GENDER IDENTIFICATION OF CHILDREN USING HIDDEN MARKOV MODEL BASED ON MEL-FREQUENCY CEPSTRAL COEFFICIENT

ADIRA BINTI IBRAHIM

FACULTY OF ENGINEERING

UNIVERSITY OF MALAYA

KUALA LUMPUR

2013

GENDER IDENTIFICATION OF CHILDREN USING HIDDEN MARKOV MODEL BASED ON MEL-FREQUENCY CEPSTRAL COEFFICIENT

ADIRA BINTI IBRAHIM

RESEARCH REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF MALAYA

KUALA LUMPUR

2013

UNIVERSITI MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Adira Binti Ibrahim I.C No:

Matric No: KGL 110004

Name of Degree: Master of Biomedical Engineering

Title of Project Paper: Gender Identification of Children Using Hidden Markov Model based on Mel-Frequency Cepstral Coefficient

Field of Study: Signal Processing

I do solemnly and sincerely declare that:

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

Candidate's Signature: Date: 15 Feb 2013

Subscribed and solemnly declared before,

Witness's Signature: Date: 15 Feb 2013

Name:

Designation:

ABSTRAK

Pertuturan adalah komunikasi antara manusia dengan pelbagai bahasa yang diterjemahkan kepada perkataan, frasa dan ayat. Isyarat pertuturan membawa kepada nada intonasi yang boleh memberi informasi seperti slanga, emosi, jantina dan umur. Walau bagaimanapun, penyelidikan dalam vokal untuk kanak-kanak mengandungi beberapa kesusahan seperti kesalahan sebutan dan ketidak fasihan bahasa. Projek ini bertujuan untuk membangunkan sistem bagi mengenal pasti jantina penutur berdasarkan isyarat ucapan dengan menggunakan Model Markov Tersembunyi (HMM) sebagai pengenal. Koefisien Cepstral Frekuensi Mel (MFCC) digunakan sebagai kaedah ciri cabutan. HMM di latih dengan algoritma Baum-Welch dan di kaji dengan algoritma Viterbi untuk mendapatkan ketepatan mengenal pasti jantina. Analisis bingkai tunggal memberi ketepatan maksimum iaitu 64.17% pada kepanjangan isyarat 30 milisaat. Untuk analisis bingkai berganda, ketepatan maksimum ialah 64.26% pada kepanjangan analisis bingkai 20 milisaat dengan peralihan 10 milisaat. Untuk analisis bingkai tunggal, ketepatan budak perempuan adalah 67.78% manakala ketepatan budak lelaki adalah 60.56%. Untuk bingkai berganda, ketepatan budak perempuan adalah 65.74% dan budak lelaki adalah 62.78%. Maka, penutur perempuan mempunyai ketepatan yang tinggi berbanding penutur lelaki. Ketepatan mengenal pasti jantina adalah bergantung kepada kepanjangan bingkai iaitu rendah dan tinggi bingkai membawa kepada ketepatan rendah.

ABSTRACT

Speech is a communication between humans using variety of language that is translated into word, phrases and sentences. Speech signal carries pitch intonation that can express information such as accent, emotion, gender, and age. However, study in vowel for children has some difficulties such as false pronunciation and disfluencies of speech. This project aims to develop a system that can identify gender of speakers based on speech signal using Hidden Markov Model (HMM) as a recognizer. Mel Frequency Cepstral Coefficient (MFCC) was applied as the feature extraction method. HMM was trained with Baum-Welch algorithm and tested with Viterbi algorithm to get the gender identification accuracy. For single frame analysis, maximum accuracy was obtained at 64.17% at signal length of 30ms. For multiple frame analysis, maximum accuracy was achieved at 64.26% at AFL 20ms with 10 ms shift. For the single frame analysis, the accuracy of female children was 67.78% while accuracy for male children was 60.56%. For the multiple frame analysis, the accuracy for female children was 65.74% and 62.78% of male children. Hence, female speakers had higher identification accuracy compare to male speakers.

ACKNOWLEDGEMENT

First of all, I would like to thank Allah for guiding me in completing this Master project. It is by HIS grace and mercy that I am able to embark on the project within such a limited time. Alhamdulillah.

Secondly, I would like to thank my project supervisor, Dr. Ting Hua Nong for his guidance and directions to improve the quality of this research. His advices and support always be with me throughout of my research. He is very helpful when I need his suggestion to completion of this project.

Then, special thanks to my friend, Mostafa Mirhassani for his guidance and take his time to help me in completing this project.

Furthermore, I would like to thank all my fellow friends for giving me a moral support and shared the ideas throughout the research project development period. Thank you to my family for their encouragement and motivation for my academic pursuit. Lastly, thank you for other contributors for assisting in this project which is involved directly and indirectly upon completing this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS OR NOMENCLATURE

AFL	Analysis Frame Length
ASR	Automatic Speech Recognition
BP	Back Propagation algorithm
CPP	Cepstral Peak Prominence
DCT	Discrete Cosine Transform
EE	Energy Entropy
EM	Expectation-Maximization algorithm
E-step	Expectation step
F_0	Fundamental Frequency
FFT	Fast Fourier Transform
GEM	Generalized Expectation-Maximization algorithm
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HNR	Harmonic-to-Noise Ratio
HTK	Hidden Markov Model Tool Kit
LPC	Linear Predictive Coding
LRB	Left-right Banded
MFCC	Mel Frequency Cepstral Coefficient
M-step	Maximization step

ms	Millisecond
NN	Neural Network
RASTA	Relative Spectral
RASTA-PLP	Relative Spectral Perceptual Linear Predictive
RBF	Radial Basis Function
SFL	Speech Frame Length
STE	Short Time Energy
SVM	Support Vector Machine
VTLN	Vocal Tract Length Normalization
WNN	Window Neural Network
ZCR	Zero Crossing Rate

CHAPTER 1

INTRODUCTION

1.1 Introduction

The human being does not only use hand to write the information and body gesture to passing the knowledge, but lip motion which is speech is also the most important way of communicating with other societies. Communicated by speech is in the form of words, phrases, and sentences by applying proper grammatical rules. The composition of human speech constructs from a succession of phonemes that can nearly identical to the sounds of each letter of the alphabet.

Speech recognition is the most significant in the application when study of speech signal and this speech recognition can give many advantages and a lot of information such as to identify the gender, age and the vowel speech. The acoustic signal obtained from a microphone or a telephone was converted and the speech recognition process gives a set of words. The speech wave can determine linguistic information when apply to the computers or electronics circuits (Gomathy et al., 2011).

The gender seems to be the important factor linked to physiological differences that create speech variability. A speaker's gender can be one of the variability adversely affecting the speech recognizer's accuracy and separating speakers can be considered as an important way of improving a speech recognizer's performance (Phoophuangpairoj et al., 2009).

