PERFORMANCE ANALYSIS OF SIGNAL PROCESSING BASED PERIODICITY MINING TOOLS

Dissertation submitted in partial fulfilment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

IN

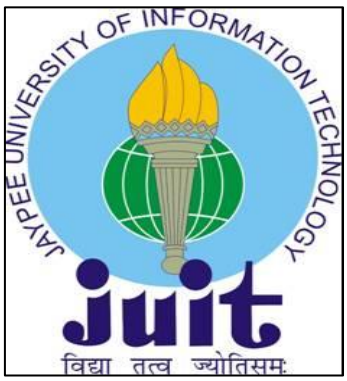
ELECTRONICS AND COMMUNICATION ENGINEERING

By

Gaurav Rangta (131073)
Soumitra Mehrotra (131109)

UNDER THE GUIDANCE OF

Dr.SunilDatt Sharma



JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

May 2017

TABLE OF CONTENTS

AKNOWLEDGEMENT

Project is the most vital part of any course, it is a link between theory and practical knowledge as well as an opportunity for hands on experience in actual environment. We, therefore consider ourself fortunate to receive the opportunity in an esteemed organization like JUIT. Yet the opportunity could not have been utilized without the guidance and support of many individuals who, although held varied positions, but were equally instrumental for completion of my project. First of all, I owe a lot of debt of gratitude to Dr.SunilDatt Sharma for his kind support and guidance. In addition, I would also like to express gratitude to the respected faculty members and ex-perts from different fields for their invaluable inputs and direction. Last but not least, I thank my family, friends, mentors and all well-wishers without whose cooperation this project would have been a dream.

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the B-Tech thesis entitled **"Performance Analysis of Signal Processing Based Periodicity Mining Tools"** submitted at Department of Electronics And Communications, **Jaypee University of Information Technology, Waknaghat India,** record of my work carried out under the supervision of **Dr.Sunil Datt Sharma.** I have not submitted this work elsewhere for any other degree or diploma.

Gaurav Rangta (131073)

Soumitra Mehrotra (131109)

Date :01/05/2017

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the B-Tech. thesis entitled **"Performance Analysis of Signal Processing Based Periodicity Mining Tools"** submitted by **Gaurav Rangta** and **Soumitra Mehrotra** at **Jaypee University of Information Technology, Waknaghat , India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

Dr.SunilDatt Sharma

Date :01/05/2017

ABSTRACT

Periodicity Mining is an area that deals with detection and identification of periodicities in any data. The work entails the analysis signal processing tools for periodicity mining and use it develop our proposed algorithm based on Intrinsic Integer-periodic function (IIPF).Our focus is to identify periodicities in a discrete integer periodic signals, whose periodic nature is far different than that of continuous time signals. The basis of the analysis is the idea that a signal or data has some hidden periodicity that are hard to estimate especially if the data is large. There are various signal processing tools for identifying periodicities in a given data or signal, like ML Techniques, CSMP Techniques, IIPF etc. We have developed Short-Time IIPF contrary to traditional IIPF. Thereafter, we have implemented and analysed the proposed algorithm on different DNA sequences.

CHAPTER 1

INTRODUCTION

A signal is said to exhibit periodic nature, when its value repeats after a finite interval of time periodically. Periodicities of some periodic mathematical functions like sinusoids are relatively easy to find out. However, finding periodicities, in signals and data, which incorporates values for a very large duration, is a tedious task. In this project, we wish to cover a broader area known as Periodicity Mining, to account for various types of periodicities.[1], [6]

1.1 Role of Periodicity

Any data that show periodic data and is a source of some useful information, can be effectively analysed by identifying its periodic nature.This can be closely justified in relation to some natural laws and phenomenons. For instance, weather prediction that is based on geographical parameters like topography and convection currents, repeats periodically. In Cryptographic technique, deciphering the cipher text involves study of patterns and repetitions in the code. If the periodicities in these patterns can be estimated, the task becomes quite easy. Periodicity Mining is an important aspect of signal processing, that involves the determination and study of periodic components in a given signal. Periodicity Mining is the study of detecting and analysing repeating sequences in a repeated pattern or data.

1.2 Types of Periodicities

We define two types of periodicities:

- **Segment periodicity** - Segment periodicity deals with the periodicity of the entire input sequence, where periodic nature of segments are being identified.[7]
- **Symbol periodicity**- Symbol periodicity implies the periodicities of the different symbols or values of the input sequence.[7]

For each periodicity type, Short time IIPF algorithm is proposed and is analysed, Furthermore, we extend the proposed algorithm to the time varying periodic signals like inverse chirp signal.

1.3 Motivation

Weather forecasting, requires the tonnes of geographical data all around the globe. This data has some hidden periodicities, which can be efficiently accounted for using periodicity mining tools. In Cryptographic technique, deciphering the cipher text involves study of patterns and repetitions in the code. If the periodicities in these patterns can be estimated, the task becomes quite easy.Various diseases like Fragile-X syndrome, Huntington's disease, Frederick's ataxia, and some other neurological, neurodegenerative and neuromuscular diseases are governed by these tandem repeats. These tandems repeat also finds application in DNA fingerprinting, behaviour of living organisms etc. The fact the variation in number of repeated patterns, their periodic behaviour and locations, possess such crucial and uncountably large information, is what motivates researchers worldwide to account and lay down for various methods for tandem repeat periodicity detection.

1.4 Objective

The objective of this project is to detect and identify hidden periodicities in a given data set, particularly a DNA sequence. To accomplish the same, we apply our proposed algorithm that is, Short time IIPF on a DNA sequence. The idea serves the sole purpose of studying the DNA sequences including mutations and its repeats, etc. using periodicity identification and their analysation.

