Republic of Iraq
Ministry of Higher Education
and Scientific Research
University of Babylon
College of Engineering

An Efficient Speech Recognition System Using Machine Learning methods

A Thesis

Submitted to the College of Engineering / University of Babylon

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Engineering / Electrical Engineering

/Industrial Electronic

By

Ayad Sobhan Obeis Motr

B.Sc. in Electrical Engineering
University of Babylon (2016)

Supervised by

Asst. Prof. Dr. Hanaa M A ALabboodi ***Asst. Prof. Dr. Hayder Mahdi Abdulridha***

2021 A.D ***1443 A.H***

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

﴿ نَرْفَعُ دَرَجَاتٍ مَنْ نَشَاءُ وَفَوْقَ كُلِّ ذِي عِلْمٍ عَلِيمٌ ﴾

(صَدَقَ اللَّهُ الْعَلِيّ الْعَظِيم)

سورة يوسف/ الآية (٧٦)

Dedication

To the Prophet Muhammad and his honorable family, may God's prayers be upon them all

To my master, the owner of the age and time, may God hasten his reappearance

For your courage, hard work and sacrifices to save our Iraq, I would like to dedicate this work to you (martyrs, fighters of the Popular Mobilization and the security forces).

Secondly, I would like to dedicate my work to:

My support and strength in life…… my father.

to

The basis of love and tenderness. my mother.

to

My jewels in life... my brothers and sisters.

to

My family and my children

and finally to

The secret of my success... My best friends.

Ayad Sobhan

2021

Abstract

The use of a speech recognition model has become extremely important due to its multitude of uses in controlling; Our research designed a noise word detection model by applying the features of speech recognition with machine learning with three types of it: deep convolutional neural learning (CNN), the k-nearest neighbor algorithm (KNN), and the support vector machine (SVM). Six words (start, stop, forward, backward, right, left) are used in the English language. I collected words from two equal parts, men and women, in our speech dataset that is used to train and test proposed machine learning networks, the data was collected in various places of the street, park, laboratory and market. The words ranged in length from 1 to 1.30 seconds for 30 people. We obtained the verbal features by Mel frequency cepstral coefficient (MFCC) method, within the properties: standard deviation, mean, the pitch, the windowing, FFT, frequency spectrum and filter bank, which were 20 filters. The data was trained and tested in three methods: SVM was used for training and testing and its efficiency reached 48%, and the Convolutional Neural Network (CNN), advanced deep classification to classify each word from our aggregate dataset as a multi-classes classification, the proposed deep neural network CNN returned 97.06% as accuracy for word classification. 97.4% efficiency for KNN were obtained. Our work is distinguished from many other research that often uses fairly consistent ready-made data of the isolated word type. While our data is collected in different noisy environments and from two types of speech, the isolated word and the continuous word. One of the important problems that we faced: collecting data and differing word lengths, which causes imbalance in the classes, in addition to the difference in pronunciation of each word for the person.

Chapter One

Introduction and Literature Survey

1.1 Introduction

Discourse has consistently been the essential method of correspondence between people. This is because of enormous data that can be moved by the speaker to the audience within a brief period of time. With the new headways, this method of discourse correspondence is being utilized to collaborate among people and machines with incredible progressions like the Alphabet's "Google Assistant" known as "Google Now" already, Apple's "Siri" and Amazon's "Alexa" where these innovations enormously affect the business in like home robotization, handheld gadgets, content inscribing for recordings and sans hands gadgets in auto. Discourse acknowledgment has been created through research works more than 50 years of headways [1]. Speech is the common natural mode of communication between humans, and ordinary people are increasingly using this type of interaction with devices. The property a message contains includes information about the person's identity, age, and language, as well as the message itself. Speech recognition (SR) is a process for automatically distinguishing who is speaking based on personal features embedded in speech signals such as base frequency, pitch, and speaking style. These features are used as distinguishing components between the different speakers [2]. Speech recognition can be achieved by studying the sound wave pattern., the unknown speech sample is processed to determining who it belongs to among the registered speakers or to find out the same spoken word, or infer that the speech sample is unknown. The SR process can be divided into three consecutive steps; pre-processing, feature extraction

and calcification, Pre-processing is the recording of speech with a sampling frequency of, for example, 16 kHz and, according to The Shannon Theorem, a bandwidth limited signal can be reconstructed if the sampling frequency is more than double the maximum frequency meaning that frequencies up to almost 8 kHz are constituted correctly. It has been shown that data transmitted over telephone network, ranging from 5 Hz to 3.7 kHz, is sufficient for recognition so 8 kHz is more than enough. All frequencies below 100 Hz can be removed as they are considered noise. One important part of pre-processing is to remove the parts between the recording starts and the user starts talking as well as after the end of speech. This is done to counter the fact that a SR system will assign a probability, even if very low, to any sound-phoneme combination making background noise insert phonemes into the recognition process. Speech signals are slowly timed varying signals and their characteristics are fairly stationary when examined over a short period of time (5-100 ms). Therefore, in the feature extraction step, acoustic observations are extracted in frames of typically 25 ms. For the acoustic samples in that frame, a multi-dimensional vector is calculated and on that vector a fast Fourier transformation is performed, to transform a function of time, a signal in this case, into their frequencies [3].

1.2 Applications of Speech Recognition

There are many applications of speech recognition for the purpose of verification or identification tasks. Some are suitable for both tasks. Some of SRS apps: Banking service by telephone, Voice mail. Remote access control, Access services for Databases, Information services. Forensics, Intelligent of machines answering and Remote the process of credit card purchases.

1.3 Literature Survey

[4] Patel and Dr. YS. 2010, Using Mel-Frequency Cepstral Coefficients (MFCC) as Feature Extraction with Frequency Sub-Band Decomposition Technique I. Patel and Dr. Y.S. Rao created an efficient speech recognition system in 2010. These improved MFCCs perform better than MFCCs that do not use sub band resolution technology. By focusing more on the signal preprocessing stage, this system can operate more accurately

[5] Anand Singh, D.K. Rajoriya and Vikas Singh. 2012, used Linear prediction cepstral coefficients LPCC as a feature extraction method and ANN as a speech recognition classifier for Hindi double-hybrid words in 2012. They found that consonant-dominated words have a higher discrimination rate than vowel-dominated words.

[6] Deng, L., Hinton, G., & Kingsbury, B. (2013). The authors arranged a session at ICASSP-2013 called "New Types of Deep Neural Network Learning for Recognition of Speech and Related Applications," and they provided an overview summary of the papers that were part of it. The report also included a timeline of the evolution of deep neural networks for acoustic models for voice detection. The overview summary focused on the various methods for improving deep learning, which were divided into five categories: improved types of network architecture and activation functions, improved optimization methods, improved methods for determining deep neural network parameters, and finally enhanced methods for leveraging multiple languages at the same time. When compared to GMM-based models, the overview demonstrated rapid continued growth in acoustic models that use deep neural networks, which may be witnessed on numerous fronts. These acoustic models can also be used to improve performance in other signal processing applications, not just speech recognition, according to the report

[7] Ge, Z., Iyer, A. N., Cheluvaraja. (2017). proposed a new framework for the classification and verification of text-independent speaker recognition based on a feed-forward neural network and Mel-Frequency Cepstral Coefficients (MFCCs) with first and second derivation, and stricter Voice Active Detection (VAD) to ensure a stronger voiced portion of speech is extracted. In the proposed system, Grid search is utilized to optimize neural network structure parameters, and dynamically lowered regularization parameters are employed to prevent training. that ends at a local minimum. The proposed system allows further training with lower costs and the performance is improved in speaker verification with normalization of the prediction score. The TIMIT corpus was employed with an 8K sampling rate. To train and test the categorization performance, the first 200 male speakers are used. The results of the proposed system showed outperform in classification and verification proposed method with less than 6% error rate.

[8] Lee, J., Kim, T., Park, J., & Nam, J. (2017)., researchers used two types of sample-level deep convolutional neural networks, which use small-granularity filters and utilize raw waveforms as input. The first is a basic model made up of convolutional layers and pooling. The second is more complex, containing residual connections, squeeze-and-excitation modules, and multi-level concatenation. It proves that the stamp is genuine. It means that the sample-level models are up to date in terms of performance. The authors compared the properties of learnt filters as well and displayed them as layers.

[9] Boles, A., & Rad, P. (2017). this work, a speaker identification system was proposed. Specifically, the Mel-Frequency Cepstral Coefficient features of human speech was investigated. It was found that using the lower 20 coefficients as input to the SVM provided the best accuracies while training and testing. To come to this conclusion, a dataset from LibriSpeech

was utilized. Additionally, the appropriate length to segment the audio files was analyzed and determined to be 2 or 4 seconds if there is sufficient data and shorter if the amount of data is limited. Using these findings, two systems were created and tested. First, speaker identification systems using 10-person and 40- person sets from the development set in LibriSpeech were trained, both with high accuracies. Second, a speaker
identification system using in-house recorded audio files was constructed and the results of this were given. The ability to handle both regular volume audio and whispering audio shows that the system is robust.

[10] Kaur, G., Srivastava, M., & Kumar, A. (2017). proposed an automated speaker and speech recognition system. The proposed system uses Mel Frequency Cepstral Coefficient (MFCC) as a feature extraction method, and for generating an appropriate feature set, Genetic Algorithm (GA) has been used. In the first phase of the proposed system, features from the speech signal are extracted using MFCC. The second phase is feature optimization using GA. And the last and third phase is classification using Deep Neural Networks (DNNs). The evaluation and validation of the proposed system model have been carried out using 5 isolated words are recorded from 4 speakers, 2 males, and 2 females. To check the efficiency of the proposed work, accuracy, sensitivity, recall rate, precision rate, and specificity parameters are calculated. The authors achieved the average of the accuracy about 97.18%.

[11] Li, X., & Zhou, Z. (2017). In this undertaking, they used 3 a small-footprint keyword spotting (KWS) mission models: Vanilla, DNN, and CNN. All of these 3 models exploit softmax classifier and MFCC feature extraction. The experiment results show that CNN model outperforms the other two models and achieves 31.43% and 66.67% relative improvement with regard to DNN and Vanilla in accuracy; 82% and 94.8% in loss;

18.6%and 72.3% in precision value; and 37.4% and 61.8% in recall value. One limitation of our CNN model is the huge number of multiplies in the second ConvLayer because of the 3- dimensional inputs spanning across time, frequency and feature maps. So our next step is to dive into some new CNN architectures with fewer multiplies and thus make it feasible to work on some power-constrained devices.

[12] Kim, T., Lee, J., & Nam, J. (2019). The authors published a study in 2019 on convolutional neural networks (CNNs) for audio classification, they examined On audio categorization, Wavelet based on CNN and spectrogram based on CNN were proposed, as well as a sample. Using three separate audio domains: music, voice, and acoustic scene sound, CNN-based wavelet audio was created. TensorFlow and Keras were utilized in this project. When using SampleCNNs with the shortest filter and stride sizes, the best results were obtained.

[13] Bunrit, S., Inkian, T., Kerdprasop, N., & Kerdprasop, K. (2019). CNN based method by training from scratch for text-independent speaker identification is proposed in this work. Each sample of signal wave is transformed to spectrogram and use as a gray scale input image to the network. Our proposed method that the spectrogram image use as an input also compare to a case when image of raw signal wave is employed to the CNN model and MFCC based method. Experiments are conducted on five speakers speak in Thai language of which voices are extracted from YouTube.

[14] Algihab, W., Alawwad, N., Aldawish, A., & AlHumoud, S. (2019). examine the neural network usage for Arabic speech recognition using a distributed word representation. Furthermore, the model of the neural network allows robust generalization and enhance the ability to fight the data sparseness problem. Also, the investigation process includes different configuration neural probabilistic model, n-gram order parameter experiment,

output vocabulary, the method of normalization, model size and parameters. The experiment has been done on the Arabic news broadcast, and conversation broadcast. As a result, some improvement has been achieved using the optimized neural network model over the 4-gram baseline model resulting in up to 0.8% absolute reductions and 3.8% relative Word Error Rate (WER). However, different parameters do not have a significant impact on model performance. The paper was based on analyzing first. Then, feature extraction. After that, modeling and finally, testing.

[15] Yang, X., Yu, H., & Jia, L. (2020). The speech recognition system designed for a group of words and used three methods of classification, including CNN, and obtained results of about 92.88%.

[16] Rownicka, J., Renals, S. (2017). the method Very Deep Convolutional Neural Networks (VDCNN) architecture for robust speech recognition has been implemented. The layers of convolutional networks in this suggested model are not fully coupled. When compared to CNN, VDCNN systems have been demonstrated to improve recognition accuracy. When the same was done using MGB-3, the findings showed that the VDCNN-based method produced the best results.

[17] Poudel, S., & Anuradha, R. (2020). Paper has been created for the recognition of Bangla speech. Bangla brief spoken commands were used as samples, and three different forms of CNN were created to recognize them. In a real-life noisy situation, ten various words of utterances situations are considered for the dataset. The model utilizing MFCC had improved accuracy in the experiments, although predicting the raw audio was difficult. Last but not least, this system was able to identify single syllable word with excellent results, but has trouble understanding multi-syllable orders.

[18] Jia, Y., Chen, X., Yu, J., Wang, L., Xu, Y., Liu, S., & Wang, Y. (2021). this paper proposes a method which uses short-time spectrogram statistics to obtain stable pronunciation features for each speaker, and then recognizes speakers with an adaptive clustering self-organizing feature map SOM (AC-SOM) neural network-based adaptive clustering method. A Chinese language database was created, which contains recordings of 100 speakers. An effective feature extraction method was proposed to obtain the speakers' characteristic spectrograms

1.4 Problem Statements

When trying to create a human voice recognition system for the purpose of using it in life matters, for example telephone devices or for the purpose of control, many problems arise. Most of these issues are due to the fact that it is difficult to pronounce a word exactly Same way on two separate occasions. How is the word pronounced easily, with emphasis on different parts of the word, background noise, the nature of the people who utter the word and whether or not they have the mother tongue, etc. Are many variables that are constantly changing. Real-time speech recognition is an important challenge and it is one of the topics that has finally caught the attention of many due to its importance in many fields such as applications that require human-computer interaction and others. Extracting and identifying features related to speech and its nature is also a major challenge. In addition to the decision-making process that takes place using one of the machine learning algorithms. For these reasons, a lot of pre-processing operations are required on the speech signal from the extraction of the most important features and data that make us an easy signal in a correct result.

1.5 Aims of the Thesis

The main objective of this thesis is to expose noise words to develop a machine learning strategy with the quality of normal machine learning such as SVM, knn and a special case of deep learning CNN, which provides an implicit learning method for audio fingerprint recognition in noisy environments and in an attempt to increase the accuracy of the system by using special structures. Neural networks were used for classification and the system's ability to handle noise words. This work goes through the following stages.

1. Get the important features of the word to complete speech recognition.

2. Working within the real word environment and its challenges and overcoming the problems that occur during audio recording.

3. Use the MFCC approach to obtain features extraction.

4. Three methods of training and classification (CNN, KNN, SVM) are used.

1.6 Outline of The Thesis

Chapter One which include a general introduction to speech recognition, an outline of the entire thesis, and a literature review of previous studies in the field of speech recognition

.**Chapter Two** provides the background of speech recognition and deep learning algorithms that have been applied in this field and how these technologies have been grown in recent years

Chapter Three explains the proposed system and all its details including system requirements and tools that are used to implement this system.

Chapter Four contains the simulation of the proposed system and the discussion and analysis of the results that are obtained through this work.

Chapter Five includes the most important work conclusions and recommendations for future work.

Chapter Two

Theoretical Foundation

2.1 Introduction

The main goal of Speech is a form of communication, represented as a signal or waveform that carries information which are sent, stored and processed in a variety of ways different ways in communication systems. A person's speech is a representational signal that changes continuously over time and is represented as x(t) where t is time as figure (2.1). To be processed, the analog signal is converted into a digital signal Represented as a series of samples x(n) where n is the number of samples Speech is a complex signal that contains information about the speaker, idea to be communicated, language, emotion, etc. Having feelings, it makes speech more realistic and natural. Humans use emotions to express their feelings through speech. Humans understand the intended message Through word recognition and audio information [18]. So, there is a need to develop speech systems that can be distinguished and used in a variety of fields. Speech recognition relies on two main aspects: the first is the type of features or parameters that can be related to speech and the second is the choice of the machine learning algorithm for classification. [19]. In this chapter, an overview of the speech recognition system and The relevant theoretical background will be clarified. A signal is simply a quantity that can be measured over a period of time [21]. This amount usually changes over time and that's what makes it interesting. The main goal of verbal signal analysis methods is to analyze the speech signal and evaluate useful parameters. Often, you need to determine the spectral content of the signal; The amount of signal strength located at a given frequency [23]

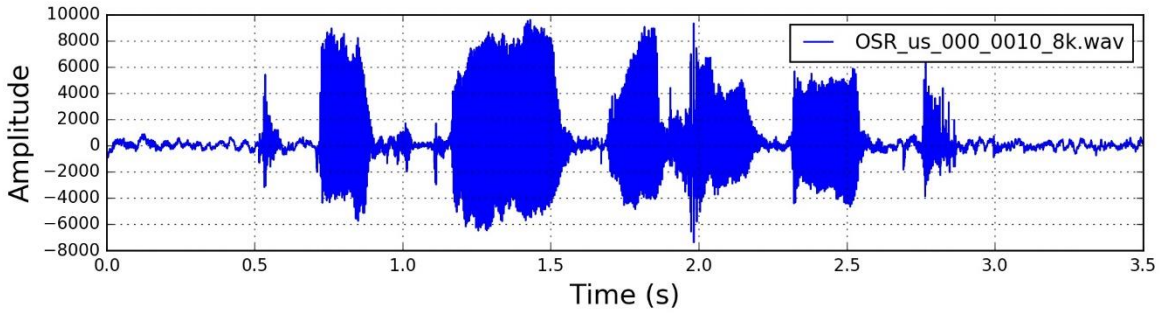


Figure (2.1): speech signal representation in time domain

2.2 Automatic Speech Recognition

Speech recognition software that works automatically is the method of translating a speech signal into a series of words using a computer program and its algorithms. The main goal of speech recognition is to allow machines to recognize sounds and act on them [2]. The ability of a computer to identify “receive and interpret” speech and translate it into readable form or text is known as Speech recognition software. Automatic Speech recognition is a term that refers to the process of recognizing and ability of a computer to understand speech as well as execute an action based on the human's instructions [21]. The phoneme has three parts of processing, identifying the spoken words and knowing the speaker and the third one is the emotional recognition. The words are displayed either in writing or by devices. They are read as a specific command as what have been done in our work, the ability to recognize the speech signal can be divided into three categories based on the type of speech signal itself and its length [33] as shown in figure (2.2).

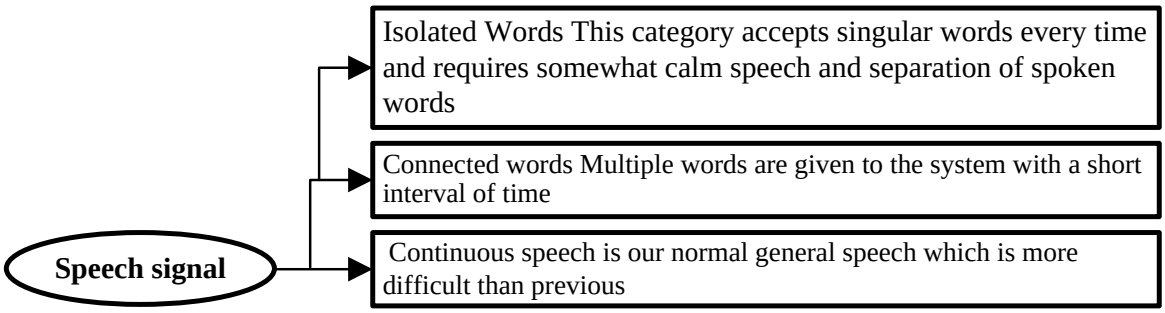


Figure (2.2): The speech types.

Sound recognition has a significant role in developing an automobile control circuit or robotic applications in household protection devices. Sound waves can be defined as a longitudinal wave that travels through an adiabatic compression and decompression phase in most cases. Longitudinal waves vibrate in the same direction as they fly. In sound processing, spectrograms are two-dimensional patterns that show frequency on the vertical axis and time on the horizontal axis, representing signal energy. In general, the movement of air in the vocal tract produces the consonant, which gives rise to the various sounds. The general path for speech recognition is explained in figure (2.3) [34].

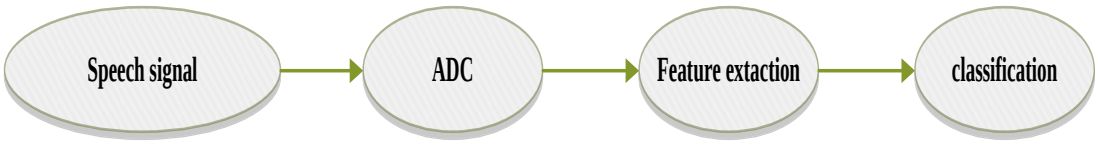


Figure (2.3): The general path for speech recognition.

The general path to completion of speech recognition theory is exhibiting, ADC (analog to digital converter) (spectral shaping), the sound waves converting it into a digital form then pre-emphasis filtering following by feature extraction stage [6]. A sound wave is a characteristic sinusoidal vibration of loud tones that vibrate quickly and with a higher frequency than low tones The vibrating sound energy is converted into electrical energy by

the microphone. The general shape of the sound wave gives an impression of the amount of energy with the amplitude of the signal [7]. The nature of the sound wave is variable frequencies in its being adherent to each other, the frequency components that make up the sound wave are analyzed by the FFT, which highlights it with a diagram called the spectrum diagram. [36].

