T.C.

ALTINBAS UNIVERSITY

Institute of Graduate Studies

Electric and Computer Engineering

ISOLATED WORD DETECTION USING LSTM BASED FEATURE EXTRACTION METHODS

Hasan Hameed Hussein AL-BAYATI

Master's Thesis

Supervisor

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Istanbul, 2021

ISOLATED WORD DETECTION USING LSTM BASED FEATURE EXTRACTION METHODS

by

Hasan Hameed Hussein Al-BAYATI

Electric and Computer Engineering

Submitted to the Institute of Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAS UNIVERSITY

2021

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

Hasan Hameed Hussein Al-BAYATI

DEDICATION

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents and my supervisor Asst. Prof. Dr. Abdullahi Abdu IBRAHIM's guidance and insight made this work possible.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Asst. Prof. Dr. Abdullahi Abdu IBRAHIM for the continuous support of my Msc study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

ABSTRACT

ISOLATED WORD DETECTION USING LSTM BASED FEATURE EXTRACTION METHODS

Hasan Hameed Hussein AL-BAYATI

M.Sc., Electric and Computer Engineering, Altınbas University,

Supervisor: Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Date: June, 2021

Pages: 57

In this study, pitch applied as feature extraction techniques and combined with RNN which these two methods are new approaches which PSO applied as RNN trainer. The Pitch based RNN presented remarkable results, which extracted features by pitch, wired to RNN and classified to seven words and presented 97.12% accuracy. The aim of applying PSO is to optimize the RNN and which find best weights and basis of the model. The presented framework presented best results than previous researches in the field of speech recognition.

Keyword: RNN, Speech Recognition, PSO, Pitch.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

LIST OF ABBREVIATIONS

SVM	:	Support vector machine
KNN	:	K-nearest neighbours algorithm
NN	:	Neural Network
RBF	:	Radial Basis Function
CNN	:	Convolutional Neural Network
DT	:	Decision Tree

1. INTRODUCTION

Speech is the dominant means of communication in humans, and while it promises to be important for human-machine communication, it can hardly be more reliable. Speech recognition converts an audible signal into a number of devices. Use words include voice commands and controls, data entry, voice user interface, operator call automation, and so on. They also serve as inputs for natural language processing. In some situations, the speech recognition problem that has consolidated over the years is an extremely concentrated computational problem. You need free computers and lots of storage. Accordingly, various attempts have been made to speed up the process using various methods. As a result, the use of speech recognition systems (SRS), which use integrated circuits (DSPs) for digital signal processing, is becoming more attractive and beautiful. The compromise for the DSP application is the technology's increased rate of expansion, long-term reliability, noise protection and the ability to perform complex, absurd calculations in the analog domain [1].

In recent years, the amount of SR software in consumer electronics products has increased rapidly. One of the greatest challenges for an integrated software developer is comparative computing power and memory limits, which must be measured early in the design phase. Hence, the superiority between the different methods will likely depend on the correspondence between cost and performance [2].

The pioneering work for ASR began in the early 1950s. The first ASR system developed at Bell Telephone Laboratories could detect different numbers from 0 to 9 for a single speaker. In 1956, Olson and Pillar created a sound generator that can recognize ten different syllables. He also needed to trust the speaker and need serious training.

This initial pattern-based recognition by ASR is based on a pattern match when the amplifier inputs have been compared to previously stored patterns or sound patterns. Pattern matching works well at the word level to identify different elements in a small vocabulary, but is less effective at identifying a larger vocabulary. Another limitation of the pattern matching is that incoming speech signals cannot be compared and compared with previously stored audio samples of different lengths. As a result, these ASR systems performed poorly as they used

acoustic approaches that only identified the main vocal units clearly expressed by the speaker [3].

Figure 1.1: Simple speech recognition system [4].

Before the machine can analyze the voice, the microphone must convert the vibration of the human voice into an electrical reflux signal. This signal is converted to a serial digital signal using system equipment such as a computer sound card. A digital signal that a speech recognition program analyzes to recognize another phonetic component of speech. Then phonemes are combined with words. However, many words sound the same, so the program must be context sensitive to select the correct word. Many programs customize text using trigram analysis, which is based on three databases of common word groups. This method assigns the probability that the third word will appear after two words. For example, if the speaker says "who am I?" However, sometimes human intervention is required to correct the error.

Single word recognition software, such as a telephone voice navigation system, such as a telephone voice navigation system, works for almost any user. On the other hand, continuous speech programs such as dictation programs must be trained to recognize patterns of human language. During training, the user reads a sample text. The improved performance of personal computers and mobile devices has greatly improved the accuracy of speech recognition today.

With a vocabulary of tens of thousands of words, the error rate dropped to about 5%. Much more precision can be achieved with a limited vocabulary for special purposes such as imaging.

Isolated word recognition (IWR) is the step of automatically selecting and classifying features of speech waveforms using electronic circuitry and computers. Isolated telephones are particularly characterized and have been used with automatic IWR technology for many years, including computers [5]. This study shows several contributions including: Developing an Efficient IWR System The main objective of this study is to develop an IWR system. Is it combined with RNN properties of extraction methods like MFCC, pitch, energy, etc.? This combination is an innovative technology and the new results announced can be applied to a wide variety of IWR applications. In this study, we applied three methods of voice activity detection (VAD) to a digital data set. Two of these three VAD methods are new and have never used VAD activities before. The purpose of using these three techniques is to show the difference between these methods and to find out which of these methods is most effective and which type of sound is most effective. Finally, a new speech recognition system was presented in our work that combines new technologies such as Bowman and Bartlett Han functions in the VAD method. Feature extraction methods such as MFCC, pitch and energy are also used in RNN for IWR.

The study is divided into 5 chapters. Introduction Related research materials and methods. Experimental results. Therefore, the first chapter contains introductions, contributions, research questions and the structure of the thesis. This chapter provides general information about this study and shows the purpose and contribution of the proposed method. Additionally, Chapter 2 explains in detail related studies and studies to assist readers who are previous studies, as well as methods and methodologies that were applied to previous studies. In addition, the MATLAB script developed in Chapter 3 “Materials and Methods for Method Development” is described and presented. In addition, the experimental results are presented and described in Chapter 4, and some of the technology related results such as VAD results and LWR results are described. Then, in Chapter 5, the conclusions and results of this study will be presented in future studies.

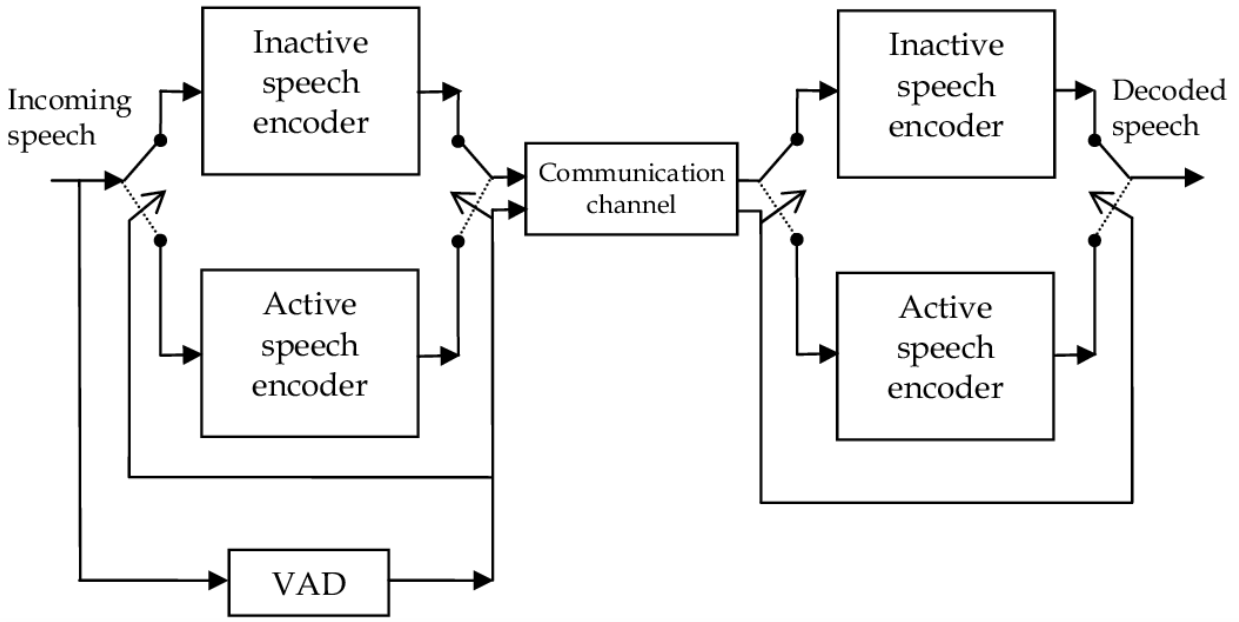

Figure 1.2: Voice activity detection (VAD) [5]

1.1 ADVANTAGES AND DISADVANTAGES OF SPEECH RECOGNITION SYSTEMS

In this section, we will present the main advantages of disadvantages of the speech recognition systems. The systems consist from number of disadvantages as listed below:

a. Configuration—some clarifications don't simply incorporate with present EHRs

b. Hardware—specific computers, medicines and additional devices make speech recognition a experiment

c. Customization—contingent on the answer used, customization may be problematic

d. Training—minus decent training, it container be firm for clinicians to accept new technology.

On the other hand, several advantages of the speech recognition systems are listed as shown below:

a. Solves Inefficiencies and Reduces Wasted Time

In the Yale Medical study, doctors used a speech recognition solution. With this software, doctors can use their voice to finish and finish appointments faster than ever before. The widespread adoption of speech recognition systems in recent years has increased efficiency and time savings in many medical reports.

According to a 2018 KLAS report, the adoption of speech recognition has improved overall productivity. Clinicians can reduce reporting time by using voice recognition tools that allow them to focus more on their patients.

b.Clinics and Hospitals Can Save Money

Many companies believe that speech recognition systems can help reduce costs. Speech recognition reduces overtime handwriting and stops outsourcing your doctor's instructions. It's more efficient for everyone and saves money at the same time.

c. Clinician Satisfaction

Doctors tend to be more satisfied with their jobs when they can complete tasks in a short period of time. They spend less time sitting at a desk writing paperwork and more time with a patient. Plus, because you don't do paperwork at night or on weekends, you can focus on your life outside of work.

1.2 THE AIM OF THIS STUDY

The aim of this study, is to developed new speech recognition system based deep learning techniques.

a. Efficient speech recognition is created. A very natural human-machine interface is provided. Nature means intuitive and easy to use, it doesn't require any special tools or machines, it just takes the power of nature that we all have. This system can be used by anyone who can speak and use large machines, especially computers.

b. This study presents a new precision technique that can be used for various HMI applications. In this study, it is one of the natural procedures for human-machine

communication that will lead to a huge leap forward in various fields such as industrial, medical and social projects.

1.3 THE THESIS STRUCTURE

In this section the structure of the thesis presented and all componts are explained:

a. Chapter 1: contain introduction which explain the speech recognition systems and definition to speech recognition systems. Furthermore, advantage and disadvantage of the speech recognition systems. Finally, the aim of the study and the quations of the thesis presented in detail form.

b. Chapter 2: this chapter includes lireture review which explain several studies presented in this field. Furthermore, several topics such as sound production and several function dealing with speech are presented.

c. Chapter 3: In this section the techniques that used in this study presented such as digital image processing, audio processing, and applications area of them. .

d. Chapter 4: In this section the experimental results of this study presented and explained in detail form. Furthermore, the discussion of results presented and explained.

e. Chapter 5: In this section, the conclusion of the whole study presented and explained.

2. OVERVIEW

2.1 LIRETURE REVIEW

Liebman [6] provides a good overview of ANN's automatic speech recognition app. One of the major problems with the development of ANN speech recognition is working with the dynamic properties of the speech signal. Perhaps the easiest way to solve this problem is to combine a static ANN model with a traditional speech recognition process model that processes dynamic information. Multilayer sensor (MLP) can be used to obtain local distance values. Programming-based dynamic recognition algorithm [7] Neurobiology-based quantitative learning [8] It can be used to create a high-performance laptop for MLP, and can also generate a discrete Hidden Markov Model (HMM) [9] HMM Viterbi Trackback used as a processor [10] Or, to specify the chronology of the instruction entry arranged by Trackback-Hash [11]. To process dynamic information using the ANN model, the TDNN model assumes latency [12]. TDNN includes short entries and delays. Hidden MLP layers that allow the node to respond. It is sent with high level neurons for different durations. Since the signal passes through TDNN, the network can only perceive the broadcast code. They have been trained using a weak back propagation algorithm [13]. We are also exploring ANNs for speech recognition through comments and feedback from neural networks (usually). Brager et al. [14] A study of Boltzmann's use of repetitive speech recognition devices, Robinson, W. [15], Anderson et al. [16] The networks were tested with redundant connections from the exit node to the entry node. Waterros and Shastri [17] suggested frequent automatic network loops on hidden nodes. Almeida [18] A modified version of the bug distribution has been proposed. Repeat the ANN learning algorithm. It is usually a mundane thing. In contrast, ANN training is more difficult to train and analyze with FEEDBACK OF ANN. Two other studies [19] suggested ANN in place of the expected nonlinear discrimination frame model. Speech recognition is provided by finding the model with the least delay error in dynamic programming procedures.

Pretty Saini has Indian voice recognition system with Barnet kernel and Dua HTK remote control [20]. The system was developed by the HTK Toolkit using the covert Markov Model. MFCC is used as a language identifier. The system is trained and can only recognize 113 words. The system works with both headphone and amplifier. Independent fund. The Indian system of

perception, which is a language of continuous practice, was developed by Gaurav Devansamunim, Shakina Devi, Gopal Krishna Sharma and Mahwah Bhattakarya [21]. It is a system based on the use of MFCC as a separate telephone language as a separate language to extract parameters and functions. Secret Markov Model. Development model. Phonetics is used to refer to words. There are 29 records currently in use in the system. In addition to the MFCC, other functions were used in the development of the sensor system. Knesset member, Linga Morty, GLN. Murphy developed an isolated word associated with a speaker in a real-time English language recognition tool [22]. The system uses linear predictive notation for feature analysis and feature vector extraction. The vector number is specified in the extracted properties to obtain the corresponding Word characters, and these characters are stored in the database. At the recognition level, a match was found for the entered word by comparing symbol similarity with letter similarity in the Similarity and Classification database. The system uses two speakers. The system dictionary consists of five words, and each word consists of ten words. Aside from this work, many popular advertising apps use speech recognition such as Dragon Naturally Speaking, IBM, and Microsoft SAPI Speech. Several studies and experiments have also been conducted to integrate MFCC with other speech functions and create an effective recognition system. Mayor R Gamet, Kinal Damilia, developed an identification system that combines MFCC and linear predictive coding (LPC) using a neural network as a classifier [23]. The system dictionary consists of English numbers from zero to nine. Each speech was shared by 28 speakers: 14 men and 14 women. Experiments are conducted for evaluation.

