

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES

AUDITORY MOTIVATED DISCRETE TIME
FREQUENCY SIGNAL REPRESENTATION AND
ITS APPLICATION TO VOWEL
CLASSIFICATION

by
Bertan KARAHODA

July, 2012
İZMİR

**AUDITORY MOTIVATED DISCRETE TIME
FREQUENCY SIGNAL REPRESENTATION AND
ITS APPLICATION TO VOWEL
CLASSIFICATION**

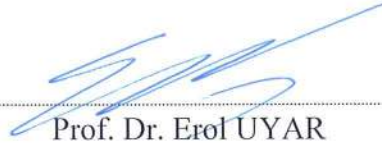
**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy
in Mechanical Engineering, Machine Theory and Dynamics Program**

**by
Bertan KARAHODA**

**July, 2012
İZMİR**

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**AUDITORY MOTIVATED DISCRETE TIME FREQUENCY SIGNAL REPRESENTATION AND ITS APPLICATION TO VOWEL CLASSIFICATION**” completed by **BERTAN KARAHODA** under supervision of **Prof.Dr. EROL UYAR** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.



Prof. Dr. Erol UYAR

Supervisor



Doç. Dr. Zeki KIRAL

Thesis Committee Member



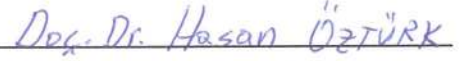
Yrd. Doç. Dr. Ahmet ÖZKURT

Thesis Committee Member



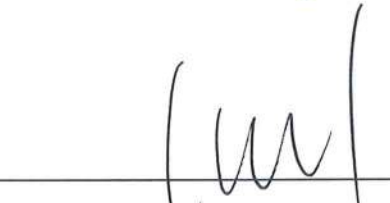
Prof. Dr. Semih BILGEN

Examining Committee Member



Doç. Dr. Hasan ÖZTÜRK

Examining Committee Member



Prof. Dr. Mustafa SABUNCU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I am greatly indebted to my thesis supervisor Prof. Dr. Erol UYAR for his help, kind interest and encouragement throughout the development of this study.

I would like to thank to thesis committee members Assistant Prof. Dr. Ahmet ÖZKURT and Associate Prof. Dr. Zeki KIRAL for their useful and helpful directives on the completion of this thesis work. I am also thankful to my family for their moral support.

Bertan KARAHODA

AUDITORY MOTIVATED DISCRETE TIME FREQUENCY SIGNAL REPRESENTATION AND ITS APPLICATION TO VOWEL CLASSIFICATION

ABSTRACT

In this thesis the Auditory Motivated Discrete Time Frequency Signal Representation method is presented. The method is simple and independent from the window function, which affect the obtained time frequency resolution in classical methods. The numerical simulations with different SNR values show that the proposed method is applicable for time frequency signal analysis. The proposed method is applied to the speech vowel signals and similar spectral shapes are obtained from the same vowel signals independent from the speakers, which is good evidence for existing similar spectral shapes inside the same vowels. The vowel classification based on vowel patterns extracted from spectral peaks distribution is performed in order to test the existence of the similar spectral shapes, and the obtained results show that the proposed method can be used to extract additional vowel patterns in speech recognition applications to improve the speech recognition performance.

Keywords : Vowel classification, spectral envelope, discrete time frequency, time frequency resolution, basilar membrane vibration, human ear.

İNSAN KULAĞI YAPISINA DAYALI AYRIK ZAMAN FREKANS SİNYAL TEMSİLİ VE SESLİ HARF SINIFLANDIRILMASINDA UYGULAMASI

ÖZ

Bu tez çalışmasında insan kulağının yapısı ve çalışmasından esinlenerek geliştirilen yeni bir ayrik zaman frekans sinyal analizi yöntemi sunulmuştur. Geliştirilen yöntem, klasik zaman frekans sinyal analizi yöntemlerinde kullanılan ve zaman frekans çözünürlüğünü etkileyen `pencere` fonksiyonundan bağımsızdır. Değişik SNR değerleriyle yapılan sayısal simülasyonların sonuçları geliştirilen yöntemin zaman frekans sinyal analizinde uygulanabilirliğini göstermektedir. Sunulan yöntem sesli harflere uygulanmış ve konuşmacıdan bağımsız aynı sesli harflerden benzer spektral şekiller elde edilmiştir, ve elde edilen sonuçlar aynı sesli harfler içinde benzer spektral şekiller olabileceğinin iyi bir kanıtıdır. Benzer spektral şekillerin varlığını test etmek için spektral tepelerin dağılımından elde edilen sesli harf örüntüleri sesleri sınıflandırmak için kullanılmıştır ve elde edilen sonuçlar geliştirilen yöntemin ses tanıma uygulamalarında tanıma başarısını arttırmak amaçlı yardımcı ses örüntüleri olarak kullanılabileceğini göstermektedir.

Anahtar sözcükler : Ünlü harf ses sınıflandırma, spektral dağılım, ayrik zaman frekans, zaman frekans çözünürlüğü, bazal membran titreşimi, insan kulağı.

ABBREVIATIONS

AMTFR – Auditory Motivated Discrete Time Frequency Signal Representation

ANN – Artificial Neural Networks

ASR – Automatic Speech Recognition

ASTFT – Adaptive Short Time Fourier Transform

DFT – Discrete Fourier Transform

FT – Fourier Transform

HMM – Hidden Markov Model

IF – Instantaneous Frequency

MFCC – Mel Frequency Cepstral Coefficients

SNR – Signal to Noise Ratio

SPD – Spectral Peaks Distribution

STFT – Short Time Fourier Transform

WT – Wavelet Transform

WVD – Wigner-Ville Distribution

CONTENTS

	Page
THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZ.....	v
ABBREVIATIONS	vi
CHAPTER ONE – INTRODUCTION.....	1
CHAPTER TWO – HUMAN EAR SOUND TRANSDUCTION	4
2.1 The Structure of the Human Ear	4
2.1.1 The Outer Ear.....	4
2.1.2 The Middle Ear	5
2.1.3 The Inner Ear (Cochlea).....	5
2.2 The Conduction of Sound into Basilar Membrane Vibrations	6
2.3 The Frequency Selectivity of Basilar Membrane	9
CHAPTER THREE – BIOLOGICALLY INSPIRED DISCRETE TIME FREQUENCY SIGNAL ANALYSIS METHOD	12
3.1 The Brief Overview of Time Frequency Signal Representations.....	12
3.1.1 The Effect of Windowing	14
3.2 The Auditory Motivated Discrete Time Frequency Signal Representation	17
3.2.1 Signal Reconstruction.....	21
3.2.2 Numerical Simulations	25
3.2.3 Auditory Motivated Discrete Time Frequency Signal Analysis Method	32

CHAPTER FOUR – SPEECH VOWEL CLASSIFICATION BY USING AUDITORY MOTIVATED DISCRETE TIME-FREQUENCY SIGNAL ANALYSIS METHOD.....	40
4.1 The Mel Frequency Cepstral Coefficients	40
4.2 Application of Auditory Motivated Discrete Time Frequency Signal Analysis Method to the Vowel Speech Signals	43
4.3 The Auditory Motivated Discrete Time Frequency Signal Analysis Method based Vowel Classification.....	68
CHAPTER FIVE – CONCLUSIONS.....	72
REFERENCES	74

CHAPTER ONE

INTRODUCTION

There are a lot of types of signals existing in nature and the signals can be classified into different classes based on their characteristics. One classification of the signals is deterministic or random signals, and the random signals have characteristics varying over time which lead to another classification of non-stationary signals (Umapathy et al.,2010).

The frequency content of signal is powerful basis for signal analyzing applications. The Fourier Transform (FT) is the popular choice for obtaining the frequency content of the signals. The FT gives the overall frequency content of the signal, and the time information is lost because the FT transform is performed over all signal duration. Therefore the FT is powerful analyzing tool for stationary signals which have the characteristics that do not change with time.

For non-stationary signals, the occurrence times of the frequencies are important because these signals have characteristics that change with time. In order to obtain the time information the Short Time Fourier Transform (STFT) is used where the fixed width window function is introduced to the FT. The STFT assumes the signals stationarity for the specific duration in time defined by the fixed window width. Therefore, for good time frequency resolutions the STFT fails in the case of non-stationary signals.

To overcome the resolution problems the multiresolution signal analysis methods are widely used. The Wavelet Transform (WT) is a powerful time frequency signal representation tool for non-stationary signals. The WT uses the variable window width for multiresolution signal analysis. The narrow window gives the good time resolution but the frequency resolution is poor, and the large window give good frequency resolution but the time information is poor (Mertins, 1999).

The above mentioned signal analysis methods are widely used by many researchers. The STFT, WT, Wigner-Ville Distributions (WVD) (Wang&Jiang, 2010) and time frequency representations based on the time frequency dictionaries (Umapathy et al.,2010) are popular choices used for time frequency signal representations. Each of these methods uses different kernels to obtain better time frequency resolutions for the analyzed signals. According to Zhong&Huang (2010) most of the scientists believe that there is no single kernel that matches best time frequency resolution for all signal types. Zhong&Huang (2010) introduced the Adaptive Short Time Fourier Transform (ASTFT) where the window width of the STFT is set according to the Instantaneous Frequency (IF) detected from the ridge of the WT. The deconvolutive STFT spectrogram (Lu&Zhang, 2009), time frequency resolutions based on Ramanujan Sums (Sugavaneswaran et al., 2012) were used to obtain better time frequency representation for the specific class of signals. According to the Heisenbergs Uncertainty Principle (Loughlin&Cohen, 2004) the time and frequency resolutions cannot be optimized at the same time, the time and frequency resolutions satisfy the condition $\Delta t \Delta f \geq 1/4\pi$ where the minimal value of $\Delta t \Delta f$ is called the Heisenberg box (Zhong&Huang, 2010).

In this thesis the Auditory Motivated Discrete Time Frequency Representation (AMTFR) method is presented. AMTFR gives the time frequency representation without the use of any windowing function. The function of the inner hair cells in the human auditory system is tried to be simulated under some assumptions. The method is simple and window independent. The numerical simulations show the effectiveness of the AMTFR. The performance of the method is tested under noisy conditions.

The speech signals fall into non-stationary signals class. Therefore the time frequency resolutions play important role for speech signal analysis. The human brain is still superior to many technical solutions. Therefore the human auditory system based feature extraction methods for automatic speech recognition (ASR) systems were widely used in literature. The Mel Frequency Cepstral Coefficients (MFCC) described in Picone (1993) is the most popular method used in this area.

Chatterjee&Kleijn (2011) give the auditory based design and optimization of feature vectors for ASR. The vowel patterns obtained from MFCC are explained well in Dusan (2007). The MFCC try to capture the frequency selectivity of human auditory system by transferring the normal frequency scale to the so called mel-scale.

Different hearing tests with subjects who have the normal hearing ability were performed in the literature to determine which of the formant frequencies are important for vowel identification (Zahorian&Zhang, 1992; Shannon et al., 1995; Sakayori et al., 2002). The formant frequencies F_1 and F_2 are important to identify the vowels (Sakayori et al., 2002). However, for the same speaker the formant frequencies F_1 and F_2 can be used to identify the vowels, but in the case of multiple speakers there exist overlap between the formant frequencies.

Zahorian&Zhang (1992) suggested that spectral envelope is important information for vowel identification. Zahorian&Jagharghi (1993) showed that computational vowel classification based on spectral envelope is superior to the information on the F_0 , F_1 , F_2 and F_3 . According to Sakayori et al. (2002) the human auditory system may identify the vowels according to the spectral shapes and formant frequencies F_1 and F_2 in the critical spectral regions.

The proposed method is applied to the vowel signals and similar spectral shapes at higher frequencies are obtained for the same vowels independent from the speakers, which suggest that spectral envelopes can be used as external cues for the classification of the vowels independent from the speakers. The proposed method is simple and independent from the window function.

The thesis is organized as follows. In chapter two the basics of sound transduction inside the human ear will be explained. In chapter three the novel discrete time-frequency signal representation method will be presented. Chapter four presents the vowel classification algorithm which is based on the spectral peaks distribution obtained from discrete time-frequency signal analysis method. In chapter five the conclusions will be given.

CHAPTER TWO

HUMAN EAR SOUND TRANSDUCTION

2.1 The Structure of the Human Ear

The human ear consists of three parts: Outer Ear, Middle Ear and Inner Ear. The schematic drawing of the human ear is given in Figure 2.1.

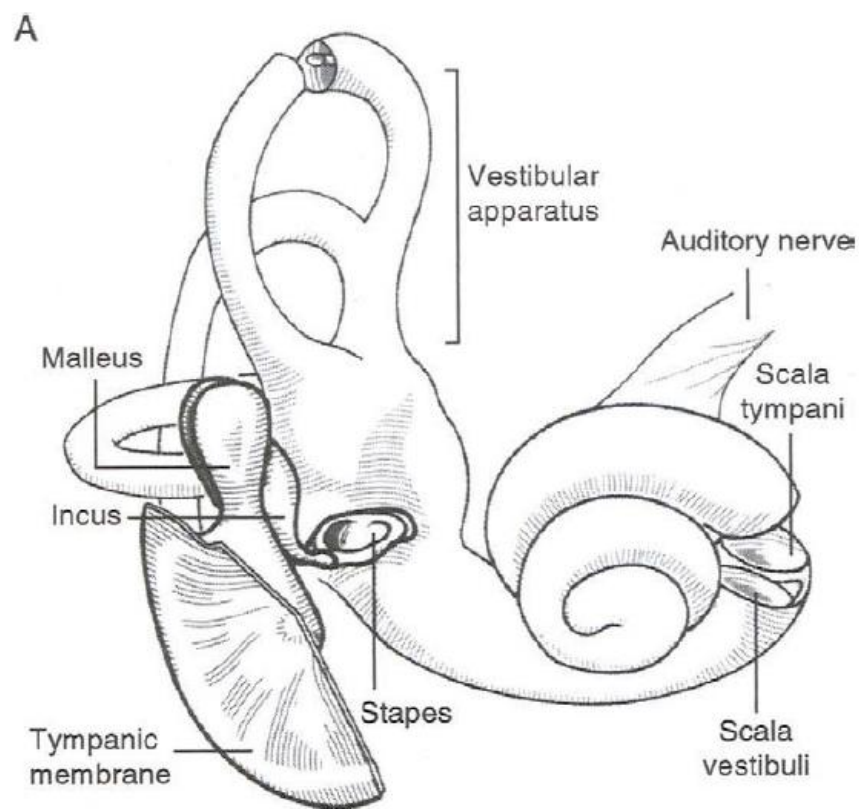


Figure 2.1 Schematic drawing of the human ear (From: Moller, A.R., “Anatomy and Physiology of sensory organs”, chapter two, pp. 38, Sensory Systems, Elsevier Inc. 2003, with permission)

2.1.1 The Outer Ear

The Outer Ear is the only part of the human ear that can be seen from the outside. The Outer Ear consists of the pinna and ear canal and the outer ear modify the spectrum of the sound according to the sound source (Moller,2003b).

2.1.2 *The Middle Ear*

The middle ear consists of tympanic membrane and three small bones (ossicles): the malleus, incus and stapes. These bones form a chain structure which is known as ossicle chain. The tympanic membrane conducts the sound vibrations to the vibrations of these bones (Moller,2003b). The footplate of stapes is located in the oval window, which is one of the openings of the cochlear structure (inner ear). The other opening of the cochlear structure is the round window. The vibrations of the stapes according to the sound waves put the fluid inside the cochlea to the motion. When oval window moves inward the round window moves outward, which allow the fluid inside the cochlea, which has rigid structure, to vibrate according to the sound waves, therefore middle ear acts as an impedance transformer which improves the sound transmission from middle ear to the inner ear (Moller,2003a).

2.1.3 *The Inner Ear (Cochlea)*

The cochlea is fluid filled and has snail-shaped structure (Figure 2.1) Its length is approximately 3.5cm. In humans the cochlea has 2.25 turns. The Figure 2.2 shows the cross section of the guinea pig cochlea.

The fluid filled structure of the cochlea is divided longitudinally into three parts: scala vestibuli, scala tympani and in the middle scala media. The scala media is separated from scala tympani and scala vestibule by the Reisnerr's membrane and basilar membrane.

The hair cells which are responsible for transferring the choclear fluid vibrations to the auditory nerve fibers are located along the basilar membrane. There are one row of inner hair cells and three rows of outer hair cells. The numbers of inner hair cells are approximately 3.500 and the numbers of outer hair cells are approximately 12.000. The inner hair cells and outer hair cells are morphologically similar but their functions are different, the only inner hair cells conduct the basilar membrane vibrations to the neural signals, however the outer hair cells participate in the motion

of the basilar membrane and amplify the basilar membrane vibrations (Moller,2003a).

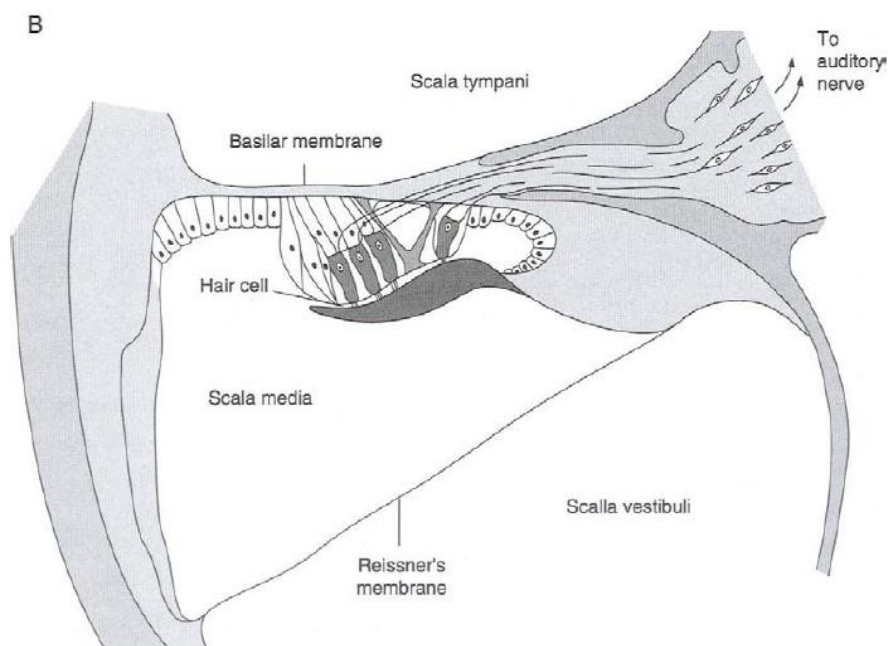


Figure 2.2 The cross section of the second turn of the guinea pig cochlea (From: Moller, A.R., "Anatomy and Physiology of sensory organs", chapter two, pp. 38, Sensory Systems, Elsevier Inc. 2003, with permission)

2.2 The Conduction of Sound into Basilar Membrane Vibrations

The outer ear and middle ear transfer the sound waves into fluid motion of the cochlea. The middle ear act as impedance matcher which improves the sound transmission to the cochlear fluid motion. This action of the middle ear causes the force over the oval window to be greater than round window. The difference between the forces over these two windows put the cochlear fluid into motion. If the sound is allowed to reach these two windows in an identical way, there would not be any fluid motion inside the cochlea (Moller,2003a).

The stapes of the oval window sets the cochlear fluid into motion. When oval window moves inward, the round window moves outward, and this allows the fluid inside the rigid structure of the cochlea to vibrate. The vibration of the fluid vibrates the basilar membrane and bends the hair cells located along the basilar membrane. The hair cells create the action potentials which are then transferred to the auditory

cortex of the human brain by auditory nerve fibers. The schematic drawing of basilar membrane vibrations is given in Figure 2.3.

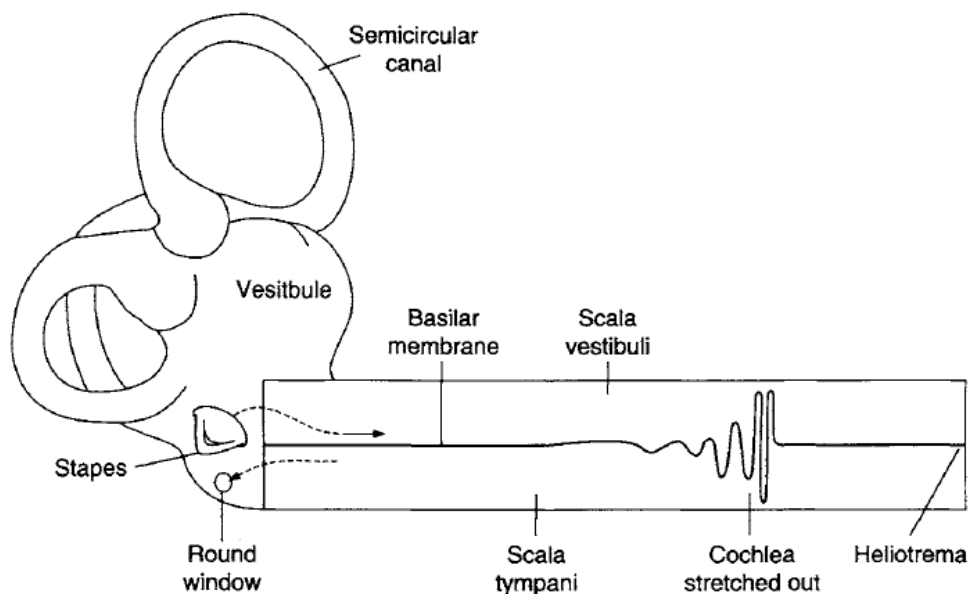


Figure 2.3 Schematic illustration of basilar membrane vibration. The cochlea is shown as straight line. (From: Moller, A.R., "Hearing" chapter five, pp. 293, Sensory Systems, Elsevier Inc. 2003, with permission)

The hair cells are mechanoreceptors that convert the vibrations of basilar membrane into neural signals. There are two types of hair cells: outer hair cells and inner hair cells. The outer and inner hair cells are similar in their structure but their functions differ from each other. The hair cells are directly connected to auditory nerve fibers, there are two types of nerve fibers: afferent and efferent nerve fibers. When the hair cells are deflected the action potentials are created. The hair cells are innervated by afferent and efferent nerve fibers (Moller, 2003b; Sumner et al., 2002). The schematic illustration of innervations of hair cells is shown in Figure 2.4.

The deflections of hair cells are bidirectional. When the hair cells are deflected in one direction, the positive receptor potential is generated and this depolarizes the hair cells. When the hair cells are deflected in opposite direction the negative receptor potential is generated and this hyperpolarizes the cell. The deflections of hair cells occur according to the basilar membrane vibrations. When the basilar membrane

vibrates the hair cells generate positive and negative receptor potentials (depolarization and hyperpolarization) continuously and generate the nerve impulses which encode the basilar membrane vibrations. The schematic illustration of deflections of hair cells and generated nerve impulses are shown in Figure 2.5.

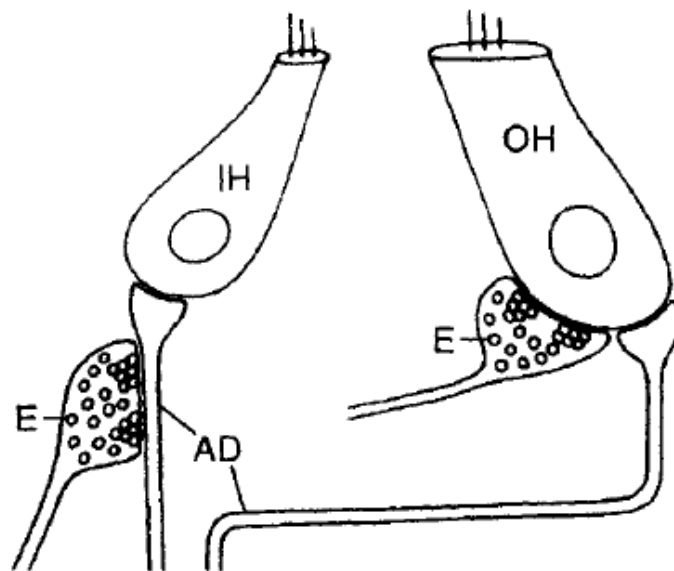


Figure 2.4 Schematic illustration of innervations of hair cells. AD:afferent dendrite, E:efferent synapse, OH:outer hair cell, IH: inner hair cell (From: Moller, A.R., "Hearing" chapter five, pp. 284, Sensory Systems, Elsevier Inc. 2003, with permission)

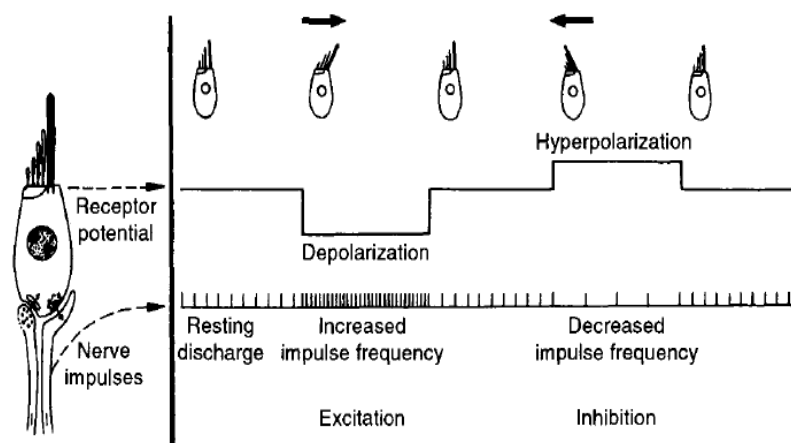


Figure 2.5 The schematic illustration of bidirectional sensitivity of hair cells (From: Moller, A.R., "Anatomy and Physiology of sensory organs", chapter two, pp. 66, Sensory Systems, Elsevier Inc. 2003, with permission)

2.3 The Frequency Selectivity of Basilar Membrane

The basilar membrane vibrations are directly related with the frequency of the sound signals. According to Loizou (1998), the different sound signal frequencies create the travelling cochlear fluid waves which cause the largest amplitude displacement of the basilar membrane at specific location along the basilar membrane; the high frequency signals create the travelling waves that displace the base of the basilar membrane with largest amplitude, and the low frequency sounds create the travelling waves that displace the apex of the basilar membrane with largest amplitudes. The mid frequency signals maximally displace the middle part between the base and apex of the basilar membrane. The base part of the basilar membrane is near to the stapes and the apex is on another ending of the basilar membrane. The maximum displacement parts of basilar membrane to the different frequency sinusoids is shown in Figure 2.6.

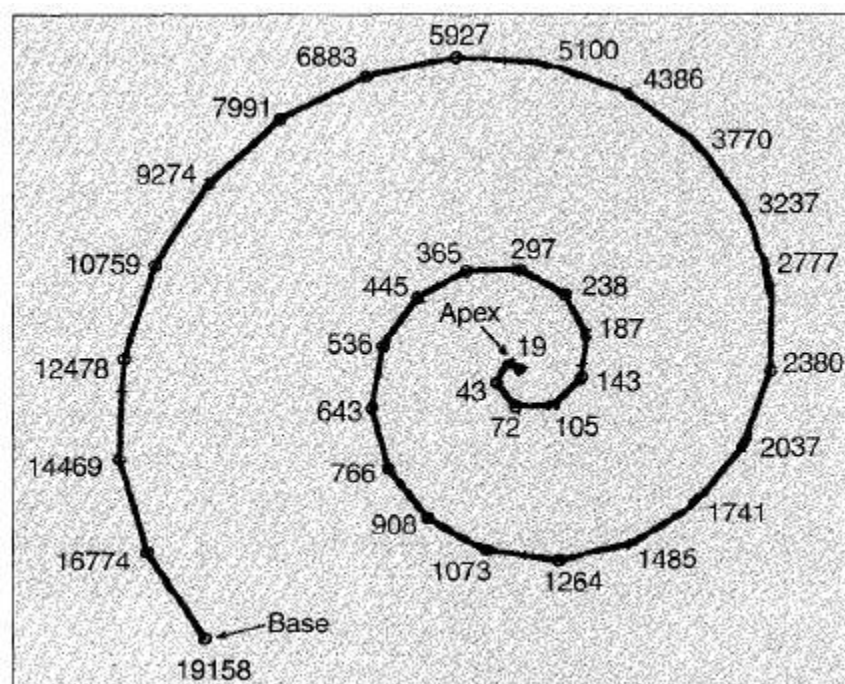


Figure 2.6 The diagram of the basilar membrane with base and apex parts. The points of maximum displacement of basilar membrane to different sinusoids with different frequencies (in Hz) are shown. (From: P.C. Loizou, *Mimicking the Human Ear*, pp.103, IEEE Signal Processing Magazine, 1998, with permission)

The cochlea is the mechanism that encodes the frequencies of the sound signal. The each location of the basilar membrane and deflection of hair cells along the basilar membrane respond with largest amplitude to specific frequencies, this is known as ‘place theory’ (Loizou,1998).

The displacement of the basilar membrane is not only the function of the sound intensity but at the same time it is the function of the frequency of the sound signal (Moller, 2003b). The width of the basilar membrane increases when moving from base to apex. The schematic drawing of basilar membrane of the human cochlea from Moller (2003a) is given in Figure 2.7.

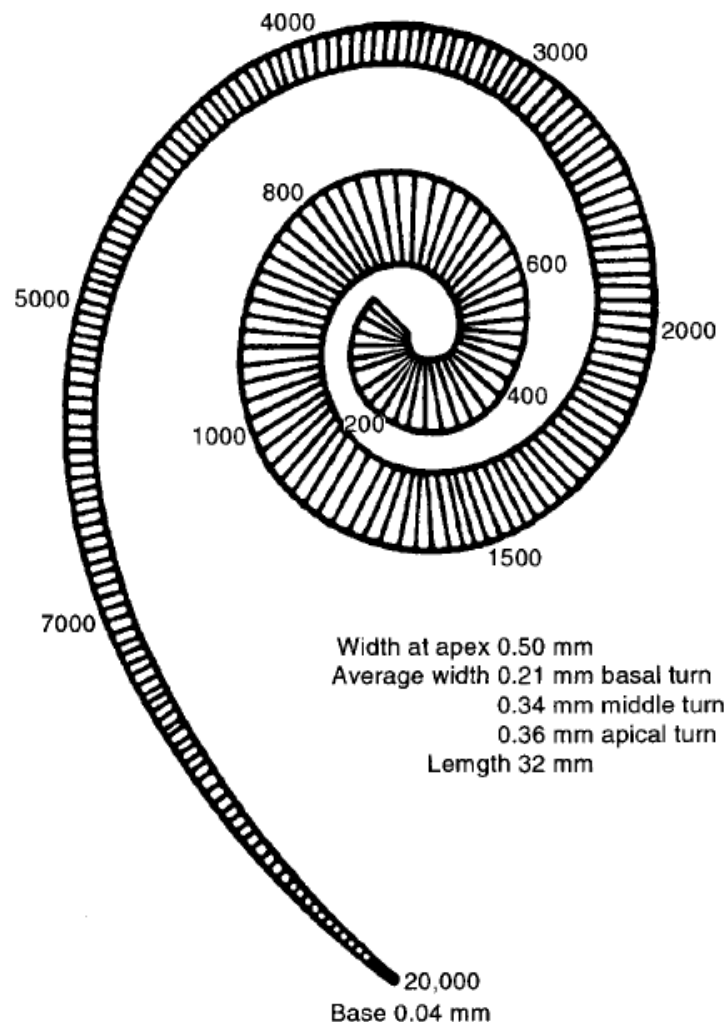


Figure 2.7 The schematic drawing of the basilar membrane of the human cochlea (From: Moller, A.R., “Anatomy and Physiology of sensory organs”, chapter two, pp. 61, Sensory Systems, Elsevier Inc. 2003, with permission)

According to Moller (2003a), the transfer of basilar membrane vibrations to hair cells deflection is a complex process; some hair cells respond to the velocity of the basilar membrane vibrations, some of them respond to the acceleration of the basilar membrane vibrations and some of them respond directly to the displacement of the basilar membrane.

The stapes sets the cochlear fluid into motion. As shown in Figure 2.7 the base of the basilar membrane is stiffer than other locations of the basilar membrane, and this facilitates the energy transfer to the basilar membrane. The vibration of basilar membrane travels from base to the apex which results in a travelling wave motion. The distance of the wave that travels along the basilar membrane is directly the function of the sound signal frequency, when the wave travels specific distance it suddenly becomes extinct (Moller,2003a).

