SPEAKER IDENTIFICATION SYSTEM ENHANCED BY OPTIMIZED NEURAL NETWORKS AND FEATURE FUSION TECHNIQUES EVALUATED BY COCHLEAR IMPLANT-LIKE SPECTRALLY REDUCED SPEECH

by

Najiya Omar

Submitted in partial fulfillment of the requirements
for the degree of Master of Applied Science

at

Dalhousie University
Halifax, Nova Scotia
February 2017

Dedication

This work is dedicated to my home country Libya and Sirte University in my hometown of Sirte. To achieve this degree is the best way that I am able to thank my hometown and university that are both experiencing great hardships, for all that they have done for me throughout my life.

Table of Contents

List of Tables . . . . . **v**ii

List of Figures . . . . . **ix**

Abstract . . . . . **xi**

LIST of ABBREVIATIONS USED . . . . . **xii**

Acknowledgement . . . . . **xiii**

Chapter 1 Introduction **1**

1.1 Background and Problem Statement 1

1.2 Motivation 3

1.3 Objective and Contribution 3

1.4 Thesis Organization 4

Chapter 2 LITERATURE REVIEW **6**

2.1 Speaker Identification System 6

2.1.1 Speech Analysis 7

2.1.2 Feature Extraction 8

2.1.3 Classification 9

2.2 Analysis of the First Stage of SI System 11

2.2.1 Cepstral and Non-Cepstral Feature Methods 11
2.2.2 Combination of Feature Extraction Approaches 14

2.3 Analysis of the Second Stage of SI System 15
2.3.1 Comparison of Different (ML) Algorithms 15
2.3.2 Performance Comparison of HMM, SVM, GMM and VQ 16
2.3.3 Enhancement of the SVM Performance using RBF 17
2.3.4 Performance Comparison of RBFNN with other Classifiers 18
2.3.5 Enhancement of RBFNN and MLP 19

2.4 Reliability Analysis of Evaluation Parameter 20

Chapter 3 SPEAKER RECOGNITION SYSTEM 21

3.1 Speaker Identification (SI) vs. Speaker Verification (SV) 21

3.2 Application of Speaker Recognition 23
3.2.1 Security Services Application 23
3.2.2 Law Services 24
3.2.3 Medicine Services 24

3.3 Mechanism of Speech: The Process Of Human Speech 25
3.3.1 The Process of Human Speech Production 25
3.3.2 The Process of Human Speech Perception 26

3.4 Speech Production Model 26
3.4.1 Larynx Stage Model 28
3.4.2 Vocal Tract Stage Model 30

3.5 Mathematics of Overall Speech Production Process 31

Chapter 4 SPEAKER IDENTIFICATION SYSTEM 35

4.1 Feature Extraction Techniques 35

4.1.1 Linear Predictive Cepstral Coefficients (LPCC) 35

4.1.2 Mel Frequency Cepstral Coefficients (MFCC) 44

4.2 Radial Basis Function Neural Network (RBFNN) 47

4.2.1 Introduction 47

4.2.2 Architecture of RBFNN 48

4.2.3 RBFN Training 49

4.3 Bacterial Foraging Optimization Algorithm (BFOA) 50

4.3.1 Introduction 50

4.3.2 The Concept of BFOA 51

4.4 Evaluation Parameters 58

4.4.1 Accuracy 58

4.4.2 Sensitivity 58

4.4.3 Specificity 58

4.4.4 Cochlear Implant-like Spectrally Reduced Speech (SRS) 59

Chapter 5 PROPOSED SYSTEM 60

5.1 Algorithm 60

5.1.1 Dataset 62

5.1.2 Proposed Features Extraction 62

5.2 Results and Discussion 66

5.2.1 LPCC_RBFNN 66
5.2.2 MFCC_RBFNN 69
5.2.3 LPCC _ MFCC, and LMACC Features Based Training RBFNN . . 71
5.3 Proposed RBFNN 72
5.3.1 LPCC Features Based Training of BFO Tuned RBFNN 73
5.3.2 MFCC Features Based Training of BFO Tuned RBFNN 75
5.3.3 LPCC & MFCC Features Based Training BFO Tuned RBFNN . . . 76
5.3.4 LMACC Features Based Training BFO 77
5.4 Summary 77
5.5 Case Study 78
5.6 Noisy Environment 80
Chapter 6 CONCLUSIONS. **83**
Bibliography **85**
Appendix A First Appendix **90**
A.1 List of Publications 90

List of Tables

Table 2.1 Summary of several research focuses on comparing cepstral and non-cepstral feature methods 13

Table 2.2 Summary of several research focuses on combination of feature extraction approaches 15

Table 2.3 Summary of several research works on comparison of HMM, SVM, GMM and VQ 17

Table 2.4 Summary of several research works on comparison of HMM, SVM, GMM and VQ 19

Table 5.1 Training Parameters of RBFNN 66

Table 5.2 LPCC Features Based RBFNN for Speaker Identification System Results 69

Table 5.3 MFCC Features Based RBFNN for Speaker Identification System Results 70

Table 5.4 Comparison between LPCC, MFCC, LPCC&MFCC, and LMACC in conventional RBFNN 71

Table 5.5 BFOA Parameters 73

Table 5.6 Comparison of RBFNN and BFO -RBFNN for SIS using LPCC Features 75

Table 5.7 Comparison of RBFNN and BFO-RBFNN for SIS using MFCC Features 75

Table 5.8 Comparison of Concatenating LPCC and MFCC Feature with Traditional and Proposed RBFNN 76

Table 5.9 Comparison of LMACC Feature with Traditional and Proposed RBFNN 77

Table 5.10 Comparison of all Feature Extraction Techniques 78

Table 5.11 Comparison of LMACC and SRS-LMACC with Traditional and Proposed RBFNN 80

Table 5.12 Comparison of all Feature Extraction Techniques with 0.1 RMS 81

Table 5.13 Comparison of all Feature Extraction Techniques with 0.5 RMS 82

List of Figures

Figure 1.1 Specturm, LPCC and MFCC Spectral Envelope. 3

Figure 2.1 Speaker Identification System procedures. 7

Figure 3.1 The two principal tasks of speaker recognition system: (a) speaker identification (SI). (b) speaker verification (SV) (Reynolds, 1995; Reynolds , 2002). 22

Figure 3.2 Schematic diagram of speech production and perception process. . . 25

Figure 3.3 Illustrative picture of speech production. The sources of the sounds are labeled periodic, impulsive and noise that can happen in the larynx or vocal tract (Quatieri, 2006.). 27

Figure 3.4 Speech production pattern in humans.(Quatieri, 2006). 28

Figure 3.5 Illustration of periodic glottal airflow velocity (Quatieri, 2002, p.62). 29

Figure 3.6 Vocal tract system. 31

Figure 3.7 Speech production model Ambikairajah (2010). 33

Figure 4.1 The idea of LPCC (Ambikairajah, 2010). 37

Figure 4.2 Procedures of LPCC. 38

Figure 4.3 Autocorrelation Method with overlapping regions Quatieri (2006). . 42

Figure 4.4 Davies and Mermelstein's Mel-scale filter-bank simulating critical band filters (Quatieri, 2006). 45

Figure 4.5 Procedures of MFCC. 46

Figure 4.6 RBFNN Architecture. 48

Figure 4.7 Swim and Tumble of a bacterium (Das et al., 2009). 52

Figure 4.8 a, b Flowchart of BFOA. 57

Figure 5.1 Flow chart for the proposed work 61

Figure 5.2 Festures Extraction Approaches 63

Figure 5.3 LPCC Features Plot of the Filtered Signal 67

Figure 5.4 RBFNN Output clusters vs. Target classes 68

Figure 5.5 LPCC_ RBFNN MSE Performance 68

Figure 5.6 MFCC Features Plot Filtered Signal 70

Figure 5.7 Comparison between LPCC, MFCC, LPCC_MFCC, and LMACC in conventional RBFNN 72

Figure 5.8 Objective function value illustrated through optimized BFO_RBFNN 74

Figure 5.9 BFO-RBF Output clusters vs. Target classes 74

Figure 5.10 LPCC_RBFNN Vs. LPCC_BFO_RBFNN 75

Figure 5.11 Comparison of RBFNN and BFO tuned RBFNN for SIS using MFCC 76

Figure 5.12 Comparison of combining LPCC &MFCC feature with Traditional and proposed RBFNN 76

Figure 5.13 Comparison of LMACC feature with Traditional and proposed RBFNN 77

Figure 5.14 SRS-LMACC feature based Evaluation of our proposed method . . . 79

Figure 5.15 SRS_LMACC-RBFNN vs SRS_BFO_LMACC_RBFNN 80

Abstract

Efficient feature extraction techniques are available in the literature in speaker identification (SI) system ; yet, their combined influences on each other as a fusion feature have not been fully investigated. Much research has been conducted in terms of enhancing only one of either the feature extraction or classification techniques. Past research has not succeeded to focus in enough detail on determining more appropriate methods of evaluating parameters. In order to enhance the features extraction of any individual's speech, we will concatenate two feature extraction techniques, along with the respective averages; Linear predictive, Mel-frequency, and the respective normalization Averages Cepstral Coefficients (LMACC) that would reflect positively on system performance. Optimized Radial Basis Function (RBF) neural network based on Bacteria Foraging Optimization Algorithm is used as a classifier to improve system performance. Cochlear implant-likespectrally reduced speech proposed in the literature will be used alongside accuracy, sensitivity, and specificity to evaluate the system.

LIST of ABBREVIATIONS USED

ANN	Artificial Neural Network
BFOA	Bacteria Foraging Optimization Algorithm
BP	Back Propagation
BPF	Band Bass Filter
DTW	Dynamic time warping
FD	Frequency Domain
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IFFT	Inverse Fast Fourier Transform
LMACC	Linear predictive, Mel-frequency, and Average of both Cepstral Coefficient
LPF	Low-Pass Filter
LPCC	Linear Predictive Cepstral Coefficients
MFCC	Mel-frequency Cepstral Coefficient
MLP	Multi Layer Perceptron
ML	Machine-Learning
MSE	Mean Squared Error
PSO	Particle Swarm Optimization
RPF	Radial Basis Function
SNR	Signal-to-Noise Ratio
SIS	Speaker Identification System
VQ	Vector Quantization
SRS	Spectrally Reduced Speech
VS	Speaker Verification

Acknowledgement

I would like to express my sincere gratitude to my thesis supervisor Dr. El-Hawary. It has been an honour for me to have someone of his calibre as my academic supervisor; I am inspired by the prodigious professional expertise and extensive knowledge that you offer. Thank you for helping to highlight my weak points during researching and writing this thesis.

I would never forget to extend a thank you to the members of the respectable committee Dr. Gu and Dr. Phillips, that have allowed me to clarify any questions that I had. Special thanks to Dr. Hamed Aly for providing me with the encouragement to continue working hard, and for clarifying and directing my navigation through graduate studies with helpful discussion and guidance.

Thank you all members in the department of Electrical and Computer Engineering and the Faculty of Graduate Studies at Dalhousie University. I am deeply appreciative of all staff at the library for allowing me to seek the previous studies and resources, which were valued and useful to this modest work.

Thank you also to all of my colleagues who have supported me along the way, offering advice, suggestions, or helping to solve a challenging problem. I would not want to forget anyone so I wish to extend this thank you to each and every one of you who I have encountered over the years, however brief.

Last and not least, I wish to give thanks to my family: my darling husband Imhamed, my lovely children. Without their support and sustained encouragement, I could not overcome the obstacles to completing this thesis/work. I would also like to thank Mr. Al-Habibi and Mr. Fowlie for their continued support and motivation during my study.

A special thank you as well to the Libyan-North American Scholarship Program and the Faculty of Graduate Studies for their financial support.

And of course, I am grateful to my parents, for they have converted me to the right way, and they were generous in praying that I might accomplish what I have intended to achieve in my life. They have helped me spiritually, despite being thousands miles away.

Chapter 1

Introduction

1.1 Background and Problem Statement

Since the 1930's many studies, where listeners are able to distinguish and identify the speakers from their voices, have been conducted. Like fingerprints, the speech signal can be a unique feature for every human individual through which the speaker can be identified and distinguished from others. In this respect, Kolokolov (2014) views the processing of speech signal as an essential procedure that takes place in recognizing speech. A speech signal conveys much information about the spoken word as well as the speaker's identity. Attempts to create rules and principles for distinguishing individuals from one another were made. Reynolds (1995) claimed that grasping the process that clarifies the information about the speaker, and attempting to imitate that method may allow for establishing the speaker recognition system.

Understanding the mechanism of speech in both production and perception process will lead to the idea of stimulating the speaker model. Speech signal is not only used as a mere procedure to identify a speaker, but has a variety of information standards. Singh and Pandey (2003) said that many levels of information are contained in the speech signal. They cited some examples of those levels; for instance, spoken words carry a message or information about the language used in communication, the emotion, gender and the iden-tity of the speakers themselves. Adding to this is the close relationship between speech and speaker recognition. The speech signal contains extraction, classification and recognition of information that work as tools through which recognizing a speaker happens automati-cally, whereas speech recognition has goals to recognize spoken words at the time of speech (op.cit).

One should, however, not forget that speaker recognition in particular is an important system due to its applications in critical aspects of everyday life. Thus, the features extracted from someone's voice are a focal stage in designing the speaker identification system. Undoubtedly, robust features extraction has a beneficial effect on performance of the classifier. The research conducted on speaker recognition system where the features extraction method MFCC has shown the best results in overall SIS improvement (Singh et al. 2012; Kumar et al. 2011; Farah and Shamim, 2013). Baidwan and Gujral (2014) suggested that future study should focus on fusion feature set. This combination of LPCC and MFCC extraction techniques has been studied (Yujin et al., 2010; Shinde and Pawar, 2013; Zhu and Liu, 2014; Nath and Kalita, 2015) previously; however, the influence of features combination has lagged behind.

The fundamental idea of LPCC and MFCC is focusing on simulating the vocal tract parameters determined by the spectral envelope of the signal' spectrum with both features generated in the quefrency domain; cepstral analysis. In the case of LPCC, the achievement of the spectral envelope (the curve linking the formants) is from finding predictor coefficients that replicate the vocal tract resonances while in MFCC, the a chievement of the spectral envelope is from applying Mel-frequency band-pass filters to the singal's spectrum. Although both techniques function different ways to extract the features related to the vocal tract, both work in the region where the formants are located (peaks of spectral envelope) as shown in Fig. 1.1; in that respect, the relationship of both features should be examined. To do so, the normalization averages of each of the features are concatenated with the original features (LMACC) is one way this could be valuable in such a system.

As has been previously studied (Finan et al., 1996; Venkateswarlu et al. 2011; Nijhawan and Soni, 2012; Anifowose, 2012), in the classification stage RBFNN performs better compared with other classifiers. Zhou and Gu (2010) suggested that the RBFNN's efficiency could be increased if its weights are properly adjusted. The BFO algorithm has been found in the literature to be the most effective method of enhancing the neural networks (Al-Hadi et al. 2011). So that future research should apply an optimization algorithm for RBFNN to increase the system performance and ultimately achieve a robust speaker identification system. In addition, the accuracy rate might be not enough to evaluate such a system (Zheng, 2015).

Thus, several carefully selected evaluation parameters should be considered such as sensitivity and specificity.

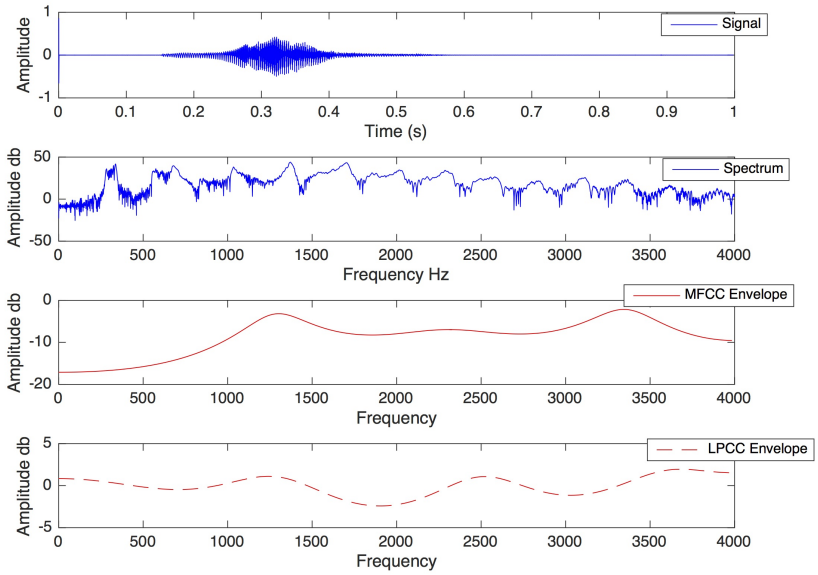

Figure 1.1: Specturm, LPCC and MFCC Spectral Envelope.

1.2 Motivation

As someone with partial hearing impairment, this has inspired me to begin research on speech production and perception, and focus in particular on this system so that I am able to gain a base knowledge of the field, which will allow me to continue forward in future speech research. Additionally, my undergraduate research focused on using artificial neural network in specific applications allowed me to develop my understanding, and ability to use more state-of-the-art systems such as RBFNN.

1.3 Objective and Contribution

Research in speech processing and artificial intelligence play an important role in the innovation of intelligent systems for many applications. One such intelligent system, Speaker

Identification System (SIS), is currently in use with highly sensitive information such as banking information, medical information, and password protection. Due to the nature of these applications, it is important that there is a high level of accuracy in system performance. The goal of this study is to understand the necessary stages of building the system and attempt to enhance the overall performance of the system by achieving the following objectives.

- Proposing new feature extraction technique that works on how the two features influence each other.
- Employing the comparison between four feature extraction techniques.
- The performance of the RBFNN will be improved through the use of an optimization algorithm.
- Different evaluation parameters will be applied to make precise assessments.
- Test our proposed system in different environmental conditions.

1.4 Thesis Organization

- Chapter 2 includes the research that has been conducted already on SIS, and the enhancements that have been achieved in both stages, as well aw how to build the speaker identification model.
- Chapter 3 discusses the speaker recognition system and its application, as well as an overview of the difference between speaker identification and speaker verification. From here, speech perception and production are discussed and the speech production model is mathematically explained.
- Chapter 4 is a more in-depth explanation of proposed system, including LPCC and MFCC as a feature extraction technique, as well as RBFNN as a classifier, and BFOA. The evaluation parameters are also discussed in this section.

- Chapter 5 includes the implementation of our proposed system and the comparative results both in clean and noisy environments.
- Chapter 6 is the conclusion and directions for future research.

Chapter 2

LITERATURE REVIEW

Past research has focused on achieving better SIS recognition performance by attempting to enhance either the Feature Extraction (Stage 1) or Classification methods (Stage 2). All previous research conducted has been assessed, with each study using different methods of evaluation. The review in this chapter focuses on understanding the effects of each stage on the overall system and how each stage can be enhanced to achieve a robust system, taking into consideration the most reliable method of evaluation. We will focus on answering the following objective questions, to help identify gaps in this area of research.

1. What methods have been used to enhance the performance of this system?
2. What evaluation methods are reliable for assessing this system?

During the author's research, the aim was to find out how the performance of this system can be enhanced, what has been done, what should be done in future work, and to determine the most reliable and consistent evaluation parameter for assessing this system.

2.1 Speaker Identification System

A speaker recognition system is based on three stages: Analysis, Feature Extraction, and Classification techniques of the speech signal as shown in Fig. 2.1.