Phoneme is the smallest structural unit that recognizes meaning. Standard Malay can expressed properly based on the combination of consonant-vowel phonemes. The proper Bahasa Malaysia has only six vowel phonemes which are /a/, /e/, /ə/, /i/, /o/, and /u/ and 18 consonants of phonemes. These consonant-vowel units have the highest frequency of occurrence among different forms of sub-word units. The system can reduce region of search and improve accuracy and time if Automatic Speech Recognition (ASR) recognizes the vowel with a good accuracy (Siraj et al., 2010).

There are two fundamental operations in speech recognition system which is signal modeling and pattern matching. Signal modeling represents the process of converting a speech signal into a set of parameters while pattern matching is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal (Yusof et al., 2007).

The signal modeling requires four basic operations which are spectral shaping, feature extraction, parametric transformation, and statistical modeling. Table 1.1 shows the basic operations of signal modeling and the descriptions of the operations.

Table 1.1: Basic operations of signal modeling (Yusof et al., 2007).

Operations	Description
Spectral Shaping	The process of converting the speech signal from a sound pressure wave to a digital signal and emphasizing frequency components in the signal.
Feature Extraction	The process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal.
Parametric Transformation	The process of converting these features into signal parameter through the process of differentiation and concatenation.
Statistical Modeling	Conversion of parameters in signal observation vectors.

1.2 Problem Statement

The speech recognition so far mostly focuses on the recognition of the speech sounds among adults. Recognition of gender and age are less studied in adults as well as in children. The identification of gender among children is more challenging than in adult due to the dynamic characteristics of speech among children. Thus, the study is carried out to investigate the ability of the state-of-the-art speech recognition technology to identify the gender of the children speakers based on the speech sounds.

1.3 Significant of the Study

Gender identification gives important of this study since the societies can identify by gender without seeing them in person and in robotic applications, the robots can interact with users by providing suitable services to females and males according to their gender information. Then, the gender identification can be used to create more security for the places that allow only for females and males when the system can recognize the voice and characteristic of the gender.

1.4 Objectives of the Study

The objective of this study is to use HMM and MFCC to perform gender identification for Malay children based on the Malay vowel sounds.

1.5 Scope of the Study

Scopes of this study include the data familiarization, feature extraction and the recognizer. For the data familiarization, the vowels speech signals such as /a/, /e/, /ə/, /i/, /o/, /u/ was analyzed for 360 Malay children which is consist of 180 males and 180 females. Then, Mel Frequency Cepstral Coefficient (MFCC) was used as a feature extraction for single and multiple frame length. Lastly, the speech signals classified the identification accuracy using Hidden Markov Model (HMM) and the result was obtained.

1.6 Organization of Thesis

Chapter one describes about the introduction of the research project. Furthermore, the research problem, significant of study, objective and scope of the study are also discussed in this section.

Chapter two covers the literature review of the previous research work regarding the different feature extraction method and the selection of the classifier is also discussed. Besides, this chapter also discussed about the classification accuracy of the selected method that have been done by the other researchers.

Chapter three includes overall methodology of this project. The tool that was used to run the project is explained in detail with methodology of the project. The method used for the feature extraction method and classification are discussed with

block diagram. Feature extraction with MFCC from the speech signals which are taken from the children with age 7 years old until 12 years old has been explained and how to classify the data as the gender identification by applying them to hidden Markov model.

Chapter four discusses about the result and discussion of this project. The results of the identification accuracy using MFCC and hidden Markov model will be discussed in this chapter. Any problem regarding to the result also will be discussed

The conclusion is covered in chapter five. The summary about overall of the project will be discussed in this chapter. Any recommendation will be brought to this chapter for future review also has been discussed.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter describes the literature review of the project. The first subsection will discuss about speech communication and production. The following subsection will discuss more on Mel Frequency Cepstral Coefficient, Hidden Markov Model and HMM Tool Kit. The next subsection will describe about the related works on this project including the feature extraction and classification accuracy of the data. The final section concludes the literature review with relevant justification for the proposed project.

2.2 Speech Communication and Production

Speech is a communication between humans with a variety of language that is translated into words, phrases and sentences. Speech sounds are developed by air pressure vibrations produced by air exhaled from the lungs through the vibrating vocal cords and vocal tract and out of the lips and nose. Speech is a sequence of acoustic sounds known as phonemes, which is 40-60 in the English language that carry the spoken form of a language. The production of each phonemic sound is affected by the context of the neighboring phonemes. Speech signal has time frequency modulation carries the

formants and pitch intonation that can express information about linguistic, paralinguistic such as accent, emotion, health of the speaker and biological features such as the speaker's identity which is include the gender and age as shown in Table 2.1 (Vaseghi, 2006).

Table 2.1: Information conveyed in speech (Vaseghi, 2006).

Information	Description
Accent	Changes in the pronunciation in the form of substitution, deletion or insertion of phoneme units in the standard transcription of words and changes in speech resonance frequencies, pitch intonation, duration and emphasis.
Emotion and health	Carried by changes in the vibration of vocal fold, vocal tract resonance, duration and stress and changes in the dynamic of pitch and vocal tract spectrum.
Speaker's identity	Conveyed by the physical characteristics of a human's vocal folds, vocal tract, pitch intonations and stylistics.
Gender	Express by the pitch, which is related to the fundamental frequency and the size and physical characteristics of the vocal tract.
Age	Conveyed by the effects of the size and the elasticity of the vocal cords and vocal tract and the pitch.

Human speech production consists of the lungs, larynx, vocal tract cavity, nasal cavity, tongue, jaw and lips. Each part of the human speech production plays an important role to produce a beautiful sound and Figure 2.1 shows the system of the human speech production.

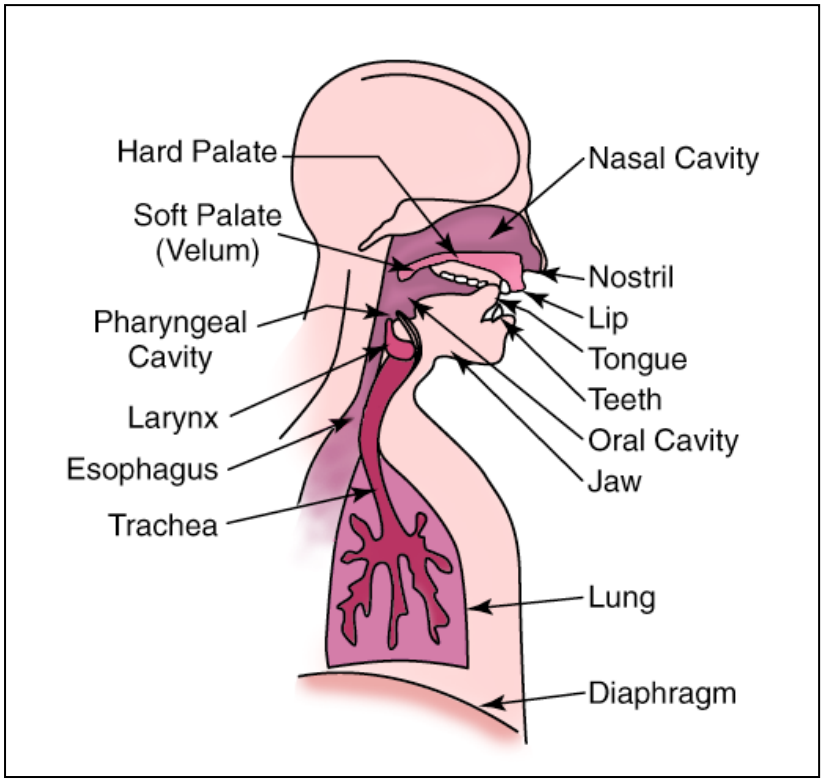

Figure 2.1: The human speech production system (Bouman, 2009)