The human ear is sensible to the frequency range of approximately from 20Hz to 20kHz. The most sensible frequency range of the human ear is from 500Hz to 6000Hz (Moller, 2003c).

CHAPTER THREE

AUDITORY MOTIVATED DISCRETE TIME FREQUENCY SIGNAL REPRESENTATION

3.1 The Brief Overview of Time Frequency Signal Representations

The Fourier Transform (FT) of any continuous time signal $x(t)$ is given in Equation 3.1:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt \quad 3.1$$

The $X(\omega)$ contains overall frequency content of the signal $x(t)$, the time information is lost because the integration is performed over all duration of the analyzed signal. The Discrete Fourier Transform (DFT) of discrete time signal $x(n)$ is given in Equation 3.2:

$$X(f) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi f n}{N}} \quad 3.2$$

The $X(f)$ contains overall discrete frequency content of the discrete time signal $x(n)$. The FT and DFT are mostly used for stationary signals, and when the occurrence time intervals of specific frequencies are not important. But in many practical applications the time information is important. In order to obtain time dependent frequency content of the analyzed signal the Short Time Fourier Transform (STFT) was developed and widely used by introducing the window function to the standard FT Equation. The continuous and discrete time forms of the STFT are given in Equations 3.3 and 3.4 respectively (Mertins,1999; Oppenheim & Schafer, 1999).

$$F(\tau, \omega) = \int_{-\infty}^{\infty} x(t) w^*(t - \tau) e^{-j\omega t} dt \quad (3.3)$$

$$F[k, \gamma] = \sum_{n=-\infty}^{\infty} x[n] w[n - k] e^{-j\gamma n} \quad 3.4$$

In Equations 3.3 and 3.4 the $x(t)$ signal is multiplied with continuous and discrete window functions respectively and the $x(t)$ signal is suppressed outside the certain region which gives the local spectra. By shifting the window and performing the FT gives the time dependent FT. In STFT the window function has fixed width which assumes the local stationarity (Oppenheim & Schaffer, 1999).

The spectrogram is another time frequency measure which has been used widely in many practical applications. The spectrogram is obtained by taking the absolute square of the STFT as given in Equation 3.5:

$$Spec[k, \gamma] = |STFT[k, \gamma]|^2 \quad 3.5$$

In order to overcome the shortcomings of the STFT the wavelet based time frequency signal representations were introduced and widely used in signal analysis applications. The continuous wavelet transform of signal $x(t)$ is given in Equation 3.6 (Mertins, 1999):

$$W(b, a) = \frac{1}{a^2} \int_{-\infty}^{\infty} x(t) \varphi^*\left(\frac{t - b}{a}\right) dt \quad 3.6$$

The $\varphi(t)$ is called the mother wavelet. The b is translation parameter and a is dilation or scale parameter which affect the center frequency and bandwidth of the mother wavelet function. The $a^{-1/2}$ is used to verify that at different scales defined with a the each mother wavelet function has the same energy. By changing the scale parameter a the multiresolution analysis is obtained.

3.1.1 The Effect of Windowing

According to the convolution theorem (Oppenheim & Schafer, 1999) if $X(e^{j\omega})$ is the Fourier Transform of $x(n)$, and $H(e^{j\omega})$ is the Fourier Transform of $h(n)$, and if:

$$y(n) = \sum_{k=-\infty}^{\infty} x(n-k)h(k) = x(n) * h(n) \quad 3.7$$

then,

$$Y(e^{j\omega}) = X(e^{j\omega}) H(e^{j\omega}) \quad 3.8$$

The convolution in time is equal to the multiplication in Fourier Domain. According to the windowing theorem (Oppenheim & Schafer, 1999), if $X(e^{j\omega})$ is the Fourier Transform of the $x(n)$, and $W(e^{j\omega})$ is the Fourier Transform of the $w(n)$, then

$$y(n) = x(n)w(n) \quad 3.9$$

$$Y(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\phi}) W(e^{j(\omega-\phi)}) d\phi \quad 3.10$$

The multiplication in time is the periodic convolution in frequency domain. Therefore, multiplying the signal with the window function in time domain, leads to the convolution of the original signal with the windowing function in the frequency domain. The window function has effect on the obtained time frequency resolution. The simplest window function is the rectangular window. The rectangular window function and its Fourier Transform are given in Figures 3.1 and 3.2 respectively. The other window functions like hamming window are used to obtain better frequency response as shown in Figures 3.3 and 3.4. As can be seen from Figures 3.1-3.4 the window function has directly effect on the obtained time frequency resolution. For rapidly changing signals the window length should be small.

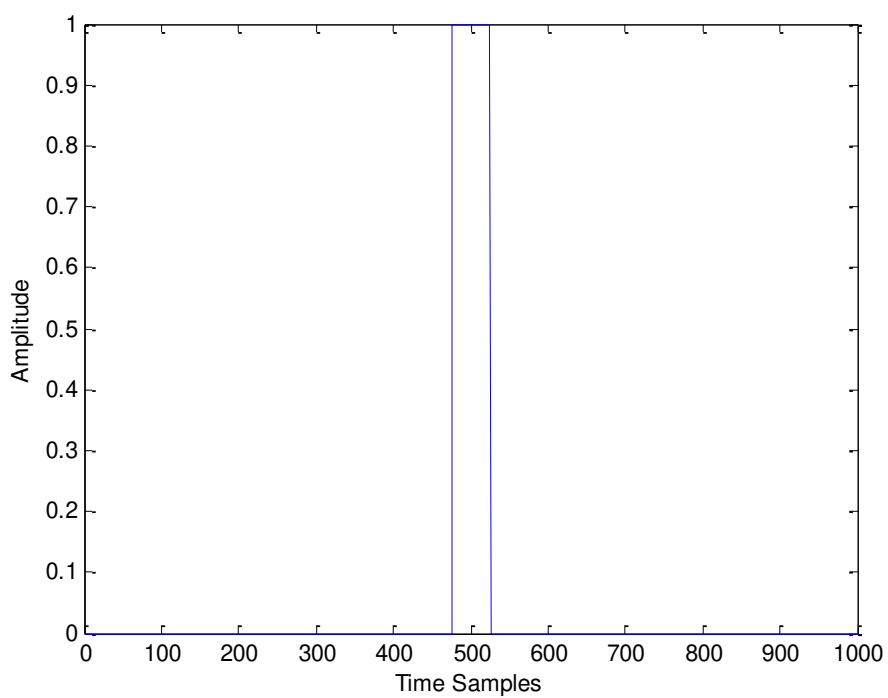


Figure 3.1 The rectangular window with 50 samples length

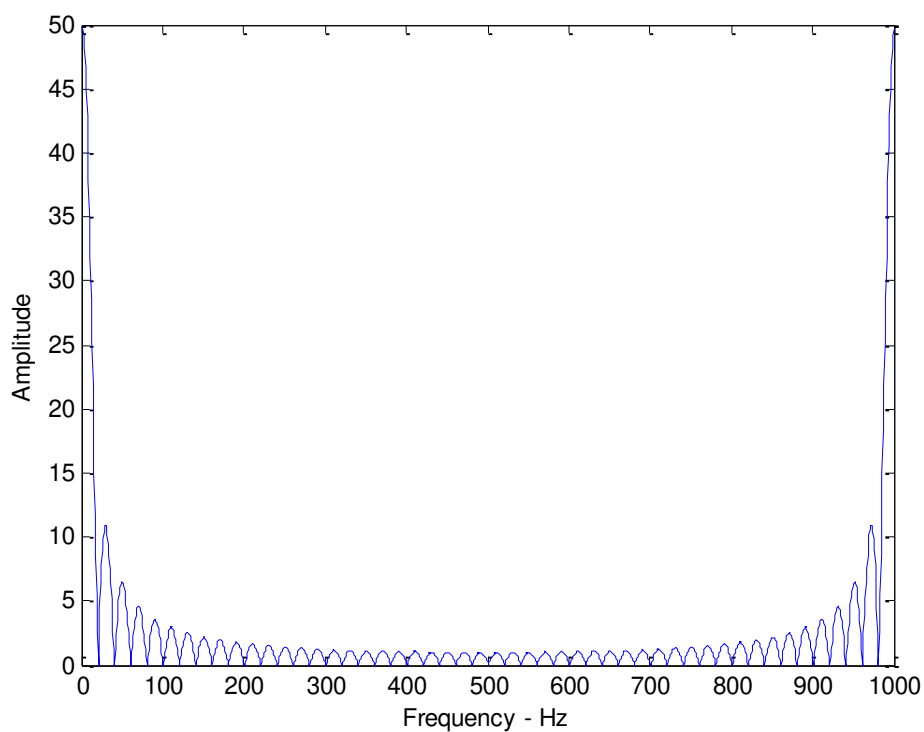


Figure 3.2 The frequency spectrum of the rectangular window with 50 samples length

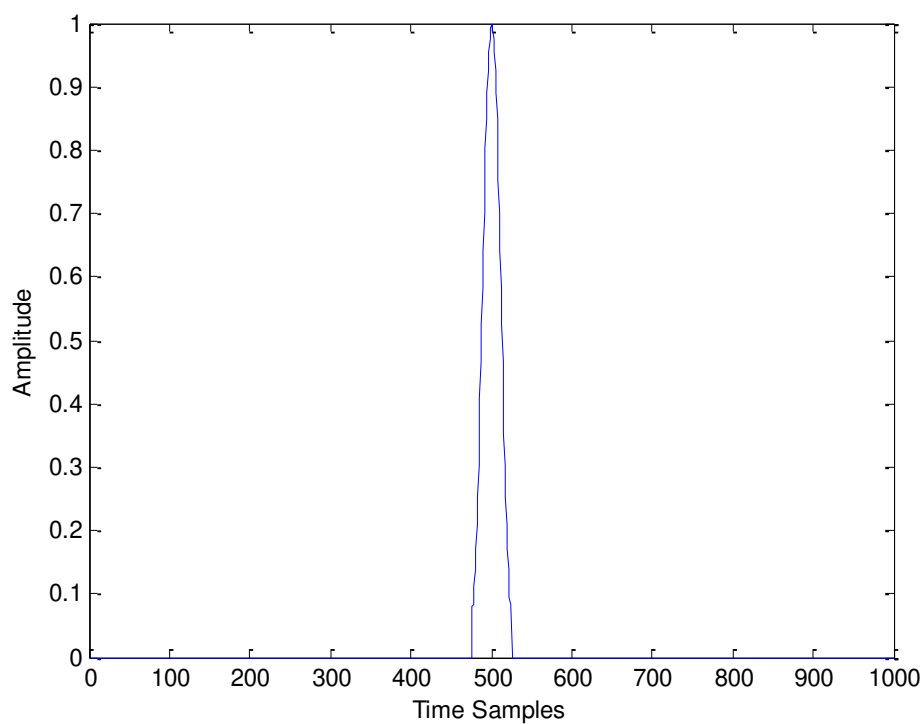


Figure 3.3 The hamming window with 50 samples length

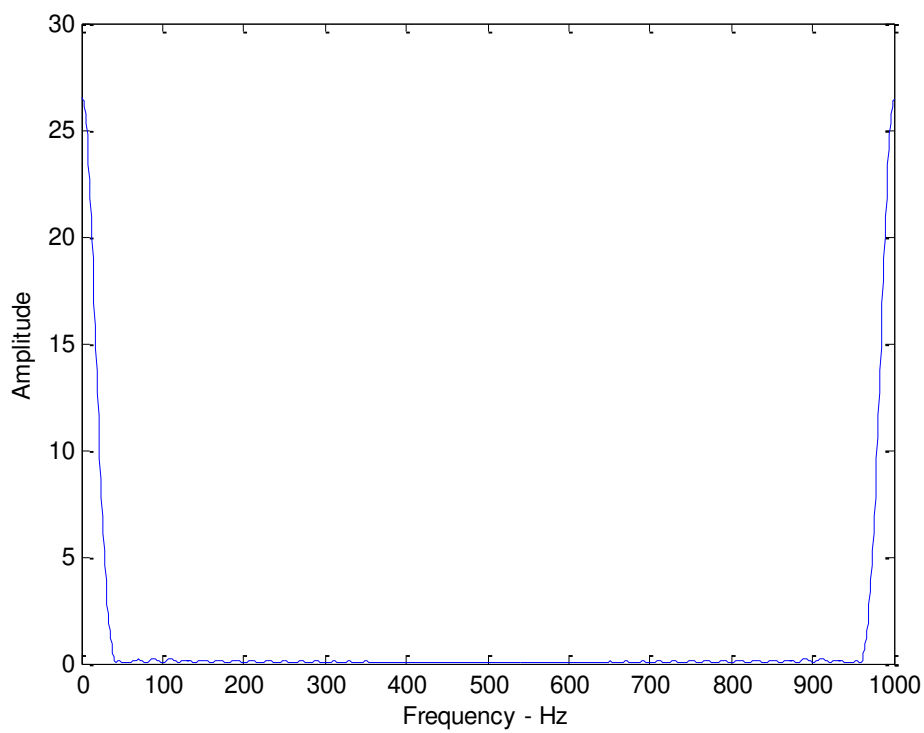


Figure 3.4 The frequency spectrum of the hamming window with 50 samples length

3.2 Auditory Motivated Discrete Time Frequency Signal Representation

In chapter two, the basics of human auditory system is given. As described in chapter two, according to Moller (2003a) some of the hair cells respond to the displacement of the basilar membrane, some of them respond to the speed of the basilar membrane, and some of them respond to the acceleration of the basilar membrane vibrations. Each hair cell responds to a specific frequency. The hair cells are depolarized and hyperpolarized according to the motion of the basilar membrane as given in Figure 2.5, chapter two. If the depolarization and hyperpolarization of hair cells are approximated as sinusoidal deflection than each hair cell has specific sinusoidal frequency to which it responds best.

In the Equation 3.2, the DFT is the response of the analyzed signal to the all discrete frequencies defined with f for all samples of the analyzed signal. At this point the discrete frequencies f may be thought as frequency responses of each hair cell. Therefore it is possible to define the internal sums of the DFT as given in Equation 3.11, which will give the time dependent responses of the each discrete frequency f :

$$X_{f,m} = \sum_{n=0}^m x_n e^{-\frac{j2\pi fn}{N}} \quad 3.11$$

Where $m = 1, 2, 3 \dots \dots N - 1$

The parameter m is introduced to the standard DFT Equation in order to obtain the internal sums of each discrete frequency component. The Equation 3.11 will give two dimensional data with discrete frequencies f and time dependent internal sums m . As a test signal, the cosine signal with 50Hz and 100Hz frequency components for 0.5s , and with 150Hz and 200Hz frequency components for the next 0.5s is used. The result of applying the Equation 3.11 to the test signal with different frequency components is given in Figure 3.5. The $N = 1000$.

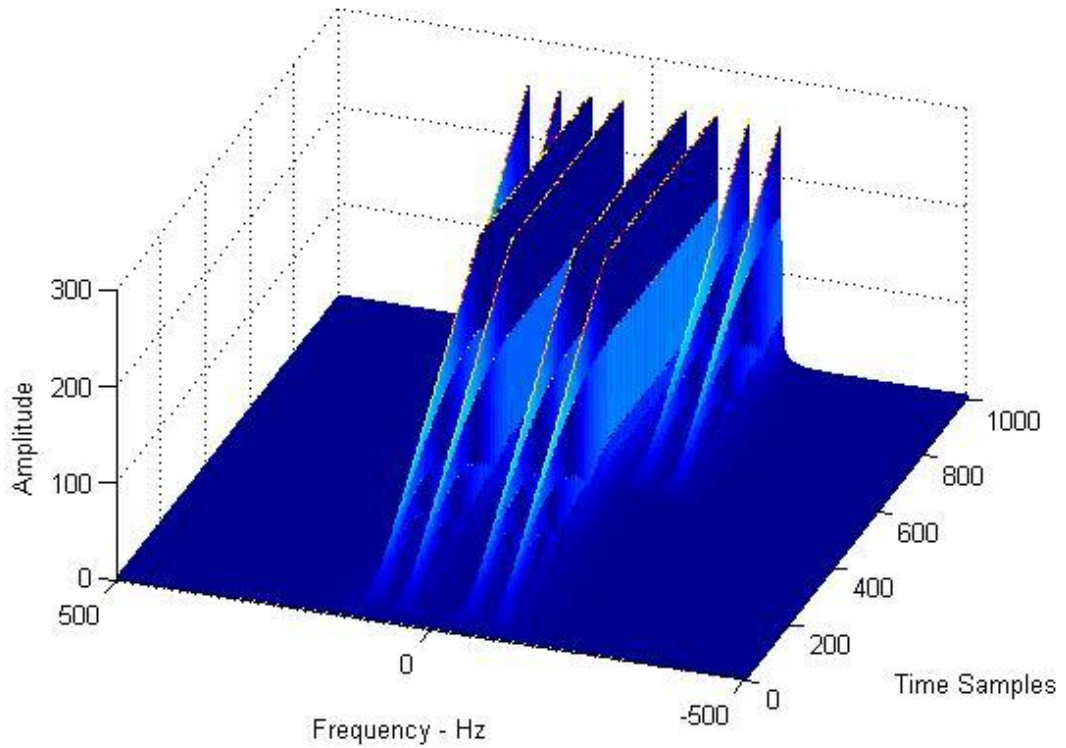


Figure 3.5 The plot of the $X(f, m)$ for cosine signal with different frequencies.

As can be seen from Figure 3.5, the internal sums will increase in the Equation 3.11 for the existing frequency components inside the analyzed signal. Therefore it is possible to define the slope Equation inside the specific time samples interval. The internal sums given in Equation 3.11 may be treated as the response of the hair cells to the displacement of the basilar membrane, and then the slope Equation will give the average speed of the basilar membrane vibrations inside the specific time interval.

$$S_{x f, m} = \frac{X f, m + \Delta k - X f, m}{\Delta k} \quad (3.12)$$

where $m = 1, 2, 3 \dots \dots N - \Delta k - 1$

The Equation 3.12 may be treated as the speed of the basilar membrane vibrations and the results obtained in Equation 3.12 will give the frequency content of the

analyzed signal. The Δk is the average speed interval and can easily be changed according to the analyzed signal. The result of applying Equation 3.12 with $\Delta k = 50$ samples to the internal sums $X f, m$ obtained for different frequency cosine signal is shown in Figure 3.6.

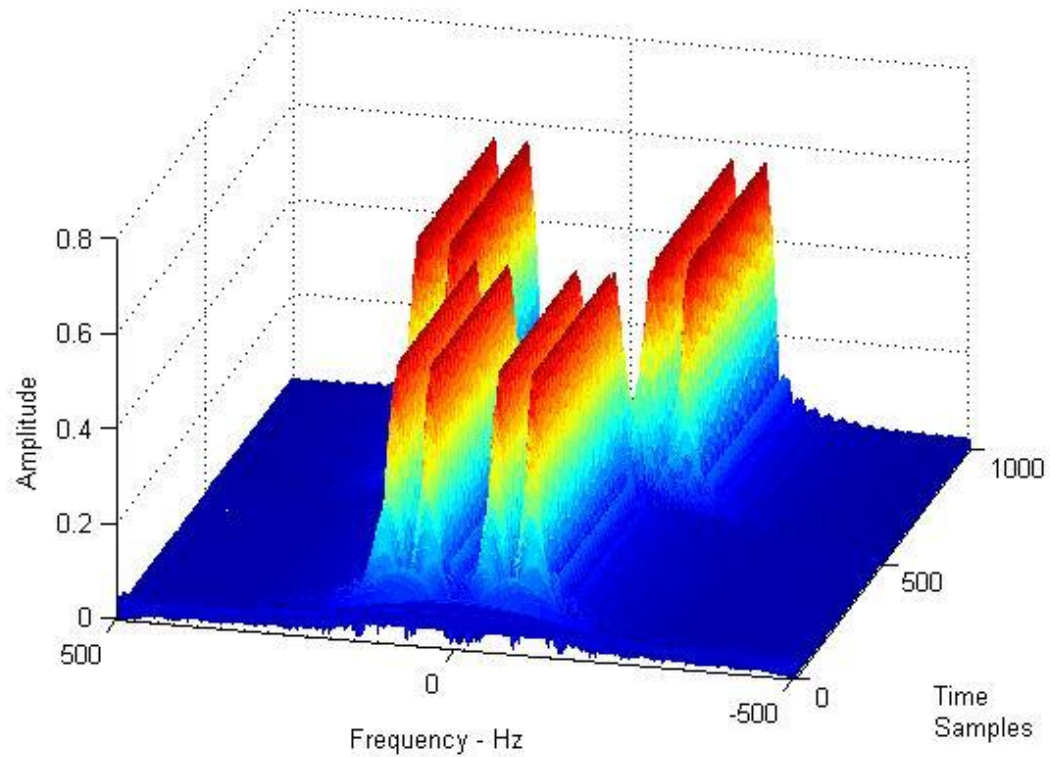


Figure 3.6 The plot of the $Sx(f, m)$ for cosine signal with different frequencies

The Figure 3.6 shows the frequency content of the analyzed signal. It is possible to define the slope Equation for the speed Equation $Sx(f, m)$ again which may be treated as the average acceleration of the basilar membrane vibrations.

$$Ax f, m = \frac{Sx f, m + \Delta k - Sx f, m}{\Delta k} \quad (3.13)$$

Where $m = 1, 2, 3 \dots \dots N - 2\Delta k - 1$

The acceleration Equation 3.13 will provide the information on the start and end times of the specific frequency components if the Equation 3.12 is treated as the frequency content of the analyzed signal. The Figure 3.7 shows the result of applying the Equation 3.13 to the results obtained in Figure 3.6.

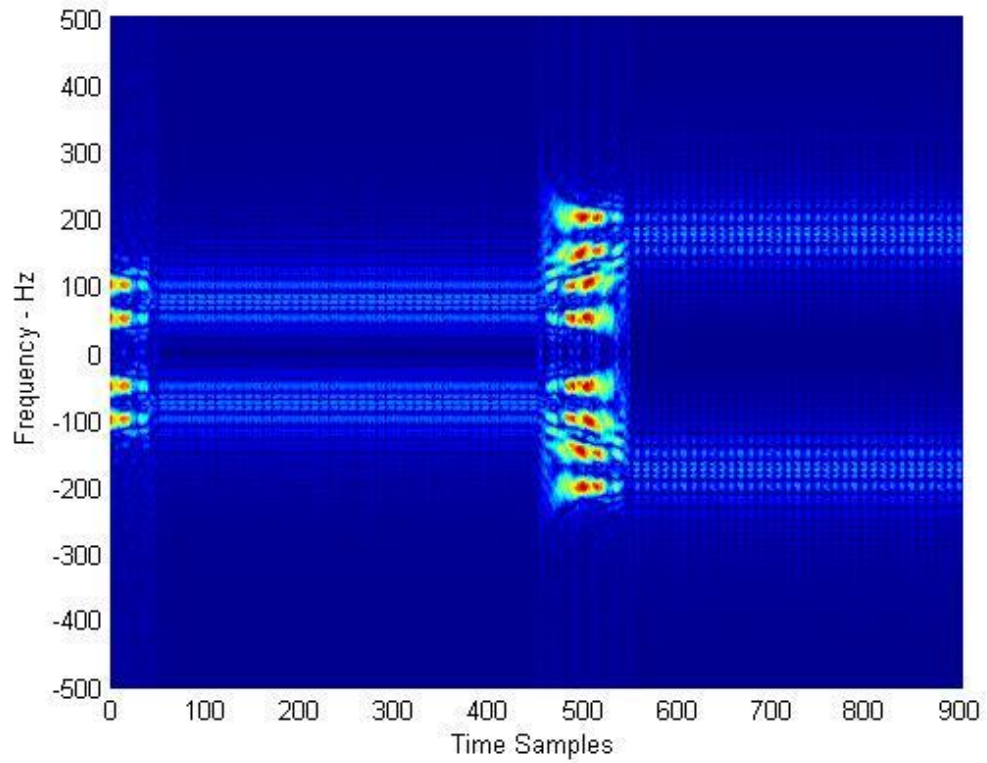


Figure 3.7 The plot of $Ax(f, m)$ applied to the signal shown in figure 3.6

As can be seen from Figure 3.7, the start and end time intervals of the specific frequency components can be obtained from the Equation 3.13.

In Equation 3.12, Δk defines the average speed interval and has directly effect on the obtained frequency resolution. The Δk can easily be adjusted according to the analyzed signal in order to obtain better time frequency resolution. The $Sx(f, m)$ is Δk samples shorter than the analyzed signal length, which may be treated as the stabilization time of the basilar membrane vibrations. Δk should not be chosen very small in order to detect the average speed. However the sample rate N has more effect on time frequency resolution, and will be discussed in the next sections.

3.2.1 Signal Reconstruction

The signal reconstruction from the Inverse Discrete Fourier Transform is given in Equation 3.14:

$$x[n] = \frac{1}{N} \sum_{f=0}^{N-1} X(f) e^{j2\pi f n / N} \quad 3.14$$

In the standard DFT Equation the $X(f)$ is the overall frequency content of the analyzed signal. The internal sums $X(f, m)$ given in Equation 3.11 is the time dependent increase in the frequency content and the overall frequency content is simply the frequency content at $m = N - 1$. The signal reconstruction from the internal sums $X(f, m)$ and the average speed $S_x(f, m)$ are given in the Equations below:

$$x[n] = \frac{1}{N} \sum_{f=0}^{N-1} X(f, m) e^{j2\pi f n / N} \quad m = N - 1 \quad 3.15$$

From the Equation 3.12:

$$X(f, m + \Delta k) = \Delta k S_x(f, m) + X(f, m) \quad m = N - \Delta k - 1 \quad 3.16$$

$$x[n] = \frac{1}{N} \sum_{f=0}^{N-1} X(f, m + \Delta k) e^{j2\pi f n / N} \quad m = N - \Delta k - 1 \quad 3.17$$

The Equation 3.17 is the signal reconstruction formula from the Equation 3.12 which give the time dependent frequency content of the analyzed signal.

The original signal can also be reconstructed directly from the average speed Equation 3.12. At any time instant m the reconstruction can be performed by using the inverse discrete fourier transform formula as given in Equation 3.18.

$$x[n] = \frac{1}{N} \sum_{f=0}^{N-1} \Delta k S_x(f, m) e^{j2\pi f n/N}, \quad m = Tn \quad (3.18)$$

When the inverse discrete fourier transform is performed for any fixed $m=Tn$ the Equation 3.18 will give the last Δk samples from Tn of the original signal. Therefore it is possible to define the fast reconstruction algorithm by applying the Equation 3.18 for specific Tn 's. The Figures 3.10 and 3.11 show the result of applying the Equation 3.18 for two different Tn .

In the standard DFT all samples of the analyzed signal must be known in order to reveal the overall frequency content of the analyzed signal. In the proposed method the analysis need not be made for all samples of the analyzed signal, because the $S_x(f, m)$ defined in Equation 3.12 give the time dependent frequency spectrum. The analysis can be made for the desired sample numbers of the original signal and the original signal can easily be reconstructed from the $X(f, m)$ or $S_x(f, m)$ when the N is known by using the Inverse Discrete Fourier Transform Equation. For the known N , the $X(f, m)$ and $S_x(f, m)$ can be performed for the desired sample numbers defined with Nd , only the analyzed signal samples can be reconstructed by using the Equation 3.18.

As an example, the original $x[n]$ defined in Equation 3.19 is used to make the analysis for the $Nd = 335$ samples of the original signal when $N = 1000$. The first 335 samples of the original signal and the frequency content obtained from Equations 3.11 and 3.12 are given in Figures 3.8 and 3.9 respectively.

$$x[n] = \cos \frac{2\pi 125n}{N} + 3 \cos \frac{2\pi 270n}{N}, \quad 0 \leq n \leq Nd - 1, N = 1000 \quad 3.19$$

The Figures 3.10 and 3.11 show the results obtained from Equation 3.18 for $Tn=300$ and $Tn=200$ respectively. The reconstructed original signal from the reconstruction algorithm is given in Figure 3.12.

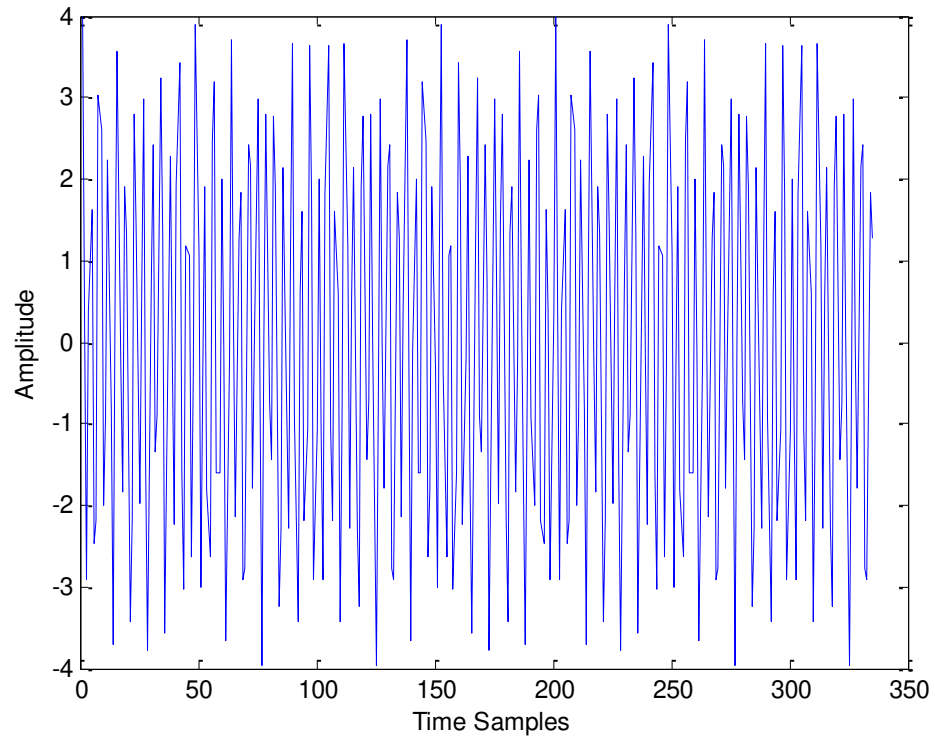


Figure 3.8 The original $x(n)$ signal defined in equation 3.19, first $Nd = 335$ samples

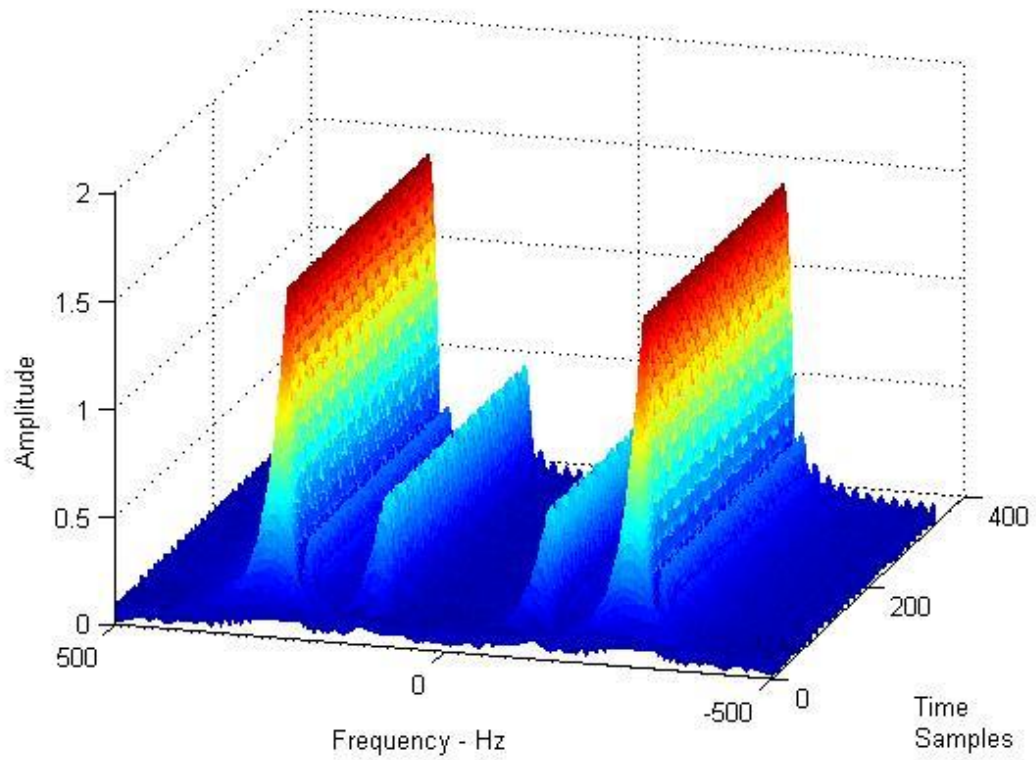


Figure 3.9 The plot of the $Sx(f, m)$ for $m = Nd - \Delta k - 1$ samples, applied to the signal shown in figure 3.8, $Nd = 335$, $N = 1000$, $\Delta k = 50$.