2.2 RELATED WORKS

2.1.1 Speech Production

How can you create a language? In this section we will look at production Words can be heard from the perspective of tone and deepen understanding The next section describes vowels and consonants. they have to say that the tongue does not start in the lungs. Start with The brain then learned with the help of psycholinguistics. After creation I need a message and grammar structure in my head Representation of a series of tones and a series of commands to do this

We need because our speech organ plays to produce expression Audio side and motor side after these metal operations, let's move on to physically creating the sound. Therefore, audio is created

by the flow of air from the lungs through the lungs. Trachea, mouth, nose. Includes 4 processes: Start, Vocalization, nasal process and joints the starting process is the time when the air leaves the water. lung. In English, the sound of a word is the result of "the progressive flow of air into the lungs." This is not true for all languages. The vocalization process takes place in the larynx. 2 in the larynx Horizontal creases in the cloth as air passes through. These are the vocal cords. Hole between these folds is called the glottis. The source of the human speech is the airstream produced by the lungs. The air flow produced by the lungs is perturbed by a constriction somewhere in the vocal tract. The air flow in the vocal tract changes the air pressure at the lip end. This, in turn, results in the radiation of acoustic waves. These radiated waves are perceived as speech by listeners. The organs in the vocal tract such as the teeth, tongue and etc. are referred as the articulators of vocal tract and they determine the type and place of constriction during speech production process. In the average male, the total length of the vocal tract is about 17 cm. The cross-sectional area of the vocal tract, determined by the positions of the articulators varies from zero (complete closure) to about 20 cm2 [24]. Figure 2.1 shows a schematic diagram of the vocal tract.

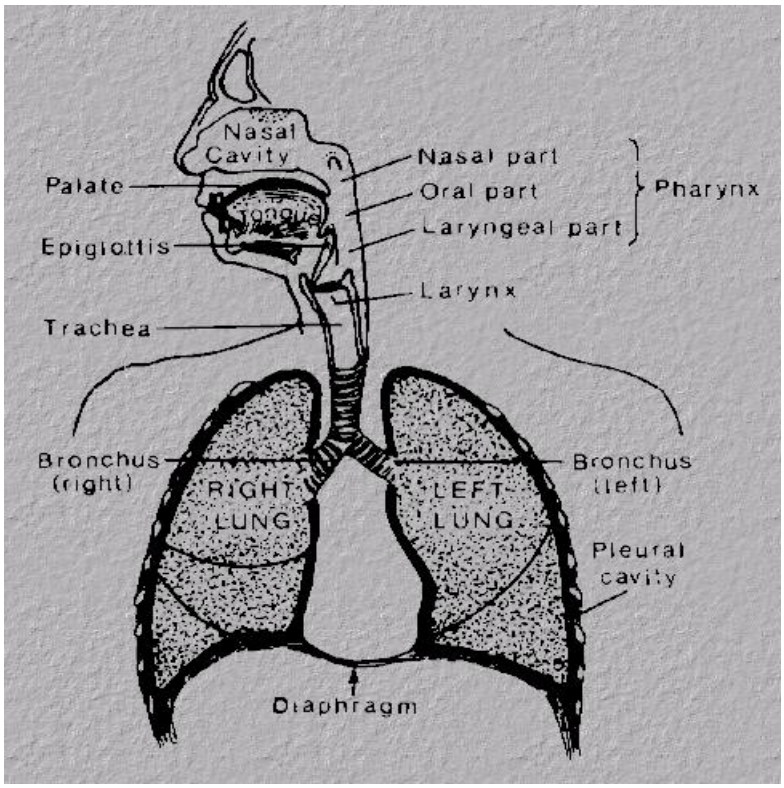

Figure 2.1: Vocal tract for human speech production [25]

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

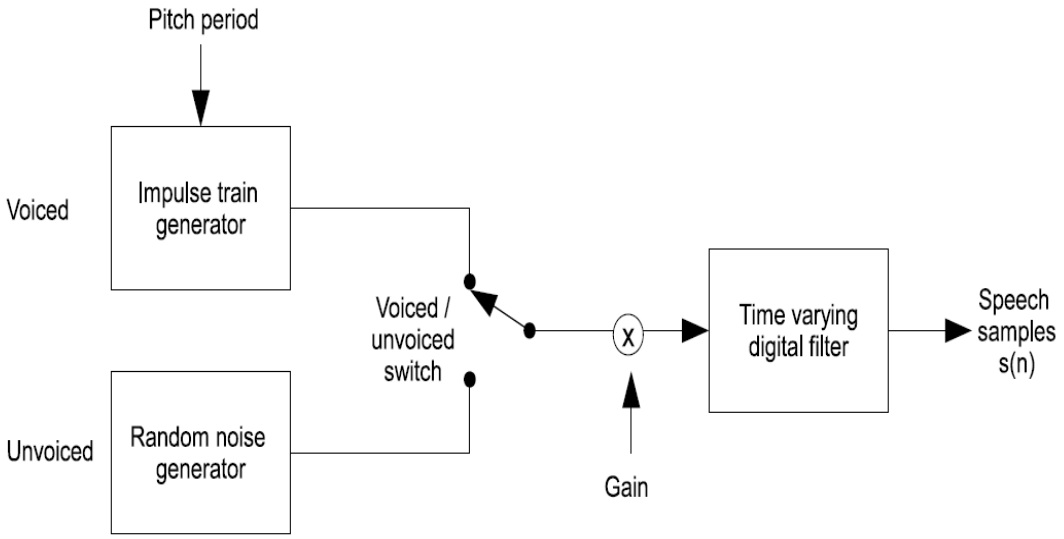

Figure 2.2: Source/filter model for speech production [25]

2.1.2 Fundamentals of Human Speech

Speech signals are time varying pressure waves that are transmitted by a speaker in order to communicate information [26]. Voice signals are composed of a sequence of tones. These tones and the transitions between them serve as a symbolic representation of information [24]. In order to apply signal processing techniques for VAD and speech enhancement, it is essential to understand the speech production process and the structure of human speech.

2.1.3 Modeling the Speech Production Process

Linguists often use symbolic expressions of the spoken language. It turns out to be a useful tool for checking word structure, etc. Level of the grammar course. Scientists are mainly interested in

the motor skills of language. However, the controls generally adopted the non-symbolic formulation presented. Dynamic access to the system and the connection. Research Foundation it became interested in speaking and started integrating these methods. Problems arising from the control and coordination of curators of the art of language. It's like the people we deal with to understand the indescribable behavior. Both The case must specify the multi-level geometry of the system as a set the corresponding reference system (coordinate system) and a number of assignments Is identified among them. You also need that dynamic. Is specified as a set of this coordinate system. At least in vocal production In general, four types of coordinate systems and the dynamics associated with them are expected with many modern designs.

2.1.4 Speech Recognition Applications

Speech recognition basically means interacting with a computer and saying what we are saying. This process works primarily with a pipeline that translates PCM (Pulse Code Modulation) digital audio from a sound card into spoken language. Development of the architecture of signal processing algorithms and hardware developments over 40 years. the systems coped and understood. Over the past 40 years, speech recognition technology has generated, solved, and solved increasingly complex problems [27].

a. Isolated word recognition: Both the speaker and the independent speaker were trained. This technology has opened a class of applications called "command and control applications". This app uses the system to recognize a symbolic command (which provides a dedicated piccoor to certain high level commands) and to recognize a nose. Respond accordingly to commands. Sensitivity and spam words were spoken as the command is male.

b. Speech Understanding Systems: Can identify the underlying message embedded in it Speak instead of just recognizing spoken words. These systems, This has just emerged and has enabled services such as customers. Care (AT&T, how can I help you with the system) and a smart worker. Systems that provide access to information sources via voice dialogs.

c. ATC training is a great application for speech recognition systems. Many ATC educational systems require the current person to act as a "dummy pilot" in order to have

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

A number of international ATC training organizations such as the USAF, USMC, United States Army, United States Navy and FAA, as well as the RAF and civil aviation authorities of Italy, Brazil and Canada are currently using voice recognition. ATC simulators. number of different suppliers.

2.2 TIME DOMAIN FEATURES

Time domain Features from the raw audio signal for ease of implementation. Ease of implementation is an advantage of audio signals, but the main disadvantage of characteristics in the time domain arises from the transient characteristics of audio signals, the statistical characteristics of which change over time, but the characteristics of the time domain remain unchanged. It is issued by signal [28]. Most of the noise captured during recording is another disadvantage of these characteristics, since the characteristics in the time domain are calculated from the signal amplitude values. The aim of this section is applied time domain functions to reduce the size of the data and extracted important features from audio data. The figure 2.3 show the transformation of high size features to the low size high level features by applying time domain functions.

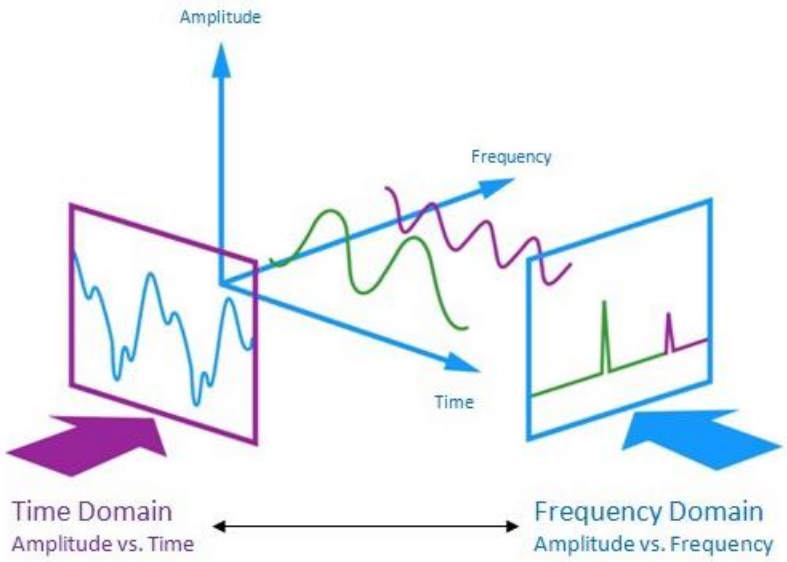

Figure 2.3: Time domain features [29]

2.2.1 Zero-crossing Rate (ZCR)

The point at which the sign of a math function changes (for example, from positive to negative). It is displayed at the intersection of the axes (zero values) of the function graph. It is a term used in electronics, math, sound and image processing. 2 zeros per sine cycle [30]. The equation of this model presented in (2.1):

$$f_0 = [ZCR * f_s]/2 \tag{2.1}$$

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

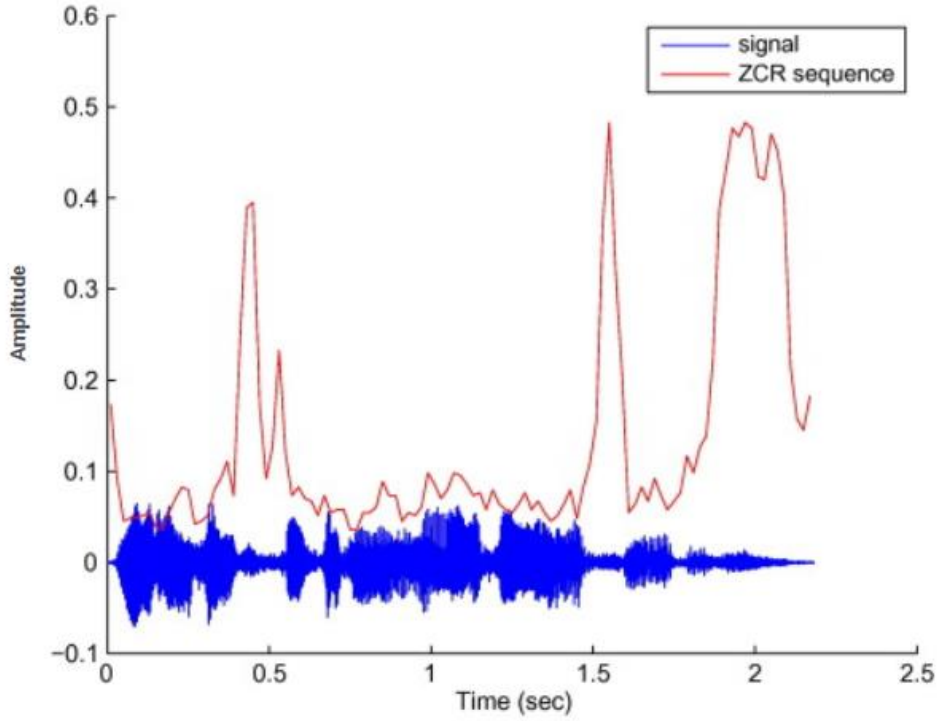

Figure 2.4: ZCR [30]

2.2.2 Short-Time Autocorrelation

Also known as serial correlation, it is the correlation between the signal and the copy delay as a function of delay. Informally, this is similar to observation due to time delays. Autocorrelation analysis is a mathematical tool for finding repeating patterns. It is often used in signal processing to analyze the same set of values or functions as a signal in the time domain. The autocorrelation function can be written as follows in Equation (2.2):

$$\emptyset_{sy}(k) = \sum_{m=-\infty}^{\infty} s(m)\, y(m-k) \tag{2.2}$$

By using this model the autocorrelation function calculated for each y (n) and s(n) with the same signal [31]. The Figure 2.5 shows the Autocorrelation.