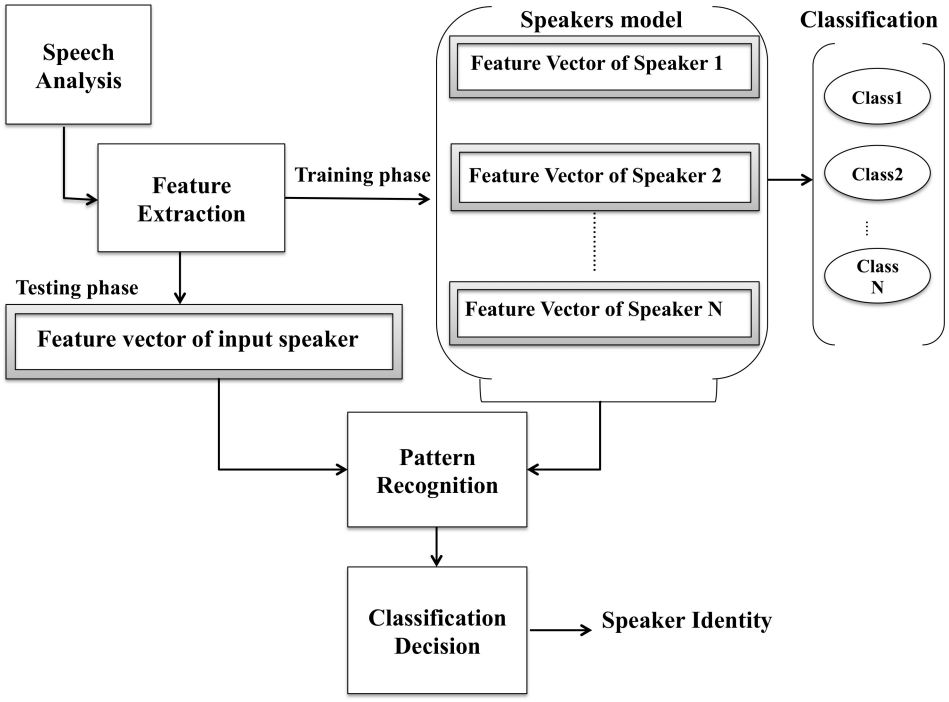

Figure 2.1: Speaker Identification System procedures.

2.1.1 Speech Analysis

A speech signal carries a large amount of data that include personal speaker's characteristics; however, a small amount is quite necessary where speaker characteristics could be recognized and easy for an algorithm to process. According to Gaikwad et al. (2010), individual speaker information included in speech signal is owing to the vocal tract mechanism. Thus, depending on the vocal tract mechanism that does not change faster than $10-20ms$, the speech signal sets in-group of frames; each of which has size duration in this interval. In the analysis stage, the result reflects meaningful information about the speech signal in order to extract features for speaker recognition (speaker identity). As reported by Singh and Pandey (2003) using a device, a microphone, the sound wave produced by human beings can be changed into an analog signal. Another device called Antialiasing filter can be used to determine the signal, and extra filtering is used in order to improve the deficiency of the channel. The Antialiasing Filter band reduces the signal of speech to almost the rate of Nyquist, which is equal to half the sampling rate. That takes place before the sampling

process. The analog signal determined is sampled by an analog –to-digital (A/D) adapter or converter so that a digital signal could be obtained. The currently used adaptors or converters for the signal of speech applications produce a resolution of 12 to 16 bits at 8000 to 20,000 samples in a second (Singh and Pandey, 2003). Nath and Kalita (2015) explain that this is all known as front-end processing; front-end processing changes the sampled signals of speech to a number of feature vectors that distinguish the properties of spoken words that allow identification of words and speakers. The front-end processing works in two ways: training and testing. The analysis stage shows that there is meaningful data about the signal of speech that helps reach the identity of speaker.

2.1.2 Feature Extraction

The extracting feature is a key step in the process prior to recognizing the pattern and the problem of machine learning. This is a technique for dimensionality reduction, which is used to reduce the large data processed by an algorithm. In this technique, the data given is changed into a number of features that lead to the related information in order to perform the required task without needing the full data size but through the reduced set. The output of the processing stage is a parameter vector that includes preliminary estimation of the signal. In theory, speech recognition is potentially obtained through the speech waveform. Yet, since speech varies in time, performing a form of feature extraction became real, and it is used in reducing the speaking signal variability. In the course of recognizing speech automatically, feature extraction retains the necessary information from the signal of speech while excluding the unnecessary information, and that brings about the analysis of speech signal. Even though the removal of unwanted information from the signal of speech may lead to the loss of some useful information, the main objective of feature extraction is to simplify the signal of speech and change it into recognizable sound components (Kurzekar et al., 2014). Gaikwad et al. (2010) depict the various feature extraction techniques: Band energies, formats and Cepstral coefficients that take the shape of a spectral feature and are derived from speaker vocal tract information. Pitch and its variation are excitation source features. For the behavior feature, the extraction technique is long term in duration and information. In actual fact, vocal tract based features such as LPCC and MFCC act as the most effective factors that led to the success of the speaker identification system. Both

features techniques work in the cepstral domain of the speech signal, represented by the cepstrum coefficients. The word cepstrum is derived from the reversal of the first four letters of the word spectrum; more details of cepstral analysis can be found in Chapter 4. Kurzekar et al. (2014), in particular, highlighted a significant comparison of different features extraction techniques in different applications using different spoken languages.

2.1.3 Classification

The last stage of the speaker recognition system is Classification, which is applied to distinguish speakers from each other. Arora and Singh (2012) claim that in addition to the artificial intelligence approach, the pattern recognition one has two methods: template and stochastic. Both approaches are two essential techniques used for solving classification problems specifically in speaker recognition system (op. cit). The aim of the classification process is to recognize the speaker depending on the feature matching techniques recognizer. That recognizer identifies the similarity between the input feature vectors and the constructed model (Yee and Ahmad, 2007). This process has two procedures called enrollment, such as the training phase, and verification such as the testing phase. The enrollment stage aims to build speaker's model contingent upon the features taken from the words produced by the speaker during the feature extraction stage. The resulting model from each speaker will be stored, as template database will eventually have *N* stored speaker models. The testing stage aims to authenticate the speaker. One of those speaker models will record speech, and transfer it to the pre-processing block to set a template based on the features. Afterwards, the experimented template will be examined against the database, and best score will be based on the fit one. The neural network is one way to use for training the database of the template; while examining, the neural network will recognize the original speaker (Dhameliya and Bhatt, 2015). Saksamudre et al. (2015) confirmed that using the artificial neural network technology has been effectively applied so as to rectify pattern classification issues. Many classification methods of speaker recognition are available in the literature, and their most commonly used ones are:

• Dynamic Time Warping

The Dynamic Time Warping (DTW) is a classifier used to determine the optimal alignment between two signals in time domains; to achieve this, one signal might be "warped" non-linearly through manipulations to the time axis. The optimized alignment can help to draw comparisons and similarities between the two signals; the alignment may also be used to discern if two waveforms share the same origin (phrase or speaker) by measuring the distance between the spoken phrase and one in the database, where a shorter distance in comparison indicates a greater degree of similarity. This minimum distance path may be determined through the use of dynamic programming (Salvador and Chan, 2007).

• Vector Quantization

In Vector Quantization (VQ), the training data is clustered to different codebooks; with each codebook representing a particular class, allowing for compression of the data by coding them in the codebook. Vectors are classified by determining the nearest codebook vector and measuring the distance to the test point, with the smallest distance indicating the identified user.

• Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is a weighted sum of Gaussian densities. Gaussian densities have parameters defined by the mean vectors and the covariance matrices. These parameters are approximated by the maximum likelihood standards. An optimal maximum likelihood may then be determined by the expectation maximization (EM) algorithm. During the training phase, each speaker is represented by a GMM; in the testing phase, the GMM for the speaker is compared to the established, speaker independent GMM models with the maximum a posteriori probability representing the client (Wan, 2003).

• Support Vector Machine

The Support Vector Machine (SVM) first creates a linear hyper-plane function and then works to maximize its margins. SVM may also work non-linearly through the use of kernel functions, in order to produce an optimal hyperplane. The optimal hyperplane is defined by support vectors, which are themselves defined as those samples with the smallest proximity to the hyperplane.

• Artificial Neural Networks

A Multi Layer Perceptron (MLP) is a multi-layer feed forward neural network model that transfers the input vectors to the corresponding output, with each node in the output being represented by a different class. During the training phase, the weight for every node is modified after each input data is received, contingent on minimizing the error in the output, relative to the target. In the output each node represents a class, with the number of nodes determined from the number of classes, so that the class of the feature input in the testing phase can be recognized (Agrawal et al., 2010). The RBFNN is the state of the art in artificial neural networks and is discussed further in chapter 4.

2.2 Analysis of the First Stage of SI System

2.2.1 Cepstral and Non-Cepstral Feature Methods

Current research in this field h as focused on comparing two systems and determining the best method of enhancing the particular stage. Baidwan and Gujral (2014), proposed the comparison of Linear Predictive Coefficients (LPC) a nd P rosodic features ($F_0, F_1, F_2,$ and F_3) with a Radial Basis Function Network (RBFN) classifier for recognition of a speaker from a sample population of 100 speakers. Results were evaluated on the basis of three parameters: accuracy, precision, and recall and showed prosodic features achieving better recognition than LPC on all three evaluation parameters that were used. Baidwan and Gujral determined prosodic features performed better than LPC because only four prosodic

features were used for recognition in contrast to 12 LPC features. Performance results were comparable between LPC and prosodic features but the latter was determined to be better due to a smaller vector dimensionality and being less computationally expensive. Baidwan and Gujral recommend that future studies should focus on a two-feature set using LPC and Prosodic features to achieve better recognition rate. The limitation of this study is that the compared results could be different in a noisy environment. These results have not been supported by all research however, as Singh et al. (2012) have found results counter to those from Baidwan and Gujral by showing that prosodic features alone require more speech samples, time consumption, and computational complexity, making them less effective because speech analysis in this case focuses on syllables of the spoken word.

Singh et al. (2012) focused on the differences between cepstral and non-cepstral feature extraction techniques. Their comparative results have presented the two feature extraction techniques for speaker recognition, Mel Frequency Cepstral Coefficient (MFCC) and Prosodic features. Accuracy was used as the sole evaluation parameter. From this they concluded in their analysis that MFCC is better than prosodic in speaker recognition to describe the signal characteristics, related to the speaker's vocal tract properties. Their assumption for these results is that short-term cepstral methods typically achieve better accuracy in the overall system since they replicate information about the physiology of the speaker, such as analysis of vocal tract characteristics, and do not require phonetic content. Prosodic features, in contrast, depend on excitation of the vocal tract and the phonetic content, such as the speaker's style, rhythmic, and intonational properties (fundamental frequency, pitch, intensity, and duration). Singh, Khan and Shree criticize using prosodic features due to impracticality as a result of the large amount of data needed to obtain effective recognition. Future research should focus on comparison between two different feature extraction techniques in the same domain.

Similar research performed by Kumar et al. (2011) compared MFCC and LPC features extraction for text dependent speaker identification with a Gaussian Mixture Model (GMM) implemented as the classifier. Their research takes into account the implications of both a clean and a noisy environment. The results have indicated recognition rate in a clear environment is 96.65% and 93.65% for MFCC and LPC respectively. It was verified that MFCC features also achieve better identification level in different Signal-to-Noise Ratio

(SNR) levels compared to LPC features. The reason they determined MFCC has better performance is because it is found in the cepstral domain, and is more resistant to noise than LPC. Features found in this domain are effective because they allow for separation of the excitation effect of the glottal from the effect of the vocal tract filter. In this researcher's point of view, the comparison should be between features evaluated in the same domain for a more reasonable evaluation. Farah and Shamim (2013) have also expanded upon this research through comparison in a speaker recognition task by using MFCC and LPC as feature extraction methods with VQ as the classification technique. The study focused on the comparison of MFCC with LPC in both text-dependent and text-independent speaker identification, considering the noisy environment and how these factors would affect the recognition rate. Results were presented that in both LPC's text-independent and MFCC's text-independent recognition, performance is lower than their text-dependent cases. Typically, the speaker recognition performance with MFCC and VQ is more accurate compared with LPC and VQ, with system accuracy decreasing for both in increasingly noisy environments.

Table 2.1: Summary of several research focuses on comparing cepstral and non-cepstral feature methods

Research	**Feature Extraction Approach**	**Classification Approach**	**Environment**	**Evaluation Approach**	**High Performance**
Kumar et al. (2011)	MFCC/ LPC	GMM	Clean/Noisy	Accuracy	MFCC
Singh et al. (2012)	MFCC/ Prosodic (F_0, F_1, F_2, F_3)	----------	Clean/Noisy	Accuracy	MFCC
Farah and Shamim (2013)	MFCC/ LPC	VQ	Clean/Noisy	Accuracy	MFCC
Baidwan and Gujral (2014)	LPC/ Prosodic (F_0, F_1, F_2, F_3)	RBFNN	Clean	Accuracy Precision & Recall	Prosodic

2.2.2 Combination of Feature Extraction Approaches

Similarly, Shinde and Pawar (2013) proposed a comparative study between two feature extraction methods, MFCC & LPC. Using Artificial Neural Network (ANN) classifier with Scaled Conjugate Gradient (SCG), a supervised learning algorithm; MFCC achieved 100% recognition rate, 50 epochs, and used 10 neurons while LPC achieved only 93.3% recognition rate, 30 epochs, and used 20 neurons. Additionally, using the two feature-extraction techniques by combining two features achieved the best performance in less time and with fewer neurons achieving a 100% recognition rate, 12 epochs, and 5 neurons.

Yujin et al. (2010) proposed a comparison of two features operating in the cepstral domain LPCC, MFCC, and a combination of both, for speaker recognition. Dynamic Time Warping (DTW) and (VQ) were applied as classifiers; research was performed in normal laboratory conditions, studying 40 participants. Accuracy was used as the sole evaluation parameter. Results showed MFCC performed on average 1% more accurately than LPCC, with results for both combined performing between 1 and 2% better than MFCC on measures of recognition rate and time.

Nath and Kalita (2015) investigated the influence of using combinations of two or three feature-extraction methods in the overall performance of speaker and speech recognition systems. Formants (*F1, F2, F3*), Linear Predictive Coefficients (LPC) and MFCC features were extracted as groups in cases of two-feature combinations, and one group containing all three, measured against all individually considered features for both systems. The LPC and MFCC combination is thought to be the best combination overall in regards to recognition rate. Nath and Kalita (2015) also say that the effect of combination of feature extraction has more improvement on speaker recognition task than on speech recognition, with accuracy rate near 100% for the former.

Table 2.2: Summary of several research focuses on combination of feature extraction approaches

Research	Feature Extraction Approach	Classification Approach	Environment	Evaluation Approach	High Performance
Yujin et al. (2010)	MFCC LPCC (LPCC+MFCC)	DTW/ VQ	Clean	Accuracy	MFCC+LPCC
Shinde and Pawar (2013)	MFCC LPC (LPC+MFCC)	ANN	Clean	Accuracy	MFCC+LPC
Nath and Kalita (2015)	MFCC LPC Formants (F1, F2, F3) (LPC+MFCC) (MFCC+ Formants) (LPC+ Formants) MFCC+LPC+ Formants	ANN	Clean	Accuracy	MFCC+LPC

2.3 Analysis of the Second Stage of SI System

2.3.1 Comparison of Different (ML) Algorithms

Agrawal et al. (2010) proposed the investigation of the four different machine-learning (ML) algorithms' performance including Multilayer Perceptron (MLP), Radial Basis Function Network (RBFN), C4.5 decision tree, and Bayes-Net for a speaker recognition task with an additional focus on gender differentiation; LPC features extraction method was applied. Results showed that MLP performs better for gender recognition while RBFN gives better performance in cases of increasing population size. Agrawal et al. (2010), through this investigation, indicated that the features extraction stage plays an essential part in increasing the classification rate and consequently improving SIS performance. From this, it was suggested that a combination using two methods of features extraction as a hybrid

feature set, such as a combination of prosodic features with MFCC or LPC with MFCC, would help to improve the recognition rate for speaker recognition task.

2.3.2 Performance Comparison of HMM, SVM, GMM and VQ

Weber and Du Preez (1993) focused on the comparison between two classifiers of speaker recognition task; Hidden Markov Model (HMM) and the VQ model. The speech dataset used in this comparison was recorded from a number of speakers over many weeks with recorded utterance modified for three different time intervals including 2,4 and 8 seconds. Some varieties of the respective models' configurations were also taken into account, including the VQ codebook of sizes 10, 16, 32, 64, and 128 as well as HMM's algorithm of 10, 16, 32, and 64 states, with difficulty executing in 128 states. System accuracy results found HMM performing better than VQ with recognition accuracy rate of 96.1% achieved in the case of 64 states and 8 second utterance, while the best accuracy of the VQ model was 93.1% on size 128 and 8 seconds. Results from former research byMatsui and Furui (1991) support the conclusion that the HMM algorithm performs better on measures of voice recognition than does VQ. Abdallah et al. (2012) proposed a comparative study providing support for Weber and Du Preez (1993), studying Text-Independent Speaker Identification System. 6 possible systems were created for comparison, using two feature extraction techniques, LPCC and MFCC, with three classifiers, HMM, Linear Discriminant Analysis (LDA) and MLP. Abdallah et al. (2012) conclude that MFCC features perform better than LPCC on measures of identification rates with all three classifiers. Their results have shown that the HMM classifier outperformed the other two classifiers, approximately 20% higher identification rates compared with MLP and around 30% compared with LDA. MFCC features performed better than LPCC when used with all 3 classifiers and, at the highest level of performance, were found to present better identification rates with 100%, compared with 94%. Abdallah et al. (2012) recommend that future research should focus on Support Vector Machine (SVM) classifier.

Related to the recommendation made by Abdallah et al. (2012), Yee and Ahmad (2007); Chandra and Kalaivani (2014) studied three different classifier techniques, with MFCC feature extraction, for speaker recognition task: Dynamic Time Warping (DTW), Gaussian

Mixture Model (GMM) and SVM. Yee and Ahmad (2007) state that for reasonable comparison of results, the same speech dataset, per-processing, and feature extractor method were used in their experiments. Although the accuracy rate of GMM model was reduced from increasing the training data, it still achieved better results compared with DTW and SVM. In addition, SVM was found to be the least effective and it is suggested that this is because SVM has shortcomings in which there are difficulties handling non-stationary signals such as speech signals, due to being restricted to work with fixed-length vectors. Both papers suggest that in future research, alternative ways of applying an SVM to fixed length vectors is to use an appropriate normalization function or kernel function capable of working with data in sequence to ensure that the SVM classifier achieves a higher accuracy rate.

Table 2.3: Summary of several research works on comparison of HMM, SVM, GMM and VQ

Research	Feature Extraction Approach	Classification Approach	Environment	Evaluation Approach	High Performance
Matsui and Furui (1991)	....	HMM/ VQ	Clean	Accuracy	HMM
Weber and Du Preez (1993)	.....	HMM/ VQ	Clean	Accuracy	HMM
Abdallah, et al. (2012)	MFCC LPCC	HMM/ MLP/ LDA	Clean	Accuracy	HMM
Yee and Ahmad	MFCC	DTW/GMM/ SVM	Clean	Accuracy	GMM
Kalaivani, et al. (2014)	MFCC	DTW/GMM/ SVM	Clean	Accuracy	GMM

2.3.3 Enhancement of the SVM Performance using RBF

The most recent research using SVM, performed by Nijhawan and Soni (2014), utilizes RBF kernel function to enhance the performance of SVM so as to determine an optimal

hyperplane separation line; this is important because failure to do this can lead to misclassification p roblems. The proposed MFCC-SVM model uses three different functions: Polynomial kernel functions (degree 2 and 3) and RBFNN kernel functions.

Results from the study indicated that RBF kernel function performs better than Polynomial kernel functions, observing that RBF kernel function deal with nonlinear separable case samples transforming into a higher dimensional space to be linear separable. The RBF kernel function was found to perform better than the polynomial kernel function because the former has fewer hyper-parameters and numerical difficulties. They suggest three directions for future research that could improve the performance and increase accuracy, eventually accomplishing a robust speaker identification system. First, neural network techniques should be used to enhance the system performance. Second, two different feature extraction techniques such as MFCC or LPC should be combined. Lastly, optimization algorithm should be applied to enhance the classifier performance.

2.3.4 Performance Comparison of RBFNN with other Classifiers

Past research has focused on RBFNN, comparing its performance relative to various other classifiers (Finan et al. 1996; Nijhawan and Soni 2012; Anifowose 2012; . Finan et al. (1996) compared the performance between MLP and RBFNN for verification and identification task of speaker recognition system. It was demonstrated that RBFNN offers better results than MLP for speaker recognition system in both tasks and has robust performance when faced with a poor training set compared with MLP. When used with a more efficient training set, RBFNN also extends a more robust improvement. Another piece of research, using the same two classifiers conducted by Nijhawan and Soni (2012) proposed performance comparison of two classifiers including MLP, using back-propagation (BP) algorithm, and RBFNN for the speaker recognition task. Three advantages were found using the RBFNN model: higher recognition rate, greater efficiency in recognizing numbers of speakers, and less computational time. RBFNN achieved 81.1% recognition rate compared to 78.1 % in BP in 50 epochs and 1400 epochs, respectively with 4 neurons in both input layers. RBFNN contained 7 neurons in 1 hidden layer, and BP contained 4 neurons also in 1 hidden layer. Anifowose (2012) proposed a similar comparative study involving two

different types of RBFNN, this time compared against GMM using the same training and testing datasets in a speech recognition task. Results from this study have shown both the traditional RBFNN and the DTREG type of RBFNN performing better than the GMM with traditional RBFNN providing the fastest recognition time of all three models.