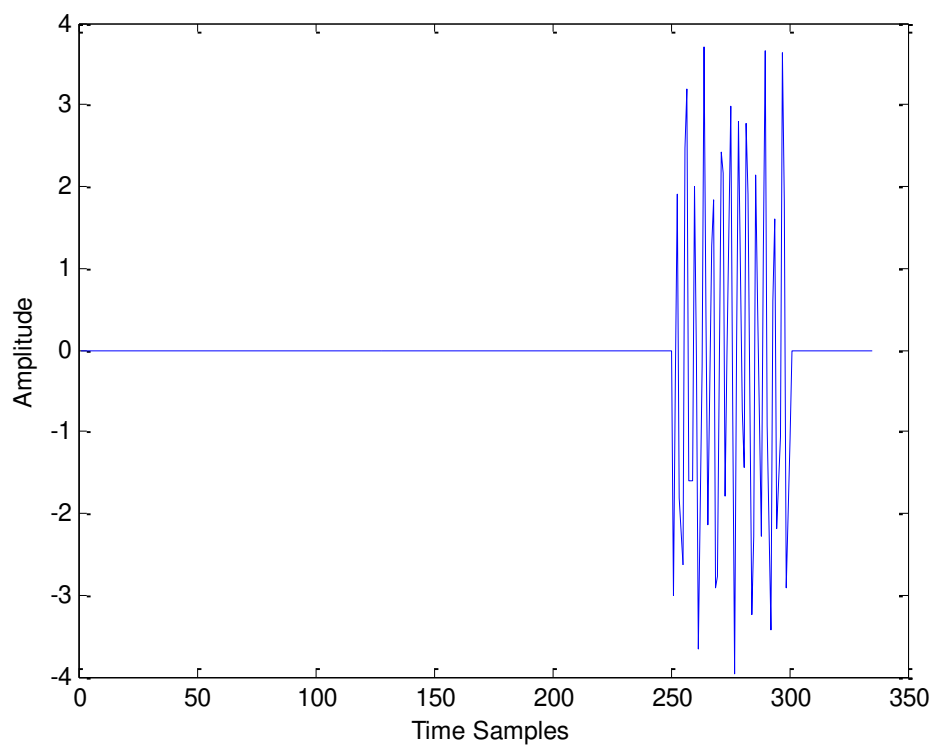


Figure 3.10 The reconstructed signal samples for $Tn=300$, $\Delta k=50$.

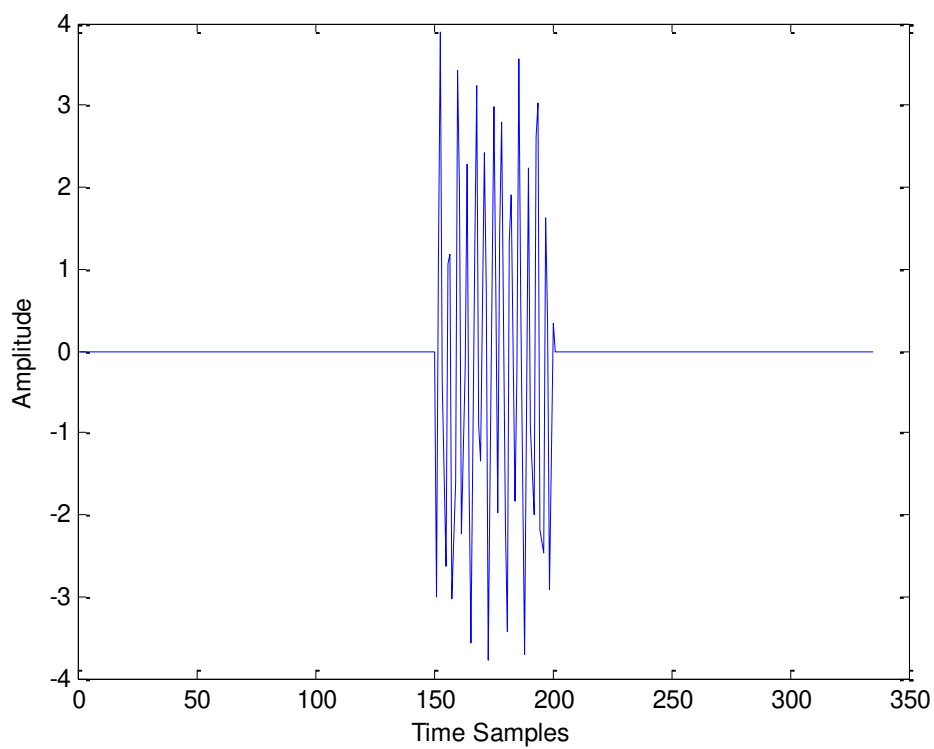


Figure 3.11 The reconstructed signal samples for $Tn=200$, $\Delta k=50$.

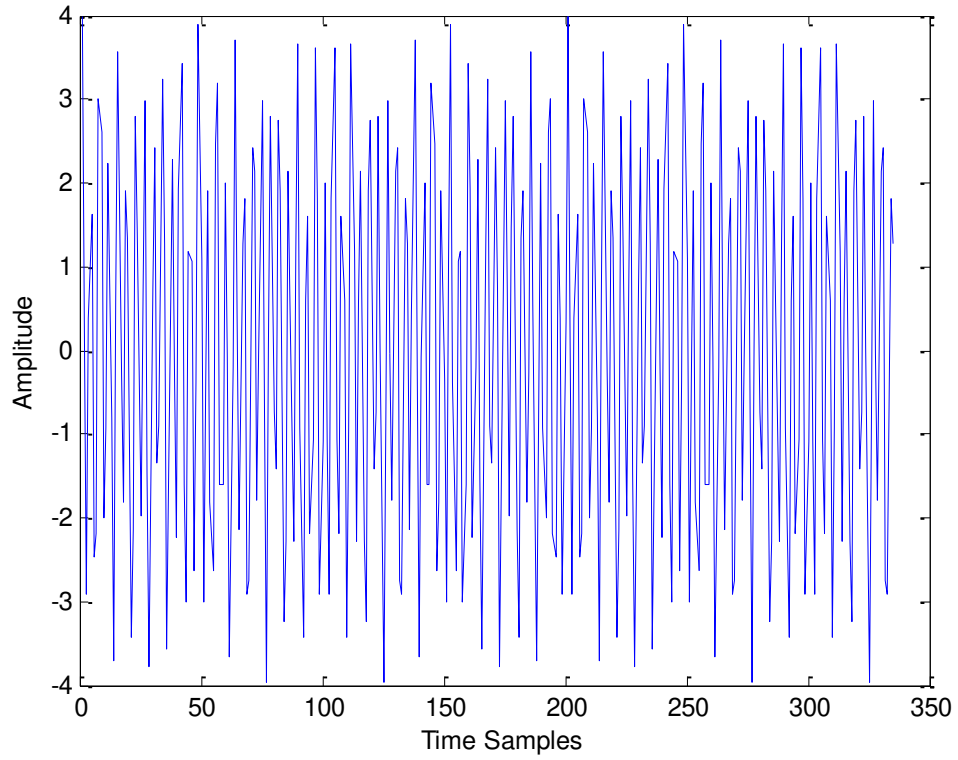


Figure 3.12 The reconstructed signal from equation 3.18, $Nd = 335, N = 1000, \Delta k = 50$.

3.2.2 Numerical Simulations

The presented Auditory Motivated Time Frequency Representation (AMTFR) method is simple and is based on the $S(f, m)$ defined in Equation 3.12. In this section the AMTFR method is applied to the different mono component and multi component signals and its performance is tested for different Δk and SNR values.

The spectrogram of the STFT defined in Equation 3.5 is applied to the $Sx(f, m)$ as given in Equation 3.20 and the results for different types of signals are shown for $Sx(f, m)$ and its spectrogram.

$$Spec(f, m) = |Sx(f, m)|^2 \quad 3.20$$

The $Sx(f, m)$ is Δk samples shorter than $X(f, m)$ because of the slope interval. In the results shown in this section, the first Δk samples of the $Sx(f, m)$ is made equal

to the first Δk samples of $X(f, m)$. The $Sx(f, m)$ is shifted Δk samples right in order to equalize the lengths.

The nonlinear chirped signal is given in the Equation below (Zhong&Huang, 2010):

$$C_n = \exp \frac{j4\pi f}{3} \frac{n-1}{N-1}^{3/2} \quad (3.21)$$

Where $N = 2000, 1 \leq n \leq N, f = 900$.

The N is the length of the signal and the f is the maximum frequency. The real part of the Equation 3.21 is taken for analysis as given in Equation 3.22:

$$c_n = \cos \frac{4\pi f}{3} \frac{n-1}{N-1}^{3/2} \quad (3.22)$$

The waveform defined in Equation 3.22 is given in Figure 3.13.

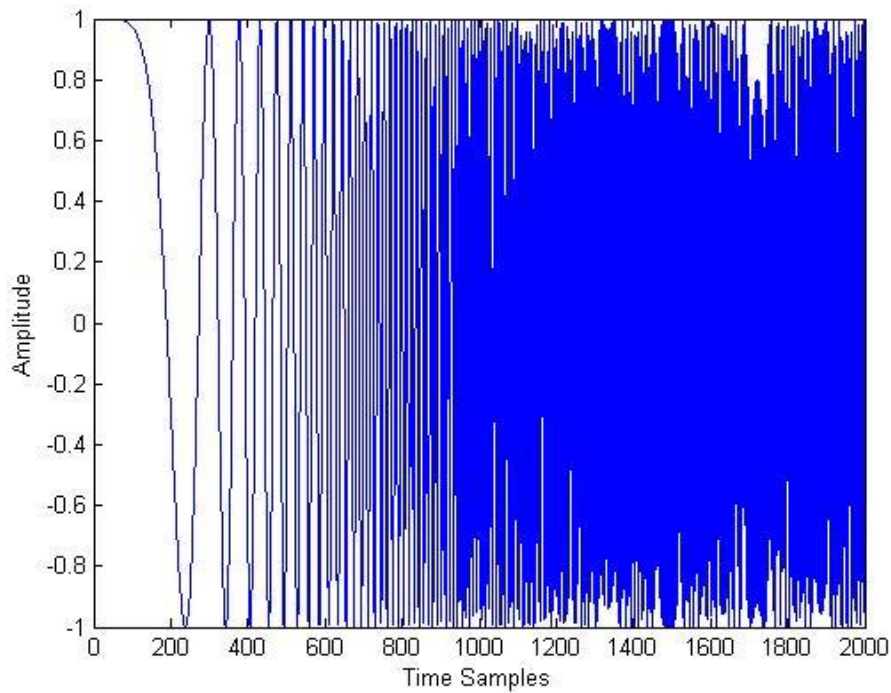


Figure 3.13 The nonlinear chirped signal with frequencies from zero to 900Hz.

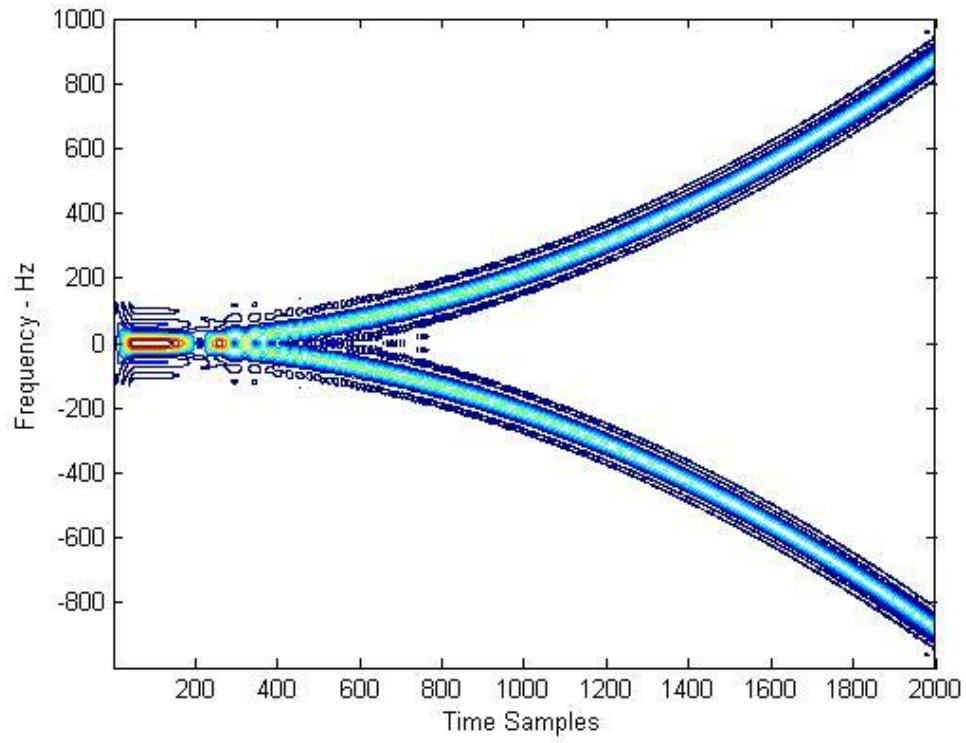


Figure 3.14 The contour plot of $Sx(f, m)$ for the nonlinear chirped signal defined in equation 3.22, $\Delta k = 50$.

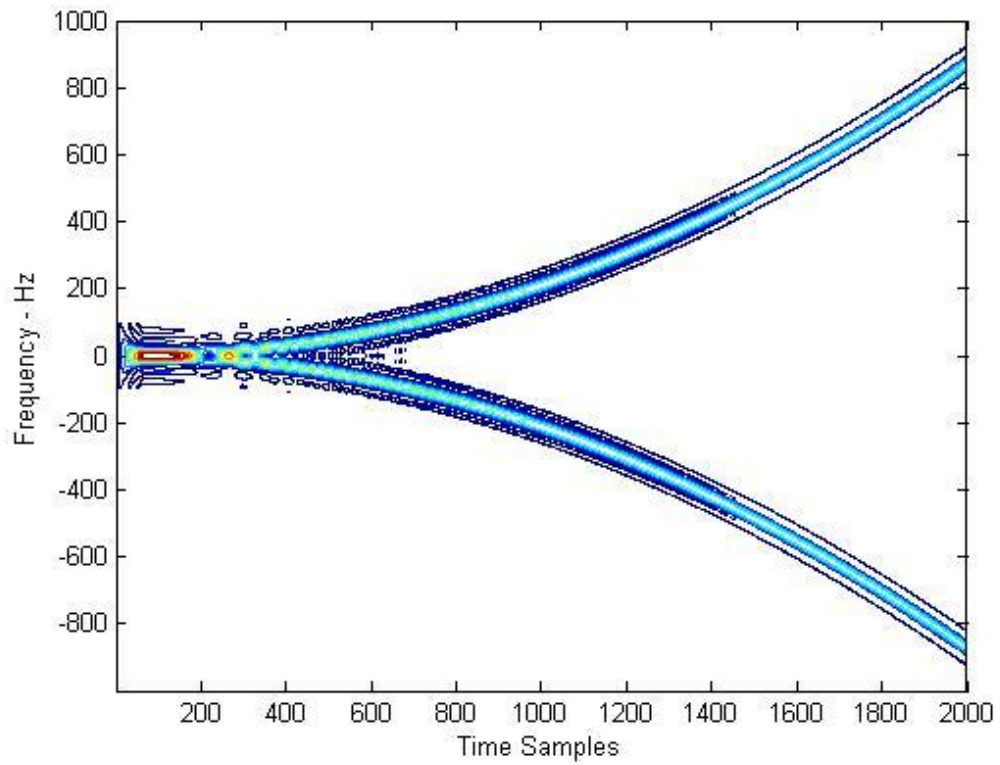


Figure 3.15 The Contour Plot of $(Sx(f, m))$ for $\Delta k = 65$.

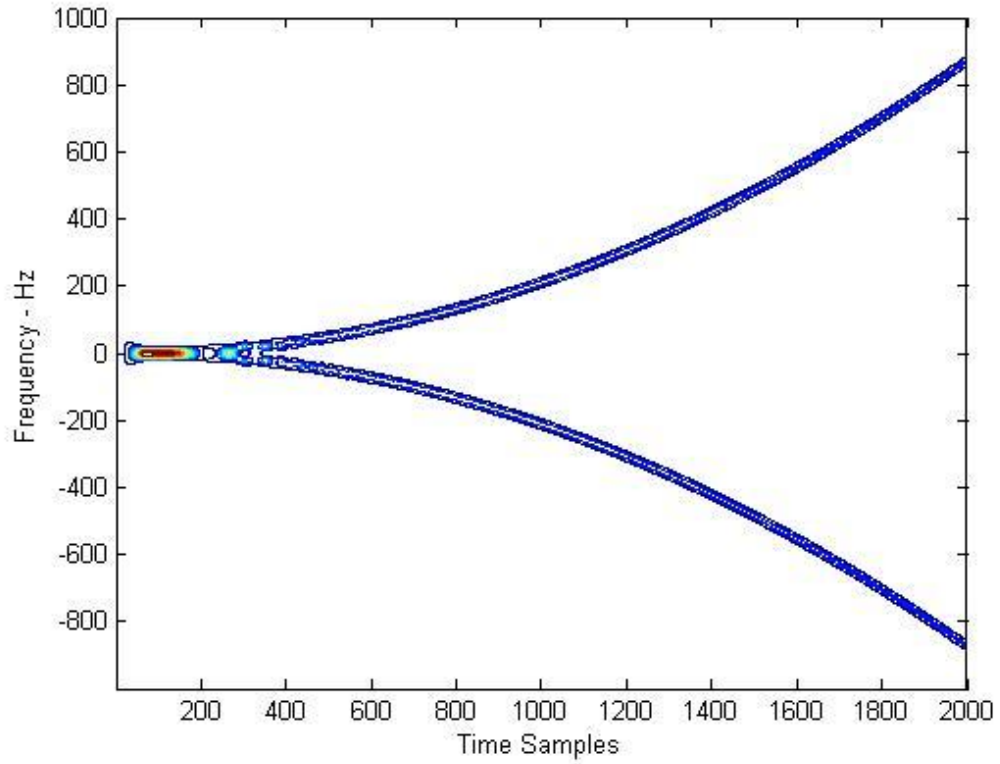


Figure 3.16 The contour plot of $Spec(f, m)$ for $\Delta k = 65$

The Figures 3.14 and 3.15 show the results obtained for different Δk values. As can be seen from Figures the Δk has effect on the obtained time frequency resolution, but it simply can be adjusted to obtain better resolution. The sample numbers defined by N has more effect on the resolution because of the discrete computation of the slope interval.

The result obtained in Figure 3.16 from $Spec(f, m)$ give better visualization of the proposed method. The obtained time and frequency resolutions are almost same at higher and lower frequencies. The results given in Figure 3.16 are comparable to the results obtained in (Zhong&Huang, 2010). In (Zhong&Huang, 2010) the local stationarity of the nonlinear chirped signal is defined from the instantaneous frequency (IF) obtained from the ridge of the Wavelet Transform, and the window length of the Adaptive Short Time Fourier Transform (ASTFT) is adjusted according to the detected IF.

The additive white Gaussian noise is added to the nonlinear chirped signal for the different SNR_{dB} values. The white Gaussian noise added to the nonlinear chirped signal given in Equation 3.22 for the $SNR=0dB$, is shown in Figure 3.17.

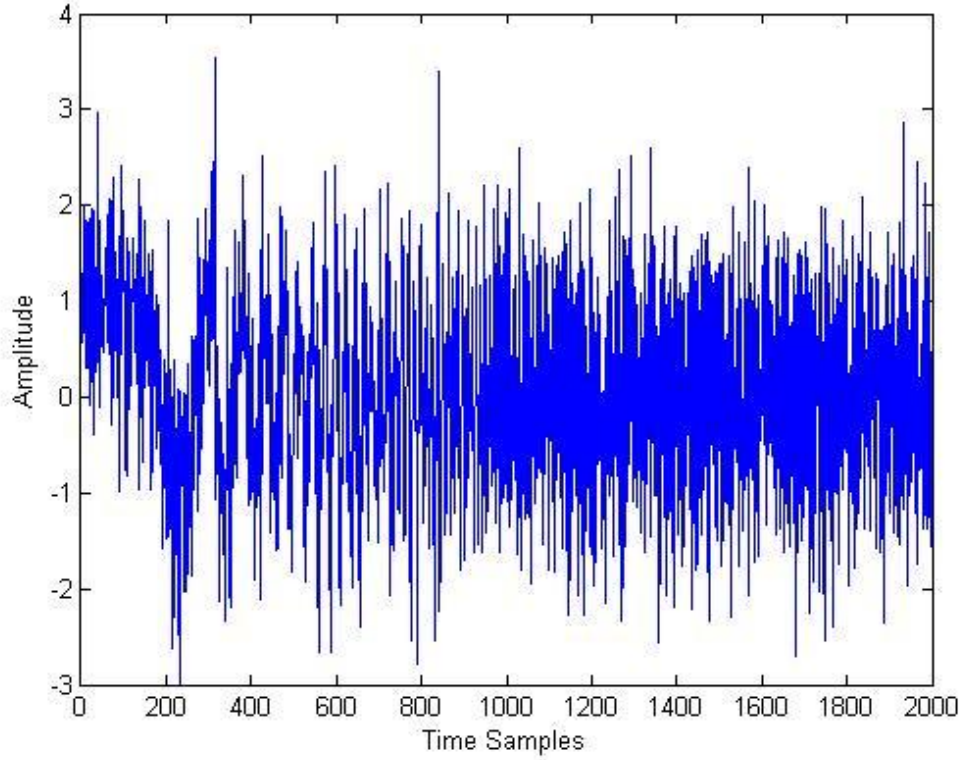


Figure 3.17 The nonlinear chirped signal with additive white Gaussian noise, $SNR=0dB$.

Figures 3.18 and 3.19 show the results obtained from $Spec(f, m)$ for the SNR values 0dB and 5dB respectively. As can be seen from Figures, the obtained time frequency resolutions are similar to the results obtained in Figure 3.16 even for low SNR_{dB} values.

The synthetic time series consisting of three linear chirps is given in the Equation below (Lu&Zhang, 2009):

$$\begin{aligned}
 x[n] = & \cos\left(2\pi\left(130 + \frac{n}{20}\right)\frac{n}{2000}\right) + \cos\left(2\pi\left(250 + \frac{n}{20}\right)\frac{n}{2000}\right) \\
 & + \cos\left(2\pi\left(400 + \frac{n}{20}\right)\frac{n}{2000}\right), \quad 1 \leq n \leq 2000
 \end{aligned} \tag{3.23}$$

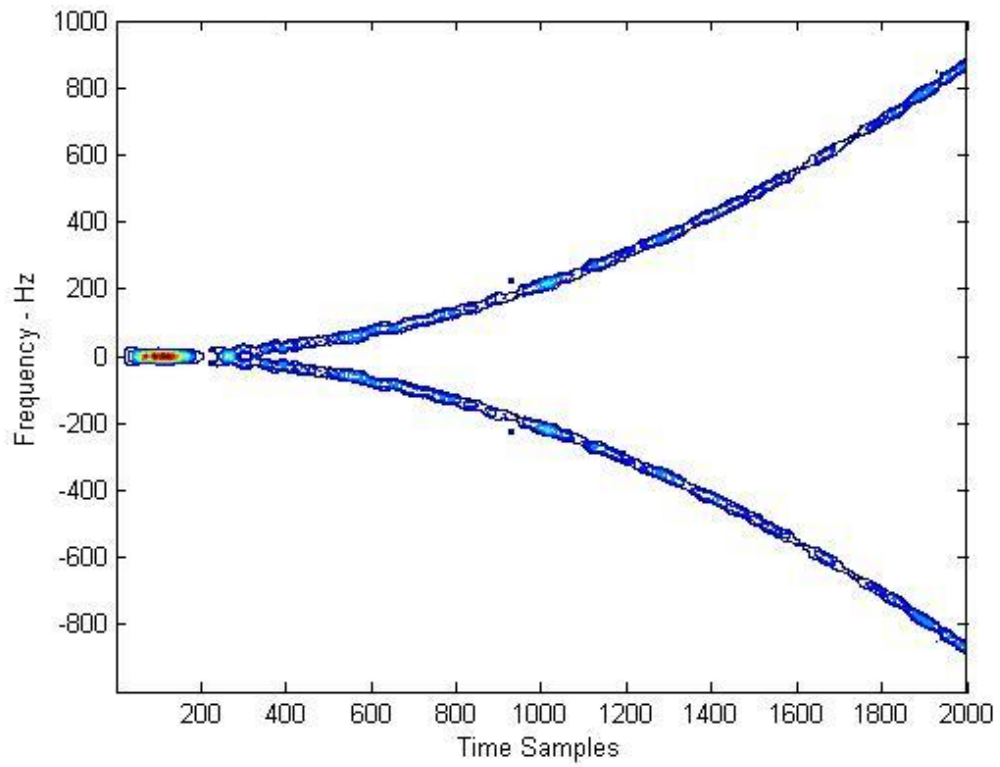


Figure 3.18 The contour plot of $Spec(f, m)$, $SNR=0dB$.

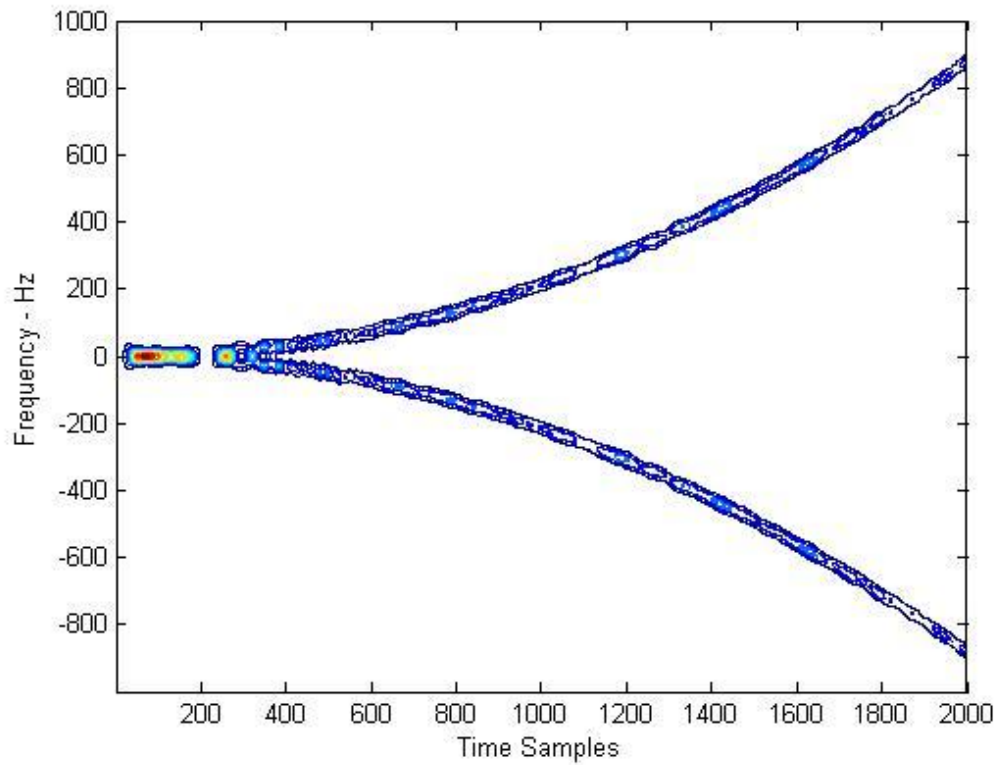


Figure 3.19 The contour plot of $Spec(f, m)$, $SNR=5dB$.

The waveform obtained in 3.23 and the $Spec(f, m)$ for $\Delta k = 100$, are shown in Figure 3.20 and 3.21 respectively.

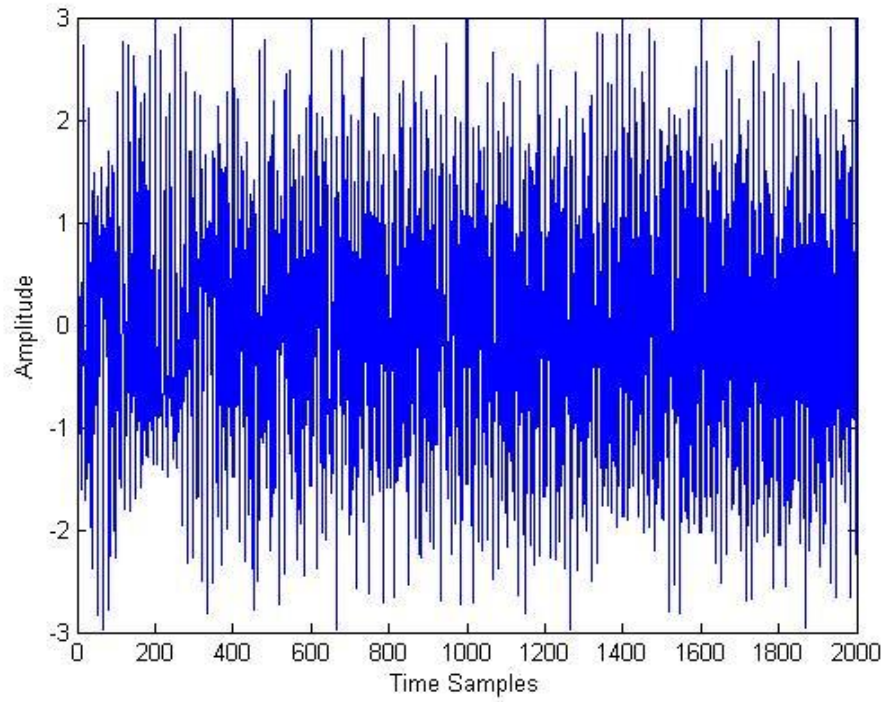


Figure 3.20 The waveform of synthetic time series defined in equation 3.23

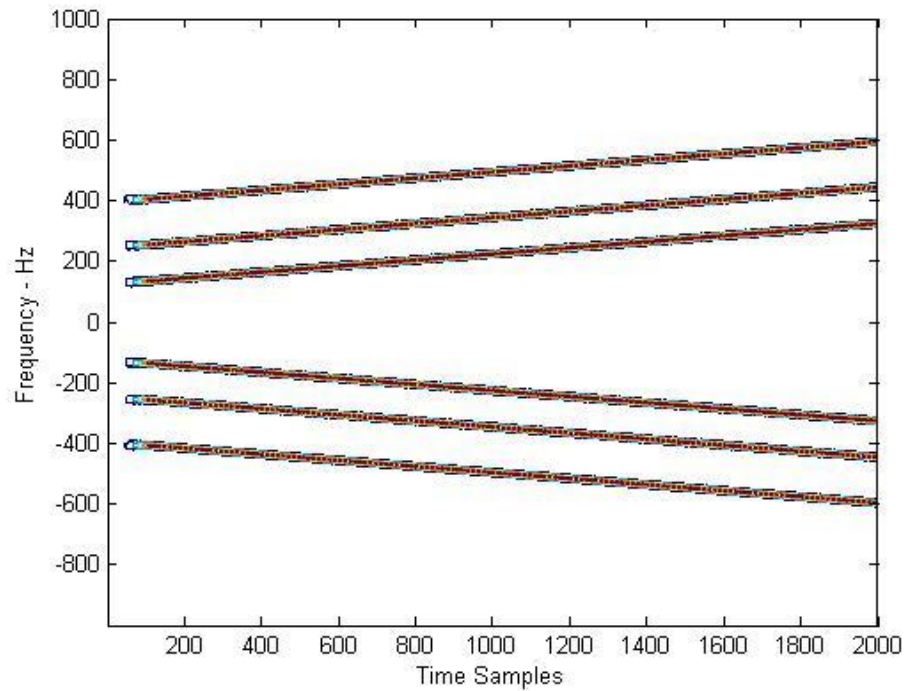


Figure 3.21 The contour plot of $Spec(f, m)$ applied to the signal shown in figure 3.20

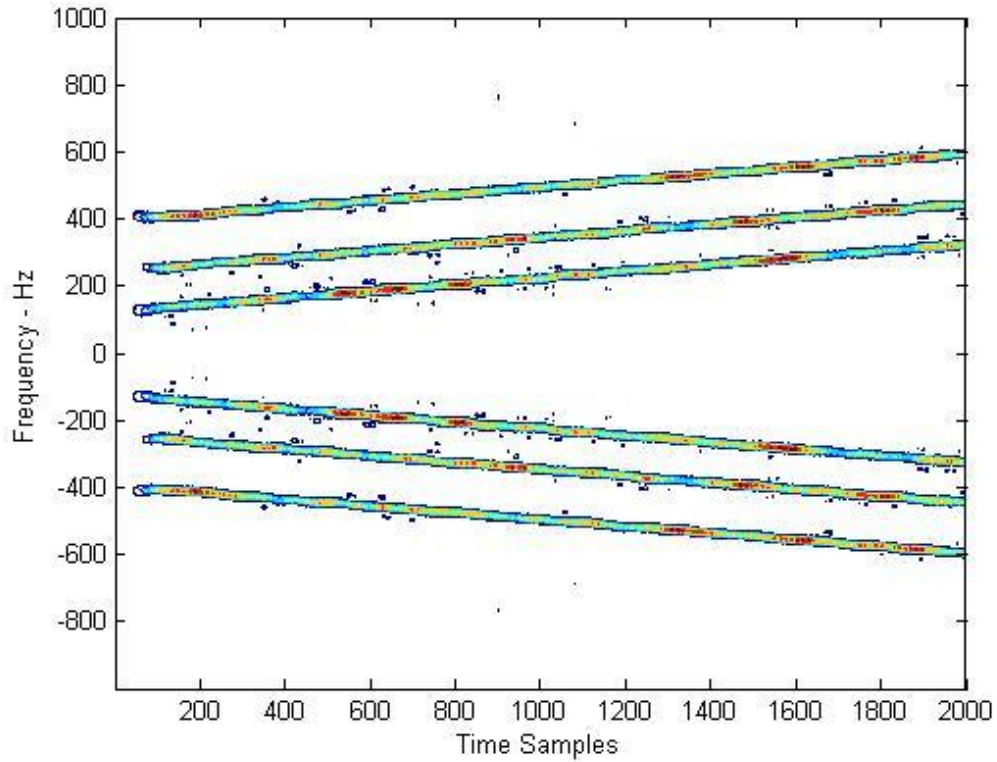


Figure 3.22 The contour plot of $Spec(f, m)$ for the $SNR = 5dB$

The Figure 3.22 shows the results obtained with additive white Gaussian noise added to the signal given in Equation 3.23 for the $SNR = 5dB$. The results obtained in Figures 3.21 and 3.22 give same time frequency resolution for low and high frequency components. In (Lu&Zhang, 2009) the Deconvolutive Short Time Fourier Transform approach is used to obtain the time frequency resolution. The results obtained from $Spec(f, m)$ are for higher values of n when compared to the synthetic signal used in (Lu&Zhang, 2009), also the frequencies are higher. Because the slope Equation is used in AMTFR method, the sample numbers play important role in the obtained time frequency resolution. The time frequency resolution obtained for low sample rate is poorer which is the main disadvantage of the AMTFR method. However at higher sample numbers the AMTFR method give comparable results to the results obtained in literature. The AMTFR is independent from the window function mostly used in time frequency signal representations, which gives main constraint to the time frequency resolution defined by the uncertainty principle (Loughlin&Cohen, 2004).

