DESIGN AND IMPLEMENTATION OF A PLATFORM INDEPENDENT AUTOMATIC SPEECH RECOGNITION SYSTEM

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

OF

ATILIM UNIVERSITY

BY

DOĞAN URGUN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

THE DEPARTMENT OF MECHATRONICS ENGINEERING

JULY 2012

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

Prof. Dr. K. İbrahim AKMAN

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof. Dr. Abdulkadir ERDEN

Head of Department

This is to certify that we have read the thesis "Design and Implementation of a Platform Independent Automatic Speech Recognition System" submitted by "Doğan URGUN" 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	Prof. Dr. Abdulkadir ERDEN
Co-Supervisor	Supervisor

Examining Committee Members

Prof. Dr. Abdulkadir ERDEN	______________________
Assoc. Prof. Dr. Fuad ALIEW	______________________
Asst. Prof. Dr. Bülent İRFANOĞLU	______________________
Asst. Prof. Dr. Çiğdem TURHAN	______________________
Asst. Prof. Dr. Hakan TORA	______________________

Date: 16.07.2012

I declare and guarantee that all data, knowledge and information in this document has been obtained, processed and presented in accordance with academic rules and ethical conduct. Based on these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Doğan URGUN

ABSTRACT

DESIGN AND IMPLEMENTATION OF A PLATFORM INDEPENDENT AUTOMATIC SPEECH RECOGNITION SYSTEM

URGUN, Doğan

M.S., Mechatronics Engineering Department

Supervisor: Prof. Dr. Abdulkadir ERDEN

Co-Supervisor: Asst. Prof. Dr. Hakan TORA

July 2012, 63 pages

In this thesis, an isolated word syllable based speech recognition system is designed and tested. The system is limited to 50 words in Turkish language. Because of the agglutinative nature of the Turkish language, a word based approach in speech recognition has a negative impact on system performance. For this reason, phoneme based structures are widely used in speech recognition but small unit size of phonemes introduces difficulty to their recognition. Hence it is decided to follow a syllable based approach in this thesis.

In the scope of this study a syllable detection algorithm is designed. Mel Frequency Cepstral Coefficients are selected to extract features of the speech signal and Artificial Neural Networks are used to classify syllables. Finally, the results of the Artificial Neural Networks are processed in a language model which is designed for this syllable based system.

The designed system is trained and tested with 10 samples of 5 different users for 50 different Turkish words. The system showed approximately 85% performance in recognition of speech. Moreover the system is tested for an untrained user and approximately 75% performance in recognition of speech is obtained.

Keywords: Syllable Based, Speech Recognition, Neural Networks, Pattern Recognition

ÖZ

PLATFORMDAN BAĞIMSIZ BİR OTOMATİK KONUŞMA TANIMA SİSTEMİNİN TASARLANMASI VE UYGULANMASI

URGUN, Doğan

Yüksek Lisans, Mekatronik Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. Abdulkadir ERDEN

Ortak Tez Yöneticisi: Yrd. Doç. Dr. Hakan TORA

Temmuz 2012, 63 Sayfa

Bu tez içerisinde, 50 kelime ile sınırlı bir ayrık kelimeli hece tabanlı konuşma tanıma sistemi tasarlanmış ve test edilmiştir. Türkçenin sondan eklemeli yapısından dolayı kelime tabanlı bir yaklaşımın konuşma tanıma performansı üzerinde negatif etkisi olacaktır. Bu sebepten ötürü fonem tabanlı yapılar konuşma tanıma içinde geniş biçimde kullanılmaktadır ancak fonemlerin küçük boyutları onların tanınmasını zorlaştırmaktadır. Bu nedenle, bu tezde hece tabanlı bir yaklaşımın takip edilmesine karar verilmiştir.

Bu tez kapsamında bir hece tespit etme algoritması tasarlanmıştır. Mel Frekansı Kepstral Katsayıları özellik çıkarmak üzere seçilmiştir ve Yapay Sinir Ağları hecelerin sınıflandırılmasında kullanılmıştır. Son olarak, Yapay Sinir Ağlarının sonuçları hece tabanlı sistem için tasarlanan dil modeli içerisinde işlenmiştir.

Tasarlanan sistem 5 farklı kişiden 50 kelime için alınan 10 örnekle eğitilmiş ve test edilmiştir. Sistem yaklaşık %85 konuşma tanıma performansı göstermiştir. Ayrıca sistem eğitilmemiş bir konuşmacı tarafından da test edilmiş ve %75 konuşma tanıma performansı elde edilmiştir.

Anahtar Kelimeler: Hece Tabanlı, Konuşma Tanıma, Yapay Sinir Ağları, Örüntü Tanıma.

To My Mother, Father and little brothers Alper, Kamil and Sinan

ACKNOWLEDGMENTS

I express my sincere appreciation to my supervisor Prof. Dr. Abdulkadir ERDEN for his guidance and insight throughout the research. Also thanks to my co-supervisor Asst. Prof. Dr. Hakan TORA.

I would like to thank Asst. Prof. Dr. Bülent İRFANOĞLU for his great advices and support.

I wish to thank Alper ERTÜRK, Kağan TELEK and Levent ÖZER for their assistance in managing sound files.

Also would like to thank my cousin, Ramazan ALPTEKİN for his assistance and friendship within this thesis.

Thanks to all my friends, especially to the ones who were always with me throughout this period: Orçun KİRİŞ and Serkan ÖZKUL.

Special thanks to my one and only Nuray ERKUTLU, with her love and generous support with this thesis.

Of course, I would like to thank my family with much love and pride. In particular my parents, Emine and Osman URGUN, who encouraged me into continuing my education and gave me the support, guidance and love throughout and being patient with me all the way.

TABLE OF CONTENTS

LIST OF TABLE

TABLE OF FIGURES

NOMENCLATURE

ANN: Artificial Neural Networks

ASR: Automatic Speech Recognition

DFT: Discrete Fourier Transform

DTW: Dynamical Time Warping

DWNN: Discrete Wavelet Neural Networks

FFT: Fast Fourier Transform

FIR: Finite Impulse Response

FSN: Finite State Network

HMM: Hidden Markov Model

LPC: Linear Predictive Coding

LSD: Linguistic Statistical Decoder

LTA: Linear Time Alignment

MCE: Minimizing Classification Error

MFCC: Mel Frequencies Cepstral Coefficients

PCM: Pulse Code Modulation

SVM: Support Vector Machine

TDNN: Time Delay Neural Networks

CHAPTER 1

INTRODUCTION

Speech is one of the most important ways of communication between humans. Technology which converts human speech into text is known as speech recognition. The idea of being able to speak with machines is not only to attract people but also to provide them with many advantages. Hands free usage, easy generation of speech and natural structure of spoken inputs are just a few of these advantages. For these reasons, speech recognition has become a popular research field for the past six decades.

Since the invention of the first speech recognizer at Bell Laboratory in 1952 (Davis, 1952), significant progress in speech recognition techniques was achieved. Researches started on speech recognition before the invention of computer however it has gained significant acceleration after the development of digital computers which could convert analog signals to digital data. Even if today's speech recognition techniques provide us with the implementation of many successful applications, it is still far away from complete, reliable and robust speech recognizers. Unsteady characteristic of speech signal, which depends on recording and transmission environment and emotional state of speakers which directly affects the speech features, are the most common problems of topic.

General description of a speech recognition system can be defined with a source-channel model as given in Figure 1 (Jelinek, 1985). In this model speaker decides the message which he/she will utter using his/her vocal apparatus. Acoustic channel is defined as composition of speech producer, noisy transmission channel and acoustic processor. At the end of the acoustic channel, acoustic processor generates acoustic observation vector 'A', after the analysis of the audio signal. Acoustic processor and linguistic decoder create the speech recognizer together. The linguistic model makes a prediction to decide message 'W' by using a previously defined language model. As a result, the most likely prediction, '$\hat{W}$', is assigned according to the analyses of 'A' in the language model.

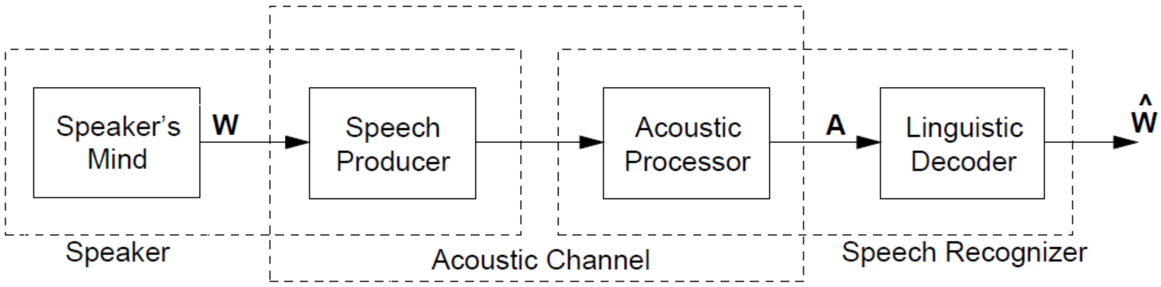

Figure 1.1 Source channel model of speech recognition (Jelinek, 1985)

There are a variety of approaches for speech recognition systems. These systems can be classified in three categories: "Template Based Classify Techniques", "Stochastic Classify Techniques" and "Connectionist Classify Techniques". In template based models, the system recognizes speech by matching the voice input to the recorded data of templates. "Dynamic Time Warping (DTW[1])" and "Linear Time Alignment (LTA)" are most known methods for template based techniques. Stochastic speech recognition technique is one of the most common techniques recently. In these techniques, voice input modeling is achieved by probabilistic models. Hidden Markov Model (HMM) is the most popular probabilistic model for speech recognition. In connectionist techniques, classification is made by neural networks.

The capacity of vocabulary which a machine can recognize is one of the differences in speech recognition systems. Increasing vocabulary size of a system also increases the system complexity. Speech recognition systems can be classified as small vocabulary, medium vocabulary or large vocabulary systems.

Another important point in speech recognition systems is spelling. A speech recognition system basically can be an isolated word system or a continuous time system. In isolated word system, speech input is a discretely spelled word sequence. The user must pause between spellings of different words. In continuous time speech recognition, user does not have to pause between words. Although these systems provide more natural use, there are some practical difficulties in detecting word boundaries. There are also hybrid approaches in spelling like connected speech recognition which are a combination of continuous and discrete time speech.

[1] All of the abbreviations are given in Appendix A with brief explanations.

Another difference in speech recognition systems is being speaker dependent or speaker independent. Speaker dependent systems are designed for specified speakers only. These systems require a training stage and they are generally more successful than the speaker independent systems. A speaker independent system is designed to perform for any speaker. These systems do not need to be trained before using, and success of the speech recognition is less satisfactory than the speaker dependent systems.

1.1 Statement of the Problem

Main problem of this thesis work is to design and implement a speech recognition system for use in the Humanoid Systems Laboratory of the Mechatronics Engineering Department at the Atılım University. The system should recognize some defined words which are detected from meaningful sentences in the Turkish language. The proposed system is planned to be integrated into projects of multipurpose assistant robots like; guide robot, guard robot, office robot, etc. Later on, the thesis will be part of the large scale humanoid robot project.

1.2 Scope of Thesis

Scope of the thesis covers the following topics:

- Speech recognition system will detect, extract and recognize the words from acquired live records in laboratory environment.
- System should recognize untrained people.
- System should recognize approximately 50 preselected words in the Turkish vocabulary.

1.3 Outline of Thesis

Outline of thesis is as follows:

- Chapter 2 introduces speech recognition system applications that are published in literature.
- Chapter 3 describes theory of speech recognition system that is based on problem statement of thesis.
- Chapter 4 summarizes the experiments performed at our laboratories and their results using the speech recognition system.
- Chapter 5 discusses and concludes research work and gives some suggested future work on the thesis topic.

CHAPTER 2

LITERATURE SURVEY

Automatic Speech Recognition (ASR) systems are one of the popular research topics for more than half of a century. This chapter gives a brief overview of the history and research state of speech recognition for both Turkish and other languages in literature.

2.1 Brief History of Automatic Speech Recognition Systems

Studies on this field was started to appear in published literature after the creation of a spoken digit recognizer in 1952 at Bell Laboratories, where a recognizer circuit was designed to recognize isolated digits for individual speakers. The system recognized digits by looking at limited number of spectral features of speech with an accuracy varying between 97-98 percent (Davis, 1952).

Another system was designed to recognize 10 distinct syllables which were defined as 10 monosyllabic words for an independent speaker in RCA laboratories in 1956 (Olson, 1956).

In 1959 a phoneme recognizer was developed at University College in England. The system was designed to recognize 4 vowels and 9 constants (Fry, 1959). They increased the overall phoneme recognition accuracy for words consisting of two or more phonemes by incorporating statistical information concerning allowable phoneme sequences in English. This work marked the first use of statistical syntax at the phoneme level in automatic speech recognition (Furui, 2005).

In the same year, a speaker-independent 10 vowel recognizer was built in MIT Lincoln Laboratories. The recognized vowels placed in /b/ vowel /t/ form (Forgie, 1959). Because of the inadequate speed of the 1960's computers, some researchers had focused on special purpose hardware for speech processing. Most prominent of these researches were the vowel recognizer designed in the Radio Research

Laboratory in Tokyo (Suzuki, 1961), the phoneme recognizer devised at Kyoto University (Sakai, 1962) and the digit recognizer which built in NEC Laboratories (Nagata, 1963).

One of the difficult problems of speech recognition is in the non-uniformity of time scale in speech events (Furui, 2005). In the 1960s some of the researchers focused on this problem and achieved some success in normalization methods of speech pattern matching.

Performance of the recognizer were significantly improved by using basic time-normalization methods which were based on start-end point detection designed at RCA Laboratories in 1964 (Martin, 1964).

Another pioneering research was published by Carnegie Mellon University in 1966. In this study, researchers worked in continuous speech recognition field by dynamic tracking of phonemes (Reddy, 1966).

In 1968, some researchers in the Soviet Union developed a system which used dynamic programming methods for time alignment between two speech patterns in order to derive a reasonable assessment of their similarity. The method is also known as Dynamic Time Warping (DTW) (Vintsyuk, 1968). In the same years, some researchers at NEC Laboratories published their research results that used dynamic programming technique to solve the non-uniformity problem in speech recognition systems as an independent study (Sakoe, 1978).

Some improvements on pattern recognition idea for isolated word speech recognition were published in Russia in 1970 (Velichko, 1970). A new approach based on using Linear Predictive Coding (LPC) as spectral parameters instead of common feature extraction methods like Fourier transform were published by Bell Laboratories in 1975 (Itakura, 1975). In the same years, a Linguistic Statistical Decoder (LSD) was designed for continuous time speech recognition in IBM laboratories. The designed structure of LSD was described in four parts:

a) A statistical model of the language being recognized.
b) A phonemic dictionary and statistical phonological rules characterizing the speaker.

c) A phonetic matching algorithm that computes the similarity between phonetic strings by using the performance characteristics of the acoustic-processor.
d) Word level search control (Jelinek, 1975).

A series of experiments were initiated at Bell Laboratories in order to develop speaker independent speech recognition systems in 1979. They used a wide range of sophisticated clustering algorithms to determine the number of distinct patterns which are needed to represent all variations of different words across a wide user population (Rabiner, 1979).

An ambitious speech understanding project was funded by the Defense Advanced Research Projects Agency (DARPA), which led to many seminal systems and technologies (Klatt, 1977). One of the first demonstrations of speech understanding was achieved by Carnegie Mellon University in 1973. Their Hearsay I system was able to use semantic information to significantly reduce the number of alternatives considered by the recognizer. CMU's Harpy system (Lowerre, 1990) was shown to be able to recognize speech using a vocabulary database which had more than one thousand words with reasonable accuracy. One particular contribution from the Harpy system was the concept of graph search, where the speech recognition language is represented as a connected network derived from lexical representations of words, with syntactical production rules and word boundary rules. The Harpy system was the first to take advantage of a finite state network (FSN) to reduce computation and efficiently determine the closest matching string (Furui, 2005). In 1979, a two level dynamic programming approach was developed by researchers in NEC laboratories. The algorithm is based on pattern matching algorithm for connected word recognition. Performance of the algorithm is tested in a talker adapted recognition experiment that digit data (one to four digits) connectedly spoken by five persons are recognized with close to 99.6 percent accuracy. Computation time and memory requirement are both proved to be within reasonable limits (Sakoe, 1978).