1.5 Methods of Periodicity Detection.

There have been various proposed techniques in the field of Periodicity detection and identification based on statistical approaches. However, our focus is on signal processing based periodicity mining tools to achieve efficient identification of hidden periods. There are various methods for periodicity mining that estimates the hidden periods in a signal and data. However, we focus on following signal processing tools to identify periodicities:

1.5.1 REPORTED TOOLS

- **Intrinsic Integer-periodic function (IIPF)-** Any discrete periodic signal can be decomposed into new class of orthogonal components known as IIPF, where in each components describes the periodic nature of a signal. [1]
- **Conjugate Subspace Matching Pursuit(CSMP)-** CSMP algorithm follows a two-step procedure for identification of hidden periodicities. In the first stage, periodicity strategy is adopted to find a dominant hidden period. Then in second stage, the dominant conjugate subspace is chosen with the energy strategy.[5]
- **Maximum Likelihood Detection Technique(MLDT)-** In MLDT technique, a statistical frame work is modelled, based on probability distribution function. Each symbol of the input sequence is assumed to occur with a certain probability, and based on above model, periodic nature is described.[11]
- **Ramanujan Fourier Transform(RFT) –** RFT is based on the famous Ramanujan summation given by Dr. Srinivasa Ramanujan in 1918. It is based on study, where various mathematical function can be expressed as the linear combination of Ramanujan sum.[2]
- **Ramanujan Fourier Banks(RFB)-** RFB are also based on Ramanujan Sum, where Filter are designed to identify the hidden periodicities. This method also results in some non existing period detection, due to comples filter structure.[3]

1.5.2 PROPOSED TOOL

For efficient detection and identification of hidden periodicities in the DNA sequence, we have proposed **Short Time Intrinsic Integer-periodic function (STIIPF).** The method requires the sequence to pass through a moving rectangular window, and then apply IIPF to it. This gives us the hidden periods along with their respective positions of occur in the sequence.

1.6 Organization of the Project

The project requires us to study the various signal processing periodicity mining tools listed above. In *Chapter 2* these methods have been discussed in detail, followed by a detail discussion and explanation on *Intrinsic Integer-periodic function* in *chapter 3.* Thereafter, we have proposed *Short Time IIPF* and discussed its results on DNA sequence, in *chapter 4.* Finally the conclusion has been explained in *chapter 5.*

CHAPTER 2

LITERATURE REVIEW

There have been various signal processing tools that has been proposed so far in context to Periodicity Mining. We wish study these methods in the following section.

2.1 Intrinsic Integer-periodic function (IIPF)

Hidden Periodicities are very hard to point out as Most real world signal do not exhibit strict periodicity. Intrinsic Integer Periodic Function(IIPF) proposed by Soo Chang Lei and Keng-Shih Lu in IEEE proceedings, acts as an important tool in finding hidden periodicities in a discrete time signal. It decomposes the given signal into mutually orthogonal and periodic functions, with the help Ramanujan Subspaces, which altogether gives the information of hidden periodic components. The form of IIPF decomposition of an arbitrary -periodic signal x[n] is

$$x[n] = \sum_{d|q} x_d[n] \tag{2.1}$$

where $x_d[n]$ is a d –IIPF.The summation holds true for every divisor d of q.

$x_d[n]$ can be found as-

$$x_d[n] = \frac{1}{q}(x \otimes c_d) \tag{2.2}$$

$$c_d(n) = \sum_{\substack{k=1 \\ \gcd(k,d)=1}}^{d} exp\left(\frac{j2\pi kn}{d}\right) \tag{2.3}$$

Where $c_d(n)$ is Ramanujan Summation and forms the basis of IIPF.The components so constituted are orthogonal are Orthogonal in nature, and hence independently useful in exploring periodic characteristics in each component that defined the overall Periodic nature of the discrete signal.To extend the novelty of IIPF, Soo Chang Lei and Keng-Shih Lu also extended IIPF to frequency domain by defining IIPF Spectrum. Later, in the results we shall see that IIPF Spectrum proves to be

better than Ramanujan Fourier Transform in terms of Time Sensitivity. Mathematically, the expression of IIPF Spectrum is –

$$y_q[r] = \frac{1}{N}\sum_{n=1}^{N} y[n]c_q[n-r+1] \tag{2.4}$$

Where $y[n]$ is the input signal, N is the length of the signal. However for better graphical view, Norm $\|y_q\|$ of the above expression plotted against q.

2.2 Conjugate Subspace Matching Pursuit

This method unlike dictionary based approaches is capable of reducing the dictionary size by using a Two stage algorithm. The method was proposed by Shi-wen Deng and Ji-qing Han wherein, they revealed the conjugate symmetry of the Ramanujan subspace with a set of complex exponential basis functions and showed that they can be represent the subspace as the union of a series of conjugate subspaces(CCSs).In the first stage, the dominant hidden period is chosen with the periodicity strategy. Then, the dominant conjugate subspace is chosen with the energy strategy in the second stage. While the most traditional methods revolve around the greedy strategies to illustrate the periodic behaviour in Dictionary based approaches, CSMP can be executed over CCSs in two stages-

Stage1: In the first stage, using periodic strategy, a dominant hidden period is estimated, which is based on signal's periodic metric corresponding to each Ramanujan subspace periodicity metric, is used to estimate the dominant hidden period, which is given as

$$P(x_q, q) \square \frac{N+q}{2q}\left\|x_p^2\right\| \tag{2.5}$$

where N is the length of signal, q is the measured hidden period, and x_q is the hidden periodic component that is the projection of the signal x onto the Ramanujan subspace S_q.It can be shown that $\left\|x_p\right\|^2$ can be iteratively calculated using-

$$\left\|x_q\right\|^2 = \left\|\hat{x}_q\right\|^2 - \sum_{p\in\Gamma}\left\|x_p\right\|^2 \tag{2.6}$$

Where,

$$\left\|\hat{x}_q\right\|^2 = \frac{q}{N}\left(\phi_x(0) + 2\sum_{l=1}^{M-1}\phi_x(lq)\right) \tag{2.7}$$

Where $\phi_x(.)$ is the auto correlation function of x ,M=[N/q].So, the dominant hidden period q^* can be chosen by,

$$q^* = \operatorname{argmin}_{q\in[1,Q]} P(x_q, q) \tag{2.8}$$

where Q is the maximum period.With the chosen hidden period q^* in the first stage, the dominant CCS is chosen from all the CCSs $\{G_{q^*,i}\}_{i=1}^{M_{q^*}}$ of the Ramanujan subspace S_{q^*}, where

$$G_{q^*,i} = span\{g(\omega_{q^*,i}), \overline{g(\omega_{q^*,i})}\} \tag{2.9}$$

Stage 2*:* In the second stage, the energy strategy is used to choose dominant conjugate subspace $G_{ql,il}$ in each iteration starting from *i=0* to *i=l-1*..The most important characteristics of CSMP comprises its ability to detect not only the hidden periods, but also the changes of period with time. To be precise, the shifted CSMP enables the demonstration of period varying with time, in a time-period plane, thereby giving a visualization of period varying. More generally, the CSMP enables the decomposition of any signal into a series of periodic components, thus providing a different point of view for analysing the signal's structure.CSMP Algorithm has the capability to identify all the hidden periods from 1 to the maximum hidden period Q of a signal of any length, without the need of truncating it and accounts for low computational cost.

