# ROBUST SPEAKER IDENTIFICATION BASED ON NEURAL RESPONSE IN CLEAN AND NOISY CONDITIONS

**MD. ATIQUL ISLAM**

**DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING SCIENCE**

**DEPARTMENT OF BIOMEDICAL ENGINEERING**
**FACULTY OF ENGINEERING**
**UNIVERSITY OF MALAYA**
**KUALA LUMPUR**

**2016**

**UNIVERSITY OF MALAYA**

**ORIGINAL LITERARY WORK DECLARATION**

Name of Candidate: Md. Atiqul Islam [redacted]

Registration/Matric No: KGA140023

Name of Degree: Master of Engineering Science

Title of Dissertation:

ROBUST SPEAKER IDENTIFICATION BASED ON NEURAL RESPONSE IN CLEAN AND NOISY CONDITIONS.

Field of Study: Signal processing.

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature [redacted] Date: 19/08/2016

Subscribed and solemnly declared before,

Witness's Signatur[redacted]

Name: Dr. Ng Siew Cheok

Designation: Senior Lecturer

[redacted]

50603 Kuala Lumpur
Tel: 03 79677627
Fax: 03 79674579
Email: lai.khinwee@um.edu.my

## ABSTRACT

Speaker identification (SID) is a biometric technique of determining an unknown speaker's identity using underlying information of his/her speech utterances. It is very essential for security, crime investigation, forensic test, and telephoning. Robust SID under noisy conditions is still a challenging topic in the field of speech processing. Most of the acoustic-feature-based methods fail to achieve robust SID scores under noisy conditions. However, human performance is very robust in noisy environments. The physiologically-based computational model of the auditory nerve (AN) proposed by Zilany and colleagues (2006), which captures almost all of the nonlinearities observed at the level of auditory periphery, was used in this study to obtain a robust SID performance. A neural-response-based novel feature was proposed in this study for both text-dependent and text-independent speaker identification systems. The proposed feature, referred to as neurogram, was computed from the output of the AN model. The training and testing speech signals were taken from three renowned text-independent datasets (YOHO, TIMIT, and TIDIGIT) and a text-dependent audio speech dataset 'UNIVERSITY MALAYA' to evaluate the performance of the proposed system. The speaker modeling was done using speech signals recorded under clean environment whereas testing was done in both clean and noisy conditions. The testing speech signals were contaminated by adding white Gaussian noise, pink noise, and street noise with signal-to-noise ratios (SNRs) ranging from -5 to 15 dB in steps of 5 dB.

To develop a speaker model, three standard classifiers were employed in this study such as the Gaussian mixture model (GMM), support vector machine (SVM), and Gaussian mixture model-Universal background model (GMM-UBM). The performance of the proposed neural-feature-based speaker identification was compared to the results from the traditional acoustic-feature-based methods, such as the Mel-frequency cepstral

coefficient (MFCC), Frequency domain linear prediction (FDLP) and Gammatone frequency cepstral coefficient (GFCC). Although the classification accuracy achieved by the proposed method was comparable to the performance of those traditional techniques in clean condition, the new neural feature was found to provide lower error rates of classification under noisy conditions, especially under white Gaussian noise. Also, the proposed neural feature provided a consistent performance across different types of noise irrespective of the speech materials used and the duration of the signal. On the other hand, the performance of other existing methods was dependent on the type of noise and the database used. In addition, the performance of the proposed method was classifier-independent, whereas acoustic-feature-based method showed a classifier-dependent performance. For the proposed feature, although it was difficult to assess the effect of each individual nonlinear phenomenon observed at the level of the auditory periphery on the identification accuracy, based on simulation results, it can be inferred that they certainly play important roles in the speaker identification tasks. The proposed feature can also be studied for speech intelligibility, speech recognition, phoneme classification, effects of channel variation on SID, and gender classification.

## ABSTRAK

Identifikasi Penutur adalah teknik biometrik yang menentukan identiti seorang Penutur yang tidak diketahui degan menggunakan maklumat asas daripada isyarat pertuturan. Ia adalah sangat penting untuk keselamatan, penyiasatan jenayah, ujian forensik serta sistem menelefon. Sistem identifikasi Penutur yang mantap dalam keadaan bising masih menjadi suatu cabaran dalam bidang pemprosesan pertuturan. Kebanyakan kaedah yang berdasarkan kepada ciri akustik gagal menunjukkan prestasi yang memuaskan dalam keadaan bising. Walau bagaimanapun, prestasi manusia adalah sangat kukuh dalam persekitaran yang bising. Model pengiraan berasaskan fisiologi daripada saraf auditori yang dicadangkan oleh Zilany dan rakan-rakan (2006), yang menunjukkan hampir kesemua ciri bukan linear yang diperhatikan pada tahap saraf auditori, telah digunakan dalam kajian ini bagi mencapai prestasi yang lebih kukuh. Ciri-ciri berdasarkan tindak balas neural dicadangkan dalam kajian ini untuk kedua-dua sistem identifikasi yang bergantung kepada teks dan teks bebas. Ciri yang dicadangkan, yang juga dikenali sebagai neurogram, telah didapati dari model pengiraan saraf auditori. Isyarat bagi proses latihan dan ujian telah diambil dari tiga set data teks bebas yang terkenal (YOHO, TIMIT dan TIDIGIT) dan satu set data teks yang bergantung kepada ucapan audio set data 'UNIVERSITI MALAYA' bagi menilai prestasi sistem yang dicadangkan. Pemodelan Penutur yang telah dilakukan dalam keadaan senyap dan bising. Untuk keadaan bising, isyarat bagi proses pengujian dihasilkan dengan penambahan bunyi bising putih, 'pink' dan, bunyi jalanan dengan nisbah isyarat-hingar dalam julat antara 5 hingga 15 dB dalam selangan 5 dB.

Bagi proses pemodelan Penutur pula, tiga pengelas piawai telah digunakan dalam kajian ini seperti Model Campuran Gaussian (GMM), Mesin Vektor sokongan (SVM) dan Model campuran Gaussian – latar belakang sejagat (GMM-UBM). Prestasi kaedah yang

dicadangkan dibandingkan dengan kaedah tradisional berdasarkan ciri akustik seperti Pekali Sepstral frekuensi Mel (MFCC), Ramalan Linear Frekuensi (FDLP) dan Pekali Sepstral Frekuensi Gammatone (GFCC). Walaupun klasifikasi ketepatan prestasi kaedah yang dicadangkan adalah setanding dengan kaedah tradisional dalam keadaan senyap, kaedah baru yang ditemui mempunyai klasifikasi kadar ralat yang lebih rendah dalam keadaan bising terutamanya dengan tambahan bunyi hingar 'white' Gaussian. Ciri neural yang dicadangkan juga telah menunjukkan prestasi yang tinggi dan konsisten dalam semua jenis keadaan bising tanpa mengira jenis hingar dan tempoh isyarat. Di samping itu, kaedah ini berbeza dengan prestasi kaedah yang sedia ada yang bergantung kepada jenis hingar dan pangkalan data yang digunakan. Untuk ciri yang dicadangkan, ia adalah pengelas bebas berbanding kaedah lain. Bagi ciri neural, walaupun peranan ciri-ciri bukan linear terhadap ketepatan identifikasi sukar dinilai, tetapi berdasarkan simulasi yang dijalankan, boleh disimpulkan bahawa ciri-ciri tersebut memainkan peranan yang amat penting dalam pengenalan tugas identifikasi. Ciri yang dicadangkan juga boleh dikaji untuk kebolehfahaman pertuturan, pengecaman pertuturan, klasifikasi fonem, kesan variasi saluran pada identifikasi Penutur serta klasifikasi jantina.

## ACKNOWLEDGEMENTS

Bismillahir Rahmanir Rahim

This thesis is the outcome of two years work experiences in auditory neuroscience laboratory (AN) in university of Malaya.

I would like to give thanks from deepened my heart to my honorable supervisor Dr. Muhammad Shamsul Arefeen Zilany and surely this written cannot elicit the true sincerity to him what he disserves. In my sight, he is the best teacher in my life who has not only given direction on relevant studies but also guided me to go forward. He always taught us to do our best that will bring blessing for mankind and forbidden to do work only for applause. He is a person with love, deep knowledge, high patience, dynamism, motivation, enthusiasm and leadership. I could have never imagine, I will get such supervisor and mentor like him. All the achievements of this study were due to his cordial guidance. I would also like to hearty thanks to Dr. Ng Siew Cheok, who has encouraged learning acoustic property of audio signal. I will keep him in my heart forever for his cordial cooperation. I am also gratitude with Dr. Wissam A. Jassim, who always helps me to check in Matlab codes and writing. He always encouraged me as an elder brother. I would like to thanks to my colleagues for their valuable advises.

Last but not least, I am very grateful to my family; especially to my mother without her love, prayers, sacrifices, it was not possible for me to complete my study from abroad.

This study was supported by the grant UM.C/625/1/HIR/152 and RG157-12AET of University of Malaya.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF SYMBOLS AND ABBREVIATIONS

| | | |
|---|---|---|
| AN | : | Auditory Nerve |
| BM | : | Basilar Membrane |
| CASA | : | Computational auditory scene analysis |
| CF | : | Characteristic frequency |
| CFCC | : | Cochlear filter cepstral coefficients |
| CNS | | Central nervous system |
| dB | : | Decibel |
| DNN | : | Deep neural network |
| DCT | : | Discrete cosine transform |
| DFT | : | Discrete Fourier transform |
| ENV | : | Envelope |
| EM | : | Expectation-maximization |
| FFT | : | Fast Fourier transform |
| FDLP | : | Frequency domain linear prediction |
| GF | : | Gammatone filter |
| GFCC | : | Gammatone frequency cepstral coefficient |
| GMM | : | Gaussian mixture model |
| IHC | : | Inner hair cell |
| $V_{IHC}$ | : | Inner hair cell potential |
| LPC | : | Linear predictive coding |
| MAP | : | Maximum a-posteriori |
| ML | : | Maximum-likelihood |
| MLLR | : | Maximum-likelihood linear regression |
| MFCC | : | Mel-frequency cepstral coefficient |

| | | |
|---|---|---|
| MMSE | : | Minimum mean square error |
| MLP | : | Multi-layer perceptron |
| NH | : | Normal hearing |
| OVO | : | One versus one |
| OVR | : | One versus rest |
| OHC | : | Outer hair cell |
| $V_{OHC}$ | : | Outer hair cell potential |
| PLDA | : | Probabilistic linear discriminant analysis |
| RBF | : | Radial basis function |
| SNR | : | Signal to noise ratio |
| SPL | : | Sound pressure level |
| SID | : | Speaker identification |
| SR | : | Spontaneous rate |
| SVM | : | Support vector machine |
| TFS | : | Temporal fine structure |
| T-F | : | Time-frequency |
| UBM | : | Universal background model |
| VAD | : | Voice activity detector |

# CHAPTER 1: INTRODUCTION

## 1.1 Introduction to Speaker Identification System

The speech audible signal conveys a number of important information. It conveys not only the message being spoken, but also the identity of the speaker. Speaker recognition is employed for a wide range of applications such as in banking over a telephone network, voice dialing, voice mailing, database access services, telephone shopping, security control for confidential information, remote access to computers, forensic tests, and information and reservation services. The application of speaker recognition can be divided into two parts: speaker identification and speaker verification. Speaker identification (SID) is a biometric modality which gives a corresponding speaker identity among a set of speakers by analyzing the target speaker utterance and comparing with a set of known speakers' model. On the other hand, speaker verification is to use the voice to verify a certain identity claimed by the speaker (Campbell Jr, 1997). The basic difference between the SID and speaker verification is in the number of decision alternatives. The SID system gives decision alternatives as many as the number of users in the system, whereas the speaker verification is the justification of the correct identity of the identified speaker. Generally, identification of talker (speaker) is done based on text-dependent and text-independent speech samples. In a text-dependent SID (Islam, Zilany, & Wissam, 2016) system, every speaker utters the same text (word, phrase, combination of digits or sentences, paragraph) for several times. When speaker utterances are not constrained to a fixed template, that SID system is known as a text-independent system (Togneri & Pullella, 2011). In a text-dependent SID system, speech recording is basically done by aligning input speech time axis with the reference template from the register speaker. The similarities between the input and template speech are accumulated from initial to final time of the utterance. Depending on the

identity of the speaker, SID system is of two types: close-set and open-set SID (Campbell Jr, 1997). In a close-set SID technique, the unknown speaker belongs to the test data-set. When the tested speaker is chosen from outside of the dataset, i.e., anyone can talk and be tested over the given speaker models is known as an open-set SID system. This study has been done based on the close-set SID technique.

Each and every identification technique consists of the training and testing sessions. Human being is also trained with their inherent language for several years and can recognize the target speaker by hearing his/her utterance speeches. Similarly, speaker identification comprises two individual parts: speaker modeling which corresponds to training part, and testing which corresponds to decision making based on target speaker voice signal score over all speaker models. Once the underlying speaker features are obtained from clean utterances, it is forwarded to a classifier to obtain the speaker training model. To make noisy conditions, test speech samples are corrupted by various stationary and non-stationary noises at various level of signal-to-noise ratios (SNRs). In this research, Gaussian Mixture model (GMM), adopted GMM-Universal background model (GMM-UBM) and Support vector machine (SVM) have been used to make behavioral speaker model to evaluate the performance of the presented SID system.

## 1.2 Research Background

A substantial research has been done on developing both text-dependent and text-independent speech-based SID system for the last few decades. However, it still remains a challenge to provide a robust SID method in noisy conditions due to the deviation of the noisy signal from the clean signal. The acoustic cepstral-feature-based Mel-frequency cepstral coefficient (MFCC) (Davis & Mermelstein, 1980) is a very common and popular system in SID study that is used now-a-days as a baseline feature to benchmark against any newly proposed SID method. Linear predictive coding (LPC)

(Makhoul, 1975) is also a very popular feature for acoustic signal analysis in SID system. For a robust SID system, a probabilistic linear discriminant analysis (PLDA) (Lei, Burget, Ferrer, Graciarena, & Scheffer, 2012) method has been proposed with an i-vector extractor paradigm (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011). Also, cochlear filter cepstral coefficients (CFCCs) (Li & Huang, 2011) based on time-frequency transformation, referred to as auditory transform (AT), achieved a substantial SID improvement over other existing methods in noisy conditions.

To improve the speech quality in noisy conditions for a SID system, a number of algorithms have been introduced. Weiss *et al.* (Weiss, Aschkenasy, & Parsons, 1974) initially proposed a spectral subtractive algorithm in the correlation domain and then in the Fourier transform domain. Boll (Boll, 1979) also developed a de-noising algorithm to improve the speech quality by removing predicted noisy spectrum from the corrupted signal. Zhao *et al.* (Zhao, Wang, & Wang, 2014) has achieved a robust identification rate in noisy conditions through binary masking using a deep neural network (DNN) classifier with bounded marginalization and direct masking. Statistical model based algorithm is one of the best algorithms for de-noising noisy speech signal with maximum likelihood approach (McAulay & Malpass, 1980). Ephraim and Malah (Ephraim & Malah, 1984) has also proposed a de-noising technique using the minimum mean square error (MMSE) estimator of the magnitude spectrum. Recently, a robust SID performance under noisy conditions has been achieved by Zhao *et al.* (Zhao, Shao, & Wang, 2012) based on the computational auditory scene analysis (CASA) and mask estimation technique, which uses a feature called Gammatone frequency cepstral coefficient (GFCC).

A classifier is used to extract speaker distinguishing parameters to make an individual speaker model. Gaussian mixture model (GMM) (D. A. Reynolds, 1995) is the mostly

used classifier for speaker identification due to its ability to capture the relevant speaker distinguishing features. GMM is adapted with universal background model (UBM) to make the speaker model and thus the classifier is named as the GMM-UBM classifier (Zheng, Zhang, & Xu, 2004). It can make very fast and more accurate speaker modeling than using the GMM alone. Support vector machine (SVM) (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998) is also a popular classifier in speaker identification for supervised learning and cross-validation algorithm. In addition, multilayer perceptron (MLP) (Mak, Allen, & Sexton, 1994), GMM-SVM (Togneri & Pullella, 2011), and GMM-i vector-PLDA (Vasilakakis, Cumani, & Laface, 2013) are also extensively used in SID system.

## 1.3 Problem Statement

In this section, the problems with the existing feature to obtain a robust SID performance are discussed. The acoustic cepstral representation-based feature, MFCC, has achieved almost 100% text-independent SID accuracy in clean (quiet) condition (Nakagawa, Wang, & Ohtsuka, 2012; Zue, Seneff, & Glass, 1990), but its performance in noisy conditions (e.g., at a 0 dB SNR) is reduced to a remarkably low level (Chi, Lin, & Hsu, 2012). Fast Fourier Transform (FFT)-based MFCC system loses its robustness with the increment of noise level, since FFT-based spectrogram gets more contaminated by noises as reported in (Chi et al., 2012).

To improve the SID performance under noisy conditions, a 2-D auto-regression-based feature, frequency domain linear prediction (FDLP) (Ganapathy, Thomas, & Hermansky, 2012), has been proposed. FDLP-based system can improve the SID performance for fluctuating noises, but still remains poor compared to the results of MFCC-based method for stationary noise.

It has been observed in (Li & Huang, 2011) that, the cubic operation instead of log-operation on spectral representation of audio signal can significantly improve the SID performance under noisy conditions. Similarly, Zhao and Wang (Zhao & Wang, 2013) has also reported an improved performance in SID score with cubic operation and proved that the performance of the MFCC-based system could have significantly increased using a cubic root operation instead of a log operation in noisy conditions. GFCC-based system has also achieved a robust SID performance by applying a computational auditory scene analysis (CASA) method (Zhao et al., 2012) with individual and combining marginalization and reconstruction module. It has been reported in (Zhao et al., 2012) that, the performance of GFCC-based system without reconstruction or marginalization module in noisy conditions is comparable to the MFCC-based system performances.

In general, the application of de-noising technique makes the system complicated, and also there is no strong evidence that the human auditory system also uses any de-noising technique for identification in noisy conditions. Moreover, most of the existing features do not reflect the non-linear properties of the cochlea, which is obviously represented in the human auditory system. To get a robust SID performance without any de-noising algorithm, cochlear filter time-frequency transform-based feature, cochlear filter cepstral coefficient (CFCC), has been proposed in (Li & Huang, 2011). However, the auditory filters used in their study are linear, and some of the parameters of this method are database-dependent.

All of the above mentioned feature-based methods have shown SID results for individual text-dependent and text-independent datasets, but human auditory system can identify the speaker irrespective of the text-dependent and text-independent speech

samples. In this study, it was also observed that the SID performance of the MFCC- and GFCC-based systems was classifier-dependent.

In a study by (Razali, Jassim, Roohisefat, & Zilany, 2014), it has been mentioned that the SID result based on temporal fine structure (TFS) neurogram (response of AN model) outperformed MFCC-based SID results whereas envelope (ENV) neurogram-based SID scores were comparable with MFCC-based SID performances. The synchronization frequency for ENV neurogram was up to ~156.25 Hz whereas the synchronization frequency for TFS was up to ~ 6.25 KHz. However, computation of TFS responses was highly expensive, and thus has a limited application in real situations.

## 1.4 Motivation

A number of investigations has indicated that the machine performance under clean condition is better than the human listener, even the speech signals are misrepresented (Gallardo, 2015; Kajarekar, Bratt, Shriberg, & de Leon, 2006), while human performance outperforms machine-based SID results under noisy conditions presented by different background noises and communication networks (Kajarekar et al., 2006), and handset variations (Campbell Jr, 1997). However, Human and machine learning method differ in audio content and amount of data. Furthermore, it is well-known that the machine-learning-based SID system performance is largely dependent on extraction of latent information from input audio signal. Feature extraction process is also known as Front-end processing which is mostly done by using various existing cepstral-based features like MFCC, Polynomial linear prediction (PLP), LPC, and a very recent GFCC.

In the Human versus machine speaker recognition study (Wenndt & Mitchell, 2012), a GMM-UBM classifier with MFCC, PLP, and LPCC was used to obtain machine

performance and compared with human performance. Additionally, RASTA filtering, cepstral means and variance normalization, voice activity detection (VAD), and Gaussian super vector were also used. The study was done for 17 known speakers (containing 425 voices) to listeners. The experiment was done for clean and noisy condition introduced by speech-shaped noise for SNR of -20 dB at four different frequency ranges. There were two clean sessions with the same testing sample to check the variation of performance of human listeners and machine.

This study implies that the human and machine performances for speaker identification in clean condition are almost comparable, whereas the performance of acoustic-property-based method drops significantly under noisy conditions. The existing acoustic cepstral feature-based systems fail to achieve a robust SID score under noisy conditions but human performances in noisy environment are comparable to the performances in clean.

The existing SID methods are mostly developed based on acoustic features but do not take into account the non-linearity or physiological behavior of human auditory system. It has been narrated in the literature that the auditory system is nonlinear, and thus the responses have the properties such as the level-dependent gain (compression) (Zhang, Heinz, Bruce, & Carney, 2001), two-tone rate suppression (Sachs & Kiang, 1968), level-dependent rate (Knight, 1972), frequency selectivity (Fletcher, 1940), nonlinear phase responses, and phase locking (Rose, Brugge, Anderson, & Hind, 1967), the shifting of best frequency (the frequency in which response is best) (Patuzzi & Robertson, 1988) to higher level for the higher sound level of auditory nerve in human auditory system.

The AN model proposed by Zilany *et al.* (Zilany, Bruce, & Carney, 2014) that reflects almost all of the above-mentioned non-linearities of human auditory system is a useful

tool for understanding the physiological and mechanical properties of the cochlea. The non-linear properties of cochlear has been motivated to explore the neural feature as a front-end in the present study. To keep pace with the human auditory system SID performance, it is desirable to develop a method that will be effective for both text-dependent and text-independent SID systems. To meet this challenge, a neural-response-based feature has been introduced in this present study.