Table 2.4: Summary of several research works on comparison of HMM, SVM, GMM and VQ

Research	Feature Extraction Approach	Classification Approach	Environment	Evaluation Approach	High Performance
Finan et al.1996	.....	MLP RBFNN	Clean	Accuracy	RBFNN
Agrawal et al. (2010)	LPC	MLP RBFNN	Clean	Accuracy Precision & Recall	MLP Gender Recognition RBFNN Speaker Recognition
Nijhawan and Soni, 2012	MFCC LPC	MLP RBFNN	Clean	Accuracy	RBFNN
Anifowose, (2012)	....	GMM RBFNN	Clean	Accuracy	RBFNN

2.3.5 Enhancement of RBFNN and MLP

All research explored in a review of the literature has provided support for the superior performance of RBFNN and MLP, taking all other relevant classifiers into consideration. Despite being considered the two most effective models for speaker recognition, RBFNN and MLP have some shortcomings that, when improved upon, will achieve a higher recognition rate and eventually enhance the recognition system. According to Al-Hadi et al. (2011), the major drawback of MLP is its slow convergence rate and its tendency to become stuck at the local minima. A proposed improvement in the MLP model has been found in the literature and has been mostly focused on optimizing the weights values of feed-forward neural network, which should achieve the target error without being stuck in the local solution. In their study, the Bacterial Foraging Optimization (BFO) and Particle Swarm Optimization (PSO) algorithms were compared; both algorithms have been used in feed-forward neural

network to optimize the learning process on measures of classification accuracy and rate of convergence. The results from this optimization show BFO algorithm to be vastly better than PSO in terms of both learning time and accuracy of classification. Zhou and Gu (2010) also outline weaknesses with the RBF network like the topology design and, more specifically, the difficulty determining the required number of hidden layer nodes, slow convergence speed, and lack of theoretical support for the setting of initial weight. They proposed in their research an optimization method for RBFN based on genetic algorithm to be applied to optimize the connected weights and topological design of RBFN; through this the repetitive nodes and connected weights are eliminated. This optimization algorithm is capable of global searching, allowing it to overcome RBFN's tendency to become stuck in a local solution. As a result, this proposed method based on genetic algorithm achieves improvement in terms of a fast learning speed and a high recognition rate.

2.4 Reliability Analysis of Evaluation Parameter

In the research, accuracy rate was the evaluation parameter used most often to assess the overall performance of SIS. The accuracy paradox is the problem with accuracy, which lies in the inability to determine false positives and false negatives in the results, stemming from the lack of distinction between class 1 and class 2, for example. Accuracy does not provide the researcher with the ability to determine the accuracy of just 1 or 2. Say, for instance, the class responsible for labeling positives identified no true positives while the class responsible for labeling negatives identified 2 700 true negatives on 3 000 total examples. Overall results for the system would show that there is a 90% accuracy rate, creating the appearance of an accurate system when it has, in fact, poor classification (Zheng, 2015). Future evaluation metrics for classification problems should be precisely chosen to correctly evaluate the system.

Chapter 3

SPEAKER RECOGNITION SYSTEM

Voice or speaker recognition refer to the automated method of identifying or confirming the identity of an individual based on his/her voice beware the difference between speaker recognition (identifying the speaking) and speech recognition (recognizing what has been said). This area of speech signal processing exploits the variability of speech model parameters across speakers. Reynolds (1995) states that scholars have developed growingly complex automatic speaker-recognition algorithms for 1960's by considering the change of the two principal tasks of speaker recognition: speaker identification (SI) and speaker verification (SV).

3.1 Speaker Identification (SI) vs. Speaker Verification (SV)

Speaker Identification is considered as an N-class decision classification problem since the idea is to model N-models as voiceprint for N numbers of speakers, in addition to comparing the speech signal being tested with those known models to identify each speaker. On the other hand, a Speaker verification system is the Binary-class decision classification problem. The idea of this system is to provide two models where one is known to the system called claimant and the other is unknown called impostor. The system will recognize the speaker known to the system, who claims to be from a large group of unknown voices (Reynolds, 1995).

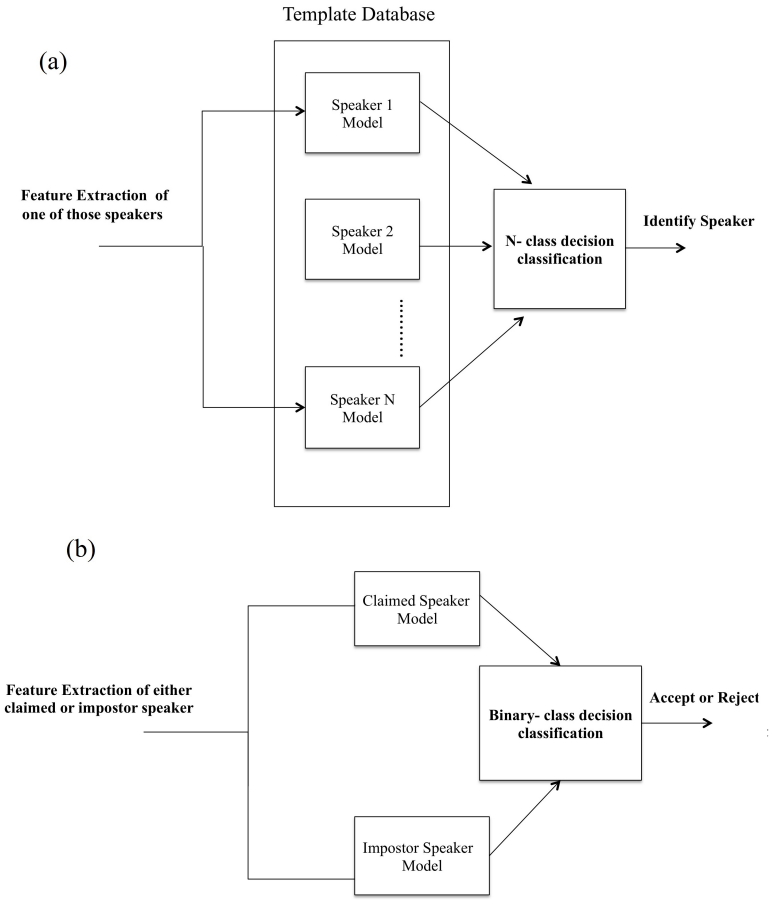

Figure 3.1: The two principal tasks of speaker recognition system: (a) speaker identification (SI). (b) speaker verification (SV) (Reynolds, 1995; Reynolds , 2002).

The speaker recognition task is categorized into two cases: a text-dependent system and a text-independent one, and that depends upon the spoken text used in those systems. In a text-dependent system called “text-constrained system”, the text has to be the same word or utterance for training and testing the system. While in a text-independent system called “text-unconstrained system”, there is no constraints for utterances that are used to train and test that system (Reynolds, 1995). The speaker identification task could be subdivided into two sets: closed and open. If the speaker to identify is registered, the recognition system is a closed- set. If they are not registered, then the issue will be named open. Generally, the open one is more difficult, while in the closed problem, a decision is enforced through the choice of the most compatible speaker from the database. Yet, with the open set, the system has to have a definition of a level of tolerance in advance in order that the degree

of similarity between the not known speaker and the most compatible one is through that tolerance level (Mezghani et al. 2010).

3.2 Application of Speaker Recognition

In reference to a new invention in the same field, Hanafi and Sukor (2012) highlight the contribution of biometric identification system which appeared as a new technology that relies on the biometric specific features of a person that differ from other people's. They emphasize that, for that reason it can be used to identify the speaker. Central to that, Anifowose (2012) states that voice, being unique to each individual as the most essential and natural method for communication, helps in different aspects of life; courts, businesses, law enforcement agencies and forensics. There are large demands in different fields for applications of speaker recognition system. Reynolds (2002) presented a variety of areas where speaker recognition technology has been applied. In general, law, medicine, and security fields widely use speaker recognition technology in different applications. Here is more information on the use of speaker recognition systems in different applications.

3.2.1 Security Services Application

Reynolds (2002) cited a number of examples where the speaker recognition system is widely used in security services. He mentioned that more recent applications such as computer networks and websites, have tended to use the speaker recognition system for access control. They have added biometric factors for creating passcodes, for example. They also used it to reset facilities or services. Another form of security service is the transaction authentication where the speaker recognition system is applied and used; bank account transactions, e-commerce and smartphone purchases are examples. Another example of using the speaker recognition system is personalization when restoring or saving personal settings for multi-user devices. As reported by Shahin (2010) the voice is used as the main proof to help the claim of the speaker in banking database services, security control for secret information, remote access to computers or tracking speakers, and for online transactions through phone calls.

3.2.2 Law Services

As for law services, Reynolds (2002) reported that law enforcement organizations have adopted a number of applications for monitoring homes such as call parolees at random times to ensure they are at home as well as prison call monitoring in order to authenticate inmate before the outbound call. There is a debate on using the automatic systems to back up oral spectral investigation of samples of voice for the purpose of forensic science (op.cit). In the same field of law services, Shahin (2010) stated that the speaker identification system can be utilized in police investigations with criminals to ensure the convicts that generated the recorded voice where the crime took place; he added that the system is usable for media civil cases such as calling radio stations, government offices and people services like insurance companies.

3.2.3 Medicine Services

Bahari (2014) indicated that the unique features of an individual's voice could be applied to diagnose, analyse and monitor some diseases such as Autism Spectrum Disorder (ASD) and Parkinson's disease in the field of medicine. Research by Oller et al. and Rektorova et al. (as cited in Bahari, 2014) has shown that extracted auditory features in a speaker's voice are helpful in identifying ASD and Parkinson's disease. In this respect, a study conducted by BioMetroSoft Company, voice analysis is applied to detect and monitor organic and neurologic disease development, rehabilitate speech therapy and patients with laryngitis. They proposed a voice analysis software known as BioMet®Engine through which the Glottal Source was extracted from the voiced speech that has a correlation with the dynamic pressure in the supraglottal of the vocal folds. That glottal source is used in diagnosing the larynx disease, and it can tell the conditions of the Neurological Paths that connects the larynx, hypothalamus and the speech neocortex.

In fact, understanding production and perception mechanisms of speech is a central concept for modeling the speaker recognition systems. The following sections attempt to shed light on the mechanisms of speech and mathematically review the speech production model.

3.3 Mechanism of Speech: The Process Of Human Speech

The mechanism of speech production and perception may be broken down into a few general processes including psychological, articulatory, and physiological processes (Cruttenden, 2014). Fig. 3.2 shows the essential process that include the speech production and perception mechanisms.

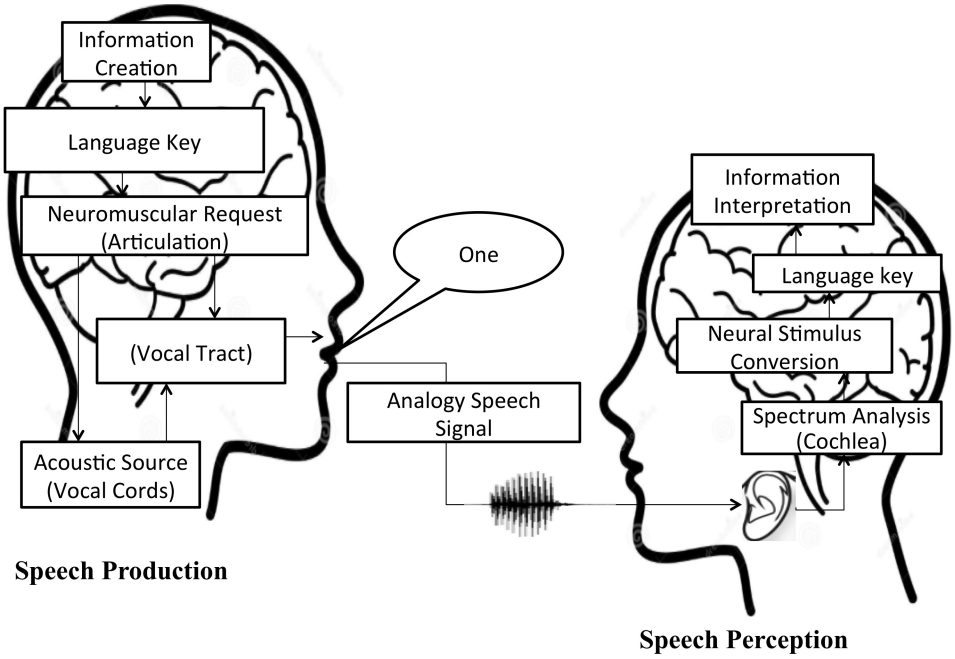

Figure 3.2: Schematic diagram of speech production and perception process.

3.3.1 The Process of Human Speech Production

In actual fact, speech production is the result of the reception of neuromuscular commands by the vocal cords, inducing vibration. Rabiner and Juang (1993) explained that the first step of speech production-speech-generation-starts in the human brain when the speaker formulates a message as a digital signal. That signal transmits speech to the listener. The message is converted into a language code; i.e. speaker's language or the wanted language. The language code is presented by a group of phoneme sequences considered as the smallest information unit in speech. Those phoneme sequences are equivalent to the

sounds forming the words. Each of those phonemes is formed by sending series of analog signals with duration, loudness and pitch of sounds appropriate to control muscles. The controlled muscles cause the vibration of vocal cords and shape vocal tract that creates the accurate speech sounds for a certain speaker. Then this acoustic signal transmits from the lips through the environmental air to the listeners' ears.

3.3.2 The Process of Human Speech Perception

The listener's ears are considered as receivers of incoming signals. The first step of speech perception process is the acoustic signal transformed by the auditory system in the inner ear - basilar membrane- resulting spectrum analysis of this signal. The spectral signal output of the basilar membrane is changed into activity signals through a neural transduction process. This process is approximately equivalent to a characteristic extraction process. In a complicated process that is not easily understood, that neural activity within the auditory nerve is transformed into a language code in high units in the brain and finally message comprehension is achieved (Rabiner and Juang, 1993).

3.4 Speech Production Model

Basically, human speech production is divided into three main parts: the lungs, larynx, and vocal tract, shown in Fig. 3.4. The lungs are a source functioning as a power supply that provides airflow to the larynx, with vocal cords that modulate the airflow produced from the stage when lungs work. Subsequently, it provides either a periodic vowel sound or a noisy constant one that transmits to the vocal tract. The vocal tract works as a filter w ith oral, nasal, and pharynx cavities. Those cavities spectrally shape the airflow s ource. At the lips, differences of air pressure take the form of aired sounds waves received and understood as speech by listeners (Quatieri, 2006, p. 55).

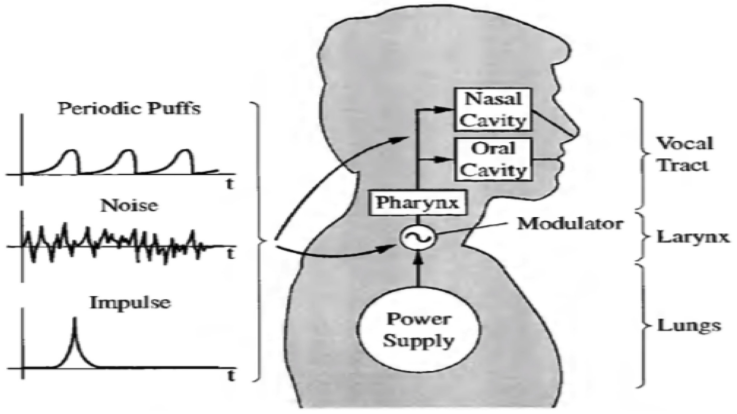

Figure 3.3: Illustrative picture of speech production. The sources of the sounds are labeled periodic, impulsive and noise that can happen in the larynx or vocal tract (Quatieri, 2006, p. 56).

Speech starts when the air enters the lungs through normal breathing. This air functions as a periodic signal that causes the vibration of the vocal cords, which gives the pitch of the sound or the so called fundamental frequency. In fact, this periodic signal is perceived as an excitation input passing through to the vocal tract or resonant cavity whose length starts from vocal cords through lips including a group of natural filters: pharynx, oral and nasal cavity. As can be seen in Fig. 3.4 The vocal tract or the resonant cavity works as a resonator that spectrally shapes a periodic input signal. Hence, from this fundamental perception of the speech production mechanism, a simple engineering model can be designed (Quatieri, 2006, p.4).

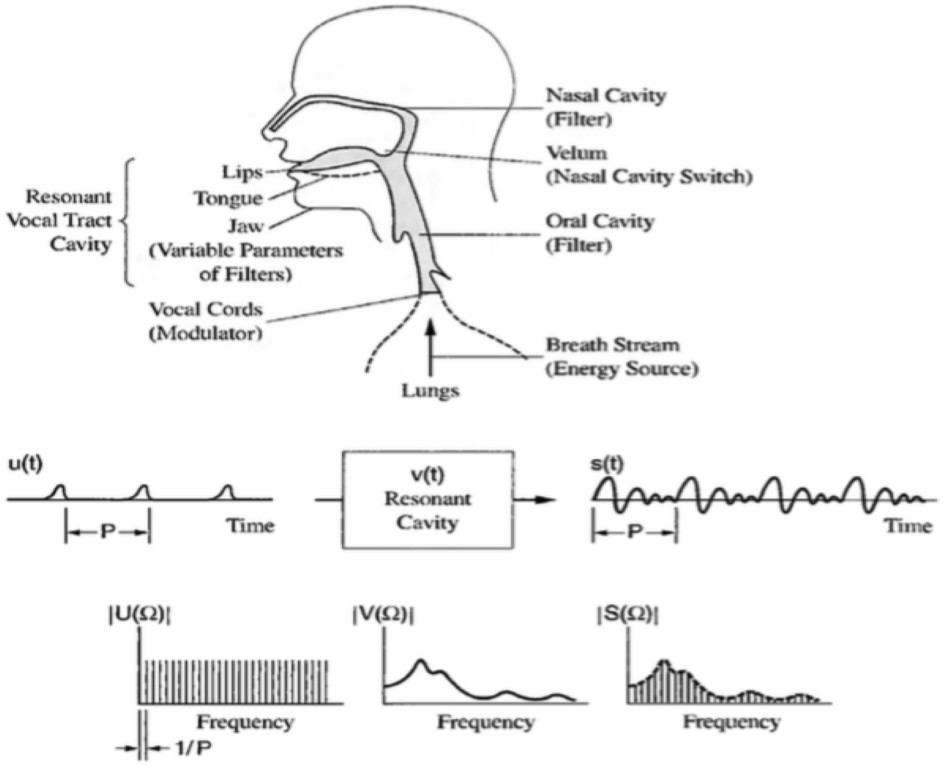

Figure 3.4: Speech production pattern in humans.(Quatieri, 2006).

The importance of the responses of vocal tract is to distinguish one person from another since each one has their own transferring function of vocal tract. This signal of transfer function has a group of peaks that indicate the fundamental frequency F_0 and resonant frequencies $F_1F_2F_3$. In fact, the transfer function of vocal tract is considered as a feature that helps us recognize the speech as well as the speaker (Quatieri, 2006, p68-69). Accordingly, there was a necessity to study the mechanism of vocal tract as well as the attempt to design a mathematical model. That model is supposed to be based on simulating the vocal tract. To design this model is the foundation of building the recognition of speech system and the distinction between speakers' voices.

3.4.1 Larynx Stage Model

The speech production happens through the control over vocal cords. In speaking, the glottis, which is the narrow passage between the two cords that turns to be small as a result

of vocal cords vibration using muscles during speech. There are three basic situations of vocal cords: breathing, voiced, and voiceless. At glottis, measuring the airflow velocity as time function, gives us approximately a periodic signal describing three stages as a time-varying system. These three stages can be seen in Fig. 3.5. The first stage is the time interval; when the vocal cords are closed, i.e. no airflow, is called “the glottal closed phase”. Then, the time interval where the vocal cords open and reach the climax of airflow velocity that is called “the glottal open phase”. The last stage is the time interval when the vocal cords sharply reduced to zero, and it is called “the glottal return phase”. Based on speaker style and speech, the shape of airflow is changing.