2.3 Maximum Likelihood Detection Technique(MLDT)

We have particular examined the Maximum Likelihood technique for periodic detection with respect to large DNA sequence. When Maximum Likelihood approach is employed on DNA sequence, it is assumed that each symbol present in a DNA has been generated from Information source, wherein there is a certain Probability

associated with each symbol that contributes to its Entropy. The number of sources signifies the periodicity of the sequence. Since the Probability of occurrence is not known to us, we use Maximum Likelihood Estimation to chart a statistical model for defining Probability mass function of the sources as well as period of the sequence. While understating any method/model which involve Probability, the it is necessary to understand the concept of Statistical Periodicity, which forms the basis of MLDT. If $D=[D_1,D_2,.....D_N]$ is the input DNA sequence, then it is mapped to DNA alphabet $S=[A,T,G,C]$ under the mapping $D:N\rightarrow S$.With an assumption of statistical/hidden period '*T*' in the input sequence D, the number of input random variables are denoted as $X_1,X_2,.....X_T$. Each of these random variable takes a value $\in S$, according to a certain Probability Mass Function(PMF) denoted by P_i. It is the P_i which is the sole purpose of MLDT. There are various approaches and models to describe the above stated PMF, each being presumptuous in nature. The method starts with a defining a following parameter,

$$\hat{\Theta}=\arg\max_{\Theta} P(D|\Theta) \tag{2.10}$$

which is an approximation of period *T*. MLDT techniques aims to identify the unknown quantity Θ. The input data sequence $D=[D_1,D_2,.....D_N]$, is made to be represented in the form of vectors given by $W=[w_1,...,w_N]$ where each w_i is an $|S|\times 1$ vector with w_{ij} = 1, where $|S|$ cardinality of the alphabet S. Also, a stochastic matrix A of order $|S|\times T$ is defined , with $A_{ij}=P(X_i=S_j)$, where *i* defines pmf's of the information sources and A_{ij} denotes the probability of j^{TH} symbol being generated by i^{TH} source.

For a fixed T the maximum likelihood estimate for A is denoted as

$$A^T=\arg\max_{A}\log P(W|A,T) \tag{2.11}$$

$\log P(W|A,T)$ is given by the following equation

$$\log P(W \mid A,T) = \sum_{i=1}^{N-MT} \sum_{j=1}^{|S|} w_{ji} \log(\mathrm{A_{ij}}) \tag{2.12}$$

With the help of the above two equations the hidden period *T* can be approximated as

$$T^{*} = \arg\max_{T} \log P(W \mid \mathrm{A}^{\mathrm{T}}, T) \tag{2.13}$$

This method eliminates the need of transforming the symbol into numerical sequences and the method has the ability to find the homological, latent and eroded Periodicities effectively.

2.4 Ramanujan Fourier Transform

One of the greatest Indian Mathematician of all time, Dr. Srinivas Ramanujan introduced Ramanujan sum [1], denoted by $\mathrm{c_q(n)}$ and defined in the following equation.

$$c_q(n) = \sum_{\substack{P=1 \\ \gcd(P,q)=1}}^{q} \exp(\frac{j2\pi Pn}{q}) \tag{2.14}$$

Where (P, q) are the co-primes.

In the late 90's, Ramanujan sum have found a great contribution in Signal processing tools and number theory which has further led to formulation of various Periodicity mining tools.

The motivation behind formulation of this sum was to be able to express various standard arithmetic functions into linear combinations of Ramanujan sum, i.e.

$$x(n) = \sum_{q-1}^{\infty} \alpha_q c_q(n) \qquad \mathrm{n} \geq 1 \tag{2.15}$$

Where α_q is the Ramanujan Series coefficient for $\mathrm{q} \to \infty$,But when limited to finite q, is called Ramanujan Fourier Transform [7]. Planat in *et.al* [7] has shown a generalized arithmetic expression for α_q given as

$$\alpha_q = \frac{1}{\phi(q)} \sum_{n=1}^{N} x[n] c_q(n) \tag{2.16}$$

Where, $\phi(q)$ is the Euler-totient function, which gives the sum of divisors of the enclosed quantity which is q here. The Ramanujan Fourier Transform(RFT) gives a spectrum varying with q, thereby providing measurements and analysis of the periodic behaviour. However, the major disadvantage of RFT is its time sensitive nature. The spectrum is seen to be adversely affected by phase and time shifts

2.5 Ramanujan Filter Banks-

Tenneti and Vaidyanathan proposed the Ramanujan Filter Bank (RFB) based on the Ramanujan subspaces to identify all the hidden periods of a signal. This method is a one step ahead of Dictionary based method of periodicity detection and overcomes the drawbacks of Discrete Fourier Transform as it is capable of determining hidden periodicities in Data whose behaviour varies significantly with time. This includes signals that are periodic only for a short duration and signals such as chirps. For such signal, time vs Period plane is used which is analogous to the traditionally used time vs frequency plane.

P.P. Vaidyanathan first introduced periodicity matrix **A** .**"**A very important property of **A** is, if a discrete vector **x** of length **P,** can be made to be expressed as **x=Ay , y** being the matrix satisfying this relation then, Then the lcm of the periods of all those columns of A that are multiplied by non-zero entries in **y** is exactly equal to the period of the signal x(n)**"** .

