbert transform approach. *Mathematical Finance*, Vol. 18, pp. 337–384.

Gurland, J. (1948). Inversion formula for the distribution of ratios, *Ann. Math. Statist.*, Vol. 19, pp. 228–237.

Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Rev. Financ. Stud.*, Vol. 6, pp. 327–343.

Hurd, T. & Zhou, Z. (2010). A Fourier transform method for spread option pricing, *SIAM J. Finan. Math.*, Vol. 1, pp. 142–157.

Lee, R. (2004). Option pricing by transform methods: Extensions, unification, and error control, *Journal of Computational Finance*, Vol. 7, pp. 51–86.

Lewis, A. (2001). A simple option formula for general jump-diffusion and other exponential Lévy processes. http://www.optioncity.net

Lord, R., Fang, F., Bervoets, F. & Oosterlee, C. W. (2008). A fast and accurate FFT-based method for pricing early-exercise options under Lévy processes. *SIAM J. Sci. Comput.*, Vol. 30, pp. 1678–1705.

Kou, S. G. (2002). A jump-diffusion model for option pricing, *Management Science*, Vol. 48, pp. 1086-1101.

Kwok, Y. K., Leung, K. S. & Wong, H. Y. (2010). Efficient options pricing using the fast Fourier transform. http://ssrn.com/abstract=1534544

Madan, D., Carr, P. & Chang, E. (1998). The variance gamma process and option pricing, *European Financial Review*, Vol. 2, pp. 79–105.

Merton, R. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* , Vol. 4, pp. 141–183.

Merton, R. (1976). Option pricing when underlying stock returns are discontinuous. *J. Financ. Econ.*, Vol. 3, pp. 125–144.

Schmelzle, M. (2010). Option Pricing Formulae using Fourier Transform: Theory and Application. http://pfadintegral.com/articles/option-pricing-formulae-using-fourier-transform/

Shephard, N. G. (1991). From characteristic function to distribution function: a simple framework for the theory, *Econometric Theory*, Vol. 7, pp. 519–529.

Stein, E. & Stein, F. (1991). Stock price distribution with stochastic volatility: An analytic approach, *Rev. Financ. Stud.*, Vol. 4, pp. 727–752.

12

Hilbert Transform and Applications

Yi-Wen Liu
National Tsing Hua University
Taiwan

1. Introduction

Hilbert transform finds a companion function $y(t)$ for a real function $x(t)$ so that $z(t) = x(t) + iy(t)$ can be analytically extended from the real line $t \in \mathcal{R}$ to upper half of the complex plane.

In the field of signal processing, Hilbert transform can be computed in a few steps: First, calculate the Fourier transform of the given signal $x(t)$. Second, reject the negative frequencies. Finally, calculate the inverse Fourier transform, and the result will be a complex-valued signal where the real and the imaginary parts form a Hilbert-transform pair.

When $x(t)$ is narrow-banded, $|z(t)|$ can be regarded as a slow-varying envelope of $x(t)$ while the phase derivative $\partial_t[\tan^{-1}(y/x)]$ is an instantaneous frequency. Thus, Hilbert transform can be interpreted as a way to represent a narrow-band signal in terms of amplitude and frequency modulation. The transform is therefore useful for diverse purposes such as latency analysis in neuro-physiological signals (Recio-Spinoso et al., 2011; van Drongelen, 2007), design of bizarre stimuli for psychoacoustic experiments (Smith et al., 2002), speech data compression for communication (Potamianos & Maragos, 1994), regularization of convergence problems in multi-channel acoustic echo cancellation (Liu & Smith, 2002), and signal processing for auditory prostheses (Nie et al., 2006).

The rest of this review chapter is organized as follows: Sec. 2 reviews the mathematical definition of Hilbert transform and various ways to calculate it. Secs. 3 and 4 review applications of Hilbert transform in two major areas: Signal processing and system identification. The chapter concludes with remarks on the historical development of Hilbert transform in Sec. 6.