In the late 1970s, some researchers used Viterbi algorithm as a dynamic programming method, which came from the Communication Theory Community at its origin, and it was recognized as an indispensable technique in automatic speech recognition (Viterbi, 1979).

In the same year, researchers in Joint Speech Research Laboratory developed a system to recognize continuous connected words by using whole word templates (Bridle, 1979). Another important study in field of connected word recognition is made in Bell Laboratory, where a frame synchronous network search algorithm was developed for recognition of connected words (Lee, 1989).

One of the key technologies developed in the 1980s is the Hidden Markov Model (HMM) approach (Ferguson, 1980 Rabiner, 1973, 1993). It is doubly stochastic process in that it has an underlying stochastic process that is not observable (hence the term hidden), but can be observed through another stochastic process that produces a sequence of observations. Although the HMM was well known and understood in a few laboratories (primarily IBM, Institute for Defense Analysis (IDA) and Dragon Systems), it was not until widespread publication of the methods and theory of HMMs in the mid 1980s that the technique became widely applied in virtually every speech recognition research laboratory all over the world (Furui, 2005).

In 1986, use of the combination of instantaneous cepstral coefficients and their first and second order polynomial coefficients as fundamental spectral features for speech recognition was proposed by Furui (Furui, 1986). The method was first proposed for speaker recognition in the late 1970s, but no one attempted to apply it to speech recognition for many years. This method is now widely used in almost all speech recognition systems (Furui, 2005).

In 1989, researchers at Carnegie Mellon University presented time delay neural networks approach for phoneme recognition. The system is characterized by two important properties; First, they used a three layer arrangement of simple computing units, a hierarchy constructed that allowed for the formation of arbitrary nonlinear decision surfaces, which the Time Delay Neural Network (TDNN) learned automatically using error back-propagation. Second, the time-delay arrangement enabled the network to discover acoustic phonetic features and the temporal relationships between them are independent of position in time and therefore not blurred by temporal shifts in the input. As a recognition task, speaker dependent recognition of the phonemes B, D, and G in varying phonetic context was chosen. Researchers compared the system with several discrete Hidden Markov Models (HMM) which were trained to perform the same task. Performance evaluation from three speakers showed that the TDNN achieves a recognition rate of 98.5% correct while the rate obtained by the best of the HMMs was only 93.7% (Weibel, 1989). The DARPA community has conducted research on large vocabulary, continuous speech recognition systems, aiming at achieving high word accuracy for a 1000 word database management task. Major research contributions resulted from efforts at CMU with the SPHINX system (Lee, 1990), BBN with the BYBLOS system (Chow, 1987), SRI with the DECIPHER system (Weintraub, 1989), Lincoln Laboratory (Paul, 1989), MIT (Zue, 1989) and AT&T Bell Laboratory (Lee, 1990). The SPHYNX system successfully integrated the statistical method of HMM with the network search strength of the earlier Harpy system. Hence, it was able to train and embed context dependent phone models in a sophisticated lexical decoding network (Furui, 2005).

In 1990, researchers at Royal Signals and Radar Establishment focused on performance problems of automatic speech recognition systems in the presence of interfering signals and noise. They proposed a new technique to increase robustness of speech recognition systems in noisy environments. The technique exploited the ability of hidden Markov models to model dynamically varying signals in order to accommodate concurrent processes, including interfering complex signals like speech (Varga, 1990).

In 1995, Researchers at Cambridge University developed a method of speaker adaptation for continuous density Hidden Markov Model (HMM). An initial speaker-independent system is adapted to improve the modeling of a new speaker by updating the HMM parameters. Statistic data were gathered by researchers from the available adaptation data and used to calculate a linear regression based transformation for the mean vectors. The transformation matrices were calculated to maximize the likelihood of the adaptation data and implemented, them for using the forward backward algorithm. Experiments performed on the ARPA RM1 database using an HMM system with cross word tri-phones and mixture Gaussian output distributions. Results showed that using 40 adaptation utterances was achieved a 37% reduction in error from the speaker independent system with supervised adaptation and a 32% reduction in unsupervised mode (Legetter, 1995).

A new method was proposed for training of speech recognition systems in 1997 that they differentiate the method of classifier design by way of distribution estimation and the discriminative method of minimizing classification in error rate (MCE). As a result of the study, the MCE method provided a significant reduction of recognition in error rate (Chou, 2003).

In 1998, a novel framework of robust speech understanding is presented by researchers in Bell laboratory. The researchers proposed a combined detection and verification strategy to realize flexible speech understanding. It extracted the semantically significant parts of information and rejected their relevant parts rather than decoding the whole utterances. The experiments showed that the proposed strategy showed an effective performance (Kawahara, 1998).

In 2005, researchers developed a structural metadata extraction which was used in the DARPA - Effective Affordable Reusable Speech to Text (EARS) rich transcription program. The goal of the research includes detection of sentence boundaries, filler words, and disfluences. Modeling approaches combined lexical, prosodic, and syntactic information, using various modeling techniques for knowledge source integration. The performance of these methods is evaluated by task, by data source and by origin of transcriptions, whether they come from humans or from an automatic speech recognizer. A representative sample of results show that combining multiple knowledge sources (words, prosody, syntactic information) is helpful, that prosody is more helpful for news speech than for conversational speech, where word errors significantly impact performance, and discriminative models generally provide benefit over maximum likelihood models (Liu, 2005).

2.2 Brief History of Automatic Speech Recognition Systems for Turkish Language

Implementing a speech recognition algorithm in another language does not always give similar results. This is because nature of the languages has their own specific rules and structures. Research on Automatic Turkish Speech Recognition have started in the past two decades and accelerated in the last decade. Some of these studies are listed here;

In 1994, researchers in Hacettepe University generated a Turkish Codebook by using feature vectors of sounds. They assumed each written letter in Turkish as Turkish

Phoneme and designed a Turkish speech phoneme clustering system which is based on self organizing feature maps (Artuner, 1994).

In 1996, a speech recognition system which was based on pattern comparison techniques was designed in Istanbul Technical University. Linear Predictive Coding (LPC) was used as feature extraction method and an acoustic model was created based on Dynamic Time Warping (Meral, 1996).

A speaker dependent speech recognition system was designed to recognize connected numerals by researchers in METU in 1997. The system used DTW method to classify sound vectors which were extracted based on LPC (Özkan, 1997).

A Turkish isolated word, speaker independent speech recognition system under noisy environments was designed by the researchers in Hacettepe University in 1999. The system is a word based structure and used Hidden Markov Model as classification method (Mengüşoğlu, 1999).

In 1999, a rule based speech recognition system was designed and implemented by researchers in Hacettepe University. In this research, Mel-cepstrum and Rasta features were tested for the phoneme-based isolated word speech recognition system. Successful classification rate of the system was 89.6% for the Mel-cepstrum and 94.3% for the Rasta features (Karaca, 1999).

In 1999, a speaker dependent large vocabulary speech recognition system for Turkish was designed in Bilkent University. They modeled the words as tri-phones and used Hidden Markov Model as classification method (Yılmaz, 1999).

In 2002, an acoustic feature analysis for robust speech recognition was made by researchers in Boğaziçi University. The system used MFCC and Rasta-Plp methods for feature extraction and Hidden Markov Model for classification (Koç, 2002).

In 2005, a speaker dependent large vocabulary isolated word speech recognition system was designed for Turkish Language by researchers in Marmara University. They developed a phoneme based system by using Hidden Markov Models (Tunalı, 2005).

In 2006, a unified language model for large vocabulary continuous speech recognition of Turkish was developed by researchers in Boğaziçi University. They used Hidden Markov Model for their study. The system was designed for both speaker dependent and speaker independent cases. Performance of the system was approximately 20% for speaker dependent system and 28% for speaker independent systems (Arısoy, 2006).

In 2007, a syllable based speech recognition system was designed and implemented for Turkish Language by researchers in Hacettepe University. They designed a syllable detection algorithm for Turkish which has showed 44% success rate in detection of syllables. Neural Networks were used as classification method (Can, 2007).

In 2007, an automatic Turkish word recognition system was designed by researchers in Fırat University. Discrete Wavelet Neural Network which based on adaptive entropy was used. The system tested on Turkish spoken words under noise. The performance of the system was 92% for 15 words (Avcı, 2008).

In 2007, Turkish speech corpora and recognition were developed by porting SONIC: Towards multilingual speech recognition designed by researchers in METU. The system designed as speaker Independent and used Hidden Markov Model tri-phone model. Feature vectors were extracted based on MFCC (Salor, 2007).

In 2007, a domain-specific keyword spotting system was designed by researchers in Boğaziçi University. They proposed an alternative language model based on Hidden Markov Model (Haznedaroğlu, 2007).

In 2008, a speaker dependent syllable based speech recognition engine was designed by researchers in Dokuz Eylül University. System used MFCC, LPC, PARCOR, RASTA methods for feature extraction. A comparison of Hidden Markov Model, Neural Networks, Dynamical Time Warping and SVM were used in different feature extraction methods. The best result obtained in MFCC and DTW combination with 94.2% (Aşlıyan, 2008).

In 2010, a computer based system control using voice input was designed by researchers in Dokuz Eylül University. The System used MFCC as feature extraction method and Dynamic Time Warping as classification method. They tested the performance of the system in both trained speaker and non-trained speakers for previously defined four commands. The performance of the system was approximately 90% for trained speaker and 84% for non-trained speaker (Karakaş, 2010).

In 2010, a continuous time digit sequence recognition system was designed by researchers in Osmangazi University. The system designed as speaker independent and phone based. Hidden Markov Model was chosen as classification method. System was trained for 50 speakers and showed approximately 74% success rate in recognition of speech (Tunca, 2010).

2.3 Current State of Automatic Speech Recognition Systems

Despite of the fact that the great progress in speech recognition technology has been achieved in the past, the technology is still open to many improvements on the topic.

Even if today's speech recognizing systems work almost perfectly in a small set of vocabulary, recognition rates decrease as vocabulary size grows. For instance, a spoken digit recognition system can be implemented perfectly today, but a system with 10000 or more words capacity might have increasing error rates depending on the vocabulary size. The errors could be observed more frequently in continuous time systems because of the difficulties in detecting word boundaries. In addition to that, some utterances might increase complexity of recognition which increases the error rate in large vocabulary systems (Dahl, 2012).

Another focus research topic in speech recognition is achieving speaker independence. Since structure of speech signal contains many speaker-specific parameters, designing speaker independent systems includes many challenging difficulties. Today's speech recognition technology is still far from developing completely speaker independent high performance large vocabulary systems (Muller, 2011).

In recent years, spontaneous speech recognition systems have also gained importance. Spontaneous speech includes many false starts or incomplete sentences which are parts of natural human speech. In order to have intelligent or humanlike interactions in dialogue applications, it is important to deal with unknown words for a system (Zeng, 2009).

Environmental conditions like noise, acoustic distortions or different usage of equipments (e.g, microphones) are also other important parameters which might affect performance of the system. Even if a speech recognition system works perfectly in laboratory environment, performance of the same system might be degraded by adverse environmental conditions. Researchers have focused on robust speech recognizers to handle these difficulties in recent years.

There is also an increasing interest towards multimodal speech recognition systems. Multimodal systems use both face information, especially lip information, and speech information together to classify the utterance. The results of the studies have shown that using both types of information provide a higher recognition performance rather than using only the audio or only the visual information. Multimodal systems are especially effective in noisy environments (Furui, 2005).

CHAPTER 3

DESIGN OF SPEECH RECOGNITION SYSTEM

In this chapter, a speech recognition system is designed for the stated purpose and basic components of the speech recognition system are presented in some detail. Main parts of our speech recognition system are described in Figure 3. 1.

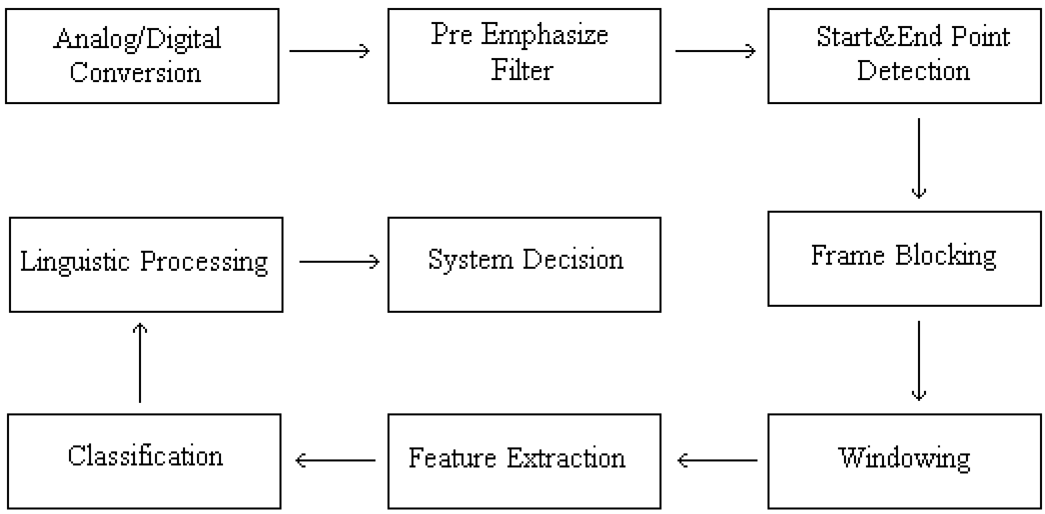

Figure 3.1 Main Steps of Speech Recognition System

3.1 Analog to Digital Conversion of the Sound

The first step of speech recognition is conversion process of input speech signal into a discrete array which includes essential information of audio signal. In this process, a sampler takes discrete instants from a continuous time speech signal. Frequency of this sampling operation is called "Sampling Frequency". The maximum recorded frequency constituent of sound signal cannot be higher than half of the sampling frequency. This rule is named as "Nyquist Rate". The Nyquist rule is formulated in equation 3.1 where F_s is sampling frequency and F_{max} is the maximum frequency constituent that can be captured.

$$F_s = 2 \times F_{max} \tag{3.1}$$

After sampling operation, captured sound information is converted into digital data. All of the conversion process is named as “Recording”.

3.2 Preprocessing of Speech

The captured sound signal is required to be processed to increase the performance of the recognition. These processes can be summarized as pre-emphasizing, word and syllable detection, frame blocking and windowing. Figure 3.2 demonstrates the main steps of preprocessing operation.

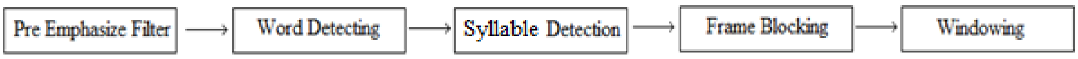

Figure 3.2 Main Steps of Preprocessing

3.2.1 Pre-Emphasize Filtering

A first order FIR filter which is also known as “Pre-Emphasize filter” is implemented to the captured audio signal. The transfer function of the filter which is represented as H(s), is given in equation 3.2.

$$H(s) = 1 - a.s^{-1} \tag{3.2}$$

In this equation “a” is a constant value and is generally taken between 0.9 and 0.95. Pre-emphasis filter offsets the natural slope of speech signal which naturally has a negative slope (Rabiner, 1978). In addition, pre-emphasis filter emphasizes the higher frequencies (above 1 KHz) of the speech signal. These advantages provide an efficient analysis for speech signal (Picone, 1993). Figure 3.3 shows a speech signal before and after implementation of pre-emphasis filter for the word “Mühendis” in Turkish which is “Engineer”.