Four processes are involved in the production of speech and the processes as follows (Giegerich, 1992):

i) Initiation – Air exhaled from the lung and if there are no modulations, the air will sound like a noise.

ii) Phonation – Occurs at larynx, which is having two folds to pass air by closing and opening of glottal fold. When the air passing through the glottis, vocal fold will vibrate and produced the sound. Then, it passes through the larynx and the pharynx and to the nasal or the oral cavity.

iii) Oral and nasal cavity – This part can differentiate between nasal consonants and other sounds.

iv) Articulators – Involve four parts such as tongue, lips, jaw, and velum which is all this part place in the mouth. This process can differentiate the speech sounds according to the place and how they are articulate.

2.3 Mel Frequency Cepstral Coefficient (MFCC)

MFCC is widely used in the speech recognition since it is based on the human peripheral auditory system. The human perception of the frequency of sounds for speech signals does not follow a linear scale. The definition of Mel Scale is when a subjective pitch is measured for each tone with an actual frequency, *f* measured in Hz. The Mel frequency scale is linear frequency spacing below 1000 Hz and logarithmic spacing above 1 kHz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels. Then, to compute the mels for a given frequency *f* in Hz, the approximate formula can be used as follows (Tiwari, 2010; Ittichaichareon et al., 2012):

$$\text{Mel}(f) = 2595 * \log10(1 + f/700) \tag{2.1}$$

The filter bank has a triangular bandpass frequency response and the spacing and bandwidth is determined by a constant mel-frequency interval. The Mel scale filter bank is a series of one triangular bandpass filters that have been designed to simulate the bandpass filtering. This represents a series of bandpass filters with constant bandwidth and spacing on a Mel frequency scale. Then, the coefficients of the MFCC were obtained after discrete cosine transform convert log mel spectrum to time domain Tiwari, 2010).

2.4 Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) provides effective algorithms for state and parameter approximation since it is a mathematical tool for modeling time series and it performs dynamic time warping for signals. In speech recognition, HMM is very useful for many pattern recognition and image analysis problems. Beside in speech recognition, the HMM has been used in a lot of area like finance, biological modeling and language.

In the early twentieth century, the name of the mathematical theory of Markov processes was given by Andrei Markov and the theory of HMMs was developed by Baum and the coworkers, which are Eagon, Petric, Soules, and Welss in the 1960s. Furthermore, in the early 1970s, Jim Baker at Carnegie-Mellon University was the first one used these HMMs for speech recognition (Blunsom, 2004; (Paul, 1990).

The elements and definition are needed to characterize an HMM completely are (Blunsom, 2004):

$$\lambda = (A, B, \pi) \tag{2.2}$$

Where,

A is a transition array, storing the probability of state *j* following state *i*. Note the state transition probabilities are independent of time:

$$A = [a_{ij}],\ a_{ij} = P\ (q_t = s_j | q_{t-1} = s_i) \tag{2.3}$$

B is the observation array, storing the probability of observation *k* being produced from the state *j*, independent of *t*:

$$B = [b_i(k)],\ b_i(k) = P(x_t = v_k | q_t = s_i) \quad (2.4)$$

π is the initial probability array:

$$\pi = [\pi i],\ \pi_i = P(q_1 = s_i) \quad (2.5)$$

S is state alphabet set, and *V* is the observation alphabet set:

$$S = (s_1, s_2, \ldots, s_N) \quad (2.6)$$

$$V = (v_1, v_2, \ldots, v_M) \quad (2.7)$$

Q to be fixed state sequence of length *T*, and corresponding observation *O*:

$$Q = q_1, q_2, \ldots, q_T \quad (2.8)$$

$$O = o_1, o_2, \ldots, o_T \quad (2.9)$$

Furthermore, there are four types of HMMs topologies which is included ergodic model, general left to right model, bakis model and linear model. The topology of HMM is defined as the statistical behavior of an observable symbol chronological sequence in term of a network of states, which acts the overall process behavior with regard to movement between states of the process, and describes the underlying variations in the behavior of the observable symbols within a state. The HMM topology consists of the number of states with different connections between states which depend on the

occurrence of the observable symbol sequences being modeled. Every state represents a similarly behaving portion of an observable symbol sequence process, such as phonemes to speech and facial features to face identification (Raymond C. Vasko et al., 1996).

The first topology is Ergodic model or it also can be called Fully Connected model. Each state of the model could be reached in a finite number of steps from every other state of the model. This model enables the returns of each state with probability one in finite intervals, by allowing non-zero state transition paths between any two states (Elmezain et al., 2009; Arica and Yarman-Vural, 1999). Figure 2.2 shows the Ergodic topology with four states:

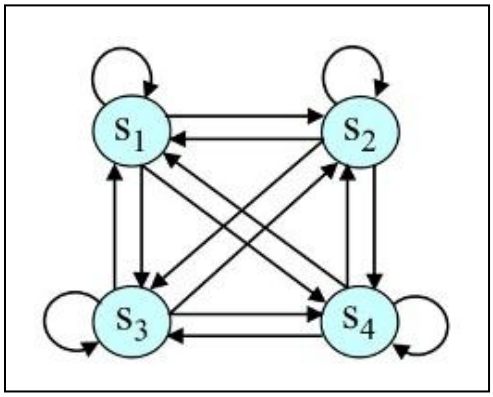

Figure 2.2: Ergodic topology with 4 states (Elmezain et al., 2009).

Then, Left-right Banded (LRB) or linear model is the second topology in the Hidden Markov Model as shown in Figure 2.3. This model underlying state sequence linked with model has the attribute that as time increases the state index increase or stays the same state which is no transitions are allowed to states whose indices are lower than the current state. For more clearer, each state of LRB model can go to the next state or to itself only (Elmezain et al., 2009).

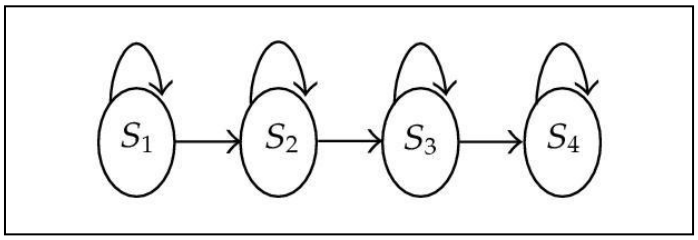

Figure 2.3: Left-right Banded topology or linear model (Wang et al., 2012).