## 1.5 Objectives of this study

The main objective of this presented study is to design a neural response-based SID system that will be equally effective for both text-dependent and text-independent speech-based speaker identification system. The time-frequency response of the AN model called envelope (ENV) neurogram has been used to obtain a robust SID score in the present study. Since the AN model can capture almost all of the nonlinearities of the peripheral auditory system, it is expected that the SID performance of the proposed method will be comparable to the human listeners' behavioral performances. The objectives of the proposed method are to:

i) Propose a new feature-based method for both text-independent and text-dependent speech signal-based robust SID system,

ii) Study the sound pressure level (SPL) effect, channel bandwidth, band of characteristic frequency (CF), and window size on the robust speaker identification system, and

iii) Make comparison among the obtain results of the presented method to the results from some existing popular feature-based SID methods like Mel-frequency cepstral coefficient (MFCC), Frequency domain linear prediction (FDLP), and Gammatone filter cepstral coefficient (GFCC).

## 1.6 Significance of this Study

Voice speech, finger print, and eye-iris are commonly used as the biometric feature for the recognition of individual user and for security issues. Speech is one of the popular biometric features which gives the talker identity, talker position, speech information, gender identity, emotion, and so on. In the modern age, most people use internet for chatting, banking, shopping, learning, and earning. Everyone has individual password or PIN number or initial respective to different website for individual's safe sign-in. Sometimes it is difficult to remember them and even it may be stolen by someone. In these cases, since the user voice is very available, a text-independent speech based identification system could be employed and there is no need to memorize the speech password. In security issues, voice plays a very vital role to identify the criminal by matching his/her voice with the existing national database including all population voice templates.

A method having similar non-linear properties of the mammal peripheral auditory system has been proposed in this study that is effective for both text-dependent and text-independent speech-based SID systems. The most significant finding of this study is that using only a portion of CF responses (below 1 kHz), a very robust system can be developed which works both in clean and under noisy conditions. In addition, the effects of individual nonlinearity on the SID tasks could be studied using the proposed method which is impossible to accomplish during physiological studies. Since the AN model employed in the proposed study can successfully replicates the responses of the AN fibers for the traumatized people, the proposed metric can be extended to observe the impact of hearing loss on speaker identification tasks.

## 1.7 Thesis formation structure

This thesis has been arranged in the following format.

Chapter 1 introduces the text-dependent and text-independent speech-based SID methods. In the beginning of this chapter, a short description of the existing SID systems and their drawbacks has been discussed and then motivation behind this study and objectives of this research has been chronologically mentioned. At the end, the findings of this study have been discussed.

Chapter 2 describes the anatomy and physiology of the peripheral human auditory system with necessary figures and the phase-locking property of the auditory system. In second stage, a brief history of auditory-nerve modeling has been narrated. The classifiers used in the SID system have been described as well. The existing features which are mostly used in SID methods have been briefly described with the necessary block diagram in the later part of the chapter.

Chapter 3 methodologically explains the presented study. A short overview of this study has been described with block diagram at first. The description and parameter selection of the AN model and construction of neurograms have been discussed. The description of speaker modeling based on GMM, SVM, and GMM-UBM classifier and a brief narration of each data set presented in this study have been given at the end of this chapter.

In chapter 4, the results for each dataset under different noises and in clean condition have been discussed. The results of the proposed method for text-independent speech-based SID system using GMM-UBM speaker modeling and for text-dependent speech-based SID scores based on GMM, SVM, and GMM-UBM classifier have been discussed sequentially. The important findings of the proposed metric have been also

discussed in this chapter. The reason behind the uniqueness and novelty of the proposed method has been discussed in this section.

The summary of the proposed study has been discussed at first in chapter 6. The application of the proposed metric, limitations of the current study and the direction for future development of this study have also been described.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Anatomy and Physiology of the Peripheral Auditory System

The hearing system is very essential for communication. Understanding of auditory system underlying mechanism is essential to develop auditory computational model, hearing aids and other commercial and security purposes. Mammal auditory system is one of the mechanically most sensitive organs in the body that responds to a vibration as small as the diameter of a Hydrogen atom and sensing capability is thousand times faster than the visual photoreceptors (Qing & Mao-li, 2009). Human auditory peripheral pathway is a complex network which acts like a signal receiver and links our sensory organs (the ears) with the auditory parts of the nervous system so that the received information can be interpreted and make intelligible. Sound is a series of pressure waves that propagate through the air with continuous condensations and rarefactions of air molecules (Figure 2.1). When these pressure waves reach into the ear, they travel through the external auditory meatus (the ear canal that accumulates wax) and to the end of the canal where the tympanic membrane (eardrum) is located. The pressure waves that strike the eardrum cause it to move: the condensations push it inward, whereas the rarefaction phase causes it to bulge outward. The basilar membrane responds to the mechanical vibration and transforms it into membrane potential of hair cells. This transformation of energy from one form (mechanical motion) into another form of energy (electrical energy as a change in membrane potential) is known as transduction. This change in the action potential of the hair cells then triggers the action potentials from the auditory nerve (AN) fibers that innervate the hair cells. The action potentials constitute the main code by which the nervous system conveys information. Anatomically, human auditory system is consisted with the outer, middle, and inner ear

that make up the auditory periphery and the final part is the central auditory nervous system. The human auditory pathway is shown in Figure 2.1.

### 2.1.1 The Outer Ear

The outer air is consisted with pinna, or auricle and external auditory meatus (ear canal). The curve shaped structure that situates on both sides of head to accept omnidirectional audio signal is known as pinna. It gives the idea of sound source and direction. After receiving audio signal through pinna it is forwarded and guided through auditory canal and the sound pressure level of that tone is raised in the frequency range of 3 to 4 kHz. The tympanic membrane's necessary humidity and temperature is controlled by auditory canal.

### 2.1.2 The Middle Ear

The Ossicular chain is house in the middle ear. The tympanic membrane and the cavity also located in the middle ear. The Tympanic membrane is a thin layer which is located

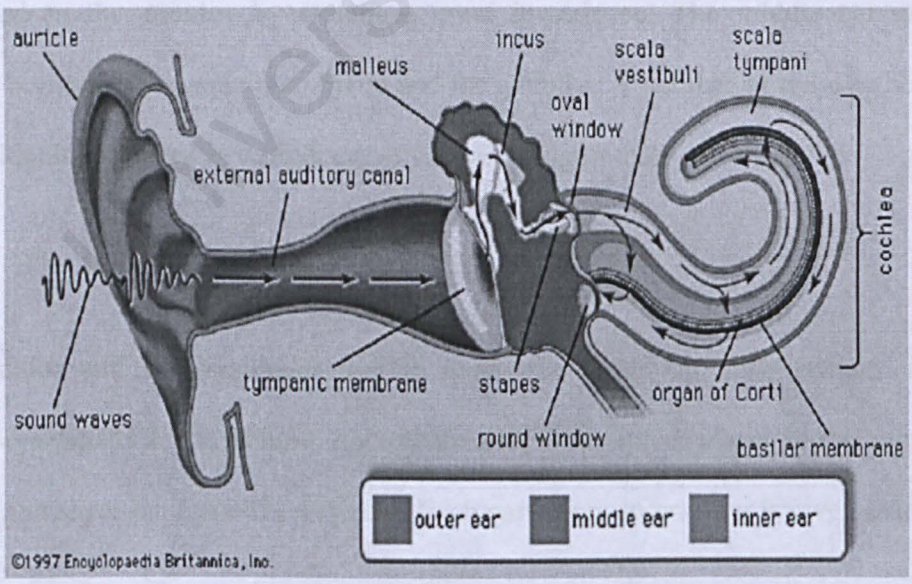


Figure 2.1: Illustration of human peripheral auditory pathway (Encyclopedia Britannica, Inc. 1997).

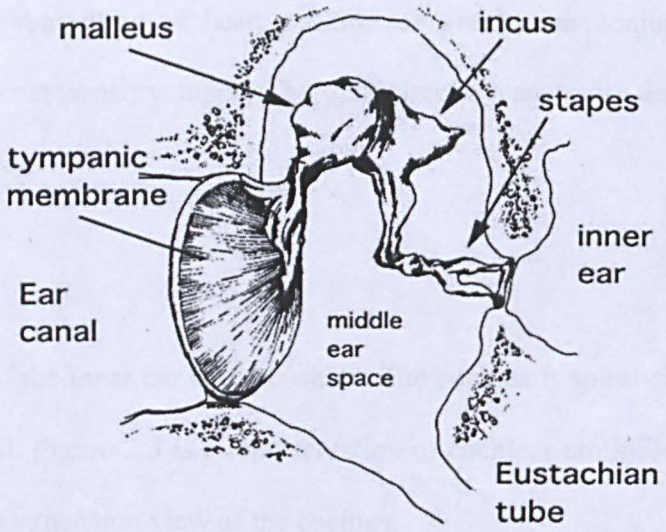


Figure 2.2: The location and conjunction between Tympanic membrane and Ossicle in Human middle ear (Encyclopedia Britannica, Inc. 1997).

between outer ear and middle ear to separate them. It's another name is eardrum. The eardrum transduces the acoustic energy into mechanical energy. Three bones (malleus, incus, and stapes) follow the eardrum and transmit the energy to the inner ear. The mechanical transduction of the Ossicle is transferred to the oval window where it is connected to the cochlea by matching input impedance. The middle ear cavity is situated between Tympanic membrane and the cochlea. A number of delicate bones are filled with air is located in middle ear as shown in Figure 2.2.

### 2.1.3 The Inner Ear

The cochlea and the vestibular system make the inner ear. The cochlea and the vestibular systems are separated, notwithstanding both are confined in the same bony capsule and share the same fluid system. Each part short description is given below.

#### 2.1.3.1 Vestibular or Balance System

The balance part of the ear is referred to the apparatus of vestibular. It is composed with three semicircular canals are located within the inner ear. The vestibular system helps to

maintain balance, regardless of head position or gravity, in conjunction with eye movement and somatosensory input. The semicircular canals are innervated by the eighth cranial nerve.

### 2.1.3.2 Cochlea

The hearing part of the inner ear is the cochlea. The cochlea is spiral-shaped, similar to the shape of a snail. Figure 2.3 is a representation of cochlear cross-sectional view and Figure 2.4 presents expansion view of the cochlea.

The cochlea is composed of three fluid-filled chambers that extend the size of the structure. The two outer chambers are filled with a fluid called perilymph. These two chambers are called Vestibular canal and Tympanic canal. Perilymph acts as a cushioning agent for the delicate structures that occupies the center chamber. It is important to note that perilymph is connected to the cerebrospinal fluid that surrounds the brain and the spinal column. The third fluid filled chamber is the center chamber, called the cochlear duct. The cochlear duct secretes a fluid called Endo-lymph.

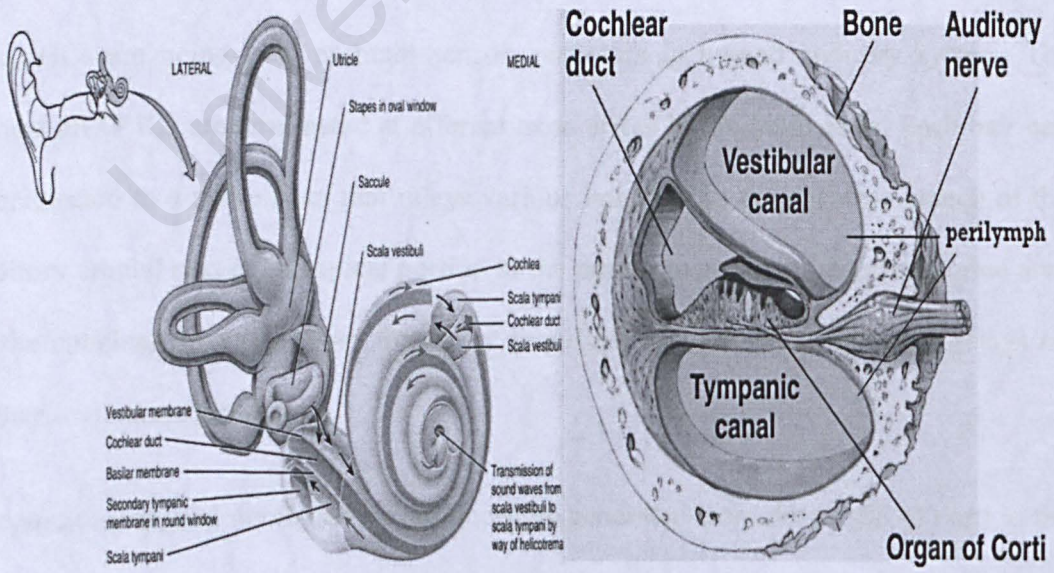


Figure 2.3: A typical mammal cochlear and it's cross-sectional view, (Encyclopedia Britannica, Inc. 1997).

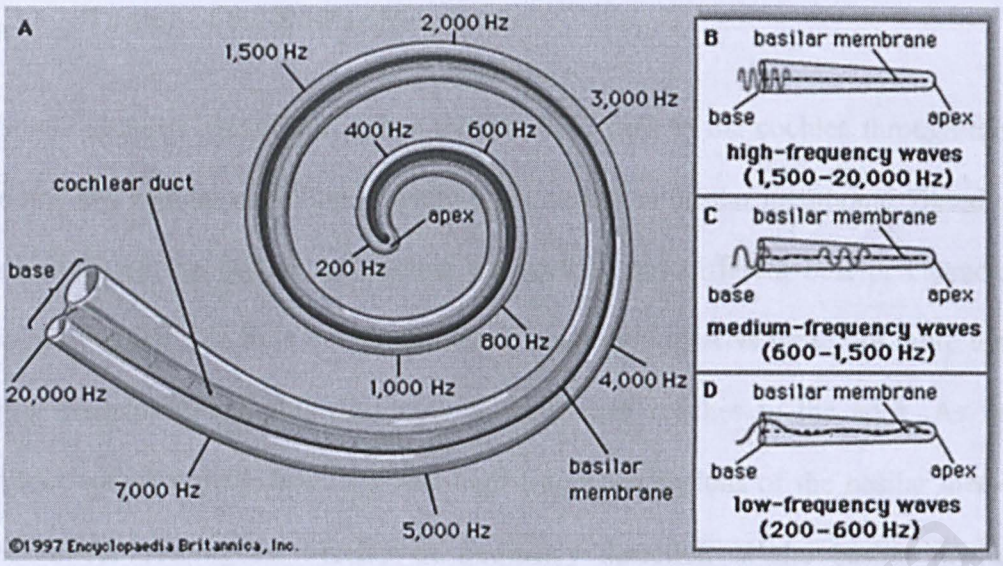


Figure 2.4: Basilar membrane motions at different frequencies (Encyclopedia Britannica, Inc. 1997).

The cochlear duct contains the Basilar membrane upon which lies the Organ of Corti. The Organ of Corti is a sensory organ essential to hearing. It consists of approximately30,000 finger-like projections of cilia that are arranged in rows. These cilia are referred to as hair cells. There are two types of hair cell in human auditory system called: inner hair cell (IHC) (one row) and outer hair cell (OHC) (three rows). The IHCs are being used as main sensory receptors in human auditory system. The almost all OHCs are terminated at efferent axon arises in the brain cells. Each hair cell is connected to a nerve fiber that relays various impulses to the cochlear branch of the auditory cranial nerve. The apical portion of the basilar membrane (the most curled area of the cochlea) transfers lower frequency impulses. The basal end response relays on higher frequency impulses.

The auditory cranial nerve carries the impulses generated from the Organ of corti to the brainstem. From the brainstem, nerve pathways extend through numerous nuclei to the cerebral cortex in the temporal lobes of the brain.

### 2.1.3.3 Tuning of the Basilar Membrane

A sound pressure waveform strikes the eardrum, pass to the cochlea through the oval window and sets up travelling pressure wave along the basilar membrane. Based on the position along the basilar membrane, the cochlea has different best or characteristic frequency (CF). The base of the cochlea is connected to ANs those are more tuned to higher frequency, and this decreases as the length reaches to the apex. As the ear receives sound stimuli, both low and high frequency regions of the basilar membrane are excited, causing an overlap of frequency detection in the basilar membrane. However, the resulting nerve spikes action potential are synchronized based on the low frequency tone (below 5 kHz) through the phase-locking process, which has been successfully captured by the AN model and will be discussed in the next chapter. The motion of the basilar membrane at different characteristic frequencies is shown in Figure 2.4.

### 2.1.4 The phase-locking, spontaneous rate and adaptation property of AN system

The Human auditory system functions are so complicated even the OHC operation can take place without any acoustic input signal through external canal. The reason behind this sudden activation of these cells is not detectable but stimulates basilar membrane. It seems at that time, the basilar membrane is responding to a sound even though there is no acoustic input into auditory system. In this situation auditory neurons start firing indiscriminately which is termed as spontaneous rate and the sound that causes the neuron spontaneous firing called Otoacoustic emission. Poisson-distribution determines auditory nerve fiver's average discharge rate in response to a single pure tone. A single nerve fiber does not fire on every cycle of the stimulus but once spike is obtained, it provides spike roughly at same time interval of the waveform which is called phase-

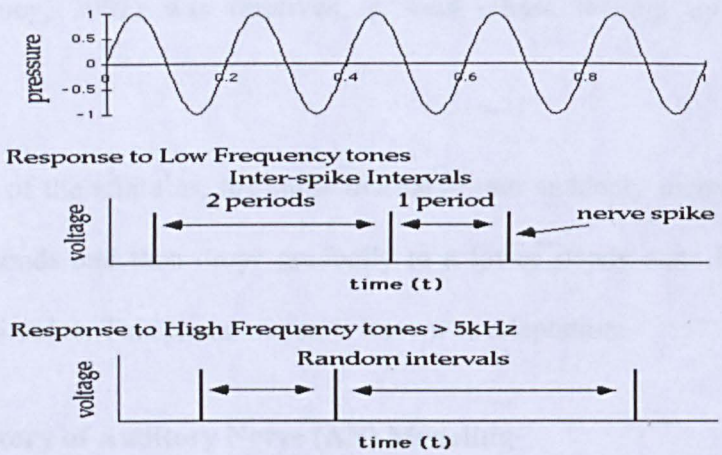


Figure 2.5: A presentation of phase locking property of AN fiber, response to a low and high frequency tone.

locking. In the audible frequency ranges, phase locking does not happen in all over the ranges (Rose et al., 1967). Once the phase locking incur, spontaneous rate become align and phase locked with a particular phase of the stimuli. At very low sound pressure levels, spontaneous activities get phase locked for low stimulus frequencies and it is a threshold at which the spontaneous rate is just increased. Liberman (Liberman, 1978) was observed, these auditory nerve spontaneous rates in human auditory system are divided into three classes. They are: high spontaneous rate-fiber with lowest threshold, medium spontaneous rate-fibers with higher threshold, and lowest spontaneous rate fibers with very highest thresholds. According to the spontaneous rate distribution, the human auditory system synapse responses were approximately 20%, 20%, and 60% to low, medium, and high spontaneous rate respectively as reported in (Liberman, 1978).

Although, the phase locking varies somewhat across species but upper frequency boundary lies at 4-5 KHz (Palmer & Russell, 1986). Javel (Javel, 1980) showed that a portion of the neural activity was phase locked to the overall repetition rate of the stimulus (equal to the absent fundamental frequency). Heinz and his colleagues (Heinz,

Colburn, & Carney, 2001) was observed, a weak phase locking up to 10 KHz frequencies.

During the onset of the stimulus, the spike discharge rate suddenly increases over the first few milliseconds and then drops gradually to a lower steady state for the whole duration on the stimulus. This phenomenon is known as adaptation.

## 2.2 Brief history of Auditory Nerve (AN) Modeling

The auditory system has been the focus of interest of substantial research more than the past half century. To describe the auditory system hearing mechanism and create computational AN model, the knowledge on psychology, physiology, and engineering has been integrated. To simulate quantitative nerve response that explain human auditory system behavior, mathematical and computational models have been built.

AN modeling is one of the useful tools for understanding and testing human auditory peripheral system's underlying mechanical and physiological mechanism. Although there are some complex process such as level dependent gain (compression) (Zhang et al., 2001), two tone rate suppression (Sachs & Kiang, 1968), level dependent rate (Knight, 1972), frequency selectivity (Fletcher, 1940), phase responses, and phase locking (Rose et al., 1967), the shifting of best frequency (the frequency in which response is best) (Patuzzi & Robertson, 1988) to higher level for the higher sound level of auditory nerve in human auditory system, it can easily recognized the talked speech whether the speech is text-dependent, or text-independent that is not subject and also recognized the talker. To capture human auditory system non-linearity and achieve same performance under clean and noisy environment several triumphs have been done to develop auditory computational model. These auditory computational modeling not

only helps for simulating auditory nerve responses but also quantifying the available information in the central nervous system (CNS).

A typical schematic block diagram for data flow in the human auditory system and the corresponding signal processing in a typical auditory computation model is illustrated in Figure 2.6. This block diagram represents the basic concept of auditory periphery pathway modeling.

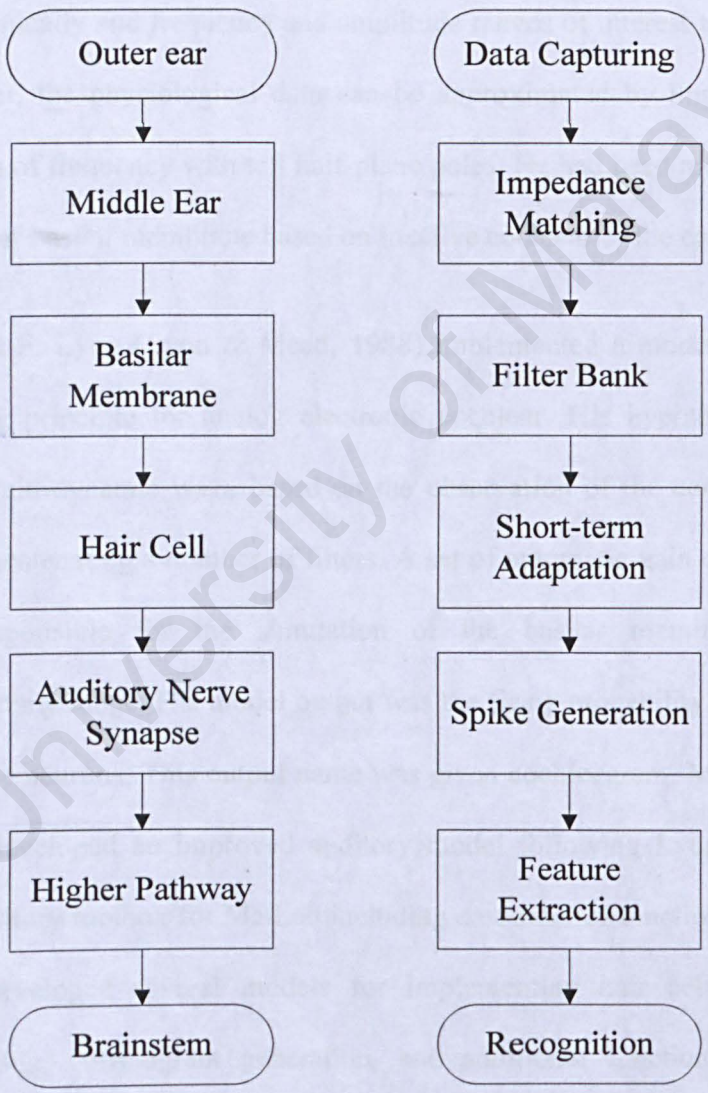


Figure 2.6: Auditory pathway: physiological and functional equivalents (Flanagan, 1960).

