MOHAMMED HAMID

ATILIM UNIVERSITY

2020

SPEAKER INDEPENDENT ISOLATED DIGIT RECOGNITION

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ATILIM UNIVERSITY

MOHAMMED SAEED HAMID

A MASTER OF SCIENCE
THESIS
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

SEPTEMBER 2020

SPEAKER INDEPENDENT ISOLATED DIGIT RECOGNITION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ATILIM UNIVERSITY

BY

MOHAMMED SAEED HAMID

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

SEPTEMBER 2020

Approval of the Graduate School of Natural and Applied Sciences, Atilim University.

_____

Prof. Dr. Ali KARA
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of **Master of Science in Electrical and Electronics Engineering, Atilim University**.

_____

Assoc. Prof. Dr. Kemal Efe ESELLER
Head of Department

This is to certify that we have read the thesis SPEAKER INDEPENDENT ISOLATED DIGIT RECOGNITION submitted by MOHAMMED HAMID and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science

_____

Asst. Prof. Dr. Hakan TORA
Supervisor

.

Assoc. Prof. Dr. Fikret ARI
Electrical-Electronics Engineering Department,
Ankara University _____

Asst. Prof. Dr. Hakan TORA
Avionics Department, Atilim University _____

Asst. Prof. Dr. Baran USLU
Electrical-Electronics Engineering Department,
Atilim University _____

Date: *8 August 2020*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Mohammed Hamid

Signature:

ABSTRACT

SPEAKER INDEPENDENT ISOLATED DIGIT RECOGNITION

Hamid, Mohammed Saeed Hamid

MSc, Electrical and Electronics Engineering

Asst. Prof. Dr. Hakan TORA

September 2020, 72 pages

In several speech signal processing applications, VAD presents an important character for splitting an audio stream into time intervals that include speech activity and time intervals where speech is absent.

In this research, we presented new approach dealing with isolated word recognition. In the first stage, three functions applied for voice activity detection (VAD) problem hamming window, Bohman function, and Bartlett-Hann function. The both Bohman function and Bartlett-Hann function are not applied in previous studies for VAD problem.

On the other hand, pitch, MFCCs, and energy applied as feature extraction techniques and combined with SOFTMAX which these two methods are new approaches. The Pitch based SOFTMAX presented remarkable results which extracted features by pitch wired to SOFTMAX and classified to seven words and presented 85% accuracy. Furthermore, energy also applied as feature extraction and the output of this function wired to the SOFTMAX. This framework easily can applied to the various isolated word recognition which only the user modified the input data easily.

The main contribution in this study, combine SOFTMAX with several feature extraction techniques. The SOFTMAX is trend probability function which analysis input features to the labels between (0,1) and used in several deep learning techniques as last layer function for classification or regression issues.

The obtained results compared with several studies presented in this field by applying several machine learning and deep learning techniques combined with audio signal processing techniques that’s applied for feature extraction.

Keywords: MFCCs, SOFTMAX, pitch, VAD, energy, MATLAB

ÖZ

HOPARLÖR BAĞIMSIZ İZOLASYONLU RAKAM TANIMA

Hamid, Mohammed Saeed Hamid
Yüksek lisans., Elektrik Elektronik Mühendisliği
Doç. Prof. Dr. Hakan TORA
Eylül 2020, 72 sayfa

Çeşitli konuşma sinyali işleme uygulamalarında VAD, bir ses akışını konuşma etkinliği ve konuşmanın olmadığı zaman aralıklarını içeren zaman aralıklarına bölmek için önemli bir karakter sunar.

Bu araştırmada, izole kelime tanıma ile ilgili yeni bir yaklaşım sunduk. İlk aşamada, ses etkinliği algılama (VAD) problem kırma penceresi, Bohman işlevi ve Bartlett-Hann işlevi için üç işlev uygulanmıştır. Hem Bohman fonksiyonu hem de Bartlett-Hann fonksiyonu VAD problemi için önceki çalışmalarda uygulanmamıştır.

Öte yandan, perde, MFCC'ler ve enerji, özellik çıkarma teknikleri olarak uygulanır ve bu iki yöntemin yeni yaklaşımlar olduğu SOFTMAX ile birleştirilir. Pitch tabanlı SOFTMAX, SOFTMAX'a bağlanan ve yedi kelimeye göre sınıflandırılan ve% 85 doğrulukla özelliklerle çıkarılan olağanüstü sonuçlar sundu. Ayrıca enerji, özellik çıkarma ve SOFTMAX'a bağlanan bu fonksiyonun çıktısı olarak da uygulanır. Bu çerçeve, yalnızca kullanıcının giriş verilerini kolayca değiştirdiği çeşitli yalıtılmış kelime tanıma işlemlerine kolayca uygulanabilir.

Bu çalışmadaki ana katkı, SOFTMAX'ı çeşitli özellik çıkarma teknikleriyle birleştirmiştir. SOFTMAX, (0,1) arasındaki etiketlere girdi özelliklerini analiz eden ve sınıflandırma veya regresyon sorunları için son katman fonksiyonu olarak çeşitli derin öğrenme tekniklerinde kullanılan trend olasılık fonksiyonudur.

Elde edilen sonuçlar, özellik çıkarma için uygulanan sesli sinyal işleme teknikleri ile birleştirilmiş çeşitli makine öğrenme ve derin öğrenme teknikleri uygulanarak bu alanda sunulan çeşitli çalışmalarla karşılaştırılmıştır.

Anahtar Kelimeler: MFCCs, SOFTMAX, pitch, VAD, energy, MATLAB.

To My Parents

My Father "Dr. Saeed Al-Nattah"

My Mother "Shaymaa Ayyed"

My Uncle "Ahmed Ayyed"

My FAMILY

ACKNOWLEDGMENTS

To the most generous, kind and mighty. Without your blessings and guidance this moment would have never come to existence. I am dedicating this humble work to you GOD.

To the best teacher/mentor a graduate student could ever ask for. It is your sharing of knowledge and guidance that have molded me into the person I am today, both scientifically and personally, and for that i will be forever in your debt.

Thank you for all the fatherly advices, patience, wisdom, and guidance that directed me to this point. I will be forever grateful to you Asst. Prof. Dr. Hakan Tora.

A special feeling of gratitude to my loving parents, whose words of encouragement and push for tenacity ring in my ears. To my brothers and sisters who have been my emotional little support squad, who kept a smile on my face and pushed me to outwork myself and reach my goal.

I dedicate my dissertation work to My family.

To the joyful spirits, to the ones who made this journey memorable, to the ones who stayed beside me throughout the ups and downs and encouraged me to pursue this dream, i will always remember every second of this wonderful journey My LOVE and My friends.

Author

Mohammed Saeed Hamid

TABLE OF CONTENTS

ABSTRACT iii

ÖZ..... iv

ACKNOWLEDGMENTS viii

LIST OF TABLES ix

LIST OF FIGURES xi

LIST OF SYMBOLS/ABBREVIATIONS xiiii

CHPTERS

1. INTRODUCTION 1

1.1 Speech Recognition Applications 4

2. LITERATURE REVIEW 8

2.1 Fundamentals of Human Speech 8

2.2 Speech Production 8

2.3 Characteristics of Speech 10

2.4 Characteristics of Noise 11

2.4.1 Statistical Properties of Noise 11

2.5 Survey 11

2.6 Time Domain Parameters of Speech Recognition 13

2.7 Speaker Independent/ Speaker dependent 15

2.9 Classification 17

3. MATERIAL AND METHODS 19

3.1 Free Spoken Digit Dataset (FSDD) 19

3.2 Feature Extraction 20

3.2.1 Mel-frequency cepstral coefficients (MFCC) 20

3.2.2 Pitch 21

3.3 Voice Activity Detection Using Energy 21

4. EXPERIMETAL RESULTS 26

3.4 Voice Activity Results 26

4.1.1 Hamming 26

4.1.2 Bohman window 32

4.1.3 Bartlett-Hann window 38
4.2. Feature Extraction Using Pitch 44
4.3 Feature Extraction Using MFCCS 50
4.4 Feature Extraction Using Energy 56
4.5 Discussion 57
5. CONCLUSION 72
REFERENCES 73

LIST OF TABLES

TABLES

Table 4.1 MFCC Parameters..... 51
Table 4.2 Results Comparisons..... 70

LIST OF FIGURES

FIGURES

Figure 1.1 Speech Recognition System [4] 2
Figure 1.2 Isolated word recognition (IWR) [6] 4
Figure 1.3 Speech Recognition System 5
Figure 2.1 Vocal tract for human speech production [10] 9
Figure 2.2 Source/filter model for speech production 10
Figure 2.3 ZCR 14
Figure 3.1 Co-Learning Algorithm 20
Figure 3.2 VAD Block Diagram 23
Figure 3.3 Our Framework 25
Figure 4.1 VAD for 0 using hamming 27
Figure 4.2 VAD for 1 using hamming 28
Figure 4.3 VAD for 2 using hamming 28
Figure 4.4 VAD for 3 using hamming 29
Figure 4.5 VAD for 4 using hamming 29
Figure 4.6 VAD for 5 using hamming 30
Figure 4.7 VAD for 6 using hamming 30
Figure 4.8 VAD for 7 using hamming 31
Figure 4.9 VAD for 8 using hamming 31
Figure 4.10 VAD for 9 using hamming 32
Figure 4.11 VAD for 0 using Bohman 33
Figure 4.12 VAD for 1 using Bohman 34
Figure 4.13 VAD for 2 using Bohman 34
Figure 4.14 VAD for 3 using Bohman 35
Figure 4.15 VAD for 4 using Bohman 35
Figure 4.16 VAD for 5 using Bohman 36
Figure 4.17 VAD for 6 using Bohman 36
Figure 4.18 VAD for 7 using Bohman 37
Figure 4.19 VAD for 8 using Bohman 37
Figure 4.20 VAD for 9 using Bohman 38
Figure 4.21 VAD for 0 using Bartlett-Hann 39
Figure 4.22 VAD for 1 using Bartlett-Hann 40
Figure 4.23 VAD for 2 using Bartlett-Hann 40
Figure 4.24 VAD for 3 using Bartlett-Hann 41
Figure 4.25 VAD for 4 using Bartlett-Hann 41
Figure 4.26 VAD for 5 using Bartlett-Hann 42
Figure 4.27 VAD for 6 using Bartlett-Hann 42
Figure 4.28 VAD for 7 using Bartlett-Hann 43
Figure 4.29 VAD for 8 using Bartlett-Hann 43
Figure 4.30 VAD for 9 using Bartlett-Hann 44
Figure 4.31 pitch Algorithm 46

Figure 4.32 Feature extraction for zero using pitch 46
Figure 4.33 Feature extraction for one using pitch 47
Figure 4.34 Feature extraction for two using pitch 47
Figure 4.35 Feature extraction for three using pitch 48
Figure 4.36 Feature extraction for four using pitch 48
Figure 4.37 Feature extraction for five using pitch 49
Figure 4.38 Feature extraction for six using pitch 49
Figure 4.39 Feature extraction for one using MFCCs 53
Figure 4.40 Feature extraction for two using MFCCs 53
Figure 4.41 Feature extraction for three using MFCCs 54
Figure 4.42 Feature extraction for four using MFCCs 54
Figure 4.43 Feature extraction for five using MFCCs 55
Figure 4.44 Feature extraction for six using MFCCs 55
Figure 4.45 Feature extraction for seven using MFCCs 56
Figure 4.46 Energy Calculation 56
Figure 4.47 Pitch confusion matrix 58
Figure 4.48 Pitch based SOFTMAX Roc curve 59
Figure 4.49 Energy based SOFTMAX confusion matrix 60
Figure 4.50 Energy based SOFTMAX Roc Curve 60
Figure 4.51 MFCCs based SOFTMAX confusion matrix with 1024 overlap length 61
Figure 4.52 MFCCs based SOFTMAX Roc curve with 1024 overlap length 62
Figure 4.53 MFCCs based SOFTMAX confusion matrix with 512 overlap length .. 63
Figure 4.54 MFCCs based SOFTMAX Roc curve with 512 overlap length 63
Figure 4.55 Pitch confusion matrix with 30 test sample 64
Figure 4.56 Pitch based SOFTMAX Roc curve with 30 test sample 65
Figure 4.57 Energy confusion matrix with 30 test sample 66
Figure 4.58 Energy Roc curve with 30 test sample 67
Figure 4.59 MFCCs confusion matrix with 30 test sample 68
Figure 4.60 MFCCs Roc curve with 30 test sample 69
Figure 4.61 Comparisons 71

LIST OF SYMBOLS/ABBREVIATIONS

MFCCs	Mel-frequency cepstral coefficients
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
TPR	True Positive Rate
TNR	True Negative Rate
ROC	Receiver operating characteristic

CHAPTER 1

INTRODUCTION

Speech is the dominant tool of communication between people and if it promises to be important for communication between people and machines it can only be made a little more reliable. Speech recognition is the process of converting an acoustic signal into a set of devices. Words applications include voice commands and control, data entry, voice user interface, automate the work of the telephone operator on the phone, etc. it also acts as an input to natural language processing. In some situation, the speech recognition problem, as established over the years, is an extremely calculation concentrated problem; it needs debauched computers, and a huge amount of memory. Several efforts have, consequently, been complete to hurry up the procedure by applying different methods. Consequently, the employment of speech recognition systems (SRS) applying digital signal processing integrated circuits (DSP's) is attractive progressively gorgeous. The compensations of applying a DSP are upsurge in speed of the technique expansion, extended term dependability, noise protection and the aptitude to achieve difficult computations absurd in the analog range [1].

The number of SR software's in customer electronic products has been cumulative quickly in the last years. One of the chief challenges confronted by the built-in software designer is the comparative calculating power and memory shortage, which must previously be measured in the early phases of technique design. Consequently, it is probable that the excellent between different techniques is depended on conciliation between cost and performance [2].

Pioneering work on ASR dates back to the early 1950s. The first ASR system developed at Bell Telephone Laboratories was able to recognize distinct numbers from 0 to 9 for a single speaker. In 1956, Olson and Pillar created a sound machine capable of recognizing ten different syllables. He also had to rely on the speaker and needed extensive training.

This initial pattern-based recognition by the ASR program is based on pattern matching as the amplifier inputs were compared to previously stored sound patterns or patterns. Pattern matching works well at the word level for recognizing various elements of a small vocabulary, but less effective for recognizing a larger vocabulary. Another limitation of pattern matching is the inability to match the input voice signals and align them with previously stored audio samples of different lengths. Therefore, the performance of these ASR systems was poor because they used acoustic approaches that identified only the main vocal units clearly expressed by the speaker [3]. The main sections of speech recognition system presented in Figure 1.1.

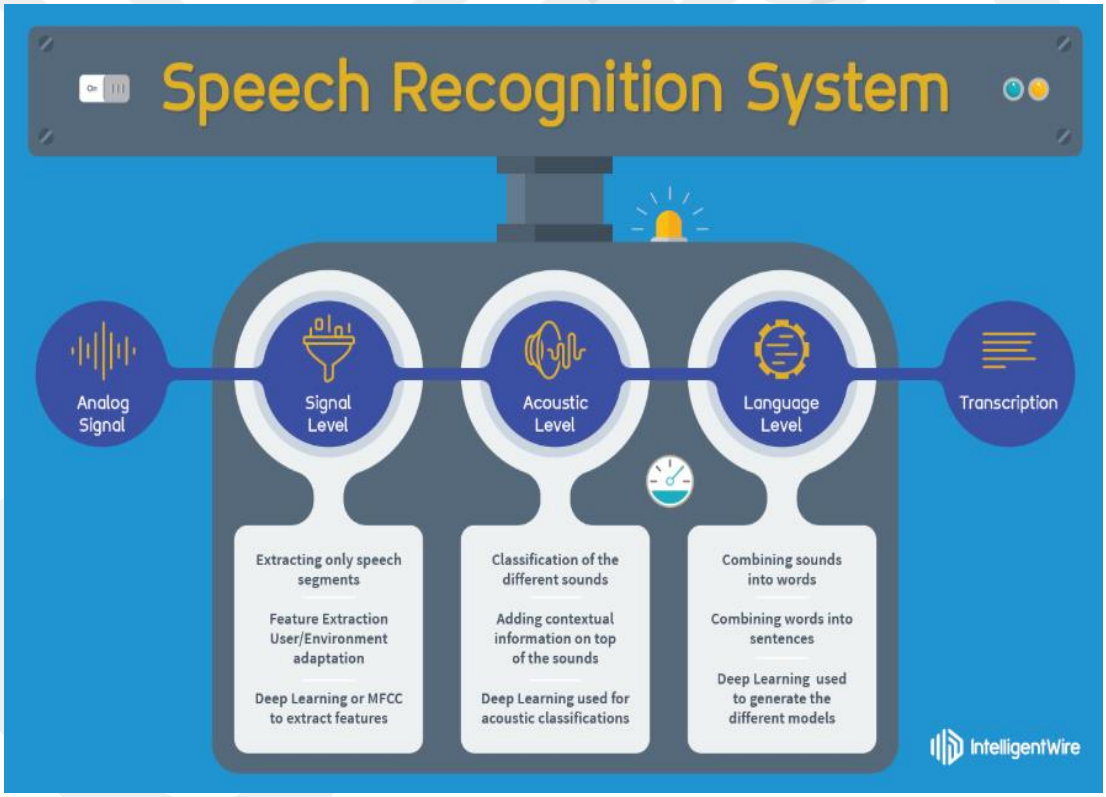

Figure 1.1 Speech Recognition System [4]

Isolated word recognition (IWR) is the procedure of automatically selecting and classifying voice wave features applying a electronic circuits and computer. The isolated phone has been extraordinarily identify and automatic IWR techniques for several years, containing computers [5]. Several contributions presented in this study such as: Developed efficient IWR system, the main aim of this study is to develop

IWR system which mean the system have ability to convert the sound that received in different Accents and different peoples. Which several feature extraction techniques such as MFCCs, Pitch, and energy combined with SOFTMAX. This combination is new method and presented new results can applied in several IWR application. In this study three Voice activity detection (VAD) applied to the digit dataset, which two of these three VAD techniques are new and not applied in any VAD problems previously. The aim of using three techniques is to Show the differences between these techniques and investigate which of these techniques is more effective and which of these techniques is more effective with which sound types. Finally, Our study presented complete new voice recognition system which combine new techniques such as Bohman and Bartlett-Hann functions as VAD techniques. Besides, the feature extraction techniques such as MFCCs, Pitch, and energy with SOFTMAX for IWR.

This study consists from five chapters: introduction, related works, material and methods, and experimental results.

Then, the chapter one consists from Introduction, Contributions, Questions Of The Study, and Thesis Structure. The aim of this chapter presents general information about this study, and presented the goal and contributions of the proposed method. Furthermore, in chapter two, the related Works, and survey studies are explained in detail form to assist the reader what is the previous studies presents and what is the techniques and methods applied to previous studies. Moreover, in chapter three, Material and methods that's used to developed our methods explained and presented also the MATLAB scripts that we developed are presented and explained. Besides, in chapter four: Experimental results are presented and discussed which several results related to our method explained such as VAD results and IWR results. Then, in chapter five we presents our Conclusion related this study and results with future works too.

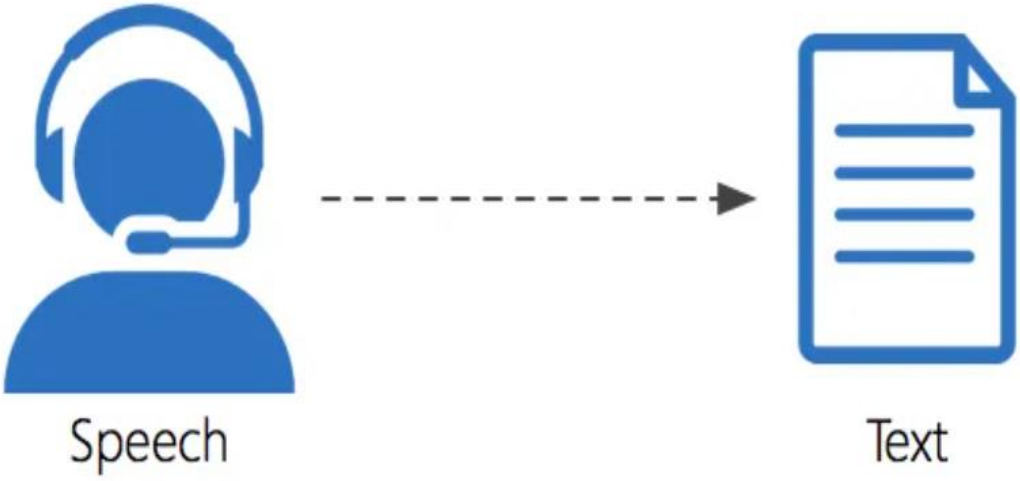

Figure 1.2 Isolated word recognition (IWR) [6]

1.1 Speech Recognition Applications

Speech recognition basically means talking to a computer and telling it what we're saying. This process mainly works as a pipeline that converts digital PCM sound (pulse code modulation) from sound card to known speech. Speech recognition technology has followed the evolving architecture of signal processing algorithms and hardware advances for 40 years. During this time, an art of laboratory curiosity developed and finally a comprehensive technology that was practiced and understood by many engineers, scientists, linguists, psychologists and system designers. In the past 40 years, speech recognition technology has evolved and been addressed and resolved, resulting in a steady stream of increasingly complex problems [7]. In figure 1.3, the Speech Recognition system presented.

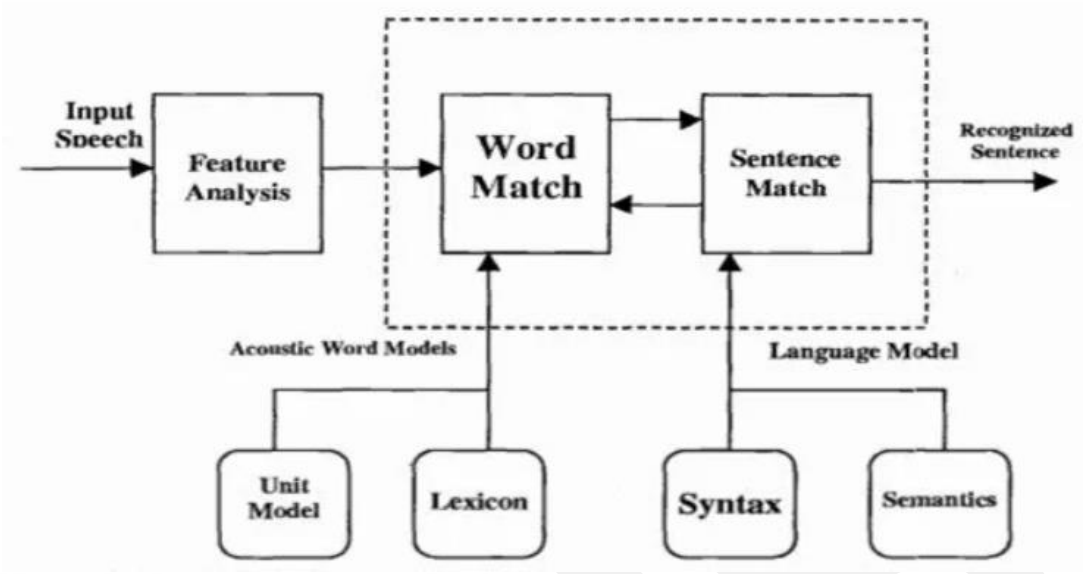

Figure 1.3 Speech Recognition System

- Isolated word recognition: Both the speaker and the independent speaker were trained. This technology has opened a class of applications called "command and control applications" through which the system can recognize a single word command (from small vocabulary to a word command) and respond appropriately to the recognized command. One of the main problems with this technique was background noise sensitivity (often recognized as misspelled words) and spam that was mistakenly pronounced using the word command. Various types of keyword detection algorithms have been developed to solve these types of problems.
- Speech Understanding Systems: Can identify the underlying message embedded in it

  Speak instead of just recognizing spoken words [8]. These systems, This has just emerged and has enabled services such as customers. Care (AT&T, how can I help you with the system) and a smart worker. Systems that provide access to information sources via voice dialogs.
- ATC training is a great application for speech recognition systems. Many ATC educational systems require the current person to act as a "dummy pilot" in order to have a voice conversation with the student's console, which simulates interaction with the pilot in a real ATC situation. The knowledge

and technology of TTS (Text-to-Speech) prevents people from becoming poor pilots and reduces the level of education and employee support. In theory, air traffic control operations have a highly structured language with basic command output, which reduces the complexity of speech recognition processing. In practice, I rarely do. FAA 7110.65 provides detailed information on the terminology used by air traffic controllers. There are fewer than 150 examples of these suggestions in this article, but the simulation vendor's voice recognition system supports over 500,000 suggestions.

USAF, USMC, the U.S. Army, U.S. Navy, and FAA, as well as a number of international ATC training organizations such as the Royal Australian Air Force and civil aviation authorities in Italy, Brazil, and Canada, are currently using ATC simulators with speech recognition from a number of different providers.

1.2 The Aim of This Study

- Efficient speech recognition is generated. A very natural human-machine interface is provided. Nature means intuitive and easy to use, it does not require any special tools or machines, it only needs the power of nature that everyone has. This system can be used by anyone who can speak and can use machines on a larger scale, especially computers.
- This study, presented new methods with high accuracy that can be used in several human-machine interface applications. This study is one of the steps in the direction of natural communication between human and machine that’s lead to great leap in various fields such as industrial, medical, and social projects.

1.3 Questions of the Study

- What are the speech recognitions? The author search about this question and presented detail definition which is the first step to understand the speech field to learn how can to presents new solutions to this field problem.
- What is the Isolated word recognition (IWR) and how developed new methods in this field? The details of IWR explained and presented to help the researcher to fix the field and the problem. Furthermore, new methods presented to presented scientific contribution in IWR field.
- What is VAD and who effective VAD techniques developed? The author applied new two techniques as VAD methods which presented satisfactory results and can be used in various VAD problems in future.
- What are the best feature extraction techniques for audio sound? Several feature extraction techniques presented in this study such as MFCCs, Pitch, and energy. The MFCCs applied in several speeches recognition and IWR studies because its presented suitable results but it's become effective in most time with LSTM. The other feature extraction techniques Pitch and energy also presented satisfactory results with several classifier such as KNN and SVM.
- How can we presented best combination feature extraction and classifier to obtain best IWR system? We presented new scientific contribution which we combined the feature extraction techniques such as MFCCs, Pitch, and Energy with SOFTMAX which is not used in previous presented studies and presented satisfactory results. The SOFTMAX is new trend classifier which used probability logic to classify the data in multiclass categories and used with deep learning techniques in the last years.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we review general concepts that are related to VAD and speech enhancement and discuss their functionality in motivating the research in the sub-sequent chapters of this thesis. Discrimination between speech and noise is crucial for both VAD and speech enhancement algorithms. In order to provide a comparative analysis, this chapter presents the characteristics of speech and noise. Due to its role in determining speech characteristics, human speech production system is also explained.

2.1 Fundamentals of Human Speech

Speech signals are time varying pressure waves that are transmitted by a speaker in order to communicate information [8]. Voice signals are composed of a sequence of tones. These tones and the transitions between them serve as a symbolic representation of information [9]. In order to apply signal processing techniques for VAD and speech enhancement, it is essential to understand the speech production process and the structure of human speech.

2.2 Speech Production

The source of the human speech is the airstream produced by the lungs. The air flow produced by the lungs is perturbed by a constriction somewhere in the vocal tract. The air flow in the vocal tract changes the air pressure at the lip end. This, in turn, results in the radiation of acoustic waves. These radiated waves are perceived as speech by listeners. The organs in the vocal tract such as the teeth, tongue and etc. are referred as the articulators of vocal tract and they determine the type and place of constriction during speech production process. In the average male, the total length of the vocal tract is about 17 cm. The cross-sectional area of the vocal tract,

determined by the positions of the articulators varies from zero (complete closure) to about 20 cm2 [9]. Figure 2.1 shows a schematic diagram of the vocal tract.

