DESIGN OF ROBUST SPEAKER IDENTIFICATION WITH BUILT IN NOISE IMMUNITY

ALI NAJDET NASRET CORAN

FEBRURAY 2021

DESIGN OF ROBUST SPEAKER IDENTIFICATION WITH BUILT IN NOISE IMMUNITY

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

ÇANKAYA UNIVERSITY

BY
ALI NAJDET NASRET CORAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF
ELCTRONICS AND COMMUNCATION ENGINEERING

FEBRUARY 2021

ABSTRACT

DESIGN OF ROBUST SPEAKER IDINTIFICATION WITH BUILT-IN NOISE IMMUNITY

CORAN, Ali Najdet Nasret

PHD, Department of Electronic and Communication Engineering

Supervisor: Prof.Dr.Hayri SEVER

Co-Supervisor: Asst. prof. Dr. Murad MOHAMMED AMIN

February 2021, 107 Pages

Speaker recognition system aims to identify the speakers by their voice imprint as these imprints are processed with supervised or unsupervised model. Speech signal is considered as time variant quantity where its frequencies are keep changing with time. So, the conventional speech recognition schemes such as number of zero crossings and Fourier Transform cannot stand with uncertain nature of speech.

This work intends to be achieved two goals. The first part is made to discuss noise resistive speaker recognition system. The proposed system is composed contained of modified mel frequency cepstrum coefficients method enhances by Fundamental frequency feature coefficient to modulate the speaker. A dataset consisting of two-hundred and fifty speech imprints are applied to the proposed system so that features matrix is constructed as the dataset elements are treated under features extraction schemes using of labelled loops.

The study involved deploying of machine learning algorithms such as Random Forest, Feed Forward Neural Network, Model Freezing Feed Forward Neural Network and Particle Swarm Optimization based Feed Forward Neural Network. Each algorithm is made to learn through the feature's matrix and then, each is tested by partial data (thirty percent of total data in features matrix). The algorithms are studied invasively in order to implement a speaker recognition model with enhanced accuracy. Performance monitoring factors (metrics) re derived for each algorithm to identify the recognition accuracy as well as the Mean Square Error, Root Mean Square Error and the time taken by the algorithm to reach that accuracy. The results

revealed that Feed Forward Neural Network based Particle Swarm Optimization algorithm is outperformed among the others. A ninety-six percent of the input are correctly recognized by this model with relatively short processing time. The results show an accuracy enhancement in identification of the speakers which is made using advance optimization algorithm that is more likely Particle swarm optimization, the same enhanced the accuracy to be ninety-six percent.

The second part of thesis is to propose a model that can focus and isolate desired voice from other voices (which is termed as Cocktail Party effect). The problem motivation is, in case there are many people talking at the same time in addition to various voices form different resources such as TVs, cars, etc, then there will be a form of cacophony and interference due to all these acoustics. In order to recognized specific voice, there is a need to shut out all other voices in the background. The proposed model is utilized deep learning that have ability to recognize each person separately. combining Fully Convolutional Network (FCN) and a Bidirectional Long Short-Term Memory (BLSTM) for source separation. The FCN utilizes a convolutional neural network (CNN) to convert image pixels to pixel classes. In contrast to the CNN, an FCN converts the width and height of the intermediate layer feature map returning to the input image size throughout the transposed convolution layer, to make sure that the predictions include a one-to-one correspondence for input image. BLSTM is an (LSTM) recurrent NN that utilizes contextual info from past and future from the input/output sequences. In which the hidden layers are BLSTM layers and LSTM is the output layer. The FCN-BLSTM network is able to captures the characteristics of spectro-temporal of the audio data much better than single model (FCN or BLSTM). In this approach the FCN is applied first to acquire an initial estimation of the magnitude spectrogram of the specific source coming from the input sequence. Then the initial estimation is passed to BLSTM network to improve the output sequence of the FCN. The results show that the system is successfully isolate desired speaker voice from other voices with good accuracy as shown from retrieve voice signal

Keywords: Mel Frequency Cepstrum Coefficients, Feed Forward Neural Network, Neural Network and Particle Swarm Optimization. Cocktail party effect, Fully Convolutional Network, Bidirectional Long Short-Term Memory

ÖZET

GÜRÜLTÜ AYIRIMA ÖZELLİKLİ HOPARLÖR TASARIMI

CORAN, Ali Najdet Nasret
PHD, Elektronik ve Haberleşme Bölümü Departmanı
Danışman: Prof. Dr. Hayri SEVER
İkinci Danışman: Yrd. Doç. Dr.Murad MOHAMMED AMIN

Şubat 2021, 107 sayfa

Bu tez çalışmasında, Konuşma Tanıma Sistemindeki ses izlerine göre tanımlanmasını amaçlanmış, ses izler denetimli veya denetimsiz model içerisinde işlenmiştir. Konuşma sinyali zaman değişken niteliği olarak kabul edilip frekansları zaman içerisinde değişmeye devam etmektedir. Yani sıfır geçiş sayısı ve Fourier dönüşümü gibi geleneksel konuşma tanımla sistemleri konuşmanın belirsiz doğasına dayanamaz.

Bu tez çalışması, iki hedefe ulaşmayı amaçlamıştır. Birincisi gürültüye dayanıklı konuşma tanıma sistemlerini ele almak için yapılmıştır. Önerilen sistem konuşmacıyı modüle etmek için temel frekans özelliği katsayısıyla artırılan değiştirilmiş MEL frekans spektrum (cepstrum) katsayıları metodunun içerilmesinden oluşur. İki yüz elli konuşma izinden oluşan veri seti önerilen sisteme uygulanır böylece veri seti elemanlarının etiketli döngüleri kullanan özellik çıkarma şemaları altında işlendiği için özellikler matrisi oluşturulur.

Bu çalışma Rastgele Orman, Besleme İleri Sinir Ağı, Model Dondurma Besleme İleri Sinir Ağı, Parçacık Yığını Optimizasyon tabanlı besleme ileri sinir ağır gibi makine öğrenme algoritmalarının uygulanmasını içerir. Her bir algoritma özellikler matrisiyle öğrenmek üzere yapılır ve daha sonra her biri kısmi verilerle test edilir (özellikler matrisindeki verilerin yüzde ellisi). Konuşma algılama modelini artan doğrulukla uygulamak üzere bu algoritmalar invazif olarak ele alınmıştır. Doğruluğa ulaşmak için algoritma tarafından alınan Ortalama kare hatası, Kök Ortalama hatası ve zamanın yanında Performans izleme faktörleri(ölçütler) her bir algoritma için

tekrar türetilmiştir. Sonuçlar Besleme İleri Sinir Ağı tabanlı Parçacık Yığını Optimizasyonu algoritmasının diğerlerinin arasında daha iyi olduğunu ortaya çıkarmıştır. Bu modelle birlikte girdilerin yüzde doksan altısı göreceli daha kısa sürede doğru şekilde tanınmıştır. Sonuçlar Çok muhtemelen Parçacık yığını optimizasyonu yöntemi kullanarak konuşmacıların tanınmasında doğrulukta artış olduğunu gösterir, aynısı doğruluğu yüzde doksan altı seviyesine artırmıştır.

Tezin ikinci aşamasında istenen sesi diğer seslerden (Kokteyl parti etkisi olarak ifade edilir) odaklayabilen ve izole edebilen model önermektir. Problem motivasyon ise aynı anda bir çok kişinin konuşması ve ilave olarak TV, araçlar vb gibi farklı kaynaklardan sesler olması durumunda tüm bu akustiklere bağlı olarak bozulma ve kakafoni(ahenksizlik) ortaya çıkmasıdır. Spesifik bir sesi algılayabilmek için arka plandaki diğer tüm sesleri susturmak gerekir. Önerilen model kaynak ayrıştırması için Tam Evrişimli Ağ (FCN) ve İki Yönlü Kısa Süreli Hafıza(BLTSM) metotlarını birleştirerek her bir kişiyi ayrı ayrı tanıyabilecek derin öğrenme kullanır. FCN görüntü piksellerini piksel sınıflarına dönüştürmek için evrişimli sinir ağı kullanır. CNN'nin aksine FCN tahminlerin girdi görüntü için bire bir karşılık içermesini sağladığından emin olmak için dönüştürüşmüş evrişim (konvolüsyon) katmanı aracılığıyla girdi görüntü boyutu elde etmek için ara katman özellik haritasının genişlik ve yüksekliğini dönüştürür. BLSTM girdi/çıktı dizilerinden geçmiş ve gelecekten içeriksel bilgileri kullanan tekrarlayan NN'dir. Burada saklı katmanlar BLSTM katmanlarıdır ve LSTM çıkış katmanınıdır. FCN-BLSTM ağır tekli modele göre (FCN veya BLSTM) ses verilerinin spektro-zamansal özelliklerini daha iyi şekilde uygulayabilir. Bu yaklaşımda ilk olarak girdi dizisinde gelen spesifik kaynak büyüklük spektrogramının ilk öngörüsünü elde etmek üzere FCN uygulanır. Daha sonra FCN çıkış dizisini iyileştirmek için ilk öngörü BLSTM'ye geçer. Sonuçlar elde edilen ses sinyalinden elde edilen doğruluğun gösterdiği gibi istenen konuşmacı ses sinyalini diğer seslerden başarılı şekilde izole edebildiğini göstermektedir.

Anahtar Kelimeler: Mel Frekans Cepstrum Katsayıları, Besleme İleri Sinir Ağı, Rastgele Orman Sinir Ağı, Model Dondurma Besleme İleri Sinir Ağu ve Parçacık Yığın Optimizasyonu

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Prof. Dr. Hayri SEVER for his supervision, special guidance, suggestions, and encouragement through the development of this thesis.

I also extend my sincere Special thanks to Dear Zuhair Shakor MAHMOOD for his objective guidance and firm support. Also, I must not forget to thank Çankaya University, for their precious encouragement and facilities. I am very grateful to my parents and all of my family for their encouragement and support as this research achieved

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

AI	Artificial Intelligence.
ANN	Artificial Neural Network.
ASR	Automatic Speech Recognition.
BatchNorm	Batch Normalization.
BPTT	BackPropagationThrough Time.
ConvNet	Convolutional Neural Network.
CPU	Central Processing Unit.
DCT	Discrete Cosine Transform.
DenseNet	Densely Connected Convolutional Network.
DFT	Discrete Fourier Transform.
DNN	Deep Neural Network.
ELM	Extreme Learning Machine.
ERM	Empirical Risk Minimization.
FER	Frame Error Rate.
GMM	Gaussian Mixture Model.
GPU	Graphics Processing Unit.
GR	Gender Recognition.
HMM	Hidden Markov Model.
HOG	Histograms of Oriented Gradient.
IID	Independent and Identically Distributed.
LOSO	Leave-One-Speaker-Out.
LSTM	Long Short-Term Memory.
MAP	Maximum A Priori.
MFCC	Mel Frequency Cepstral Coefficient.

MFSC	Mel Frequency Spectral Coefficient.
MLE	Maximum Likelihood Estimation.
MLP	Multi-Layer Perceptron.
MSE	Mean Squared Error.
NN	Neural Network.
PCA	Principal Component Analysis.
PER	Phone Error Rate.
RBM	Restricted Boltzmann Machine.
ReLU	Rectified Linear Unit.
RNN	Recurrent Neural Network.
SER	Speech Emotion Recognition.
SGD	Stochastic Gradient Descent.
SIFT	Scale Invariant Feature Transform.
SR	Speaker Recognition.
SVM	Support Vector Machine.
UAR	Unweighted Average Recall.
UE	Unweighted Error.
VC	Vapnik-Chervonenkis.

CHAPTER 1
INTRODUCTION

1.1 Preface

Speech is very first method of communication and very convenient and reliable way of information exchanging for humankind. Countless information can be passed through speech among human as compared to other methods of communication such as writing or coding. Well, for different geographical lands that exhibited by man, they started communicating using particular (unique) codes and samples which is later on termed as language [1] [2].

Due to it vital role in human's life progress, languages became paramount tool to survive. More specifically, speech irrespective of its kind is the focal of this study. For instant, speech is producible as analogue signal from the so-called vocal cords vibration. [3] Human by whole is more likely recognizable by their speech since the shape of vocal track is unique in every person the voice produced from there is differs from person to person.

Speaker recognition is a long stand approach which is still under progression where large population of signal processing community are paying extended efforts to modulate the so-called vocal track to recognize the speech (voice) of individuals. Till date, general (common) speaker identification model which can recognize any speaker has not came to the light [4].

The speaker recognition systems are limited to the available data for training and so-far can not accommodate other inputs unless similar data is given to the model at the time of training (learning). The most common term in such system called as

supervised learning approach where known dataset is provided to the classifier where it can formulate the model in post learning [5].

Speech signal is defined as time variant signal which involves different frequencies that is probably changed with time. Speech is recognized at frequency in limit of eight kilohertz which represents the hear able frequency that human ear can recognize.

The term mel frequency is invented to understand the effect of different local frequencies on human ear. It represents the unit of frequency that human ear can recognize. The mel scale is shown in Figure 1 which is differed from local speech frequency. It can be produced using the so called filter banks in mel scale.

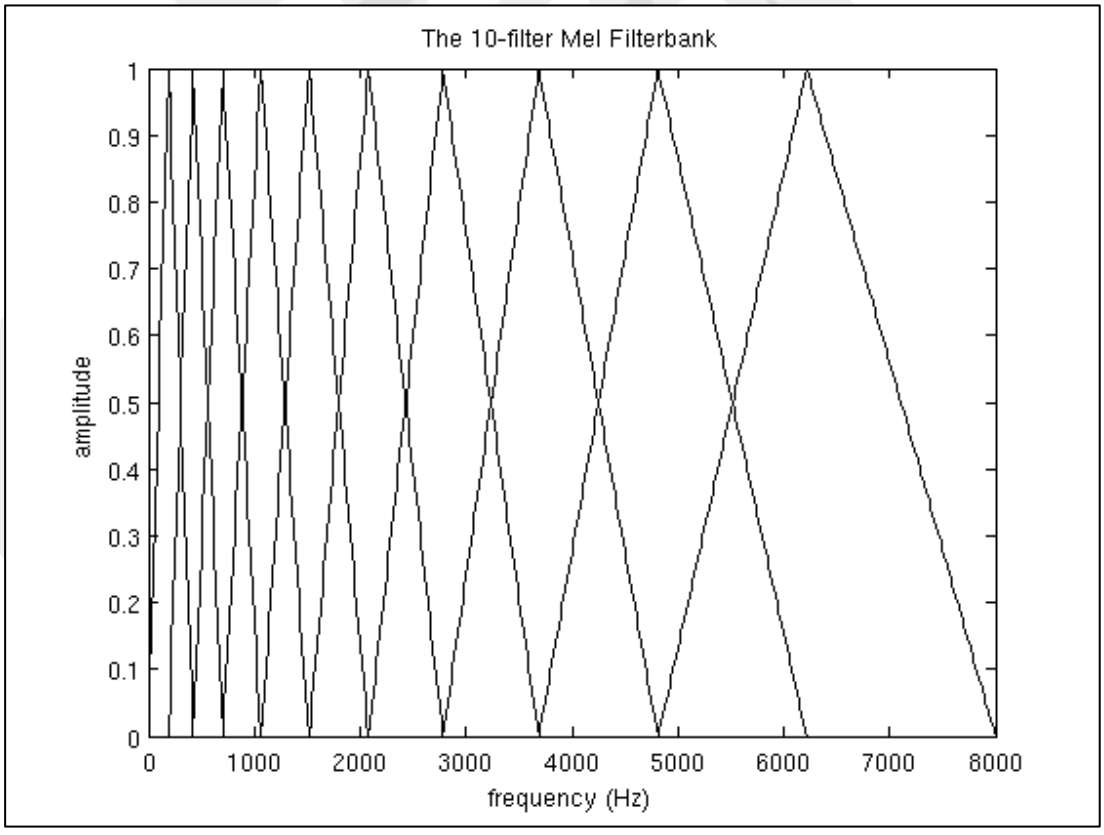


Figure 0: Mel frequency outlook in speech signals context.

The mel scale participate efficiently to understand the different voice signals impact on human. The Figure 1 depicts horizontally discriminative regions which diverge and converge frequency (can be represented according to time as well). the amount of divergence and convergence reflects the perception of human ear to the voice [6].

The random process are bigger threat (obstacles) in the way of speak signal processing where background information which are not relevant to the speaker are interventional with speech information and acts as noise. This process can not be predicted in most of cases more likely when speaker presented in outdoor environments such as roads or cafes where the surrounding voices are keep changing with time (frequencies are in continuous modification) [7].

To get away with speech processing in presence of noise, several approaches are made traditionally more likely applying several microphones to detect voices from different locality in the same area where those frequencies will be isolated from the main speaker microphone.

Here, different approaches are available to process speech signal which made great contribution in extraction the features of speech efficiently. Those approaches involves time domain signal processing and frequency domain signal processing or even more consistently hybrid signal processing which made under both frequency and time domains [8].

Several terminologies are considered as plus point while dealing with speech signals such as spectrogram and periodogram. The first (spectrogram) is important to represent the filter banks output in bother time and frequency domains. Whereas, periodogram is used to plot the frequencies information for each time period in layer wise using the color coding so, the frequencies will appear as red, dark yellow, yellow green and so on according to the frequency level [9].

Machine learning is recently exploited in context of speech signal processing and proven noticeable performance. The algorithms of machine learning are trained to (made to learn) the acoustic information of speech signal and hence to map that information to their particular source (speaker) [10].

1.2 Problem Formulation

According to the literature survey and the practical observations while dealing with speech signal, set of obstacles are realized in terms of their electrical characteristics. Speech signal processing encountered large challenges related to formulation a model of speakers (in speaker recognition applications). Two kinds of speaker recognition is realized practically: text-dependant and text independent speaker recognition.

The text independent speaker recognition model (identification) involves recognizing speakers according to their voice imprints where speaker need not to provide the same data in both training and testing. This is usually taking place using time domain features extractions. From the other hand, text-dependant speaker recognition model emphasis data similarity in both training stage and testing stage.

The problem raised when dealing with text dependant speaker identification model which strictly demand same data content to be provided while testing. The recoding of same speech signal that preserve same information twice is a truly difficult and practically can't be guaranteed as speaker may use the system in different environments where different back ground information are there.

From the other hand, frequencies participated in speech signal are directly impacted by the emotional status of the speaker. More likely, if same voice content is spoken when speaker is in normal mental or emotional conditions and same thing is spoken

when speaker is in up normal emotional status. The spectrum of both signals will never match even if speaker is same and contents are identical.

Somehow, the conventional speaker identification approaches are failed to perform in the above circumstance, traditional approaches such as Frequency domain information or time domain information more likely Fourier Transform of number of zero crossings cannot reveal much about voice signal. Speech is time variant signal where frequency components are continuously changing with time and using approaches such as Fourier transform of full speech signal is likely price less.

1.3 Research Motivations

In current days, large technology revolution is made on digital signal processing, the speech signal processing was in the lead of those advancement. Today's world is tending to depend speech in essential application of human life including automobiles (self-driven cars), smart home control system, big security systems (using the voice imprint to reveal the speaker character), applications of voice to text and vice versa, human emotional status investigations using the speech analysis, voice based costumer service (by recognize the speech and acting accordingly in fully computerized costumer care centers), more applications which is under research will come to the light soon after.

Technically, the accuracy of speaker identification approaches and application of speech signal analysis are largely depending on how pre-processing is effectively performed also depends on the amount of corruption in the signal due to ambient noise. The insisting demand of consistent speaker recognition system with accepted level of noise immunity is become high in many civil and industrial applications.

Data technology has drawn great development in current years and hence, it was used in many engineering applications to tackle complex time variant problems. The advancement of data science and machine learning field is promoted deploying this large facility in processing speech signals.

The speech imprint to identify the people is enough robust likely as other personal verification means such as eye imprint and thumb imprint. Speech imprint is paramount security method which ensure large level of authenticity as compare with aforementioned methods of speaker recognition.

1.4 Research Objectives

In accordance with research problem and motivations to form this thesis, there are some headlines of the major objects targeted in the next chapters of this thesis report. First of all, it is made in mind to derive a method to recognize the speakers according to their voice imprint using text dependant speaker recognition system.

In that data (irrespective to their noise level) are planned to be treated under consistent features extractions algorithm such as mel frequency cepstrum coefficients to derive noise resistive speaker recognition model as speech signal will be undergone the effects of pre-emphasis filter to enhance their signal to noise ration.

Secondly: implementation of the model that demonstrates the timely frequency information using a hybrid spectrum approaches and filter banks with different transfer functions.

Thirdly: derivation of common model to evaluate the speech signal impacts on human ear and how dose human perception will look like to different speech signals.

Fourthly: recalling time domain analysis to participate the overall features obtained from mel frequency cepstrum coefficients which will ack as secondary gate to verify the speakers.

Fourth: utilized machine learning algorithms to learn the acoustic behaviours of speakers and hence to identify the speakers according to their feature's vectors.

Last and not least; evaluate the speaker recognition paradigm with best possible accuracy. For that performance metrics such as accuracy, MSE, RMSE and time are to be derived for each proposed algorithm.

The last gaol of this study is to isolate desired voice from other voices (which is termed as Cocktail Party effect) by design a model that utilized deep learning in order to separate human voices and recognize each person separately, this done by combining Fully Convolutional Network (FCN) and a Bidirectional Long Short-Term Memory (BLSTM) for source separation.

1.5 Thesis Organization

This thesis report is consisting of several technical chapters that made to provide a detail explanation of speaker recognition system. Six chapters are folded under this thesis which made as following:

Thesis is beginning with Chapter one named as "Introduction" which contained the preface discussion on the speech signal processing model structure and discussing the detailed problem statement along with the motivations and objectives behind commencing this dissertation.

Chapter two entitled as “Literature Survey” which is made to provide the historical evidence of the similar attempts of research in the context of speaker recognition.

Chapter three named as “Methodology” which stated the complete methodology of features extractions of speech signal and conventional speaker modelling approaches.

Chapter four named as “Empirical Model” which is made to detail the proposed system including the proposed methodology of features extractions.

Chapter Five name as “Speech Classification” involves the machine learning paradigms that used to supports speaker identification.

Chapter six named as “Discussion” involves all result which was compared each other’s

Chapter Seven is named as “Conclusion” which enlist the paramount facts after completing the dissertation.

The referenced used in the whole dissertation work is enlisted on the final section of this report under the References heading.

CHAPTER 2

LITERATURE SURVEY

In the modern era, there are a large number of technologies that ensure privacy. One of those aspects of privacy is the identity of the speaker. Most likely, anonymity may be required in some applications such as: recording medical data where patients speak, or training call centers where the identity of the caller must be hidden. In such circumstances, the identity of the speaker should be concealed [11].

Technology of signal processing servers the speaker de identification and speaker re identification process. Speaker identification involves conversion of speech from its original form into other form (targeted form) so that the resultant speech will appear as it was spoken by the targeted form. In study, researchers attempt to perform speaker de identification by Appling the frequency wrapping technique. Form conversion performance evaluation, (how well speaker's character is hidden), entropy or GINI index can be used. Chain Corpus dataset is used in this study to examine the system.

whispering is another tradition technique of privacy delivering which is used between speakers while they are conveying a confidential information. For example, when user is speaking over the phone and he required to share credit card details or other information of such nature, he might speak (whisper) to the other party to prevent public snooping [12]. From signal processing point of view, the nature of whisper speech can be described as fundamental frequency absence while harmonic

excitation of the voice and the lower region of frequency is inclusive with formant shifting.

For discovering the identity of speaker over a whisper speech, spectrum analysis shown that both natural speech and whisper speech dominate a different identity due to the difference in their spectrum. In this study, features mapping and modification of feature extraction using LPCC algorithm is presented. This study involved using dataset of UT-Whisper corpus.

At [13], a study demonstrated the Model of Gaussian Mixture (GMM) to extract features in text-independent speech. It is however stated that model of Gaussian Mixture is outperformed to process the text-independent input speech. Speaker recognition with GMM model is taking place by using multiple components of GMM more likely thirty-four component. It is realized from this study that GMM model is good option for speaker recognition unless the large complexity of computational cost in this method. This model is tested by using dataset of eight hundred and sixty-three, the signals are recorder from forty candidates (test subjects), twenty males and twenty females.

At [14], the process involved in speaker identification is about comparing one speaker with many speakers (usually with speakers' dataset) and hence, uncovering the similarity between (matching) existence if signal matched any of dataset contents. Speaker identification is taken place either when speaker speaks particular phrase (sentence) and system is aware about this sentence, the same is called text dependent speaker identification.

Otherwise, when the speaker can speak any phrase and system is not aware about the content of this phrase but still can recognize the speaker, this is termed as text-independent identification. This study is also stated that three possibility of identification model can be derived according to the speaker language more likely in crosslingual system, the speaker language can be different in both testing and training, whither, the speaker language can be one in during training phase and multiple in testing phase which is termed as multilingual system. Ultimately, another system called monolingual system which forces single language of speaker in both training phase and testing phase. The researches in this study have mentioned that monolingual system is the common language model used in many speaker identification systems.

The problem is raised speaker's identification system is being deployed country with many local languages. For that, this study attempts to figure out the issues in cross lingual model of speaker identification. However, the state of art for processing a speech signals as it proposed by the authors of this study involve using the MFCC algorithm for feature extraction of speech. MFCC is performing by evaluation of fast Fourier transform of moving window throughout the speech signal samples. Dataset is prepared to server in this study by recording a speech using five recorders and two languages with several environment such as (laboratory, collators, rooms). The dataset is made available on IITG inventory, the recording of speech is done by sampling the speech signal at 16 KHz.

