# SPEAKER RECOGNITION USING NEURAL NETWORK

ZUL RASYIED BIN AB. RASAT

This thesis is submitted as partial fulfillment of the requirements for the award of the Bachelor of Electrical Engineering (Electronics)

Faculty of Electrical & Electronics Engineering
Universiti Malaysia Pahang

JUNE, 2012

# ABSTRACT

Speaker recognizer may be employed as part of a security system requiring user authentication. Mostly, years by years there have been several techniques being developed to achieve high success rate of accuracy in the identification and verification of individuals for authentication in security systems. A major application area of such systems would be providing security for telephone-mediated transaction systems where some form of anatomical or "biometric" identification is desirable. In daily usage, it maybe can be used in car environment or building to voice control non-critical operation such as open the gate to ensure a maximum control to the car or building and enhance the safety. This project is emphasizes on speaker recognizer system which is automatically verifying or recognizing the identity of the speakers based on voice unique characteristics. This system will be focusing on text-dependant method and in open-set situation. The feature extraction is done by Mel Frequency Cepstral Coefficients (MFCC). The feature for the speaker who has to be identified are extracted and compared with the stored templates by using Back-propagation Algorithm. After that, the trained network corresponds to the output meanwhile the input is the extracted features of the speaker are identified. The best match of recognize voice is found the speaker identity after the network has done its weight adjustment. For the decision-making purpose, Artificial Neural Network (ANN) method will be used. This project will be done by applying Neural Network Toolbox in MATLAB software. Lastly, the system should be able to automatically accept or reject the voice of different person based on voice unique characteristic.

## ABSTRAK

Pengesan Suara boleh digunakan sebagai sebahagian daripada sistem keselamatan yang memerlukan pengesahan pengguna. Dari tahun ke tahun terdapat beberapa teknik yang dibangunkan untuk mencapai kadar kejayaan yang tinggi terhadap ketepatan dalam pengenalan dan pengesahan individu untuk pengesahan dalam sistem keselamatan. Satu kawasan permohonan utama sistem itu akan menyediakan jaminan bagi sistem transaksi telefon-pengantara di mana beberapa bentuk anatomi atau pengenalan "biometrik" adalah wajar. Dalam penggunaan harian, ia mungkin boleh digunakan dalam persekitaran kereta atau bangunan untuk menyuarakan kawalan operasi yang tidak kritikal seperti membuka pintu untuk memastikan kawalan maksimum untuk kereta atau bangunan dan meningkatkan keselamatan. Projek ini memberi penekanan kepada sistem pengesan suara yang secara automatic akan mengesahkan atau mengiktiraf identiti individu berdasarkan ciri-ciri suara yang unik. Sistem ini akan memberikan tumpuan kepada kaedah yang bergantung kepada teks dan dalam keadaan set terbuka. Pencarian sifat dilakukan oleh *Mel Frequency Cepstrum Coefficient* (MFCC). Ciri-ciri untuk pembesar suara yang dikenal pasti diekstrak dan dibanding dengan suara sedai ada yang disimpan dengan menggunakan *backpropagation* algorithm. Selepas itu, rangkaian yang terlatih sepadan dengan output Sementara itu, input adalah ciri-ciri yang diekstrak penceramah dikenal pasti. Perlawanan terbaik suara mengiktiraf didapati identiti pembesar suara selepas rangkaian telah dilakukan pelarasan berat. Bagi tujuan membuat keputusan, kaedah *Artificial Neural Network* (ANN) akan digunakan. Projek ini akan dilakukan dengan menggunakan Toolbox Rangkaian Neural dalam perisian MATLAB. Akhir sekali, sistem sepatutnya boleh secara automatik menerima atau menolak suara orang yang berbeza berdasarkan ciri-ciri suara yang unik.