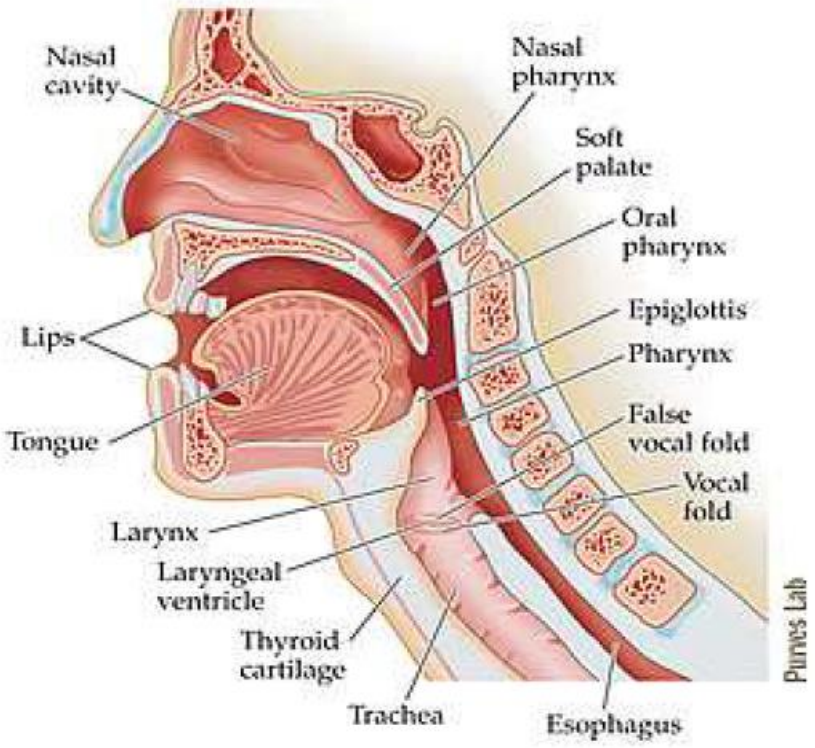

Figure 2.1 Vocal tract for human speech production [10]

Human speech production system can be modelled by the well-known source/filter production model. Figure 2.2 depicts the source/filter production model. The model contains a time varying digital filter which is driven by an excitation function. The excitation function is a periodic impulse train with a period of the pitch period for voiced sounds. For unvoiced sounds, the excitation function is a random noise generator. The changes in the shape of vocal tract caused by articulators to produce different sounds are modelled by representing the vocal with a time varying digital filter. A variable gain factor determines the intensity of the produced speech.

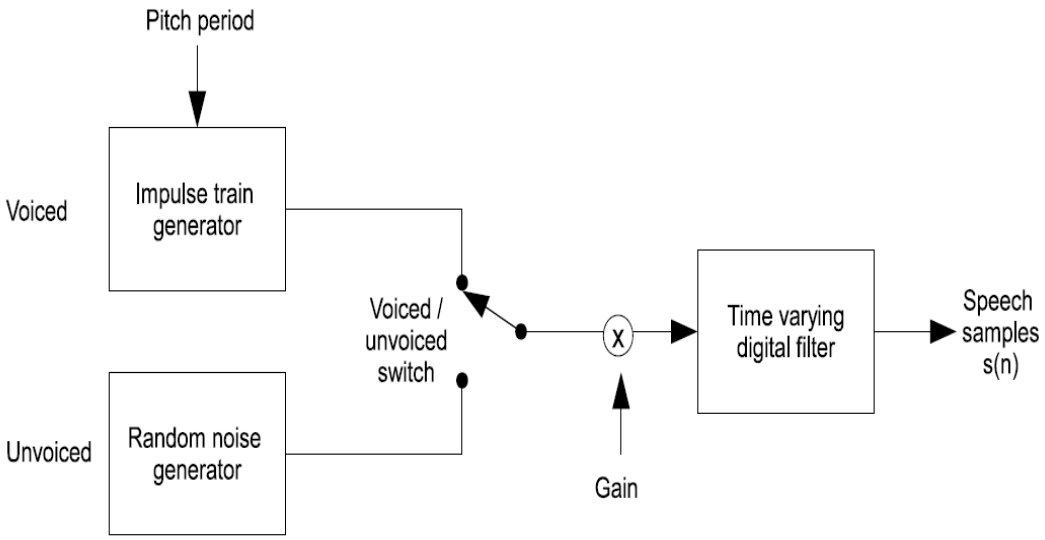

Figure 2.2 Source/filter model for speech production

2.3 Characteristics of Speech

The frequency spectrum of human speech is shaped by the frequency selectivity of the vocal tract. The periodic air flow associated with voiced sounds results in the formation of the resonant frequencies (formant frequencies in speech processing terminology) in human speech spectrum [9]. Peterson et al. [11] demonstrated that the first three formants of vowels in English are located at frequencies lower than 3 kHz. A similar study was done in Türk et al. [12] for Turkish vowels and this study indicated that the first three formants of Turkish vowels are also located at frequencies lower than 3 kHz. The first three formants contain much of the energy of voiced sounds and they are generally enough to characterize voiced speech.

Spectrogram plots are frequently used in the analysis of the time-varying characteristics of human speech. In a spectrogram plot, the horizontal dimension corresponds to time and the vertical dimension demonstrates the energy distribution over frequency. The darkness of the spectrogram plot is proportional to the signal energy [9]. In a typical spectrogram, voiced sounds are characterized by a striated appearance due to the periodicity of the time waveform. On the other hand, unvoiced sounds are more solidly filled in due to their broader spectrum [9].

2.4 Characteristics of Noise

In general terms, noise refers to the unwanted random signal added to some desired signal. Noise added to human speech degrades the perceptual aspects of speech such as quality or intelligibility [13]. Noise may be either correlated or uncorrelated with the actual speech signal. In this thesis, we will only be dealing with the additive noise problem, where the speech signal and noise are uncorrelated.

2.4.1 Statistical Properties of Noise

Since noise is a random phenomenon, noise suppression algorithms use various statistical properties of noise for analysis. The most important characteristics of noise for analysis are stationary and spectral energy distribution over frequency. In broadest terms, a random process is called stationary if its distribution functions or certain expected values are invariant with respect to a translation of the time axis. The degree of stationary for a random process ranges from stationary in strict sense to a less restrictive form of stationary called wide-sense stationary.

2.5 Survey

Liebman [l4] has a good review of the ANN application on automatic speech Recognition. One of the main problems when designing ANN's Speech recognition is the way to deal with dynamic characteristics from the speech signal. Possibly the easiest way to solve the problem is merging a fixed ANN model with a traditional model the speech recognition process that processes dynamic information. Multilayer Perceptron (MLP) can be used to provide local distance values dynamic programming-based detection algorithm [15] Vector quantitative learning is neurologically inspired [16] Can be used to provide a high performance notebook for MLP can also make a separate hidden markov model (HMM) [17] It can be used as a therapist to make a classification decision For the time of the input statements, which are organized by HMM Viterbi Track back [18] or track back hash [19]. To process dynamic information using the ANN model, The neural network model (TDNN) was a time lag [20] suggest. TDNN contains short entries and delays Hidden layers of MLP so the node responds Over several time periods they are sent together to

Neurons in the upper layer. Because the signal is flowing inside TDNN is only a feed character the network can do they are trained by the wrong back propagation algorithm [21]. ANN'S with feedback links and feedback (frequent) Neural networks) were also examined for speech recognition Purposes. Brager et al. [22] Examine the use of repetition Boltzmann speech recognition machines, Robinson et The. [23] and Anderson et al. [24] Studied networks with Repeated connections from the output node to the input node Watros and Shastree [25] suggested frequent networks Self-loop connections on hidden nodes and exit. Almeida [26] A modified version of the error spread is proposed Repeat the ANN training algorithm. Generally common By comparison, ANNs are more difficult to train and analyze With ANN'S FEEDBACK. Two recent studies have proposed [27] ANN instead of expecting a non-linear model from Distinguish language frames. Speech recognition was this is done by finding the model with the least expectation Error in dynamic programming routines.

Pretty Saini, Barnet Core and erased Dua have one Indian speech recognition system with HTK [28]. The system was developed using the hidden Markov model HTK Toolkit. MFCC is used as an identifier by talk. The system is trained and can recognize only 113 words. The system works in both headphone and speaker dependent formats Independent context. A continuous language for the application the detection system in Hindi was developed by Gaurav Devanesamonim, Shakina Devi, Gopal Krishna Sharma and Mahua Bhattacharya [29]. It is a separate phone a system based on the use of MFCC as a distinct language Parameters and hidden Markov model for feature extraction Development of the model. Phonics is used to represent words. That the system currently includes the use of 29 sounds. Other characteristics In addition to MFCC, it was also used in development Detection system. MK Linga Murthy, GLN. Murphy They has developed a word dependent on the speaker isolated in real time English language recognition tool [30]. The system uses linear predictive coding for feature analysis and extraction of feature vectors. The quantity of the vector is determined in Extracted features to get the corresponding icons Word and these icons are stored in the database. During The recognition level found a match for the word entered by Compare the similarities between the code and the symbols in Similarity database and classification are created. The system uses 2 speakers. System vocabulary it consists of five words and

each word has ten words registered. Apart from this business, many popular commercials Speech recognition applications such as Dragon Naturally Speaking, IBM via Voice and Microsoft SAPI. Several studies and experiments were also conducted to integrate MFCC with other language features to create efficient detection system. Mayur R Gamit, Kinnal Damilia developed a detection system that combined Both MFCC and LPC (linear predictive coding) using neural Network as a workbook [31]. The vocabulary of the system consists of English numbers zero to nine. Register for each Word Made by 28 speakers - 14 men and 14 women. The Experiments are conducted to evaluate the performance of a system that uses only MFCC and a system that combines MFCC and LPCs.

2.6 Time Domain Parameters of Speech Recognition

The implementation of the time domain analysis is simple. Time domain analysis transformation the audio signal of a series of parametric signals changes very slowly over time as before Signal. This means that the associated language parameters can be saved or edited more efficiently. Original signal. Several parameters are required to cover the relevant aspects of the speech. This can be done by sampling the signal at low speed. Short-term treatment method. It delivers parameter signals in the form of Q (n) in the time and frequency domain [32]. Where Q (n) Written in Equation 2.1:

$$Q\ (n) = \int_{-\infty}^{\infty} T[s(m)]w(n-m) \qquad (3.1)$$

$Q\ (n)$ is the output of the model, The speech signal S(n) undergoes transformation T.

- **Short-Time Averaging Zero-crossing Rate (ZCR)**

You need one or four frequencies to measure the spectrum of an audio signal. Becoming one With a simple spectral measurement technique called zero crossing speed, suitable spectral information can be generated from S (n) signals at low cost. The signal traverses the time axis and the code changes. 2 zeros per sine cycle [33]. The equation of this model presented in (3.2):

$$f_0 = [ZCR * f_s]/2 \qquad (3.2)$$

Which the ZCR choice between unvoiced and voiced, and the f_0 is the frequency sampling. The Figure 2.3 presented the example of the ZCR executed in Matlab**.**

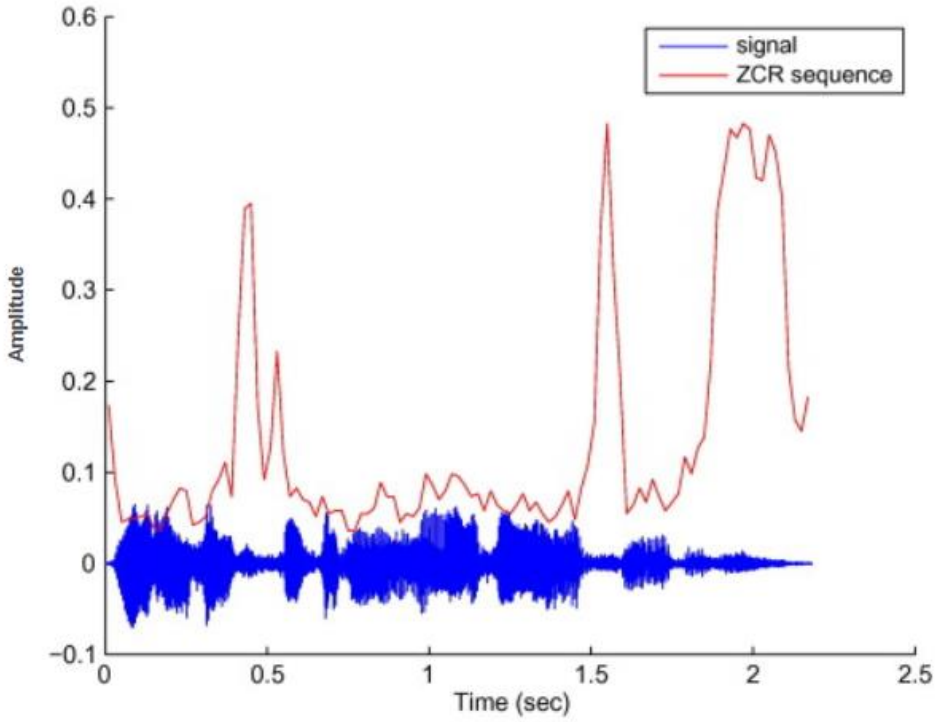

Figure 2.3 ZCR

- **Short-Time Autocorrelation**

The amplitude of the Fourier spectrum of the speech signal S (n) and Step. The time signal r (k) is called for the inverse Fourier transformation of the energy spectrum Autocorrelation function S (n). The function r (k) stores type information Harmonic periodicity and amplitude S (n). R (K) ignores the phase value S (N). Information below the spectral value. The autocorrelation function can be written as follows in Equation (3.3):

$$\emptyset_{sy}(k) = \sum_{m=-\infty}^{\infty} s(m)\, y(m-k) \qquad (3.3)$$

By using this model, the autocorrelation function calculated for each y (n) and s(n) with the same signal [34]. The Figure 2.4 show the Autocorrelation.

2.7 Speaker Independent/ Speaker dependent

Speech recognition software It depends on the knowledge of the specific vocal characteristics of the speaker. The software learns the speaker's vocal characteristics through language training (or recording). This type of system is trained so that a specific user recognizes what has been said. This type of system works well when only one user is communicating with the system.

Furthermore, Speaker Independent Software that does not require training is independent of the presenter. This system is used for the automatic telephone interface. The system can be used by a large number of untrained people to identify the type of language each person is using.

Speaker recognition Identify people based on their voice characteristics. With the recent proliferation of smart devices, a lot of research has been done on how to identify speakers in order to provide users with a personalized experience. However, most searches are based on the fact that the user utters a keyword to activate the identified device (text-dependent speaker recognition). This creates a little monotonous interaction with the device. This research was conducted to overcome these key and language barriers and help users understand what they are saying.

The human body consists of several main sensations. Visualization of sensory perception with the eyes on the skin ear with body window Communicate with the basic physiological environment to understand this meaning inspired by the environment and the result converts organic stimuli into a series of signals treatment (chemistry and physics). The nervous system it therefore receives complex and dynamic signals. Nonlinear analysis to do what you need after this. But the auditory nerve a complex human sensory system that empowers people acoustically it's a lot it is important in the learning process and in everyday life. The physiological process of the ear route to understand how signals are converted the auditory canals then connect with what you need late [35].

2.8 Feature Extraction

Since the data properties are not necessarily non-correlated (array X is not a complete rank), they share some information and therefore there is usually fake information in the data pattern. Therefore, when plotting x ∈ R d 7 → y ∈ R p is usually p p or at least p <d, where p is called the fundamental dimension of the data. This is also called the diversity hypothesis, which states that data points are located on a sub-distributor or subspace of lower dimensions. This is the reason why feature extraction and dimensional reduction are often used interchangeably in the literature. Usually the primary goal of dimension reduction is either a better representation / differentiation of data or a simpler visualization of the data. It is worth noting that the area of R p is referred to as the feature area (i.e. feature extraction), the embedded area (i.e. embedding), and the encoded area.

(I.e. coding), subspace (i.e. learning in subspace), low-dimensional space (i.e. dimensional reduction), subsection (i.e. diversified learning) or representational space (i.e., representative learning) in literature . Dimensional reduction operations can be divided into two main categories i.e. observed and unmonitored actions. Observed methods take into account the names and categories of data samples, while unattended methods are based on the variety and style of data. Another classification for feature extraction consists of the division of methods into linear and non-linear. The former assumes that the data falls on a linear subspace or that data categories can be distinguished linearly, while the latter assumes that the data type is more complex and that it is present on a nonlinear sub-distributor. Various methods for extracting the feature are examined below. For the sake of brevity, we mention the basics and concede some very detailed methods and improvements to these methods, for example b. Use the Nystrom method to include data outside the sample.

Feature extraction is a process of reducing attributes. Unlike the selection of entities where the existing attributes are classified according to the meaning of the prediction, the attributes are actually transformed during the extraction of entities. The transformed attribute or characteristic is a linear combination of the original attributes.

The feature extraction process makes the attribute set much smaller and more extensive. Models based on extracted entities can be of high quality because the data is described with less significant attributes.

Feature extraction projects a larger dimension data set into a smaller number of dimensions. Therefore, reducing a complex dataset to 2D or 3D can be useful for visualizing data, which is useful for visualizing data.

The uses of feature extraction are potential semantic analysis, data compression, data decomposition and projection, and pattern recognition. Function extraction can be used to improve the speed and efficiency of supervised learning.

You can use the extraction function to extract topics from your document collection. Documents are represented by a number of key words and their frequency. Each argument (function) is represented by a combination of keywords. The documents in the collection can be expressed according to the topic found [36,37,38].

2.9 Classification

Classification is a key topic in machine learning that has to do with learning machines to group data according to certain criteria. Classification is the process by which computers group data based on certain characteristics; This is called supervised learning. There is an automatic version of the classification called grouping, in which computers find common characteristics for which data can be grouped when no category is specified [39].

A common example of classification is spam detection. To write a spam filter program, a computer programmer can train a machine learning algorithm with a series of spam emails identified as spam and normal emails identified as spam. Spam is marked. The idea is to create an algorithm that can learn the properties of spam emails from this training set so that spam emails can be filtered when new emails are displayed.

Classification is an important tool in today's world, where big data is used to make all kinds of decisions in government, business, medical, and other settings.

Researchers have access to large amounts of data, and classification is a tool they can use to understand data and find models.

While machine learning classification requires the use of complex algorithms (sometimes), classification is something that people naturally do every day. When they classify things, they are simply grouped by similar characteristics and attributes. If you go to a grocery store, you can roughly group foods by food group (cereals, fruits, vegetables, meat, etc.). Machine learning in classification involves teaching computers to do the same [40].

CHAPTER 3

MATERIAL AND METHODS

In this section, several techniques presented which used in this study. The presented Isolated Word Recognition consists from three parts. Data preparing, feature extraction, and classification.

3.1 Free Spoken Digit Dataset (FSDD)

In in study, Free Spoken Digit Dataset (FSDD) which published in Kaggle are used as meterics dataset to vallidate our methods. A humble audio/speech dataset involving of footages of spoken digits in wav files at 8kHz. The records are clipped so that they have close nominal stillness at the early stages and ends. The all cases that are between 0 and 9 are presented as audio sound. Three person which as called Jackson, nicolas, and theo are the persons who repeated the each number 50 time. The whole dataset consist from 1500 instances each digit repeated 150 time which each person repeated each digit 50 time.

The main contribution in this dataset is that it's the first English pronunciations that's contain large number of instances which assist the researcher to validate the presented methods in the speech recognition and voice activity recognition fields.

The dataset also become very important in the last previous years because of the increasing the applications dealing with the smartphones which the developed applications need to developed and validated before presented as commercial products. In this study, we applied 10 instances from each case for testing which mean 100 sample used for teasting in complete.

3.2 Feature Extraction

3.2.1 Mel-frequency cepstral coefficients (MFCC)

Mel-frequency cepstrum is a short-term depiction of the power spectrum of a audio signal. The MFC depend on linear cosine convert of a log power spectrum on a nonlinear mel-scale of frequency. MFCC is the coefficients composed applying MFC investigation which are derivative from a cepstral kind depiction of a audio signal. The modification a,omg MFC and cepstrum is; the frequency bands spread out on mel-scale in MFC which stimulated from the human aural system's reply carefully extra than linear frequency bands in the normal cepstrum [41]. The procedure of MFCC features extraction shown in figure 3.1.

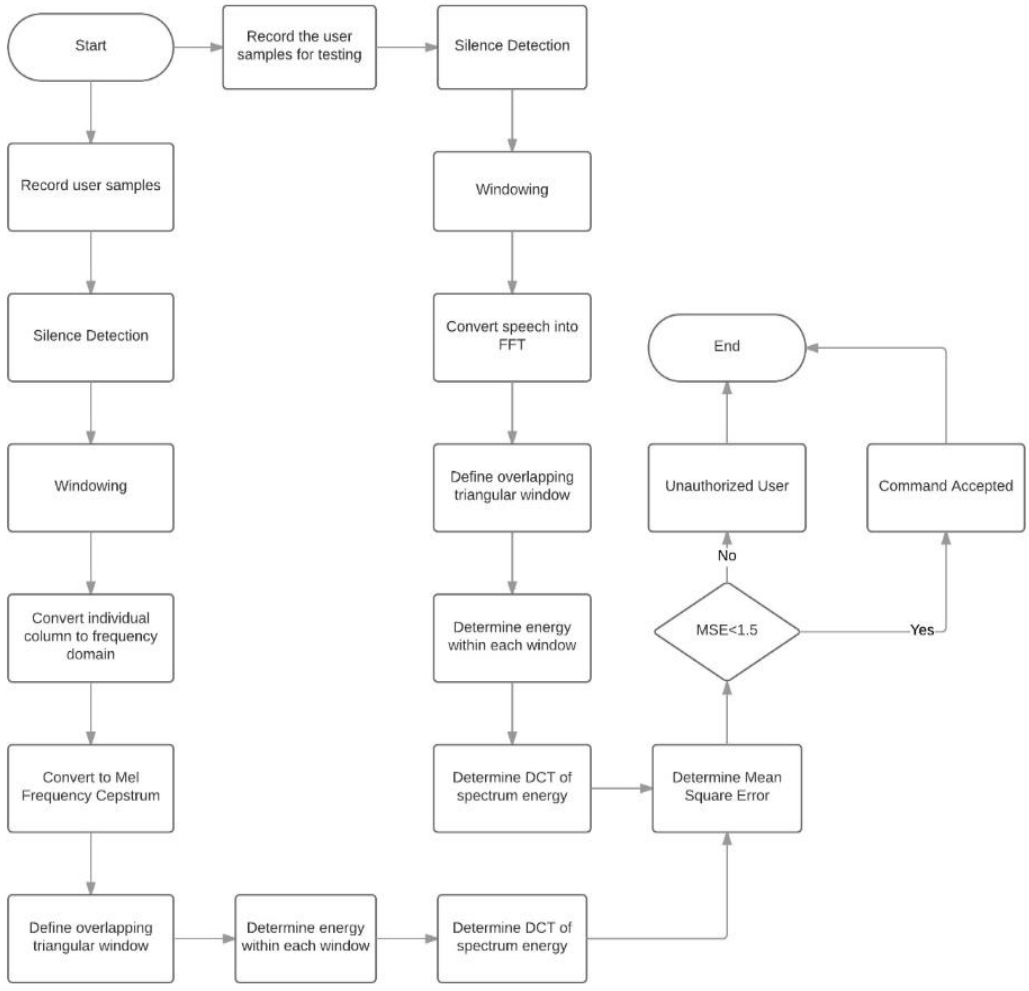

Figure 3.1 Co-Learning Algorithm

3.2.2 Pitch

Pitch is a perceptual characteristic that classifies tones on a scale related to frequency. The sound can be quantized as a frequency called the fundamental frequency (F0). Changes in the pitch and pitch of words form the pitch of a tonal language such as Chinese [42].

RAPT is a time domain algorithm for estimating F0 on which the Kaldi tone function is based. Most F0 estimators have very important steps:

- Preprocessing, counting DC-eliminating and inclosing, etc.;
- Applicant making, main technique to predicate the F0;
- Post processing, selecting the best applicant and cleansing the outcome yield.

3.3 Voice Activity Detection Using Energy

Many speech processing algorithms require significant resources and require significant processing power or transmission bandwidth. However, the language is intermittent, so there are often pauses between sentences and gaps, even within sentences. In addition, in a dialogue, the speaker generally uses an educated sequence so that the others remain silent when a person speaks. No resource-intensive handling is required for language interruptions. Therefore, there is great potential to save resources by disabling advanced speech processing methods when the input signal contains no voices.

Voice activity detection (VAD) relates to the task of determining whether a signal contains speech or not. So, it's a binary solution. A related task is to determine the probability that the input signal contains or does not contain speech, which is referred to as the probability of speech presence (SPP). The SPP is therefore generally expressed in terms of a probability between 0 and 1. The likelihood of speech being present is generally an intermediate step in the detection of speech activity, so that the classification of speech activity is obtained by defining a threshold at the output of the output signal to estimate the likelihood of speech being present.

In general, algorithms for recognizing speech activities are relatively simple, so more complex activities such as speech recognition should only be used in the presence of speech. Similarly, when coding speech, we should only transmit speech in the presence of speech, and we can reduce the bit rate if there is no speech.

Voice Activity Detection (VAD) is a technology that detects the presence or absence of human speech. Detection can be used to start a process. VAD has been used in voice-controlled applications and devices such as smartphones that can be operated using voice commands. According to the Disability report issued by the Statistics Agency of Namibia (NSA), the 2011 census showed that 98,413 people with disabilities live in Namibia. This report also said that about 64% of people with disabilities use radio as an ICT resource for communication, which means that these people are not using other technologies available today that can automate home operations. Most voice-based home automation systems rely on remote control or smartphones for home control or rely on commercial ASR-API (automatic speech recognition), intended for general use and therefore not specifically developed for home use automation orders [44,45].

Most of the proposed techniques for VAD can be separated into two processing steps:

- First, features are selected from the loud speech signal in order to realize a distinction between noise and speech.
- In the second phase, the recognition pattern is used to the characteristics that lead to the final decision.

Classification becomes more difficult with multi-dimensional features. Progressive classifiers such as support vector machine (SVM) or naïve Bayes classifier (NB) can be trained to classify noise and speech based on feature vectors. Finally, a binary and modified decision for these features is also made.

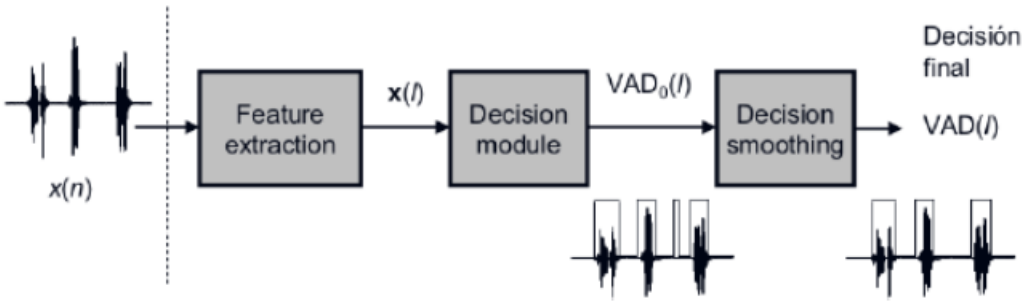

Figure 3.2 VAD Block Diagram

The energy-based VAD methods are the greatest famous methods and are extensively applied in speech recognition software. A change of energy-based algorithm approaches has been recently presented for healthy VAD. With very little computational complexity, for clean speeches or speeches with less noise, these methods have good performance. Energy is a humble amount of the power of the signal. We can adopt that speech is continuously brasher than related noise, and we can allocate the high-energy frames to speech and lesser ones to noise for VAD. When the SNR is low, the simple energy feature cannot separate speech and noise. Energies of sub-bands were used as features in earlier work on energy based VAD to increase the discriminative power of the VAD. Another approach to increase noise robustness is to combine energy-based features with other features, like zero-crossing rate (ZCR), or the line spectral frequency (LSF). In general, these methods work well with clean speech or high SNR conditions. But, their discriminative power falls significantly under high noise level such as when SNR falls below 10 dB. Nevertheless, through with their low computation complexity, energy-based methods are still developed by some standards and different real-world applications. The mathematical model of energy represented in the Equation 3.1.

$$E = \int_{-\infty}^{\infty} x(t)^2 \, dt \tag{3.1}$$

Where the ∞ and $-\infty$ represented the upper and lower limit for the frame, $x(t)$ represented the audio data that it energy will be calculated.

3.4 Isolated Word Recognition Framework Based Deep Learning

In this section, new approach presented Isolated Word Recognition using pitch feature extraction based SoftMax. In the first stage, the dataset prepared and read by using MATLAB2020. The read of data section is very important which the data also arranged and divided into train and test section. The aim of dividing data into two groups is to train the system with a group of data and test with another group to calculate the real capability of the proposed method.

In the second step, three VAD techniques proposed to detect the voice activity which is the more critical issue in this study. the both Bohman window and Bartlett-Hann are two new ideas applied to VAD problem and also Hamming window also used but the Hamming used in previous studies too which is used only to compare with our proposed VAD techniques.

On the other hand, three techniques applied as features extractor in the same system such as: pitch, MFCC, and energy. These techniques are applied in several studies as features extraction from audio data. These techniques presented high and remarkable results in several audio data classification and processing problems. The pitch and energy function represented best results than MFCC with Softmax classifier because the SoftMax deal with fixed size samples features across the LSTM and RNN that are deal effectively with time series data and presented suitable results in audio recognition data exactly when combined with MFCCs because MFCCs presented time series output with multiple dimension and not fixed features sizes.