This property undoubtfully played an important role in periodic measurements. But this technique is found to be useful only if we want to find the periodicity in a certain ν_P as concluded, and will unfortunately fail if ν_P is not known to us. To overcome this, a method aiming at combining different periodicity matrix in dictionary based approach was introduced. Vaidyanathan also concluded **y** as

$$y=D^{-2}A^{T}(AD^{-2}A^{T})^{-1}x \qquad (2.17)$$

where, using a simple convex program with close form solution.

Vaidyanathan, introduced Filter Banks on foresaid matrix P based on its interesting row pattern. He noticed that left-inverse seem to be periodic, with exactly φ(P) rows with period P, which was similar to the Dictionary itself. Thereby, concluded that

following expression produced equally fair results as (1) and could be easily implemented using Filter Banks Approach.

$$y = D^{-1}A^{T}x \tag{2.18}$$

Successive N-sized input blocks, each shifted by one sample, were employed to process a particular input signal. The final Filter Bank output is given by the following equation-

$$y_P = \sum_{i = n-\Phi(P)+1}^{n} \left| \frac{(x * h_P)(n)}{f(P)} \right|^2 \tag{2.19}$$

Where $h_P = \{c_P(0)\ \ c_P(1)\ \ c_P(2). . . .\ \ c_P(LP-1)\}$, $c_P(n)$ is the P^{th} Ramanujan Sum and $f(P)$ is the Penalty Function. Ramanujan Filter eliminates the drawback of Dictionary based approaches where Period size is very fat and its ability to find Period only when ν_P is known to us. However complex design structure of these filter banks cause undesired overlapping of the signal component, resulting in some false hidden periods

2.6 Summary of Periodicity Mining Tools

Table 2.1: Comparison between different Periodicity Mining Signal Processing Tools

S.no	Signal Processing Tools	Advantages	Disadvantages
1	Maximum Likelihood Periodicity Detection Technique	Ability to find homological, latent and eroded Periodicities effectively.	Does not perform well in identification of ncRNAs because most of them preserve their secondary structures.
2	Ramanujan Fourier Transform	Provides for a spectrum which accounts for the hidden periodicities	Ambiguous Physic Significance,Time shift Senstivity.
3	Ramanujan Filter Banks	Eliminates the drawback of Dictionary based approaches having fat Period size and its ability to find Period only in region ν_P.	Complex Structures leading to overlapping Periodicities.
4	Conjugate Subspace Matching Persuit Algorithm	Ability to detect not only the hidden periods, but also the changes of period with time, Low computational cost.	Complex Mathematical Structure.
5	Intrinsic Integer Periodic Functions	Eliminates Time and phase Sensitivity, simple.	High Computational cost

CHAPTER 3
STUDY OF IIPF FOR PERIODICITY DETECTION

3.1 Periodicity Present in Signal

Most real-world signals do not possess strict periodicity. There are some partial periodic patterns, and hidden periodicity in a signal, due to multiple periodic components. These signals cannot be easily analysed by mere observation. The main objective of this project is the study of discrete periodicity. As discrete signals are indexed by integer numbers, a strict period of a discrete signal must be an integer. This results in differences between periodic identification of a discrete signal from that of a continuous signal.

3.2 Intrinsic Integer-Periodic Function (IIPF)

To put into simpler words IIPF is a class of functions, into which, if a discrete integer signal, can be made made to split into, its Periodic behaviour can be analysed efficiently. The function forms the orthogonal component of a discrete signal.

Let us consider a signal

$$x[n]=\overline{[6,-6,6,-6,6,-4]} \tag{3.1}$$

where the top bar means signifies the expansion of the specified six numbers (x[1] to x[6]) in a periodic manner consisting an infinite-length 6-periodic sequence. Hence in a strict sense, the period of x[n] is 6 as it repeats after every 6 samples. However, x[n] seems to oscillate between 6 and -6: it is very close to

$$\overline{[6,-6]}= \overline{[6,-6,6,-6,6,-6]}. \tag{3.2}$$

An different and more logical way to represent x[n] is,

$$x[n]= \overline{[6,-6]} + \overline{[0,0,0,0,0,-4]} \tag{3.3}$$

So, we can see that the signal gets decomposed into mutually orthogonal and periodic functions.

A mathematical model of IIPF decompositions can be achieved with the help of Ramanujan Subspaces, which altogether gives the information of hidden periodic components. The form of IIPF decomposition of an arbitrary -periodic signal x[n] is

$$x[n] = \sum_{d|q} x_d[n] \tag{3.4}$$

where $x_d[n]$ is a d –IIPF.The summation holds true for every divisor d of q. $x_d[n]$ can be found as-

$$x_d[n] = \frac{1}{q}(x \otimes c_d) \tag{3.5}$$

$$c_d(n) = \sum_{\substack{k=1 \\ \gcd(k,d)=1}}^{d} exp\left(\frac{j2\pi kn}{d}\right) \tag{3.6}$$

Where $c_d(n)$ is Ramanujan Summation and forms the basis of IIPF.

3.3 IIPF Spectrum

Mathematically, the expression of IIPF Spectrum is –

$$y_q[r] = \frac{1}{N}\sum_{n=1}^{N} y[n]c_q[n - r + 1] \tag{3.7}$$

The IIPF spectrum is an important tool, that gives the hidden periods in a discrete signal. They are less prone time sensitivity variations unlike Ramanujan fourier Transform. Where $y[n]$ is the input signal, N is the length of the signal. However for better graphical view, Norm $\|y_q\|$ of the above expression plotted against q.