At [15], speech signal by using his info can be recognized by its language information and speaker vocal cords info. The speech which was produced from human vocal cords vibration is impacted by the shape of the vocal cords , the nature

of speech path such as moth , teeth, etc. which all collect have been formed the final content of speech signal. In recent years the technology advancement involves many applications that enabled the human to control machines remotely such as bank application and reliability by voice retrieving. Hence, in order to recognize a particular speaker, the characteristic of his vocal track should be modelled so that machine interface human application will work efficiently.

This study involves introduction of Gaussian Mixture components model to modulate the text-independent speech. This work included digit wise utterances whereas other works concentrates on full sentence or phrase utterances. Hence, training a system to recognize speech by detection the spoken utterances may require large Gaussian Mixture model with large number of components to accommodate all the possibilities of single utterance pronunciation.

In order to tackle the large training tasks, background universal model (BUM) is used. data produced from utterances of all speakers are pooled together and processed using MLLR technique in order to evaluate the background universal model parameters such as weight, mean and variance.

At [16], the first step in speaker identification system involves from end signal processing and feature extraction tasks which is directly impact the quality of recognition. In other word, voice imprint of the speaker can be recognized depending on this front-end process. Speech features are likely to be estimate at short window of time where the nature of those features is constant during that time frame.

Experiments revealed that speech signal is constant (stationary) at time frames of ten to twenty milliseconds. Dataset is prepared to serve in this study by participation of fifty-one candidates more likely a thirty-five males and sixteen females. Each

speaker was asked to provide two signals so that one will be utilized for training and other for testing. The training speech signal is made for twenty seconds and testing signal is made for ten seconds. Recording of the signal is performed as many other studies with sampling rate of 16 KHz with resolution of 16 bit per sample.

This study involves disturbing the testing signal with unity variant white gaussian noise (WGN) for testing the different techniques efficiency to process the said noisy input speech. Since the speaker might be presented at uncontrol environments more likely, another voices might be presence when speech signal is recorded, the testing signal in this study is combined with noise to meet the ground environments. Furthermore, features are extracted using MFCC and PLP methods. Results revealed that MFCC is outperform over the other technique.

At [17], in many cases where speaker recognition is required, speaker might be under stress where the resultant utterances are influenced by this stress. This study is looking after stressed speaker recognition using the aggregation method which is reported as best method for performing decision making process in such kind of input.

The overall system proposed in this study involved a five sections more likely speech pre-processing where the signal is prepared for features extraction which dominate the second section of this system. The third section is represented by modelling the voice signal by training the system to recognize the extracted features. The last section of the system involves the classification model which made for mapping the speech signals to the appropriated target after performing the training process. Study involved using of MFCC algorithm as underlying model and results

revealed that system run time is reduced relatively by thirty-five percent as compared to the similar studies.

At [18], another study is demonstrated to analysis the whisper speech signal, the researchers have stated that whispered speech is kind of regular means of communication and the spectra of this speech is completely different than natural speech spectrum where most of the spectrum contains are not similar; for that reason, recognition of whispered speech is having different aspects to be followed.

Where it is already revealed by the previous studies that voiced part excitation is totally disappeared in such signals with speech formants are being shifted to the region of low frequency; this study has observed the slop of spectrum is more-flatter in whispered speech than normal speech. In order to recognize the speaker by depending of whisper speech, several features must be depended such as: features of high-level processing, features of spectrum temporal, features from the source of voice (source unique features) and prosodic features.

The mentioned features can be extracted using three different approaches: Exponential frequency cepstral coefficients (EFCC), Linear frequency spectral coefficients (LFCC) and Mel frequency spectral coefficients (MFCC). However, the dataset served in this prototype is consisted of 125 cells where each cell is about single word, single digit or continuous sentence utterance. The data set is constructed by participation of five subjects in age range of ten years to 54 years. Subjects have been asked to utter five different sentences in normal speech and six sentences in whisper speech.

At [19], the pre-processing of people's speech signal is playing a big role in the accuracy of recognition process in whole. In practical situation (environments), speech signal may be wrapped with unwanted speech (background interference) so, the preprocessing may take place to isolate the unwanted speech or any component of none-speech nature from the input signal. One of popular algorithms that perform pre-processing is called as end point detection (EPD) algorithm. The major of this algorithm involves the discovery of the beginning point (starting) and end point in speech signal.

EPD algorithm is also determining the zero-crossing number in the said speech signal and hence it detects the noise region of the speech. All these features (boundaries, zero-crossing and noise locations) are critical to the accuracy of speaker identification system as revealed by the author of this study. The further step of speech recognition and speaker identification as well is about good features extraction, the best definition of the feature extraction is more likely conversion the speech signal into compacted vector that includes all the important speech characteristics.

In this study, authors defined the unit of Mel as human ear level of preservation which is equivalent to lesser than one-kilo-Hertz of linear spacing and bigger than one-kilo-Hertz in logarithmic scale. Author has mentioned that MFCC algorithm is outperform in speech features extraction. Post the feature extraction step, speech might be undergo machine algorithms, usually, discreet matching algorithm is used where similar features from the testing phase and training phase are analyzed discretely (point by point) in order to evaluate the minimum matching between them.

In the practical model of this study, author mentioned that forty-five objects have been participate the study and speech signals are collected from them in such way

every speaker is speaking two words. Ultimately, ninety speech signals are gathered which used as forty-five for testing and forty-five for training. In here it's noteworthy to state that recording of speech signals during the experimental part of this study is performed with sampling frequency of eight kilo-Hertz.

At [20], A speaker identification experiments were conducted on the speech data extracted from the extended M2VTS database (XM2VTSDB). Authors divided a total of 293 speakers which means one in this database into two subsets. The rate of accuracy in recognition process is given in the following formula.

Accuracy rate = the number of correct test utterances / the number of total test utterances.

As the speech signals are found too many, author utilized a Gaussian Mixture theory to perform feature extraction, GMM is integrated with Universal background Method (UBM). From the other hand and for sake of performance comparison, author of this study used LLR method for verification of speaker. All data are made as text-independent where speakers can use two different speech imprints during the training and testing phases.

At [21], wavelet transform is employed for speaker identification system backend process, author mentioned that continuous wavelet transform is used to formulate two methods of identification. The first approach in this study involves modulation of male and female identification system, unlikely, second approach is proposed to modulate a gender independent speaker identifications system. Author stated that wavelet based speaker identification system can optimize the recognition accuracy of

overall system. More likely, the system implemented in this study has reported an accuracy which equal to ninety-eight percent.

The wavelet theory applied on finite discreet function $f_k = [f_1 f_2 f_3 f_4 f_5]$ is firstly required to evaluate the weight coefficients which form the weight vector $W_k = [W_1 W_2 W_3 W_4 W_5]$, the discreet wavelet transformation will attempt to approximate the input function; hence, according the example f_k input, the result will be approximation input $A_k = [A_1 A_2 A_3 A_4 A_5]$. Mallat theory is usually used for evaluating the weight coefficients. However, this study involves eight objects (candidates) where four of them are males and rest are females. Every object is asked to produce a ten different speech imprints so that, eighty signals are totally served as a dataset for this study.

At [22], segmentation of speech signal is vital to the success of further (high-level)speech processing in ant speaker identification system. This study keen on implementation of enhance segmentation method to be applied on speech signal to optimize the correlation between the practical (real) characteristics of speech and the likelihood vector of utterances.

Authors of this study revealed that speaker gender information availability is directly influence the performance of segmentation and hence influence the recognition accuracy of the speech. The study has utilized large elements dataset called as Farsdat which is being analysed for speaker recognition using multiple component gaussian mixture method (GMM). At the end, author made a note that segmentation process were taken place by hamming window of twenty-four milliseconds with shifting of ten seconds and overlapping of ten seconds.

At [23], text independent speaker recognition is required by majority of virtual applications such as those applications involves voice interaction between human and virtual machine. Human (user) voice should be well recognized by such application in order to ensure the security of information and authenticity.

Modification of other technologies such as data mining techniques and artificial intelligence has made the advancement of those technique usable by all engineering sectors. Authors here has demonstrated the big role of neural network to classify the voice data. Voice data can be flexibly classified using the probabilistic neural network for the application adopted in this study.

The results of this study contained an accuracy of recognition equal to ninety-six percent by using short voice imprint during the testing phase. Populist neural network was said as best option in this study since the design simplicity and straightforward functionality in classification tasks.

Form the other hand, authors of this study have stated that the probabilistic neural network is fit for classification of complex problems with many advantages such as ability to upgrade the data any time and the accuracy in the results of classification. One realized drawback is the longer time of processing due to the complexity of its internal operations.

Data is collected from twenty-eight candidates (twenty-one female and seven males). Each speaker is asked to provide the voice clip twice so that one will be used for training and other for testing. However, the voice is recorded from all candidates using the same microphone and hence the voice is converted into digital format using ADC with sampler of 16 KHz. It is noteworthy that text dependant speaker identification system is implemented in this study.

At [24], relatively new approach is proposed in this study unlikely as other previous research activates in speaker recognition systems. Speck signals clustering is another approach of recognition which is made base on support vector machine (SVM) algorithm which attempts to cluster the group of speakers in several iterations where each iteration involves elimination of similar elements (signals) as sub-cluster so that, only single cluster will be residual at the end of the process.

This approach is shown moderated accuracy for large speaker's recognition system (above seventy speakers) as compared with the traditional Gaussian Mixture Method (GMM). The time consumption of the last i.e., GMM is relatively high. The segmentation is made as twenty-five milliseconds hamming window with ten milliseconds of shifting (repetition).

The dataset served in this study is chosen in random fashion from the corpus of NIST-2002. Experiment is commenced by detection of silences in the speech signals and removing of it using the voice detection-based energy level method. The number of selected speakers from the mentioned dataset is only thirteen speakers. However, Mel Frequency spectral coefficients is used for feature extraction.

At [25], study emphasised that speech is vital for human communication and daily life information conveying. In order to recognize the speaker, several approaches are proposed depending on the nature or physical prosperities of the speech signal. Those techniques are essentially studying these prosperities and extracting it to be unique identifiers for the signal.

The study has listed those technologies as per the previous researches as Mel frequency spectral coefficients, Gaussian Mixture Method and Bark frequency spectrum coefficients. However, the author mentioned that Mel frequency spectral

coefficients and Gaussian Mixture method are outperformed with respect with the others. This study involved a twenty speaker as test speakers where each speaker is providing five voice clips for training and other five voice clips for testing. The candidates of this study were ten males and ten females.

At [26], the features extraction from speech signal may involves several features; so, as features are extracted, a vector may be formed to accommodate those features (called as features vector). The recognition process is taking place by finding the similarity level or the probability of speaker detection. It is achievable by measuring the distance between the features in both dataset features vector and the unknown speaker's vector of features. This study were keen on development of text dependant speaker identity system that recognize the speaker as they provide the digit wise password to the system such as two, three, etc. the speaker is hereby required to utter a single digit at the system.

The base of this study lies on how accurate the results will be after using the distance estimation approach to find the minimum distance between the test features vector and training features vector. The cosine distance evaluation method is employed in this study to find the distance between the vectors.

A forty-five candidates are called to participate the experiment where each candidate is been asked to provide a one voice imprints (single digit word) for the training phase and same for the testing phase. Eventually, the experiment is conducted by using a ninety speech signals. The recording of the voice imprints was done at eight kilo-Hertz sampling frequency.

At [27], Indonesian Database for Speaker Recognition (IDSR) is introduced during this paper, that is associate Indonesian corpus collected mistreatment multi-languages in Indonesia, which are Bahasa, Javanese, and Sundanese. Speech is recorded using multi-recording devices planning to bring session variability downside toward the building of speaker's model. Session variability issues is that the main focus in our experiment using IDSR. The experiment performed shows that session variability may be handled using correlational analysis approaches. correlational analysis approaches show promising results for each case data} sort employed in this research. it's argued that to handle session variability problem effectively, particularly within the case wherever multi-channels included, varied information which are collected from completely different channels got to be provided. However, if the necessity couldn't be met, the information uses to make the model ought to be recorded in high quality. within the future, IDSR are accustomed analyze the language impact towards the performance of speaker recognition. The examination towards however well the system performs once the speaker is talking mistreatment over one language will be done. Furthermore, additional data has to be collected to get style of speakers. he recording was taken in a very laboratory surroundings to attenuate the doable noise. There are 3 totally different recording devices used: AT2005USB electro-acoustic transducer (cardioid), Scorpion- H8633 receiver (omnidirectional), and commonplace hand phone (LG-H502F). These devices begin and end the recording within the same time. The microphone and also the headset are connected to Macintosh laptop computer for recording, whereas the hand phone uses it own OS recording software package once the recording is being held. All of the recordings are recorded and sampled at 44.1 kc and 16-bit rate.

At [28], in the laboratory environment, speaker recognition has made great progress. But when compare it with real life the performance of automatic speaker recognition system is depended on the different factors, the very important is a healthy condition and environmental noise. In this study have been experiment the performance of speaker identification system when the test side is suffering from the cold. The cold direction to swelling of the nasal cavity and induce inflammation, then convert the modulation of nasals to sound supply excitation signal and makes the speaker's voice changed. it had been found by previous researches that speaker recognition system's performance considerably decreases once taking traditional speech spoken by the healthy persons as train speech, whereas cold speech by persons who are catching cold as take a look at speech. In this study , through studying the composition of nasal and comparing the frequency domain properties of cold speech and normal speech, the authors find the cold makes the LF components larger and HF components smaller. So authors propose the method using various pre-emphasis filter process normal speech and cold speech. Experimental results show that in this method can improve the performance of the speaker identification system by 6% compared it with general method all speeches are processed with the same type filter.

At [29], performance monitoring of the speaker identification system is vital for the system optimization and further enhancement which to be done on the system. The monitoring of this system specifically about determining the error rate or error probability in speaker recognition. Sue to the leak of diversity in the techniques to monitor the performance, this task is considered very tough in practical situation.

In this study, performance is assessed using logarithmic likelihood method for estimation the error in speaker recognition through conducting several tests

(experiments) and in each experiment the estimator is applied. The strength of this estimator lies on the dependency of number of tests, in other word, estimator may reach a minimum variance with only five test. So, the method of logarithmic likelihood estimation is tackle the problem of test (experiments) shortage.

The experiments were conducted using the YOHO database. However, the speech is recorded from one-handed and thirty seven speakers in such way every speaker will speak twice so that one speech imprint will be used for training and other will be used for the testing. The study is covered the pre-processing part as author mentioned that speech signals are undergone silences removing process from the both ends. Hence, signals are segmented where each segment is of twenty-five milliseconds length and ten milliseconds overlapping.

At [30], a multi-model system to recognize speakers is introduced in this study using the Gaussian mixture method (GMM) and supported by universal background model. The study attempts to describe the importance of using the same for speaker identification system. Two different datasets are used to train and test the model respectively, the experiments had taken place at three locations (that is about allowing the speaker to generate the voice signal in three different environments). The first dataset which is used for training is TIMIT-corpus and the other dataset that used for testing is NIST-corpus. The accuracy achieved over each environment is differs by twenty-four percent. Every dataset involved using various recording ways such that in TIMIT corpus, one hundred of speakers are spoken in the first environment (laboratory room) which is used for training. While the other dataset (NIST corpus) is used for training and recorded when same number of speakers provide their voices through a mobile phone calling.

In other hand there is an issue for voice recognition that is known as “The cocktail party effect”, which is that the most tough challenges in audio signal processing for over sixty years [106]. during this tackle , the goal is to separate and acknowledge every speaker in extremely overlapped speech recordings, as frequently happens in an exceedingly cocktail party. though humans will solve this downside naturally while not a lot of effort, it's very tough to make an efficient system to model this process. thanks to the big variation in commixture sources, the party downside remains unsolved. before the deep learning era [107] several makes an attempt were made. The approaches projected are often divided into 2 categories: mono systems and multi-channel systems, wherever the difference lies within the variety of recording microphones involved. In single-channel systems, the separation method entirely depends on the spectral properties of speech, love pitch continuity, harmonic structures, common onsets etc., and this will be achieved by victimization applied mathematics models [108], rule-based models or decomposition based mostly models .In multi-channel systems, the separation method will exploit the special properties of every source. numerous beam forming, or a lot of exactly special filtering, strategies were projected by using, for example, freelance element Analysis [107]. Alternatively, clustering-based algorithms conceive to cluster time-frequency bins to individual sources by using special features. there's conjointly a clustering-based approach that uses both special and spectral options. However, no matter the amount of microphones being used, most existing systems work just for rather easy scenarios, for instance mounted speakers, restricted vocabulary, mixtures of various genders etc., and can't generate satisfying performance for general cases. The booming of deep learning has brought progresses during this problem. completely different from most different deep learning tasks, multi-talker separation has 2

distinctive problems: a permutation downside associate degreed an output dimension problem [108]. The permutation problem arises thanks to the very fact that the majority deep learning algorithms need estimation targets to be fixed, whereas in multi-talker separation, arbitrary permutations of the separated sources are equivalent. The output dimension problem refers the very fact that the amount of blending speakers varies in numerous samples, that creates problem in learning as a result of a neural network typically needs a set spatial property at its output layer. 3 single-channel neural network models were proposed, particularly the deep bunch (DC)[22], deep attractor network(DAN) [23] and permutation invariant training(PIT) [24]. In DC and DAN, every time-frequency bin in mixture spectrogram is mapped into the next dimension representation, i.e. embedding, where the bins from an equivalent speaker are closely situated to every other. the 2 problems are resolved by the affinity learning in DC and DAN. PIT was first of all proposed in and was shown to attain comparable separation performance in [108]. PIT follows the mask learning framework [109], wherever the network 1st generates the output mask for every target speaker, followed by AN complete search of combination between the output and therefore the clean regard to fix the permutation problem. The 3 algorithms mostly boost the state of the art in speech separation. The analysis showed they achieved similar performance for two speaker and 3 speaker separation on common knowledge sets. though the deep learning-based ways achieved breakthrough within the party problem, it is still tough to use them to world applications as a result of 2 reasons. Firstly, their separation power has inherent limitation. For example, when there are 4 speakers, even for the most tractable scenario, i.e., two males and two females, single-channel separation seems almost impossible because the mixture is so complex that each speaker's voice is mostly masked by other speakers. Secondly,

the current single-channel systems are usually vulnerable to reverberation. This would be because the reverberation blurs speech spectral cues which the single-channel separation systems leverage on to isolate each speaker.

At [106], proposed a model based on utilized a complex convolutional DT approach to cocktail party source separation using spectrograms. The model is featuring both parametric statistical estimation of spectrogram magnitude and circular statistical estimation of phase. The convolutional DNN was trained on two minutes of the speech of two speakers and tested on 10 seconds of new speech from the same speakers. The separation results show it are on a par with equivalent binary-mask based non-complex separation approaches and the separation quality is similar to binary mask based convolutional DNN approaches but features slightly improved artefact performance. Although the DNN employed here is of only 3 layers, if we consider the degree of abstraction already provided by the STFT and inverse STFT.

At [107], proposed a methodology composed by a listener and a speller that is capable of processing audio input and outputting a sequence of words contained in the audio. The listener component is a pyramidal Bidirectional Long Short-Term Memory (pBLSTM) neural network. Basically, the BLSTM inputs audio through filter-bank-spectra frames and reduces its dimensionality using a pyramidal approach in BLSTM.

At [108], presented a novel combined three-dimensional convolutional architecture for audio and visual stream networks along with convolutional fusion in secular dimension (by making use of three dimensional convolutional and pooling operations) and combining between the networks. The experimental results of various data sets approved that the proposed architecture beats the other present

methods for audio and visual matching, in addition reduces the number of parameters considerably in comparison to the previously proposed methods. the performance results of their model present the effectiveness of the learning of when employing CNN. The integrated of local speech representative characteristics are proven to be more probable for audio and visual recognition by using CNN.

At [109], propose an audio separation model combine fully convolutional neural networks (FCN) and Bidirectional long short-term memory (BLSTMs) which is a type of recurrent neural networks. They advantage of used both networks are that the FCNs are effective for extracting useful features out from the audio data while BLSTMs are suitable at modelling the temporal structure of the audio signals. The experimental results of their model show that the combining of BLSTMs and FCNs achieved much better separation and high performance than using every model individually.

CHAPTER 3

METHODOLOGY

3.1 Outline

Recognition of speakers in oral communication is essential routine of humans where speakers are recognizable from their voice. Such is natural activity in mankind where oral conversation is most dependent and popular kind of communications. Since 1970s, computer-based speaker recognition approaches began to practice. The main goal of those computerized systems is the establishment of digital speaker identification systems capable to recognize different speakers by analyzing their voice imprints [31].

Speaker recognition system can be split into two parts: speaker identification and speaker verification. During the identification process, speaker may provide a voice imprint to the system and hence system will analysis that. The acoustic analysis during speaker identification will be made to determine information such as zero crossing, voice part, pitch period, mel coefficients, etc. which will act a speaker identity and used to recognize a particular speaker [32].

By whole, speaker verification is vital process in speaker recognition system, it is used to map a known speaker information (a mentioned above) to their real speaker. In other word, speaker verification is processed to shortlist a speaker from a group of speakers according to his/her aquatic information [33].

Speaker recognition systems are also categorized according to speech information nature. More likely, it can be said as text dependent speaker recognition system if and only if the same speech signal is used in both identification and verification

process. Otherwise, system may be called as text independent speaker recognition system if different speech imprints are deployed for identification and verification [34].

However, in text independent speaker recognition, different signals to be used while identification and verification. Known that those signals may relate to same speaker, in other word, speaker can be recognized irrespective to what he said [35].

This chapter demonstrates the method of conventional speaker identification, furthermore, a framework of the proposed speaker identification system is described. The methods of testing and training for implementing speaker verification (mapping) paradigm are also detailed along with different machine learning approached which are being used to enhance mapping paradigm [36].

3.2 Speech Signal

The speech signal is physical quantity produced from human as air flows over the vocal cords. The vibration of the cords in human through is reducing the so-called voice [37]. The final voice of human is formed after passing from vocal cords into the mouth. Final form or hearable rhythm of human voice is differing from human to another since the shape of vocal track is not identical in all humans. Furthermore, the physical structure of mouth such as (tooth presence, tongue shape) is directly influencing voice rhythm and essentially contributes formation of final voice. The process of voice production in human vocal track is demonstrated in Figure 2.

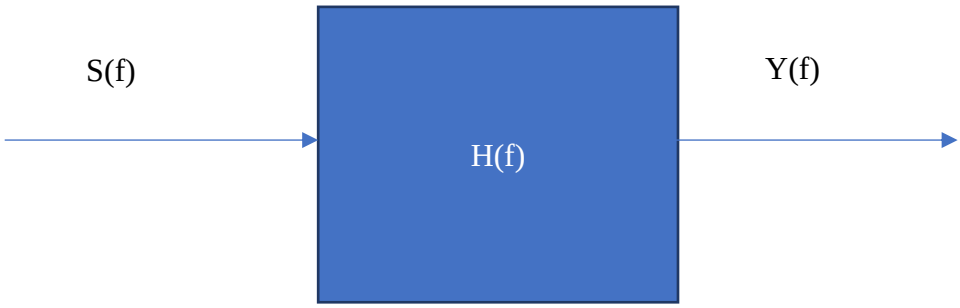


Figure 2: Modelling of voice production in vocal track.

If input signal (time domain nominated signal) is passed through a vocal track the following expressions can be made.

$$S(f) = \sum_{n=-\infty}^{\infty} s(t)e^{-jnwt} \quad (3.1)$$

$$H(f) = \sum_{n=-\infty}^{\infty} h(t)e^{-jnwt} \quad (3.2)$$

$$Y(f) = H(f).S(f) \quad (3.3)$$

The term $H(f)$ is representing the influence of vocal track on the said voice signal $S(f)$ which produces an output equal to $Y(f)$. In order to realize the process and keep track of each point of the signals, signals are being sampled with appropriate sampling frequency. Samples can be tracked during the frequency domain as well [38].

The time domain natural speech signal produced by human vocal track can be depicted in Figure 2. Where the x-axis is representing time and y-axis is representing the amplitude of time components. Figure reveals a voice signal sampled at sixteen kilohertz, different amplitudes as depicted in the figure corresponding to different samples [39].