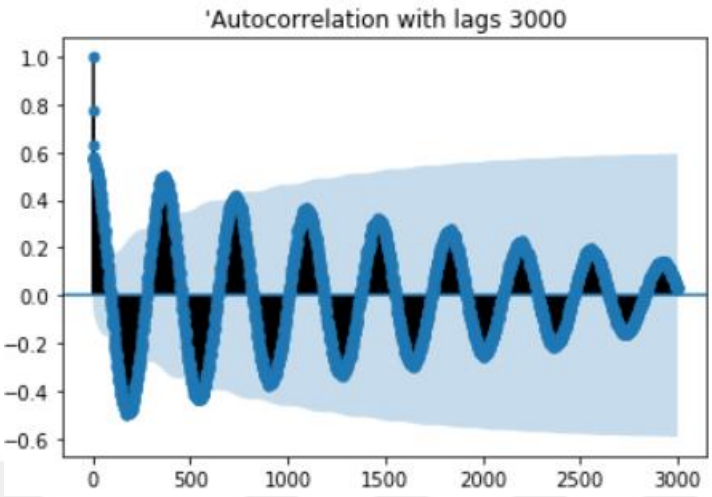

Figure 2.5: Autocorrelation

2.2.3 Hjorth Parameters

The Hjorth parameters is a statistical attribute indicator for signal processing in the time domain, which was introduced by Bo Hjort in 1970 [32]. Parameters: activity, maneuverability, difficulty. They are commonly used when analyzing audio signals to distinguish features. The parameter is the standard gradient descriptor (NSD) used in the audio. In addition, in the field of robotics, the Hjorth parameter is used to process touch signals to determine properties of physical objects, such as: B. the recognition of surface / material structures and the classification of the touch mode using artificial robot skins.

$$Activity = var\left(y(t)\right) \tag{2.3}$$

$$Mobility = \sqrt{\frac{var\left(y^{-}(t)\right)}{var\left(y(t)\right)}} \tag{2.4}$$

$$Complexity = \frac{var\left(y^{-}(t)\right)}{var\left(y(t)\right)} \tag{2.5}$$

Although these three parameters contain information It is also useful to use the frequency spectrum of the signal. Analyze the signal in the time domain. Also the bottom row You can use it to get computational complexity.

2.3 NONLINEAR-BASED FEATURES

2.3.1 Petrosian Fractal Dimension (PFD)

The fractal dimension is a chaotic method for calculating the complexity of the signal [33]. You can use PFD to quickly calculate the fractal size. PFD does this process by converting the signalIn a binary sequence. It can be estimated as follow facial expression as shown in equation (2.6):

$$PFD = {log_{10^k}}/{log_{10^k} + log_{10^{(\frac{k}{k}+0.4N)}}} \tag{2.6}$$

where k is the amount of signal's examples and Nδ is the amount of sign variations in the signal derivative.

2.3.2 Mean Energy

Mean energyin physics, the ability to work. It can exist in various potential forms, kinetic, thermal, electrical, chemical, nuclear, and others. There is also warmth and work - i.e. H. Energy is transferred from one body to another. Once transferred, the energy is still labeled according to its type. Thus, the transferred heat can be converted into heat energy and the work performed can manifest itself as mechanical energy.

$$ME[n] = \frac{1}{N} \sum_{m=k-N+1}^{k} x[m]^2 \tag{2.7}$$

where x[m] is an audio time series, N is the window length and k is the last example in the epoch.

2.3.3 Mean Curve Length (MCL)

The curve length specified by the parameter can be set as the limit for the sum of the line segment lengths for a regular partition if the number of segments approaches infinity.

$$ME[n] = \frac{1}{N} \sum_{m=k-N+1}^{k} x[m] - x[m-1] \tag{2.8}$$

where x[m] is an audio time series, N is the window length and k is the latter example in the period [34].

2.4 FREQUENCY DOMAIN

In physics, electronics, control engineering, and statistics, the frequency range is a mathematical function and analyzes a signal as a function of frequency, not time. In simple terms, a time history diagram shows how a signal changes over time, and a frequency domain diagram shows the variation of a signal in a frequency band in a given frequency band. The frequency domain representation may also contain information about the phase changes that should be applied to each sine wave, so that the frequency components can be recombined to reconstruct the original time signal [35].

2.4.1 Fourier Series

In mathematics, a Fourier series is a periodic function consisting of rational symmetric sine waves related to the weighted sum. With appropriate weights, additional time (or time) can be used to approximate a function within that range (or the entire function if the function is periodic). Here is a summary of all the other jobs. An example of a Fourier series is an example of a discrete time Fourier transform. The process of obtaining weights that represent a particular function is a form of Fourier analysis. Analytical and synthetic isotopes of functions at infinite periods are Fourier transform and inverse transformation [36]. The Fourier series is represented as shown in the equation (2.9):

$$Sn(x) = \sum_{n=-N}^{N} c_n \cdot e^{i\frac{2\pi nx}{P}} \tag{2.9}$$

2.4.2 Wavelet Transform

A wave series in mathematics is a representation of the integral square function (real or complex) of a given orthogonal series resulting from a wave. This article provides a formal mathematical description of an orthogonal signal and an integral wavelet transform [37].

All row of features contain wavelet to these rows to extracted features by using Eq (2.10).

$$DWT(i,k) = \frac{1}{\sqrt{{a_0}^j}} = \sum_{n=-\infty}^{\infty} f(n)\psi\left(\frac{\left(n - {a_0}^j\, k\, b_0^{\cdot}\right)}{{a_0}^j}\right) \tag{2.10}$$

Where j, k, $n \in Z$ and $a_0 > 1$.

2.4.3 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-frequency cepstrum is a Brief display of the power spectrum of the audio signal. MFC is based on a linear cosine transformation of the logarithmic power spectrum on a non-linear Mel frequency scale. MFCC are the coefficients that implement the MFC study derived from the description of the type of cepstral audio signal. A, omg MFC and cepstrum modification; Frequency bands that are scattered along the Mel scale in the MFC are stimulated more strongly in relation to the response of the human hearing system than the linear frequency bands in the normal cepstrum [38]. The following are the extraction properties of the MFCC process figure 2.5.

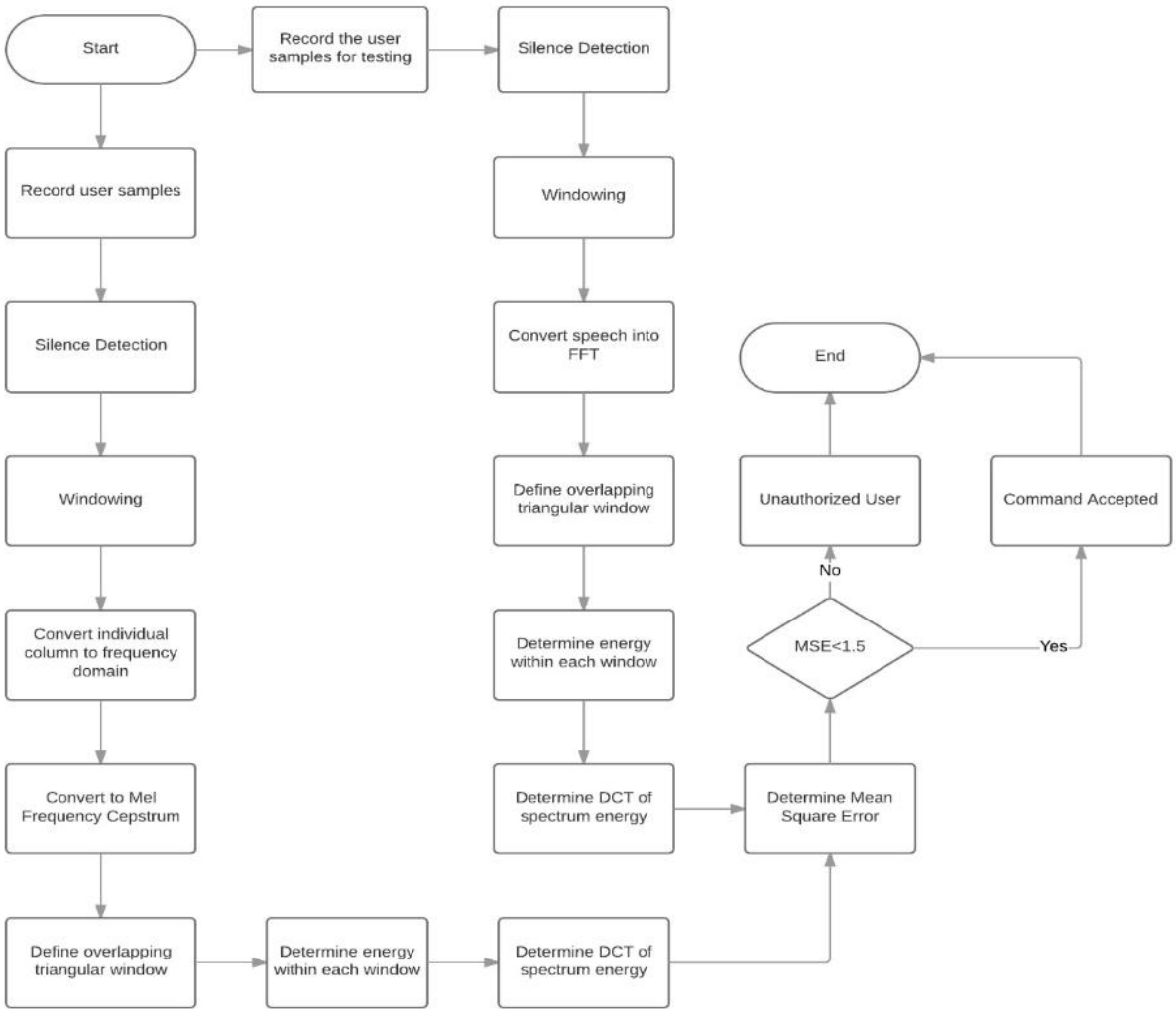

Figure 2.6: MFCC [39]

3. MATERIAL AND METHODS

3.1 DIGITAL SIGNAL PROCESSING

Digital signal processors (DSPs) record signals from the real world and process them mathematically, e.g. B. numbered tones, tones, videos, temperatures, pressures or positions. DSPs are designed to perform mathematical functions such as plus, minus, multiplication and division very quickly.

The signal must be processed so that the information contained in it can be viewed, analyzed or converted into other types of signals that may be useful. In the real world, analog products recognize and control signals such as sound, light, temperature or pressure. As a digital converter, an analog-to-analog converter collects the actual signal and converts it to digital formats 1 and 0. The DSP receives and processes the digital information from there. The digital information is then returned for use in the real world. This is done in two ways, digital or analog, using a digital-to-analog converter. Everything happens very quickly.

To illustrate this concept, the following illustration shows how DSP is used with an MP3 player. During the recording phase, analog audio is received through the receiver or another source. This analog signal is then converted to digital using an analog-to-digital converter and sent to the DSP. DSP encodes the MP3 file and stores the file in memory. The currently playing file can be recalled from memory and then converted to an analog signal with the DSP decoded and played through the speaker system. In a more complex example, the DSP can perform other functions such as volume control, balance and user interface.

For example, DSP information can be used to monitor the security of computers, phones, home theater systems, and video compression. You can compress the signal and move it from one location to another faster and more efficiently (for example, you can transfer audio and video over a phone line and during a conference call). Signals can be amplified or manipulated to improve quality or provide information that a person cannot see (eg, echo cancellation for cell phones or medical images on a computer). Actual signals can be processed in analog format, but digital signal processing has advantages such as high speed and high accuracy.

Continuous temporal signals are defined as continuity in time and are expressed as a single continuous variable. Continuous time signals are often referred to as analog signals.

These signals have both amplitude and temporal continuity. It still has values. The sine and cosine functions are the best examples of continuous time signals.

Figure 3.1: Continuous time signal

The above signal is an example of a continuous time signal, the value of which can be obtained at any time.

Signals defined at different times are called discrete signals. Therefore, each independent variable has its own value. Thus, it is displayed as a series of numbers.

The advantage of audio and video signals is that they are offered in a variety of formats with continuous timing. Under certain conditions, it is the same. The containers have discrete characteristics. Digital signals are a good example. Its amplitude and time are different.

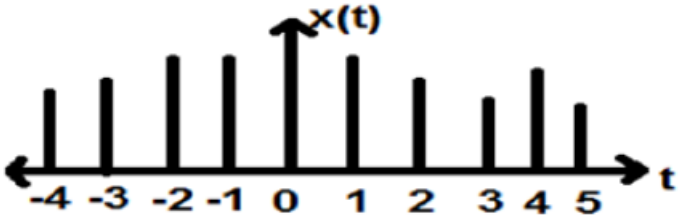

Figure 3.2: Discrete signal

3.1.1 Audio Processing

Sound processing includes many different areas, each related to the presentation of sound to the listener. The focus is on three areas: (1) For example, playing hi-fi music. Audio CDs have (2)

speech, another name for telephone networks, and (3) synthetic speech, in which computers form and recognize human speech patterns. ... While these practices have different goals, they have a common arbiter: the human ear. Digital signal processing has revolutionized these and other areas of sound processing.

Audiophiles demand better sound quality and all other factors are secondary. If you had to describe your thought in one word, it would be: exaggeration. In addition to responding to the capabilities of the human ear, these systems are designed for more than just hearing. This is the only way to ensure that the music played is perfect. Digital audio from CDs or CDs. It was a revolution in music; The sound quality of a CD system is much higher than that of older systems such as vinyl records and cassettes. DSP has been at the forefront of this technology.