2.3 Signal Representation

Vibrations in the vocal cords generate changes in air pressure that generates a sound wave. This signal can be represented by time domain and frequency domain [18].

2.3.1. Time Domain and Frequency Domain

The time domain is the domain in which all the signals are represented. Time domain signal can be tested or verified with the use of oscilloscope. In time domain signals are represented by amplitude on Y axis and time on X axis. Frequency domain is useful to do deeper analysis of the time domain signal [31]. Frequency domain helps study frequency contents of the discrete time domain signals as well as continuous time domain signal. The frequency domain signal can be analyzed with the use of spectrum analyzer. In frequency domain signals are represented by power(amplitude)^2 on Y axis and frequency on X axis [36]. Time domain signal can be converted to Frequency domain signal with the use of Discrete Fourier Transform or Fast Fourier Transform(FFT) [38].

2.4. Feature extraction

The process of acquiring various features such as power, tone, and the formation of a sound signal in the speech signal. Prior to the collection of audio signal, the features in that audio signal were first released and later selected [39]. Feature releases are performed to reduce or decrease the amount of data and to select various features from the specified features. The input signal was analyzed by the output feature in which various components such as MFCC, Pitch, sample frequency, sound, volume, etc. were extracted [22]. Audio Signal systems typically use two types of features: perceptual and physical.as show in figure (2.4) [24]

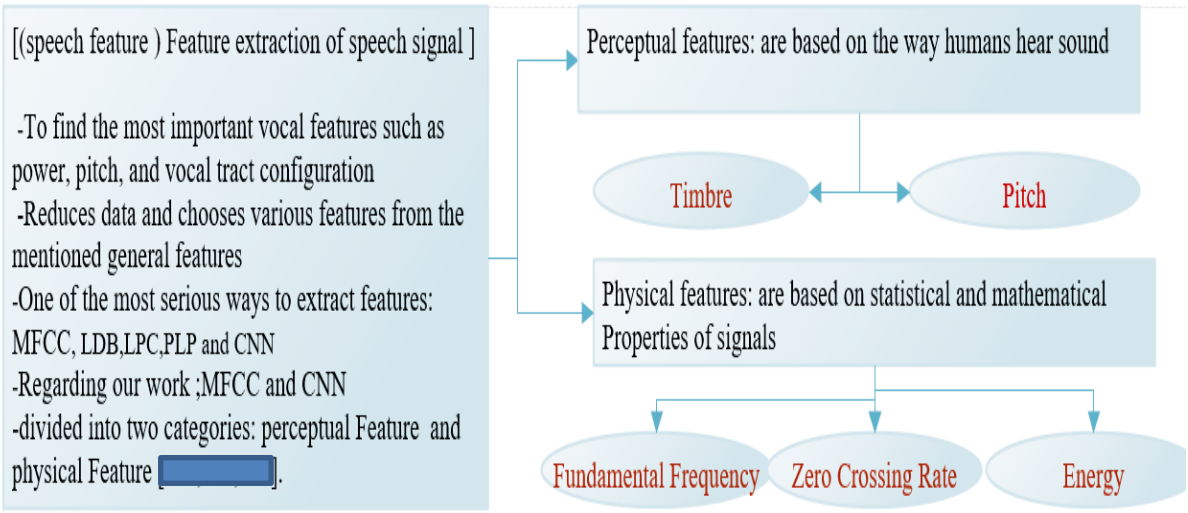


Figure (2.4). types of features [24]

A. Perceptual features

The way humans perceive sound determines perceptual qualities. Pitch and timbre are examples of perceptual properties.

1) **Pitch:** The sound that comes through the vocal cords starting from the neck, where the vocal cords are located, and ending with the mouth. Vocal cord vibration and vocal cord construction can be monitored by brain sensors. This is the sound that we produce, and it can be determined by

the sound and volume of the sound. When producing unnamed sounds, the vocal cords do not vibrate, but remain open, and when the signal sounds, it means that the vibrations also produce what is known as a voice pulse. The wrist is the sum of the sine wave basis, the frequency of harmonics (Size decreases as frequency increases) [32]. The basic frequency of the glottal pulse is known as the vocal cords. Pitch is a perceptual attribute of sounds that allows them to be ordered on a frequency-related scale, or, to put it another way, pitch is the quality that allows humans to judge sounds as "higher" or "lower" in the sense of musical melodies. [24]

Although they are linked, fundamental frequency and pitch are not the same. The first distinction is that pitch is a subjective quality, whereas f_0 is a physical attribute. The *log* of the f_0 values is related to pitch values., with an increase in volume about an octave with all the doubling. The relationship is not a direct logarithmic, as the frequency doubled above 1000 Hz corresponds to a break time slightly below the octave. This relationship is also changing with power. The visible sinusoid volume increases sharply when the sinusoid is above 3000 Hz, and a sinusoid with a frequency of less than 2000 Hz is seen as a decrease in pitch as the power increases [24].

2) Timbre: timbre is a feature which is used to differentiate a particular musical sound, even when they have the same pitch and loudness. Timbre depends primarily upon the frequency spectrum; it also depends upon the sound pressure and the temporal characteristics of the sound [28]

B. Physical features

Physical characteristics of signals are dependent on statistical and mathematical properties. Fundamental frequency, Zero-Crossing Rate (ZCR), and Energy are examples of physical characteristics [24].

1) Fundamental Frequency: In a sound or multimedia database The characteristic f_0 is critical for discriminating between sound signals. The fundamental frequency, often referred to simply as the fundamental, is defined as the lowest frequency of a periodic waveform. In terms of a superposition of sinusoids (e.g. Fourier series), the fundamental frequency is the lowest frequency sinusoidal in the sum. The Fundamental frequency is usually abbreviated as f_0 to indicate that it is the lowest frequency counting from zero. The fundamental frequency is defined as: $f_0=1/T$ [4].

2) Zero Crossing Rate (ZCR): ZCR is the average number of times a signal value falls to zero axes. It is easy to measure ZCR signals. Frequently a cyclic sound has a lower value than a high-frequency sound. To quantify approximately the basic frequency, ZCR is used for marked signals. [24].

3) Energy: One of the most significant physical features is the measurement of energy based features, that is, it is a measure of how much the signal is at a given time. The energy will be used to detect the silence of the signal. The signal energy is often calculated over a short period of time as a basis, due to windowing the signal with a certain amount of time, as well as the surface of the samples and the average load. The square root of this result is what the engineering value is known as the root mean square. [4].

2.4.1 Mel frequency Cepstrum coefficient (MFCC)

The MFCC method is used to extract different aspects of a person's voice. Represents the human voice's short-term energy spectrum. Based on the cosine line variation of the log power spectrum on Mel scale, it is utilized to determine the coefficients that reflect the cepstral frequency of these

coefficients. The Mel Scale is a logarithmic transformation of a signal's frequency. The core idea of this transformation is that sounds of equal distance on the Mel Scale are perceived to be of equal distance to humans. the mel scale is a scale of pitches judged by listeners to be equal in distance one from another [25]. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 mel. Below about 500 Hz the mel and hertz scales coincide; above that, larger and larger INTERVALs are judged by listeners to produce equal pitch increments. Voice functions that are present in the MFCC by splitting the speech signal into frames, and these must be inserted into windows and Fourier transforms of the signal [4]. Mel frequencies are obtained using a Mel filter, or an upper-pass filter in the transformed signal. Summing up, it should be noted that changing the shape using DCT reveals small-cepstral factors, like the features of the human voice. To perform extraction, MFCC has five process steps that are sequentially processed, namely Frame Blocking, Windowing, Fast Fourier Transform, Mel-Frequency, and cepstrum [25]. The Steps of MFCC Method:

- **Frame Blocking and Windowing:** Because speech is time-varying in the sense that the vocal-tract configuration changes over time, an appropriate set of predictor coefficients must be calculated adaptively over short time frames (usually 10ms to 30ms) in which time-invariance is assumed. As a result, the continuous voice signal is divided into frames of N samples, with M (over lapping) (M< N) between consecutive frames, normally the overlapping region ranges from 0 to 75% of the frame size. Speech is a non-stationary signal due to differences in phoneme spectral characteristics, changes in prosody, and random variations in the vocal tract [26] .Windowing **:** At this stage, the sound signal that has been divided into

several frames is carried out by windowing to minimize signal discontinuity, the windowing

$$w(k) = 0.54 - 0.46\cos\left(\frac{2\pi k}{N-1}\right). \qquad (2.1)$$

where N is the length of the filter and $k = 0.1. \ldots . N - 1.$ [26]

- **Pre-Emphasize and Normalization**:

Pre-emphasis: is a technique used in speech processing to enhance high frequencies of the signal, Pre-emphasis is practice in speech research to apply a pre-emphasis filter before windowing in order to boost high frequency. Usually a pre-emphasis filter is a first-order high-pass filter by 6" " dB /octave, [27]

$H(Z) = 1 - az^{-1}$ with α in [0,1] (2.2)

Benefits of Pre-Emphasize Filter: The pronounced parts of the natural The spectral slope of voice signals is negative. (reduction of 20 dB for ten years due to the physical condition of the speech production system. Hearing is more sensitive than the 1-kHz region of the spectrum. The preemphasis filter expands this spectrum area. This assists the spectral analysis algorithm in modeling the most important visual objects. [27]

Normalization: Because differing recording levels might cause the volume to vary dramatically from word to word, normalization is a method for correcting the loudness of audio files to a standard level. Sound samples with varying loudness and maybe some DC offset were recorded [27]

- **Fast Fourier Transform (FFT) and Spectrogram** : FFT is the process of converting each frame sample N from the time domain to the frequency domain, calculated using

$X(k) = \sum_{x=0}^{N-1} x(n)\, W_N^{KN}$ where $k = 0.1.2. \ldots N-1$ ... (2.3)

Consider only absolute values of X_k which are Complex numbers integral[38]. A spectrogram is a method of using discrete Fourier transforms (FFT) to translate a one-dimensional time sequence, into a two-dimensional image containing the same information, but organized in a way that makes locality meaningful in two dimensions rather than just one. Each column of the image represents the magnitude of the Fourier transform of the time sequence taken in a different contiguous window of time; the image as a whole represents a two-dimensional function from time and frequency to intensity rather than a one-dimensional function from time to absolute amplitude. Thus, in the resulting image, it is possible to directly see which frequency components change and how the frequency spectrum of the audio changes with time [24]

- **Mel filter or triangular band pass filter [Mel - Frequency Wrapping]**: Mel-filter bank-a set of triangular filter banks with a Mel scale. The 10 filters below 1000 Hz can be divided linearly, while the remaining filters at the top are 1000 Hz and represent a separate logarithm. To create a filter bank with M filters. The filter triangle must be used, and, with respect to the center frequency f(M)), M=1,2..., M, M, often 22nd-26th". (the number of filters, as well as the number of critical bands). Using MFCC as an acoustic element, and will not depend on the difference in the input range. In addition, the computer's capabilities may be limited. The Mel scale is used to map the signal frequency scale to a logarithmic scale for frequencies higher than 1kHz. This scale makes the spectral frequency of the signal following

human hearing. This scale is defined by Stanley Smith, John Volkmann, and Edwin Newman on (2.4) [29]

$$\text{Mel}(f) = 2595\log10(1 + f/700) \quad (2.4)$$

Mel filter bank, this is very important because one of the following reasons: Use of a fine-frequency scale-a scale of perception that helps to simulate the work of the human ear. This is consistent with much better resolution at low frequencies, as well as at high ones. Using a triangular filter bank helps capture the energy in each critical band, and it gives you a rough estimate of the volume and shape, and this changes the harmonic structure of the spectrum. In theory you could manipulate on raw DFT bins, but then you are not reducing the dimensionality of your features - this is the whole point of doing filter-bank analysis, to capture the spectral envelope. [29]

- **Spectrum [Result After FFT] and Cepstrum [After DCT]:** A spectrum is a 1D plot and a spectrogram is a 2D plot. Cepstrum is derived from the word "spectrum" when the first four letters are reversed, suggesting a modified spectrum. The independent variable associated with the cepstrum transform is known as "quefrency," and because it is so closely tied to time, it is appropriate to refer to it as "time." The cepstrum stage converts log Mel spectrum into the time domain by using Discrete Cosine Transform DCT (2.5) [32]. Spectrum [After FFT] and Cepstrum [After DCT]

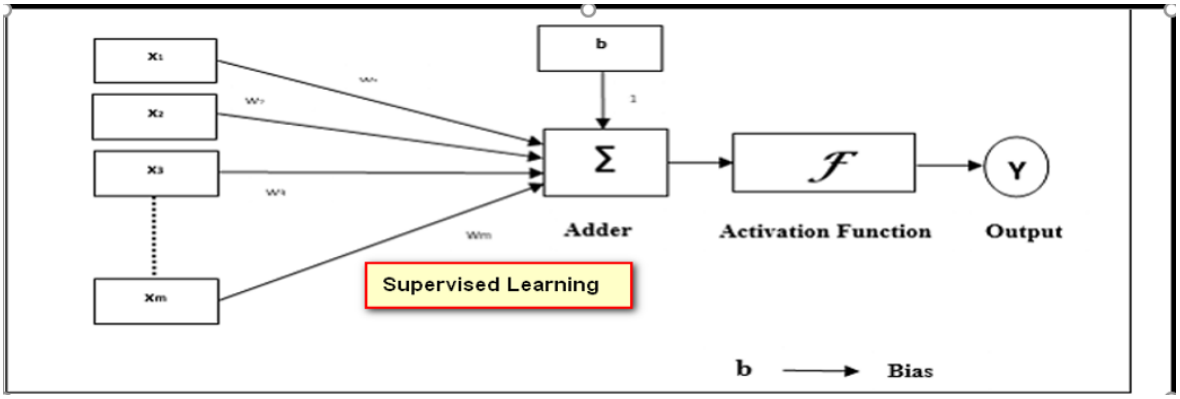
$$C_m = \sum_{k=1}^{N} E_k \cdot \cos\left[m * \left(k - \frac{1}{2}\right)\frac{\pi}{N}\right] \; m = 1.2. \ldots . L \quad (2.5)$$

where N is the number of spectrum coefficients and L is the number of the Melscale cepstral coefficients (usually the maximum number of L is 13). E_k is the log energy. [23]

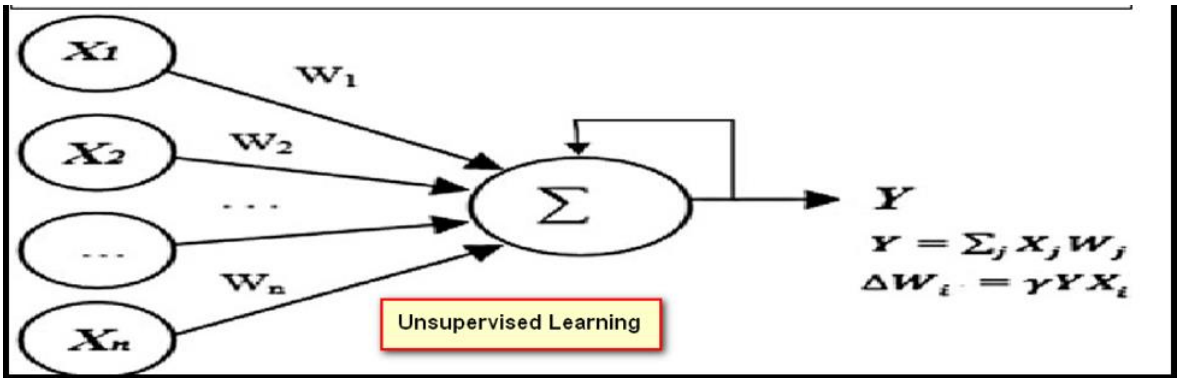
2.5 Machine Learning (ML)

Machine learning is one of the branches of artificial intelligent that are interested in providing the necessary algorithm, applications and frameworks that make computers able to learning by achieving more precise prediction and valuable results from analysis the input data. Applying machine learning algorithms to large databases is called data mining. Machine learning ML also used to find solution to many problems in speech recognition and robotics. The statistical theory has been used in machine learning to construct the mathematical models because the basic task of ML is inference from a sample. The role of ML is two folds: First, in training the efficient algorithm is used to optimize a problem as well as to store and process, its own data. Second, the model is learned for one time. The representation and algorithmic solution for inference needs to be effective [30].

The predictive accuracy of any algorithm is measured by complexity which is mean space and time. Machine learning algorithms are classified into Supervised Learning, Unsupervised Learning as figure (2.5) and Reinforcement Learning. In supervised learning, the input and output need to exist as well as the accuracy of the predictions during the algorithm training. In the case of unsupervised learning, there is no output need in training algorithm. In other words, unsupervised have only input data. The supervised algorithms are used to observe the complicated learning system [33]. Figure (2.6) show the main Machin learning methods [34]. In this thesis, the SVM and KNN are employed for extraction proposed, in addition to CNN.



(a)



(b)

Figure (2.5): (a). Supervised machine learning, (b). Unsupervised machine learning [41]

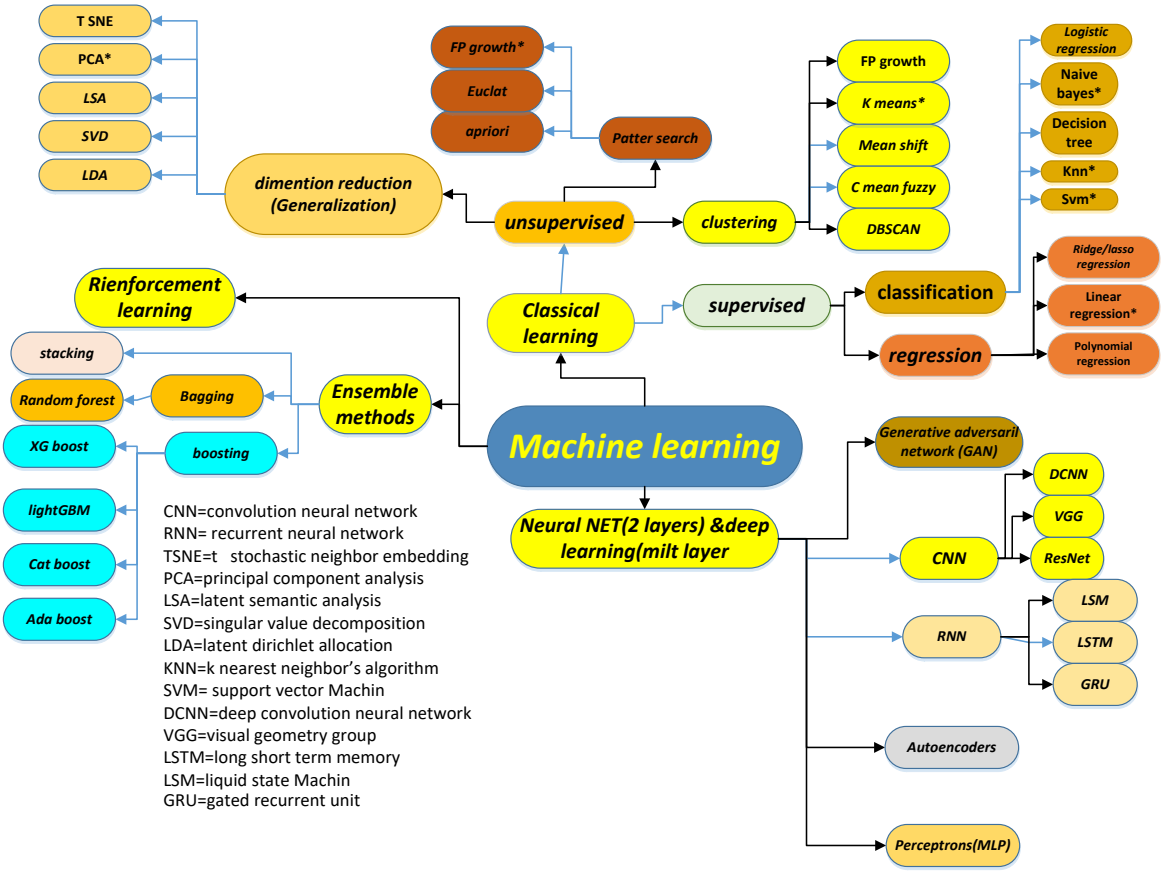


Figure (2.6). The main Machin learning methods [41,33]

2.5.1 Support Vector Machine (SVM)

Supportive One of the most widely used machines is the vector machine. classifiers based on the linear discrimination function, and it is very suitable for binary classification. Its preference is due to the excellent software packages that have been developed in recent times. SVM uses two types of data, linear data separable and non-linear data separable .as figure (2.7) [33]

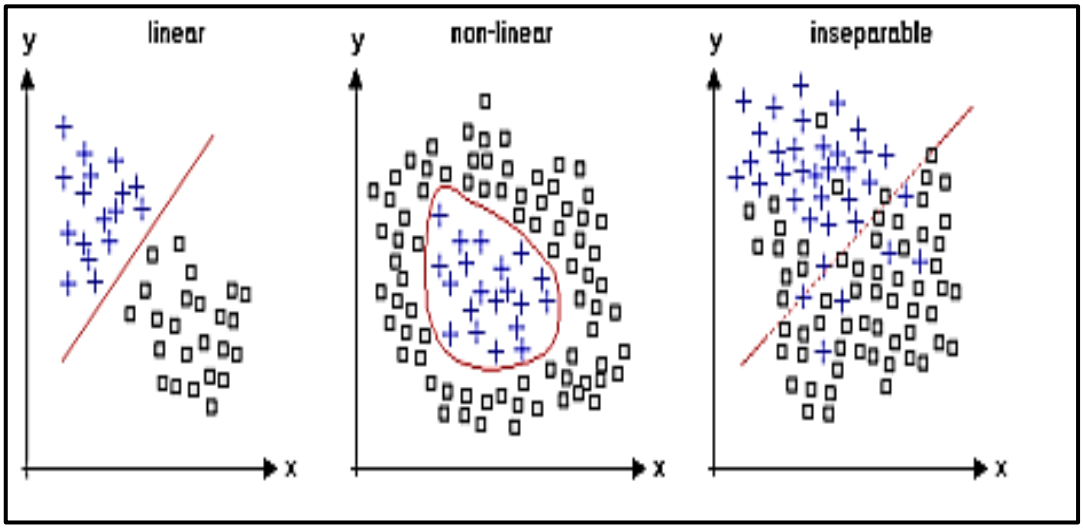


Figure (2.7): Linear, nonlinear and inseparable data

In a linear SVM classifier model, the data is separated by a visible space. SVM Principle Finding a hyperplane that divides data into two classes. In drawing the plane, the primary focus on superlative is on maximizing the distance from the hyperboloid to the closest data of any class. In a non-linear SVM classifier, the data set is generally scattered. It is not appropriate to plot a straight linear plane to separate this data. [35]

To construct an optimal hyperplane, it must firstly compute the weight vector by the following Equation (2.6) which is a linear combination of support vectors:[35]

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \qquad (2.6)$$

w is a weight vector, x is input vector and y_i =+1 or -1. Then in the feature space, the best hyperplane is defined by: [35]

$$(w \cdot x) + b = 0 \qquad (2.7)$$

Where x is the row vector of corresponding speech sample, w is the weight vector as mention in Equation (2.6) and b is the bias [37].