# TABLE OF CONTENT

# LIST OF FIGURES

## LIST OF ABREVIATION

| | | |
|---|---|---|
| AI | - | Artificial Intelligence |
| ANN | - | Artificial Neural Network |
| NN | - | Neural Network |
| LMS | - | Least Mean Square |
| MSE | - | Mean Square Error |
| MFCC | - | Mel Frequency Cepstrum Coefficient |
| FFT | - | Fast Fourier Transform |
| DCT | - | Discrete Cosine Transform |
| SR | - | Speaker Recognition |
| DTW | - | Dynamic Time Warping |
| MLP | - | Multilayer Perceptron |

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

The term *biometrics* comes from the ancient Greek *bios* is "life" and *metron* is "measure". Biometrics refers to the entire class of technologies and techniques to uniquely identify humans [1]. Biometric technology has been used in many fields as a means of protecting assets. Many organization has use it to enhance physical and logical access controls for decades such as military, intelligence and law enforcement.

Biometric systems work through enrolling users by measuring and storing their particular biometric and then later comparing the stored biometric data with data from unverified subjects to determine whether they should be allowed to access a system or location. The entire processes in detail are:

1. Enrollment.

   Before a user can start to use biometric system, the user must complete the enrollment process which is to provide the user's biometric data. Usually the biometric system will request several samples so that the system can determine an average and deviation.

2. Usage

   When the user wishes to access a system or building guarded with biometrics, the user authenticates according to procedure. In this process the biometric system will compare the sample with data stored at enrollment process and make decision reject or accept on whether the biometric data matches or not.

3. Update.

   For the type of biometrics that change slowly over time (such as handwriting or facial recognition), the biometric system may need to update the data originally submitted at enrollment. The biometric system may perform this update with each subsequent measurement (thereby increasing the number of samples, with emphasis on the newer ones), or it may utilize a separate update process.

Biometric Authentication methods are divided into two categories which are [2]:

1. *Behavioral-based authentication methods* perform the identification task by recognizing people's behavioral patterns, such as signatures, keyboard typing, and voice print.
2. *Physiological-based authentication methods* verify a person's identity by means of his or her physiological characteristics such as fingerprint, iris pattern (eye blood vessel pattern), palm geometry, DNA or facial features.

Speaker Recognition (SR) is one of the biometric technologies that have been developed almost 3 to 4 decade ago. It is divided in behavioral-based authentication methods in biometric authentication families. Speaker recognition is a generic term which refers to any task which discriminates between people based upon their voice characteristics [3]. It is generally divided into two categories: speaker verification and speaker identification.

Speaker identification determines the identity of an unknown speaker from a group of an unknown speaker from a group of known speakers [2]. Meanwhile speaker verification authenticates the identity of a speaker based on or his own voice. A speaker that claiming his or her identity is called a claimant and an unregistered speaker pretending to be a registered speaker is called an impostor [2]. The goal is to automatically accept impostors (false acceptance), not reject registered speakers (false rejection) or reject the unregistered speakers that are claimed their identity.

## 1.2 Problem Statement

Nowadays, the security system mostly used the old technology for security system. This old security system is no longer efficient today. This is because the old identification or verification that we used as the "PIN" could be lost, stolen, forgotten or misplaced. It is also cannot differentiate between genuine person and the impostor.

Beside, when we used knowledge as the authentication mechanism, it is difficult to remember the password although it is easily recallable password such as pet's name or birthday date. It is because that entire kind authentication can easily been guessed by the theft.

With the advanced technologies that have been achieved today, most of the researchers have founded that using biometric as the security system is the most efficient method. It is because biometric can only verify the unique characteristic that we have such as voice, fingerprint, iris and et cetera.

Voice Recognition or SR is one of the biometric security systems that use voice as the identification. Voices have its unique characteristic. It cannot be copy or steal by anyone and always with us. There never been two people that have the same voice. If someone want copy it by recording, the machine or equipment that been used must have a very high quality.

## 1.3 Project Objective

The main objectives of this project are:

1. To build a system that automatically accept or reject an identity that is claimed by the speaker.
2. To create voice for the word spoken by using MATLAB software.
3. To recognize voice from feature extracted data for the recognition analysis purpose.