3.4 Results and Discussion

For easy understanding we have taken following examples:

3.4.1 Example 1: $x[n] = \cos\left(\frac{2\pi n}{6}\right) + 3\sin\left(\frac{2\pi n}{14}\right)$

signal x[n] consists of periods 6 and 14,

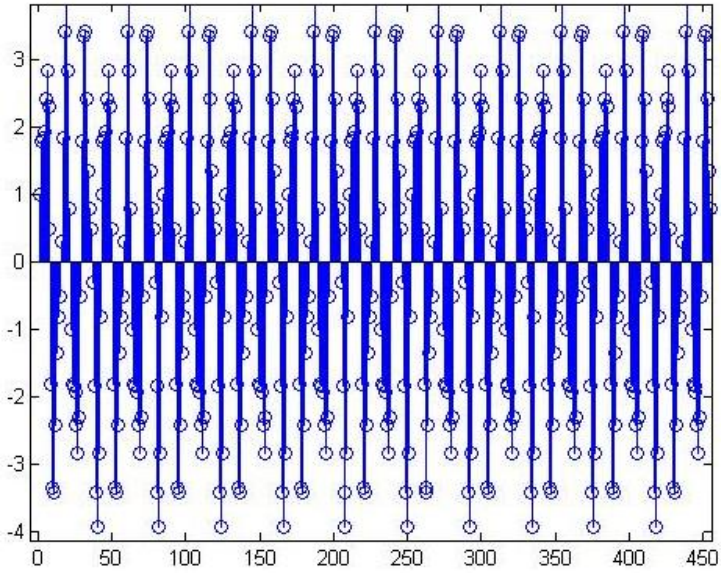


Figure3.1:x[n]

From the figure of x[n], it is very difficult to observe the period 6 and 14 individual, although you can easily identify the total period that is *lcm*(6,14). Now, we plot its IIPF spectrum, which gives us the following plot,

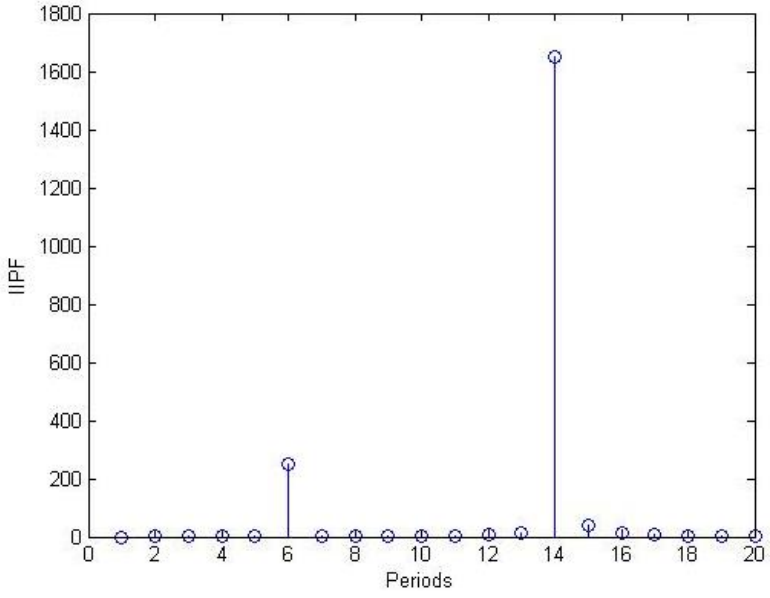


Figure3.2:IIPF of x[n]

From above plot we can see the peaks occurring at 6 and 14, signifying the presence of these periods in the sequence.

3.4.2 Example 2

Now, we take another example to address an important aspect of Short Time IIPF. It is known that some signals exhibit periodic nature whose periodicity varies with time, such as in a speed, music and in an inverse chirp signal. The periods of these signals can change with time or be present for a short duration. So, we take an inverse Chirp signal x(t) to demonstrate its periodic nature.

$$x(t) = \sin\left(\frac{1}{at}\right),\ t \in [2,10] \tag{3.8}$$

Where a=0.01/2π. The discrete chirp is obtained by sampling x(t) at every 0.01 seconds.