2.5.2 Theory of K-nearest neighbors KNN

K-Nearest Neighbor KNN is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity [2]. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification. [47]

2.6 Deep Learning (DL)

A part of machine learning is Deep Learning. Deep Learning is influenced by the structure and potential of the Artificial Neural Network, a human neuron. Artificial Neural Networks are systems that, without an explicitly specific program, learn to take actions based on examples. The architecture of ANN consists of three layers, namely input, output and one hidden layer. The foundation of deep learning is ANNs [42]. Since the 1950s, a little subset of Artificial Intelligence (AI), frequently called Machine Learning (ML) has changed a few fields over the most recent couple of decades. Artificial Neural Networks (ANN) is a subfield of ML and it was

this subfield that produced -DL. Since its beginning, DL has been making ever bigger disturbances and demonstrating extraordinary accomplishment in every application area. DL utilizes either profound models of learning or progressive learning draws near. [42]

2.6.1 Convolutional Neural Network (CNN)

Convolutional networks were the beginnings Hubel and Wiesel who found that a single network architecture could reduce complexity in the feedback neural network when studying neurons used for local sensitivity and orientation selection in the cerebral cortex of cats. CNN is often used with image processing that requires a two-dimensional matrix containing features and may be three-dimensional, the pixel values are in the horizontal and vertical coordinate indicators. CNN is a neural network model. Its architecture has three main ideas as Fig (2.8). Each one of them has the susceptibility to improve speech recognition performance. [43]

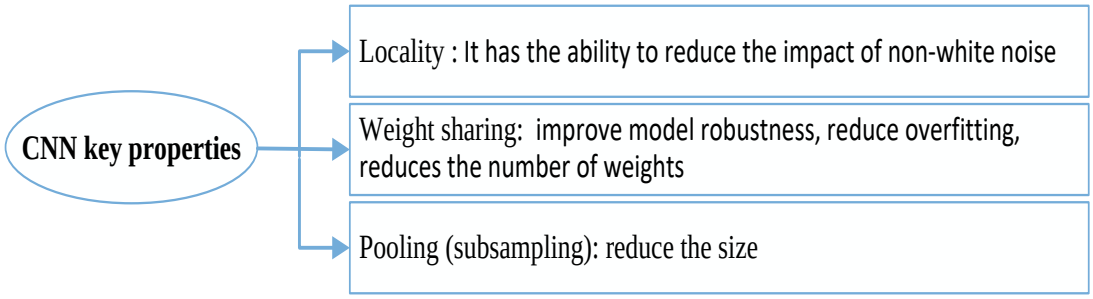


Figure (2.8): Architecture of CNN properties [44]

CNN has a filter that shifts over the image to produced feature map at convolution layers, through this window or filter, the weights of the network can identify the different features of the incoming image. The activation function decides if a particular feature is present at a particular location in the image. Usually uses a lot of filters over the image to find the necessary features [44]. CNN is often called the local network because the individual

units computed in a specific location of the window depend on the local area that the window is currently looking at [45]. Convolutional architecture is coordinated by three main layers arranged in the forward feed structure. The convolutional layer for feature extraction, sub-sampling layers, the aggregation(pooling) layer, to reduce the dimensions of the input data and the output which a fully-connected layer for final classes prediction [47]. linear filter and a nonlinear activation function, One of the most important elements. In a convolutional layer, each plane is connected to one or more feature maps of the preceding layer .an activation function is applied on to the result obtain the plane's output. The plane output is a 2-D matrix called a feature map; this name arises because each convolution output indicates the presence of a visual feature at a given pixel location.

A convolution layer produces one or more feature maps. Each feature map is then connected to exactly one plane in the next sub-sampling(pooling) layer [44]. Sharing of weights and location are essential to the properties of the pooling, feature values computed at different locations are grouped together and represented by a single value in order to minimize differences in the extracted features along the frequency dimension when the input patterns are shifted. This is important when dealing with the small frequency shifts common in speech resulting from different path lengths vocal. CNN used activation faction as figure (2.9) and table (2.1) explain properties of CNN layers.

After converting the audio into a Spectrograms as an image, let's say we have an image of dimensions N = 6 * 6, and the filter was F = 3 * 3 kernel filter, in example below and after making the convolution (*) the result is 4 *4 according to the formula Out =N-F+1.

N=6, F=3 with padding P=0 and Strided convolution S=1, the out =4=n-f+1. table (2.1) explain some properties of CNN layers.

Table 2.1: Illustrates the properties of CNN layers. [44,47]

Convolution layer	Pooling layer	Full connected layers
Filters are included to find features of an image.	Reduce dimensionality.	Aggregation formation from final feature.
The filter consists of small kernels (number of kernels) one bias per filters.	Maximum or average area is extracted.	General final classification.
For every value of feature map must apply activation function.	Sliding window approach.	Parameters full connected, (number of nodes, activation function; usually changes depending on role of layers. RELU used for aggregating information, and SOFTMAX for producing final multi-classification)
parameters of CONV layers ,(size of kernels ,activation function ,stride, padding and regularization type and value)	Parameters of pooling, (stride and size of window).	

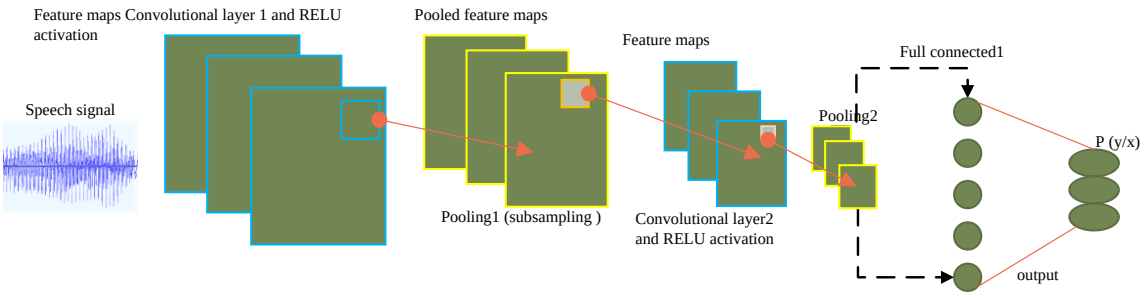


Figure (2.9). CNN layers and Activation Function.

2.6.2 Activation Function

In a neural network, which is a weighted sum of its inputs as well as biases, an activation function will be calculated and used to determine

whether the neuron will cause a fire or not. It processes the data using gradient processing, mainly gradient shutter release, and then outputs a neural network that contains information about the settings. In some literature, this playlist will be referred to as a transfer function. [48]**.** The activation function can be linear or nonlinear, depending on the function it represents, and can be used to control the flow of neural networks, as in all areas ranging from object recognition and segmentation, speech recognition, speech, scene understanding and description to machine translation systems-test-to-speech (TTS), cancer detection systems, fingerprint detection, weather forecast, autonomous cars and other areas to name just a few. In most circumstances, the affine transformation provides a linear translation of an input function to an output, as executed in the hidden layers before the final prediction of class score for each label. The input vectors x transformation is given by [48]

$$f(x) = w^T x + b \qquad (2.8)$$

where $x =$ input, $w =$ weights, $b =$ biases.

Furthermore, the mappings from equation (2.8) generate linear outcomes, necessitating the use of the activation function to turn these linear outputs into non-linear outputs for further computing, particularly to discover patterns in data. These models' outputs are given by [48]:

$$y = (w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b) \qquad (2.9)$$

Each layer's outputs are fed into the next consecutive layer until the ultimate output is obtained in multilayered networks like deep neural networks, though they are linear by default. The type of activation function to use in a given network is determined by the intended outcome. These activation function (AFs) are transfer functions that are applied to the linear model outputs to yield transformed non-linear outputs that can be processed further. After applying the AF, the non-linear output is given by [48]:

$$y = \alpha(w_1x_1 + w_2x_2 + .. + w_nx_n + b) \quad (2.10)$$

Where α is the activation function.

The input layer receives data through multiple neural network training methods, including images, videos, text, voice, sounds, and numerical data, while the hidden layers are usually composed of convolutional layers and pooling, the convolutional layers recognize local patterns and insert of inputs from the previous layers and bring them to the same elements as the image, while the integration layers intertwine logically the comparative features into one. The output layer reflects the network effects commonly affected by AF, such as segmentation or predictions, and related opportunities. The role of AF in network structure determines where it is located in the structure; for example, when AF is placed behind hidden layers, it translates a readable map into forms that are not in the distribution line, and when placed in the output layer, it makes predictions. [48]

The deep architectures are composed of several processing layers with each, involving both linear and non-linear operations, that are learned together to solve a given task. These deeper networks come with better performances though common issues likes vanishing gradients and exploding gradient arises, as a result of the derivative terms which are usually less than 1. With successive multiplication of this derivative terms, the value becomes smaller and smaller and tends to zero, thus the gradient vanishes. Consequently, if the values are greater than 1, successive multiplication will increase the values and the gradient tends to infinity thereby exploding the gradient. Thus, the AFs maintains the values of these gradients to specific limits. These are achieved using different mathematical functions and some of the early proposals of activation functions, used for neural network computing were explored by Elliott, 1993 as he studied the usage of the AFs in neural

network. The compilation of the existing activation functions is outlined with the advantages offered by most of the respective functions as highlighted by the authors as found in the literature [48]

2.6.3 Activation Functions kinds

A. Softmax Function

Another form of activation function utilized in neural computing is the Softmax function. It's used to make a probability distribution out of a set of real numbers. The Softmax function returns a range of values between 0 and 1, with the probability total equal to 1. The connection is used to calculate the Softmax function.

$$f(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad \text{, where x = input data to AF} \qquad (2.11)$$

In multi-class models, the Softmax function returns probabilities for each class, with the target class having the highest likelihood. [48]

B. Rectified Linear Unit (ReLU) Function

Nair and Hinton (2010) proposed the rectified linear unit (ReLU) activation function, which has subsequently become the most extensively used activation function for deep learning applications with state-of-the-art results. The ReLU is a faster learning AF that has shown to be the most popular and successful. In deep learning, it outperforms the Sigmoid and Tanh activation functions in terms of performance and generalization. The ReLU represents a nearly linear function and so keeps the features of linear models that make them easier to optimize using gradient-descent methods The ReLU activation function applies a threshold operation to each input element, setting any values less than zero to zero, resulting in the ReLU. [48]

$$f(x) = \mathrm{m}\ (0.x) = \begin{cases} x_i. & \text{if } x_i \geq 0 \\ 0. & \text{if } x_i < 0 \end{cases} \qquad (2.12)$$

This function corrects the values of the inputs that are less than zero, driving them to zero and removing the vanishing gradient problem that plagued previous forms of activation functions. Within the hidden units of the ReLU function has been used. Deep neural networks with additional AF, utilized in the network's output layers, with common applications in object categorization and speech recognition**.** [48]

2.7 Confusion matrix:

Confusion Matrix is a summary used to measure the performance of a classification model in machine learning, figure 2.10

True positive (TP): is a correctly assigned positive class label.

False positive (**FP**)**:** is an incorrectly assigned positive class label.

False negative (FN): is a document that actually belongs to the category, but was not recognized as such by the classier**.**

True negative (**TN**)**:** is a document not belonging to the class, and the classier correctly did not place in the class.

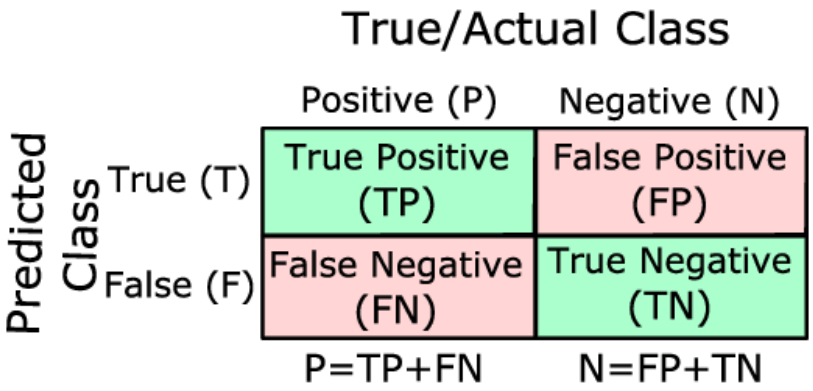


Figure (2.10). Confusion Matrix construction [49]

Accuracy (**AC**): The model's correct (or incorrect) forecast rate on a database. In most cases, it is assessed using an independent test set that is not used during the learning period. [49]

$$\text{Accuracy } (AC) = \frac{TP+TN}{TP+TN+FN+FP} \quad (2.13)$$

Precision (**P**)**:** is the likelihood that an instance allocated to a class actually belongs to that class. Precision is calculated by dividing the number of true positives by the sum of true and false positives. [49].

$$\text{Percision } (P) = \frac{TP}{TP+FP} \quad (2.14)$$

Recall (**R**)**:** the number of true positives divided by the total number of true positives and false negatives. [49]

$$\text{Recall}(R, \text{ true positive rate}) = \frac{TP}{TP+FN} \quad (2.15)$$

2.10 Training Parameters

In deep neural network training, the parameters that are utilized to tune the training are as follows: Training parameters are a type of parameter that aids in improving accuracy and performance. It's time for some practice. The parameters for training are as follows: [50]

a) Batch Size

The batch size is the number of datasets delivered to a network every iteration; the number of batch sizes influences model training time; if a complete dataset is passed through the network on each iteration, training time may be reduced. However, as the network becomes more extended to the dataset that we already have, the accuracy will suffer. [50,51]

b) Epoch

Epochs are the number of times the entire dataset is sent to the network; for example, if the dataset is 200 bytes and the batch size is 10, one epoch will take 20 iterations to complete. Epochs allow you to train with the same datasets over and over again. [51]

c) Loss

A value that is calculated after each iteration to determine the error in question in the loss. [51]

d) Learning Rate

The learning rate is the component that will be used by the optimizer to change the weight function. With a low learning rate, which means that the model will take longer to complete, but it will be more accurate. Higher education is often a good place to start learning. With a lower learning rate, it will help us improve accuracy when the model reaches an acceptable level of accuracy. [51]

Chapter three

The proposed system

3.1 Introduction

The study approach, solution methods, development, and algorithm provided to handle the complex problem of speech recognition are presented in this chapter. The steps of the overall research technique are depicted in the flow chart below.

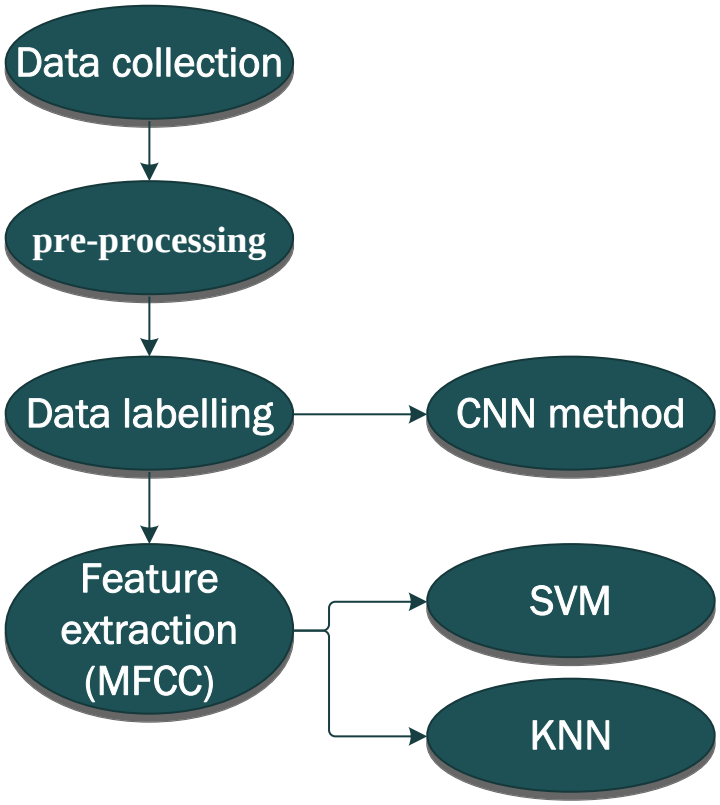


Figure (3.1): The general structure of the thesis

3.2 Data Collection

In this thesis, six words (start, left, right, backward, forward and stop) for thirty people are collected. Each person uttered these six words once and in English languages. The people were half men and the other half were women. The six words were recorded in different locations. In the laboratory, on the street, in the garden, in the market, in places free from noise and in other noisy areas. The recorded words are different in length of rely on the word itself. Also, some words differ in length from one person to another depend on the speaker himself and his pronounce ways. These conditions made the work more complicate especially for the training and classification process. Audacity and Adobe Audition were used to display the signal and determine the start and end point in addition to seeing the volume of silence between the spoken lettersand word. The input words have length in ranges from (1s to 1.35 s). The data have been classified into seven classes according to the words that required to conduct to the control circuit. The recorded words classes are forward class, backward class, start class, stop class, left class, right class and unknown class. The unknown class includes the words intended to complete the training process and these words are (yes, no, are, is, friend, hello, he and she).

The process of record data was a difficult task in terms of providing people and record mechanisms. One of the problems was the standardization of the data type in terms of mono or stereo, this problems of recorded type appeared to us when extracting features and in reading audio files in CNN.

What distinguishes our work from other works in this field is the data that we collected ourselves in different recording areas with high, medium and little noise, for example in the market, home, garden, and laboratory. As for the other research, it mostly uses ready data, simple and clean, recorded in

noise-free areas and secondly we used two types of words (figure 2.2) isolated words such as (stop and start). And connected words such as (backward) (figure 3.2), and other research have used separate words only for the most part, can see kind in figure 3. The third advantage of our work was that we were able to obtain good efficiency compared to other research by working the algorithm that deals with various data and filtering them in a way that enables the classifier to distinguish words with high accuracy.