Then, the output of the features extraction function wired to the SoftMax probability function which applied as classifier. The SoftMax function calculates the probability distribution for “n” different cases. The function as a whole calculates the probabilities of individual label classes of all possible label classes. The calculated probabilities are useful for choosing a label class for each item.

The main advantage of using SoftMax is the likelihood of leaving the region. The domain ranges from 0 to 1, and the overall gain is 1. If the SoftMax function is used as a model for multiple classifiers, the probability of each label is generated, and the probability of a good label is highest.

The mathematic model of SoftMax function [49, 50, 51] is presented below, where the vector of inputs to the output layer represented by z [74]. And again, j indexes the output units, so i = 1, 2, ..., K see equation 3.2.

$$SoftMax(y) = \frac{e^{zi}}{\sum_{k=1}^{k} e^{zak}} \tag{3.2}$$

Our method presented in the flowchart that explains detail of the isolated word recognition and the system presented in the Figure 3.3.

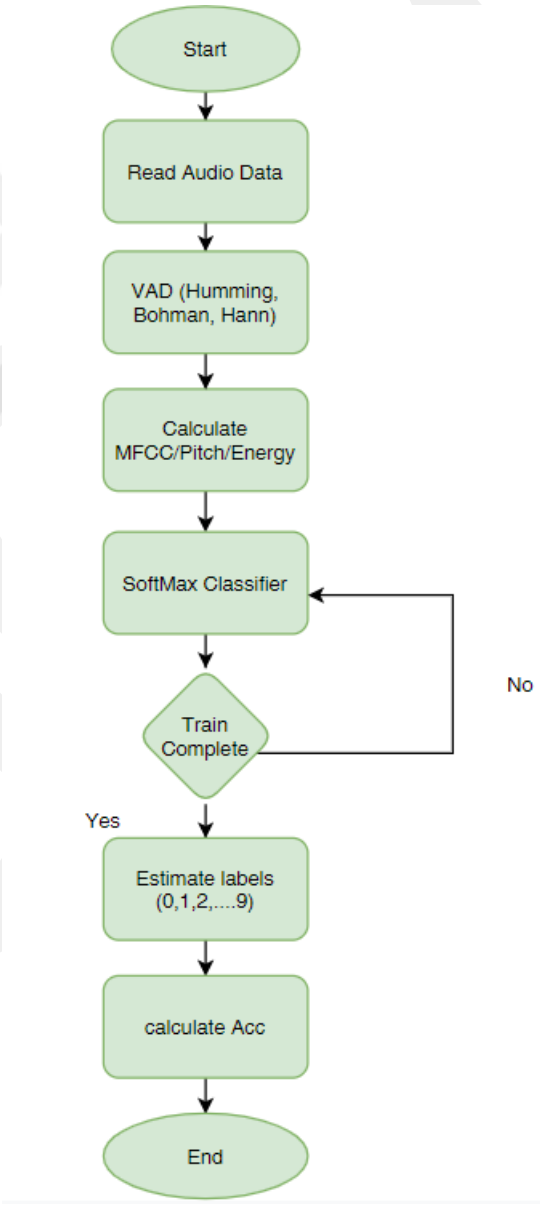

Figure 3.3 Our Framework

CHAPTER 4

EXPERIMETAL RESULTS

In this section several techniques are presented dealing with voice activity detection, and isolated word recognitions. Results about these techniques explained and discussed.

3.4 Voice Activity Results

In this section, three VAD techniques applied to the audio digit recognition data and the results presented and discussed. The three techniques are hamming, Bohman, and Bartlett-Hann.

4.1.1 Hamming

hamming code applied for voice activity detection we applied the hamming code with ten data types such as: 0, 2, 3, 4, 5, 6, 7, 8, and 9. The VAD of all words presented in the Figures (4.1-4.10) The frequently used in signal processing applications hamming window, as an argument only It has a window length, and with Equation (4.1) it is defined:

$$w_h(\mathrm{n}) = 0{,}54 - 0{,}46\cos\frac{2\pi\mathrm{n}}{\mathrm{N}-1} \qquad (4.1)$$

The 0,54 is the a_0 value that's in Hamming window its setting to 0,54 or exactly 25/46. The hamming window also called hamming blip when used in pulse shaping.

We conclude number of advantages in hamming window which Computers cannot compute on an infinite number of data points, so all signals are "clipped" at both ends. This causes ripples on each side of the peak you see. The Hamming window reduces this ripple and gives you a more accurate picture of the frequency spectrum of the original signal.

It is observed that the value of $w_h(n)$ in the active area is always larger than the inactive voice area. In the absence of noise, the inactive speech area (s) is zero and the active speech area can be easily found by checking whether $w_h(n) > 0$. However, if the voice is contaminated with noise (s), the adaptive thunderstorm increases regularly. "AWT" is zero even in areas where the voice is not active.

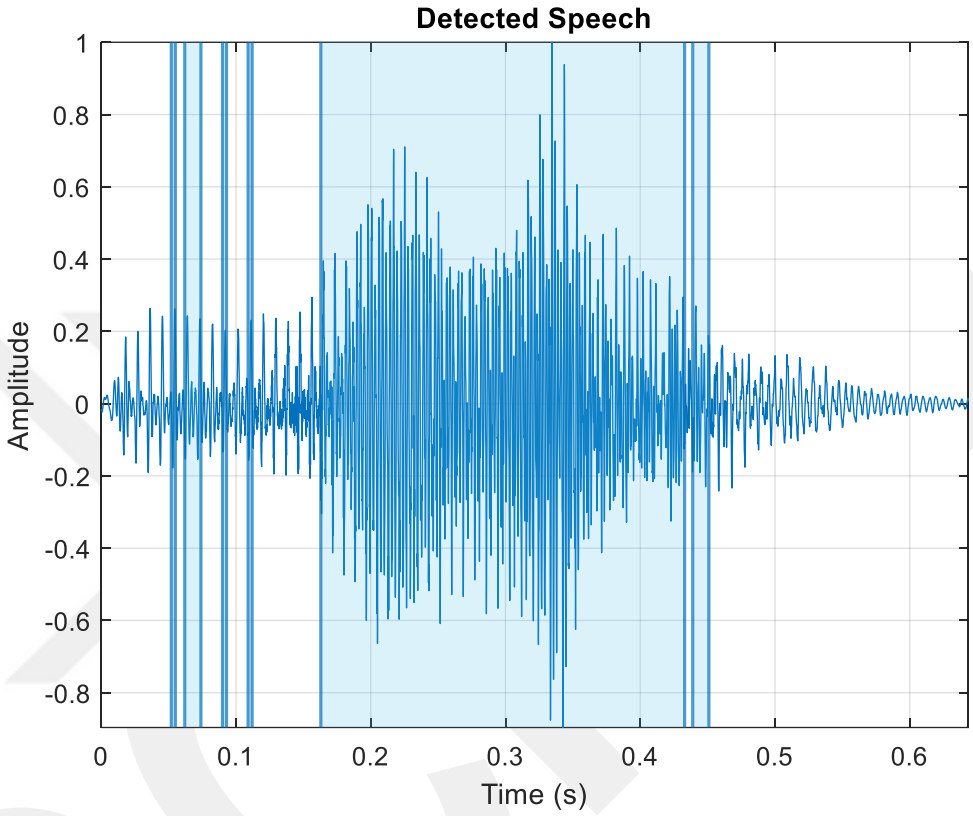

Figure 4.1 VAD for 0 using hamming

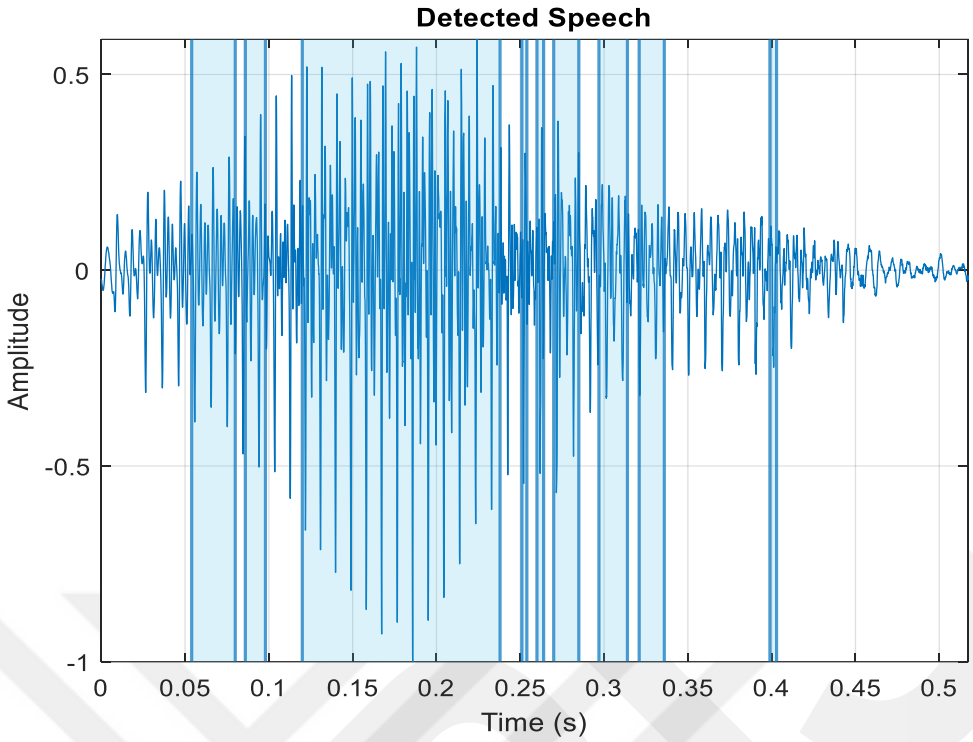

Figure 4.2 VAD for 1 using hamming

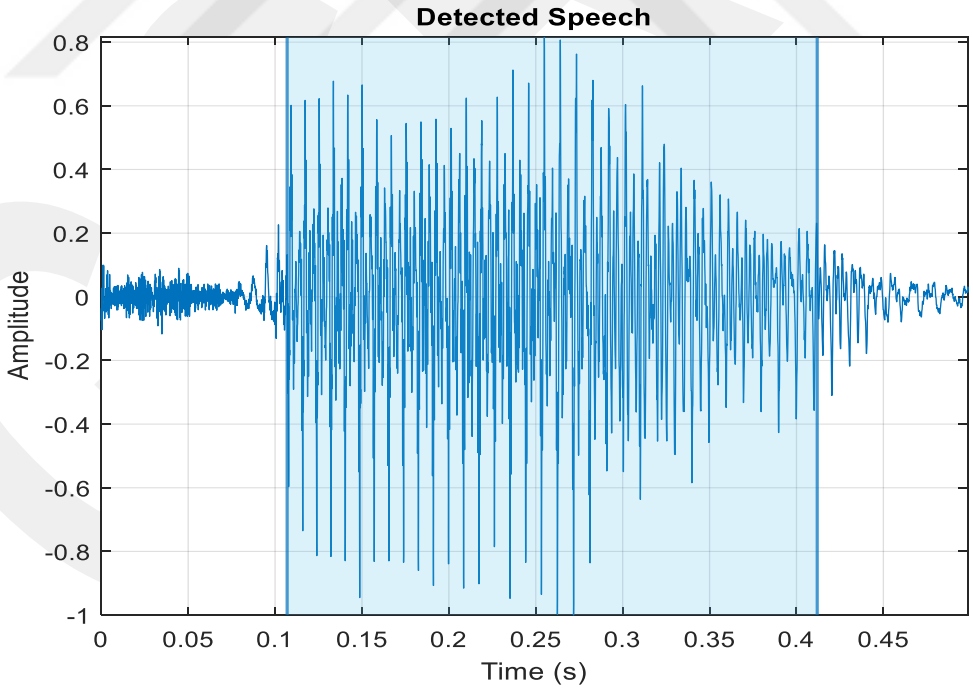

Figure 4.3 VAD for 2 using hamming

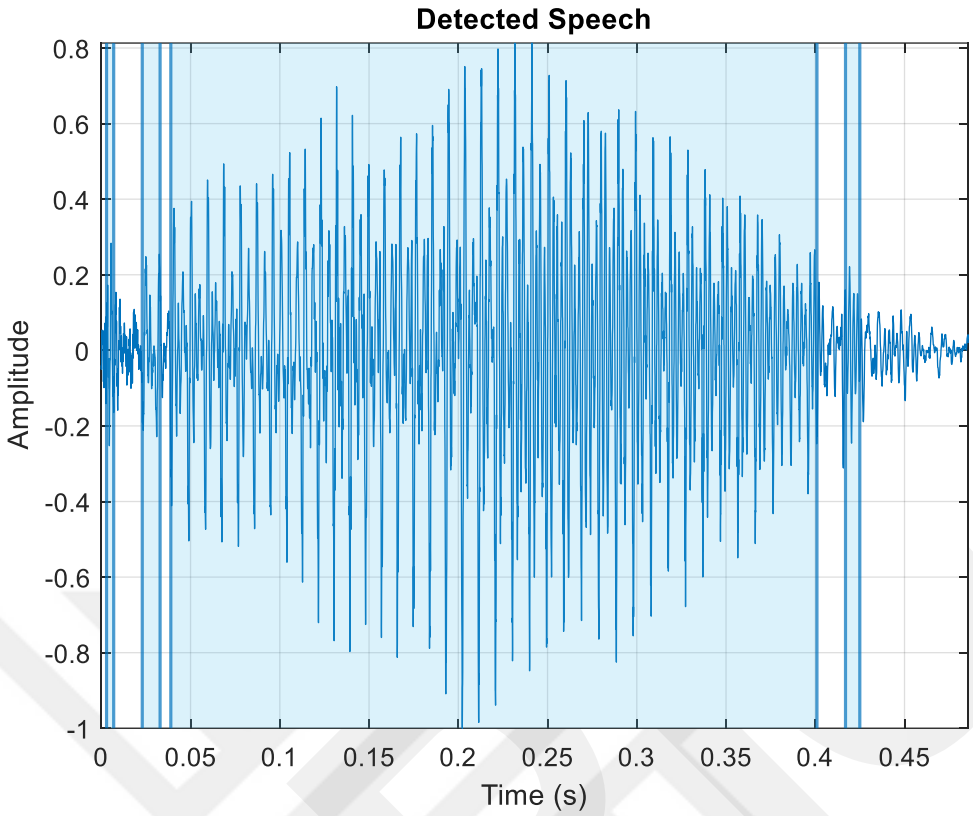

Figure 4.4 VAD for 3 using hamming

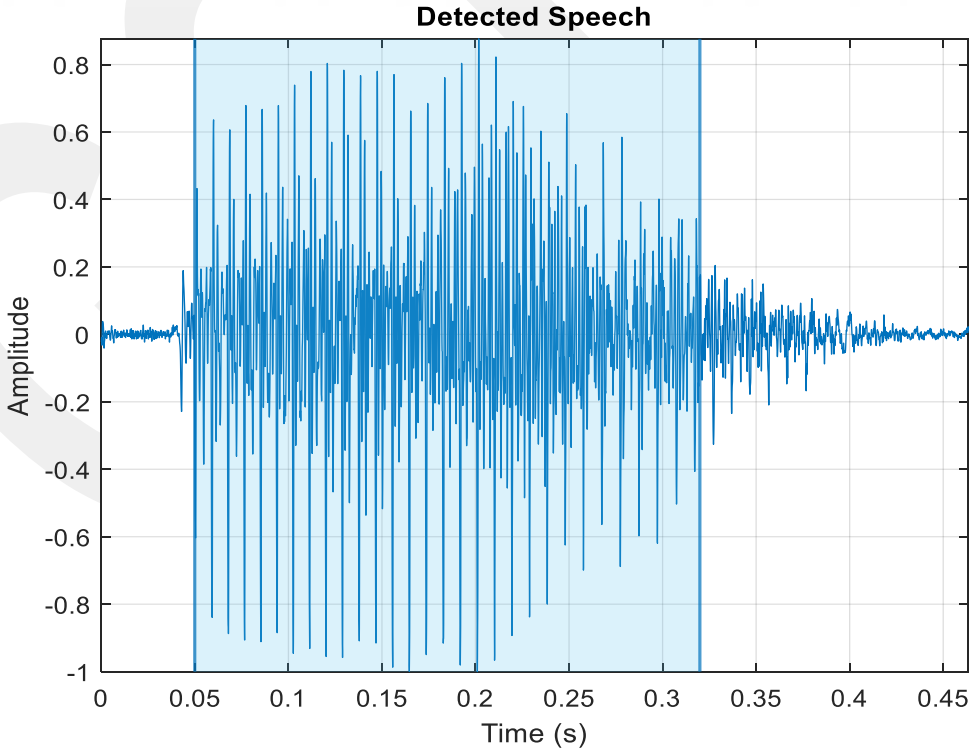

Figure 4.5 VAD for 4 using hamming

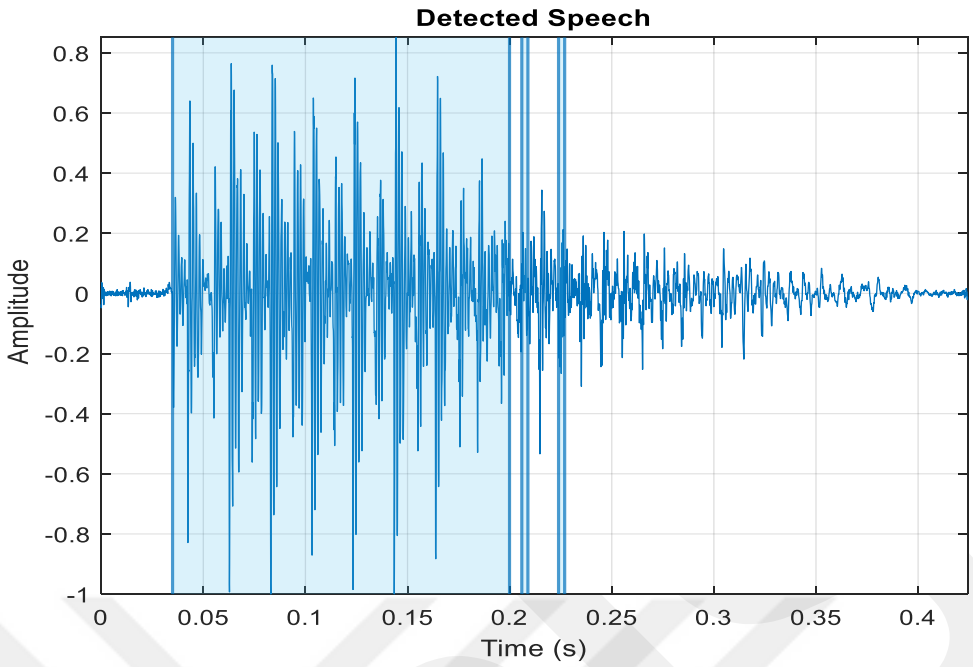

Figure 4.6 VAD for 5 using hamming

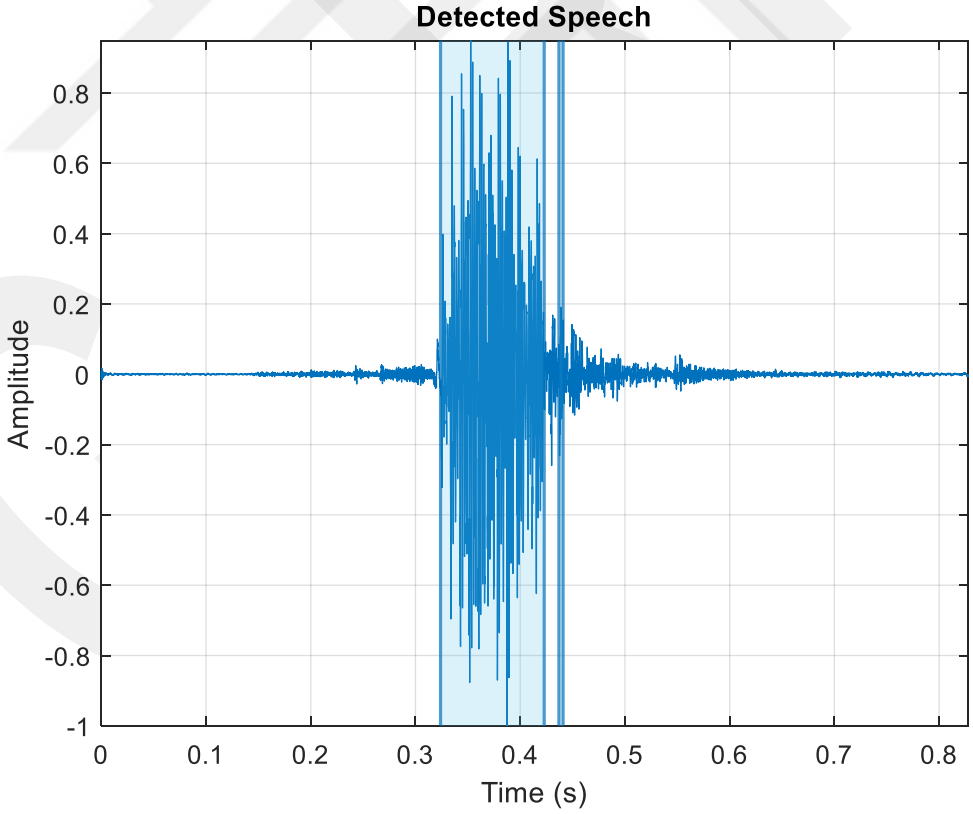

Figure 4.7 VAD for 6 using hamming

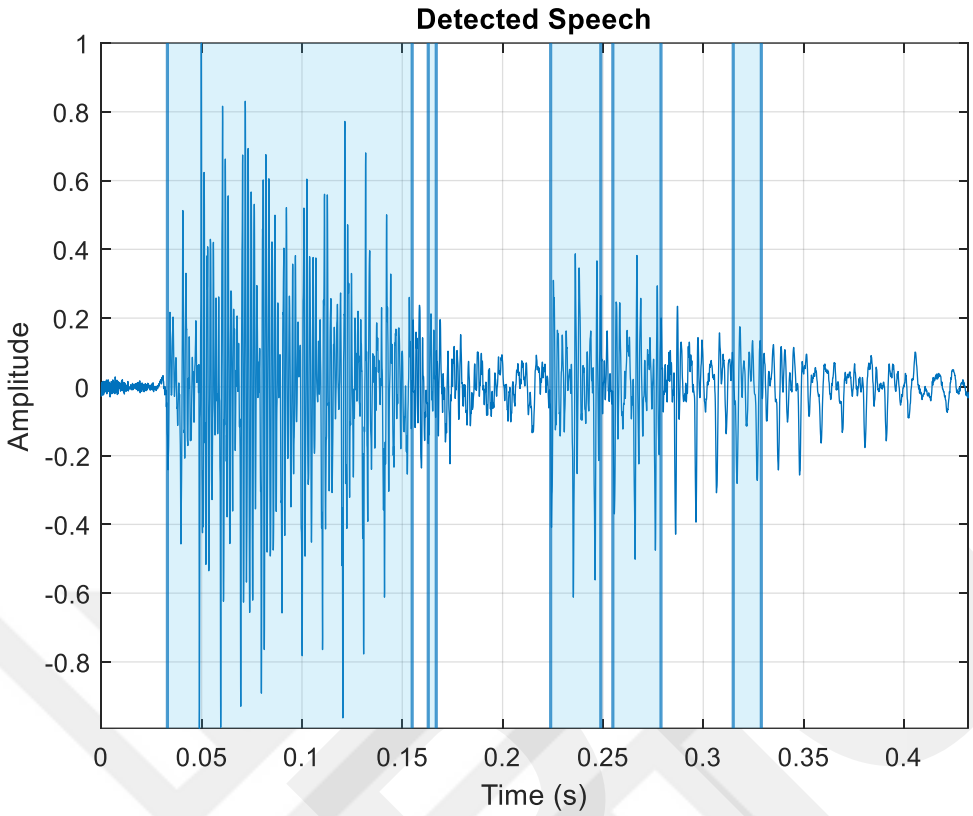

Figure 4.8 VAD for 7 using hamming

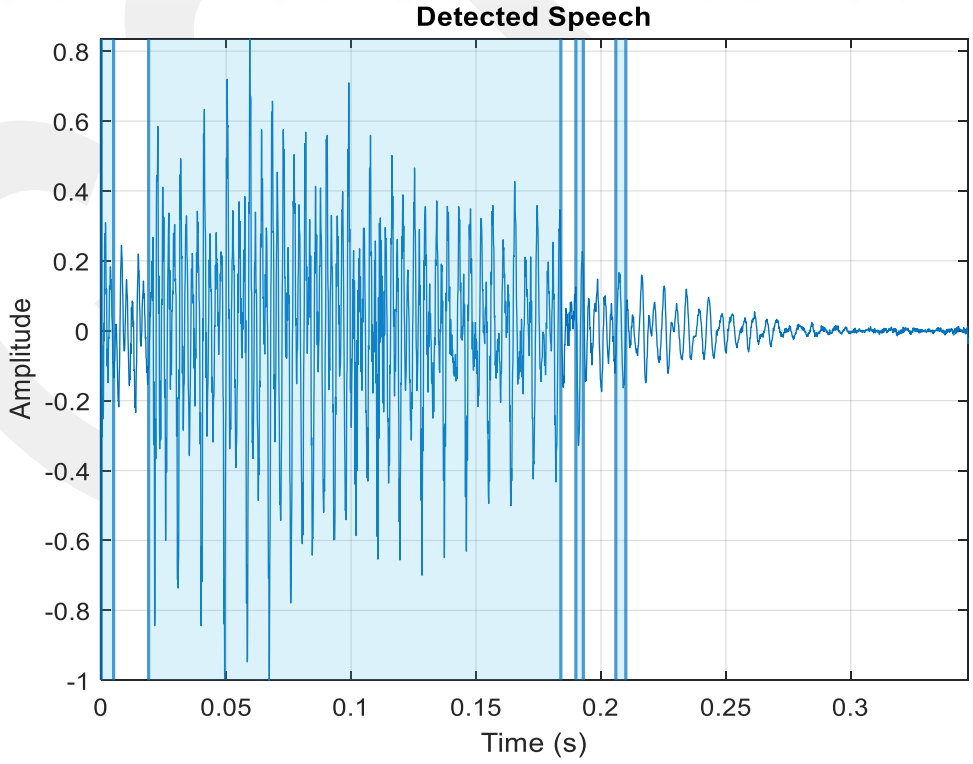

Figure 4.9 VAD for 8 using hamming

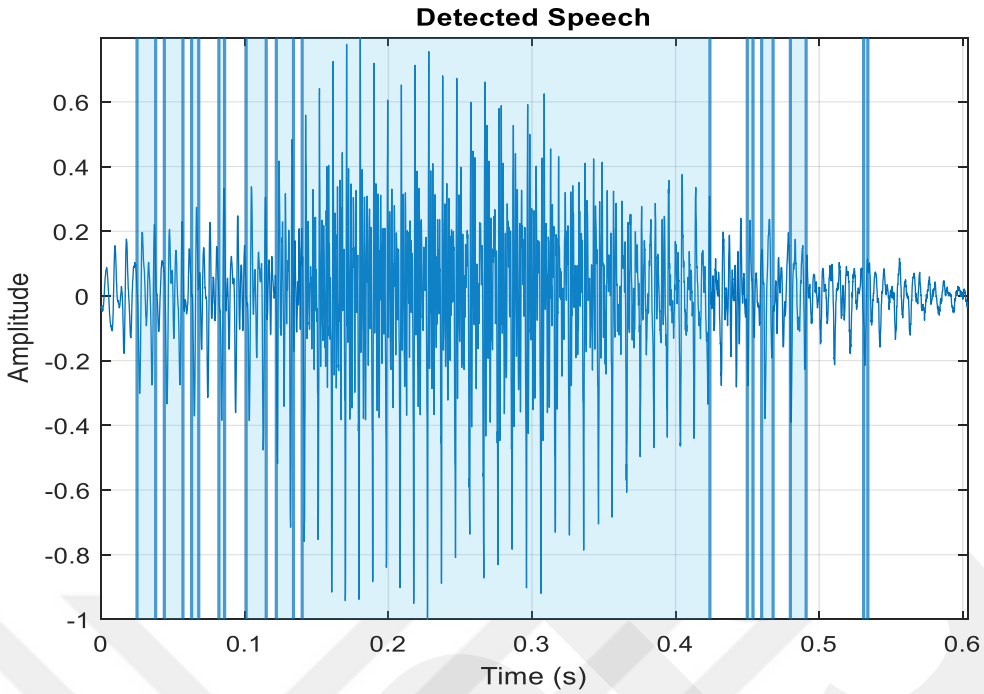

Figure 4.10 VAD for 9 using hamming

4.1.2 Bohman window

The bohman window is one of the power functions that we applied it to VAD problem. Where X is the complex valued input sequence and the Windowed X is the input signal with the window used. A Bohman window is the convolution of two half-duration cosine parts. In the time domain, it is the produce of a trilateral window and a lone series of a cosine with a period added to group the first derived to zero at the margin. Mathematically the Bohman window can represented as shown below in (4.2):

$$w(n) = 0.1 - \left[\frac{\left(n-\frac{n}{2}\right)x^2}{\frac{n}{2}}\right] cos\left[\pi\frac{\left(k-\frac{n}{2}\right)}{\frac{n}{2}}\right] + \frac{1}{\pi} sin\left[\pi\frac{\left(k-\frac{n}{2}\right)}{\frac{n}{2}}\right] \quad (4.2)$$

Where 0<=/K/<=N

By analysing the results of VAD for Bohman window we can conclude number of points:

The conversion area and the chief lobe are wider than those for Hamming window but the stop band reduction is upper, consequently. The reduction of the chief side lobe for Bohman window is 46dB, while the filters considered with Bohman window have the stop band reduction of 51dB. See Figure 4.11-4.20.

