

SPEAKER VERIFICATION  
USING NEURAL RESPONSES FROM  
THE MODEL OF THE AUDITORY  
SYSTEM

NOOR FADZILAH BT RAZALI

SUBMITTED TO THE  
DEPARTMENT OF BIOMEDICAL  
ENGINEERING  
FACULTY OF ENGINEERING  
UNIVERSITY OF MALAYA, IN PARTIAL  
FULFILMENT OF THE REQUIREMENT  
FOR THE DEGREE OF MASTER OF  
BIOMEDICAL ENGINEERING

2014

## Abstract

Speaker verification is the process of authenticating a person's identity. Most of the available speaker verification systems have been implemented today is based on the step by step analysis of the acoustical signal itself. However, they are very sensitive to noise and work only at very high signal to noise ratio (SNR). On the other hand, the neural responses under noise are very robust, and the behavioral responses are also robust under diverse background noise. Therefore, a speaker verification system is proposed using the neural responses at the level of the auditory nerve (AN). For this, a very well-developed AN model by Zilany and colleagues (Zilany et al. 2009) is employed to simulate the neural responses on verifying a speaker. For this project, the feature extraction of the speech is analysed using the responses from the AN model, where the output is in the form of synapse output. A neurogram is constructed from the synapse responses of neurons with a wide range of characteristic frequencies. The neurogram's average discharge or envelope (ENV) is then calculated. The resulted vector is then used to train the system using Gaussian Mixture Model (GMM) classification technique. Features are then extracted for testing data set and compared to the vectors for each of the trained speakers in order to verify a particular speaker. The speaker database is made up of recordings in a quiet room of 10 speech samples with 8 kHz sampling rate from 39 different speakers. Out of them, 70% speech samples of the speaker are used as the training set and the remaining 30% are for testing. As the neural responses are very robust to noise, speaker verification using AN model responses can substitute or outperform the current technology and thus improve performance for application such as in security processing.

## Abstrak

Proses ‘pengesanan suara’ adalah proses untuk mengenal pasti sama ada seseorang disahkan benar megikut identiti diri seperti yang didakwa. Kebanyakan sistem pengesanan suara yang ada telah dilaksanakan hari ini adalah berdasarkan kepada analisis isyarat akustik itu sendiri. Walau bagaimanapun, alatan ini sangat sensitif terhadap bunyi bising. Sebaliknya, tindak balas saraf dengan percakapan bersama bunyi latar belakang adalah sangat teguh, dan mempunyai tindak balas yang pelbagai. Oleh itu, satu sistem pengesanan suara dicadangkan dengan menggunakan system saraf auditory ‘*Auditory Nerve*’ (AN). Untuk ini, model AN oleh Zilany dan rakan-rakan (J. Acous. Soc. Am., 2009) digunakan untuk mensimulasikan jawapan neural untuk proses pengesanan suara. Untuk projek ini, pengekstrakan ciri ucapan dibincangkan dengan menggunakan hasil daripada model AN, di mana hasilnya adalah dalam bentuk output sinaps. Neurogram yang terhasil daripada neuron sinaps dihasilkan daripada pelbagai ciri frekuensi. Penghasilan purata neurogram atau ‘*envelope*’ (ENV) kemudiannya dikira. Ciri vektor yang terhasil kemudiannya digunakan untuk melatih sistem menggunakan ‘*Gaussian Mixture Model*’ (GMM). Ciri vector yang selebihnya digunakan sebagai data untuk ujian pengesanan. Pangkalan data pengesanan suara adalah terdiri daripada 10 rakaman ucapan dengan kadar pensampelan 8 kHz daripada 39 individu yang berbeza. Daripada semua sampel suara, 70% sampel digunakan sebagai set latihan manakala baki 30% adalah untuk tujuan pengujian. Memandangkan AN adalah teguh kepada bunyi, sistem pengesanan menggunakan model AN adalah diharapkan dapat menggantikan teknologi semasa dan dengan itu meningkatkan prestasi untuk aplikasi contohnya pemprosesan keselamatan.

## **Acknowledgement**

Bismillahirrahmanirrahim & Alhamdulillah;

My first and foremost appreciation would be for my supervisor, Dr Muhammad Shamsul Arefeen Zilany, who gives me an opportunity to do this project under his supervision. It is also for his thorough guidance, brilliant ideas, and wide knowledge in the field which he passed along that enables me to complete the task in doing the project in the given time.

A heartfelt gratitude goes towards Dr Wissam A. Jassim who is a research fellow from Biomedical Engineering Department who dutifully and patiently taught me the hands on application in developing the software and giving out ideas on solving the problem arises in completing the project.

My greatest appreciation also goes to my friends in the Auditory Neuroscience Lab (ANL) especially to Mr Nursadul Mamun whose help in giving out ideas and comments during our weekly discussion to improve the project output and my knowledge in the field.

Last but not least, for my loving family; although far apart never stopped showing their care and support emotionally through this hard but wonderful time in completing my thesis.

# Table of Contents

Abstract .....	ii
Abstrak .....	iii
Acknowledgement .....	iv
Table of Contents .....	v
List of Figures .....	vii
List of Tables .....	ix
List of Symbols and Abbreviations.....	x
Chapter 1. INTRODUCTION .....	1
1.1. Speaker verification.....	2
1.2. Human Auditory Nerve Pathway .....	4
1.3. Objective .....	7
1.4. Scope of study .....	7
1.5. Outline of the report .....	8
Chapter 2. LITERATURE REVIEW .....	9
2.1. Auditory Nerve (AN) modelling .....	9
2.2. Speaker verification/identification based on AN response .....	20
2.3. Classification for speaker verification.....	24
2.4. Feature extraction .....	35
Chapter 3. METHODOLOGY .....	41
3.1. System design.....	41
3.2. Verification System.....	43
3.3. Robustness of the system .....	53
3.4. System performance .....	53
3.5. Statistical analysis .....	55

Chapter 4. Result & Discussion.....	57
4.1. Pre-processing using DTW .....	57
4.2. AN model response .....	58
4.3. Krawtchouk Polynomial Feature Extraction .....	59
4.4. Training using GMM classification .....	60
4.5. System Performance.....	62
4.6. Statistical analysis .....	65
Chapter 5. Conclusion .....	72
References.....	75

# List of Figures

Figure 1.1 Typical Speaker verification system. T=threshold value. ....	3
Figure 1.2 A cross section of auditory pathway of the organ ear. ....	5
Figure 1.3 Pictorial representation of cochlea and the frequency band accepted on different parts of the basilar membrane from base (highest frequency) to apex (lowest frequency). ....	6
Figure 2.1 The signal flow of Meddis inner hair cell model.....	10
Figure 2.2 AIM model representation (functional) compared to actual physiological activity of AN in human in three stages; spectral analysis, neural encoding and auditory image. ....	12
Figure 2.3 Zilany-Bruce AN model pathway .....	14
Figure 2.4 Zilany-Bruce AN model (2009) with added IC model (shaded) .....	18
Figure 2.5 Speaker identification rate versus SNR.....	22
Figure 2.6 GMM distribution in 1 dimension .....	26
Figure 2.7 (a) Full covariances (b) diagonal covariances .....	29
Figure 2.8 Comparison of distribution modeling. (a) histogram of a single cepstral coefficient from a 25 second utterance by a male speaker (b) maximum likelihood uni-modal Gaussian model (c) GMM and its 10 underlying component densities (d) histogram of the data .....	31
Figure 2.9 EM algorithm depicted as data observed log-likelihood as a function of the iteration number .....	34
Figure 2.10 Krawtchouk polynomials plots for different values of polynomial order n, with Krawtchouk coefficients for different values of order n and p. Note that, N = 100. ....	38
Figure 3.1 Flow-chart of speaker verification .....	42
Figure 3.2 ENV neurogram blocks transformation to ENV moment neurogram .....	48
Figure 3.3 Flowchart of the GMM training process. *ENV represented all output of AN model (including synapse output and tsf), N=number of speakers. ....	50
Figure 3.4 Flowchart of the PDF testing process of one speech sample.....	52
Figure 4.1 (Top) Original speech sample of a speaker. (Bottom) Warped version of the same speech sample of the words ‘University Malaya’. ....	57

<b>Figure 4.2 Output neurograms of the AN model a) Synapse Output b) Envelope (ENV), c)</b>	
<b>Temporal fine structure (TFS).....</b>	<b>58</b>
<b>Figure 4.3 The result of original ENV Neurogram (top) compared to result of Krawtchouk</b>	
<b>Neurogram after transition from time-domain to moment-domain (bottom) .....</b>	<b>60</b>
<b>Figure 4.4 Graph of a speech sample test for different components (K) vs Speakers. The number</b>	
<b>shown (1.32E-208, 1.37E-208) belongs to K=16 and (2.02e-211, 1.32E-208) belongs to K=128</b>	
<b>for Speaker #2 and Speaker #10 respectively. ....</b>	<b>61</b>
<b>Figure 4.5 Error bar graph (for 95% CI) Speaker identification accuracy for with (red) and without</b>	
<b>(blue) feature extraction.....</b>	<b>71</b>



## List of Tables

Table 1.1 List of previous speaker verification studies using AN model response .....	24
Table 4.1 Speaker verification system performances .....	62
Table 4.2 Table of the accuracy of the system compared to different K components of the GMM .....	63
Table 4.3 Speaker identification system performances .....	64
Table 4.4 The Spearman's Correlation coefficient tested among the three speech samples .....	66
Table 4.5 Correlation coefficient for GMM using different K values .....	67
Table 4.6 Statistical data on internal reliability for 10 repeats of verification system .....	68
Table 4.7 Statistical on test-retest reliability for 10 repeats of verification system .....	68
Table 4.8 Standard deviation and mean for 10 repeated tests.....	70

## List of Symbols and Abbreviations

Auditory Nerve	<b>AN</b>
Basilar membrane motion	<b>BMM</b>
Complementary metal–oxide–semiconductor	<b>CMOS</b>
Discrete Cosine Transform	<b>DCT</b>
Dynamic Time Warping	<b>DTW</b>
Direct Current	<b>DC</b>
Expectation Maximization	<b>EM</b>
Excitation Pattern	<b>EP</b>
Envelope	<b>ENV</b>
Equal Error Rate	<b>ERR</b>
False Rejection Rate	<b>FRR</b>
False Acceptance Rate	<b>FAR</b>
Gaussian Mixture Model	<b>GMM</b>
Hidden Markov Model	<b>HMM</b>
Inner hair cells	<b>IHC</b>
Linear Predictive Coding	<b>LPC</b>
Neural Activity Pattern	<b>NAP</b>
Outer hair cells	<b>OHC</b>
Peristimulus Time Histogram	<b>PSTH</b>
Range of Interest	<b>ROI</b>
Short Time Fourier Transform	<b>STFT</b>
Spontaneous Rate	<b>SR</b>
Temporal Fine Structure	<b>TSF</b>

Total Success Rate

**TSR**

Universal Background Model

**UBM**

Vector Quantization

**VQ**

## **Chapter 1. INTRODUCTION**

Speaker recognition is a field of speech analysis as a biometric modality that uses the individual's speech for the purpose of recognition. The voice which is the feature representing a speaker is influenced by both how the speech was formed (individual characteristics) and on how it is physically formed (vocal tract and air pathways).

The human body consists of several main sensory organs such as the eyes for visualization, skin for touch perception and ear for hearing, as the main window for the body to communicate with its environment. Basic physiological understanding of these sensory organs mainly involve in the process of receiving stimulus from the environment whereby the organs then convert the stimulus into a series of signal processing (chemically and physically). The nervous system will then receive the signals through complex dynamic and nonlinear interpretation so that necessary action could be made after. Having been said that, the auditory nerve is one of the human's sophisticated sensory system that enable human to receive information acoustically, and it becomes very important in the process of learning and everyday life. The physiological process of the auditory pathway is crucial in understanding how the signals are transformed along the auditory pathway and thus subsequently result in the required perception.

### ***1.1. Problem statement***

The availability of using real-time telecommunication services nowadays (e.g. telephone networking, internet, etc) enables the user to use voice as a feature for remote authentication. However, this will also increase the susceptibility to transmission

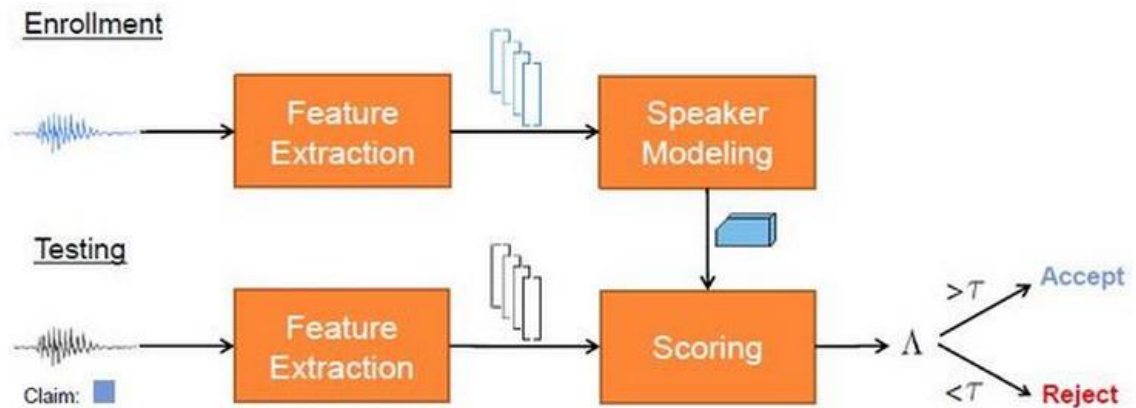
channel noise and microphone variability during speech recording especially when the training process used the clean recording while the speech sample used for testing is recorded with noisy background.

## **1.2. Study significance**

The motivation behind the undertaken research is based on the knowledge of the ability of human to identify and simply isolate the owner of a certain speaker just by hearing the speech sample even with the presence of background noise. As the model of the auditory system possesses the required characteristics of the system, the adoption of AN model in developing a biometric system might be a way to increase the speaker verification performance overall.

## **1.3. Speaker verification**

Speech analysis uses the speech itself as a biometric identity that is unique for each individual. The study of speaker recognition is widely used in various applications such as in forensics, banking and commerce, security enforcement and etc. It can be subdivided into two applications; speaker identification and speaker verification. Speaker identification is a system to determine who is the speech signal belongs to by comparing a list in a database, while the later which is going to be to be addressed in this project is a process to authenticate whether the person is who he/she claims to be compared to the one stored in the system.



**Figure 1.1: Typical Speaker verification system. T=threshold value (Reproduced from Dikici, 2000).**

A typical speaker verification system is as depicted in the Fig. 1.1 above. It has two main phases of verification process. The first step is the *training* in which speech samples of the speakers are enrolled in the system's database where its feature is extracted to be trained in the speaker modelling using classification technique. The second process, *testing*, is where any attempt to access the system is made by providing his/her speech where its feature is also extracted so that scoring or comparison with the threshold value set by the training phase can be made. This is where the "likelihood-ratio" is found in which if a certain similarity threshold or likelihood that the speaker is who he/she claims to be, is achieved then the user can access the system or otherwise.

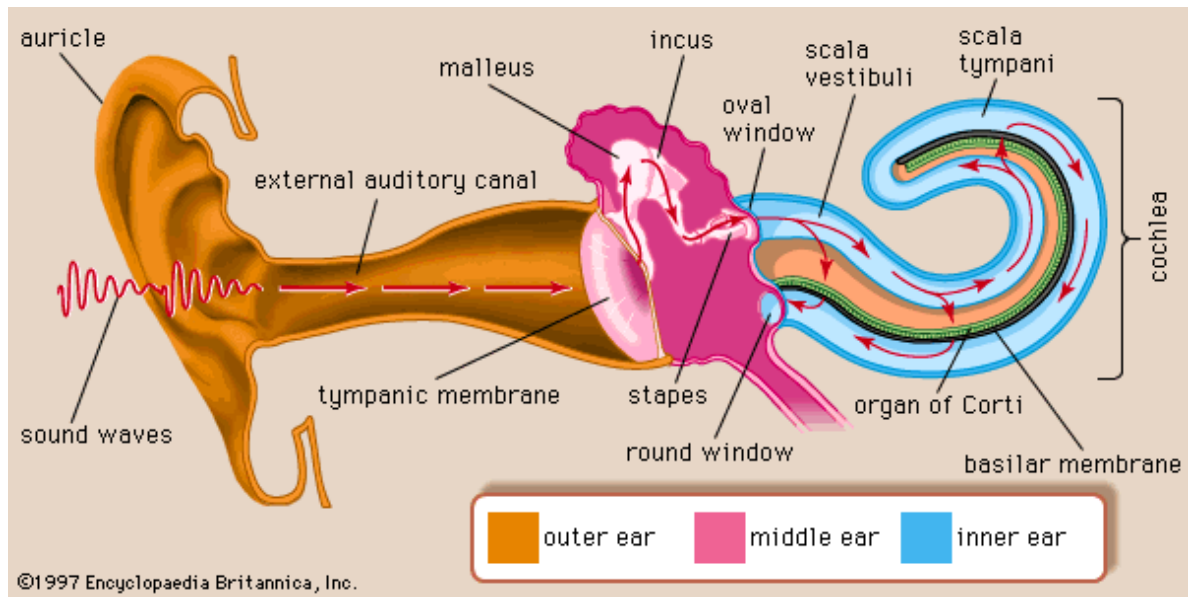
In speaker identification process, an input speech data must be compared to all speaker models in the database of the system thus increasing the number of authenticated speakers. This in turn might lower the performance speed of the system. Meanwhile this is not true for the case of speaker verification, where the system just needs to make comparison with only the claimed identity whether or not the claimant speaker is who he/she is, therefore increasing the system speed. The application of speaker verification

as opposed to identification usually is in the form of security reasoning where an individual is usually asked to verify whether he/she is the individual claimed to be (e.g. to access a bank account).

The key performance of a speaker verification system is often measured through false rejection rate (FRR) and false acceptance rate (FAR) where both of the rates should be set as equals, known as equal error rate (EER) to make sure a fair way in determining the acceptance/rejection threshold. The threshold value is the point to accept or reject the claimant identity and must be set carefully since setting the value too high might cause the system to be too strict, otherwise it might be too easy to break.

#### ***1.4. Human Auditory Nerve Pathway***

The auditory nerve (AN), also known as the cochlea nerve is a complex network that links our hearing system (ear) with the nervous system so that the information received can be interpreted. The structure of the hearing system consists of three main parts; namely the outer ear, middle ear and finally the inner ear, with each consists of separate types of membrane (outer: tympanic; middle: oval window; inner: basilar membrane). All these parts play an important role in the auditory system.