2. Mathematical foundations of Hilbert transform

The desire to construct the Hilbert transform stemmed from this simple quest: Given a real-valued function $f : \mathcal{R} \to \mathcal{R}$, can we find an imaginary part ig such that $f_c = f + ig$ can be analytically extended? For example, if $f(x) = \cos(x)$, then by inspection we can find $g(x) = \sin(x)$ such that $f_c(x) = f + ig = \exp(ix)$. This function can obviously be extended analytically to the entire complex plane by replacing the real variable x with the complex variable z in the expression; the result is $f_{\text{ext}}(z) = \exp(iz)$ and we have

$$\text{Re}\{f_{\text{ext}}(z)\}|_{z=x} = f(x), \tag{1}$$

which states that real part of the extended function is equal to the original given function $f(x)$ on the real line. The companion function $g(x)$ is called the Hilbert transform of $f(x)$.

This simple example brings forth a few questions. First, is analytic extension always possible for reasonably smooth $f(x)$? If so, is the companion function $g(x)$ unique, in the sense that Eq. (1) would hold? The answer to the second question is a definite 'No', because any $f_c(x) = f(x) + i\{g(x) + g_0\}$ also satisfies Eq. (1), where g_0 is an arbitrary constant. Therefore, the original simple quest needs to be refined as follows.

2.1 Hilbert transform as a boundary-value problem

To establish the uniqueness of the companion function, we first note that any analytic function $f_{\text{ext}}(z) = f_R(z) + if_I(z)$ defined on the complex plane $z = x + iy$ must satisfy Cauchy-Riemann equations,

$$\frac{\partial f_R}{\partial x} = \frac{\partial f_I}{\partial y},$$
$$\frac{\partial f_I}{\partial x} = -\frac{\partial f_R}{\partial y}.$$

Consequently, both f_R and f_I satisfy Laplace's equation,

$$\frac{\partial^2 F_R}{\partial x^2} + \frac{\partial^2 F_R}{\partial y^2} = 0, \tag{2}$$
$$\frac{\partial^2 F_I}{\partial x^2} + \frac{\partial^2 F_I}{\partial y^2} = 0 \tag{3}$$

over the region where $f_{\text{ext}}(z)$ is analytic. Conventionally, by requiring $f_{\text{ext}}(z)$ to be analytic in the upper half-plane, the quest of finding the Hilbert transform for any given function $f(x)$ can be formulated as a boundary value problem (Scott, 1970). By specifying the boundary conditions that

- $f_R(x, 0) = f(x)$, and that
- $f_R(x, y) = 0$ as $x \to \pm\infty$ or $y \to \infty$,

$f_R(x, y)$ can be uniquely determined by solving Laplace's equation (Eq. 2) in the upper half plane. Then, through Cauchy-Riemann equations, $f_I(x, y)$ can be calculated in the entire upper half plane and in particular on its boundary the x-axis (Scott, 1970). Thus $g(x) = F_I(x, 0)$ is the Hilbert transform of the given function $f(x)$.

2.2 Calculation through improper integrals

The above formulation of Hilbert transform as a boundary-value problem is hardly mentioned in recent texts [except as an exercise problem in Oppenheim & Schafer (2010)].[1]: Instead,

[1] The boundary-value-problem formulation is missing for a reason — it does not tell us how to *compute* the Hilbert transform.

Hilbert transform is commonly introduced and defined through an improper integral [e.g., (Hahn, 1996)]:

$$g(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(u) \frac{1}{x-u} du. \tag{4}$$

Here, note that the convolution kernel function $h(x) = 1/\pi x$ is singular at $x = 0$. Therefore, the integral in Eq. 4 is improper in the sense of Cauchy's principal value:

$$g(x) = \lim_{\epsilon \to 0} \left(\int_{-\infty}^{x-\epsilon} + \int_{x+\epsilon}^{\infty} \right) f(u) \cdot h(x-u) du. \tag{5}$$