## 1.4 Project Scope

The limitations of this project are listed below:

1. Conducted in controlled environment.
2. The speaker's voice is in good health.
3. Unavailable to person who have mute problem.
4. Decision will be made in offline mode.

## 1.5 Expected Outcome

To build a system that can identify and verify automatically the voice of different person based on its unique characteristic.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Speaker Verification and Speaker Identification

There are two types of speaker verification [7]:

1. Text dependant – Speaker speech corresponds to known text, cooperative user, 'PIN' type application.
2. Text independent – No constraints on what the speakers speaks, potentially uncooperative user.

Meanwhile, there are two types of speaker identification [9]:

1. Closed-set – Identifies the speaker as one of those enrolled, even if he or she is not actually enrolled in the system.

2. Open-set – The system should be able to determine whether a speaker is enrolled or not (impostor) and, if enrolled the system will determine his or her identity.

## 2.2 Previous Research

Ehab F. M. F. Badran and Hany Selim [9] have came out with the statistic for text-dependent and text-independent speaker recognition. For the average result of text-dependent, the verification correct percentage is 95.67% with false rate 4.33% compared to text-independent which is only 92.2% with false rate 7.78%. For the average result of identification, the system just only confused 3.33% in confusing of text-dependent compared to text-independent that have 6.377% of confusing in the voice signal. It shown that text-independent speaker recognition have more higher accuracy in speaker recognition.

Anjali Bala, Abhujeet Kumar and Nidhika [6] said that both MFCC Coefficient and Dynamic Time Warping (DTW) algorithm have been worked out for same speech signals as well for different speech signal and found that if both speech signal are same the cost will be 0 and if the speech signal are different voices the cost then definitely will have some value to show that there is mismatching. From the research it shown that by using this both algorithm it will worked very well in this project also.

Cheang Soo Yee and Abdul Manan Ahmad [7] said that accuracy correctly predicted for Artificial Neural Network (ANN) is 90% for 10 speakers, 84% for 20 speaker and 74% for 30 speakers respectively. From their research it shown that ANN is one of the most suitable method to use as the decision-making step as to get this project expected result.

Based on Yutai, W., Bo, L., Xiaoqing, J., Feng, L. and Lihao, W. [10] claims that Mel Frequency Cepstrum Coefficient (MFCC) is the technique that have been developed to recognize the voice owner and also claims that the MFCC is the most widely used in speech and speaker recognition. However the experiment of this paper showed that the

parameterization of the MFCC which is best for discriminating speakers is different from the one usually used for speech application.

## 2.3 Approaches to Speech Recognition

There are three categories of Speech Recognition according to the research paper Speaker Identification Using Neural Network [7]:

1. The Acoustic Phonetics Approaches

   This approach is based upon the theory of acoustic phonetics that postulate that there exist a set of finite, distinctive phonetic units in spoken language and that the phonetic units are broadly characterized by a set of properties that can be seen in the speech signal, or its spectrum, over time. It is assumed that the rules governing the variability are straightforward and can readily be learned and applied in practical situations. The first step in this approach is called segmentation and labeling phase. The second step is required just for speech recognition which it is attempt to determine a valid word from the sequence of phonetic labels produced in the first step.

2. The Pattern Recognition Approaches

   The Pattern Recognition approach to speech is basically one in which the speech patterns are used directly without explicit feature determination and segmentation. There two steps in this approach; training of speech patterns and recognition of patterns via pattern comparison. The concept of this approach is that if enough version of pattern to be recognized are included in the training set provided to the algorithm, the training procedure should be able to adequately characterize the acoustic properties of the pattern.

3. The Artificial Intelligence Approaches

   This approach is the hybrid of the last two approach before which is acoustic phonetic approach and the pattern recognition approach in which its ideas and concepts. The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies intelligence in visualizing, analyzing and lastly making decision on the acoustic features.