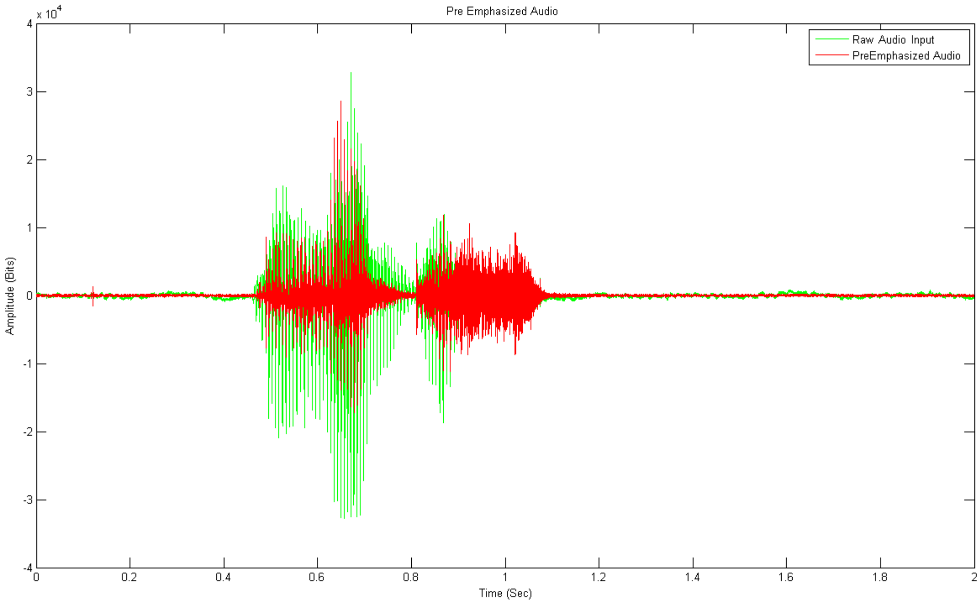

Figure 3.3 Pre Emphasize Filter output for word 'Mühendis'

3.2.2 Start and End Point Detection

Detecting the presence of speech in a background noise is an important problem in speech recognition which is often referred to as the end-point location problem (Rabiner, 1974). Detecting the boundaries of speech from non-speech parts of the sound has an essential importance for design of a reliable speech recognition system.

Developing a word based speech recognition system will not be efficient because of the agglutinative nature of the Turkish. Another popular approach in speech recognition system is using phonemes as patterns. But negative effects of smaller unit sizes of the phonemes make it difficult to recognize these patterns. While the negative effects of these approaches are considered, designing a syllable based speech recognition system was decided. The most important problem of syllable based systems is detecting syllables. To overcome this problem a syllable detection algorithm is designed. After the spoken word detected and separated from the background noise this syllable detection algorithm is applied to the spoken word.

3.2.2.1 Word Detection

After preprocessing of the sound data, a word detection algorithm is applied to the record. Aim of the algorithm is to detect the exact spoken work with minimum information lost. The system detects the spoken word signal in three main steps. Figure 3.4 demonstrates the schematic demonstration of the word detection algorithm.

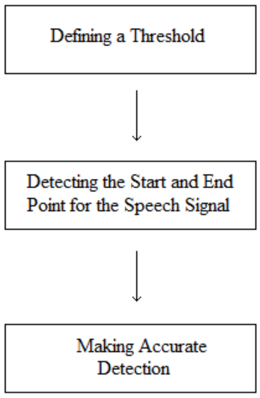

Figure 3.4 Schematic Demonstration of the Word Detection Algorithm

In the first step a threshold value defines the brink between noise and speech signal. If the first 10 ms of the sound is considered as background noise, the maximum value of the absolute of that noise can be defined as threshold. Figure 3.5 explains the detection of background noise threshold.

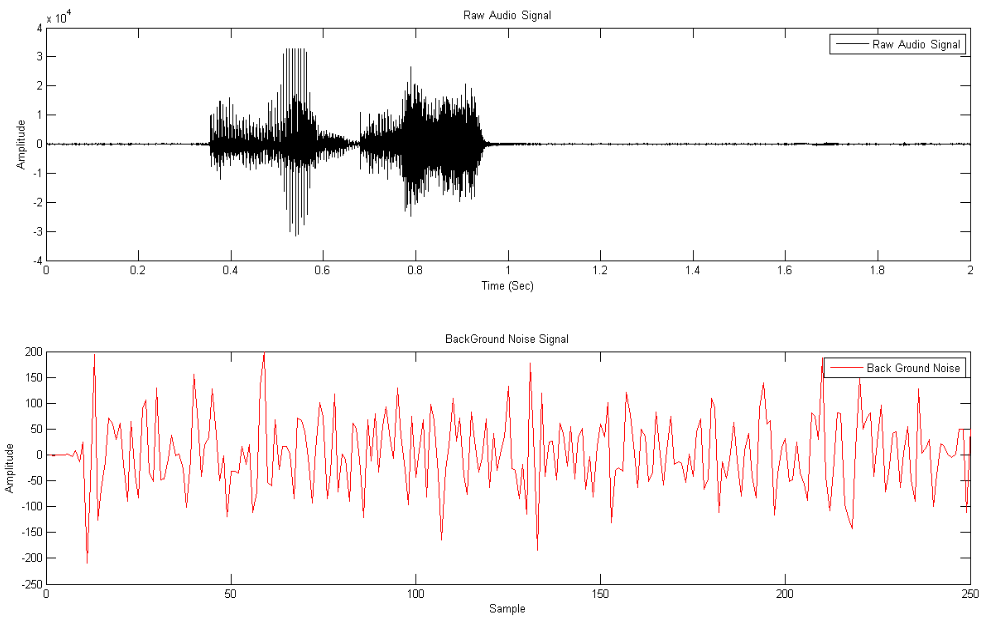

Figure 3.5 Detection of Background Noise Threshold

In second step, the samples which are below the threshold value are equalizing to zero. The beginning point and end point of the remaining signal can be defined as word boundaries. Figure 3.6 demonstrates the zero equalization process for the word “Mühendis”.

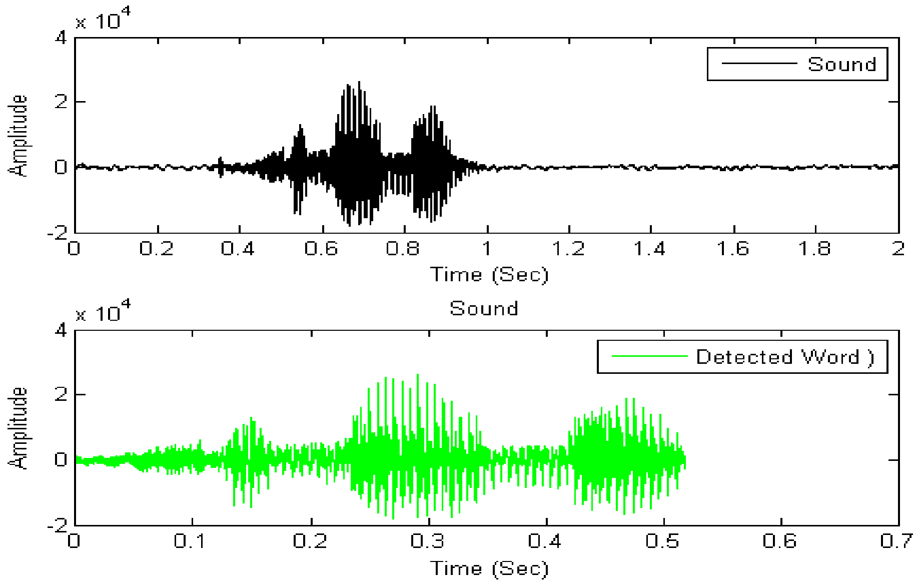

Figure 3.6 Equalizing to zero the samples below threshold value

The last step is applied to make the detection more accurate. If the signal is considered as frames the beginning and ending of the samples which consist of the boundaries defined as new beginning and ending points. Figure 3.7 demonstrates the exact word detection process.

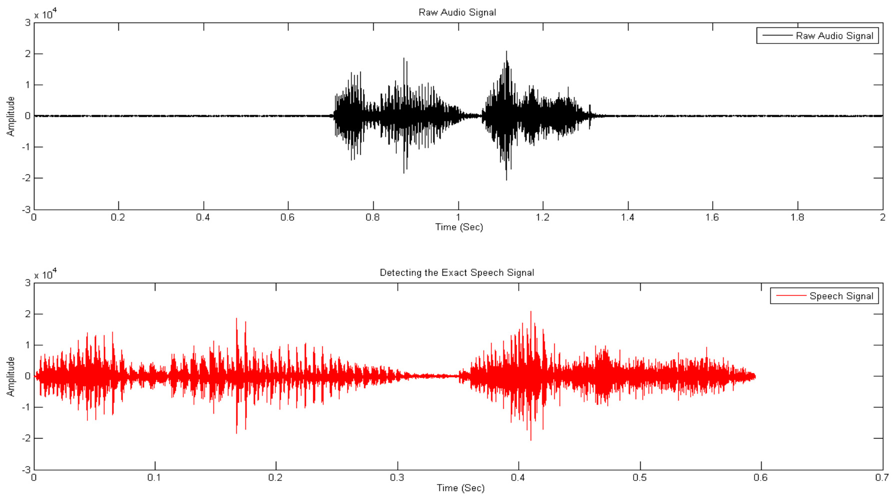

Figure 3.7 Detecting the Exact Spoken Word

3.2.2.2 Syllable Detection

After word detection process, a syllable detection algorithm is applied to the speech signal. This algorithm is designed in two main steps. The first step is Energy level analysis and second step is zero crossing analysis. These main steps are explained below;

3.2.2.2.1 Energy Level Analysis

Each syllable must consist of only one vowel according to the definition of a syllable in Turkish. This rule causes a reduction in energy of signal at each transition between syllables. So, the first step of detecting syllable boundaries defines these reductions on speech signal. The local minimum of the absolute average speech signal for each frame gives the transition point. Figure 3.8 shows the word “Depo” which means “Storage” in English. The local minimum which showed on the graph is the boundary between the syllables “De” and “Po”.

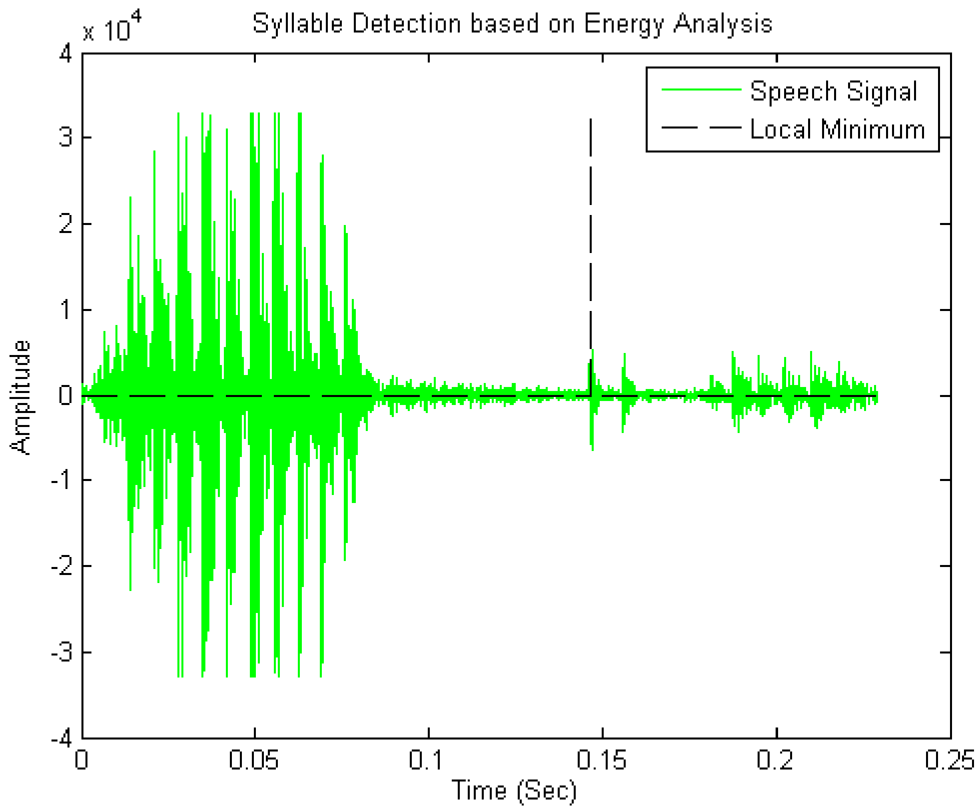

Figure 3.8 Detecting Syllable Boundary by Energy Level Analysis

3.2.2.2.2 Zero Crossing Analysis

Analyzing energy reductions of a signal might not always be enough to define syllable boundaries. There are some utterances which might be defined as a syllable by looking at their energy even if they are not. Especially the plosive consonants (like ç, k, p, t) cause these kinds of mistakes. To overcome this problem, a zero crossing analysis is applied to the signal. Each utterance is divided into 20 ms blocks. Zero crossing rates of these blocks change according to the energy level. In other words there is an inverse proportion between energy level and zero crossing rate of an utterance. So, if any of the zero crossing rates of the signal blocks is lower than a threshold value, that utterance is consumed as a constant. Else, it considered as a new syllable. Figure 3.9 shows both of the energy analysis and zero crossing analysis for the three syllable word “Mühendis” which means “Engineer” in Turkish.

In the figure 3.9 the word “Mühendis” is divided into four different utterances in energy analysis. While the first three utterances correspond to the syllable “Mü-hen-di”, and the last utterance is extension of the constant “s”. Because of the high zero crossing level of the second utterance, the utterance is assumed as a constant. Figure 3.10 gives the algorithm which defines syllable detection process.

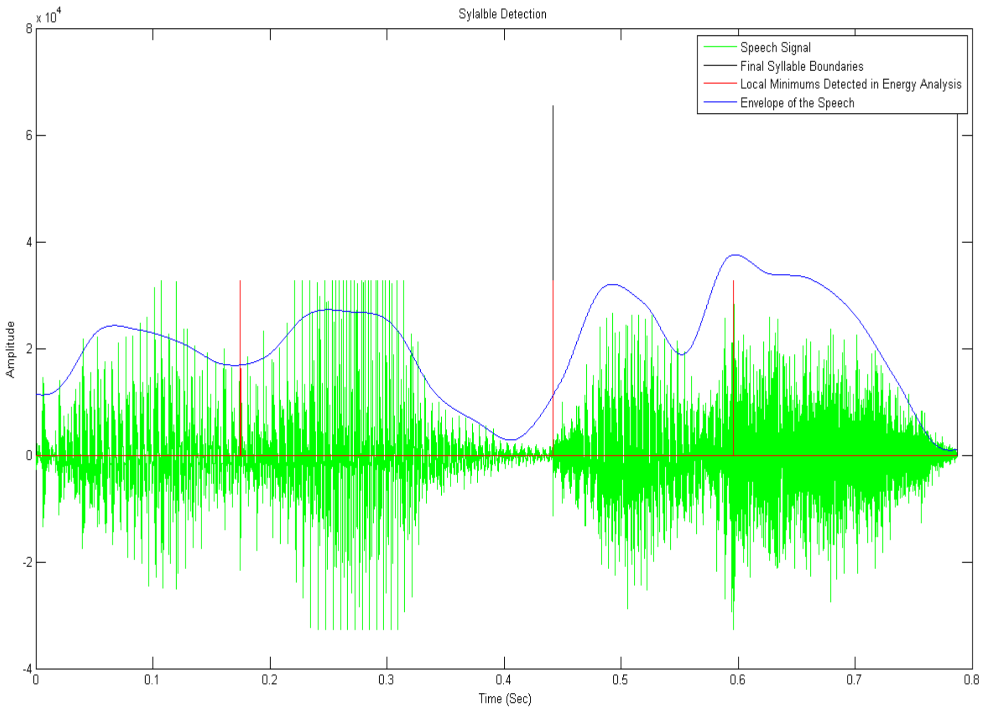

Figure 3.9 a) Demonstration of Energy Analysis of Syllable Detection Process

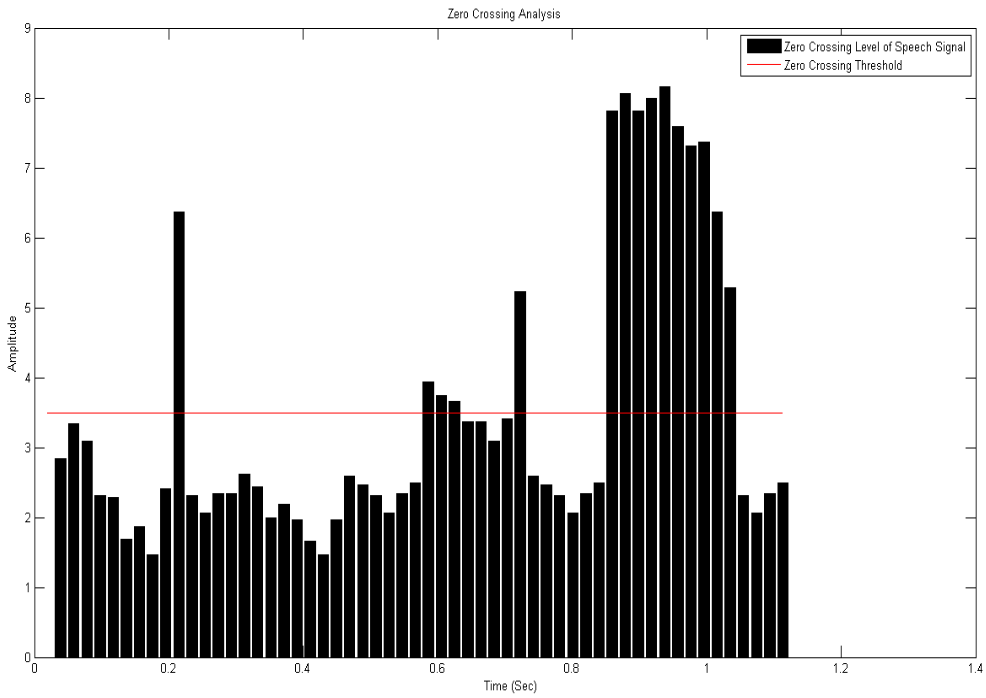

Figure 3.9 b) Demonstration of Zero Crossing Analysis of Syllable Detection Process