Bakis model as shown in Figure 2.4 is also the topology of Hidden Markov Model which is this model is stays at the same states, move and skip the state. The comparison between the LRB and bakis model is this model can skip one state and the transition repeat until finish in N state (Elmezain et al., 2009).

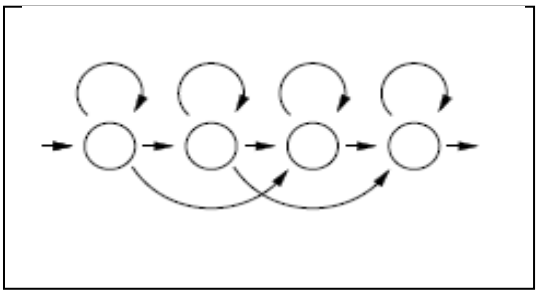

Figure 2.4: Bakis model (Fink, 2008).

Lastly, general left-right topology includes any number of skipping transitions and parallel paths but there is no backward loop that involves more than one state and this model same as the previous three topologies, it also can stay unchanged as the time increases (Wang, 1994). From the Figure 2.5, left-right transition topology shows that it includes all the topologies used in speech recognition and this topology was used by the researchers because its underlying structure that can model the temporal flow of speech signals over time (Abdulla and Kasabov, 1999).

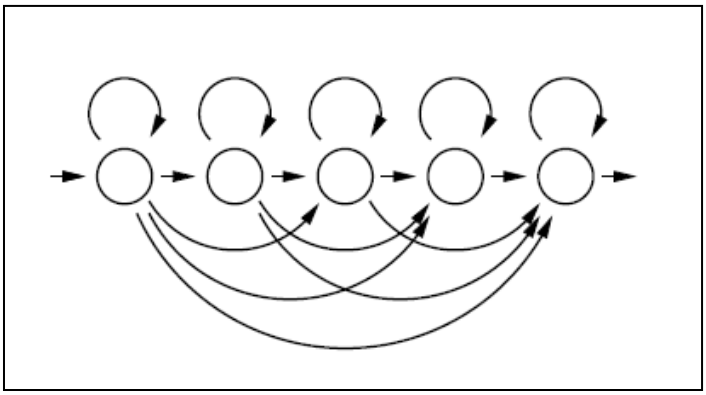

Figure 2.5: Left-right topology (Fink, 2008).

2.5 HMM Algorithms

Hidden Markov Model has three algorithms which are include Viterbi algorithm, the expectation-maximization (EM) algorithm and Baum-Welch algorithm invented by Leonard E Baum and Lloyd R. Welch (E.Baum et al., 1970). These algorithms provide a powerful tool for tailoring HMM topologies to data for use in knowledge discovery and clustering. Moreover, these training algorithms update HMM parameters based on new samples compared with the supervised training algorithm, it's only use statistics of known samples.

2.5.1 Baum-Welch algorithm

The Baum-Welch algorithm is applied to find the unknown parameters of a hidden Markov model (HMM) and it is a particular case of a Generalized Expectation-Maximization (GEM) algorithm (Kouemou, 2011). Churbanov (Churbanov and Winters-Hilt, 2008) suggested when using Baum-Welch algorithm to train the HMM, the expected probabilities of being at a certain state at a certain time-point using the forward-backward procedure need to find (Rosdi, 2008).

2.5.2 Viterbi algorithm

The idea of the Viterbi algorithm is come from Andrew Viterbi in 1967 for convolution codes over noisy digital communication links and it used for space, satellite program, military communications and cellular telephone systems in that time (Bell, 2006). Now, the Viterbi algorithm can use for worldwide application in decoding the convolution codes such as speech recognition, dial-up modems, computational linguistics, and bioinformatics. The Viterbi algorithm can be considered as the dynamic programming algorithm applied to the HMM since this algorithm can find and recalls the best path although a state sequence is hidden in the HMM framework (Kouemou, 2011; Rosdi, 2008).

2.5.3 Expectation-Maximization (EM) algorithm

Expectation-Maximization (EM) algorithm is applied in statistics for finding maximum likelihood estimates of parameters in probabilistic models. Training process estimates the HMM parameters in the most appropriate way since the speech signals can differ considerably for several acoustic surroundings. Although the convergence of EM algorithm can be slow, this algorithm for HMM's simple, well defined, and stable.

Kouemou (2011) gave an overview that EM algorithm has two stages to estimate the HMM parameters which is an Expectation step (E-step) and Maximization step (M-step). The E-step computes an expectation of the likelihood by including the latent variables while the M-step computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E-step. To begin another E-step, the parameters found on

the M-step are used until the process is repeated (Kouemou, 2011; Gotoh et al., 1998).

For HMM based speech recognition, the acoustic models are trained with Baum-Welch training algorithm, in which each observation is attributed to a set of acoustic models with weights. Only a portion of each observation, equal to its posterior probability, is associated with each model. Hence, Baum-Welch algorithm is the most suitable to train acoustic models and it have smoother convergence property than the Viterbi algorithm which is each observation is assigned to a single acoustic model (Shu et al., 2003).

2.6 Hidden Markov Model Toolkit (HTK)

HTK is a portable software toolkit for building and manipulating systems that use continuous density Hidden Markov Models (HMMs). HTK is mainly designed for building HMM based speech processing tools, in specific speech recognizers. This HTK can be used to perform a variety of tasks including isolated or connected speech recognition using models based on all word or sub-word units, it is especially suitable for performing large vocabulary continuous speech recognition (Wiggers and Rothkrantz, 2003).

Four approaches to build speech recognition systems have been discussed by SJ Young (1994). First approach is HTK is restricted to continuous density systems in preference to discrete systems because, it have a number of mathematically suitable properties in research. Then, second approach is parameter tying is regarded as being an essential necessity and HTK provide a generalized mechanism which allows tying at all levels.

Thirdly, HTK further an incremental approach to model building whereby a system of HMMs is refined through a number of stages involving provided model manipulation and model re estimation. Lastly, the fourth approach is building a complex HMM based system involves manipulating a diverse range of data including speech, transcriptions, and dictionaries. Furthermore, it provides many sets of integrated tools to facilitate these activities.

Moreover, there are two major processing stages required in HTK which is training phase and recognition phase. The training phase includes a training tools that are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions while the recognition phase is an unknown utterances are transcribed using the HTK recognition tools. Figure 2.6 shows that a processing stages in HTK.