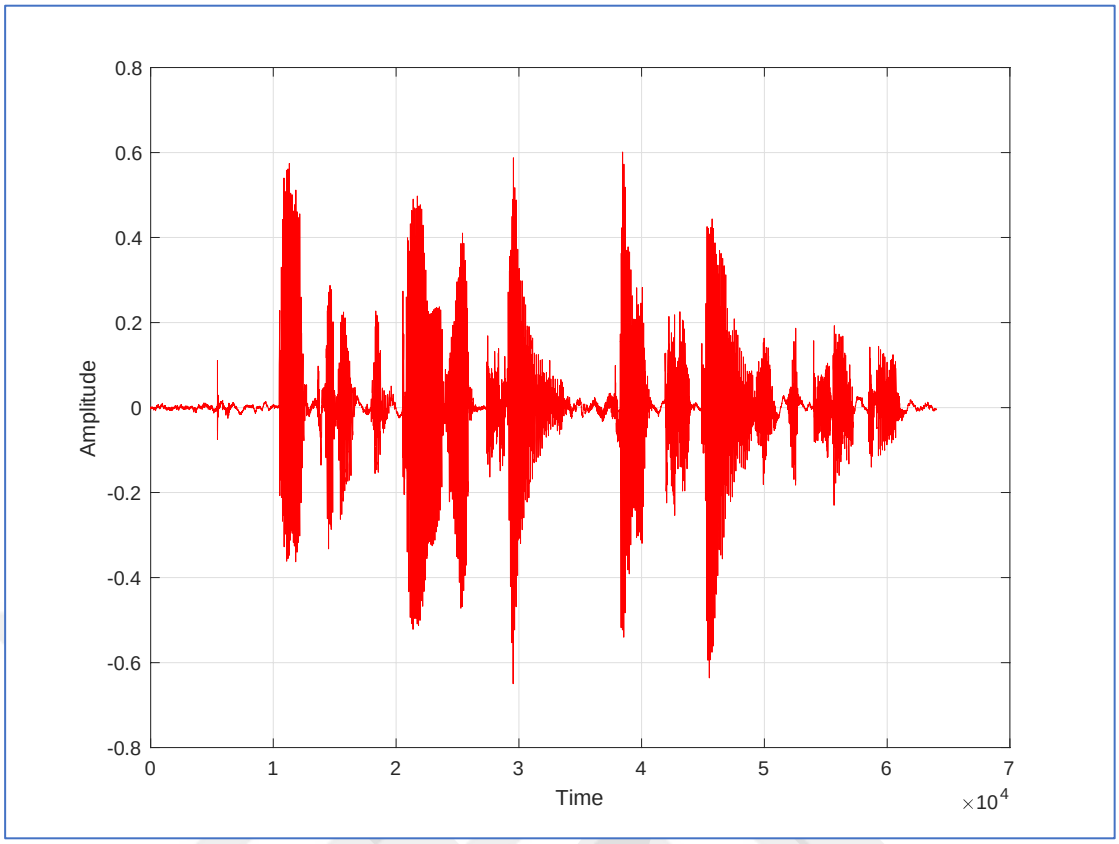


Figure 0: Waveform of natural speech signal.

Large set of information can be derived from speech signal, some of interested properties in speech signal which can participate the speaker identification process are listed herein.

1. Unvoiced samples: which stands for the region available in speech signal but not vital to speech information. The unvoiced parts are usually taking place due to conversation or speech breaking which might happen during the speech recording.

In order to uncover and identify the unvoiced parts, the following algorithm cab be applied. First of all, signal is assumed to be sampled at known sampling frequency and accordingly sample by sample voice processing may take place with light of the algorithm demonstrated in Figure 2.

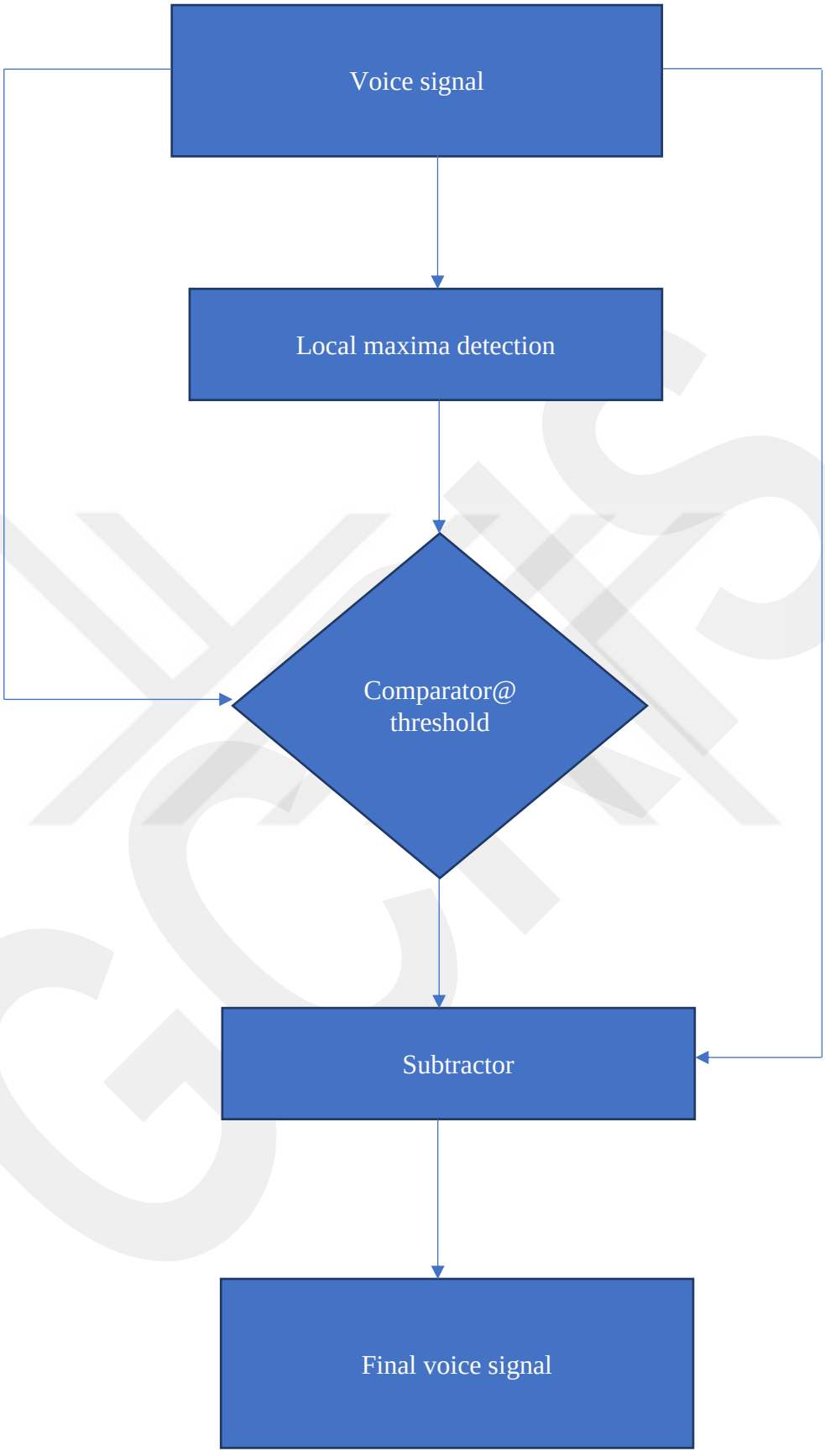


Figure 4: Voiced only signal-unvoiced part filtration prototype.

2. Zero crossings: voice signal is attributed as none stationary time variant signal, the virtual analysis of the signal revealed that different frequencies can be detected in the signal. However, number of zero crossings in the signal may reveal the frequencies a combined with the said signal. Zero crossing is paramount method of digital signal processing which uncover the numbers that signal cross the x-axis. Figure 5 depicts segment of speech signal that shows different frequencies combination of same speech imprint. This method of zero crossing determination is outperform in finding the frequencies associated in speech imprint in time domain analysis without involving complex calculations [40] [41].

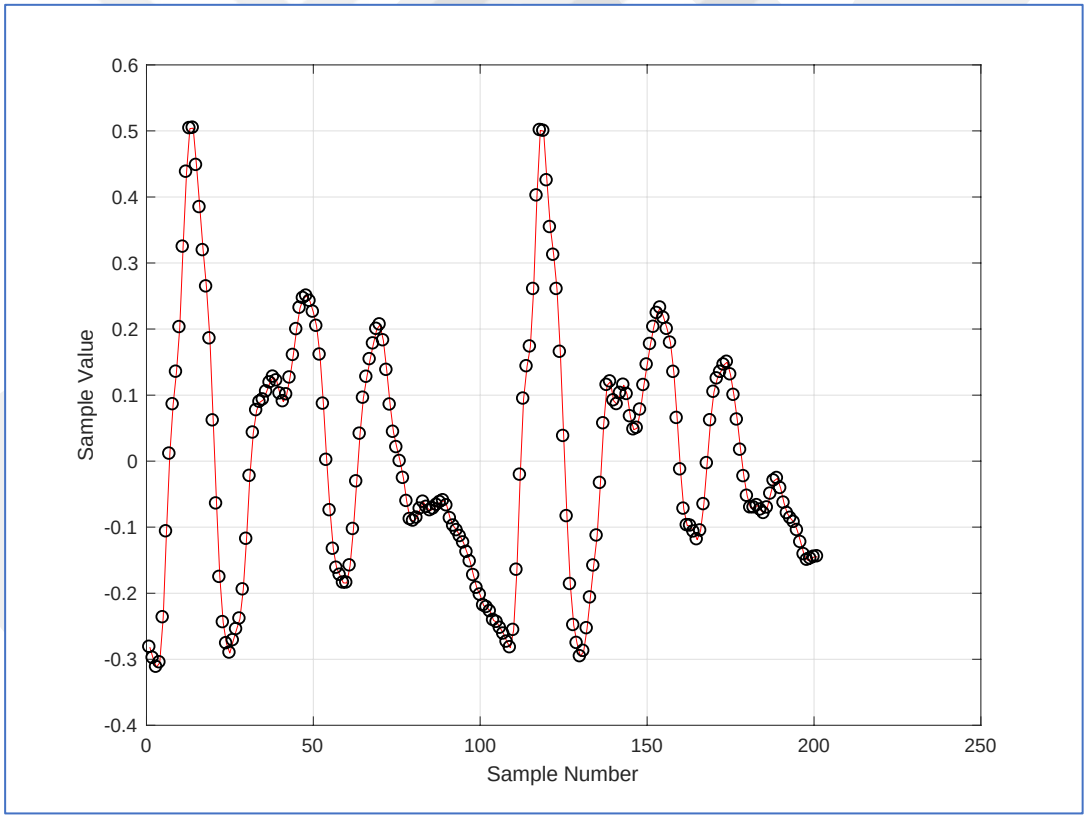


Figure 5: Sampled segment of speech signal shows several zero crossings.

3. Another way to decompose the frequency components of speech signal is by Fast Fourier Transform (FFT). The general formula of FFT is used in previous section

and given in Eq. 4. For x_n is time domain voice signal, X_w will be same signal represented in frequency domain [42] [43].

$$X_w = \sum_{n=-\infty}^{\infty} x_n . e^{-jnwt} \quad (3.4)$$

The resultant signal from Fast Fourier Transform is demonstrating all frequencies participated the signal, Figure 6 depicting the frequency domain of the voice signal shown in Figure 5.

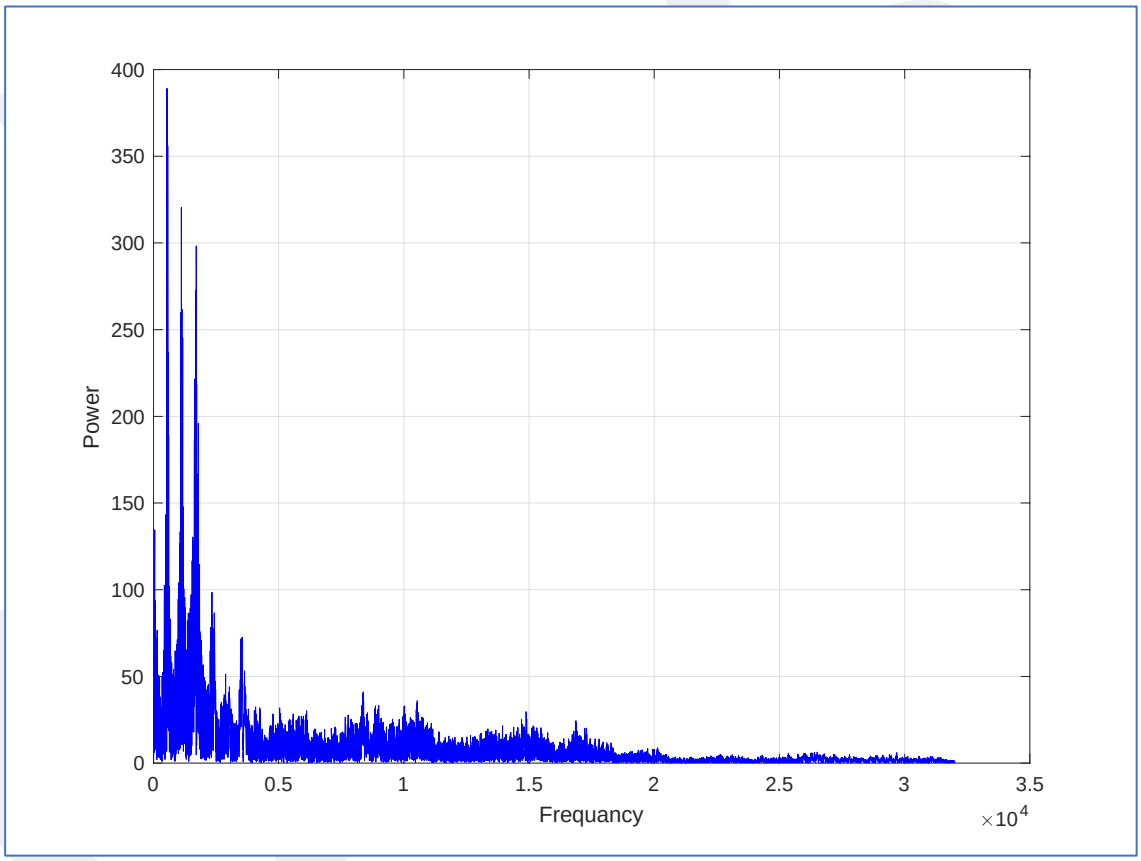


Figure 6: Frequency domain representation of speech signal.

Figure 6 depicts power spectrum distribution on each frequency, so-to-say; higher power lies on period between zero to five kilohertz. The frequencies in range of twenty kilohertz and above contains minimum power spectrum density. Usually

those lower frequencies involve unvoiced parts which are not required while the analysis [44].

The unvoiced of very small power components of the speech signal can be removed using the algorithm detailed above. Those samples are removed from whole signal so that only speech information will remain in the signal. With FFT, signal can be viewed according to their frequency information and hence further signal processing such as filtration, modification, modulation, etc. can be achieved smoothly [45].

3.3 Conventional Approach

Conventional approach of speaker recognition relies directly on the data and parameters obtained from speech signal (imprint) assuming that signal is interferences consistence data. So-to-say, speaker model is to be driven from the obtained coefficients and hence the model will work to recognize the speaker if and only if intrinsic speech signal is feed in. Noise influence of such signals may need to remodel the system to recognize the noisy influenced signal [46].

This system is attributed by its dependency on none analytical parameters of the signals which considered as plus point for the same. From the other hand, signals fluctuations for any reason may be required to rebuild (redesign) the model.

The conventional model of speaker recognition consists of two essential stages, more likely; identification stage and mapping stage. the both stages contained of regular analytical blocks that constructs whole speaker recognition systems [47].

During speaker identification process, model will be formulated for every speaker. Model is then known as speaker model which stands as speaker identity. This model contains all speech signal coefficients derived from signal processing.

In conventional speaker recognition system speech imprint may go through several steps (blocks) of processing in order to derive the signal coefficients used in speaker identification. The first step is signal recording. A recorder with known sapling rate is to be implemented to intake the speech from the speaker [48].

Speech signal is usually recorded using sixteen kilohertz sampling frequency and sometimes it can be recorded by thirty kilohertz of sampling rate. Higher sampling frequency results more accurate information of speech and however, it may cause computational complexity. For that reason, most of speech processing applications are intaking speech in sixteen kilohertz of sampling [49].

If the voice frequency that affect human hear (he arable frequency) is said to be eight kilohertz, the sampling frequency (minimum sampling rate) that required to retrieve the same signal is given in Eq. 3.5.

$$F_s \geq 2\, f_t \tag{3.5}$$

Where F_s is sampling frequency and f_t is the natural local frequency of the speech signal. However, as signal is sampled with suitable sampling frequency, the following steps are followed in hereafter:

1. Pre-processing: in conventional speaker identification system, voice might be reordered directly (on field process) and hence the voice signal is set undergo several stages aiming to isolate other signals interferences such as noise and unwanted background voice. In pre-processing and as discussed in early sections, speech signal is to be prepared for noise removing or silence elimination which might mitigate the computational cost. The common expression of frontend speech signal is given in Eq. 6 where $y(n)$ is the final speech signal at the first stage of system and $x(n)$ is original speech signal without ambient noise and $k(n)$ is a function that represents the ambient noise [50]; so, the final speech signal can be represented as:

$$y(n) = x(n) + k(n) \quad (3.6)$$

Pre-processing is responsible to produce a form of signal that is ready for features extractions by removing the ambient noise and background unwanted information. In order to mitigate the noise (ambient noise) in audio signal, noise must be monitored efficiently. The common ways to monitor noise in speech signals can be such as: monitoring the frequency relying on frequency domain information in speech; this can be achieved using the Fast Fourier Transform; from the other hand, in most speech processing projects, two microphones are placed to detect the speech and to detect the background voice respectively. In order to measure the effect of noise in the signal, both noise and speech power must be measured and substituted in signal to noise ratio formulate given in Equation 3.7.

$$SNR\ (watt) = \frac{P_s}{P_n} \quad (3.7)$$

$$SNR\ (dB) = 10\ log_{10} \frac{p_s}{p_n} \quad (3.8)$$

2. Features extraction: in this stage of speech processing, features vector will be formed in order to be used in further identification process. Features that most useful in speech processing more likely, zero crossings, speech signal energy and power, signal entropy, auto correlation etc.; are being detected from the said voice signal and merged in one vector called as features vector. In order to derive a general form of features extraction process, Figure 7 is depicted for that purpose [51].

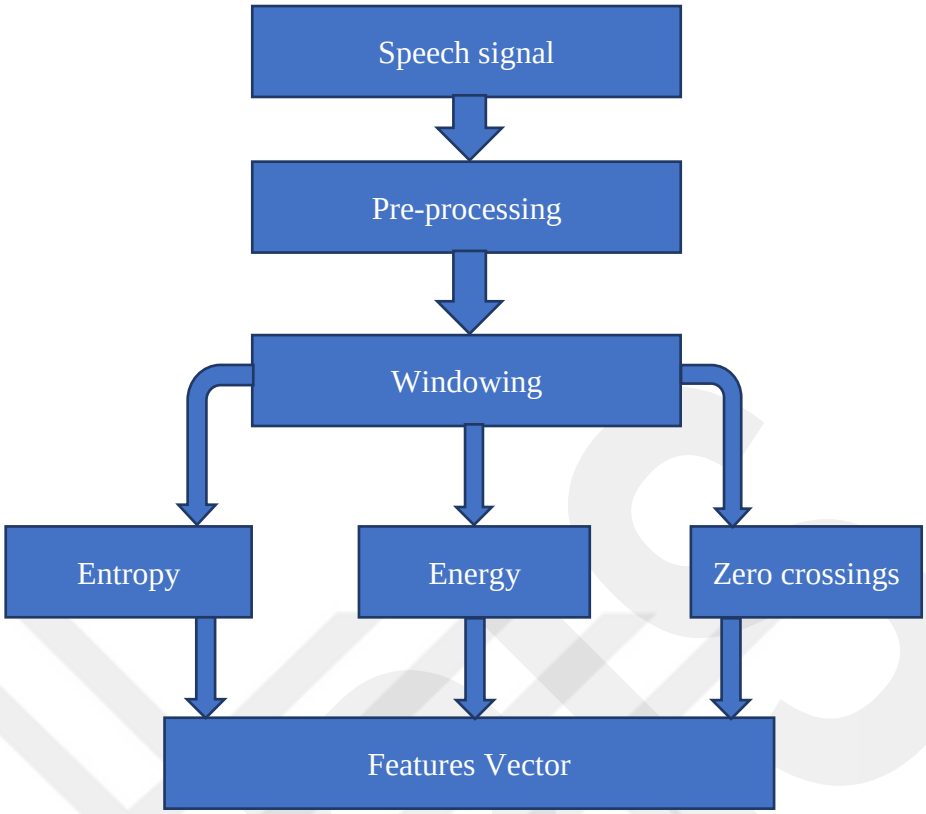


Figure 7: Features extraction framework.

Measure of signal energy is a key point of features vector which reveals the amount of energy preserved by the signal as given in Equqation3.9.

$$E = \sum_n |x(n)^2| \quad (3.9)$$

For several voice signal, features vector is calculated for each signal and hence the final dataset that correspond for features vectors of all candidates in the system is formulated. Eventually, a matrix called features matrix is produced. In order to implement a consistence voice recognition system, two tools are mostly required: features matrix and target vector. Those tools serve in mapping process where a classifier is trained to route particular vectors from features matrix into their corresponding target.

3.4 Speaker modelling

Speaker recognition system is playing vital role in recognition of large number of speakers which is completely differs from single task system that is capable to recognize a single speaker. In order to implement such system, every speaker must be modelled according to its aquatic information that gathered in features matrix. With pool of speakers, large features matrix is formed and however, recognition system must rely on this matrix to discover the particular speaker among set of speakers [52].

Since speech signal is time variant and no fixed procedure is available to tackle the uncertainty of speech. So single speaker model can't be used to accommodate all other speakers due to the nature of voice signals. The following attributes can be overseen in all audio signals [53].

1. One-way analysis: which means that recoding same signal for more than once is not compulsory yield same features [54].

2. Background information: speech signal is usually containing an information related to other sound sources more likely, the voices surround the speaker. Such information may disgrace the concern speaker information [55].

3. Some categories of voice more likely whispering voices are not having clear spectrum which make then untraceable while spectrum analysis.

4. Speech signal can preserve some features in they would analysed in very short window such as ten micro seconds [56] [57].

5. According to the literature, mel frequency cepstrum coefficients (MFCC) features vector is played paramount role in recognizing the speakers. This algorithm is time preserved and relied contained most efficient signal processing approaches such as

pitch period calculation, auto correlation, digital filters and flexible windowing which make it consistence enough for speaker identification [58].

6. machine learning algorithms are paved the way for flexible and reliable speech analysis independent of number of features. Those approaches have tackled the time and accuracy limitations in voice recognition methods. The most reliable way to perform voice recognition in current days is achievable by merging the digital signal processing schemes and machine learning algorithms [44].

3.5 Speakers Mapping

In order to develop speaker recognition system, speech signals are analyzed to derive their features and to establish what so called features matrix. This matrix includes large number of information which represents all speaker's acoustic information. In the proposed model, speakers are modelled according to their mel frequency information and according to their pitch period [55].

The overall framework of the speaker modelling is illustrated in Figure 8. The herein points are detailing the process of establishment the speaker model:

3.5.1 Dataset preparation

Generally, dataset is to be established by participation of several candidates (known as objects) where everyone provides his speech so that it can be recorded and stored. Dataset volume in speech applications or speech-oriented signal processing projects is slightly high since large number of objects are participating the experiment.

Dataset is changed base on the project requirement or the nature of application. In some applications, speech data need to be text dependent and in other applications it should not be so. Furthermore, special kind of speech data might be required in some

experiments such as whispering speech or gender related speech or age-related speech (e.g. children speech recognition) [55-61].

Eventually, voices from all candidates are captured in such way that every object is asked to speak particular text for known time. The recorder must be set for well sampling frequency (double or greater than double than local speech frequency). Speech (voice signal) to be recorded using one of voice format (encryption) more likely (.wav, .mp3, etc.), single or dual channel to be used while recording more likely, stereo or mono channel [63].

A noteworthy feature in dataset preparation is that higher sampling frequency may not always lead a success in speech processing system. Higher sampling will defiantly increase the number of captured points (information) in the speech signal which intern complicate the so-called computational process. In order to avoid such occurrence, proper sampling frequency must be set. Known the fact of sampling frequency must satisfy the equation (6) and the hearable voice frequency is eight kilohertz, a proper sampling frequency can be sixteen kilohertz [64].

Well, this sampling frequency is deployed in the majority of speech recognition projects as seen in the literature. Unless application requirement of particular sampling frequency, a sixteen kilohertz is safer and accurate enough for the most applications.

For large speaker identification systems, off-line recognition is required to develop a consistence speakers' models, dataset is outsourced from previous experiments or from well-known data banks. The process of deployment large speech dataset in voice recognition model is done by implantation of proper indexing loops as illustrated in Figure 8.

Dataset is firstly made ready along with their sampling frequency, since the number of voice imprints is high and all voice imprint must be feed into the recognition system, indexing loops is used. all speech imprints are labelled by particular simples i.e. names or numbers and those labels are collected under array called indexing array.

The further step of voice processing is features extraction in order to form the features vector and then features matrix. All the analysis procedures must fold under the indexing loops in order to extract all features details for every object in the dataset.

Some ready-made datasets which is available in online sources might be found will missing data more likely in case of voice imprints, some imprints are not available and this will cause un traceable error in the further process. So, it is highly recommended to review all voice imprints and match them with data indexing array prior to go ahead. In case of missing voice imprint, the name or serial number of that missing clip must be omitted from the inventory array (indexing array) [70].