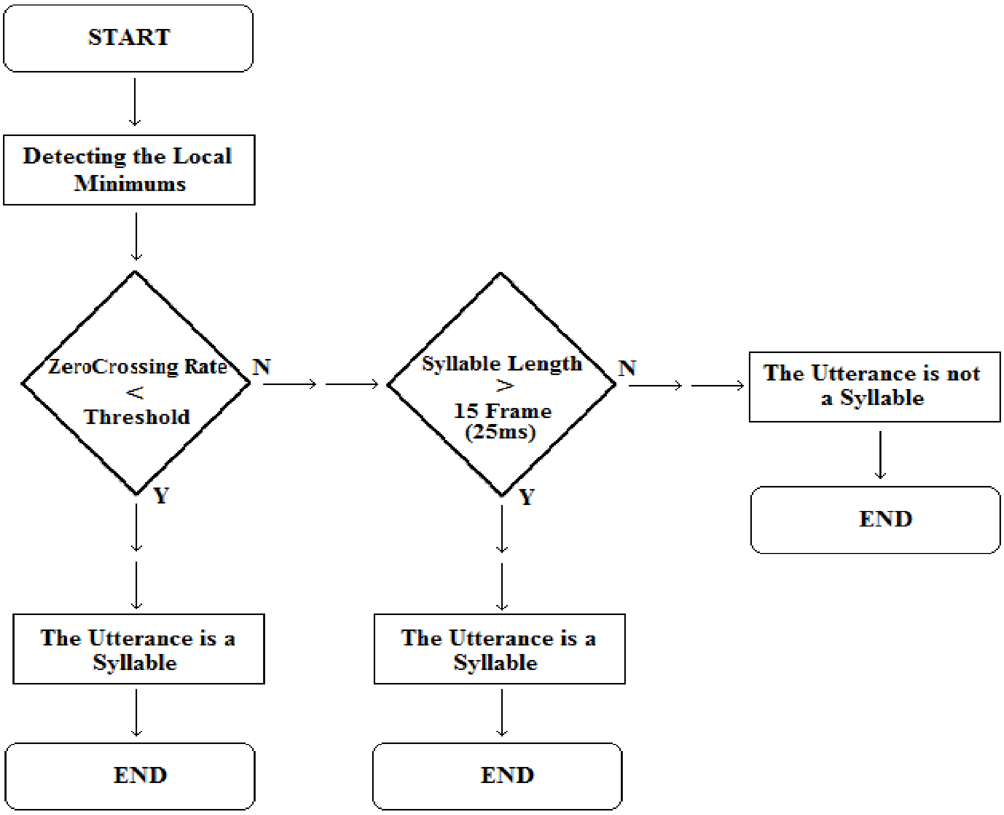

Figure 3.10 Algorithm designed for Syllable Detection

3.2.3 Frame Blocking

In this step speech signal is converted to frames of 'N' samples. The first frame is defined as 'N' sample from the beginning of the speech signal and following frames start after 'M' (M<N) sample. Thereby, the frames overlap by N-M samples. Processing small particles of signal provides us better results in analysis on feature extraction rather than processing whole signal. Figure 3.11 demonstrates framing process for a speech signal.

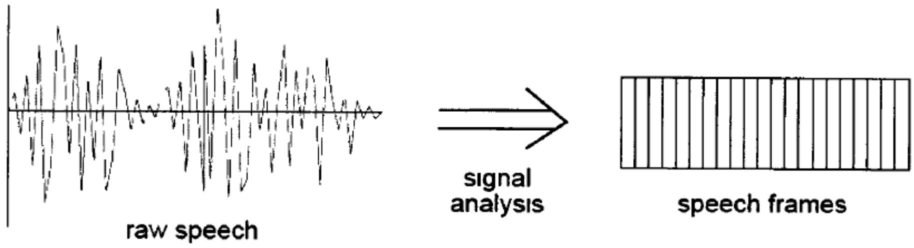

Figure 3.11 Schematic Description of the Frame Blocking Process (Savoji, 1989)

3.2.4 Windowing

After dividing the signal into frames, each frame is applied to a window function. The most common windowing functions are given below;

Rectangular Window;

$$S\ (i) = \begin{cases} 1, if\ 0 < i < N \\ 0, else \end{cases} \tag{3.3}$$

Hamming Window;

$$S\ (i) = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi\ .\ i}{N}\right), if\ 0 < i < N \\ 0, else \end{cases} \tag{3.4}$$

Hanning Window;

$$S\ (i) = \begin{cases} \frac{1}{2} - (\cos\left(\frac{2\pi\ .\ i}{N}\right), if\ 0 < i < N \\ 0\ , else \end{cases} \tag{3.5}$$

Barlett Window;

$$S(i) = \begin{cases} \frac{2i}{N} & , if\ 0 \leq i \leq \frac{N}{2} \\ 2 - \frac{2i}{N} & , if\ \frac{N}{2} \leq i \leq N \\ 0 & , else \end{cases} \tag{3.6}$$

Blackman Window;

$$S(i) = \begin{cases} 0.42 - 0.5 \cos\left(\frac{2\pi \,.\, i}{N}\right) - 0.08 \cos\left(\frac{4\pi \,.\, i}{N}\right), if\ 0 < i < N \\ 0, else \end{cases} \tag{3.7}$$

Hamming window is preferred to apply to the frames because of advantages in reducing discontinuity of speech signal. Figure 3.12 shows results of hamming windowing process for word 'Mühendis'.

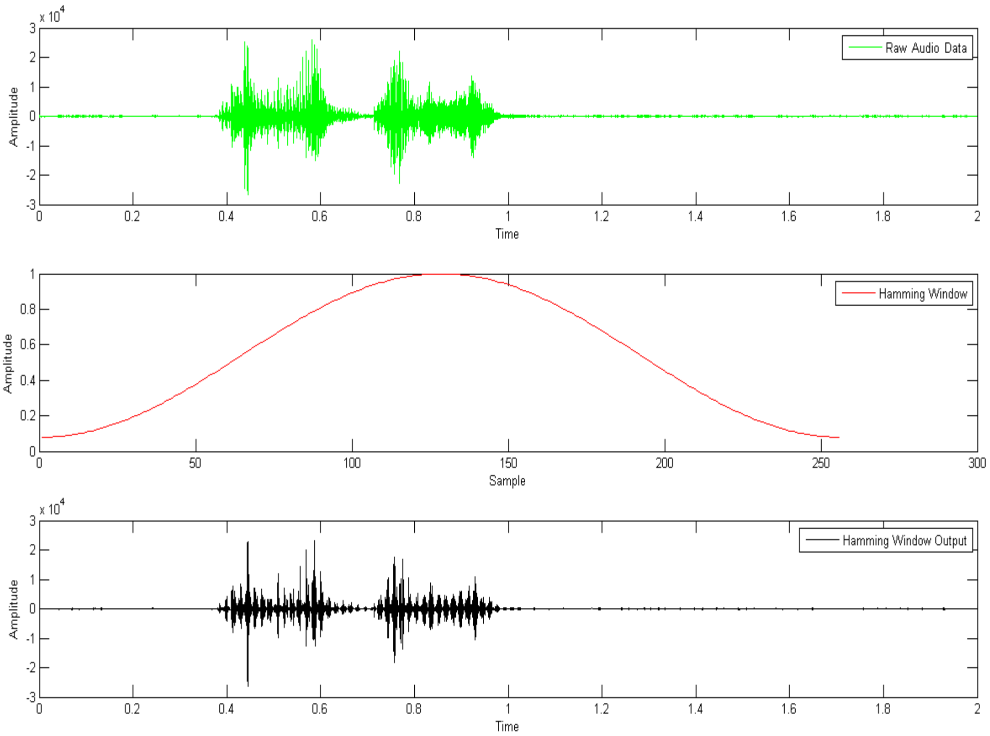

Figure 3.12 Hamming Windowing for word 'Mühendis'

3.3 Feature Extraction

Feature extraction process detects the unique characteristic information of the preprocessed utterances. With this operation, performance of classification is enhanced. Two of the most common and efficient feature extraction techniques, Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) are discussed in the next paragraph.

3.3.1 Linear Predictive Coding Coefficients

The method of LPC is a widely used technique in speech recognition, encoding and compression. In this method, an all-pole system is used to model a vocal tract which is represented in Figure 3.13.

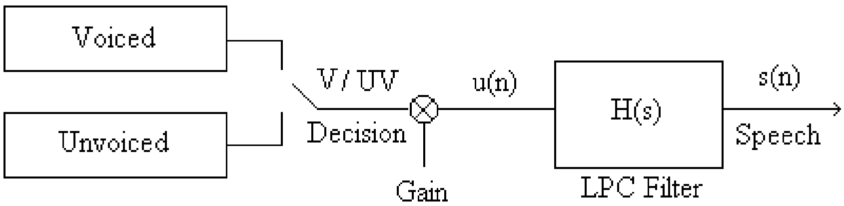

Figure 3.13 Schematic Representation of LPC Model

In this model, u(n) is the source which can be considered as periodic, as in the case of voiced utterances or noise which is represented by the case of unvoiced utterances. H(s) is an all-pole filter which is formulated in equation 3.8. In this formulation, ‘a’ coefficients are the digital filter coefficients. In other words, speech signal can be considered as a sound which is applied to a digital filter and scaled by a gain in vocal chords.

$$S(n) = -\left(a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p)\right) + u(n) \qquad (3.8)$$

Levinson- Durbin algorithm is commonly used for solving the parameters of this model (Aidazade, 2006). These coefficients provide a compressed representation of the speech signal. The frequency domain of the H(s) filter has peaks in representative frequencies which are named as formants. Applying H(s) filter to the speech encodes the key speech information (Belilovsky, 2011).

3.3.2 Mel Frequency Cepstral Coefficients

Another well known method in feature extraction is MFCC. MFCC is based on an auditory model which is very approximate to the human auditory system. Auditory mechanism in a human ear is more sensitive to the lower frequency components of the speech information rather than higher ones. This approximation is provided by Mel frequency scale. Procedure of MFCC is summarized below;

The speech utterance information which is in time domain required to be converted in frequency domain. This operation is done by Fast Fourier Transform (FFT). FFT is a fast working algorithm designed to evaluate Discrete Fourier Transform (DFT) of a signal. The FFT operation is formulated as;

$$\mathrm{Xa}(k) = \sum_{n=0}^{N-1} \mathrm{X}(\mathrm{n}) e^{-j\pi k/N} \quad , \qquad 0 \leq \mathrm{k} \leq \mathrm{N}\text{-}1 \qquad (3.9)$$

In this sequence, n=0 is called zero frequency. The interval between 1 and $\frac{N-1}{2}$ are called positive frequencies and can be placed in $0 \leq f \leq \frac{Fs}{2}$ in frequency domain while the interval between $\frac{N}{2}$ and N-1 are called as negative frequencies and can be placed in $\frac{-Fs}{2} \leq f \leq 0$.

The frequency data which are converted into frequency domain from time domain are applied to Mel-Scale. By applying this scale, an auditory model similar to the human ear is implemented.

Figure 3.14 illustrates the Mel frequency scale. In this graph, scale has linear frequency spacing below 1 kHz and logarithmic frequency spacing above 1 kHz in various intervals. The formulation to define Mel-scale is given below;

$$\mathrm{Mel}\ (f) = 2595.\log_{10} (1+\frac{f}{700}) \qquad (3.10)$$

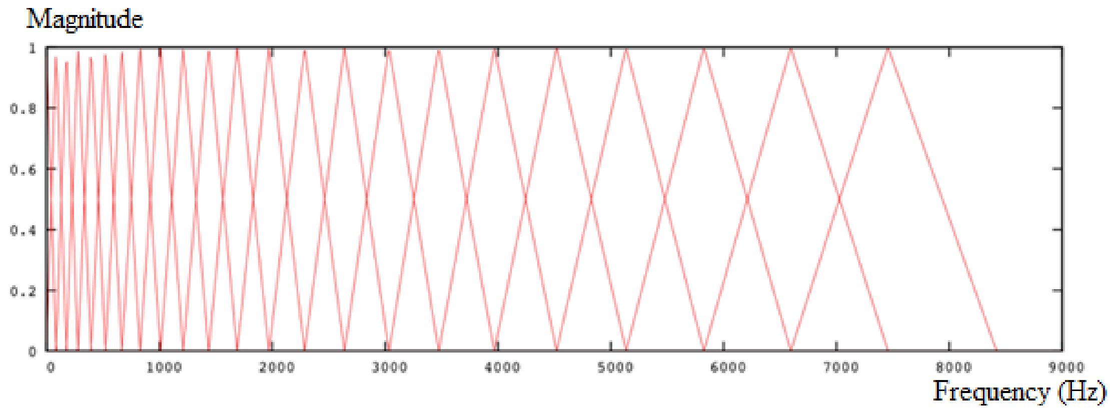

Figure 3.14 Mel Frequency Scale (Anonymous)

In the last step, scaled Mel spectrum coefficients need to be converted back into the time domain. Discrete Cosine Transform is applied to the Mel spectrum coefficients for converting back. Taking logarithm of the Mel spectrum coefficients before DCT provides more robust coefficients. The final results of this operation are called as MFCC. The equation which is used for converting back mel scale coefficients to time domain is given below;

$$\mathrm{Cn} = \sum_{k=1}^{K} (\log_{10} S_k)\, Cos\left(n\left(k - \tfrac{1}{2}\right)\tfrac{\pi}{k}\right) \tag{3.11}$$

Because of the robust structure and performance advantages, MFCC is selected as feature extraction method in this study (Furui, 1986).

3.4 Classification Methods

There are several methods which are widely used for classification of extracted feature vectors. Dynamic Time Warping, Hidden Markov Model and Artificial Neural Networks are the most common classification techniques in speech recognition problems.

3.4.1 Dynamic Time Warping

Dynamic Time Warping is a template based classification method. Even if the same person voices an utterance twice, length of the speech signal can be different. The main idea in DTW is aligning starting time and ending time of the both reference signal and input signal nonlinearly by stretching or shrinking it along its time axis. DTW method is widely used in speech recognition applications.

The distance which is represented with “D” in DTW between two time series M and N can be calculated by using a dynamical programming approach which is formulated in section 3.12. M = (X_1, $X_{2,}$ X_3... X_M) and N= (Y_1, $Y_{2,}$ Y_3... Y_N) and considered as two time series where the “X” and “Y” are elements of sequence;

$$D(i,j) = minimum \begin{cases} D(i,j-1) \\ D(i-1,j) \\ D(i-1,j-1) \end{cases} + \; d(X_i, Y_i) \qquad (3.12)$$

Calculating the distance between templates is required to determine similarities of speech templates. Classification of these templates is made according to these similarities. Distance measurement is efficiently done by Euclidian distance measurement which is formulated below;

$$d(x,y) = \sqrt{\textstyle\sum_i (X_i - Y_i)^2} \,, \qquad (3.13)$$

The template which provides minimum distance path between reference templates and input template corresponds to the best matching template.