To be convinced that Eq. 4 indeed produces the Hilbert transform, we need to think about the effects of Hilbert transform in the frequency domain. First, for any frequency k, note that the Hilbert transform of $f_k(x) = \cos(kx)$ is $g_k(x) = \sin(kx)$. So, we can understand Hilbert transform as a *phase shifter* which gives every sinusoidal function -90 degrees of phase shift. Therefore, in the frequency domain, we have

$$G(k) = F(k) \cdot (-i \cdot \text{sgn}(k)), \tag{6}$$

where $G(k)$ and $F(k)$ are the Fourier transform of $g(x)$ and $f(x)$, respectively, and $\text{sgn}(x)$ is the sign function (i.e., $\text{sgn}(k) = 1$ if $k > 0$ and $\text{sgn}(k) = -1$ if $k < 0$.) Therefore, if we think of $H(k) = -i \cdot \text{sgn}(k)$ as the transfer function of a phase-shift kernel $h(x)$, the kernel can be written as the inverse Fourier transform of the transfer function; that is,

$$h(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(k) e^{ikx} dk. \tag{7}$$

Note that $H(k) = -i$ for $k > 0$ and $H(k) = i$ for $k < 0$. Therefore, $H(k)$'s first derivative with respect to k is

$$\frac{\partial H}{\partial k} = -2i\delta(k), \tag{8}$$

where $\delta(k)$ is the Dirac delta function. Since the operator $\partial/\partial k$ in the frequency domain corresponds to multiplication by $-ix$ in the space domain, we can take the inverse Fourier transform on both sides of Eq. 8 and obtain the following,

$$-ix \cdot h(x) = -2i \cdot \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \delta(k) e^{ikx} dk \right) = \frac{-i}{\pi}. \tag{9}$$

Dividing both sides by $-ix$, we conclude that the convolution kernel $h(x) = 1/\pi x$.

2.3 The notion of Hilbert transform "pairs"

The phase-shifter interpretation of Hilbert transform leads to the fact that if $f(x)$'s Hilbert transform is $g(x)$, then $g(x)$'s Hilbert transform is $-f(x)$; in this sense, $f(x)$ and $g(x)$ form a *Hilbert transform pair*.

This symmetric property can be understood as follows. Note that the $H^2(k) = -1$ for all k since $H(k) = \pm i$. This means that if we take the Hilbert transform twice, the result would be the original function with a negative sign.

This makes sense because Hilbert transform introduces a 90-degree phase shift to all simple harmonics. Therefore, Hilbert transform repeated twice introduces a 180-degree phase shift to all simple harmonics, which means multiplication of the original function by -1.

A table of commonly used Hilbert transform pairs can be found in the Appendix of Hahn (1996) for applications in signal processing. A thorough 80-page table of Hilbert transform pairs can be found in the Appendix of King (2009b) and transform pairs are also plotted in a 20-page atlas.

2.4 The convolution kernel $h(x)$ as the Hilbert transform of $\delta(x)$

We now introduce another way to derive the convolution kernel of Hilbert transform. To begin, note that Eq. 5 essentially states that Hilbert transform is a filtering process which is characterized by its impulse response $h(x)$. Therefore, $h(x)$ must be regarded as the Hilbert transform of the impulse function $\delta(x)$. Then, it is of our interest to check that

$$f_c(x) = \delta(x) + ih(x)$$

can be regarded as an analytic function in the sense of Eq. 1. To see it, consider a family of complex analytic functions $f(z) = i/\pi(z + i\eta)$ parametrized by a variable $\eta > 0$. Since the only singularity of $f(z)$ is at $z = -i\eta$, $f(z)$ is analytic in the entire upper half plane. Therefore, the real part and imaginary part of $f(z)$ form a Hilbert transform pair on the real line $x \in \mathcal{R}$. With a little algebra, the real and imaginary parts can be written as

$$f(x) = \frac{i}{\pi(x + i\eta)} = f_R(x) + if_I(x), \tag{10}$$

where

$$f_R(x) = \frac{\eta}{\pi(x^2 + \eta^2)}$$

and

$$f_I(x) = \frac{x}{\pi(x^2 + \eta^2)}$$

form a Hilbert transform pair for any $\eta > 0$.