Flanagan (Flanagan, 1960) first proposed a mathematical and computational auditory model based on the physiological data surveyed by Békésy. His model was consisted with two major parts: one was the middle ear and another comprised Basilar membrane. The input at middle ear section was the sound pressure at eardrum and output was stapes displacement. The output from middle ear section was the input to Basilar membrane section. The output of the computational model was the displacement of basilar membrane at a specific distance from stapes. Flanagan hypothesized the ear to be passive in mechanically and frequency and amplitude ranges of interest to be linear. He was assumed that, the physiological data can be approximated by implementing the rational functions of frequency with left half-plane poles. He had been modeled inactive cochlear and linear basilar membrane based on inactive cochlear of the cadaver.

In 1988, Richard F. Lyon (Lyon & Mead, 1988) implemented a model based on the cochlear working principle for analog electronic cochlear. His hypothesized was to model cochlea fluid-dynamic wave based on the observation of the cochlear medium properties by concatenating a number of filters. A set of automatic gain control (AGC), which were responsible for the simulation of the basilar membrane dynamic compression intensity range. His model output was the firing probability along time for the auditory nerve neurons. This output name was given cochleagram. Malcolm Slaney (Slaney, 1998) developed an improved auditory model following Lyon's model. He developed an auditory toolbox for MatLab including a number of functions for auditory modeling. He developed several models for implementing hair cell transduction, cochlear processing, correlogram generation, and additional functions for spectral analysis.

To reflect human cochlear non-linearity like two-tone rate suppression, compression, and suppression of tones by noises; Robert and Erikson (Robert & Eriksson, 1999) proposed an auditory model. In order to mimic the cochlear response to complex stimuli, they had taken special care. Their implementation of auditory modeling was simple but provided good approximation about the cochlear mechanism. Their proposed model application area covered development of hearing aids, understanding of signal coding in the cochlea and auditory nerve as well as speech analysis. Figure 2.7 illustrates Robert & Erikson proposed auditory nerve model.

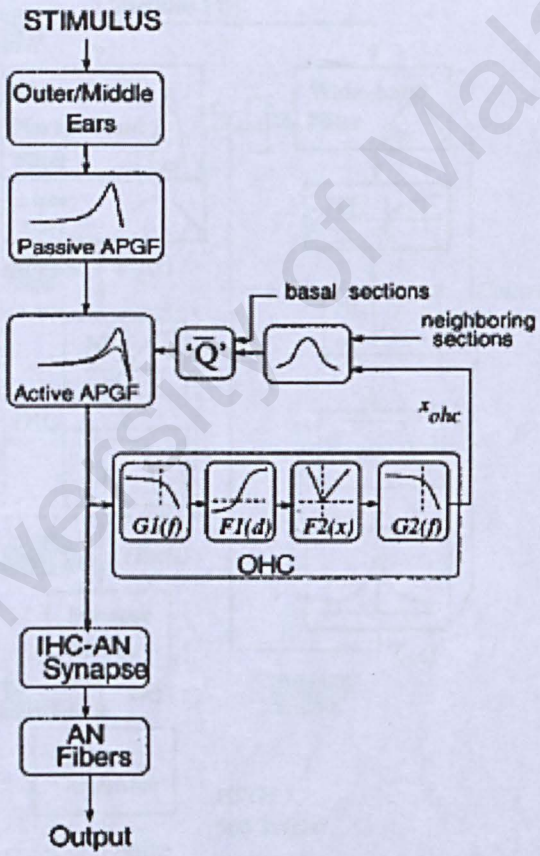


Figure 2.7: Block diagram of AN model of the mammal auditory periphery system (Robert & Erikson, 1999).

A phenomenological model based on AN fibers responses has been implemented by Xuedong Zhang and his colleagues (Zhang et al., 2001). This model provides a useful for understanding the mechanism of cochlear compression, two-tone rate suppression, bandwidth effect, and phase properties in the encoding of population of AN fiber responses to simple and complex tone. The development of this model provides a comparatively simple phenomenological explanation of a single mechanism that contains various essential AN fibers nonlinear response properties. The AN model was developed by (Zhang et al., 2001) has been depicted in Figure 2.8.

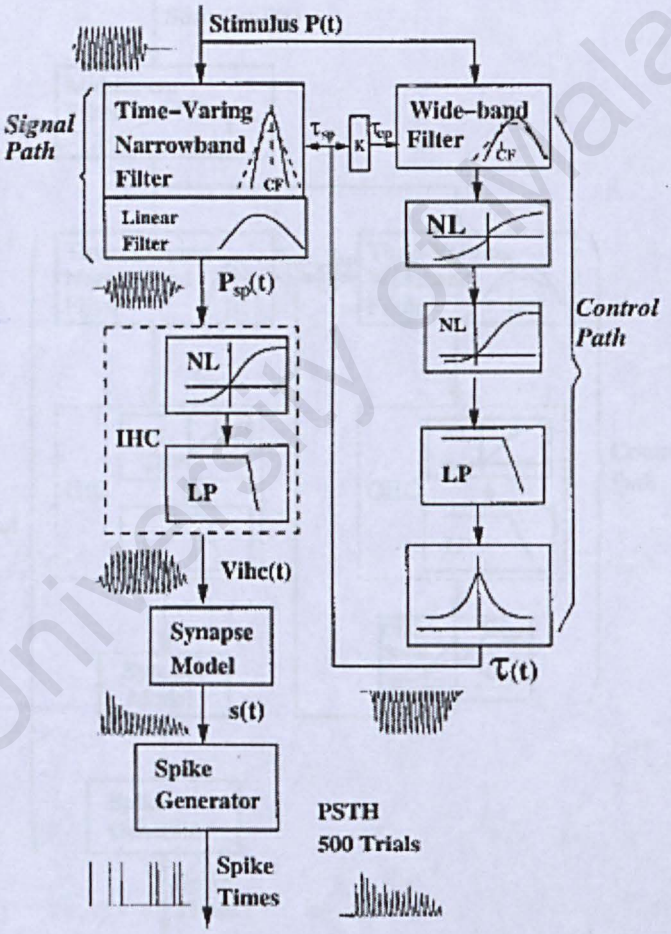


Figure 2.8: Block diagram of auditory nerve (AN) model with non-linear tuning properties of cochlea (Zhang *et al.*, 2001).

Including the effect of acoustic trauma on AN responses Bruce *et al.* (I. C. Bruce, Sachs, & Young, 2003) expanded the previous AN model. They addressed hearing impairment problem and developed a model for both normal and impaired listener. results proved the impairment of both IHC and OHC reduces the tonotopic representation of the auditory nerve formant frequencies. Figure 2.9 illustrates the auditory periphery model was proposed by Bruce *et al.* was implemented an auditory nerve model that can capture almost all the non-linearity properties of human auditory system and was the update version of Bruce *et al.* (I. C. Bruce et al., 2003).

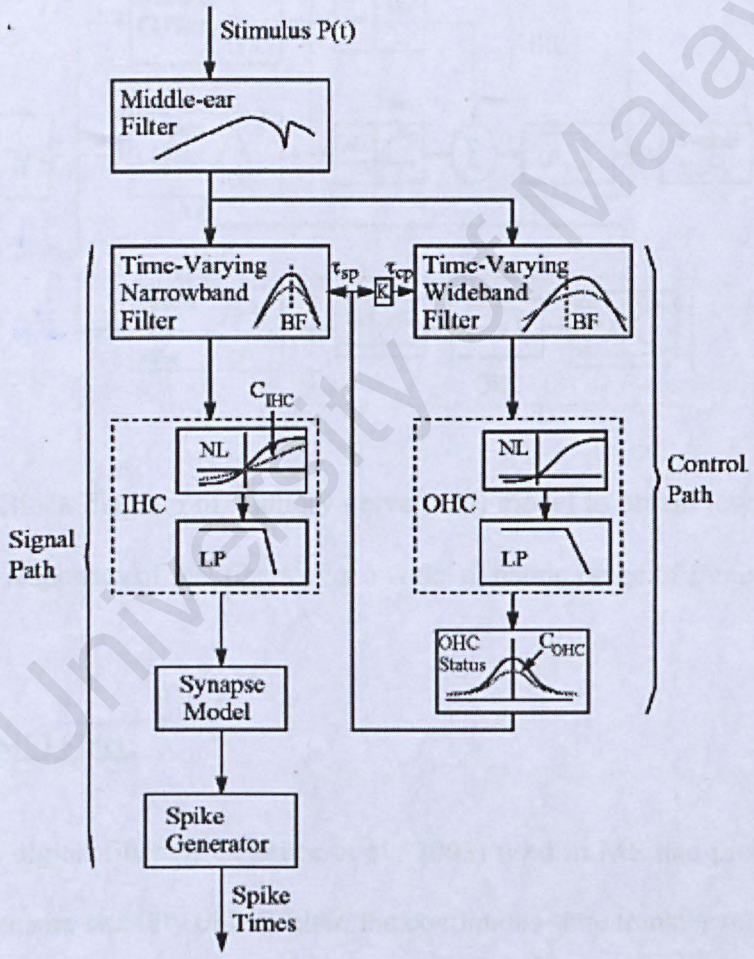


Figure 2.9: Block diagram of auditory nerve (AN) model including acoustic trauma on AN fibers responses (Bruce, Sachs, & Young, 2003a).

### 2.2.1 Description of AN model proposed by Zilany *et al.*

In this section, auditory nerve model has been proposed by Zilany *et al.* (Zilany & Bruce, 2006) has been discussed. Since acoustic sound signal was used as input to this model that was the function of the outer ear, so the model included the phenomenological description of auditory system from middle ear to auditory nerve response. It is to be recalled that, this model has been used in this study to obtain robust SID performance. The description of each part of this model is given below.

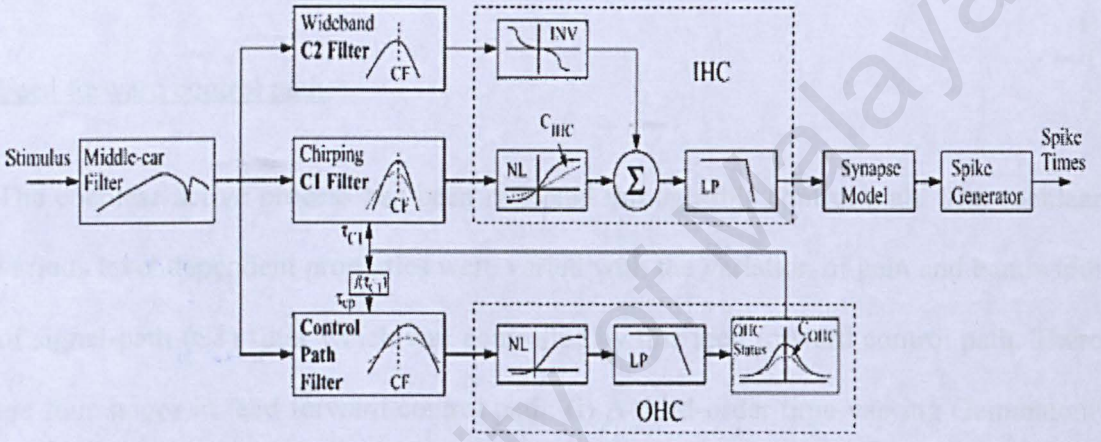


Figure 2.10: Block diagram of auditory nerve (AN) model to obtain low, moderate and high synapse responses of AN fibers over a wide dynamic range of frequencies. (Zilany *et al,*. 2006).

Middle Ear (ME) Filter:

The previous digital filter (I. C. Bruce et al., 2003) used in ME had problem to ensure stability. To ensure stability of ME filter, the continuous-time transfer function has been used to minimize the system order to fifth from eleventh. A bilinear transformation has been applied to create fifth-order digital filter. The instantaneous sound pressure wave has been used as input to the middle ear (ME), sampled at 100 KHz. Because of stability

problem, fifth order filter was implemented by cascading second order system as following.

$$ME1(z) = 0.0127\left(\frac{1.0000 + 1.0000z^{-1}}{1.0000 - 0.9986z^{-1}}\right) \quad (2.1)$$

$$ME2(z) = \frac{1.0000 - 1.9998z^{-1} + 0.9998z^{-2}}{1.0000 - 1.9777z^{-1} + 0.9781z^{-2}} \quad (2.2)$$

$$ME3(z) = \frac{1.0000 - 1.9943z^{-1} + 0.9973z^{-2}}{1.0000 - 1.9856z^{-1} + 0.9892z^{-2}} \quad (2.3)$$

Feed forward control path

The cochlear active process has been reflected through this control path. The cochlear various level dependent properties were varied with the variation of gain and bandwidth of signal-path (C1) filter which was controlled by this feed forward control path. There are four stages in feed forward control path: (i) A third-order time-varying Gammatone filter which bandwidth was broader than the signal-path C1 filter was implemented following (Zhang et al., 2001) and (I. C. Bruce et al., 2003), (ii) a non-linear function was created by Boltzmann function, (iii) a second-order low pass filter, and (iv) a non-linear function to get time varying time constant for C1 filter from low-pass filter.

The advantage of having broader bandwidth of Gammatone filter was that, it can produce two-tone suppression rate property of cochlear at the model output. Depending upon the tuning property of signal-path filter, the tuning of the control-path Gammatone filter has been fixed. The maximum and minimum time constant of the control path filter has been computed by following equations.

$$\tau_{cpmax} = \tau_{wide} + 0.2 \times (\tau_{narrow} - \tau_{wide}) \quad (2.4)$$

$$\tau_{cpmin} = \tau_{cpmax} \times \text{ratio} \quad (2.5)$$

Where $\tau_{narrow}$ and $\tau_{wide}$ are indicating the signal-path Gammatone filter maximum and minimum time constant from earlier model (I. C. Bruce et al., 2003) and determined as follows.

$\tau_{narrow} = 2Q_{10}/(2\pi CF)$, where CF is in Hz unit and $\tau_{wide} = \tau_{narrow} \times$ ratio,

where ratio $= 10^{-gain_{CA}^{(CF)/(20 \times 3.0)}}$.

The gain of the control path filter was varied with the variation of characteristic frequency (CF) according to following equation.

$$gain_{CA}(CF) = \max\{15,52(\tanh(2.2\log_{10}(CF/10^3) + 0.15) + 1.0)/2\} \quad (2.6)$$

Here, CF is given in Hz unit. The Gammatone filter asymmetric non-linearity was done with second-order Boltzmann function by following equations.

$$BN(V) = (1/\ 1\text{-shift}_{cp}) \times \{(1/(1 + e^{-(V-x0)/s0}(1+e^{-(V-x1)/s1}))) \text{-shift}_{cp}\} \quad (2.7)$$

Where BN and V represent the output of Boltzmann function and wide-band filter respectively. A second-order low pass filter with cut-off frequency 600 Hz has been introduced in feed forward control path. This filter's output was used to generate time varying time constant for signal-path filter through non-linear function to keep patch with (I. C. Bruce et al., 2003; Zhang et al., 2001). Time varying time-constant of the control path filter was computed through following equation

$$\tau_{cp} = f(\tau_{C1}) = a.\ \tau_{C1} + b \quad (2.8)$$

Where a and b can be found through $a = (\tau_{cpmax} - \tau_{cpmin})/(\tau_{C1max} - \tau_{C1min})$ and $b = \tau_{cpmax} - a.\tau_{C1max}$.

The control path output at low sound pressure level was almost equal to estimated time constant of C1 filter at low level was such that the gain was high, sharp tuning and response was linear. The control path output was deviated largely from signal path estimated time constant at low level for moderate sound pressure level. The C1 filter become effectively linear and reduced gain at low level at high sound pressure level when control signal saturates. To reflect cochlear suppression and compression non-linearity, the signal-path filter C1 tuning become broader and reduces the gain.

C1 Filter

The input for the inner hair cell (IHC) transduction function was obtained from the model Basilar membrane (BM) response which tuning properties are determined by signal-path C1 filter. This filter was implemented with a fifth-order zeros on the real axis and two second-order poles and one first-order pole, their complex conjugates on the imaginary axis. This filter configuration was selected by following the AN model proposed by Tan and Carney (Tan & Carney, 2003). The sharpness of tuning of a filter was affected by its order. C1 filter provides fairly sharp tuning even for high-sound pressure level stimuli or in the case of outer hair cell (OHC) impairment when the filter order is too high. The filter order has been reduced from 20 to 10 to make the impaired and high-level tuning more reasonably broadly tuned. The auditory nerve (AN) model has been presented here can simulate AN response with CFs up to 40 kHz which was ten (10) times more than the simulated response in model proposed by (Tan & Carney, 2003). The CF-dependent pole-zero locations for low-level stimuli are governed by the required $Q10$ values and tuning curve tail shape govern the locations of CF-dependent pole and zero. To replicate the tuning of auditory nerve, the C1 filter's poles and zeros positions have been carefully chosen.

The shape of tuning curves of neural gradually changes with CF as physiological studies showed in (N. Y.-S. Kiang, 1965). A shallow, symmetric tuning have for low-CF and high-CF fibers have sharp, asymmetrical tuning with extended low-frequency tails. Single AN fibers therefore appear to behave as band-pass filters, with asymmetric filter shape.

In this presented AN model, the C1 filter frequency glides in the impulse response show downward glides for CFs below 750 Hz, constant glides for CFs ranging from 750 to 1500 Hz, and upward glides for CFs above 1500 Hz, which are qualitatively consistent with the AN data (Carney, McDuffy, & Shekhter, 1999).

C2 Filter

To reflect the Kiang two-factor cancellation hypothesis a filter parallel to C1 filter has been used in this presented model to accomplish peak-splitting phenomena and C1/C2 transition. The C2 filter has two important features that support a number of literatures (Gifford & Guinan Jr, 1983; M. C. Liberman & Klang, 1984; Sewell, 1984a; Wong, Miller, Calhoun, Sachs, & Young, 1998).

First, The tuning curve of C1 filter was almost flat as reported by (M. C. Liberman, and Kiang, 1984). The levels at which the responses of the fiber experience an sudden phase shift of about 180° as a function of the stimulus frequency defines the thresholds of a C2 tuning curve for a particular fiber. A tenth-order C2 filter has been chosen to achieve the broadest possible form of signal-path C1 filter. To achieve this broadest possible C1 filter, C2 filter's poles and zeros has been placed at same position of C1 filter at the complete OHC impairment condition.

Second, The C2 responses were less sensitive to decrease in endocochlear potential (EP) production by furosemide as showed by (Sewell, 1984b). The crossed olivocochlear

bundle stimulation can put down C1 responses whereas C2 responses were not affected as reported by (Gifford & Guinan Jr, 1983). All of these studies support that the C2 filter response was independent of OHC function whereas C1 response was dependent. To implement these findings in AN model, C2 filter was made linear, static, and is followed by a separate IHC transduction function.

Inner Hair Cell (IHC)

The conversion of BM mechanical response to electrical potential that leads to neurotransmitter release across the IHC-AN synapse was done by IHC function. The C1 response was generated by the tallest row of IHC stereocilia whereas the shorter IHC stereocilia is responsible for C2 response as found in acoustically traumatized cats (M. C. Liberman, and Kiang, 1984). So, acoustic trauma has little effect on C2 responses but can dropped or removes C1 response. The IHC model was the modified version of previous models (I. C. Bruce et al., 2003; Zhang et al., 2001). The IHC section was the combination of two portions. First, the summation of two transduction outputs from C1 and C2 together and pass the output through low-pass filter IHC to provide potential Vihc, of IHC . A seventh-order low-pass filter has been used in IHC section with a cut-off frequency of 3800Hz.

Synapse model and discharge generator

The synapse model and discharge generator was the replicate form of previous model proposed by (Zhang et al., 2001). The synapse model was responsible to reflect the adaptation properties, spontaneous rate, and the rate-level behavior at the output of AN model. The ratio of the IHC potential to the synaptic release rate was varied with the variation of CF to adapt the empirical data in AN model.

The discharge generator section has been adopted in this model from previous models proposed by (Zhang et al., 2001) and (I. C. Bruce et al., 2003).

Modeling OHC and IHC impairment

It has been observed in a number of literature (N. Y. Kiang, Liberman, & Levine, 1976; M. C. Liberman & Dodds, 1984) that the tuning curves of AN fiver thresholds are broaden and arisen up due to damage of OHC stereocilia but the damage of IHC stereocilia is only liable for lifting of the tuning curve as observed by (M. C. Liberman & Dodds, 1984). The OHC and IHC section of AN model has been modified for hearing impairment in the cochlea (I. C. Bruce et al., 2003). The effects of the OHC status were incorporated in the model by introducing a scaling factor $0 \leq C_{OHC} \leq 1$ to the control path output as proposed by (I. C. Bruce et al., 2003). $C_{OHC}=1$ simulates the normal functioning of OHC in the model and $C_{OHC}=0$ indicates complete impairment in the OHC. Depending upon the degree of impairment, control signal has been modified according to following equation

$$T_{C1\text{-impaired}} = C_{OHC}\,(\tau_{C1} - \tau_{C1min}) + \tau_{C1min} \qquad (2.9)$$

The C1 transduction function is affected by the impaired cochlea, and thus impairment in the IHC has been addressed in the model only for C1 transduction function by introducing a scaling constant $0 \leq C_{IHC} \leq 1$, comply to (I. C. Bruce et al., 2003). The value of constant $C_{IHC}=1$, corresponds to the normal functioning of IHC and $C_{IHC}=0$ indicates complete impairment.