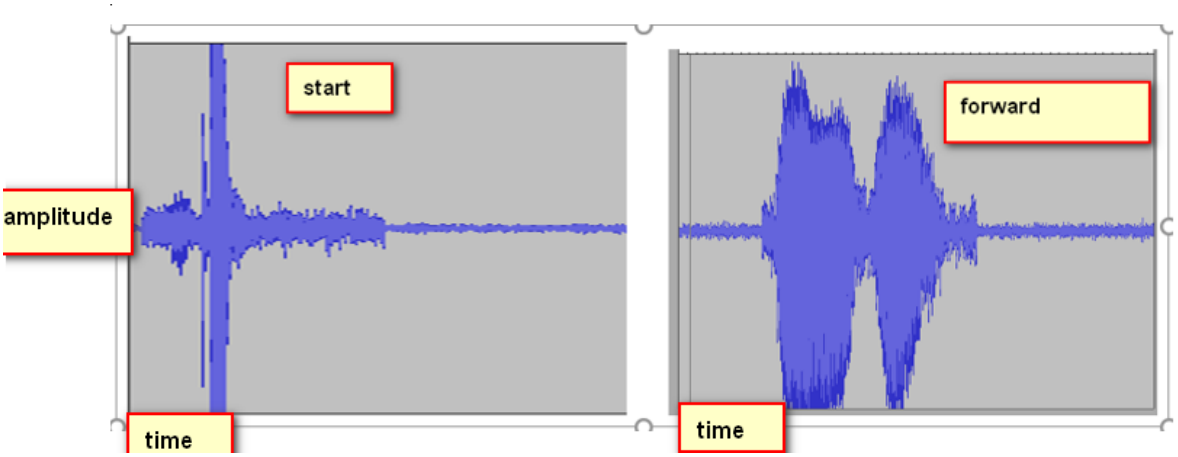


Figure (3.2): start and forward in Audacity program

As the data was collected in several locations, each place had its own type of noise. Some speech samples had very little noise (in-room recordings), whereas others had a lot of noise (on road recordings). Because the noise varied for each speech sample, it was necessary to develop an accurate system. Each speaker delivered the samples in a batch of one voice note., so they had to be as noise-free as possible before being divided into separate files with each file consisting of a group of similar words for different people. These processing techniques have the potential to MATLAB. One of the important problems faced by the research is that the neural network works. With fixed and equal dimensions of the training data and therefore requires a specific arrangement that works to prepare the data in

the appropriate format that ensures the equality of dimensions for each word that will be tested and trained on. We also worked on calculating the SN ratio. The ratio was very bad, until it reached (-9) as average, which indicates the amount of high noise relative to the signal as show in figure (3.3) SNR for right and forward words

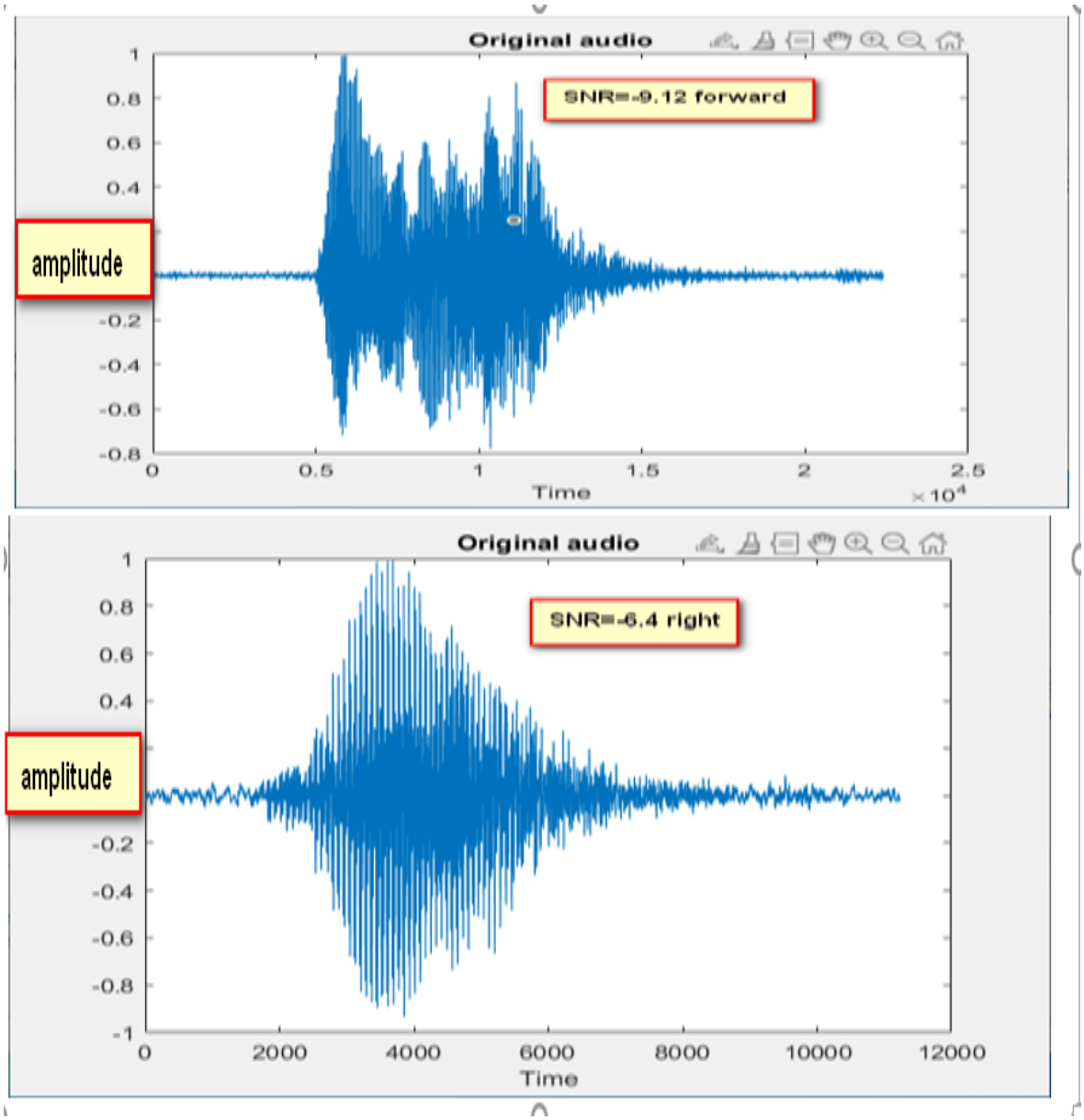


Figure (3.3): SNR for right and forward words

3.3 Data Labelling

The fact of speech recognition is knowing the entered word in order to transfer it to any practical application. The process of labeling the data was done based on the ability of the neural network to deal with this data. The words we have are six words in order (backward, forward, left, right, start, stop). Each word and the like is arranged in a special file that bears the name of the words inside it, and therefore we will have six files, each file bearing the name of one of the six words, and each file contains 30 registered words from thirty people. In the feature extraction process by MFCC these files were read according to the naming pattern of each file. As for the training and classification process, which relied on MFCC as an input to KNN and SVM, each of the six files was given a special number representing the special word in the pattern 1, 2, 3 ..... to 6 as figure 3.4. Added as a column in the characteristics matrix represents the label for each word during the training and classification processes. As for the method of classifying and training CNN, it was named as it is according to the name of each word

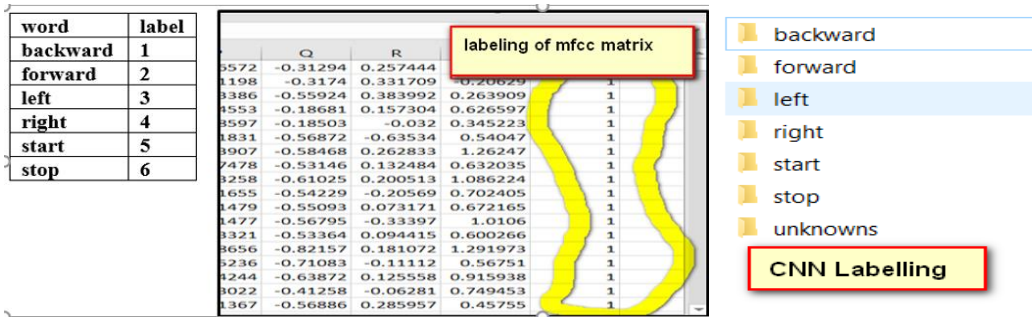


Figure (3.4): Labelling of KNN, SVM and CNN data set training

3.4 pre-processing

One of the general steps pre-processing: remove silence, normalization and deduction of end point, we remove silence by selecting the maximum amplitude about 0.3, and for normalization, an equal amplitude is equal to 1. Determining the ending points was done by defining a fixed time based on the longest words in the data set as show in figure (3.5) and figure (3.6).

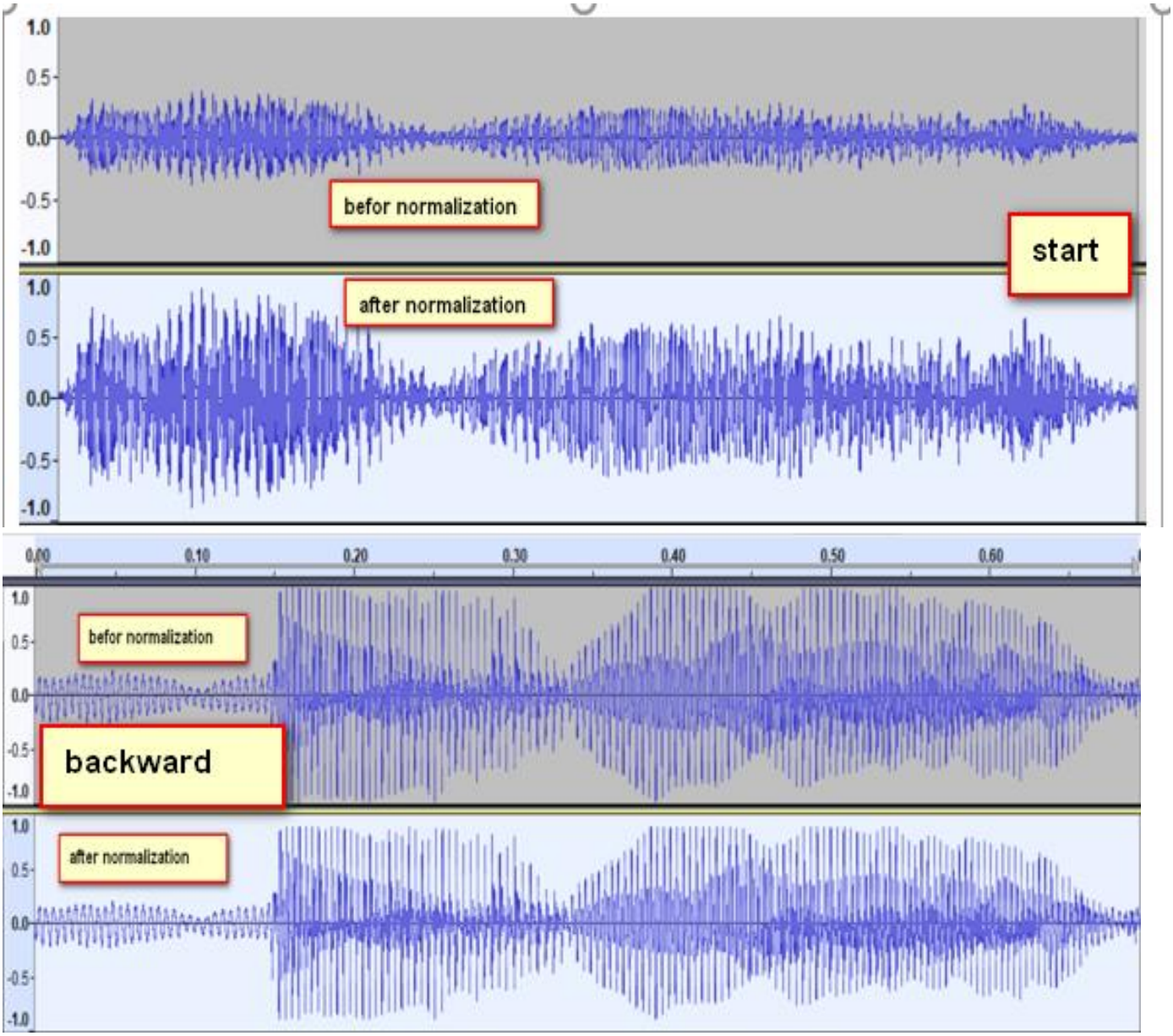


Figure (3.5): Normalization for start and backward

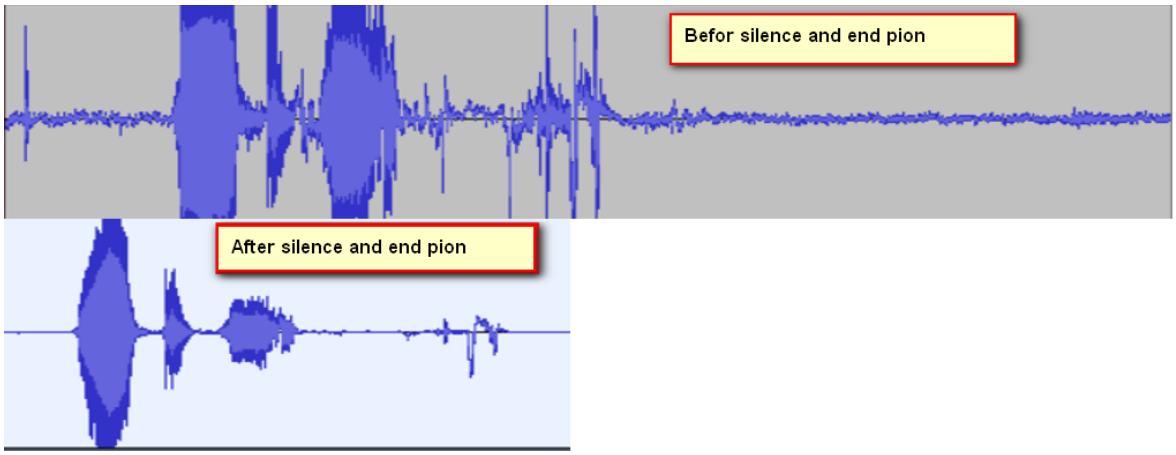


Figure (3.6): pre-processing, silence and end pion for Backward

3.5 Feature Extraction

An important step in the process of speech recognition is to find the main features of each word. In our work, we used the MFCC theorem to find the features of speech, the summary of which is shown in the Fig 3.7.

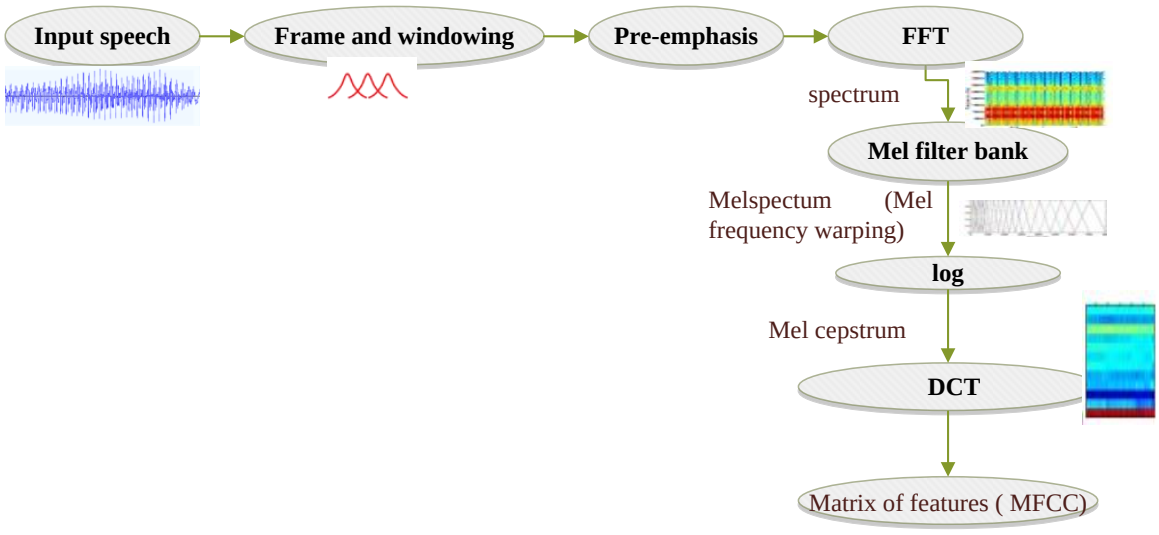


Figure (3.7): MFCC steps.

1. Use MFCC to extract exclusive speech features
2. Perform short term power spectrum of speech

3. By MFCC calculated The coefficients that represent the frequency cepstral
4. Use M=2595log(1+f/700) to convert frequency to the Mel scale
5. For each 1000 Pitch represents 1000 Hz tones,
6. The best representation of speech is Mel frequency warping that is produced from set of filter bank

Table (3.1): MFCCs computation parameters

Frequency sampling (fs)	16kHZ
Silence Removal	0.3 as maximum of amplitude
frame duration	25ms
Overlap Length	(10ms)
Number of filters bank	20

3.6 Framing, Windowing, FFT and DCT

To complete the process of extracting audio features, several steps are taken in speech for the purpose of obtaining features that give high efficiency in the training and classification process .The steps are carried out by removing the signal from noise and dividing it into short time limits, where the duration of each frame was 25ms and overlap between frames was 10ms, see to figure 3.8. The frame process is one of the initial steps of the MFCC process, through which the signal is divided in order to facilitate taking the features from the speech signal well.

The study discovered that the speech frequency spectrum features arose in the 30 ms region while certain physical characteristics remained essentially the same. As a result, the stationary process processing method and theory can be applied to short-term speech signal processing, and the speech signal can be separated into several short-term speech segments, each of which is referred to as an analysis frame. Processing a frame of voice signal in this manner is

equal to processing a continuous signal with fixed characteristics, the frame can be continuous or separated into frames, or it can be overlapped and divided into frames. The frame length is typically 10-30ms. Weighting with a movable window of finite length, that is, multiplying S(n) with a specific windowing function w, results in framing (n). We can consider the voice signal stable if it is between 10 and 30 milliseconds long. To process the speech signal, we must first add a window to it, and then process only the data in the windowing at a time. We cannot and do not need to process very much data at once because the actual speech signal is very long. The smart solution is to evaluate one piece of data at a time, then evaluate another piece of data. The data takes on a more regular structure after the Hamming window is added.

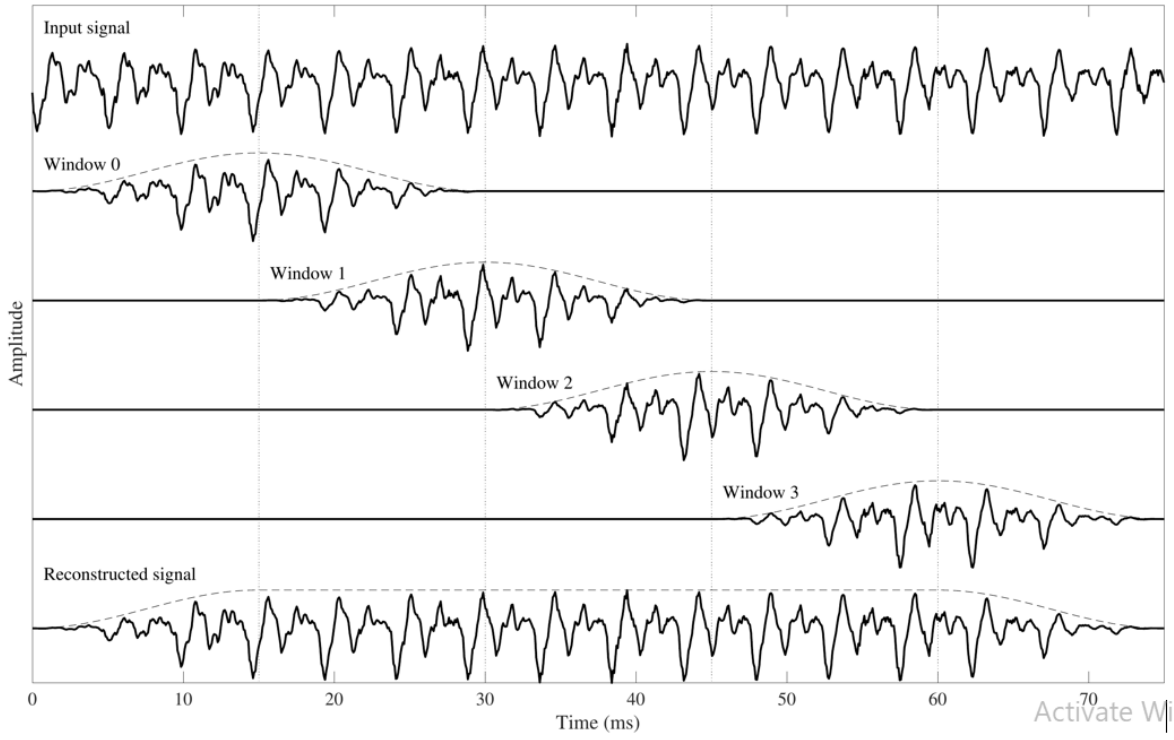


Figure (3.8): Hamming window and framing method

FFT: Because it is often difficult to see the properties of a signal when it is transformed in the time domain, where the energy is converted across the frequency domain for inspection. The features of different voices can be represented by distinct energy distributions. As a result, each frame must perform fast Fourier transform after multiplying the Hamming window. To acquire the frequency spectrum of each frame, after dividing a reference into frames, these frames are processed by FFT with length 1024. The power spectrum of the speech signal is obtained by taking the modulus square of the frequency spectrum of the spoken signal.

while collecting the audio's energy spectrum, use a mel filter bank and do a dot product operation with the energy spectrum. The mel filter convert the energy spectrum into a mel frequency that is more in line with the theory of human headphones. The MEL scale is designed to resemble the human ear, which has more auditory features. for Fig (3.9) and Fig (3.10) show FFT display of start and right words.