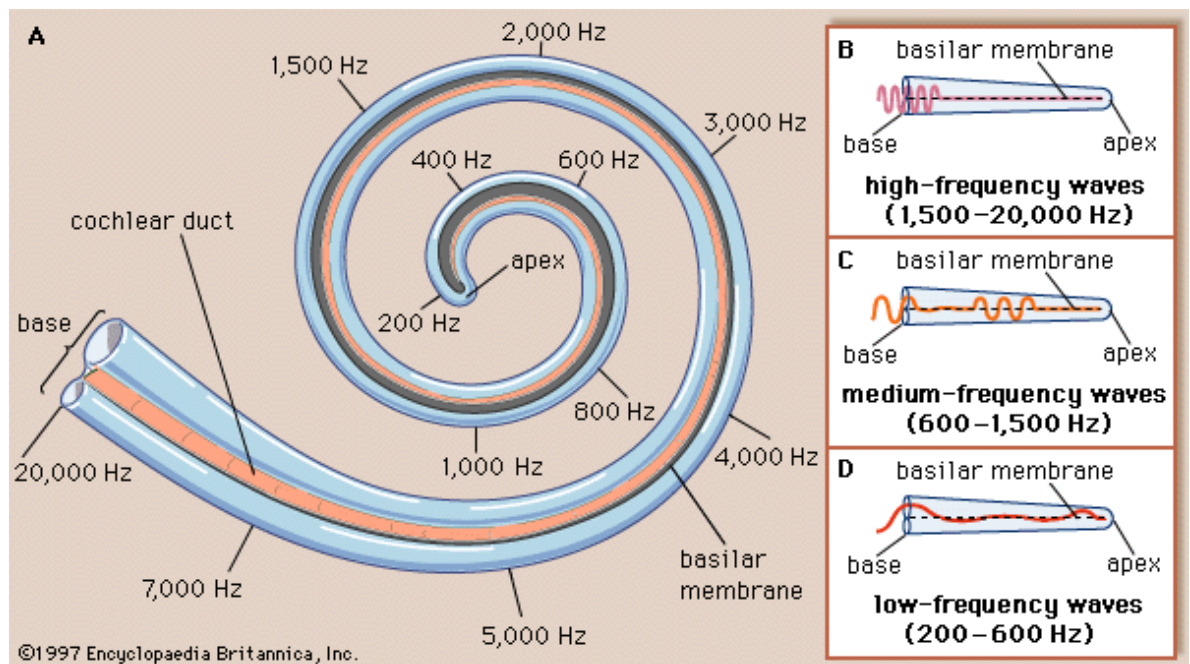


**Figure 1.2: A cross section of auditory pathway of the organ ear (Reproduced from Encyclopedia Britannica, Inc. 1997).**

The auditory pathway can be understood as depicted in the Fig. 1.2 above. The sound waves arrive at the outer ear, which consists of the pinna (earlobe), auricle, and external auditory meatus (ear canal). The pinna acts as an antenna to catch the sound waves that are received as a pressure wave and sends the acoustical energy wave into the external auditory meatus where at its end is the tympanic membrane (ear drum), which is a part of the middle ear. The main function of the outer ear, aside from catching the sound wave, is to act as a 'pre-amplifier' of the sound wave to around 3 kHz, which is the optimal sound frequency, and it further increases the amplification in the ear canal to about 12 kHz. Furthermore, the pressure wave received by the tympanic membrane is transferred to the middle ear section, which consists of the ossicle bones through vibration (mechanical) energy. The mechanical transduction of the ossicles is transferred to the oval window, where it connects to the inner ear part, which is the cochlea. The vibration causes the fluid inside the cochlea to move, therefore causing the neural receptors; outer hair cells (OHC) and inner hair cells (IHC), which connect to the basilar membrane to



bend, subsequently causing the transformation to the neural spikes of the auditory nerve bundle.



**Figure 1.3: Pictorial representation of cochlea and the frequency band accepted on different parts of the basilar membrane from base (highest frequency) to apex (lowest frequency) (Reproduced from Encyclopaedia Britannica, Inc. 1997).**

The auditory nerve that lies along the basilar membrane have different best or characteristic frequency (CF) based on its different position along the basilar membrane. This can be further understood by referring to the Fig. 1.3 above. The base of the basilar membrane has auditory nerve that is more tuned to higher frequency, and this decreases as the basilar membrane reaches 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 5kHz) through phase-locking process (Gold, 2000), which has been successfully captured by the AN

model and will be discussed in the next chapter . Detailed discussion of the role of cochlea as a filter will be discussed in Chapter 2.

### **1.5. Objective**

The main goal of this study is to develop a neural response-based speaker verification system instead of systems based on the properties of the acoustic signal. As the AN model by Zilany and colleagues (Zilany et al., 2009) captures most of the nonlinearities observed at the level of the AN, the performance of the proposed system is expected to be comparable with the behavioural performance of human subjects. The objectives of the project are:

- to get the AN model response for speaker verification process for speech processing technique using Matlab ®.
- to test the system accuracy by increasing the performance of the GMM distribution.
- to test the system's robustness by introducing Gaussian white noise into the tested speech samples.

### **1.6. Scope of study**

The study involves the development of text-dependent speaker verification system. The speech samples were recorded to be used as the database corpora for this study. The speech samples are analysed starting from pre-processing, getting the AN neural response, applying feature extraction method, training the system using GMM classification technique and finally to test the verification system using the computed GMM models for each speakers. Robustness of the system is also tested by using

simulated noise speech sample in the program. The performance of the system in clean and noisy conditions are calculated initially based on different Gaussian components numbers and for both verification and identification system. Finally, the reliability of the system is tested to make sure that the instrumentation used gives out reliable output using statistical analysis.

### ***1.7. Outline of the report***

This research report consists of 5 chapters. In chapter one, the introduction of the auditory pathway is discussed. A brief introduction on speech or speaker recognition system is provided that is going to be implemented in this study. The scope of the study and the main objectives are also covered in this chapter.

In chapter two, a historical background and study of speaker verification using an AN model is briefly discussed. The background theory of GMM as the classification technique is described, as well as feature extraction method.

In chapter three, the methodology involved in accomplishing this project is discussed in details. Four stages such as pre-processing, AN model response with or without transformation, training, and finally testing with the GMM will be elaborated in this chapter.

Chapter four discusses the result of the speaker verification system along with the system's overall performance and accuracy with the support of statistical tests.

Finally in chapter five, the report is concluded with some discussion about the limitation and future work.

## **Chapter 2. LITERATURE REVIEW**

This chapter will discuss on previous study that has been made in developing the AN model to understand the origin and the background theory used. The background theory of the AN model used in this project (Zilany et al., 2009) is then discussed in detail. Next, several speaker verification studies done in the past using different AN models are also discussed to illustrate the potential of applying the model in speaker recognition system. The background theory of GMM classification technique is also explained, and finally, the theory behind the Krawtchouk polynomials to be used in feature extraction is discussed.

### ***2.1. Auditory Nerve (AN) modelling***

Computational auditory modelling has been proposed by researchers since 1960 (Flanagan et al., 1960). The model is loosely correlated with the physiological study of human basilar membrane and cochlea stimulation of the cadaver. The first attempt on modelling the AN by Flanagan and colleagues is through computational model of the middle ear and the basilar membrane. The author uses the assumption that the sound wave is perceived by the basilar membrane in the pressure waveform and from then, the energy is transduced through mechanical energy by the ossicles. The obvious problem with Flanagan's model is known years after that; the initial hypothesis was that the cochlea is linear and thus making assumption that the cochlea itself is passive (Flanagan,1960) is not true. This is mainly caused by the inactive cochlea of the cadaver used during modelling.

Further development of the AN model is realized through hardware implementation using complementary metal–oxide–semiconductor (CMOS) technology. According to Lyon, (1988), the auditory model moved to modelling the non-linearity of the system through the fluid-dynamic wave of the cochlea. In this model, the OHC model is included by using a set of automatic gain controls to simulate the dynamic compression of the OHC on the basilar membrane that give the neural spike output. The gain of the OHC is the analogy of the gain for the CMOS transistor in the model. The active gain resulted from the OHC model is used as a control unit for the IHC part. IHC modelling is also implemented in the model that acts as a half-wave rectifier, where physiologically, the IHC only generates neural spike when the IHC stereocilia is deflected in one way only and not the other way around (Lyon, 1988). A further in-depth study for AN model particularly for connection of the IHC in the synaptic cleft is done by Meddis in 1986.

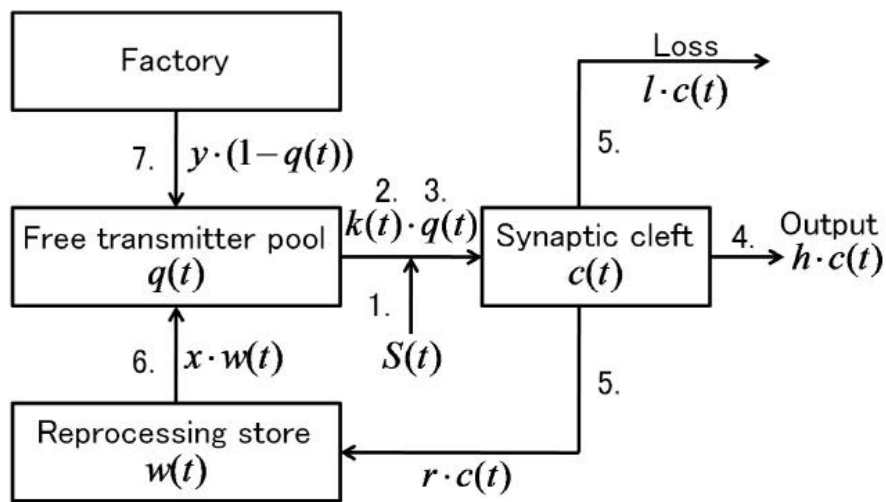
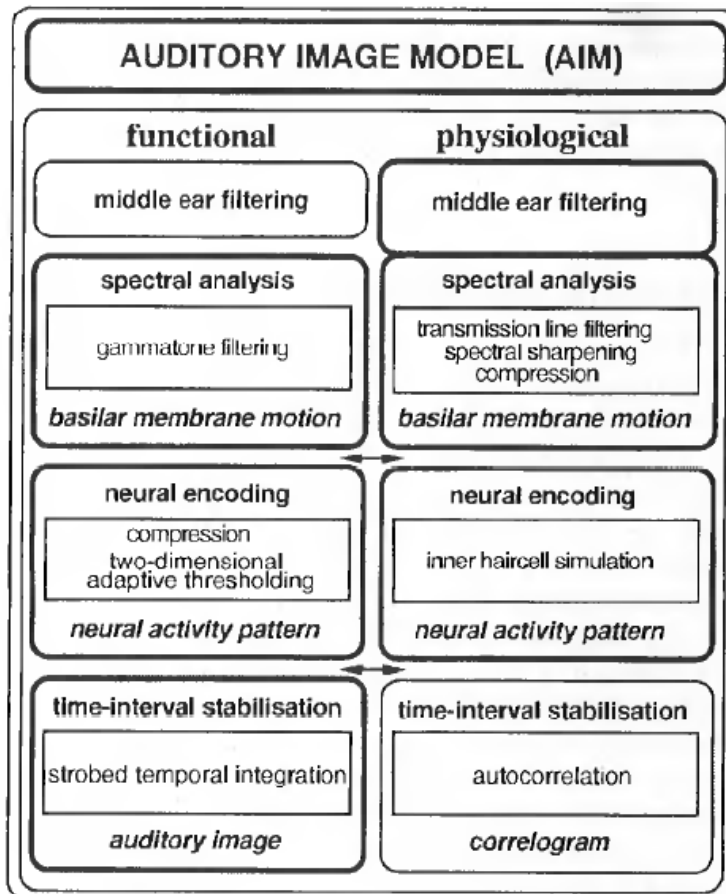


Figure 2.1: The signal flow of Meddis inner hair cell model (Reproduced from Meddis, 1986).

The signal flow of Meddis IHC model is depicted in Fig. 2.1. Based on the physiological activity of neurotransmitter released in the synaptic cleft between the IHC and AN fiber, the release of transmitter is controlled by using a function that is released from the transmitter pool on the IHC side to the synapse (Meddis, 1986) where the series of functions is as in Fig. 2.1. The AN activity are represented by the amount of neurotransmitter released, thus the amplitude and frequency of the spike that produced from the output of the AN model will be generated after the excitation triggered by neurotransmitter.

Another study of AN modelling by Patterson et al. (1995) is focused on the output of the model itself in the form of auditory imagery, where the supposedly produced sound is processed in term of graphical image of itself. For example, visualizing the pitch, loudness and tempo in different note and frequency level processed automatically in the brain.



**Figure 2.2: AIM model representation (functional) compared to actual physiological activity of AN in human in three stages; spectral analysis, neural encoding and auditory image (Reproduced from Patterson, 1995).**

Patterson et al. (1995) uses the Meddis IHC modelling in his model to form a time-domain model for auditory processing technique known as the Auditory Image Model (AIM) (Patterson et al., 1995). The structure of the AIM model composed of three stages is as illustrated in Figure 2.2. First, the basilar membrane motion is analysed using auditory filterbank (Gammatone filter) as a result of sound produced in the cochlea in the middle ear. In the second stage, a bank of neurotransmitter holding functions are activated (as been investigated by Meddis et al. 1986) that converts the basilar membrane motion (BMM) to neural activity pattern (or generating of action potential) by rectifying and compressing the BMM where suppression and time

adaptation is applied across the frequency to produce the neural activity pattern (NAP). Finally, the temporal activity of the NAP channel that might have repeating patterns is stabilized and summarized by applying strobes temporal integration to produce the auditory image of the sound (Lyon et al., 2010; Patterson et al., 1995).

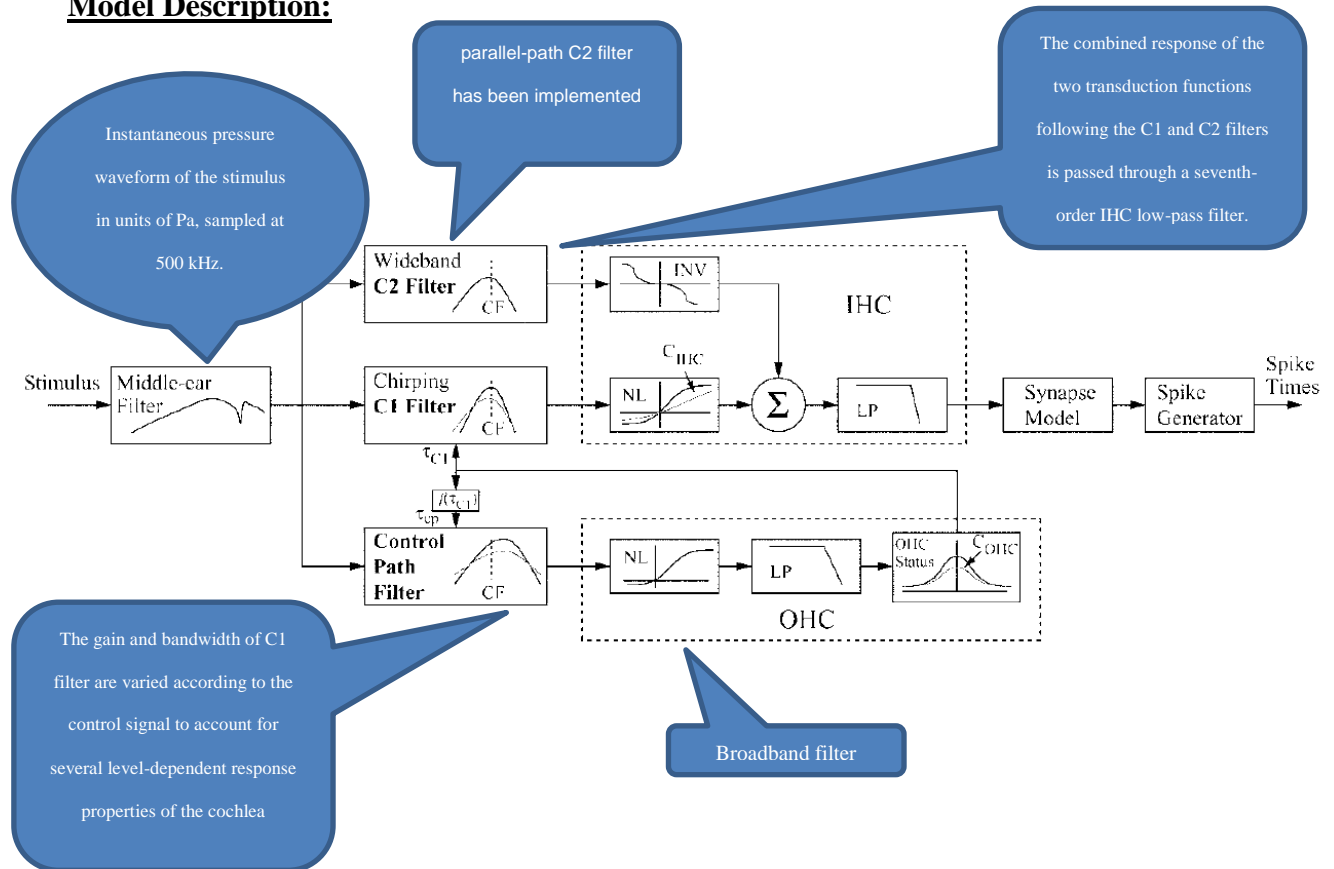
Another AN model based on the Gammatone filter has been introduced as a cochlea model to model the cochlea's forward transfer function (Johannesma, 1972). The model however was improved and re-named to auditory transform (AT) combining both forward transform; where the speech was decomposed through a bank of cochlea filters into several frequency bank readings; and inverse transform which is where the original speech signal are reconstructed based on the decomposed bandpass signals in the first step to retain information that might loss during the forward transform (Li, 2009; Li & Huang, 2010).

### 2.1.1. AN Model

Compared to previously mentioned AN models, the Zilany and Bruce (2006, 2007) model was improved by introducing two modes of basilar membrane that includes the inner and outer hair cells resembling the physiological basilar membrane function in two filter components C1 and C2. The IHC corresponds to component C1 where it filters low and intermediate responses. Meanwhile, C2 corresponds to OHC which filters high response and then followed by C2 transduction function to produce high-level effects and transition region. This feature in the Zilany-Bruce model causes it to be more effective on wider dynamic range of bands of frequency model of the basilar membrane compared to previous AN models. (Zilany & Bruce, 2006, 2007; Zilany et al., 2009). Figure 2.3 below shows the model of the auditory periphery model by Zilany-Bruce:



### Model Description:



**Figure 2.3: Zilany-Bruce AN model pathway (reproduced and edited from Zilany and Bruce, 2006).**

The model consists of four main filters path, namely the middle ear (ME) filter, two parallel filter (C1 and C2) paths, the feed-forward control filter path that is controlled by the C1 filter, and finally the IHC and OHC filters paths. Any speech or stimulus input is made through the ME filter, in which where the signals is measured in the unit of Pascal (Pa) and is sampled again at 500 kHz to match the overall frequency response of the AN model at 1 kHz (Zilany et al., 2006). The output of the ME filter is then used for the C1 filter. In the next section, the functions of C1, C1 with feed forward control path, C2, IHC and OHC filters will be discussed separately to ease the understanding of the system.