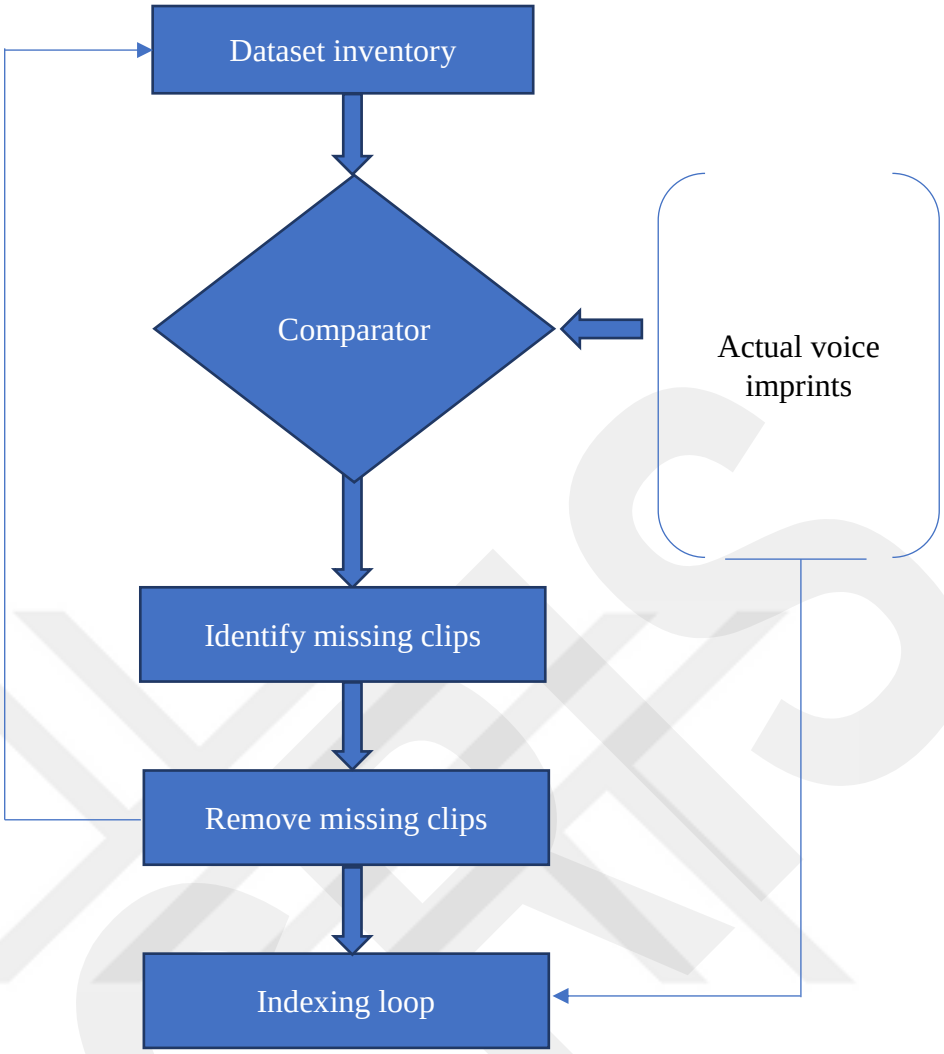


Figure 8: Dataset preparation corpora.

The voice inventory data which forms the index array is demonstrated in Figure 9., figure shows the general appearance of the indexing array. In case of this project, two-hundred and fifty voice imprints available in this inventory/index array.

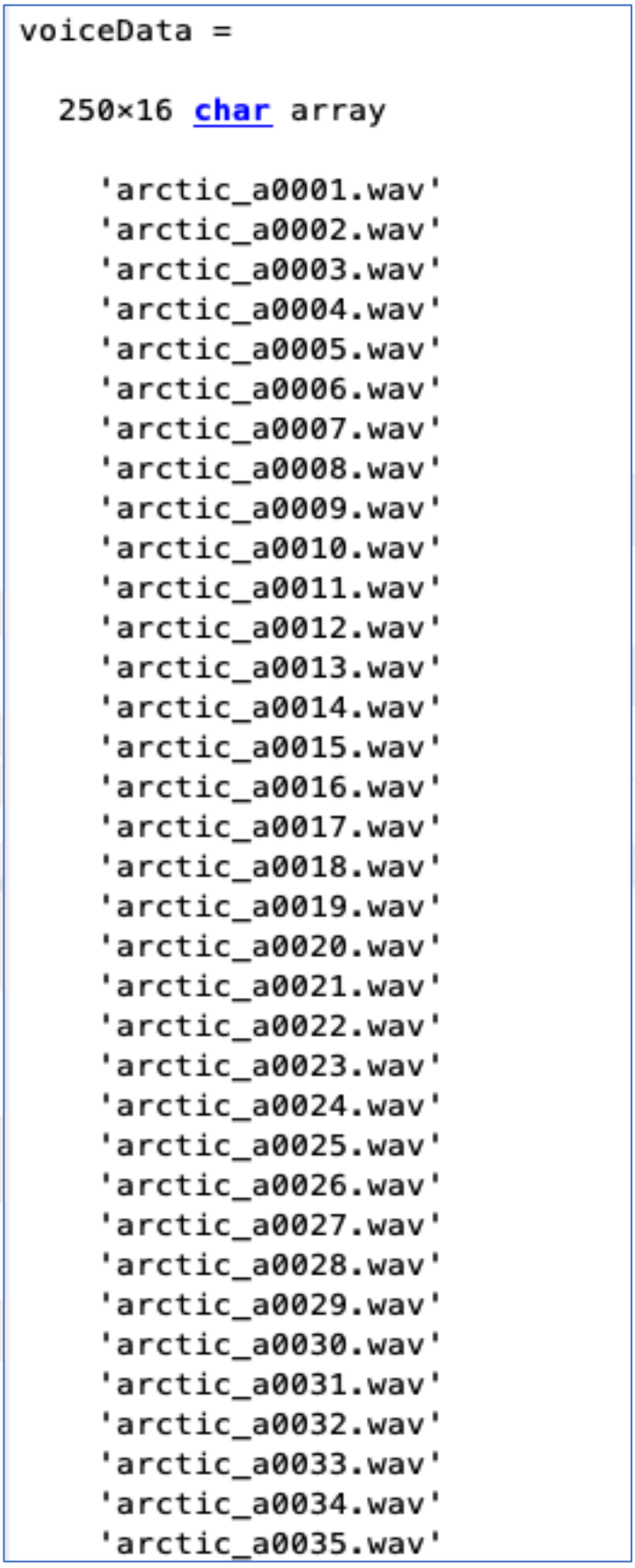


Figure 9: Dataset indexing/inventory array.

3.5.2 Features Extraction

Features as was discussed in preceding sections is termed to acoustic properties of the said speech imprint. However, speech signals will be analyzed acoustically under indexing loops for extraction of their acoustic features and then to formulate the features matrix. Speech features is extractable in both time domain and frequency domain. Time domain features more likely number of zero crossings, auto correlation and signal energy are useful in traditional speech processing [71].

Form the other hand, frequency domain will yield a lot about speech nature as signal power spectrum densities will be virtualized in accordance to the number of frequencies resides in the said speech. The overall matrix of features will be constructed to obtain the same features for each speech signal at the inventory matrix. Considering the nature of speech and the uncertainty of the environments where the test set and train set (inputs) are presents, at some points, time and frequency domain information will not stand to model the speaker [72].

Other methods are said consistence to tolerate the uncertainty of training and testing data. In other word, speaker identification system must be trained with particular data (speeches) and to be test with different set of data but at preserved nature. In order to understand the problem of uncertainty in speech identification and modelling, the following observation can be mentioned: [73-82]

1. Speech as it is; containing of data say (voltage) or power (analogue data) that varies versus with time so it is known as time variant signals which means that speech properties are keeping change with time. This is most reliable interpretation behind speech identification challenges.

2. The physical properties are not similar in both training and testing data: it can be interpreted further by descripting the problem with more details, so-to-say, recording a speech signal with same sampling frequency for same time will not produce similar features by any extent. It is actually noticed that voice signals will not preserve the same characteristics in terms of zero crossing and Fourier representation.

3. In speech corpora, training signals might feed to the model so that model will learn about the training data structure in accordance to the features matrix. In most of speech projects, testing data might be re-recorded and supplied to system. Recording the same voice by same speaker in different environments (background effects is different) may lead to major change in speech nature.

4. Speech signal may differ at the time of training and testing not only due to noise but because their major structural differences.

The general frame work of speaker modelling in speaker identification system can be demonstrated in below Figure.

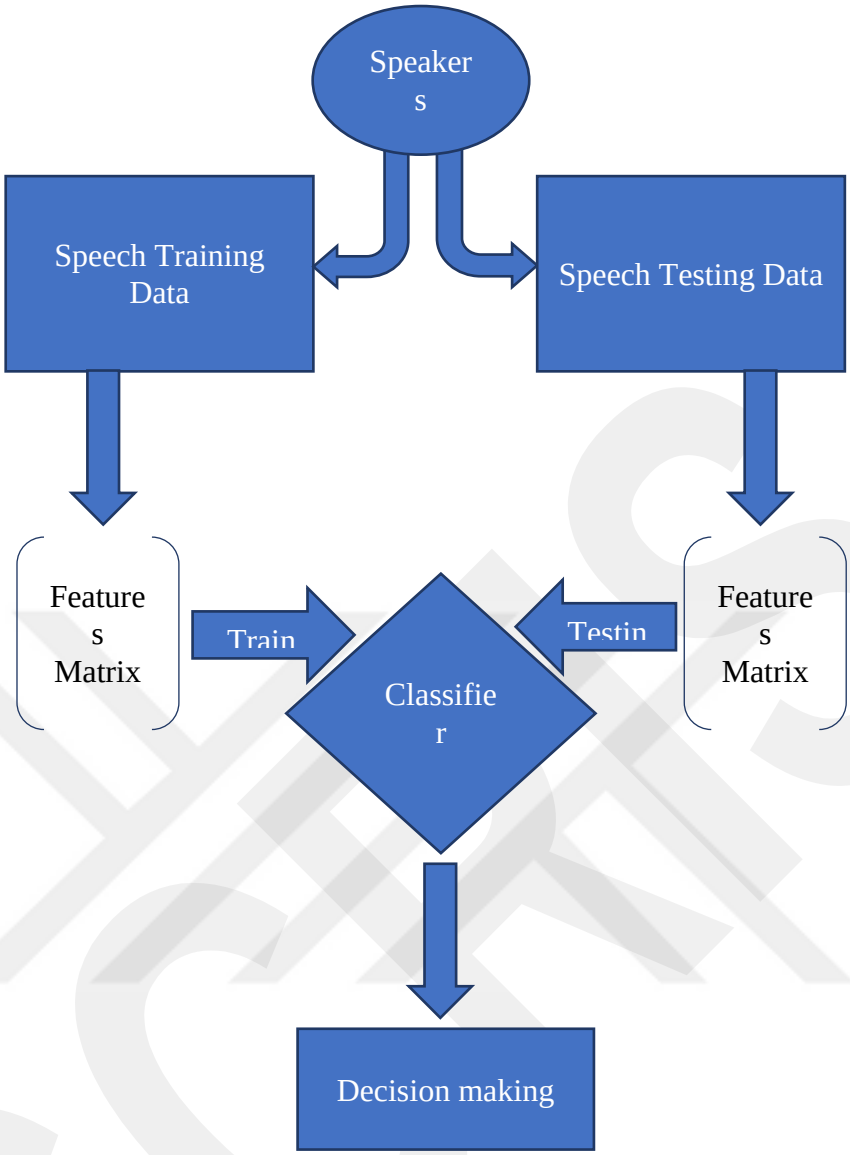


Figure 10: Speaker modelling prototype, depicts of training and testing models.

CHAPTER 4

EMPIRICAL MODEL

4.1 Overview

In this chapter, the speech practical model that implemented for the purpose of identifying speakers through-out learning their acoustic characteristics is discussed. In the early stages of this chapter, the features extraction paradigms will be described with more details. The following sections will release the methods (algorithms) that used to establish the so-called classifier. Speech classifiers are usually used when number of speakers in the dataset is high; so, machine learning based speech classification with light of features matrix is proposed.

The features extraction is made as combination between two main speaker modelling methods called as fundamental frequency coefficient and mel frequency cepstrum coefficients respectively. However, the merge between these two essential approaches in speech modelling is expected to plot extended accuracy in speaker identification.

The features which planned to be extracted from the both algorithms are brought together using a features array. The fundamental frequency method will produce single coefficient corresponding to the pitch period in the speech signal. From the other hand, mel frequency cepstrum coefficient is made to provide tewleve coefficients for each speech signal. From the above explanation. The array of features will be having thirteen coefficients which supposed to give as accurate as possible information of the speaker.

The corner stone of this study is using the machine learning to evaluate whither the both data in training model and testing model are relative. The process of machine learning is began with Random Forest Algorithm through deployment of other more advanced algorithms such as Neural Networks. Optimization algorithm such as Particle Swarm Optimization PSO and Optimized Particle Swarm Optimization OPSO are used to enhance the performance (accuracy) of speaker identification in neural network.

The second part is to propose a model that can solve cocktail party effect. The proposed model is utilized deep learning that have ability to recognize each person separately. combining Fully Convolutional Network (FCN) and a Bidirectional Long Short-Term Memory (BLSTM) for source separation. The FCN utilizes a convolutional neural network (CNN) to convert image pixels to pixel classes. In contrast to the CNN, an FCN converts the width and height of the intermediate layer feature map returning to the input image size throughout the transposed convolution layer, to make sure that the predictions include a one-to-one correspondence for input image. BLSTM is an (LSTM) recurrent NN that utilizes contextual info from past and future from the input/output sequences. In which the hidden layers are BLSTM layers and LSTM is the output layer. The FCN-BLSTM network is able to captures the characteristics of spectro-temporal of the audio data much better than single model (FCN or BLSTM). In this approach the FCN is applied first to acquire an initial estimation of the magnitude spectrogram of the specific source coming from the input sequence. Then the initial estimation is passed to BLSTM network to improve the output sequence of the FCN.

4.2 Mel Frequency Cepstrum Coefficients (MFCCs)

In speech recognition applications, the term mel scale is very popular due to its impact to formulate a lot of acoustical characteristics for the speech signal. In mel frequency cepstrum coefficient algorithm, speech signal is set to be undergone various of processing in order to obtain the features vector. It was denoted that features vector in this study contained of eight characteristics from mel frequency cepstrum coefficients and one characteristic from the fundamental frequency [81] [88] [90].

The stages of processing in this algorithm are began with pre-emphasis which take the speech signal and amplify the value (power) of some segments of this speech. It is realized that higher frequency components in the speech signal are seen with lower power. however, the pre-emphasis process is made using special filter called as pre-emphasis filter and aimed to amplify the power of the high frequency regions in speech signal.

The advantages of using the pre-emphasis prior to further process in mel frequency cepstrum coefficients can be made as following:

1. Prevention of computational problems in the further steps of mel frequency cepstrum coefficients algorithm especially those come with Fourier Transform.
2. Derivation of power (amplitude) balanced distribution amongst all frequency components on the speech signal.
3. Improvement of signal to noise ratio since the low power in the signal will be amplified, the signal power will be compensated against noise power.

Implementation of pre-emphasis filter in the empirical model is about producing a new speech signal with same characteristics of the previous one (before applying the pre-emphasis filter), the new version of this signal will have amplified power in high

frequency region. The overall power distribution in pre-emphasis signal is relatively even between the high frequency regions and low frequency regions.

Let the signal $S(t)$ to be the time domain sampled speech signal in the input gate on pre-emphasis filter, the output of this $Z(t)$ filter will be given in Equation 4.1.

$$Z(t) = S(t) - m.S(t-1) \quad (4.1)$$

Where m is the pre-emphasis filter coefficient which is ordinary equal to 0.95 or 0.97, the virtual investigations between the input signal and output signal to the pre-emphasis filter is seen in Figures 11 and 12 respectively.

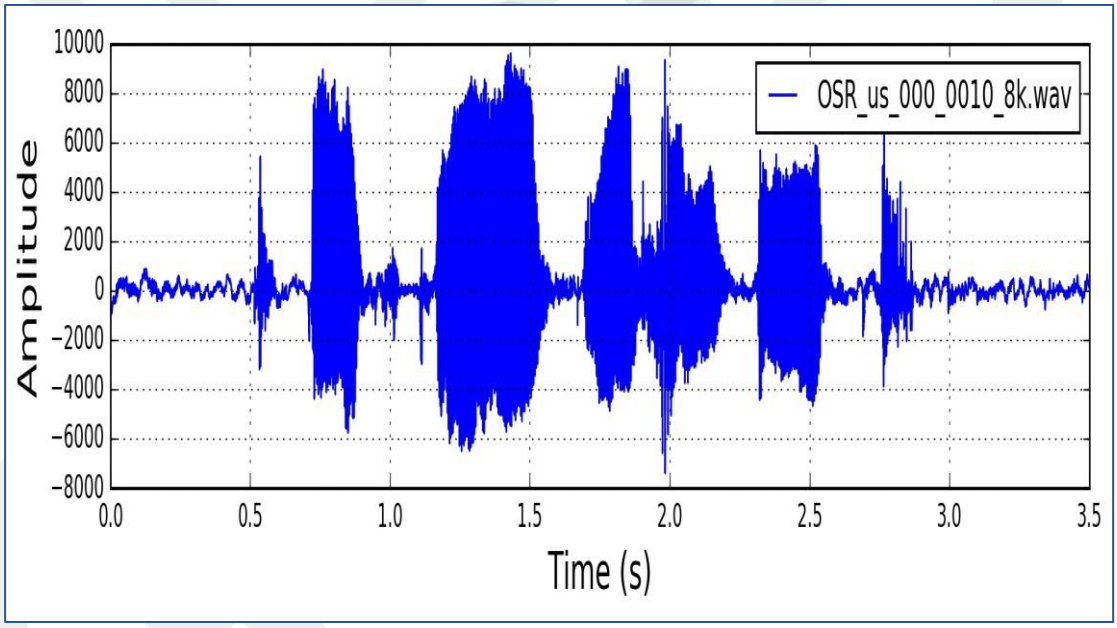


Figure 11: The input time domain speech signal into pre-emphasis filter.

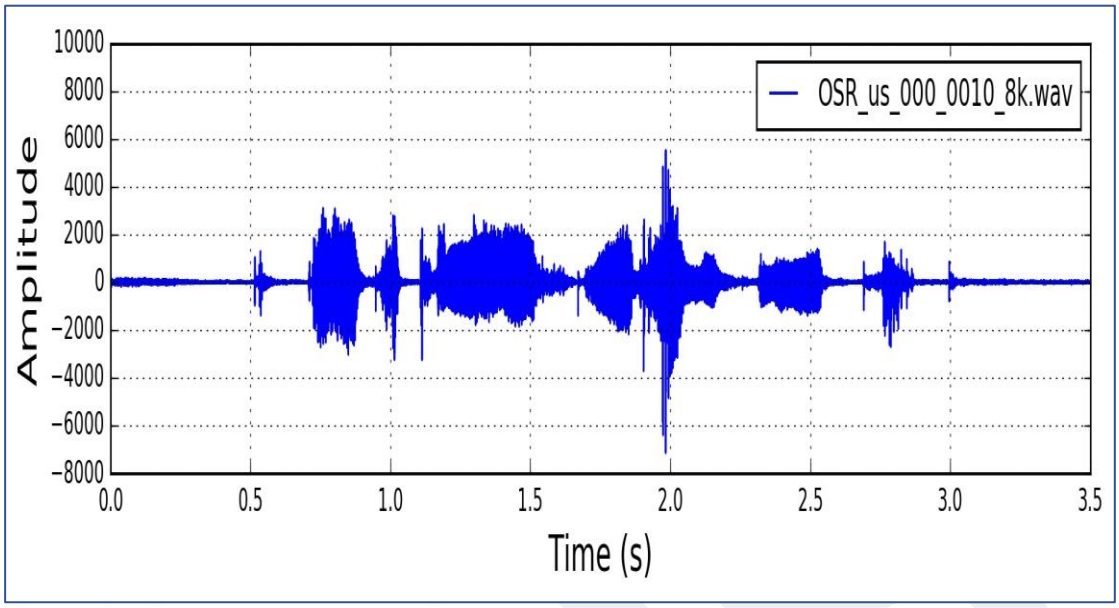


Figure 12: The output time domain speech signal into pre-emphasis filter.

The coming stage of mel frequency cepstrum coefficient method is to forward the pre-emphasis filter output into splitter where full frames of signal is segregated into small frames. Since the speech signal frequency is changing be the time, the frequency virtualization is not worthy or spectrum of full speech signal cannot be relied in the further process as frequency keep changing with time.

In order to tackle this problem, an assumption is made that speech signal remains as time invariant for very short time duration fall in the range of twenty-three to twenty-five milliseconds. So, the signal will be framed into small periods frames in twenty-five milliseconds. In order to achieve more accurate approximation, overlapping frames are chosen with overlap of ten milliseconds. Figure 13 demonstrates the process of framing.

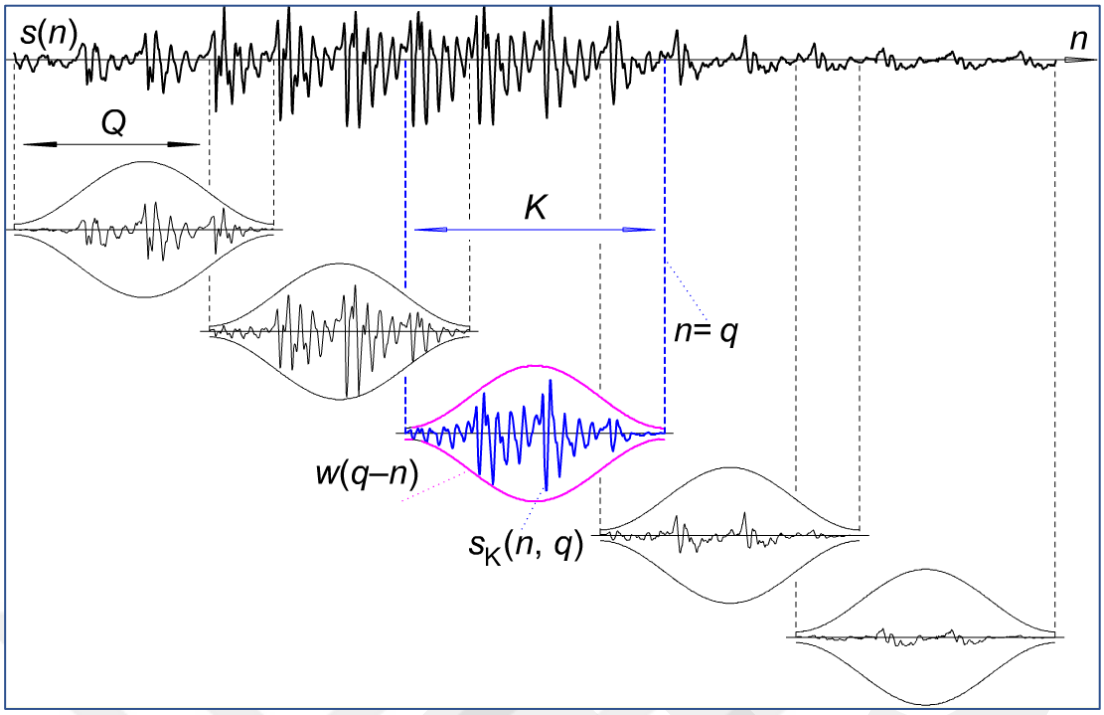


Figure 13: Overlapping framing of 25 milliseconds in speech signal.

Prior to compute the power spectrum density (Fast Fourier Transform), each available frame will be windowing by surrounding it by hamming window. By other means, frames period is been identified with respect to the time (duration of each frame) and the overlapping factor of ten milliseconds. However, in order to extract those frame in more effective way, hamming window is used instead of rectangular window. The hamming window is illustrated in Figure 4.4 according to the hamming function in Equation 4.2, noting that windowing process to be applied in sampled version of the speech signal unlikely the pre-emphasis filtering.

$$F[n] = -0.46\cos\left(\frac{2.pi.n}{N-1}\right) + 0.54 \qquad (4.2)$$

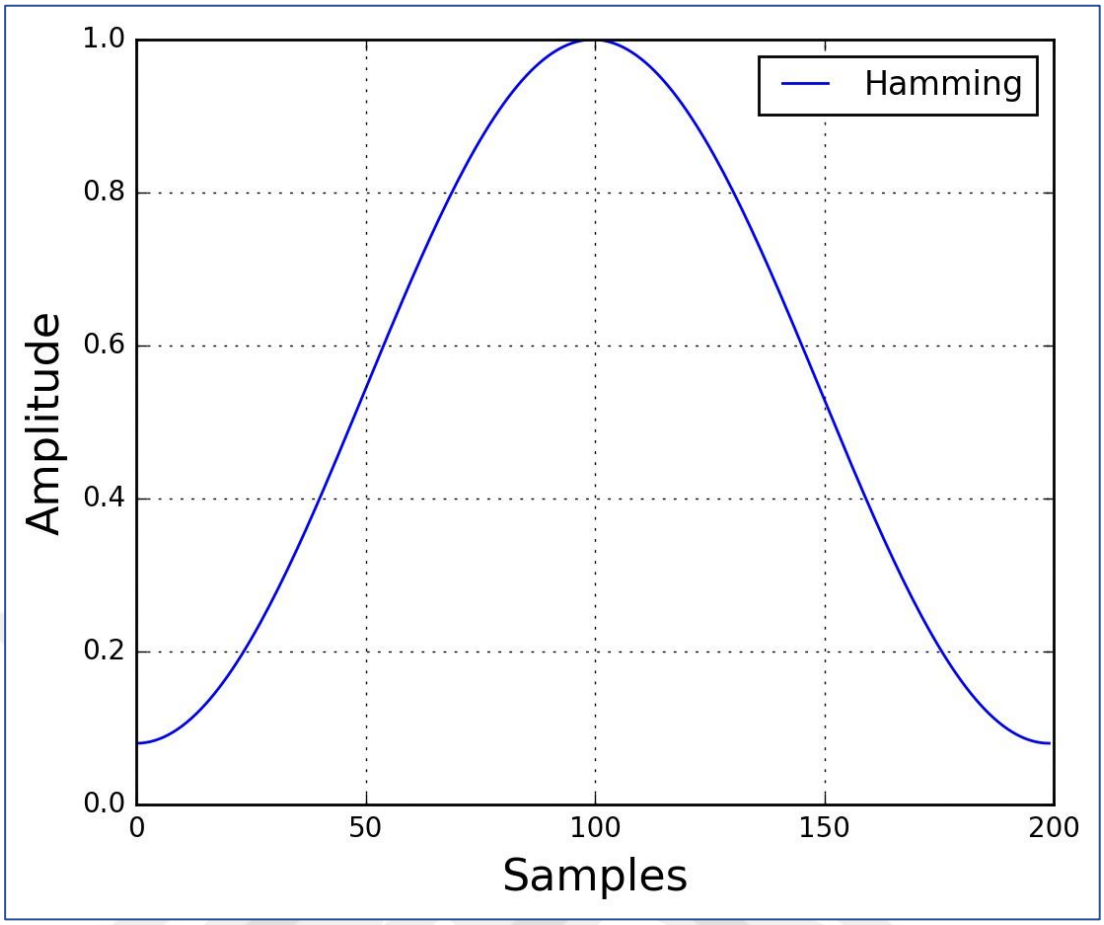


Figure 14: The Hamming window outlook.