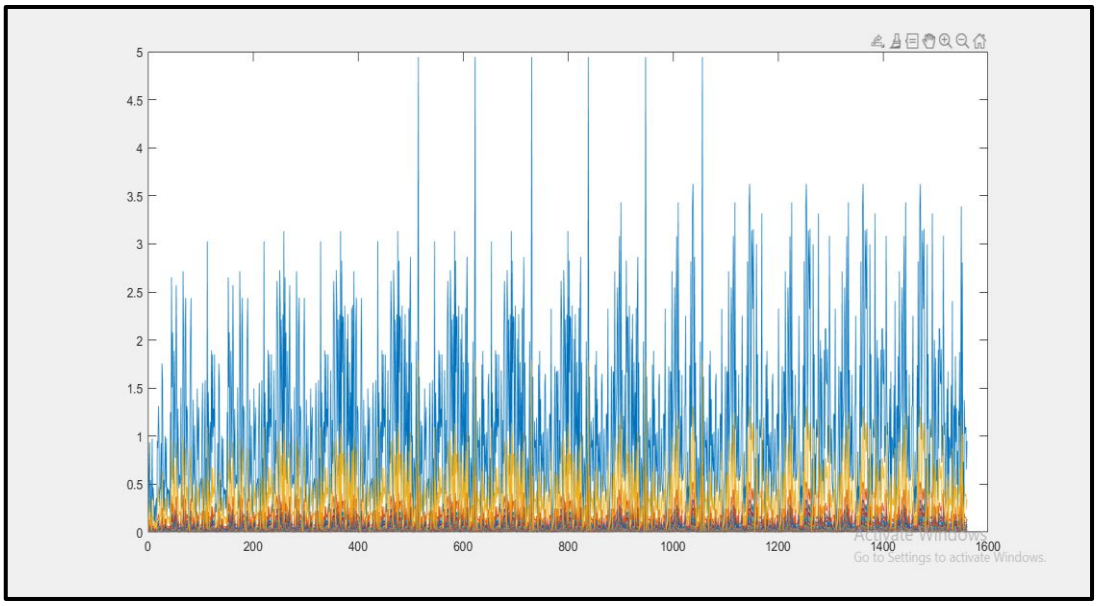


Figure (3.9): FFT for start ward

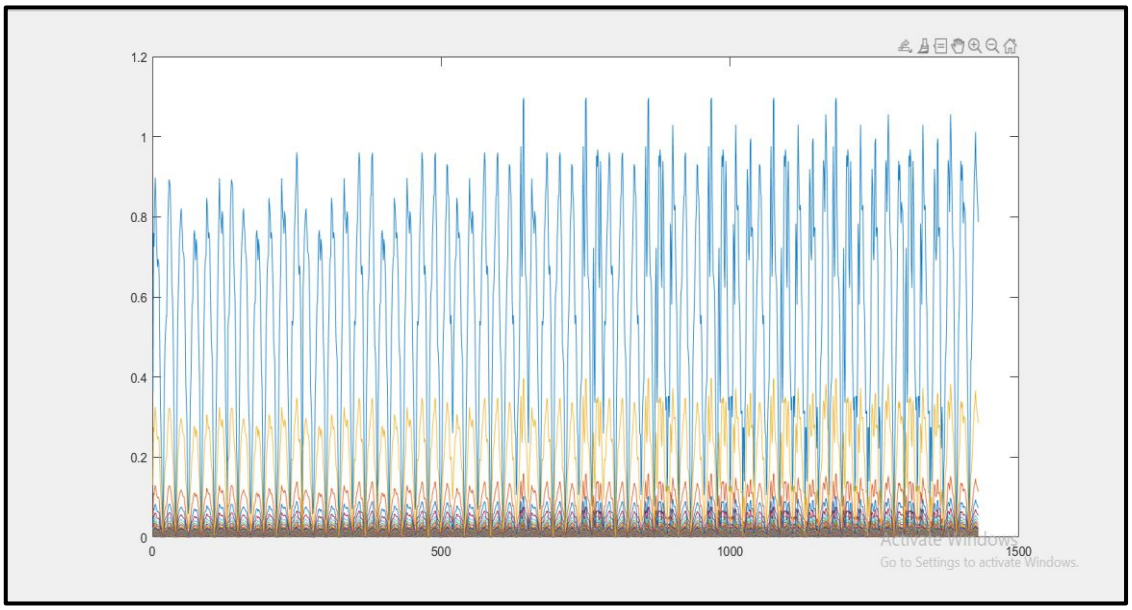


Figure (3.10): FFT for right word

Discrete Cosine Transform (DCT): DCT is the final step of MFCC feature extraction procedure. The primary principle of DCT is to correlate mel spectrum values in order to generate a fair representation of spectral local attribute.

3. 7 CNN Method Construction

CNN is often used with image processing that requires a two-dimensional matrix containing features, the pixel values are in the horizontal and vertical coordinate indicators. CNN is a neural network model. CNN has a filter that shifts over the image to produced feature map at convolution layers, Convolutional architecture is coordinated by three main layers arranged in the forward feed structure, the convolutional layer for feature extraction, sub-sampling layers, the aggregation(pooling) layer, to reduce the dimensions of the input data and the output which a fully-connected layer for final classes prediction. We have used these steps:

- Create audio data store to manage importing and processing files without overloading memory.
- Split to training and validation.
- Preprocess and convert to spectrogram.
- Use CNN for classification.
- Evaluate results.

Supervised learning of with suggested deep neural model have been used to train and test the convolution neural networks CNN. The suggested CNN structure has 13 layers. In CNN, sounds file has been entered directly to the designed network structure that contained multilevel learning procedure to achieve the model multi-classification task to classify six labels for words (start, stop, right, left, forward, backward). The size of data has been enlarged by increasing the number of training data through a process augmentation. The number of augmentations are three for every image(spectrogram) of audio. Auditory Spectrograms in frequency sampling 16KHZ has been computed with segmentation duration equal 1.8 and frame duration 0.02Ms, FFT length 1024 and number of band 64 over melspectum have been applied. Table (2.1) explain all details of our CNN layers. The model learning process started from acquiring the corresponding learning patterns in the input speech spectrum. The learning features go through our suggested network layers and training parameters have been updated through learning process. the words classification accuracy for both training and validation data have been enhanced by increasing the number of iterations to reach the better accuracy with 275 iterations. On the other hand, the misclassification data are decreased throughout the training and validation progress. The model consists of 13 layers, the first layers has 179 neurons that corresponded to the speech samples in the input matrix and the final layer which is the

classification layer has 7 neurons that refer to our seven classes (targeted labels).

- Split to training and validation: Data partitioning is an important step in the classification and training process in deep learning .In our work, we have tried more than one ratio of data division and its effect on the efficiency of the system, and this is what we will discuss in the results chapter.
- Preprocess and convert to spectrogram: One of the initial steps in the theory of CNN, which deals with images in general. In our work, we have done a lot in the process of converting sound into a two-dimensional in the frequency domain in the form of a frequency spectrum.
- CNN for classification: The process of classifying data and entering it into N-layers, whose nature consists of three main layers: Convolution layer, pooling layer and connection layer The first layer responsible for accommodating the size of the dimensions of the incoming image, whose dimensions we had were 179 * 64 * 1 for the input image to the network, and the dimensions of the convolutional layer were 179 * 64 * 10, with a weight of 3 * 3 * 1 * 10 and a basis of 1 * 1 * 10, followed by the activation layer ReLU 1. - With the same dimensions, as for Maxpool 1 with a size of 90 * 32 * 10 Which worked to reduce the dimensions of the image and thus easier to deal with, after taking 3 * 3 kernels and 2 * 2 Stride and padding. convolution layer2 (90*32*20) with RelU activation function 90*32*20, maxpooling2 reduce the dimension to 45*16*20. Fig (3.11) and table (3.2) for all details of cnn layer construction

Table (3.2): The layers of CNN for our proposed model

CNN layers (our work)	Description
1st layer	Image input layer to adjust image dimensions by (number of hopes and number of bands).
2nd layer	To add filter size of pixels (padding equal 3).
3rd layer	To balance the data and put mean and standard deviation equal to zero and do smoother gradients ,faster training and better generalization accuracy [normalization] with RelU layer.
4th layer	Add pooling to reduce size with 3 stride and 2 padding.
5th layer	2*number of filter(numF) for padding [number of filter =10].
6th layer	Batch normalization layer with RelU layer and maxpooling2dlayer with stride 3 and padding 2.
7th layer	Convolution 2D Layer (3,4*numF, 'Padding', 'same'), numf=10.
8th layer	Batch Normalization Layer (RelU layer).
9th layer	Max Pooling 2D Layer([timePoolSize,1]).
10th layer	Dropout Layer (dropout rob), dropout, prevent overfitting.
11th layer	Fully Connected Layer (Number of Classes).
12th layer	Softmax Layer , compute probability of each label.
13th layer	Classification Layer, classify based on softmax, cost will be x-entropy.

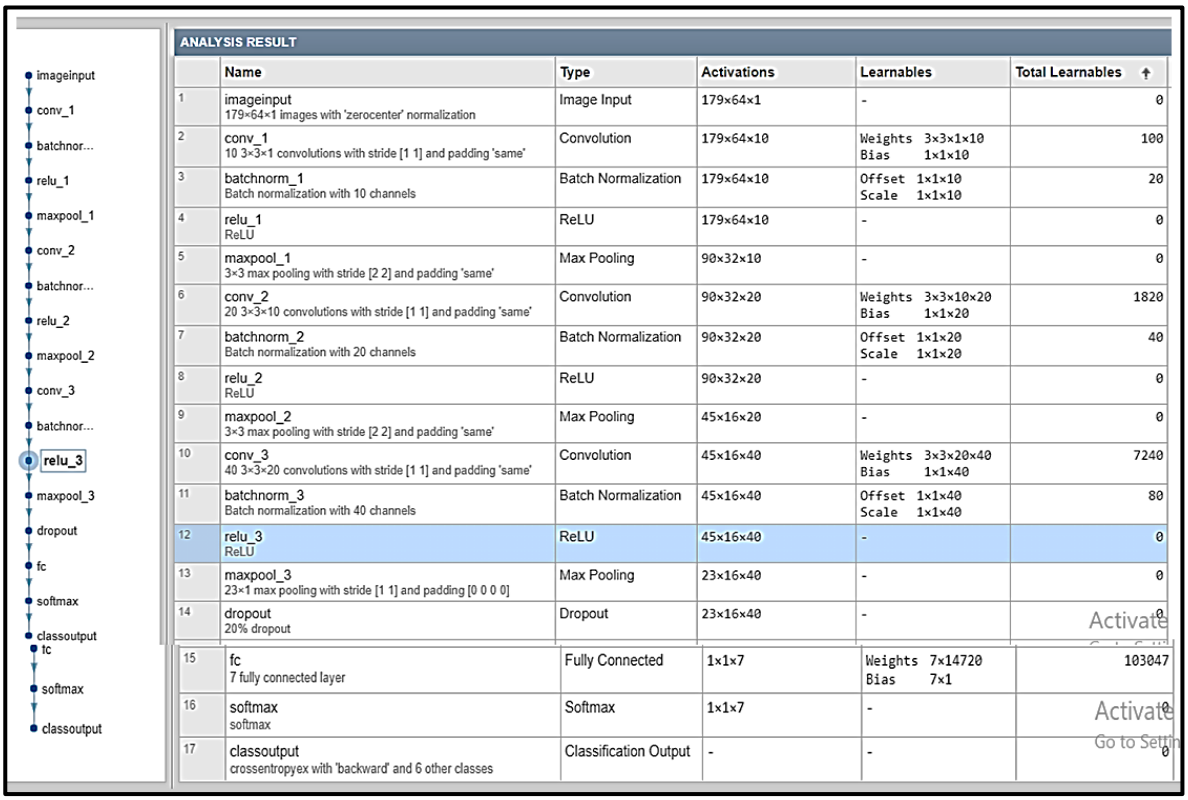


Figure (3.11): Illustrates CNN layers.

3.8 KNN

KNN the classification scenario, the k-Nearest neighbor technique effectively boils down to creating a vote by a majority of the k most comparable examples to the 'unseen' observation. A distance measure between two data points is used to define similarity. A Euclidean distance is a popular choice. We frequently employ the formula for the Euclidean distance, especially when measuring distance in the plane. We used k=5. figure3.12 explain the knn algorithm

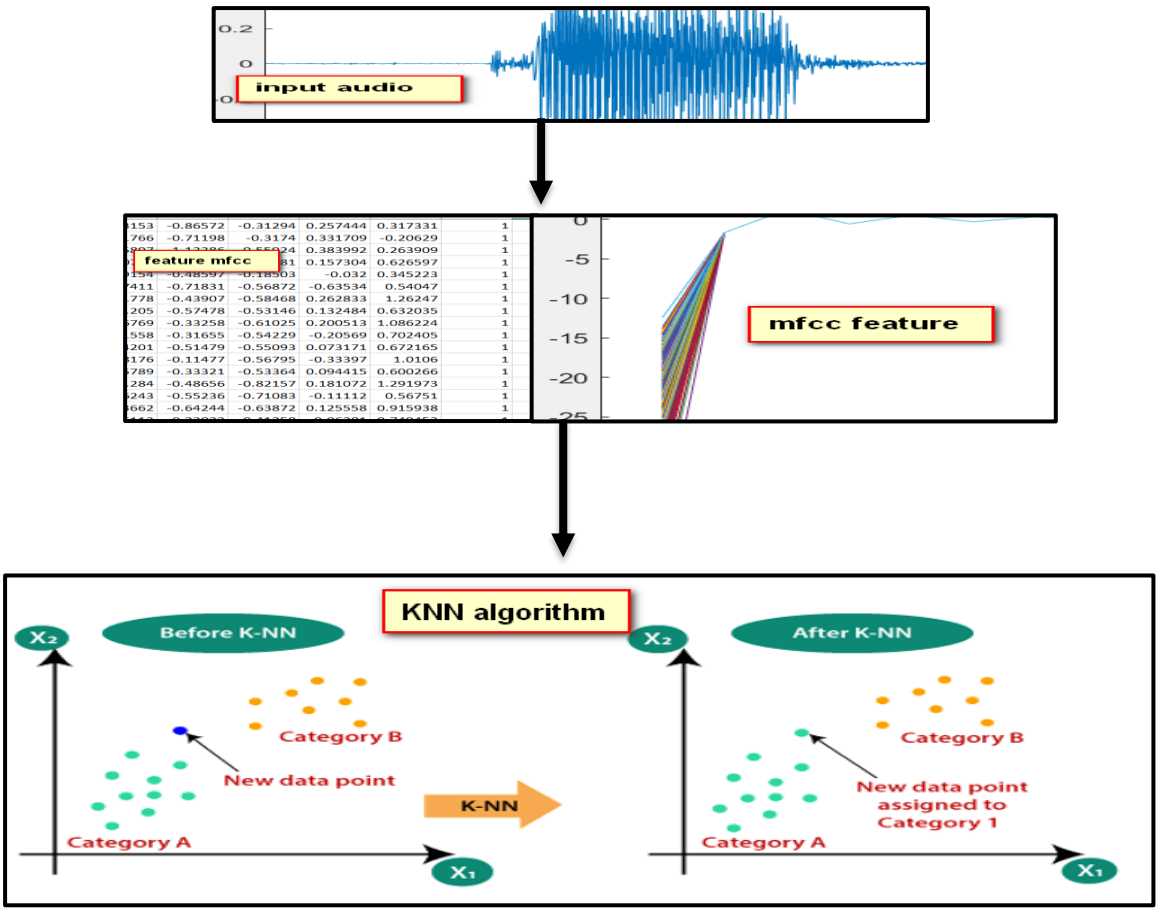


Figure (3.12): Explain the KNN algorithm

KNN's work was completed by reading the excel file that contains the features extracted by the MFS method. The stored MFCC matrix in SCV format consisted of 48155 * 20 dimensions. The main purpose of the thesis was to study the work of machine learning on the recorded data with low

accuracy and completely inaccurate pronunciation and very noisy. The data was divided into training and testing --- the division ratios were done in three different ratios until we determined the most accurate possible ratio of efficiency.15% and 85%, 20% and 80%, 25% and 75%.

3.9 SVM Method

SVM's work was completed by reading the excel file that contains the features extracted by the MFCC method. The stored MFCC matrix in SCV format consisted of 48155 * 20 dimensions. Use SVM for the purpose of categorizing with six-class entries as many words as we own. The property matrix was of equal fixed dimensions for all six words. The data was divided into training and testing --- the division ratios were done in three different ratios until we determined the most accurate possible ratio of efficiency.15% (testing) and 85% (training), 20% and 80%, 25% and 75%

The ratio	the description
15%	testing
20%	testing
25%	testing
75%	training
80%	training
85%	training

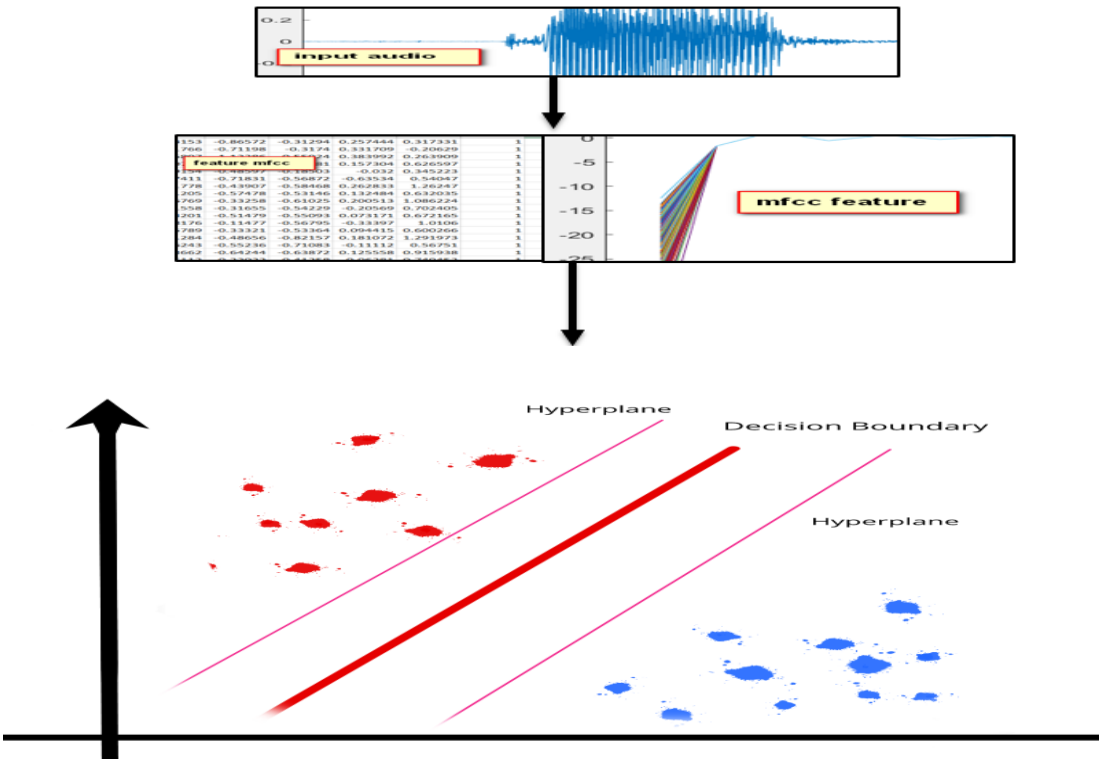


Figure (3.13): Explain the SVM algorithm

Chapter four

Results and Discussion

4.1 Introduction

In this chapter, we will discuss the results of the data that we used with the three methods. According to the order

- feature extraction results
- KNN results
- SVM results
- CNN results

4.2 Feature Extraction Results of MFCC Method

The verbal features were found using the method used by MFCC, which requires cutting the verbal signal into frames within a window that smoothest that signal from both sides, in addition to several determinants of these properties such as standard deviation, mean and pitch, final feature matrix was 48155 rows and 20 columns.

4.2.1 Number of Frames for Each Word

The data we have 30 people for each person 6 words arranged (Backward, forward, left, right, start and stop. Frame duration was 25ms, and frequency sampling (fs) was 16Kh with overlap length 10ms because the data we have is inefficient data as a result of the registration sites and the people from whom the data was collected speak the words in different forms and in many patterns, we worked to increase the number of frames and overlap duration, to find the most important possible properties of the properties of the word sign, all the features that have been identified from the standard deviation, the mean, the pitch within the windowing, FFT, the frequency spectrum

and the bank filter, which were 20 filters. Our dependence on the frames and the number of their division was based on trials and errors to obtain the highest percentage of efficiency in the classification process.

4.3 KNN Result

The data was divided into training and testing, the division ratios were done in three different ratios until we determined the most accurate possible ratio of efficiency, (15% and 85%), (20% and 80%), (25% and 75%) as table (4.1), and the best efficiency was with (15% testing 80%training) and figures (4.1).

Table (4.1): KNN accuracy result

Percentage	accuracy
15% testing , 85% training,	97.4%
20% testing , 80% training,	96.6%
25% testing , 75% training	95%

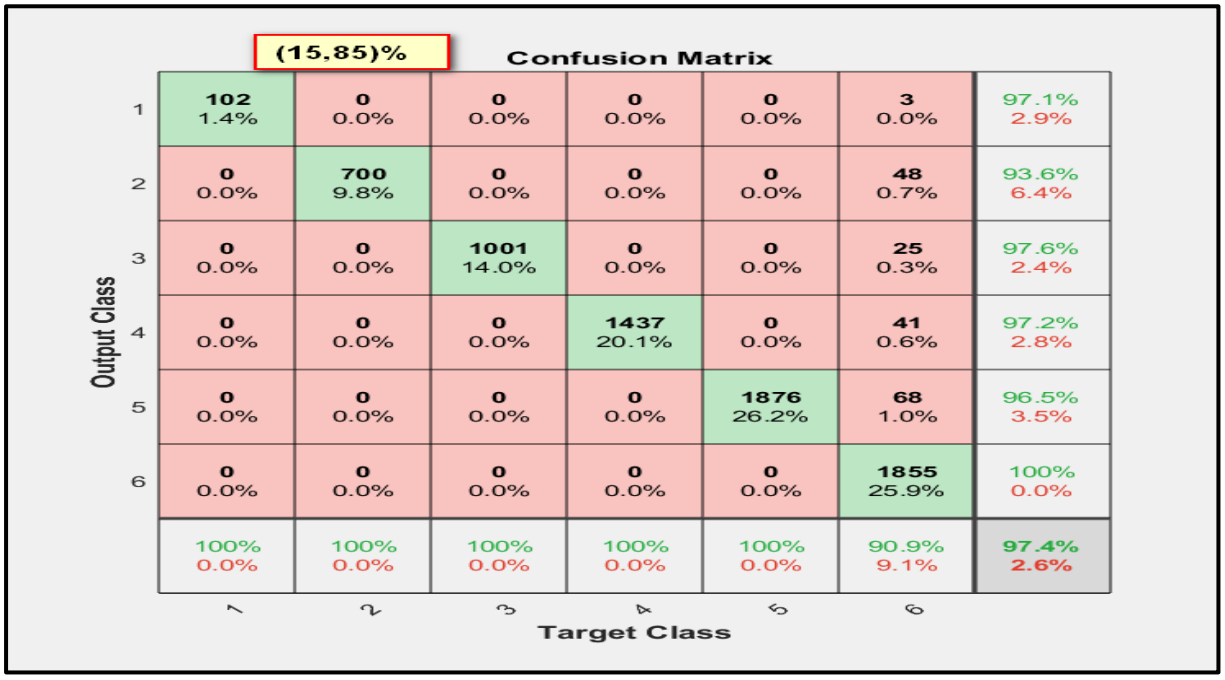


Figure (4.1): KNN confusion matrix with training and testing rate (15,85%)

- Matrix for feature extraction consisting of (48155 * 20) and the results of the first set [(15 testing, 85 training) %], in the figure (4.1), for KNN

1. Figure (4.1), which represents the confusion matrix, which represents the percentage of [(15 testing, 85 training) %]. The overall accuracy was 97.4%. Where the number samples for the test part was 15%, which is equivalent to about 7223 samples of the number of frames for the randomly chosen word from the features matrix. In the confusion matrix above, we notice a multi-class classification as much as the six input words where the number one represents the word backward, the number 2 represents the word forward, and the number 3 for left, 4 for right, 5 for start and 6 for stop. As it is known that the MFCC matrix is represented by the number of samples. Through machine learning, the process of reading these samples, which represent the words.