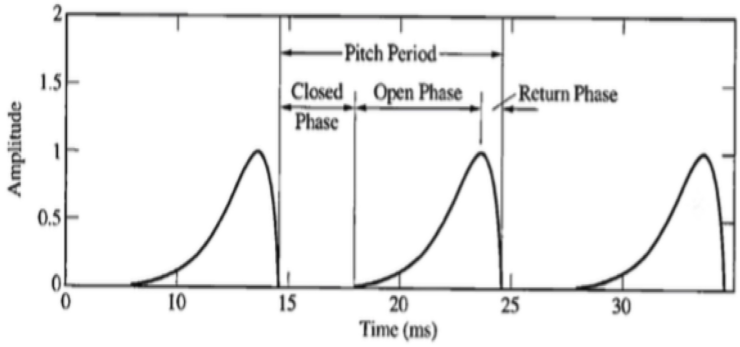

Figure 3.5: Illustration of periodic glottal airflow velocity (Quatieri, 2006, p.62).

In this system, the pitch period is the time interval in one glottal cycle. In fact, the pitch period change depends on the vocal cords muscle tension. The relationship between tension and pitch is exponential, i.e. when tension increases, the pitch does. Whereas the relationship between vocal cords mass and pitch is the inverse. In other words, when the vocal cords mass decreases, the pitch increases (Quatieri, 2006). Typically, vocal cords vibrate at fundamental frequency F_0 of the sound, which can be categorized as in male speakers 50-200 Hz, while in female speakers it is 150-200 Hz while in a child it is 200-400 Hz (Ambikairajah, 2010). According to Quatieri (2006, p. 62), the excitation process is the basic mathematical framework of glottal flow and is the convolution of a periodic impulse train with the waveform of glottal airflow is given by :

$$u[n] = g[n] * e[n] \tag{3.1}$$

$$e[n] = \sum_{n=-\infty}^{\infty} \delta[n - kp] \tag{3.2}$$

Where: $e[n]$ is impulse train spacing by pitch period p and k= 0,1,2...

3.4.2 Vocal Tract Stage Model

According to Quatieri (2006, p. 66-68), this stage is an acoustic tube that includes the nasal and the oral cavity through the larynx to the lips. The velum is a valve between the two cavities. The vocal tract has many different lengths and many cross-sections related to the positions of jaw, tongue, and lips. It also has the average length, around $17cm$ in an adult male and less in a female. In fact, the output of vocal cords is pressure wave entering the vocal tract. The vocal tract has two main purposes; it is a new source of sound production, and it spectrally colors the pressure wave, which is important in making distinction between the speech sounds. The vocal tract configuration with jaw, tongue, lips change is contingent upon different phonemes resulting resonance frequencies called formants $F_1F_2F_3$. These formants can be clearly determined by the peaks of sound spectrum in Frequency Domain (FD), and it is essential in speech and speaker recognition as well as the emotional status of the person. The vocal tract formants are organized from low to high frequencies. Those frequencies increase as the length of vocal tract decrease; thus, the male speaker seems to have lower formants compared to female; whereas the child appears to have more formants than the females do. Since the supposition of vocal tract, time-invariant all-pole linear system, has a sound source from the glottis as an input, and it has the system response, known as the vocal tract, $h[n]$. Therefore, the output of vocal tract is speech sound giving by:

$$x[n] = h[n] * (g[n] * e[n]) \tag{3.3}$$

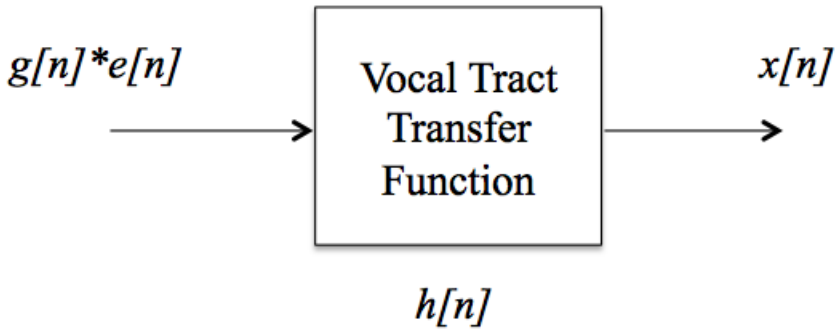

Figure 3.6: Vocal tract system.

Premised on the human production of speech, it is as necessary to create a digital filter that has a transfer function, which illustrates the way of how the system responds to inputs. Finding the coefficients of that digital filter can be used to emulate the vocal tract response.

3.5 Mathematics of Overall Speech Production Process

As reported by Ambikairajah (2010), the voiced and unvoiced sounds will be generated from being impulse train at the Excitation source. This shows the first stage of that process. In reality, the stimulation of the voiced speech is a sequence of impulses, while the stimulation to produce unvoiced speech takes the shape of random noise. Applying the Z-transform to equation 3.2 produces.

$$E(z) = Z\{e(n)\} \tag{3.4}$$

$$E(z) = \sum_{n=-\infty}^{\infty} e(n)z^{-n} \tag{3.5}$$

$$E(z) = 1 + z^{-p} + z^{-2p} + \ldots \tag{3.6}$$

$$E(z) = \frac{1}{1 - z^{-p}} \tag{3.7}$$

The formulated pulse resulted from the glottis acts as a low pass filter and is thought to be the following stage of excitation process used for voiced speech; thus, the glottal model with transfer function is given as in equation 3.8. In instances of unvoiced speech, the transfer function $G(z) = 1$.

$$G(z) = \frac{1}{(1 - z^{-1})^2} \tag{3.8}$$

The amplitude tuned by a gain factor and the result of excitation stage is given

$$u[n] = Ae[n] * g[n] \tag{3.9}$$

$$U(z) = AE(z)G(z) \tag{3.10}$$

The vocal tract model is seen as an all-pole model with transfer functions, which represent the digital filter function that contains resonant frequencies; and each formant has to have two poles.

$$V(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \quad (p - \text{order filter}) \tag{3.11}$$

Lip radiation can be modeled as a high pass filter using the following equation:

$$R(z) = 1 - 0.98z^{-1} \approx 1 - z^{-1} \tag{3.12}$$

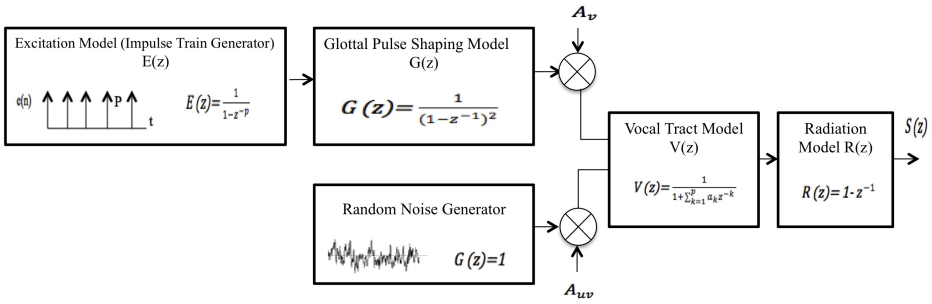

Figure 3.7: Speech production model Ambikairajah (2010).

The result of the speech signal is represented by:

$$S(z) = AE(z)G(z)V(z)R(z) \tag{3.13}$$

The following transfer function can be used to represent the speech production model:

$$\frac{S(z)}{E(z)} = AG(z)V(z)R(z) \tag{3.14}$$

The transfer functions for voiced and unvoiced speech are given by:

$$\frac{S(z)}{E(z)} = \frac{A_v}{1+\sum_{k=1}^{p+1} a_k z^{-k}} \tag{3.15}$$

$$\frac{S(z)}{E(z)} = \frac{A_{uv}}{1+\sum_{k=1}^{p+2l+2} a_k z^{-k}} \tag{3.16}$$

Transfer functions for both voiced and unvoiced speech are modeled by the equation $H(z)$:

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1+\sum_{k=1}^{q} a_k z^{-k}} \tag{3.17}$$

Where: $q = 10, 12,$ or 14 poles, with $q = 12$ providing sufficient information processing for modeling voiced and unvoiced speech.

a_k : Digital filter coefficients of the vocal tract; speakers will generally have distinct coefficients from one another.

Chapter 4

SPEAKER IDENTIFICATION SYSTEM

The preceding chapter examined research conducted on speaker recognition systems. A decision was made on the basis of this research to develop an SIS to include LPCC and MFCC as feature extraction techniques, RBFNN as a classifier, and to apply BFOA as a way of enhancing the classifier. The following sections describe the concepts of LPCC, MFCC, RBFNN and BFOA. The evaluation parameters selected are also described.

4.1 Feature Extraction Techniques

4.1.1 Linear Predictive Cepstral Coefficients (LPCC)

• The LPCC principle

The LPCC principle is designed upon the model of human speech production. It applies transfer functions of the glottal, vocal tract, and lips' filters into one all-pole filter that simulates the vocal tract (Shrawankar and Thakare, 2013).

$$\frac{E(z)}{U(z)} = \frac{G}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{4.1}$$

where, a_k : Digital filter coefficients of vocal tract.

Essentially, the idea of LPCC is to minimize the squared error between the actual speech-sample *s(n)* and the estimated signal $s^{\wedge\wedge}n)$ within $20ms$ frame size. This linear combination of the past p speech samples will give a set of predictor coefficients α_k. (Shrawankar and

Thakare, 2013). Ambikairajah (2010) note that applying the inverse z- transform to equation 4.1

$$s(n) = \sum_{k=1}^{p} a_k S[n-k] + Gu[n] \tag{4.2}$$

Where, $s(n)$: Current speech sample

a_k: The actual digital filter model coefficient of vocal tract, prior to calculation.

$s[n-k]$: Past sample

Assuming $u[n]$= 0, the estimated speech signal is given by:

$$s^{\wedge}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{4.3}$$

Where, $s^{\wedge}(n)$: estimated speech (current speech)

α_k :the desired predicted coefficient.

From this, the prediction error for a finite duration is given as:

$$e(n) = s(n) - s^{\wedge}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{4.4}$$

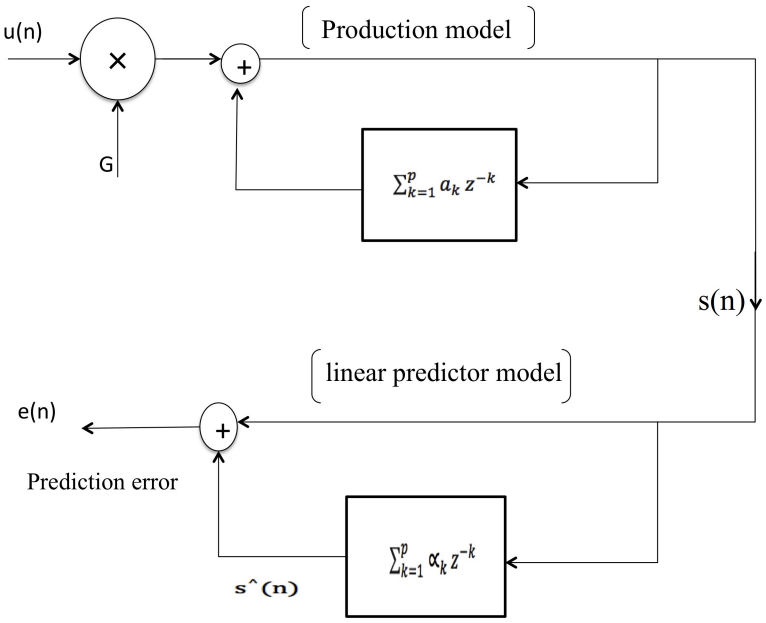

Figure 4.1: The idea of LPCC (Ambikairajah, 2010).

The vocal tract production model contains an actual coefficient a_k for a specific frame. The output of this production model speech signal $s(n)$ serves as the input for the prediction model, containing the predicted coefficient α_k , thus the optimal prediction is when $a_k = \alpha_k$.The Infinite Impulse Response (IIF filter) is a mathematical representation of the vocal tract model (all pole filter) and is analyzed by the transfer function:

$$\frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{4.5}$$

Similarly, the Finite Impulse Response (FIR filter) is a mathematical representation of the prediction model, also known as the all zero filter or the vocal tract model, and is analyzed by the transfer function:

$$\frac{E(z)}{U(z)} = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{4.6}$$

In case of optimal prediction: $a_k = \alpha_k$

$$\frac{E(z)}{U(z)} = G \tag{4.7}$$

• LPCC Procedures

The processor of LPCC can be summarized in six steps seen in the block diagram of Fig. 4.2:

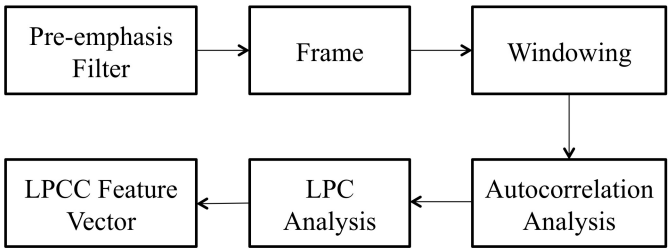

Figure 4.2: Procedures of LPCC.

• Pre-emphasis

From the transfer function of the vocal tract, it can be seen that the glottis has the effect of making the prediction of coefficients not isolated to that of the vocal tract. From the speech production model obtained by the previous equation 3.14.

$$\frac{S(z)}{E(z)} = AG(z)V(z)R(z)$$

$$= \left(\frac{1}{\left(1-z^{-1}\right)^2}\right)\left(\frac{1}{1+\sum_{k=1}^{p}\alpha_k z^{-k}}\right)\left(1-z^{-1}\right) \approx \left(\frac{1}{1-z^{-1}}\right)\left(\frac{1}{1+\sum_{k=1}^{p}\alpha_k z^{-k}}\right)$$

Therefore, the pre-emphasis filter serves to counter the glottal effect from the transfer function of the vocal tract. Countering of this effect can also be achieved by implementing high pass filtering, denoted either by difference equation or transfer function (Ambikairajah, 2010).

$$h(n) = s(n) - as(n-1) \tag{4.8}$$

$$H(z) = 1 - 0.99z^{-1} \approx 1 - z^{-1} \tag{4.9}$$

• Framing

In order to determine these predictor coefficients at a given time n of the speech signal, it must be estimated from a short segment of this signal. Therefore, the basic idea for finding this set of coefficients that minimize the mean squared error $e(n)^2$ is to carry out frame-by-frame examination of speech with $10-20ms$ frame size until we have all speech samples (Ambikairajah, 2010). According to Quatieri (2006), the mean-square production error is shown by:

$$E = m = \sum_{m=-\infty}^{\infty} e(m)^2 = \sum_{m=-\infty}^{\infty} \big(s(m) - s^{\wedge}(m)\big)^2 \tag{4.10}$$

We have to work on short time sequence of speech signal $s_n(m)$ with time interval $[n-M, n+M]$, so we can rewrite the equation 4.10.

$$E_n = \sum_{m=n-M}^{n+M} e_n^2(m) \tag{4.11}$$

From equation 4.3 we can obtain:

$$e_n^2(m) = \begin{cases} \left[s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k)\right]^2 & n-M \leq m \leq n+M \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

α_k minimizes the mean-square error and can be obtained by differentiation of the total error E with respect to $\alpha_1, \alpha_2 \ldots$

$$\frac{\partial E}{\partial \alpha_i} = 0, \ i = 1, 2, 3, \ldots p \tag{4.13}$$

Where: p indicates the order of LPC analysis

$$\begin{aligned} \frac{\partial E}{\partial \alpha_i} &= \frac{\partial}{\partial \alpha_i} \sum_{m=-\infty}^{\infty} \left[s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \right]^2 \\ &= 2 \sum_{m=-\infty}^{\infty} \left[s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \right] \left[-\frac{\partial}{\partial \alpha_i} \sum_{k=1}^{p} \alpha_k s_n(m-k) \right] \end{aligned} \tag{4.14}$$

Note that:

$\frac{\partial}{\partial \alpha_i} \sum_{k=1}^{p} \alpha_k s_n(m-k) = -s_n(m-i)$

since $\alpha_k s_n(m-k)$is constant with respect to i when $i \neq k$

From equation 4.13 we obtain:

$\frac{\partial E}{\partial \alpha_i} = 2 \sum_{m=-\infty}^{\infty} \left[s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \right] \left[-s_n((m-i)) \right] = 0$

$= \sum_{m=-\infty}^{\infty} s_n((m-i) s_n(m)) = \sum_{k=1}^{p} \alpha_k \sum_{m=-\infty}^{\infty} s_n(m-i) s_n(m-k)$

The summation of two outputs with different shifts is referred to as a correlation, which can be expressed as :

$\Phi_n[i,k] = \sum_{m=-\infty}^{\infty} s_n(m-i) s_n(m-k)$

is correlation where $1 \leq i \leq P,\ 1 \leq k \leq P$

$\Phi[i,0] = \sum_{m=-\infty}^{\infty} s_n(m-i) s_n(m)$

From this we develop the equation

$$\sum_{k=1}^{p} \alpha_k \Phi_n[i,k] = \Phi_n[i,0],\ i = 1,2,3,\ldots p \tag{4.15}$$

This equation is referred to as the matrix form of the normal equation:

$\Phi \alpha = b$

The former equation presents Φ as a matrix $(p \times p)$, b as a vector $(1 \times p)$ represented by $\Phi_n[i,0]$, and vector α as α_i (Quatieri, 2006).

• **Windowing**

Upon blocking of the frame and to reduce the disruption of the signal at either the beginning or end of the frame, each frame must be windowed (Wijoyo et al. 2011).Quatieri (2006) states that in order to apply the autocorrelation method, we should assume that the samples outside the time interval $[n-M, n+M]$ or our windowing size $N_w = 2M+1$ are equal to zero, in our output signal $s[m]$ starting at n- time and ending at $n+N_w-1$. In other words, we are shifting the output signal by n samples and windowing by size N_w window given by $s[m] = s[m+n]w[m]$. The resulting segment we used as two signals to be correlated from above 4.15.

$\sum_{k=1}^{p} \alpha_k \Phi_n[i,k] = \Phi_n[i,0],\ i = 1,2,3,\ldots p$

The time interval with window is $(N_w + n - 1)$

$\Phi_n[i,k] = \sum_{m=0}^{N_w+n-1} s_n(m-i)\, s_n(m-k)\ 1 \le i \le P,\ 1 \le k \le P$

The overlapping regions of both signals are the only determinants of $\Phi_n[i,k]$ with interval $[i, k+N_w-1]$

$$\Phi_n[i,k] = \sum_{m=i}^{k+N_w-1} s_n(m-i)\, s_n(m-k) \tag{4.16}$$

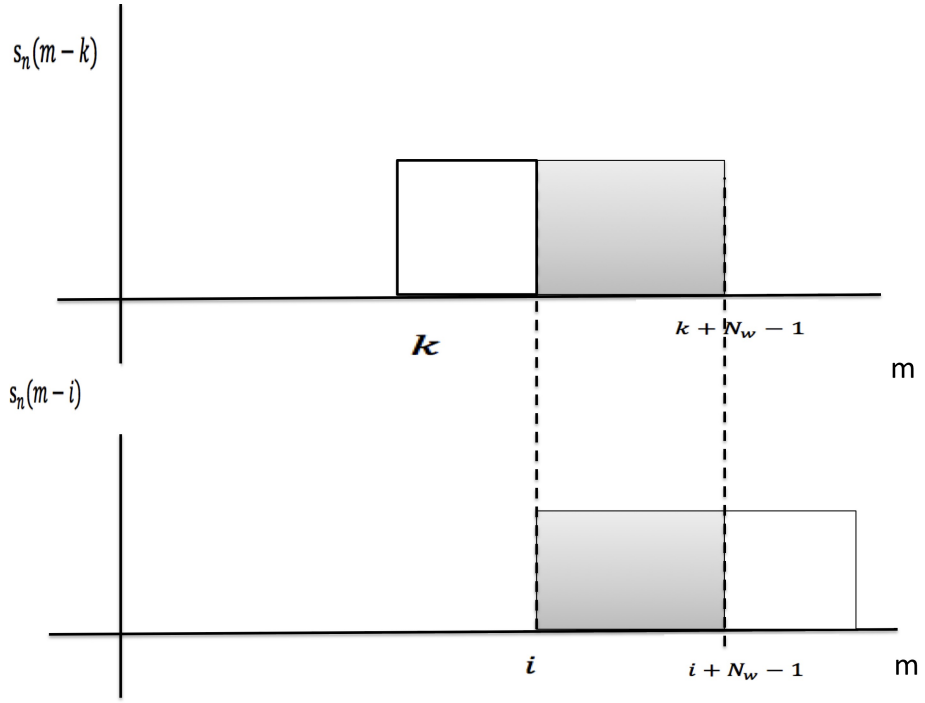

Figure 4.3: Autocorrelation Method with overlapping regions Quatieri (2006).