This model was further developed in 2009 (Zilany, Bruce, Nelson, & Carney, 2009) and 2014 (Zilany et al., 2014) respectively.

## 2.3 Neurogram and Spectrogram

The cochlea's auditory nerve filters decompose the complex broadband sounds into a series of relatively narrowband signals, each of which can be considered as a slowly varying envelope (ENV) superimposed on a more rapid temporal fine structure (TFS). Although TFS information depends on phase locking to individual cycles of the stimulus waveform, both ENV and TFS information are represented in the timing of neural discharges (Moore, 2008). Depending on the dominant fluctuation rates of audio

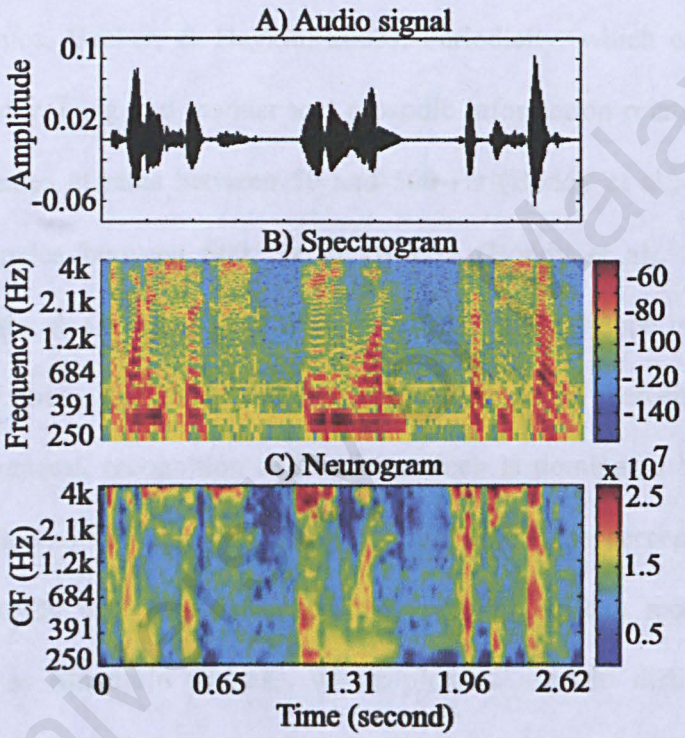


Figure 2.11: Time-frequency representations of speech signals. (A) a typical speech waveform taken from the YOHO database (to produce spectrogram and neurogram of that signal), (B) the corresponding spectrogram responses, and (C) the respective neurogram responses.

signals, (Rosen, 1992) separates the temporal features of speech into three primary categories: envelope, periodicity, and TFS. AN model has been used for robust phoneme classification (Alam, Jassim, & Zilany, 2014), speaker identification (Islam et al., 2016) and verification (Razali et al., 2014), and speech intelligibility (Mamun, Jassim, & Zilany, 2015).

Envelope, which fluctuates at a rate (modulation frequency) between 2 and 50 Hz, conveys information of manner of articulation, voicing, vowel identity, and prosodic cues (Bondy, Bruce, Becker, & Haykin, 2003). Periodicity, which carries segmental information about voicing and manner and prosodic information relating to intonation and stress, fluctuates at rates between 50 and 500 Hz (Bondy et al., 2003). TFS has dominant frequencies between 600 Hz to 10 kHz (Bondy et al., 2003). Smith *et al.*(Smith, Delgutte, & Oxenham, 2002) observed that, the envelope is most important for speech perception, and the TFS is important for pitch perception and sound localization. In general, recognition of English speech is dominated by the envelope, whereas recognition of melody is dominated by the TFS. Pitch perception should also help convey prosody cues in speech and may enhance speech reception for tonal languages, such as Mandarin Chinese, where pitch is used to distinguish different words. It was observed in (Xu & Pfingst, 2003) that, the lexical-tone recognition depends on fine structure, not on envelope, when the number of frequency bands was between 4 and 16.

All of aforementioned information was motivation to explore the ENV neurogram in robust SID system. The basic difference between spectrogram and neurogram is that, the spectrogram is the frequency spectrum respective to time for an acoustic signal using FFT whereas neurogram is the collective 2-D neural responses i.e. the collection

of responses to corresponding CF. The neurogram and spectrogram presentation to an acoustic signal is illustrated in Figure 2.11.

## 2.4 Classifiers in Speaker Modeling

The most challenging part in SID is speaker modeling. The most successful classifiers those are substantially used in speech and speaker recognition are SVM, GMM, GMM_UBM, DNN, HMM and SVM_GMM. Among the mentioned classifier DNN (deep neural network) is most computationally expensive but provide comparatively better recognition performance. In the following session, the most commonly used classifier SVM, GMM and GMM-UBM-based speaker modeling will be discussed.

### 2.4.1 Support Vector Machine (SVM)

Support vector machine (SVM) is a good classifier for speaker recognition and guileless compare to neural network (Wang, Liu, Xing, & Li, 2008). A Support Vector Machine (SVM) is a discriminative statistical classifier formally defined by a separating hyper-plane. It effectively constructs boundaries for linear or nonlinear classification and is able to yield a sparse solution through the so-called support vectors. The support vectors are the observation data which indicates that, it is perfectly classified data within boundary or not on the classification boundary. The extracted features (train-data and test-data) are forwarded to SVM classifier for speaker modeling and testing the test samples with the speaker models for speaker identification.

SVM is basically a binary classifier, which compares each sample to another sample. There are two types of classification mode in SVM: one versus one (OVO) and one versus rest of samples (OVR). The advantage of OVO classification system is that, it takes less time since the problems to be solved are smaller in size. In OVR system, each test instance is compared with the target one samples and rests all instances at same

time. In practice, heuristic methods such as the OVO and OVR approaches are mostly used than other multiclass SVM implementations. The big reason to use them havoc, there are several online software packages available that efficiently solve the binary SVM, such as (Cortes & Vapnik, 1995).

The performances of SVM classification is largely depend upon the scaling of data. The number of observation should be arranged in row wise for whole set of data. In SVM classification, training data set and testing data set were distinguishingly scaled using 'mapminmax' which normalized data on feature-row based. The default minimum and maximum value is 0 and 1 respectively. The used data should be scaled by frame otherwise there have most chance to be biased rest of samples data by first one. The train-data and test-data should be normalized by same ranges otherwise the performance of classification will reduce significantly.

The SVM is a supervised classifier because it produces a speaker model on the basis of training data with their corresponding target values and gives a target value for a tested sample as well as identification rate. Every instance or samples whether from training or testing has a target value (class label). Initially training data maps into a higher dimensional spaced by the kernel function. Four types of kernel function are used in SVM. The default type of kernel function (Radial basis function (RBF)) is has been used in presented SID classification. The RBF function between two samples $x_i$ and $x_j$ can be mathematically defined as following form

$$K(x_i, y_i) = e^{(-\gamma \|x_i - y_i\|^2)} \tag{2.10}$$

Here $x_i$ and $y_i$ are different samples where $i = 1,2,\ldots\ldots l$ is the number of instances, $\gamma$ is the gamma parameter defines as how far a single training instance reaches i.e. for low value of $\gamma$ refers to far and close of influence of training instances with higher value.

RBF maps training data into high dimensional space by nonlinearly and can handle attribute which were nonlinearly related to target values. Another reason to use RBF function is that, it needs comparatively less number of hyper-parameters than polynomial kernel which classifies training data from each other.

There are two parameters C and $\gamma$ are associated to RBF kernel. The C (penalty parameter) prevents misclassification of training data against plainness of decision surface. A small value of C is liable for ingenuousness of decision surface while large value gives freedom to select large data as support vector to the model (Soong, Rosenberg, Juang, & Rabiner, 1987). Too small value of gamma ($\gamma$) makes the model constrain that cannot handle complex data. It implies that, the classification performance of SVM model with RBF kernel fully depend on a proper selection of C and $\gamma$.

Using cross validation (Suykens, Van Gestel, De Brabanter, et al., 2002) algorithm, best C and $\gamma$ has been chosen in which it given best accuracy. Cross validation is a classifying procedure in which each instance is independently classified. In cross validation technique, the train data is divided randomly in roughly equal size to 10 (K) parts. Then, each of these parts is tested once as a test set against k-1 set, where k-1 sets are combined in a training set. The algorithm is repeated for 10 times for each fold with trained on the training set and tested on the test set.

Over fitting of training data is a problem to better classification as well as better accuracy. This over fitting problem can be obstructed using cross validation procedure. It has been noticed in this study that, C and $\gamma$ values give better performance for 1and 3 respectively. In recall, training data set are partitioned into equal D subset and each subset was tested with rest D-1 subsets. Testing data set was compared with speaker model to predict the testing instance lied in which target value (class label). The

obtained target value for each instance was compared with the given test label and the correctly identified target value was given the identification score.

The above described classification gives correct identification rate for single dimension with single observation. When multi-dimension data are used, the embedded system treats each observation as an instance and don't consider the dominating number of instances. To overcome this problem with multi-class SVM a confusion matrix has been added and each test data sample length was saved.

### 2.4.2 Gaussian Mixture Model (GMM)

The most generic statistical speaker modeling paradigm one can adopt and extensively used for speaker modeling is Gaussian Mixture Model (GMM) (D. A. Reynolds, Quatieri, & Dunn, 2000). A GMM classifier assumes the feature vectors follow a Gaussian distribution, characterized by a mean and a deviation about the mean. With advances in the parameter estimation, computations and scoring of these models, they remain one of the most widely used now-a-days (D. A. Reynolds et al., 2000). The Gaussian mixture model for speaker $j$ and a weighted sum of $M$ component densities $\lambda j$ is governed by the output probability expression (for a given feature vector, $\overrightarrow{x_t}$).

$$p(\vec{x}_t|\lambda_j) = \sum_{i=1}^{M} g_i \mathcal{N}(\vec{x}_t; \vec{\mu}_i, \sum_i) \tag{2.11}$$

Where $g_i$ are the mixture weights satisfying $\sum_{i=1}^{M} g_i = 1$.

The GMM model for speaker, $\lambda j$ is parameterized by the mean vectors, covariance matrices and mixture weights from all $M$ component densities.

Why does GMM classifier so successful in speaker recognition? The first and main reason is that, nevertheless with enough mixtures (on the order of 64 or more), the

component densities may be able to represent the individual speaker's broad phonetic class distribution. Under the GMM framework the overall model provides a smooth transition from one acoustic class (or mixture) to the other via the linear weighting function thus making the system text-independent in nature.

Another reason for the success of the GMM is the availability of a powerful and versatile parameter estimation paradigm: the expectation-maximization (EM) algorithm. A key feature of the EM algorithm is that it can guarantee monotonic convergence to the set of optimal parameters (in the maximum-likelihood (ML) sense) in only a few (5 or so) iterations.

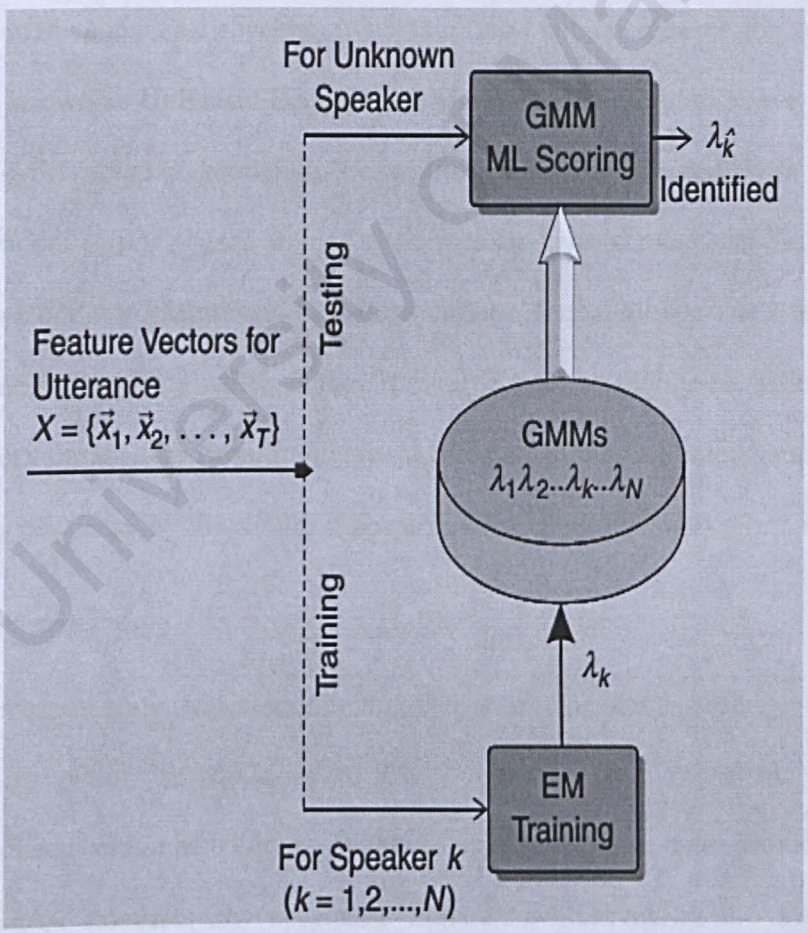


Figure 2.12: Block diagram of speaker training and testing framework for N speakers using GMM modeling. N represents the number of speakers used in speaker identification (SID) system (Togneri & Pullella, 2011).

Although GMMs are quite powerful in speaker identification, they do suffer from two important drawbacks. First, it needs enough training data to properly estimate the model parameters. A common trick is to use diagonal covariance matrices rather than "full" covariance matrices to solve above mentioned problem. The second problem with a GMM, as with any generative modeling paradigm, is that data unseen in the training which appears in the test data will trigger a low score on that data and degrade the overall system performance. The whole operation technique for speaker identification using GMM classification technique has been illustrated in Figure 2.12.

### 2.4.3 Gaussian Mixture Model-Universal Background Model (GMM-UBM)

Sometimes GMM models all speakers other than the claimed speaker for verification, and then it is known as Universal Background Model (UBM). Due to above mentioned criteria of GMM speaker modeling in section 2.3.3, is adapted with Universal background model (UBM)-based trained each speaker data to make the system faster, stable and have better performance. UBM can capture almost all speaker's feature with inadequate training samples which makes it reliable than any other classifier. EM can derive necessary parameters from little amount of data and the estimated parameters can be adapted to new data by maximum a-posteriori (MAP) adaptation (S. Young et al., 2002).

The most successful statistical classifier that can adopt speaker modeling paradigm is GMM. The component distributions of GMM classifier can represents individual speaker's phonetic class distribution without enough mixture and provide smooth transition through mixtures by weighting function which makes the system text-independent as defined by the equation (2.12).

$$p(\vec{x}_t|\lambda_j) = \sum_{i=1}^{M} g_i \mathcal{N}(\vec{x}_t; \vec{\mu}_i, \sum_i) \qquad (2.12)$$

The application of expectation maximization (EM) algorithm in GMM-based speaker modeling is another reason to make this classifier successful in SID system. The main advantage of EM is that, it can convergence speaker's feature data to optimal parameters set in very little iteration.

Due to aforementioned criteria of GMM in speaker modeling in literature review section, it is adapted with universal background model (UBM)-based trained each speaker data to make the system faster, stable and have better SID performance. UBM can capture almost all speaker's feature with inadequate training samples which makes it reliable than any other classifier (Togneri & Pullella, 2011).

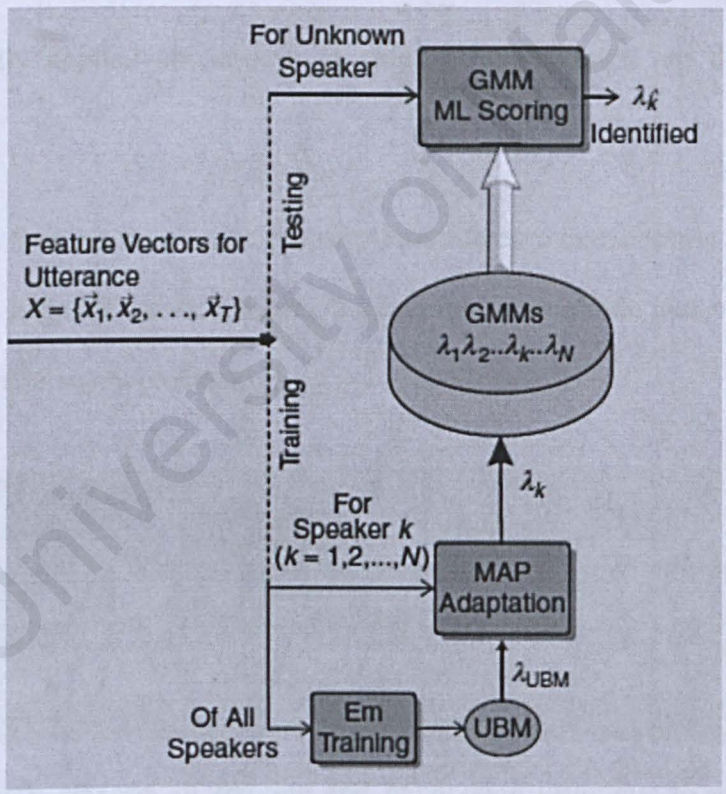


Figure 2.13: Block diagram of speaker training and testing framework for N speakers using GMM-UBM speaker modeling. N represents the number of speakers used in speaker identification (SID) system (Togneri & Pullella, 2011).

## 2.5 Existing Metrics in Speaker Identification

Speaker identification has been an important topic for research for last few decades in the speech processing field and still is a challenging issue to recognize an unknown speaker in noisy condition. Linear predictive coding (LPC) (Makhoul, 1975), and prosodic polynomial coefficient are very popular feature to acoustic signal analysis for SID system. In SID robustness issue, probabilistic linear discriminant analysis (PLDA) (Lei et al., 2012) method has been proposed with i-vector extractor paradigm (Dehak et al., 2011) and cochlear filter cepstral coefficients (CFCCs) (Li & Huang, 2011) based on time-frequency transformation named auditory transform (AT) got substantial improvement in noisy condition. In this session MFCC, FDLP, GFCC, and CFCC; which are mostly applied now-a-days in speaker identification are described very shortly.

Windowing and framing in acoustic property-based feature extraction is very common for rounding the edge information. Figure 2.14 represents the basic idea of window and frame application in signal processing.

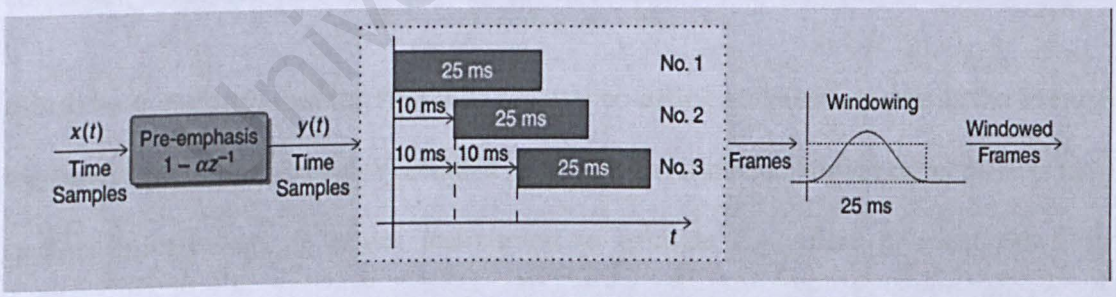


Figure 2.14: Block diagram of framing and windowing technique to an input audio waveform.

### 2.5.1 Mel-frequency Cepstral Coefficient (MFCC)

The acoustic cepstral-based feature Mel-frequency cepstral coefficient was first introduced by Davis and colleagues in 1980 (Davis & Mermelstein, 1980). The spectral-

based features which are commonly derived by the direct application of Fast Fourier transform (FFT) to the acoustic signal had become very popular over two past decades. It was shown in (D. Reynolds, 1994) that the same feature is equally successful in speech as well as in speaker recognition. This feature is spectral logarithmic energy-based Mel-frequency cepstral coefficient (MFCC). MFCC-based feature extraction is very easy and fast. It performance in quiet condition is also satisfactory that makes it popular and mostly used as a baseline feature to compare with newly introduced method. Figure 2.15 represents the derivation procedure of MFCC- feature from a typical acoustic sound waveform. Initially a FFT is applied to each frame to obtain complex spectral features. Normally 512-points FFT is applied to derive 256-points complex spectral without considering phase information. To make more efficient information representation only 30 or so smooth spectrum per frame is considered and scaled logarithmically to Mel or Bark scale to make those spectrum more meaningful. In MFCC derivation linearly spaced triangular filter-bank is used. The final step is to convert the filter output to cepstral coefficient by using discrete cosine transform (DCT). Spectral coefficients are highly correlated but the cepstral coefficients are de-correlated which are very essential for speech or speaker recognition.

It is to be mentioned that the first ($C_0$) cepstral co-efficient which represents the average log-power of the frame is also included in MFCC coefficients. However, as there is little speaker information, it is not uncommon to exclude $C_0$; where in most cases, its inclusion is by "default". The first derivative and fifth derivative of extracted cepstral is taken which are generally named 'del' and 'ddel' respectively. Sometimes, these del features are called acceleration feature. The acceleration features are mostly affected by noises. The derived cepstral coefficient without any derivation is known as static feature which is less affected by noise compare to acceleration features. Generally, the static feature contains 13 features per frame. Normally at the end, the static and accelerated

features are added to get 39 features dimension. at the end the MFCC-feature size is $l$ x 39; where $l$ is the number of frame.