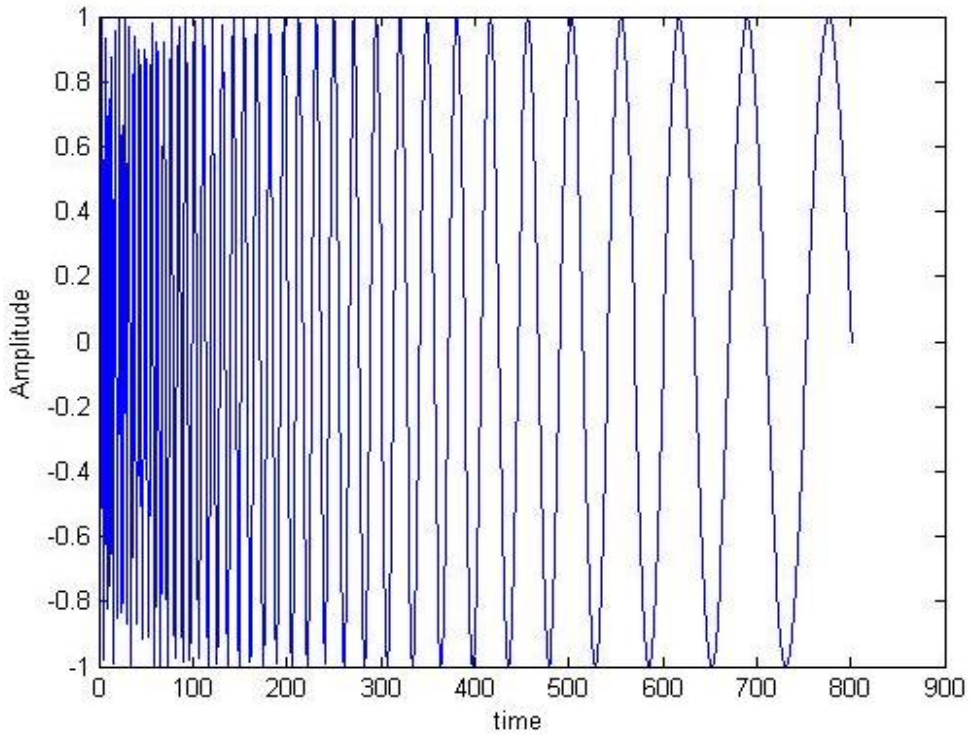


Figure 3.3: Inverse Chirp Signal

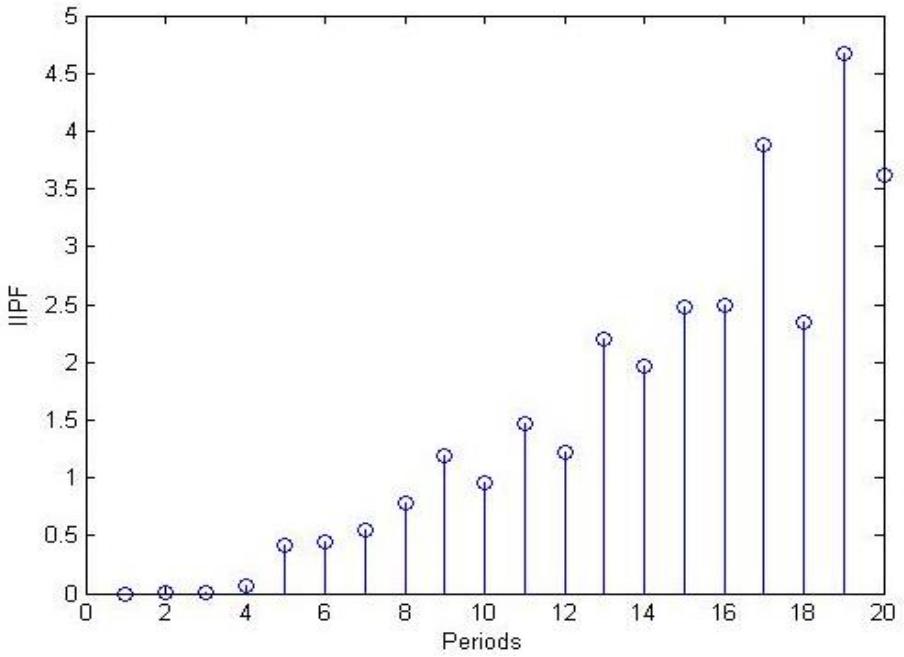


Figure 3.4: IIPF of Inverse Chirp Signal

In the IIPF of inverse chirp signal, we can see the periodic elements changing with time, which justifies the property of inverse chirp signals.

CHAPTER 4

PERIODICITY DETECTION IN DNA USING SHORT TIME IIPF

4.1 DNA Sequence

The genome comprises of the deoxyribose nucleic acid (DNA). There are two DNA strands which are termed as polynucleotides. They are composed of simpler monomer units termed as nucleotides. Each nucleotide is composed of one of the nitrogen containing nucleobases cytosine (C), guanine (G), adenine (A),or thymine (T) as well as a sugar called deoxyribose, and a phosphate group. These nucleotides are joined to one another by covalent bonds in chain like structure, between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. The polynucleotide strands have their nitrogenous bases bound together , according to base pairing rules (A with T, and C with G), with hydrogen bonds to make double-stranded DNA. The tandem repeat patterns are classified in different types subject to range of base pairs. Tandem repeat pattern of the nucleotides 1-6bp is called short tandem repeats (STRs) and are categorised into perfect and imperfect STRs. The perfect STRs consists of exact copies of the repeated patterns while the imperfect STRs consists inexact copies of the repeated patterns which arises due to insertion, deletion and substitution.

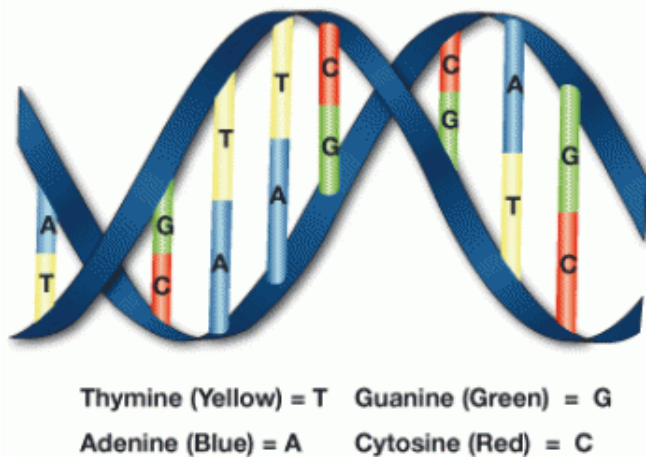


Figure4.1:DNA Sequence

4.2 Short Time IIPF For Repeat Detection(ST-IIPF)

The Short time IIPF is one step ahead of standard IIPF. The ST-IIPF is performed iteratively in two stages.

Stage 1- In the first stage, the DNA sequence is passed to moving rectangular window of Unity gain, whose size shall remain fixed. Theory suggests that window size is should approximately be the double of the number of maximum periods, one wishes to measure upto. Mathematically, if x[n] is the input signal of length N and w[n] is the rectangular window of length L,

$w[n]'=\sum_{i=1}^{N-L-1} w[i]$ -moving rectangular window (4.1)

$x[n]'=x[n]* w[n]'$ -passing the signal through rectangular window (4.2)

Stage 2- In the second stage, the IIPF is applied to the sequence obtained for every position of the rectangular window, where the window is being moved by one position at every iteration of the loop. Mathematically, First, we define Ramanujan sum as follows:

$$C_p = \sum_{\substack{k=1 \\ \gcd(k,p)=1}}^{p} \exp(\frac{j2\Pi kn}{p}) \tag{4.3}$$

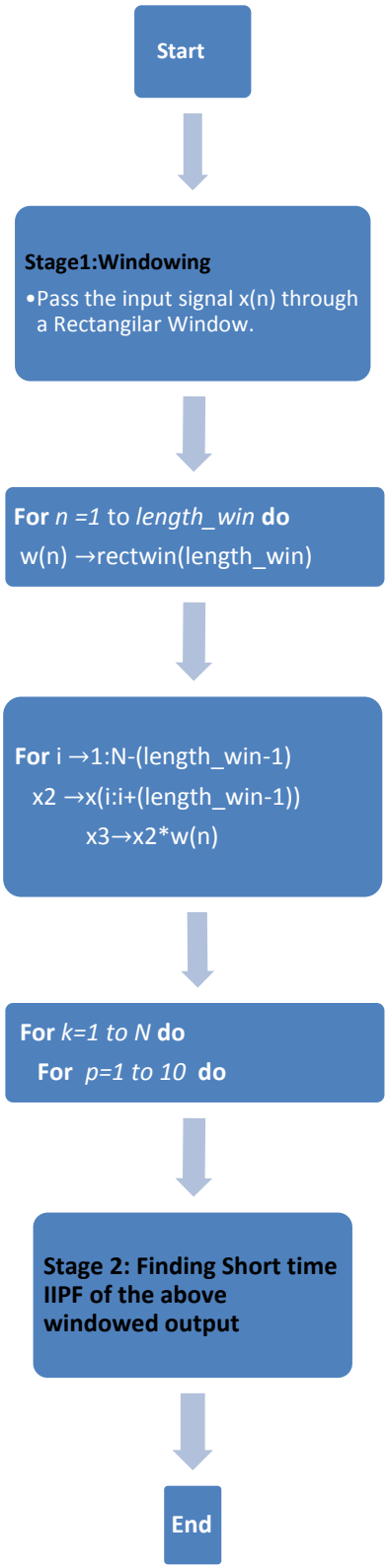
Then, we calculate IIPF Spectrum as follows:

$$y_p(k) = \frac{1}{N}\sum_{n-1}^{N} x(n)C_p(n-k+1) \tag{4.4}$$

At last, we compute the norm of the IIPF spectrum, which is given by:

$$\gamma[p] = \sqrt{\frac{1}{N}\sum_{n=1}^{N} y_p[n]^2} \tag{4.5}$$

4.3 Algorithm for Repeat Detection in DNA



4.4 Results and Discussion

4.4.1 Synthetic Data

x(n)=['ATATATATATATATATATATCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGA
TGCATGCATGCATGCATGCATGCATGCATGCATGCATGCCGGGACGGGACGGGAC
GGGACGGGACGGGACGGGACGGGACGGGACGGGATACTTTTACTTTTACTTTTACT
TTTACTTTTACTTTTACTTTTACTTTTACTTTTACTTTGGGCTTCGGGCTTCGGGCTTCG
GGCTTCGGGCTTCGGGCTTCGGGCTTCGGGCTTCGGGCTTCGGGCTTC']

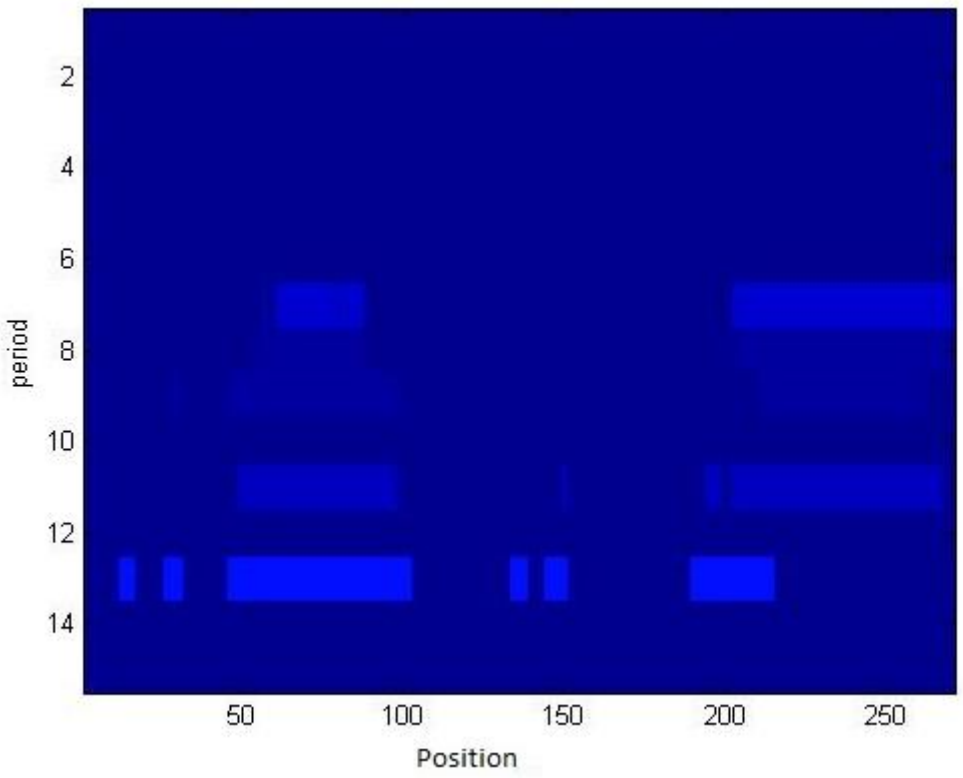


Figure 4.2:Periods present in Synthetic data

In the figure 4.2, the highlighted Blue regions, signifies the periods and their respective position in the synthetic sequence taken into consideration. There is also some overlapping periods a same locations, the brightness of Blue colour signifies the intensity level of these periods.

4.4.2 Real Data

To demonstrate the Short time IIPF for a real data, we have picked the sequence U73920 that is, Apis mellifera (Honey Bee) microsatellite repeat region.