Now we let η approach zero and observe $f_R(x)$ and $f_I(x)$. Note that $\int_{-\infty}^{\infty} f_R(x)dx = 1$ regardless of the value of η, and that $f_R(0) = 1/\pi\eta$ approaches infinity as $\eta \to 0$. So we can claim that

$$\lim_{\eta \to 0} f_R(x) = \delta(x). \tag{11}$$

Meanwhile, it is trivial that

$$\lim_{\eta \to 0} f_I(x) = 1/\pi x. \tag{12}$$

From the arguments above, we can be convinced that the Hilbert transform of $\delta(x)$ is indeed $1/\pi x$, for $f_c(x) = \delta(x) + i \cdot (1/\pi x)$ is equal to the limit function of $f(x)$ as $\eta \to 0$ (Scott, 1970).

3. Applications in signal processing

Signal processing is nowadays conveniently conducted in the digital domain. The first step of signal digitization involves sampling an analog signal $x(t)$ at a constant rate $f_s = 1/T$ where

T is the sampling period. In this section, we denote the sampled waveform as $x[n] = x(nT)$, using the square brackets $[\cdot]$ to indicate that the signal is sampled in discrete time. So the *discrete-time Fourier transform* (DTFT) is defined as follows:[2]

$$X(j\omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}.$$

Note that $X(j\omega)$ is periodic at every 2π in the frequency domain.[3] Next, we show how Hilbert transform can be defined in discrete time.

3.1 The discrete-time Hilbert transform and Hilbert transformers

Recall that the Hilbert transform introduce a 90-degree phase shift to all sinusoidal components. In the discrete-time periodic-frequency domain, the transfer function of Hilbert transform is specified as follows,

$$H(j\omega) = \begin{cases} -j, & 0 < \omega < \pi \\ j, & -\pi < \omega < 0 \end{cases} \tag{13}$$

The convolution kernel for $H(j\omega)$ can be calculated through inverse Fourier transform (Oppenheim & Schafer, 2010):

$$h[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(j\omega)e^{j\omega n}d\omega \tag{14}$$

$$= \begin{cases} \frac{2}{\pi}\frac{\sin^2(\pi n)}{n}, & n \neq 0 \\ 0, & n = 0 \end{cases} \tag{15}$$

Note that $h[n]$ has a infinite support from $n = -\infty$ to ∞. In practice, the entire function can not be stored digitally. To circumvent this difficulty, we now discuss two major methods for calculating the discrete-time Hilbert transform.

3.1.1 The MATLAB approach

The universally popular scientific-computing software MATLAB (MathWorks, Natick, Massachusetts, USA) has a `hilbert()` function that "computes the so-called discrete-time analytic signal `X = Xr + i*Xi` such that `Xi` is the Hilbert transform of `Xr`".[4] MATLAB's implementation of the `hilbert()` function takes advantage of the fast Fourier transform (FFT). Essentially, the `hilbert()` function completes the calculation in three steps:

- Do the FFT of `Xr`.
- Set the elements in FFT which correspond to frequency $-\pi < \omega < 0$ to zero.
- Do the inverse FFT.

[2] We now switch to the electrical-engineering convention of using j to refer to $\sqrt{-1}$.

[3] Hereafter, we use capital letters $X, Y, Z, H, ...$ to denote spectrums in the frequency domain, and lowercase letters x, y, z, h for signals in the time domain.

[4] This is what MATLAB's help file shows.

That the three steps above can work is a consequence of the fact that, if $x_i[n]$ is the discrete Hilbert transform of $x_r[n]$, the Fourier transform of $x_r[n] + jx_i[n]$ vanishes for all negative frequencies $-\pi < \omega < 0$. This can be verified by inspecting the definition given in Eq. 13.

Remarks: Perhaps because of MATLAB's popularity and easiness to use, many have allegedly mis-regard the returned vector `X=hilbert(Xr)` as the Hilbert transform of `Xr`. One must be aware of these deviations from the conventional definition to avoid unnecessary confusion.

3.1.2 Hilbert transform as a filter design problem

Thanks to FFT, MATLAB's implementation of Hilbert transform is very efficient. However, it should be used with caution — note that the Hilbert transform defined in Eq. 13 has a discontinuity at $\omega = 0$. Consequently (due to Gibb's phenomenon), the convolution kernel $h[n]$ in Eq. 15 has an infinite support in time. As a result, when implemented through FFT, the Hilbert transform kernel wraps around itself and time-domain aliasing comes in (Oppenheim & Schafer, 2010). The time-domain aliasing could be perceived as artifacts in applications such as audio and video signal processing.