To achieve robust SID score and to compare with a robust SID method only static features are used. It has been reported in (Zhao et al., 2012) that, without any derivation MFCC-based SID attained comparable performance under noisy condition to GFCC-based system.

It has mentioned earlier that, the conventional MFCC-based method only takes into account magnitude of spectral coefficients and does not consider phase information but speech signal contains both magnitude and phase information. It has been reported in (Nakagawa et al., 2012) that, the MFCC-based system performance can be improved significantly including speech magnitude with phase information.

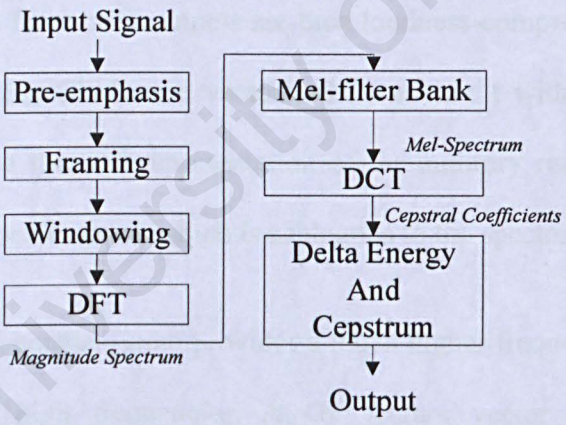


Figure 2.15: Block diagram for MFCC feature derivation technique of an input audio speech signal.

### 2.5.2 Gammatone Frequency Cepstral Coefficient (GFCC)

The Gammatone frequency cepstral coefficient (GFCC) is an acoustic cepstral-based feature which is derived from Gammatone feature (GF) using Gammatone filter-bank. According to the physiological observation, Gammatone filter-bank is most resemble to

cochlear filter-bank (Patterson, Nimmo-Smith, Holdsworth, & Rice, 1987). To extract GFCC, the audio signal is initially synthesized using a 128-channel or 64-channel Gammatone filter-bank. Its center frequencies are quasi logarithmically spaced from 50 Hz to 8 KHz (or half of sampling frequency of the input signal), which models human

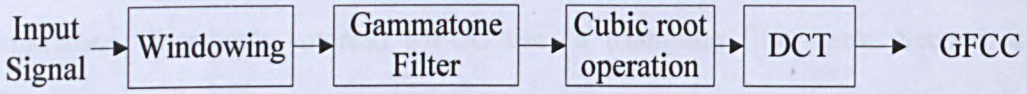


Figure 2.16: GFCC feature extraction block diagram using Gammatone filter and cubic root operation.

cochlear filtering (Shao, Srinivasan, & Wang, 2007). The filter-bank outputs are then down sampled to 100 Hz in time dimension, corresponding to a frame rate of 10 ms, which is used in many short-term speech feature extraction algorithms. The magnitudes of the down-sampled filter-bank outputs are then loudness-compressed using cubic root operation. The resulting GF feature vectors, Gf(t) at time t with component index of frequency f, comprise the T-F representation of the auditory response. This response matrix is called the cochleagram, which is analogous to the spectrogram.

Analogous to MFCC, cochlea-gram provides a much higher frequency resolution at low frequencies than at high frequencies. A GF feature vector contains 64 or 128 components depending upon the number of channels are used. The obtained GF feature dimension is higher than the features are normally used in a typical speech or speaker recognition system. To reduce the GF and de-correlate coefficient components of this feature DCT is applied. The lowest 23-order GFCCs among 64-order GFCCs or 30-order GFCCs among 128-channel-based GFCC are normally used for speaker identification since there retain most information of GF due to energy compaction property of DCT as mentioned in (Shao et al., 2007). So the size of the GFCC feature is

$m \times 23$ or m × 30; where $m$ is the number of frames. It has been observed in (Zhao et al., 2012) that, the first GFCC-coefficient was mostly tentative to noise corruption. It has been also observed in this study; the inclusive of first-order of GFCC coefficient reduced the SID performance under noisy condition but improved SID score in clean.

There are two main differences between GFCC and MFCC. First, GFCC uses a Gammatone filter-bank whereas MFCC uses a triangular filter-bank. Second, a log operation is applied in deriving MFCC whereas a cubic root operation is used in GFCC derivation. Figure 2.16 represents the GFCC derivation procedure.

### 2.5.3 Frequency Domain Linear Prediction (FDLP)

Ganapathy *et al. (Ganapathy, Thomas, & Hermansky, 2010)* proposed auto-regressive model based frequency domain coefficient using linear prediction technique is known as frequency domain linear prediction (FDLP). This feature-based SID method acheived improved performance compare to MFCC-based method as reported in (Ganapathy et al., 2012). This feature is newly developed and used in speech (Ganapathy, Thomas, & Hermansky, 2008) and speaker recognition.

FDLP feature has been designed based on high-energy peaks in the T-F domain. The derivation of this feature is done in several steps. Initially, the sub-band Hilbert envelopes is derived using frequency domain linear prediction auto-regressive (AR) model. Then, The FDLP envelopes in each sub-band are integrated in short-term frames (25ms window with a shift of 10ms i.e. 40% overlap of adjacent frame). These all-pole envelopes from each sub-band are converted to short-term energy estimates and the energy values across various sub-bands are used as a sampled power spectral estimate for the second AR model. The output prediction coefficients from the second AR model are converted to cepstral coefficients and are used for speaker recognition.

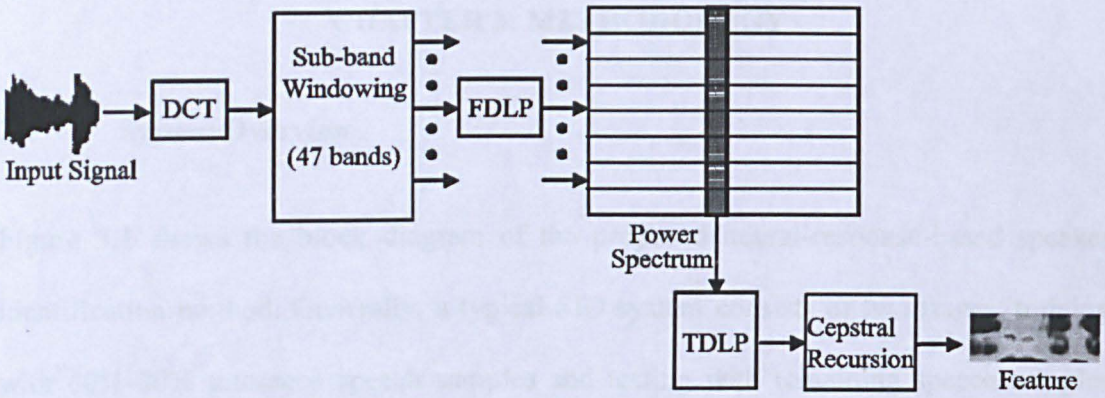


Figure 2.17: Block diagram of FDLP feature extraction using 2-D auto-regressive modeling

# CHAPTER 3: METHODOLOGY

## 3.1 System Overview

Figure 3.1 shows the block diagram of the proposed neural-response-based speaker identification method. Generally, a typical SID system consists of two stages: training with 60%-80% utterance speech samples and testing with remaining speech samples from each speaker. In the training stage, the processed unvoiced free speech is applied to a computational model of the human auditory periphery to generate neurogram feature. The GMM-UBM classifier is used to train the proposed SID system using neurogram coefficients extracted from the train samples of each speaker and GMM and SVM classifier are used specially for text-dependent speech-based SID system. As a result, an identity model is generated for each speaker and saved for identification process. In the testing stage, the extracted features of a test signal (speech of unknown speaker) are used as an input to each model. The model that gives maximum similarity in terms of probability score determines the identity of the test sample talker. The performance of the prosed method was tested in clean and noisy conditions as described in chapter 4 in this presented study. The noisy signal was produced by masking with stationary and non-stationary noise to ensure that, the proposed system is eventually efficient for any types of noise. The computation procedure of the proposed SID method has shown in block diagram in Figure 3.1 will be discussed in the following subsections step by step.

## 3.2 Pre-processing

Pre-processing is refer to a step that encompasses aligning of speech sample to a specific template, removal of silent period, maintaining the starting and ending time of speech signal recording, and so on.

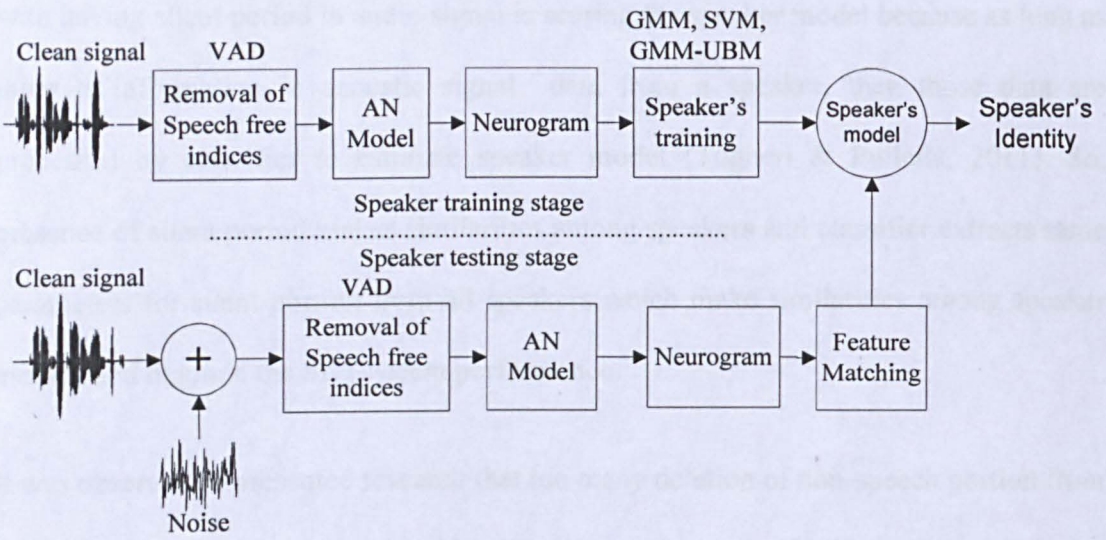


Figure 3.1: Block diagram of the proposed method for robust speaker identification.

As a pre-processing step in this study, the voice activity detector (VAD) statistical algorithm (Brookes, 1997)is used to remove the unvoiced signal portion from input speech samples which is important to improve the classification performance.

To obtain unvoiced free speech signal, the clean and noisy signal has been forwarded through VAD algorithm. This algorithm has been applied to detect unvoiced signal indices by computing low energy level of the input speech signal and delete those indices to provide unvoiced free speech signal output.

The silent period is very essential for separating words from each other and speech intelligibility for human being. There could be a considerable amount silence period between the activation and actual speech in a push-to-talk system, which does not carry any information about the speaker and make similarities among speakers. It was observed in presented study that the silent period between phrases or words was highly affected by noises and reduced the proposed SID method performances. So it was given importance on removing silent period (unvoiced portion) from speech signal to put down the effect of noise as well as improve the SID performance. The another problem

with having silent period in audio signal is scoring the speaker model because as long as there is information in acoustic signal data from a speaker, then those data are processed by classifier to estimate speaker model (Togneri & Pullella, 2011). So, presence of silent period makes similarities among speakers and classifier extracts same parameters for silent portion from all speakers which make similarities among speaker models and degrade the SID system performance.

It was observed in presented research that too many deletion of non-speech portion from audio signal reduced the speaker identification performance for both clean and noisy conditions. So it was given importance to select a standard algorithm that would not affect to speech intelligibility and delete unvoiced portion from target signal. Another advantage of taking away non-speech from acoustic signal, it made the proposed system fast due to smaller size of input signal.

## 3.3 Auditory Nerve (AN) Model and Neurogram

The AN model developed by Zilany *et al.* (Zilany & Bruce, 2006) is a useful implement for studying the principle physiological and mechanical mechanism in human auditory periphery. The schematic block diagram of this AN model is shown in Figure 1 in (Zilany & Bruce, 2006). This model represents the method of encoding the simple and complex sounds in the auditory periphery. Each block of the AN model represents a phenomenological description of the major functional component of the auditory periphery from the middle ear to the auditory nerve. This study results were significantly improved compare to results of previous study (Islam et al., 2016) using update version of AN model (Zilany et al., 2014). That's why the AN model with updated parameters (Zilany et al., 2014) was used in this study to extract the proposed feature for robust SID system. The input to the model is an instantaneous pressure waveform in Pascal, and the output is the spike times.

Four different conditions of IHC and OHC combinations in cochlear were considered in AN model for human and cat. In speaker identification, the AN model for human cochlea with normal inner hair cell (IHC) and normal outer hair cell (OHC) are used. Human conversation sound pressure level (SPL) is 65 dB to 70 dB (Dubno, Horwitz, & Ahlstrom, 2005; Studebaker, Sherbecoe, McDaniel, & Gwaltney, 1999). So, 70 dB SPL is used as reference in AN model for speaker identification.

The first stage of the model filters the input signal to simulate the response properties of the middle ear. After the middle ear module, the model splits into three paths. A narrowband component 1 (C1) filter mimics the response properties of the basilar membrane. The feed-forward control-path regulates the gain and bandwidth of the C1 filter to account for level-dependent properties associated with the outer hair cells (OHCs) such as compression, suppression, and nonlinear phase responses in the cochlea. The C1 and component 2 (C2) filters in the signal path interact to account for the effects associated with the AN responses at high sound levels such as peak splitting and the C1/C2 transition (N. Y.-s. Kiang, 1990). The third stage simulates inner-hair-cell (IHC) mechanisms with a static nonlinearity followed by a seventh-order low-pass filter. The IHC output drives the model for IHC-AN synapse which includes exponential as well as power-law adaptations (Zilany et al., 2009). The model synapse output represents the probability of instantaneous discharge rate of AN fibers as a function of time which was used in this study to construct neurograms. Finally, the discharge times are produced by a renewal process that includes refractory effects.

In light of the current debate on human cochlear tuning (Oxenham & Shera, 2003; Shera, Guinan Jr, & Oxenham, 2010) some parameters of the most recent version of this model are adjusted to better match human anatomy and physiology. These

modifications include the middle-ear filtering, the cochlear place-frequency map, and the sharpness of cochlear frequency tuning (Ibrahim & Bruce, 2010).

In this study, the neural responses were simulated for 32 CFs logarithmically spaced between 250 and 8000 Hz. It is mentioned that, each CF is considered as a single auditory nerve and behaves like a band-pass filter with asymmetric filter shape. In this presented study, each CF's neural responses have been simulated only for single repetition for each stimulus. A single nerve fiber does not fire on every cycle of the stimulus but once spike is obtained, it obtains roughly at same phase of the waveform at repetitive time which is called phase locking. It has been mentioned in literature review section that, the phase locking varies somewhat across species but upper frequency boundary lies at 4-5 KHz (Palmer & Russell, 1986) and Heinz and colleagues (Heinz et al., 2001) was observed, a weak phase locking up to 10 KHz frequencies. That's why; the maximum limit of CF ranged are commonly used 8 KHz in AN model.

The sampling rate of the input speech signal was resampled to 100 kHz. A high sampling rate is required by the AN model in order to faithfully replicate the frequency response properties of different parts of the model (V. Bruce, Green, & Georgeson, 2003; Zilany et al., 2009).The output at each AN fiber represents the instantaeous discharge rates in response to the acoustic signal. For each CF, three types of spontaneous rates (SR) of fibers (high, medium and low) were considered in this study. Consistent with the distribution of SR of AN fibers (M. C. Liberman, 1978), the maximum weight (0.6) was given to high spontaneous rate fibers, and the weight given to medium and low spontaneous rate fibers was 0.2 each. The neural responses for each CF were then binned with a 100 μs bin-width, and then a Hamming window of 420 points was applied with a 60% overlap among adjacent frames to smooth the neural responses (i.e., the effective frame length was 25.2 ms).

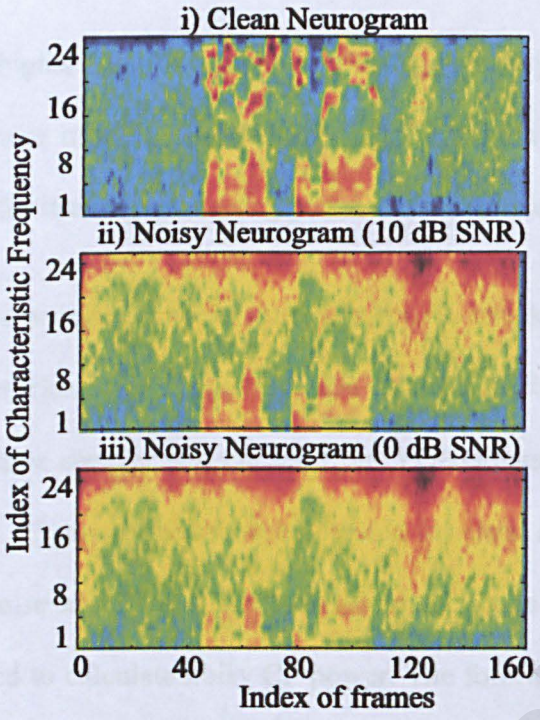


Figure 3.2: Illustration of the effects of noise on the neural responses. Neurogram responses are shown for a typical speech signal taken from the UM dataset. The neurogram to the clean speech signal is shown in the panel A, and the two neurograms in response to speech signal distorted by two levels of white Gaussian noise are shown in panels B (10 dB SNR) and C (0 dB SNR).

## 3.4 Neurogram feature dimension

In this study, the 2-D ENV neurogram coefficients have been extracted through AN-model for speech samples and used as features for the proposed SID method. The proposed neural feature average size was 190 x 32 for YOHO dataset; where 190 is number of frames and 32 is number of CFs and feature size was varied with different dataset.

Figure 3.2 illustrates clean and noisy neurogram with 32 CFs. It was observed that, with the increment of noise level, the higher CFs (above 12 CFs) were distorted by noise havoc whereas lower CFs (1-12 CFs) are less affected by noise compare to higher CFs.

In Figure 3.2 (C), the higher CFs contains most of energy at noisy level. It was observed that, the SID performance of the proposed method fallen quite a lot when 32CFs-based neurogram was taken for training and testing due to their mismatch condition.

The power of clean and noisy neurogram was calculated to check the correlation among CFs coefficients under clean and noisy condition as depicted in Figure 3.3. A single speech sample from same speaker was chosen from YOHO dataset and power of each CF was computed. So, 32 power points were obtained for clean neurogram for 32 CFs. After adding white noise at different SNRs level to the same speech sample, same procedure was followed to calculate noisy CF power. The following equation was used to compute the power of the neural responses per CF

$$X(i) = \frac{1}{N}\sum_{j=1}^{N} A_i^2(j), \qquad i = 1,2,3,\ldots\ldots\ldots,32 \tag{3.1}$$

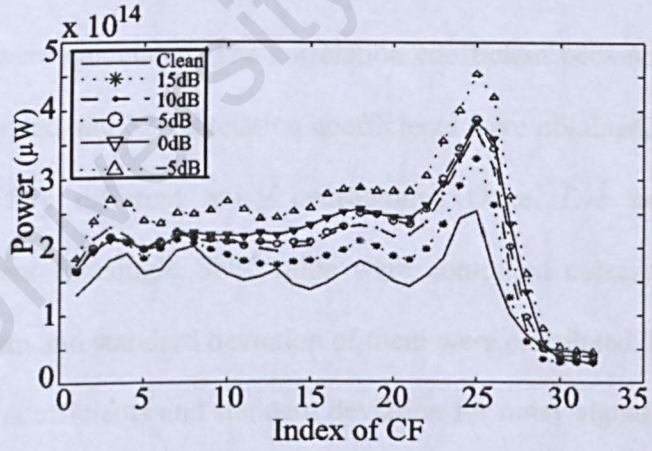


Figure 3.3: The correlation measure among CFs at different SNR in term of power as a function of CF.

Where $A_i(j)$ is the ENV or TFS responses vector as a function of time for the $i$th value of CF, and N is the length of $A_i$. It is seen in Figure 3.3 that the power of neurogram at different SNR level are almost consistent up to twelve (12) CFs at clean and noisy condition; whereas higher CFs power are deviated mostly for different SNRs.

It is seen from Figure 3.3 that, the upper CFs after 26 are more correlated. The matter of fact was that, YOHO dataset sampling frequency 8 kHz and has no speech energy above 4 kHz. However, the AN model has been used to simulate speech signal up to 8 kHz irrespective to different datasets and 26 number CF represents above 4 kHz. So, 26 CF to 32 CF power were more correlated.

The correlation coefficient among each CF for different noises at different SNRs were calculated to find which CFs contents are more correlated and which are not under noisy condition compare to clean neurogram CFs contents. To test this CF correlation, five different samples from same speaker were taken from YOHO dataset and clean and noisy neurograms were calculated. The correlation coefficient between each clean and noisy CF was computed and 32-correlation coefficients were obtained. Same procedure was repeated for five different noisy neurograms. Once, five noisy neurograms correlation coefficient for single SNR value were computed corresponding to clean neurograms and mean and standard deviation of them were calculated. Figure 3.4 shows the CFs correlation coefficients and standard deviation for noisy signal's neurograms. It is noted that, the synapse response has been used in this study which has dc values at frequencies> 4 kHz and correlation coefficients are high. So, the higher CF>4 kHz has been excluded from Figure 3.4.

It is seen from Figure 3.4 that, the first 12 CFs at 15 dB SNR irrespective to noises are correlated with clean neurogram on average 85% and low standard deviation indicates the proposed feature's consistency among speech samples from same speaker. With the

increment of SNR level, correlation between clean and noisy neurogram was degrading but still first 12 CFs are more correlated compare to other CFs with clean neurogram.