### C1 Filter:

The C1 filter is a type of linear chirping filter, which is used as a feed-forward control path where the output is used to tune the gain and bandwidth in making it the same as the cochlea's level-dependent frequency response. This tuning property is controlled by C1 transduction function which is then used as the input of the C1 IHC transduction function. The C1 filter's configuration is made with an asymmetrical orientation of the second order poles and zeros different damping coefficients in the complex plane in the impulse response of the C1 filter. To enable the filter to tune broadly, the C1 filter order is set at 10<sup>th</sup> order and this resulted in the filter to simulate AN CF fiber up to 40kHz compared to previous design by Tan & Carney (2003), thus increasing the tuning range.

### Feed forward control path (including OHC):

The function of this path is to reflect the active processes in the cochlea by regulating the bandwidth and gain of the BM by using the output of the C1 filter based on different level of stimuli. This is where the nonlinearity of the AN model that represents an active cochlea is modelled. There are three main stages involves in this path, which are:

- a) Stage 1: Gammatone filter (A gammatone filter is a linear filter described by an impulse response that is the product of a gamma distribution and sinusoidal tone) that has a broader bandwidth than C1 filter.
- b) Stage 2: Boltzmann function followed by a third order lowpass filter that controls the time course and dynamic range of compression.
- c) Stage 3: a nonlinear function that converts the lowpass filter output in stage 2 to a time-varying time constant for the C1 filter

Any impairment of the OHC is controlled by the  $C_{OHC}$  function in stage 3 and the output is used to control the nonlinearity of the cochlea as well. Moreover, the nonlinearity of cochlea is controlled inside the feed forward control path based on different type of stimulus of the sound pressure levels:

- a) low stimulus: The control-path output is almost equal to when the control path output is maximum, in which it has high gain and sharp tuning point, causing the filter to act linearly.
- b) moderate stimulus: The control-path output signal deviates substantially from the maximum control path output that dynamically varying between maximum and minimum output value. The C1 tuning filter broadens, while the gain reduced and resulting in the filter to behave nonlinearly.
- c) High stimulus: The control-path output signal saturates, that equals to minimum control path output. The C1 filter is again effectively linear with broad tuning and low gain.

#### C2 Filter:

C2 filter is a wideband pass band filter in which it is similar to the C1 filter with its broadest possible tuning (i.e. at 40 kHz). The implementation of C2 filter is based on Kiang's two-factor cancellation hypothesis, in which the level of stimuli will affect the C2's transduction function followed after C2 filter's output. The hypothesis states that 'the interaction between the two paths produces effects such as the C1/C2 transition and peak splitting in the period histogram' (Zilany et al., 2006). The transduction function gives off the output based on sound pressure levels that affect the C1/C2 interactions at:

- a) low sound pressure levels, its output is significantly lower than the output of the corresponding C1 response.

- b) high sound pressure levels, the output dominates and the C1 and C2 outputs are out of phase.
- c) Medium sound pressure levels, where C1 and C2 outputs are approximately equal and tend to cancel each other.

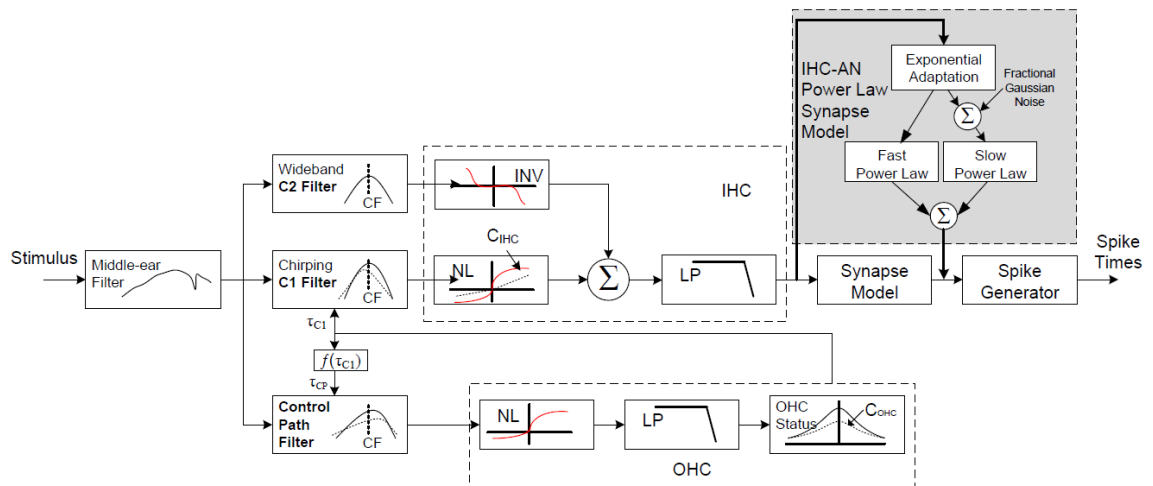
Furthermore, the C2 response is not subject to rectification, unlike the C1 response (at high levels) such that the peak splitting phenomenon also results from the C1/C2 interaction. Poor frequency selectivity of AN fiber is caused by too many frequency components consists in a speech stimuli. This is overcome by increasing the order of C2 up to 10<sup>th</sup> order, which compensate the order of C1 filter.

#### IHC:

The IHC is modelled by a low pass filter that functions to convert the mechanical energy produced by the basilar membrane to electrical energy that stimulates the neurotransmitter to be released in the IHC-AN synapse. Two types of IHC; tallest and shorter types; generate the C1 and C2 responses respectively and were controlled by both C1 and C2 transduction functions. C1 transduction function uses the output of the C1 filter and is related to high-CF model fibers to produce the direct current (DC) components of the electrical output. Meanwhile, the C2 transduction function uses the C2 filter output that is first transformed to increase towards 90-100 sound pressure level at low and moderate-CF level. Finally, the C1 and C2 transduction function outputs,  $V_{ihc,C1}$  and  $V_{ihc,C2}$  are summed and resulted to the overall potential of  $V_{ihc}$  output after passing through the IHC lowpass filter.

The spontaneous rate, adaptation properties, and rate-level function of the AN model are determined by the model of the IHC-AN synapse. The spike timings are provided by a non-homogenous Poisson process driven by the synapse output.

Finally, discharge times are produced by a renewal process that includes refractory effects and is driven by the synapse output. The output of the AN model simulates multi-dimensional pulse signals from each channel that is obtained by means of its statistical characteristics of the pulse signals called the peristimulus time histogram (PSTH).



**Figure 2.4: Zilany-Bruce AN model (2006) with added PLA model (shaded) (Reproduced from Zilany et al., 2009).**

Figure 2.4 shows the same model by Zilany-Bruce as in 2006 but with the additional rate-adaptation model which is the IHC-AN Power- Law Synapse Model (PLA) indicated in the shaded area in the figure (Zilany et al., 2009). In this model, the introduction of the PLA model is used to further adapt and shape the output of the IHC exponentially into two separate fast and slow adapting responses. These responses further made the AN output to improve the AN response after stimuli offset, in which

the person could still hear a persistent or lingering effect after the stimuli has past and also to adapt to a stimuli with increasing or decreasing amplitude.

The adapting power-law adaptation in the synapse model significantly increases the synchronization of the output to pure tones, and therefore, the adapted cut-off frequency is matched with the maximum synchronized output of the AN fiber for pure tones as a frequency function. In previous model (Zilany & Bruce, 2006) the model output only simulates a single repetitive stimulus of the synapse. Whereas in the 2009 model, the PLA model simulates repetitive of the stimulus output of the synapse into a single IHC output. Because of the discharge generator has quite a relatively long lifetime emission dynamics and can be extended from one stimulus to the next, a series of the same output synapses were formed through a combination of repetitive stimulus and silences between each stimuli. Moreover, generally the model synaptic PLA also has memory that exceeds the repetition duration of a single stimulus (Zilany et al., 2009).

### 2.1.2. Envelope (ENV) and Temporal Fine Structure (TFS)

The output of an AN model are typically visualized through an electrical recording of the peripheral (auditory) nerve, called the neurogram. The neurogram in describing speech contents are typically represented by two types of measurements, called the temporal envelope (ENV) and temporal fine structure (TFS). The difference between the two is that ENV averages the PSTH output of AN model intensity at each CF over a number of time frames and the speech is represented in a smooth average discharge rate. ENV usually translates into how the speech is articulated, vowel identity, prosody of speech and voicing manner of the speaker (Hines & Harte, 2012). Meanwhile, the TFS contains the fine timing structure of the AN spikes that happens between periods of a periodic signal that usually carries the formant information of the speech. TFS

neurograms preserve spike timing information and the synchronisation to particular stimulus phase, or phase-locking phenomenon. Both features are provided useful in measuring the intelligibility index of a speech (Hines & Harte, 2012). The AN model takes the speech stimulus as an input and produces synapse response as output for a range of CFs and depicted in the forms of ENV and TFS.

## ***2.2. Speaker verification/identification based on AN response***

Researchers have tried to implement the use of auditory model response in the field of speaker recognition as early in the 90s. A study by Colombi et al. (1993) uses KING database corpus (English) implementing an earlier AN model based on physiological data developed by Payton (1988) to generate the output. This model like Flanagan's (1960), assumes that the basilar membrane frequencies are linear and does not incorporate filter banks causing more unnecessary speech features to be included in the AN response, therefore reducing its speed performance. Colombi et al. uses vector quantization (VQ) codebook classification technique by applying self-organizing mapping process called the Kohonen map with and without neighborhood; and also Linde-Buzo-Gray (LBG) algorithm to design the VQ speaker codebook for training in this study. 10<sup>th</sup> order linear predictive coding (LPC) analysis was also done to compare the result with the AN model response for speaker identification. The result shows an increase of 5% accuracy rate for VQ algorithm applied to Payton's AN model response compared to LPC cepstral coefficient method (Colombi et al., 1993).

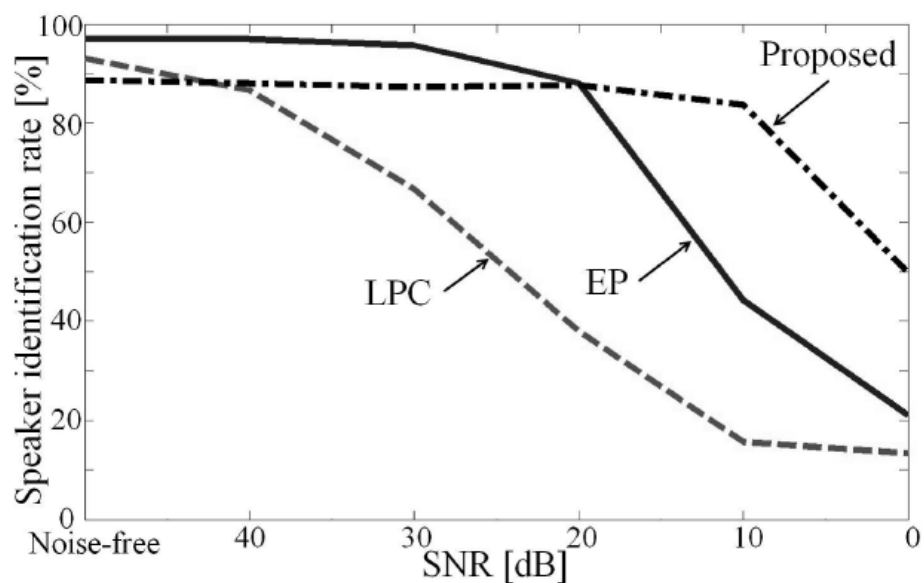
Meanwhile, a series of studies done by Li, Q. from the year 2003 on implementing the AN model for speaker recognition field. In 2010, the author proposed to use his

developed AN model in 2003 to be implemented into a speaker identification system, by naming their technique as Auditory Transform (AT) that is based on both forward and inverse algorithm using Gammatone filter as the cochlea model for their project in 2010. Their AN model is based on AN model by Johannesma, (1972). They intended to value the accuracy of the system by calculating the output of the AT algorithm into a cochlea feature cepstral coefficient (CFCC) as opposed to the mel feature cepstral coefficient (MFCC) using general LPC technique. 34 speakers from the Speech Separation Challenge database is used in this project and the effect of using applying different signal to noise ratio (SNR) levels on the AN model response is also studied. The result of the study shows the CFCC outperforms the MFCC based identification system having 90.3% accuracy using the CFCC compared to only 42.1% resulted using common MFCC technique (Li, 2003; Li, 2010).

Another study by Abuku et al. in 2010 also uses AN model in speaker recognition field. The feature vector used in this study is extracted directly from the PSTH of AN model based on Meddis, (1986) IHC model that is enhanced with phase-locking model by Maki et al., (2009) based on 12 Japanese speakers (vowels). Two additional steps to increase the system's accuracy are applied on the training data: standardization and normalization. The output of the speaker recognition study was compared to conventional method by using LPC analysis. A pattern recognition method (Nearest neighbour method) was applied instead of typical classification technique. The result of the study shows that the average of the speaker identification accuracy are highest by using standardization and normalization of the PSTH output (86.6%) compared to LPC analysis (80.6%) (Maki et al., 2009; Abuku et al., 2010; Azetsu et al., 2012) .



A similar study done based on Abuku in 2012 is done to detect the speaker identification accuracy of using the same AN model in Abuku et al., (2010). The robustness performance is tested by setting the threshold of the action potential of the AN model, much the same as done by Abuku et al., but more in depth on the threshold factor in determining the system's accuracy. The vector feature resulted from the PSTH was then classified using Difference of Gaussian method that also increases the frequency resolution of the training data. Three types of noises (white, pink and blue) are induced and this time, subspace method of pattern recognition is applied as the feature extraction method (Azetsu et al., 2012).



**Figure 2.5: Speaker identification rate versus SNR (reproduced from Azetsu, 2012).**

Figure 2.5 shows the speaker identification rates versus signal to noise ratio (SNR) by each method under noises in the study done by Azetsu, 2012. In case when the noise level is low, the human peripheral auditory model has a less performance than the other methods. However it is better than the other methods as SNR decreases the EP

(Excitation pattern), LPC (LPC Spectrum) performance compared to the proposed method (AN model output) when the noise level is increased.

The result clearly shows higher accuracy rates in proposed method (75%) compared to EP (55.3%) and finally LPC by 11.3% for speaker identification rate with induced white noise in the testing set of the speakers.

Robustness of the speaker recognition system is also the main point addressed by authors Shao and Wang in their paper in 2007. The authors also uses Gammatone filter to model the human cochlea filtering process and the feature extracted from the filter is called the Gammatone feature (GF) in which 32 orders of GFCC is derived from. The authors also applied a previous method also proposed by the same author called the missing data method, where the noise in a speech sample is treated as a missing data in the feature of the speaker that requires the application of binary mask to conclude it is a missing feature or not. The assumed missing or corrupted GF is reconstructed based on the *a priori* data derived from the speech training set which is similar to using a Universal Background Model (UBM). To extract the GFCC, discrete cosine transform (DCT) is applied to the GF to its cepstral domain, much similar to how MFCC is obtained in spectral analysis. The GFCC is then trained using GMM classification and is compared to several changes in feature extraction parameters as shown in Figure 3 in the particular study. The accuracy of the proposed GFCC-based features is shown as the highest compared to the others having  $\pm 55\%$  accuracy at  $-6$  dB SNR level (Shao & Wang, 2007).

Table 1.1 summarizes previous studies on speaker verification or identification using AN response.

**Table 1.1: List of previous speaker verification studies using AN model response.**

No	Study	Database	Feature Extraction for AN response	Classification Type
1	Columbi, J.M. (1993)	KING	AN+ Kohonen Mapping; LBG Algorithm	VQ vs LPC
2	Abuku, M. (2010)	Japanese speakers	AN + normalization/ standardization	LPC
3	Li, Q. (2010)	Speech Separation Challenge Database	AN+CFCC	GMM vs LPC
4	Azetsu, T. (2012)	Japanese speakers	AN+Difference of Gaussians	Subspaced method
5	Shao, Y. (2007)	-Not stated	Feature Extraction for AN response	GMM

For conclusion, all of the past studies that use AN model response in speaker verification/identification system shows a higher accuracy result compared to the accuracy of using conventional LPC/MFCC based spectral analysis.

### **2.3. Classification for speaker verification**

In order for the speaker verification process to work, a classification technique as a supervised machine learning process is required for the system to be trained so that a

series of observations can be made with a training set whose membership is known to the system. Compared to unsupervised learning process, classification technique requires known data from a category for training process. In a common speaker verification process, the system will initially need several speeches of the speakers for training data during the system set up.

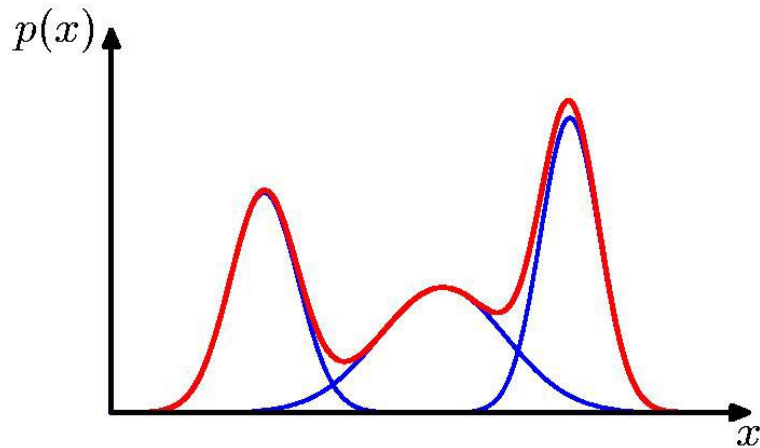
An algorithm that specialized in classification is called a classifier, which is comprised of mathematical functions that works with the raw data that classify or ‘maps’ a training set to a particular category or clusters. The classifiers are usually based on statistical analysis that gave outputs based on the highest probability of a feature vector to belong in a certain class.

There are a lot of classifiers that can be used depending on the type of training set, including neural networks, Gaussian mixture model (GMM), Hidden Markov Layer (HMM), k-nearest neighbours, Bayes classifier, Support Vector Machine (SVM) etc. For this project, GMM classification technique is chosen as it should be better to represent a model with high computational input data which is true for speech signal through AN model response.

### 2.3.1. Mixture Model: Gaussian Mixture Model

GMM is one of the first and mostly used technique for training model in speaker recognition field (Li & Huang, 2010; Reynolds et al., 2000; Reynolds, 1995). It can be used as itself or with the combination with other classification techniques depending on the number of observations and tuning factors, such as with Support Vector Machine (SVM) (Togneri & Pullella, 2011). The GMM itself is the combination of Gaussian density that corresponds to a class and the Expectation Maximization (EM) algorithm to solve a database that consists of parameter-estimation problem for data training. The

Gaussian model for finding densities assumes that the feature vectors of the training data follow a Gaussian distribution.