Fast working algorithm and easy implementation are the most advantageous features of Dynamic Time Warping method. However the performance in DTW decreases while training set increases. For this reason, this model is not preferred for this system.

3.4.2 Hidden Markov Model

Hidden Markov Model has become increasingly popular during the last few decades. Very rich mathematical structure of the models and high performance results when implemented in practice are the two strong reasons of why HMM is widely used in literature.

The HMM is a finite set of states which are connected to each other with a probability transition. The HMM model is in consistence of transaction probabilities and emission probabilities. The figure 3.15 shows the main parameters of a HMM. Main parameters of HMM;

N is the number of states in the model, $\{S_1, S_2, S_3 \dots S_N\}$

M is the number of distinct observation symbols per state, $\{V_1, V_2, V_3 \dots V_M\}$

A is the transition probability from state i to state j, $\{a_{ij}\}$

B is the emission probabilities in state j, $\{b_j(V_k)\}$ where $1 \leq k \leq M$

And initial state distribution, π_i

By the given information, HMM mathematically can be represented as $\lambda = (A, B, \pi)$. (Rabiner, 1989)

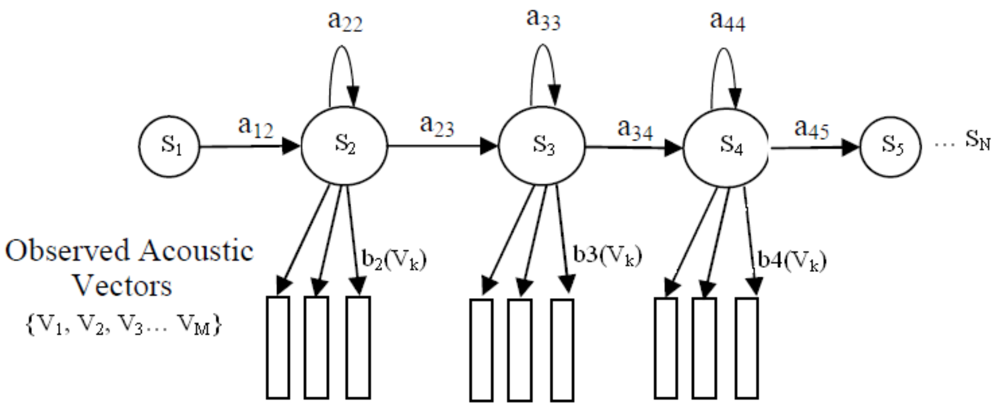

Figure 3.15 Model of HMM (Inspired from Haznedaroğlu, 2007)

In speech recognition, the observation vectors defined with the feature vectors and states can be expressed by the utterance like phoneme or syllable. Finding the hidden state sequence by using observations is the main goal of this process. For instance, in a phoneme based system, state sequence can be represented as phoneme sequences. This requires a new model definition for each phoneme. If these phoneme models are concatenated side by side, a transition between the end of a model and beginning of a next model is observed. Therefore all utterances can be defined as a Markov process.

Three basic problems must be solved before a useful real world application is developed in HMM;

i. Probability of the observation sequence for a specified HMM needs to be computed. This step is known computation.

ii. State sequence which corresponds to the given observation sequence needs to be found. This step also called as Analysis.

iii. At last, parameters of a model need to be defined to reach high probability in a sequence of specified observation. This step is named training (Can, 2007).

HMM model requires very large assumptions while defining transition probabilities, enormous number of parameters. In addition, there is some uncertainty while defining the number of states and transitions and hence these parameters are defined by trial and error or guesses. Therefore, even if HMM has one of the most wide spread usage in speech recognition, it is not preferred to be used in our speech recognition system.

3.4.3 Artificial Neural Networks

Artificial Neural Network is widely used classification method which is inspired from biological human brain. Artificial neural networks have a flexible and adaptive learning structure and show high performance in speech recognition applications, the method is preferred as classification method in this study.

ANN includes large numbers of simple processing units, called neurons, similar to the neurons in the brain. These neurons are connected to each other in various ways and generate a network. Basically, all of these artificial neurons give an output which is obtained by implementation of a transfer function to the sum of the multiplication of each input data and their weights. Figure 3.16 illustrates basic structure of an artificial neuron (Dahl, 2012).

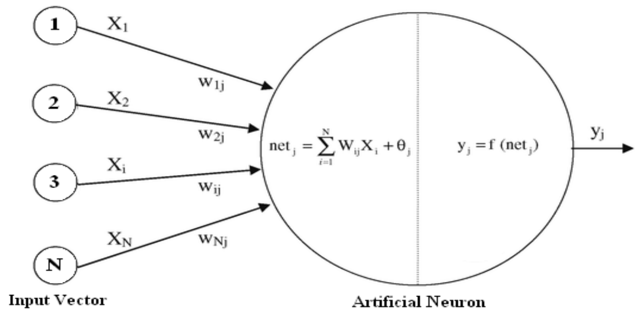

Figure 3.16 General Structure of an Artificial Neuron

This network can be examined in two parts. In the first part, the incoming information is multiplied with its weight parameter. In this figure the Nth input parameter which represented as X_N is multiplied with Nth weight of jth neuron that represented as W_{Nj}. After each input data is multiplied with its weight, summation of all multiplications is taken. There might be an additional bias function which is represented with θ_j. This process is formulated in equation 3.14.

$$Net_j = \sum_{i=1}^{N} W_{ij} X_i + \theta_j \tag{3.14}$$

In the second part, a transfer function is applied to each summation which is obtained in input part. The equation 3.15 formulates this operation.

$$y_j = f(Net_j) \tag{3.15}$$

There are various types of transfer function which might be used in a neuron. Hard limit, linear, hyperbolic tangent and sigmoid are the most popular transfer functions. Sigmoid transfer function is used in this thesis.

Sigmoid Function;

Sigmoid function has a non linear characteristic. The output of the function takes a value between 0 and 1 for any input value. Formulation of a sigmoid function is given in equation 3.14.

$$H(s) = \frac{1}{1+e^{-s}} \tag{3.16}$$

This function is highly preferred in dealing with non linear problems. Figure 3.17 illustrates a graph of a sigmoid function.

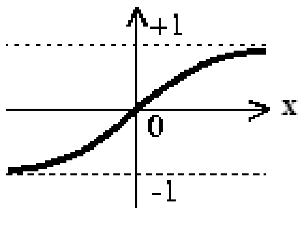

Figure 3.17 Sigmoid Transfer Function

Even if there are many different types of ANN, there are basic features which are shared by all ANN models. Firstly, all neural networks consist of a set of processing units which are referred as layers. There is a set of connections between these processing units. Computational procedures are performed by each layer of network. At last, a training procedure is required for all artificial neural networks. Basically, Artificial Neural Networks are considered in two basic models;

Feed Forward Neural Networks

In feed forward neural networks there are basically three different layer types. Input layer, hidden layer and output layer. In this model, information is propagated from input layer to output layer. Input layer is where the information enters the system, hidden layer is where the learning process is done, and the output is the layer that gives the results of the system. Feed Forward Neural Networks shows high performance in pattern recognition applications so these networks are used very popularly in speech recognition systems. Figure 3.18 shows the general structure for the feed forward neural networks.

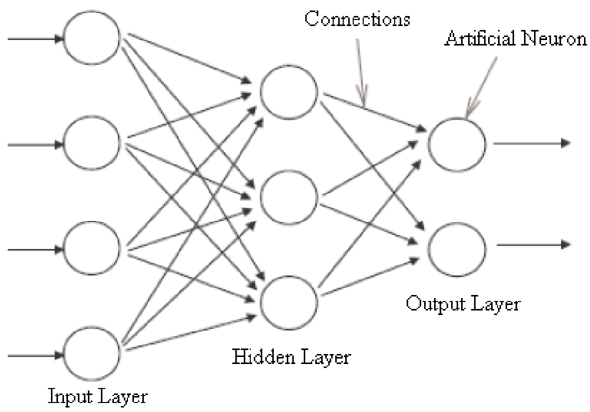

Figure 3.18 Structure of a Feed Forward Neural Network

Input layer defines our raw information so number of neurons in this layer is defined according to input data. In a speech recognition application, input layer can be considered as a feature vector thereby, number of the neurons in this layer should be the same with the number of coefficients in the feature vector of speech. Number of neurons in hidden layer is highly dependent to the application. Neural number in output layer is defined according to the number of possible classes for the input pattern. In speech recognition process number of neurons in an output layer is equal to the possible classes of the utterances to be recognized.

Recurrent Neural Networks

In recurrent neural networks there is a feedback connection between output and input. Recurrent neural networks have ability to keep information in its internal memory for a defined period. There are many different kinds of recurrent neural network, Hopfield, Elman, Jordan networks are the most common of them all. Figure 3.19 illustrates a general structure of recurrent neural networks.

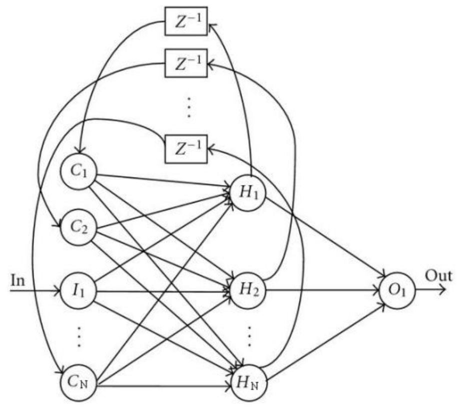

Figure 3.19 General Structure of Recurrent Neural Networks

Training of Artificial Neural Networks

In training of neural networks, the weights are arranged for taking desired output from given input. ANN’s can be trained in an iterative period because of their nonlinear and layered structure. The comparison of the signals in output and input layers gives an error rate defined as ε in each iteration. The aim of the training process is minimizing ε by changing weights. When ε minimized to desired value, the weights are kept and training is complete. This period in which weights are changed is named learning. The value of ε has a critical importance for training period. If ε is chosen too small, the training process might take infinite time. In contrast, if defined ε is too big, information looses is occurred.

Back Propagation Learning Algorithm

Back-Propagation learning algorithm is a widely used algorithm in training of feed forward neural networks. The main idea in this algorithm is that weights of the network are changed in each iteration until the error rate is reduced to the desired value. The error rate is tried to reduce in an inverse sequence that is why this algorithm named back- propagation.

3.5 Linguistic Model

After classification process is completed a linguistic process is required for determining the spoken utterance. Classification methods are generating probabilities according to the similarities between reference set of information and spoken input information. Linguistic models are used for processing these probabilities to determine the final output of the recognition system. Although language models differentiate it depends on the recognized language highly, N-gram language models have a wide spread usage in language processing applications. This process can also be named as post processing in speech recognition. Figure 3.20 explains the language model designed within the scope of this thesis

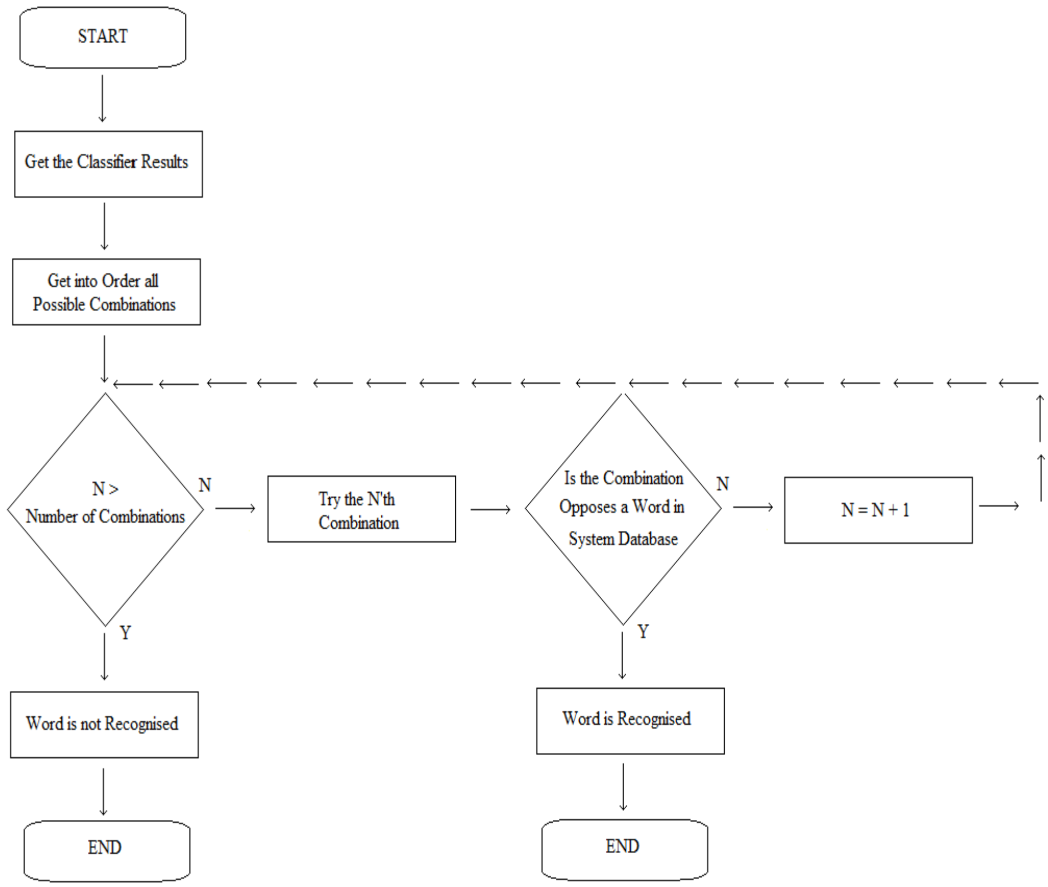

Figure 3.20 Schematic Demonstration of Algorithm for Post Processing

The language model uses the output of the classifier. The output vector is put in an order and each possibility is evaluated to determine the spoken word. For instance, the output of the neural network for a three syllable word is assumed as below;

Table 3.1 Word "Gömülü" is determined by analyzing the assumed possibilities

0.91 Gö	0.12 lü	0.99 lü
0.21 Bö	0.10 rü	0.01 ri
0.01 Ro	0.01 mü	0.00 -

The language model considers each combination for the given syllables and determines that any of the generated combinations corresponding to a word which is known by the system. In the table given above, the system examines each combination and gives the word "Gömülü" as an output. Figure 4.11 explains the language model algorithm.

CHAPTER 4

EXPERIMENTS & RESULTS

In this study, an isolated word syllable based speech recognition system is designed and tested limited to 50 words in the Turkish language. In the application MFCC is used for feature extraction and Artificial Neural Networks is chosen as the classification method. Designed system is tested on both trained and untrained speakers. This system is designed to be used in Guide Robot, Guard Robot, Humanoid Robot projects in Atılım University, Mechatronics Engineering Department. In this chapter, the system is described in details.

4.1 Hardware and Software

The system is tested on a computer which has; 2.3 Intel i5 CPU, 4 GB Ram, 250 GB Hard Disk, Windows XP Operating System, a soundcard and a headset microphone are used to design the program. The algorithms are implemented and tested on MATLAB R2012a software.

4.2 Database of the System

System dictionary consists of 50 different Turkish words. Two different databases are created for the study. The first database is for training and the second database is created for testing the system. The speech data is collected from 5 people. The total size of the database capacity is 100 MB which may store approximately 2 hours of speech data. Speakers spelled out each word in the database 20 times. 10 samples are used for the training and 10 samples are used for testing.

4.3 Data Acquisition and Preprocessing

A headset microphone is used for recording process. The utterances are recorded in a WAVE file format. The waveforms are encoded with Pulse Code Modulation (PCM) coding format. Sampling frequency of the system is defined as 16 KHz (16384Hz) and 16 Bit per sample. Each utterance is recorded in 2 seconds time duration.

After the utterances are captured, a first order FIR which is also known as Pre Emphasize filter is applied to the sound data. This process emphasizes the higher frequencies of the speech. The transfer function of the Pre Emphasize filter which is represented as H(s) is formulated below

$$H(s) = 1 - 0.95.s\text{-}1 \quad (4.1)$$

The figure 4.1 shows the pre emphasize process for word “Ses” which means “Voice”.