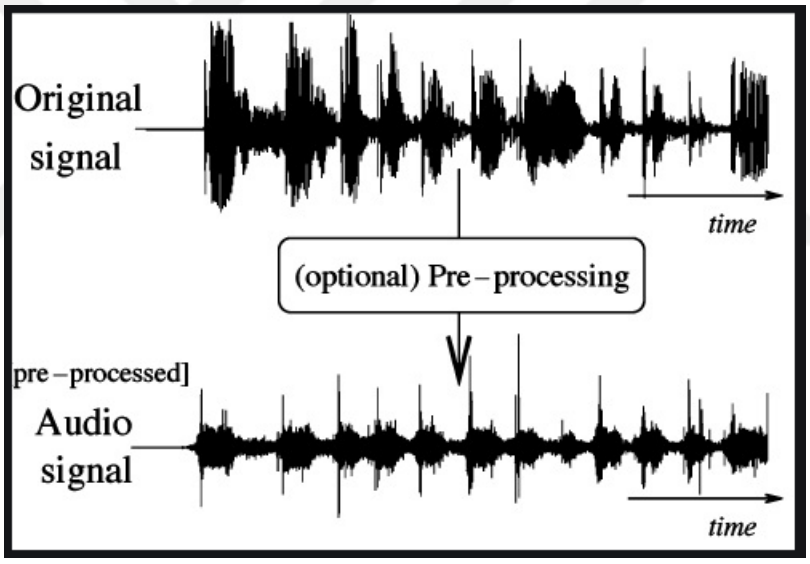

Figure 3.3: Discrete signal

3.1.2 Audio Signal Characteristics

The audible sounds are caused by pressure fluctuations in the ears. Human hearing aids respond to frequency noise in the 20 Hz to 20 kHz range with an intensity that exceeds the frequency dependent "audible threshold". The sound range is about 120 dB, which is the gap between the bottom of the blades and the combat. Dislocation Figure 1 illustrates the human auditory field in terms of frequency intensity. The sound captured by the microphone is a passing waveform. The air pressure changes at the position of the microphone in the sound field. Digital audio signals

are captured by taking accurate samples and estimating the electrical microphone output. Sampling frequency. Above 40kHz is sufficient to cover the entire audible frequency range. The common sampling rate is 44100Hz according to recorded data. Audio and video data must be synchronized. "CD quality" refers to 44.1 kHz. 16-bit digital audio and preview.

3.1.3 Application Areas

Processing and usage methods include storage, data compression, music search, speech processing, translation, speech recognition, transmission, noise reduction, speech printing, speech recognition, synthesis and optimization.

a. Audio broadcasting

While many countries around the world have a digital radio standard that provides digital audio radio services, HD radio is the digital radio standard in North America. The DAB standard emerged in the 1980s as a European research project. The Norwegian Broadcasting Corporation (NRK) launched the world's first DAB channel (NRK Klassisk) on June 1, 1995 [2], while the BBC and Swedish Radio (SR) launched the first digital broadcast in DAB format. September 27, 1995 DAB receivers have been available in most countries since the late 1990s.

b.Audio synthesis

Sound synthesis is an electronic musical instrument that produces sound signals. Synthesizers create sounds using techniques such as subtractive synthesis, additive synthesis, and frequency modulation synthesis. These sounds can be shaped and edited using components such as filters, envelopes, and LFOs. Synthesizers are usually played on the keyboard or are usually controlled by sequencers, software, or other instruments via MIDI.

3.4 Speech Recognition Based Optimized Deep Learning

In this section, the presented method pitch based RNN and PSO proposed. The speech data first analyzed using pitch function the aim of this step is to extracted to reduce the size of input

features and increase accuracy of the framework. In the second step, Then, the mathematical model of the pitch presented as shown below Equation (3.1) and Equation (3.2):

$$Max\ number\ of\ semitones\ lowered:\ -12 * log2(numel(Window) - OverlapLength) \tag{3.1}$$

$$Max\ number\ of\ semitones\ raised:\ -12 * log2\left(numel(Window) - \frac{OverlapLength}{numel(window)}\right) \tag{3.2}$$

In the first iteration, $EnvX_a$ is fitted to X. Then the output features wired to the RNN which is powerfull time series analyzing technique applied to classify the input features. The aim of the PSO in this study is to rain RNN model and optimize the weight and basis that can be presented best accuracy. The flowchart of this study presented in Figure 3.3.

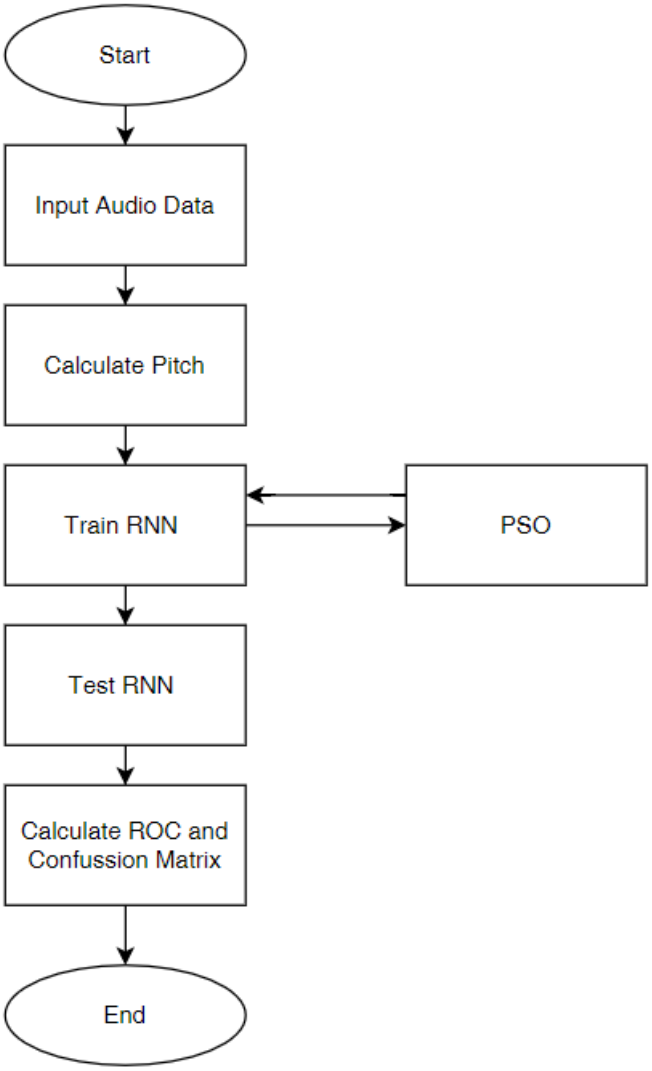

Figure 3.4 : Our framework

4. EXPERIMENTS AND DISSCUSION

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

4.1 FEATURE EXTRACTION USING PITCH

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

Then, the mathematical model of the pitch presented as shown below Equation (4.1) and Equation (4.2):

$$Max\ number\ of\ semitones\ lowered\colon\ -12 * log2(numel(Window) - OverlapLength) \tag{4.1}$$

$$Max\ number\ of\ semitones\ raised\colon\ -12 * log2\left(numel(Window) - \frac{OverlapLength}{numel(window)}\right) \tag{4.2}$$

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

a. To obtain new predicate lowpass filter applied as cepstral representation to the EnvX_a.

b. To update the existing finest fitting, the algorithm obtain the element-by-element extreme of the present spectral cover predicate and the earlier spectral cover predicate:

$$\mathrm{EnvX}_a = \max \mathrm{EnvX}_a, \mathrm{EnvX}_b \tag{4.3}$$

c. The loop finishes if whichever a supreme sum of iterations (100) is grasped, or if entirely baskets of the predictable log wrapper are inside a assumed broadmindedness of the unique log spectrum. The broadmindedness is set to $\log(10^{1/20})$.

d. Then, the process scalars the spectrum of the pitch-shifted audio by the relation of predicated wrappers, element-wise:

$$\mathrm{Y} = Y * \left(\frac{\mathrm{EnvX}_b}{\mathrm{EnvY}_b}\right) \tag{4.4}$$