It is observed that the value of $w(n) =$ in the active voice area is always larger than the inactive voice area. In the absence of noise, the inactive voice area (s) is zero and the active voice area can be easily found by checking whether $w(n) =>$ 0. However, if the voice is contaminated with noise (s), the adaptive thunderstorm increases regularly.

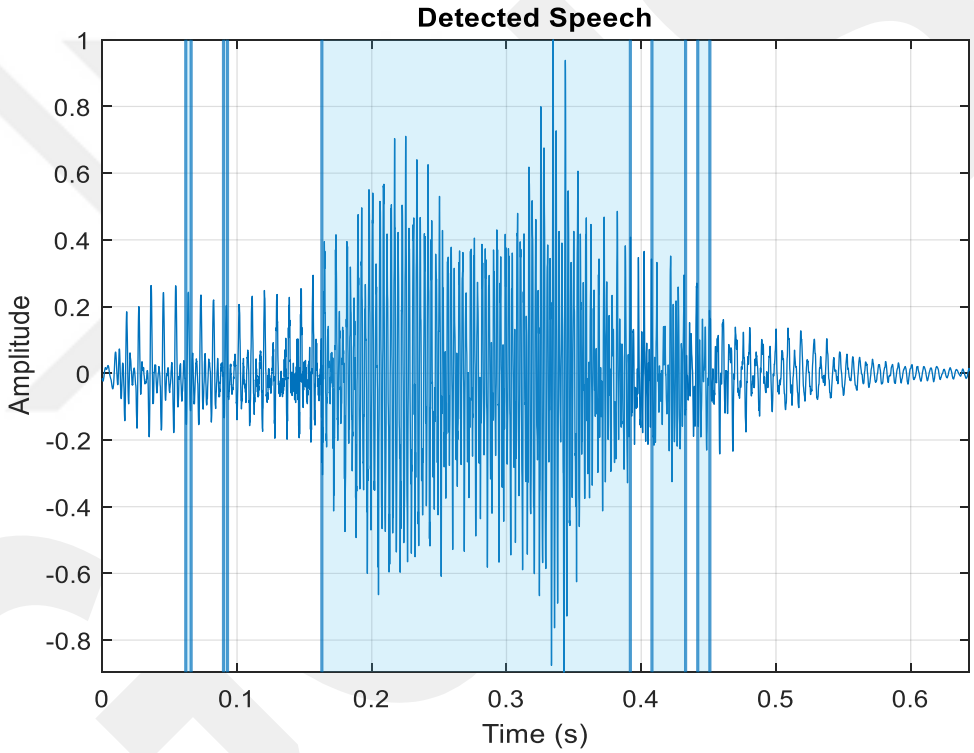

Figure 4.11 VAD for 0 using Bohman

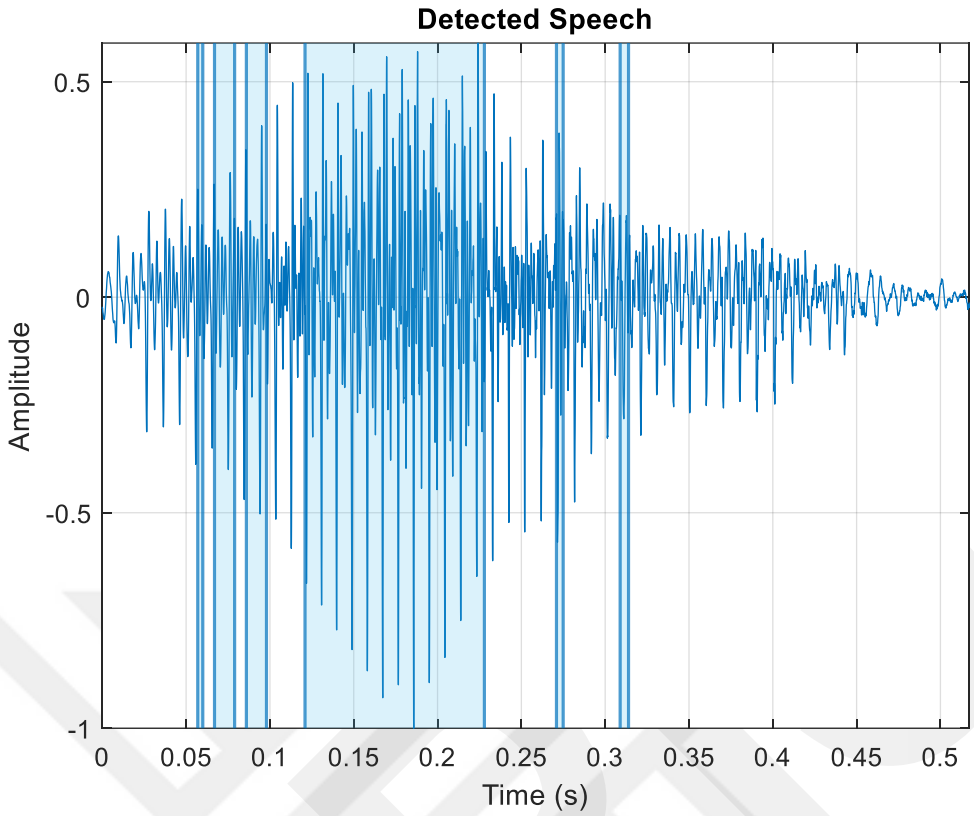

Figure 4.12 VAD for 1 using Bohman

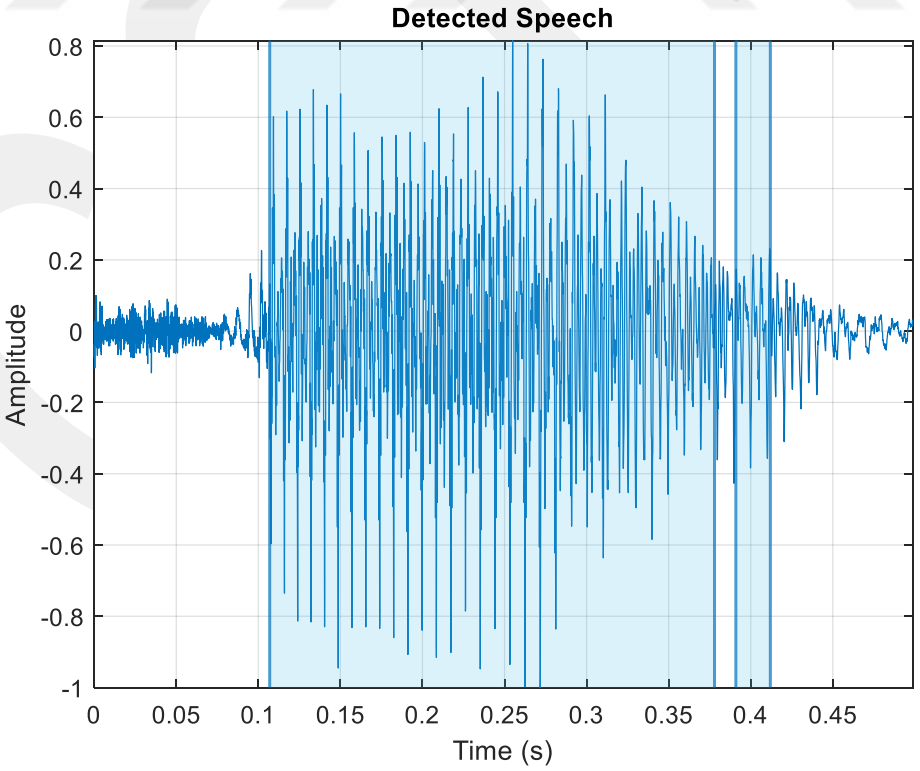

Figure 4.13 VAD for 2 using Bohman

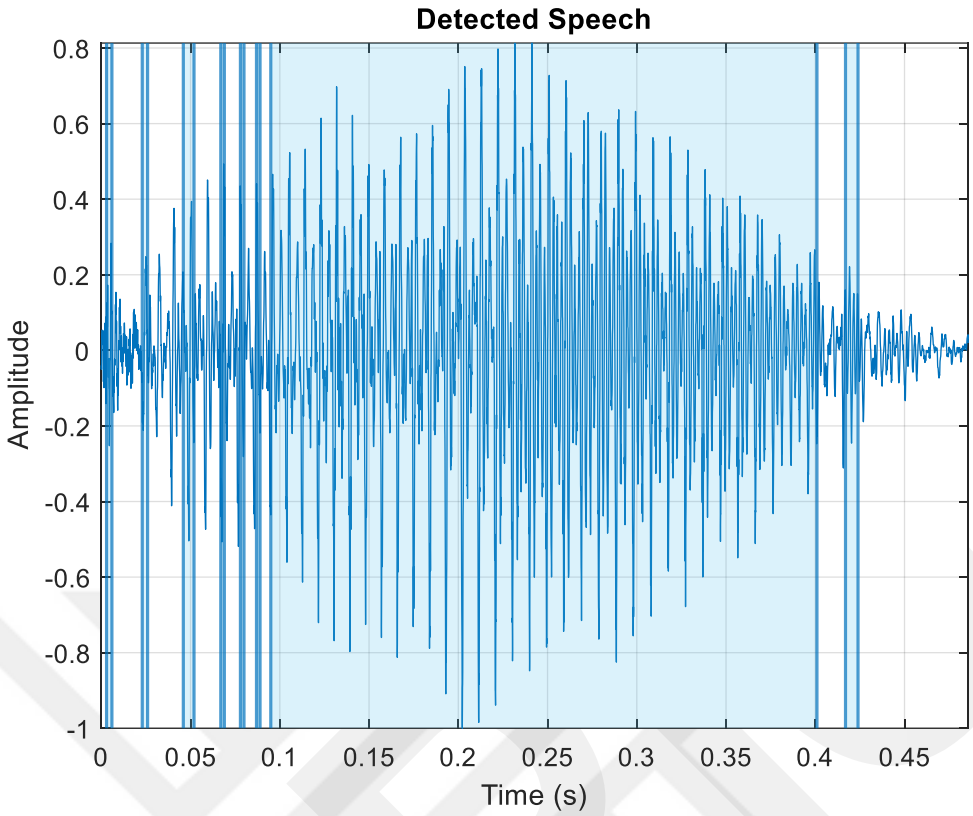

Figure 4.14 VAD for 3 using Bohman

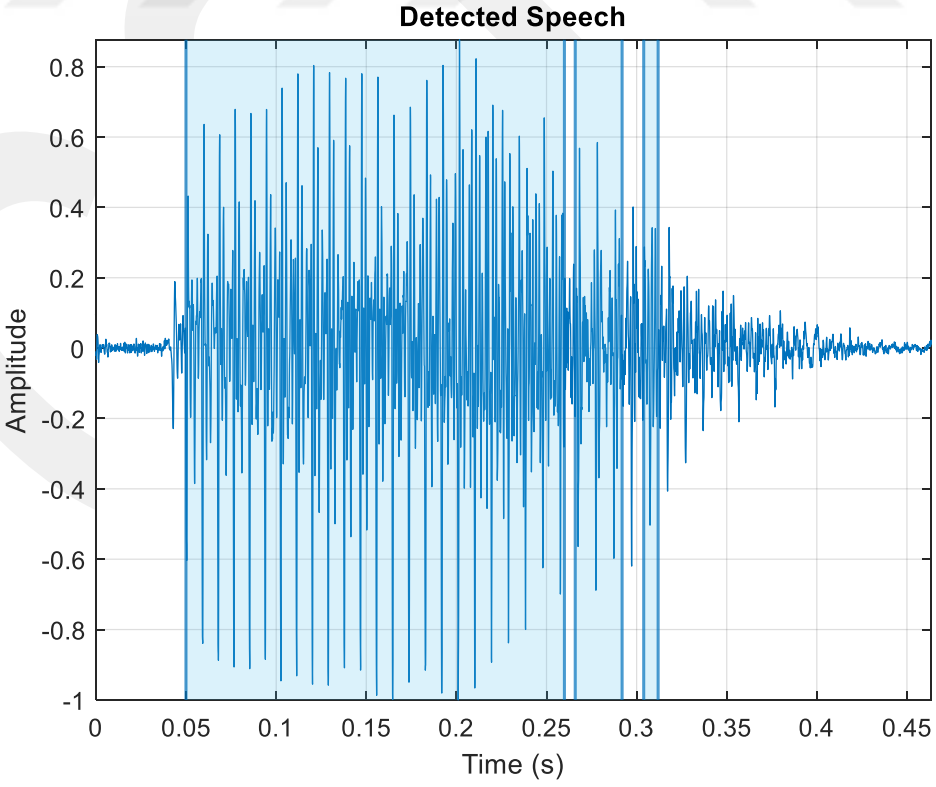

Figure 4.15 VAD for 4 using Bohman

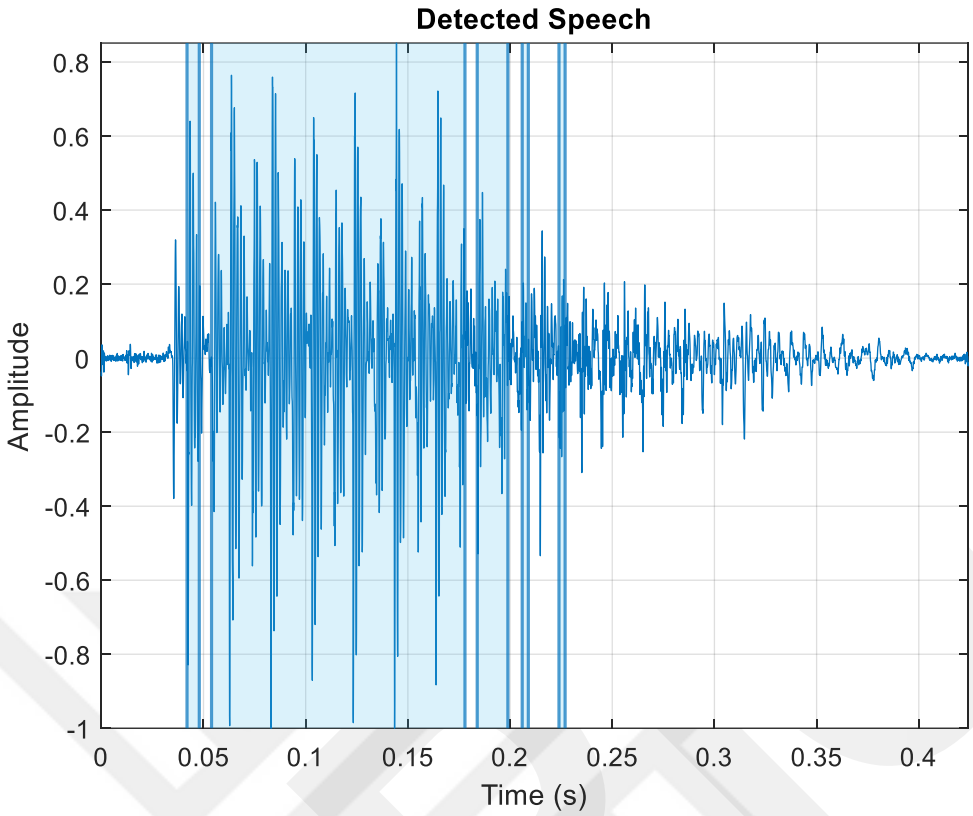

Figure 4.16 VAD for 5 using Bohman

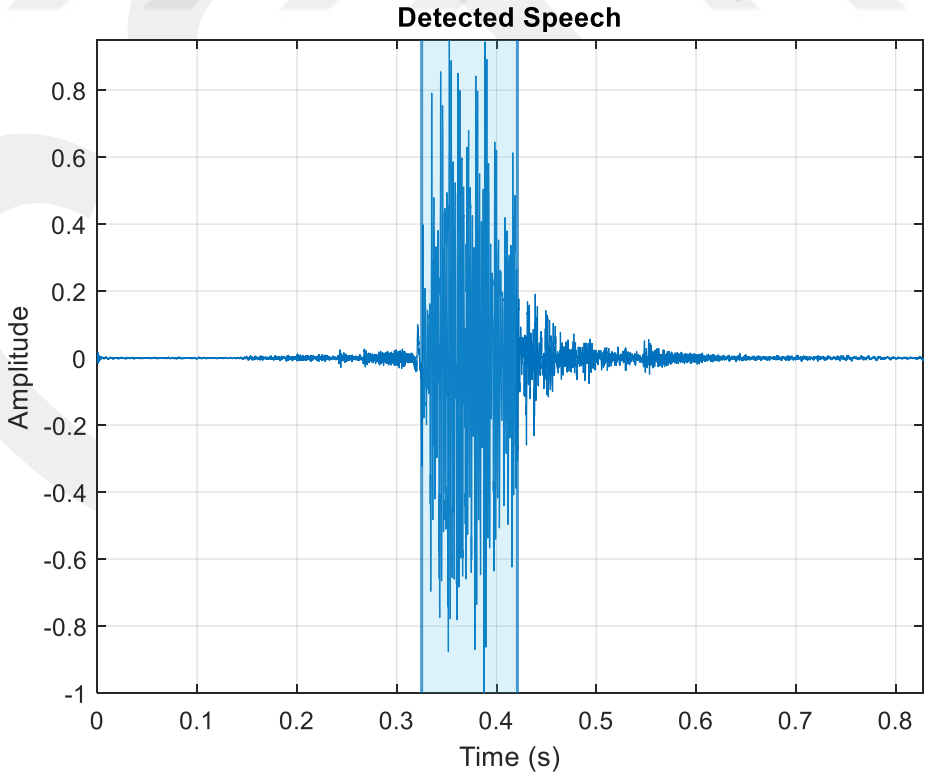

Figure 4.17 VAD for 6 using Bohman

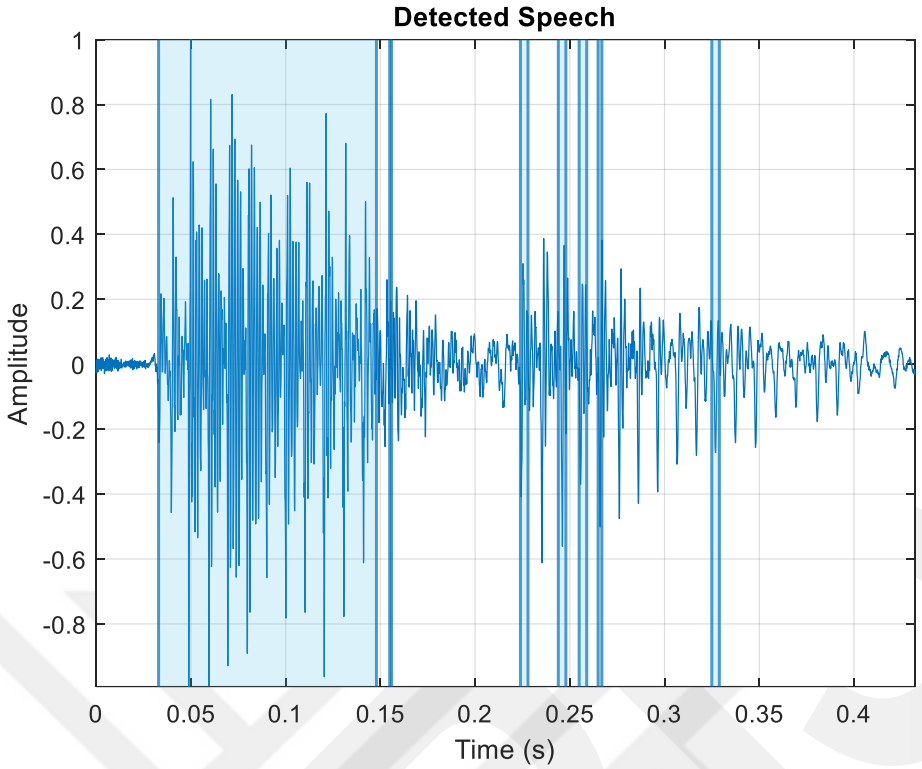

Figure 4.18 VAD for 7 using Bohman

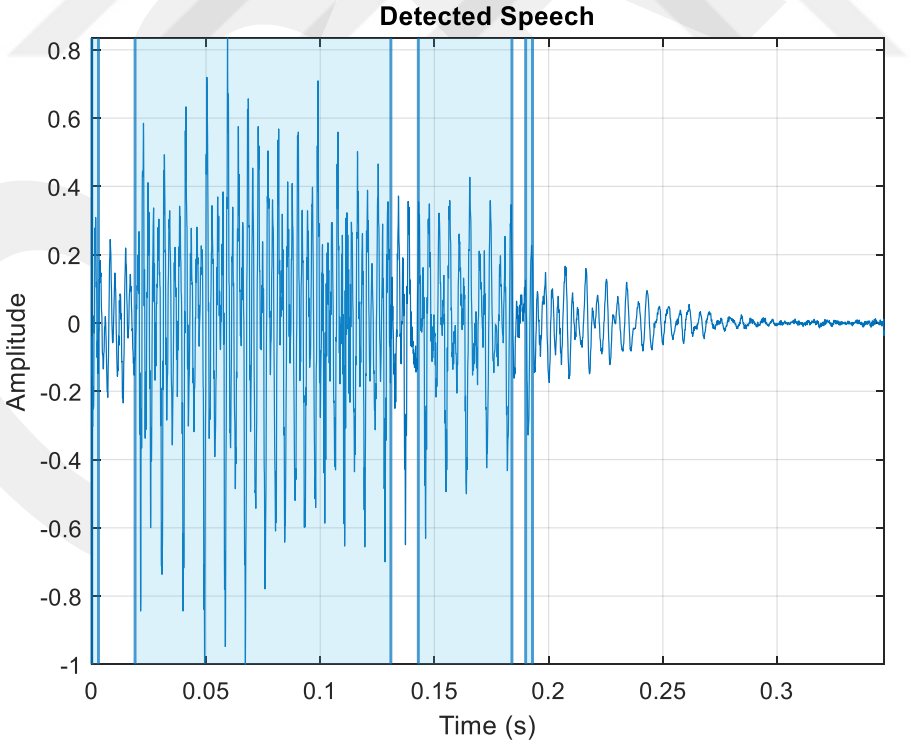

Figure 4.19 VAD for 8 using Bohman

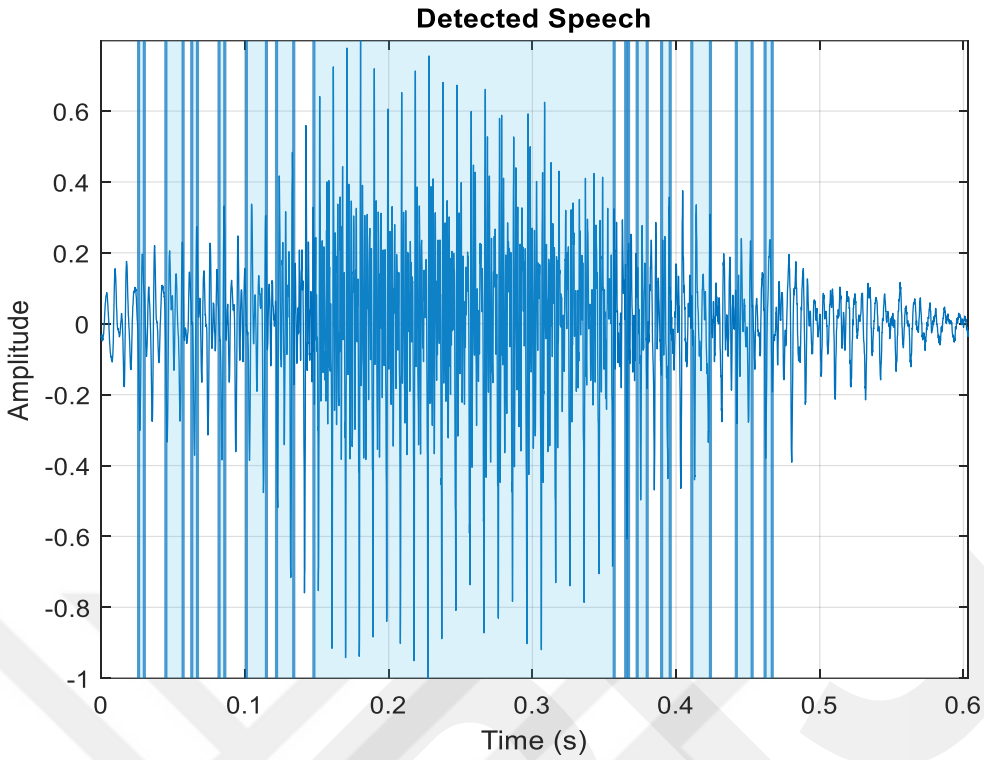

Figure 4.20 VAD for 9 using Bohman

4.1.3 Bartlett-Hann window

The Bartlett-Hann window was called for Julius von Hann, an Austrian Scientist. It is also recognized as the Cosine Bell. Some novelists favour that it be named a Hann window, to aid avoid confusion with the very like Hamming window. The mathematical model of hamming presented below in Equation (4.3):

$$\mathrm{w}(n) = \mathrm{a}_0 - (1 - \mathrm{a}_0).\cos\frac{2\pi n}{N}, 0 \leq \mathrm{n} \leq \mathrm{N}) \tag{4.3}$$

Then, the zero-phase form of this equation presented in Equation (4.4):

$$\mathrm{w}_0(n) = w\left[n + \frac{N}{2}\right] = a_0 + a_1 \cos\frac{2\pi n}{N}, \frac{n}{2} \leq n \leq \frac{n}{2} \tag{4.4}$$

Setting $a_0 = 0.5$ outcome a Hann window in Equation (4.5):

$$\mathrm{w}(n) = 0.5\left[1 - \cos(\frac{2\pi n}{N})\right] = \sin^2\left(\frac{\pi n}{N}\right) \tag{4.5}$$

After analysing of the mathematical model and the VAD results we can conclude the superior Related to $w_0(n)$ is one part higher cosine model. The model combine both sum of sine and sum of consine families. Reverse the Hamming the end points of the Hann window just touch zero. Furthermore, the Hamming window only uses the cosine function. See Figure 4.21- Figure 4.30.