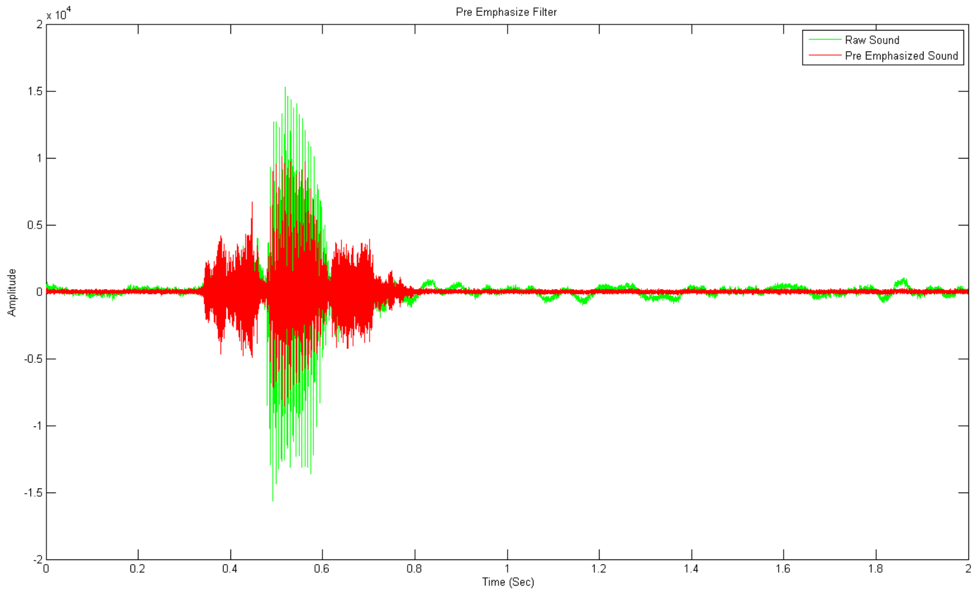

Figure 4.1 Pre Emphasize Filter output for word “Ses”

4.4 Start and End Point Detection

Separating the speech from background noise is an important problem in speech recognition. This problem is defined as “end point detection” (Rabinner, 1975). The reason for requiring an effective end-point algorithm is that the computation for processing the speech would be minimum when the end-points are accurately located (Savoji, 1989).

4.4.1 Word Detection

After preprocessing the sound data, a word detection algorithm is applied to the record. Aim of the algorithm is to detect the exact spoken work with minimum information lost. A speech data of 2 seconds, which is recorded in 16 KHz, includes 32768 samples. The first 250 samples of this data, approximately corresponds to 8 ms, which can be considered as background noise. Maximum value of this signal can be considered as noise level threshold of the environment. This sound level is used for detection of the spoken part. If the amplitude of the signal is under the threshold value, that sample can be considered as noise and equalized to zero. Finally, the spoken word is clearly separated from the background noise. Figure 4.2 illustrates the word detection process for the word “Sekiz” which means eight in English.

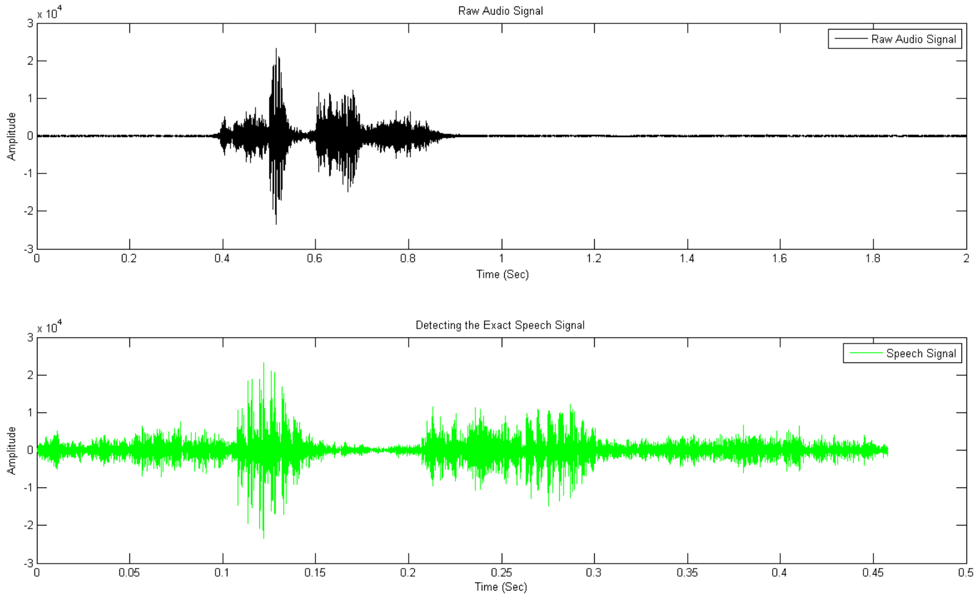

Figure 4.2 Word Detection Process for the Word “Sekiz”

4.4.2 Syllable Detection Algorithm

After word detection process, a syllable detection algorithm is applied to the speech signal. This algorithm is designed in two main steps. The first step is Energy level analysis and second step is zero crossing analysis.

Figure 4.3 and Figure 4.4 explains the syllable detection algorithm on two one syllable words “Üç” and “Beş” which respectively mean three, five. The words are consumed as two syllable word according to changes in their energy. The defined syllables in first step are “Ü-Ç” and “Be-Ş”. When these utterances are examined by their zero crossing rates, the algorithm determines that they are not a syllable. This is because the zero crossing rates of the second syllable of these words exceed the threshold value.

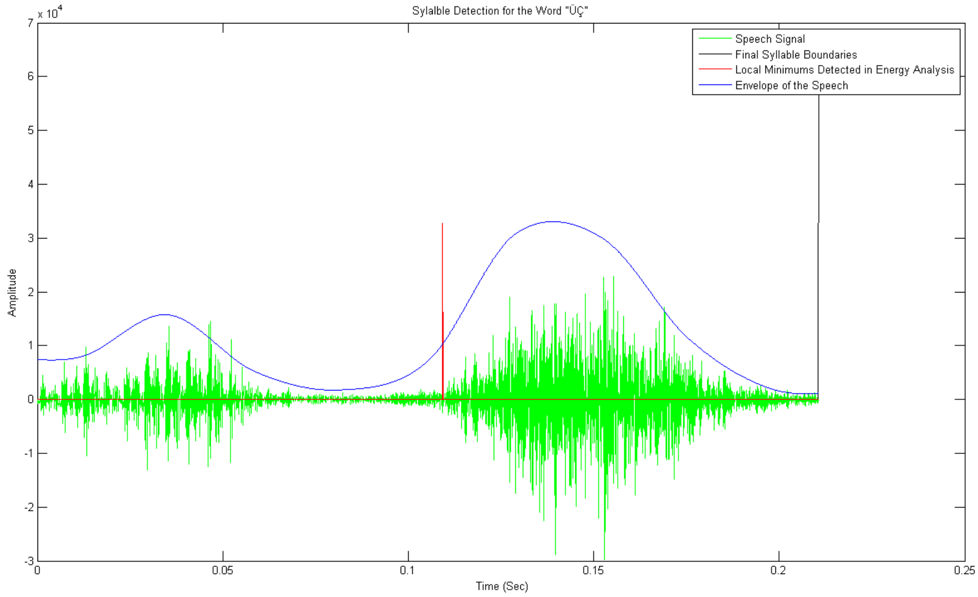

Figure 4.4 a) Explanation of the Energy Analysis Part of Syllable Detection Algorithm on the Word "Üç"

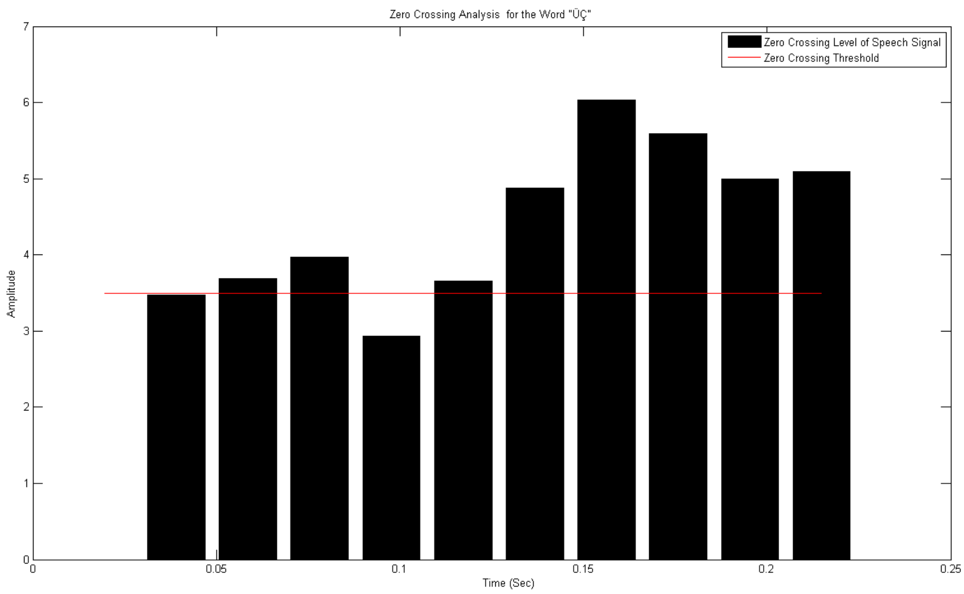

Figure 4.3 b) Explanation of Zero Crossing Analysis Part of Syllable Detection Algorithm on the Word "Üç"

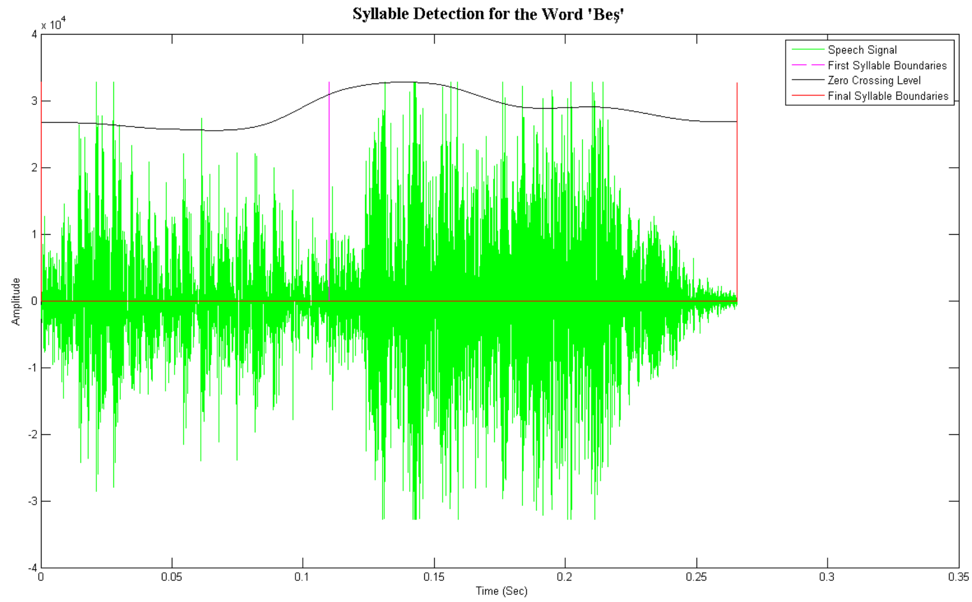

Figure 4.4 a) Explanation of the Energy Analysis Part of Syllable Detection Algorithm on the Word “Beş”

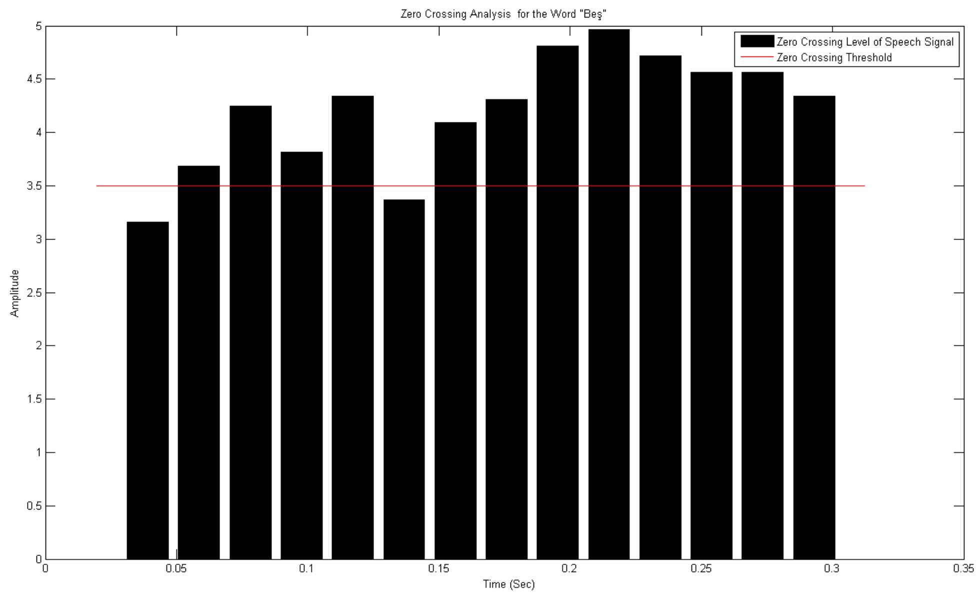

Figure 4.4 b) Explanation of Zero Crossing Analysis Part of Syllable Detection Algorithm on the Word “Beş”

Figure 4.5 and 4.6 explains the syllable detection algorithm on two one syllable words “Altı” and “On” which respectively mean six and ten. The word “Altı” is naturally two syllable word and separated into two syllables in energy level analysis. The defined syllables of this word are below the threshold zero crossing rate value so algorithm approves these utterances as different syllables. The word “On” is naturally a one syllable word. The algorithm determines that the utterance is a one syllable word after both energy level and zero crossing analysis.

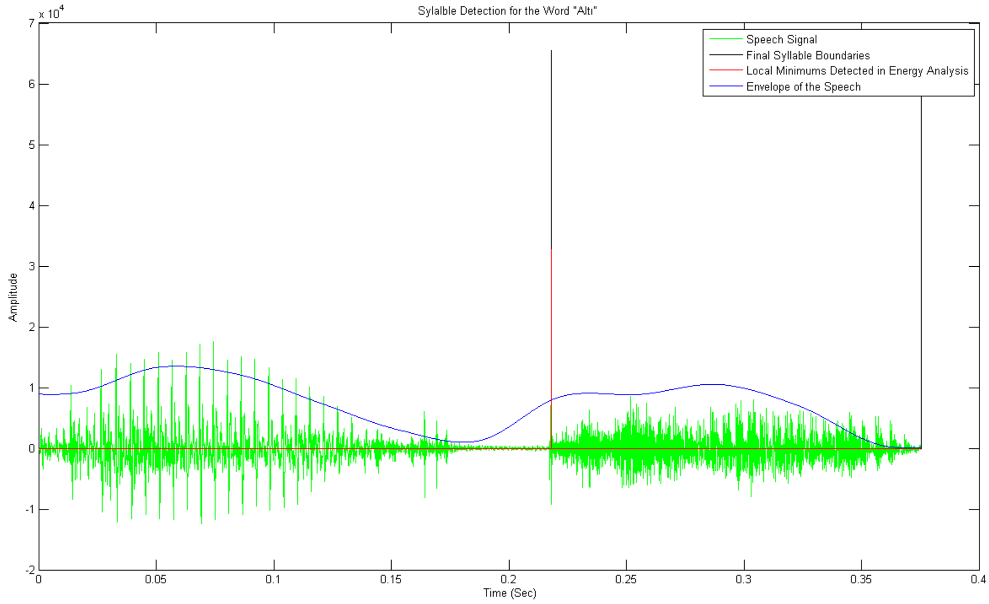

Figure 4.5 a) Explanation of the Energy Analysis Part of Syllable Detection Algorithm on the Word “Altı”