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

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

Figure 4.1: Pitch algorithm

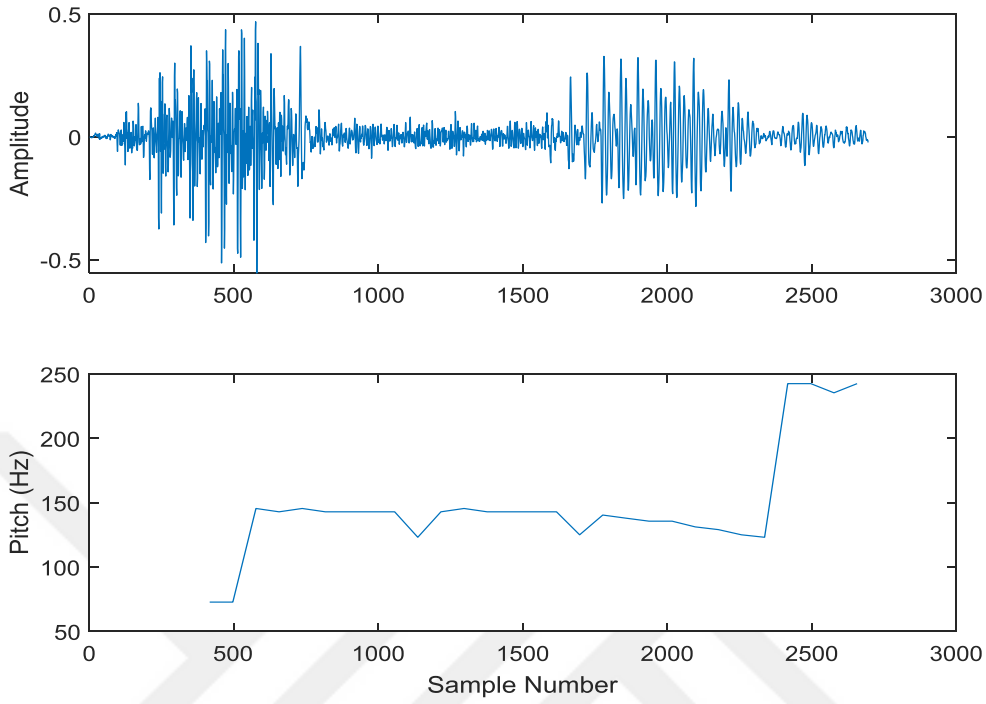

Figure 4.2: Feature extraction for zero using pitch

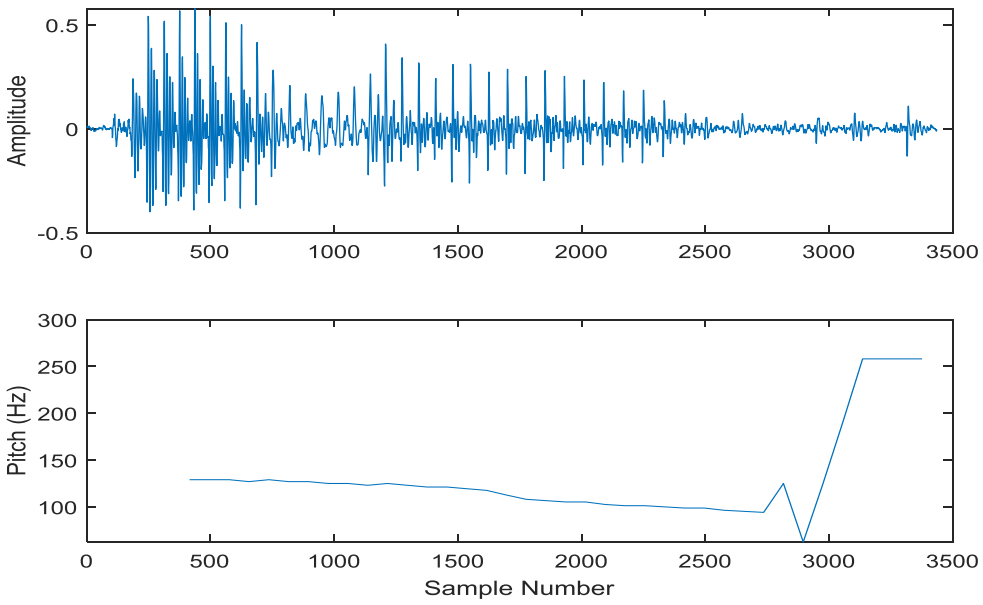

Figure 4.3: Feature extraction for one using pitch

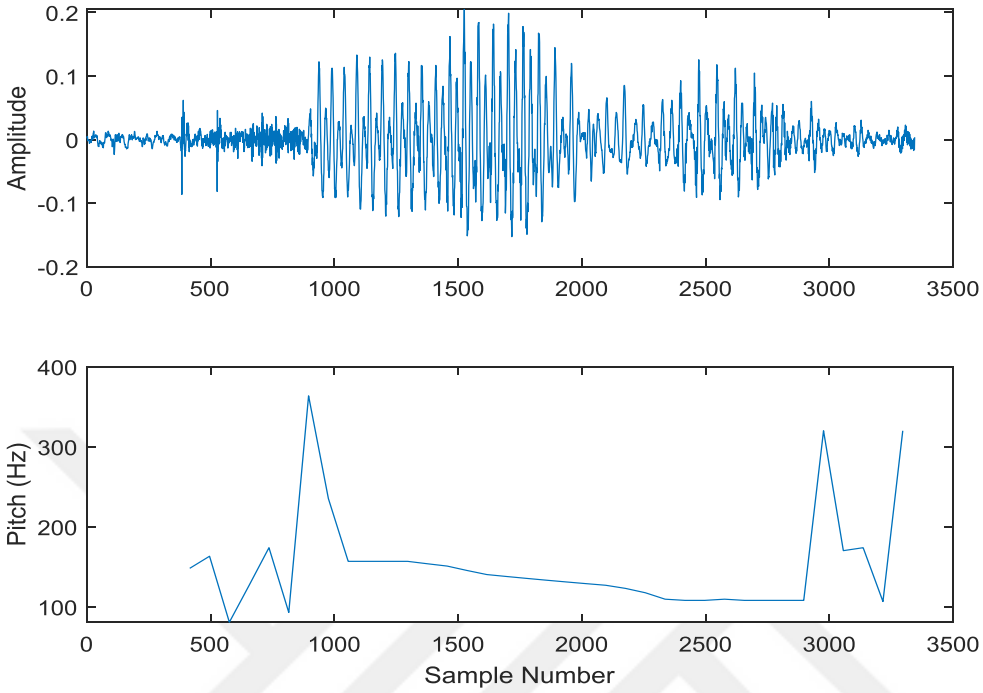

Figure 4.4: Feature extraction for two using pitch

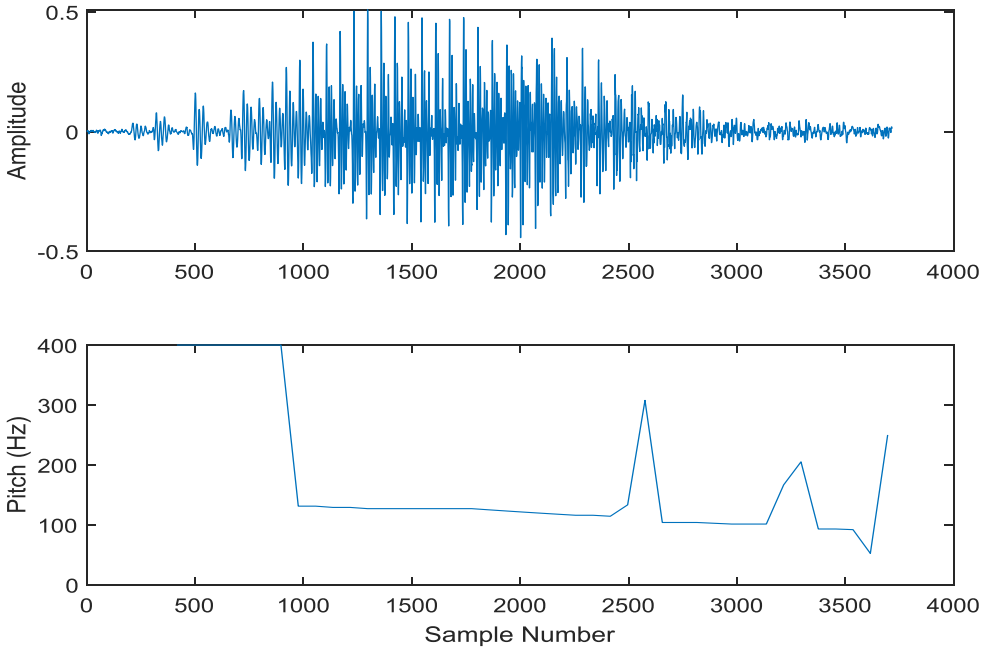

Figure 4.5: Feature extraction for three using pitch

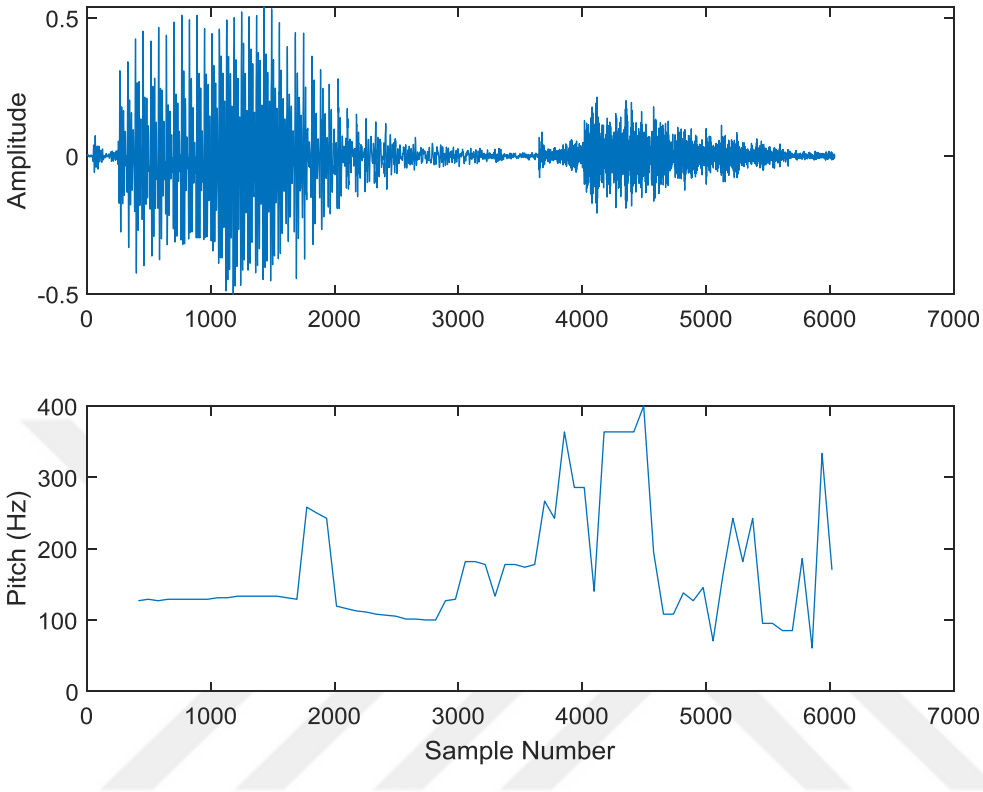

Figure 4.6: Feature extraction for four using pitch

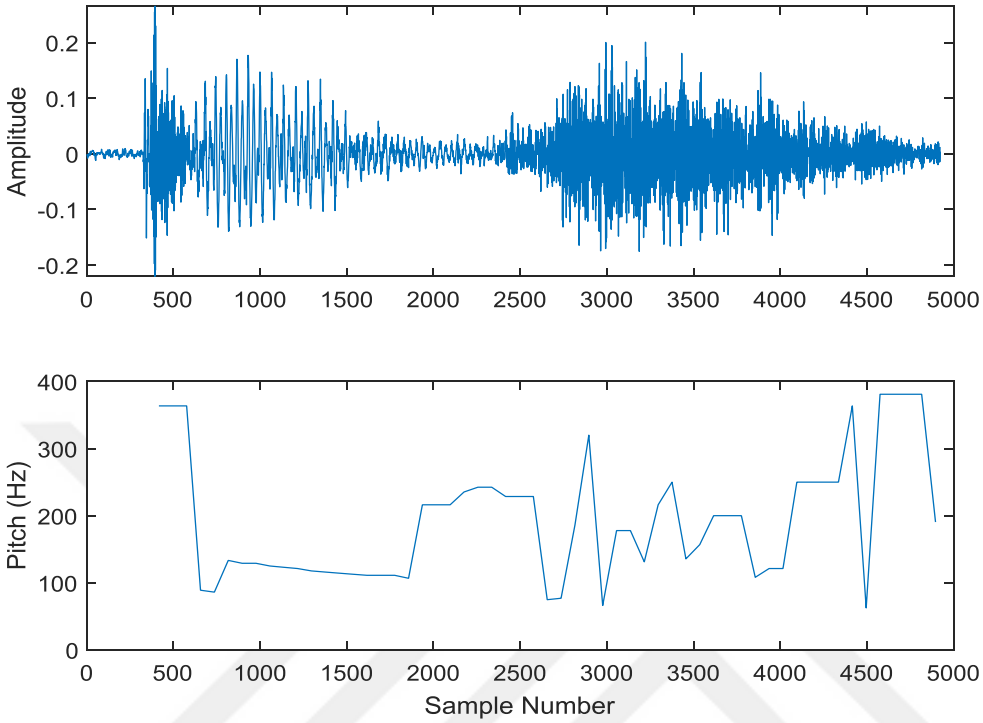

Figure 4.7: Feature extraction for five using pitch

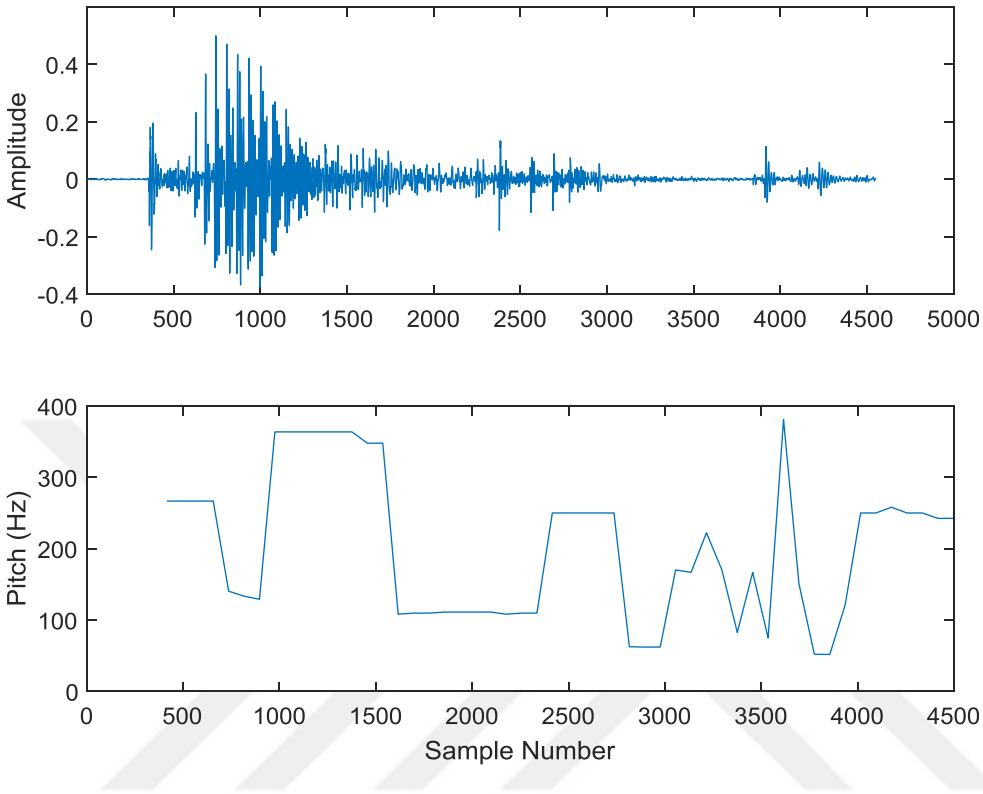

Figure 4.8: Feature extraction for six using pitch

4.2 Feature Extraction Using MFCCS

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

The computation of MFCC feature can be explained as follows as illustrated in Table 4.1.

Table 4.1: MFCC Parameters

Parameter	Value
Window length	25
Window overlap	10
Cepstral coefficients	12
Delta cepstral coefficients	12
Double delta cepstral coefficients	12
Energy coefficient	1
Delta energy coefficient	1
Double delta energy coefficient	1

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

$$y_k = log(\Upsilon/x_{k-1}/^2 + /x_k/^2 + /x_{k+1}/^2 + £) \quad (4.5)$$

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

$$\text{Power spectrum of signal} = /F^{-1}\{f\{x(t)\}\}/^2 \quad (4.6)$$

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

$$m = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \quad (4.7)$$

Then, willingly derived the inverse transform as:

$$f = 700\ \left(10^{\frac{m}{2595}} - 1\right) \quad (4.8)$$

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

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

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

$$0 \qquad otherwise \tag{4.9}$$

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

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

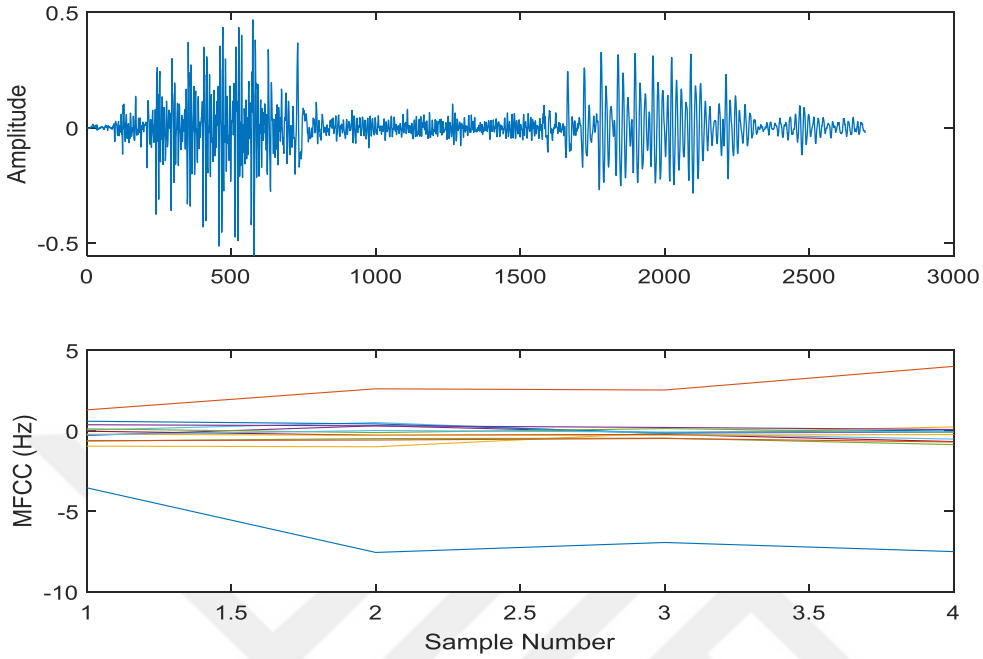

Figure 4.9: Feature extraction for one using MFCCs

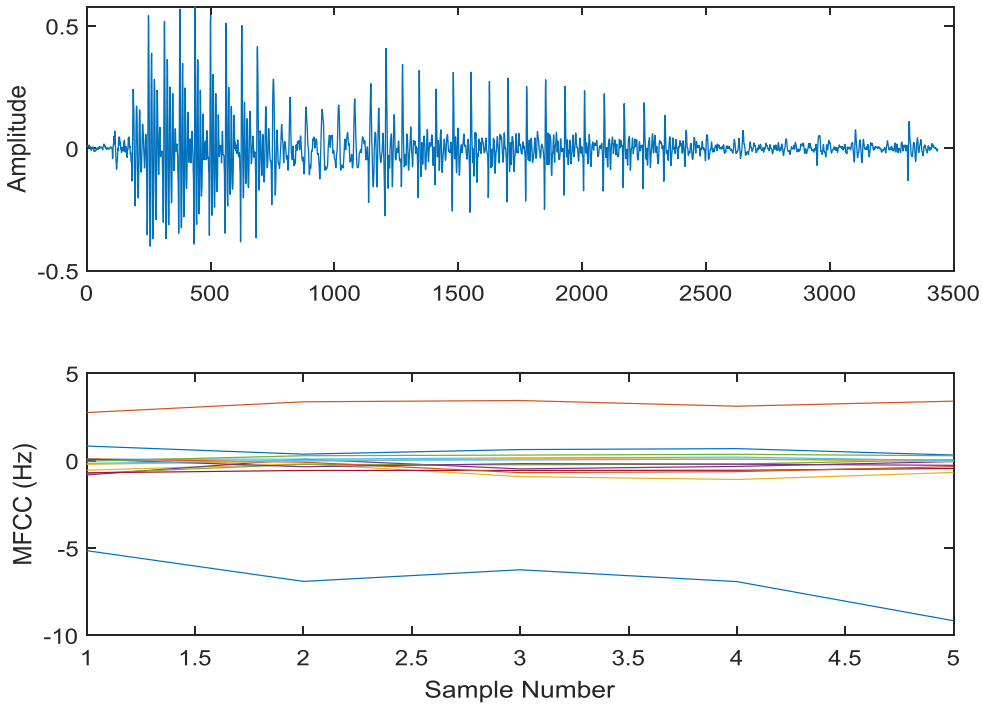

Figure 4.10: Feature extraction for two using MFCCs

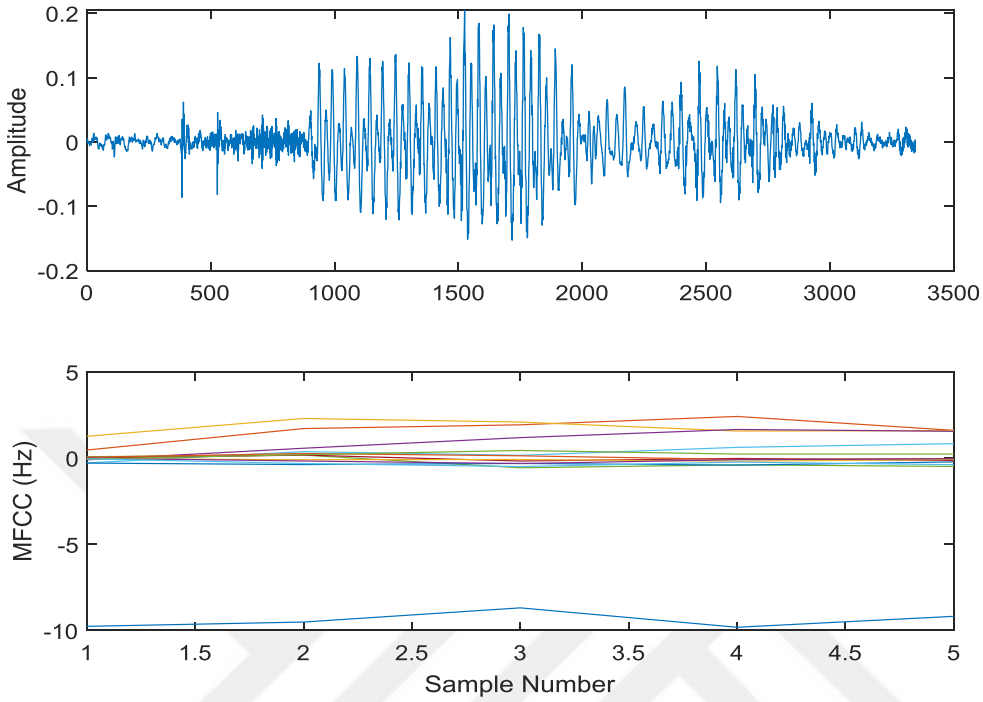

Figure 4.11: Feature extraction for three using MFCCs

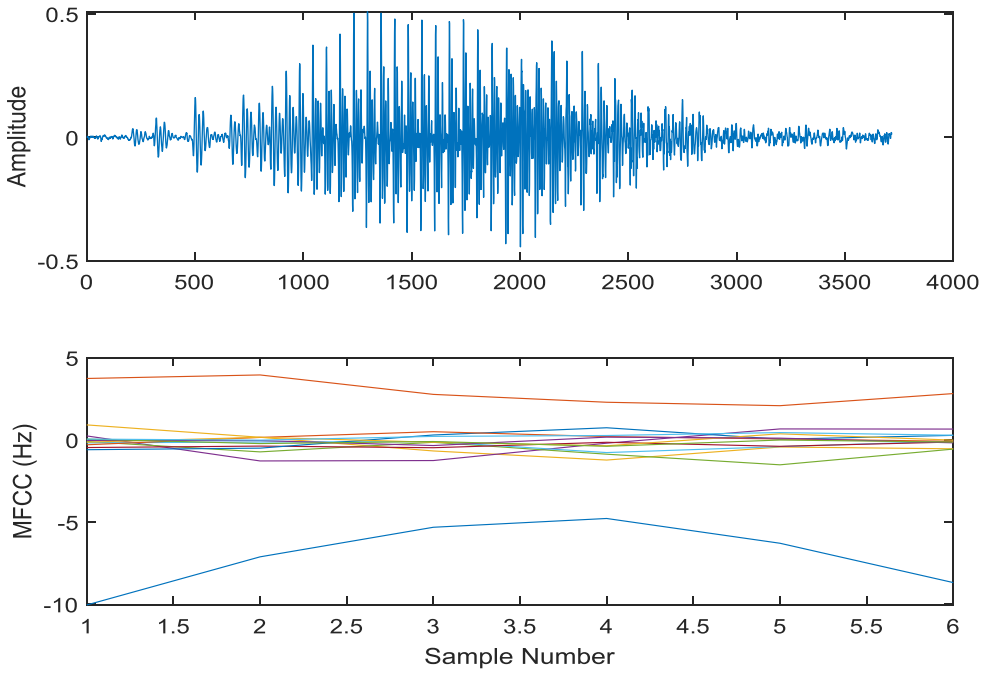

Figure 4.12: Feature extraction for four using MFCCs

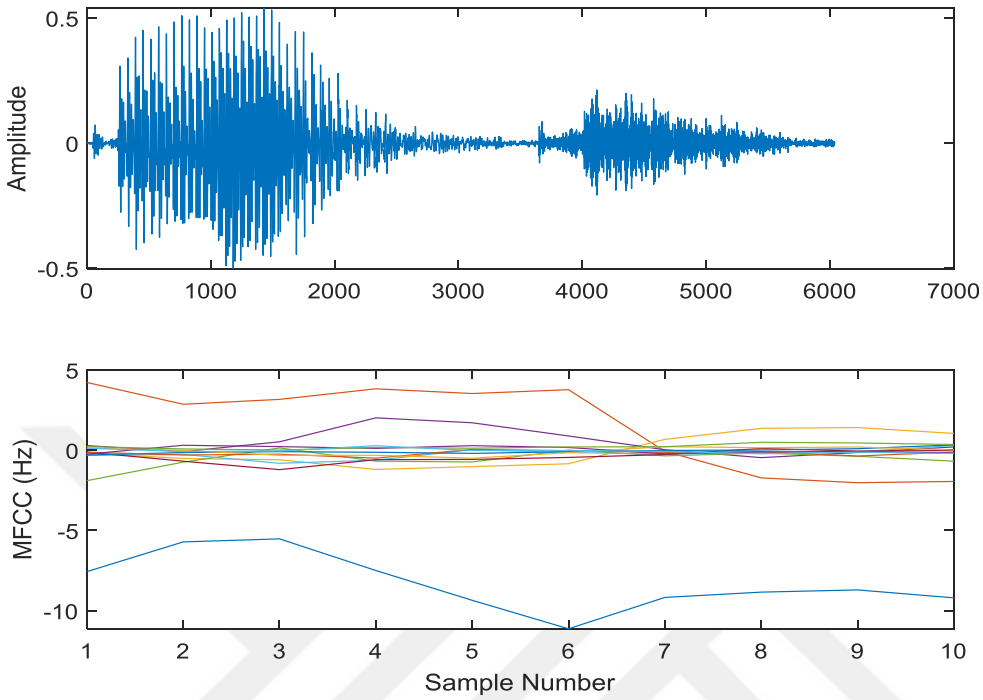

Figure 4.13 Feature extraction for five using MFCCs

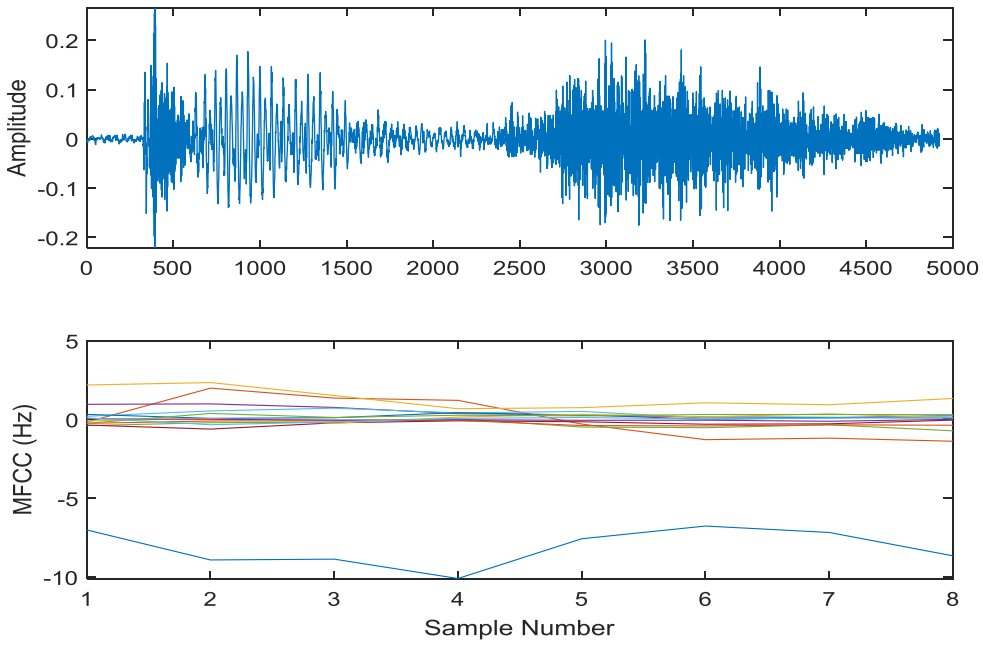

Figure 4.14: Feature extraction for six using MFCCs

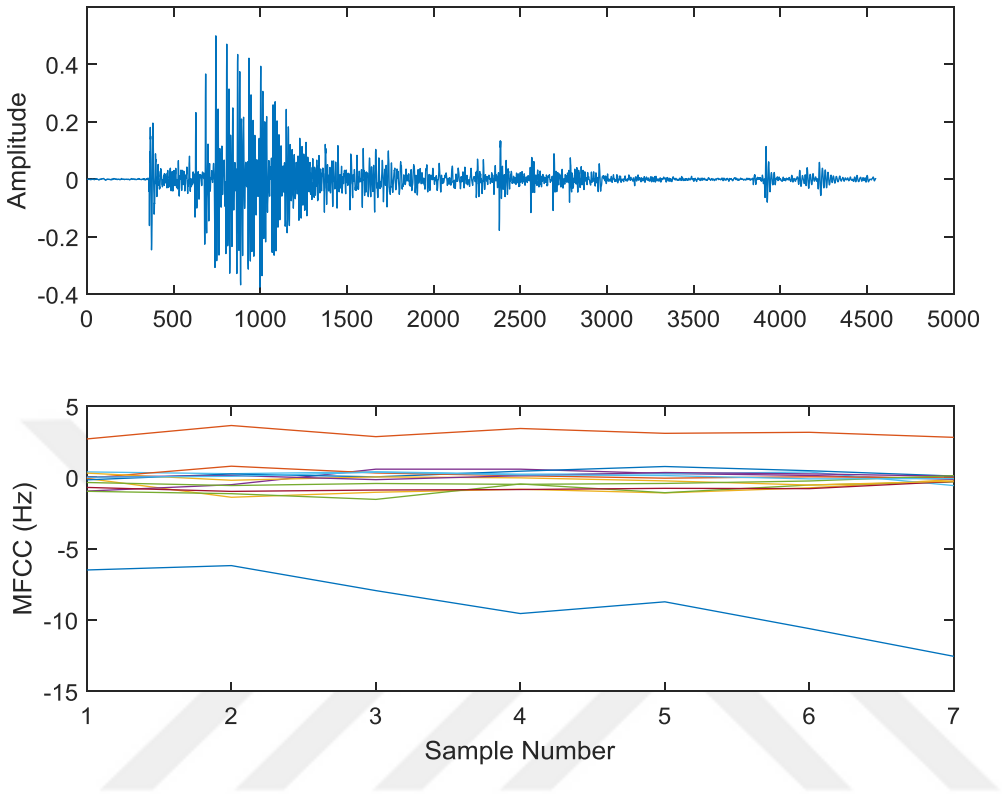

Figure 4.15: Feature extraction for seven using MFCCs

4.3 Feature Extraction Using Energy

In this section, energy of each signal is calculated the main idea of this method is calculated energy for each frame (epoch). Then, the data classified to the epochs and the energy of each epoch calculated we can visualize this idea in the Figure 4.16.

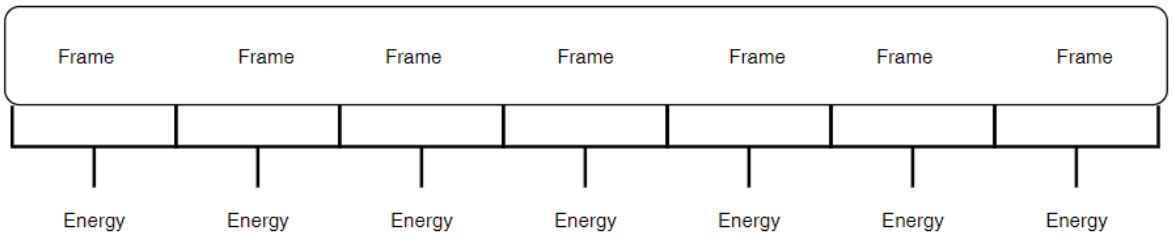

Figure 4.16: Energy calculation

Mathmaticaly we can represented this step as shown in equation (4.10):

$$E = \int_{-\infty}^{\infty} x(t)^2 \, dt \tag{4.10}$$

Where the ∞ and -∞ represented the upper and lower limit for the each frame, x(t) represented the audio data that it energy will be calculated. If the input signal in the frequency domain then the model represented as shown in the equation (4.11):

$$ESD = \int_{-\infty}^{\infty} x(f)^2 \, df \tag{4.11}$$

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

4.4 Discussion

In this section IWR results presented by using confusion matrix and roc curve for evaluate the results. In the first experiment pitch based RNN presented and the results show in Figure 4.17 and 4.18.