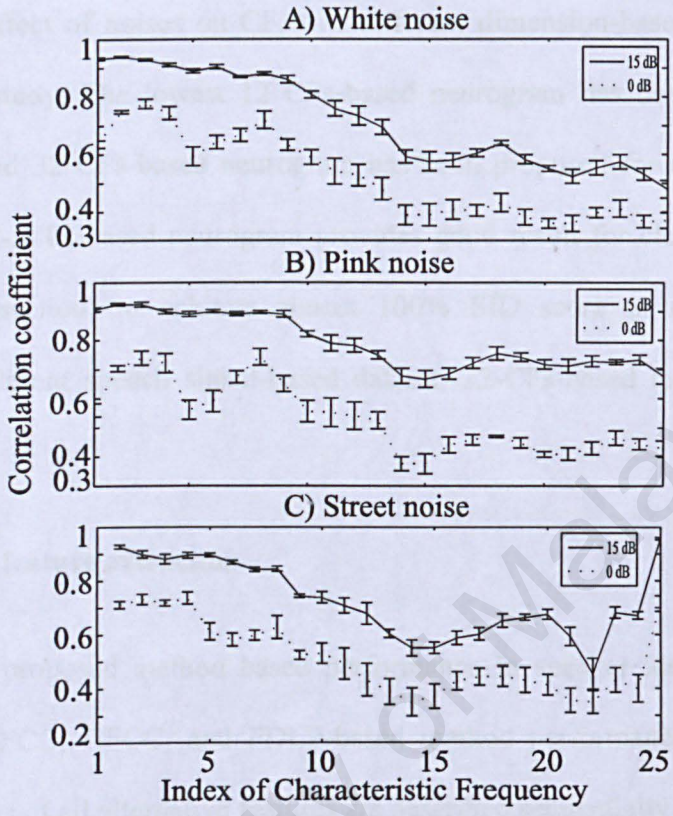


Figure 3.4: Clean to noisy neurogram correlation measure for three different noises at 15 dB and 0 dB SNRs.

Since, the most similarities between clean and noisy neurogram lied in first twelve (12) CFs as illustrated in Figure 3.2, Figure 3.3, and Figure 3.4; so 12-CFs-based feature has been proposed for noisy condition in this study. However, it was observed in this study, there are most similarities between noisy and clean neurogram across 32 CFs from 20 dB SNR level.

To check the stability of the proposed method, four different samples from each speaker were chosen from UM dataset. Three speakers with same four samples were chosen. The mean value of each CF was calculated. So, each speech sample was presented with only twelve points. It was observed that, the speech samples with same speaker are

consistent but different among speakers. It implies that, the proposed SID method can extract speaker distinguishing feature from speech signal very accurately.

Considering the effect of noises on CF, two different dimension-based features have proposed in this study. The lowest 12 CFs-based neurogram has been proposed for noisy condition and 32 CFs-based neurogram has been proposed for clean condition. Although, only 12- CFs-based neurogram provides good result for clean condition as shown in result section; to achieve almost 100% SID score in clean condition irrespective to different speech signal-based dataset, 32-CFs-based method has been suggested.

## 3.5 Baseline feature extraction

In this study, the proposed method based performance in speaker identification was compared with MFCC, GFCC, and FDLP-based method performances. The feature extraction processes of all alternative features are described sequentially in this session.

### 3.5.1 MFCC derivation

Mel-frequency cepstral coefficient (MFCC) is a short-time cepstral coefficient which is substantially used in the field of speech processing. In this study, the RASTAMAT toolbox (Ellis, 2005) was used to extract MFCC features from each speech frame. The Hamming window with a size of 25 ms and 60% overlap was used for dividing speech signals into frames. The 39 log-energy based MFCC coefficients were then computed for each frame. This set of coefficients consists of three groups: Ceps (Mel-frequency cepstral coefficients), Del (derivatives of ceps) and Ddel (derivatives of del) with13 features per group. The procedure is usually followed to extract MFCC feature has been described in detail in literature review chapter.

### 3.5.2 GFCC derivation

In this study, GFCC has been derived following code from Shao *et al.* (Shao et al., 2007) which is available on their official website. A fourth-order 64-channel Gammatone filter was used which center frequencies were quasi logarithmically spaced from 50 Hz to half of speech signal sampling frequency to simulate the Human cochlear filter-bank. The energy of each frame of each channel was taken which is called T-F based cochleagram. To make difference from MFCC, cubic root instead of log operation was perform on response of Gammatone filter. To obtain GFCC, DCT was applied on cubic operation-based spectral to convert into cepstral. According to the suggestion of Zhao *et al.* (Zhao et al., 2012), only 22-dimensional GFCC at a range from 2-23 coefficients were used in this study.

### 3.5.3 FDLP derivation

The FDLP feature extraction procedure based on 2-D autoregressive model has been shown in Figure 2.17. The acoustic time domain signal was transformed into frequency domain using DCT. The frequency domain signal was windowed into 47 linear sub-bands in maximum frequency range of 4 kHz. Then, linear prediction was applied to each sub-band to obtain Hilbert envelope. In this study, 160 order poles per sub-band per second were used. A 25 ms length of window with 40% adjacent frame overlap was used to compute 13 cepstral coefficients including zero ($C_0$) order coefficient. Delta and acceleration coefficients were also included. So, 39 cepstral coefficients-based FDLP feature was derived. The detail derivation process can be found in (Ganapathy et al., 2012).

## 3.6 Speech Dataset

Three renowned text-independent databases and a local text-dependent dataset have been used in this study to check the consistency in SID evaluation of the proposed method. The result has been simulated for text-independent YOHO, TIMIT, and TIDIGT databases and text-dependent UNIVERSITI MALAYA (UM) dataset. A brief description of each dataset, their field of application and contribution of them in this study are discussed in this section.

### 3.6.1 YOHO Dataset

YOHO dataset is a very famous dataset used for speaker verification (Campbell & Higgins, 1994). A large scale, high quality data contain in YOHO dataset, collected by ITT technical institute in 1989 under the contract of USA Government (Campbell & Higgins, 1994). In this study, to test the robustness of the proposed feature-based SID system, YOHO database from which 137 speakers (106 males and 31 females) out of 138 speakers were chosen (due to feature extraction problem from rest one speaker). In YOHO dataset each speaker has four sets of enrollment session with 24 independent utterances (with three two digits number (27-82-39, pronunciation twenty-seven eighty-two thirty-nine)) for each enrollment session.18 samples among 24 samples were selected for training and rest 6 samples were used for testing for each speaker. It is mentioned that, the randomly choosing of samples for each speaker did not affect the proposed SID method performances.

### 3.6.2 TIMIT Dataset

TIMIT dataset is the combine triumph output of Texas Instruments and Massachusetts Institute of Technology (MIT) that implies the corpus name (Fisher, Doddington, & Goudie-Marshall, 1986). This dataset is used for general linguistic research and text-

independent speaker recognition (Bimbot, Magrin-Chagnolleau, & Mathan, 1995). TIMIT database is the combination of 630 speakers (438 males and 192 females) with eight major (eight directories) American English dialect having 10 different phonetically rich sentences for each speaker.100 speakers out of 630 speakers were disorderly selected to get different regions dialect combinations. In this study, among 10 sentences 8 samples and 2 samples were selected for training and testing respectively under different noisy condition.

### 3.6.3 TIDIGITS dataset

Texas Instruments, Inc. (TI) designed and collected speech data contain in TIDIGT corpus to design and evaluate speaker-independent recognition for connected digit sequences (Leonard & Doddington, 1993). There have 55 male speakers and 57 female speakers. Each speaker has 77 digit-utterance samples which have different length. Forty (40) speakers (20 males and 20 females) were indiscriminately taken from 112 speakers for TIDIGIT corpus to have a standard amount of training and testing samples. Each speaker contains 77 different length audio speeches. 50 samples were selected for training and 27 samples were tested for clean and noisy conditions from each speaker.

### 3.6.4 UM Dataset

To show text-dependent speaker identification performance with the proposed method UM dataset is used. A speech database with 39 speakers (25 males and 14 females) from the auditory neuroscience lab in the University of Malaya has been used. Every speaker was asked to utter 'Universiti Malaya' for 10 times. The signals were recorded in a sound proof booth with a sampling frequency of 8 kHz.

## 3.7 Speaker Modeling

To train speaker identification system three different classifiers have been used in this study. GMM and SVM have been used only for text-dependent speaker identification. GMM_UBM has been used for text-independent speaker identification.

### 3.7.1 GMM

The application of GMM classification technique is very renowned and well established in speaker identification. The expectation maximization (EM) algorithm (Bilmes, 1998) makes the GMM as a successful classifier in speaker identification. Mixture weights, mean vectors, and covariance matrices from all component densities, parameterized speaker GMM model.

In this study, GMM –based speaker modeling has been done only for text-dependent technique-based SID. All neurograms of the training set from each speaker of UM dataset were combined together to form an input array for training. Thirty nine (39) GMM models were generated for UM dataset using EM algorithm. Diagonal covariance matrices were used to reduce the computational overhead and thus make the system faster. Ten (10) mixture components were used to obtain 12 dimensional feature vectors, and it was noticed that the proposed system performance dropped gradually when the number of mixture component was increased above or below 10. Speaker identification process was accomplished by comparing each unknown test utterance to all GMM speaker models and calculating the maximum likelihood values using the probability density function (pdf).

The same methodological steps have been followed for all others substitute feature-based GMM speakers modeling except number of mixture of components. It was

observed that, 32-mixture components made better speaker training models for MFCC and GFCC-based method and given best SID score.

### 3.7.2 SVM

The Matlab libsvm toolbox (Chang & Lin, 2011) was used to generate 39 SVM speaker models for speaker identification. In this study, the one-versus-rest (OVR) classification technique was employed to train the SVM models. Note that the size of the feature vector determines the dimensional space of the classifier. In this study, 7 samples were used to train the SVM models, and the remaining 3 samples were used to evaluate the performance of the system. In this study, the feature size of the proposed method was m × 12, where m was the number of envelope points in the neurogram. For the MFCC-, FDLP-, and GFCC-based methods, there were $n \times 39$, n × 39, and $n \times 22$ features for each speech sample; where $n$ is the number of frames of the speech signal.

In the proposed method, the features were normalized in such a way that training data array was bounded to a mean value of zero (0) and a standard deviation of one (1). The default type of kernel function (Radial basis function, RBF) was used in this study. Using a cross validation algorithm (Suykens, Van Gestel, De Moor, & Vandewalle, 2002), the best C and $\gamma$ (associated to the RBF kernel) were chosen in such a way that it provided best accuracy. Once the training model was obtained, it was saved and tested against test samples for range of SNRs. In this study, the following parameters: cost function (c), gamma (g), SVM type (s), shrinking parameter (h) was set to 4, 1, 0, and 0, respectively for the proposed and GFCC-based method to achieved maximum performance. It was observed, 39-dimentional MFCC and FDLP-based system provides better speaker identification accuracy with the following parameters: c = 1 and g = 0.0125. The speaker identity of the unknown test speech was determined by the model that given a maximum decision value.

### 3.7.3 GMM-UBM

In GMM-UBM (D. A. Reynolds et al., 2000), the GMM classifier-based speaker modeling is done by optimizing the range of speaker distinguishing feature with the help of universal background model (UBM). So, another name of this classifier is adapted GMM. Statistical models like the GMM are not only able to be estimated directly using a powerful technique like the EM algorithm, but with small amounts of data the parameters can be further *adapted* to the new data using either Maximum Likelihood Linear Regression (MLLR) or Maximum A-Posteriori (MAP) adaptation (E. D. Young & Sachs, 1979). In this study, training data was adapted using MAP technique.

To achieve maximum speaker identification with the proposed method, GMM classifier was adapted with UBM using 128 mixture components. It was observed; MFCC and FDLP-based method provided better SID score at 128 mixture components. GFCC-based speaker modeling was also done using 64 mixture components to get better SID performance. In GMM-UBM speaker modeling, features are arranged in nDim × nFrames format. All testing and training data are arranged at a time in a file by cell wise. It was observed in this study, putting the sub-sampling factor and map adaptation relevance factor to 10 improve speaker identification performance. Initially, UBM parameters (mean, covariance, and weight) were calculated from training data to adapt with GMM classifier. In this study, 'wmv' configuration was used to adapt with weight, mean, and covariance.

## 3.8 The Experimental Setup

The experimental set up has two stages: feature extraction and speaker modeling. Feature extraction block diagram is exactly same as shown in Figure 2.10 and the speaker modeling has been shown in Figure 3.1. The neural responses were simulated to

construct neurograms in response to speech signals from the databases. The overall sound pressure level (SPL) of each speech signal was normalized to 70 dB before applying to the AN model. For both in quiet and under noisy conditions, only responses of the first 12 CFs (<~1 kHz) were used in both training and testing phases, and the results are shown in Figures. 5-8. However, Table 1 summarizes the speaker identification performance of the proposed method in quiet when all the 32 AN fiber responses were considered for training and testing.

In the training stage, generally 70% to 80% of clean speech samples for each speaker were used for GMM-UBM speaker modeling. In the testing stage, the rest of the 20% to 30% of speech samples for each speaker was used directly or corrupted by three types of noise (white Gaussian noise, pink noise, and street noise) with a range of SNRs. The performance of the proposed system was evaluated for three times (by randomly selecting the training and testing samples) for each condition, and it was found that the SID score varied less than 1%.

# CHAPTER 4: RESULTS AND DISCUSSION

In this section, the proposed SID feature robustness has been methodically explained based on GMM, SVM and GMM_UBM classifier. The proposed SID system was tested in clean and noisy conditions. The simulated SID results were compared with the baseline feature MFCC, FDLP, and GFCC-based SID results using GMM-UBM speaker modeling. GMM and SVM were used only for text-dependent SID purposes. In this section the obtained results of this study has been compared with baseline feature-based performances. The findings of this study has been also analyzed along with results

## 4.1 Evaluation Results

It is to be mentioned that the proposed feature was applied in both text-dependent and text-independent SID system to check the performance of the proposed method. So, this section presents results for both text-dependent and text-independent SID system.

### 4.1.1 Text-independent Speaker Identification Results

In Figure 4.1, the performance of the proposed SID method for YOHO dataset is illustrated for three different types of noise. It is seen in Figure 4.1 that, with the increment of noise level, all baseline features-based system performances were degraded quite a lot compare to proposed SID method. The proposed method-based SID scores were also dropped with the increment of noise level due to the degradation of relation between noisy and clean signal as shown in Figure 3.4. The proposed system's performances are almost constant irrespective to noises whereas other baseline feature-based methods performances were varied quite a lot response to various noises. GFCC-based SID system was biased with pink noise and provided significant improved SID

result; but the proposed method's performances under pink noise still comparable with GFCC-based SID scores.

In Figure 4.1(A), it is seen that the proposed system's performance slightly less compare to other system performances under clean environment but still comparable. The reason of degradation of SID score in quiet condition was that the information above 1 KHz were not taken into consideration to make the dimension size same for clean and noisy conditions. It is noted that, the proposed method SID score was 99.51%

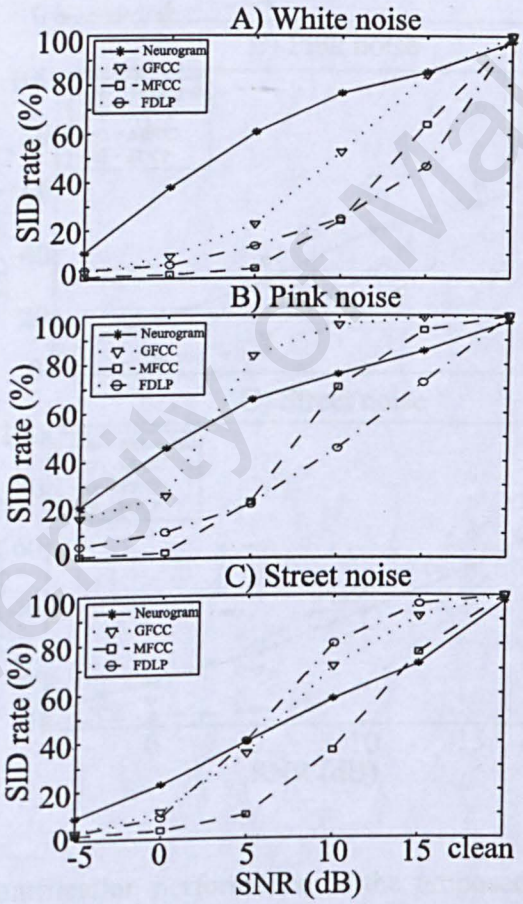


Figure 4.1: Speaker identification performance of the proposed and existing methods using YOHO database. Results are shown as a function of SNR with three different types of noise (A: white Gaussian noise, B: pink noise, and C: street noise). Speech samples from 137 speakers were used for evaluation and comparison of the performance of the methods.

including 32 CFs in clean condition. In case of non-stationary (street) noise, most of speech energies lied in lower frequencies that affected the proposed system performances under noisy conditions.

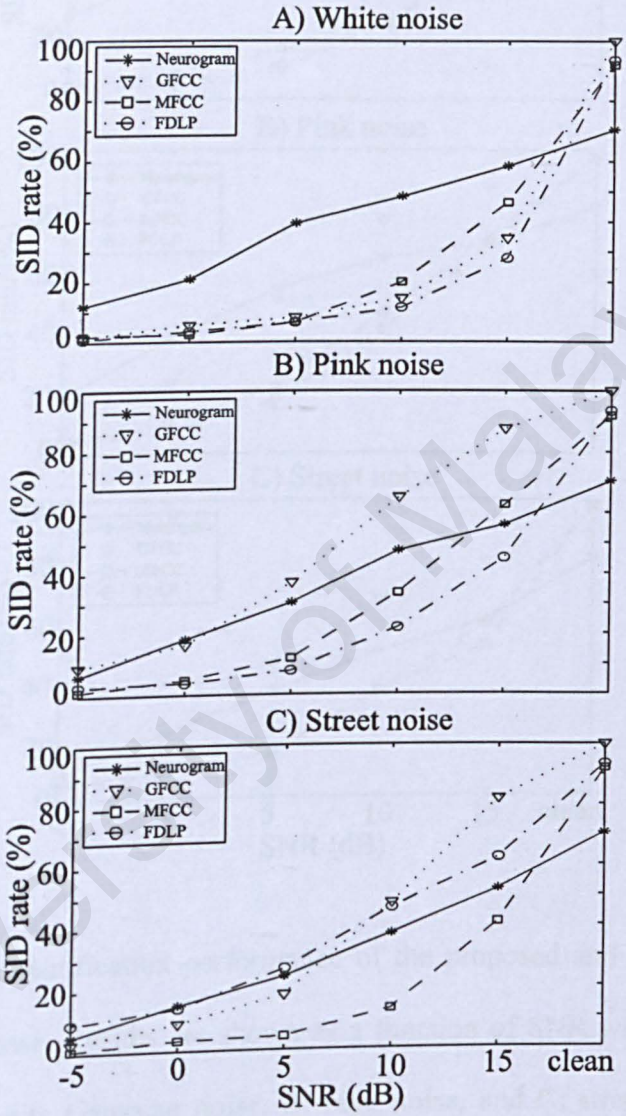


Figure 4.2: Speaker identification performance of the proposed and existing methods using TIMIT database. Results are shown as a function of SNR with three different types of noise (A: white Gaussian noise, B: pink noise, and C: street noise). Speech samples from 100 speakers were used for evaluation and comparison of the performance of the methods.

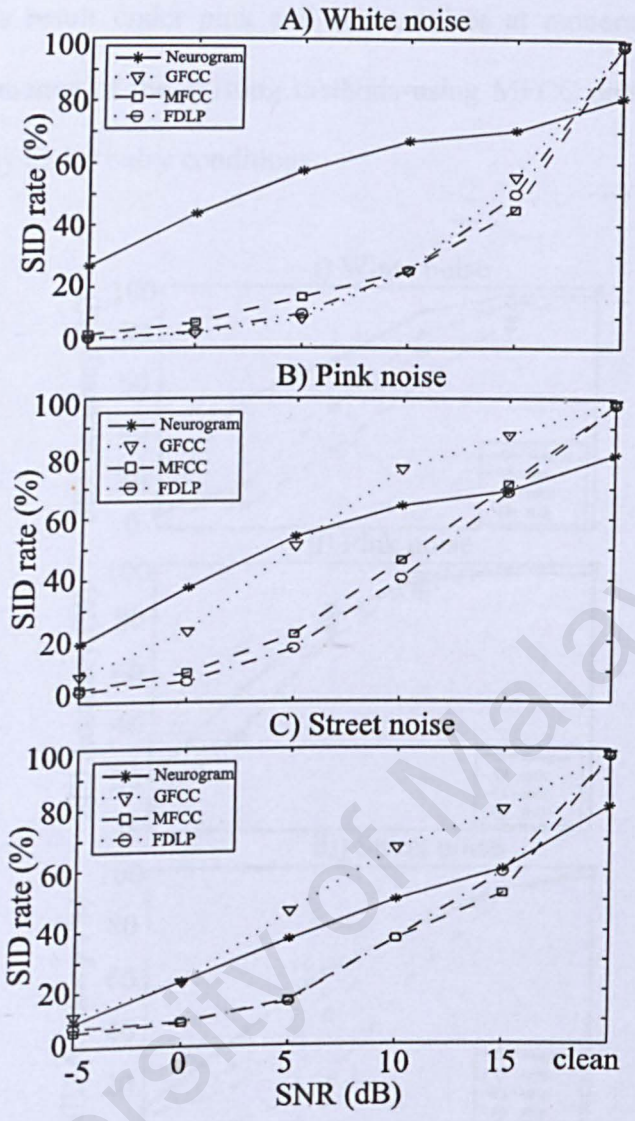


Figure 4.3: Speaker identification performance of the proposed and existing methods using TIDIGIT database. Results are shown as a function of SNR with three different types of noise (A: white Gaussian noise, B: pink noise, and C: street noise). Speech samples from 40 speakers were used for evaluation and comparison of the performance of the methods.

The speaker identification performance of the proposed and several existing methods are shown in Figure 4.2 and Figure 4.3 for the text-independent TIMIT and TIDIGIT databases, respectively. In general, the proposed method outperformed all other methods under white Gaussian noise, whereas GFCC performed slightly better than the

proposed methods result under pink and street noises at moderate levels of SNRs. Again, the performance of the existing methods using MFCC and FDLP coefficients declined drastically under noisy conditions.