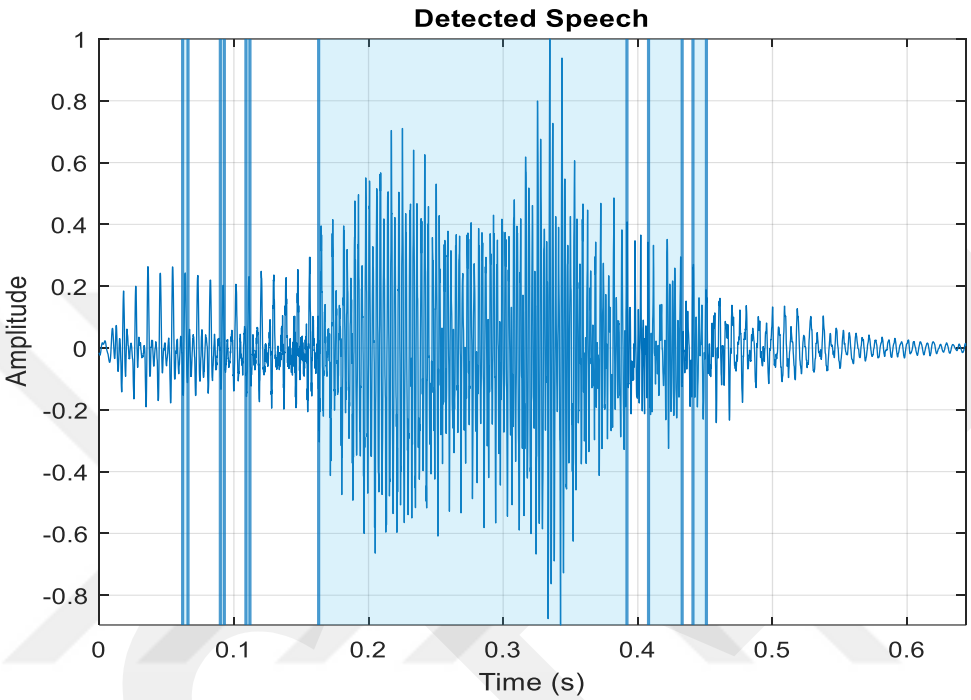

Figure 4.21 VAD for 0 using Bartlett-Hann

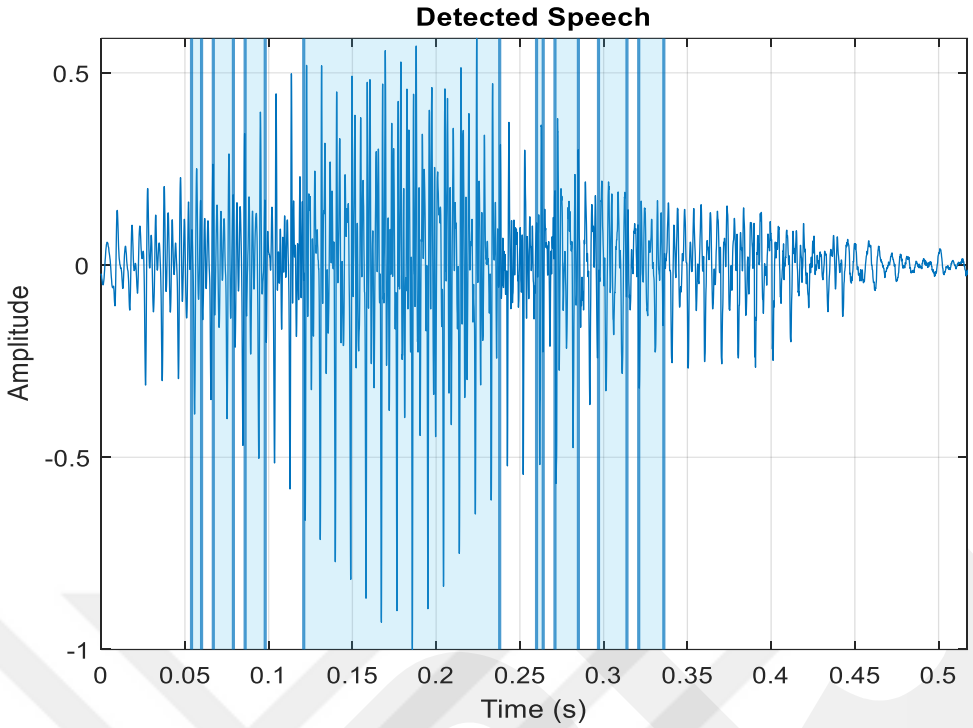

Figure 4.22 VAD for 1 using Bartlett-Hann

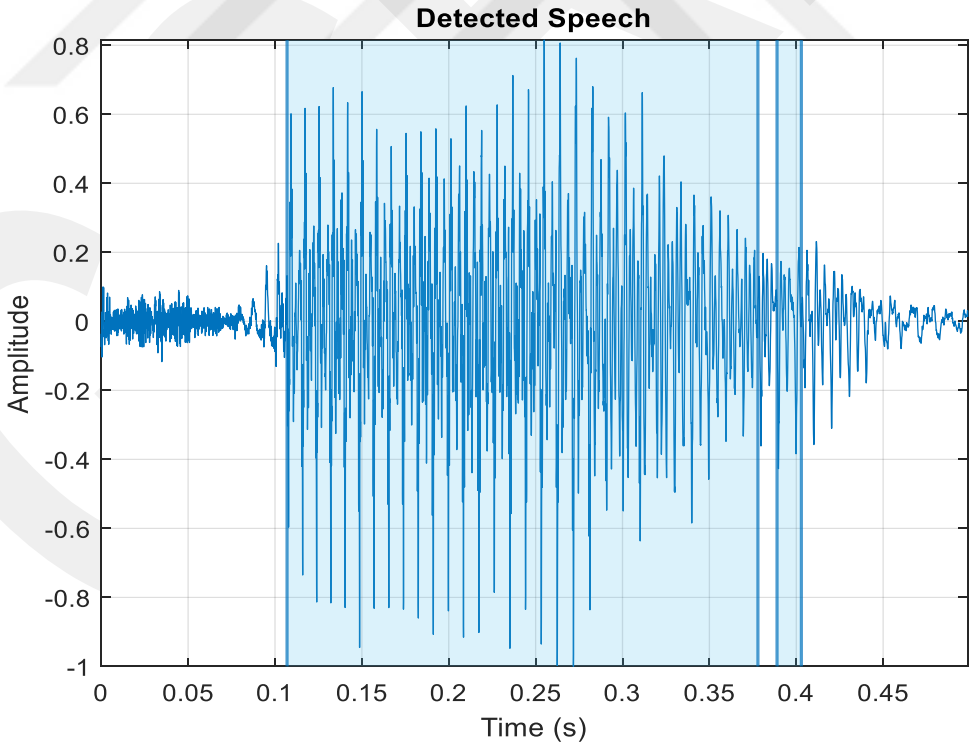

Figure 4.23 VAD for 2 using Bartlett-Hann

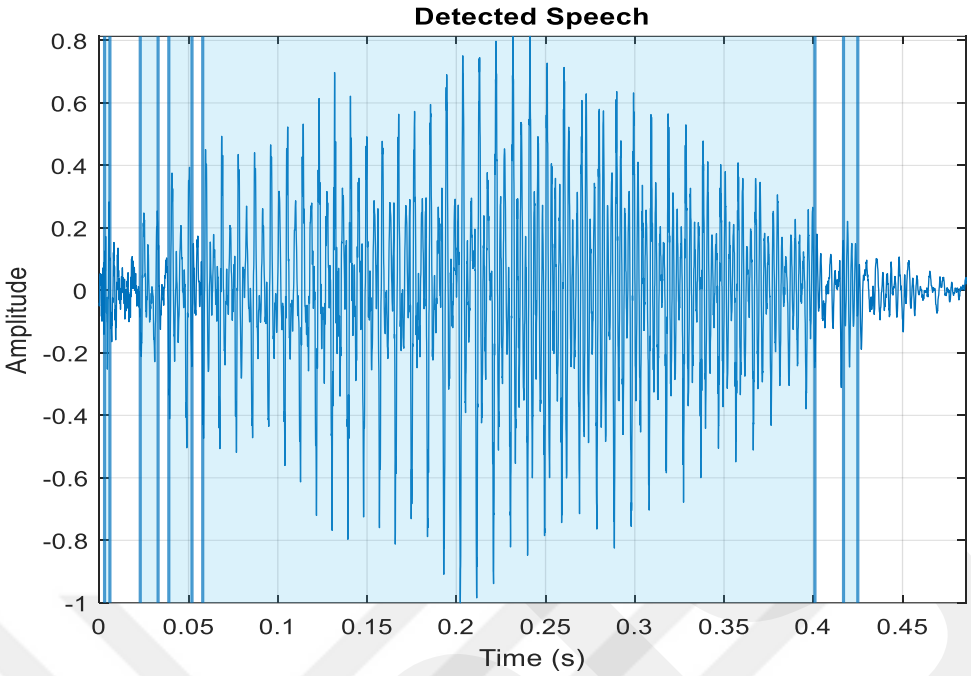

Figure 4.24 VAD for 3 using Bartlett-Hann

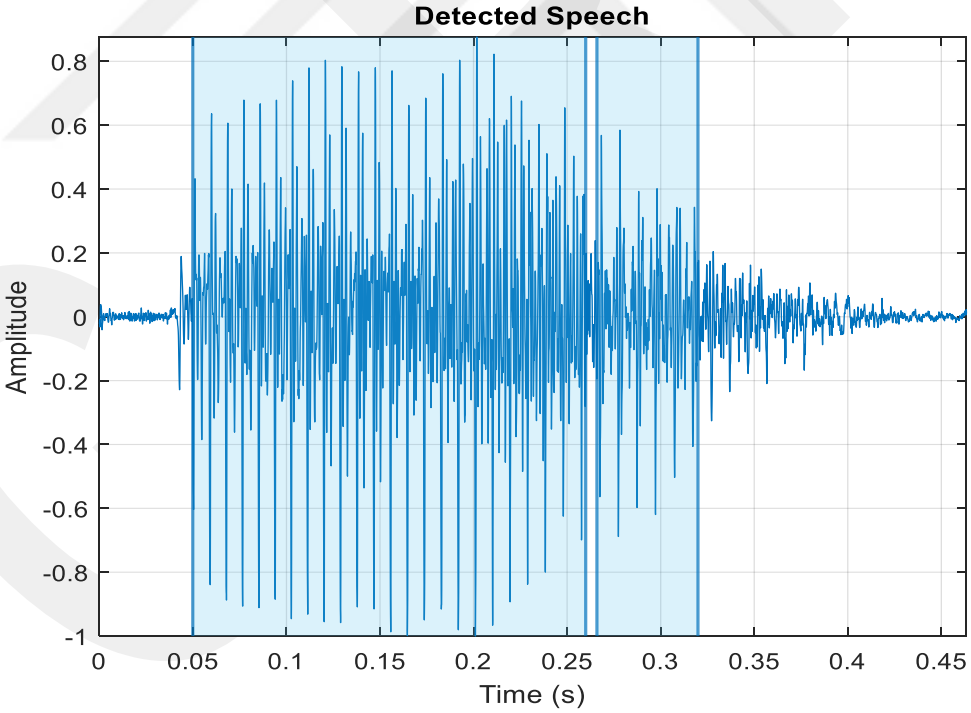

Figure 4.25 VAD for 4 using Bartlett-Hann

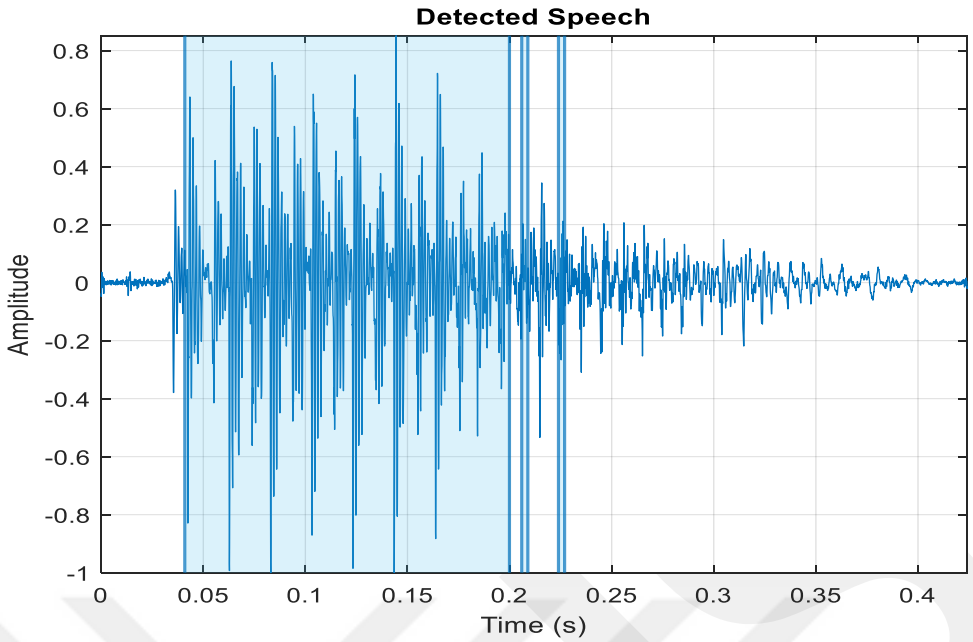

Figure 4.26 VAD for 5 using Bartlett-Hann

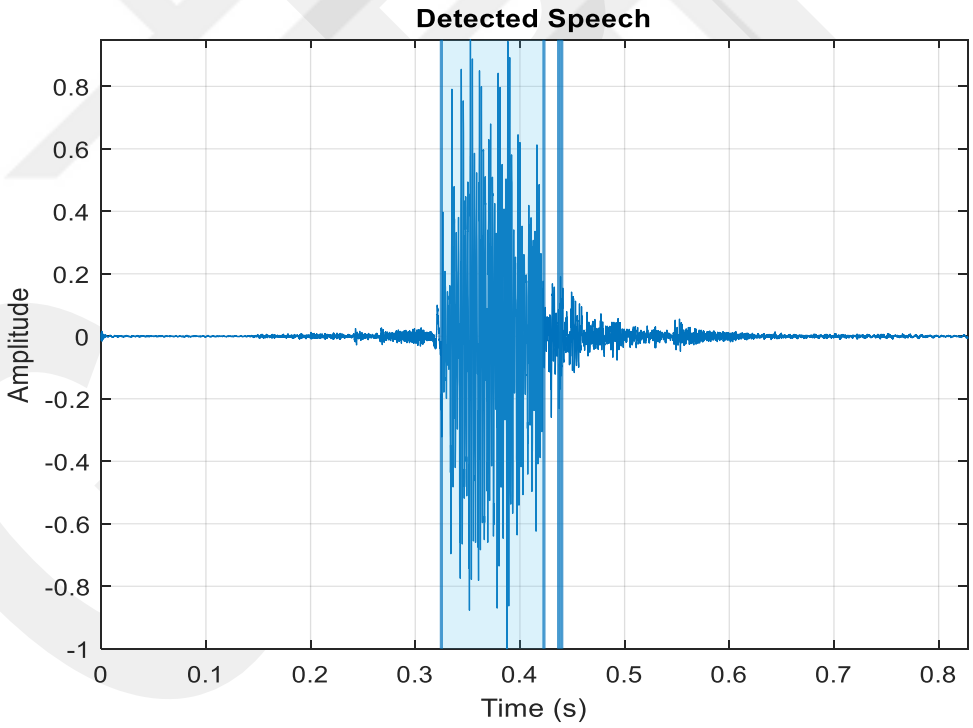

Figure 4.27 VAD for 6 using Bartlett-Hann

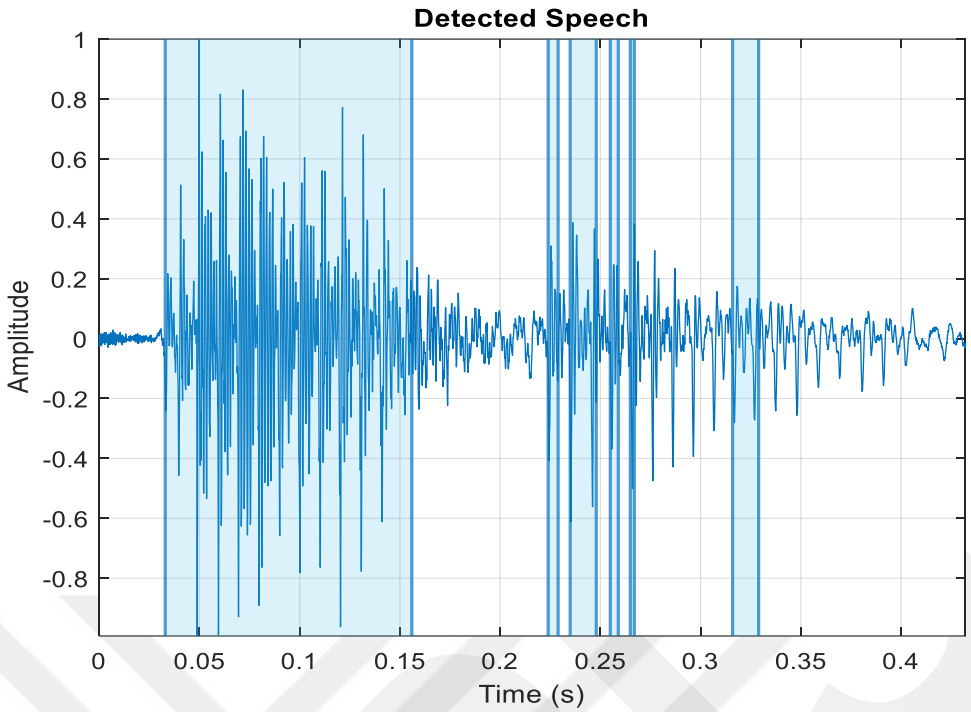

Figure 4.28 VAD for 7 using Bartlett-Hann

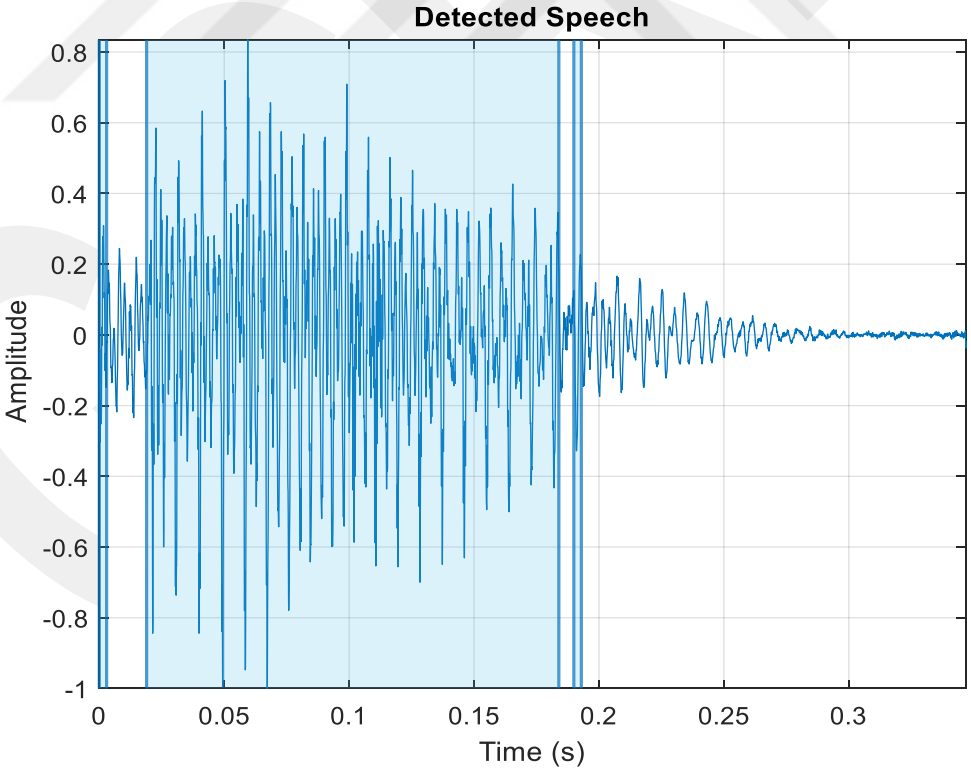

Figure 4.29 VAD for 8 using Bartlett-Hann

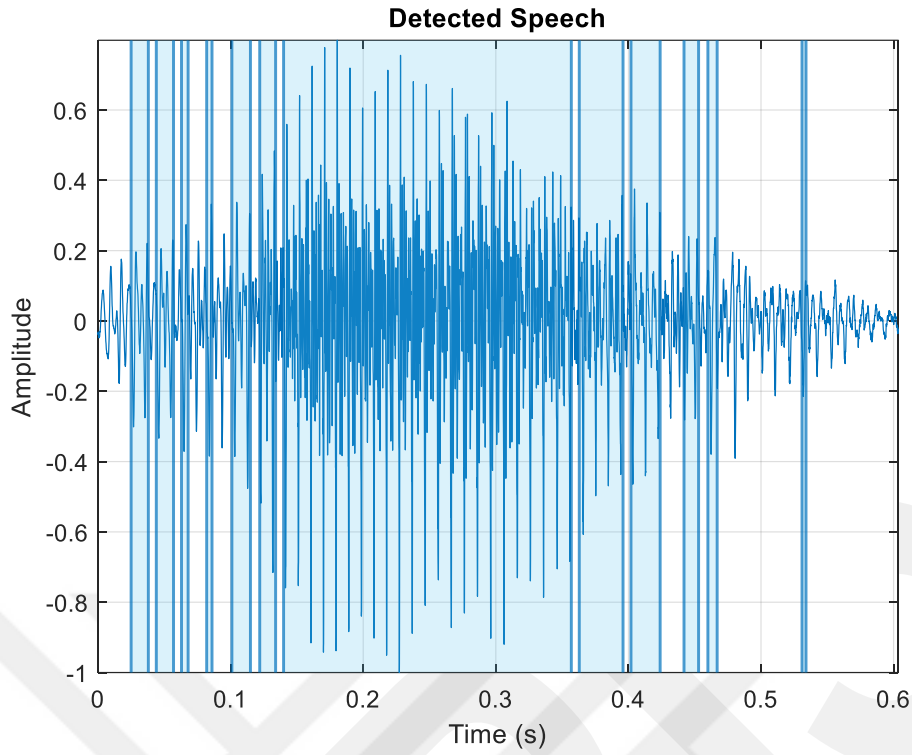

Figure 4.30 VAD for 9 using Bartlett-Hann

4.2. Feature Extraction Using Pitch

The presented sensitive features compared with other feature extraction techniques. The pitch parameters really presented different parameters from case to other. For example there is really great difference between pitch of zero and one or two with three. Besides, the mathematical model of the pitch is very simple and executed in low time which we can gain in the execution time compared to MFCC and other techniques. Furthermore, the pitch produce linear one dimension array which is the extracted features from complex data (audio). See the pitch Figures from Figure 4.31-Figure 3.37.

Then, the mathematical model of the pitch presented as shown below Equation (4.6) and Equation (4.7):

$$Max\ number\ of\ semitones\ lowered: -12 * log2(numel(Window) - OverlapLength)$$

(4.6)

$$Max\ number\ of\ semitones\ raised: -12 * log2\left(numel(Window) - \frac{OverlapLength}{numel(window)}\right)$$

(4.7)

In the first iteration, $EnvX_a$ is fitted to X. The Pitch technique eiterate these two procedure in a loop:

- To obtain new predicate lowpass filter applied as cepstral representation to the EnvX_a.
- To update the existing finest fitting, the algorithm obtain the element-by-element extreme of the present spectral cover predicate and the earlier spectral cover predicate:

$$\mathrm{EnvX}_a = \max \mathrm{EnvX}_a, \mathrm{EnvX}_b$$

(4.8)

- The loop finishes if whichever a supreme sum of iterations (100) is grasped, or if entirely baskets of the predictable log wrapper are inside a assumed broadmindedness of the unique log spectrum. The broadmindedness is set to $\log(10^{1/20})$.
- Then, the process scalars the spectrum of the pitch-shifted audio by the relation of predicated wrappers, element-wise:

$$\mathrm{Y} = Y * (\frac{\mathrm{EnvX}_b}{\mathrm{EnvY}_b})$$

(4.9)

1. **Result**: Fundamental frequency (F0)
2. Input: Audio signal *x* and sampling frequency *fs*.
3. Find the pitch of the audio signal