The effectiveness of the AMTFR method is also tested for the multicomponent nonlinear chirped signal which has the three nonlinear chirped signal components with different frequencies. The signal is created with the *chirp* function of the Matlab, the logarithmic sweep method is used to create the three nonlinear chirped signal with the following start f_0 and end f_1 frequencies: $f_0 = 50, f_1 = 700$; $f_0 = 600, f_1 = 30$; $f_0 = 200, f_1 = 950$. The waveform of the original signal and the noisy signal with $SNR = 3dB$ are shown in the Figure 3.23 and 3.24 respectively.

The results obtained from $Spec\ f, m, \Delta k = 50$, for the original and noisy signal are shown in Figures 3.25 and 3.26 respectively. The time frequency resolution is same for low and high frequency components also in the case of multicomponent signals, and are comparable to the results obtained in (Zhong&Huang, 2010) for the two component nonlinear chirped signal by using the adaptive short time fourier transform. Only sample rate used for the simulations in the AMTFR method is higher.

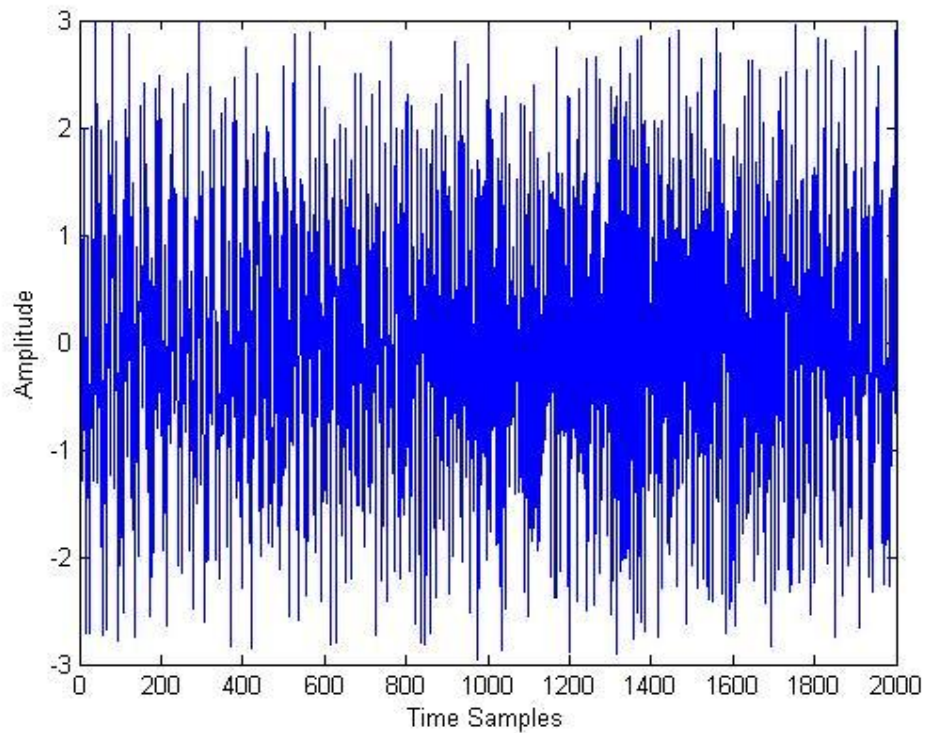


Figure 3.23 The original waveform of the multicomponent nonlinear chirped signal

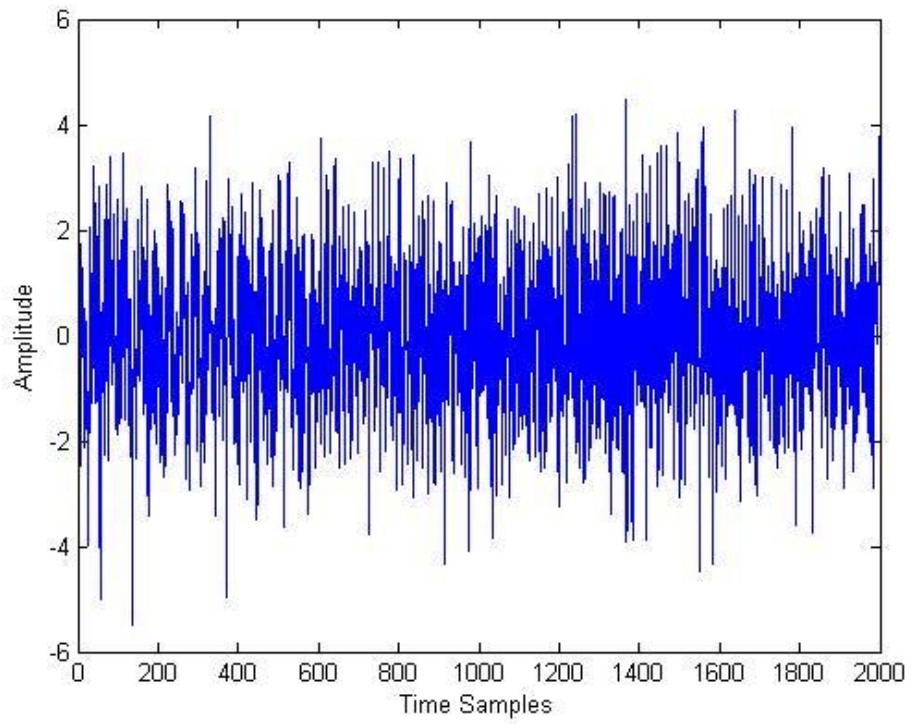


Figure 3.24 The noisy waveform of the multicomponent nonlinear chirped signal, $SNR = 3dB$.

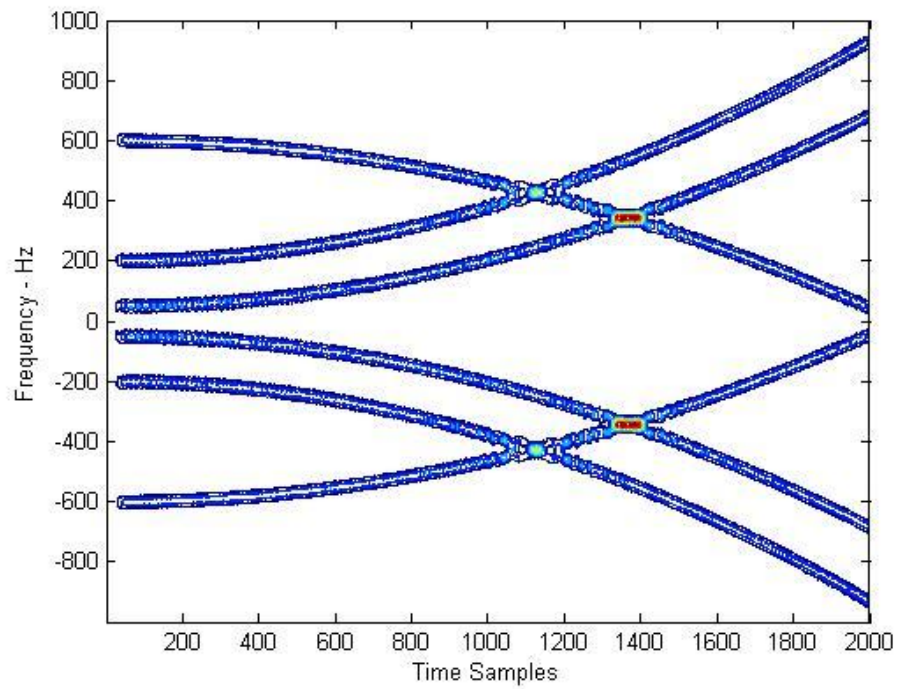


Figure 3.25 The contour plot of $Spec(f, m)$ applied to the original multicomponent nonlinear chirped signal, $\Delta k = 50$.

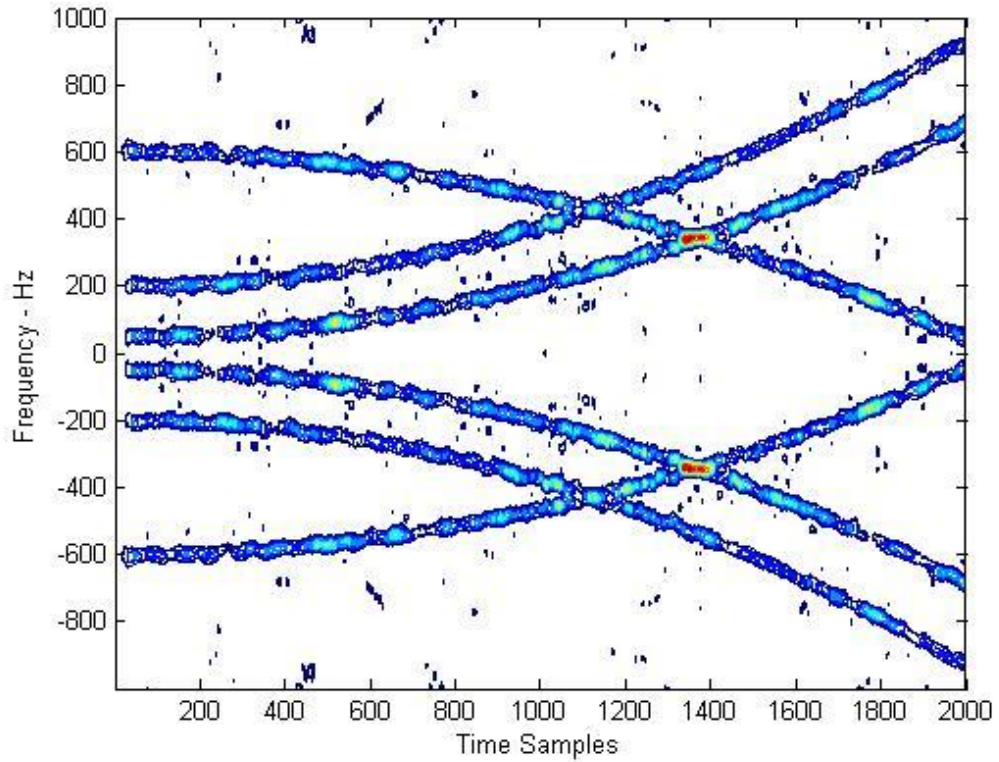


Figure 3.26 The contour plot of $Spec(f, m)$ applied to the noisy multicomponent nonlinear chirped signal, $\Delta k = 50$, $SNR = 3dB$.

In this section the different chirped signals were used to show the effectiveness of the AMTFR method. In all simulation results fixed Δk values were used. The performance of the AMTFR method was tested also for noisy signals with different SNR_{dB} values. The nonlinear chirped signals have the characteristics that change over time, but within the small time intervals their characteristics are stationary. Therefore for the non-stationary signals those have the local stationary properties the fixed Δk can be used for the time frequency analysis. The sample rate is more important than Δk for the performance of the AMTFR method which is the main disadvantage of the proposed method. However when the sample rate is high enough, the results are comparable to the results obtained in literature.

For the non-stationary signals without local stationary properties the variable Δk must be used which will be explained in the next sections.

3.2.3 Auditory Motivated Discrete Time Frequency Signal Analysis Method

The AMTFR method described in section 3.2 can be used for time frequency representation, which is simply based on the internal sums of the Discrete Fourier Transform (DFT). The average speed Equation 3.12 detects the increase in the internal sums of the DFT obtained in Equation 3.11 for fixed Δk . However, for fixed Δk , the sudden increase in the internal sums Equation 3.11 cannot be easily detected because the higher frequency components last in shorter time intervals. Therefore it is important to define the variable Δk for each frequency component. In the case of continuous time signals the variable Δk can be defined as given in the Equations below:

$$X(f, t) = \int_{-\infty}^t x(t) e^{-j2\pi ft} dt \quad (3.24)$$

$$S_x(f, \tau) = \frac{X(f, t + \tau) - X(f, t)}{\tau} \quad 3.25$$

$$\tau = \frac{2}{f} \quad 3.26$$

The average speed of each frequency component can be calculated in two periods of the analyzed frequency. Therefore the Equation 3.25 will give the better time frequency resolution compared to the results obtained in section 3.2.2. However, because of the discrete computation of the computers, it is impossible to define variable discrete slope intervals for each frequency component. In order to overcome this problem, it is possible to define the complex exponentials those have the discrete periods, and the variable Δk slope Equation can be applied easily to these discrete time intervals, as given in the Equations below:

$$X(n, m) = \sum_{n=0}^m x(n) e^{-\frac{j2\pi n}{S_n}} \quad 3.27$$

$$Sx_{Sn,m} = \frac{X_{Sn,m+2Sn} - X_{Sn,m}}{2Sn} \quad 3.28$$

Where $m = 1, 2, 3 \dots N - 2Sn - 1$.

The variable Sn is the discrete period of the each complex exponential function defined in Equation 3.27 and its frequency correspondence is dependent on the sample rate N of the analyzed signal. For $N = 1000$, the frequency correspondence f_c of the Sn is plotted in Figure 3.27.

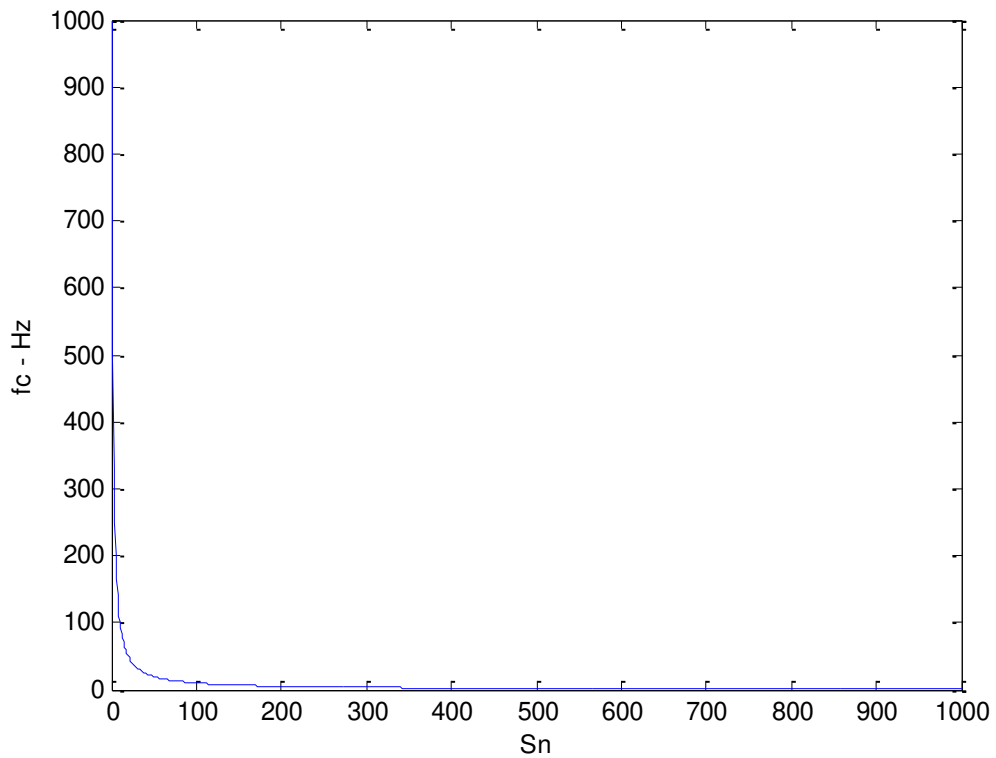


Figure 3.27 The plot of f_c versus Sn for $N = 1000$

The frequency resolution obtained for f_c is poor at higher frequencies, lower Sn values. However, the frequency resolution can easily be increased by increasing the sample rate N . In the proposed method the frequency resolution is dependent on the sample rate. The following examples give the results obtained with fixed Δk and

variable Δk for the 160ms duration of the vowel /a/ speech signal shown in Figure 3.28, where $N=8000$.

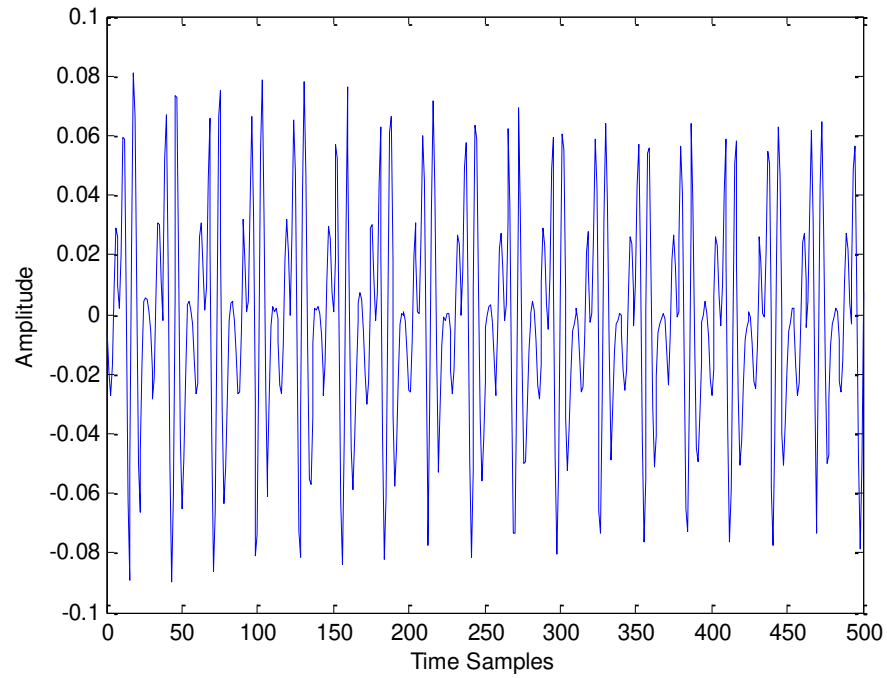


Figure 3.28 The original speech waveform for vowel /a/.

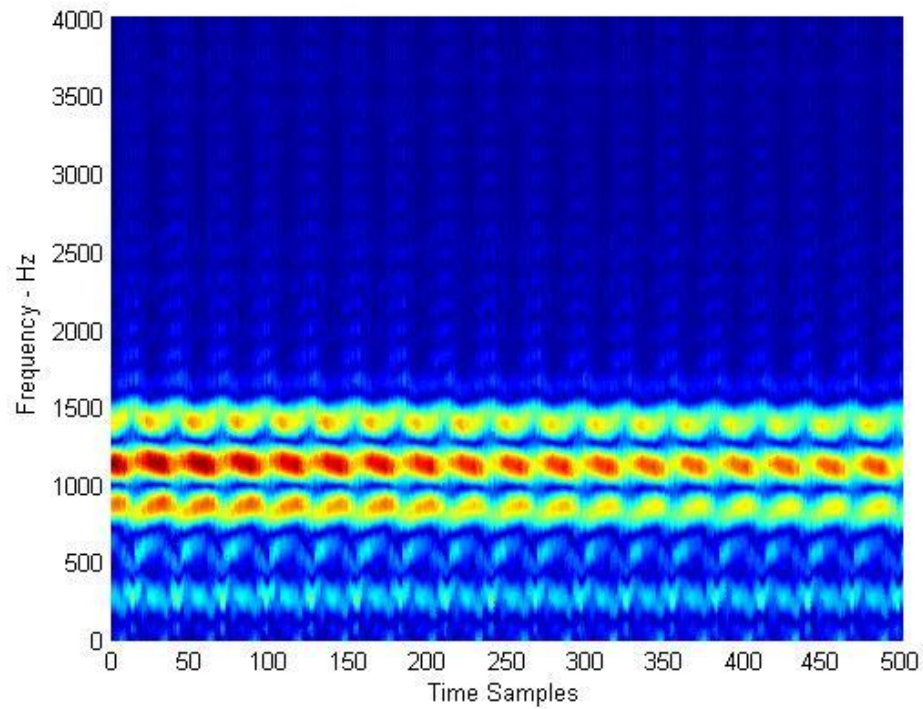


Figure 3.29 The frequency resolution obtained with fixed Δk .

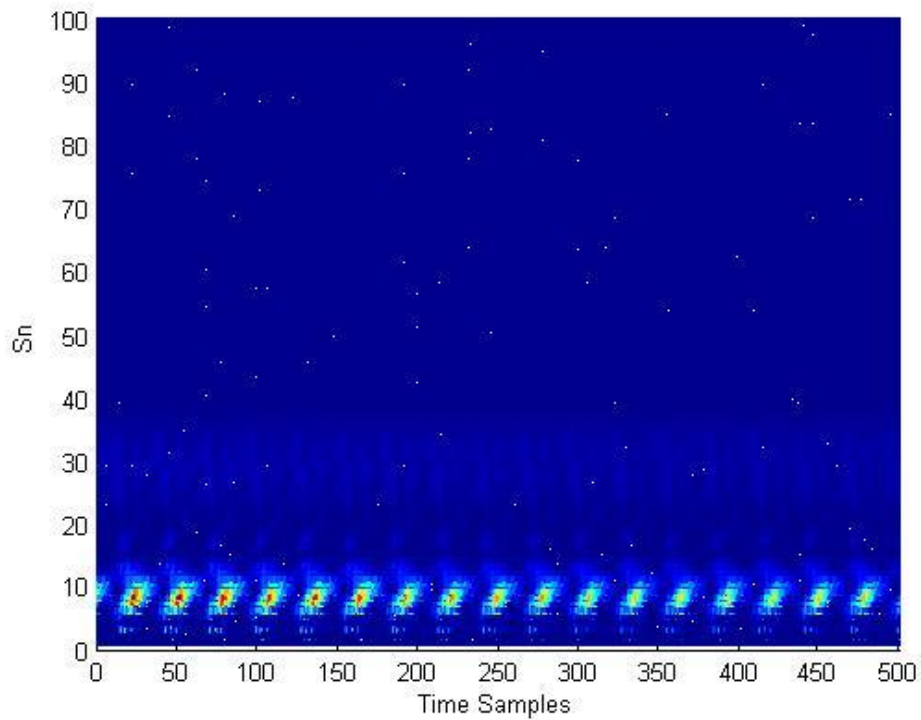


Figure 3.30 The frequency resolution obtained with variable Δk .

The Figures 3.29 and 3.30 show the time frequency resolution obtained for fixed and variable Δk respectively, the S_n can easily be converted to the corresponding frequencies f_c in order to compare the results obtained in Figures 3.29 and 3.30. The frequency resolution obtained for fixed Δk is better than variable Δk , however, the time resolution is better for variable Δk and the frequency resolution is dependent on the sample rate N which is defined with S_n .

The variable Δk give different approach to the time frequency signal analysis, and will be used to obtain the time frequency resolutions for speech vowel signals which are non-stationary signals, and will be explained in the next chapter.

CHAPTER FOUR

SPEECH VOWEL CLASSIFICATION BY USING AUDITORY MOTIVATED DISCRETE TIME-FREQUENCY SIGNAL ANALYSIS METHOD

4.1 The Mel Frequency Cepstral Coefficients (MFCC)

The speech signals have high variability due to the speakers, therefore performing the speech recognition by computer system is very difficult task (Rabiner&Yuang,1993; Kasabov,1996). The human brain is still superior to many technical solutions. Therefore the human auditory system based methods are widely used for speech recognition applications. The most popular one is the Mel Frequency Cepstral Coefficients (MFCC).

The Mel frequency is the scale derived from auditory system. The relation between normal frequency and mel frequency scale is defined in the Equation 4.1 (Picone, 1993).

$$f_{mel} = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (4.1)$$

The plot of the mel scale versus normal frequency scale is given in Figure 4.1. In the derivation of MFCCs the triangular filters are used based on the mel scale obtained in Equation 4.1. The mel scale filter banks and block diagram for the computation of the MFCCs are shown in Figures 4.2 and 4.3 respectively.

Chatterjee&Kleijn (2011) generalized the Mel frequency scale given in Equation 4.1 by introducing the α which is the warping factor and affects the extent of warping.

$$f_{mel} = 2595 \log_{10} \left(1 + \frac{f}{\alpha} \right) \quad (4.2)$$

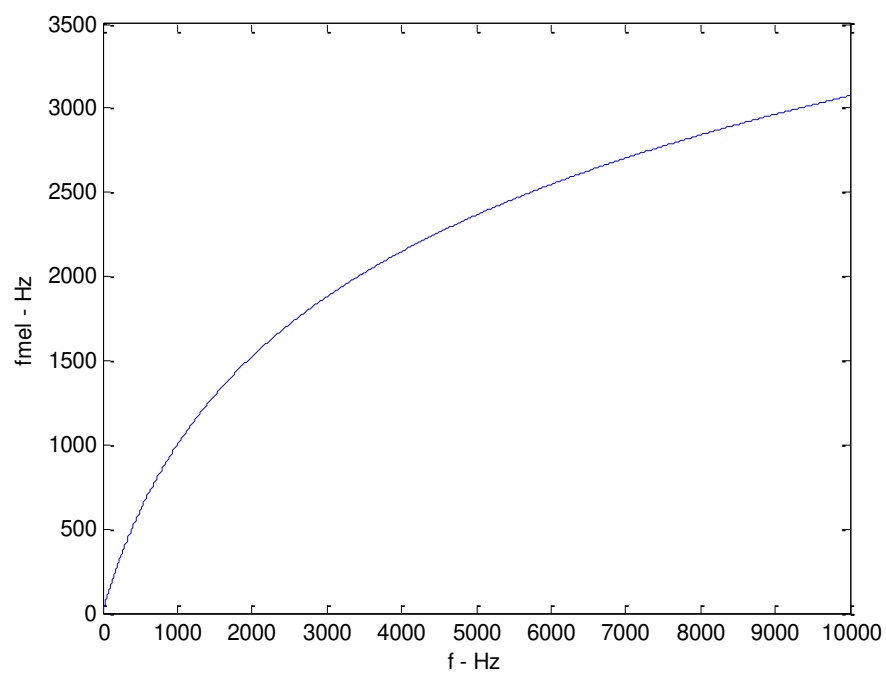


Figure 4.1 The plot of the mel frequency scale versus normal frequency scale f .

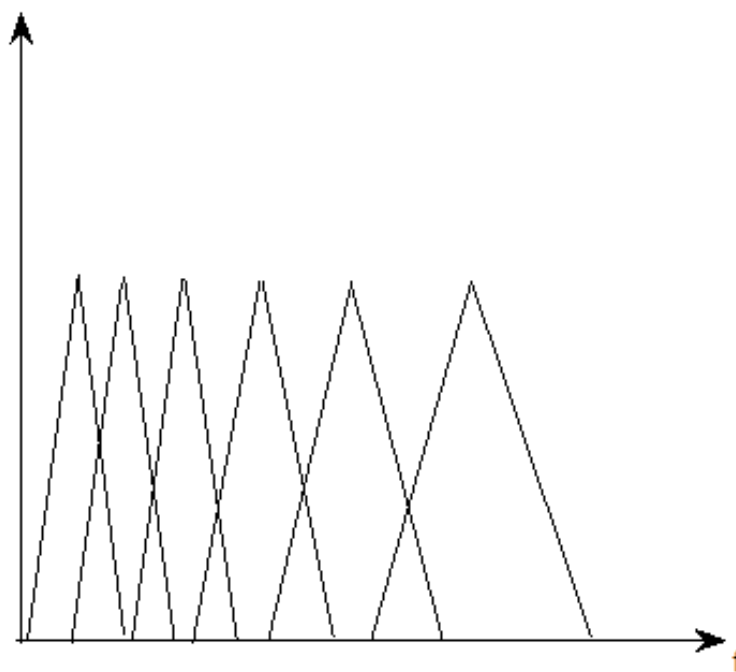


Figure 4.2 Approximate Mel filter banks

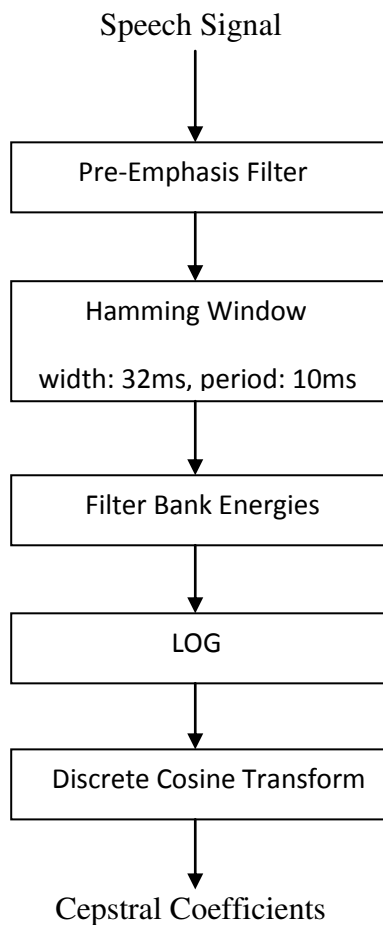


Figure 4.3 The standard approach for acoustic feature extraction
(Redrawn from: (Pavez&Silva, 2011))

In the standard acoustic feature extraction approach, the speech signal is pre-emphasized by the filter, then the hamming window is used to extract the features for the short duration in time. The mel filter bank energies together with logarithm and discrete cosine transform are employed to obtain the feature vectors.