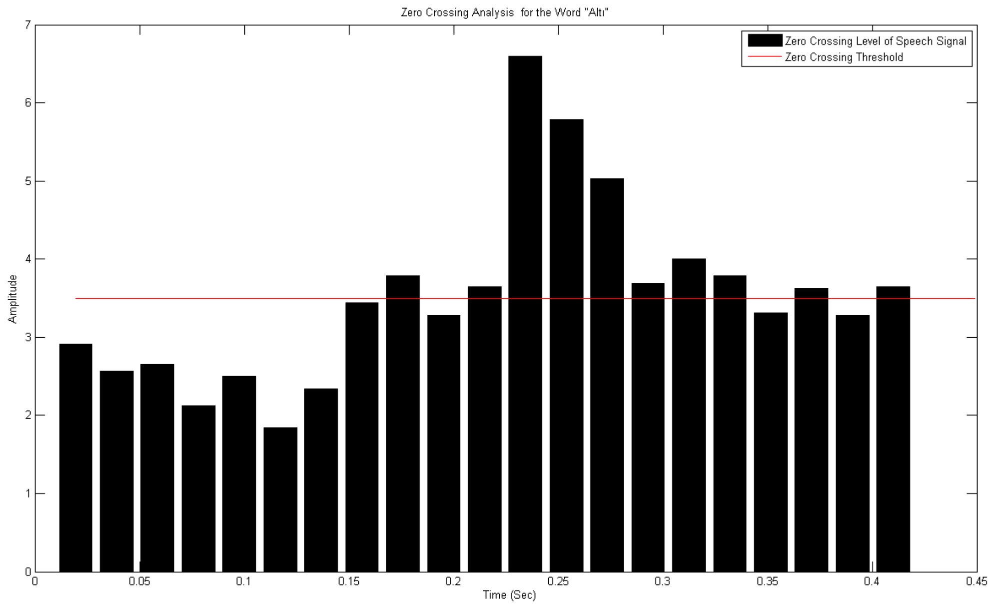

Figure 4.5 b) Explanation of Zero Crossing Analysis Part of Syllable Detection Algorithm on the Word “Altı”

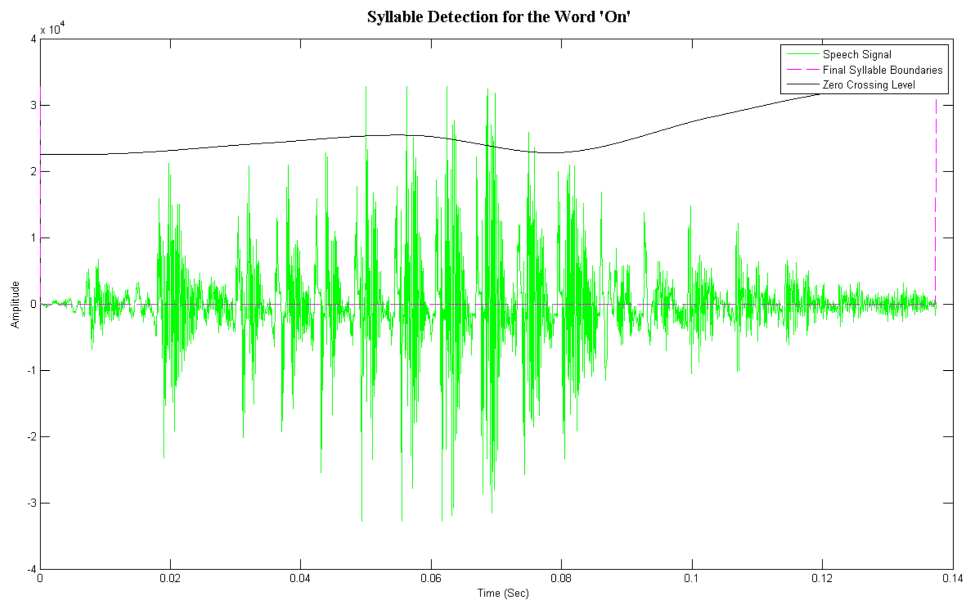

Figure 4.6 a) Explanation of the Energy Analysis Part of Syllable Detection Algorithm on the Word “On”

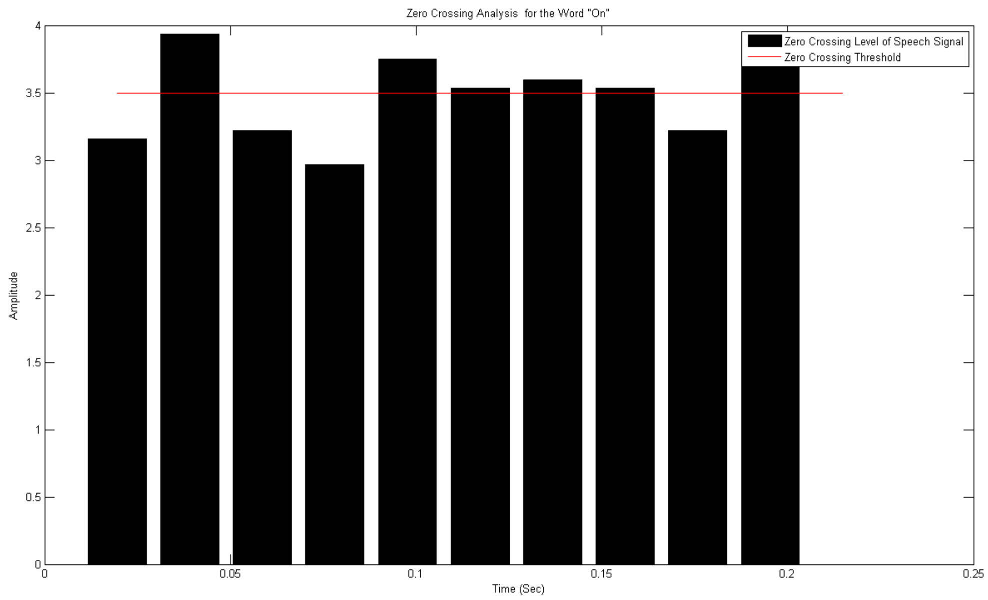

Figure 4.6 b) Explanation of Zero Crossing Analysis Part of Syllable Detection Algorithm on the Word “On”

4.4.2.1 Limitations of the Syllable Detection Algorithm

The designed algorithm shows an acceptable performance in detecting syllables. However, there are limitations of algorithm. These main limitations of the syllable detection algorithm are given below:

Requires an Accurate Pronunciation: The words are required to be pronounced accurately and users should emphasize the syllables clearly. The designed algorithm is not appropriate for the spontaneous speech. The word “Proje” is an instance for this situation.

The Performance is not high for the words are not Turkish in its origin: The systems show a low performance in detecting the words which are adapted to Turkish from other languages.

Turkish Vowel Harmony is Highly Effective on the Performance: The algorithm has a low detecting rate for the words which are not appropriate to either major or minor Turkish vowel harmony.

High Zero Crossing Rate of “i”: The vowel “i” generally has a high zero crossing rate. This reduces the performance of the algorithm in syllables which consist of both the vowel “i” and other constants.

4.5 Frame Blocking and Windowing

Characteristic features of the detected syllables need to be extracted for recognition of speech. Frame blocking and windowing processes are important steps in analysis these characteristic features. As a result of performance analysis in the experiments, size of each frame is defined as 256 samples. This value is approximately 15 milliseconds for a speech data which recorded in 16 kHz.

Hamming window is preferred for windowing operations because of the advantages in speech recognition performance. To overcome discontinuity problems, an overlap is applied to the signal. The overlap ration is selected as 50% in this application. Figure 4.7 represents the frame blocking and windowing operation with 50% overlapping for the word “Akşehir”.

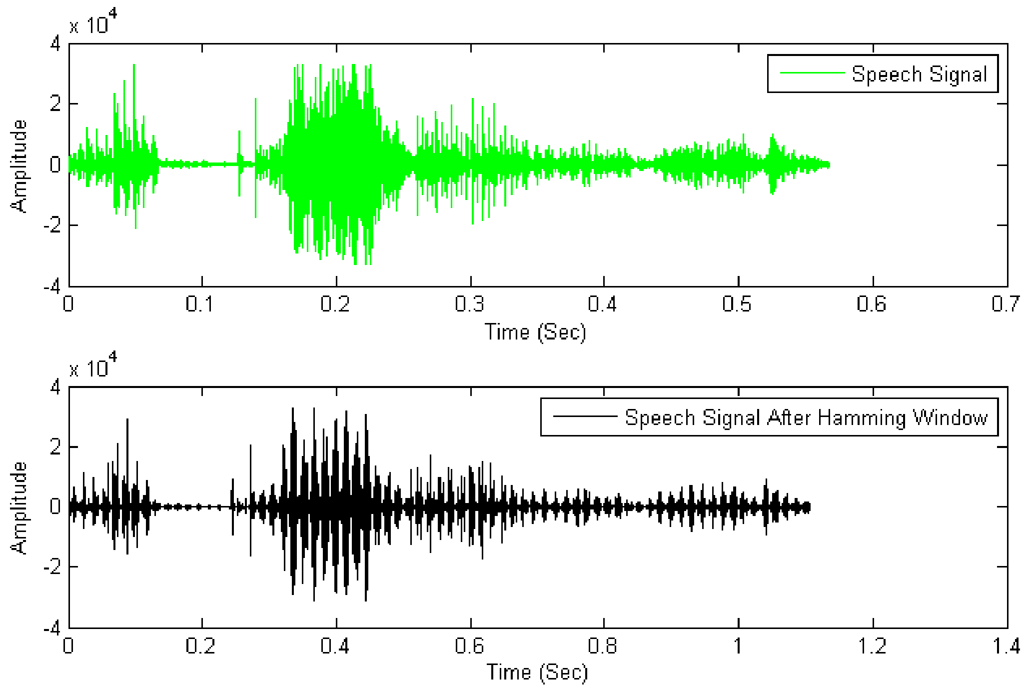

Figure 4.7 Implementation of Hamming Window to the Word “Akşehir” with 50% Overlap

4.6 Feature Extraction

MFCC is a feature extraction method inspired from the human ear which is more sensitive to the lower frequencies. Because of human like auditory mechanism, MFCC provides high performance in speech recognition. For this reason, mel-frequency cepstrum coefficients are selected as speech features. First 13 coefficients are used for recognition operation.

4.7 Classification

4.7.1 Artificial Neural Network

Artificial Neural Network is a widely used classification method which is developed by inspiring human brain. Artificial Neural networks offer a number of advantages, including requiring less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables, ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms [Tu, 1996].

Sigmoid function is selected as transfer function which is applied to the sum of each input of a neuron to generate new output. Back propagation neural networks require to be trained before classifying patterns is widely used in classification operations. In this study back propagation network is used to classify speech patterns.

4.7.1.1 Zero Matching

In neural networks, each input must be defined as a vector and all of these vectors must be the same length. Because of the uncertain nature of the speech, length of the speech signal can be different even if the same person's voice utters twice. The length of this signals needs to be equalized before generating the neural network. Zero matching method is used to satisfy this condition.

In this method, length of the maximum length vector is chosen as a reference value, and lengths of the other vectors equalizing by adding zeros to their ends. For instance, if A, B, C and D are considered as input vectors which have respectively 6, 7, 8 and 9 elements. Figure 4.8 describes zero matching operation for defined vectors.

$$
\begin{array}{cccc}
A_1 & B_1 & C_1 & D_1 \\
A_2 & B_2 & C_2 & D_2 \\
A_3 & B_3 & C_3 & D_3 \\
A_4 & B_4 & C_4 & D_4 \\
A_5 & B_5 & C_5 & D_5 \\
A_6 & B_6 & C_6 & D_6 \\
0 & B_7 & C_7 & D_7 \\
0 & 0 & C_8 & D_8 \\
0 & 0 & 0 & D_9
\end{array}
$$

Figure 4.8 Zero matching operation for input vectors A, B and C

In our system MFCC features are rearranged as one dimensional vector and considered as input of the system. The neurons in the hidden layer differentiate depending on the training parameters. At last, the number of the neurons in the output layer is defined according to the syllable variety in each training set. Each neuron addresses to a syllable. Generally, output of the network is a one dimensional vector in which each element indicates a probability between 0 and 1. The highest probability indicates that the pattern addressed by the neuron is the most similar pattern to the input signal. Figure 4.9 shows the neural network parameters for designed speech recognition system.

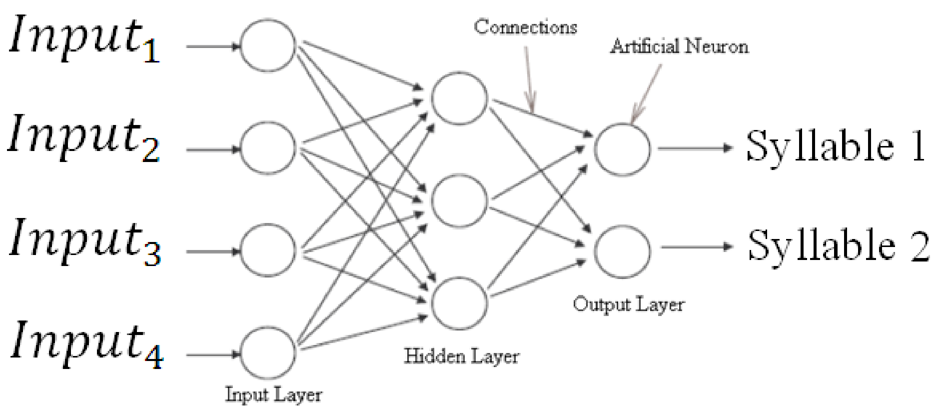

Figure 4.9 Schematic demonstration of ANN

4.8 Language Model

After the classification is made by ANN, the results are processed by the language model. The language model uses the output of the classifier. The output vector is put in an order and each possibility is evaluated to determine the spoken word. Figure 4.10 gives an assumed output of the neural network for a three syllable word.

0.0	0.0	0.01
0.0	0.02	0.0
0.01	0.0	0.0
0.09	0.0	0.0
0.99	0.07	0.35
0.0	0.0	0.0
0.0	0.79	0.0
0.0	0.0	0.0
0.0	0.0	0.0
0.45	0.0	0.0
0.0	0.0	0.0
0.59	0.0	0.28
0.0	0.0	0.0

Figure 4.10 Assumption on output possibilities at the of Neural Network

The linguistic model arranges the possible syllables and decides the spoken word according to highest possibilities. Table 4.1 shows the syllables which correspond to the possibilities given and demonstrate the word detection process.

Table 4.1 Demonstration of Word Detecting Process

0.99-Ro	0.79-Ler	0.35-Lar
0.59-A	0.07-Bot	0.28-Lı
0.45-Kim	0.02	0.01-Ci
0.09-Gö	0	0
0.01Sis	0	

4.9 Experimental Results

The designed speech recognition system is tested by both trained and non-trained speakers for 50[2] words. The words have different syllable rounds numbers between 1 and 7. Each word is spelled by 20 times by 5 speakers. 10 of them are used for training and 10 of them are used for testing the system. The system is trained and tested for 4 people in speaker dependent case. After that, the system is tested for one non trained speaker in the speaker independent case. At the end, all evaluations are repeated for a word based system. The test results are given in Table 4.2. The best results are obtained in seven syllable words and the worst results are achieved in three syllable words.

4.10 Limitations of the System

The main limitations of the general speech recognition system are explained as follows:

High Background Noise: The recognition performance of the system decreases as the background noise increases. That is the reason why getting records in quiet environments will provide better results.

System is designed for Only One Person: The system cannot recognize speech of multiple people at the same time. Only one user should record for the defined time interval.

[2] A complete list of words is given in Appendix B

Isolated Word: The system is designed as an isolated word system. For this reason, the user must pause between words for a defined time period.

Table 4.2 Test Results of the System

Syllable Number	Speaker Dependent Case		Speaker Independent Case	
Model Type	Syllable Based	Word Based	Syllable Based	Word Based
One Syllable Words	95	95	82.5	82.5
Two Syllable Words	85.5	80	78.182	72.3
Three Syllable Words	86.2	77	71.6	65
Four Syllable Words	88.34	81	80	71
Five Syllable Words	86	74	76	61
Six Syllable Words	100	80	80	64
Seven Syllable Words	100	67	93	60
Total	90	79	80	68

CHAPTER 5

DISCUSSION & CONCLUSION

5.1 Discussion

In this study, a syllable based speech recognition system is designed and implemented. The system consists of data acquisition, pre-processing, feature extraction, classification and linguistic processing parts.

A headset microphone is used for data acquisition process. 16 KHz is selected as sampling frequency and the samples are coded in 16 bit with linear PCM. MATLAB software is used for audio acquisition process.

After the speech is captured, a pre emphasize filter is applied to accentuate high frequency components of speech. A word detection algorithm is applied to the filtered sound signal to separate speech from the background noise. A word detection algorithm is used to sense accurately, the start and end points of the word. After this step, a syllable detection algorithm is used to detect syllable boundaries. Syllables are detected with a combination of energy analysis and zero crossing analysis.