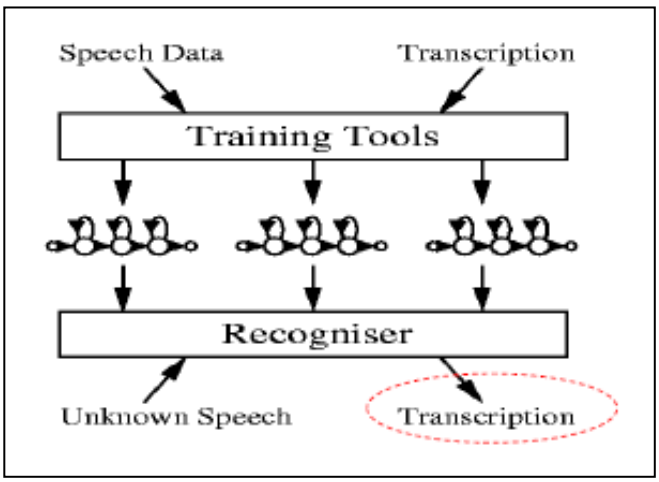

Figure 2.6: Processing stage in HTK (Young et al., 2009).

HTK tools introduced the processing steps involved in building a sub-word based continuous speech recognizer. There are four main phases in these steps as follows:

Table 2.2: Four main phases in processing steps (Young et al., 2009).

Phases	Description
Step 1 Data Preparation	- A set of speech data files and associated transcriptions are needed to build a set of HMMs. It converted into an appropriate parametric format and converted the associated transcriptions of the speech data files into an appropriate format which consists of the required phone or word labels. - The example of tools in this phase is HSLAB, HCOPY, HLIST, HLED, HLSTATS, and HQUANT
Step 2 Training	- Define the topology required for each HMM by writing a prototype definition and allow HMMs to be built with any desired topology. HINIT and HREST tools can be used to train each sub-word if the training speech files are equipped the sub-word limits. - HINIT – computes an initial set of parameter value using the segmental k-means training procedure. - HREST – to re estimate the HMM parameters that were computed by HINIT and Baum-Welch re estimation procedure was used. - HEREST – to perform embedded training on the whole set of the HMMs simultaneously.
Step 3 Testing/Recognition	- The recognition phase involves HVITE which is to perform the Viterbi based speech recognition. This tool describes the allowable word sequences, how each word is pronounced and a set of HMMs as inputs. - It supports cross-word triphones, run with multiple tokens to generate lattices containing multiple hypotheses and configured to rescore lattices and perform forced alignments. - The other tool is HBUILD and HPARSE are provided to create the word networks.
Step 4 Analysis	- HRESULTS was used to evaluate the performance by comparing the recognition results with the correct reference transcriptions. - HRESULTS are compatible with those used by the US National Institute of Standards and Technology (NIST).

Figure 2.7 shows the overall tools in HTK processing stages that involves the data preparation, training, testing and analysis.

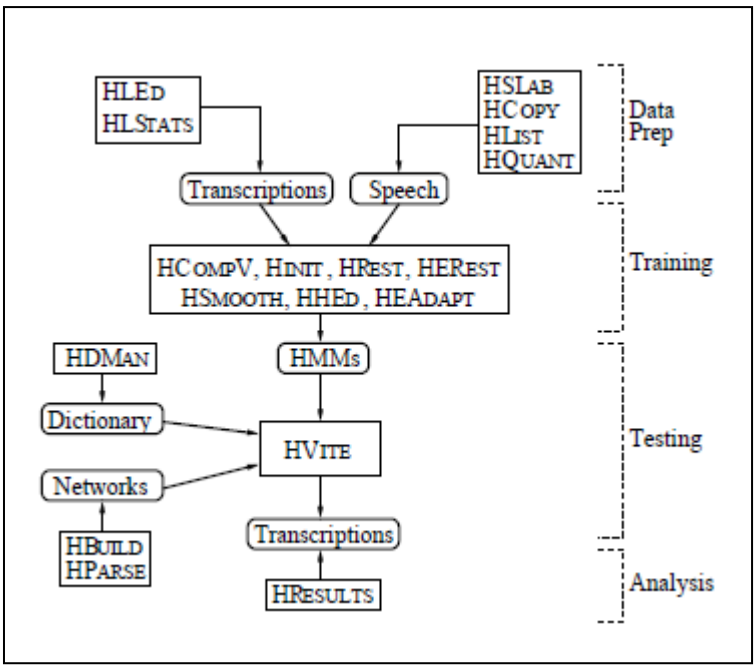

Figure 2.7: HTK processing stages (Young et al., 2009).