By canceling the shift of $s_n(m-i)$ by adding $(-i)$, we can rewrite the equation:

$$\Phi_n[i,k] = \sum_{m=0}^{N_w-1-(i-k)} [s_n(m)\, s_n(m+(i-k))] \tag{4.17}$$

• Autocorrelation Method

The function $\Phi_n[i,k]$ in the autocorrelation function 4.17 can be rewritten as the difference between i and $k(i-k)$

$r_n[i-k] = \Phi_n(i,k)$, where $i-k=\tau$

$r_n[\tau] = \Phi_n(i,k)$

$r_n[\tau] = \sum_{m=0}^{N_w-1-\tau} [s_n[m]\, s_n[m+\tau]]$

$= s_n[\tau] * s_n[-\tau]$

By letting $\Phi_n(i,k) = r_n[i-k]$ we can rewrite equation 4.15 as:

$\sum_{k=1}^{p} \alpha_k r_n[i-k] = r_n[i-0] \quad 1 \leq i \leq p$

$\sum_{k=1}^{p} \alpha_k r_n[i-k] = r_n[i] \quad 1 \leq i \leq p$

Presenting the linear equation

$R_n \alpha = r_n$

The matrix below indicates the simultaneous equations. The matrix R_n is known as a Toeplitz Matrix, which is symmetric about the diagonal, and all diagonal elements are equal.

$$\begin{bmatrix} r_n[0] & r_n[1] & r_n[2] & \cdots & r_n[p-1] \\ r_n[1] & r_n[0] & r_n[1] & \cdots & r_n[p-2] \\ r_n[2] & r_n[1] & r_n[0] & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ r_n[p-1] & r_n[p-2] & r_n[p-3] & \cdots & r_n[0] \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \alpha_p \end{bmatrix} = \begin{bmatrix} r_n[1] \\ r_n[2] \\ \cdot \\ \cdot \\ r_n[p] \end{bmatrix}$$

• LPC Analysis

In order to solve these equations, known as the Toeplitz Matrix system, and to find the predicted coefficients $\alpha_1, \alpha_2 \ldots \alpha_p$, the Durbin's algorithm, also known Levinson recursion, is one method of solving this system. LPC analysis is the next step in processing and uses Durbin's method as a way of transforming each frame of p autocorrelations into an LPC parameter set (Wijoyo et al. 2011). According to Quatieri (2006), the following steps are shown in the Levinson recursion algorithm:

1. Set all initial predicted coefficients and predicted error of the transfer function to zero. $\alpha_0^0 = 0, E^0 = r_n[0], i = 0,1,2,...p$
2. Partial correlation coefficients (PARCOR) k_i are calculated by the following equation $k_i = \left(r[i] - \sum_{j=1}^{i-1} \alpha_j^{i-1} r[i-j]\right) / E^{i-1}$

3. Where, $r[i]$: Actual correlation$r[i-j]$: The shifted correlation E^{i-1}: The previous energy. Set $\propto$ equal to the partial correlation coefficient, $\alpha_j^i = k_i$; $\alpha_j^i = \alpha_j^{i-1} - k_i\alpha_j^{i-1}$, $1 \leq j \leq i-1$

4. The minimum mean-squared prediction error is recalculated by the equation: $E^i = (1 - k_i^2)E^{i-1}$

5. Repeat the above steps until p to obtain optimal coefficients for minimizing error $\alpha^* =_j \alpha_j^p, 1 \leq j \leq p$

P is the order of the LPC analysis. LPC is then converted to the Cepstral domain by applying the log of the power spectrum before the IFFT, to obtain the LPCC coefficient (Wijoyo et al. 2011). To convert the LPC to LPCC, the recursive method is applied as defined Rabiner and Juang (1993):

$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k},\ 1 \leq m \leq p$

$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k,\ m>p}$

Where; a_m : LPC coefficients.

k_m : Partial correlation coefficients (PARCOR).

c_m : Cepstral Coefficients.

P : LPC order.

4.1.2 Mel Frequency Cepstral Coefficients (MFCC)

• The MFCC principle

According to Dave (2013), MFCC is one of the most common methods of feature extraction that has been used in speaker recognition in the frequency domain; it utilizes the Mel scale whose concept depends on the human ear scale. MFCC simulates the model of the human auditory system in which human hearing perception has better recognition in low

frequencies. Human hearing perception has better recognition in low frequencies because the hearing system acts as a group of filters that are more sensitive in low frequencies (around 1KHz) and less in high frequencies. Shinde and Pawar (2013) explain that these filters, known as mel-scale filterbanks, are linear when the frequency is lower, below 1 KHz, and logarithmic when the frequency is higher, above 1KHz. Speech understanding by the hearing system can be simulated by filters that have center frequencies corresponding to the critical band filters. Fig. 4.4 is an example of a mel-scale filterbank, cited by Quatieri (2006) as having been introduced in past research by Davies and Mermelstein. The filterbank uses 24 triangular filters to simulate the critical band filters that encompass a 4KHz range. Similar to the above filters discussed by Shinde and Pawar (2013), these filters are linear when the frequency is lower and exponential at higher frequencies.

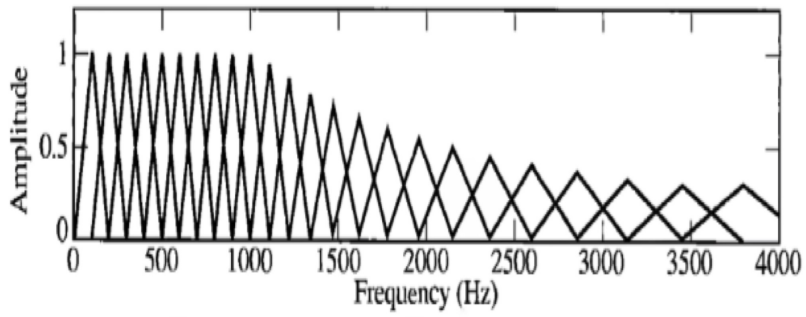

Figure 4.4: Davies and Mermelstein's Mel-scale filter-bank simulating critical band filters (Quatieri, 2006).

• MFCC procedure

MFCC arises from the combination of the advantages of both cepstrum analysis and a perceptual frequency scale, based on critical bands (Yankayiş). This process can be summarized through the block diagram of Fig. 4.5 where the pre-emphasis, frame and windowing steps are the same as mentioned in the LPC process.

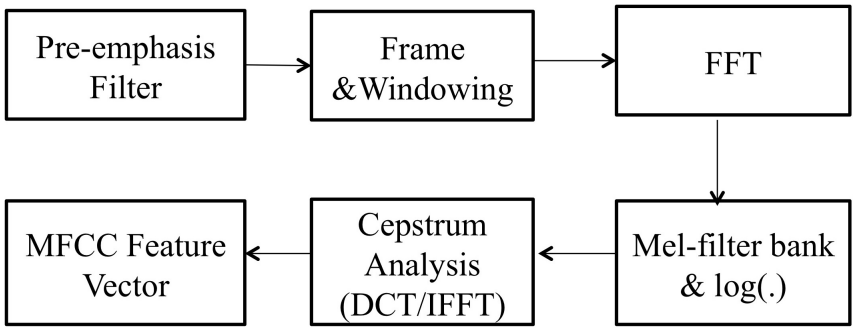

Figure 4.5: Procedures of MFCC.

• Fast Fourier Transform

After applying short time analysis, Fast Fourier Transform (FFT) for each window is applied to allow conversion of each frame into the frequency domain from the time domain; upon conversion from the time domain, the convolution process of the glottal pulse and the vocal tract impulse will be transformed into a multiplication function in the frequency domain (Yankayiş). This conversion into the frequency domain allows use of a spectrum for representation, where the formants are indicated by the peaks of spectral envelope. The spectrum involves low frequency regions representing the spectral envelope, which is known to be an important feature for speaker recognition. The high frequency regions here represent the spectral details (Prahallad, 2011).

• Mel filter-bank

In agreement with perceptual observation, the human ear relies on certain ranges instead of using the entirety of the spectral envelope. Mel frequency analysis simulates human speech perception, i.e. human ear mechanics acting as a group of filters as seen in F ig. 4 .4. As a result of applying Mel filter-bank to the spectral envelope, the Mel-spectrum is formed. The following equation is used to calculate the relationship between the Mel frequency and a given frequency (Yujin et al. 2010).

$M = 2595 log_{10}(1 + f/700)$

• Cepstral Analysis

Cepstral analysis refers to taking the Log of the spectral envelope, producing the Mel-Frequency Cepstral Coefficients (MFCC) (Prahallad, 2011). The multiplication process of the glottal pulse and the vocal tract impulse will be transformed into a summation function in the quefrency domain. Low-Pass Filter (LPF) can be applied to the spectral envelope to get the vocal tract features and exclude the excitation effect. In the final step of cepstrum analysis, the cepstrum is reconverted to a standard time scale by and then applying IFFT. The resulting spectrum offers a better representation of spectral information for the sig-nal, allowing for recognition and representation of speaker characteristics (Kurzekar et al. 2014).

4.2 Radial Basis Function Neural Network (RBFNN)

4.2.1 Introduction

A Radial Basis Function (RBF) is a developed type of neural network. Proposed in the 1980s by Moondy and Darken, RBFNN was introduced with the intention of improving MLP's neural network performance (Wang and Xu, 2013). TIAN et al. (2007); Svozil et al.(1997) have identified MLP back propagation (BP) learning algorithm as having several flaws; problems include long learning time because of the large number of weights, and slow converging speed stemming from no use of physical knowledge in developing an ap-proximating mapping of parameters. Additionally, the function of hidden layers' neurons cannot often be interpreted which can lead to difficulty explaining the prediction and ulti-mately less accuracy through the training process. It may then be considered that RBFNN is a powerful solution to MLP; more specifically, past research has found that RBFNN performs better than the MLP-BP network in function approximation and classification problems in solving nonlinear problems (Wang and Xu, 2013).

4.2.2 Architecture of RBFNN

Radial Basis Function Network includes an input layer, a hidden layer and an output layer as can be seen in figure.

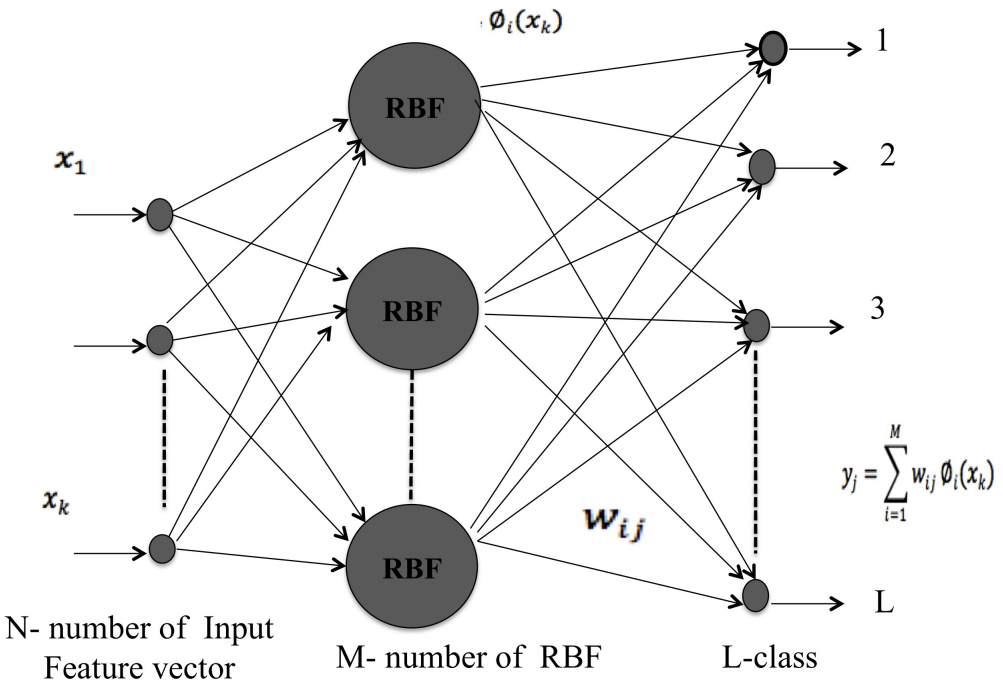

Figure 4.6: RBFNN Architecture.

The input layer is comprised of N-number of nodes, where N is dependent on the N-dimensional input vector that is presented. The hidden layer includes M-number of nodes with each node computing its RBF, where the main function of the hidden layer nodes is to transfer the input data space to higher dimension space. It is important to note here that M must be much greater than N in order for this operation to function properly. Increasing the dimension of the features vector from an N-dimension space to a higher dimension space serves to allow a features vector to become linearly separable where this was previously not possible. After achieving linear separability of the feature vector, the linear computation of the hidden layer output determined by the weight vector and performed by the nodes of the output layer, becomes more likely to give a class belonging to each node in the output layer. The output layer contains L-number of nodes, directly corresponding to the number

of classes; based on the value of the previously mentioned linear computation, the output layer nodes will then determine the appropriate class for the input vector; if feature n belongs to *j* class, the output of the respective node will have its highest relative value equal to 1 and the other output nodes will have their lowest relative values equal to 0 (Biswas, 2014).

4.2.3 RBFN Training

As reported by Anifowose (2012), in the hidden layer nodes each node performs RBF, which serves to transmit the input vector to the hidden space. The activation function of hidden nodes depends on the Euclidean distance between the input vector and the center of each node, while the output of each hidden unit depends only on the radial distance between these two points. The output of each hidden node is given by:

$$H_i(x_k) = \phi \left\| (x_k - c_i) \right\| \tag{4.18}$$

Where x_k : the input vector

c_i : the center of i neuron in the hidden layer

ϕ : the activation function

$(x_k - c_i)$: the distance between

x_k and c_i

The generally used non-linear activation function in RBFNN is Gaussian, which is given by:

$$\phi_i(x_k) = exp\left[\frac{-\left\|(x_k - c_i)\right\|^2}{2\sigma_i^2}\right] \tag{4.19}$$

Where σ_i : the width (spread) of the radial basis function.

According to Biswas (2014), there are two levels of RBFNN training: training the hidden layer and training the output layer. Training hidden layer nodes requires the determination of the center of each node, represented by RBF; the conventional way to find the center of each hidden layer node, as outlined by Bapat (2013), is to apply clustering algorithms such as k-means clustering. Training of the output layer requires computation of the connecting weights between the hidden and output nodes. The output of RBFN is presented by the linear computation between the output of hidden nodes and the connected weights and is given by the equation:

$$y_i = \sum_{i=1}^{M} w_{ij}\phi_i(x_k) \tag{4.20}$$

The most common method of estimating the connecting weights is through application of gradient descent (least mean squares algorithm) in order to minimize error and determine the class of each input.

Anifowose (2012) indicates that RBF has great potential as a neural network model for use in pattern classification as well as speech and speaker recognition, utilizing many useful properties such as interpolation and design matrices. In the literature, we have found that RBFN has better classification accuracy compared to many other pattern recognition methods (Finan et al. 1996 ; Agrawal et al. 2010 ; Venkateswarlu et al. 2011; Nijhawan and Soni, 2012 ; Baidwan and Gujral, 2014), however, RBFN defect is that it is easily stuck in a local solution, and the setting of initial weight is not theoretically backed (Zhou and Gu, 2010). In an attempt to improve upon these problems, RBFN optimization scheme based on Bacteria Foraging Optimization Algorithm (BFO-RBF algorithm) should be proposed.

4.3 Bacterial Foraging Optimization Algorithm (BFOA)

4.3.1 Introduction

The social behaviour and intelligence of some animals, as well as the study of mimicking their intelligence and social behaviour is known as swarm intelligence and focuses on such

phenomenon as bird flocking patterns, fish schools, and collective ant navigation. Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) are optimization algorithms inspired by the natural social behavior swarms of these animals. According to Das et al. (2009), such inspired algorithms have demonstrated their efficiency in solving complex problems in optimization in different areas. BFOA is a state-of-the-art algorithm proposed by Passino Das et al. (2009); formed as an extension of the swarm-based algorithms, this algorithm is highly effective in solving complex optimization problems. BFOA applies the group foraging strategies of a particular strain of *E. Coli* in multi-optimal function optimization, in which the bacteria attempt to maximize energy per unit of time in their search for nutrients, as the basis for its algorithm and then communicate with other bacteria by sending signals. A bacterium makes its decisions in foraging for nutrients after the consideration of two previous considerations. Chemotaxis, the method of movement employed by *E. coli*, involves the search for nutrients in small steps; mimicry of this chemotactic movement in the problem search space is the underlying principle behind BFOA. Li et al. (2014) explain that such a self-organized inspired system forms the basis for an algorithm that is advantageous because it includes parallel distributed processing, has little or no sensitivity to the initial value, and uses global optimization.

4.3.2 The Concept of BFOA

Locomotion of the *E. coli* for the gathering of nutrients is achieved either through tumbling or swimming, two functions performed by the tensile flagella. Tumbling as a means of locomotion is achieved through the clockwise rotation of the flagella; the flagella experiences tension after rotating in this manner, forcing the tumbling of the bacterium. The bacterium does not tumble as frequently in favourable environments but will tumble a great deal in those environments that will be harmful to it. Rotation of the flagella in the counterclockwise direction creates propulsion for the bacterium, allowing it to swim towards the nutrient gradient in a process previously described as chemotaxis. Fig. 4.8 outlines the tumbling and swimming movements of the *E. coli*. There are four central mechanisms that can describe the foraging strategies of *E. coli* and are mimicked by BFOA to solve the problem of non-gradient optimization, including chemotaxis, swarming, reproduction, and

elimination-dispersal (Das et al. 2009). These processes will be discussed below in further detail.

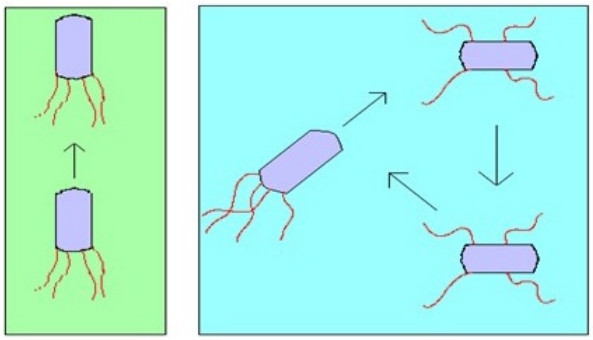

Figure 4.7: Swim and Tumble of a bacterium (Das et al., 2009).

• Chemotaxis

According to Li et al. (2014), chemotaxis describes the movement of the model *E. coli* and occurs through swimming or tumbling. Upon entering a favourable environment (one rich in nutrients), the bacterium will begin swimming via counterclockwise rotation of the flagella; if the location remains favourable then the bacteria will continue the swimming process until chemotaxis is complete. If the bacterium encounters an unfavourable, or noxious, environment it will tumble by rotating its flagella in a clockwise fashion to allow it to change direction so that it is able to reach a location rich in nutrients again. A bacterium position represents a candidate solution of the problem and information of the bth bacterium with d-dimensional vector represented as

$$\theta^b = [\theta_1^b, \theta_2^b, \ldots \theta_d^b] \tag{4.21}$$

Where: $b = 1, 2 \ldots, Nb$

Nb number of bacteria have certain parameters, represented as *Nc* –the number of chemotactic processes, *Nr* –the number of reproductive processes, *Ne* –the number of elimination and dispersal processes shown by a d-dimensional vector where d represents the dimension of the search space and the number of parameters to be optimized. In computational chemotaxis, the movement of each bacterium is shown by the equation:

$$\theta_b(c+1,r,e) = \theta_b(c,r,e) + C(b)\phi(c) \tag{4.22}$$

Where:

$C(b)$: the run length unit, dictated by the size of the tumble

$\Phi(c)$: the random direction in which the bacterium tumbles

• The Reproduction

The below cost function is used to calculate the cost of each chemotactic operation on the bacterium and the resulting health of the bacterium.

$$ob - fun_{health}(b) = \sum_{c=1}^{N_n} ob - fun_{less}(b,c,r,e) \tag{4.23}$$

From this equation, the location of healthier bacteria can be known; to optimize the search and increase efficiency, a larger number of bacteria must be placed in these locations in the reproduction stage. Bacteria are first arranged in order from most fit to least fit (lowest to highest cost function values). In order to maintain a constant population size, the healthiest half of the bacteria at the end of chemotaxis, as determined by the lowest numbers from the above function, remain alive and reproduce; after reproducing, there are two bacteria at the location of the original healthy bacteria, leaving a net amount of bacteria that is constant (Li et al. 2014; Al-Hadi et al. 2011).