The MFCCs are based on short term speech analysis methods where the speech signal is assumed as stationary for short time durations. Because they use auditory frequency scale the features extracted with MFCC are used widely for speech recognition applications. In order to classify the speech signals the Hidden Markov Model (HMM) or Artificial Neural Networks (ANN) are widely used in literature (Chatterjee&Kleijn, 2011; Dusan, 2007; Ali et al., 2002; Zahorian&Nossair, 1999).

4.2 Application of Auditory Motivated Discrete Time Frequency Signal Analysis Method to the Vowel Speech Signals

The spectral information is widely used in speech recognition applications. The short term Fourier coefficients are used to extract the patterns from the speech signals. The vowel signals are characterized by the fundamental frequency F_0 , and upper formants F_1 , F_2 and F_3 .

Different hearing tests with subjects who have normal hearing ability were performed in the literature to determine which of the formant frequencies are important for vowel identification (Zahorian&Zhang, 1992; Shannon et al., 1995; Sakayori et al., 2002). The formant frequencies F_1 and F_2 are important to identify the vowels (Sakayori et al., 2002). However, for the same speaker the formant frequencies F_1 and F_2 can be used to identify the vowels, but in the case of multiple speakers there exist overlap between the formant frequencies, which makes it difficult to identify the vowels based on the formant frequencies. Therefore additional information is needed to identify the vowels. According to Sakayori et al. (2002) in order to minimize the change of phonetic quality, each formant frequency should be moved by %1-12 upward by one octave increase in the fundamental frequency F_0 . This shows that vowel patterns can be dependent on the fundamental frequency F_0 . However Shannon et al. (1995) suggested that F_0 cue is not essential for vowel identification based on the hearing tests performed with different listeners.

Zahorian&Zhang (1992) suggested that spectral envelope is important information for vowel identification. Zahorian&Jagharghi (1993) showed that computational vowel classification based on spectral envelope is superior to the information on the F_0 , F_1 , F_2 and F_3 . According to Sakayori et al. (2002) the human auditory system may identify the vowels according to the spectral shapes and formant frequencies F_1 and F_2 in the critical spectral regions.

Kameoka et al. (2010) suggested the joint estimation of the fundamental frequency F_0 and spectral envelope. In the following Figures the results of the proposed method to the different speech vowel signals are given.

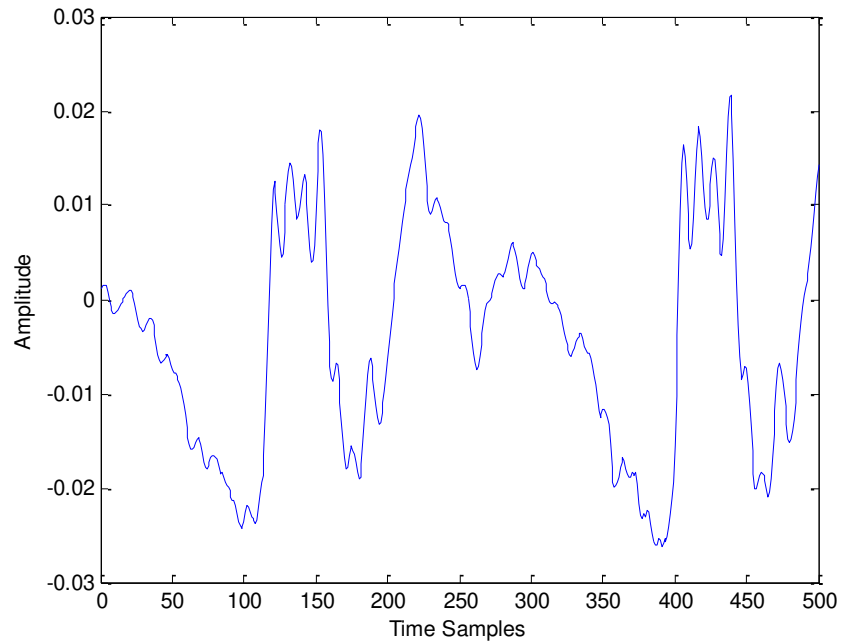


Figure 4.4 The vowel /i/ signal recorded from male speaker

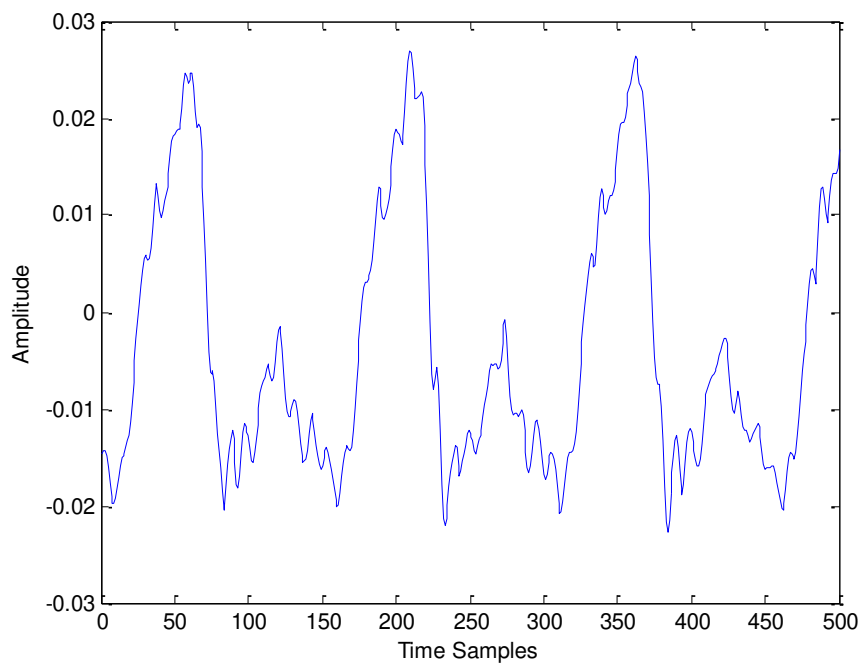


Figure 4.5 The vowel /i/ signal recorded from female speaker

The Figures 4.4 and 4.5 show the vowel /ii/ signals for 40ms duration, the signal is sampled at $N=20000$ sample rate. The frequency scale fc obtained for the Sn values from 1 to 200 is given in Figure 4.6.

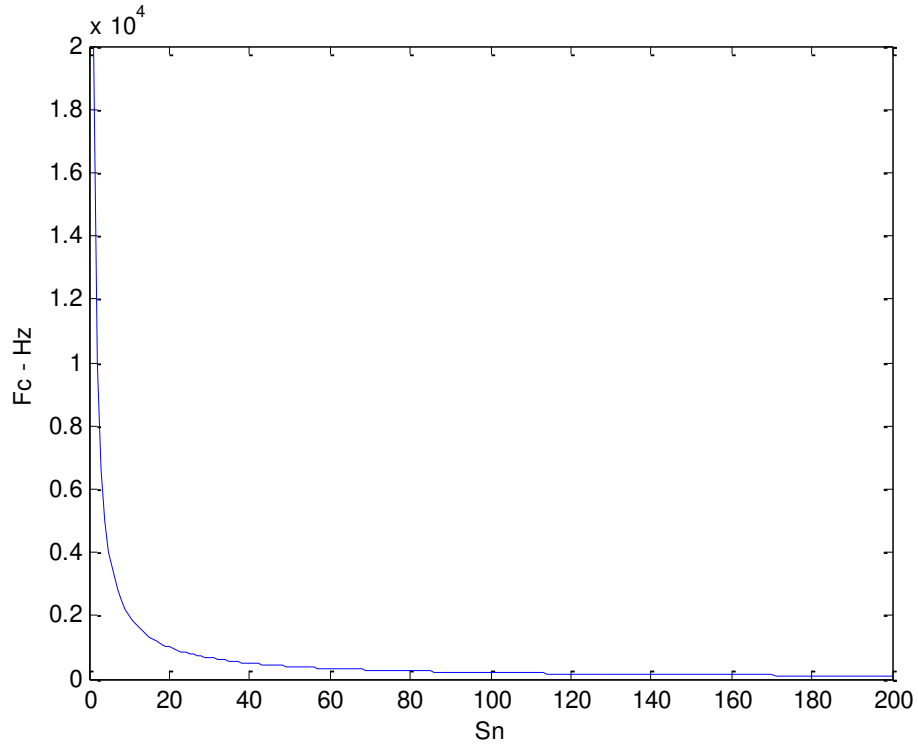


Figure 4.6 The plot of fc versus Sn values from 1 to 200, $N=20000$.

The frequency scale fc is similar to the auditory motivated mel frequency scale given in Figure 4.1, where the low frequencies are sampled closer than high frequencies.

In order to compare the results obtained with the proposed method, the analysis is made also with wavelet transform, where the ‘haar’ and ‘morlet’ mother wavelet functions shown in Figures 4.7 and 4.8 are used to obtain the time frequency resolutions.

The following Figures give the results obtained with the proposed method and with wavelet transform.

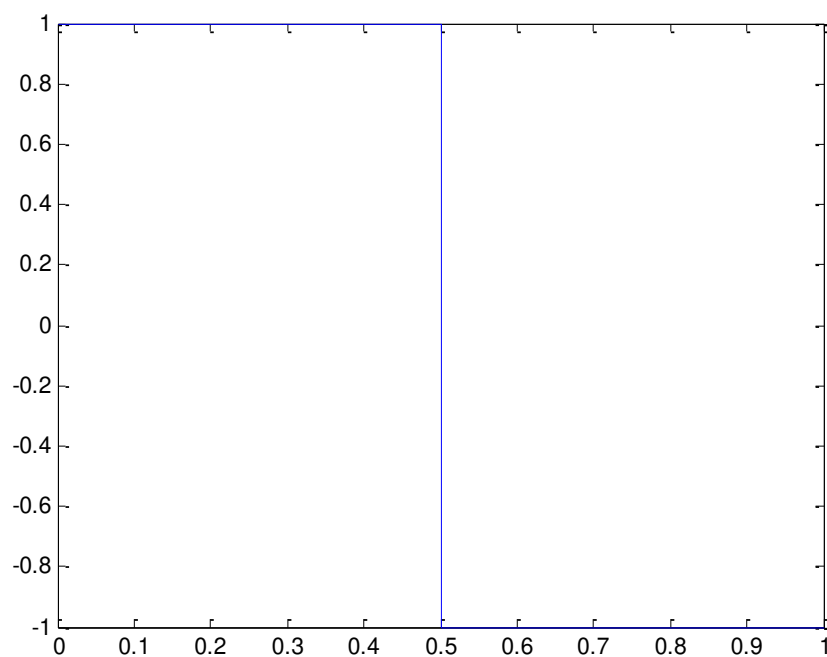


Figure 4.7 The 'Haar' Mother Wavelet Function.

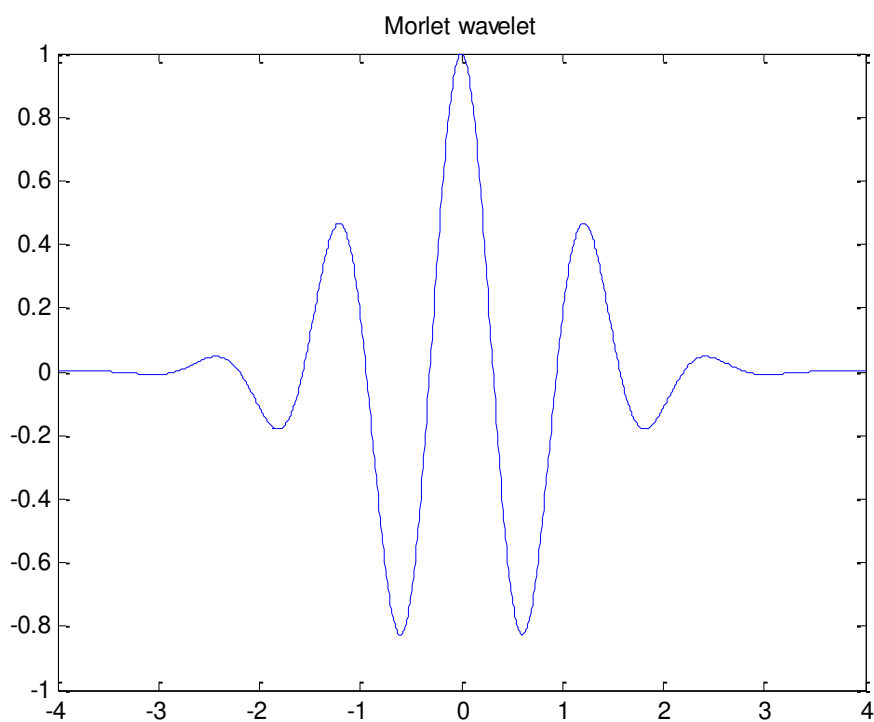


Figure 4.8 The 'Morlet' Mother Wavelet Function

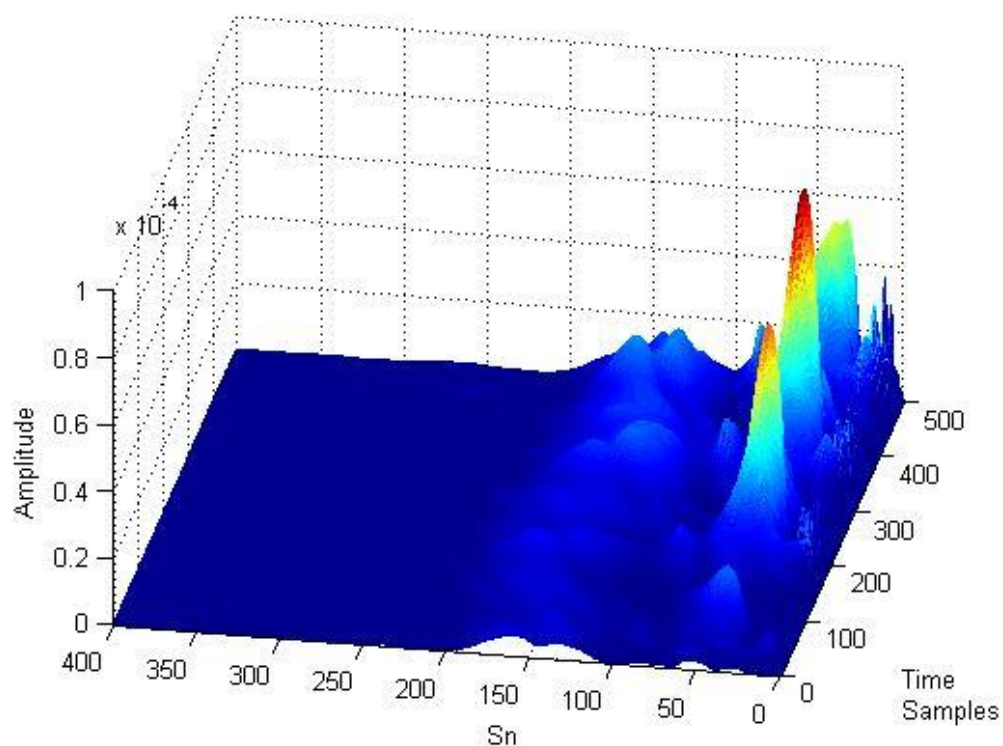


Figure 4.9 The Time Frequency resolution obtained with the proposed method, applied to the speech signal recorded from male speaker.

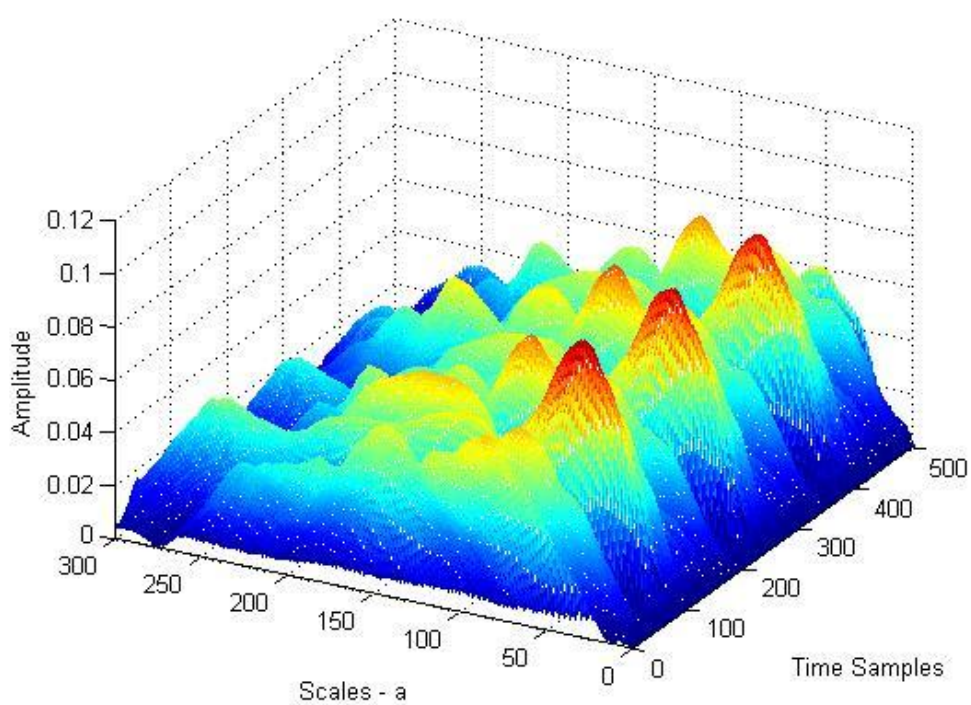


Figure 4.10 The time frequency resolution obtained with 'Haar' wavelet transform, applied to the speech signal recorded from male speaker.

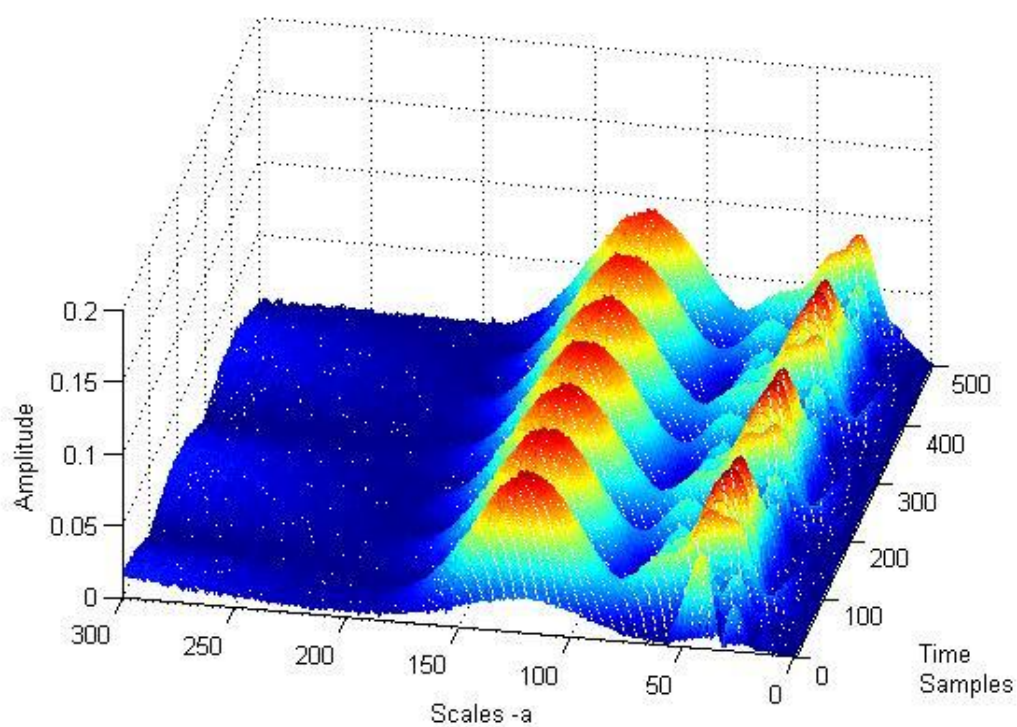


Figure 4.11 The time frequency resolution obtained with 'Morlet' wavelet transform, applied to the speech signal recorded from male speaker.

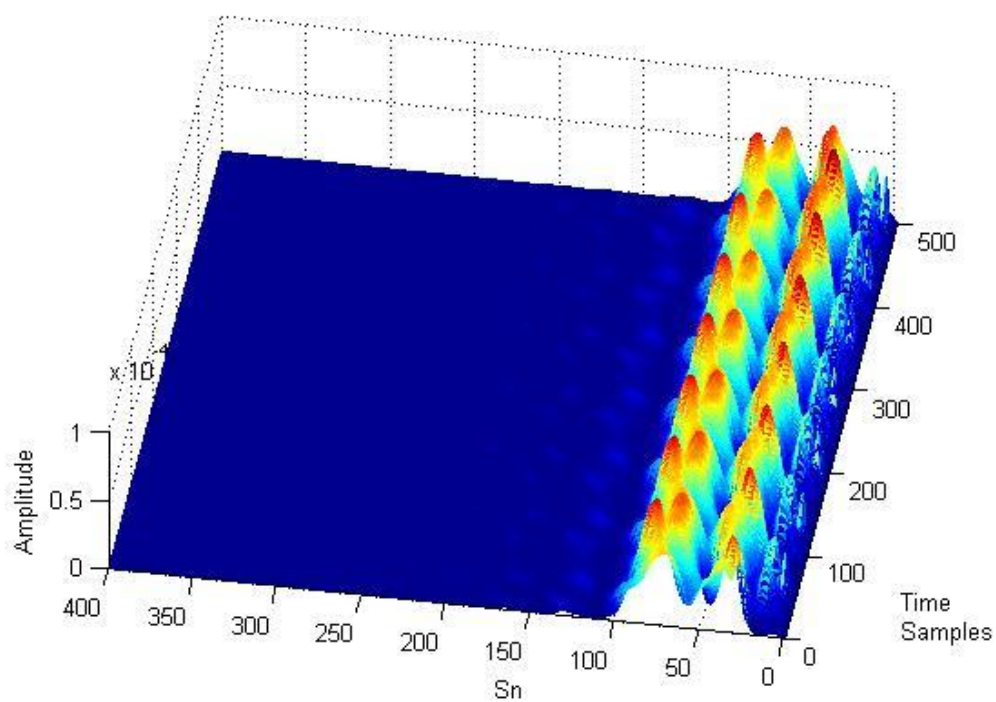


Figure 4.12 The Time Frequency resolution obtained with the proposed method, applied to the speech signal recorded from female speaker.

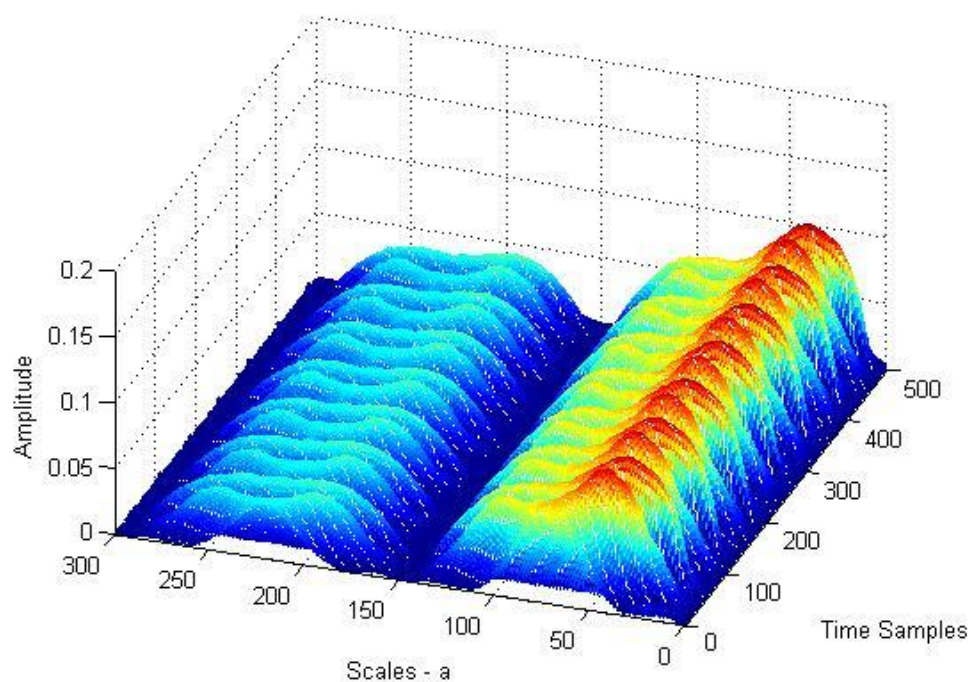


Figure 4.13 The time frequency resolution obtained with 'Haar' wavelet transform, applied to the speech signal recorded from female speaker.

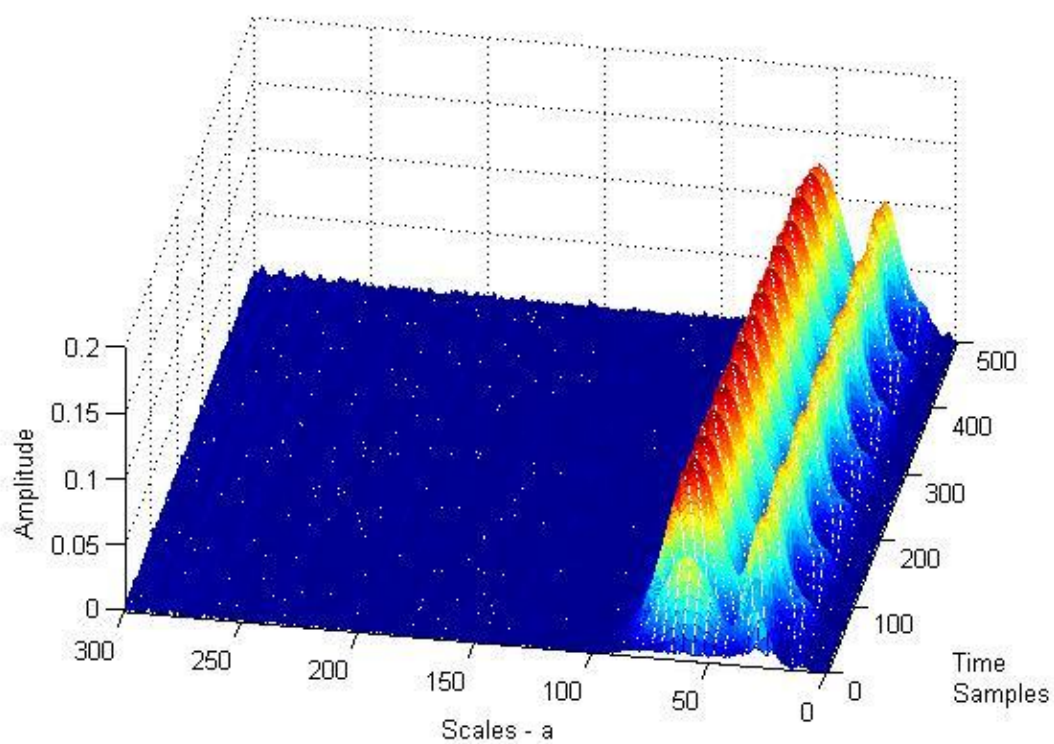


Figure 4.14 The time frequency resolution obtained with 'Morlet' wavelet transform, applied to the speech signal recorded from female speaker.

As can be seen from Figures 4.9-4.14, the window function has directly effect on the obtained time frequency resolution. Because the proposed method is independent from the window function, the results are directly based on the recorded signal structure. In the results obtained with the proposed method, for two different speakers the similar spectral envelope is obtained at higher frequencies, lower Sn values for the same vowel signal, where in the case of wavelet transform, the spectral envelopes are different for two different speakers.

In order to detect the spectral envelope, the spectral peaks obtained for each Sn value can easily be determined by taking the maximum values of the amplitudes inside 40ms duration, which gives the spectral peaks distribution over specific time duration, and can be thought, as approximate spectral envelope. The following Figures show the detected spectral peaks distribution for both speakers, with the proposed method and wavelet transform.

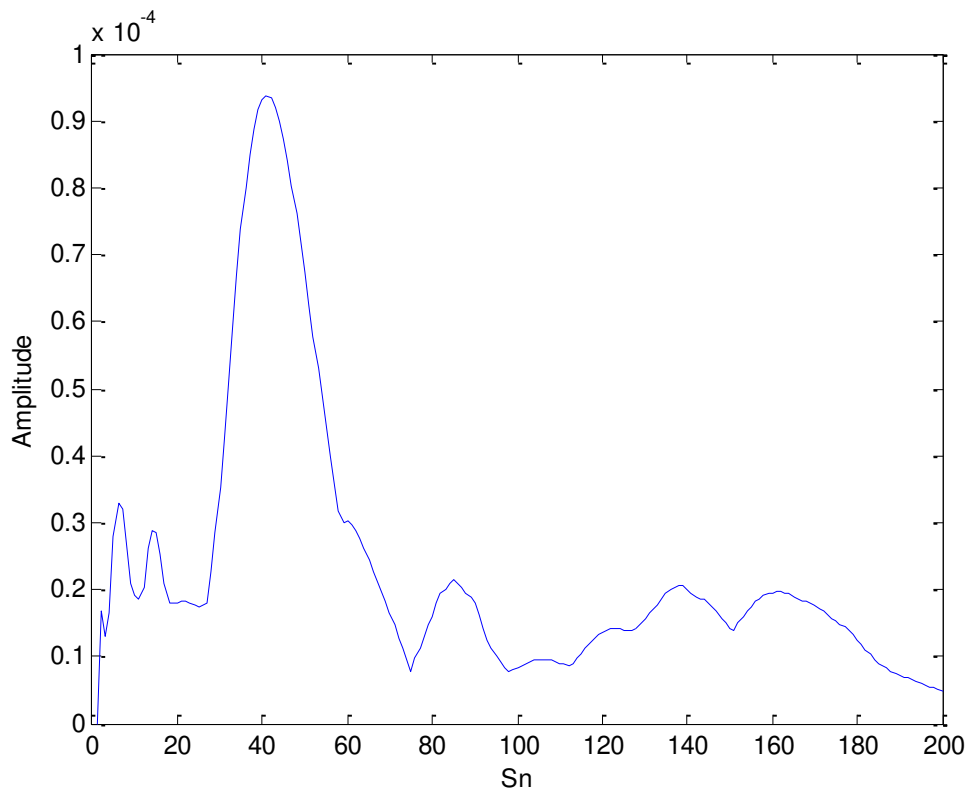


Figure 4.15 The detected spectral peaks distribution with the proposed method, for male speaker.

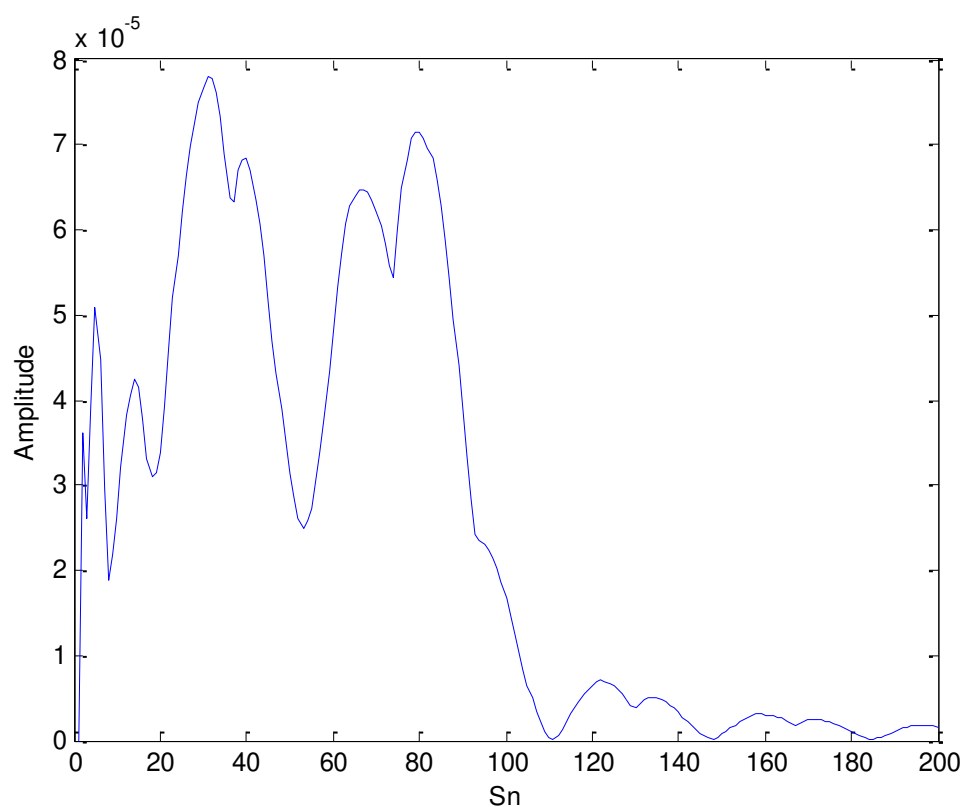


Figure 4.16 The detected spectral peaks distribution with the proposed method, for female speaker.