## 2.3 Mel Frequency Cepstral Coefficients (MFCC)

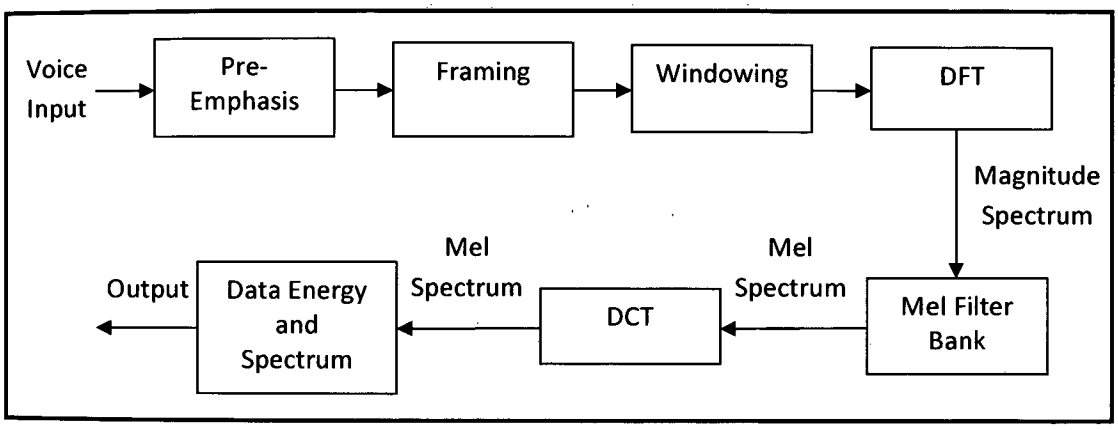


Figure 2.1: MFCC Block Diagram

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency [8]. Meanwhile MFCC is the coefficient that collectively makes up an MFC and it was derived from a type of cepstral.

MFCC is based on human hearing perceptions which cannot perceive frequencies over 1KHz and its have two types of filter which are spaced linearly at low frequency below 1000Hz and logarithmic spacing above 1000Hz [6]. MFCC consist of seven process or steps as shown in Figure 2.1

1. Pre-Emphasis

   This first steps is refers to a system process designed to increase, within a band of frequencies, the magnitude of some (usually higher) frequencies with respect to the magnitude of the others (usually lower) frequencies in order to improve the overall SNR. Then it emphasizes higher frequencies of the signal that passing through a filter. Pre-emphasis will increase the level of energy of signal at higher frequency.

2. Framing

   In this process, the speech samples that have been obtained will be segmenting into a small frame from Analog to Digital Converter (ADC) within the range of 20 to 40 millisecond of length. The adjacent frames are being separated by $M$ $(M<N)$ and the typical value that were used are $M$=100 and $N$=256.

3. Windowing

   For the windowing process, Hamming windowing will be used. It is used as window shape by considering the next block in feature extraction processing chain and integrals all the closest frequency lines. If the windowing defined as $W(n)$, $0 \leq n \leq N-1$.

   $N$ = number of samples in each frames.
   $Y[n]$= Output signal
   $X[n]$= Input signal
   $W[n]$= Hamming window

$$Y[n] = X(n) * W(n) \tag{2.1}$$

4. Discrete Fourier Transform (DFT)

   The next step is the Fast Fourier Transform (FFT), which converts each frame of $N$ samples from the time domain into the frequency domain. The FFT is a fast

algorithm to implement the DFT and is used to convert the convolution of the glottal pulse *U[n]* and the vocal tract impulse response *H[n]* in the time domain.

$$Y(w) = FFT[h(t)*X(t)] = H(w)*X(w) \quad (2.2)$$

If *X(w), H(w)* and *Y(w)* are the fourier transform of *X(t), H(t)* and *Y(t)* respectively.

5. Mel filter bank

   The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the Centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then each filter output is sum of its filtered spectral components. The equation below is used to compute Mel for given frequency,*f* in Hz.

$$F(Mel) = [2595*\log 10(1+f/700)] \quad (2.3)$$

6. Discrete Cosine Transform (DCT)

   In this process, the log Mel spectrum will be converted into time domain using DCT. The result of the conversion is called MFCC. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

7. Data Energy and Spectrum

   The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time. 13 delta or velocity feature (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for signal $x$ in a window from time sample *t1* to time sample *t2*, is represented as shown below.

$$Energy = \sum X^2[t] \tag{2.4}$$

When $X[t] = signal$

Each of the 13 delta features represent the change between frames corresponding to cepstral or energy feature, while each of 39 double delta features represent the change between frames in the corresponding delta features.