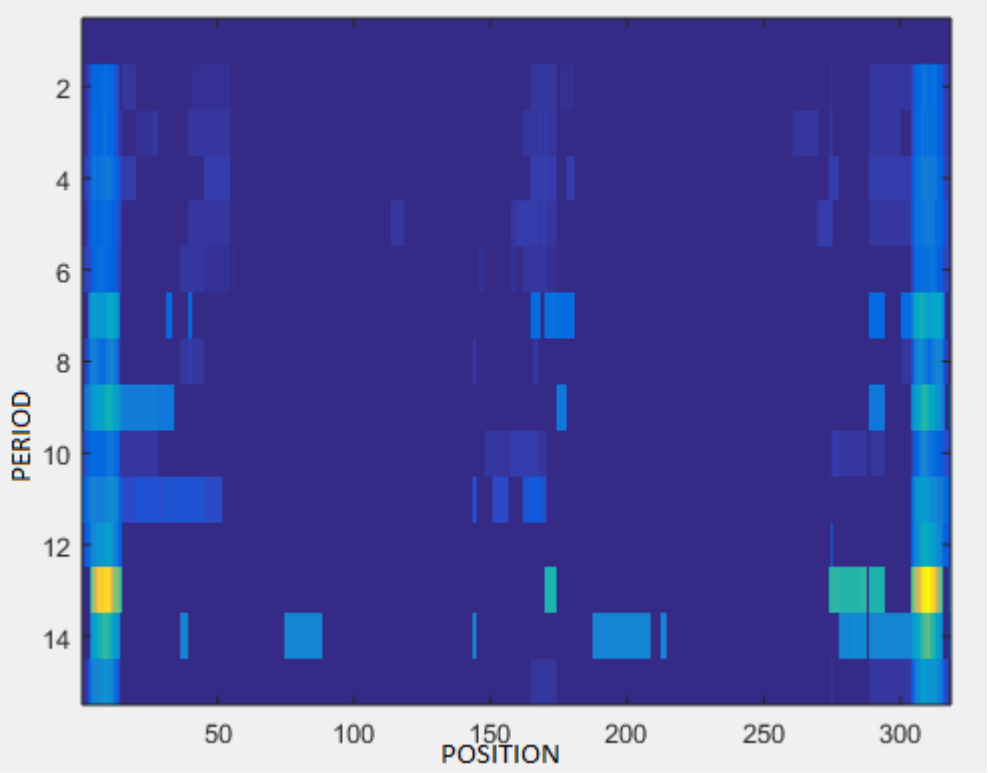


Figure 4.3 Periods present in U73920

Table 4.1: Periods present in U73920

INDEX	PERIOD	Nucleotide Position
1	7	147-169 417-433 450-466 512-535 552-573
2	9	194-223 389-417 540-570 750-767
3	11	512-573 633-655 713-768
4	12	532-571
5	13	257-364
6	14	246-381 643-692

In Figure 4.2 and Figure 4.3, we can see the various periods and their positions present in U73920. These periods and their positions are being highlighted with the different colour spectrum, indicating the intensity level of periods. For instance, the yellow colour in period 7 in figure 4.3, signifies the intensity of that period.

CHAPTER 5
CONCLUSION AND FUTURE WORK

In this project we were successfully able to apply our proposed algorithm in real and synthetic DNA sequences as well as in Inverse chirp signal, with satisfactory results. The result shown above can be easily comprehended along with period and nucleotide position highlighted in the results. There is a certain Computational cost associated with the project, which can be worked upon, so that operation becomes less time consuming.

REFERENCES

[1] S. C. Pei, and K. S. Lu, "Intrinsic integer-periodic functions for discrete periodicity detection," IEEE Signal Processing Letters., vol. 22, no. 8, pp. 1108–1112, 2015

[2] P. P. Vaidyanathan, "Ramanujan-sum expansions for finite duration (FIR) sequences," inProc. IEEE Int. Conf. Acoust. Speech, and Signal Proc., 2014.

[3] P. P. Vaidyanathan, "Ramanujan sums in the context of signal processing; part i: Fundamentals,"IEEE Trans. Signal Process., vol. 62,no. 16, pp. 4145–4157, Aug. 2014.

[4] P. P. Vaidyanathan, "Ramanujan sums in the context of signal processing; part ii: Fir representations and applications," IEEE Trans.Signal Process., vol. 62, no. 16, pp. 4158–4172, Aug. 2014.

[5] "Signal periodic decomposition with conjugate subspaces" Shi-wen Deng*, Ji-qing Han*, Member, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. X, NO. X, XX 2016

[6] P. P. Vaidyanathan and S. Tenneti, "Properties of Ramanujan filterbanks," in Proc Signal Processing Conference (EUSIPCO), 2015, pp. 2816–2820.

[7] S.D Sharma, Rajiv Saxena, S.N.Sharma, A.K.Singh (2015), "Short Tandem Repeats Detection in DNA Sequences Using modified S-Transform,"International Journal of Advances in Engineering and Technology, vol.8.issue2, April, 2015.

[8] M. Nakashizuka, H. Okumura, and Y. Iiguni, "A sparse periodic decomposition and its application to speech representation," in Proc. 16th European Signal Processing Conference, 2008, pp. 1–5.

[9] L. Toth and P. Haukkanen, "The discrete fourier transform of r-even functions,"ActaUniv.SapientiaeMath.,vol.3,no.1,pp.5–25,2011.

[10] M. Planat, H. C. Rosu, and S. Perrine, "Ramanujan sums for signal processing of low frequency noise," Phys. Rev. E, vol. 66, p. 056128, 2002.

[11] P. Vaidyanathan and P. Pal, "The Farey-dictionary for sparse representation of periodic signals," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 360–364.

[12] A.V.Oppenheim,R.W.Schafer,andJ.R.Buck,Discrete-Time Signal Processing, 2nd Ed.ed.Upper Saddle River, NJ,USA:Prentice-Hall,1999.

[13] J.H.McClellanandT.W.Parks,"Eigenvalueandeigenvectordecomposition of the discrete fourier transform," IEEE Trans. Audio Electroacoust., vol. AU-20, no. 1, pp. 66–74, Mar. 1972

[14] S. V. Tenneti, and P. P. Vaidyanathan, "Nested Periodic Matrices and Dictionaries: New Signal Representations For Period Estimation," IEEE Transactions on Signal Processing, vol. 63, no. 14, pp. 1–1, 2015.

[15] P. P. Vaidyanathan, "Multidimensional Ramanujan-sum expansions on non separable lattices." Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 3666–3670.

[16] L. Sugavaneswaran, S. Xie, K. Umapathy, and S. Krishnon, "Timefrequency analysis via ramanujan sums," IEEE Signal Process. Lett., vol. 19, no. 6, pp. 352–355, 2012.

[17] L.T.Mainardi,L.Pattini,andS.Cerutti,"Application of the ramanujan fourier transform for the analysis of secondary structure content in amino acid sequences," Meth. Inf. Med., vol. 46, no. 2, pp. 126–129, 2007.

[18] M. M. Goodwin and M. Vetterli, "Matching pursuit and atomic signal models based on recursive filter banks," IEEE Transactions on Signal Processing, vol. 47, no. 7, pp. 1890-1902, 1999.

[19] S.-W. Deng, and J.-Q. Han, "Voice activity detection based on conjugate subspace matching pursuit and likelihood ratio test," Eurasip Journal on Audio Speech & Music Processing, vol. 2011, no. 1, pp. 1–12, 2011.

[20] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dic tionaries," IEEE Trans. on Signal Process., vol. 41, no. 12, pp. 3397– 3415, 1993.