**Figure 2.6: GMM distribution in 1 dimension.**

Figure 2.6 depicted three different feature vectors observations of  $x$  (blue lines) normally distributed based on its probability  $p(x)$  and whereas the overall probabilities could be represented by combining all three Gaussians into a single mixture of Gaussian density (red line) through its probability density function (PDF) of the original observation.

The mixture of the Gaussians of the feature vectors forms a distribution for a particular speaker. Meanwhile, the Gaussian densities are characterized by estimating three parameters; the means, variances and deviation about the mean. The GMM algorithm needs to estimate these quantities but there is no way to know which features belong to which Gaussian distribution. Therefore, the combination of EM method is added as an optimization of the Gaussian mixture that will maximize the likelihood of the observed data.

Assuming that a GMM model for the speaker  $j$  represented by  $\lambda_j$ , is defined as the sum of all  $K$  components of the Gaussian densities for the feature vectors ( $x_t$ ) of that particular speaker. Defining the probability of  $x_t$  based on the GMM model or its weighting probability function as:

$$p(x_t | \lambda_j) = \sum_{i=1}^K g_i \mathcal{N}(x_t; \mu_i, \Sigma_i) \dots \dots \dots \text{Eq. (1)}$$

$\mu_i$  = mean for feature vectors

$\Sigma_i$  = covariance matrix

$\mathcal{N}$  = individual component densities parameterized by the feature vector, mean vector and covariance matrix for a  $D$ -variate Gaussian function. Meanwhile, the GMM model for speaker  $j$  is defined as:

$$\lambda_j = (\mu_i, \Sigma_i, g_i)_j, \quad i = 1, 2, 3, \dots, K. \dots \dots \dots \text{Eq. (2)}$$

with  $g_i$  is the mixture weight. The linear weighting function (Equation 1) can be used as a function or controlled constant to use as a transition form from an acoustic class to another. With enough number of Gaussian density components  $K$  particularly for text-dependent case (where there was not enough cumulative values of the overall phonemes to be trained), the GMM is able to pool all possible features from a single speaker into their respective distinct phonetics features for a particular speaker. Text-dependency however is used in this project to allow lesser number of training data used and using specific prompted utterances for verification (Reynolds, 1995; Togneri & Pullella, 2011).

A Gaussian mixture model for a concatenated values for training samples ( $X$ ) can be defined as the weighted sum of  $K$  component Gaussian densities as given by the equation,

$$P(X|\lambda) = \sum_{i=1}^K w_i g(X|\mu_i, \Sigma_i) \dots \dots \dots \text{Eq. (3)}$$

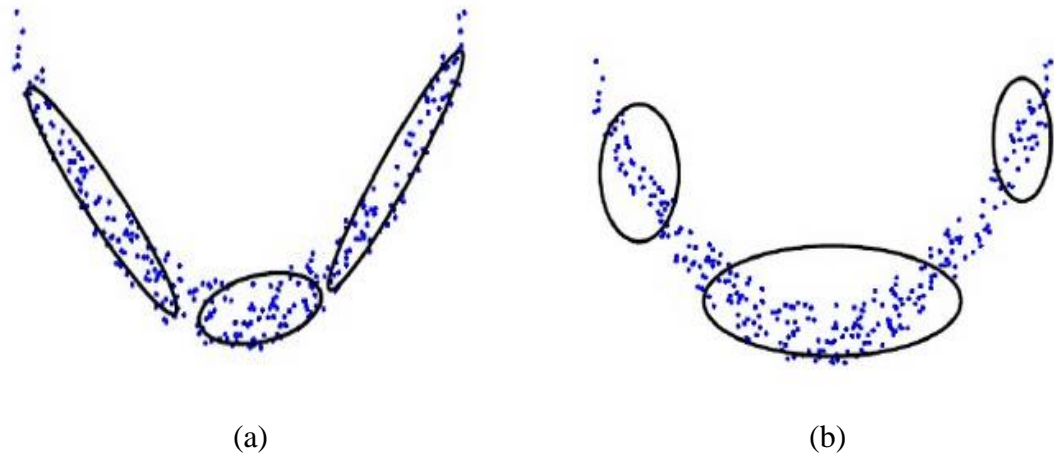
where  $w_i$ ,  $i = 1, \dots, K$ , are the mixture weights, and  $g(X|\mu_i, \Sigma_i)$  with  $i = 1, \dots, K$ , are the component Gaussian densities. Each component density is a  $D$ -variate Gaussian function in the form of,

$$g(X|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) \right\} \dots \dots \dots \text{Eq. (4)}$$

with mean vector  $\mu_i$  and covariance matrix  $\Sigma_i$ . The mixture weights satisfy the constraint that  $\sum_{i=1}^M w_i = 1$ .

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by Equation 2 mentioned before.

Although the use of GMM is a powerful classification technique, however, there are some disadvantages of the GMM classification depending on the type of application it can be used. First is the appropriate type of covariance matrix  $\Sigma_i$  that should be used inside the model.



**Figure 2.7:** (a) Full covariances, (b) diagonal covariances.

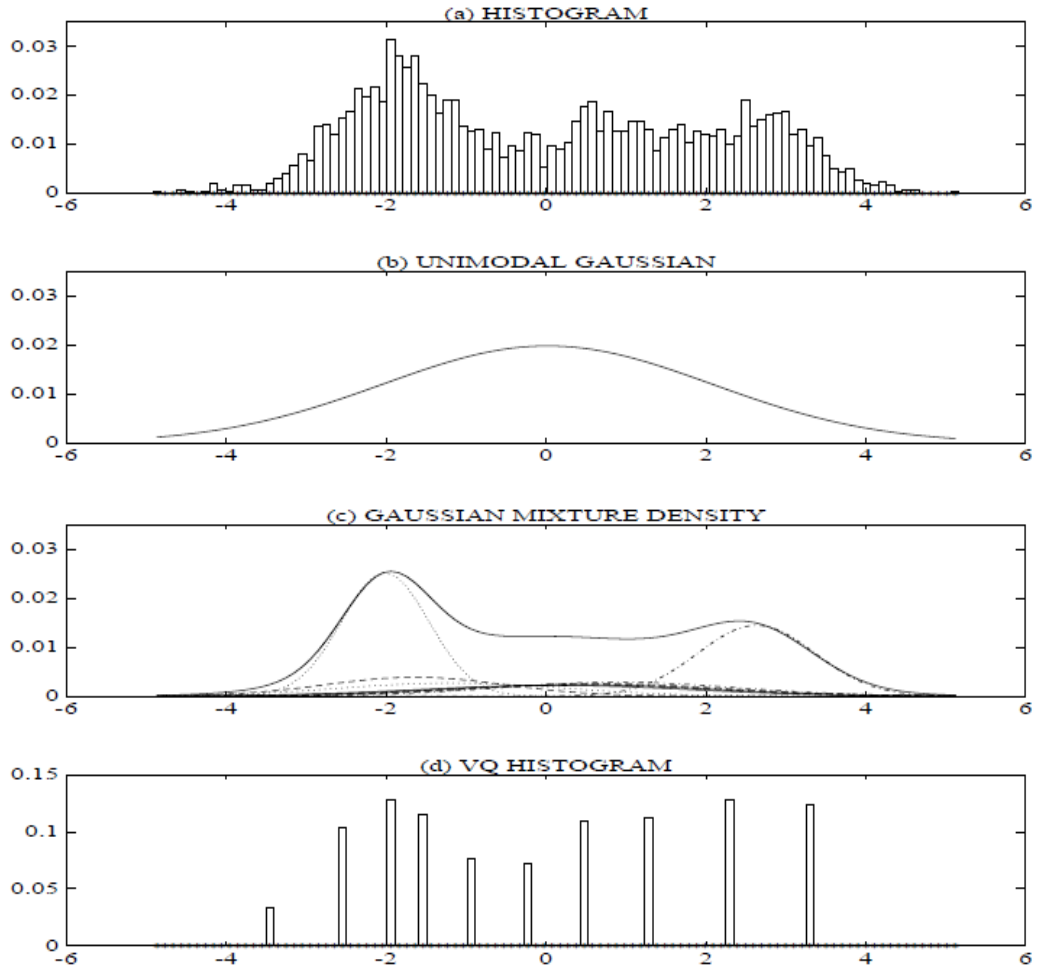
Figure 2.7 (a) shows that the default setting of the GMM in Matlab is in ‘full’ covariance type which means that all data in the speech samples are covered by the model, however this could lead to there was not enough data to be covered by all Gaussian density components ( $K$ ) provided in that time. One solution for this is to change the setting to ‘diagonal’ covariance matrix as in Fig. 2.7 (b) setting so that the system could place the feature observations into a less specific area. This somehow in turn, reduced the quality of the mixture as it might not cover all the observations. However, increasing the number of training data could overcome this problem without having to change it to diagonal covariance matrix. The data could also be decorrelated beforehand (normalization, etc) to overcome this problem (Togneri & Pullella, 2011; Reynolds et al., 2000).

Second, is the singularity problem faced, which is the unseen data that is hidden the training could ‘pop out’ during testing data procedure that could lead to degrading the system’s performances. This usually resulted in very low resulting probabilities during testing phase. This problem is countered by introducing an additional computational method of UBM where in speaker verification, instead of just comparing the intended



speaker GMM model, the speech would also be compared to its UBM, also known as the 'imposter GMM' model. However, this method involves a greater number of Gaussian component densities  $K$  as with the increasing number of imposter speeches that is used, therefore reduced the speed performance. Finally, choosing the wrong number of Gaussian densities  $K$  could lead to improper training of getting the maximum likelihood on each iterations resulting in failure of getting global likelihood for all training data to be not specific (Reynolds, 1995; Reynolds et al., 2000).

The application of GMM in biometric system is often used especially in speaker recognition field due to its ability to represent a large scale of spectral features of a speaker into GMM model. The GMM is also powerful in term of its ability to smoothly approximate any features densities distributed in any shape. The use of GMM could be depicted as the combination of the classical unimodal Gaussian with the use of nearest neighbour algorithm by each mixture have their own covariance matrix, and mixture weights for a better modelling capacity.



**Figure 2.8: Comparison of distribution modeling. (a) histogram of a single cepstral coefficient from a 25 second utterance by a male speaker (b) maximum likelihood uni-modal Gaussian model (c) GMM and its 10 underlying component densities (d) histogram of the data (reproduced from Reynolds,1995).**

Figure 2.8 compares the densities obtained using a unimodal Gaussian model, a GMM and a VQ model. In (a), the histogram plot shows the original distribution of the all observations of the speaker from a 25 second utterance by a speaker. Meanwhile, plot (b), (c) and (d) shows the similar distribution based on the data in plot (a) using unimodal Gaussian, GMM and 10-element codebook VQ histogram respectively. The data clearly shows the shape of the GMM in (c) provides a smooth overall distribution

fit and closely follows the nature of the original density of the histogram of the observations as in (a) compared to (b) and (d) distributions (Reynolds, 1995).

The use of a GMM for representing feature distributions in a biometric system may also be motivated by assuming that the individual component densities may model some underlying set of *hidden* classes. For example, in speaker recognition, it is reasonable to assume the acoustic space of spectral related features corresponding to a speaker's broad phonetic events, such as vowels, nasals or fricatives. These acoustic classes reflect some general speaker dependent vocal tract configurations that are useful for characterizing speaker identity. The spectral shape of the  $i$ th acoustic class can in turn be represented by the mean  $\mu_i$  of the  $i$ th component density, and variations of the average spectral shape can be represented by the covariance matrix  $\Sigma_i$ . Because all the features used to train the GMM are unlabeled, the acoustic classes are hidden in that the class of an observation is unknown. A GMM can also be viewed as a single-state HMM with a Gaussian mixture observation density, or an ergodic Gaussian observation HMM with fixed, equal transition probabilities. Assuming independent feature vectors, the observation density of feature vectors drawn from these hidden acoustic classes is a Gaussian mixture (Reynolds, 1995; Reynolds et al., 2000).

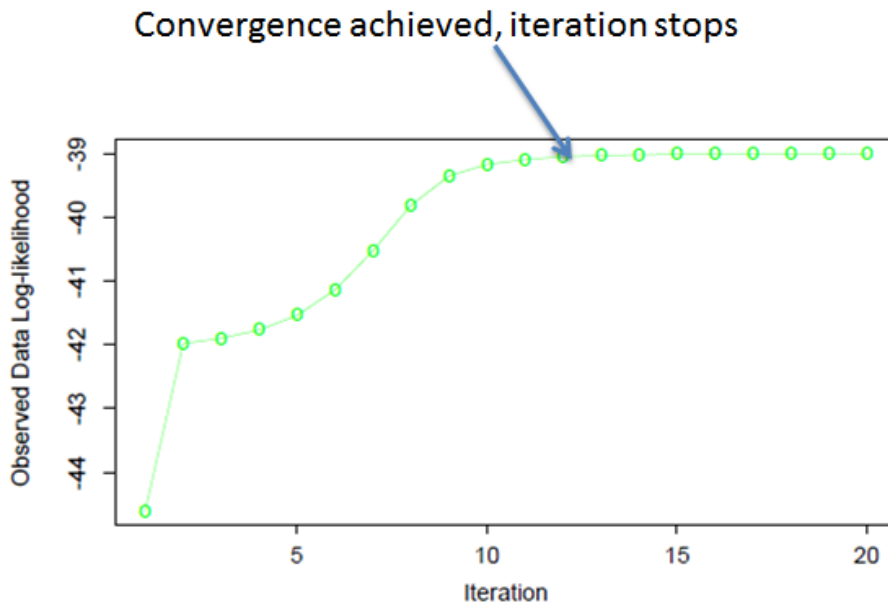
### **Maximum Likelihood Parameter Estimation**

It is now defined that the GMM model of a speaker could be represented by its parameters as defined in Eq. 2. However the estimation of these parameters should be made based on the given training feature vectors, by using a parameter estimation algorithm, the maximum likelihood estimation (MLE). There are also other algorithms that can be used for example Maximum *A Posteriori* method. However, MLE method

is considered to be well-established and the algorithm is automatically related to the GMM modelling aspect in the Matlab® itself. The aim of MLE estimation is to find the model parameters which maximize the likelihood of the GMM given the training data. For a sequence of T training vectors  $X = \{x_1, \dots, x_T\}$ , the GMM likelihood, assuming independence between the vectors, can be written as,

$$p(X|\lambda) = \prod_{t=1}^T p(x_t|\lambda) \dots\dots\dots \text{Eq. (5)}$$

The parameters used in MLE could be obtained by applying the iterative method of EM (Dempster, 1977). In this method, the first step which is the *Expectation*, *E* begins with an initial model  $\lambda$  as the value of log-likelihood using a randomly chosen data as the starting parameters of the weighting function. This step will evaluate the responsibilities which are defined as the conditional probability defined in Eq. 5 using the current (or initial) parameter values. This value it is used to estimate a new model  $\bar{\lambda}$  such that  $P(X|\bar{\lambda}) \geq p(X|\lambda)$ . The second step, the *Maximization*, *M* re-estimates the parameter using the current new responsibilities of the new model  $\bar{\lambda}$ .



**Figure 2.9: EM algorithm depicted as data observed log-likelihood as a function of the iteration number.**

As in Figure 2.9, this step will again evaluate the log-likelihood by checking the convergence of either parameters or the log-likelihood criterion is satisfied and the iteration will stop. If it is not, the algorithm will return to the *Expectation* step. With this, the overall likelihood increases at each iteration step. The alternating *Expectation* and *Maximization* process is repeated until some convergence threshold is reached to at least a local maximum likelihood (global is better, based on K number). The new parameters based on the new  $\bar{\lambda}$  are defined as:

*Mixture Weights*

$$\bar{w}_i = \frac{1}{T} \sum_{t=1}^T Pr(i|X_t, \lambda) \dots \dots \dots \text{Eq. (6)}$$

*Means*

$$\bar{\mu}_i = \frac{\sum_{t=1}^T Pr(i|X_t, \lambda) X_t}{\sum_{t=1}^T Pr(i|X_t, \lambda)} \dots \dots \dots \text{Eq. (7)}$$

*Variances (for diagonal covariance)*

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^T Pr(i|X_t, \lambda) X_t^2}{\sum_{t=1}^T Pr(i|X_t, \lambda)} - \bar{\mu}_i^2 \dots\dots\dots \text{Eq. (8)}$$

Where  $\bar{\sigma}_i^2, X_t$  and  $\mu_i$  refer to arbitrary elements of the vectors  $\bar{\sigma}_i^2, X_t$  and  $\mu_i$  respectively (Reynolds, 1995).

## **2.4. Feature extraction**

Feature extraction is one of the fundamental components needed in any speaker verification system in order to find the simplest feature vectors that can represent the whole speech regarding a particular speaker's identity. Getting a simplest form of feature vectors that does not include other unrelated noise is the main goal for getting a better system performance. Both high-level and low-level features has been proposed throughout decades although a trade-off for better result for high-level features compared to speed performance has to be done due to extensive computational effort. Usually, high level feature extraction involves modelling the AN and the human voice production, which is based on deriving the cepstral coefficient from linear prediction method while the former can be based on both Fourier transform and auditory filter bank (Li & Huang, 2010). The combination of both AN feature response and common extraction method does lower the speed performance as expected from this project. There are a lot of available feature extraction methods that can be used in speech recognition field generally such as mel-scale cepstral to get the MFCC features.

### **2.4.1. Krawtchouk Orthogonal moment for feature extraction**

Orthogonal moments is one of the method used in image processing that uses an image signal into a set of coordinates in the orthogonal polynomial basis. The polynomial is

used to compact data from time domain to moment domain that acts as a good signal descriptor for the speech signal. The scalar moments are then can be used to create a function that represents the related feature of the particular coordinate for that image. Orthogonal moment has been widely used in both image (Yap & Paramesran, 2003) and writing character recognition (Duval et al., 2010). Since the result of the AN model response is in the form of neurogram, orthogonal moments polynomial is considered to be applied in this project to extract usable feature of the neurogram in speech signal processing technique. An example of using such method has been made by using Chebychev polynomial in speech recognition by Carballo et al. (2001).

Due to their inherent properties such as translation invariance, rotation invariance, oscillating kernels, its ability to compact information in the selected range of interest (ROI) (by varying constant  $p$ ), and the ability to contain phase information of an image, orthogonal moments have successfully been employed in recognition applications (Rani & Devaraj, 2012). Furthermore, the computational load could be reduced because of the symmetrical property of the Krawtchouk orthogonal polynomials. Orthogonal moments have the ability to represent a signal using a limited number of moments without compromising signal quality. Different components such as plain, edge and texture of an image can be extracted using different types of filtrated procedure (Jassim et al., 2012; Yap & Paramesran, 2003).