4. Pitch will give an estimate of the fundamental frequency of the signal *x*.

Figure 4.31 Pitch algorithm

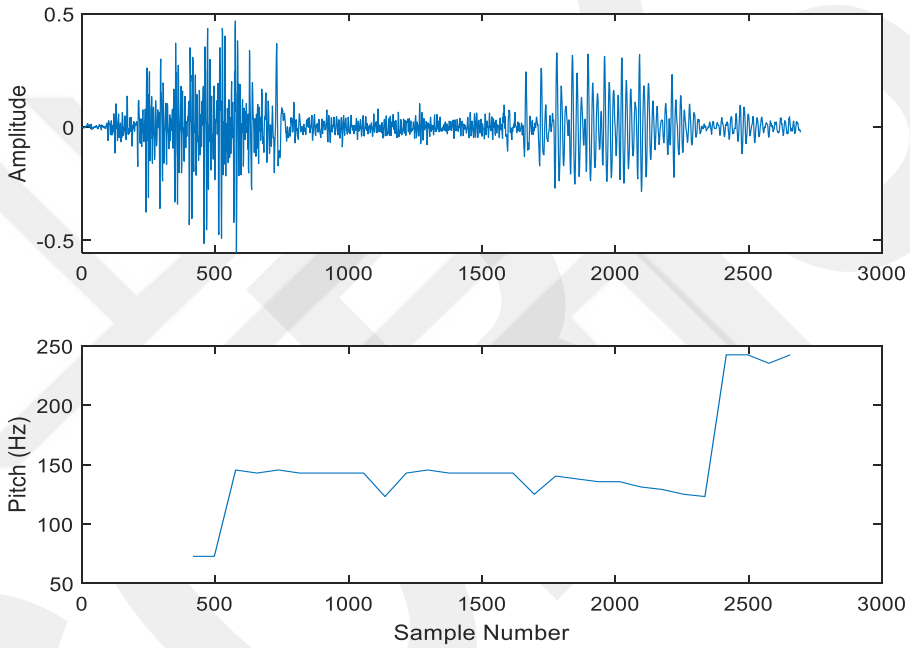

Figure 4.32 Feature extraction for zero using pitch

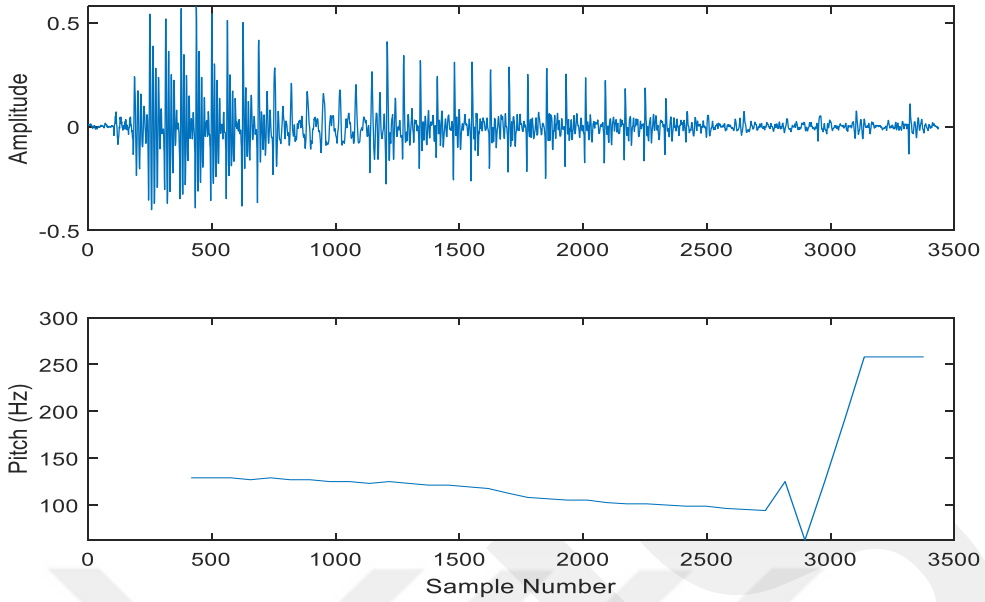

Figure 4.33 Feature extraction for one using pitch

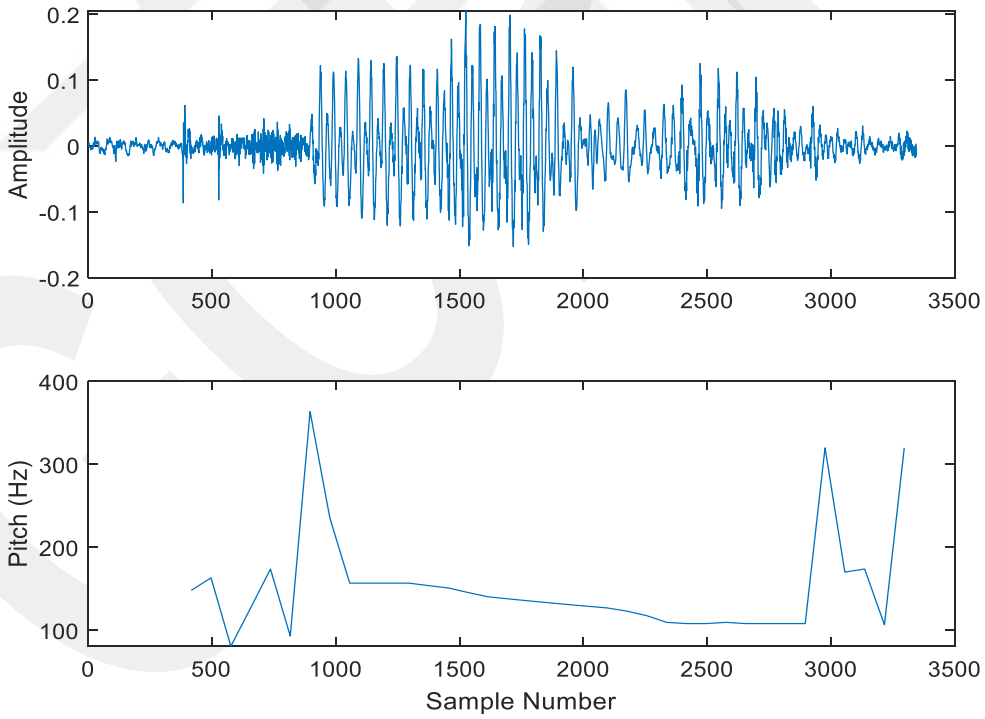

Figure 4.34 Feature extraction for two using pitch

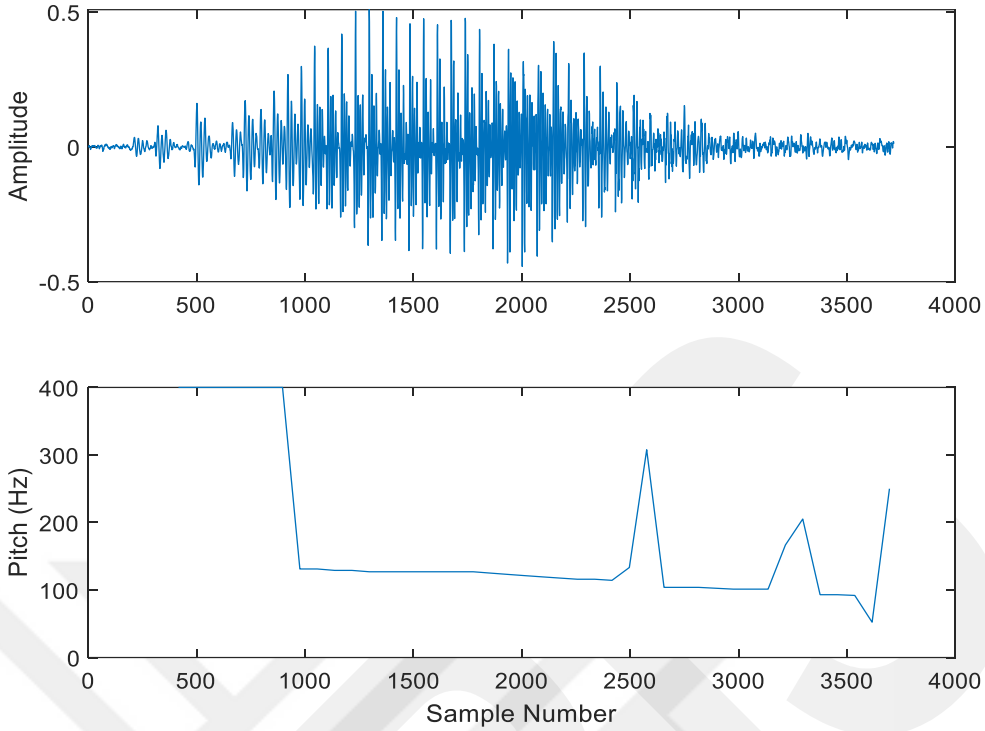

Figure 4.35 Feature extraction for three using pitch

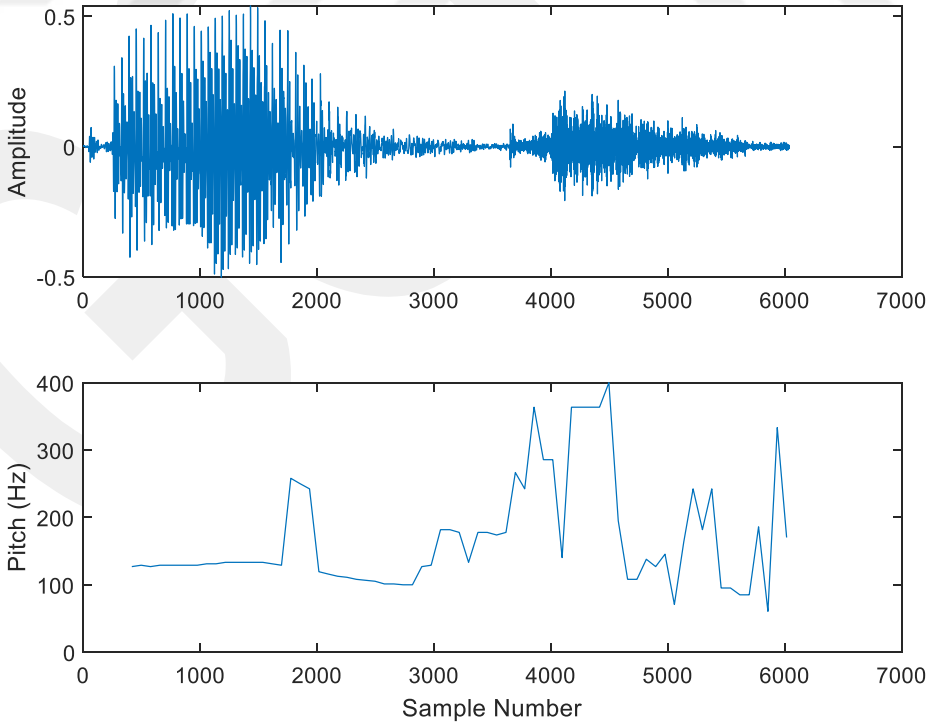

Figure 4.36 Feature extraction for four using pitch

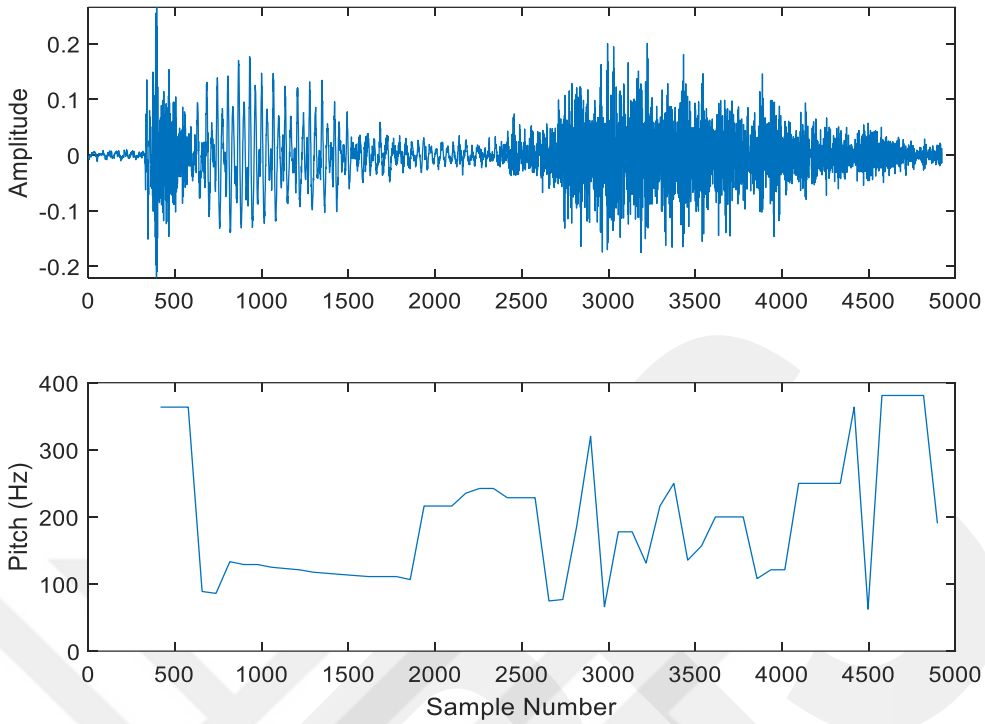

Figure 4.37 Feature extraction for five using pitch

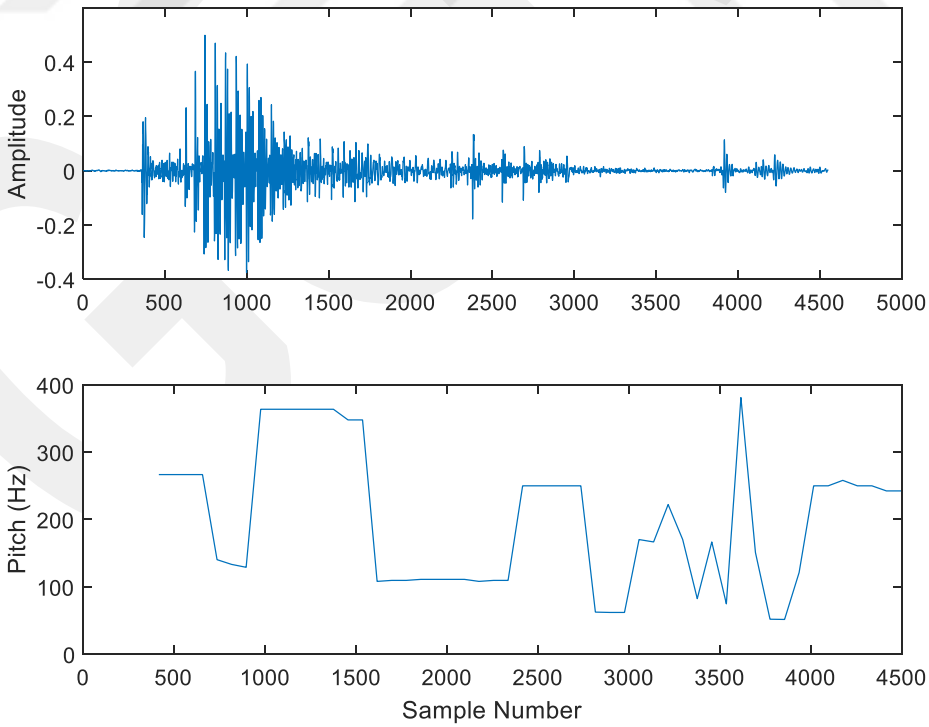

Figure 4.38 Feature extraction for six using pitch

4.3 Feature Extraction Using MFCCS

As shown in the script the MFCC calculated by using the windowing period 1024, this value differs from data to another. On the other hand, the overlapped length is fitted to 512, the overlap length is the critical section of implementing the MFCCs which is affected the output size of the function and the execution time of the script. Furthermore, the coefficients that's presented in the Figure 4.27 called as multidimensional features (coefficients).

The computation of MFCC feature can be explained as follows:

- Specify the window size and skip rate to process the speech file (usually 25 ms window size with 10 ms skip rate).
- For every frame:
  1) Perform data extraction, optimal dithering, pre-emphasis and dc offset removal, and multiply it by a windowing function (Hamming windowing is preferred);
  2) Calculate the energy at this point (log energy was used);
  3) Perform FFT and compute the power spectrum;
  4) Compute the energy in each Mel bin;
  5) Compute the log of the energies and take the cosine transform, keep a specified number of coefficients (usually 13 coefficients are kept)
  6) Optionally do cepstral liftering; this is just a scaling of the coefficients, which ensures they have a reasonable range. For this thesis, 13 Mel-frequency cepstral coefficients are computed for each respective frame. Then 13 delta and 13 delta-delta coefficients are appended to the feature vector resulting in an observation vector of 39-dimension. Details of the MFCC feature extraction process is illustrated in Table 4.1.

Table 4.1 MFCC Parameters

	Parameter	Value
1	Window length	25
2	Window overlap	10
3	Cepstral coefficients	12
4	Delta cepstral coefficients	12
5	Double Delta cepstral coefficients	12
6	Energy coefficients	1
7	Delta energy coefficients	1
8	Double delta energy coefficients	1

The mathematical model of MFCC can represented as shown in blow equations:

$$y_k = log(\Upsilon/x_{k-1}/^2 + /x_k/^2 + /x_{k+1}/^2 + £) \quad (4.10)$$

Where scalar $0<\Upsilon<1$, x_{k-1}, x_k, and x_{k+1} are coefficients very near to zero. Then, the power spectrum of signal can calculated as equation below:

$$\text{Power spectrum of signal} = /F^{-1}\{f\{x(t)\}\}/^2 \quad (4.11)$$

Which the Fourier transform represented by $F\{.\}$ And F^{-1} represented the inverse. A traditional estimate is to describe the frequency-to-mel convert function for a frequency f as:

$$m = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \quad (4.12)$$

Then, willingly derived the inverse transform as:

$$f = 700\ \left(10^{\frac{m}{2595}} - 1\right) \quad (4.13)$$

By using the previous Equation, we can find the frequency point of f_k by obtaining equally spaced point m_k.

Then, the weighting coefficients $w_{\mathrm{k,h}}$ are typically selected as triangular functions as:

$$w_{\mathrm{k,h}} f(x) = \begin{cases} \dfrac{h - f_{k-1}}{f_k - f_{k-1}} \; for \;\; f_{k-1} < h < f_k \\ \dfrac{f_{k+1} - h}{f_{k+1} - f_k} \; for \;\; f_k < h < f_{k+1} \end{cases}$$

$$0 \qquad otherwise$$

(4.14)

The main problem of the MFCCs is that presented multidimensional features which cannot applied easily and solved with classical machine learning techniques such as SVM, SoftMax and neural network easily. This need to create one-dimension features by converting two dimensions to one-dimension features and applied classical machine learning techniques for this problem.

We conclude that's techniques such as RNN and LSTM are only deep learning techniques which provided remarkable results and applied easily to time series data which size of each case differ from other and not affected the process of LSTM or RNN. Other machine learning and deep learning techniques require fixed size of features for all cases.

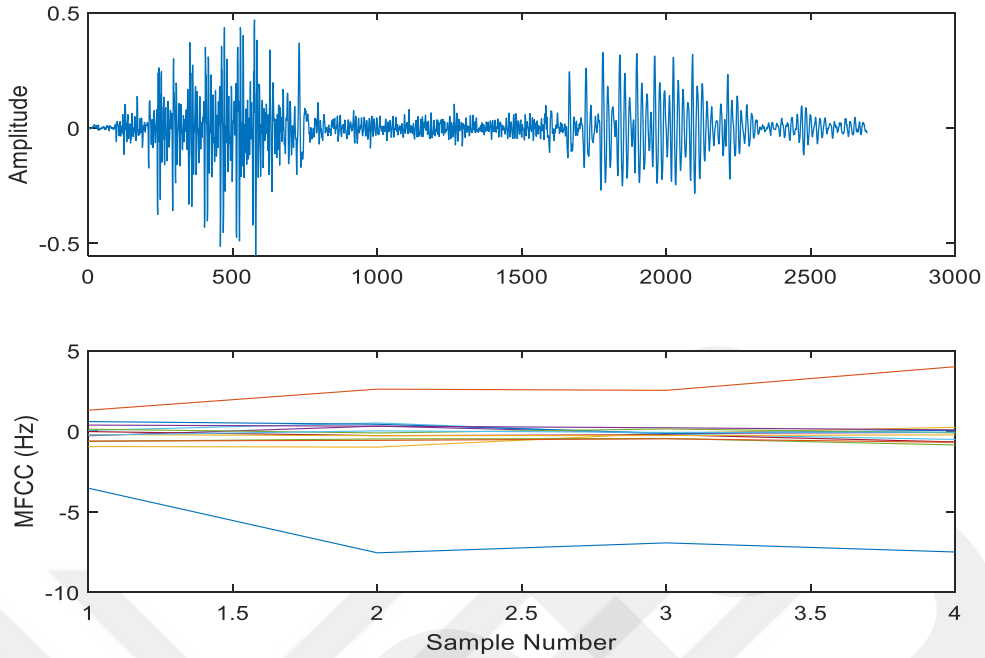

Figure 4.39 Feature extraction for one using MFCCs

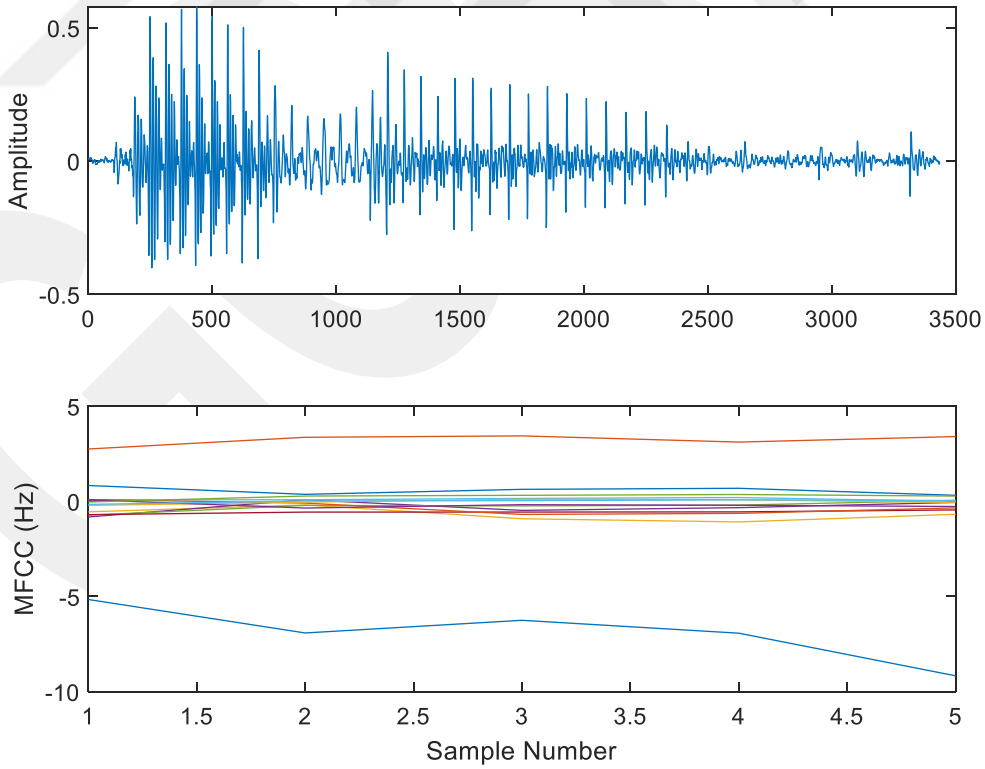

Figure 4.40 Feature extraction for two using MFCCs

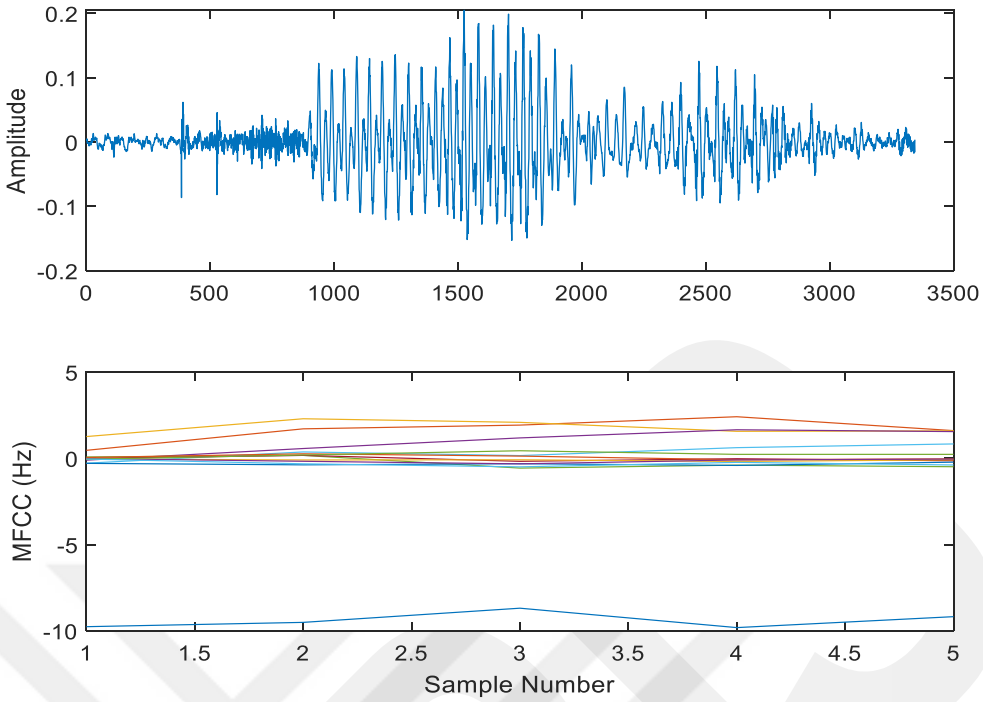

Figure 4.41 Feature extraction for three using MFCCs

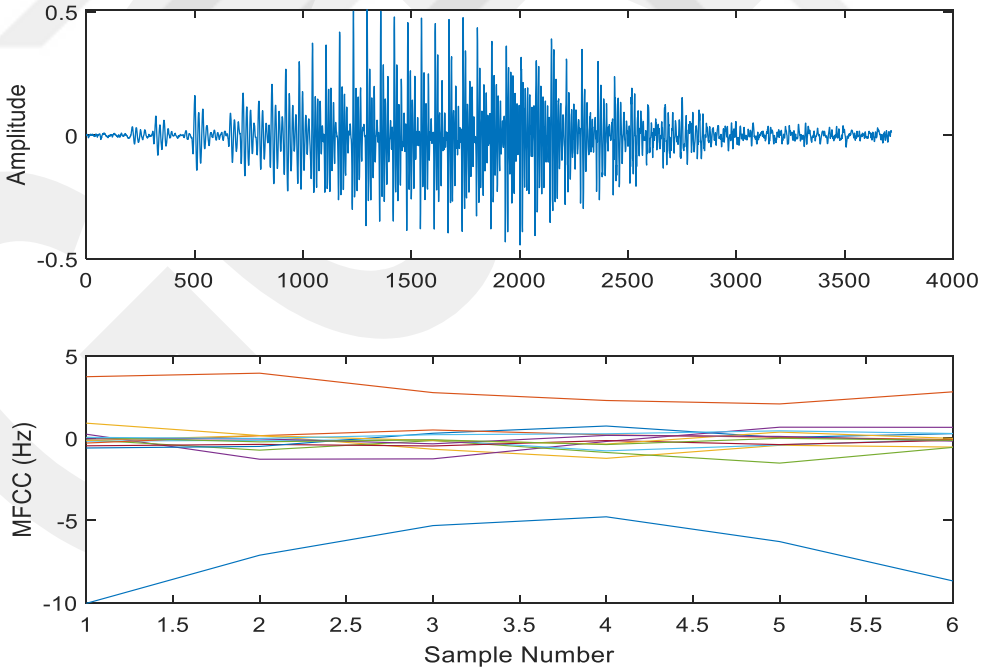

Figure 4.42 Feature extraction for four using MFCCs

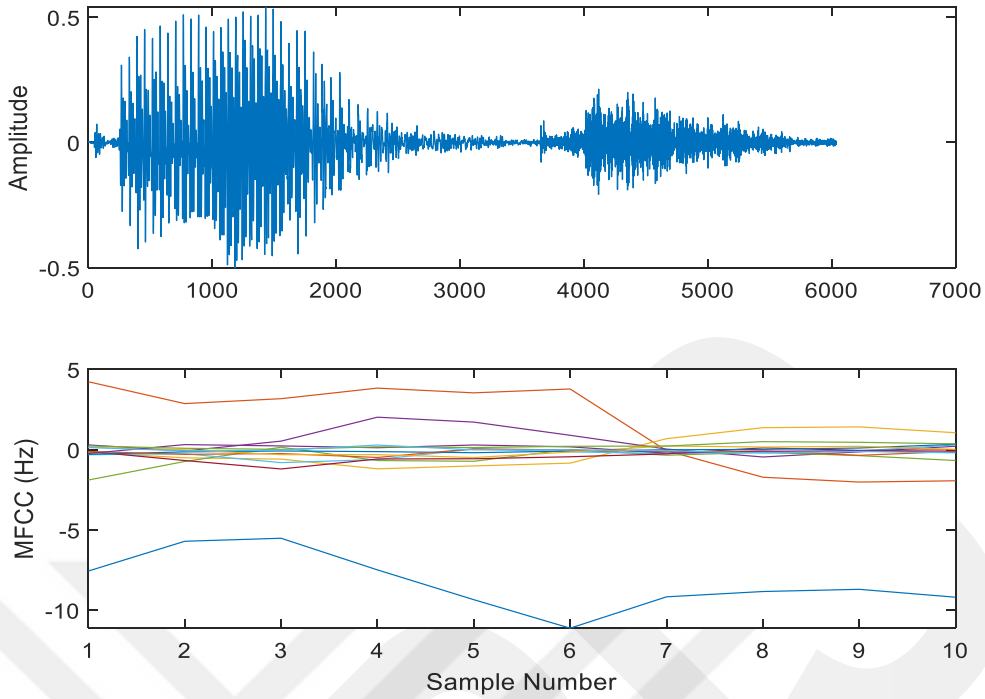

Figure 4.43 Feature extraction for five using MFCCs

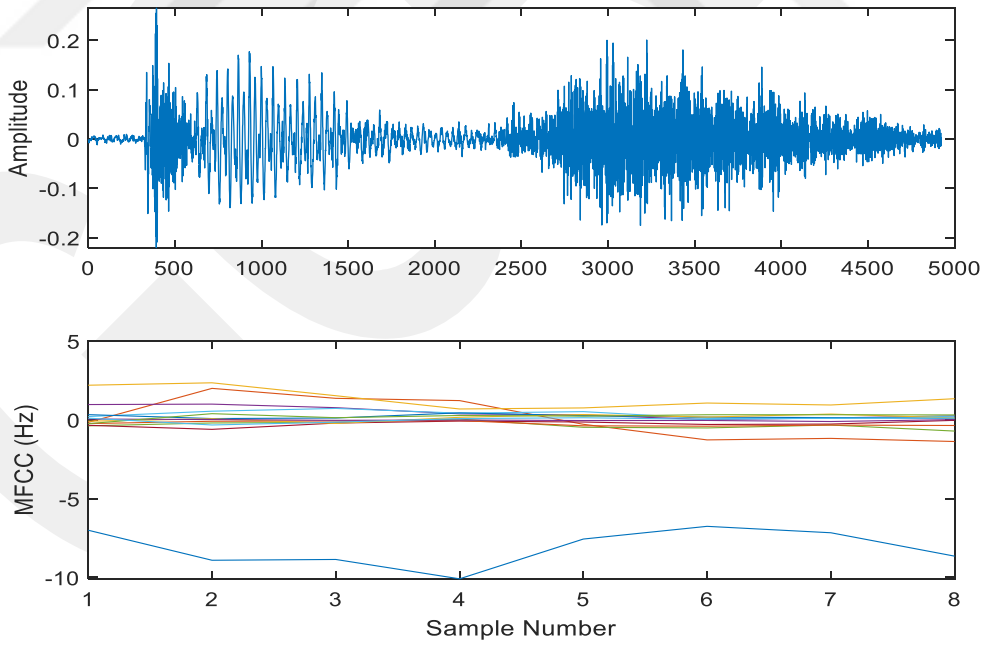

Figure 4.44 Feature extraction for six using MFCCs

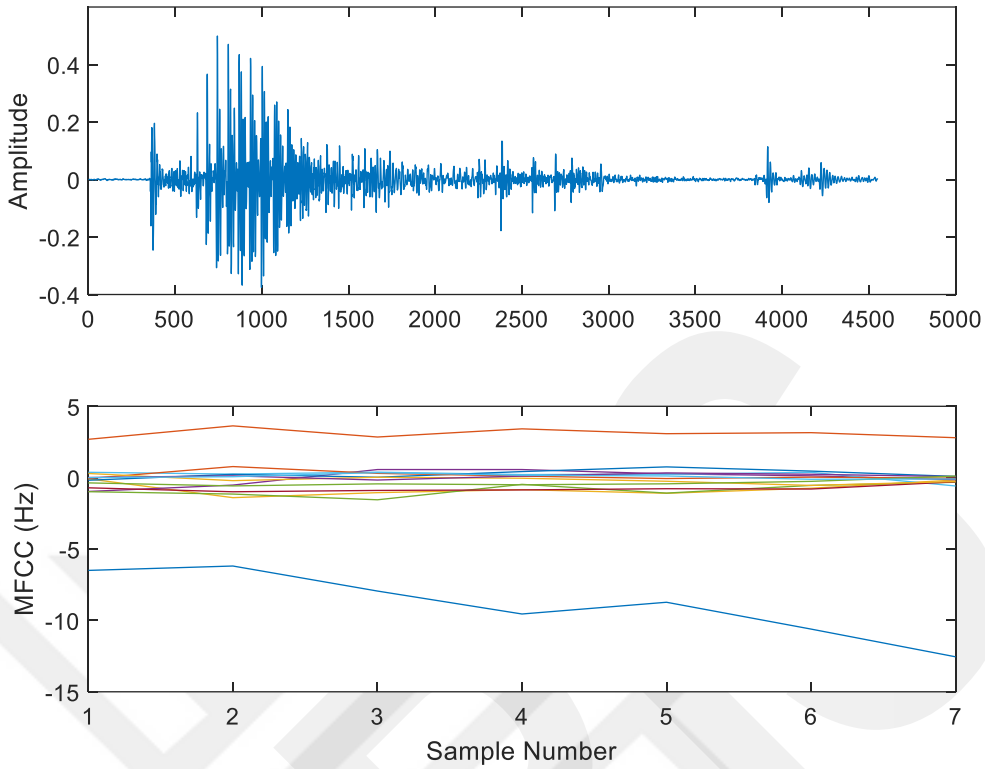

Figure 4.45 Feature extraction for seven using MFCCs

4.4 Feature Extraction Using Energy

In this section, energy of each signal is calculated the main idea of this method is calculated energy for each frame (epoch). Then, the data classified to the epochs and the energy of each epoch calculated we can visualize this idea in the Figure 4.45.