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

a. TP: True Positive: estimated prices appropriately estimated as real positive

b. FP: estimated prices falsely predicted an real positive.
c. FN: False Negative: Positive values estimated as undesirable
d. TN: True Negative: estimated values correctly estimated as an actual

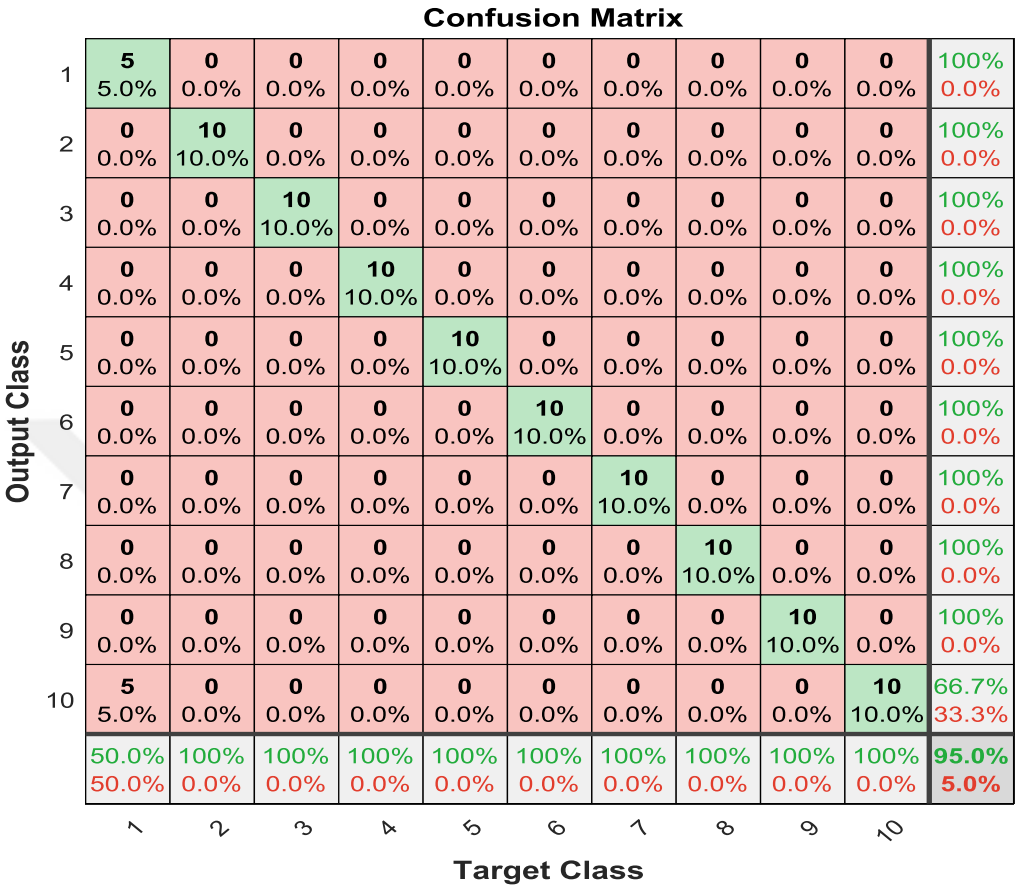

Figure 4.17: Pitch confusion matrix

The roc curve of the pitch based RNN also presented and shown in the Figure 4.18. The ROC curve is mainly used to graphically represent clinical sensitivity and specificity relationship / compensation for testable limits or combinations. In addition, the area under the ROC curve shows the advantage of using the test provided. The ROC curve is used in clinical biochemistry to select the most appropriate test limits. The best cutoff is the highest true positive rate and the lowest false positive rate.

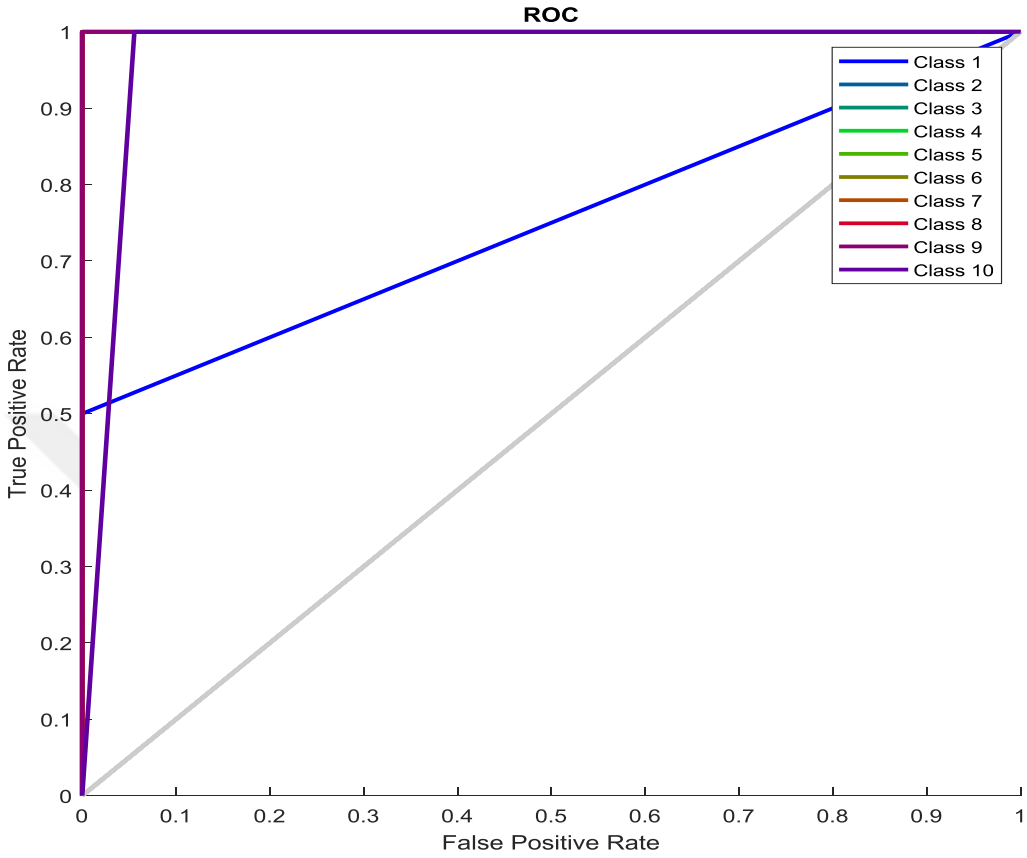

Figure 4.18: Pitch based RNN Roc curve

The pitch based RNN presented high results in fast execution time. The execution time is very low compared with other techniques because its mathematical model is very simple which presented and explained above. The second experiment combined the energy with RNN. The energy of each audio calculated then the energy signal wired to the RNN. The presented energy based RNN results presented in Figure 4.19 and 4.20.

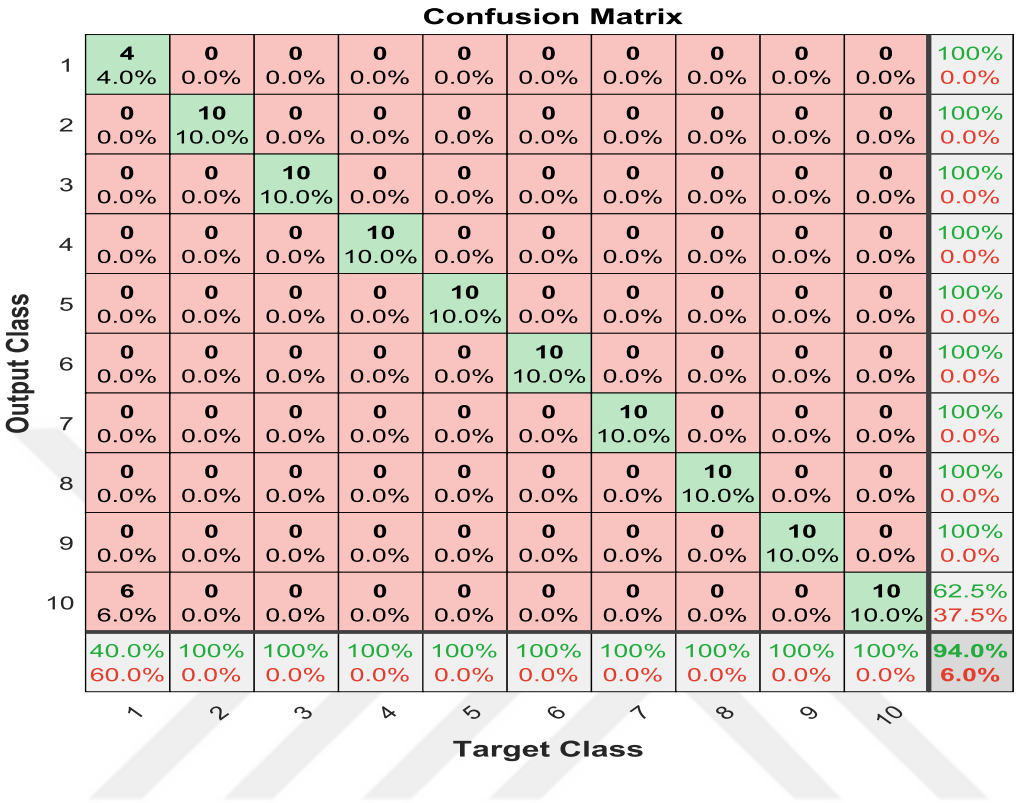

Figure 4.19: Energy based RNN confusion matrix

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

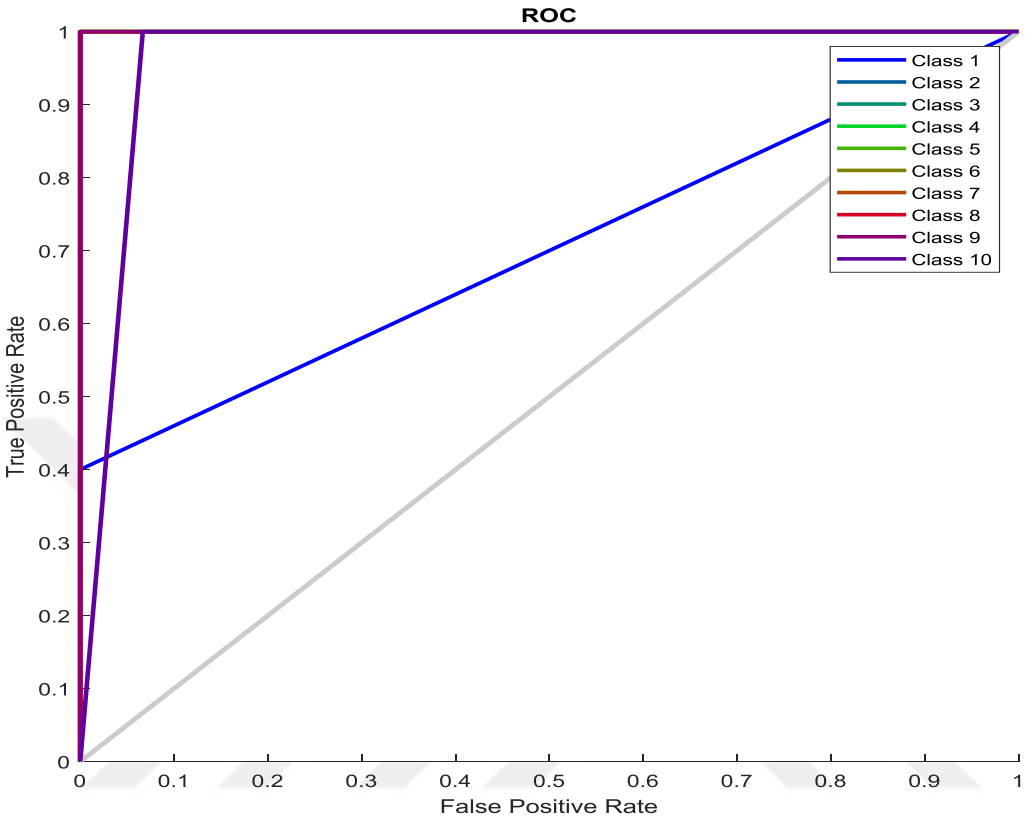

Figure 4.20: Energy based RNN Roc Curve

The MFCCs based RNN applied to the same digit audio dataset with 1024 windowing size. The main issue in MFCCs topic is the length overlap which we set two overlapped length: 1024 and 512. The results affected with the vary of overlap length which with overlap 1024 the model presented low results with low execution time and the reason of this because the window cover the whole signal with low number of iteration because the window in each iteration jump 1024 and the results of this step presented in 4.21 and 4.22.

Figure 4.21: MFCCs based RNN confusion matrix with 1024 overlap length

Furthermore, the roc curves of this case presented in the Figure 4.23.

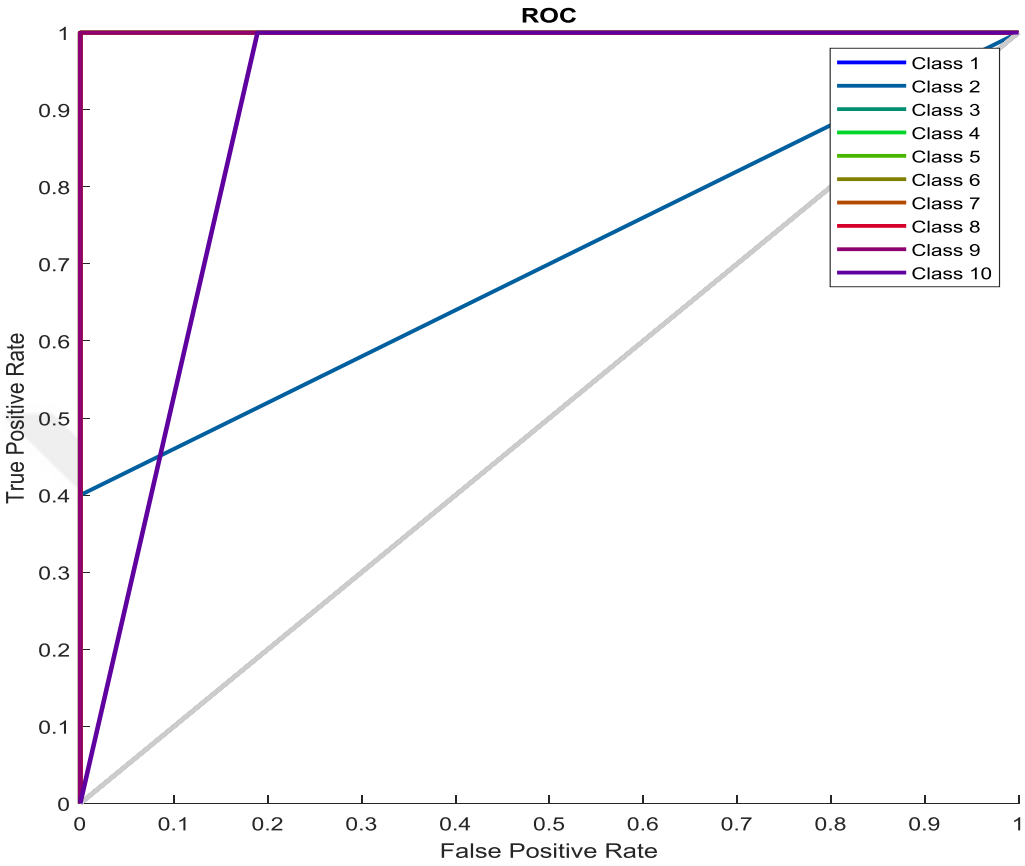

Figure 4.22: MFCCs based RNN Roc curve with 1024 overlap length

On the other hand, the 512 overlap length applied which lead to presented best results but with high execution time. The presented method results shown in the Figure 4.23 and 4.24.