Comparing the window of hamming shape given in Equation 4.2 with rectangular window as in Figure 4.5, the reason of selecting a hamming window over rectangular window is to preserve minimum spectrum leakage. From the other hand, the sharp window is more susceptible to noise as compare to hamming window [85].

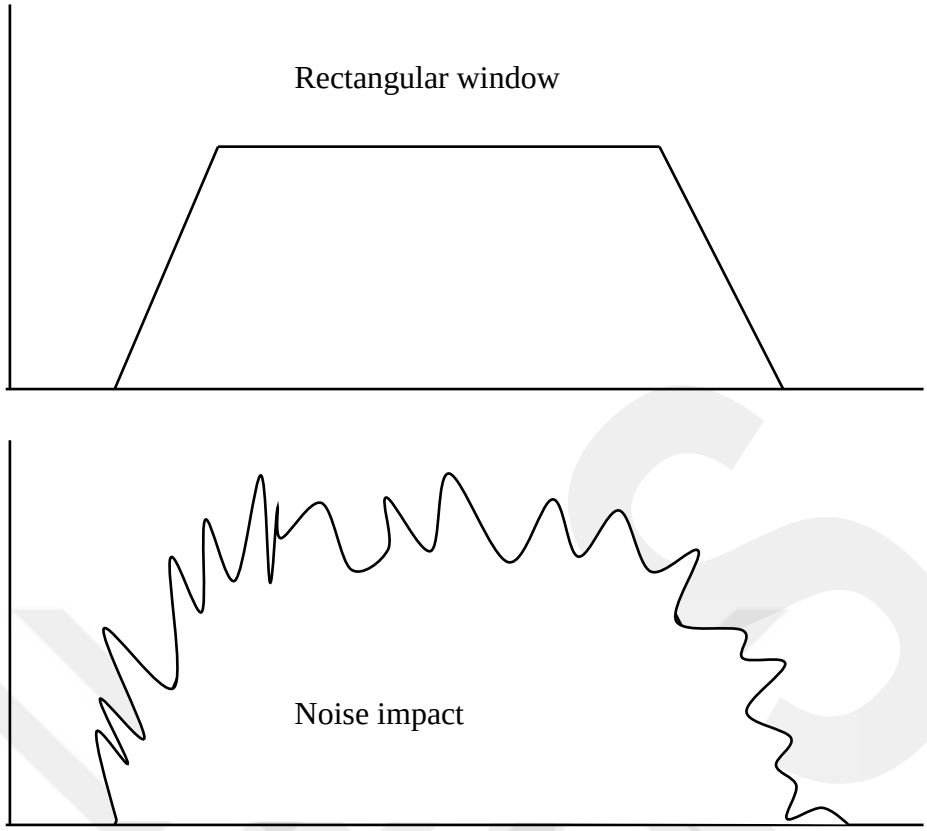


Figure 15: Rectangular window vs. noise effects.

At the end of this stage signal will be ready for spectrum analysis using the Fast Fourier Transform.

As in short time Fast Fourier Transform, N points are to be identified in order to compute the speech power spectrum. This stage of signal processing is aimed to compute the power spectrum in the speech signal using N-points short time Fast Fourier Transform [86].

The number of points that are usually selected in this analysis is two-hundred and fifty-six or five hundred and twelve points. The more accurate analysis is achievable using five-hundred and twelve-point short time Fast Fourier Transform. The power spectrum can be calculated using the Equation 4.3.

$$PS = \frac{|FFT(F_n)|^2}{N} \quad (4.3)$$

Up to here, signal frames are analyzed in spectrum domain and power Periodogram is obtained.

The further step is applying Filter-Banks in order to simulate the actual perception of human ear. In order to simulate the human voice sense perception, two factors are needed to be understood:

1. Human perception is more sensible to lower frequency within voice signal than it is in higher once. So, model should act as discriminative in response to the low and high frequencies.

2. Human perception to the voice (sound/audio) can be measured in mel scale where the same is achievable using the Equation of 4.4.

$$f_m = 2595 \log(\frac{f_n}{700} + 1) \quad (4.4)$$

The inverse of the Equation 4.4 will yield the original frequency of the speech signal. So, the same is given in Equation 4.5.

$$f_n = 700\ (-1 + 10^{\frac{m}{2529}}) \quad (4.5)$$

However, in order to simulate the human perception to particular voice signal, the first step is converting the frequencies into mel scale using the Equation 4.4. As signal is converted to the mel scale, triangular filter is applied on the signal in order to discriminate the frequency fluctuation which reflects the perception of human ear. The triangular filter will be designed to have five responses according the mel frequency which are given in Equation 4.6.

The different five transfer functions cannot be achieved using a single filter, for this reason, five filters with different frequency responses are form what is so called as filter bank.

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases}$$

The filter banks transfer functions can be plotted in the Figure 16.

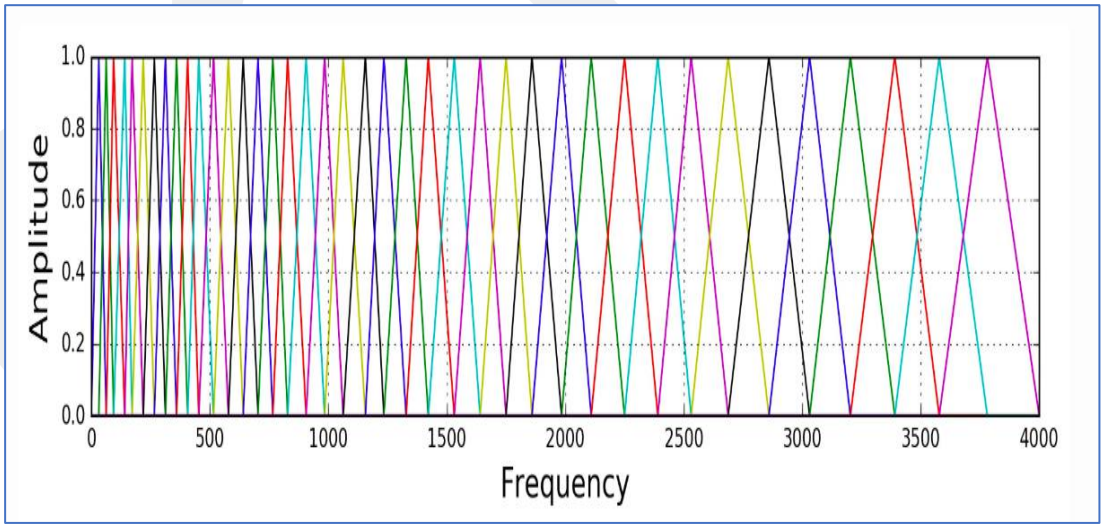


Figure 16: Filter bank responses to different frequency (mel) scales.

The signal at this stage of analysis can be plotted using the spectrogram which can yield the time-frequency representation of the speech signal [87].

The following notations can be realized from the so-called spectrogram after the speech signal passed from the triangular filter banks.

Speech signal as it cross the stage of filter banks, the resultant of this stage is human ear perception to the speech signal at different frequencies. Now, in order to explore more and more information in the speech signal, time domain information is also needed [88].

This can be interpreted as following: since speech signals is suffering from continuous changes in their frequency component over the time (as time expanded), the solely frequency representation of the signal is not sufficient. Frequency-Time representation of the speech signal can be produced using the spectrogram [90]. For the same signal, the spectrogram is illustrated in Figure 17.

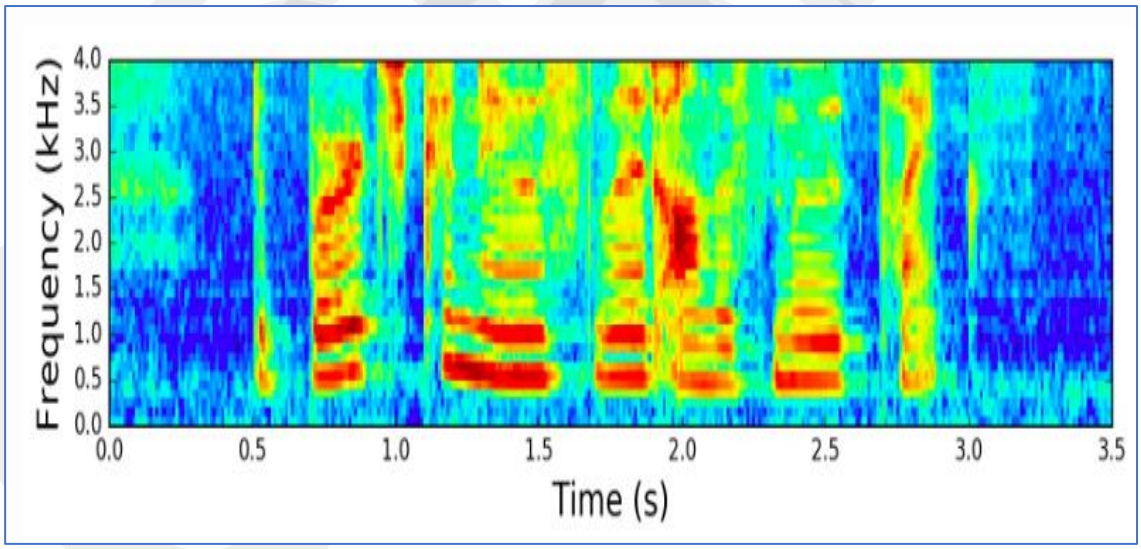


Figure 17: The spectrogram of the voice signal.

Looking at this illustration of Frequency-Time of speech signal, the x-axis demonstrates the time while y-axis demonstrate the frequency. The very high frequency regions can be seen from the Figure as dark yellow reaching to red color. Spectrogram shoes that in each time period how the frequency is changed [91].

Two last stages are remained to finalize the mel frequency spectrum coefficients method: the second last stage is applying the discreet cosine transfer into the resultant signal of the filter bank.

The purpose of applying the discrete cosine transform on the filter banks output is reduce the coefficients resulted in the filtering process (compression of the coefficients). The compression of the signal coefficients is performed to reduce the said coefficients number into only thirteen coefficients. The previous experiments revealed that the coefficients from thirteen onward are not related to the MFCC process so it can be neglected [92].

The first last stage is liftering which is another kind of filtering but in the cepstrum domain. This is made as an attempt to enhance the signal to noise ratio as a top of signal pre-emphasizing filter.

The final result of this method which corresponds to the mel frequency cepstrum coefficients of the speech signal is illustrated in Figure 18.

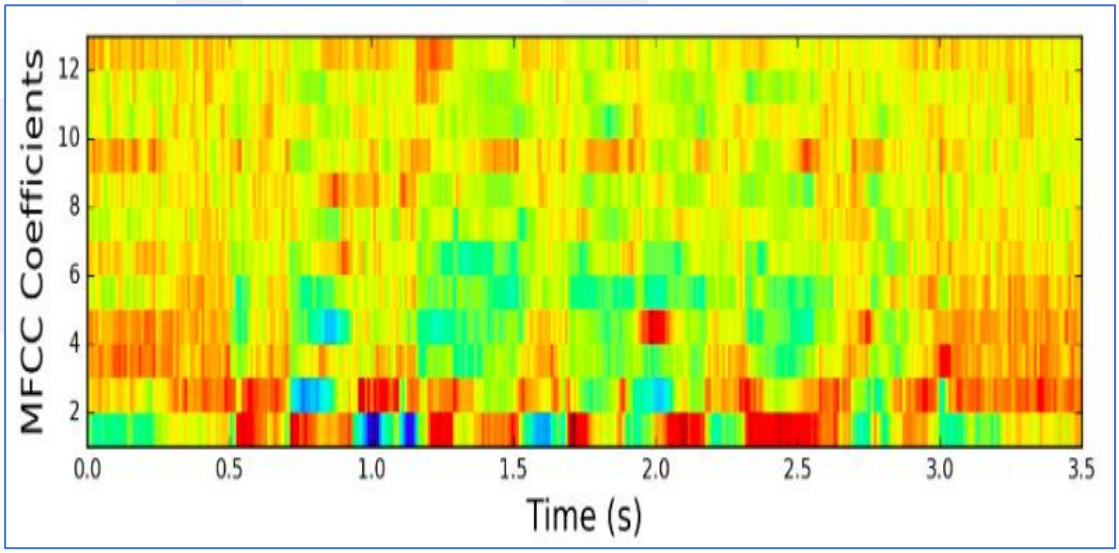


Figure 18: Mel Frequency Cepstrum Coefficients for speech signal.

The outcomes of this methods can be discussed in the following points: at the end of discreet cosine transfer, the speech signal which results from triangular filter banks is

compressed for reducing the number of coefficients (selecting only those which related and served in MFCC). Eventually, since the coefficients are reduced to twelve, the mel frequency spectrum coefficients will be twelves that corresponds to the input speech signal [93].

Now for speech analysis purpose, the filter banks output is reminded enough to understand the acoustic characteristics of the speech signal. From the other hand, mel frequency cepstrum coefficient is the further process after the filter banks which formulate the coefficients of speech signal where they can be used to recognize the speaker in machine learning algorithms.

The entire method of speaker recognition using the MFCC method can be graphically demonstrated in the Figure 19.

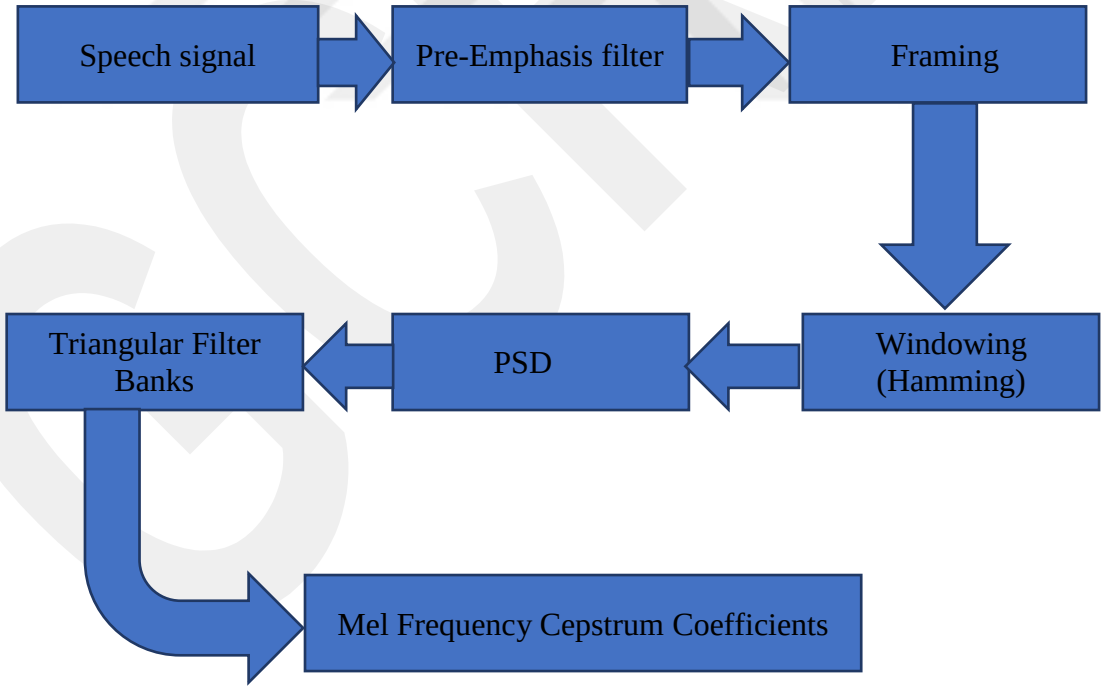


Figure 19: Mel Frequency Cepstrum Coefficients work flow.

4.3 Fundamental Frequency Feature

The acoustic information which acts as real identifiers of the speech signals are obtained through the so-called Acoustic model. The dataset contents are feed to this model for determination of their pitch period and hence to calculate the Mel frequency of each utterance. In order to do so; each signal is cross correlated with its same copy for defining the top maximum peaks in their correlation resulted signal. The cross-correlation can be evaluating as per the following derivation [94].

Let $x[n]$ is the input speech signal, and let $x'[n]$ is the same copy of the signal. Hence, m, is the number of the samples in the signal. The cross correlation can be derived using the Equation 4.7.

$$Cor = \sum_{n=1}^{n=m} x[n] \,.\, x'[n] \tag{4.7}$$

The term Cor is shown in the Figure 20. Here, it is noteworthy to mention that the resultant signal (cross-correlation result) is dominant a double length of the original signal.

$$L_{cor} = 2.L_x \tag{4.8}$$

The main reason of calculation the cross-correlation is to evaluate the so-called local maxima, in other word, the peak amplitude if the cross correlated signal as Figure 20 and Figure 21 depict is located at the center of the plot and represents the maximum similarity of the signals participated in correlation function Equation . The local first maxima is depicted in the Figure 22. (Red highlighted).

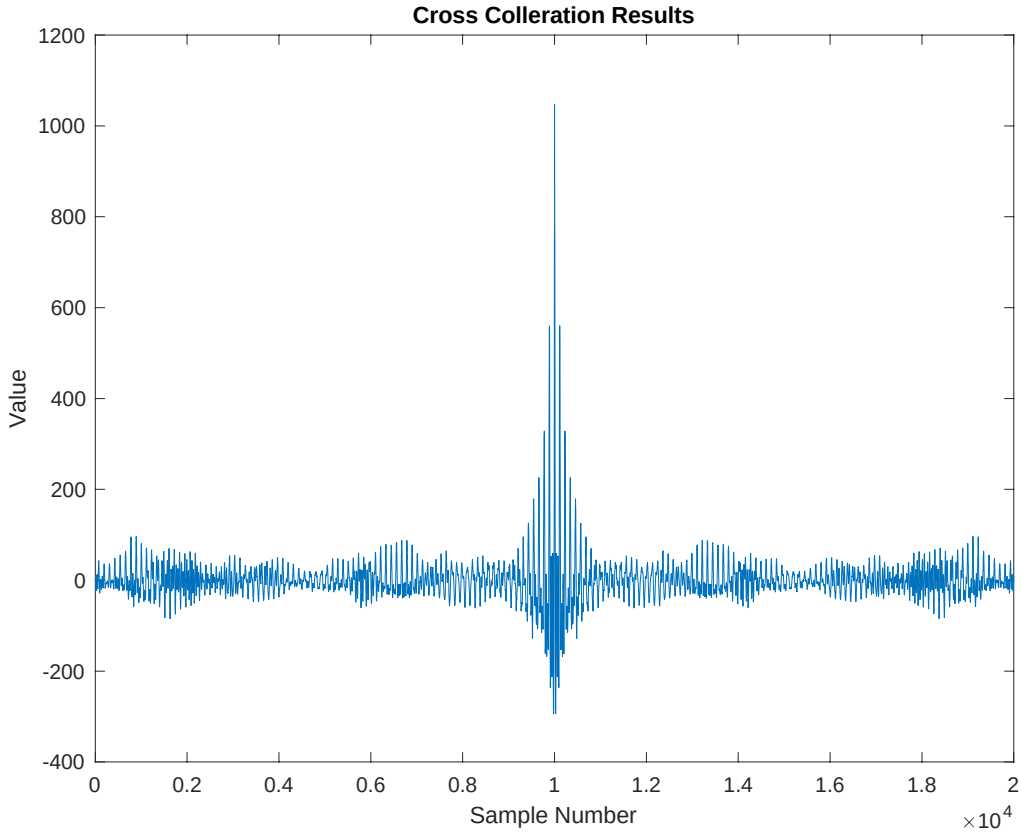


Figure 20: The cross-correlation resulted signal.

The process of evaluation the fundamental frequency can be given as:

Let $Cor\,[fm]$ is the cross-correlation function at the first local maxima (first peak) and $Cor\,[sm]$ is the cross-correlation function in the second local maxima (second peak). Then pitch period can be calculated as per Equation 4.9 and Equation 4.10, the same is depicted in Figure 22.

$$P_{lagging} = |fm - sm| \tag{4.9}$$

$$P_{leading} = |sm - fm| \tag{4.10}$$

$$F_f = \frac{P^{-1}}{F_s} \tag{4.11}$$

Where, F_f is the fundamental frequency and F_s is the sampling frequency.

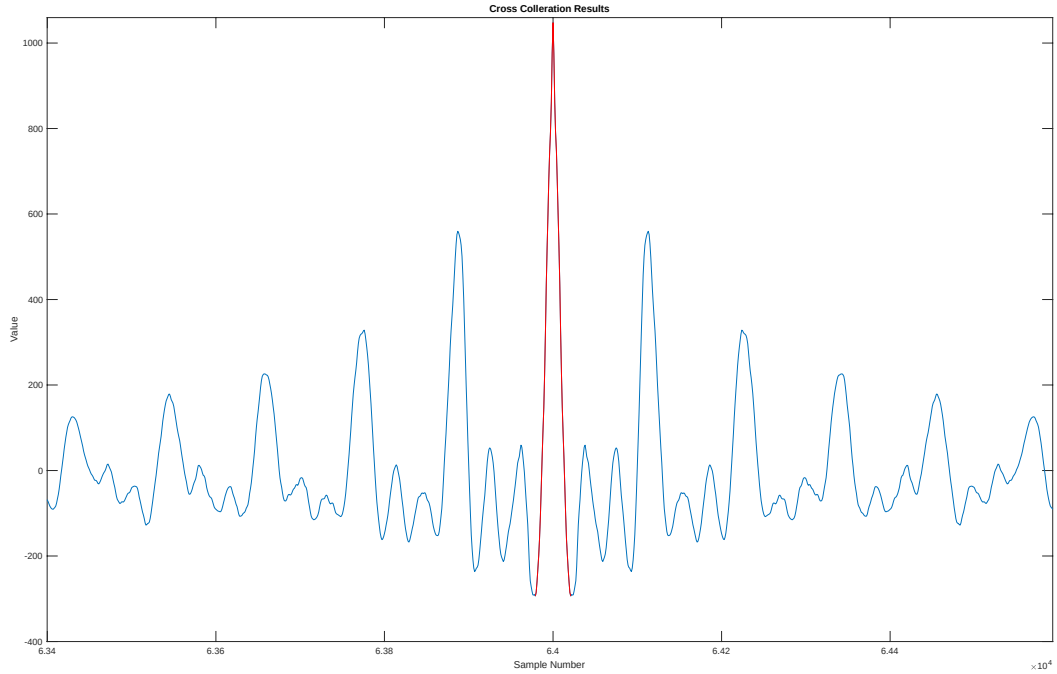


Figure 21: Local maxima peak of the cross-correlation function.

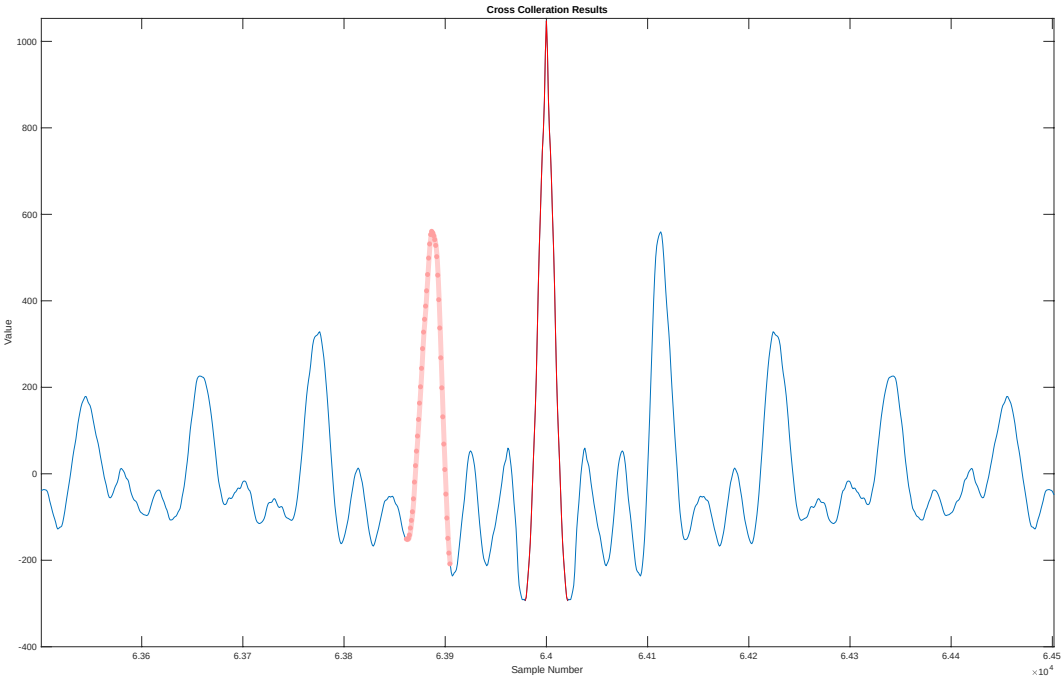


Figure 22: A depict of local maxima is cross-correlation function.