• Elimination and Dispersal

BFO focuses on simulation of the chemotactic and reproductive processes as a way of searching locally for the optimal positions. Chemotaxis and reproduction are not sufficient for optimal global searches because it is possible that bacteria become stuck in local optima. To prevent being stuck in such a local optima, BFO simulates a dispersion event

that kills some random amount of bacteria, as determined by a preset probability Ped, and then moves the replacement bacteria to a new space. This dispersion is designed to simulate the real-world situations that may be encountered by the bacteria, such as a suddenly unfavourable location due to a decrease in available nutrients (Das et al. 2009; Li et al. 2014).

• The Swarming

E.coli bacteria has a governing, sensing, and anticipatory mechanism that allows it to determine the proper location and communicate this to other bacteria. When the bacterium moves, it releases a chemical known as an attractant that signals other bacteria to follow it; simultaneously, there is another chemical released known as a repellant that tells other bacteria to maintain a safe distance from one another. BFO simulates these attractive and repellant processes, which known cell-cell effect, by the below equation:

$$\begin{aligned} ob-fun_{eff}(\theta,\theta^b(c,r,e) &= \sum_{b=1}^{N_b} ob-fun_{eff}(\theta,\theta^b(c,r,e) \\ &= \sum_{b=1}^{N_b}[-d_{attractant}exp(-w_{attractants}\sum_{s=1}^{d}(\theta_s-\theta_s^b)^2)] \\ &+\sum_{b=1}^{N_b}[h_{repellant}exp(-w_{repellants}\sum_{s=1}^{d}(\theta_s-\theta_s^b)^2)] \end{aligned} \tag{4.24}$$

Where:

$ob-fun_{eff}$: objective function value with take in account cell-cell effect.

$d_{attractant}$: the depth of the attractant.

$w_{attractant}$: the width of the attractant

$h_{repellant}$: the height of the repellant

$w_{repellant}$: the width of the repellant

All those coefficients are carefully selected. The objective function value is to be added to the actual objective function and minimized, in order to present a time varying cost function (Li et al. 2014). The steps for BFOA is given:

1. Initial variables are defined as:

1.1 Nb Amount of bacteria

1.2. The step size C(b)

1.3. Number of parameters (d) to be optimized

1.4. Swimming distance Ns

1.5. Total iterations per chemotatic loop N_c

1.6. Reproductive abundance Nr

1.7. Abundance of elimination and dispersal events Ne Probability of elimination and dispersal Ped Location of each bacterium.

2. Elimination and dispersal loop $e = e + 1$

3. Reproduction loop $r = r + 1$

4. Chemotactic loop $c = c + 1$

4.1 For b =1, 2 ,..., Nb

4.1.1 Simulation of RBFNN using random weight values (random initial position and direction).

4.1.2 Calculation of the first objective function value for each bacterium. Save the objective function value as $ob_{fun1}(b,c,r,e)$ in case a better value is found

4.1.3 Loop termination

4.2 For b =1,2,..., Nb the tumbling/swimming decision is made

4.2.1 Tumble:

4.2.1.1 A random vector is generated by the equation $\theta_b(c+1,r,e) = \theta_b(c,r,e) + C(b)\phi(c)$

4.2.1.2 Using $\theta_b(c+1,r,e)$ to compute the new location taking into account the cell-cell effect as equation 4.24 and save it $ob_{fun2}(b,c,r,e)$

4.2.2 Swim:

4.2.2.1 Initiate swim loop $s = s + 1$, While $s < Ns$

4.2.2.2 Compare the two objective functions If $(ob_{fun2}) < (ob_{fun1}), let(ob - fun_{less}) = (ob_{fun2})$ Else $s = Ns$

4.2.3 If $b \neq Nb$ the succeeding bacterium begins $(b+1)$

4.2.4 If $c < Nc$, the chemotactic loop resumes (the bacterium is still alive)

5. Reproduction loop

5.1. For each bacterium (b), as well as for each of the reproduction loops (r) and elimination and dispersal events (e), compute the health of each bacterium from summation of the minimum objective functions.

$ob - fun_{health}(b) = \sum_{c=1}^{N_c} ob - fun_{less}(b,c,r,e)$

5.2. Let $br = b/2$ with the half of the bacterial population with the highest $ob - fun_{health}$ (lowest health) being allowed to die and the other half spilt into two, allowing the population of the bacteria to remain constant.

5.3. If $r < Nr$, initiate a new chemotactic step.

6. Elimination and dispersal loop For each bacterium with a probability, p_{ed} ,eliminate and disperse each bacterium so the total population of the bacteria remains constant. To achieve this constancy, if a bacterium is eliminated, another should be randomly dispersed in the optimization domain. If $e < Ne$, proceed to step 2; otherwise terminate.

A flow chart for the aforementioned procedure is represented in Fig. 4.8 (a, and b).

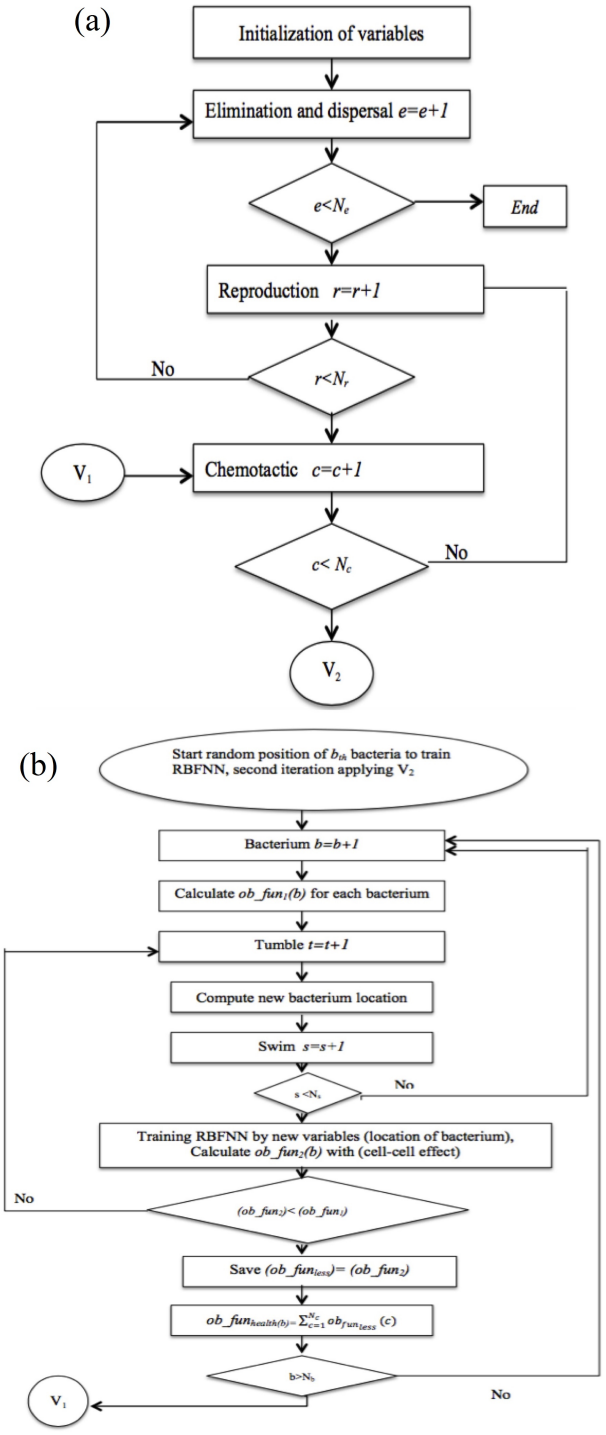

Figure 4.8: a, b Flowchart of BFOA.

4.4 Evaluation Parameters

4.4.1 Accuracy

According to Costa et al. (2007), speaker recognition is one task of pattern recognition problems. Accuracy is analyzed by arranging correctly and incorrectly identified examples in something called a confusion matrix for classification problems. The confusion matrix is a 2×2 matrix composed of a positive or negative true class and a positive or negative predicted class. This creates 4 possible results in determining the accuracy of the system: A true positive [TP] is the correct identification of a speaker, a false positive [FP] is the incorrect identification of a speaker, a true negative [TN] is the correct identification of the incorrect speaker, a false negative [FN] is the incorrect identification of the incorrect speaker. Currently, accuracy is the most common evaluation parameter used; accuracy measures the classifier's performance by determining the ratio of correct ("true") results over all cases, which are achieved by the classifier.

$$Accuracy = \frac{|TN| + |TP|}{|TN| + |TP| + |FN| + |FP|} \tag{4.25}$$

4.4.2 Sensitivity

Sensitivity is another parameter used to evaluate the performance of a classifier and is represented as a ratio of true positives over all actual positives. This measure gives a better representation of the actual performance of the classifier because the ratio represents only the cases that were actually positive (TP and FN are truly positive).

$$\text{Sensitivity} = \frac{|TP|}{|TP| + |FN|} \tag{4.26}$$

4.4.3 Specificity

Similar to Sensitivity is the evaluation parameter of Specificity. It is a proportion of the number of true negatives over all actual negatives.

$$\text{Specificity} = \frac{|TN|}{|TN| + |FP|} \tag{4.27}$$

4.4.4 Cochlear Implant-like Spectrally Reduced Speech (SRS)

The cochlear in human inner ear acts as a narrowband filter ranged (20Hz-20KHz). The processes of auditory perception in much research have considered that the signal resulting from cochlear filtering should be perceived as a sinusoidal combing of two principal parts: Amplitude Modulation AM called the temporal envelope, and frequency modulation FM called the temporal fine structure. Both are essential in speech perception (Lorenzi, 2014). In speech signal processing, the AM component (the envelope) is an important process for speech recognition and pitch perception e.g. speaker identification. The FM compo-nent (the fine structure) is more significant for the music sound perception. The reason why the envelope is a useful strategy for speaker identification that the formant frequencies of the vocal tract resulting from the peaks in the spectral envelope that can be extracted by low pass filters; $F_1, F_2, \ldots F_N$, considered as features that help us identify the speakers. Therefore, the capable auditory human system simulation to extract the AM modulations in speech signal is the fundamental approach to achieve a robust speaker identification sys-tem. Do and Barras (2012) proposed an algorithm called Cochlear implant-like spectrally reduced speech (SRS). It is a re-synthesized signal from the basic speaking one employing the idea of audio perception of temporal modulations in speaking. The speaking signals pass through group of filters to dismantle the speaking signals to N sub-narrowband ones. Applying rectification, sub-band temporal envelopes are extracted. Modulating a cosine signal with fundamental frequencies can be obtained through those low-bandwidth sub-band temporal envelopes. That modulated signal is equal to cut-off frequency of BPF. As a result, spectrally reduced speech (SRS) is known from the addition of those signals to-gether. Cochlear implant-like speech processing will be used to evaluate and check our proposed SI system.

Chapter 5

PROPOSED SYSTEM

In this research, LPCC, MFCC and their concatenation LPCC-MFCC, as well as this concatenation with their respective normalized cepstral averages (LMACC) are extracted from the speech data sets and then used to train and test the dataset to evaluate the classifier performance. For the features matching stage, the radial basis neural network (RBFNN) is used as a classifier, and then it will be optimized using Bacterial Foraging Optimization (BFO) as mentioned earlier. Accuracy, sensitivity, and specificity will be used alongside cochlear implant-like spectrally reduced speech (SRS) to evaluate the systems.

5.1 Algorithm

The Bacterial Foraging Optimization Algorithm (BFOA) is used to tune the weights of the RBFNN. In fact, BFO algorithm has an objective function whose value is manipulated to achieve minimum value. In our proposed system, the evaluation parameters' values include accuracy, sensitivity, and specificity; those three have to be increased using BFO in order to be accurately modified RBFNN. As a result of the equation, inverting sensitivity, as a minimized objective function shown below, would positively affect the other values since it will assist in giving the maximum number TP (avoiding stuck in accuracy paradox). Sensitivity gives the most effective evaluation of the system because this ratio represents only the cases that are actually positive.

$$\text{objective function} = \frac{1}{\text{sensitivity}}$$

In our algorithm, a bacterium position will be changed to accomplish the minimum objective function; that position will be defined by the adjusted variables of RBFNN. In those variables, weights of the output layer will be adjusted to correctly achieve the target class;

thus, the weights will act as controllers of bacteria dynamics. The Total number of weights presents dimensions searching space and the change in RBFNN weights occurs when the position of bacteria is updated. This idea of connection weights optimization also employed in (Zhou and Gu, 2010; Al-Hadi et al. 2011); Kaur and Kaur, 2015) to enhance ANN' performance. Fig. 5.1 illustrates a flow chart for the proposed work.

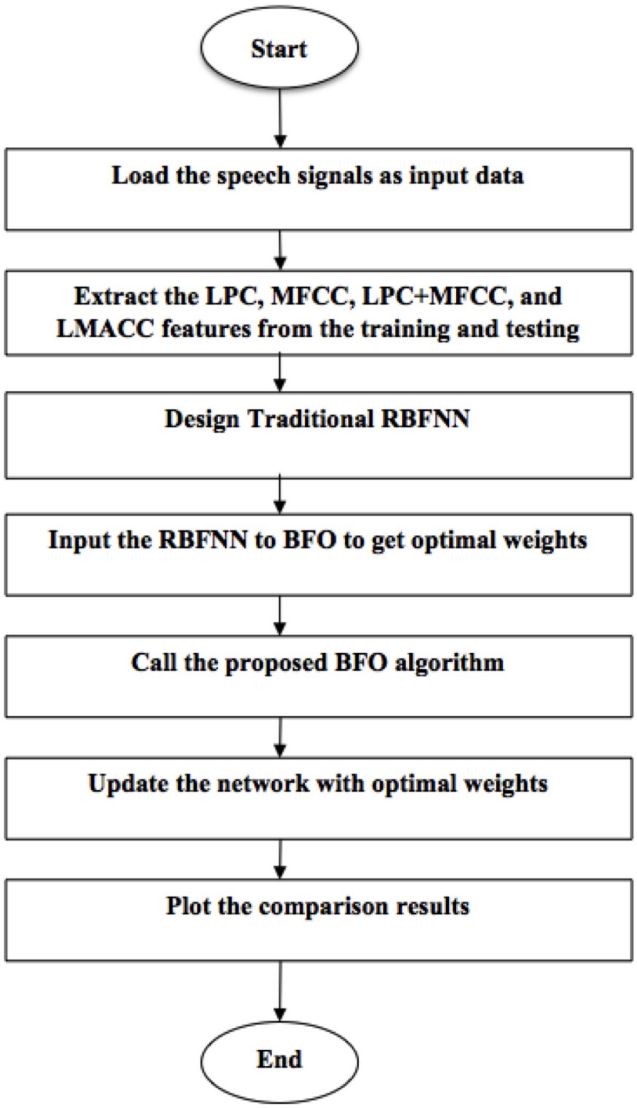

Figure 5.1: Flow chart for the proposed work

5.1.1 Dataset

We have recorded the voice of four speakers in both a quite calm environment and a noisy one using PRAAT software (Boersma, 2016) .The participants are two adult and two chil-dren with both genders. Each participant spoke four different words including the numbers: one, two, three, and four. To create the training dataset, the speakers repeated each word twenty times. Therefore, the training features matrix consists of 320 rows for total features and columns equal to the orders of LPCC and MFCC, which are 12. To create the test-ing data set, each speaker produces the numbers word ten times, and then both LPCC and MFCC features will be extracted from that. The testing matrix will have 160 features and an equal number of orders, the same as for training dataset.

5.1.2 Proposed Features Extraction

The front end of the analysis is the same for both LPCC & MFCC as shown in Fig. 5.2.

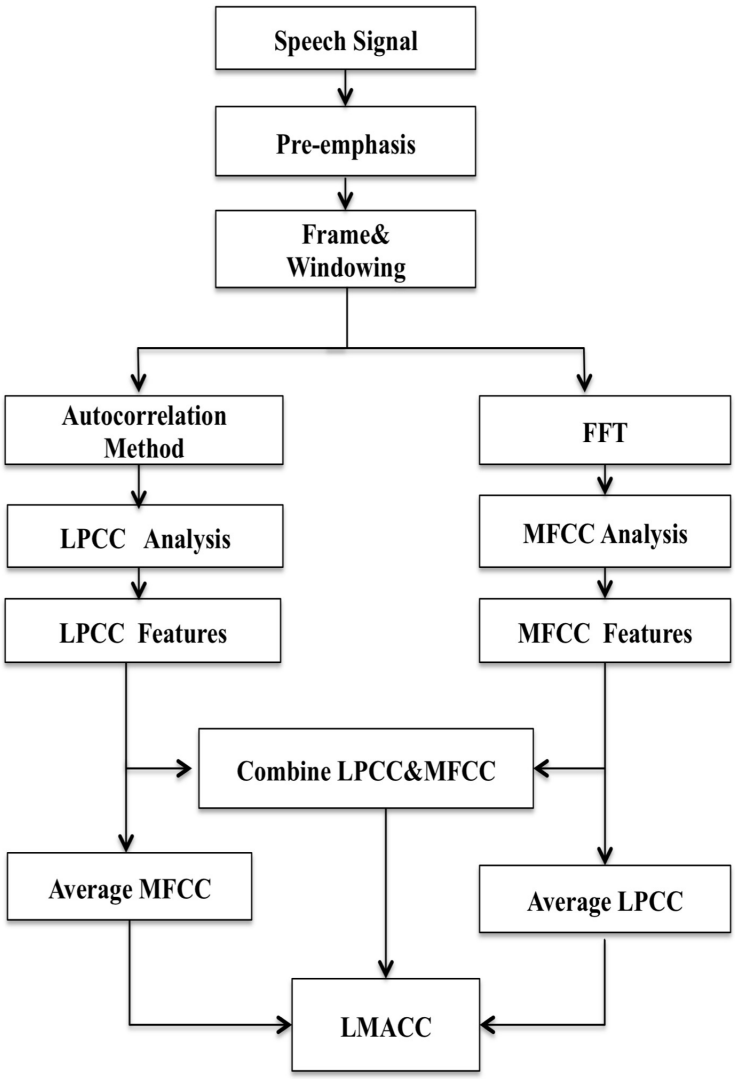

Figure 5.2: Festures Extraction Approaches

• Pre-emphasis

The effect of glottis is making the prediction of coefficients not belonging to vocal tract. Thus, the idea of pre-emphasis filter is to cancel the glottal effect from the transfer function of vocal tract by applying high pass filtering.

• Framing and Windowing

According to the Nyquist sampling theorem; the recorded speech signal of each speaker will be sampled using the 16 kHz sampling frequency, which should be at least twice the maximum frequency. As known, human speech frequencies range between 100Hz and 8KHz; so a 16KHz sampling frequency is sufficient for speech and speaker recognition tasks. This speech signal will be framed in the form of segments, each of which is $25ms$ in length, which will include $0.025 \times 16000 = 400$ samples. Each frame, then, is windowed with overlap by 30 % in order to avoid the signal discontinuities. ENFAME MATLAB code is used for framing and windowing the recorded speech signals.

• LPCC Analysis

Each windowed signal frame will be auto-correlated, and then each frame of autocorrelations is converted into LPC coefficients by using Durbin's algorithm. To convert the LPC to LPCC, the recursive method is applied as defined in chapter 4.

• MFCC Analysis

Each window will be converted to a signal' spectrum by using FFT. Spectral analysis is carrying important information of speech signal, so it is considered an essential step in feature extraction process. Applying Log and IFFT from the signal' spectrum, called Cepstrum analysis, will get a spectral envelope. Obtaining Mel spectrum will be through applying Mel filters in resulting spectrum, and getting Mel spectrum envelope will be through the application of Cepstrum analysis whose result will be Mel Frequency Cepstral Coefficient (MFCC). MELCEPST MATLAB function is used in our experiment.

• Concatenating LPCC & MFCC Features

Both feature vectors: LPCC and MFCC will be examined as concatenated, which means both are combined in a matrix. Therefore, the training features matrix will consist of 640

rows for total features and columns equal to12. The testing matrix will have 320 rows and 12 columns.