   In the first class, 105 samples were randomly read from the word backwards, through which the KNN system was able to identify 102 samples and made a mistake with three samples, with recall was 97.1% and precision was 100%. as these samples are supposed to be for the word Backward, but the system classified it in the number six box, which represents the word stop.

2. In the second class 748 samples were read randomly from the word forward (test set), KNN could identify 700 samples and miss 48 samples, its recall is about 93.6% for the word forward, 48 samples were assumed that these samples are for the word forward, and the fact of the matter is that the system classified them as samples belonging to the word stop.
2. In the third class, 1026 samples were randomly read from the word left (test group), through which the KNN system was able to identify 1001

samples and made a mistake with 25 samples, so its recall was about 97.6% for the word left and about 2.4 % of its error rate, as these samples are supposed to be for the word left, but the system classified it in the number six class, which represents the word stop.

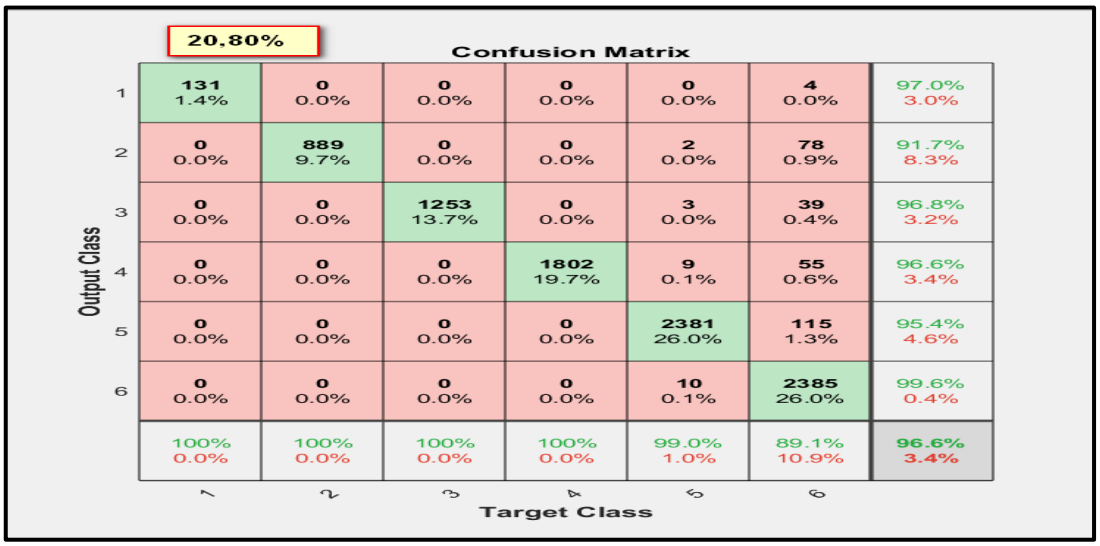


Figure (4.2): confusion matrix with training and testing rate (20,80%)

- For Figure (4.2), which represents the second set (20,80) %:

1. The confusion matrix, which represents the percentage of (20,80) %. In the ranks of the first out, 135 samples were randomly read from the word Backward (test group), through which the KNN system was able to identify 131 samples and made a mistake with 4 samples, so its recall was 97 % for the word Backward
2. In second class, 969 samples were randomly read from the word forward (test group), through which the KNN system was able to identify 889 samples and made a mistake with 80 samples, so its recall was 91.7% for the word forward,

3. In the fourth class, 1866 randomly selected samples were read from the data for the word right (test set), the KNN system was able to identify 1802 samples of the right row, and it was wrong with 64 samples, so its retrieval was about 96.6% for the word right, as this is assumed to be The samples that he misspelled for the word forward, but the system classified them in the number six, which represents the word stop, and the number five, which represents the word start. In the sixth class, 2395 samples were randomly read from the word stop (test group). The KNN system correctly identified 2385 samples while making a mistake with 10 others. The word halt has a 99.6% recall rate and a precision rate of 89.1%. The third set of KNN (25,75)% is illustrated in Figure 4.3.

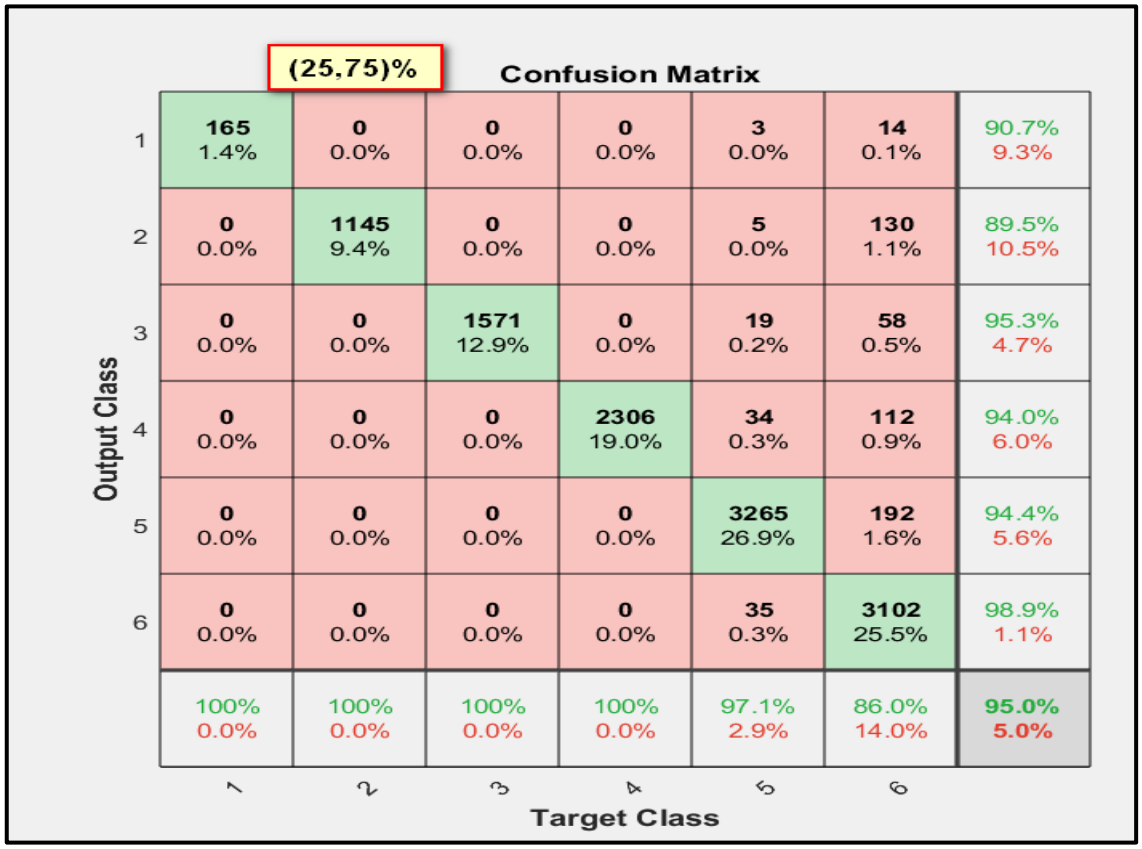


Figure (4.3): Confusion matrix with training and testing rate (25,75%)

4.4 CNN Results

One of the most important steps in speech recognition is the verbal data used and its nature, and the fact that we look at the importance of the topic as the data we collected aims to determine the ability of machine learning to deal with this data and to find the highest possible accuracy. We have tried to find different results depending on the difference in the data splitting ratio and the change in some characteristics of CNN, in addition to other methods used KNN and SVM, all in order to study the data we collected and the ability of machine learning to train on it. So, regarding CNN, we worked on finding about six different results based on the division ratio and the change in the properties of CNN as in table (4.2) and table (4.3). The best accuracy we could get was 97.06% in Figure (4.2).

The first set of CNN result was based on the three partition ratios [(15 testing ,85 training), (20 testing,80 training), (25 testing,75 training)] %, with two constant characteristics, Max Epoch = 25 and mini batch size = 50, table (4.2) shows the results of this group.

The second set of CNN result was based on three-level division ratios [(15 testing ,85 training), (20 testing,80 training), (25 testing,75 training)] %, with two constant properties, namely, Max Epoch =35 and Mini Batch size =65, table (4.3) shows the results of this group

The third set of CNN result was [(15 testing ,85 training), (20 testing,80 training), (25 testing,75 training)] %, with, Max Epoch =25 and Mini Batch size =32.

1. **The first set of CNN result:** [(15 testing ,85 training), (20 testing,80 training), (25 testing,75 training)] %, Max Epoch = 25 and mini batch size = 50**.**

I. (15 testing ,85 training) %:

Figure 4.4, which contains the confusion matrix and the training process, in which: the training accuracy, the amount of time the system took to complete the process, and the number of iteration =150 with frequency =6 iteration (iteration per epoch)

Figure 4.4, with two properties: Max Epoch =25 and MiniBatch size =50. As for the confusion matrix, which shows the words used in the validation process for each of the six class. Backward class have 8 word, the system was able to recognize 6 words with precision =66.7%, recall =75% and 2 words were missed.

Forward class have 8 word; the system was able to recognize 8 words with precision =100%, recall =100%. Left class have 8 word, the system was able to recognize 8 words. Start class have six words, the algorithm was able to detect two of them and misspelled four others with precision =100%, recall =33.3%. The system was able to recognize 5 words and misspelled 1 words in the stop class, which included a total of 6 words. The system has an average accuracy of approximately 88.33 %.

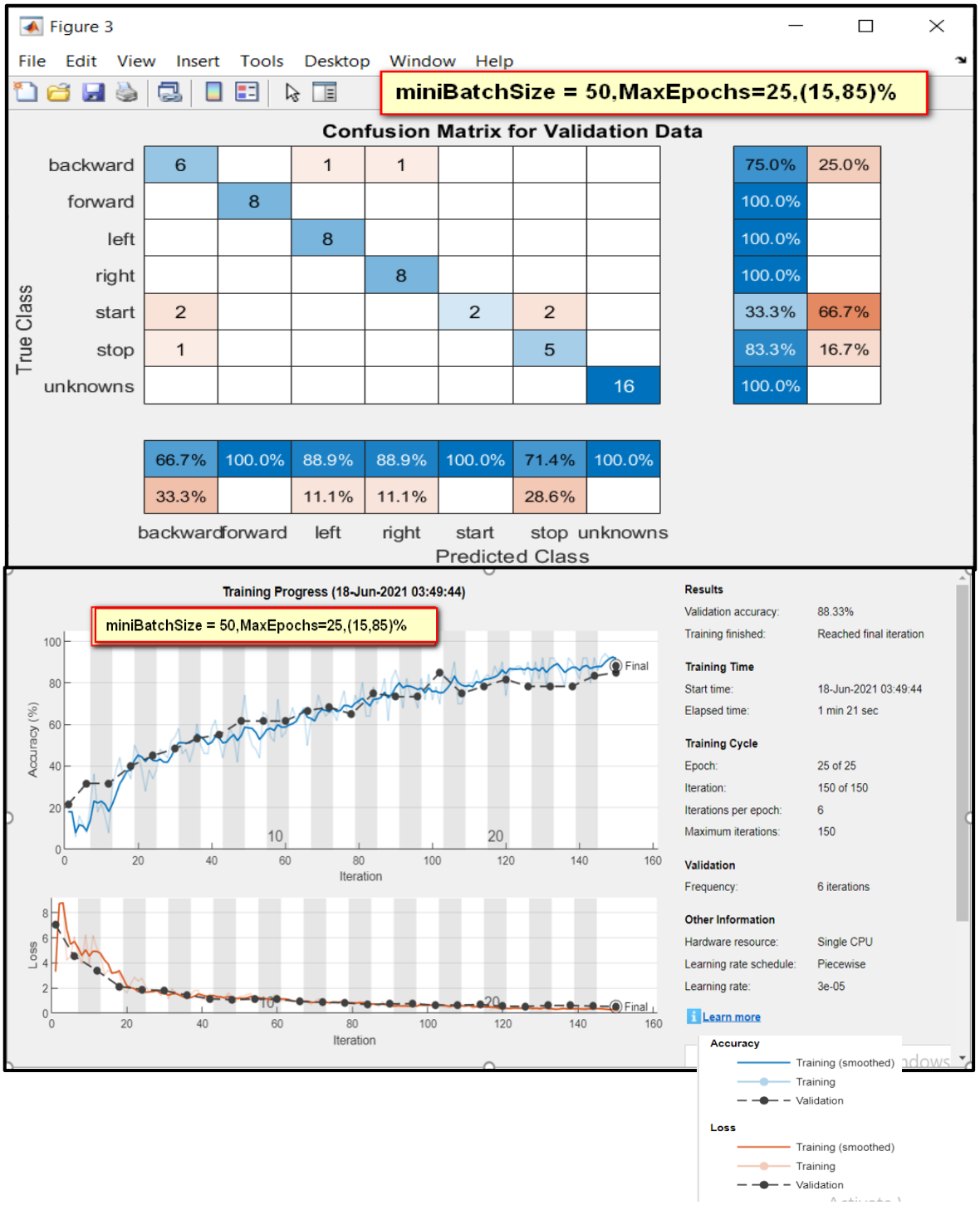


Figure (4.4): CNN training, validation and confusion matrix with 15%testing and 85% training

II. **(20 testing,80 training) %:**

Figure (4.5): Max Epoch =25 and Mini Batch size =50, (20,80) % division ratio, the number of iteration =150 with frequency =6 iteration (iteration per epoch). As for the confusion matrix, which shows the words used in the validation process for each of the six class, with the apparent percentage of error for each word.

Backward class include ten words in figure (4.5), and the system was able to recognize eight of them while making mistakes in tow of them from the testing group. In class tow (forward class), the system was able to recognize 6 words and make mistakes in 4 words with precision 100% and recall equal 60%.

The left class contain ten words, of which the machine correctly identified 9 and misspelled one. The algorithm recognized 2 of the 8 words in the start class and misspelled 6. The stop class included ten words, of which the classifier recognized eight and misspelled two with precision 53.3% and recall equal 80%. The system has a 78.75 % accuracy rate.

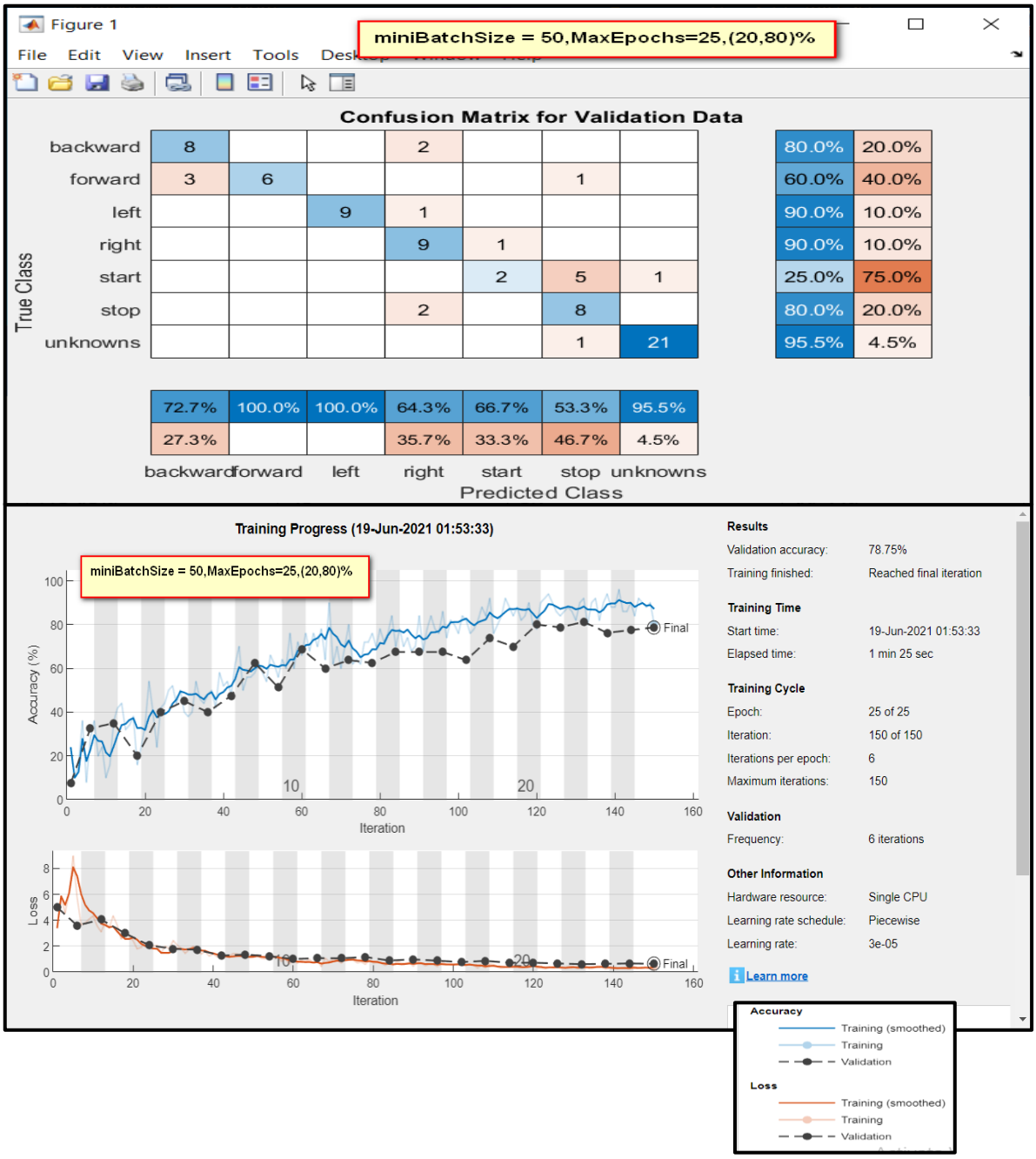


Figure (4.5): CNN training, validation and confusion matrix with 20%testing and 80%training

III. (25 testing,75 training) %:

Figure 4.6, Max Epoch =25 and MimBatch size =50, (25,75) % division ratio, which contains the confusion matrix and the training process, in which the training accuracy, the amount of time the system took to complete the process, and the number of iteration =150 with frequency =6 iteration (iteration per epoch).

The confusion matrix, on the other hand, displays the words used in the validation process for each of the six class, along with the apparent percentage of inaccuracy for each. The algorithm was able to recognize 10 words from the testing group in the backward class. The algorithm recognized four words and made six mistakes in the forward class, which totaled ten words. The machine was able to distinguish 10 words in the left class. The system recognized four words and misspelled four others in the start class, which totaled eight words. The stop class included ten words, of which the machine recognized five and misread five. The method has a 76.25 % accuracy rate.

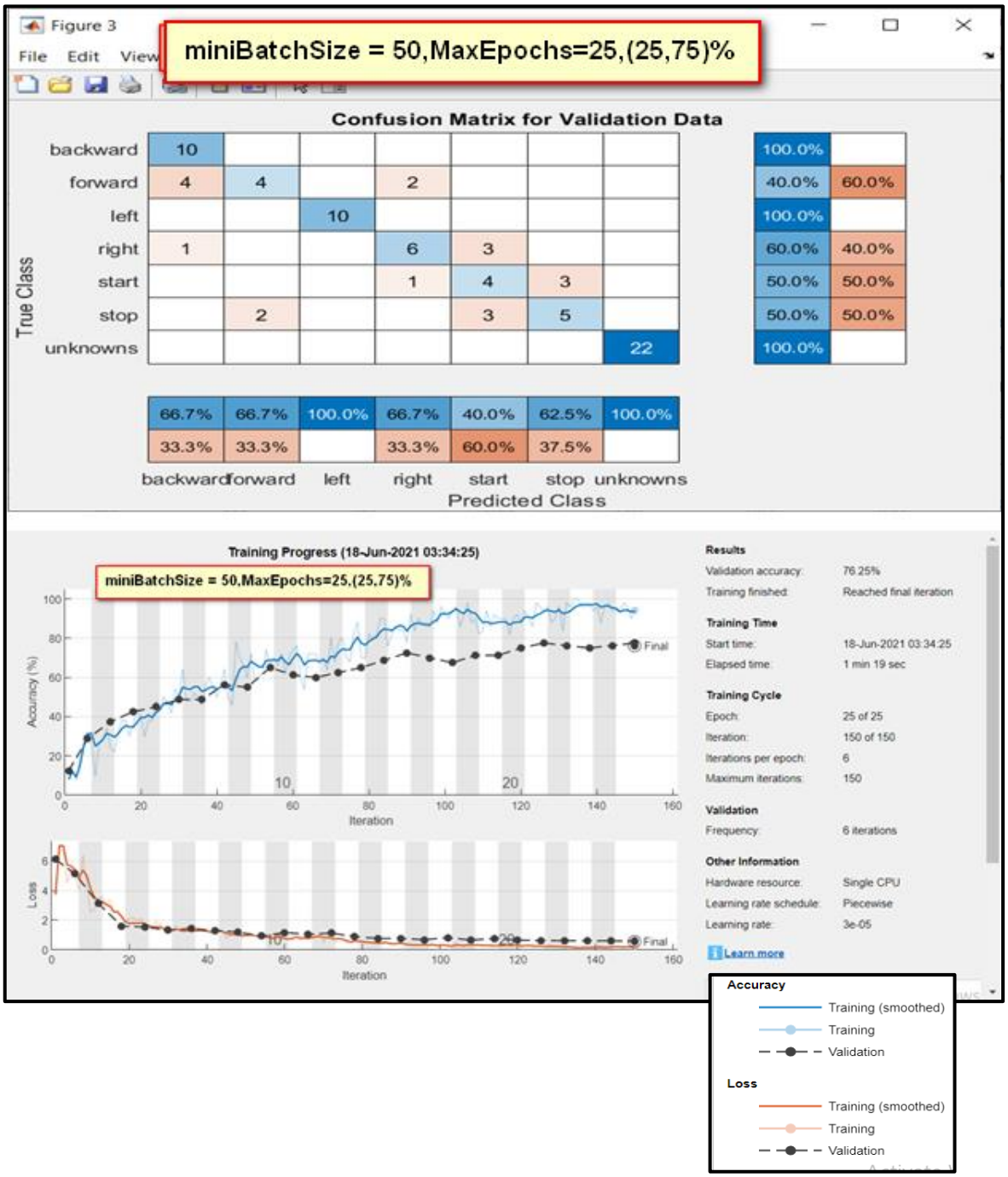


Figure (4.6): CNN training, validation and confusion matrix with 25% testing and 75%training

2. **The second set of CNN result:** [(15 testing ,85 training), (20 testing,80 training), (25 testing,75 training)] %, with two constant properties, namely, Max Epoch =35 and Mini Batch size =65

I. (15 testing ,85 training) %:

Figure 4.7, which depicts the confusion matrix and the training process, has two fixed properties: Max Epoch =35 and MiniBatch size =65, division ratio (15,85) percent, and the number of iterations =175 with frequency =5 iteration. (iteration per epoch).