4.4 Features Matrix

This study involves designing of speaker identification system to identify speakers using their voice imprints. However, the system is text dependent where it concerns about same imprint spoken at the time of training should be provided at the time of testing [95].

Dataset of speaker with two-hundred and fifty speech signals is provided to the system. With help of indexing loops, system is able to produce the acoustic characteristics of each speaker. It is however required to list all the acoustic information in a single array correspond to single speaker characteristics.

Those arrays are merged together to form what so-called features matrix which will be used in machine learning algorithms to classify the speakers. This matrix consists of N rows where N corresponds to the number of speakers participated in the system and K columns where K corresponds to the number of acoustic information.

The noteworthy point in the number of columns is involved the mel frequency cepstrum coefficients which are twelve elements vector (array) and fundamental frequency coefficient which is single element array.

Total thirteen element will be there as matrix columns, the number of column element will be same for every speaker so that the matrix will contain of two-hundred and fifty rows by thirteen columns (250 X 13). The total number of cells in the features matrix will be three-thousand and two-hundred and fifty cells corresponding to the total number of features for two-fifty speakers in the system.

Figure 23 demonstrates the process of constructing the so-called features matrix by referring the information from features extraction algorithms for each speaker.

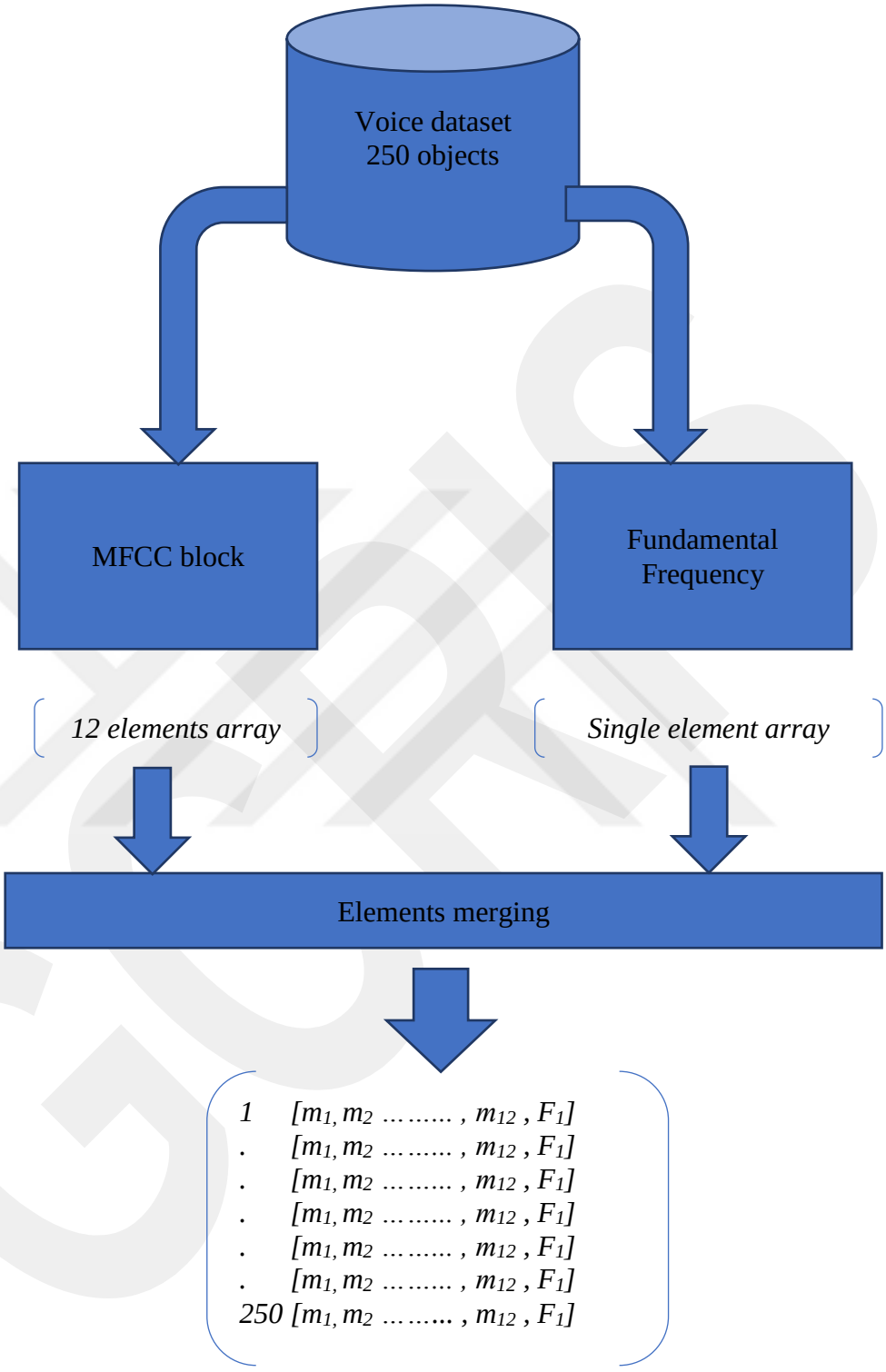


Figure 23: Features matrix establishment process.

CHAPTER 5

SPEECH CLASSIFICATION

5.1 Outline

The said speech signals are produced from human vocal track as vocal cords vibrates inside the vocal track. In the context of speaker recognition, speaker model is trained on particular voice imprint in order to be able to simulate the vocal track process. It was prescribed earlier in this thesis that simulating the vocal track process is not an easy task due to uncertainty and random variables. As speech signal is time variant, the frequency components of the said speech is on continues changing as time expands.

Mel frequency cepstrum coefficients model is used to simulate the human ear perception of voice which indicates how human ear will respond to particular voice signal. It is well known that auditable (hearable) frequencies is falling in range of eight kilohertz, so in this range human ear is more responding to the lower frequencies i.e.> 8000 Hz and less responding to frequencies higher of equal to 8000 Hz.

Mel frequency cepstrum coefficients model is outperformed in modulating the human ear process in mel scale so that speech signals as they treated by the Mel frequency cepstrum coefficients model, they can be identified by a numerical vector of twelve called as features array and corresponds to the coefficients of filter banks in MFCC algorithm [97] [97] [98] .

Furthermore, one speech characteristic resulted from the time domain analysis of speech signal (using the cross-correlation) called as fundamental frequency (pitch

period) is also considered as one of vital features to simulate the voice imprints. This feature will help to verify the speech information in both training and testing. In other word, fundamental frequency is single numerical identifier for the speech imprints similarity.

Speech classification will be discussed in this chapter, as signals are processed in the aforementioned methods and features matrix is established, speech features classification is pledged by machine learning algorithms. Three machine learning setups are made more likely: Random Forest, Feed Forward Neural network and Particle Swarm Optimization combination with Feed Forward Neural Network. The results from each classifier is monitored with performance metrics identifiers such as Recognition Accuracy, Training Mean Square Error, Training Root Mean Square Error, Approximation Time.

5.2 Classifiers

As the acoustic information gained from the aforementioned procedure; it is now necessary to map each fundamental frequency to its own speaker if a set of speakers willing to get identified (in testing part). Hence, the state of the art in this work is using more than one data classification algorithm to implement these requirements. Prior to get in deep with the said classifier, the data is divided in two parts as 70 percent of data is for training of the classifier and rest 30 percent of the it is for testing the performance of the trained classifier. The classification is examined using Random Forest Algorithm and Feed Forward neural Network. In each classifier, performance metrics are evaluated, so-to-say, the following terms are being calculated: Mean Square Error (MSE), Root Mean Square Error (RMSE), Recognition Accuracy and Time taken by classifier to perform the approximation.

So, those classifiers are made to learn from the acoustic information and predict the speaker identity [99] [100].

5.2.1 Random Forest

It is classification algorithm works to segregate the data into their classes base of the target information. In our case, the target is prepared to give the serial number of the speech signal as per the dataset index. The random forest is established firstly by feeding the dataset (acoustic information) along with the target into the algorithm workplace. The process of Random Forest Algorithm can be illustrated in Figure 24.

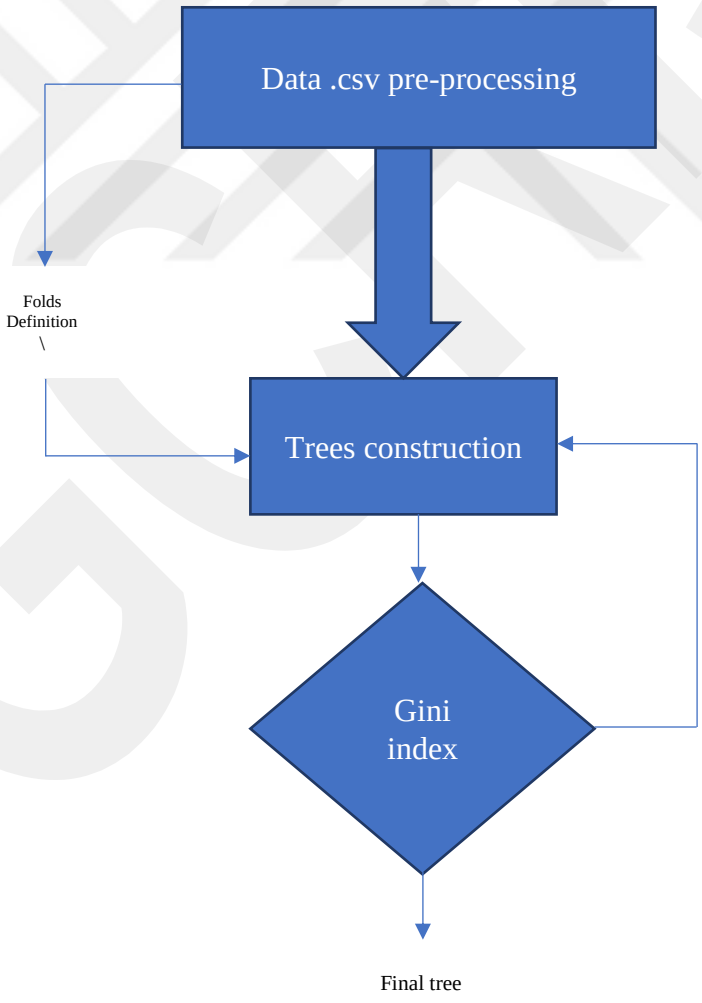


Figure 24: Working mechanism of Random-Forest model.

The data is feed as coma separated values which is needed to be converted into lists of float values. As soon as the data is prepared, it is splitted into number of folds to construct the trees. In our prototype, optimum number of folds was found equal to FIVE folds [101]. As per the Random Forest rules, data is segregated into number of trees where each tree must represent the data related to same class. In other, word, tree wise data need to be of the same nature and to measure the similarity index. The frame work example of Random Forest Algorithm is demonstration Figure 25.

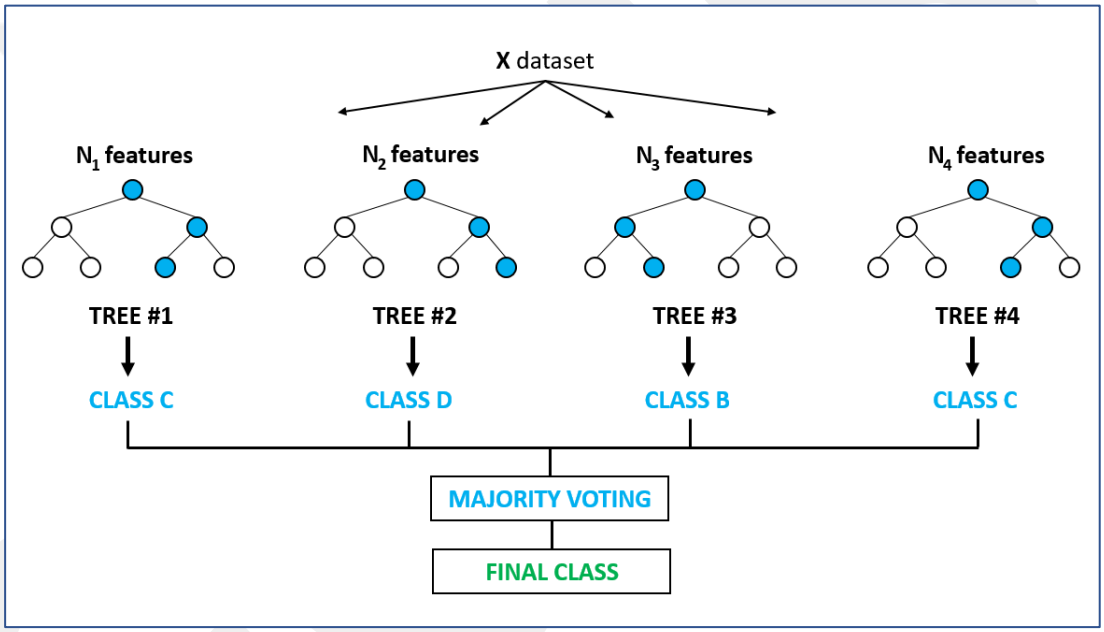


Figure 25: Example of Random Forest Algorithm Frame Work.

After applying the features matrix to the Random Forest algorithm, the performance metrics are monitored as given in the Table 1. The classification of speech signals in Random Forest Algorithm yielded a ten percent of accuracy which is very far from the required accuracy. Table 1 is involved the other performance metrics such as Root Mean Square Error which is reflecting the amount of average error exists while the training of classifier.

Table 1: Random Forest Classifier performance results.

Metric	Value
Accuracy (percentage)	10
Time	6.17362
Mean Square Error	68.9
Root Mean Square Error	8.3006

5.2.2 Feed Forward Neural Network

FFNN is popular type of neural network able to learn a complex problem of real-life applications. This kind of neural network is outperformed in learning the problems that independent of time unlikely the Current Neural Network that typically made to learn out timely problems. It is obvious, the speaker recognition process of this project and majority of similar projects are producing a time independent data so that, best type of neural network with minimal computational cost is chosen to be Feed Forward Neural network [102] [103]. The model parameters are tabulated as following:

Table 2: FFNN model parameters.

Particle	Details
Number of Layers	3
Nodes (Layer Wise)	30, 10, 1
Learning method	LM
Minimum Gradience	1e-29
Iterations	100

The model is fed with data for a very first experiment and hence all the said performance metrics are obtained. The experiment is repeated for100 times due to the random nature of weigh/bias assignment. The results of plain Feed Forward Neural Network are given n the following Table.

Table 3: Feed Forward Neural Network Classifier performance results.

Metric	Value
Accuracy (percentage)	91.493
Time	1.3248
Mean Square Error	1.23E-27
Root Mean Square Error	3.5071e-14

5.2.3 Model Freezing (MFFNN)

As demonstrated in Figure 26, neural network classifier is constructed from three functional layers namely: input layer, hidden layer and output layer. Nodes are residing in each layer and connected to each other by means of neurons. The neurons

are connecting the nodes at each layer with the other layers nodes in a way similar to the brain neurons.

Two paramount elements are considered while dealing with neural network classifiers namely as number of hidden layers, mathematical representation of neurons connecting the layers. The classical structure of neural network layers are one hidden layer and two layer for the input and output respectively.

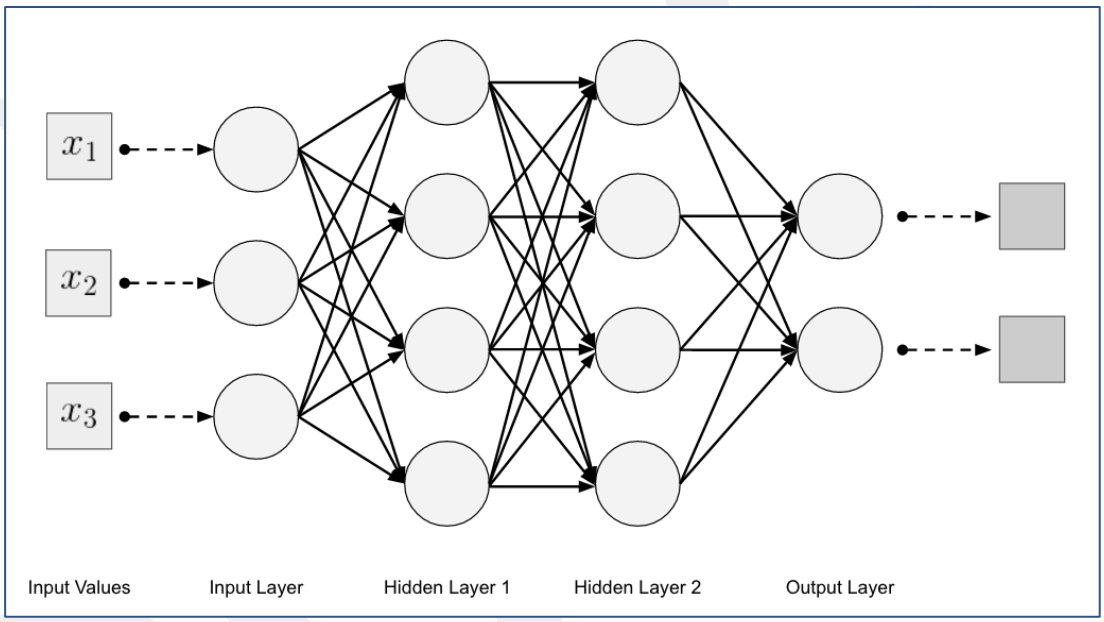


Figure 26: Structural diagram of the neural network classifier.

The Figure above shows a four-layer neural network classifier which has two hidden layers instead on one hidden layer as stated above. The most vital procedure to ensure the accuracy in the neural network classifier is well tuning of the neurons in the network.

Neurons of classifier are represented mathematically by particular values (numerical values) that used as scales or biases to approximate the input data and map them to appropriate output (target). These neurons are called as weights and tuning of neural

network is about selecting the perfect value of this weight so that input data is approximated and mapped accurately to a appropriate output.

However, weights values are randomly allotted by the learning algorithm inside corpus of neural network classifier. Levenberg-Marquardt algorithm is used inside the neural network classifier as an optimization algorithm that provides a random value to the weights and attempts to reduce the training error. In order to understand the random nature of training in Feed Forward Neural Network model, the Table 5.2model is implemented and experiment is repeated for one-hundred repetitions, in each reparation, performance metrics are monitored more likely, Mean Square Error, Root Mean Square Error, Time and Accuracy.

Freezing the model is about selecting the wright vector that produce the minimum error in the training and yields the maximum accuracy or recognition.

Table 4 yields the results of classification the speech signal in Model freezing Feed Forward Neural Network. The accuracy obtained from the Feed Forward Neural Network is approximately ninety-one percent. So-to-say, Feed Forward Neural Network classifier is outperformed in classifying the acoustic data.

The training quality of Feed Forward Neural Network is seen better that in Random Forest. The quality of training is monitored using the Mean Square Error and Root Mean Square Error.

Table 4: Modified Feed Forward Neural Network Classifier performance results.

Metric	Value
Accuracy (percentage)	94.6667
Time	0.9831
Mean Square Error	4.03E-28
Root Mean Square Error	2.01E-14

5.2.4 Particle Swarm Optimization

Particle Swarm Optimization is used here to search the optimum weight vector and apply it to the Feed Forward Neural Network so that accuracy of classification can be enhanced. The Particle Swarm Optimization can be illustrated in Figure below [104] [105]. The algorithm is producing population which equivalent to number of weight and all the produced populations are tested with fitness function to evaluate the error. The population with minimum error will consider as model weight.

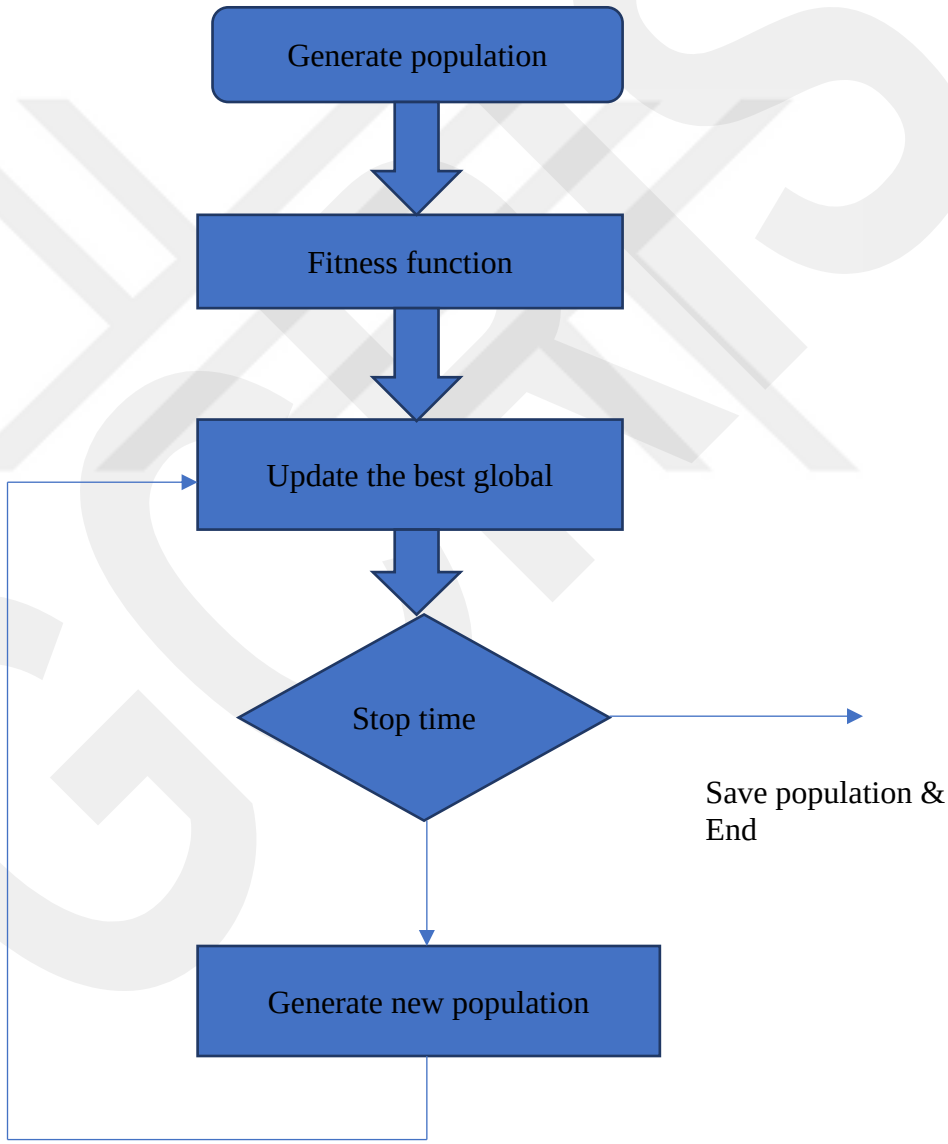


Figure 27: PSO optimization algorithm frame work.

Perfect tuning of neural network model is obtained after optimizing the weight using the Particle Swarm Optimization. The results of performance metrics are tubulated in Table 5. it is realized that speaker identification accuracy of the Feed Forward Neural Network combined by Particle Swarm Optimization has yielded an accuracy equal to ninety-six percent.

Table 5: PSO combination with Feed Forward Neural Network Classifier performance results.

Metric	Value
Accuracy (percentage)	96
Time	1.4769
Mean Square Error	2.70E-28
Root Mean Square Error	1.64E-14

5.3 Cocktail Party Effect

5.3.1 Outline

The Cocktail Party effect is the capability human to focus on voice of his attention in a noisy environment and neglect the rest of them [1]. the equivalent task is very difficult when using computers to do same task, mainly when using just single microphone for recording the mixed speech. Hence, many studies focused on solving this issue by utilized several techniques such as support vector machine, neural network, etc.