• LMACC Features

The newly created feature's matrix includes both LPCC and MFCC, as well as their respective normalized averages for each cepstral feature (LMACC). This matrix allows a large number of features. The training features matrix will consist of 960 rows for total features and columns equal to12. The testing matrix will have 480 rows and 12 columns.

• Features Normalization

The proposed feature extraction method: LMACC will be rescaled in order to enhance the performance of RBFNN in terms of the convergence speed, improving the numeric calculation accuracy, and avoiding existence in ill-conditioned neural network. The normalization method we applied is *z*-score, defined by

$$x_{norm} = \frac{(x-\mu)}{\sigma}$$

where:

μ: Mean of each speaker 's feature

σ: Standard deviation of each speaker 's feature

- **RBFNN Design**

The structure of RBFNN is considered for the classification stage to match LPCC and MFCC filtered speech signals. Our designed network consists of 12 input nodes with 202 hidden nodes and one output node for the speaker identification system. The parameters used for the RBFNN model are formulated as shown in Table 5.1.

Table 5.1: Training Parameters of RBFNN

RBFNN input nodes	12
Target Error	0.02
Spread Constant	1
Hidden Layer nodes	202

5.2 Results and Discussion

The comparison of feature approaches will be performed twice: with conventional RBFNN and again with the optimized RBFNN.

5.2.1 LPCC_RBFNN

• LPCC Features Extraction

We use the LPC function available in the MATLAB's speech processing and neural network toolbox to extract LPCC features, and an estimated signal is filtered using these LPCC coefficients as shown in Fig. 5.3. The speech signal recorded is used as an input signal. For each signal recorded, LPCC coefficients are extracted and the signal is filtered out to generate an estimated signal. The error of the filtered signal with the original one is shown.

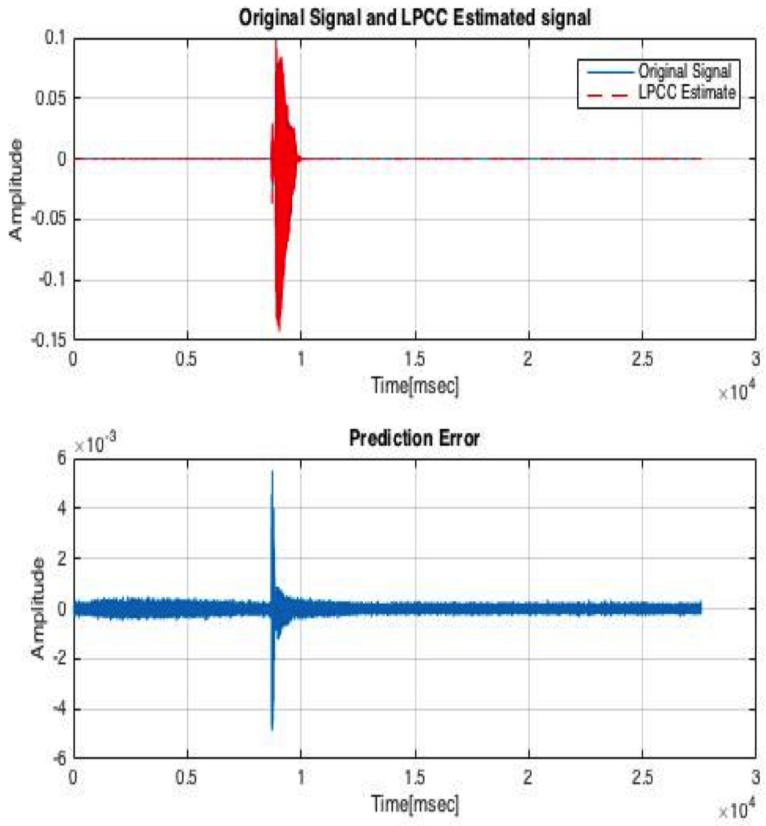

Figure 5.3: LPCC Features Plot of the Filtered Signal

• LPCC Features Based Training RBFNN

During the feature extraction phase we designate classes for each feature to which a speaker belongs. Fig. 5.4 shows these classes in different plus sign colors red for class one, green for class two, blue for class three, and magenta for class four. After simulating RBFNN with a testing extracted feature, the output is plotted as a square symbol. The square symbol indicates the actual output of our system. If the color of the square symbol fits the color of the plus sign that means the correct feature has been identified, if this is not the case then the color of the square symbol indicates the actual class that this feature should belong to this speaker. For example, if the square is green with a red plus sign, this means it is identified as speaker one but it is actually speaker two. The feature without a square is considered to be an unknown class to the system. As a whole, there are 4 clusters considered for 4-recorded speakers' speech signals.

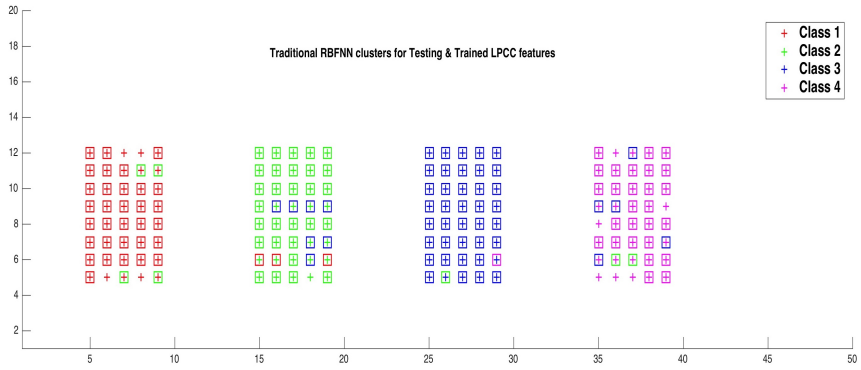

Figure 5.4: RBFNN Output clusters vs. Target classes

Figure 5.5 shows MSE graph with 202 epochs, which indicates that the RBFNN completed learning after those epochs in order to achieve the MSE with 0.0204, which almost reaches the error goal 0.02.

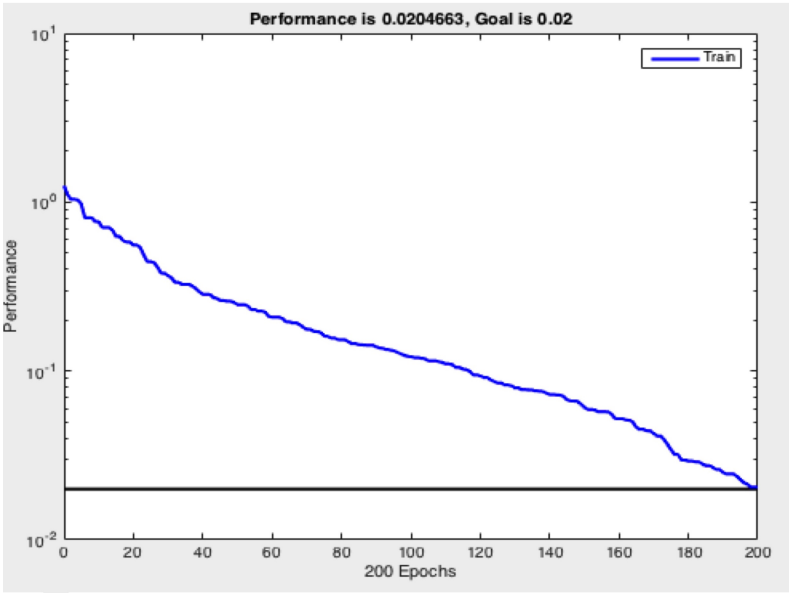

Figure 5.5: LPCC_ RBFNN MSE Performance

Premised on the output accuracy, sensitivity, and specificity values are calculated to evaluate the system. Those evaluation parameters for LPCC-RBFNN are indicated in Table 5.2.

Table 5.2: LPCC Features Based RBFNN for Speaker Identification System Results

Approach	LPCC_RBFNN
Accuracy	78.75
Sensitivity	78.9
Specificity	55.3693
MSE	0.0204
Epochs	202

5.2.2 MFCC_RBFNN

• MFCC Features Extraction

From the dataset of the same spoken words, MFCC features are extracted. Fig. 5.6 shows the MFCC Filtered signal, and the error prediction, which indicates that the error in MFCC features output is higher than LPCC filter as it is changing approximately between 0.4 and -0.4. In the LPCC, it is changing between 0.0075 and -0.0075; therefore, LPCC performs better than MFCC in the prediction of the speech signal.

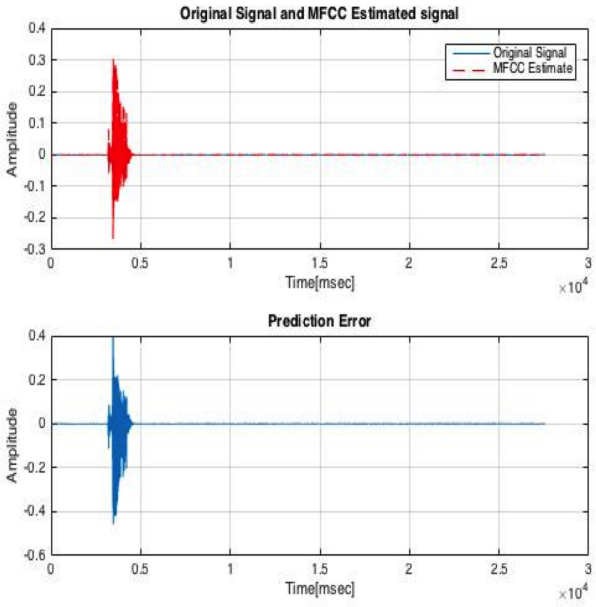

Figure 5.6: MFCC Features Plot Filtered Signal

• MFCC Features Based Training RBFNN

Inspite of repeating the training RBFNN using MFCC instead of LPCC, and that the results explaining that the MFCC prediction error is higher than the LPCC filtered signal, MFCC features work better for the training RBFNN classifier based on the same evaluation parameters. Table 5.3 shows the MFCC_RBFNN evaluation parameters model.

Table 5.3: MFCC Features Based RBFNN for Speaker Identification System Results

Approach	MFCC_RBFNN
Accuracy	82.5
Sensitivity	82.6
Specificity	61.1538
MSE	0.02
Epochs	200

5.2.3 LPCC _ MFCC, and LMACC Features Based Training RBFNN

Both feature vectors: LPCC and MFCC will be examined as cascading, which means both are combined in a matrix. LPCC features followed by MFCC are a cascade used for training RBFNN as happens in an individual case. This combination will be used again with the average of both features in one matrix, as LMACC.

Results from this case show that combination between LPCC and MFCC features perform better than the MFCC and LPCC features independently. Yet, LMACC gives a better performance in our identification system. Thorough values are represented in table 5.4. For clearer demonstration, Fig. 5.7 is plotted to illustrate the comparison between the four identification systems based on the three previously mentioned evaluation parameters. That concludes that the LMACC_RBFNN model is the best mode among all checked models once the dataset, the filter order and the RBFNN parameters are stabilized.

Table 5.4: Comparison between LPCC, MFCC, LPCC&MFCC, and LMACC in conventional RBFNN

Approach	LPCC_RBFNN	MFCC_RBFNN	LPCC&MFCC_RBFN	**LMACC_RBFN**
Accuracy	78.75	82.5	84.687	**86.667**
Sensitivity	78.9	82.6	84.8	**86.8**
Specificity	55.3693	61.1538	64.844	**68.541**

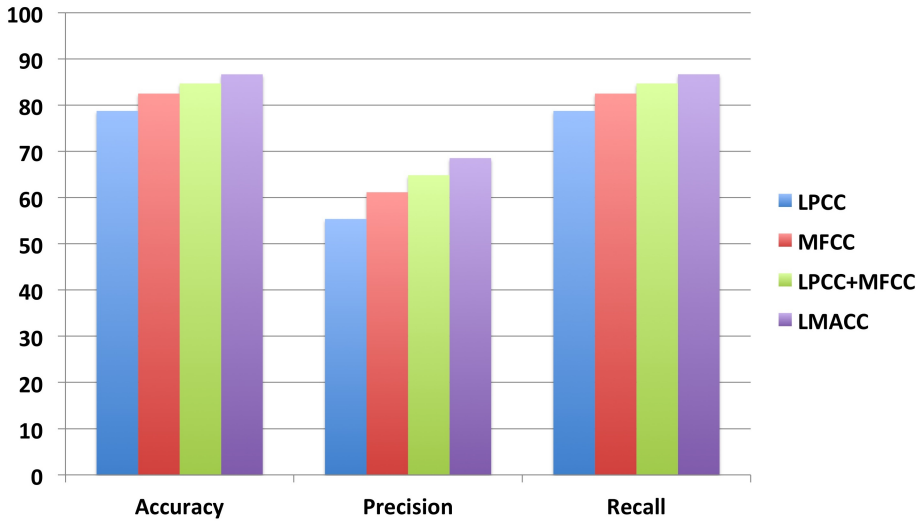

Figure 5.7: Comparison between LPCC, MFCC, LPCC_MFCC, and LMACC in conventional RBFNN

5.3 Proposed RBFNN

As mentioned earlier, the problem in RBFNN is within-space limit searching algorithm. In addition, the setting of the initial weight is not theoretically backed. Those two defects lead to inefficient identification and less accuracy particularly in SIS. In our proposed scheme, more overall performance can be achieved by applying global optimization algorithm to overcome this problem. Thus, to enhance RBFNN, its weights are adjusted through bacterial foraging optimization (BFO). In our research, the considered BFO's parameters are formulated in Table 5.5.

Table 5.5: BFOA Parameters

BFOA parameters	Values
N_b Amount of bacteria	10
d-Dimensions searching space	2828
Ns -Swimming distance	4
N_c - chemotactic loop	8
N_r-Reproductive abundance	4
N_e - elimination and dispersal	2
$C(b)$ - the run length unit	0.05
P_{ed}–Probability of elimination and dispersal	0.25
Number of Iterations	350

The next four sections will include applying the four different extracted features as training and testing data for BFO-RBFNN, then comparing our proposed RBFNN with conventional RBFNN in each feature extraction technique.

5.3.1 LPCC Features Based Training of BFO Tuned RBFNN

In each iteration, the bacteria change their position. The movement ranges between -4 and 4 for each bacterium, and 10 bacteria are considered (shown in table 5.5). Around 350 iterations are executed in total to reach a minimum value of the objective function. In our optimization, a model provides the minimum value after 25 iterations and then it approximately remains steady with minor fluctuations (shown in Fig. 5.8). The trained BFO-RBFNN is consequently used to simulate the test signal and output clusters with the desired target (shown in Fig. 5.9), which clearly shows the reduction of the incidence of misclassification or non-existent classification.

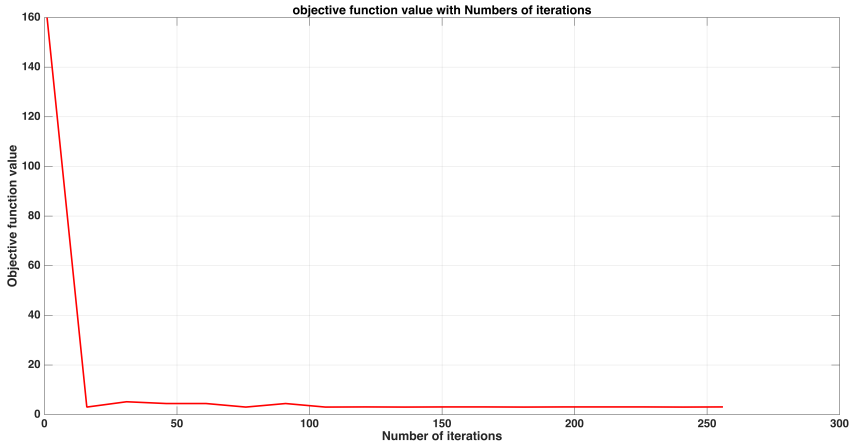

Figure 5.8: Objective function value illustrated through optimized BFO_RBFNN

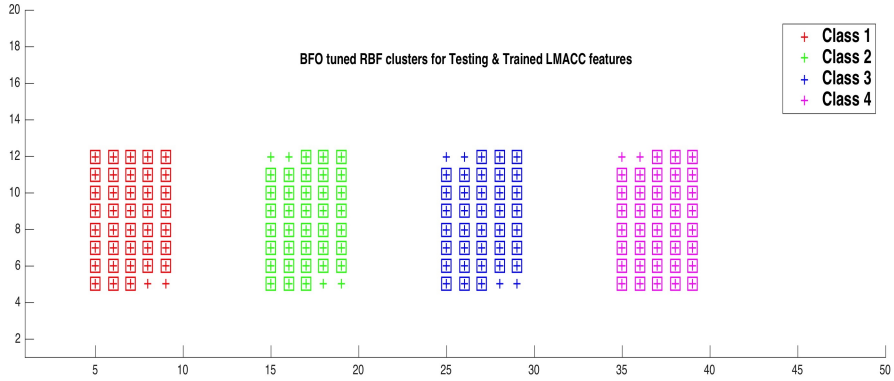

Figure 5.9: BFO-RBF Output clusters vs. Target classes

In table 5.6, the comparison between both networks is done based on the three evaluation parameters. For clear comparison, a graph for two systems is shown in Fig. 5.10 illustrating that our proposed algorithm approximately achieves improvement of 18 % of system overall.

Table 5.6: Comparison of RBFNN and BFO -RBFNN for SIS using LPCC Features

Approach	LPCC_RBFNN	**LPCC_BFO_RBFN**
Accuracy	78.75	**92.5**
Sensitivity	78.9	**92.7**
Specificity	55.3693	**80.492**

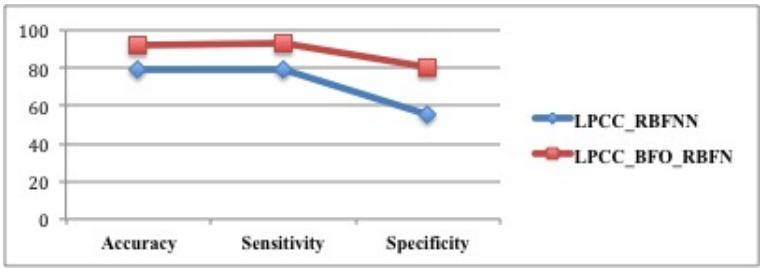

Figure 5.10: LPCC_RBFNN Vs. LPCC_BFO_RBFNN

5.3.2 MFCC Features Based Training of BFO Tuned RBFNN

After training and testing BFO_RBFNN using MFCC, we reached Table 5.7 that compares MFCC_RBFNN and the MFCC_BFO_RBFN. It is clearly shown in Fig. 5.11 that there is about 17% improvement in the results of the optimized BFO_RBFNN algorithm compared to the improvement in the Traditional RBFNN.

Table 5.7: Comparison of RBFNN and BFO-RBFNN for SIS using MFCC Features

Approach	MFCC_RBFNN	**MFCC_BFO_RBFN**
Accuracy	82.5	95
Sensitivity	86.8	**98.6**
Specificity	61.1538	86.471

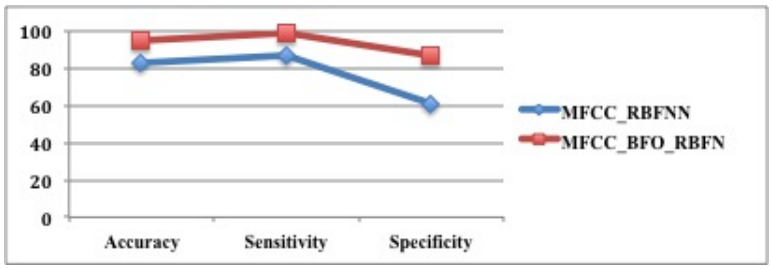

Figure 5.11: Comparison of RBFNN and BFO tuned RBFNN for SIS using MFCC

5.3.3 LPCC & MFCC Features Based Training BFO Tuned RBFNN

After analyzing LPCC and MFCC as Cascading features for training and testing proposed RBFNN by BFO, Tables 5.8 illustrate a comparison between features in conventional RBFNN and proposed RBFNN. For clearer demonstration, Fig. 5.12 shows that our proposed algorithm approximately achieves an improvement of 16 % of the overall performance of the Speaker Identification system.