The confusion matrix, shows the words used in the validation process for each of the six class, along with the apparent proportion of inaccuracy for each. Backward class have 8 words, and the system was able to recognize 6 words and make a mistake in 2 word with precision 60% and recall equal 75% from the testing group for (backward).

The left class has 8 words, were detected. Three of the six words in the start class were identified by the algorithm, while three were misspelled. The stop class consisted of 6 words, 4 of which the classifier recognized and two of which it misspelled with precision 57.1 percent and recall of 66.7 percent. The system achieves an average accuracy of about 78.33%

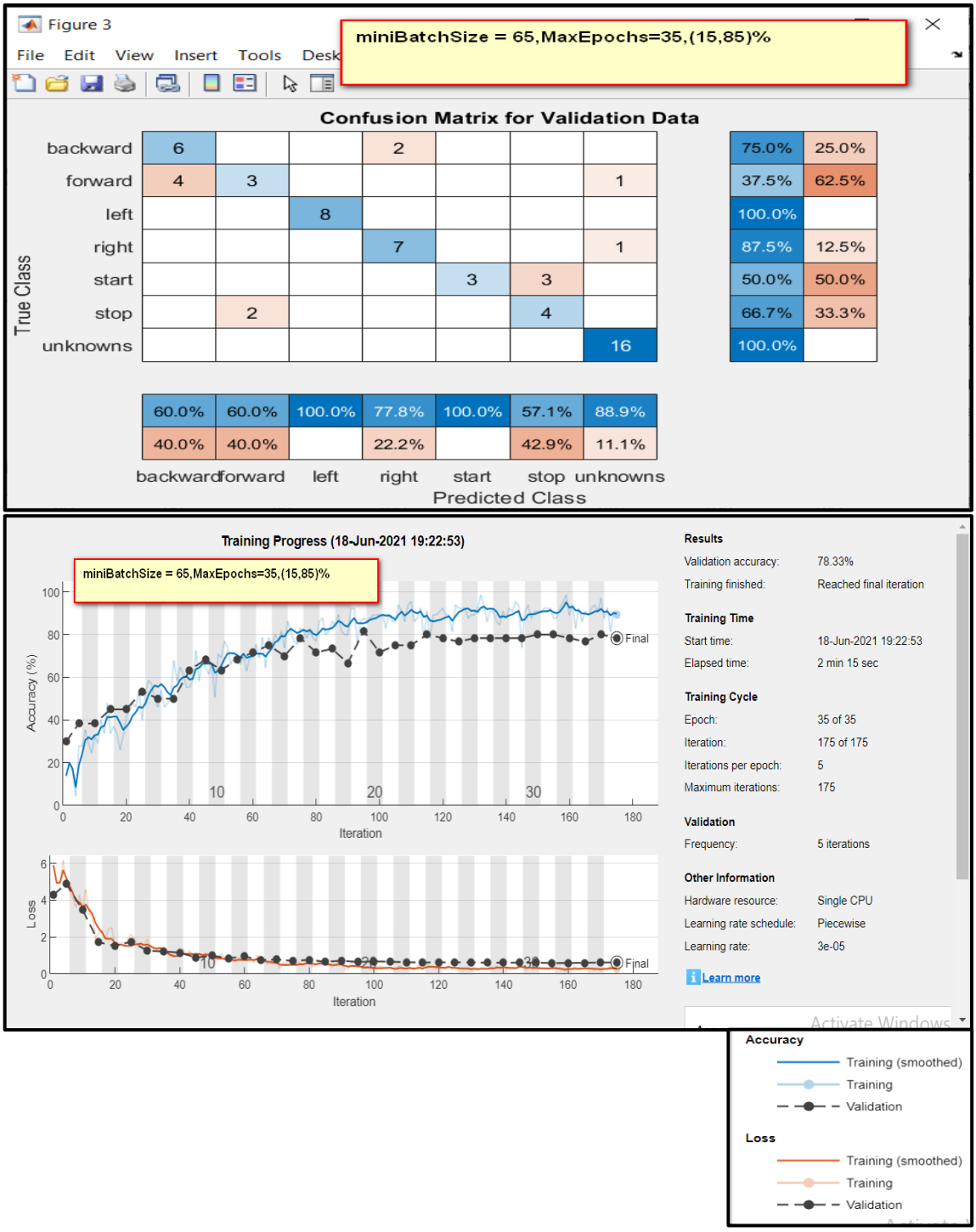


Figure (4.7): CNN training, validation and confusion matrix with 15% testing and 85%trainin

II. (20 testing,80 training) %:

In Figure. (4.8), the confusion matrix and the training process are shown, together with the training accuracy, the time it took the system to complete the procedure, and the number of iteration. With, Max Epoch =35 and Mini Batch size =65, division ratio (20,80) %, and the number of iteration =140 with frequency =4 iteration (iteration per epoch). the confusion matrix, which displays the words used in validation for each of the six classes, together with the apparent percentage of mistake for each of them.

Figure (4.8), the backward class has ten words, and the system was able to recognize seven of them, with three errors from the testing group (backward), with precision 58.3 percent and recall of 70 percent

Forward class has ten words, of which the system was able to recognize eight and make two mistakes with precision =72.7% and recall =80%. Left class had 10 words, of which the system recognized eight and misspelled two. Start class with eight words, the algorithm was able to recognize four and misspell four. Stop class had ten words, of which the system recognized six and misspelled four with precision =66.7% and recall =60%. The system has a 71.25 % accuracy rate on average.

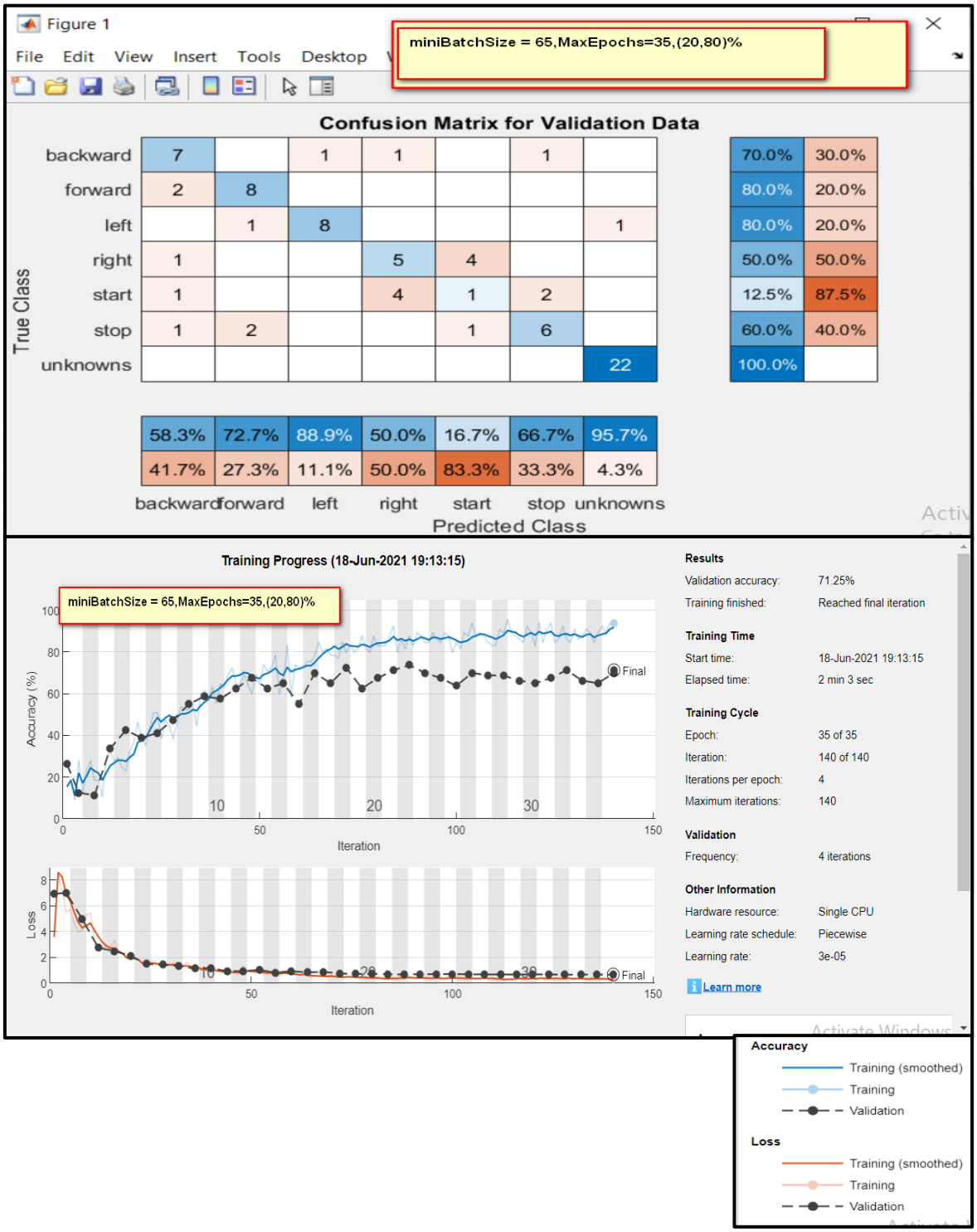


Figure (4.8): CNN training, validation and confusion matrix with 20%testing and 80%training

III. (25 testing,75 training) %:

The training process is depicted in Figure 4.9, which includes the confusion matrix and the validation accuracy, as well as the time it took the system to finish the process and the number of iterations.

Figure 4.9, with two constant properties, namely, Max Epoch =35 and MiniBatch size =65, division ratio (25,75) %, number of iteration =140 with frequency =4 iteration (iteration per epoch).

The confusion matrix, on the other hand, displays the words used in the validation process for each of the six class, in addition to the error ratio for each of them. Figure 4.10 shows that the backward class of twelve words, of which the system classify 5 and misread 7, with precision =38.5% and recall =41.7%.

The forward class has 12 words, of which the system recognized four and misspelled eight. The system was able to recognize 9 words and misspelled three terms in the left class, which included a total of 12 words. Starting with ten words, the machine was able to recognize five and misspell five. Stop class had eleven words, of which the system recognized 8 and misspelled three. The method has an accuracy (68.37 %).

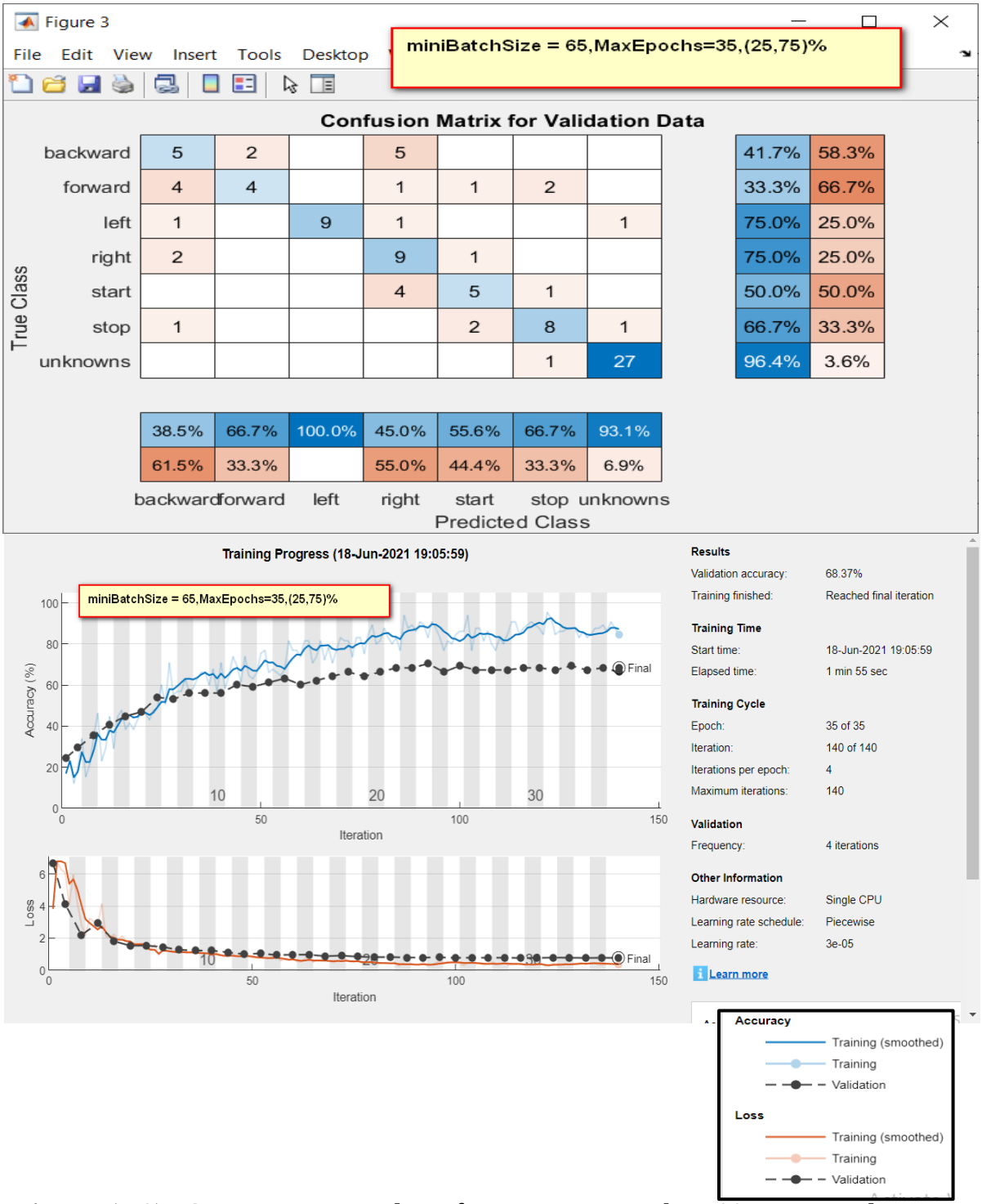


Figure (4.9): CNN training and confusion matrix with 25%testing and 75%training

3. The third set of CNN result:

Figure (4.10) The best result in the third group of CNN training and testing. Its accuracy was about 97.06%, with, Max Epoch =25 and MiniBatch size =32, (15 testing ,85 training) %. Figure (4.10) which contains the training process, in which: the training accuracy, the amount of time the system took to complete the process, and the number of attempts per batch appear and the number of iteration =275 with frequency iteration (iteration per epoch) =11.

Figure (4.11), representing the confusion matrix showing the testing aspect of the data, with, Max Epoch = 25 and MiniBatch size = 32, with a division ratio (15 testing ,85 training) %.

As for the confusion matrix, which shows the words used in the validation process for each of the six classes, with the apparent percentage of error for each word.

Figure (4.11), backward class have 4 word, the algorithm was able to recognize 4 words from the testing group for this word (backward). Forward class have 4 word; the system was able to recognize 4 words from the testing group for this word (forward).

Left class have 4 word, the system was able to recognize 4 words. start fields have 4 word, the system was able to recognize 4 words. stop box have 4 word, the system was able to recognize 4 words. The system achieves an average accuracy of about 97.06%.

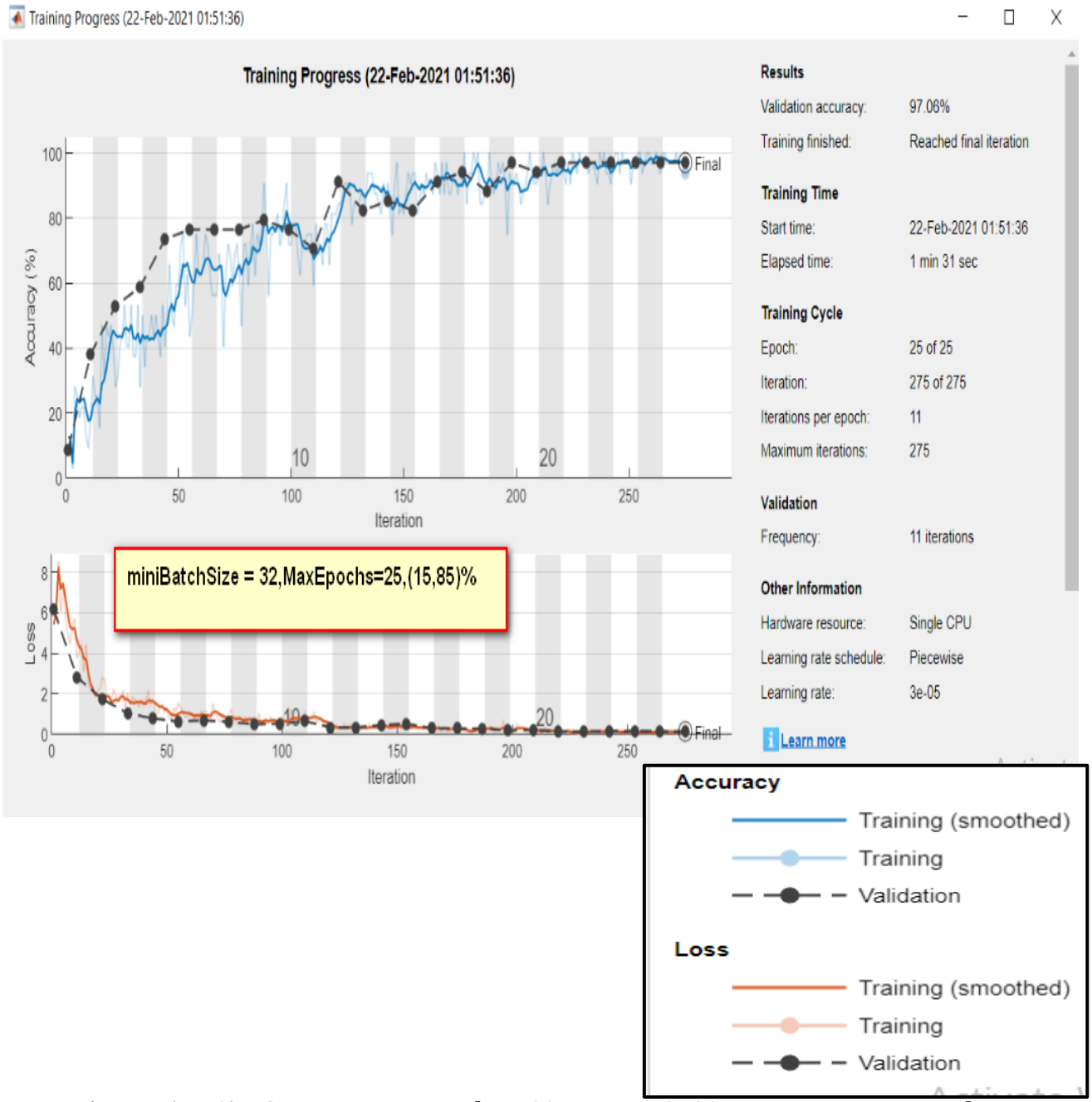


Figure (4.10): CNN training with 15% testing, 85% training, Max Epoch =25 and Mini Batch size =32

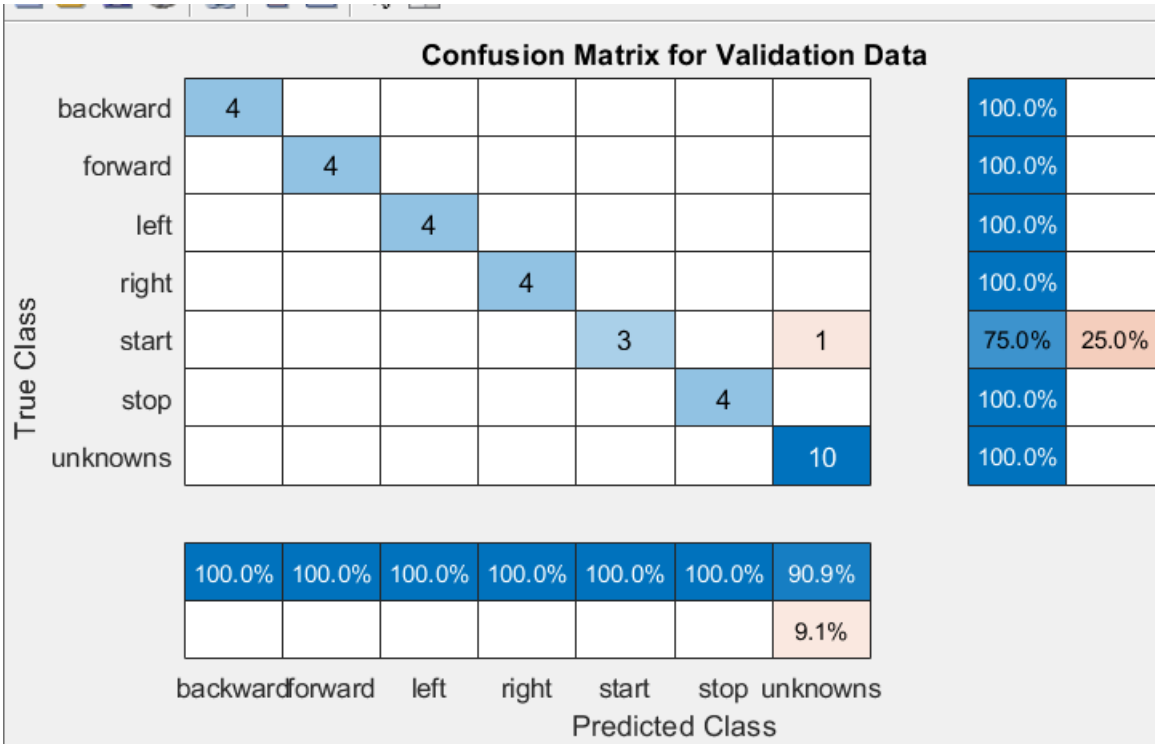


Figure (4.11): CNN confusion matrix for validation with 15% testing, 85% training and Max Epoch =25 and Mini Batch size =32

Table (4.2) and table (4.3) illustrate the results of CNN

Table (4.2): CNN accuracy result

Percentage of training and testing%	Max Epoch	Mim Batch size	Accuracy
(15,85)%	25	50	88.33%
(20,80)%	25	50	78.75%
(25,75)%	25	50	76.25%

Table (4.3): CNN accuracy result

Percentage of training and testing%	Max Epoch	Mim Batch size	Accuracy
(15,85)%	35	65	78.33%
(20,80)%	35	65	71.25%
(25,75)%	35	65	68.37%

4.5 SVM Results

We have previously mentioned that our study dealt with three proposed models of modern neural networks. Concerning the third model, SVM, which results are shown in Figure 4.12 and table (4.4).