## 2.4 Dynamic Time Warping (DTW)

DTW algorithm is based on Dynamic Programming and it is used for measuring similarity between two time series which may vary in time or speed. Beside, this technique is also used to find the optimal alignment between two times series may be warped non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or two determine similarity between the two time series. An *n*-by-*m* matrix where the ($i^{th}$, $j^{th}$) element of the matrix contains the distance $d$ ($qi$, $cj$) between the two points $qi$ and $cj$ is constructed to align two sequence using DTW. By using Euclidean distance, the absolute distance between the values of two sequences is calculated by equation below

$$D\ (qi,\ cj) = (qi - cj)2 \tag{2.5}$$

Each matrix element ($I$, $j$) corresponds to the alignment between the points $qi$ and $cj$. Then, accumulated distance is measured by equation below

$$D\ (I, j) = min[D(I-1, j-1)],\ D(i-1, j),\ D(I, j-1)] + d(I,) \tag{2.6}$$

## 2.6 Artificial Neural Network (ANN) Properties

An ANN, usually called Neural Network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data. The original inspiration for the term ANN came from examination of central nervous systems and their neurons, axons, dendrites, and synapses, which constitute the processing elements of biological neural networks investigated by neuroscience. In an artificial neural network, simple artificial nodes, variously called "neurons", "neurodes", "processing elements" (PEs) or "units", are connected together to form a network of nodes mimicking the biological neural networks — hence the term "artificial neural network". Because neuroscience is still full of unanswered questions, and since there are many levels of abstraction and therefore many ways to take inspiration from the brain, there is no single formal definition of what an artificial neural network is. Generally, it involves a network of simple processing elements that exhibit complex global behavior determined by connections between processing elements and element parameters. While an artificial neural network does not have to be adaptive per se, its practical use comes with algorithms designed to alter the strength (weights) of the connections in the network to produce a desired signal flow. These networks are also similar to the biological neural networks in the sense that functions are performed collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which various units are assigned. Currently, the term ANN tends to refer mostly to NN models employed in statistics, cognitive psychology and artificial intelligence. Neural network models designed with emulation of the Central Nervous

System (CNS) in mind are a subject of theoretical neuroscience and computational neuroscience.

### 2.6.1 Multi Layer Perceptron (MLP)

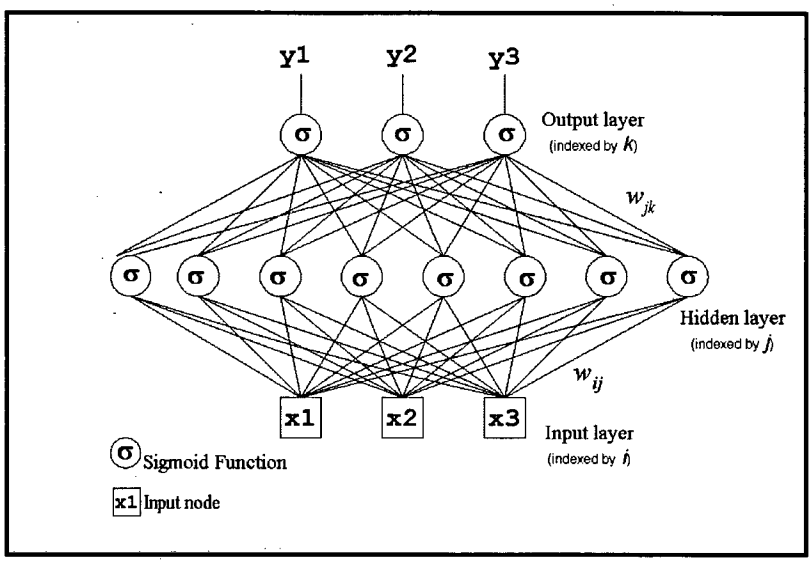


Figure 2.2: Multilayer Perceptron (MLP)

The MLP is a hierarchical structure of several perceptrons, and overcomes the shortcomings of these single-layer networks. The multilayer perceptron is an ANN that learns nonlinear function mappings. The multilayer perceptron is capable of learning a rich variety of non-linear decision surfaces. Nonlinear functions can be represented by MLP with units that use nonlinear activation functions. Multiple layers of cascaded linear units still produce only linear mappings.