In the training process, the model was programme to define the topology and all of the HMM parameters which are means and variances of Gaussian distribution are ignored only with the exception of the transition probability (Young et al., 2009). Figure 2.8 shows the definition for simple left-right topology in HMM.

```
~h "hmm1"
<BeginHMM>
    <VecSize> 4 <MFCC>
    <NumStates> 5
    <State> 2
        <Mean> 4
            0.2 0.1 0.1 0.9
        <Variance> 4
            1.0 1.0 1.0 1.0
    <State> 3
        <Mean> 4
            0.4 0.9 0.2 0.1
        <Variance> 4
            1.0 2.0 2.0 0.5
    <State> 4
        <Mean> 4
            1.2 3.1 0.5 0.9
        <Variance> 4
            5.0 5.0 5.0 5.0
    <TransP> 5
        0.0 0.5 0.5 0.0 0.0
        0.0 0.4 0.4 0.2 0.0
        0.0 0.0 0.6 0.4 0.0
        0.0 0.0 0.0 0.7 0.3
        0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

Figure 2.8: Definition for simple left-right HMM (Young et al., 2009).

The simple left-right HMM is a continuous density HMM with in total 5 states and 3 of them are emitting. The symbol ~h indicates the following string is the name of a macro of type h which means that it is a HMM definition. The global features of the HMM are in the first line and these features will be the same for any system of HMMs. The global definitions indicate the observation vectors have 4 components (<VecSize> 4) and the MFCC coefficients (<MFCC>).

The number of states indicates in the next line. Then it followed by a definition for each emitting state *j*, each of which has a single mean vector μ_j configure by <Mean> and a diagonal variance vector Σ_j configure by <Variance>. Lastly, the definition ends with the transition matrix $\{a_{ij}\}$ introduced by keyword <TransP>.

2.7 Speech Encoding Process

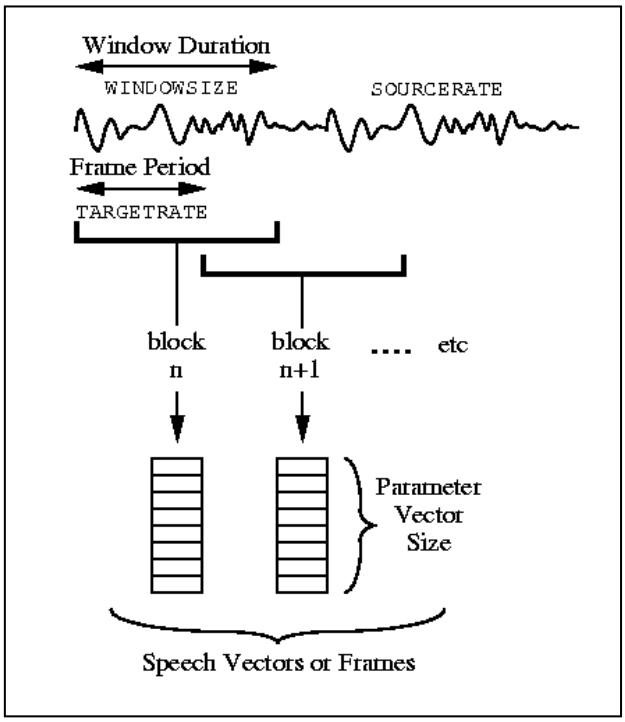

Figure 2.9: Speech Encoding Process (Young et al., 2009).

In general, HTK regards both waveform files and parameter files as being just sample sequences, the only difference being that in the former case the samples are 2-byte integers and in the latter they are multi-component vectors. Figure 2.9 shows the speech encoding process. The input file can determine the sample rate of the input waveform and it can be set using the configuration parameter SOURCERATE. Furthermore, TARGETRATE configure the period between each parameter vector and it determines the output sample rate (Young et al., 2009).

The segments of a waveform or known as window used to determine each parameter vector and its size is set by the configuration parameter WINDOWSIZE. The window size will be larger than the frame rate hence, successive windows overlap since the window size and frame rate are independent (Young et al., 2009).

2.8 Related Works

DeMarco et al. (2008) presented the comparison between three classifiers to identify gender which is context based classifier, pitch-shifting loopback classifier and baseline MFCC classifier. The researchers used TIMIT, ABI-1 and WSJCAM0 database for training and testing the gender classification. Pitch-shifting loopback classifier give a better performance of classification accuracy which is 95% from the pitch-based distortions of the speech signal and followed by context based classifier which is 93% - 94% accuracy. The Gaussian Mixture Models (GMMs) of the pitch value linked with each Mel Frequency Cepstral Coefficient (MFCC) vector included in the calculation of the centroid give poor results in the gender identification which is 30% classification accuracy since the value of the number of cluster centroids, k only increases from 2 to 16, but it drops when k > 16. Gender identification of female speech is more difficult when compare with male speakers because it is ambiguous and utterances sound in pitch.

Gaurav (et al., 2012) proposed the Hindi Continuous Speech Recognition System in primary education and it involved 29 context-dependent Hindi phonemes from the 43 distinct Hindi sentences. This project used MFCC to extract the original speech signals and these feature extractions are used to estimate the parameters of HMMs. Furthermore, the researchers construct prototype models by applying nine iterations of the standard Baum-Welch embedded training procedure to specify the overall characteristics and the topology of the HMM. Julius recognizer was used for decoding since it is language-independent decoding program, real-time, high speed and gives accurate recognition. The recognition accuracy of words for the male speech signals; 92.72% is more than females, which is 84.90% while the percentage of correct sentences for male is 76.84% and for the female is 60.28%.

A new method for gender identification proposed by (Meena et al., 2011) using three feature extraction which is Short Time Energy (STE), Zero Crossing Rate (ZCR) and Energy Entropy (EE) with Harvard-Haskins database. The role STE of the speech signal is to increase in energy signal while the ZCR is defined as to be the ratio of the number of time domain zero crossings occurred to the frame length. Then, the EE speech signal is defined as the sudden different changes in the energy level of a signal and the testing result for these three features are female speakers are high and continuous while male speakers is low and distributed. This is because the pitch value, which is depends on the frequency sound of the female is higher than male. These three feature values give as an input to two classifiers which is fuzzy logic and neural network. These two classifiers can give the percentage accuracy for the gender classification. The new method of three feature extraction gives high percentage accuracy, which is 65% compared with other method like fuzzy logic; 50% and neural network is 60%.

Deiv (et al., 2011) gave an overview of the approaches to automatic gender recognition can be separated into three classes. The first approach is about gender dependent features such as pitch because it has more information about gender identification of male and female speakers. The Cepstral features like Mel-Frequency Cepstral Coefficients (MFCCs) is another approach for Pattern Recognition since it yields a robust system in noise condition and it is widely used in other application like automatic speech recognition. Then, the last approach is the combination of features like pitch and MFCC that contribute to the enhancement of the Performance of the Gender Recognition.

The researchers found vowels and nasals are more useful in the study of gender identification because it is easy to identify in the speech signal and their spectra contain features that can distinguish the genders. The speaker for Hindi speech sound such as ten Hindi vowels and five nasals has been studied and this paper used the Euclidean distance for feature matching. The feature matching shows that the female gender recognition is better than male then HMM Tool Kit (HTK) was used to parameterize raw speech into sequence feature vectors. The identification rate for this project is between 97%-100% for mix two Hindi phonemes like vowel and nasal (Deiv et al., 2011).

Furthermore, Gaikwad (et al., 2012) explained some challenges for gender identification while using pitch period which are a clean signal can only get from a good estimate of the pitch period and the overlap of pitch values between male and female. Then, MFCC also has several limitations like MFCC captures speech information at a very short time scale and increase in computation complexity. It also has a problem of over training, so the performance of MFCC can affect by recording conditions and it gives inaccurate results. The isolated word, 12 MFCC combined with energy and pitch values for the feature extraction and Support Vector Machine (SVM) classifier was used in this study and the accuracy shows that the accuracy male speakers; 93.22% is higher than accuracy female speaker which is 86.90%.