Krawtchouk Orthogonal moment is a discrete orthogonal polynomial based on discrete probability distribution (binomial) of the data as oppose to more classical types of polynomial (Chebichev, Jacobi, Legendre, etc) that relies on continuous probability data (Jassim et al., 2012). Orthogonal in the sense of word means that the discrete data measurement is aligned orthogonally from polynomial model. Krawtchouk moments set

is formed based on Krawtchouk polynomials introduced by Mikhail Krawtchouk in 1929. The mathematical modelling of the Krawtchouk polynomials of a function  $f(x,y)$  is:

$$Y_{pq} = n_p n_q \iint_{\Omega} P_p(x) P_q(y) f(x,y) dx dy \quad p, q = 0, 1, 2, \dots \dots \dots \text{Eq. (9)}$$

where,  $n_p, n_q$  are the normalizing factors and  $\Omega$  is the area of orthogonality. The neurogram image  $f(x, y)$  are then scaled such that its support is contained in  $\Omega$ . The  $n$ th-order Krawtchouk polynomial is defined as (Yap & Paramesran, 2003):

$$K_n(x; p; N) = \sum_{k=0}^N a_{k,n,p} x^k = {}_pF_q \left( -n, -x, -N; \frac{1}{p} \right); \quad x, n = 0, 1, 2, \dots, N, N > 0, p \in (0,1) \dots \dots \dots \text{Eq. 10}$$

where the orthogonal polynomials can be defined using hypergeometric function,  ${}_pF_q$ , defined as (Jassim et al., 2012):

$${}_pF_q(a_1, \dots, a_p; b_1, \dots, b_q; z) = \sum_{n=0}^{\infty} \frac{(a_1)_n (a_2)_n \dots (a_p)_n z^n}{(b_1)_n (b_2)_n \dots (b_q)_n n!}, \dots \dots \dots \text{Eq. (11)}$$

Where  $(a)_n$  is a Pochhammer symbol given by:

$$(a)_n = a(a+1) \dots (a+n-1) = \frac{(a+n-1)!}{(a-1)!} = \frac{\Gamma(a+n)}{\Gamma(a)}, \dots \dots \text{Eq. (12)}$$

Meanwhile,  ${}_2F_1(a, b; c; z)$  derived from equation 11 is defined as:

$${}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n z^n}{(c)_n n!} \dots \dots \dots \text{Eq. (13)}$$

The polynomial  $r_n(x)$  can be defined as the sets of discrete orthogonal polynomials, such as Krawtchouk, with weight function  $\varrho(x)$ , within the interval  $[s_1, s_2]$  that satisfies the following orthogonality relation:



$$\sum_{x=s_1}^{s_2-1} r_n(x)r_m(x) \varrho(x) = v_n^2 \delta_{nm}, \quad s_1 \leq n, m \leq s_2 - 1, \dots \text{Eq. (14)}$$

where  $v_n^2$  indicates the square of the norm, and  $\delta_{nm}$  denotes the Dirac function. The normalization by the squared norm  $v_n^2$  is the traditional approach of avoiding numerical instabilities for coefficients computation. The weighted polynomial  $\tilde{r}_n(x)$  is defined as:

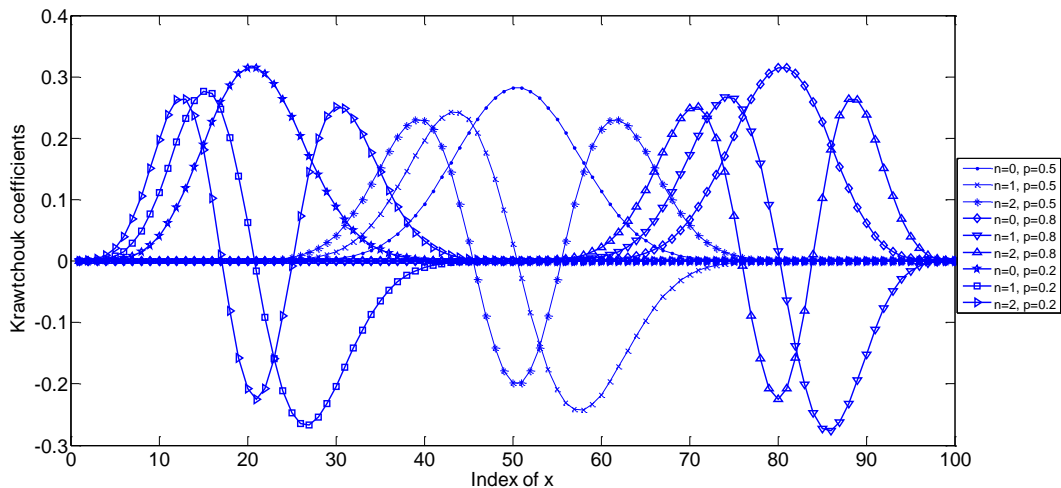
$$\tilde{r}_n(x) = r_n(x) \sqrt{\frac{\varrho(x)}{v_n^2}}, \quad \dots \text{Eq.(15)}$$

Therefore, the orthogonality condition defined based on Equations 14 and 15 becomes:

$$\sum_{x=s_1}^{s_2-1} \tilde{r}_n(x)\tilde{r}_m(x) = \delta_{nm}; \quad s_1 \leq n, m \leq s_2 - 1 \dots \text{Eq. (16)}$$

Note that,  $s_1 = 0, s_2 = N$  in the applications of 1 dimensional signal such as speech, where  $N$  is the typical number of samples in a speech frame. The polynomial coefficients calculation can be derived in both directions of  $x$  and  $n$ .  $k_n(x)$  is used as the normalized orthogonal for Krawtchouk polynomials.

**Krawtchouk polynomial:**



**Figure 2.10: Krawtchouk polynomials plots for different values of polynomial order n, with Krawtchouk coefficients for different values of order n and p . n =moment order, p=ROI constant.**

Krawtchouk polynomial can be represented by 2-D arrays with a controllable parameter,  $p$  used to emphasize a certain ROI on time frame of the signal. The value of  $p$  controls the moment's localization on the ROI. When  $p = 0.5$ , the ROI will be located in the middle of the signal frame. If  $p < 0.5$  the ROI is shifted to the left, and for  $p > 0.5$ , the ROI is shifted to the right. Plots for the Discrete Krawtchouk Transform (DKT) matrix for few values of  $n$  and  $p$  are shown in Fig. 2.10. This figure illustrates the effect of the parameter  $p$  on the position of the range of interest within the signal frame with different Krawtchouk coefficients.

The recurrence algorithms of the Krawtchouk polynomial,  $k_n(x;p,N)$  of the  $n$ -th order are given as follows

$$k_n(x; p, N) = \sqrt{\frac{\binom{N-1}{x} p^x (1-p)^{N-1-x}}{(-1)^n \left(\frac{1-p}{p}\right)^n \frac{n!}{(-N+1)_n}}} {}_2F_1\left(-n, -x; -N+1; \frac{1}{p}\right) \dots \dots \dots \text{Eq. (17)}$$

$$\gamma_1 k_n(x+1; p, N) = \gamma_2 k_n(x; p, N) + \gamma_3 k_n(x-1; p, N) \dots \dots \dots \text{Eq. (18)}$$

$$n = 0, 1, 2, \dots, M-1; \quad x = 1, 2, \dots, \frac{N}{2} - 2; \quad p \in (0, 1)$$

Where

$$\gamma_1 = \sqrt{p(N-x-1)(1-p)(x+1)}$$

$$\gamma_2 = -n + p(N-x-1) + x(1-p)$$

$$\gamma_3 = \sqrt{x(1-p)p(N-x)}.$$

The initial conditions are

$$k_0(0; p, N-1) = \sqrt{(1-p)^{N-1}}$$

$$k_n(0; p, N) = \sqrt{\frac{(N-n)p}{n(1-p)}} k_{n-1}(0; p, N), \quad n = 1, 2, \dots, M-1,$$

$$k_n(1; p, N) = \frac{-n+p(N-1)}{p(N-1)} \sqrt{\frac{(N-1)p}{(1-p)}} k_n(0; p, N) \quad \dots\dots\dots \text{Eq. (19)}$$

$$n = 1, \dots, M - 1.$$

The following symmetry condition can be applied for any value of  $p$  by terminating the recursion at  $x=N/2$  in Equation 19, to evaluate the polynomial values, where  $x$  is in the range  $[N/2, N-1]$ .

$$k_n(x; p, N) = (-1)^{n+x+1} k_{N-n-1}(N-x-1; p, N), \quad \dots\dots\dots \text{Eq. (20)}$$

**Orthogonal Transformation:**

The DKT for a neurogram  $f(x, y)$ , which is an array of  $(N \times M)$ , are defined as

$$\phi_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} t_n(x) t_m(y) f(x, y), \quad \dots\dots\dots \text{Eq. (21)}$$

$$\psi_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} k_n(x) k_m(y) f(x, y)$$

$$\text{where, } n = 0,1,2, \dots, N - 1, \quad m = 0,1,2, \dots, M - 1$$

$$x = 0,1, \dots, N - 1, \quad y = 0,1, \dots, M - 1,$$

Plot of the DKT matrix is also shown in Fig. 13 for different value of polynomial order.

After the neurogram has been applied with the polynomial, it could be reconstructed back by using inverse transformation of:

$$\tilde{f}(x, y) = \sum_{n=0}^{M-1} \sum_{m=0}^{N-1} t_n(x) t_m(y) \phi_{nm}, \quad \dots\dots\dots \text{Eq. (22)}$$

$$\tilde{f}(x, y) = \sum_{n=0}^{M-1} \sum_{m=0}^{N-1} \psi_m \psi_n k_n(x) k_m(y),$$

$$x = 0,1,2, \dots, N - 1; \quad y = 0,1,2, \dots, \dots, \dots, M - 1.$$

## **Chapter 3. METHODOLOGY**

### ***3.1. System design***

#### **3.1.1. General**

The purpose of a speaker verification system is to make sure that the claimed identity from a speaker does belong to the claimed speaker model. Therefore, all 3 speech samples from that speaker will be tested against the GMM model of the claimed speaker. If a genuine speaker's probability value resulted from the PDF is higher than the threshold value, the speaker is authenticated, however if an imposter speaker's tested probability higher than the threshold value, then an imposter has been authenticated.

The project is sub divided into two parts, with the first part is to train the system using the AN model with GMM classifier so that the speaker models can be saved in the database; while the second part involves in testing the system itself by using two types of speeches (the original person and imposters) to try to verify which one is the true speaker. The project will be assembled using Matlab ® (Matlab 2013a, The Mathworks Inc.).

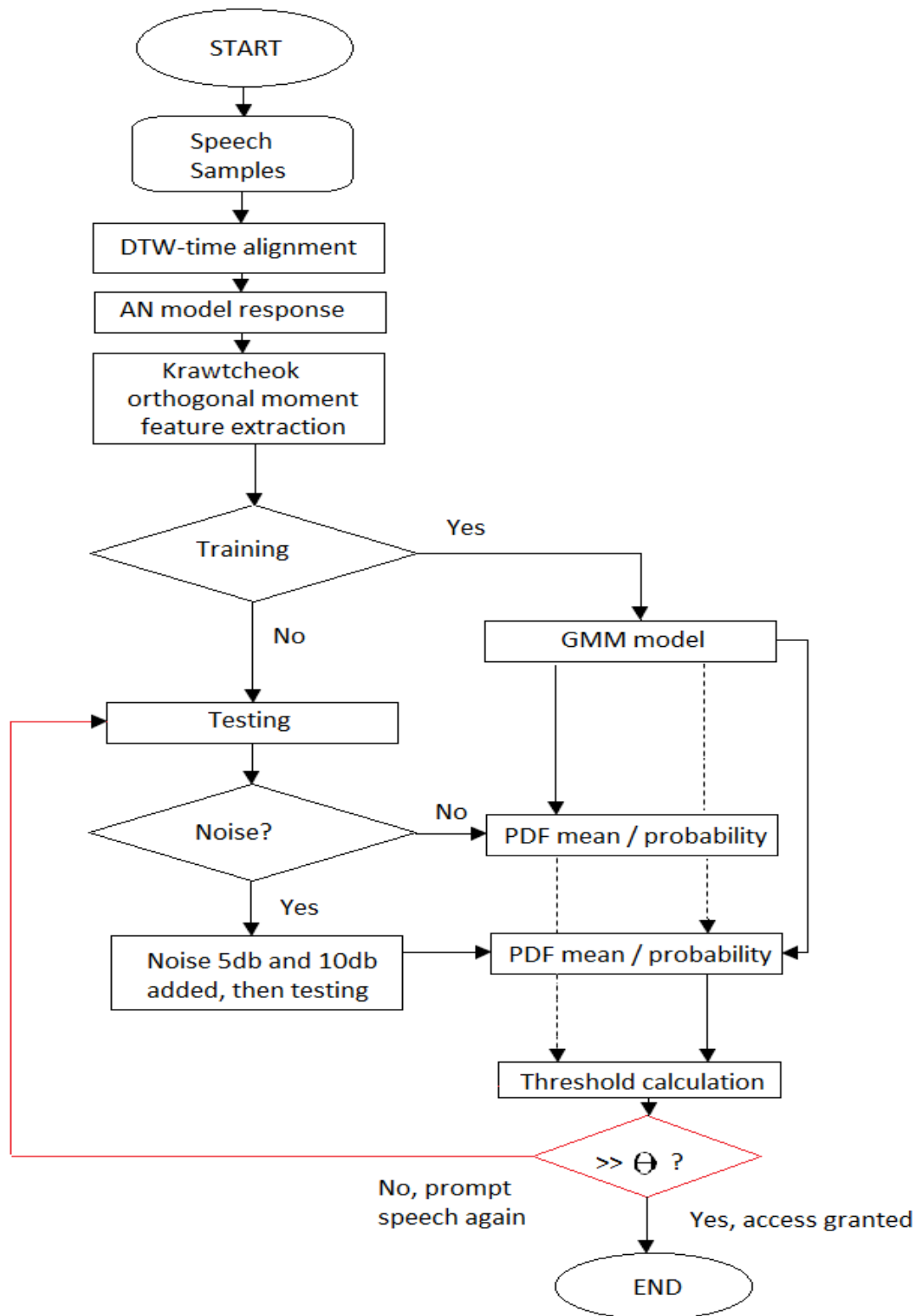


Figure 3.1: Flow-chart of the speaker verification system.

Figure 3.1 shows the flowchart of the overall process of training and testing the system. The program starts by analysing the speech by deleting the silences (not shown in flowchart) before time-alignment can be made. All speech samples are then run with the AN model to produce the synapse output, ENV and TFS features. ENV is chosen to both train and test the system. For training, 70% of the speech samples are used for each speaker to find their respective GMM models. The remaining 30% is then used as testing speech samples with PDF values are then calculated according to GMM models. The resulting values are used to calculate the threshold (discussed in Section 3.8) which is the crucial value in determining the system's performance. If the calculated PDF for testing is higher than the threshold, the speaker is verified. Otherwise, the system will prompt speech to be input again or considered to be rejected.

## **3.2. Verification System**

### **3.2.1. Phase 1: Speech sampling and pre-processing**

For the purpose of speaker verification, a database of text-dependent speech samples is used where it consists of 39 different speakers (25 males, 14 females) among students aged 22 to 24 years old were recorded using a microphone with 16-bit quantization rate and 8kHz sampling rate in a quiet room. Each speakers are asked to say 'University Malaya' 10 times in different recording sessions, and the speech samples are recorded and saved in the database. There was no added artificial noise into the speech samples nor was it recorded in a recording booth for this project. The performance of AN model used in this project is expected to be noise robust and works nonlinearly as discussed in Chapter 2.

The pre-processing of the speech samples in this project consists of three steps; first, deleting the silences (start and end of samples) for all 10 speech samples of the speaker by using Adobe Audition ® (Adobe Systems Inc., United States) software; second, applying the *specgram* function in Matlab and finally time/temporal alignment for each samples of the speakers.

After the silences have been deleted, short time Fourier transform (STFT) features for all speeches that will be used in DTW are applied with Hamming window with 25% overlapped between segments. The FFT length is set to 512 with the sampling frequency is set to 384 (25% overlaps 512 samples Hamming)

### **Time alignment – Dynamic Time Warping (DTW)**

The speech samples taken from the speakers is in the form of word utterance, therefore the possibility of the sample length to be different is higher compared to single pronunciation samples (e.g. phonemes, vowels, etc). Previous study (Pandit & Kittler, 1998) uses DTW as a classification technique that is used after feature selection process in order to combine all possible intra-speaker variances in a single speaker to better optimize the performance of the system. The optimal path distance between the reference speeches is calculated with imposter's optimal path to determine which sample is closest to the reference speech. However, in this project, the DTW methodology is applied simply as a pre-processing step for the speech samples using a reference sample before getting the response from the AN model.

DTW is applicable to align between two speech samples, whereby a reference speech sample is chosen out of the 7 training samples so that the remaining 6 is aligned to the reference speech sample. The length for all 7 speech samples were analysed and the

mean of the length is calculated where the time function and feature parameters between the training sample and reference sample is averaged and registered. The speech sample having a length nearer to the mean value is selected as a reference speech sample for DTW aligning.

The cosine distance between the magnitude of the STFT for both the reference and tested samples are calculated with 25% Hamming window overlap in previous step to get the local match. Once STFT is applied, the time-domain speech samples are converted to its time-frequency phase to allow modification of the magnitude in further steps. Then dynamic programming algorithm is applied to the framed speech samples to find the shortest path between the speech frames of the two speech samples for optimization process. The optimal path (shortest) between the speech sample and the reference is determined, and the words in speech samples are aligned so that it is warped at the exact timing of the reference sample. This is done by calculating the frames in the tested samples to match each of frames in the reference sample. To resynthesize a warped version of the speech sample, the number of frames in the speech sample that matches the referenced sample is calculated using zeros function to get the warped version of the speech samples. Finally, for the phase-transformed and warped speech sample needed to be transformed back to its time domain for it to be usable in the next step using inverse STFT.

DTW is then applied on the remaining 6 samples and the warped version for all speech samples in .wav format are saved in the Matlab environment. The length of all speech samples belong to a particular speaker should be the same, but not necessarily for different speakers. The speech samples for all speakers are divided into 70%-30% where 7 speech samples are used to train the classifier and the remaining 3 are used for



testing. The methodology involving the use of DTW is adopted from (Turetsky & Ellis, 2003).

### 3.2.2. Phase II: Simulation of AN model responses

An established AN model is used for the purpose of this project, which is a widely known and well-developed AN model (Zilany et al., 2009) will be employed to simulate the neural responses on verifying a speaker. The use of AN model subsequently avoid the use of common technique in acoustical speech signal processing, that is usually viewed from the speech production process of different speaker itself. For example, the use of common technique in speech processing that usually involves in analysing the speech parameters of a particular speaker through techniques such as LPC and getting the feature selection via cepstrum coefficient and its derivative orders are not used in this study.

The output of the DTW step in phase I that is saved before is used as the input of the AN model. All the initial parameters of the AN model is set as, with the sound pressure level equals to 74dB which is the level specified by the microphone's manufacturer. Meanwhile, the sampling frequency is upscaled to 100 000 Hz for the AN model. The average normal human hearing range is set at 250 to 8kHz to mimic the normal human speech fundamental frequency (Aalto et al., 2013). Along this range, the hearing frequency range is logarithmically spaced in 32 characteristic frequencies (CF)s. There are also other parameters of the AN model that could be controlled, for example setting the level of impairment for both inner and outer hair cell, the type of intended species to be modelled (human or cat), and adding additional noise to the input audio. But as this is not required for the current project, these parameters were set as default.