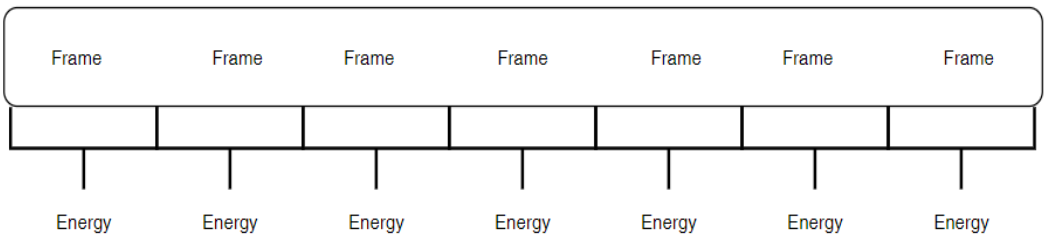

Figure 4.46 Energy Calculation

Mathematically, we can represented this step as shown in Equation 4.15.

$$E = \int_{-\infty}^{\infty} x(t)^2 \, dt \tag{4.15}$$

Where the ∞ and -∞ represented the upper and lower limit for the each frame, x(t) represented the audio data that it energy will be calculated. If the input signal in the frequency domain then the model represented as shown in the equation **(4.16)**

$$ESD = \int_{-\infty}^{\infty} x(f)^2 \, df \tag{4.16}$$

Where the ∞ and -∞ represented the upper and lower limit for each frame, $x(f)$ represented the audio data that its energy spectral density then calculated.

4.5 Discussion

In this section IWR results presented by using confusion matrix and roc curve for evaluate the results. In the first experiment pitch based SOFTMAX presented and the results show in Figure 4.47 and 4.48.

A confusion matrix is a summary of forecasting results for classification problems. The number of correct and false predictions is summarized in numbers and classified into classes. This is the core of the confusion matrix. The confusion matrix shows the degree of confusion in the classification of the predictive model. This not only provides information about the mistakes made, but also tells us about the most important types of mistakes. Several parameters are created the confusion matrix which are:

- TP: True Positive: estimated prices appropriately estimated as real positive
- FP: estimated prices falsely predicted a real positive.
- FN: False Negative: Positive values estimated as undesirable
- TN: True Negative: estimated values correctly estimated as an actual

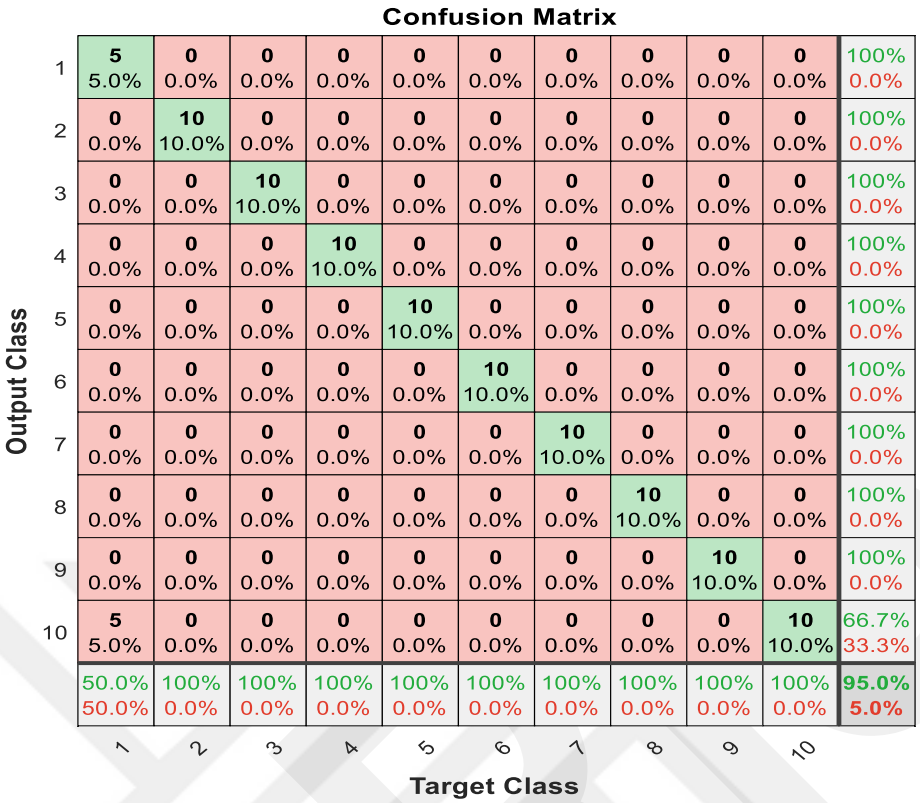

Figure 4.47 Pitch confusion matrix

The roc curve of the pitch based SOFTMAX also presented and shown in the Figure 4.48.

The ROC curve is mainly used to graphically represent clinical sensitivity and specificity relationship / compensation for testable limits or combinations. In addition, the area under the ROC curve shows the advantage of using the test provided.

The ROC curve is used in clinical biochemistry to select the most appropriate test limits. The best cut off is the highest true positive rate and the lowest false positive rate.

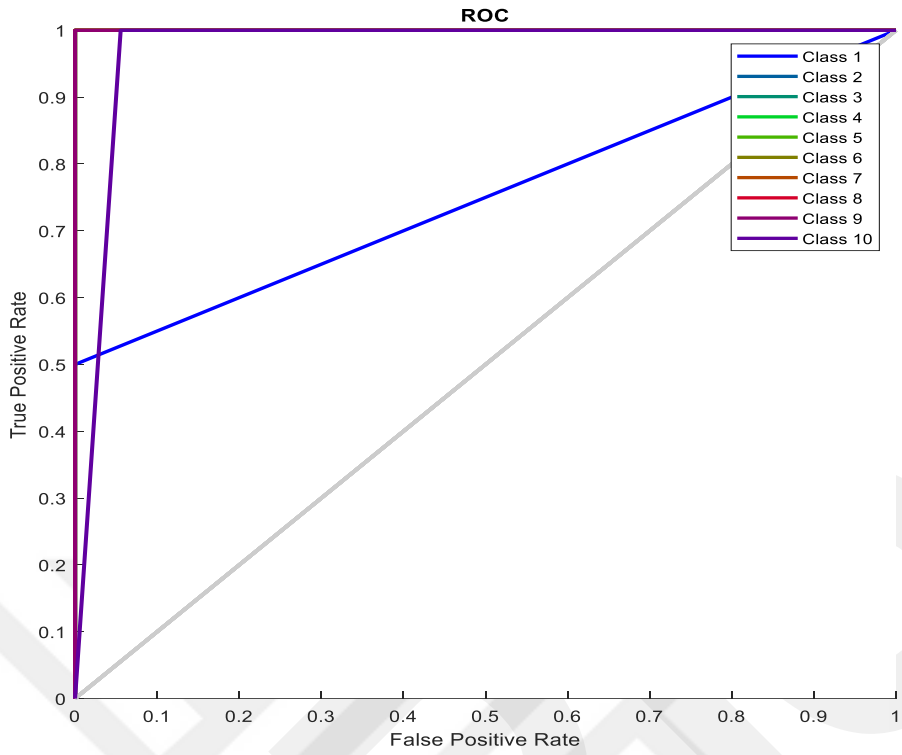

Figure 4.48 Pitch based SOFTMAX Roc curve

The pitch based SOFTMAX presented high results in fast execution time. The execution time is very low compared with other techniques because its mathematical model is very simple which presented and explained above. The second experiment combined the energy with SOFTMAX. The energy of each audio calculated then the energy signal wired to the SOFTMAX. The presented energy based SOFTMAX results presented in Figure 4.49 and 4.50.

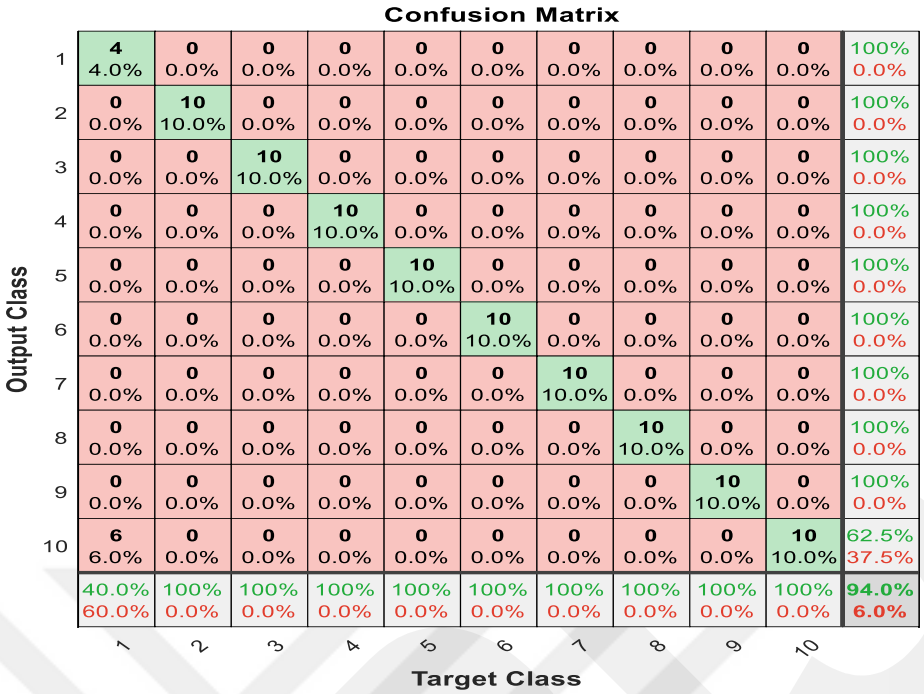

Figure 4.49 Energy based SOFTMAX confusion matrix

The roc curve of the energy based SOFTMAX presented in the Figure 4.50,

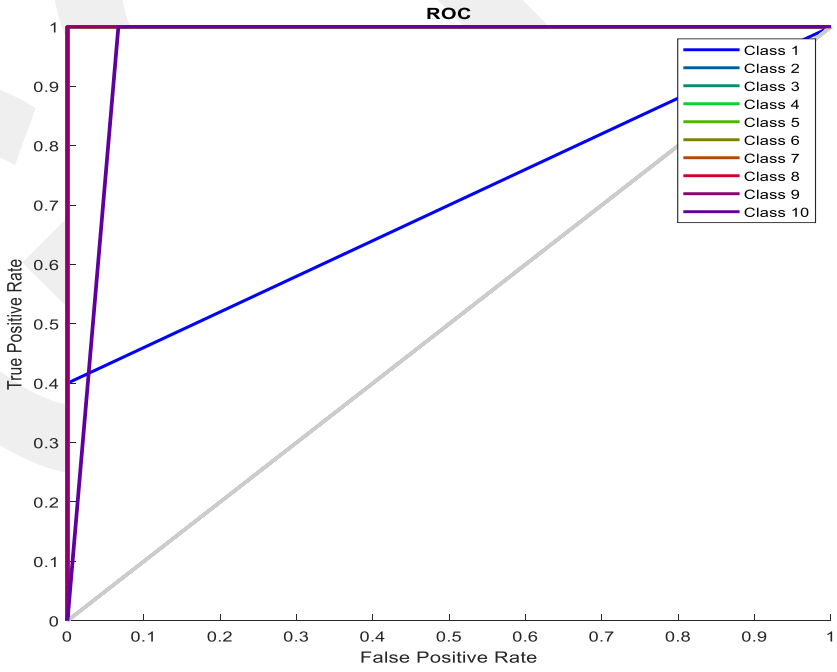

Figure 4.50 Energy based SOFTMAX Roc Curve

The MFCCs based SOFTMAX applied to the same digit audio dataset with 1024 windowing size. The main issue in MFCCs topic is the length overlap which we set two overlapped length: 1024 and 512. The results affected with the vary of overlap length which with overlap 1024 the model presented low results with low execution time and the reason of this because the window cover the whole signal with low number of iteration because the window in each iteration jump 1024 and the results of this step presented in 4.51 and 4.52.

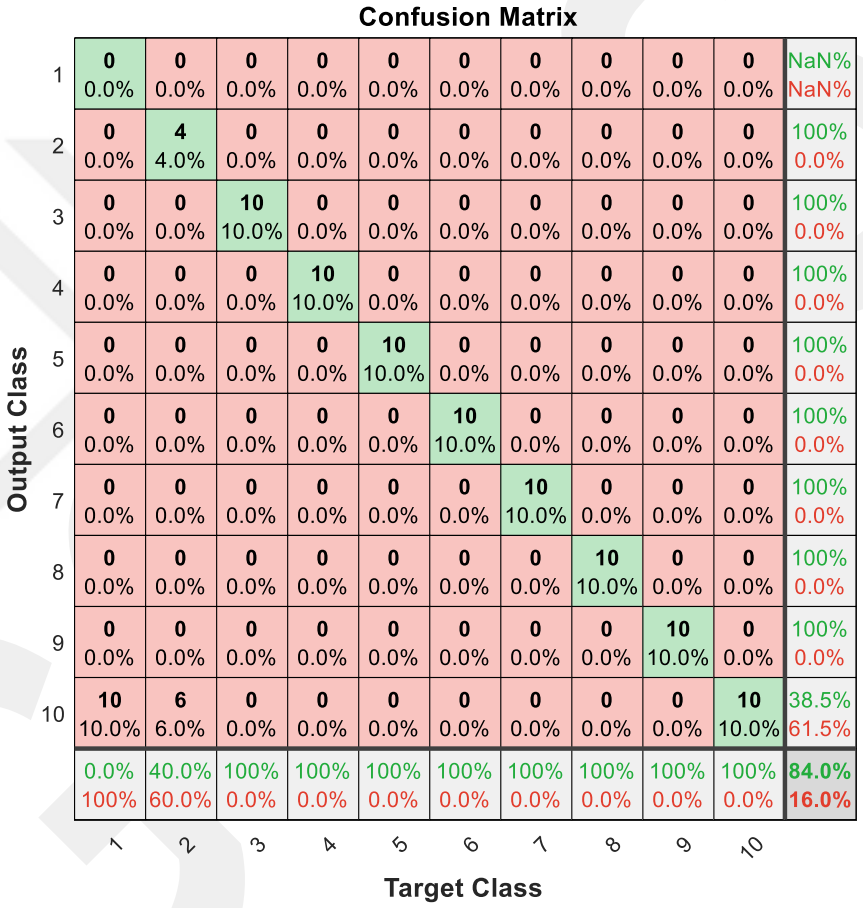

Figure 4.51 MFCCs based SOFTMAX confusion matrix with 1024 overlap length.

Furthermore, the roc curves of this case presented in the Figure 4.49.

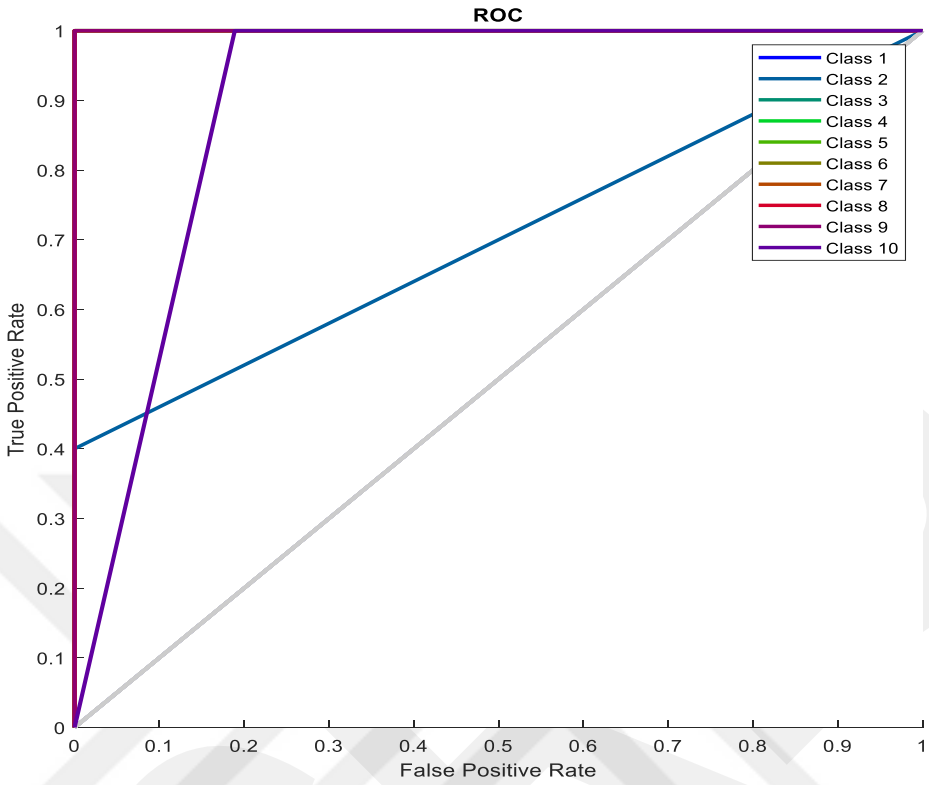

Figure 4.52 MFCCs based SOFTMAX Roc curve with 1024 overlap length

On the other hand, the 512 overlap length applied which lead to presented best results but with high execution time. The presented method results shown in the Figure 4.50 and 4.51.

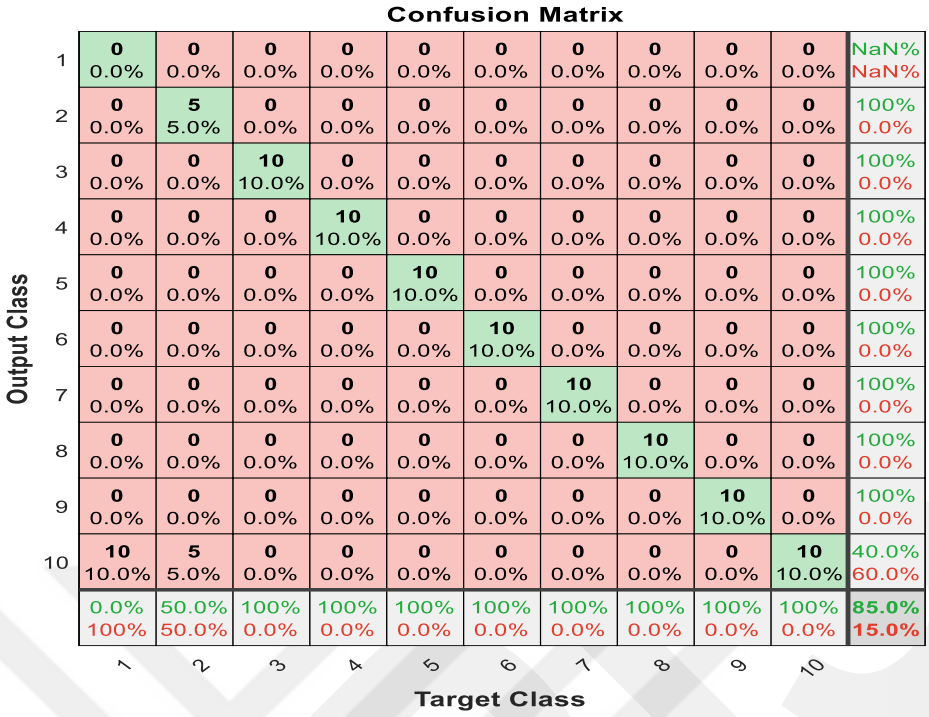

Figure 4.53 MFCCs based SOFTMAX confusion matrix with 512 overlap length

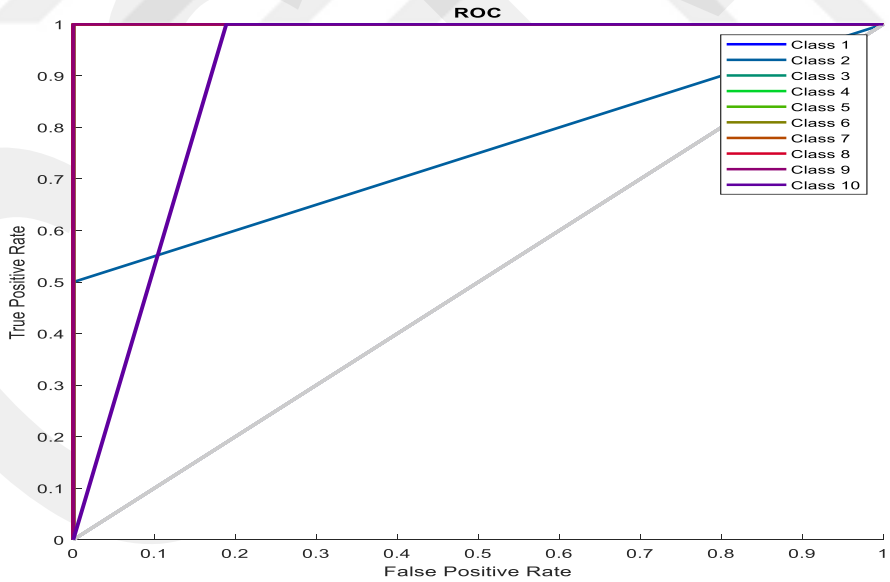

Figure 4.54 MFCCs based SOFTMAX Roc curve with 512 overlap length

Then, the results of the IWR are presented in the Table 4.2 to compare all the results.

On the other hand, 30 sample used for testing and 120 for training for each class. This lead to increase the difficulty of the estimation part. The reason is that the increase of the number of testing sample lead to reduce the training samples.

By combining pitch with SOFTMAX the model presented 93.7 % which the results visualized in the Figure 4.55 and 4.56.

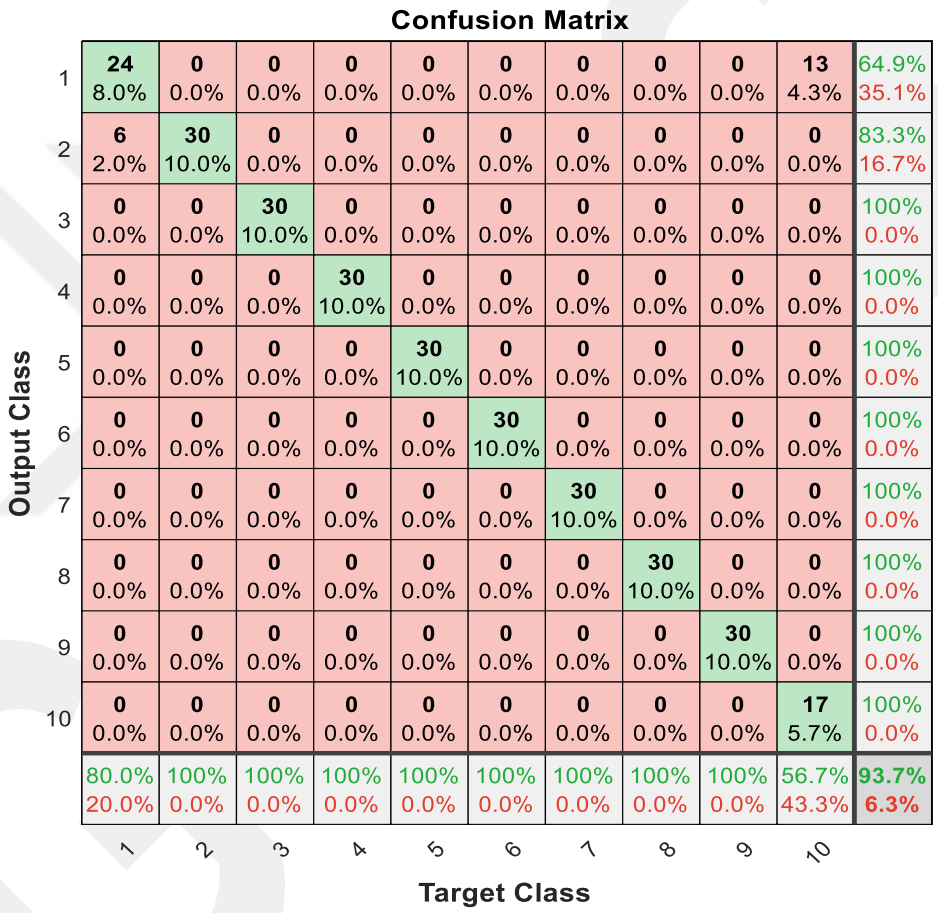

Figure 4.55 Pitch confusion matrix with 30 test sample

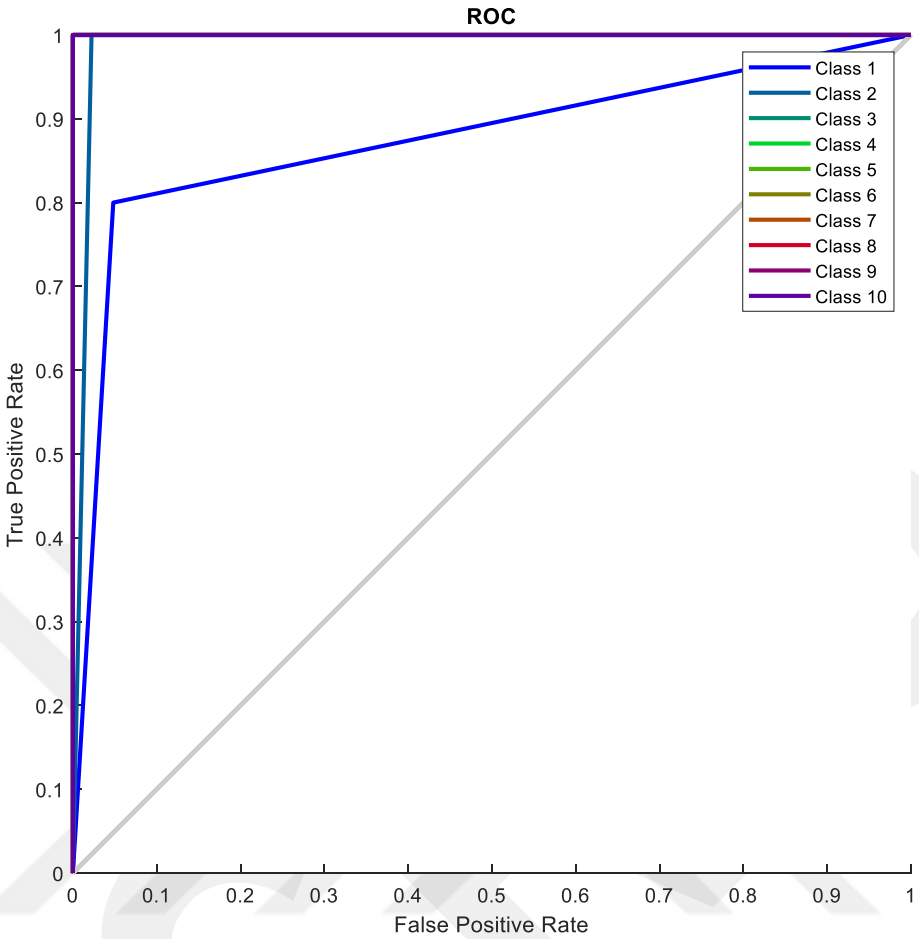

Figure 4.56 Pitch based SOFTMAX Roc curve with 30 test sample

Furthermore the energy based SOFTMAX also tested using 30 sample of testing data. The energy based SOFTMAX presented 92.7% accuracy and the results presented in the Figure 4.57 and Figure 4.58.