Chen (et al., 2010) suggested Cepstral Peak Prominence (CPP) and Harmonic-to-Noise Ratio (HNR) are most useful in improving gender classification accuracy for children's speech since it is a bit difficult to identify due to the fundamental frequency (F_0) and formant frequencies are not easily distinguishable between boys and girls. This project was used an SVM classifier with a Radial Basis Function (RBF) kernel then the

result of the SVM were compared with traditional MFCC features. The difference between male and female speakers in CPP increases with the age while the F_0 values do not help differentiate between the genders. The classification accuracy for the girls; 72% - 96% is higher than boys group which is 69% - 93% due to the CPP measures that highly correlated with breathiness.

Zeng (et al., 2006) gave an overview the relative spectral (RASTA) uses filtering in the log domain of the power spectrum to compensate for the channel effects in recognizers, which is demonstrated to be more robust for noisy speech recognition, hence pitch and relative spectral perceptual linear predictive (RASTA-PLP) for the feature extraction in the gender classification was used to get the cleanest and degraded speech signals. The features from the extracted signals then set as input to the GMM classifier and the classification accuracy is above 97% for both genders such as female is 98.6% and male is 97.7%.

Gender classification for the children speech recognition has some difficulties and the accuracy slightly lower because the F_0 and formant frequencies of the children are higher than adults. Then, false pronunciation, disfluencies, breath noise is also another cause that it is hardly to classify them. Lastly, the duration of some vowels is longer and more variable compare with the adults (Tabrizi et al., 2011).

Tabrizi (et al., 2011) presented the solution for the variability of speech recognition for Persian children using adaptation techniques. This adaptation is referring to the Vocal Tract Length Normalization (VTLN) because it can be used for the feature extraction to control for the differing vocal tract length of speakers and indirectly it can reduce the variability. The nonlinear feature transformation approach was used as the

VTLN warps to create new acoustic models which are normalized based on specific speaker or group. The MFCC as a feature extraction and HMM as a recognizer with a standard left to right model also was used to get better result and the classification accuracy for boys is 71.36% and girls are 67.2%.

Gurgen (et al., 2006) proposed the MFCCs and neural networks (NN) classifier to identify speaker gender and phoneme-based features like a few sentences, vowels and consonants as a database. The NN as a classifier was used because it has ability to distinguish the gender when the unknown speech signal is applied to the network. The Window Neural Network (WNN) was used as trained in the identification of non-linguistic features using phoneme samples while NN to test the feature in unknown phoneme samples and a back propagation (BP) based recognition algorithm was used to train the networks.

The researchers investigated the effect of the number of coefficients by using the vowels and consonants. The vowel provided better identification accuracy; 97.4% compared with the consonant which is 67.5%. Then, the vowel /a/ and /i/ was used as a training data for gender identification. The single features affected the accuracy of the identification since the performance was improved which is for the vowel /a/ is 98.1% and vowel /i/ is 96.3% (Gurgen et al., 2006).

Fundamental frequency or pitch gives an important feature to identify male and female gender since it has average speaking fundamental frequency. The speaking fundamental frequency of men is between 100 and 146 Hz, whereas for females is between 188 and 221 Hz. Furthermore, resonance also can contribute to gender identification because it is a function of the supralaryngeal vocal tract. Vocal tract

resonances are frequently studied in terms of vowel formant frequencies. The vocal tract for men is about 15% longer than the female, hence the speech of men have a lower formant frequencies than the female (Gelfer and Mikos, 2005).

Fundamental frequency of 120 Hz and 240 Hz, first three formant frequencies and bandwidths of isolated vowels speech like /i/, /u/, and /ê/ was used to analyze the accuracy of gender identification by Gelfer and the coworker (Gelfer and Mikos, 2005). The researchers were used Fast Fourier Transform (FFT) as a pre-processing and Linear Predictive Coding (LPC) to derive the frequencies and amplitudes of the lowest three formants. From the paper, 120 Hz fundamental frequency can identify more male speakers, which is 84.2% while female speakers have high accuracy; 73.8% at 240 Hz fundamental frequency. Table 2.3 shows the summary of the previous studies.

Table 2.3: Summary of the previous studies

Author	**Language**	**Speech Sounds**	**Database**	**Feature Extraction**	**Recognizer**	**Accuracy**
DeMarco et al.	American English	TIMIT – 10 short utterances ABI-1 – 3 accent diagnostic passages WSJCAM0 – 5 utterances	TIMIT, ABI-1 and WSJCAM0. 100 females and 100 males	MFCC	Pitch-shifting loopback, context based, and GMM	Pitch shifting loopback –95%, context based– 93%-94% and GMM – 30%.
Gaurav et al.	Hindi	Hindi sentences	Text corpus - 12 females and 18 males	MFCC	HMM	Male – 92.72% Female – 84.90%
Meena et al.	American English	10 different sentences	Harvard-Haskins	Short Time Energy. Zero Crossing Rate and Energy Entropy	Fuzzy Logic and Neural Network	65%
Deiv et al.	Hindi	Vowels and Nasals	10 vowels with 10	MFCC	Euclidean distance and	94% - 100%

			females and 10 males 5 nasals with 5 males and 5 females		HMM	
Gaikwad et al.	Indian	Isolated words and natural continuous sentences	Collected from students of Department of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad - 8 males and 12 females	MFCC	SVM	Male – 93.22% Female – 86.90%
Chen et al.	English	20 utterances of this form with different vowels.	CID Database	Fundamental frequency and first three formant frequency, CPP and HNR.	SVM classifier with a Radial Basis Function kernel	Female: 72% – 96% Male: 69% – 93%
Zeng et al.	English, German, Japanese, and Italian.	Sentences	TIMIT with 182 males and 133 females	Parameters of pitch and relative spectral perceptual linear predictive (RASTA-PLP)	GMM	Female: 98.6% Male: 97.7%
Tabrizi et al.	Persian	Digit strings	2 different Persian speech database - 42 children	Vocal tract length Normalizatio n (VTLN) and Maximum Likelihood Linear Regression (MLLR) with MFCC	HMM	Boy: 71.36% Girl: 67.20%
Gurgen et al.	Australian English	Few sentences, vowels and consonants	200 rich sentences. 3 speakers	MFCC	Window Neural Network and Neural Network	Vowel /a/ - 98.1% and vowel /i/ - 96.3%
Gelfer	American	3 sustained	10 men and	Fundamental	LPC	120 Hz –

and Mikos	English	vowels - /i/, /u/, and /ê/	10 female	frequency and first three formant frequency		84.2%. 240 Hz – 73.8%

2.9 Chapter Conclusion

From the literature review, the speech signal has time frequency modulation that carries the formants and pitch intonation that can give information about linguistic, paralinguistic. Then, the production of the system involves four processes which is initiation, phonation, oral and nasal cavity and lastly articulators. The HMM provides effective algorithms for state and parameter approximation since it is a mathematical tool for modeling time series and perform dynamic time warping for signals.