According to Liberman (1978), the distribution AN fibers shows approximately 61% of high spontaneous rate (SR), 23% medium SR fibers, and 16% of low SR fibers. To take into account of this physiological observation, the PSTH is computed as

$$PSTH = 0.6 * (High\ synchronization) + 0.2 * (Medium\ synchronization) + 0.2 * (Low\ Synchronization)..... Eq. (23)$$

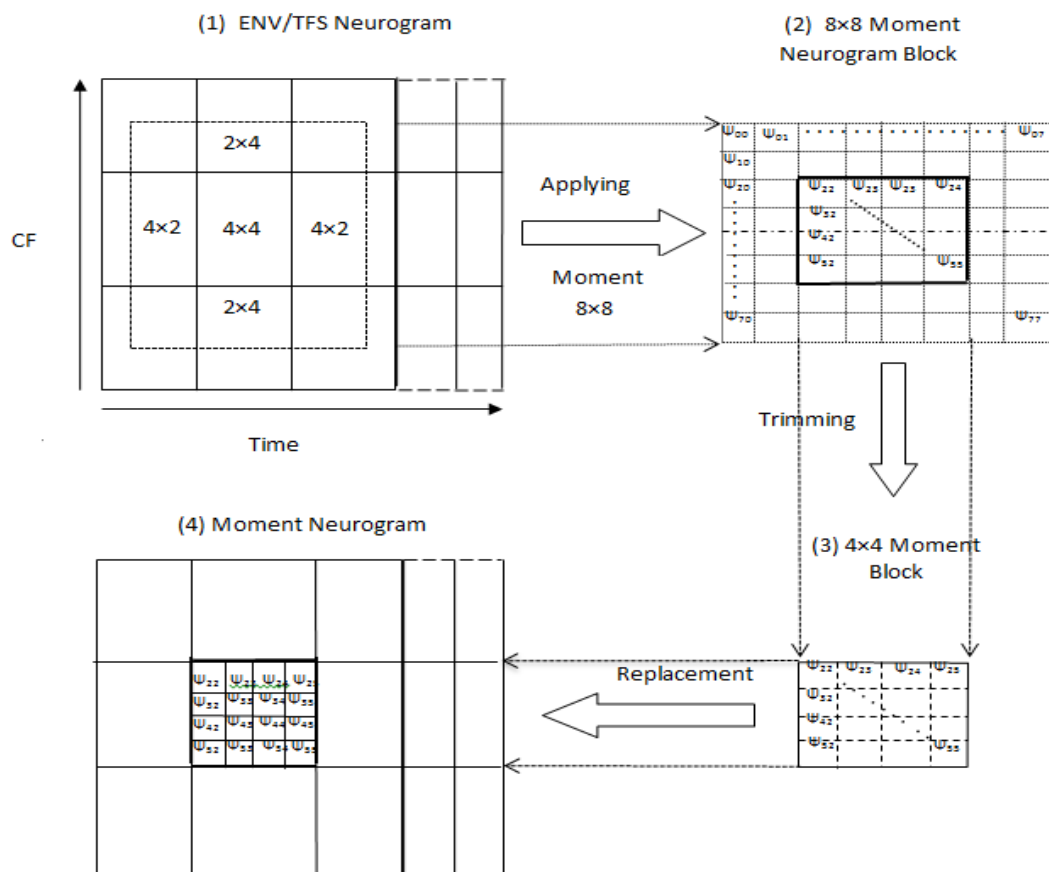
where the resulting PSTH output is based on the weighted matrix that contains 60% of high SR fibers and 20% for both medium and low SR fibers. There are three types of outputs that were processed for AN model responses; the first is the synapse output which develops from the synapse output based on the PSTH, ENV and TFS. All responses for each speaker are saved in the database for further analysis.

### 3.2.3. Phase III: Krawtchouk Orthogonal moment feature extraction

As speech is considered to be a one dimensional signal, the orthogonal moment technique is applied to further process the neurogram. The output for the AN model response is changed from the time-domain speech to moment-domain by using orthogonal transformation of the Krawtchouk polynomials of the neurograms. However applying the algorithm on the overall ENV neurogram itself does not clearly represents the computed orthogonal moment function of the ENV neurogram. Therefore, it is necessary to re align the overall ENV neurogram into several blocks of 4x4 and overlapped with 50% with each other. The overlapped blocks portions resulting in the additional of two extra rows on top and bottom and also on both sides of the 4x4 blocks subsequently creates a new 8x8 blocks representing the original ENV neurogram where the moment features are computed in each blocks. In the end, the additional border size of the resulted moment feature of the ENV is removed to match the original neurogram size. Krawtchouk polynomial constant  $p = 0.9$  is used to indicate the ROI for calculating

the moment. The resulted features are saved for all speech samples used for training and testing purposes.

In this project, feature extraction method is also applied after getting the AN response of the original speech .wav files in this project. Common initial method in speech feature extraction method that is generally used such as pre-emphasizing, framing, and windowing is applied during time alignment step (Section 3.2.1) and the feature extraction step is applied after getting the AN response of the speech.

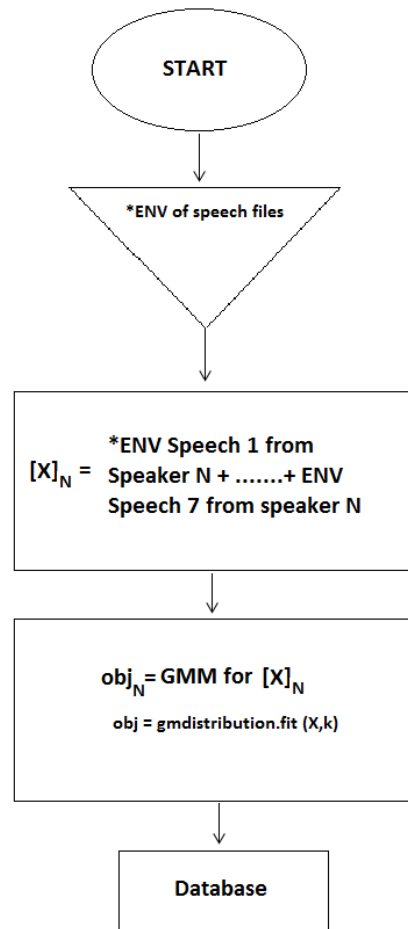


**Figure 3.2: ENV neurogram blocks transformation to ENV moment neurogram (reproduced from Mamun, personal communication).**

Figure 3.2 loosely illustrated on how the original ENV neurogram is divided into 4x4 blocks and the overlapped 50% in Fig. 3.2(1) and when the moment is applied, each blocks are realigned to 8x8 moment neurogram block Fig. 3.2 (2). The moment blocks containing necessary information is reshaped or trimmed into the original size in Fig. 3.2 (3) and finally the resulted moment neurogram Fig. 3.2 (4) is saved as the extracted features for the particular speech sample. In this project, window size 8x8 and constant  $p=0.9$  is used for the applied moment.

#### 3.2.4. Phase IV: Training using classification technique – Gaussian Mixture Model (GMM)

Each transformed speech samples of the AN model output in section 3.2.3 were saved in .mat files in the Matlab environment, where they were called in the main programming of the project. Fig. 3.3 below shows the flowchart for the training process of the algorithm:



**Figure 3.3: Flowchart of the GMM training process. \*ENV is the response of AN model.**

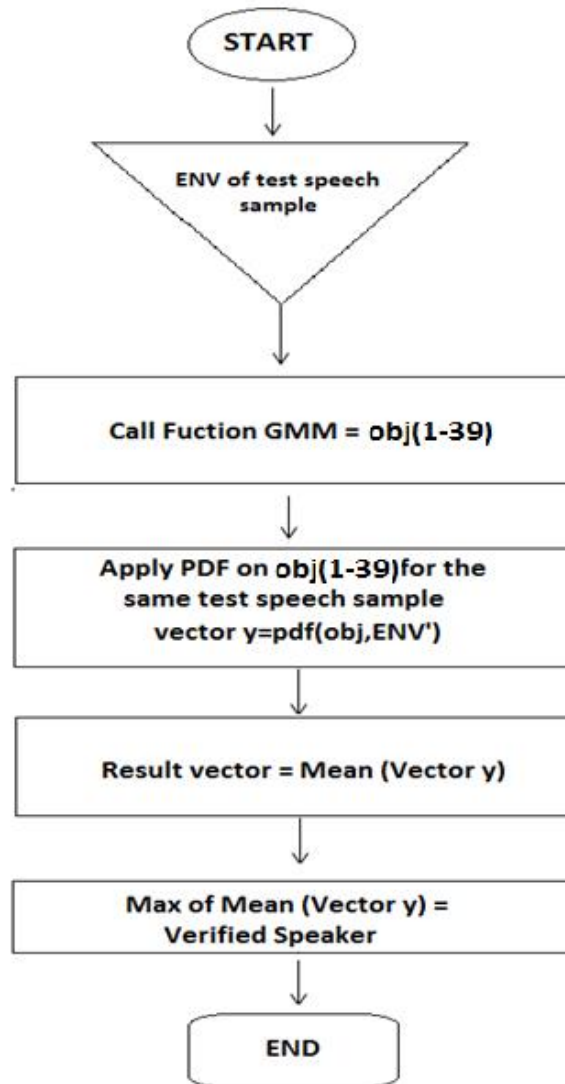
**N=number of speakers.**

For the first step, the algorithm of the training phase starts by using the output of ENV of the AN model from the original speech files. The ENV of one speech sample should be in the form of matrix of  $(d \times n)$  with  $d$ , dimension = 32 CFs  $\times$   $n_1$ , which is the number of data in the first speech sample. In order for the matrix data can be used in the GMM algorithm, the ENV matrix should be converted from  $(d \times n)$  to  $(n \times d)$  using inverse matrix in the program. For training purpose, 7 inverse matrices ENVs of the speech samples are randomly selected and concatenated with each other to form a single matrix of  $(n \times d)$ ,  $[X_N]$ ; with a fix number of  $d=32$  and  $n=n_1 + n_2 + n_3 \dots + n_7$ .  $[X_N]$  is then used in the GMM distribution function in the Matlab environment using  $obj =$

gmdistribution.fit ( $X_N$ ,  $K$ ), with the value of the distribution of components ( $K$ ) = 16, 32, 64, 128 and 256. This process is repeated with a loop for all 39 speakers and the output of Gaussian mixture distribution of the particular speakers (obj1 – obj39) is saved in the database to be used in the testing stage of the speaker verification.

### 3.2.5. Phase V: Testing using probability density function (PDF)

In the previous training phase, the GMM models for all speakers were saved as functions so that they can easily called by the system to run the test. 3 remaining speech samples of each speaker are used to test the reliability and accuracy of the overall design itself. To test the accuracy of the design, the project is now treated as speaker *identification* (instead of *verification*), where the speech samples are tested for all 39 speakers instead of only one in the case of verification. Fig. 3.4 shows the flowchart of the testing phase:



**Figure 3.4: Flowchart of the PDF testing process of one speech sample**

The algorithm of the training process starts with calling the ENV of the testing speech sample of a speaker into the program. The function of obj from the training phase is called for all 39 speakers into the environment and the testing sample is applied with PDF of the Gaussian mixture distribution to get a vector with same dimension and length of the  $[X_N]$  in the training stage. However, 39 vector outputs of the PDF function is produced for each speaker for a single testing speech sample. Therefore, the mean value of the vector is taken as the result. After the program calculates the mean vectors

for all speakers, the maximum vector is chosen as a reference point and verified as the speaker. Any other vector values lower than the reference point is considered as the imposter for that model. For the purpose of discussion of the project, all vector values for 39 speakers are saved and ranked to calculate the accuracy of the overall system for speech verification and identification.

### **3.3. Robustness of the system**

One of the problems faced in speaker recognition once it has been implemented in real life application is the presence of additional environmental noise during the system-prompted to get the speech of the user. In order to prove the use of AN model response robustness through mismatched acoustic condition, the 3 speech samples used for testing are introduced with additional noise simulated with white Gaussian noise with the function *awgn* in the Matlab environment for all speakers. For this test, the original GMM model of the speakers in section 3.2.4 is used. The added noise is varied into two values (5 dB and 10 dB) and is added separately. Finally the accuracy result of the system is calculated and compared with the baseline system result.

### **3.4. System performance**

For this project, both performance of speaker verification and speaker identification is done although the methodology is mainly based on the verification performance.



### 3.4.1. Speaker Verification

The identity of the speaker is assumed to be known and belongs to the GMM model. If the probability value of the GMM is larger than the threshold value, then the speaker is authenticated. The determination of threshold value is crucial in this system. The performance of the system is tested on how much an imposter gained access to the system and how much genuine speaker was rejected from the system.

The performance calculated from the total success rate (TSR) based on the probability of incorrect acceptance; false acceptance rate (FAR) and the probability of correct acceptance; false rejection rate (FRR) by analysing the result using the speech samples as an imposter or the genuine speaker. The performance of the speaker verification system is tested based on the study done by Ilyas et al., (2007). The EER, or the initial threshold ( $\theta$ ) of the speaker model is calculated as:

$$EER, \theta = \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_1 + \sigma_2} \dots\dots\dots \text{Eq. (24)}$$

Where  $\mu_1$  and  $\sigma_1$  are the mean and standard deviation of the distribution of probabilities resulted from the true speech tested against the GMM model of the intended speaker, while  $\mu_2$  and  $\sigma_2$  is based on the distribution of probability tested by speeches of the imposters against the same intended speaker. The EER for all 39 speakers were calculated to get the threshold value for the system to accept or reject any attempted genuine or impostor speaker in the system. Meanwhile, for calculating the speaker verification system performance TSR, FAR and FRR are defined as:

$$FAR = \frac{\text{Number of accepted imposter claims}}{\text{Total number of imposter accesses}} \times 100 \dots\dots\dots \text{Eq. (25)}$$

$$FRR = \frac{\text{Number of rejected genuine claims}}{\text{Total number of genuine accesses}} \times 100 \dots\dots\dots \text{Eq. (26)}$$

The overall performance of the system can be calculated by combining the FAR and FRR to gain the TSR as in the equation:

$$TSR = 100\% - \left( \frac{FAR + FRR}{200} \right) \times 100 \dots\dots\dots \text{Eq. (27)}$$

Higher value of TSR indicates higher accuracy of the system performance.

### 3.4.2. Speaker Identification

An unknown speech identity is tested against all possible GMM models available in the system. If the probability value is the maximum value for a particular GMM model, then the unknown speech is identified to belong as the speaker for that GMM model. Otherwise, mismatch happens and the speech is falsely identified as other person. The accuracy of the GMM for speaker identification is calculated for all 39 speakers. Accuracy based on rankings is done just to visualize the effect of choosing the number of Gaussian components K that will be discussed in Chapter 4.

## 3.5. Statistical analysis

To make sure the speech data for the speakers are correlated with each other, Spearman's correlation coefficient statistical analysis is applied to determine the significance difference between the PDF of three speech samples used in the training stage. Two-tailed non-parametric test with 95% confidence of intervals is applied during the test.

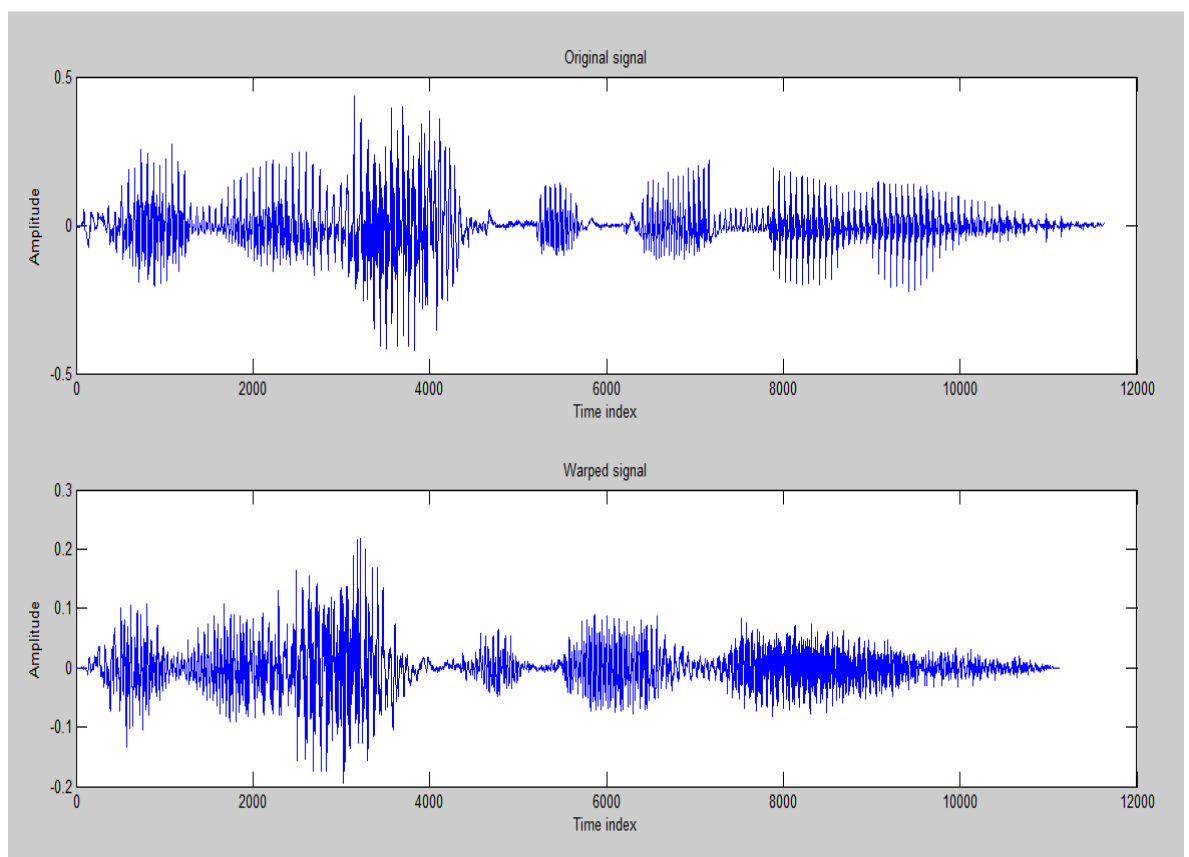
Standard deviation within a speaker's resulted PDFs (probabilities) is also tested to make sure how spread is the resulted values from the mean value of the group to determine its variability. High variability (low standard deviation, less spread values off the cluster mean) is expected for the system. Furthermore, a statistical test to determine the system's internal reliability is also made by repeating the procedure 10 times using randomly selected user. The reliability test is based on the value of Cronbach's alpha coefficient. Higher coefficient value indicates higher reliability of the system result which is also expected for this test. Pearson's correlation coefficient 'test-retest' statistical analysis is also calculated when the same verification system is run with the speech samples for 10 times. Higher coefficient value also shows high correlation between the data and low variability exists between retesting.