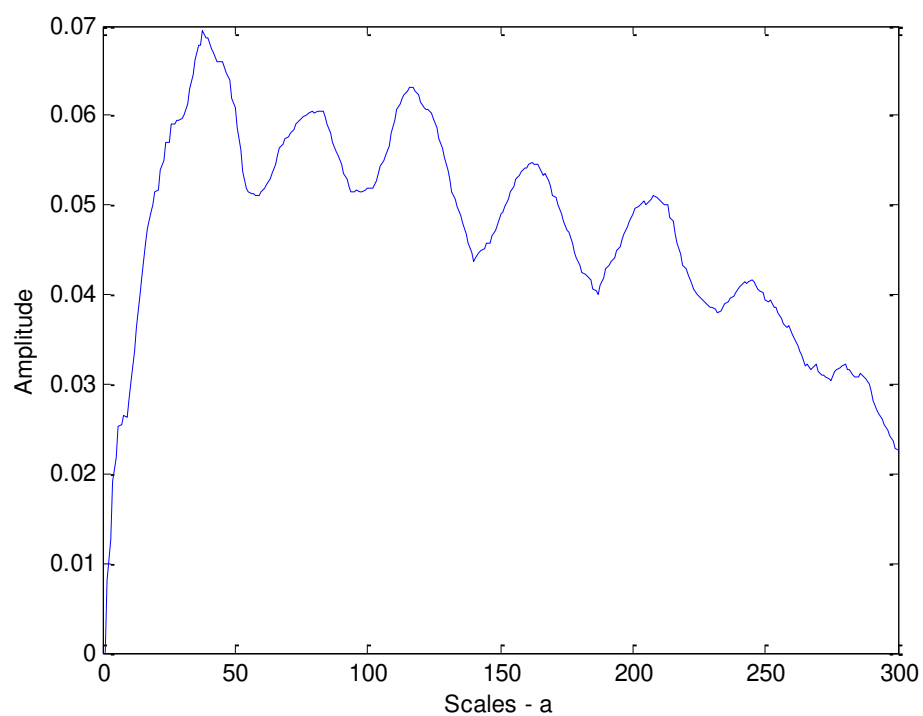


Figure 4.17 The detected spectral peaks distribution with the 'Haar' wavelet transform, for male speaker.

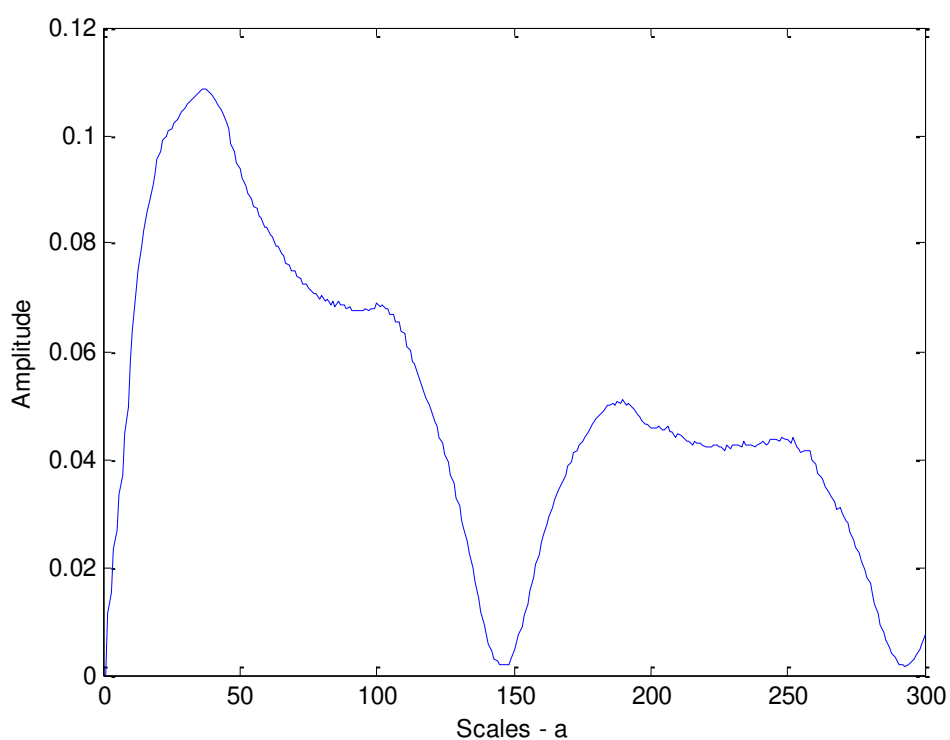


Figure 4.18 The detected spectral peaks distribution with the 'Haar' wavelet transform, for female speaker.

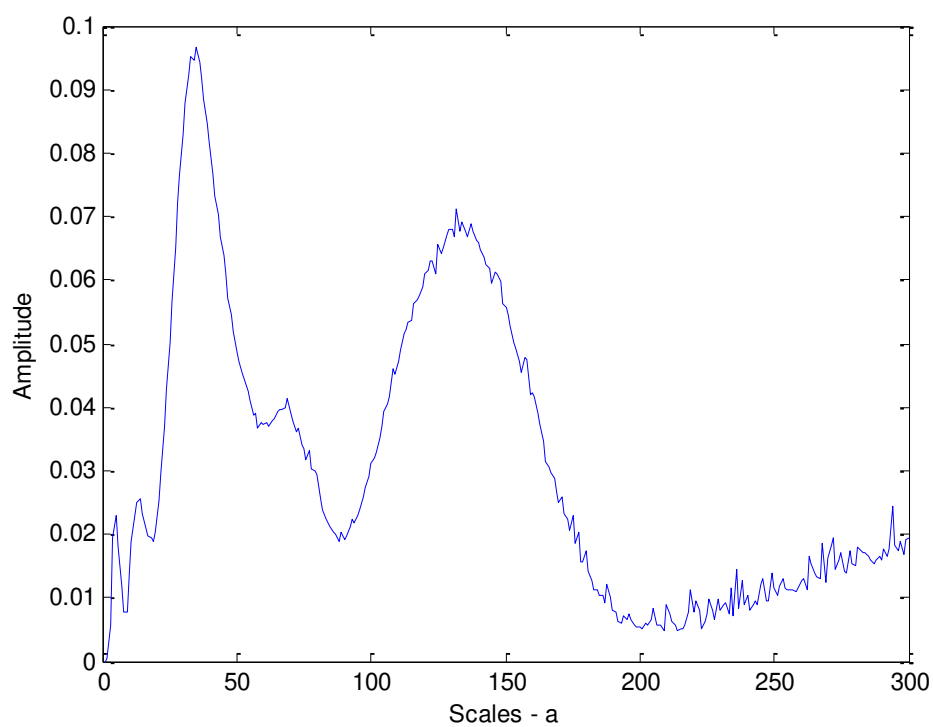


Figure 4.19 The detected spectral peaks distribution with the 'Morlet' wavelet transform, for female speaker.

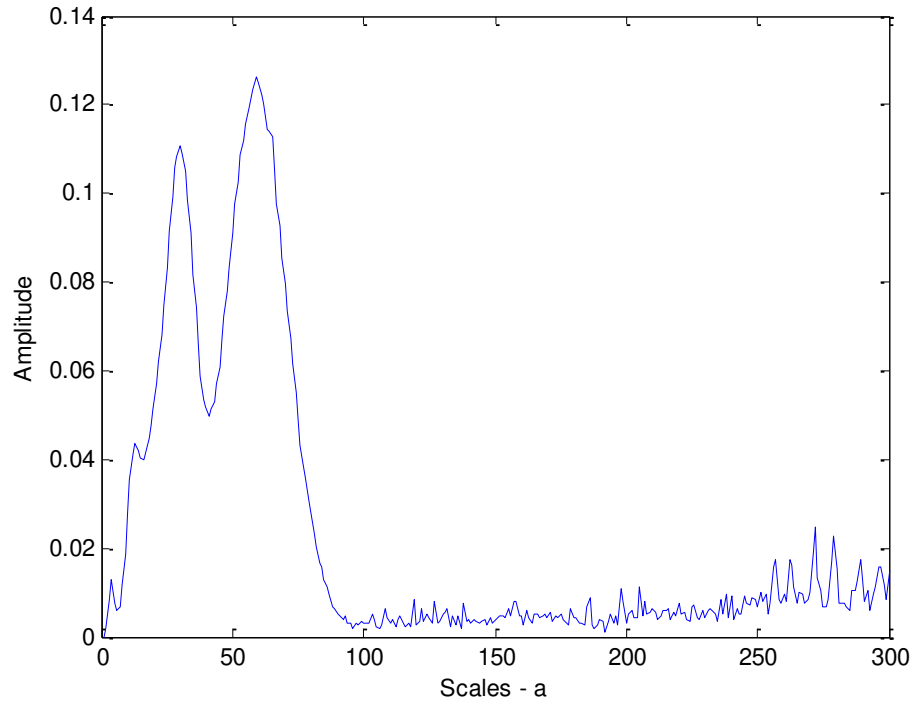


Figure 4.20 The detected spectral peaks distribution with the 'Morlet' wavelet transform, for female speaker.

The Figures 4.15-4.20 show the spectral peaks distribution obtained for two different speakers. In the case of proposed method the similar spectral peaks distributions are obtained easily for two different speakers, for higher frequencies. The distributions of spectral peaks are directly relational to the fundamental frequency F_0 which supports the Sakayori et al. (2002). In the case of higher F_0 the spectral peaks distribution is narrow, and in the case of lower F_0 the distribution is wider. The results obtained with the proposed method are good evidence for the existing similar spectral envelope for the identification of the vowels.

The proposed method is applied to the remaining Turkish vowels /a/, /o/, /u/, /e/, /i/, /ö/, /ü/, recorded from male and female speakers, and the following Figures show the obtained time frequency representations and the detected spectral peaks distribution (SPD) for each vowel.