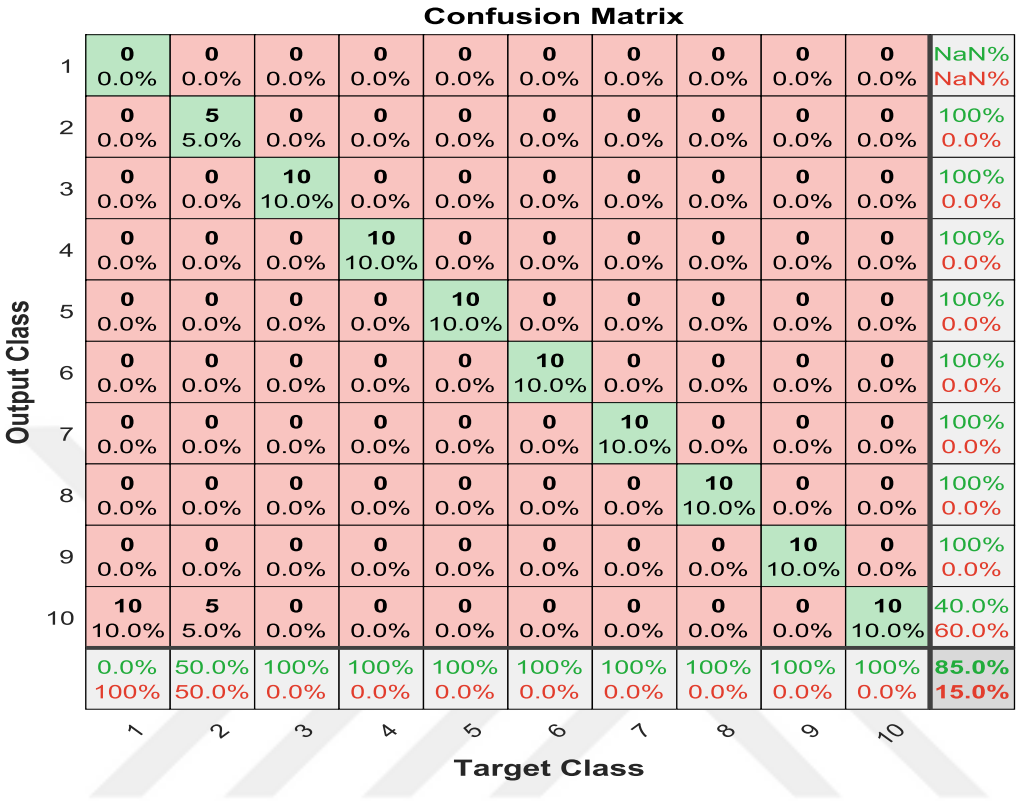

Figure 4.23: MFCCs based RNN confusion matrix with 512 overlap length

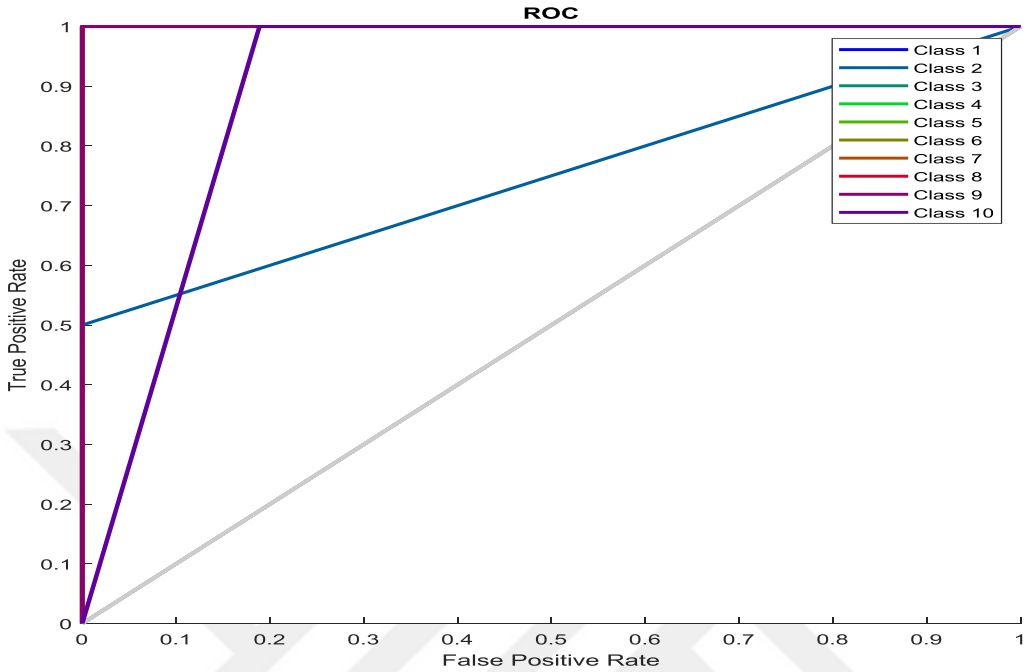

Figure 4.24: MFCCs based RNN Roc curve with 512 overlap length

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

Table 4.2: Results Comparisons

Method	Methods	Acc (%)
Meutzner et al. [40]	MFCC+SVM	88.72
Meutzner et al. [40]	Energy + SVM	92.32
Gurban and Thiran [41]	Model Based Entropy	81.32
Estellers et al. [42]	A second-order exponential function	81.21
Our Method	Pitch+PSO+RNN	95

Form Table 4.2, we can prove that our method presented best results than previous studies. In [40] presented two methods MFCC+SVM and Energy + SVM the proposed methods presented

88.72 % and 92.32 respectively. Gurban and Thiran [41] presented new method based Model Based Entropy technique this method presented 81.32 % accuracy which is suitable. Estellers et al. [42] presented new method based A second-order exponential

Function this method is new and interested and presented 81.21% accuracy. On the other hand, our method Pitch+RNN presented best results than all previous studies which presents 95%. Furthermore, Energy+RNN is presented best results than previous studies which presented 94%. The important point that energy+RNN presented best results than Energy + SVM which mean the RNN more effective these types of problems.

5. CONCLUSION

In this study, we presented a new approach to speech recognition. At the first stage, the Bohman function was used for the VAD (voice activity detection) task. The output of Boman's two functions is connected to an RNN that has been trained with the PSO. The method, tested on a publicly available dataset, shows that the proposed method shows remarkable and high quality results when compared to previous studies. The main advantages of the PSO algorithm are as follows: simple concept, easy implementation, reliability of control parameters and computational efficiency compared to the mathematical algorithm and other methods of heuristic optimization. maximum number of repetitions, the number of repetitions available. Furthermore, The main advantage of RNN over ANN is that RNN can simulate a series of data sets (i.e. temporal aggregation) so that each model can be assumed to depend on the previous ones. Recurring neural networks are also used with convolutional layers to augment the powerful pixel environment. The training RNN with PSO realy lead to remarkable results when compared with previous stdeis and other techniques. The PSO applied to optimize the performance of the RNN by updating the weight and basis and applying to the PSO to maximize the accuracy of the model.

For future work, we recommend that researchers apply these presented methods to other datasets, such as digital speech recognition. We also recommend combining the MFCC with the LSTM, which in our opinion works best with other classifiers, since the MFCCs had multidimensional properties with different sizes, which led to problems and low classification rates with RNNs. and other classifiers not dealing with time series problems.

REFERENCES

[1] R. P. Lippmann, ”Review of neural networks for speech recognition,” Neural Computation, vol. 1, no. 1, pp. 1-38, 1989.

[2] H. Bourland and C. J. Wellekens, ”Speech pattern discrimination and multi-laver perceotrons,” Tech. Reo. M-211, PhilliDs Res. Lab., Bruseels, Belgium, 1988. ’

[3] T. Kohonen G. Bama, and A R. Chrisley, “Statistical pattern recognition . 1 with neural networks: Benchmarking studies,” in Proc. \€€E Ann. Int. Conf: Neural Networks, July, 1988.
[4] H. Iwamida, S. Katagiri, E. McDermott, and Y. Tohkura, “A hybrid speech recognition system using HMMs with an LVQ-trained codebook,” J. Acoustic Soc. Japan, vol. 11, no. 5, pp, 277-285, 1990.

[5] Y. Q. Gao, T. Y. Huang, and D. W. Chen. “HMM-based warping in neural networks,” in Proc. IE€E Int. Conf Acoustics Speech Signal Processing, 1990, pp. 501-504.

[6] J. Tebelskis and A. Waibel, “Large vocabulary recognition using linked predictive neural networks,” in Proc. IEEE Int. Conf Acoustics Speech Signal Processing, 1990, pp. 437-440.

[7] K. I . Is0 and T. Wdtanabe, “Speaker-independent word recognition using a neural prediction model,” in Proc. /€€E Int. Conf Acoustics Speech Signal Processing, 1990, pp. 441-444.

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

[9] Gaurav, Devanesamoni Shakina Devi, Gopal Krishna Sharma, Mahua Bhattacharya, “Development of Application Specific Continuous Speech Recognition System in Hindi”, Journal of Signal and Information Processing, 2012, 3, 394-401.

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

[11] Mayur R Gamit, Kinnal Dhameliya, "Isolated Word Recognition Using MFCC, LPC AND Neural Network", International Journal of Research.

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

[13] Manoj Kumar, So Hyun Kim, Catherine Lord, Thomas D. Lyon, Shrikanth Narayanan, Leveraging Linguistic Context in Dyadic Interactions to Improve Automatic Speech Recognition for Children, Computer Speech & Language, Volume 63, 2020, 101101, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2020.101101.

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

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

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

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

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

[19] Georgios Kontonatsios, Sally Spencer, Peter Matthew, Ioannis Korkontzelos, Using a Neural Network-based Feature Extraction Method to Facilitate Citation Screening for Systematic

Reviews, Expert Systems with Applications: X, 2020, 100030, ISSN 2590-1885, https://doi.org/10.1016/j.eswax.2020.100030.

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

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

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

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

[24] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, “Phoneme recognition using time-delay neural networks,” IEEE Trans. Acoustics Speech Signal Processing, vol. 37, no. 1, pp. 328-339, 1989.

[25] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representation by error propagation,” in Parallel arid Distributed Processing, Volume I: Foundations (J. L. McClelland, Ed.). Cambridge, MA: MIT Press, 1986.

[26] R. W. Prager, T. D. Harrison, and F. Fallside, “Boltzmann machines for speech recognition,” Computer Speech Language, vol. 1, pp. 2-27, 1986.

[27] A. J. Robinson and F. Fallside, “Static and dynamic error propagation networks with application to speech coding.” in Neural Information Processing Systems (D. Anderson, Ed.). New York: Amer. Inst. Phys., 1988, pp. 632441.

[28] Senirkentli GB, Ekinci F, Bostanci E, Güzel MS, Dağli Ö, Karim AM, Mishra A. Proton Therapy for Mandibula Plate Phantom. *Healthcare*. 2021; 9(2):167. https://doi.org/10.3390/healthcare9020167.

[29] Manoj Kumar, So Hyun Kim, Catherine Lord, Thomas D. Lyon, Shrikanth Narayanan, Leveraging Linguistic Context in Dyadic Interactions to Improve Automatic Speech Recognition for Children, Computer Speech & Language, Volume 63, 2020, 101101, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2020.101101.

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

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

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

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

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

[35] Karim, A. M., Güzel, M. S., Tolun, M. R., Kaya, H., & Çelebi, F. V. (2019). A new framework using deep auto-encoder and energy spectral density for medical waveform data

classification and processing. Biocybernetics and Biomedical Engineering, 39(1), 148-159. doi:10.1016/j.bbe.2018.11.004.

[36] A. M. Karim, F. V. Çelebi, and A. S. Mohammed, "Software Development for Blood Disease Expert System," Lecture Notes on Empirical Software Engineering, vol. 4, no. 3, pp. 179–183, 2016.

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

[38] Karim, A. M., Güzel, M. S., Tolun, M. R., Kaya, H., and Çelebi, F. V., "A New Generalized Deep Learning Framework Combining Sparse Auto-encoder and Taguchi Method for Novel Data Classification and Processing," pp. 1–22.

[39] Karim, A.M.; Kaya, H.; Güzel, M.S.; Tolun, M.R.; Çelebi, F.V.; Mishra, A. A Novel Framework Using Deep Auto-Encoders Based Linear Model for Data Classification. Sensors 2020, 20, 6378.

[40] César Montenegro, Roberto Santana, Jose A. Lozano, Analysis of the sensitivity of the End-Of-Turn Detection task to errors generated by the Automatic Speech Recognition process, Engineering Applications of Artificial Intelligence, Volume 100, 2021, 104189, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2021.104189.

[41] Ziping Zhao, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Shihuang Sun, Haishuai Wang, Jianhua Tao, Björn W. Schuller, Self-attention transfer networks for speech emotion recognition, Virtual Reality & Intelligent Hardware, Volume 3, Issue 1, 2021, Pages 43-54, ISSN 2096-5796, https://doi.org/10.1016/j.vrih.2020.12.002.

[42] Peter Smit, Sami Virpioja, Mikko Kurimo, Advances in subword-based HMM-DNN speech recognition across languages, Computer Speech & Language, Volume 66, 2021, 101158, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2020.10115.