5.3.1 Proposed Model

There are several pre-process approaches can be used to achieved project goal, including modified hyperbolic tangent, STFT and iSTFT, complex mask, power-law

compression, Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), Signal to Artifact Ratio (SAR), Short Time Objective Intelligibility Measure (STOI) etc. The proposed methodology for audio data can be described in follows:

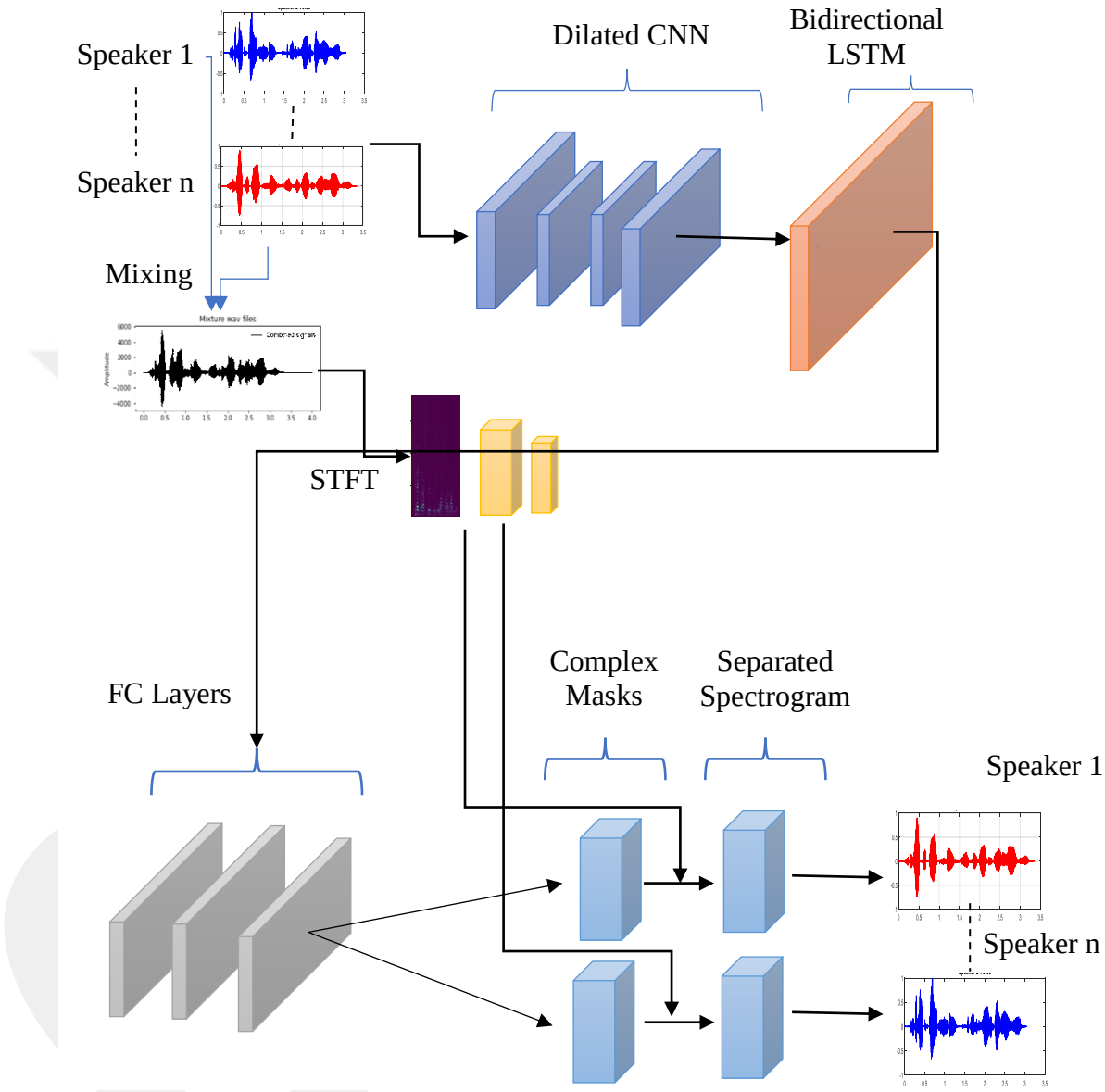


Figure 28. Proposed system structure

From figure 28, the proposed audio model gets each person voices. The voices signals are mixed then use Bidirectional Long Short-Term Memory (BLSTM) algorithm to generate spectrogram. In addition, each person spectrum image is trained using Dilated CNN. The next step is to use BLSTM and got FC layer then

generate complex mask then retrieve the spectrogram of each person voice and the result detect each person with his voice.

- **Input features:** The proposed model has involved auditory features as input. Considering an audio file contains several speakers, we need to calculate the short-time Fourier transform (STFT) of related seconds of audio segments. In this case each time-frequency consists of the imaginary and real parts of a complex number, each are used as input. However, we can utilize power-law compression in order to avoid loud audio from complicated soft audio. An identical processing will also be utilized for both the clean reference and noisy signals. At inference time, the proposed separation model could be applied to randomly long segments of audio. In cases where more than a single speaking is detected in a frame, the prosed model should be accepting multiple voice streams as input.
- **ANN model:** We have utilized deep learning network that is based on work done by [3], that combining Fully Convolutional Network (FCN) and a Bidirectional Long Short Term Memory (BLSTM) for source separation. The FCN utilizes a convolutional neural network (CNN) to change image pixels to pixel classes. In contrast to the CNN, an FCN converts the W and H of the intermediate layer feature map returning to the input image size throughout the transposed convolution layer, to make sure that the predictions include a one-to-one correspondence for input image. BLSTM is an (LSTM) recurrent NN that utilizes contextual info from past and future from the input/output sequences. In which the hidden layers are BLSTM layers and LSTM is the output layer. The FCN-BLSTM network is able to captures the characteristics of spectro-temporal of the

audio data much better than single model (FCN or BLSTM). In this approach the FCN is applied first to acquire an initial estimation of the magnitude spectrogram of the specific source coming from the input sequence. Then the initial estimation is passed to BLSTM network to improve the output sequence of the FCN.

- The audio stream initially calculates the STFT of the input audio signal to get a spectrogram, and after that learns an audio representation making use of a dilated convolutional neural network.
- A mixing of audio-visual representation is afterward generated by concatenating the learned audio and visual features and is consequently additional processed making use of a bidirectional LSTM and 3 fully connected layers.
- The outputs of network are a complex spectrogram mask for every person that is multiplied with the noisy input and transformed back to waveforms to acquire a separated speech signal for every person.

The data set steps, and parameter used in our model are described as in follows:

1- **Dataset:** In this work we have used "CSR-I (WSJ0)" [108], The dataset includes 990 voices as trained sets, 247 as development sets and 148 as evaluation. The all voices of datasets have 16kHz sampling rate.

2- **Preprocessing of input files:** This process converted the WSJ0 audio files that is in SPH format into wav format.

3- **Generate Mixed and Target Speech:** In this part we used two speakers voices and we mixed two speaker audios with and we made sample rate be 8000.

4- **Features Extraction:** STFT is used to extract features, then it converted to the tfrecords format that need to use as input for Tensorflow. The steps as in follows:

- Calculates the STFT frames taken from samples in time domain
- Reads the speakers wav file, then converts it to 32bit float values, followed by reshapes it depending on the number of channels.
- Compute the short time Fourier transformation (STFT) of a multi-channel and multi-speaker time signal. It can add more zeros for fadein and fadeout and need to generate an STFT signal that makes it possible for best reconstruction.
- Compute the inverse STFT to exactly reconstruct the time signal

5- **Training network**. In this part we are apply BLSM with RNN to multi-speaker speech separation. in which two speeches (two wave files) have been used as an input, and we use an Ideal Binary Mask to separate them again. The Neural Network is Trained using Keras. Every pickle file includes a dictionary with keys X, S, N, where every key stores a v list, in which columns are the features.

5.3.3 Separating Results

For testing model, we have used 2 speakers voices for testing the proposed model.

1- **Input data (Persons 1 and Persons 2 Signal)**

At first, we take wav files of two persons that talked in low noise environment. Each person's voice print is taken as an input. Figure 29 shows the speakers voice single.

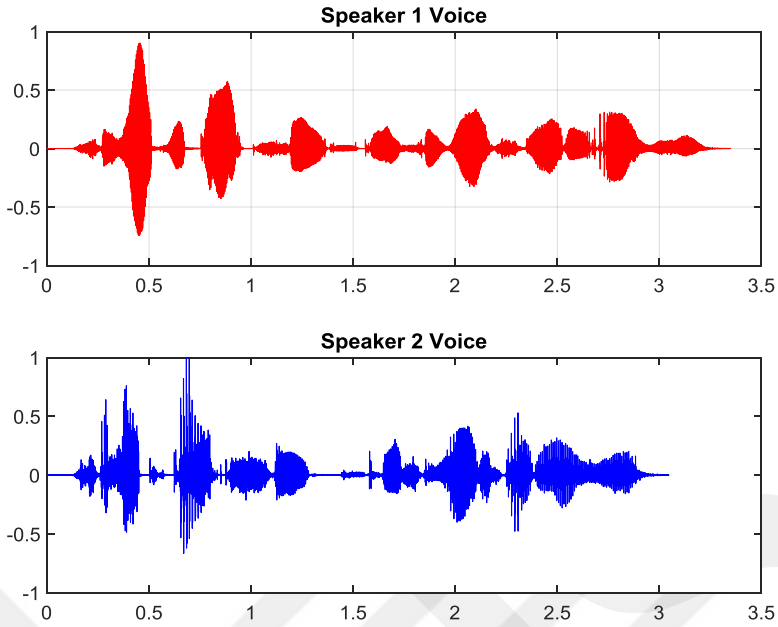


Figure 29 . two person voice signal

2- Mixed speech Signal

The second part is two mix voices of two persons and generate mixture voice signal as shown in figure 30.

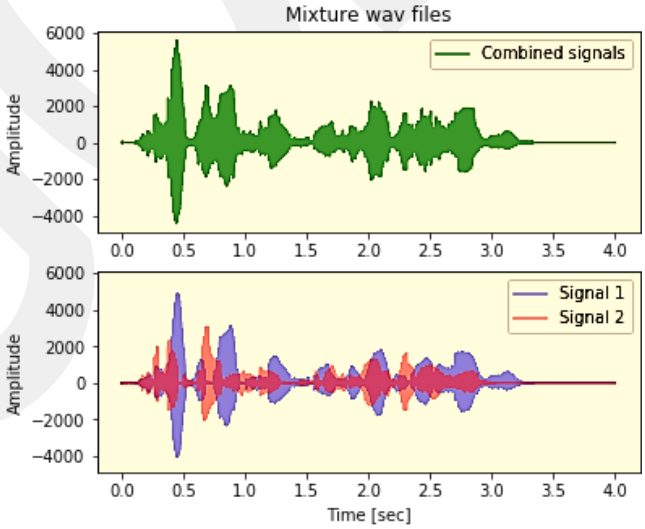
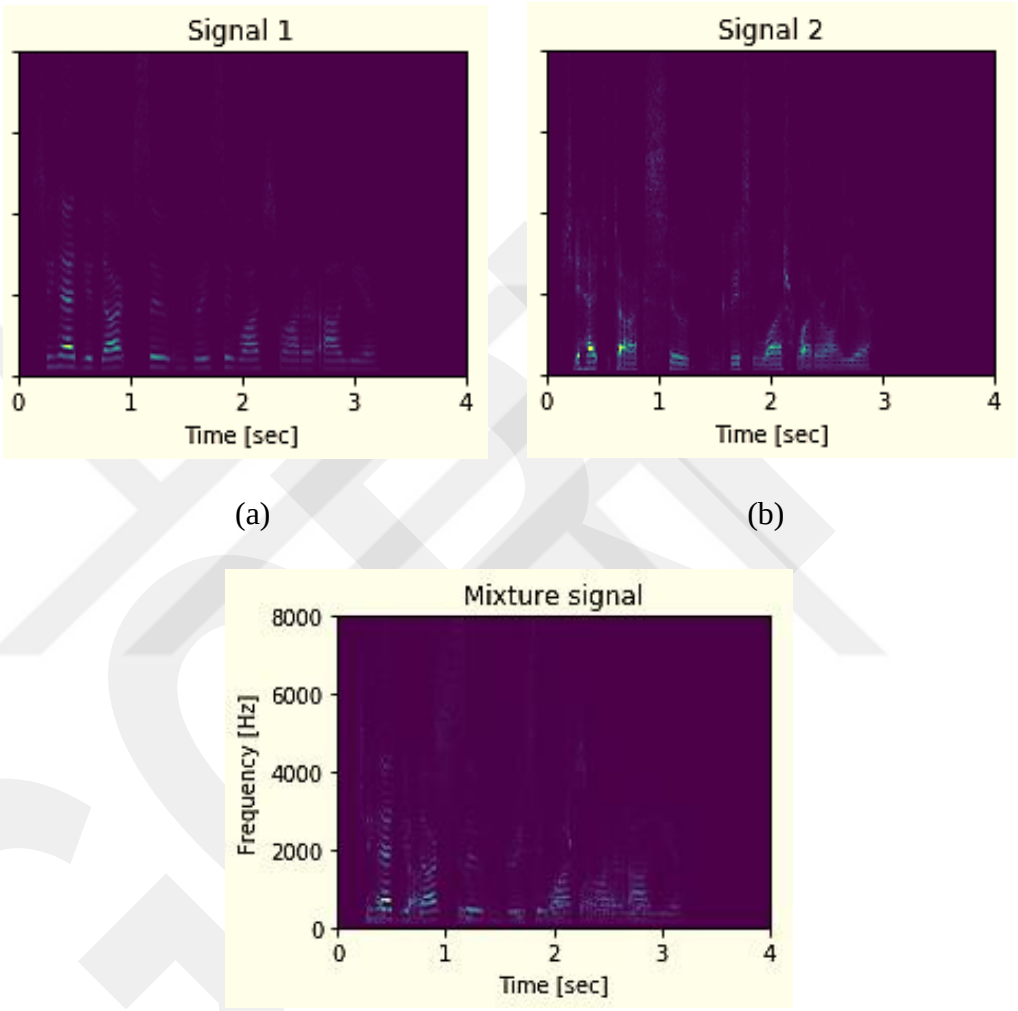


Figure 30. Mixing voice of two persons.

3- Spectrogram

The next stage is to get spectrogram signal of each person's voices and the mixing signal, figure 31 shows the spectrum signals of two person and the mixing signal.



(a) (b)

(c)

Figure 31. Spectrum signal of (a) first person, (b) second person, (c) mixing voices of both

4- Masks

The next process is to get voice masks of two person and the mixing signal to used then as input to network training. Figure 32 shows the voices masks results.

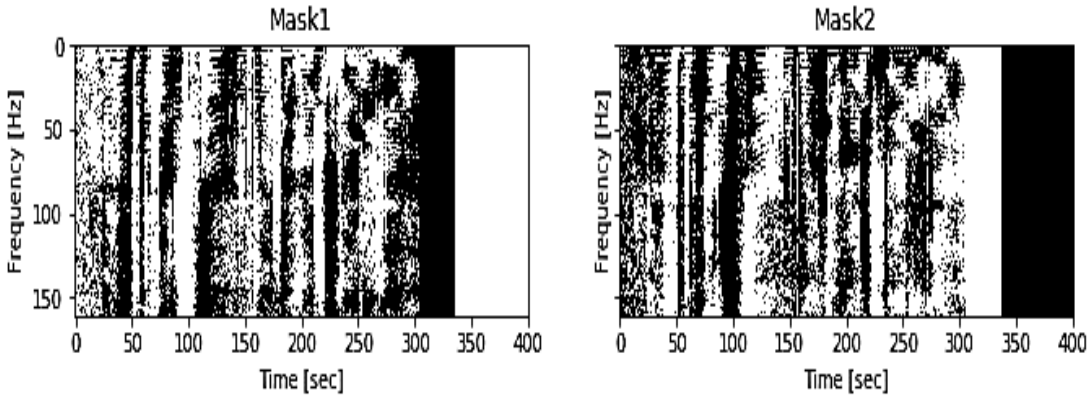


Figure 32**.** voice masks of each person

4- Recognized signal Results

This part recovers the original signal of each person by separation it from mixing (and/or noisy environment)

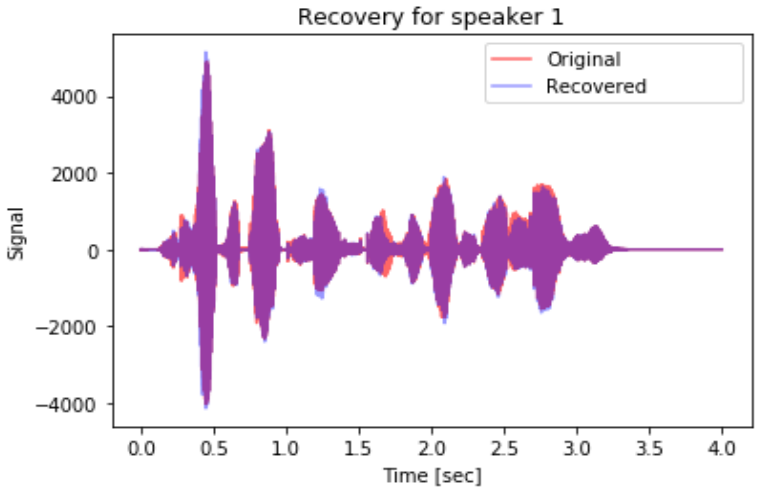


Figure 33**.** Recovered signal for speaker one

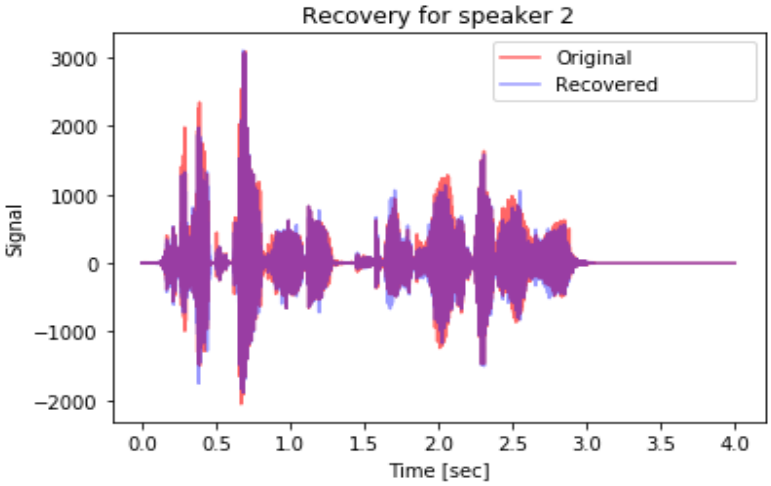


Figure 34**.** Recovered signal for speaker two

5.3.4 Comparison Between Proposed Model with the Other Related Works

This section concludes with some of the key aspects of the relevant work reported in this study. Table 6 shows the main features of the related works.

Table 6: Comparison between the proposed model and some of related works.

Comparison criterion	Methodology	Achievement	Similarities and differences from our study
Work in [106]	Utilized complex convolutional DT of 3 layers.	The separation results show it are on a par with equivalent binary-mask based non-complex separation approaches and the separation quality is similar to binary mask based convolutional DNN approaches but features slightly improved artefact performance	**Similarities** Utilized convolutional Neural network **Differences** Not used Bidirectional Long Short-Term Memory
Work in [107]	Utilized pyramidal Bidirectional Long Short-Term Memory (PBLSTM) neural network	Pyramidal Bidirectional Long Short-Term Memory (PBLSTM) neural network	**Similarities** Utilized Bidirectional Long Short-Term Memory (BLSTM)

			Differences Not used Dilated CNN
Work in [108]	three-dimensional convolutional architecture for audio and visual stream networks along with convolutional fusion in secular dimension	Architecture beats the other present methods for audio and visual matching, in addition reduces the number of parameters considerably in comparison to the previously proposed methods. the performance results of their model present the effectiveness of the learning of when employing CNN	**Similarities** Utilized convolutional Neural network **Differences** Using audio and visual for speech isolation
Work in [109]	Combine fully convolutional neural networks (FCN) and Bidirectional long short-term memory (BLSTMs) which is a type of recurrent neural networks	The proposed model that based on the combining of BLSTMs and FCNs achieved much better separation and high performance than using every model individually	**Similarities** Utilized BLSMs **Differences** Limited for single channel audio source separation (SCSS) system not for audio cocktail parity effect or speech separation Not utilized Dilated CNN

CHAPTER 6

DISCUSSION

6.1 Discussion

As machine learning tools are used to predict speaker according to his/her voice imprint, the performance of deployed tools are monitored in order to select best accuracy of prediction (identification). The very first machine learning deployment is made using the Random Forest Algorithm which yielded accuracy of ten percent only and the same was reported as minimum ever as compared to other experiments (tools).

From the other hand Feed Forward Neural Network is used as an option to enhance the accuracy of identification, for the plain (common) settings of Neural Network classifier, accuracy was swinging between high and low values, the fluctuation in the results of this classifier was termed to the nature of tuning the machine which is happening randomly (random numbers of weights every time machine is restarted). Eventually, an average accuracy could obtain from this classifier equal to ninety one percent approx. Figure 28 demonstrates the accuracy of all the algorithms used for speaker identifications.

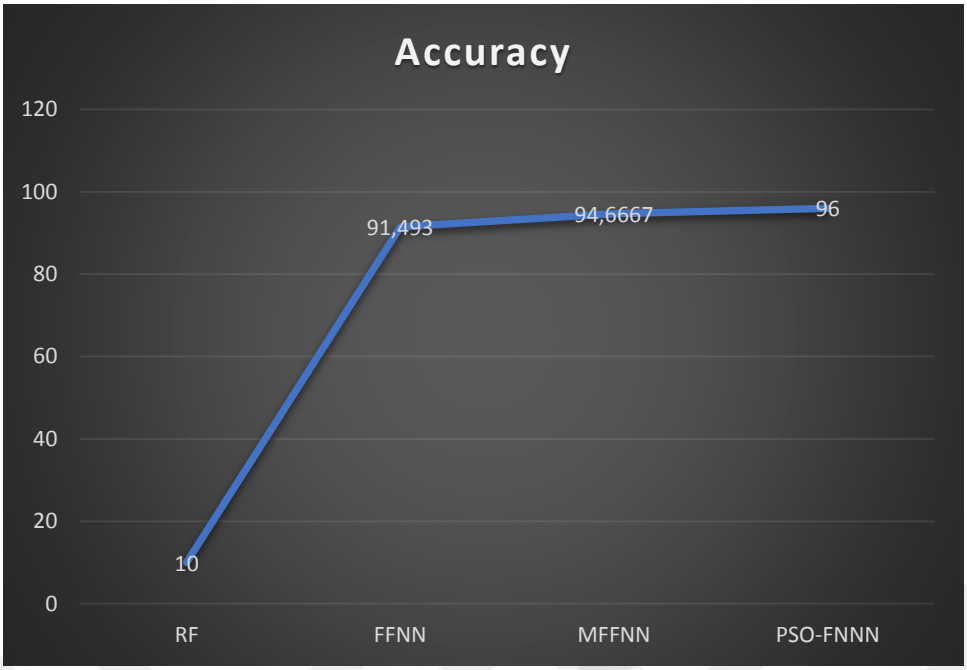


Figure 35: A plot of speaker identification accuracy all machine learning tools.

With effort to ensure more enhancement of the results, the current (used) model of Feed Forward Neural Network classifier is treated as frozen weight machine. This means, tuning of neural network classifier for several iterations say one-hundred! Has yielded different accuracies so, in frozen model, only single experiment is made to identify the weight vector that yielded the beset accuracy and hence the vector is set as a permanent weight of the machine so that, no random allotment is to be made on further actions and algorithm will consider the allotted weight only wherever the classifier came to action. The result of this (third experiment has yielded an accuracy of ninety-four percent approx.

The last trial to enhance the accuracy of identification the speakers is made using advance optimization algorithm more likely Particle swarm optimization, the same enhanced the accuracy to be ninety-six percent.

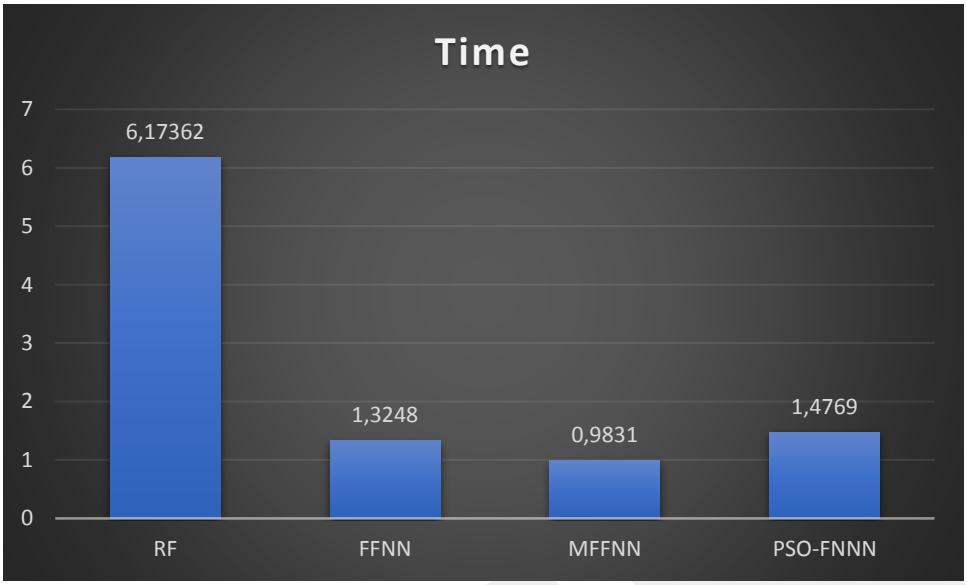


Figure 36: Time consumption of the machine learning algorithms used to identify the speaker.

Figure 29 depicts the time consumption of the four machine learning experiments, time of Particle Swarm Optimization combined by Feed Forward Neural Network higher that the time in plain FFNN and modified FFNN i.e experiments two and three. This time is far less than the time in Random Forest algorithm. From the time graph, there is a noteworthy tradeoff between accuracy and time in machine learning algorithm used in this study.

The next argument in performance monitoring is related to the training accuracy (learning efficiency). The mean square error is representing the average error square in the post the training which fulfil the Equation 5.1 and 5.2.

Let the vector V_o is the results of neural network after it gets trained and let the V_t is the correct (expected) results which may be called as a target. The so-called error is represented as E which can be given in the Equation below.

$$E = V_t - V_o \tag{5.1}$$

Mean square error is counted for the error vector E has a length of k is given in Equation (5.2). similarly, the root mean square error is given in Equation (5.3).

$$MSE = \frac{\sum_{n=1}^{k}(E[n])^2}{k} \tag{5.2}$$

$$RMSE = \left(\frac{\sum_{n=1}^{k}(E[n])^2}{k}\right)^{0.5} \tag{5.3}$$

The same is obtained for each machine learning algorithm and the results are demonstrated in Figure 30 and 31.