To avoid time-domain aliasing, one can formulate discrete Hilbert transform as a filter-design problem. The ideal transfer function is specified by Eq. 13, and there are standard techniques to design finite impulse response (FIR) or infinite impulse response (IIR) filters that appraoch the ideal transfer function. For example, one can truncate the ideal impulse response by multiplying it with a window function $w[n]$ which has a finite support (let's say from $n = -N$ to N). Then the resulting function $w[n]h[n]$ yields an approximate magnitude response $H_w(j\omega)$ that has a smooth transition between negative and positive frequencies as well as ripples in both regions. The height of the ripples can be reduced by selecting the window function wisely, while the transition bandwidth is inversely proportional to the window length. Interested readers can refer to Chapter 7 and 12 of Oppenheim & Schafer (2010).

The FIR or IIR filters designed to approximate Hilbert transform are called *Hilbert transformers*. Next, we discuss applications of Hilbert transformers in communication and in biomedical engineering.

3.2 Sampling of bandpass signals for communication

An important application of Hilbert transformers is in sampling bandpass signals.[5] To explain this, let us assume that a bandpass signal $s(t)$ is has a region of support $f_c \leq f \leq f_c + \Delta f$ in the frequency domain, where $\Delta f = 0.2 f_c$. Based on Nyquist's theorem, the sampling rate needs to be at least two times the highest frequency, or $2.4 f_c$, to avoid frequency-domain aliasing. However, the bandwidth of this signal is really $\Delta f = 0.2 f_c$, so $f_s = 2.4 f_c$ is in fact an *oversampling*.

To take advantage of the narrow bandwidth, we initially need to sample at $2.4 f_c$ to obtained an oversampled signal $s_r[n] = s(nT)$, where $T = 1/f_s$. Then, we can use a Hilbert transformer to obtain $s_i[n]$ such that the Fourier transform of $s[n] = s_r[n] + js_i[n]$ has no components at negative frequencies $\pi < \omega < 0$. Now, $S(j\omega)$'s region of support is $\omega \in (5\pi/6, \pi)$. Therefore, we can down sample $s[n]$ by a factor of 6 and there would be no frequency-domain aliasing.

[5] This part is adapted from Sec. 12.4.3 of Oppenheim & Schafer (2010)

Transmitting $s_d[n] = s[6n]$ is more efficient than transmitting $s_r[n]$ because the sampling rate is lowered.

To reconstruct $s_r[n]$ from $s_d[n]$, we can do the following:

- Expand $s_d[n]$ by a factor of 6; i.e., construct $s_e[n] = \begin{cases} s_d[n/6], & \text{if } n = 0, \pm 6, \pm 12, ... \\ 0, & \text{otherwise.} \end{cases}$
- Filter $s_e[n]$ with the passband of $(5\pi/6, \pi)$.
- Take the real part, and the result is a reconstructed copy of $s_r[n]$.

In practice, since all Hilbert bandpass filters are not ideal, one needs to consider sampling at slightly higher than $2.4 f_c$. That would protect the passband from ripple interference both during downsampling and during signal reconstruction.

3.3 AM-FM decomposition for auditory prostheses

A cochlear implant device consists of up to tens of electrodes which are inserted as an array to the cochlea to stimulate auditory nerves by electrical currents (Zeng et al., 2008). Each channel (electrode) represents an acoustic frequency band; the amount of currents sent to an electrode should faithfully reflect how acoustic energy entering through the ear microphone varies in time in the corresponding frequency band. Typically, acoustic waveforms are processed with a bank of filters and the resulting envelopes control the electrical currents sent to individual electrodes.