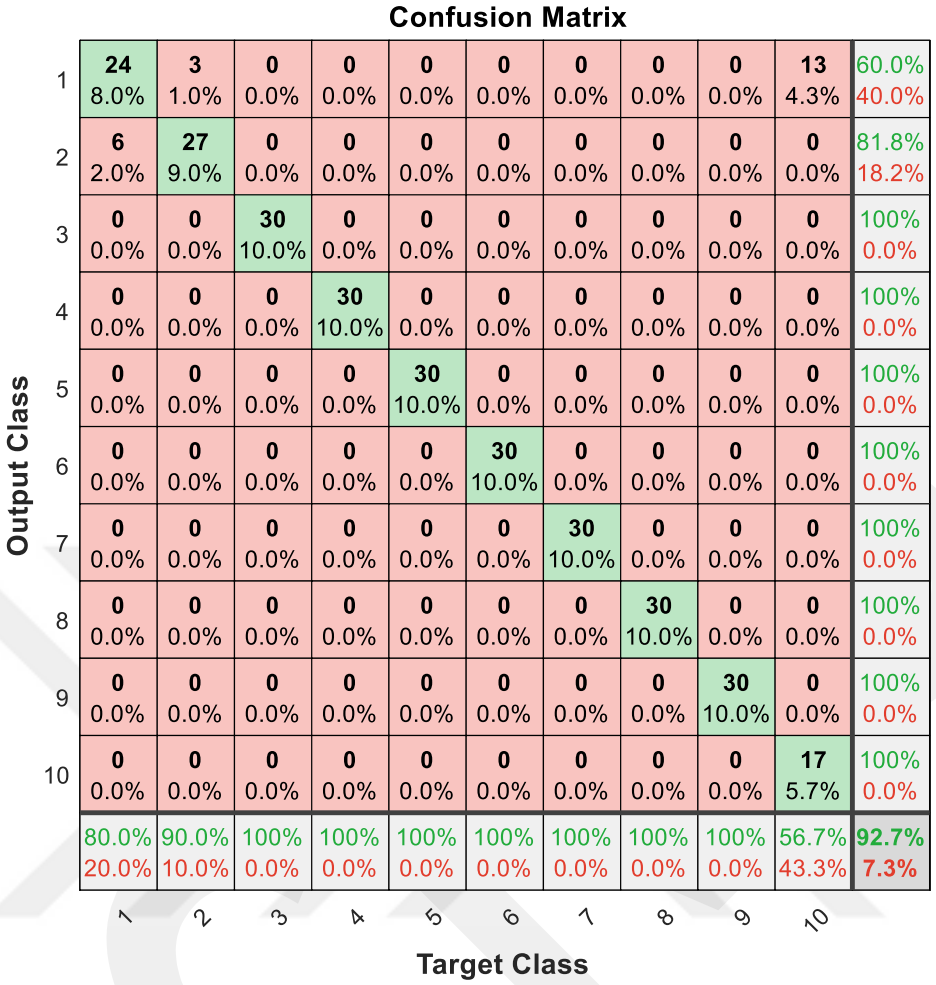

Figure 4.57 Energy confusion matrix with 30 test sample

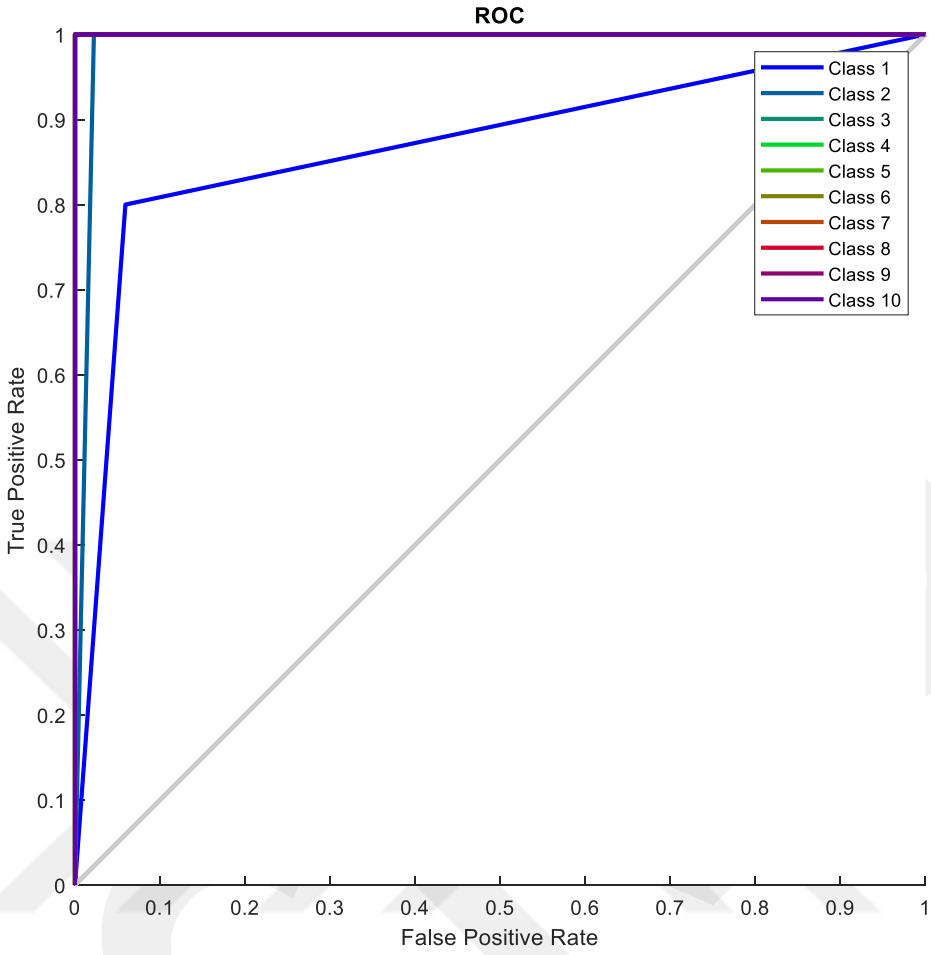

Figure 4.58 Energy Roc curve with 30 test sample

Finally, the MFCCs combined with SOFTMAX which presented lower results than previous two methods 83.7 % accuracy. The confusion matrix of this method presented in 4.58 and the roc curve presented in 4.59.

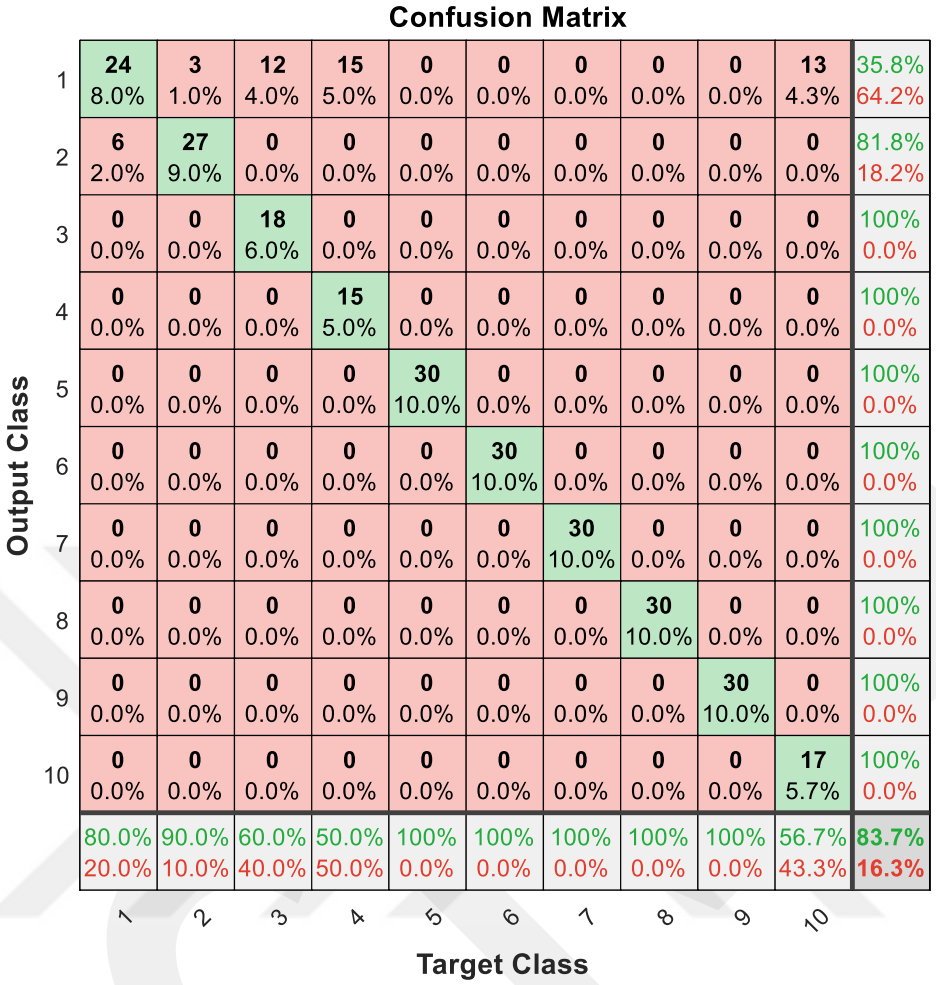

Figure 4.59 MFCCs confusion matrix with 30 test sample

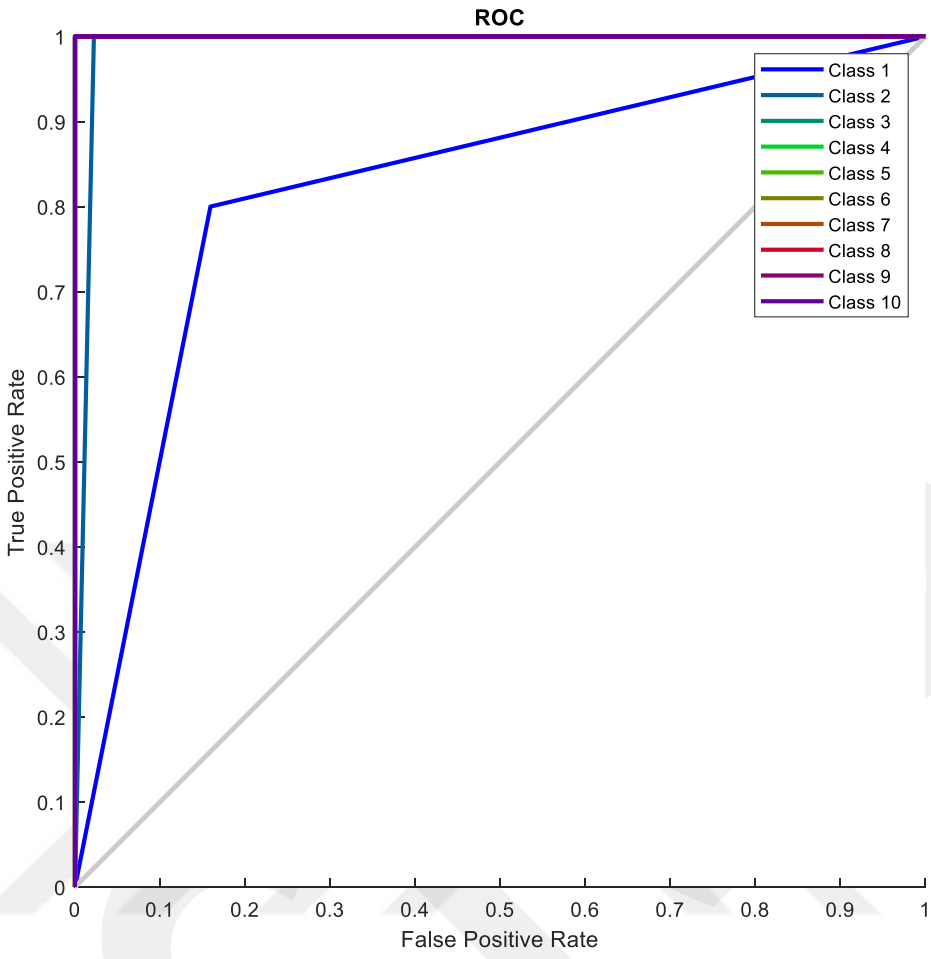

Figure 4.60 MFCCs Roc curve with 30 test sample

Table 4.2 Results Comparisons

METHOD	TEST (%)	WINDOW SIZE	METHODS	ACC (%)
MEUTZNER ET AL. [46]	10	-	MFCC+SVM	88.72
MEUTZNER ET AL. [46]	10	-	Energy + SVM	92.32
GURBAN AND THİRAN [47]	10	-	Model Based Entropy	81.32
ESTELLERS ET AL. [48]	10	-	A second-order exponential function	81.21
OUR METHOD	10	32	Pitch+SOFTMAX	95
OUR METHOD	10	32	Energy+SOFTMAX	94
OUR METHOD	10	32	MFCCs+SOFTMAX	85
OUR METHOD	30	32	Pitch+SOFTMAX	93.7
OUR METHOD	30	32	Energy+SOFTMAX	92
OUR METHOD	30	32	MFCCs+SOFTMAX	85.7

Form Table 4.2, we can prove that our method presented best results than previous studies. In [46] presented two methods MFCC+SVM and Energy + SVM the proposed methods presented 88.72 % and 92.32 respectively. Gurban and Thiran [47] presented new method based Model Based Entropy technique this method presented 81.32 % accuracy which is suitable. Estellers et al. [48] presented new method based A second-order exponential function this method is new and interested

and presented 81.21% accuracy. On the other hand, our method Pitch+SOFTMAX presented best results than all previous studies which presents 95%. Furthermore, Energy+SOFTMAX is presented best results than previous studies which presented 94%. The important point that energy+SOFTMAX presented best results than Energy + SVM which mean the SOFTMAX more effective these types of problems. Furthermore, with the 30% of test data our model presented 93.7 % with Pitch+SOFTMAX, 92 % with Energy+SOFTMAX, and 85.7% with MFCCs+SOFTMAX. Moreover, 32 point of windowing applied for all windowing techniques that are used for VAD techniques such as hamming, bohman, and Bartlett windowing techniques.

Finally, the results of that compared in Table 4.2 visualized and presented in the Figure 4.61.

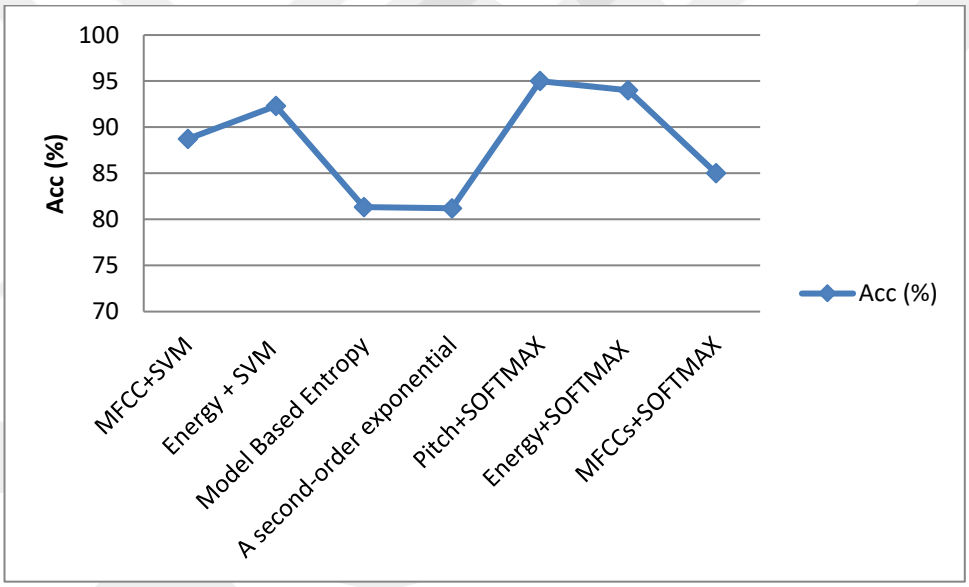

Figure 4.61 Comparisons

CHAPTER 5

CONCLUSION

The presented study consists from number of investigations such as VAD and IWR problems. The author presented three window functions which all of them presented remarkable results and good performance for in VAD problem.

Furthermore, we presented two new approaches pitch based SOFTMAX based energy based SOFTMAX for IWR problem. The pitch based SOFTMAX presented satisfactory results and energy based SOFTMAX also presented good results. One of the best conclusions in this study is that the SOFTMAX presented best results than SVM exactly when combined with energy as feature extraction method.

In this research, we presented new approach dealing with isolated word recognition. In the first stage, three functions applied for voice activity detection (VAD) problem hamming window, Bohman function, and Bartlett-Hann function. The both Bohman function and Bartlett-Hann function are not applied in previous studies for VAD problem. The method tested in two scenarios in the first one 10% of the data applied as test data and in the second scenario 30% of the data applied as the test data. The aim of applying different scenarios is to show that the proposed method have ability to deal with different sizes of training and testing and presented satisfactory results.

As future work, we advise researcher to apply these the presented techniques to other datasets such as digit voice recognition. We also advise to combine MFCCs with LSTM which we estimate that presented best results with other classifiers because MFCCs presented multi dimension features with different sizes which lead to problems and low classification rates with SOFTMAX and other classifiers that not deal with time series problems.

REFERENCES

[1] V. Mitra, H. Franco, M. Graciarena and D. Vergyri, "Medium-duration modulation cepstral feature for robust speech recognition," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 1749-1753, doi: 10.1109/ICASSP.2014.6853898.

[2] O. Hazrati, S. Ghaffarzadegan and J. H. L. Hansen, "Leveraging automatic speech recognition in cochlear implants for improved speech intelligibility under reverberation," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, 2015, pp. 5093-5097, doi: 10.1109/ICASSP.2015.7178941.

[3] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5135-5139, doi: 10.1109/ICASSP.2017.7953135.

[4] Y. Suh ., "Development of distant multi-channel speech and noise databases for speech recognition by in-door conversational robots," *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, Seoul, 2017, pp. 1-4, doi: 10.1109/ICSDA.2017.8384419.

[5] P. Nidadavolu, V. Iglesias, J. Villalba and N. Dehak, "Investigation on Neural Bandwidth Extension of Telephone Speech for Improved Speaker Recognition," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 6111-6115, doi: 10.1109/ICASSP.2019.8682992.

[6] L. Li, W. Xu, J. Wu, S. He and X. Li, "The Hokkien isolated word recognition system based on FPGA," *2014 International Conference on Anti-Counterfeiting, Security and Identification (ASID)*, Macao, 2014, pp. 1-5, doi: 10.1109/ICASID.2014.7064971.

[7] W. Shu-Guang, Z. Xiang-Yang and W. Qiang, "Isolated word recognition in reverberant environments," *2011 IEEE International Conference on Signal*

Processing, Communications and Computing (ICSPCC), Xi'an, 2011, pp. 1-4, doi: 10.1109/ICSPCC.2011.6061575.

[8] S. Sawant and M. Deshpande, "Isolated Spoken Marathi Words Recognition Using HMM," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697457.

[9] F. Ang, H. Tsutsui and Y. Miyanaga, "Incorporation of time-varying LP cepstral features in HMM-based isolated word speech recognition," *2015 International Symposium on Signals, Circuits and Systems (ISSCS)*, Iasi, 2015, pp. 1-4, doi: 10.1109/ISSCS.2015.7204030.

[10] S. Sajjan and C. Vijaya, "Comparison of DTW and HMM for isolated word recognition," *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, Salem, Tamilnadu, 2012, pp. 466-470, doi: 10.1109/ICPRIME.2012.6208391.

[11] S. Masood, M. Mehta, and D. R. Rizvi, "Isolated word recognition using neural network," *2015 Annual IEEE India Conference (INDICON)*, New Delhi, 2015, pp. 1-5, doi: 10.1109/INDICON.2015.7443697.

[12] M. Sher, N. Ahmad and M. Sher, "TESPAR feature based isolated word speaker recognition system," *18th International Conference on Automation and Computing (ICAC)*, Loughborough, 2012, pp. 1-4.

[13] C. Wei and Y. Yang, "Mandarin isolated words recognition method based on pitch contour," *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, Hangzhou, 2012, pp. 143-147, doi: 10.1109/CCIS.2012.6664385.

[14] G. Wang, Y. Zhang, M. Sun, X. Wang and Y. Zhang, "Speech signal feature parameters extraction algorithm based on PCNN for isolated word recognition," *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, Shanghai, 2016, pp. 679-682, doi: 10.1109/ICALIP.2016.7846618.

[15] M. Raczynski, "Speech processing algorithm for isolated words recognition," *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, Swinoujście, 2018, pp. 27-31, doi: 10.1109/IIPHDW.2018.8388238.

[16] P. Saini, Parneet Kaur, M. Dua, “Hindi Automatic Speech Recognition Using HTK,” International Journal of Engineering Trends and Technology (IJETT) – Volume4 Issue6- June 2013.

[17] X. Gaurav, D. Devi, K. Sharma, M. Bhattacharya, “Development of Application Specific Continuous Speech Recognition System in Hindi”, Journal of Signal and Information Processing, 2012, 3, 394-401.

[18] M. Murthy, G. Murthy,"Isolated Word Recognition Using LPC & Vector Quantization", International Journal of Research in Engineering and Technology(IJRET)-Volume1 Issue3- Nov2012.

[19] L. Xu and M. Ke, "Research on isolated word recognition with DTW-based," *2012 7th International Conference on Computer Science & Education (ICCSE)*, Melbourne, VIC, 2012, pp. 139-141, doi: 10.1109/ICCSE.2012.6295044.

[20] P. Shivakumar, P. Georgiou, Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations, Computer Speech & Language, Volume 63, 2020, 101077, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2020.101077.

[21] M. Kumar, S. Kim, C. Lord, D. Lyon, S. Narayanan, Leveraging Linguistic Context in Dyadic Interactions to Improve Automatic Speech Recognition for Children, Computer Speech & Language, Volume 63, 2020, 101101, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2020.101101.

[22] T. de Lima, M. Da Costa-Abreu, A survey on automatic speech recognition systems for Portuguese language and its variations, Computer Speech & Language, Volume 62, 2020, 101055, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2019.101055.

[23] H. Heidari, S. Gobee, Isolated Word Command Recognition for Robot Navigation, Procedia Engineering, Volume 41, 2012, Pages 412-419, ISSN 1877-7058, https://doi.org/10.1016/j.proeng.2012.07.192.

[24] T. Fukuda, O. Ichikawa, M. Nishimura, Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition, Speech Communication, Volume 98, 2018, Pages 95-103, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2018.01.008.

[25] S. Shahmoradi, S. Shouraki, Evaluation of a novel fuzzy sequential pattern recognition tool (fuzzy elastic matching machine) and its applications in speech and handwriting recognition, Applied Soft Computing, Volume 62, 2018, Pages 315-327, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2017.10.036.

[26] Z. Tan, A. Sarkar, N. Dehak, rVAD: An unsupervised segment-based robust voice activity detection method, Computer Speech & Language, Volume 59, 2020, Pages 1-21, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2019.06.005.

[27] G. Kontonatsios, S. Spencer, P. Matthew, I. Korkontzelos, Using a Neural Network-based Feature Extraction Method to Facilitate Citation Screening for Systematic Reviews, Expert Systems with Applications: X, 2020, 100030, ISSN 2590-1885, https://doi.org/10.1016/j.eswax.2020.100030.

[28] J. Zhang, L. Liu, Ling Zhen, L. Jing, A unified robust framework for multi-view feature extraction with L2,1-norm constraint, Neural Networks, 2020, ISSN 0893-6080,https://doi.org/10.1016/j.neunet.2020.04.024.

[29] T. Bikmukhametov, J. Jäschke, Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models, Computers & Chemical Engineering, Volume 138, 2020, 106834, ISSN 0098-1354, https://doi.org/10.1016/j.compchemeng.2020.106834.

[30] J. Waring, C. Lindvall, R. Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, Artificial Intelligence in Medicine, Volume 104, 2020, 101822, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2020.101822.

[31] A. Roohi, K. Faust, U. Djuric, P. Diamandis, Unsupervised Machine Learning in Pathology: The Next Frontier, Surgical Pathology Clinics, Volume 13, Issue 2, 2020, Pages 349-358, ISSN 1875-9181, ISBN 9780323756068, https://doi.org/10.1016/j.path.2020.01.002.

[32] Z. Ren, Junchi Yan, X. Yang, A. Yuille, H. Zha, Unsupervised learning of optical flow with patch consistency and occlusion estimation, Pattern Recognition, Volume 103, 2020, 107191, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2019.107191.

[33] B. Olasege, S. Zhang, Q. Zhao, D. Liu, H. Sun, Q. Wang, P. Ma, Y. Pan, Genetic parameter estimates for body conformation traits using composite index, principal component, and factor analysis, Journal of Dairy Science, Volume 102, Issue 6, 2019, Pages 5219-5229,

ISSN 0022-0302, https://doi.org/10.3168/jds.2018-15561.

[34] X. Huang, Y. Ye, H. Guo, Yi Cai, H. Zhang, Y. Li, DSKmeans: A new kmeans-type approach to discriminative subspace clustering, Knowledge-Based Systems, Volume 70, 2014, Pages 293-300, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2014.07.009.

[35] M. Mrówczyńska, J. Sztubecki, A. Greinert, Compression of results of geodetic displacement measurements using the PCA method and neural networks, Measurement, Volume 158, 2020, 107693, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2020.107693.

[36] X. Wang, X. Lin, X. Dang, Supervised learning in spiking neural networks: A review of algorithms and evaluations, Neural Networks, Volume 125, 2020, Pages 258-280, ISSN 0893-6080, https://doi.org/10.1016/j.neunet.2020.02.011.

[37] M. Zhang, J. Wu, A. Belatreche, Z. Pan, X. Xie, Y. Chua, G. Li, H. Qu, H. Li, Supervised Learning in Spiking Neural Networks with Synaptic Delay-Weight Plasticity, Neurocomputing, 2020, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2020.03.079.

[38] C. Gong, Z. Su, P. Wang, Q. Wang, Cumulative belief peaks evidential K-nearest neighbor clustering, Knowledge-Based Systems, 2020, 105982, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2020.105982.

[39] M. Khalid, S Nisar, S. Khan, M. Khan, W. Troesch, Simulation of blended nonlinear hydrodynamics forces using radial basis function in uniform moving

frame, Ocean Engineering, Volume 198, 2020, 106994, ISSN 0029-8018, https://doi.org/10.1016/j.oceaneng.2020.106994.

[40] N. Karimi, S. Kazem, D. Ahmadian, H. Adibi, L.V. Ballestra, On a generalized Gaussian radial basis function: Analysis and applications, Engineering Analysis with Boundary Elements, Volume 112, 2020, Pages 46-57, ISSN 0955-7997, https://doi.org/10.1016/j.enganabound.2019.11.011.

[41] G. Ren, Y. Wang, J. Ning, Z. Zhang, Using near-infrared hyperspectral imaging with multiple decision tree methods to delineate black tea quality, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, Volume 237, 2020, 118407, ISSN 1386-1425, https://doi.org/10.1016/j.saa.2020.118407.

[42] H. He, D. Han, J. Dezert, Disagreement based semi-supervised learning approaches with belief functions, Knowledge-Based Systems, Volume 193, 2020, 105426, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2019.105426.

[43] J. Gordon, J. Hernández-Lobato, Combining deep generative and discriminative models for Bayesian semi-supervised learning, Pattern Recognition, Volume 100, 2020, 107156, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2019.107156.

[44] D. Wu, M. Shang, X. Luo, J. Xu, H. Yan, W. Deng, G. Wang, Self-training semi-supervised classification based on density peaks of data, Neurocomputing, Volume 275, 2018, Pages 180-191, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2017.05.072.

[45] J. Jiang, H. Gan, L. Jiang, C. Gao, N. Sang, Semi-supervised Discriminant Analysis and Sparse Representation-based self-training for Face Recognition, Optik, Volume 125, Issue 9, 2014, Pages 2170-2174, ISSN 0030-4026, https://doi.org/10.1016/j.ijleo.2013.10.043.

[46] H. Meutzner, V. Nguyen, T. Holz, D. Kolossa, Using Automatic Speech Recognition for Attacking Acoustic CAPTCHAs: The Trade-off between Usability and Security, in: Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, ICASSP 2007, IEEE, 2007, pp. 945–948.

[47] M. Gurban, J.-P. Thiran, Using entropy as a stream reliability estimate for audiovisual speech recognition, in: 2008 16th European Signal Processing Conference, IEEE, 2008, pp. 1–5.

[48] V. Estellers, M. Gurban, J.-P. Thiran, On dynamic stream weighting for audiovisual speech recognition, IEEE Trans. Audio Speech Lang. Process. 20 (2012) 1145–1157.

[49] A. Karim, S. Güzel, M. Tolun, H. Kaya, and F. Çelebi, A New Generalized Deep Learning Framework Combining Sparse Autoencoder and Taguchi Method for Novel Data Classification and Processing, Mathematical Problems in Engineering, vol. 2018, Article ID 3145947, 13 pages, 2018. https://doi.org/10.1155/2018/3145947.

[50] A. Karim, S. Güzel, M. Tolun, H. Kaya, F. Çelebi, A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing, Biocybernetics and Biomedical Engineering, Volume 39, Issue 1, 2019, Pages 148-159, ISSN 0208-5216, https://doi.org/10.1016/j.bbe.2018.11.004.

[51] A. Karim, Ö. Karal, and F. Çelebi, "A New Automatic Epilepsy Serious Detection Method by Using Deep Learning Based on Discrete Wavelet Transform," no. 4, 15–18, 2018.