Matrix for feature extraction consisting of (48155 * 20)., and the results were at Figure 4.12

1. When looking at Figure (4.12), which represents the confusion matrix. Where the sample number for the test part was 15% testing, which is equivalent to 7223.25 samples of the number that randomly chosen words from the features matrix. In the confusion matrix, we notice a multi-class classification as much as the six input words where the number (1) represents the word backward, the number (2) represents the word forward, and the number (3) for left, (4) for right, (5) for start and (6) for stop. As it is known that the MFCC matrix is represented by the number of samples. Through machine learning, the process of reading these samples, which represent the sum of special words. In the ranks of the first out, 0 samples were randomly read from the word backwards, through which the SVM system was able to identify 0 samples and made a mistake with three samples, so its accuracy was about 0.

2. In the second class, 5 samples were randomly read from the word forward (test group), through which the SVM system was able to identify 2 samples

and made a mistake with 2 samples, so its recall was 40% for the word forward and 0.1% of its precision, as these samples are supposed to be for the word forward, but the system classified it in the six class, which represents the word stop.

3. In third class, 2166 samples were randomly read from the word left (test group), through which the SVM system was able to identify 1035 samples and made a mistake with 1131 samples, so its recall was about 47.8% for the word left and about 54.2 % of its precision rate, as these samples are supposed to be for the word left, but the system classified it in another boxes.

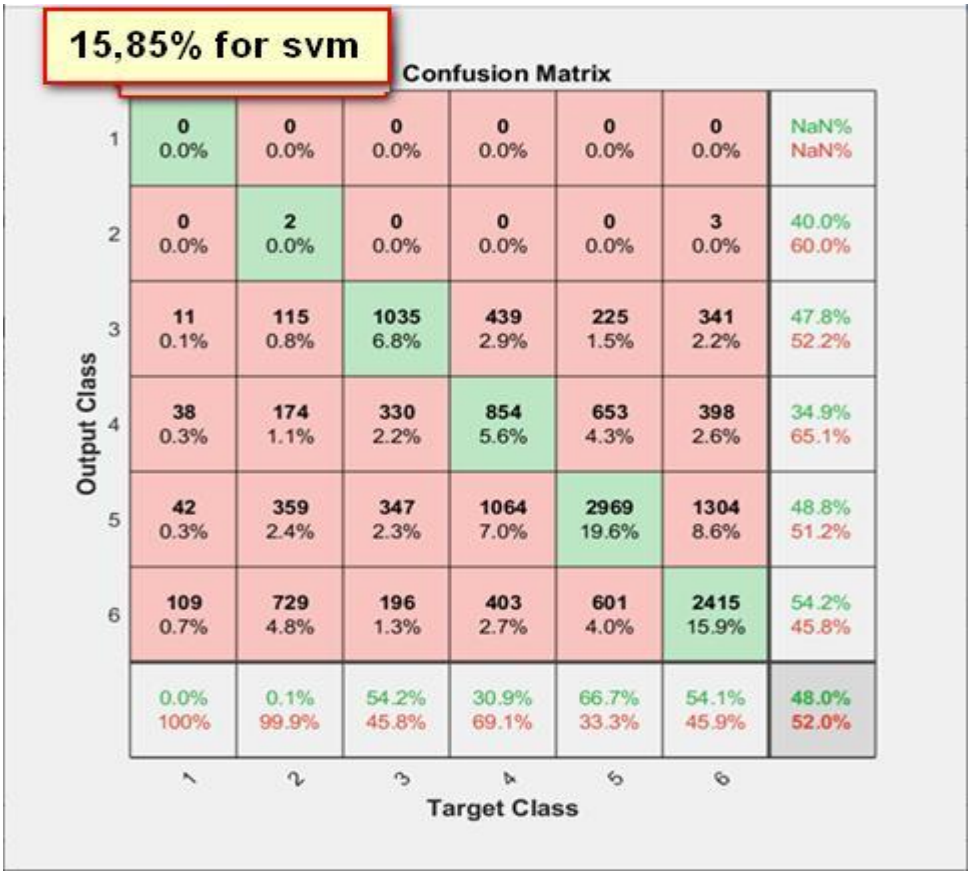


Figure (4.12): SVM confusion matrix with 15%testing and 85%training

Table (4.4): SVM accuracy results

Percentage of training and testing%	accuracy
15% , 85%,	48%
20% , 80%,	46.9%
25% , 75%	46.3%

Table (4.5) explain the results for three methods:

Table (4.5): SVM, KNN, CNN (for one state) accuracy result comparison

Model	**Percentage of training and testing%**	**accuracy**
KNN	15% , 85%,	97.4%
CNN	15% , 85%,	97.06%
SVM	15% , 85%,	48%
Model	**Percentage of training and testing%**	**accuracy**
KNN	20% , 80%	96.6%
CNN	20% , 80%	85%
SVM	20% , 80%	46.9%
Model	**Percentage of training and testing%**	**accuracy**
KNN	25% , 75%	95%
CNN	25% , 75%	81.63%
SVM	25% , 75%	46.3%

4.6 Discussion

In this thesis, three methods were used to detect noise words, KNN , SVM and CNN were used. Good results were obtained with KNN. The efficiency of KNN was 97.4% with a test 15 % and a training rate of 85%. KNN was able to reach this efficiency depending on the value of K = 5, and depending on KNN’s work mechanism, which is summarized by measuring the distance to

the nearest Five neighbors therefore it is possible to recognize the input and ease of identification between the classes and Figure 4.1 shows the result.

As for the results of SVM, Figure 4.4 shows the final efficiency results of the system.. From the figure, we notice that the network was not able to take sufficient data for the first class, which represents the word Backward due to the imbalance of data and thus model biased towards the more data class, which requires stratify, and due to the nature of SVM's work in searching for the similarity ratio between classes, which makes balancing between classes necessary to obtain high efficiency.

As for the CNN method, it was able as a result of its structure and ability to deal with data that SVM was not able to deal with, since CNN works on determine the spectrum in fixed dimensions, and passes a kernel for a filter of suitable dimensions with a spectrum shape, it also coordinates the data by adding padding

Chapter Five
Conclusions and Future Work

5.1 Conclusions

1.This research succeeded in speech recognition processes to detect noise words using three methods of machine learning networks :Support Vector Machine (SVM), k-nearest neighbor (KNN) and Convolutional Neural Network (CNN).

2. I collected datasets containing six words from thirty people: (start, stop, forward, backward, right, left). I collected the words in different places: in the market, in the street, in the laboratory, in the house, in the car. Half of the dataset are men and the other half are women. The words were recorded in English via laptops and phones.

3.The Signal-to-noise ratio SNR was calculated for the dataset and the average noise rate was about (-10dB).

4. Speech features were obtained using MFCC, the feature matrix was trained by KNN and SVM.

5. The speech sound was converted into a frequency spectrum and divided into frames, the problem of different word lengths was solved in Matlab, thus we were able to read the features at fixed lengths.

6.We performed the speech recognition process using the KNN method with three ratios for data [(20 testing, 80 training), (15 testing, 85 training) and (25 testing, 75 training)] % and the best efficiency was with (15 testing, 85 training) %, where the efficiency with it reached About 97.4%. Table (4.1) and Figure (4.1) show the details of the results

9. CNN The second proposal, CNN was able to deal with the data we have from several aspects, through the number of layers, division ratios and training properties. We ended up with 13 layers of CNN as the best possible result, we trained the data with a CNN with 3 partition ratios as well as in SVM and KNN, and with each partition ratio we changed many properties , like Max Epoch and MiniBatch size to see how it affects the rating, we showed the best efficiency with split ratio and characteristics of (15 testing,85 training) %, Max Epoch = 25 and MimBatch size = 32, where the efficiency 97.06%. Table (4.2), Table (4.3), Figure (4.10) and Figure (4.11), show the details of the results.

10. Concerning the proposal of SVM, we performed the same division process as in KNN and CNN, for the same features that we found with MFCC and the efficiency ratios were few compared to KNN and CNN

5.2Future Works

1. Designing a practical circuit that implements the speech recognition process
2. Collect a suitable dataset in Arabic and do the speech recognition process
3. Do speaker recognition for Arabic dataset

References

[1] Stenman, M. (2015). Automatic speech recognition an evaluation of Google Speech.

[2] Halageri, A., Bidappa, A., Arjun, C., Sarathy, M. M., & Sultana, S. (2015). Speech recognition using deep learning. Int. J. Comput. Sci. Inf. Technol, 6(3), 3206-3209..

[3] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). Ieee.

[4] Patel, I., & Rao, Y. S. (2010). Speech recognition using hidden Markov model with MFCC-subband technique. In 2010 International Conference on Recent Trends in Information, Telecommunication and Computing (pp. 168-172). IEEE.

[5] Singh, A., Rajoriya, D. K., & Singh, V. (2012). Broad Acoustic Classification of Spoken Hindi Hybrid Paired Words using Artificial Neural Networks. International Journal of Computer Applications, 52(12).

[6] Deng, L., Hinton, G., & Kingsbury, B. (2013, May). New types of deep neural network learning for speech recognition and related applications: An overview. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8599-8603). IEEE.

[7] Ge, Z., Iyer, A. N., Cheluvaraja, S., Sundaram, R., & Ganapathiraju, A. (2017). Neural network based speaker classification and verification systems with enhanced features. In 2017 Intelligent Systems Conference (IntelliSys) (pp. 1089-1094). IEEE.

[8]. Lee, J., Kim, T., Park, J., & Nam, J. (2017). Raw waveform-based audio classification using sample-level CNN architectures. arXiv preprint arXiv:1712.00866

[9] Boles, A., & Rad, P. (2017). Voice biometrics: Deep learning-based voiceprint authentication system. In 2017 12th System of Systems Engineering Conference (SoSE) (pp. 1-6). IEEE.

[10] Kaur, G., Srivastava, M., & Kumar, A. (2017). Speaker and speech recognition using deep neural network. International Journal of Emerging Research in Management and Technology, 6, 8.

[11] Li, X., & Zhou, Z. (2017). Speech Command Recognition with Convolutional Neural Network. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[12] Kim, T., Lee, J., & Nam, J. (2019). Comparison and analysis of SampleCNNs architectures for audio classification. IEEE Journal of Selected Topics in Signal Processing, 13(2), 285-297.

[13] Bunrit, S., Inkian, T., Kerdprasop, N., & Kerdprasop, K. (2019). Text-independent speaker identification using deep learning model of convolution neural network. International Journal of Machine Learning and Computing, 9(2), 143-148.

[14] Algihab, W., Alawwad, N., Aldawish, A., & AlHumoud, S. (2019). Arabic speech recognition with deep learning: A review. In International Conference on Human-Computer Interaction (pp. 15-31). Springer, Cham.

[15] Yang, X., Yu, H., & Jia, L. (2020). Speech recognition of command words based on convolutional neural network. In 2020 International Conference on Computer Information and Big Data Applications (CIBDA) (pp. 465-469). IEEE.

[16] Rownicka, J., Renals, S., & Bell, P. (2017). Simplifying very deep convolutional neural network architectures for robust speech recognition. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 236-243). IEEE.

[17] Poudel, S., & Anuradha, R. (2020). Speech Command Recognition using Artificial Neural Networks. JOIV: International Journal on Informatics Visualization, 4(2), 73-75..

[18] Jia, Y., Chen, X., Yu, J., Wang, L., Xu, Y., Liu, S., & Wang, Y. (2021). Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network. Complex & Intelligent Systems, 7(4), 1749-1757.

[19] Sehgal, P., & Jain, R. K. (2013). Speech Processing. International Journal of Engineering Sciences and Emerging Technologies 2, 5, 83-87.

[20] Davletcharova, A., Sugathan, S., Abraham, B., & James, A. P. (2015). Detection and analysis of emotion from speech signals. Procedia Computer Science, 58, 91-96.

[21] Shaikh Naziya, S., & Deshmukh, R. R. (2016). Speech recognition system—a review. IOSR J. Comput. Eng, 18(4), 3-8.

[22] Lokesh, S., & Devi, M. R. Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method Cluster Comput (2017).

[23] Kesarkar, M. P., & Rao, P. (2003). Feature extraction for speech recognition. Electronic Systems, EE. Dept., IIT Bombay.

[24] Vaishnavi, V., Suveetha, P., & Bharathi, R. (2018). A Survey on Features of Sound Signals. Journal of Applied Science and Computations, 13(2), 0076-5131.

[25] Yang, C. H. H., Qi, J., Chen, S. Y. C., Chen, P. Y., Siniscalchi, S. M., Ma, X., & Lee, C. H. (2021, June). Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6523-6527). IEEE.

[26] Hamid, O. K. (2018). Frame blocking and windowing speech signal. Journal of Information, Communication, and Intelligence Systems (JICIS), 4, 87-94.

[27] Supinator, A. (2019, June). Control of Electronic Devices Using Neural Network Based Sundanese Speech Recognition. In 2019 IEEE

International Conference on Innovative Research and Development (ICIRD) (pp. 1-6). IEEE.

[28] Lartillot, O., & Toiviainen, P. (2007, September). A Matlab toolbox for musical feature extraction from audio. In International conference on digital audio effects (Vol. 237, p. 244).

[29] Jasleen & Dilber, D. (2016). Feature Selection and Extraction of Audio Signal. International Journal of Innovative Research in Science Engineering and Technology, 4(2), 2319-8753.

[30] Abdelmoneim, M., & Subair, S. (2019). Comparison and Relationship between Big Data Analytics and Machine learning (No. 2143). EasyChair.

[31] Davletcharova, A., Sugathan, S., Abraham, B., & James, A. P. (2015). Detection and analysis of emotion from speech signals. Procedia Computer Science, 58, 91-96.

[32] Gerhard, D. (2003). Pitch extraction and fundamental frequency: History and current techniques (pp. 0-22). Regina, SK, Canada: Department of Computer Science, University of Regina.

[33] Sah, S. (2020). Machine Learning: A Review of Learning Types.

[34] Murty, M. N., & Devi, V. S. (2015). Introduction to pattern recognition and machine learning (Vol. 5). World Scientific.

[35] Aida-zade, K., Xocayev, A., & Rustamov, S. (2016). Speech recognition using support vector machines. In 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-4). IEEE.

[36] Ernawan, F., & Abu, N. A. (2011). Efficient discrete tchebichef on spectrum analysis of speech recognition. International Journal of Machine Learning and Computing, 1(1), 1.

[37] Ganapathiraju, A. (2002). Support vector machines for speech recognition. Mississippi State University.

[38] Alcaraz Meseguer, N. (2009). Speech analysis for automatic speech recognition (Master's thesis, Institutt for elektronikk og telekommunikasjon).

[39] Glittas, A. X., & Gopalakrishnan, L. (2021). A low latency modular-level deeply integrated MFCC feature extraction architecture for speech recognition. Integration, 76, 69-75.

[40] Chen, Y. Q., Damper, R. I., & Nixon, M. S. (1997). On neural-network implementations of k-nearest neighbor pattern classifiers. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 44(7), 622-629.

[41] Becker, S., & Plumbley, M. (1996). Unsupervised neural network learning procedures for feature extraction and classification. Applied Intelligence, 6(3), 185-203.

[42] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. Neurocomputing, 234, 11-26.

[43] Poudel, S., & Anuradha, R. (2020). Speech Command Recognition using Artificial Neural Networks. JOIV: International Journal on Informatics Visualization, 4(2), 73-75.

[44] Kwon, S. (2020). A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors, 20(1), 183.

[45] Yang, X., Yu, H., & Jia, L. (2020). Speech recognition of command words based on convolutional neural network. In 2020 International Conference on Computer Information and Big Data Applications (CIBDA) (pp. 465-469). IEEE.

[46] Kim, T., Lee, J., & Nam, J. (2019). Comparison and analysis of samplecnn architectures for audio classification. IEEE Journal of Selected Topics in Signal Processing, 13(2), 285-297.

[47] Passricha, V., & Aggarwal, R. K. (2019). Convolutional support vector machines for speech recognition. International Journal of Speech Technology, 22(3), 601-609.

[48] Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378.

[49] Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. MAICS, 710, 120-127.

[50] Goceri, E., & Gooya, A. (2018). On the importance of batch size for deep learning. In Int Conf on Mathematics (ICOMATH2018), An Istanbul Meeting for World Mathematicians, Istanbul, Turkey.

[51] Abdul Aziz Saleh, M. B. W. (2018). Spoken Arabic digits recognition using deep learning/AbdulAziz Saleh Mahfoudh Ba Wazir (Doctoral dissertation, University of Malaya).

الخلاصة

أصبح استخدام نموذج التعرف على الكلام في غاية الأهمية نظرًا لتعدد استخداماته في التحكم ؛ صمم بحثنا نموذجًا للتعرف على الكلمات من خلال تطبيق ميزات التعرف على الكلام مع التعلم الآلي مع ثلاثة أنواع منها: التعلم العصبي التلافيفي العميق (CNN) ، وخوارزمية k-الأقرب (KNN) ، وآلة متجه الدعم (SVM) . يتم استخدام ست كلمات (بدء ، توقف ، الأمام ، الخلف ، اليمين ، اليسار) في اللغة الإنجليزية. جمعت كلمات من جزأين متساويين ، رجال ونساء ، في مجموعة بيانات الكلام الخاصة بنا والتي تُستخدم لتدريب واختبار شبكات التعلم الآلي المقترحة ، وتم جمع البيانات في أماكن مختلفة من الشارع والمتنزه والمختبر والسوق. تراوحت الكلمات في الطول من ١ إلى ١,٣٠ ثانية لـ ٣٠ شخصًا. حصلنا على الميزات اللفظية بطريقة معامل ميل تردد التردد (MFCC) ، ضمن الخصائص: الانحراف المعياري ، المتوسط ، الملعب ، النافذة ، FFT ، طيف التردد وبنك المرشح ، والتي كانت عبارة عن ٢٠ مرشحًا. تم تدريب البيانات واختبارها بثلاث طرق: تم استخدام SVM للتدريب والاختبار ووصلت كفاءته إلى ٤٨٪ ، والشبكة العصبية التلافيفية (CNN) ، تصنيف عميق متقدم لتصنيف كل كلمة من مجموعة البيانات المجمعة لدينا كتصنيف متعدد الفئات ، الشبكة العصبية العميقة المقترحة CNN عادت بنسبة ٩٧,٠٦٪ كدقة لتصنيف الكلمات. تم الحصول على ٩٧,٤ ٪ من كفاءة KNN. يتميز عملنا عن العديد من الأبحاث الأخرى التي غالبًا ما تستخدم بيانات جاهزة متسقة إلى حد ما من نوع الكلمة المعزول. بينما يتم جمع بياناتنا في بيئات صاخبة مختلفة ومن نوعين من الكلام ، الكلمة المعزولة والكلمة المستمرة.

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية الهندسة / قسم الهندسة الكهربائية

نظام فعال لتميز الكلمات باستخدام طرق التعلم الالي

رسالة

مقدمة الى كلية الهندسة في جامعة بابل

كجزء من متطلبات نيل درجة الماجستير

.في الهندسة \ الهندسة الكهربائية \ الكترونيك صناعي

من قبل

اياد سبحان عبيس مطر

بكلوريوس هندسة كهرباء جامعة بابل (٢٠١٦)

اشراف

ا.م.د. هناء محسن علي **ا.م. د. حيدر مهدي عبد الرضا**

١٤٤٣هـ ٢٠٢١م