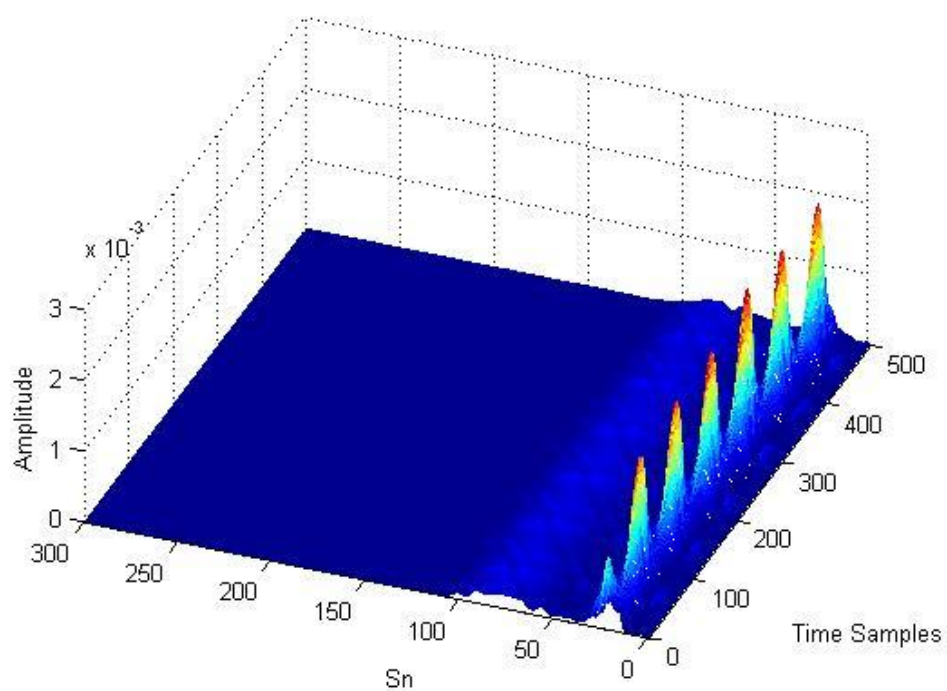


Figure 4.21 The time frequency resolution for vowel /a/, female speaker

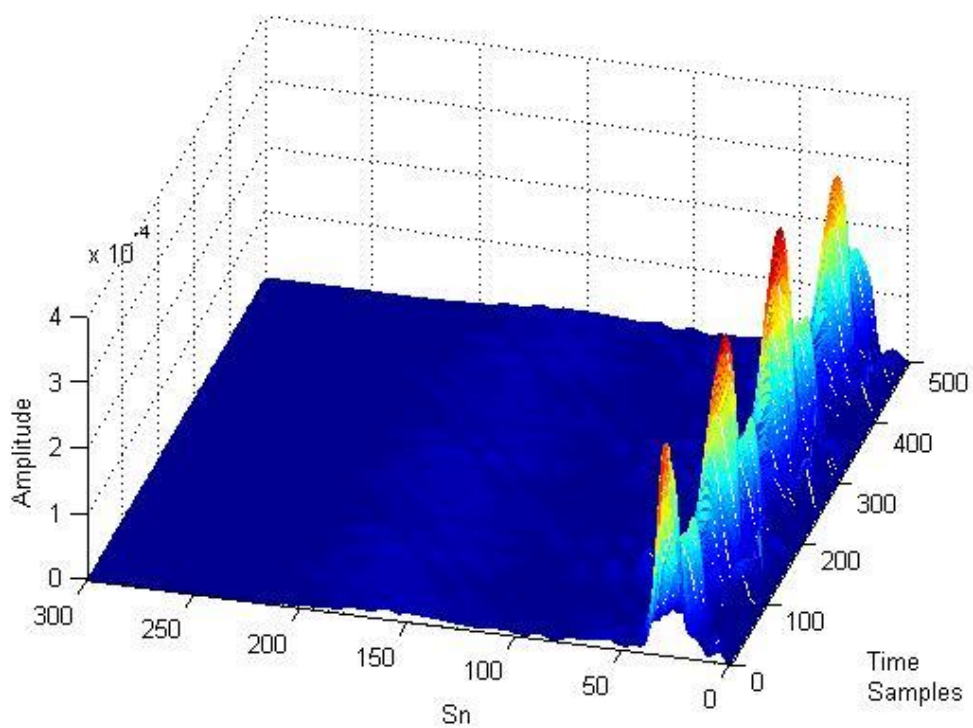


Figure 4.22 The time frequency resolution for vowel /a/, male speaker

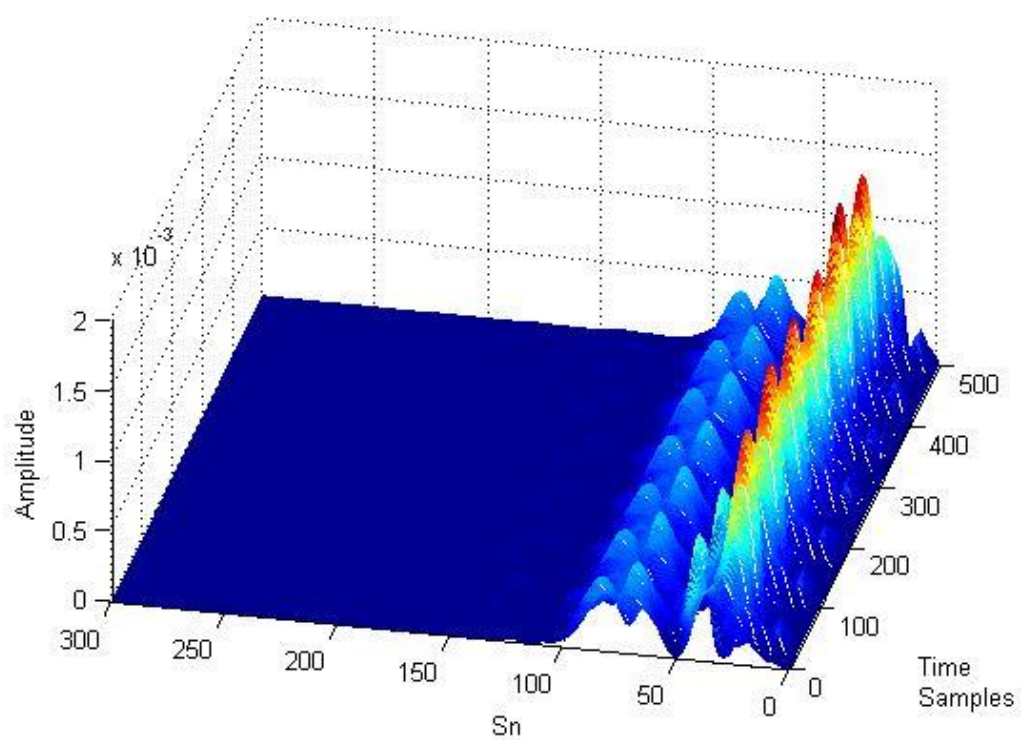


Figure 4.23 The time frequency resolution for vowel /o/, female speaker

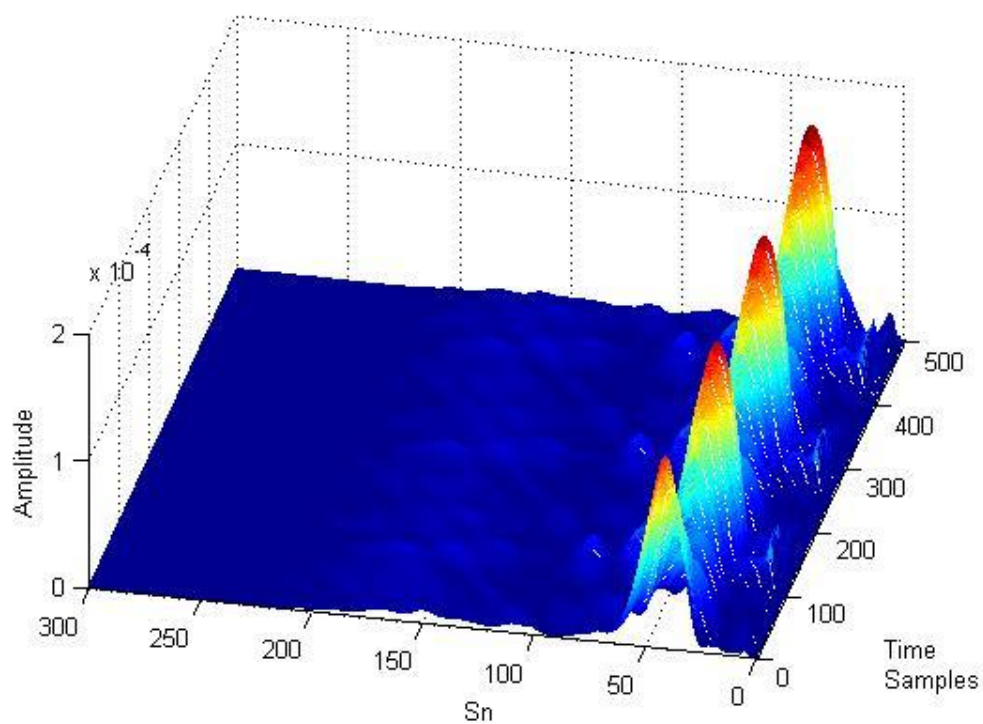


Figure 4.24 The time frequency resolution for vowel /o/, male speaker

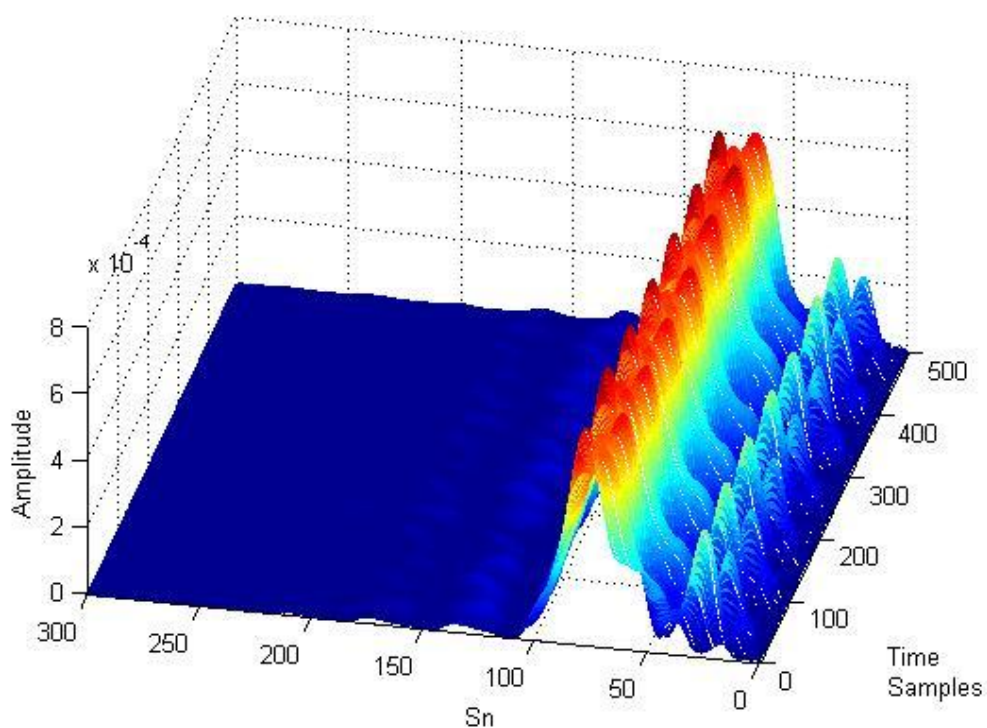


Figure 4.25 The time frequency resolution for vowel /u/, female speaker

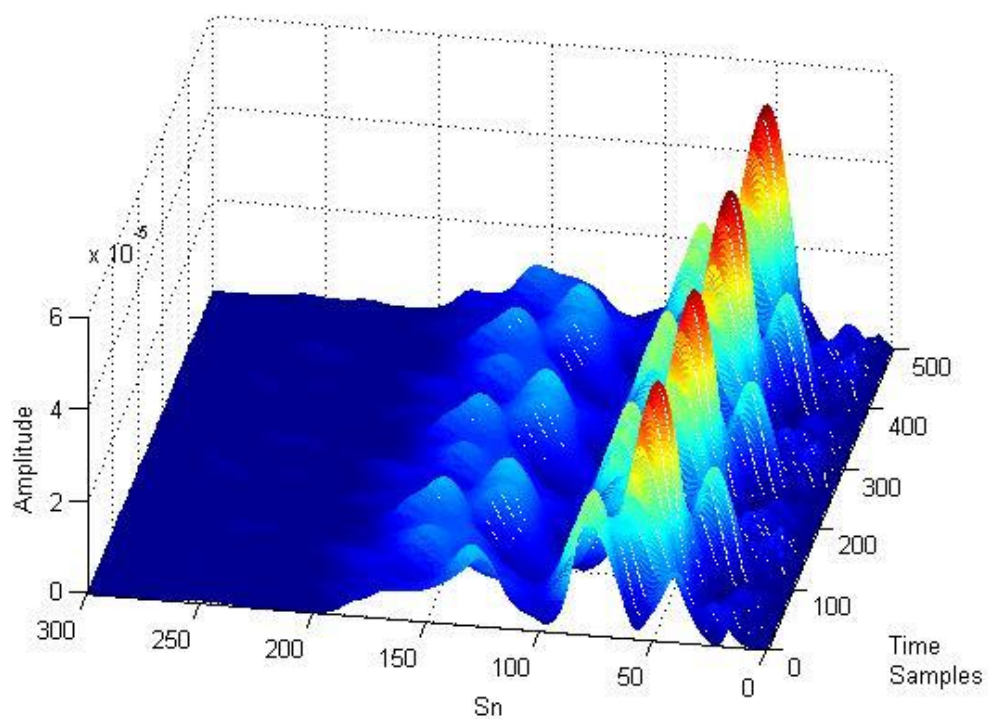


Figure 4.26 The time frequency resolution for vowel /u/, male speaker

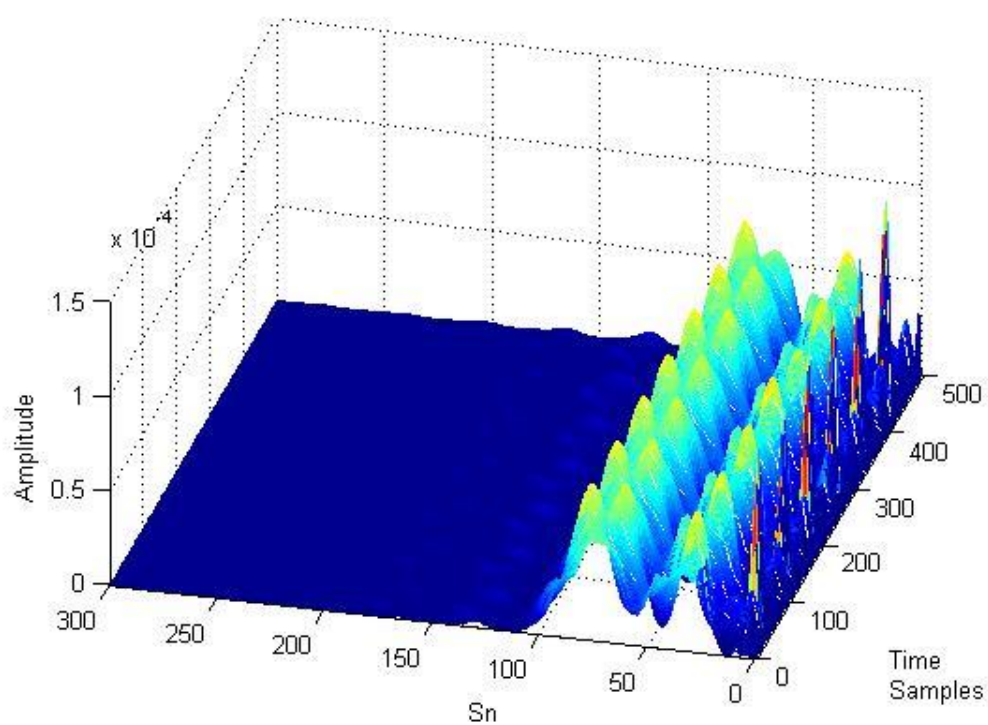


Figure 4.27 The time frequency resolution for vowel /e/,female speaker

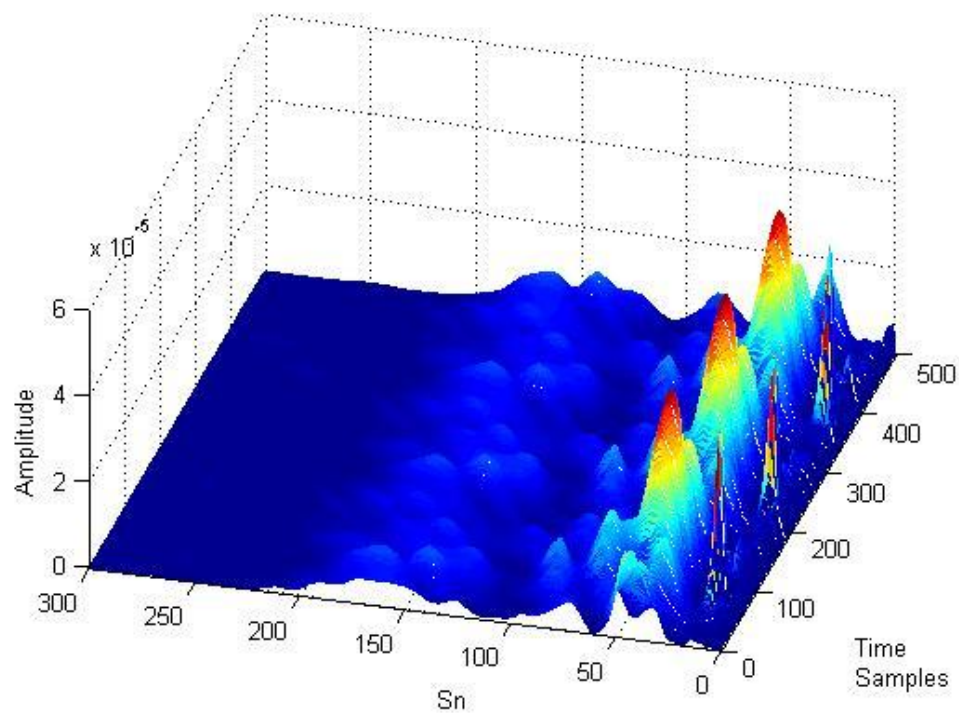


Figure 4.28 The time frequency resolution for vowel /e/, male speaker

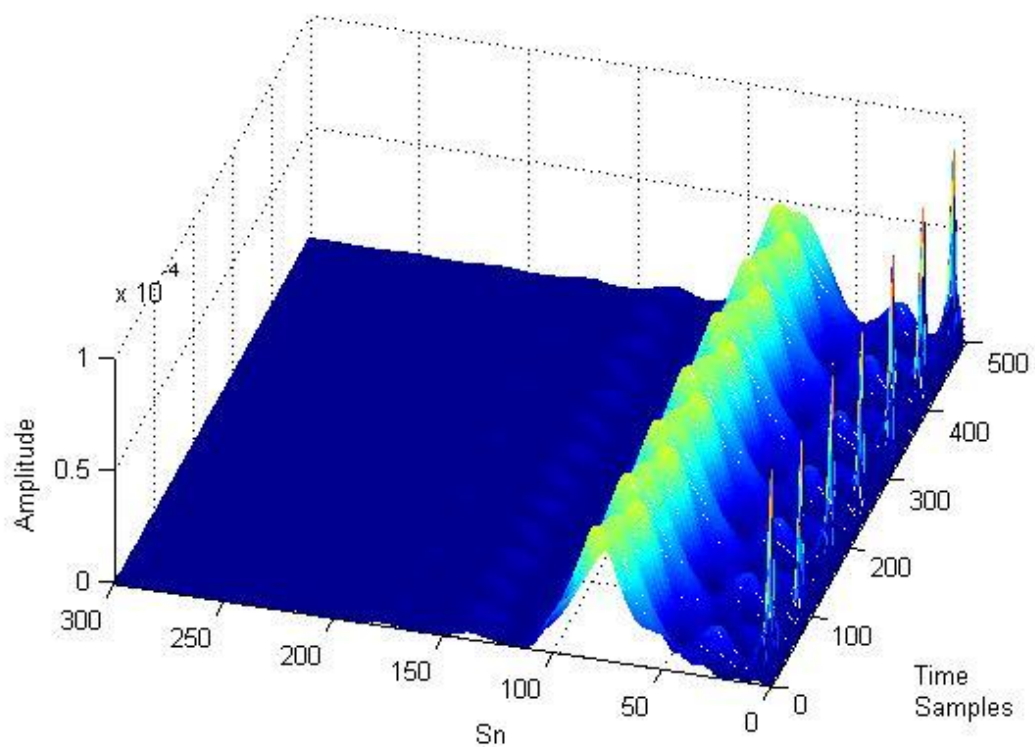


Figure 4.29 The time frequency resolution for vowel /i/, female speaker

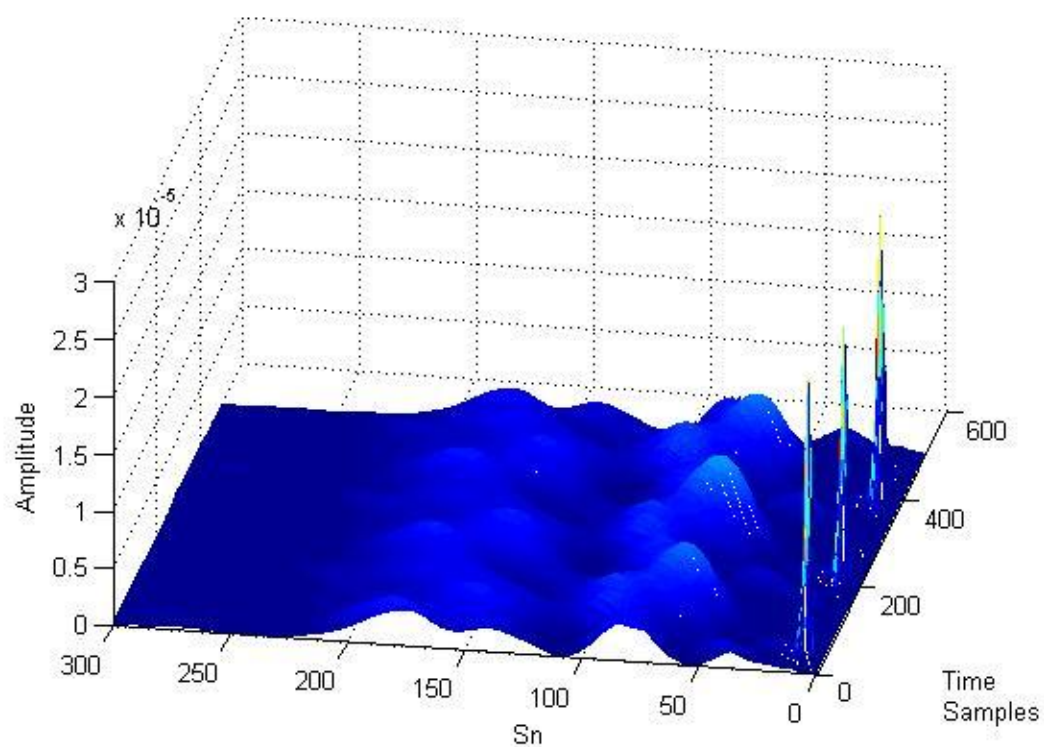


Figure 4.30 The time frequency resolution for vowel /i/, male speaker

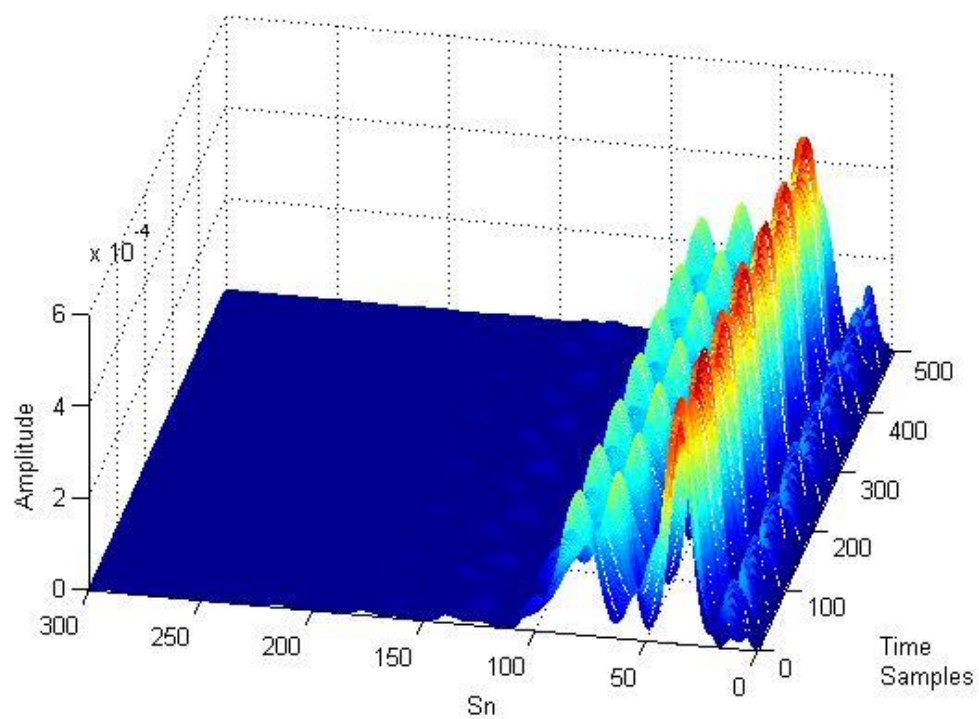


Figure 4.31 The time frequency resolution for vowel /*ö*/, female speaker

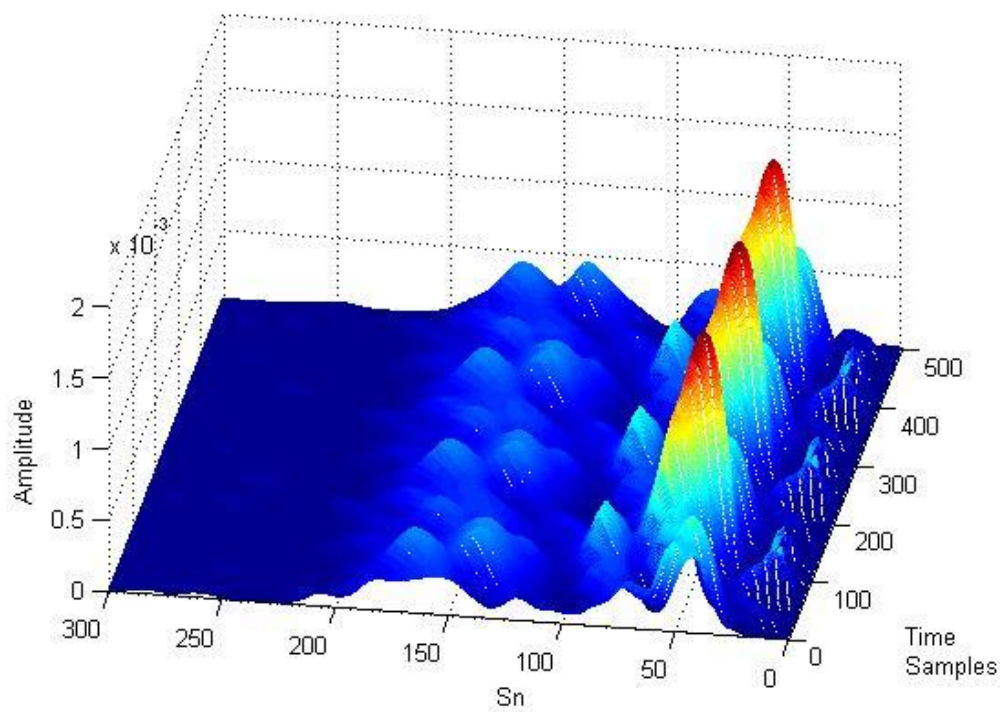


Figure 4.32 The time frequency resolution for vowel /*ö*/, male speaker

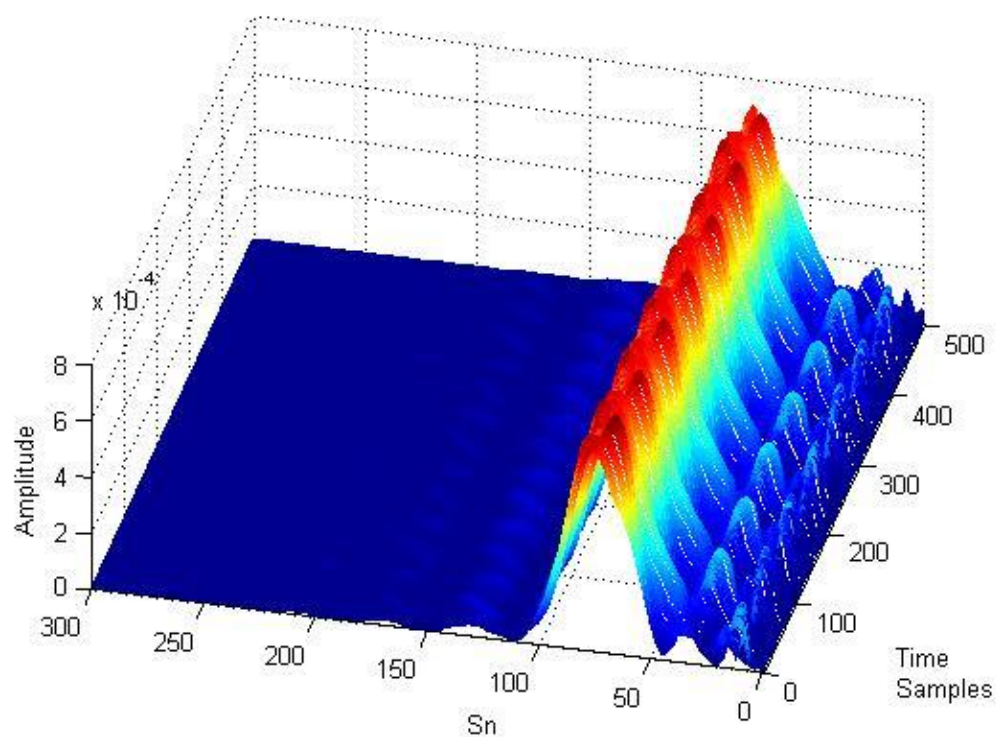


Figure 4.33 The time frequency resolution for vowel /ü/, female speaker

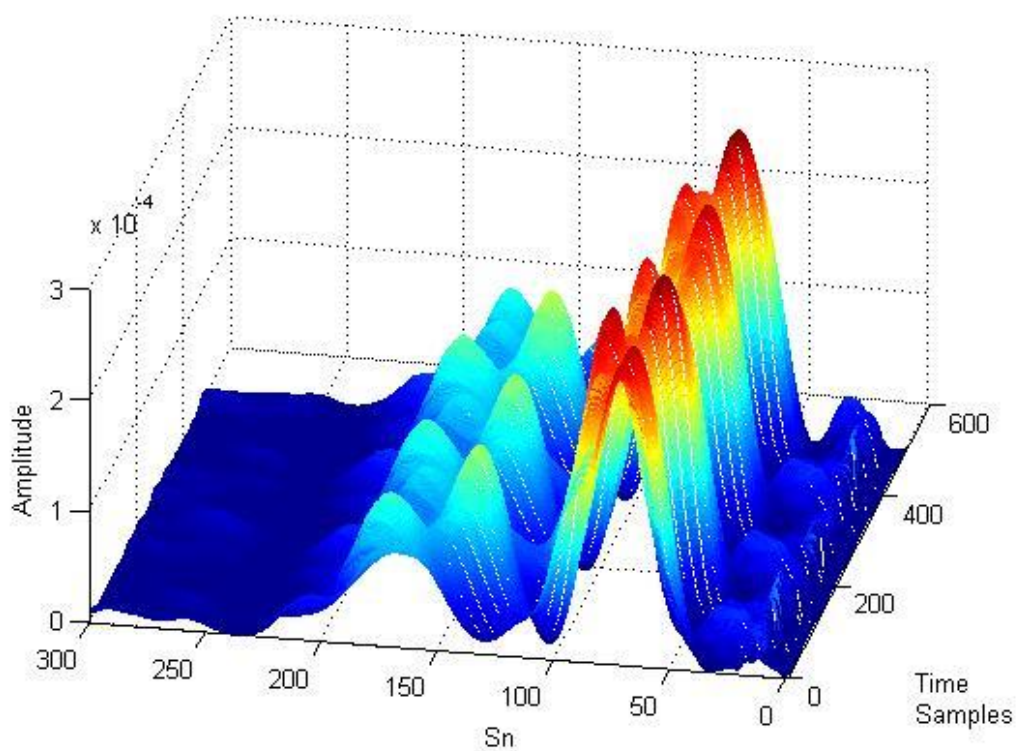


Figure 4.34 The time frequency resolution for vowel /ü/, male speaker

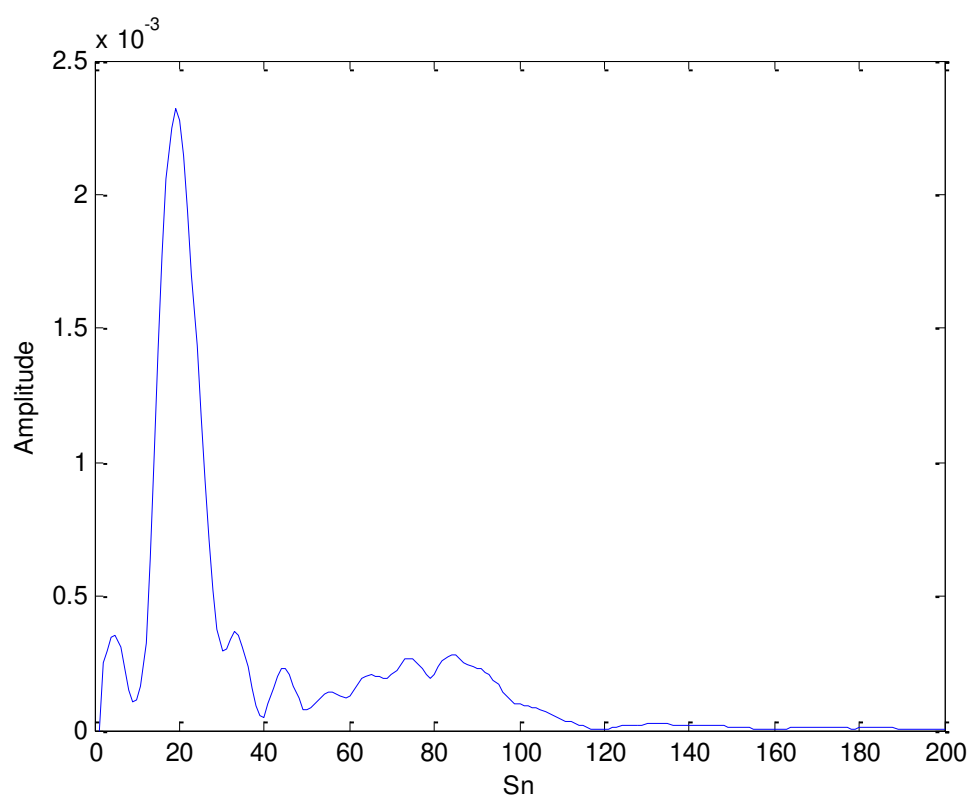


Figure 4.35 The detected SPD for vowel /a/, female speaker

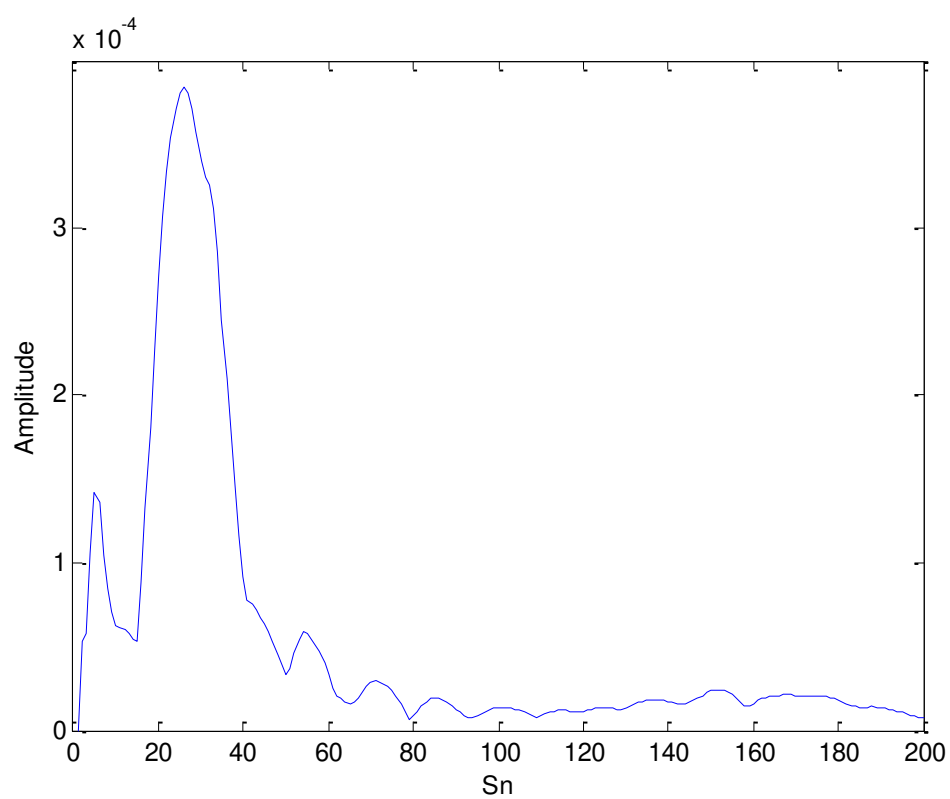


Figure 4.36 The detected SPD for vowel /a/, male speaker

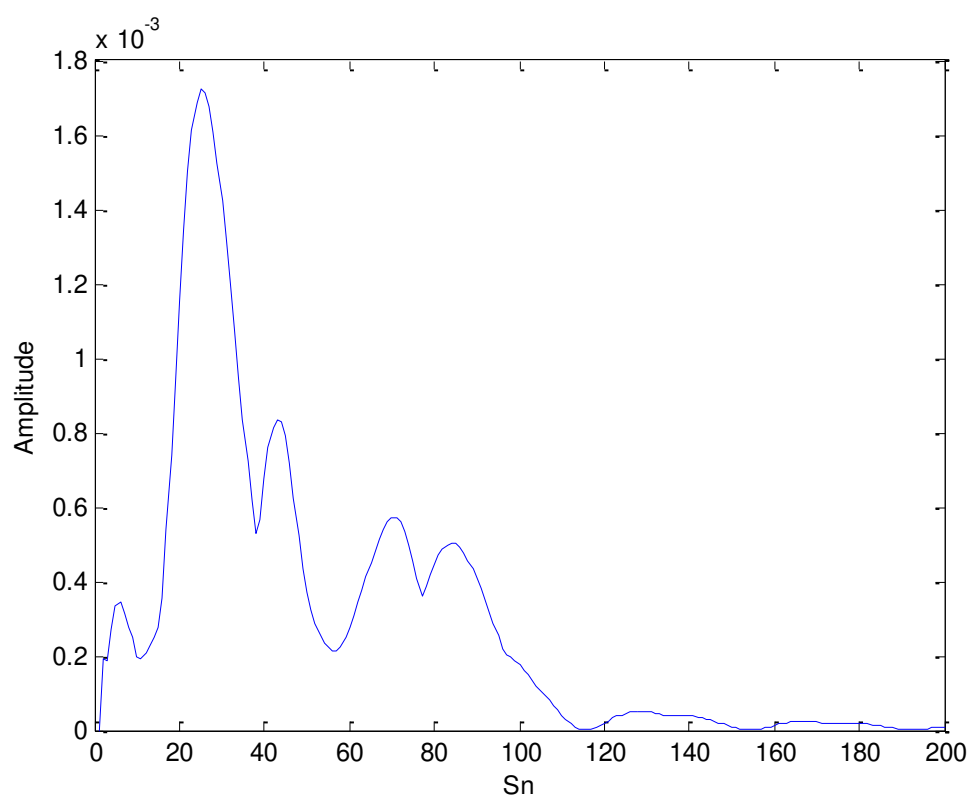


Figure 4.37 The detected SPD for vowel /o/, female speaker

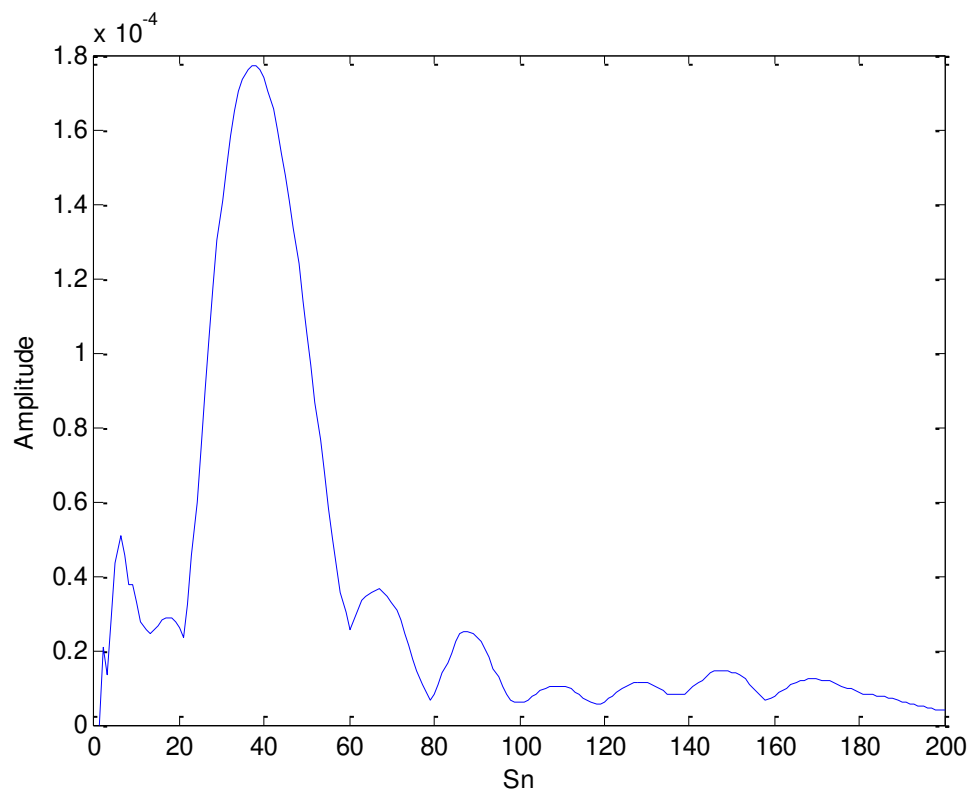


Figure 4.38 The detected SPD for vowel /o/, male speaker

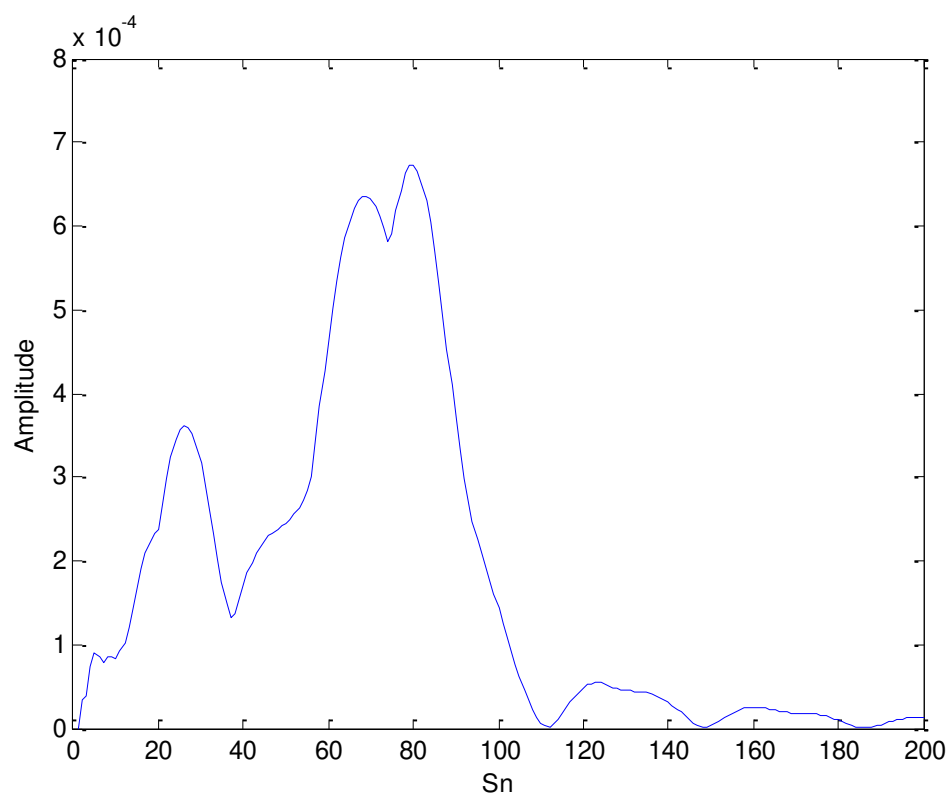


Figure 4.39 The detected SPD for vowel /u/, female speaker

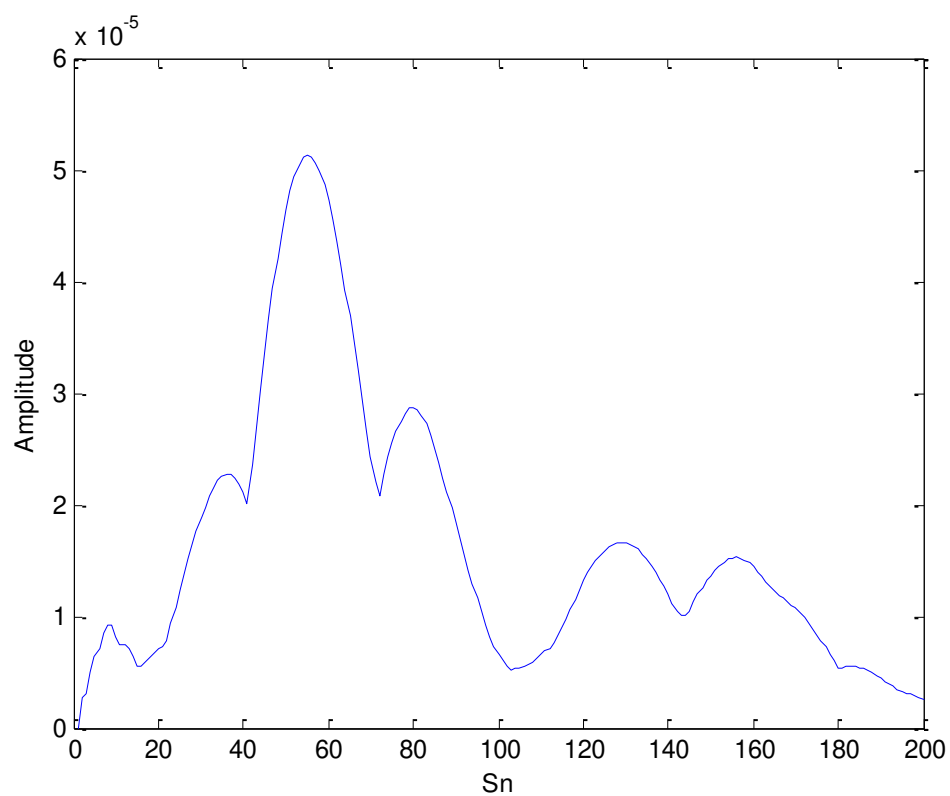


Figure 4.40 The detected SPD for vowel /u/, male speaker

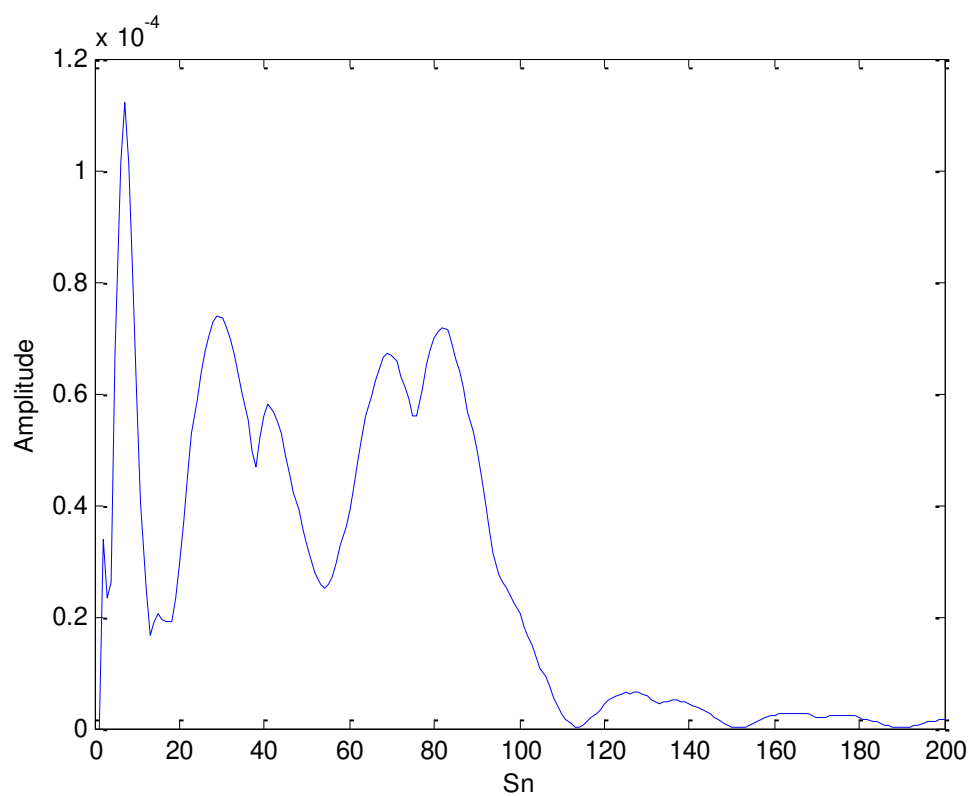


Figure 4.41 The detected SPD for vowel /e/, female speaker

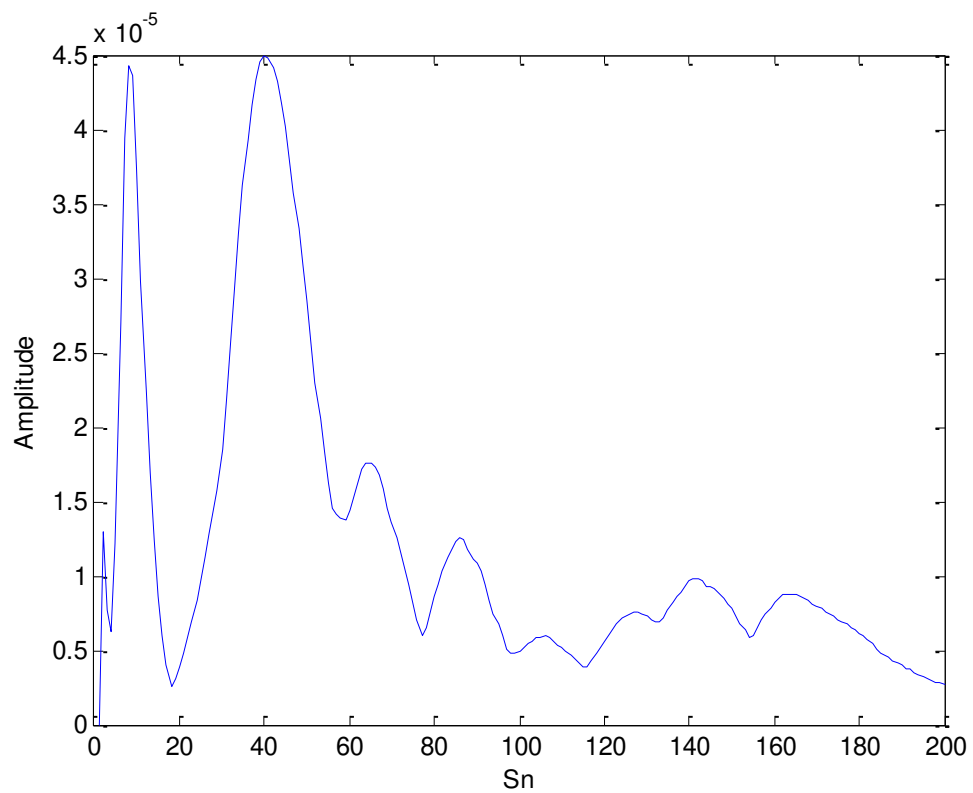


Figure 4.42 The detected SPD for vowel /e/, male speaker

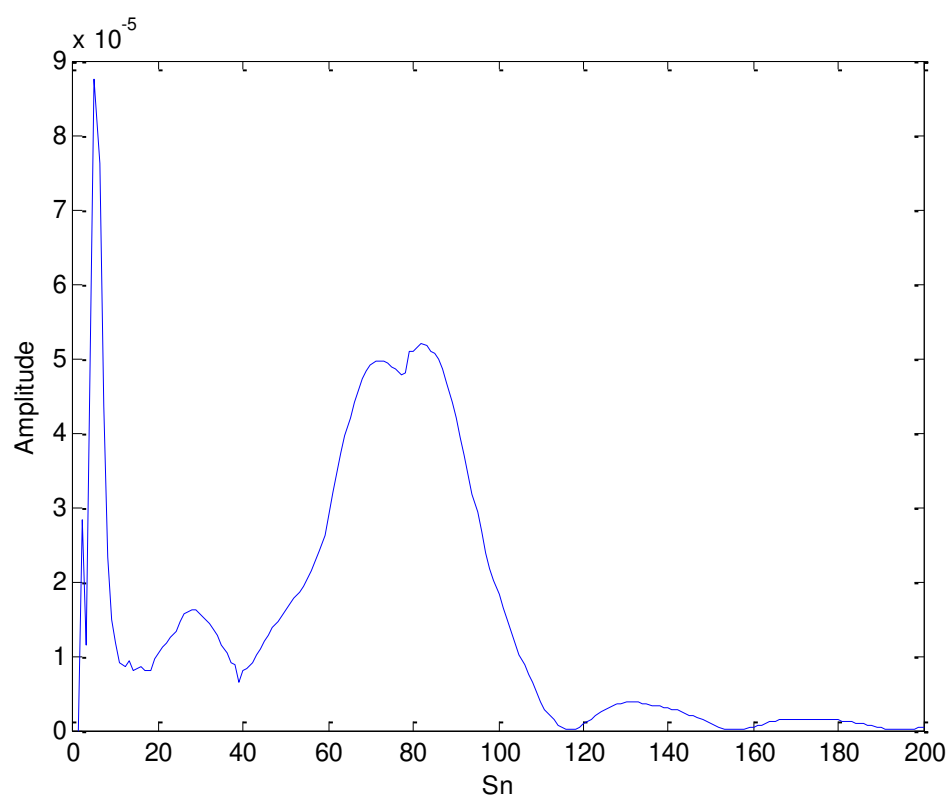


Figure 4.43 The detected SPD for vowel /i/, female speaker

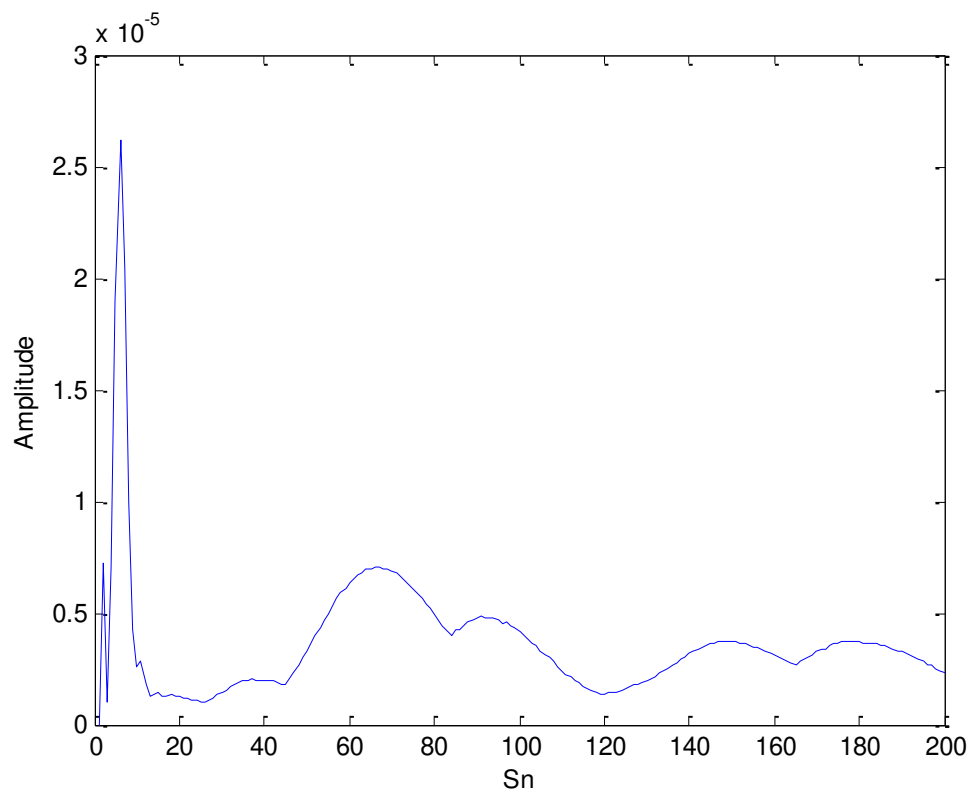


Figure 4.44 The detected SPD for vowel /i/, male speaker

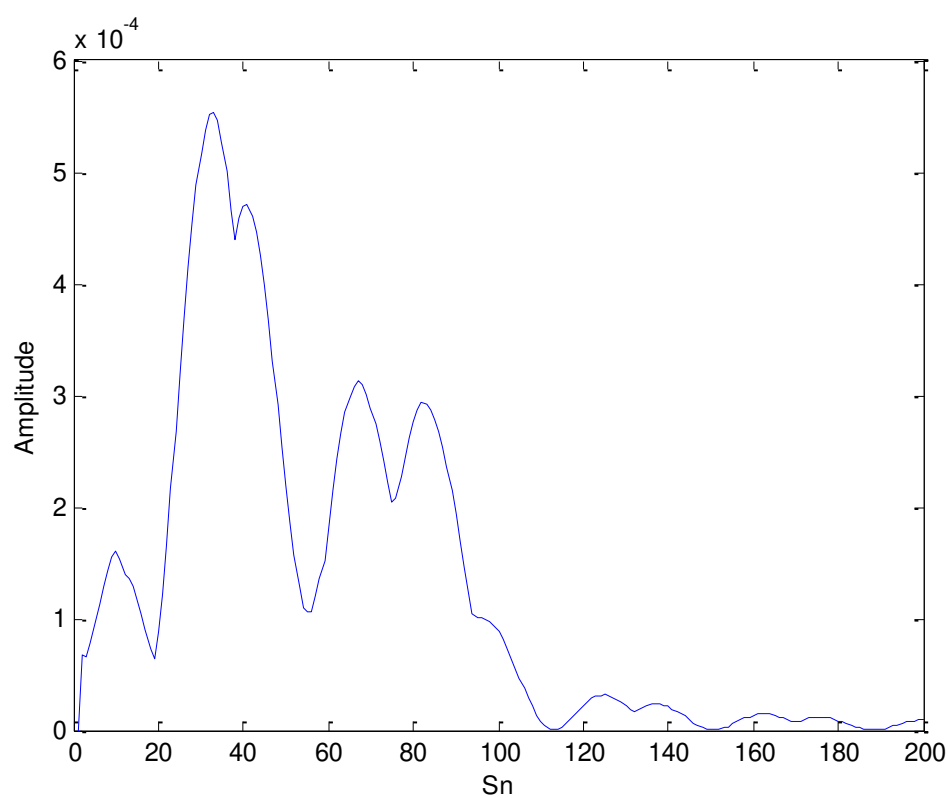


Figure 4.45 The detected SPD for vowel /ö/, female speaker

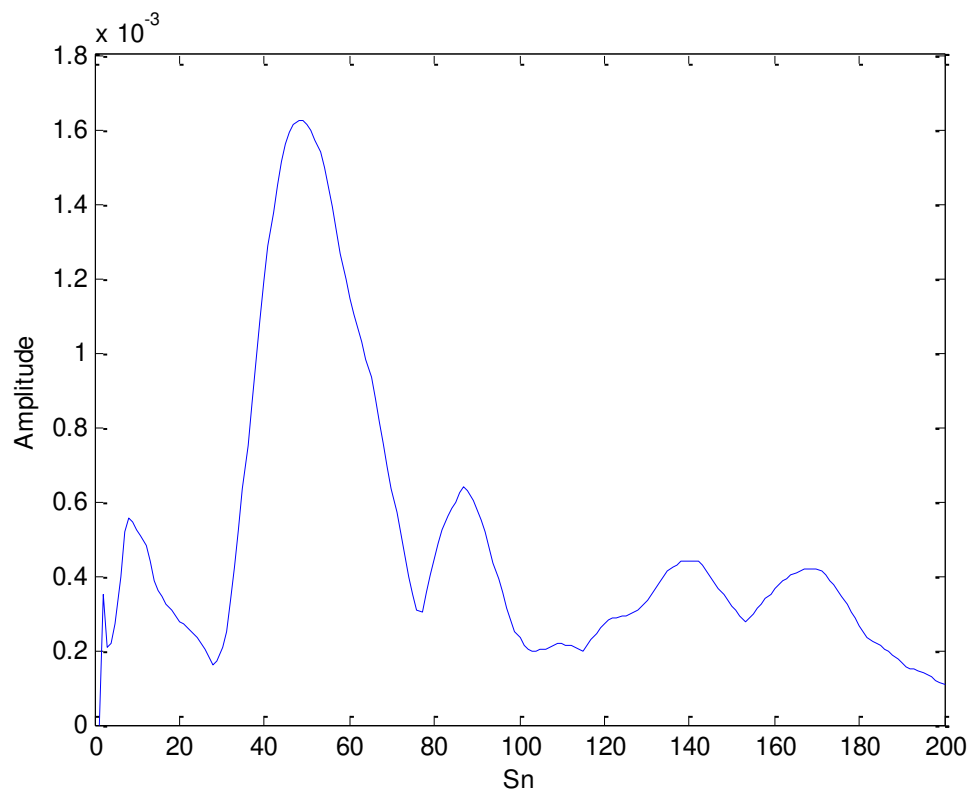


Figure 4.46 The detected SPD for vowel /ö/, male speaker

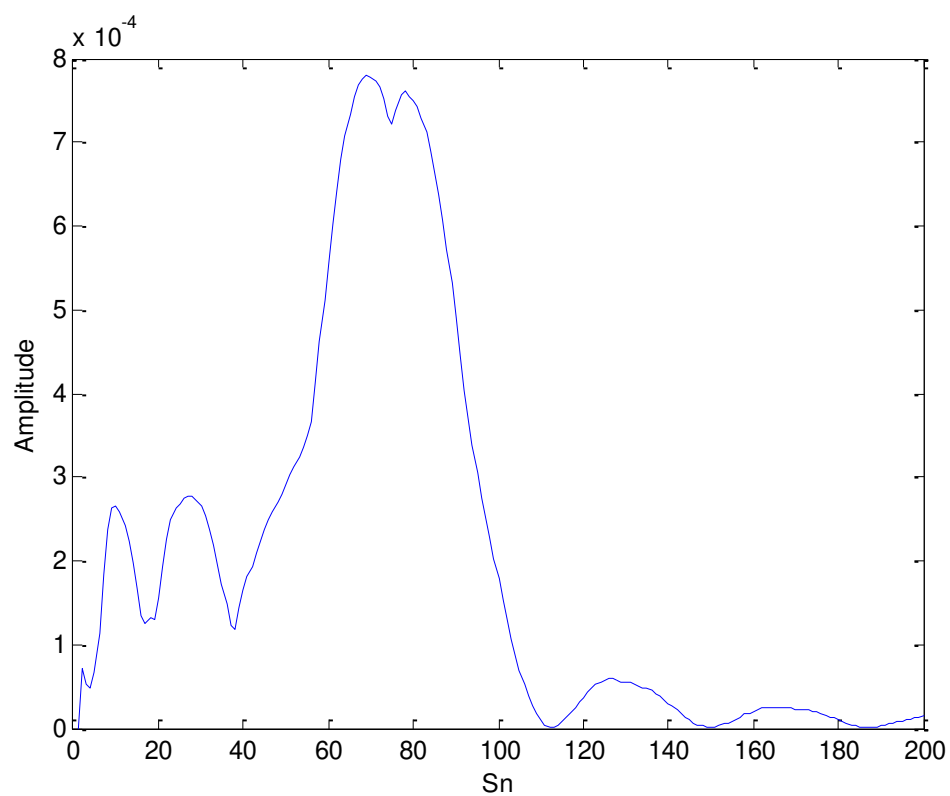


Figure 4.47 The detected SPD for vowel /ü/, female speaker

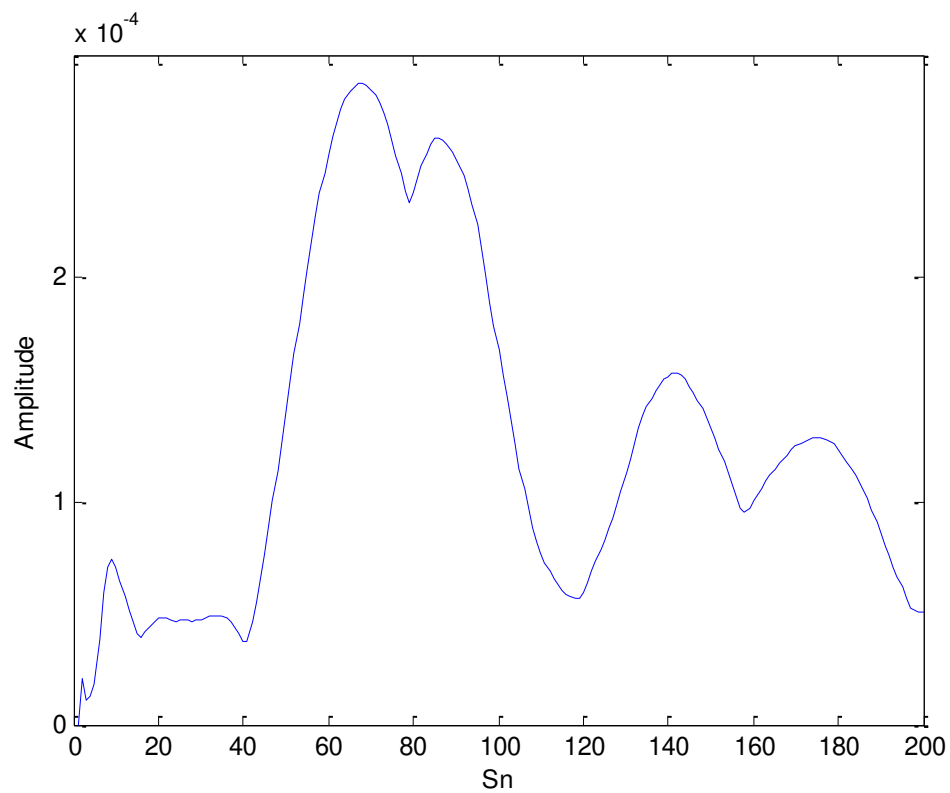


Figure 4.48 The detected SPD for vowel /ü/, male speaker

As can be seen from Figures 4.35-4.48 the similar SPDs are detected for the same vowel independent from the speakers. The similar distribution of spectral peaks occurs at higher frequencies (lower S_n) and the width of the distribution is dependent on the fundamental frequency F_0 .

4.3 The Auditory Motivated Discrete Time Frequency Signal Analysis Method Based Vowel Classification

In order to test the existence of the similar spectral peaks distribution (SPD) for the same vowels independent from the speakers, the vowel classification tests are performed. According to the results obtained in section 4.2 the vowel patterns should be fundamental frequency dependent.

The vowel classification tests are performed for the Turkish vowels /a/, /ı/, /o/, /u/, /e/, and /i/, which are also common in most of the languages. The vowel patterns are extracted from the SPDs which are easily detected from the time frequency distribution.

The detection of the vowel signal endpoints is easily performed with simple thresholding. In the case of consonant-vowel speech signals detection of the endpoints is more complex procedure. The SPDs are also independent from the duration of the vowel signals because it gives the instant amplitudes of the frequencies inside the short time duration.

After detecting the vowel signal, the signal is lowpass filtered in order to detect the approximate fundamental frequency of each vowel signal. The transfer function of the filter is given in Figure 4.49. The time frequency resolution and detected spectral peaks distribution after low pass filtering are given in the Figures 4.50 and 4.51 respectively. As can be seen from Figure 4.51, the S_n value at which the amplitude is maximum gives the approximate fundamental frequency, which will be used as scaling factor in the extraction of the vowel patterns from SPDs.

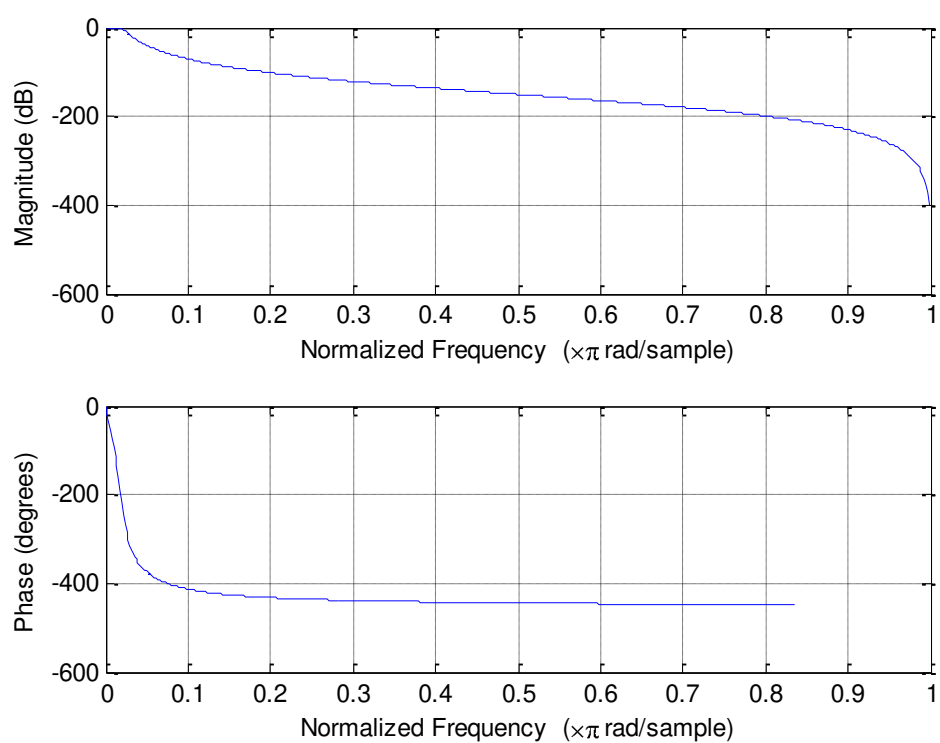


Figure 4.49 The frequency response of the used low pass filter

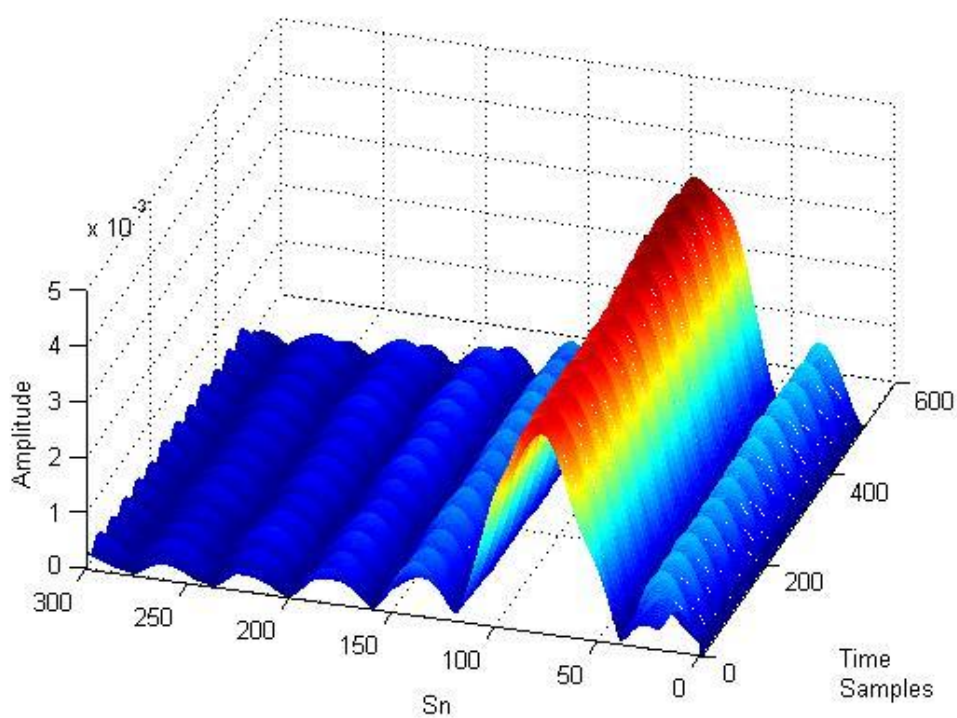


Figure 4.50 The obtained time frequency resolution after low pass filtering

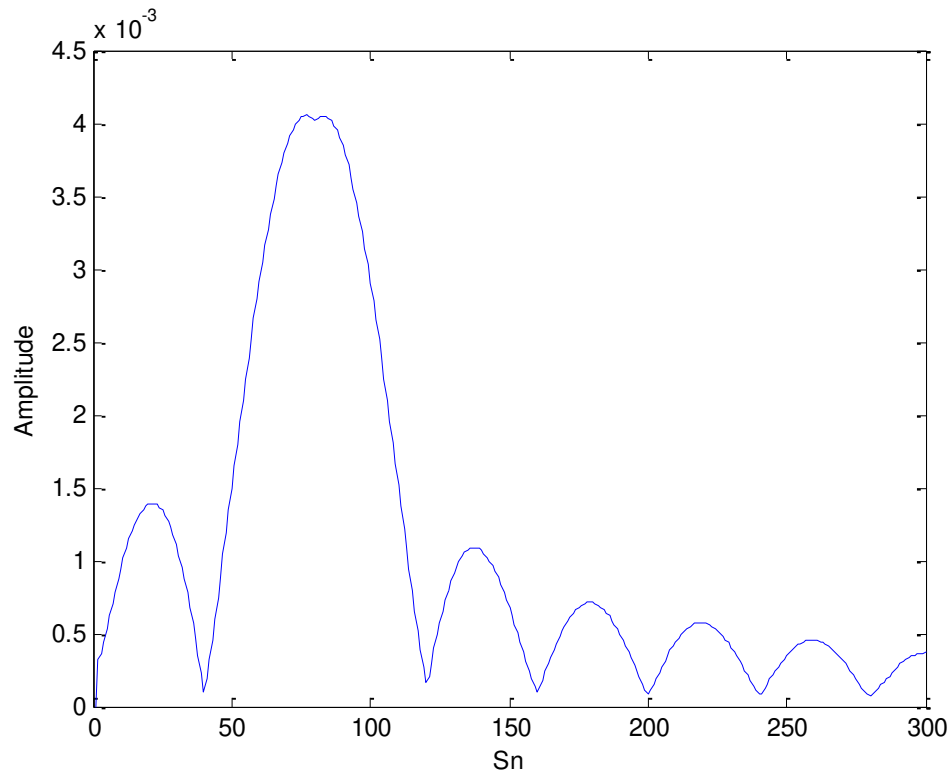


Figure 4.51 The detected SPD after low pass filtering

According to the results obtained in section 4.2 the similar SPD occur at higher frequencies and the spread of the similar SPD is dependent on the fundamental frequency. Therefore in the extraction of the vowel patterns the fundamental frequency dependent vowel patterns are extracted with different lengths.

In order to test the existing similar SPD for vowel signals taken from different speakers, the correlation coefficient which gives the similarity measure is used to obtain the best match of the current SPD to the reference SPDs obtained from the vowel signals.

The algorithm is tested with 4 male and 4 female speakers saying each vowel 5 times. For each vowel experiment is repeated 40 times and the total of 240 tests are performed to test the existence of the similar SPDs. The classification results are given in Table 4.1.

Table 4.1 Classification performance.

Actual	Classified						Performance %
	/a/	/ı /	/o/	/u/	/e/	/i/	
/a/	40	0	0	0	0	0	%100
/ı /	0	36	2	2	0	0	%90
/o/	0	0	39	0	1	0	%97.5
/u/	0	0	2	35	3	0	%87.5
/e/	0	0	0	0	38	2	%95
/i/	0	0	0	3	2	35	%87.5

The overall classification performance is %93, which is the good evidence of the existing similar SPD in each vowel signal.

In Yavuz&Topuz (2010) the speaker dependent Turkish vowel classification is performed by using the probabilistic neural networks approach. The classification performance obtained in Yavuz&Topuz (2010) is higher than %95. In the results given in Table.1, the %93 performance is obtained for speaker independent vowel classification without any training.

CHAPTER FIVE

CONCLUSION

The sound transduction process inside the human ear is a complex procedure. The outer hair cells control the function of the inner hair cells which suggests that every process inside the human ear is an adaptive process which makes it very difficult to simulate by the aid of computer algorithms.

However some basic operations of the human ear can be simulated under some assumptions. In this thesis the auditory motivated discrete time frequency signal representation method is presented. The method is motivated from the structure and operation of the basilar membrane and inner hair cells under some assumptions.

The proposed method is independent from the window function and obtained discrete time frequency resolution is directly dependent on the signal shape. At higher sample rates, the proposed method gives comparable results to the results obtained in literature. The bandwidth and center frequency properties of the windowing functions affect the obtained time frequency resolutions. The numerical simulations at low SNR_{dB} values show that the method can be used to obtain the discrete time frequency resolution easily from the analyzed signal.