In this application scenario, a Hilbert transformer has been found useful for envelope extraction. This can be understood by noting that the Hilbert transform produces a $\sin(\omega t)$ for every $\cos(\omega t)$. For a narrow-band signal $s_R(t)$, we can factor it as a product of slow-varying *envelope* $A(t)$ and fast-varying *fine structure* $f(t)$:

$$s_R(t) = A(t) f(t) = A(t) \cos[\phi(t)]$$

where $d\phi(t)/dt$ can be regarded as an *instantaneous frequency* of the signal. It can be shown that, if $A(t)$ varies sufficiently slowly, the Hilbert transform produces approximately $s_I(t) = A(t) \sin[\phi(t)]$. Then, the envelope $A(t)$ can simply be estimated by taking the root of sum of squares:

$$A(t) \simeq \sqrt{s_R^2(t) + s_I^2(t)}. \tag{16}$$

Clinically, encoding the currents based on $A(t)$ provides clear speech perception for cochlear-implant users (Nie et al., 2006). Moreover, auditory pitch information can be extracted by taking the time-derivation of $\phi(t)$, which can be determined by the relative phase between s_I and s_R:

$$\phi(t) = \tan^{-1}\left(\frac{s_I(t)}{s_R(t)}\right).$$

Pitch is arguably the utmost important feature for music perception as well as understanding tonal languages. Here, we see Hilbert transform can serve as an efficient computation tool to extract it.

4. Applications in system identification

Hilbert transform relates the real part and the imaginary part of the transfer function of any *physically viable* linear time-invariant system. By "physical viability" we mean a system should be stable and causal. Stability requires the systems to produce bounded output if the input is bounded [so called the bounded-input bounded-output criterion, (Oppenheim & Schafer, 2010)]. Causality prohibits the system from producing responses before any stimulus comes in. Denote the impulse response as $h(t)$ and its Laplace transform as $H(s)$. The above conditions requires that

- $h(t) = 0$ for all $t < 0$ (causality)
- All sigularities of $H(s)$ are located in the left half-plane (stability).

The two conditions above ensure that $H(s)$ converges and is analytic in the entire right half-plane, and in particular on the imaginary axis $s = j\omega$. Therefore, the real and imaginary part of $H(j\omega) = H_R(\omega) + jH_I(\omega)$ are inter-dependent in term of the *Kramers-Krönig relations* (King, 2009b):

$$H_I(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{H_R(u)du}{\omega - u} \tag{17}$$

$$H_R(\omega) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{H_I(u)du}{\omega - u} \tag{18}$$

which is basically Hilbert transform in its time-frequency dual form.

The Kramers-Krönig relations govern how physical viable transfer functions can vary in frequency. For instance, the real and imaginary part of an electromagnetic wave propagation function defines the attenuation and the wavenumber per unit length as a function of frequency. The frequency dependence of both functions is referred to as *dispersion*, for in optics it describes how red light travels faster than violet light in water — so we behold the beauty of rainbows after raining. It is now intriguing to realize that the attenuation and velocity of light are two inter-dependent functions of frequency. More physics-inclined readers can refer to King's thorough discussion on dispersion relations in electrodynamics and optics (King, 2009b).

To a certain extent, the concept that the real and the imaginary parts are inter-dependent similarly applies to the magnitude and phase of transfer functions of a physically viable system. Note that any transfer function $H(j\omega)$ can be decomposed logarithmically into magnitude and phase:

$$\log H(j\omega) = \log|H(j\omega)| + j\angle H(j\omega).$$

This shows that the log-magnitude and the phase are real and imaginary parts of the log-spectrum, respectively. It might appear that they must satisfy the Kramers-Krönig relations. Unfortunately, this is a wishful thinking since apparently $\tilde{H}(j\omega) = \exp(-j\omega\tau)H(j\omega)$, where τ is a constant, would have the same magnitude as $H(j\omega)$ but a different phase response.[6]

It turns out that, for any given magnitude response, the uniqueness of phase response can be established if the transfer function satisfies a *minimum-phase* criterion; the criterion requires

[6] in fact, $\tilde{h}(t) = h(t - \tau)$.

that all zeros and poles of the transfer function $H(s)$ to be located in the left-half plane. This criterion ensures that all the singularities of $\log H(s)$ are located in the left-half plane so the real and imaginary parts of $\log H(s)$ become a Hilbert transform pair. Otherwise, any transfer function can be uniquely factorized as a product of a minimum-phase function $M(j\omega)$ and an all-pass function $P(j\omega)$. It is noteworthy that the system whose transfer function is $M(j\omega)$ has the minimal energy delay among all linear time-invariant systems of the same magnitude response. Further readings are recommended in Chapter 5 and 12 of Oppenheim & Schafer (2010).