ANN with one or more layers of nodes between the input and the output nodes is called multilayer network. The multilayer network structure, or architecture, or topology, consists of an input layer, two or more hidden layers, and one output layer. The input nodes pass values to the first hidden layer, its nodes to the second and so on till producing outputs. A network with a layer of input units, a layer of hidden units and a layer of output units is a two-layer network. A network with two layers of hidden units

is a three-layer network, and so on. A justification for this is that the layer of input units is used only as an input channel and can therefore be discounted.

A two-layer NN that implements the function:

$$f(\mathbf{x}) = \sigma\left(\sum_{j=0}^{J} w_{jk}\sigma\left(\sum_{i=0}^{I} w_{ij}x_i + w_{0j}\right) + w_{0k}\right) \quad (2.7)$$

where: $\mathbf{x}$ is the input vector,

$w_{0j}$ and $w_{0k}$ = thresholds,

$w_{ij}$ = weights connecting the input with the hidden nodes

$w_{jk}$ = weights connecting the hidden with the output nodes

$s$ = sigmoid activation function.

These are the hidden units that enable the multilayer network to learn complex tasks by extracting progressively more meaningful information from the input examples. The multilayer network MLP has a highly connected topology since every input is connected to all nodes in the first hidden layer, every unit in the hidden layers is connected to all nodes in the next layer, and so on. The input signals, initially these are the input examples, propagate through the neural network in a forward direction on a layer-by-layer basis, that is why they are often called feedforward multilayer networks.

Two kinds of signals pass through these networks:

1. function signals: the input examples propagated through the hidden units and processed by their activation functions emerge as outputs;
2. error signals: the errors at the otuput nodes are propagated backward layer-by-layer through the network so that each node returns its error back to the nodes in the previous hidden layer.

## 2.7 Equal Error Rate (EER), False Accept Rate (FAR) and False Reject Rate (FRR).

Equal Error Rate (EER) for the voice recognition is less than 1% and its Failure to Enroll Rate was only 2%. The EER occur when the decision threshold of a system is set so the proportion of False Accept Rate (FAR) approximately equal to False Reject Rate (FRR). More low the EER, the higher the accuracy of the biometric system. Failure to Enroll Rate is the failure of the biometric system to extract data. So, the more low the Failure to Enroll Rate, the more higher the system able to extract template data.

Table 2.1: Difference between biometric

| | **Finger** | **Voice** | **Iris** | **Face** |
|---|---|---|---|---|
| **Type** | Physical | Behavioral | Physical | Physical |
| **Method** | Active | Active | Active | Passive |
| **Equal Error Rate** | 2-3.3% | <1% | 4.1-4.6% | 4.1% |
| **Failure to Enroll** | 4% | 2% | 7% | 1% |
| **Nominal False Accept Rate** | 2.5% | <1% | 6% | 4% |
| **Nominal False Reject Rate** | 0.1% | <1% | 0.001% | 10% |
| **Liveness Aware** | No | Yes | Bo | Possible |
| **System Cost** | High | Low | Very High | High |

Table 2.1 above show the difference between biometric from the view of EER, failure to enroll, FAR, FRR and system cost. From here we can decide the most efficient biometric.

# CHAPTER 3

# METHODOLOGY

## 3.1 Introduction

This chapter explained the methodology that has been used in this project. The method has been divided into two parts. The first part is Enrollment Process and the second part is Verification/Identification Process. Figure 3.1 show the entire process for this project.