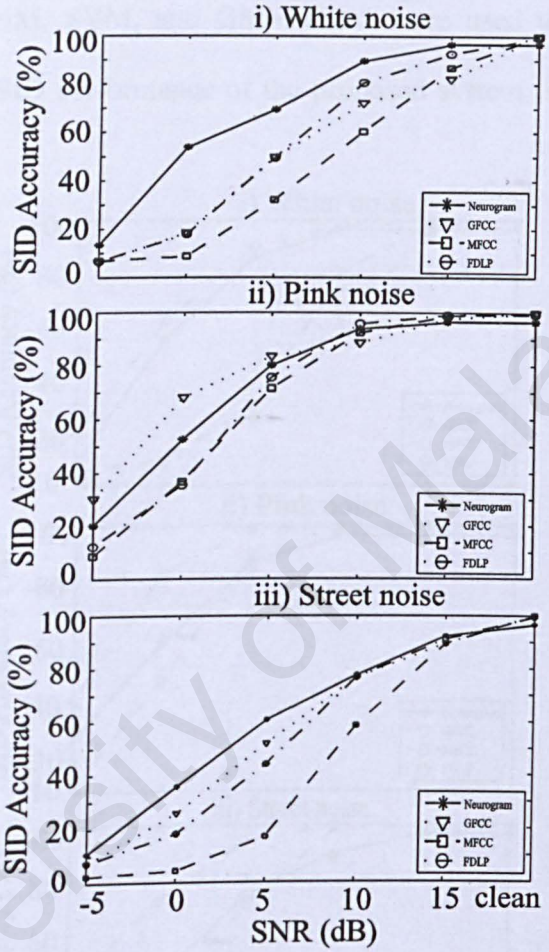


Figure 4.4: Text-dependent speaker identification performance of the proposed and existing methods for UM database using GMM classifier. Results are shown as a function of SNR with three different types of noise (A: white Gaussian noise, B: pink noise, and C: street noise). Speech samples from 39 speakers were used for evaluation and comparison of the performance of the methods.

### 4.1.2 Text-dependent Speaker Identification Results

A text-dependent dataset named UNIVERSITI MALYA (UM) was used to evaluate text-dependent SID system performance using the proposed method. Three different types of classifier GMM, SVM, and GMM-UBM were used to obtain SID results. Figure 4.4 shows the SID performance of the proposed system (solid line) along with

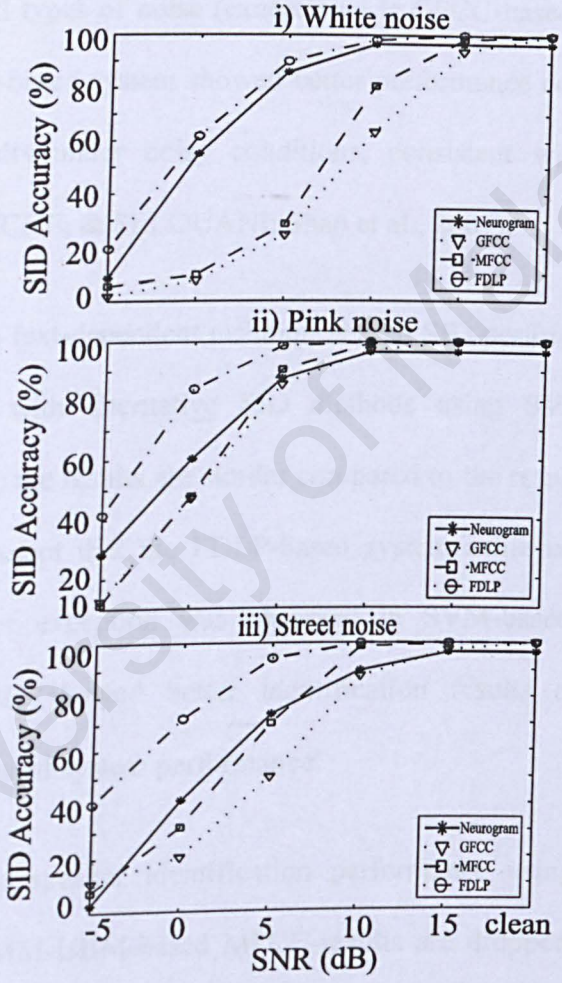


Figure 4.5: Text-dependent speaker identification performance of the proposed and existing methods for UM database using SVM classifier. Results are shown as a function of SNR with three different types of noise (A: white Gaussian noise, B: pink noise, and C: street noise). Speech samples from 39 speakers were used for evaluation and comparison of the performance of the methods.

the performance of the baseline-feature-based methods using GMM as a classifier based on text-dependent speech. The performance is shown as a function of SNR. It is clear that the performance of all the systems in quiet was almost 100% and very comparable to each other. As more noise was added to the speech signal, the identification results performance of all systems degraded accordingly. However, the neural response-based proposed system outperformed the traditional acoustic-feature-based systems at all SNRs studied for all types of noise (exceptional is GFCC-based performance for pink noise). Also, GFCC-based system showed better performance compared to the MFCC-based systems results under noisy conditions, consistent with the observation in (KROBBA, DEBYECHE, & SELOUANI; Shao et al., 2007).

Figure 4.5 shows the text-dependent technique-based SID performance of the proposed SID method along with alternative SID methods using SVM speaker modeling technique. In general, the results are similar compared to the results described in Figure 4.4 (using GMM) except that the FDLP-based system achieved significant improve performance. Another exception was observed in SVM-based SID technique the MFCC-based methods showed better identification results at all SNRs studied compared to GFCC-based system performance.

Figure 4.6 shows the speaker identification performance using GMM-UBM-based speaker modeling. GMM-UBM-based MFCC-results are dropped quite a lot whereas the proposed and GFCC feature-based method has got improved performance. The proposed method-based performance was very robust irrespective to SNR using GMM-UBM classifier and outperformed other baseline feature-based system performances under noisy conditions.

For the text-dependent UM database, the overall performance of all methods was higher compared to the corresponding results for the text-independent cases. Also, FDLP-based

method in this case showed a slightly better performance compared to the identification accuracy of other methods for different types of noise. it has been observed in Figure 4.4, Figure 4.5, and Figure 4.6 that the performance of the proposed system is almost independent of the classifier, whereas the GFCC-, FDLP-, and MFCC-based system performances are dependent on the classifier.

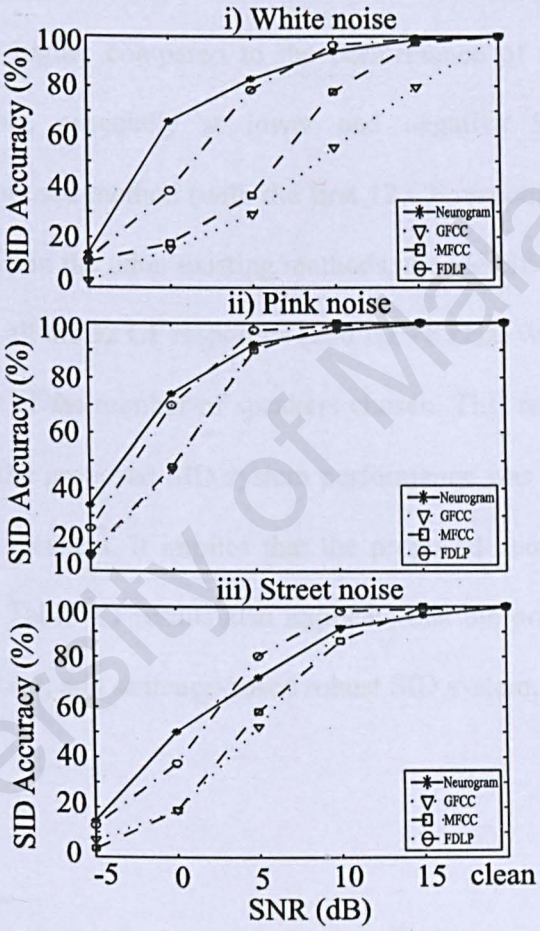


Figure 4.6: Text-dependent speaker identification performance of the proposed and existing methods for UM database using GMM-UBM classifier. Results are shown as a function of SNR with three different types of noise (A: white Gaussian noise, B: pink noise, and C: street noise). Speech samples from 39 speakers were used for evaluation and comparison of the performance of the methods.

In general, the proposed system's performance (pattern and value of accuracy as a function of SNR) was strikingly similar across all types of noises for each database, whereas the SID accuracies of other methods fluctuated substantially when the speech signals were degraded by different types of noise. This observation may imply that mostof the existing baseline features are very sensitive to noise, whereas the neural-response-based features are relatively robust against noise. Thus the SID accuracy of the proposed method was higher compared to the performance of the existing methods under noisy conditions, especially at lower and negative SNRs. Although the performance of the proposed method (with the first 12 CF responses) in quiet was a bit lower than the results from the other existing methods, the identification accuracy of the proposed method with all the 32 CF responses (250 Hz - 8 kHz) was nearly 100% for all databases, irrespective of the number of speakers chosen. This result is summarized in Table 4.1 shows that the proposed SID system performance was less affected with the increment number of speakers. It implies that the proposed neural feature-based SID system is very stable. Table 4.1 results also implicate that the proposed feature is very efficient for digits, phrase, and sentence-based robust SID system.

Table 4.1: Speaker identification (SID) accuracy of the proposed method using responses from 32 CFs for YOHO, TIMIT, TIDIGIT, and UM dataset. A subset (results italized) of the TIMIT(100/630) and TIDIGIT(40/112) dataset was randomly taken to compare the performance with the existing studies. The performances were evaluated using GMM-UBM speaker modeling technique.

| Dataset | Number of Speakers | SID Accuracy (%) |
|---|---|---|
| YOHO | 137 | 99.51 |
| TIMIT | *100* | *100* |
| | 630 | 98.1 |
| TIDIGIT | *40* | *98.46* |
| | 112 | 96.15 |
| UNIVERSITI MALAYA | 39 | 100 |

## 4.2 Analytic Study

This study investigated the implications of the proposed neural-response-based feature in the speaker identification task both in quiet and under noisy conditions. The neural features were derived from the responses of a physiologically-based model of the auditory periphery. Different databases with different combination of speech materials (sentences, words, and digits) were used to test the proposed SID method. The most important finding of this study was that the proposed neural feature resulted a consistent performance across different types of noise irrespective of the speech materials and the duration of the signal. On the other hand, most of the baseline feature-based (such the MFCC, GFCC and FDLP coefficients) systems produced quite different results for different types of noise (white Gaussian vs. other noises). Based on simulation results, the proposed system outperformed most of baseline feature-based systems under noisy

conditions, especially at lower and negative SNRs. Also, the proposed system outperformed other existing methods under white Gaussian noise.

The proposed text-dependent SID system has got significant improve SID performance as shown in Figure 4.5 and Figure 4.6 compare to previous study (Islam et al., 2016). It implies that the modified parameter-based AN model (Zilany et al., 2014) can capture speaker distinguishing feature more accurately compared to previous model (Zilany & Bruce, 2006), which is another significant outcome of this study.

To estimate optimize parameters in neurogram extraction and to study the effect of those parameters on overall speaker identification performance, each module in neurogram feature extraction was developed separately with change parameter and experimented. In this study a number of effects such like the filter width, CF dynamic range, various window length were investigated to achieve overall better SID performance.

Effect of CF dynamic range

In this study, the performance of the proposed system was evaluated using two different simulation conditions. The responses of the first 12 CFs (250 Hz – <~ 1 kHz) or all the 32 CFs (250 Hz – 8 kHz) were used. The response of only first 12 CFs (out of 32 CFs) which ranges lie within 1 KHz was chosen for robust SID performance since higher frequencies information were largely affected by noises as illustrated in Figure 3.2, Figure 3.3, and Figure 3.4 respectively. So, the dimension of the proposed feature was very small compared to other baseline features which made the SID system very fast and took on average less than half time compare to alternative baseline feature-based systems to obtain speaker identity.

It is seen from the results that the performance of the proposed SID system in quiet condition is not satisfactory with 12-CFs but the proposed feature-based system perform equally well when all the 32 CF responses are considered. The SID performance including 32 CFs coefficients was almost 100% irrespective to nature of speech materials and number of speakers. In the AN model, each CF response is considered as the response of neuron. It can be said based on the results of this study that the responses from 32 neurons (out of 30,000 neurons in the human auditory periphery) is adequate to confidently identify any speaker in quiet environment irrespective to the text of the speech material.

It was observed that the identification score for the proposed method using responses of first 12 CFs in quiet was nearly 100% for YOHO and UM datasets, whereas for TIMIT and TIDIGIT, the accuracy was ~70-80%. It is to be recalled that, the TIDIGIT dataset speech materials have different length whereas TIMIT dataset has sentences with different combination of words from different region. In order to explore the variation in performances across databases, the correlation coefficient between each corresponding CF responses for four speakers has been shown in Figure. 4.7 A typical speech sample (same text) spoken by all four speakers was selected from each database, and the corresponding neural responses were simulated. Panel A shows the correlation measure as a function of CF among four speakers chosen from the YOHO database. In general, the AN responses among speakers showed a correlation measure of less than 0.5 for the first 12 CFs, and thus the SID score was nearly 100% in quiet. On the other hand, the correlation measure of CF responses among speakers was relatively higher and in some cases became indistinguishable for TIMIT (panel B) and TIDIGIT (panel C) datasets. Panel D represents the similar correlation measure for the text-dependent UM dataset.

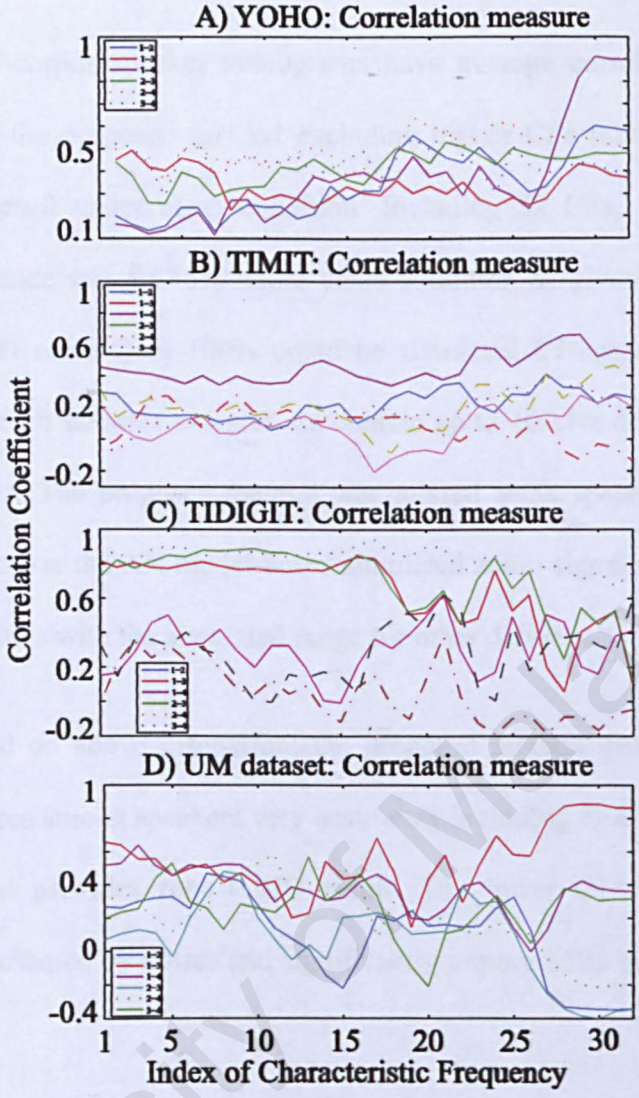


Figure 4.7: Illustration of the correlation of AN fiber responses between speakers to the same text sample. Panel A shows the correlation measure as a function of CF among four speakers in response to the text "26-81-57" from YOHO database. Similarly panel B, C, and D show the results for TIMIT ("She had your dark suit in greasy wash water all year"), TIDIGIT ("12"), and UM ("University Malaya") databases.

It was observed that the correlation measure among four speakers was on average ~0.5 for the first 12 CFs, and the obtained SID score was almost 100% in clean condition. However, taking into consideration the responses of 32 CFs, the identification performance was nearly ~100% for all databases, as reported in Table 4.1.

Specially, TIDIGIT-corpus speaker neurograms have average correlation as seen in Figure 4.7 (C) and the proposed method excluding higher CFs (above first 12 CFs) provided average result under clean condition. Including 32 CFs, TIDIGIT-corpus-based SID performance was 98.46% under clean condition as given in Table 4.1. A reason of fallen SID score from 100% could be simulated CF-responses range. The TIDIGIT corpus speech samples energies are contain up to 10 kHz due to its sampling frequency is 20 kHz. The proposed method was missed some speech information of TIDIGIT corpus because the AN model-based simulated audio signal responses up to 8 kHz to keep similarities with the simulated range for other datasets.

It can be said based on above discussion; the proposed method can capture speaker distinguishing features among speakers very accurately including 32-CFs coefficients in clean condition and provides robust SID result. The lower CF's coefficients are comparatively less affected by noises and significantly improve SID performance under noisy conditions.

The proposed method stability

In this study, four datasets were used which speeches were completely different from each other. In UM, YOHO, and TIMIT database, speech lengths were almost constant but TIDIGIT's speech samples were combinations of different length and mixed of digit and alphabet. One of the causes to performance variations of all SID method could be the variation in nature of input speech from various databases. Another reason to the fluctuation of performances of all SID methods could be the application of the same speaker modeling parameters to all databases. Irrespective to speech materials and number of speakers, the proposed method based SID results were very robust under clean and noisy conditions.

In this study, the performance of the proposed method was evaluated and reported for speech signals corrupted by three different types of additive noise. In order to test the robustness of the proposed method against other types of distortion, the HTIMIT dataset was used to evaluate the speaker identification performance. The HTIMIT corpus is a recording of a subset of the TIMIT corpus through 10 different telephone handsets. Using the same classification technique and experimental setup for TIMIT database, the proposed method produced an accuracy of 94.3% for 100 speakers randomly chosen from the database.

Phase-locking properties

The robustness of the proposed neural-response-based system could lie on the underlying physiological mechanisms observed at the level of the auditory periphery. Since the AN model used in this study is nonlinear (i.e., incorporates most of the nonlinear phenomena), it would be difficult to tease apart the contribution of each individual nonlinear mechanism towards SID performance. However, it would not be unwise to shed some light on the possible mechanism towards the identification task, especially under noisy conditions. The AN fiber tends to fire at a particular phase of a stimulating low-frequency tone, meaning that it tends to give spikes at an integer time of period of that tone. It has been reported that the magnitude of phase-locking declines with frequency and the limit of phase locking varies somewhat across species, but the upper frequency boundary lies at ~4-5 kHz (Palmer & Russell, 1986). Thus, it is not surprising that in Figure 4.7, the correlation between the noisy and clean responses declines as a function of CF, and the lower CFs (<1 kHz) show higher correlation coefficients due to the phase locking property of AN model.

SPL effect

One of the great achievements of this study was to include the effect of SPL on SID system. In general, with the increment of speech level above the normal conversation level (~65-70dB), speech as well as speaker recognition capability decreases in clean condition as reported by (Dubno et al., 2005; Studebaker et al., 1999). In addition to the broadened bandwidth of the AN fibers at higher levels, the potential mechanism underlying degraded performance at higher levels is also hypothesized to be related to the loss of synchrony capture by the second formant while synchrony to the first formant increases at higher sound levels(Zilany & Bruce, 2007). It has been observed in this study that, the SID score was decreased slightly at 98.1% when SPL level was increased from 70 dB to 90 dB for people with normal hearing but improved for people with moderate hearing loss in quiet state including 32-CFs. It is known the quietest sound level is 50dB that can understand human with normal hearing. To check this observation, the proposed SID method has been run for 40 dB SPL with 32-CFs and the performance was reduced to 53.41% from 99.51% under clean condition for YOHO dataset. It is mentioned that, MFCC-and FDLP-based SID system has no SPL effect but GFCC-based system is concern to speech loudness but there have no effect on SID performance. The GFCC-based system was run for 40 dB, 60 dB and 90 dB to check the SPL effect on SID and the SID scores were 93.5%, 93.5% and 96.5% respectively under quiet condition for TIMIT dataset. It can be said based on GFCC-based obtained SID result that, there is negligible effect of SPL on SID system which does not reflect human auditory properties. So, only the proposed feature-based system can handle the effect of SPL in SID system. The SPL effect discussion directs, with the proper adjustment of hearing loss level SID performance can be improved.

To investigate the SPL effect on text-dependent SID performance the proposed method was run for 50 dB, 70 dB, and 90 dB SPL using UM dataset materials. The audio signals were corrupted with white Gaussian noise at 0 dB and 10 dB SNR. To be consistent with aforementioned results, the best SID performance was obtained at 70 dB SPL irrespective to different instances of background. It is shown in Figure 4.8, the best SID performance was obtain at 70 dB SPL.

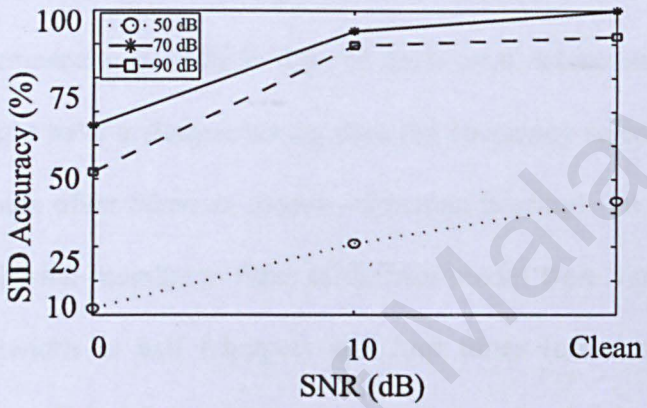


Figure 4.8: Sound pressure level (SPL) effect on robust text-dependent speaker identification performance in matched and noisy conditions.

Non-linearity

A narrowband component 1 (C1) filter mimics the response properties of the basilar membrane. The feed-forward control-path regulates the gain and bandwidth of the C1 filter to account for level-dependent properties associated with the outer hair cells (OHCs) such as compression, suppression, and nonlinear phase responses in the cochlea. The C1 and component 2 (C2) filters in the signal path interact to account for the effects associated with the AN responses at high sound levels such as peak splitting and the C1/C2 transition (N. Y.-s. Kiang, 1990). The third stage simulates inner-hair-cell (IHC) mechanisms with a static nonlinearity followed by a seventh-order low-pass

filter. The IHC output drives the model for IHC-AN synapse which includes exponential as well as power-law adaptations (Zilany et al., 2009).