All statistical analyses are done by using Statistical Packaging for the Social Science (SPSS) (PASW Statistics for Windows, Version 18.0. Chicago: SPSS Inc.).

## Chapter 4. Result & Discussion

### 4.1. Pre-processing using DTW

As discussed in Chapter 3, the raw data of speech signals need to be pre-processed before the data can be used for the system. DTW is applied on the original speech signal to get the warped version of the speech sample to align it at the same timing and temporally for all speeches of the same speaker.



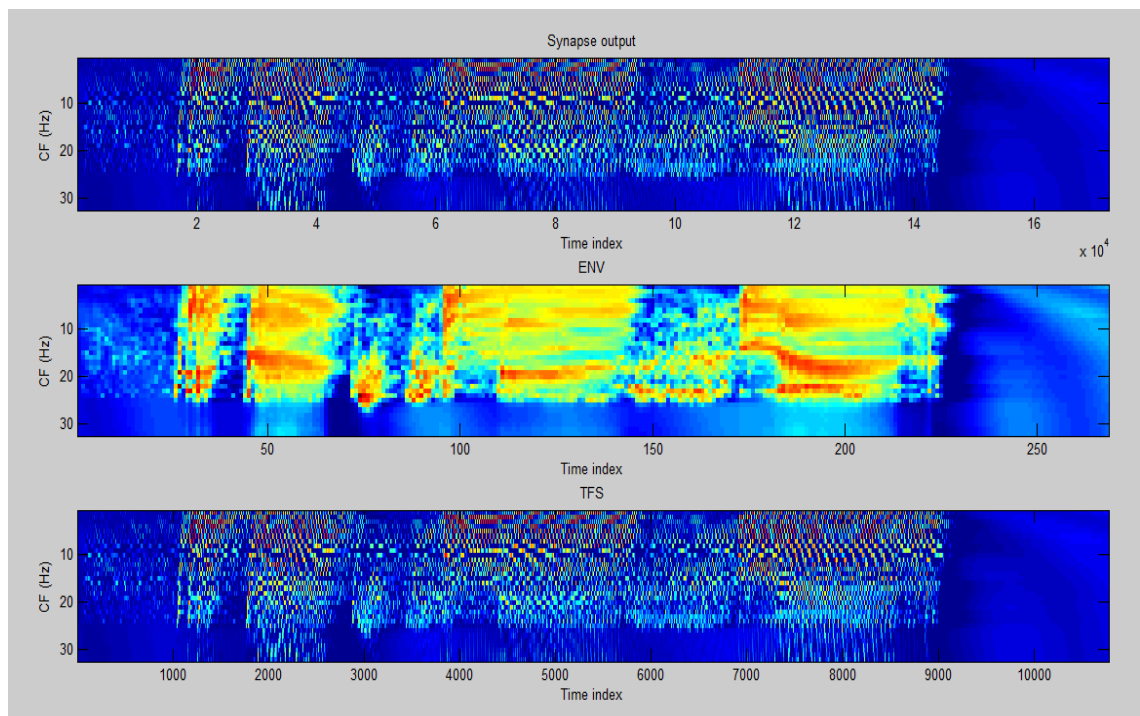
**Figure 4.1:** (Top) Original speech sample of a speaker. (Bottom) Warped version of the same speech sample of the words ‘University Malaya’.

Figure 4.1 shows both plotted original speech sample (top) and its warped version (bottom). Once DTW is applied on the speech sample based on a reference speech sample, both amplitude and the timing of the speech is aligned for the preparation of the

training process. The warped signal clearly shows lowering of the original maximum amplitude from 0.45 to 0.20. The same case is resulted on the remaining 7 speech samples for training of the same speaker resulting with all speeches have the same length, which is very crucial if the classification technique involves the use of neural network.

## 4.2. AN model response

The output of the AN model response is simulated and shown in Figure 4.2:



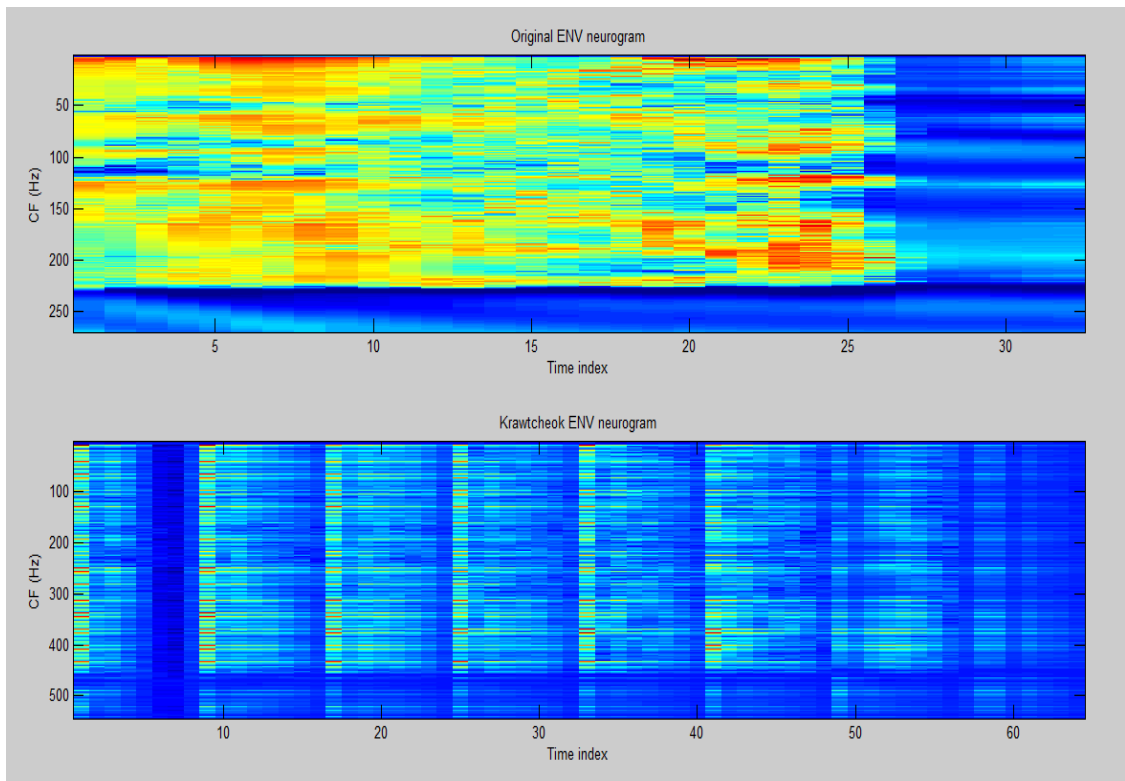
**Figure 4.2: Output neurograms of the AN model a) Synapse Output b) Envelope (ENV), c) Temporal fine structure (TFS).**

The graph shows the results of synapse output in Fig. 4.2 (top), ENV Fig. 4.2 (middle) and TFS Fig. 4.2 (lower), are plotted as neurograms showing different colours depicting the distribution of the frequency of the speech sample. Synapse output and TFS are visually the same except that the time index is subsequently compressed to a lesser

value in TFS compared to its original speech features as in the synapse output. Meanwhile, the ENV shows a lesser value of frequency distributed as shown with its prominence features of lighter colours and having lesser values compared to both synapse output and TFS. The ENV is chosen in the speaker verification process as its nature of lower number of observations creating lesser mathematical computation compared to the other two which will take longer time. However, the TFS is used in this project for comparison only to make sure that choosing ENV with lower data does not affect the overall result.

### ***4.3. Krawtchouk Polynomial Feature Extraction***

In feature extraction method, the original ENV neurogram is applied with Krawtchouk polynomials moment and it is divided into 8x8 block as can be seen in the patterns in the bottom panel of Fig. 4.3.

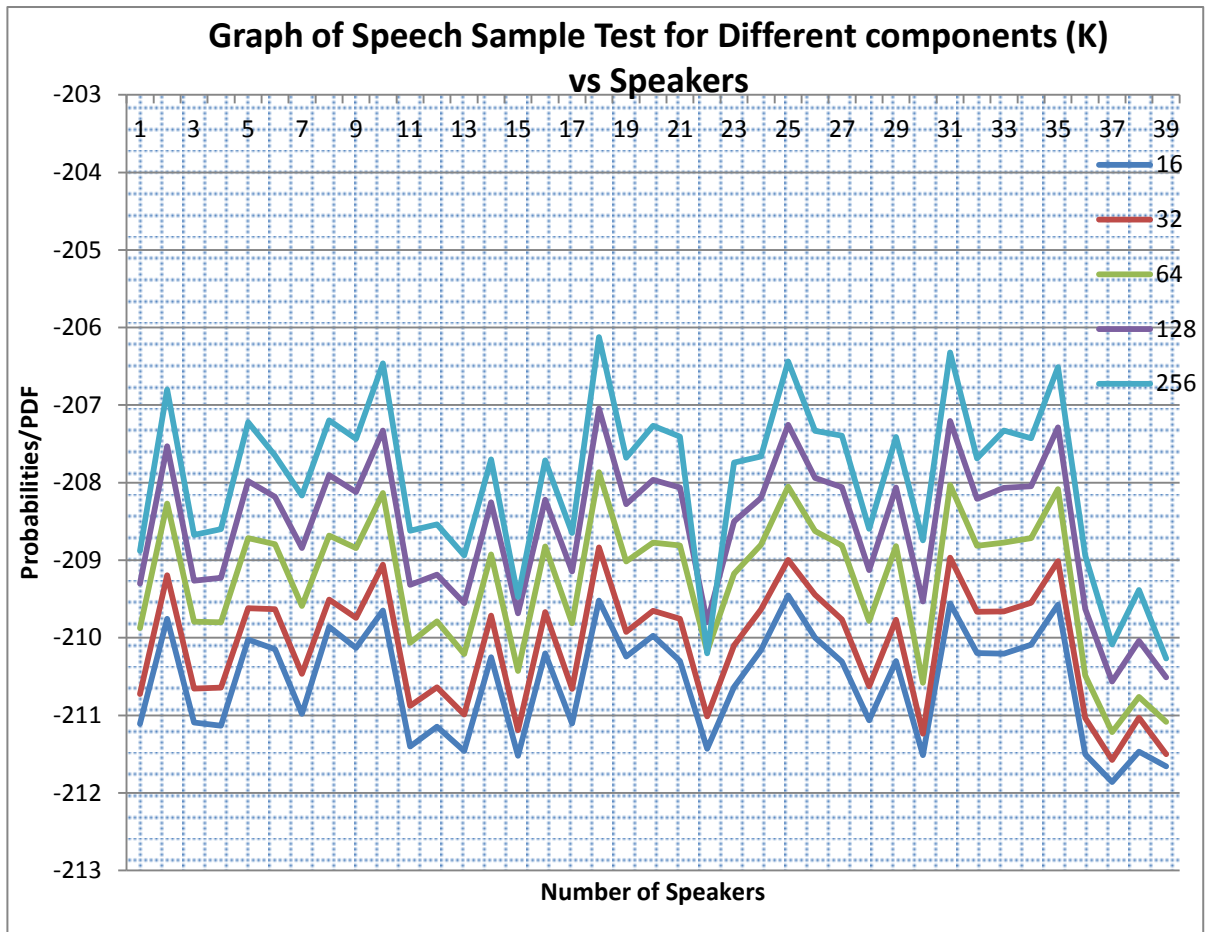


**Figure 4.3: The result of original ENV Neurogram (top) compared to result of Krawtchouk Neurogram after transition from time-domain to moment-domain (bottom)**

The use of ROI region at  $p=0.9$  and moment order  $N=8$  that shifted the frequency distribution in the original ENV to the right of the  $8 \times 8$  window frame (blue, darker colour) where most of the important feature is contained. These feature vectors are then used for both training and testing processes.

#### **4.4. Training using GMM classification**

The graph in Fig. 4.4 shows a speech sample from Speaker #18 is taken and then compared with all 39 speakers for 5 different components values (16, 32, 64, 128 and 256) that were used during the training stage.



**Figure 4.4: Graph of a speech sample test for different components (K) vs Speakers samples against GMM Speaker#18.**

The result clearly shows by increasing the number of K, the accuracy of the system increased. For example, taking the result of K=16, the system identifies that the speech belongs to Speaker #25 (-209.455 > -209.521) having larger value compared to original Speaker #18. However when the K number is increased to 256, the accuracy of the system was improved by verifying that the speech really belongs to Speaker #18. The same test was also compared to all testing speeches and the overall accuracy of the system increases to 97.7% in Table 4.2. Increasing the number of Gaussian components will cause more data to be fit in available components, therefore causing it to be more precisely trained. When a test data is compared to the trained model, it will results in a higher probability matching rate thus increased the performance's accuracy.



## 4.5. System Performance

### 4.5.1. Speaker verification

The system performance of the speaker verification is calculated based on the value of TSR as defined in Eq. 27. The EER value is used as the threshold in which it is the same for both FAR and FRR acceptance rate. The original threshold (EER) calculated was too high and only applicable if the testing speech is in clean condition. Therefore, a low level round-off error  $\pm 0.004\%$  is allowed for the threshold value to compensate the nature of low values of PDF produced. The new threshold value is used with 5 dB and 10 dB noise speech samples and the system's performance is recorded. Table 4.1 shows the TSR value for speaker verification comparison for all types of test.

**Table 4.1: Speaker verification system performances accuracy.**

Feature	Clean Speech (%)	Noisy Speech (SNR = 10 dB) (%)	Noisy Speech (SNR = 5 dB) (%)
No feature extraction	99.7	98.8	98.3
With feature extraction	N/A	N/A	N/A

The result shows that the system performance is already high even when no feature extraction method is applied at 99.7% using  $\pm 0.004\%$  threshold error. For the case of added white Gaussian noise, the result shows that it only lowers the TSR value by less than 1.5% (98.8% for 10 dB; and 98.3% for 5 dB).

For speaker verification, the use of AN response (without feature extraction) indicates the robustness of the system. This might be caused by the fact that the AN model

response could handle resolution and frequency distribution on both linear and non-linear scales during ‘hearing’ or sound input of the system, which is not applicable for acoustic analysis using Fourier transform alone (Li, 2003; Ilyas et al., 2007).

Previous study has also concluded that using auditory-based feature in speech recognition field compared to Fourier transform method has better accuracy performances by concluding that the results coming from the FFT transformation contains higher noise level computationally and more distorted compared to auditory-based transform (Li, 2009; Li & Huang, 2010). Changing the noise level to 5 dB does also lower the system’s performance but only to a little extent.

Feature extraction method was not applied in this step and considered to be redundant as the original result using ENV only already shows high TSR value, and adding the extraction method will just increase the computational load of the system.

#### 4.5.2. Speaker identification

Table 4.2 shows the accuracy of the system based on rankings with increasing number of components for GMM distribution for speaker identification.

**Table 4.2: Table of the accuracy of the system compared to different K components of the GMM.**

Gaussian Distribution Components (K)	Accuracy based on rankings (%)
16	92.3
32	95.0
64	95.7
128	95.7
256	97.7

Increasing the accuracy with increasing the number of K component shows that the more data can be classified when they are fit into the additional number of Gaussians K provided in the algorithm. However, the downside of this is the increasing computational time for log-likelihood iterations, which is needed to achieve the ideal value for convergence for the GMM. This result is also supported by a study by de Lima, et al. (2001) that shows higher accuracy result for number of Gaussian 32 compared to 8 for text-independent speaker verification using GMM (de Lima et al., 2001).

Table 4.3 shows the result of the system performance for speaker identification system using the AN model.

**Table 4.3: Speaker identification system performances.**

Accuracy, with K=128	Clean Speech (%)	Noisy Speech (SNR = 10 dB) (%)	Noisy Speech (SNR = 5 dB) (%)
No feature extraction	45.7	42.0	32.4
With feature extraction	92.4	63.8	54.4

In this step, the test speech is tested against all available 39 GMM models, which generally takes longer time than that of verification. Initially, the system was only able to identify correctly 45.7% of the speech samples. To improve this identification score, feature extraction method using Krawtchouk orthogonal moment is applied on both the training data and the testing data for  $p = 0.9$ . The result of the proposed method shows the increase of accuracy level to 92.4%. The result is also comparable to a study by Li, Q (2003) that uses GMM classification using AN response vs feature extraction based

on MFCC results only an accuracy of 41.2%. In general, when noise is added to the signal, the accuracy of the proposed system decreases, however, the accuracy is improved from 42% to 63.8% for 10 dB noise level and 32.4% to 54.4% for 5 dB when Krawtchouk polynomial feature extraction method is applied.

Some of the reason that causes Krawtchouk moments as features for both training and testing is that when low-order moment applied, it caused smoothing effect on the line output of the moments that subsequently cancelled off the noise in the ENV neurogram. This line output is explained in detail in a study by Rani & Devaraj (2012). The smoothing effect will yield a global characteristics of the neurogram at ROI  $p=0.9$ . Orthogonal moments is when the neurogram input is aligned orthogonally with the applied Krawtchouk polynomials. The orthogonality of the moment also contributes by making the neurogram output to be less correlate as opposed to non-orthogonal moment (Rani & Devaraj, 2012). Varying ROI value could change the output of the overall system's performance.

## **4.6. Statistical analysis**

### **4.6.1. Correlation between tested data**

Since there are 3 speech samples used in the testing stage of the system, the correlation between the three results obtained are tested using Spearman's correlation coefficient test using SPSS® software. Table 4.4 shows the result of Spearman's correlation coefficient for one speaker chosen randomly out of 10 speakers to show the correlation result between tested speech data.

**Table 4.4: The Spearman's Correlation coefficient tested among the three speech samples**

			Correlations		
			Speech_Te st_1	Speech_Te st_2	Speech_Test_3
Spearman's rho	Speech_Test_1	Correlation Coefficient	1.000	.988**	.894**
		Sig. (2-tailed)	.	.000	.000
		N	10	10	10
	Speech_Test_2	Correlation Coefficient	.988**	1.000	.936**
		Sig. (2-tailed)	.000	.	.000
		N	10	10	10
	Speech_Test_3	Correlation Coefficient	.894**	.936**	1.000
		Sig. (2-tailed)	.000	.000	.
		N	10	10	10

The Spearman's Correlation coefficient tested among the three speech samples shows high correlation between the three variables shown by 0.988 and 0.894 for first speech, 0.988 and 0.936 for the second speech and 0.894 and 0.936 for the third speech; which is significant beyond the 0.01 confidence interval level.