Each syllable blocked to 256 sample frames and hamming window is applied with 50% overlap. Mel Frequency Cepstral Coefficients are calculated for these syllables and the first 13 coefficients are selected to generate feature vectors.

This feature vectors are classified by Artificial Neural Networks. Two layer feed forward neural networks are used for classification. Back propagation algorithm is selected to train the networks. Sigmoid function is preferred for activation function in neurons. These types of network structures perform well in pattern recognition application.

A language model is designed to determine spoken words in Turkish by using results of the networks. The possible combinations are examined by the algorithm and the best possible choice is defined as spoken word.

5.2 Conclusion

Speech recognition systems have become increasingly significant as a research area in recent years. These systems can be implemented in various areas like voice messages, guide or assistance systems or applications where user needs to be hand-free. The speech recognition system will be part of humanoid robot project at Atılım University.

The major objective of the thesis is to design and implement of speech recognition system in humanoid systems laboratory of the Department of Mechatronics Engineering. An isolated word syllable based system is preferred to reach this objective. After the speech is divided into syllable accurately, feature vectors extracted by using MFCC method. Feed Forward Neural Networks is used to classify speech patterns. In the last step, a language model is applied to the results obtained from neural networks.

The recognition performance of the system is tested on different users in case of both trained users and non trained users. 20 samples collected from different users for 50 words. In trained user case 10 of them are used for training the system and 10 of them are used for testing the system. In non-trained user case 20 samples are used to test the system which is trained for other people. In the end, approximately 85% performance is obtained in trained case and approximately 75% performance is obtained in non-trained case. The main reason of getting low recognizing performance in the test of non-trained users is the inadequate number of samples.

5.3 Future Work

The design, implementation and test of the speech recognition system are completed. On the other hand, system has some future work for improvements and implementation on humanoid robot project.

In that manner;

- Sample database of the system might be extended to increase performance.
- The performance of the syllable detection algorithm might be increased.

- A hybrid acoustic model like ANN/HMM, might be used to increase classification performance of the system.
- The syllables in the Turkish language might be examined deeply and an extended Turkish word model might be designed which includes all syllables in Turkish.
- This system will be integrated to humanoid robot or narrower applications so redesign of this algorithm on an embedded system might be considered. Digital Signal Processors and Field Effect Gate Arrays are appropriate for this object.

REFERENCES

Aidazade, K., Ardil, C., Rustamov, S., 2006, Investigation, World Academy of Science, Engineering and Technology.

Arısoy, E. and Dutağacı, H, 2006, "A Unified Language Model for Large Vocabulary Continuous Speech Recognition of Turkish", Signal Processing 86(10), 2844-2862.

Arısoy, E., 2009, "Statistical and Discriminative Language Modeling for Turkish Large Vocabulary Continuous Speech Recognition", Ph. D. Thesis, Boğaziçi University, İstanbul.

Artuner, H., 1994, "The Design and Implementation of a Turkish Speech Phoneme Clustering System", Ph. D. Thesis, Hacettepe University, Ankara.

Aşlıyan, R., 2008, "Design and Implementation of Turkish Speech Recognition Engine", Ph. D. Thesis, Dokuz Eylül University, İzmir.

Avcı, E., 2008, "An Automatic System for Turkish Word Recognition Using Discrete Wavelet Neural Network Based on Adaptive Entropy", Arabian Journal for Science and Engineering, 32, 239-250.

Belilovsky, E., 2011, "Convolutional Neural Networks for Speaker Independent Speech Recognition", M. S. Thesis, The Cooper Union for the Advancement of Science and Art.

Bridle, J. S. and Brown, 1979, M. D. “Connected Word Recognition Using Whole Word Templates”, Proc. Inst. Acoust. Autumn Conf., pp. 25-28.

Can, B., 2007, “A Syllable-Based Speech Recognition System Design and Implementation for Turkish Language”, M. S. Thesis, Hacettepe University, Ankara.

Chou, W., 2003, “Minimum Classification Error (MCE) Approach in Pattern Recognition, W. Chou and B. H. Juang (Eds.)”, Pattern Recognition in Speech and Language Processing, CRC Press, pp. 1-49.

Chow, Y. L., 1987, et. al. “BYBLOS, the BBN Continuous Speech Recognition System”, Proc. ICASSP, pp. 89-92.

Dahl, E. G., Yu, Dong., Deng, Li., Acero, A., 2012, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition”, IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, No. 1.

Davis, K. H. et. al., 1952, “Automatic Recognition of Spoken Digits”, J.A.S.A., 24 (6), pp. 637-642.

Ferguson, J. Ed., 1980, “Hidden Markov Models for Speech”, IDA, Princeton, NJ.

Forgie, J. W. and Forgie, C. D., 1959, “Results Obtained from a Vowel Recognition Computer Program”, J.A.S.A., 31 (11), pp. 1480 1489.

Fry, D. B., 1959, “Theoretical Aspects of Mechanical Speech Recognition”; and P. Denes, “The Design and Operation of the Mechanical Speech Recognizer at University College London”, J. British Inst. Radio Engr., 19, 4, pp. 211-229.

Furui, S., 2005, “50 Years of Progress in Speech and Speaker Recognition Research”, Ecti Transactions on Computer and Information Technology Vol. 1, No.2 November.

Furui, S., 1986, “Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum”, IEEE Trans. Acoustics, Speech, Signal Processing, ASSP-34, pp. 52-59.

Haznedaroğlu, A., 2007, “Domain Specific Keyword Spotting”, M. S. Thesis, Boğaziçi University, İstanbul.

Itakura, F., 1975, “Minimum Prediction Residual Applied to Speech Recognition”, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-23 (1), pp. 67-72.

Jelinek, F. et. al., 1975, “Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech”, IEEE Trans. Information Theory, IT-21, pp. 250-256.

Jelinek, F., 1985, “The development of an experimental discrete dictation recognizer”, Proc. IEEE, 73 (11), pp. 1616-1624.

Jelinek, F., 1998, “Statistical Methods for Speech Recognition”, Cambridge: MIT Press.

Karaca, N., 1999, “Realization of a Turkish Isolated Word Speech Recognition System Under Noisy Environments”, Ph. D. Thesis. Hacettepe University, Ankara.

Karakas, M., 2010, “Computer Based System Control Using Voice Input”, M.S Thesis, Dokuz Eylül University, İzmir.

Kawahara, T. et al., 1998, “Key-Phrase Detection and Verification for Flexible Speech Understanding”, IEEE Trans. Speech and Audio Proc., 6, 6, pp. 558-568.

Klatt, D., 1977, “Review of the ARPA Speech Understanding Project”, J.A.S.A., 62(6), pp. 1324-1366.

Koç, A., 2002, “Acoustic Feature Analysis for Robust Speech Recognition”, M. S. Thesis. Boğaziçi University, İstanbul.

Lee, C. H. and Rabiner, L. R., 1989, “A Frame Synchronous Network Search Algorithm for Connected Word Recognition”, IEEE Trans. Acoustics, Speech, Signal Proc., 37 (11), pp. 1649-1658, November.

Lee, C. H. et. al., 1990, “Acoustic Modeling for Large Vocabulary Speech Recognition”, Computer Speech and Language, 4, pp. 127-165.

Lee, K. F. et. al., 1990, “An Overview of the SPHINX Speech Recognition System”, Proc. ICASSP, 38, pp. 600-610.

Lee, C. H., 2001, “Statistical Confidence Measures and Their Applications”, Proc. ICSP, pp. 1021-1028.

Leggetter, C. J. and Woodland, P. C., 1995, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, Computer Speech and Language, 9, pp. 171-185.

Liu, Y. et. al., 2005, “Structural Metadata Research in the EARS Program”, Proc. ICASSP, V-957.

Lowerre, B., 1990, “The HARPY Speech Understanding System”, Trends in Speech Recognition, W. Lea, Ed., Speech Science Pub., pp. 576-586.

Martin, T. B. et. al., 1964, “Speech Recognition by Feature Abstraction Techniques”, Tech. Report ALTDR- 64-176, Air Force Avionics Lab.

Mengüşoğlu, E., 1999, “Rule Based Design and Implementation of a Speech Recognition System for Turkish Language”, M. S. Thesis, Hacettepe University, Ankara.

Meral, O., 1999, “Speech Recognition Based on Pattern Comparison Techniques”. M. S. Thesis, İstanbul Technical University, İstanbul.

Muller, F., Mertins, A., 2011, “Contextual Invariant-Integration Features for Improved Speaker-Independent Speech Recognition”, Speech Communication, Volume: 53, Issue: 6, Pages: 830-841.

Nagata, K., Kato, Y. and Chiba, S., 1963, “Spoken Digit Recognizer for Japanese Language”, NEC Res. Develop., No. 6.

Olson, H. F. and Belar, H., 1956, “Phonetic Typewriter”, J.A.S.A., 28 (6), pp. 1072-1081.

Özkan, Ö.,1997, “Implementation of Speech Recognition for Connected Numerals”, M. S. Thesis, Middle East Technical University, Ankara.

Paul, D. B., 1989, “The Lincoln Robust Continuous Speech Recognizer”, Proc. ICASSP, pp. 449-452.

Picone, J. W., 1993, "Signal Modeling Technique in Speech Recognition", Proc. Of the IEEE, vol. 81, no.9, pp.1215-1247.

Rabiner, L. R. and Sambur, M. R., 1974, “An Algorithm for Determining the Endpoints of Isolated Utterances”, The Bell System Technical Journal, Vol:54, No:2.

Rabiner, L. R., and Schafer, R. W., 1978 “Digital Processing of Speech Signals. Englewood Cliffs”, New Jersey: Prentice-Hall.

Rabiner, L. R., 1989, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proc. IEEE, 77 (2), pp. 257-286.

Rabiner, L. R. and Juang, B. H., 1993, “Fundamentals of Speech Recognition”, Prentice-Hall, Englewood Cliff, New Jersey.

Rabiner, L. R. et. al., 1979, “Speaker Independent Recognition of Isolated Words Using Clustering Techniques”, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27, pp. 336-349.

Reddy, D. R., 1966, “An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave”, Tech. Report No. C549, Computer Science Dept., Stanford University.

Sakai J. and Doshita, S., 1962, “The Phonetic Typewriter, Information Processing”, 1962, Proc. IFIP Congress, Munich.

Sakoe, H. and Chiba, S., 1978, “Dynamic Programming Algorithm Optimization for Spoken Word Recognition”, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26 (1), pp. 43-49.

Sakoe, H., 1979, “Two level DP Matching - a Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition”, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27, pp. 588-595.

Salor, Ö and Pellom, B. L., 2007, “Turkish Speech Corpora and Recognition Tools Developed by Porting SONIC: Towards Multilingual Speech Recognition”, Computer Speech and Language. 21(4), 580-583.

Savoji, M.H., “End-Pointing of Speech Signals”. Speech Communication, 8(1), 46-60.

Suzuki, J. and Nakata, K., 1961, “Recognition of Japanese Vowels—Preliminary to the Recognition of Speech”, J. Radio Res. Lab, Vol. 37, No. 8, pp. 193-212.

Tappert, C. C. et. al., 1971, “Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recovery”, Rome Air Dev. Cen, Rome, NY, Tech. Report TR-71-146.

Tebelskis, J., 1995, “Speech Recognition Using Neural Networks”, M. S. Thesis, Carnegie Mellon University, Pittsburgh.

Tu, J. V., 1996, “Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes”, Elsevier, Volume 49, Issue 11, Pages 1225-1231.

Tunalı, V., 2005, “A Speaker Dependent, Large Vocabulary, Isolated Word Speech Recognition System for Turkish”, M.S. Thesis. Marmara University, Istanbul.

Tunca, A., 2010, “Digit Sequence Recognition Using Hidden Markov Models and Continuous Speech Recognition Technique”, M.S. Thesis, Osmangazi University, Eskişehir.

Varga, A. P. and Moore, R. K., 1990, “Hidden Markov Model Decomposition of Speech and Noise”, Proc. ICASSP, pp. 845-848.

Velichko, V. M. and Zagoruyko, N. G., 1970, “Automatic Recognition of 200 Words”, Int. J. Man- Machine Studies, 2, pp. 223.

Vintsyuk, T. K., 1968, “Speech Discrimination by Dynamic Programming”, Kibernetika, 4 (2), pp. 81- 88.

Viterbi, A. J., 1967, “Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm”, IEEE Trans. Information Theory, IT- 13, pp. 260-269.

Weibel, A. et. al., 1989, “Phoneme Recognition Using Time-Delay Neural Networks”, IEEE Trans. Acoustics, Speech, Signal Proc., 37, pp. 393-404.

Weintraub, M.et al., 1989, “Linguistic Constraints in Hidden Markov Model Based Speech Recognition”, Proc. ICASSP, pp. 699-702.

Yılmaz, C., 1999, “A Large Vocabulary Speech Recognition System for Turkish”, M. S. Thesis, Bilkent University, Ankara.

Zeng, Z., Pantic, M., Roisman, G. and Huang, T. S., 2009, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions”, IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 31, NO. 1.

Zue, V. et. Al., 1989, “The MIT Summit Speech Recognition System, a Progress Report”, Proc. DARPA Speech and Natural Language Workshop, pp. 179- 189.

APPENDIX A

ABBREVIATIONS

The abbreviations used in this thesis is listed and explained briefly below;

ANN (Artificial Neural Networks): Artificial Neural Network is a soft simulation of biological human brain. The Artificial Neural Networks recognize the patterns which they learnt before.

ASR (Automatic Speech Recognition): The ARS defines the systems that convert human speech to text.

DFT (Discrete Fourier Transform): DFT defines discrete time analysis of Fourier Transform.

DTW (Dynamical Time Warping): A template based classification method. DTW measures the similarities of two patterns which might vary in time.

DWNN (Discrete Wavelet Neural Networks): Special form of artificial neural networks.

FFT (Fast Fourier Transform): An algorithm to evaluate fourier transform of a signal.

FIR (Finite Impulse Response): A filter whose response is form of its current and finite number of past values.

FSN (Finite State Network): A network which has a single start and various final states.

HMM (Hidden Markov Model): A stochastic classification method.

LPC (Linear Predictive Coding): A feature extraction method which is popularly used in speech recognition, encoding and compression.

LSD (Linguistic Statistical Decoder): Statistical language processing process.

LTA (Linear Time Alignment): The process of matching two signals in time domain.

MCE (Minimizing Classification Error): The process of reducing error rate which reasons from classification.

MFCC (Mel Frequencies Cepstral Coefficients): A feature extraction method which uses specific mel frequency scale.

PARCOR: A feature extraction method.

PCM (Pulse Code Modulation): Modulation method used to digitally embody sampled analog signals.

RASTA-PLP: A feature extraction method used in speech analysis.

SVM (Support Vector Machine): A method which is used to analyze data and recognize patterns.

TDNN (Time Delay Neural Networks): A version of the Neural Networks which is effective on continuous time data.

WAVE: Audio file format which mainly used for raw and uncompressed audio.

APPENDIX B

COMPLETE LIST OF WORDS

One Syllable Words

A
B
C
E
G
BİR
ÜÇ
DÖRT
BEŞ

Two Syllable Words

SIFIR
İKİ
ALTI
YEDİ
SEKİZ
DOKUZ
DEPO
UÇAN
SÜRÜ
GÖRÜ
ROBOT

Three Syllable Words

ÖĞRENCİ
NEREDE
SİSTEMLER
GÖMÜLÜ
ROBOTLAR
ASANSÖR
NELERDİR
ROBOTSAL
DAVRANIŞ
KİMLERDİR
ROBOTİK

AKILLI
KONULAR

Four Syllable Words

ARAŞTIRMA
NEREDELER
YÜRÜTÜLEN
OTOMASYON
SİSTEMLERİ
ÇALIŞMALAR

Five Syllable Words

BÖLÜMÜNÜZDE
GÖREVLİLERİ
ARAŞTIRMALAR
ARAŞTIRILAN
ÇALIŞMALARI

Six Syllable Words

LABORATUVARI
BÖLÜMÜNÜZDEKİ
GÖREVLİLERİN

Seven Syllable Words

LABORATUVARLARDA
LABORATUVARINDA
LABORATUVARLARI