The application of the AMTFR method with variable Δk give the time frequency representation for non-stationary signals, and the obtained frequency values are similar to the frequency selectivity of the basilar membrane.

For the speech vowel signals, the proposed method give the detailed spectral shapes and because the method does not employ any windowing function, the obtained time frequency resolution is directly dependent on the signal shape. Most of the spectral feature extraction methods use the energies of the frequency bands. However in the proposed method the spectral envelope like spectral peaks distributions are used as feature vectors, and the results show that inside the speech

signals there may exist similar spectral envelopes for the same speech signals independent from the speakers. These features may be directly used as feature vectors or they may be used as additional cues for speech recognition applications.

The speaker independent vowel classification scores based on spectral peaks distribution is a good evidence of existing similar spectral envelopes. The classification scores for Turkish vowels obtained in (Yavuz&Topuz, 2010) are speaker dependent, however the classification results obtained in chapter 4 are speaker independent. The data set was larger in (Yavuz&Topuz, 2010) and the classification is performed for all Turkish vowels. In the proposed method, the classification is performed to test the existence of the similar spectral peaks distribution. The results support the hearing experiments made in (Sakayori et al., 2002) which show that the spectral envelopes are important for phonetic quality. Also in (Zahorian&Jagharghi, 1993) it is shown that spectral envelope based vowel classification is superior to the classification based on the formant frequencies.

The proposed method and the results obtained for vowel speech signals show that the proposed method is applicable to spectral feature extraction, and discrete time frequency signal analysis of stationary and non-stationary signals. According to Crick (2003) the human brain like the database search engine tries to find the best match of the all current sensory input features to the existing features saved from the previous experiences of the human brain, which shows that there may exists similar information inside the sensory signals. The obtained results show that the similar features for the same class of signals may be extracted by simulating the human sensory organs under some assumptions.

The proposed method can be applied and its performance can be tested for speech signals from different languages in the future works. The proposed method is used to obtain the spectral peaks distributions for short durations in time obtained from vowel speech signals. It is important to notice that the method can be expanded to obtain the spectral peaks distributions for consonants and vowels, and the obtained features can be used for word recognition applications.

REFERENCES

- Ali, A.M.A., van der Spiegel, J., & Mueller, P. (2002). Robust auditory-based speech processing using the average localized synchrony detection. *IEEE Transactions on Speech and Audio Processing*, 10(5), 279-292.
- Benesty, J., Chen, J., & Huang, Y. (2008). On the Importance of the Pearson Correlation Coefficient in Noise Reduction. *IEEE Transactions on Speech and Audio Processing*, 16(4), 757-765.
- Chatterjee, S., & Kleijn, W.B. (2011). Auditory model-based design and optimization of feature vectors for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 19(6), 1813-1825.
- Crick, F. (2003). *Şaşırtan Varsayım* (10. Baskı). (Say, S., Çev.). Türkiye Bilimsel ve Teknik Araştırmaları Kurumu Yayınları. (Orijinal çalışma basım tarihi 1990).
- Dusan, S. (2007). On the Relevance of Some Spectral and Temporal Patterns for Vowel Classification. *Speech Communication*, 49(2007), 71-82.
- Kameoka, H., Ono, N., & Sagayama, S. (2010). Speech Spectrum Modeling for Joint Estimation of Spectral Envelope and Fundamental Frequency. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6), 1507-1516.
- Kasabov, N. (1996). *The Foundations of Neural Networks and Fuzzy Systems, and Knowledge Engineering*. Massachusetts: The M.I.T. Press.
- Loizou, P.C. (1998). Mimmicking the Human Ear. *IEEE Signal Processing Magazine*, 15(5), 101-130.

- Loughlin, P.J., & Cohen, L. (2004). The uncertainty principle: global, local, or both? *IEEE Transactions on Signal Processing*, 52(5), 1218-1227.
- Lu, W., & Zhang, Q. (2009). Deconvolutive short time fourier transform spectrogram. *IEEE Signal Processing Letters*, 16(7), 576-579.
- Mertins, A. (1999). *Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications*, Wiley.
- Moller, A.R. (2003a). Anatomy and Physiology of sensory organs. In *Sensory Systems* (33-74), Elsevier Inc.
- Moller, A.R. (2003b). Hearing. In *Sensory Systems* (271-371), Elsevier Inc.
- Moller, A.R. (2003c). Basic Psychophysics. In *Sensory Systems* (1-31), Elsevier Inc.
- Oppenheim, A.V., & Schafer, R.W. (1999). *Discrete Time Signal Processing* (2nd ed.), Prentice Hall.
- Oppenheim, A.V., & Whillsky, A.S. (1997). *Signals & Systems*, (2nd ed.), Prentice Hall.
- Painter, T., & Spanias, A. (2000). Perceptual Coding of Digital Audio. *Proceedings of the IEEE*, 88 (4), 451-515.
- Pavez, E., & Silva, J.F. (2012). Analysis and design of Wavelet-Packet Cepstral coefficients for automatic speech recognition. *Speech Communication*, 54(6), 814-835.
- Picone, J.W. (1993). Signal Modeling Techniques in Speech Recognition. *Proceedings of the IEEE*, 81(9), 1215-1247.

- Polikar, R. (January 12, 2001). Wavelet Tutorial. Retrieved September 3, 2011 from <http://users.rowan.edu/~polikar/wavelets/wttutorial.html>
- Rabiner, L., & Juang, B.H. (1993). *Fundamentals of Speech Recognition*, Prentice Hall.
- Sakayori, S., Kitama, T., Chimoto, S., Qin, L., & Sato, Y. (2002). *Neuroscience Research*, 43(2002), 155-162.
- Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, 270(5234), 303-304.
- Sugavaneswaran, L., Xie, S., Umapathy, K., & Krishnan, S. (2012). Time-Frequency Analysis via Ramanujan Sums. *IEEE Signal Processing Letters*, 19(6), 352-355.
- Sumner, C.J., Lopez-Poveda, E.A., O'Mard, L.P., & Meddis, R. (2002). A revised model of the inner-hair cell and auditory-nerve complex. *Journal of Acoustic Society America*, 111(5), 2178-2188.
- Umapathy, K., Ghoraani, B., & Krishnan, S. (2010). Audio Signal Processing Using Time-Frequency Approaches: Coding, Classification, Fingerprinting, and Watermarking. *EURASIP Journal on Advances in Signal Processing*, 2010, 1-28.
- Wang, Y., & Jiang, Y.C. (2009). New time–frequency distribution based on the polynomial Wigner–Ville distribution and L class of Wigner–Ville distribution. *IET Signal Processing*, 4(2), 130-136.
- Yavuz, E. & Topuz, V. (2010). Recognition of Turkish Vowels by Probabilistic Neural Networks Using Yule-Walker AR Method. Proceedings of the 5th International Conference on Hybrid Artificial Intelligence Systems HAIS'10, Springer-Verlag Berlin, Volume Part I, 112-119.

- Zahorian, S. A., & Jagharghi, A.J. (1993). Spectral-Shape Features versus Formants as Acoustic Correlates for Vowels. *Journal of Acoustic Society America*, 94(4), 1966-1982.
- Zahorian, S.A., & Nossair, Z.B. (1999). A Partitioned Neural Network Approach for Vowel Classification Using Smoothed Time/Frequency Features. *IEEE Transactions on Speech and Audio Processing*, 7(4), 414-425.
- Zahorian, S.A., & Zhang, S.J. (1992). Perception of Vowels Synthesized from Sinusoids that Preserve either Formant Frequencies or Global Spectral Shape. *Journal of Acoustic Society America*, 92.
- Zhong, J., & Huang, Y. (2010). Time-Frequency Representation Based on as Adaptive Short-Time Fourier Transform. *IEEE Transactions on Signal Processing*, 58(10), 5118-5128.