Bandwidth effect

Frequency selectivity in the inner ear is fundamental to hearing and plays a critical role in the ability to distinguish and segregate different sounds perceptually. This implies that the cochlear-filter bandwidth might have a crucial effect on the speaker identification performance, especially in light of the current debate on human cochlear tuning (humans might have a sharper tuning than the frequency selectivity of most of the laboratory animals often taken as models of human hearing). To address this, the Q10 values of the basilar-membrane filter of the AN model were varied to adjust the cochlear-filter bandwidth to half (sharper) and four times (broader) of the normal values. It is to be noted that the tuning parameters (normal values) of the AN model used in this study were implemented based on the physiological data in cats. The performance of the proposed method in quiet was evaluated for YOHO database for three different values of cochlear-filter bandwidths. The obtained SID score was 99.51%, 99.31%, and 53.77% for normal, half, and four times broader bandwidth, respectively. This clearly suggests an important role of frequency selectivity in the inner ear towards the speaker identification task. However, exploring the detail contribution of each nonlinear phenomenon observed at the peripheral level of the auditory system on SID task is beyond the scope of this study and could be pursued as a future work.

Window resolution effect

To study the effect of neurogram resolution in robust SID system, three different window lengths (128 pints, 200 points and 420 points) were experimented for 32 speakers from YOHO dataset. Same speaker was simulated for three different window

sizes for three different noises with SNR from -5 dB to 15 dB at a step of 5 dB. It was observed that, the window length of 128 points having 50% overlap provides comparatively low and very little change in SID score with the reduction of noise level as shown in Figure 4.4. It was observed in this study, the performance of the proposed method with 200 window points was good for digit-based dataset as given in Table 4.2 but worse for TIMIT-dataset. However, 420 points of window length with 60% shifting was used in this study which performance is eventually good for all datasets and provided significant improved SID performances compare to 128-points window-based method.

Table 4.2: Effect of window size on text-independent SID performances.

| Noise Type | Window Size | SNR -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | Clean |
|---|---|---|---|---|---|---|---|
| White Noise | 128 points | 27.08 | 48.77 | 77.06 | 89.06 | 95.83 | 100 |
| | 200 points | 27.08 | 57.38 | 79.69 | 90.63 | 94.79 | 100 |
| | 420 points | 22.92 | 55.21 | 75.00 | 85.94 | 93.75 | 100 |
| Pink Noise | 128 points | 25.52 | 51.04 | 78.13 | 87.50 | 94.27 | 100 |
| | 200 points | 32.81 | 62.50 | 85.42 | 90.10 | 97.40 | 100 |
| | 420 points | 30.21 | 58.33 | 79.17 | 87.50 | 94.27 | 100 |
| Street Noise | 128 points | 15.52 | 26.04 | 48.44 | 67.71 | 83.85 | 100 |
| | 200 points | 17.71 | 32.29 | 51.56 | 72.40 | 84.90 | 100 |
| | 420 points | 15.10 | 34.38 | 56.77 | 69.79 | 85.94 | 100 |

# CHAPTER 5: CONCLUSION

## 5.1 Conclusion

This study proposes a neural response-based novel metric for robust speaker identification for both text-dependent and text-independent speech samples. The feature (neurogram) was derived from the responses of a renowned and well-established physiologically-based computational model of the auditory nerve. Neurogram was constructed from the simulated responses of a wide range of CFs to input speech samples. The proposed neural feature successfully captured the important distinguishing information about speakers to make the system relatively robust against different types of degradation of the input acoustic signals.

In this study, SID system's performances were evaluated in clean and noisy conditions in order to provide an evidence of the robustness of the proposed feature. The clean signals were corrupted by adding white Gaussian noise, pink noise and street noise with a range of SNRs from -5 dB to 15 dB at 5 dB interval. The obtained results have been compared to the performance of three traditional acoustic MFCC, FDLP and GFCC feature-based systems. While the performance of the proposed system with the responses from 12 CFs was comparable or slightly poor to the results from alternative existing methods in clean condition, the proposed system outperformed all other systems under noisy conditions, especially for white Gaussian noise. However, although the SID accuracy of the proposed 12-CFs coefficient-based method was not satisfactory for some databases in clean, the proposed system outperformed all baseline systems' performances by including responses from all the 32 CFs. Also the performance of the neural-response-based system was almost consistent across databases and different noises, whereas alternative SID system's performances were fluctuating substantially for different databases and noises. In addition, the performance of the proposed system

was independent of the classifier used in the study (GMM, SVM, and GMM-UBM), whereas the baseline systems heavily depend on the classifier employed for the tasks. Although it is difficult to infer about the contribution of individual nonlinear phenomenon towards SID performance, the robustness of the proposed system might arise from the nonlinear representation of the acoustic signals in the auditory system.

## 5.2 Application

This study investigated whether a computational model of the auditory system could be applied to evaluate the SID performance in quiet and noisy conditions. The proposed metric successfully handled the task by providing a robust representation of the acoustic signal in the neural responses. Although the proposed method was developed and tested for listeners with normal hearing, it can easily be extended for listeners with hearing loss by simulating the responses of impaired AN fibers. The proposed metric-based system can be used instead of pin, password, finger-print, and punch-card in office, mobile and other related applications under noisy conditions for a wide range of SNR. The proposed neural feature can also be employed for speech recognition, phoneme classification, speaker verification, gender classification, and speech intelligibility.

## 5.3 Limitations and Future Work

Most of the practical noises show energies at lower frequencies in the power spectral density of the signal. So, the performance of the proposed method could be degraded significantly under practical noise conditions. Two different range of CFs have been proposed in this study depending upon the presence of noise level. The proposed method with 32 CFs provides almost 100% SID accuracy in clean environment irrespective of the speech materials and speaker numbers. The work presented in this study presents a variety of potential areas for future study. The proposed metric has

been used to identify speaker based on digit-based and phonetically enriched sentences, and the result shows that it can achieve a substantially improved performance under noisy conditions. The obtained result can outperform all baseline features under noisy condition but clean condition performance. So, this metric can be further improved to achieve better performance in quiet condition as well as under noisy conditions by choosing the best CFs among a range of CFs. The proposed metric used the same speaker modeling parameters to different dataset and hence the performance was affected. The performance would be better and provide better approximation by showing results for individual dataset using the best set of parameters for speaker modeling. A very interesting but challenging topic is the speaker identification with the data from handset variation. The proposed method can be further studied for speaker recognition using channel variation speech data. Taking individual section output (like signal path filter output, IHC model output) from AN model to make the system faster for a robust SID system could be an attractive topic for future studies as well. There are many non-linearity existed in the cochlea such as compression, suppression, nonlinear tuning, and phase-locking which are captured successfully by the AN model. The effects of individual nonlinear properties on the robust SID system could be another significant study with this proposed metric in future.

## LIST OF PUBLICATIONS AND CONFERENCE PROCEEDINGS

### LIST OF PUBLICATION

i. Islam, M., Zilany, M., Siew-Cheok, Ng, & Wissam, A. *A Robust Speaker Identification System using the Responses from a Model of the Auditory Periphery.* Plose One (2016).

### LIST OF CONFERENCE PROCEEDING

i. Islam, M., Zilany, M., & Wissam, A. (2016). *Neural-Response-Based Text-Dependent Speaker Identification Under Noisy Conditions.* Paper presented at the International Conference for Innovation in Biomedical Engineering and Life Sciences.

## REFERENCES


Alam, M. S., Jassim, W. A., & Zilany, M. S. (2014, December). Neural response based phoneme classification under noisy condition. In *Intelligent Signal Processing and Communication Systems, International Symposium on* (175-179). IEEE.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute, 4*(510), 126.

Bimbot, F., Magrin-Chagnolleau, I., & Mathan, L. (1995). Second-order statistical measures for text-independent speaker identification. *Speech communication, 17*(1), 177-192.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 27*(2), 113-120.

Bondy, J., Bruce, I., Becker, S., & Haykin, S. (2003). Predicting speech intelligibility from a population of neurons. In *Advances in Neural Information Processing Systems,* (Vol. 16), MIT Press, Cambridge, MA.

Brookes, M. (1997). Voicebox: Speech processing toolbox for matlab. *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html.*

Bruce, I. C., Sachs, M. B., & Young, E. D. (2003). An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. *The Journal of the Acoustical Society of America, 113*(1), 369-388.

Bruce, V., Green, P. R., & Georgeson, M. A. (2003). *Visual perception: Physiology, psychology, & ecology*: Psychology Press.

Campbell, J., & Higgins, A. (1994). YOHO speaker verification. *Linguistic Data Consortium, Philadelphia.*

Campbell Jr, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE, 85*(9), 1437-1462.

Carney, L. H., McDuffy, M. J., & Shekhter, I. (1999). Frequency glides in the impulse responses of auditory-nerve fibers. *The Journal of the Acoustical Society of America, 105*(4), 2384-2391.

Chang, C.C., & Lin, C.J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 27.

Chi, T.S., Lin, T.H., & Hsu, C.C. (2012). Spectro-temporal modulation energy based mask for robust speaker identification. *The Journal of the Acoustical Society of America, 131*(5), 368-374.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 28*(4), 357-366.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on, 19*(4), 788-798.

Dubno, J. R., Horwitz, A. R., & Ahlstrom, J. B. (2005). Word recognition in noise at higher-than-normal levels: Decreases in scores and increases in masking. *The Journal of the Acoustical Society of America, 118*(2), 914-922.

Ellis D.P.W (2005) PLP and Rasta and MFCC and inversion in Matlab. Web resource, retrieved July 11, 2007, from http://labrosa.ee.columbia.edu/matlab/rastamat/, URL http://labrosa.ee.columbia.edu/matlab/rastamat/.

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 32*(6), 1109-1121.

Fisher, W. M., Doddington, G. R., & Goudie-Marshall, K. M. (1986, February). The DARPA speech recognition research database: specifications and status. In *Proceedings DARPA Workshop on speech recognition* (93-99).

Flanagan, J. L. (1960). Models for approximating basilar membrane displacement. *The Journal of the Acoustical Society of America, 32*(7), 937-937.

Fletcher, H. (1940). Auditory patterns. *Reviews of modern physics, 12*(1), 47.

Gallardo, L. F. (2015). *Human and Automatic Speaker Recognition over Telecommunication Channels*: Springer NewYork.

Ganapathy, S., Thomas, S., & Hermansky, H. (2008). Front-end for far-field speech recognition based on frequency domain linear prediction. In *Interspeech 2008* (2008-063).

Ganapathy, S., Thomas, S., & Hermansky, H. (2010). Temporal envelope compensation for robust phoneme recognition using modulation spectrum. *The Journal of the Acoustical Society of America, 128*(6), 3769-3780.

Ganapathy, S., Thomas, S., & Hermansky, H. (2012, June). Feature extraction using 2-d autoregressive models for speaker recognition. In *Odyssey* (229-235).

Gifford, M. L., & Guinan Jr, J. J. (1983). Effects of crossed-olivocochlear-bundle stimulation on cat auditory nerve fiber responses to tones. *The Journal of the Acoustical Society of America, 74*(1), 115-123.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications,13*(4), 18-28.

Heinz, M. G., Colburn, H. S., & Carney, L. H. (2001). Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve. *Neural Computation, 13*(10), 2273-2316.

Ibrahim, R. A., & Bruce, I. C. (2010). Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues. In *The neurophysiological bases of auditory perception* (429-438). Springer New York.

Islam, M. A., Zilany, M. S. A., & Wissam, A. J. (2015, December). Neural-Response-Based Text-Dependent Speaker Identification Under Noisy Conditions. In *International Conference for Innovation in Biomedical Engineering and Life Sciences* (11-14). Springer Singapore.

Javel, E. (1980). Coding of AM tones in the chinchilla auditory nerve: implications for the pitch of complex tones. *The Journal of the Acoustical Society of America, 68*(1), 133-146.

Kajarekar, S. S., Bratt, H., Shriberg, E., & De Leon, R. (2006, June). A study of intentional voice modifications for evading automatic speaker recognition. In *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop* (1-6). IEEE.

Kiang, N. Y. S. (1965). *Discharge patterns of single fibers in the cat's auditory nerve* (No. STR-13). Masschusetts institute of technology Cambridge research laboratory of electronics.

Kiang, N. Y.S. (1990). Curious oddments of auditory-nerve studies. *Hearing research, 49*(1), 1-16.

Kiang, N. Y., Liberman, M. C., & Levine, R. A. (1976). Auditory-nerve activity in cats exposed to ototoxic drugs and high-intensity sounds. *Annals of Otology, Rhinology & Laryngology, 85*(6), 752-768.

Knight, B. W. (1972). The relationship between the firing rate of a single neuron and the level of activity in a population of neurons experimental evidence for resonant enhancement in the population response. *The Journal of general physiology, 59*(6), 767-778.

Krobba, A., Debyeche, M., & Selouani, S.A. Comparison of Auditory Feature Based GFCCs and MFCCs for Robust Speaker Identification in Noisy Environment applied to Arabic Speech.

Lei, Y., Burget, L., Ferrer, L., Graciarena, M., & Scheffer, N. (2012, March). Towards noise-robust speaker recognition using probabilistic linear discriminant analysis. In *2012 IEEE international conference on acoustics, speech and signal processing* (4253-4256). IEEE.

Leonard, R. G., & Doddington, G. (1993). Tidigits. *Linguistic Data Consortium, Philadelphia.*

Li, Q., & Huang, Y. (2011). An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. *Audio, Speech, and Language Processing, IEEE Transactions on, 19*(6), 1791-1801.

Liberman, M. C. (1978). Auditory-nerve response from cats raised in a low-noise chamber. *The Journal of the Acoustical Society of America, 63*(2), 442-455.

Liberman, M. C., and Kiang, N. Y. S. (1984). Single-neuron labeling and chronic cochlear pathology. IV. Stereocilia damage and alterations in rate and phase-level functions, *Hearing research 16*(1), 75–90.

Liberman, M. C., & Dodds, L. W. (1984). Single-neuron labeling and chronic cochlear pathology. III. Stereocilia damage and alterations of threshold tuning curves. *Hearing research, 16*(1), 55-74.

Lyon, R. F., & Mead, C. (1988). An analog electronic cochlea. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 36*(7), 1119-1134.

Mak, M.-W., Allen, W., & Sexton, G. (1994). Speaker identification using multilayer perceptrons and radial basis function networks. *Neurocomputing, 6*(1), 99-117.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE, 63*(4), 561-580.

Mamun, N., Jassim, W., & Zilany, M. S. (2015). Prediction of Speech Intelligibility Using a Neurogram Orthogonal Polynomial Measure (NOPM). *Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 23*(4), 760-773.

McAulay, R., & Malpass, M. (1980). Speech enhancement using a soft-decision noise suppression filter. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 28*(2), 137-145.

Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology, 9*(4), 399-406.

Nakagawa, S., Wang, L., & Ohtsuka, S. (2012). Speaker identification and verification by combining MFCC and phase information. *Audio, Speech, and Language Processing, IEEE Transactions on, 20*(4), 1085-1095.

Oxenham, A. J., & Shera, C. A. (2003). Estimates of human cochlear tuning at low levels using forward and simultaneous masking. *Journal of the Association for Research in Otolaryngology, 4*(4), 541-554.

Palmer, A., & Russell, I. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing research, 24*(1), 1-15.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1987, December). An efficient auditory filterbank based on the gammatone function. In *a meeting of the IOC Speech Group on Auditory Modelling at RSRE* (Vol. 2, 7).

Patuzzi, R., & Robertson, D. (1988). Tuning in the mammalian cochlea. *Physiological reviews, 68*(4), 1009-1082.

Qing, Z., & Mao-li, D. (2009). Anatomy and physiology of peripheral auditory system and common causes of hearing loss. *Journal of Otology, 4*(1), 7-14.

Razali, N. F., Jassim, W. A., Roohisefat, L., & Zilany, M. S. (2014, December). Speaker recognition using neural responses from the model of the auditory system. In *Intelligent Signal Processing and Communication Systems, 2014 International Symposium on* (076-079). IEEE.

Reynolds, D. (1994). Experimental evaluation of features for robust speaker identification. *Speech and Audio Processing, IEEE Transactions on, 2*(4), 639-643.

Reynolds, D. A. (1996), Automatic speaker recognition using Gaussian mixture speaker models, *The Lincoln Laboratory Journal,* 8 (1996), 173–192.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing, 10*(1), 19-41.

Robert, A., & Eriksson, J. L. (1999). A composite model of the auditory periphery for simulating responses to complex sounds. *The Journal of the Acoustical Society of America, 106*(4), 1852-1864.

Rose, J. E., Brugge, J. F., Anderson, D. J., & Hind, J. E. (1967). Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *Journal of Neurophysiology, 30*(4), 769-793.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society B: Biological Sciences, 336*(1278), 367-373.

Sachs, M. B., & Kiang, N. Y. (1968). Two-tone inhibition in auditory-nerve fibers. *The Journal of the Acoustical Society of America, 43*(5), 1120-1128.

Sewell, W. F. (1984a). The effects of furosemide on the endocochlear potential and auditory-nerve fiber tuning curves in cats. *Hearing research, 14*(3), 305-314.

Sewell, W. F. (1984b). Furosemide selectively reduces one component in rate-level functions from auditory-nerve fibers. *Hearing research, 15*(1), 69-72.

Shao, Y., Srinivasan, S., & Wang, D. (2007, April). Incorporating auditory feature uncertainties in robust speaker identification. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol. 4, 277). IEEE.

Shera, C. A., Guinan Jr, J. J., & Oxenham, A. J. (2010). Otoacoustic estimation of cochlear tuning: validation in the chinchilla. *Journal of the Association for Research in Otolaryngology, 11*(3), 343-365.

Slaney, M. (1998). Auditory toolbox. *Interval Research Corporation, Tech. Rep, 10*, 1998.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416*(6876), 87-90.

Soong, F. K., Rosenberg, A. E., Juang, B.-H., & Rabiner, L. R. (1987). Report: A vector quantization approach to speaker recognition. *AT&T technical journal, 66*(2), 14-26.

Studebaker, G. A., Sherbecoe, R. L., McDaniel, D. M., & Gwaltney, C. A. (1999). Monosyllabic word recognition at higher-than-normal speech and noise levels. *The Journal of the Acoustical Society of America, 105*(4), 2431-2444.

Suykens, J. A., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., Suykens, J., & Van Gestel, T. (2002). *Least squares support vector machines* (Vol. 4): World Scientific.

Tan, Q., & Carney, L. H. (2003). A phenomenological model for the responses of auditory-nerve fibers. II. Nonlinear tuning with a frequency glide. *The Journal of the Acoustical Society of America, 114*(4), 2007-2020.

Togneri, R., & Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *Circuits and Systems Magazine, IEEE, 11*(2), 23-61.

Vasilakakis, V., Cumani, S., & Laface, P. (2013). Speaker recognition by means of deep belief networks. In *Proceedings of Biometric Technologies in Forensic Science,* Nijmegen.

Wang, Y., Liu, X., Xing, Y., & Li, M. (2008, October). A Novel Reduction Method for Text-Independent Speaker Identification. In *International Conference on Natural Computation* (Vol. 4, pp. 66-70). IEEE.

Weiss, M. R., Aschkenasy, E., & Parsons, T. W. (1974, April). Processing speech signals to attenuate interference. In *Proceedings IEEE Symposium Speech Recognition* (292-293).

Wenndt, S. J., & Mitchell, R. L. (2012, March). Machine recognition vs human recognition of voices. In *International Conference on Acoustics, Speech and Signal Processing* (4245-4248). IEEE.

Wong, J. C., Miller, R. L., Calhoun, B. M., Sachs, M. B., & Young, E. D. (1998). Effects of high sound levels on responses to the vowel/ɛ/in cat auditory nerve. *Hearing research, 123*(1), 61-77.

Xu, L., & Pfingst, B. E. (2003). Relative importance of temporal envelope and fine structure in lexical-tone perception (L). *The Journal of the Acoustical Society of America, 114*(6), 3024-3027.

Young, E. D., & Sachs, M. B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *The Journal of the Acoustical Society of America, 66*(5), 1381-1403.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2002). Hidden Markov Model Toolkit (HTK) Version 3.2. 1 User's Guide. *Cambridge University Engineering Department, Cambridge, MA.*

Zhang, X., Heinz, M. G., Bruce, I. C., & Carney, L. H. (2001). A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *The Journal of the Acoustical Society of America, 109*(2), 648-670.

Zhao, X., Shao, Y., & Wang, D. (2012). CASA-based robust speaker identification. *Audio, Speech, and Language Processing, IEEE Transactions on, 20*(5), 1608-1616.

Zhao, X., & Wang, D. (2013, May). Analyzing noise robustness of MFCC and GFCC features in speaker identification. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (7204-7208). IEEE.

Zhao, X., Wang, Y., & Wang, D. (2014). Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech and Language Processing, 22*(4), 836-845.

Zheng, R., Zhang, S., & Xu, B. (2004, December). Text-independent speaker identification using GMM-UBM and frame level likelihood normalization. In *Chinese Spoken Language Processing, 2004 International Symposium on* (pp. 289-292). IEEE.

Zilany, M. S., & Bruce, I. C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *The Journal of the Acoustical Society of America, 120*(3), 1446-1466.

Zilany, M. S., & Bruce, I. C. (2007). Representation of the vowel/ɛ/in normal and impaired auditory nerve fibers: model predictions of responses in cats. *The Journal of the Acoustical Society of America, 122*(1), 402-417.

Zilany, M. S., Bruce, I. C., & Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *The Journal of the Acoustical Society of America, 135*(1), 283-286.

Zilany, M. S., Bruce, I. C., Nelson, P. C., & Carney, L. H. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. *The Journal of the Acoustical Society of America, 126*(5), 2390-2412.

Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication, 9*(4), 351-356.