Table 5.8: Comparison of Cascading LPCC and MFCC feature with Traditional and Proposed RBFNN

Approach	LPCC&MFCC _RBFN	LPCC&MFCC_BFO_RBFN
Accuracy	84.687	**96.25**
Sensitivity	86.8	**98.6**
Specificity	64.844	**89.5456**

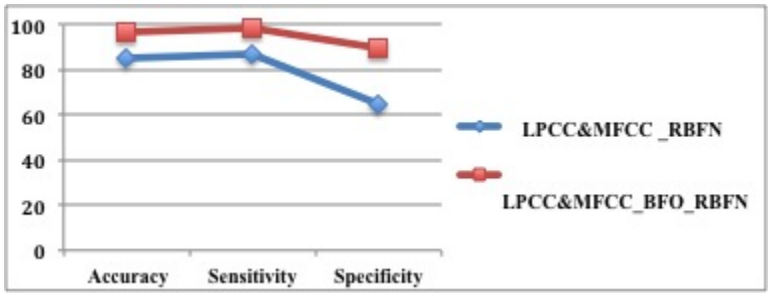

Figure 5.12: Comparison of combining LPCC &MFCC feature with Traditional and proposed RBFNN

5.3.4 LMACC Features Based Training BFO Tuned RBFNN

Table 5.9 illustrates a comparison between the combination feature, LMACC, and both the conventional and proposed RBFNN. Graph 5.13 demonstrates the overall increase in system performance after their application.

Table 5.9: Comparison of LMACC Feature with Traditional and Proposed RBFNN

Approach	LMACC _ RBFN	LMACC_ BFO_RBFN
Accuracy	86.667	**98.333**
Sensitivity	86.8	**98.6**
Specificity	68.451	**95.178**

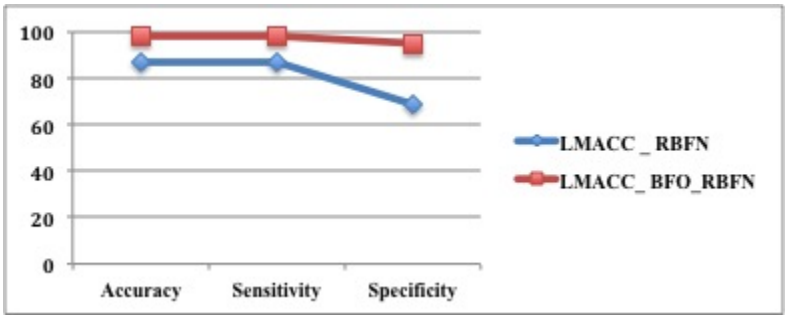

Figure 5.13: Comparison of LMACC feature with Traditional and proposed RBFNN

5.4 Summary

In conclusion, after evaluating and comparing the performance of four extraction methods with the Traditional RBFN and Optimized BFO_RBFN for SIS, are shown to be more effective than the traditional method. The proposed LMACC performs optimally with the optimized BFO_RBFNN.

Table 5.10: Comparison of all Feature Extraction Techniques

Approach	**Accuracy**	Sensitivity	Specificity
LPCC_RBFNN	78.750	78.9	55.3693
LPCC_BFO_RBFN	92.500	92.9	80.492
MFCC_RBFNN	82.50	82.6	61.1538
MFCC_BFO_RBFN	95.00	95.8	86.471
LPCC&MFCC _RBFN	84.687	84.8	64.844
LPCC&MFCC_BFO_RBFN	96.250	96.6	89.5456
LMACC _RBFN	86.667	86.8	68.451
LMACC_BFO_RBFN	**98.333**	**98.5**	**95.178**

5.5 Case Study

When experimenting with the best evaluated features extractor with the Cochlear implant-like spectrally reduced speech (SRS) proposed in the literature Do and Barras (2012), it is assumed that our proposed SI system, LMACC based on Optimized BFO-RBFN algorithm has better recognition performance in terms of the three previously mentioned evaluation parameters.

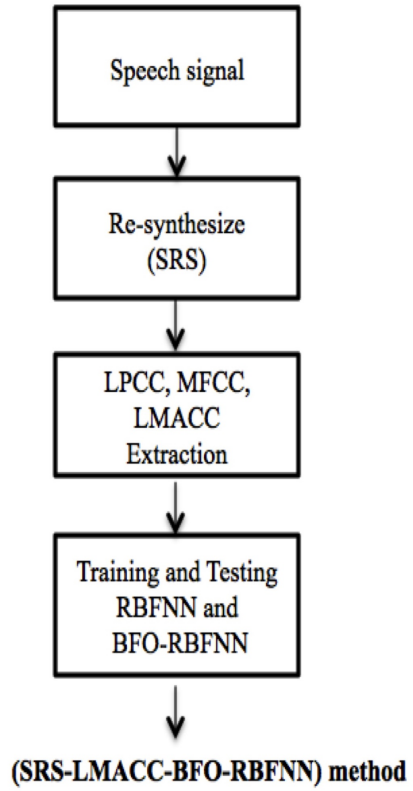

Figure 5.14: SRS-LMACC feature based Evaluation of our proposed method

Cochlear implant-like SRS will be used to evaluate and check our proposed SI system, LMACC_BFO_RBFNN as shown steps in Fig. 5.14. Enhanced LMACC features from the spoken words will be extracted first, based on the SRS algorithm then used for training and testing the RBFNN and BFO-RBFNN. Table 5.11and Fig. 5.15 show that the overall performance of SRS with our proposed method optimized RBFNN, gives better recognition than SRS without it. As a result, SRS_LMACC_BFO-RBFN is the best solution for speaker identification if the above methods implemented.

Table 5.11: Comparison of LMACC and SRS-LMACC with Traditional and Proposed RBFNN

Approach	LMACC _RBFN	LMACC_BFO_RBF	SRS_ LMACC _RBF	SRS_LMACC _BFO
Accuracy	86.667	98.541	87.083	**98.75**
Sensitivity	86.8	98.6	87.2	**98.9**
Specificity	68.451	95.178	69.335	**96.35**

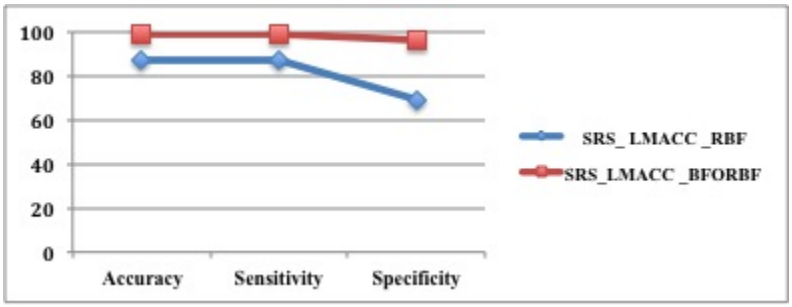

Figure 5.15: SRS_LMACC-RBFNN vs SRS_BFO_LMACC_RBFNN

5.6 Noisy Environment

All previously mentioned methods were also tested when noise was added to the recorded speech signals. The PRATT software was used under an available function called random-Gauss. We used noise whose amplitudes are 0.1 and 0.5, measured by root- mean -square (RMS). The results are shown in Tables 5.12 and 5.13, which indicate that in both cases LPCC is highly sensitive to noise compared to MFCC. Our proposed feature LMACC is immune to noise and still performs better than others even in noisy environments.

Table 5.12: Comparison of all Feature Extraction Techniques with 0.1 RMS

Approach with 0.1 RMS noise	**Accuracy**	Sensitivity	Specificity
LPCC_RBFNN	68.987	68.942	42.542
LPCC_BFO_RBFN	87.974	87.900	71.558
MFCC_RBFNN	80.379	80.224	57.955
MFCC_BFO_RBFN	89.240	89.182	74.088
LPCC&MFCC _RBFN	82.278	82.187	60.847
LPCC&MFCC_BFO_RBFN	91.455	91.474	78.627
LMACC_ RBFN	83.966	83.894	63.653
LMACC_ BFO_RBFN	**92.616**	**92.628**	**80.721**

Table 5.13: Comparison of all Feature Extraction Techniques with 0.5 RMS

Approach with 0.5 RMS noise	**Accuracy**	Sensitivity	Specificity
LPCC_RBFNN	53.459	53.397	22.709
LPCC_BFO_RBFN	80.503	80.625	58.808
MFCC_RBFNN	77.358	77.307	53.258
MFCC_BFO_RBFN	85.534	85.609	66.800
LPCC&MFCC _RBFN	79.874	79.791	57.081
LPCC&MFCC_BFO_RBFN	87.421	87.403	70.618
LMACC _RBFN	83.428	83.413	62.758
LMACC_ BFO_RBFN	**90.985**	**91.023**	**77.281**

Chapter 6

CONCLUSIONS

With the increasing trend in the use of smart systems, speaker recognition system is a trusted technology for different aspects of everyday life such security, law and medicine applications. To become ubiquitous in such applications, the research seeks more accuracy to add to the performance of the speaker identification system in order to achieve the more specified r ecognition. A chieving t his g oal w ill r equire a f ocus o n t he t wo stages in the system: feature extraction and feature match. Furthermore, evaluation of the system should be more reliable to accomplish precise evaluation. A new feature extraction technique was proposed (LMACC) based on concatenation of LPCC and MFCC as well as other features related to their cepstral means. Optimized RBFNN by using BFOA enhanced the classification s tage. The comparison was drawn between both features independently, a concatenation of both features, and this concatenation with the respective normalized averages in conventional and optimized RBFNN; the LMACC achieved a high recognition rate in optimized BFO_RBFNN case. The evaluation was made based primarily on the three standard evaluation parameters in classification t asks: accuracy, sensitivity, and specificity. After evaluation, the proposed system (LMACC_BFO_RBFNN) was verified using Cochlear implant-like spectrally reduced speech (SRS) algorithm so that the original recorded signal was resynthesized based on an acoustic simulation of the cochlear implant. i.e. human speech perception. Our proposed system demonstrated maximal relative performance in environments on measures of recognition rate, compared with other methods and covering a range of different noise levels. This research takes into consideration many factors including sex, age, changes in dialect and noise of environment. It would be beneficial for future research to focus on different factors for recording speech such the health of the speaker, the emotional status of the speaker, multilingual speakers, and to attempt to guard against voice mimicry. Future research should also focus on the comparison of this

study, which employs single cost function as means of maximizing sensitivity, and the established Pareto optimization, which utilize multi-cost functions; maximizing of sensitivity and maximizing of specificity for example.

Bibliography

Abdallah, S. J., Osman, I. M., and Mustafa, M. E. (2012). Text-independent speaker identification using hidden markov model. *World of Computer Science and Information Technology Journal*, 2(6):203–208.

Agrawal, S., Shruti, A., and Krishna, C. R. (2010). Prosodic feature based text dependent speaker recognition using machine learning algorithms. *International Journal of Engineering Science and Technology*, 2(10):5150–5157.

Al-Hadi, I. A. A., Hashim, S. Z. M., and Shamsuddin, S. M. H. (2011). Bacterial foraging optimization algorithm for neural network learning enhancement. In *Hybrid Intelligent Systems (HIS), 2011 11th International Conference on*, pages 200–205. IEEE.

Ambikairajah, E. (2010). Speech and audio processing 1,2, 3: Speech analysis ï¿œ pro-fessor ambikairajah , retrieved from: https://www.youtube.com/watch?v=xjzm7s$_{kbu.}$

Anifowose, F. A. (2012). A comparative study of gaussian mixture model and radial basis function for voice recognition. *arXiv preprint arXiv:1211.2556*.

Arora, S. J. and Singh, R. P. (2012). Automatic speech recognition: a review. *International Journal of Computer Applications*, 60(9).

Bahari, M. H. (2014). Automatic speaker characterization; automatic identification of gender, age, language and accent from speech signals (automatische sprekercharacterisatie; automatische identificatie van geslacht, leeftijd, taal en accent uit stemopnamen).

Baidwan, V. S. and Gujral, S. (2014). Comparative analysis of prosodic features and linear predictive coefficients for speaker recognition using machine learning technique. In *Devices, Circuits and Communications (ICDCCom), 2014 International Conference on*, pages 1–8. IEEE.

Bapat, A.U., . M. S. (2013). Clustering algorithms for radial basis function neural network. *ITSI Transactions on Electrical and Electronics Engineering (ITSI-TEEE)*, 1(1):2320–8945.

Biswas, P. (2014). Mod-01 lec-28 rbf neural network (contd.) [video file]. retrieved from https://www.youtube.com/watch?v=not$_v$7$ndmleandretrievedfromhttp$: $//textofvideo.nptel.iitm.ac.in/117105101/lec29.pdf$.

Boersma, Paul & Weenink, D. (2016). Praat: doing phonetics by computer [computer program]. version 6.0.21, retrieved 25 september 2016 from http://www.praat.org/.

Chandra, E. and Kalaivani, K. M. M. (2014). A study on speaker recognition system and pattern classification techniques.

Costa, E., Lorena, A., Carvalho, A., and Freitas, A. (2007). A review of performance evaluation measures for hierarchical classifiers. In *Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop*, pages 1–6.

Cruttenden, A. (2014). *Gimson's pronunciation of English*. Routledge.

Das, S., Biswas, A., Dasgupta, S., and Abraham, A. (2009). Bacterial foraging optimization algorithm: theoretical foundations, analysis, and applications. In *Foundations of Computational Intelligence Volume 3*, pages 23–55. Springer.

Dave, N. (2013). Feature extraction methods lpc, plp and mfcc in speech recognition. *International journal for advance research in engineering and technology*, 1(6):1–4.

Dhameliya, K. and Bhatt, N. (2015). Feature extraction and classification techniques for speaker recognition: A review. In *2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO)*.

Do, C.-T. and Barras, C. (2012). Cochlear implant-like processing of speech signal for speaker verification. In *SAPA@ INTERSPEECH*, pages 17–21.

Farah, S. and Shamim, A. (2013). Speaker recognition system using mel-frequency cepstrum coefficients, linear prediction coding and vector quantization. In *Computer, Control & Communication (IC4), 2013 3rd International Conference on*, pages 1–5. IEEE.

Finan, R., Sapeluk, A., and Damper, R. (1996). Comparison of multilayer and radial basis function neural networks for text-dependent speaker recognition. In *Neural Networks, 1996., IEEE International Conference on*, volume 4, pages 1992–1997. IEEE.

Gaikwad, S. K., Gawali, B. W., and Yannawar, P. (2010). A review on speech recognition technique. *International Journal of Computer Applications*, 10(3):16–24.

Hanafi, D. and Sukor, A. S. A. (2012). Speaker identification using k-means method based on mel frequency cepstral coefficients. *i-Manager's Journal on Embedded Systems*, 1(1):19.

Kaur, R. and Kaur, B. (2015). Bacterial foraging optimization algorithm for evolving artificial neural networks. *vol*, 8:16–19.

Kolokolov, A. S. (2014). A method for speech signal processing based on band filtering of the logarithmic spectrum. *Automation and Remote Control*, 75(3):496–502.

Kumar, P., Jakhanwal, N., and Chandra, M. (2011). Text dependent speaker identification in noisy environment. In *Devices and Communications (ICDeCom), 2011 International Conference on*, pages 1–4. IEEE.

Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., and Shrishrimal, P. P. (2014). A comparative study of feature extraction techniques for speech recognition system. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(12):18006–18016.

Li, J., Dang, J., Bu, F., and Wang, J. (2014). Analysis and improvement of the bacterial foraging optimization algorithm. *Journal of Computing Science and Engineering*, 8(1):1–10.

Lorenzi, C. (2014). Auditory perception of temporal modulations in sounds [video file]. retrieved from http://savoirs.ens.fr/expose.php?id=1841.

Matsui, T. and Furui, S. (1991). A text-independent speaker recognition method robust against utterance variations. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 377–380. IEEE.

Mezghani, D. B. A., Boujelbene, S. Z., and Ellouze, N. (2010). Evaluation of svm kernels and conventional machine learning algorithms for speaker identification. *International Journal of Hybrid Information Technology*, 3(3):23–34.

Nath, D. and Kalita, S. K. (2015). Composite feature selection method based on spoken word and speaker recognition. *International Journal of Computer Applications*, 121(8).

Nijhawan, G. and Soni, M. (2012). A comparative study of two different neural models for speaker recognition systems. *International Journal of Innovative Technology and Exploring Engineering, ISSN*, pages 2278–3075.

Prahallad, K. (2011). Spectrogram, cepstrum and mel-frequency [video file] retrieved from https://archive.org/details/spectrogramcepstrumandmel-frequency$_6$36522.

Quatieri, T. F. (2006). *Discrete-time speech signal processing: principles and practice*. Pearson Education India.

Rabiner, L. R. and Juang, B.-H. (1993). Fundamentals of speech recognition.

Reynolds, D. A. (1995). Automatic speaker recognition using gaussian mixture speaker models. In *The Lincoln Laboratory Journal*. Citeseer.

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. In *Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on*, volume 4, pages IV–4072. IEEE.

Saksamudre, S. K., Shrishrimal, P., and Deshmukh, R. (2015). A review on different approaches for speech recognition system. *International Journal of Computer Applications*, 115(22).

Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580.

Shahin, I. (2010). Employing second-order circular suprasegmental hidden markov models to enhance speaker identification performance in shouted talking environments. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1):862138.

Shinde, R. and Pawar, V. (2013). Fusion of mfcc & lpc feature sets for accurate speaker identification. *International JOurnal of Current Engineering and Technology*, 3:1763–1766.

Shrawankar, U. and Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. *arXiv preprint arXiv:1305.1145*.

Singh, N., Khan, R., and Shree, R. (2012). Mfcc and prosodic feature extraction techniques: A comparative study. *International Journal of Computer Applications*, 54(1).

Singh, S. and Pandey, P. P. (2003). Features and techniques for speaker recognition. In *M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay submitted Nov*, volume 3.

Svozil, D., Kvasnicka, V., and Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1):43–62.

TIAN, Y., ZHANG, X., and ZHU, R. (2007). Design of waveguide matched load based on multilayer perceptron neural network. *Proceedings of ISAP, Niigata, Japan*.

Venkateswarlu, R., Kumari, R. V., and Jayasri, G. V. (2011). Speech recognition using radial basis function neural network. In *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, volume 3, pages 441–445. IEEE.

Wan, V. (2003). *Speaker verification using support vector machines*. University of Sheffield.

Wang, H. and Xu, X. (2013). Determination of spread constant in rbf neural network by genetic algorithm. *Int. J. Adv. Comput. Technol.(IJACT)*, 5(9):719–726.

Weber, D. and Du Preez, J. (1993). A comparison between hidden markov models and vector quantization for speech independent speaker recognition. In *Communications and Signal Processing, 1993., Proceedings of the 1993 IEEE South African Symposium on*, pages 139–144. IEEE.

Wijoyo, S. et al. (2011). Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot. In *International Conference on Information and Electronics Engineering IPCSIT*, volume 6.

Yee, L. M. and Ahmad, A. M. (2007). Comparative study of speaker recognition methods: Dtw, gmm and svm.

Yujin, Y., Peihua, Z., and Qun, Z. (2010). Research of speaker recognition based on combination of lpcc and mfcc. In *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on*, volume 3, pages 765–767. IEEE.

Yankayis, M. Feature Extraction: Mel Frequency Cepstral Coefficients (MFCC). Retrieved fromhttp://citeseerx.ist.psu.edu/viewdoc/downloadoi=10.1.1.701.6802&rep=rep1&type=pdf

Introduction to Computer Programming with MATLAB [Lecture]. Retrieved from http://www.phon.ucl.ac.uk/courses/spsci/matlab/lect10.html

BioMetroSoft. (2014). The Technology BioMet®Engine: Voice Biometry applied to Security, Law and Medicine.Retrieved from http://www.glottex.com/files/28/Informacion_Tecnica/5/BioMetroSoftSL-poster-14.pdf

Nijhawan, G., & Soni, M. K. (2014). Speaker recognition using support vector machine. International Journal of Computer Applications,87(2).

Zheng, A. (2015). *Evaluating Machine Learning Models: A Beginnerâs Guide to Key Concepts and Pitfalls*. OâReilly Media, Inc.

Zhou, Y. and Gu, Y. (2010). Human speaker recognition based on the integration of genetic algorithm and rbf network. In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2010 2nd International Conference on*, volume 1, pages 239–242. IEEE.

Zhu, J., & Liu, Z. (2014, November). Analysis of hybrid feature research based on extraction LPCC and MFCC. In Computational Intelligence and Security (CIS), 2014 Tenth International Conference on (pp. 732-735). IEEE.

Appendix A

First Appendix

A.1 List of Publications

The two papers titled "*Feature Fusion Techniques Based Training MLP for Speaker Identification System*" and "*Optimizing Classifier Performance for Parkinson's Disease Detection*", have been submitted to the 30th annual IEEE Canadian Conference on Electrical and Computer Engineering (IEEE 2017 CCECE) for publication.