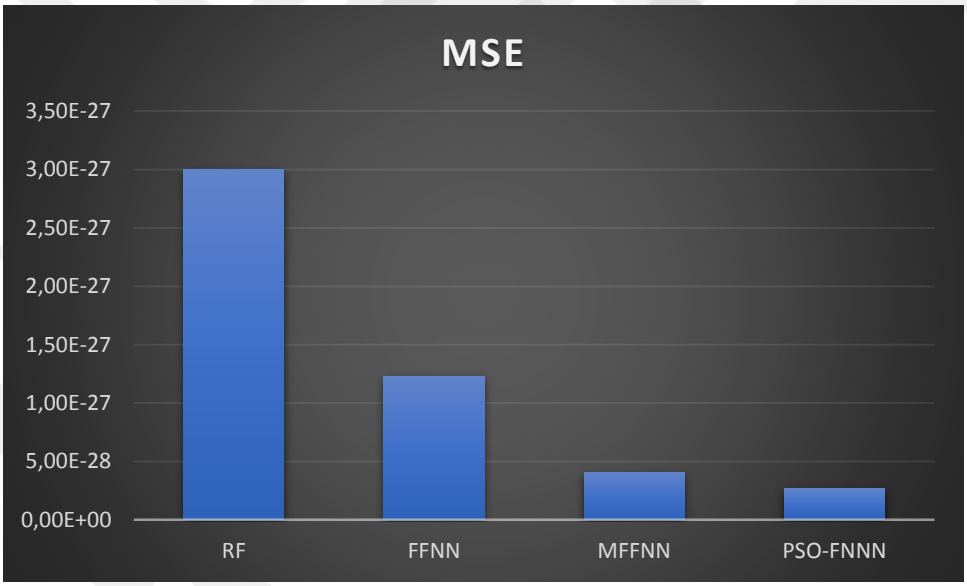


Figure 37: Mean Square Error in each machine learning algorithm.

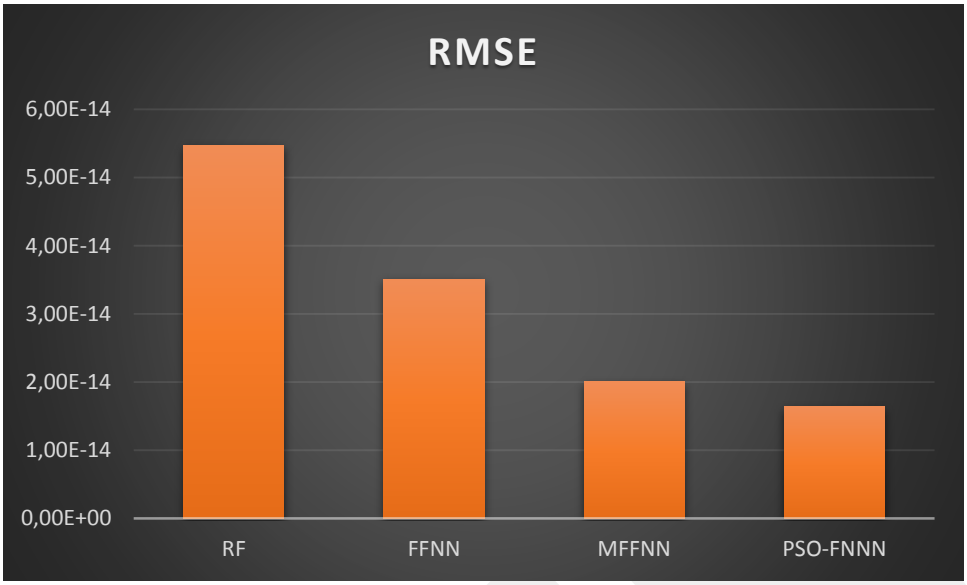


Figure 38: Root Mean Square Error in each machine learning algorithm.

The last argument in performance monitoring is the number of epochs taken by particular algorithm to reach the required accuracy. Epochs is also termed as trial or iteration which is corresponded to number of attempts made by the algorithm to reach the final results (approximation).

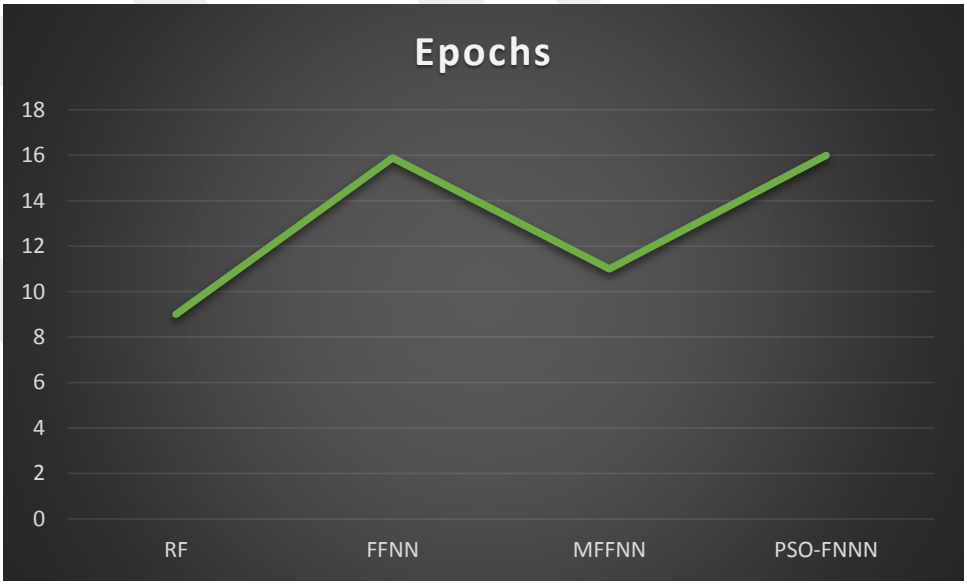


Figure 39: Epochs of each machine learning algorithm used to recognize the speaker.

CHAPTER 7

CONCLUSION

Machine learning based speaker identification system is outperformed over the traditional models. It can be used to learn the uncertain behaviors in the speech signal such as background fluctuation and noise impact. Enhance speech features are required for consistent speaker recognition in text dependent model.

Mel frequency cepstrum coefficients are extracted for each speech signal, a total number of twelve coefficients are made available after processing the speech data using Mel frequency cepstrum coefficients algorithm. In order to verify the speakers as second check active, one feature is recalled from time domain analysis namely as fundamental frequency coefficient. This is added to Mel frequency cepstrum coefficients to form a final features vector of thirteen elements.

Three machine learning approaches are made to enhance the speaker recognition task namely as Random Forest, conventional model of Feed Forward Neural Network, freezing model of Feed Forward Neural Network and ultimately, Particle Swarm optimization-based Feed Forward Neural Network.

Each algorithm was descried in the preceding section of this dissertation, from the performance point of view, the least model which called as PSO-FFNN is outperformed as compared to the others.

PSO-FFNN has yielded verification (recognition) accuracy of ninety-six percent in relatively short time.

For solving cocktail party effect, we have proposed a model based on utilized deep learning network that combining Fully Convolutional Network (FCN) and a Bidirectional Long Short-Term Memory (BLSTM) for source separation. The FCN is applied first to acquire an initial estimation of the magnitude spectrogram of the specific source coming from the input sequence. Then the initial estimation is passed to BLSTM network to improve the output sequence of the FCN. The results shows that the proposed FCN-BLSTM network model is able to captures the characteristics of spectro-temporal of the audio data much better than single model (FCN or BLSTM) where its apereas from retrieve voice signals of desired speaker.

REFERENCES

1. BAI Jun-mei, ZHANG Shi-Iei, ZHANG Shu-wu and XU Bo, "Robust Speaker Recognition in Noisy Environment," Journal of Chinese Information Processing. Vol. 20, pp. 91-97, February 2006.

2. Z. H. Chen, Y. F. Liao and Y. T. Juang, "Prosodic rrodeling and Eigen-Prosody Analysis for Robust Speaker Recognition," Proc. ICASSP 2005. PA. USA. vol. I, pp. 185-188, March 2005.

3. Gong, W.-G., Yang, L.-P., and Chen, D. Pitch synchronous based feature extraction for noise-robust speaker verification. In Proc. Image and Signal Processing (CISP 2008) (May 2008), vol. 5, pp. 295-298.

4. DENG Jing, ZHENG Fang, LIU Jian and WU Wenhu, "Using subband Mel-spectrum centroid and Gaussian mixture correlation for robust speaker identification ," Acta Acustica. Vol. 5, pp47 1-475, 2006.

5. H. Hermansky, N. Morgan. "RASTA processing of speech signal," IEEE Trans. On speech and Audio Processing. vol. 4, pp. 578-589, February. 1994.

6. F.H. Liu, A. Acero, R Stem. "Efficient joint compensation of speech for the effects of additive noise and linear filtering," Proc. Of IEEE ICASP. pp. 257-260. January 1992.

7. L. R Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Prentice-Hall, NJ, 1993.

8.Nosratighods, M. Ambikairajah, E. and Epps, J., "Speaker Verification Us ing A Novel Set of Dynamic Features," Pattern Recognition, 2006. ICPR 2006. 18th International Conference on Hong Kong. China, vol. 4, pp. 266-269, September 2006.

9. J.H.L Hansen, L.M. Arslan. Robust feature-estimation and objective quality assessment for noisy speech recognition using credit card corpus [J]. IEEE Trans. On Speech and Audio Processing, 1995, 3(3): 169-184.

10. A. P. Varga, H. J. M. Steeneken, M. Tomlinson, D. Jones. "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Documentation included in the NOISEX-92 CD-ROMS, 1992

11. G.Hinton*etal*,"Deepneuralnetworksforacousticmodelinginspeech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

12. T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Cˇernocky´, "Em- pirical evaluation and combination of advanced language modeling tech- niques," in *Proc. Interspeech*, 2011, pp. 605–608.

13. G. Saon*et al.*, "The IBM 2016 English conversational telephone speech recognition system," in *Proc. Interspeech*, 201, 2017.

14. W. Xiong*et al.*, "The Microsoft 2016 conversational speech recogni- tion system," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5255–5259.

15. H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, vol. 247, Berlin, Germany: Springer, 2012.

16. A. Graves, S. Ferna´ndez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recur- rent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

17. Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2015, pp. 167–174.

18. A. W. Senior and A. J. Robinson, "Forward-backward retraining of re- current neural networks," in *Adv. Neural Inform. Process. Syst.*, 1996, pp. 743–749.

19. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

20. D Bahdanau, J. Chorowski, D. Serdyuk, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4945–4949.

21. H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic- to-word LSTM model for large vocabulary speech recognition," in *Proc. Interspeech*, 2017, pp. 3707–3711.

22. H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recur- rent neural network acoustic models for speech recognition," in *Proc. Interspeech*, 2015, pp. 10–13.

23. K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, and D. Na- hamoo, "Direct acoustics-to-word models for English conversational speech recognition," in *Proc. Interspeech*, 2017, pp. 959–963.

24. D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2338– 2347, Nov. 2011.

25. K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, and B. Kings- bury, "End-to-end ASR-free keyword search from speech," in *Proc. ICASSP*, 2017, pp. 4840–4844.

26. T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2009, pp. 421–426.

27. Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via seg- mental DTW on Gaussian posteriorgrams," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2009, pp. 398–403.

28. G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spot- ting using long short-term memory networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5236–5240.

29.K.Levin,K.Henry,A.Jansen,andK.Livescu,"Fixed-dimensionalacous-tic embeddings of variable-length segments in low-resource settings," in Proc. Int. Conf. Acoust., Speech, Signal Process., 2013, pp. 410–415.

30. Y. Chung, C. Wu, C. Shen, H. Lee, and L. Lee, “Audio Word2Vec: Un-supervised learning of audio segment representations using sequence-to- sequence autoencoder,” in *Proc. Interspeech,* 2016, pp. 410–415.

31. H.Kamper,W.Wang,andK.Livescu,“Deepconvolutionalacousticwordembeddings using word-pair side information,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4950–4954.

32. W. He, W. Wang, and K. Livescu, “Multi-view recurrent neural acoustic word embeddings,” in *Proc. ICLR*, 2017.

33. D. Palaz, G. Synnaeve, and R. Collobert, “Jointly learning to locate and classify words using convolutional networks,” in *Proc. Interspeech,* 2016, pp. 2741–2745.

34. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in *Proc. NIPS Workshop on Deep Learning*, 2014.

35. D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. ICLR,* 2014.

36. Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, “Character-aware neural language models,” in *Proc. AAAI,* 2016, pp. 2741–2749.

37. J.Cui*etal.*,“Multilingualrepresentationsforlowresourcespeechrecog- nition and keyword search,” in *Proc. IEEE Autom. Speech Recognit. Un- derstanding Workshop*, 2015, pp. 259–266.

38. F.Chollet,“Keras,”2015.[Online].Available:https://github.com/fchollet/ keras

39. Theano Development Team, “Theano: A Python framework for fast computation of mathematical expressions,” *arXiv e-prints,* abs/1605.02688, May 2016.

40. N.MorganandH.Bourlard,“Generalizationandparameterestimationinfeedforward nets: Some experiments,” in *Proc. Neural Inf. Process. Syst.*, 1989, pp. 630–637.

41. J. Martens, “Deep learning via Hessian-free optimization,” in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 735–742.

42. B. Kingsbury, T. N. Sainath, and H. Soltau, “Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization,” in *Proc. Interspeech*, 2012, pp. 10–13.

43. F.Jelinek,“Continuousspeechrecognitionbystatisticalmethods,”Proc. IEEE, vol. 64, no. 4, pp. 532–556, Apr. 1976.

44. G. Hinton et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

45. T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying conditional ran- dom fields to Japanese morphological analysis.” in Proc. Conf. Empirical Methods Natural Lang. Process., 2004, vol. 4, pp. 230–237.

46. S. Bird, “NLTK: The natural language toolkit,” in Proc. Joint Conf. Int. Committee Comput. Linguist. Assoc. Comput. Linguist. Interact. Presen- tat. Sessions, 2006, pp. 69–72.

47. M. Mohri, “Finite-state transducers in language and speech processing,” Comput. Linguist., vol. 23, no. 2, pp. 269–311, 1997.

48. T. Hori and A. Nakamura, Speech Recognition Algorithms Using Weighted Finite-state Transducers. Williston, VT, USA: Morgan & Claypool, 2013.

49. J.Chorowski,D.Bahdanau,K.Cho,andY.Bengio,“End-to-endcontin- uous speech recognition using attention-based recurrent nn: First results,”

in NIPS 2014 Workshop Deep Learning, Dec. 2014.

50. A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in Proc. Int. Conf. Mach. Learn., 2014, pp. 1764–1772.

51. J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 577–585.

52. W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 4960– 4964.

53. L. Lu, X. Zhang, and S. Renals, “On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recogni- tion,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 5060–5064.

54. D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in Proc. Int. Conf. Learn. Repre- sentations, 2015.

55. Y. Wu et al., “Google’s neural machine translation system: Bridging the gap between human and machine translation,” arXiv:1609.08144, 2016.

56. J.ChorowskiandN.Jaitly,“Towardsbetterdecodingandlanguagemodel integration in sequence to sequence models,” in Proc. Interspeech, 2017, pp. 532–527.

57. A. Graves, S. Ferna´ndez, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recur- rent neural networks,” in Proc. Int. Conf. Mach. Learn., 2006, pp. 369–376.

58. Y. Miao, M. Gowayyed, and F. Metze, “EESEN: End-to-end speech recog- nition using deep RNN models and WFST-based decoding,” in Proc. IEEE Workshop Autom. Speech Recognit. Understanding, 2015, pp. 167–174.

59. D. Amodei et al., “Deep speech 2: End-to-end speech recognition in English and Mandarin,” Int. Conf. Mach. Learn., 2016, pp. 173–182.

60. H. Soltau, H. Liao, and H. Sak, “Neural speech recognizer: Acoustic- to-word LSTM model for large vocabulary speech recognition,” in Proc. Interspeech, 2017, pp. 3707–3711.

61. S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2017, pp. 4835–4839.

62. T. Hori, S. Watanabe, and J. R. Hershey, “Joint CTC/attention decoding for end-to-end speech recognition,” in Proc. Assoc. Comput. Linguist., 2017, pp. 518–529.

63. N. Xue et al., “Chinese word segmentation as character tagging,” Comput. Linguist. Chin. Lang. Process., vol. 8, no. 1, pp. 29–48, 2003.

64. H. Bourlard and N. Morgan, Connectionist Speech Recognition: A Hybrid Approach. Norwell, MA, USA: Kluwer, 1994.

65. T. Mikolov, M. Karafia´t, L. Burget, J. Cernocky´, and S. Khudanpur, "Recurrent neural network based language model," in Proc. Interspeech, 2010, pp. 1045–1048.

66. D. Povey et al., "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in Proc. Interspeech, 2016, pp. 2751–2755.

67. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

68. A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in Proc. IEEE Workshop Autom. Speech Recognit. Understanding, 2013, pp. 273–278.

69. N. Kanda, X. Lu, and H. Kawai, "Maximum a posteriori based decoding for CTC acoustic models," in Proc. Interspeech, 2016, pp. 1868–1872.

70. S.Tokui,K.Oono,S.Hido,andJ.Clayton,"Chainer:Anext-generation open source framework for deep learning," in Proc. Workshop Mach. Learn. Syst., Annu. Conf. Neural Inf. Process. Syst., 2015.

71. D. Povey et al., "The Kaldi speech recognition toolkit," in Proc. IEEE Workshop Autom. Speech Recognit Understanding, 2011.

72. L.Lu,L.Kong,C.Dyer,andN.A.Smith,"Multi-tasklearningwithCTC and segmental CRF for speech recognition," in Proc. Interspeech, 2017, pp. 954–958.

73. D.Bahdanau,J.Chorowski,D.Serdyuk,P.Brakel,andY.Bengio,“End-to- end attention-based large vocabulary speech recognition,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2016, pp. 4945–4949.

74. A. Graves, “Supervised sequence labelling with recurrent neural net- works,” Ph.D. dissertation, Dept. Informat., Technische Univ. Mu¨nchen, Munich, Germany, 2008.

75.L.D.Consortium,CSR-II(WSJ1)Complete,vol.LDC94S13A.Philadel- phia, PA, USA: Linguistic Data Consortium, 1994.

76. J. Garofalo, D. Graff, D. Paul, and D. Pallett, CSR-I, (WSJ0) Complete, vol. LDC93S6A. Philadelphia, PA, USA: Linguistic Data Consortium, 2007.

77. E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” in Proc. Comput. Speech Lang., 2017, pp. 535–557.

78. K. Maekawa, H. Koiso, S. Furui, and H. Isahara, “Spontaneous speech corpus of Japanese,” in Proc. Int. Conf. Lang. Resour. Eval., 2000, vol. 2, pp. 947–952.

79. Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, “HKUST/MTS: A very large scale Mandarin telephone speech corpus,” in Proc. Chin. Spoken Lang. Process., 2006, pp. 724–735.

80. I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 3104–3112.

81.D.Bahdanau,J.Chorowski,D.Serdyuk,P.Brakel,andY.Bengio,“End-to-endattention-based large vocabulary speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 4945–4949.

82. T. Moriya, T. Shinozaki, and S. Watanabe, “Kaldi recipe for Japanese spontaneous speech recognition and its evaluation,” in Proc. Autumn Meet. Acoust. Soc. Japan, 2015, pp. 155–156.

83. T. Hori, S. Watanabe, Y. Zhang, and W. Chan, “Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM,” in Proc. Interspeech, 2017, pp. 949–953.

84. Y. Miao, M. Gowayyed, X. Na, T. Ko, F. Metze, and A. Waibel, “An empirical exploration of CTC acoustic models,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 2016, pp. 2623–2627.

85. W. Chan, Y. Zhang, Q. Le, and N. Jaitly, “Latent sequence decomposi- tions,” in Proc. Int. Conf. Learn. Representations, 2017.

86. Z.-H. T. M. G. C. S. H. J. Mohamed Abou-Zleikha, "A DISCRIMINATIVE APPROACH FOR SPEAKER SELECTION IN SPEAKER DE-IDENTIFICATION SYSTEMS," 23rd European Signal Processing Conference (EUSIPCO), Vols. 978-0-9928626-3-3/15/$31.00 ©2015 IEEE, 2015.

87. X. F. a. J. H. Hansen, "SPEAKER IDENTIFICATION WITH WHISPERED SPEECH BASED ON MODIFIED LFCC PARAMETERS AND FEATURE MAPPING," 978-1-4244-2354-5/09/$25.00 ©2009 IEEE.

88. J. Z. X. P. B.-C. L. BO WANG, "A NOVEL SPEAKER CLUSTERING ALGORITHM IN SPEAKER RECOGNITION SYSTEM," IEEE Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 2016.

89. B. G. Nagaraja, "Efficient Window for Monolingual and Crosslingual Speaker Identification using MFCC," IEEE International Conference on Advanced Computing and Communication Systems, 2015.

90. S.-A. S. D. O. MdFoezur Rahman Chowdhury, "DISTRIBUTED AUTOMATIC TEXT-INDEPENDENT SPEAKER IDENTIFICATION USING GMM-UBM SPEAKER MODELS," 978-1-4244-3508-1/09/$25.00 ©2016 IEEE, 2016.

91. N. E. m. El bachirTazi, "An Hybrid Front-End for Robust Speaker Identification Under Noisy Conditions," 978-1-5090-6435-9/17/$31.00 ©2017 IEEE, 2017.

92. M. M. L. S. Y. M. N. L. B. Y. Roman Martsyshyn, "Technology of Speaker Recognition of Multimodal Interfaces Automated Systems Under Stress," CADSM 2013, 19-23 February, 2013, Polyana-Svalyava (Zakarpattya), UKRAINE, 2013.

93. R. S. H. a. T. K. B. Naresh P. Jawarkar, "Speaker Identification using Whispered Speech," IEEE International Conference on Communication Systems and Network Technologies, 2015.

94. N. A. A. A. M. R. A. H. A. MAAZOUZI, "MFCC and Similarity Measurements for Speaker Identification Systems," IEEE International Conference on Electrical and Information Technologies, 978-1-5386-1516-4/17/$31.00 ©2017 IEEE, 2017.

95. Y.-H. Chao, "Speaker Identification Using Pairwise Log-Likelihood Ratio Measures," International Conference on Fuzzy Systems and Knowledge Discovery, 978-1-4673-0024-7/10/$26.00 ©2012 IEEE, 2012.

96. W. A.-S. A.-R. A.-Q. a. I. N. A.-I. Khaled Daqrouq, "Speaker Identification Wavelet Transform Based Method," Speaker Identification Wavelet Transform Based Method, 978-1-4244-2206-7/08/$25.00 ©2015 IEEE, 2015.

97. R. S. Mohsen Bazyar, "A New Speaker Change Detection Method in a Speaker Identification System for two-Speakers Segmentation," IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), 978-1-4673-5159-1/14/$26.00 ©2014 IEEE, 2014.

98. M. S. a. K. I. SerhanDagtas, "A Multi-modal Virtual Environment with Text-independent Real-Time Speaker Identification," Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering (ISMSE'04) 0-7695-2217-3/04 $20.00 © 2004 IEEE.

99. V. R. A. a. P. L. D. Leon, "SUPPORT VECTOR MACHINE BASED SPEAKER IDENTIFICATION SYSTEMS USING GMM PARAMETERS," 978-1-4244-5827-1/09/$26.00 ©2016 IEEE, 2016.

100. F. u. R. S. K. A. M. &. G. S. Chandar Kumar, "Analysis of MFCC and BFCC in a Speaker Identification system," International Conference on Computing, Mathematics and Engineering Technologies, 978-1-5386-1370-2/18/$31.00 ©2018 IEEE, 2018.

101. N. A. M. R. A. H. Abd-ErrahimMaazouzi*, "A Speaker Recognition System Using Power Spectrum Density and similarity measurements," 978-1-4673-9669-1/15/$31.00 ©2015 IEEE, 2015.

102. D. L. Ahmad Shahab, "An Investigation of Indonesian Speaker Identification for Channel Dependent Modeling using I-vector," Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA), 978-1-5090-3516-8/16/$31.00 ©2016 IEEE, 2016.

103. Y. S. a. Q. Zhu, "Speaker Identification under The Changed Sound Environment," 978-1-4799-3903-9/14/$31.00 ©2014 IEEE, 2014.

104. T. E. Guillermo Garcia, "A STATISTICAL APPROACH TO PERFORMANCE EVALUATION OF SPEAKER RECOGNITION SYSTEMS," 1-424407281/07/$20.00 ©2017 IEEE, 2017.

105. V. R. A. a. P. L. D. L. AditiAkula, "SPEAKER IDENTIFICATION IN ROOM REVERBERATION USING GMM-UBM," 978-1-4244-3677-4/09/$25.00 ©2009 IEEE, 2009.

106. Simpson AJ. Deep transform: cocktail party source separation via complex convolution in a deep neural network. arXiv preprint arXiv:1504.02945. 2015 Apr

107. Chan W, Jaitly N, Le QV, Vinyals O. Listen, attend and spell. arXiv preprint arXiv:1508.01211. 2015 Aug 5.

108. Torfi, A., Iranmanesh, S.M., Nasrabadi, N. and Dawson, J., 2017. 3d convolutional neural networks for cross audio-visual matching recognition. IEEE Access, 5, pp.22081-22091.

109. Grais, Emad M and Plumbley, Mark (2018) Combining Fully Convolutional and Recurrent Neural Networks for Single Channel Audio Source Separation In: AES 144th Convention, 23-26 May 2018, Milan, Italy.