A series of Spearman's correlation coefficient are also tested on the PDF of a speech samples to relate is there any significance different if the increasing value of components  $K=16,32,64,128$  &  $256$  during the GMM distribution. The result of the test is tabulated in Table 4.5:

**Table 4.5: Correlation coefficient for GMM using different K values**

			<b>Correlations</b>				
			PDF_for_ K16	PDF_for_ _K32	PDF_for_ K64	PDF_for_ _K128	PDF_for_ _K256
Spearman 's rho	PDF_for_K16	Correlation	1.000	.939**	.830**	.903**	.939**
		Coefficient					
		Sig. (2-tailed)	.	.000	.003	.000	.000
		N	10	10	10	10	10
	PDF_for_K32	Correlation	.939**	1.000	.903**	.939**	.903**
		Coefficient					
		Sig. (2-tailed)	.000	.	.000	.000	.000
		N	10	10	10	10	10
	PDF_for_K64	Correlation	.830**	.903**	1.000	.952**	.867**
		Coefficient					
		Sig. (2-tailed)	.003	.000	.	.000	.001
		N	10	10	10	10	10
	PDF_for_K128	Correlation	.903**	.939**	.952**	1.000	.964**
		Coefficient					
		Sig. (2-tailed)	.000	.000	.000	.	.000
	N	10	10	10	10	10	
PDF_for_K256	Correlation	.939**	.903**	.867**	.964**	1.000	
	Coefficient						
	Sig. (2-tailed)	.000	.000	.001	.000	.	
	N	10	10	10	10	10	

\*\* . Correlation is significant at the 0.01 level (2-tailed).

The result shows significant positive relationship between all difference K components with  $r_s(10) = \{0.939, 0.830, 0.903, 0.939, 0.939, 0.903, 0.939, 0.903, 0.830, 0.903, 0.952, 0.867, 0.903, 0.939, 0.952, 0.964, 0.939, 0.903, 0.867, 0.964\}$ ,  $p < 0.01$ . The result does not indicate whether or not increasing the number of Gaussian components K shows a better performance of the system using this test.

#### 4.6.2. Reliability Test

To check whether the running system is consistent, an internal reliability test is done using Cronbach's alpha statistical test. A 'test-retest' reliability test is done in a single GMM model speaker. Table 4.6 shows the result of Cronbach's alpha value to test the reliability between 10 repeated measurements using randomly selected speech sample from the same speaker. The correlation coefficient is 0.924, which suggests a very high internal consistency of the system.

**Table 4.6: Statistical data on internal reliability for 10 repeats of verification system.**

Reliability Statistics	
Cronbach's Alpha	N of Items
.924	10

Meanwhile, Table 4.7 shows the result of 'test-retest' reliability when the same measurement is repeated 10 times using the same speaker verification GMM model and speech test to test its consistency.

**Table 4.7: Statistical on test-retest reliability for 10 repeats of verification system.**

		Correlations									
		Retest 1	Retest 2	Retest 3	Retest 5	Retest 4	Retest 6	Retest 7	Retest 8	Retest 9	Retest 10
Retest 1	Pearson Correlation	1	.989**	.997**	.986**	.989**	.992**	.988**	.996**	.998**	.988**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	39	39	39	39	39	39	39	39	39	39
Retest 2	Pearson Correlation	.989**	1	.990**	.998**	.998**	.998**	.998**	.994**	.994**	.998**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000	.000	.000
	N	39	39	39	39	39	39	39	39	39	39

**Table 4.7, Continued**

3	Retest Pearson	.997**	.990**	1	.985**	.987**	.991**	.988**	.997**	.998**	.987**
	Correlation										
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000	.000	.000
	N	39	39	39	39	39	39	39	39	39	39
5	Retest Pearson	.986**	.998**	.985**	1	.999**	.996**	.998**	.989**	.990**	.999**
	Correlation										
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000	.000	.000
	N	39	39	39	39	39	39	39	39	39	39
4	Retest Pearson	.989**	.998**	.987**	.999**	1	.996**	.997**	.990**	.992**	.999**
	Correlation										
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000	.000	.000
	N	39	39	39	39	39	39	39	39	39	39
6	Retest Pearson	.992**	.998**	.991**	.996**	.996**	1	.996**	.993**	.995**	.997**
	Correlation										
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000	.000	.000
	N	39	39	39	39	39	39	39	39	39	39
7	Retest Pearson	.988**	.998**	.988**	.998**	.997**	.996**	1	.992**	.992**	.998**
	Correlation										
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000	.000	.000
	N	39	39	39	39	39	39	39	39	39	39
8	Retest Pearson	.996**	.994**	.997**	.989**	.990**	.993**	.992**	1	.998**	.992**
	Correlation										
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000		.000	.000
	N	39	39	39	39	39	39	39	39	39	39
9	Retest Pearson	.998**	.994**	.998**	.990**	.992**	.995**	.992**	.998**	1	.992**
	Correlation										
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000		.000
	N	39	39	39	39	39	39	39	39	39	39
10	Retest Pearson	.988**	.998**	.987**	.999**	.999**	.997**	.998**	.992**	.992**	1
	Correlation										
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	
	N	39	39	39	39	39	39	39	39	39	39

\*\* . Correlation is significant at the 0.01 level (2-tailed).

The result shows high positive relationship between all PDF result of repeated measurements with  $r_p$  values range from  $r_p=0.990$  to  $0.999$  for all speech samples from



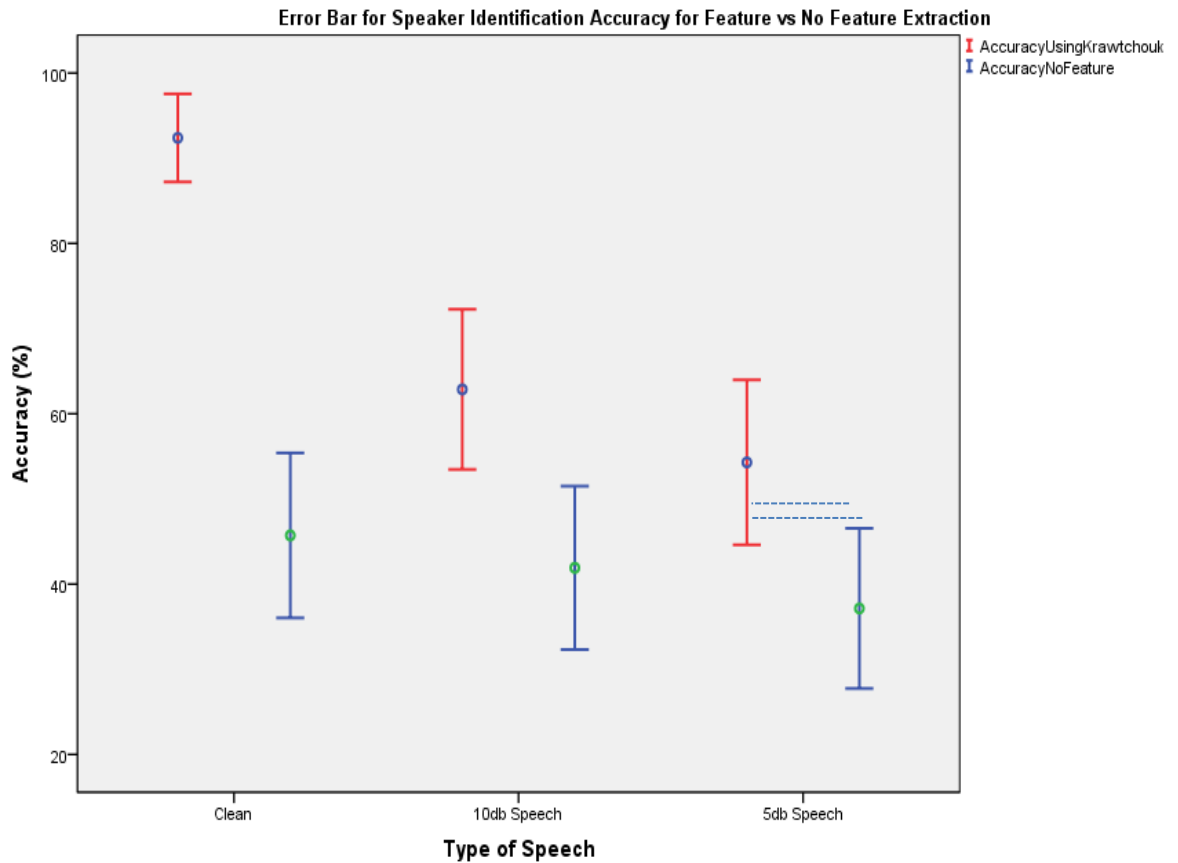
39 speakers shown by the value of Pearson correlation coefficients. This also indicates that only low level of variability affects the consistency of the system. The Sig. (2-tailed) shows value lower than 0.05 level of confidence interval, concluding that there is a statistically significant correlation between all the retest conditions.

Meanwhile Table 4.8 shows the result of standard deviation and mean for 10 repeated tests to check the variability between the results.

**Table 4.8: Standard deviation and mean for 10 repeated tests.**

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Retest1	39	-211.30539	-206.21109	-207.9906256	.95419865
Retest2	39	-210.57233	-206.51532	-208.3340480	.86891193
Retest3	39	-211.03943	-206.55284	-208.3148852	.95132245
Retest4	39	-211.30539	-206.21109	-208.0120592	.95023707
Retest5	39	-210.95013	-206.53490	-208.3568728	.90960353
Retest6	39	-210.27602	-206.50212	-208.2694090	.87582205
Retest7	39	-211.30539	-206.21110	-208.0017466	.94933678
Retest8	39	-210.27601	-206.50212	-208.3263034	.84085251
Retest9	39	-210.57232	-206.51532	-208.2776469	.90175304
Retest10	39	-211.30539	-206.21110	-208.0070549	.95910665
Valid N (listwise)	39				

A very low coefficient of variation (COV) (<0.5%) calculated from the standard deviation divided by the mean from the table above for each retests shows that the data is less spread out from the mean point suggesting low variability. This also suggests that the verification system's consistency is reliable.



**Figure 4.5: Error bar graph (for 95% CI) Speaker identification accuracy for with (red) and without (blue) feature extraction.**

Figure 4.5 shows the result of error bar graph comparing the accuracy between clean, 10 dB and 5 dB noise-altered speech quality using group differences of 95% confidence interval. The error bars between clean and 10 dB conditioned shows no overlapping between accuracy with and without feature extraction method, suggesting that the result between them is statistically significant at  $p < 0.05$ . This suggests that both tests are predicted to unlikely occurred by chance alone. Meanwhile, for 5 dB condition case, the error bar can be shown overlapped with each other that does not relate to any statistical reference conclusion. However, the overlapped less than 25% region between the error bars could be suggesting that it is also statistically significant but only to a lower extent (Belia et al., 2005).

## Chapter 5. Conclusion

The task of speaker verification is usually done by implementing the classical method of acoustical analysis such as LPC, MFCC, etc. However, this project explores an alternative yet more realistic approach for speaker verification task. The proposed method employs a model of the auditory system that mimics the processing strategy undertaken by human for the similar tasks. A physiologically-based computational model of AN provides responses to speech stimuli for a range of characteristic frequencies from which features are extracted or the AN output has been used directly for the verification and identification task.

The use of output directly from the AN response which is ENV only was not good when applied in speaker identification but not in the case of verification. This is because the test data are compared to all possible GMM model and depend on the value of threshold as opposed to choosing the maximum GMM probability in the identification case. However, the accuracy increases from only at 47.7% to 92.4% when Krawtchouk feature extraction is applied. Meanwhile, the system's robustness was also remained relatively unaffected (for speaker verification) although lessened the performance by less than 1.5%.

### Limitation and Future Work

The limitation of this project is that the simulation of the AN model response takes a long time in addition to the time required for training using higher number of Gaussian components. In order to get a good prediction for AN responses, the AN model employed in this study requires a very high sampling rate of 100 kHz. Furthermore, as the output of the AN model is in the form of neurogram that provides detailed activities

of the neurons in terms of spikes, it becomes hard to find a suitable way to extract features without including redundant information in it. It is somehow depicted as an image, and thus common feature extraction method cannot be fully utilized. Therefore orthogonal moment method (an image feature extraction method) is applied as a feature extractor in this project. The result does show improvement in the system performance from 45.7% to 92.4% when Krawtchouk moment is applied (for identification). However, this also causes additional time to be added in the overall computational time. It is hard to implement the system in a real time application as the training time itself is more than 15 minutes for a single speaker.

The acoustic data used in this study were not recorded in a quiet environment. Should the recording take place in a proper recording booth with proper recording instruments, it might make the result more accurate for clean condition. Furthermore, a standard database that has been widely used such as the TIMIT's and KING's speech corpuses could be used in the future to get a more reliable result.

To properly show the benefit of using AN model in speaker verification, a side study using conventional method of acoustic analysis (e.g. LPC) could be used in future to make a better comparison. Furthermore, other feature extraction and classification techniques could be applied to get to know which method shows higher compatibility with using the AN model response. Moreover, better representation of the system's performance such as using a Detection Error Tradeoff (DET) or Receiving Operating Characteristic (ROC) curves could be applied. The system could also be implemented with hardware design for study purpose and to make it able to observe its performance in real-time application. Finally, the training process could also be improvised by using

text-independent speaker verification as opposed to text-dependent to increase the system's performance.

In conclusion, despite the limitation and further improvement that can be done to further strengthen the system, the objectives of the project are achieved.

## References

- Aalto, D., Simko, J., Vainio, M. (2013). *Language background affects the strength of the pitch bias in a duration discrimination task*. Paper presented at the Proceedings of the Interspeech 2013 : 14th annual conference of the international speech communication association.
- Belia, S., Fidler, F., Williams, J., Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10 (4), 389-396
- Carballo, G., Alvarez-Nodarse, R., Dehesa, J. S. (2001). Chebychev Polynomials in a Speech Recognition Model. *Applied Mathematics Letters*, 14, 581-585.
- Colombi, J. M., Anderson, T. R., Rogers, S. K., Ruck, D.W. and Warhola, G. T. (1993). *Auditory model representation for speaker recognition*. Paper presented at the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, USA.
- Duval, M. A., Vega-Pons, S., Garea, E. (2010). Experimental Comparison of Orthogonal Moments as Feature Extraction Methods for Character Recognition. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 64(19), 394-401.
- de Lima, C., Alcaim, A., Apolinario, J. *Text independent speaker verification using GMM*. IEEE South-American Workshop on Circuits and Systems, Rio de Janeiro, Brazil, 2001.
- Dkici, E., Saracla, M. (2009). *Investigating the Effect of Training Data Partitioning from GMM Supervector Based Speaker Verification*. Paper presented at the 24<sup>th</sup> Symposium on Computer and Information Sciences, Northern Cyprus.
- Hines, A. and Harte, N. . (2012). Speech intelligibility prediction using a Neurogram Similarity Index Measure. *Speech Commun*, 54(2), 306-320.
- Ilyas, M. Z., Samad, S. A., Hussain, A., Ishak, K.A. (2007). *Speaker Verification using Vector Quantization and Hidden Markov Model*. Paper presented at the 5<sup>th</sup> Student Conference on Research and Development, Malaysia.
- Flanagan, J. L. (1960). Models for approximating basilar membrane displacement. *The Bell System Technical Journal*, 1163-1191.
- Jassim, W.A., Raveendran, P., Mukundan, R. . (2012). New orthogonal polynomials for speech signal and image processing. *IET Signal Processing*, 6(8), 713-723.
- Johannesma, P. I. M. (1972, June 1972). *The pre-response stimulus ensemble of neurons in the cochlear nucleus*. Paper presented at the The proceeding of the symposium on hearing Theory.
- Maki, K., Akagi, M., Hirota, K. (2009). Functional model of auditory peripheral system: Modeling phase-locking properties and spike generation process of auditory nerves. *J. Acoust. Soc. Jpn*, 65(5), 239-250.
- Li, Q. (2003). "Solution for pervasive speaker recognition," SBIR Phase I Proposal, Submitted to NSF IT.F4, Li Creative Technologies, Inc., NJ, June 2003
- Li, Q. (2009). *An auditory-based transform for audio signal processing*. Paper presented at the Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY.
- Li, Q., Huang, Y. (2010). *Robust speaker identification using an auditory-based feature*. Paper presented at the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Dallas, TX.

- Liberman, M. C. (1978). Auditory nerve response from cats raised in a low noise chamber. *J. Acoust. Soc. Am.* 63, 442–455.
- Lyon, R. F. (1988). An analog Electronic Cochlea. *Transactions on Acoustics, Speech, and Signal Processing*, 6(7), 1119-1134.
- Lyon, R.F., Katsiamis, A.G., Drakakis, E.M. . (2010, May 30-June 2 2010). *History and future of auditory filter models*. Paper presented at the International Symposium on Circuits and Systems (ISCAS) Proceedings of 2010 IEEE, Paris.
- Mamun, N. M. *Prediction of Speech Intelligibility using a Neurogram Orthogonal Polynomial Measure (NOPM)*. Unpublished Manuscript.
- Abuku, M., Azetsu, T., Uchino, E., Suetake, N. (2010, Dec 15-17). *Application of peripheral auditory model to speaker identification*. Paper presented at the 2010 Second World Congress on Nature and Biologically Inspired Computing, Fukuoka, Japan.
- Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. *J. Acoust. Soc. Am.*, 79(3), 702-711.
- Pandit, M., Kittler, J. (1998). *Feature selection for a DTW-based speaker verification system*. Paper presented at the Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, USA.
- Patterson, R. D, Allerhand, M. H., Gigure, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *J. Acoust. Soc. Am.*, 98(4), 1890-1894.
- Rani, J. S., and Devaraj, D. (2012). Face recognition using Krawtchouk moments. *Sa-dhana* . 37(4), 441-460.
- Reynolds, D. A., Quatieri, T. A., and Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10, 19-41.
- Reynolds, D. A., Rose, R. C. (1995). Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*, 3(1), 72-83.
- Shao, Y., Srinivasan, S., Wang, D. (2007). *Incorporating Auditory Feature Uncertainties in Robust Speaker Identification*. Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. Honolulu.
- Azetsu, T., Abuku, M., Uchino, E., Suetake, N. (2012). *Speaker Identification in Noisy Environment with Use of the Precise Model of the Human Auditory System*. Paper presented at the Proceedings of the International Multiconference of Engineers and Computer Scientists, Hong Kong.
- Togneri, R., Pullella, D. (2011). An Overview of Speaker Identification: Accuracy and Robustness Issues. *IEEE Circuits and Systems Magazine*, Second Quarter 2011, 23-61.
- Turetsky, R., Ellis, D. (2003, Sept 9, 2012). Dynamic Time Warp (DTW) in Matlab. Retrieved August 1, 2013, from <http://www.ee.columbia.edu/ln/labrosa/matlab/dtw/>
- Yap, P.T., Paramesran, R. (2003). Image Analysis by Krawtcheok Moments. *IEEE Transactions on Image Processing*, 12(11), 1367-1377.
- Zilany, M. S. A., & Bruce, I. C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *J Acoust Soc Am*, 120(3), 1446-1466.
- Zilany, M. S. A., & Bruce, I. C. (2007). Representation of the vowel /epsilon/ in normal and impaired auditory nerve fibers: model predictions of responses in cats. *J Acoust Soc Am*, 122(1), 402-417.

Zilany, M. S. A., 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. *J Acoust Soc Am*, *126*(5), 2390-2412.