5. Concluding remarks and historical developments

In this chapter, we first presented Hilbert transform as an analytic extension problem. Hilbert transform uniquely exists due to Cauchy-Riemann equations. We then reviewed several different ways to calculate Hilbert transform. A few important points stand out: First, Hilbert transform can be regarded as a 90-degree phase shifter. Secondly, the real part and imaginary part of a physically viable transfer function must satisfy Kramers-Krönig relations, which is the Hilbert transform applied in time-frequency duality. A good reference on these topics can be found in Oppenheim & Schafer (2010).

The construction of Hilbert transform pairs through Cauchy-Riemann equations in Sec. 2.1 was found in the appendices of an old text on microwave electronics (Scott, 1970). The original formulation was stated in terms of Kramers-Krönig relation, and in this chapter that formulation is adapted so the signal is defined on the real line instead of the frequency axis $j\omega$.

As a matter of fact, the definition of Hilbert transform was not given by David Hilbert himself. The name "Hilbert transform" was first given by the British mathematician G. H. Hardy in honor of Hilbert's pioneering work on integral equations (King, 2009a). Hardy's early work established mathematical rigor of the transform (Hardy, 1932), which we now apply in various areas such as physiology and telecommnication. A review of these contributions from brilliant mathematicians reminds us that the transform really is a heritage from the 20th century. Nevertheless, it is also amusing that nowadays we calculate Hilbert transform with super fast computers, which might never had been envisioned by 20th-century pioneers, including Hilbert, Hardy, Scott, or even Oppenheim.

6. Acknowledgement

This work is supported by Taiwan's National Science Council through research grant No. 100-2221-E-007-011.

7. References

Hahn, S. L. (1996). *Hilbert Transforms in Signal Processing*, Artech House, Inc., Norwood, MA, USA.

Hardy, G. H. (1932). On hilbert transforms, *Quart. J. Math. (Oxford)* 3: 102–112.

King, F. W. (2009a). *Hilbert Transforms, Vol. 1*, Cambridge University Press, Cambridge, UK.

King, F. W. (2009b). *Hilbert Transforms, Vol. 2*, Cambridge University Press, Cambridge, UK.

Liu, Y.-W. & Smith, J. O. (2002). Perceptually similar orthogonal sounds and applications to multichannel acoustic echo canceling, *Proc. Audio Eng. Soc. 22nd Int. Conf.*, Espoo, Finland.

Nie, K., Barco, A. & Zeng, F.-G. (2006). Spectral and temporal cues in cochlear implant speech perception, *Ear Hear.* 27: 208–217.

Oppenheim, A. V. & Schafer, R. W. (2010). *Discrete-Time Signal Processing, 3rd Edition*, Pearson, Boston.

Potamianos, A. & Maragos, P. (1994). A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation, *Signal Processing* 37(1): 95 – 120.

Recio-Spinoso, A., Fan, Y.-H. & Ruggero, M. (2011). Basilar-membrane responses to broadband noise modeled using linear filters with rational transfer functions, *Biomedical Engineering, IEEE Transactions on* 58(5): 1456 –1465.

Scott, A. (1970). *Active and Nonlinear Wave Propagation in Electronics*, Wiley-Interscience, New York.

Smith, Z. M., Delgutte, B. & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception, *Nature* 416: 87–90.

van Drongelen, W. (2007). *Signal Processing for Neuroscientists: Introduction to the analysis of physiological signals*, Academic Press, London.

Zeng, F.-G., Rebscher, S., Harrison, W., Sun, X. & Feng, H. (2008). Cochlear implants: System design, integration, and evaluation, *IEEE Rev. Biomed. Eng.* 1: 115–142.