In this work, a proposed method is used mel frequency cepstral coefficient (MFCC) as a feature extraction are widely applied in speech recognition for solving many various real-life problems. The MFCC has been particularly successful in speech recognition since it is based on the human peripheral auditory system. Furthermore, hidden Markov model is used to classify speech signal into two categories which is male and female since it is capable to statistically model the variability in speech. The hidden Markov model considered in this paper is based on HMM Tool Kit which is trained with Baum-Welch algorithm because it can compute maximum likelihood estimates and posterior mode estimates for the parameters of an HMM. Then, Viterbi algorithm was used as testing since it is a dynamic programming for finding the most likely sequence of hidden states.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter discusses about the data and methods used in developing the project consist of feature extraction and classification. To run the project, Hmm Tool Kit (HTK) was used as the tool for this method. Figure 3.1 shows the flow of the block diagram of the proposed method.

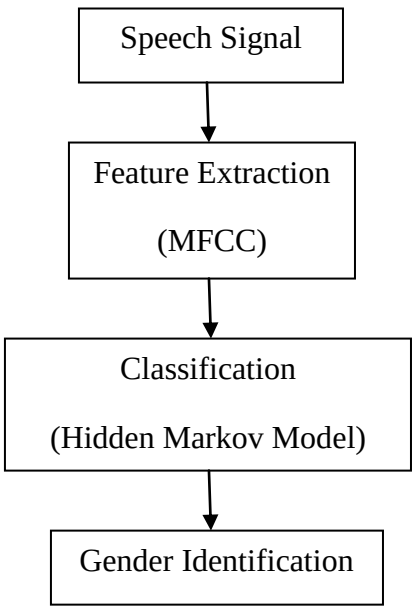

Figure 3.1: Block diagram of the proposed method

3.2 Speech Dataset

The database for this study consists of three hundred sixty normal Malaysian Malay children aged between 7 and 12 years old where each group divided into 30 males and 30 females. These children were asked to pronounce sustained vowel sounds such as /a/, /e/, /ə/, /i/, /o/, /u/. Hence, there are 2160 vowel sounds to be as a dataset for this study.

3.3 Feature Extraction

In this project, the method was proposed using 39 Mel Frequency Cepstral Coefficient (MFCC) to extract the original speech signals for each dataset. MFCCs are widely used features for automatic speech recognition systems to transform the speech waveform into a sequence of discrete acoustic vectors. To get the feature extraction, HMM Tool Kit was used for this MFCC. The Figure 3.2 shows the block diagram of the MFCC feature extraction.

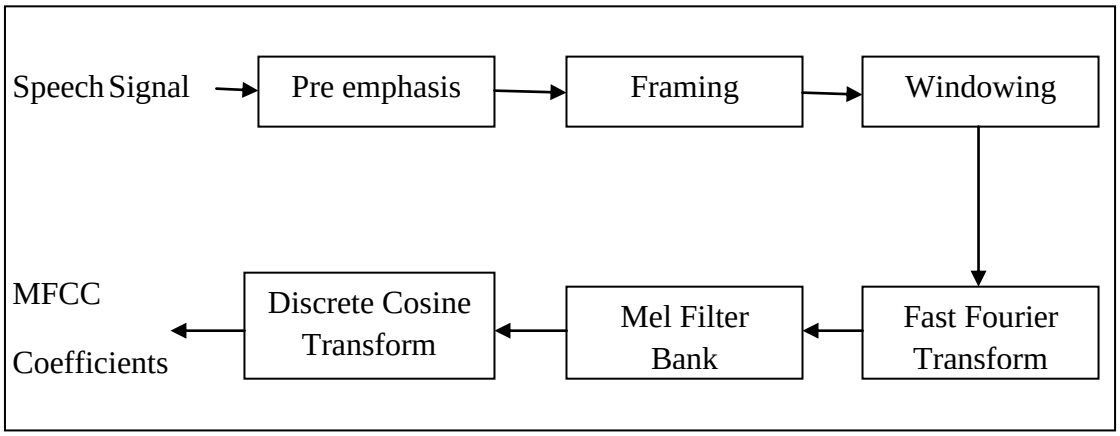

Figure 3.2: Block diagram of MFCC

The MFCC consist of six segments to get the coefficients from the speech signals and the parameters for this study, which are:

3.3.1 Pre emphasis

The pre emphasis functions as a filter which emphasizes higher frequencies of the original speech signal. This process will increase the energy of the signal at higher frequency. The first order pre emphasis of this study was applied using a coefficient of 0.97 because the pole of the filter is at zero hertz and the pre emphasis filter looks like a spectral tilt.

3.3.2 Framing

Framing process is to segment the digitized speech signals into small frames with a length within the range of 10 to 30 ms. The frame period for this study is 10 ms for each speech signal.

3.3.3 Windowing

The windowing as a window shape that is consider next frame in feature extraction process chaining and to determine each parameter vector. In this project, the Hamming window was used.

3.3.4 Fast Fourier Transform (FFT)

The FFT is used to convert each frame on N samples from time domain into frequency domain. The components of the magnitude spectrum from the analyzed signal are calculated (Deiv et al., 2011).

3.3.5 Mel Filter Bank

Compensation for non-linear perception of frequency is implemented by the bank of triangular band filters with the linear distribution of frequencies along with the so called mel-frequency range. Linear deployment of filters to mel-frequency axis results in a non-linear distribution for the standard frequency axis in hertz (Deiv et al., 2011). In this study, 20 channels of filter bank were used.

3.3.6 Discrete Cosine Transform (DCT)

This step is to calculate the logarithm of the output of filters. Then, the log mel spectrum is converted back to time domain (Deiv et al., 2011). Lastly, the result is obtained which is Mel Frequency Cepstral Coefficients. This DCT was applied to calculate the 12 Cepstral coefficients. Furthermore, the normalized log energy is added to the 12 MFCCs to form a 13-dimensional (static) vector. Then, it expanded to produce a 39-dimensional vector which includes static coefficients (MFCC 0 =13), delta coefficients (+13) and acceleration coefficients (+13).

In this study, there are two framing analysis was investigated which is a single frame analysis and multiple frame analysis. Single frame analysis of 10 ms, 15 ms, 20 ms, 25 ms, 30 ms, 35 ms, 40 ms, 45 ms, and 50 ms were used for 7 years old until 12 years old. For the multiple frame analysis, analysis frame sizes of 20 ms with a shift of 10 ms and analysis frame length of 10 ms with a shift of 10 ms were used for speech length between 50 ms and 150 ms also for the overall ages. The program was run for these two frame analysis to get the coefficients of the MFCC. Moreover, the coefficients of MFCC were obtained separately from the single age such as the coefficients for 7 years old until 12 years old to compare the result with the overall ages. These coefficients of

speech signal for 360 speakers were set as input in training and testing to the hidden Markov model.

3.4 Classification

For this project, hidden Markov model was used in the classification process using the HMM Tool Kit. The MFCC feature vectors that extracted from speech signals and their associated transcriptions was used to estimate the parameters of HMMs to get the performance of gender identification.

Three fold cross validation was used in this classification since training and testing sets are both large and each data point was used for both training and validation on each fold. The database was divided into three equal parts for training and testing the system. For each trial, two third of the data is taken for training and the remaining one third of the data was used for testing.

In this project, three fold cross validation was obtained and each fold or set assign data points to three parts which is has 360 speakers. For each part, it consists of 120 speakers so that these three parts are equal size. In set 1, part 1 and part 2 of the data which is 140 speakers is given for training. Then part 3 of the database which is 120 speakers was used for testing the trained system. In set 2, part 2 and part 3 of the database was used for training and part 1 of the database was used for testing. Lastly, for set 3, part 1 and part 3 of the database is taken for training and the system is tested with part 2 of the database. These three fold cross validation and classification was applied to the different frame analysis which is a single frame analysis and multiple frame analysis.

For the single and multiple frame analysis, two kinds of experiment of each frame analysis were done. Firstly, the database was trained with 360 speakers and second experiment is the database was trained with 60 speakers for each age, which are 30 males and 30 females. The simple left-right HMM with 5 states and 3 of them are emitting was used. Furthermore, Baum-Welch algorithm was used for training the database while the Viterbi algorithm was used for testing the trained data.

The performance in term of gender identification accuracy was obtained from each set with two kinds of experiment of single and multiple frame analysis. Then, the average of all the performance accuracy for all three sets was evaluated.

3.5 Chapter Conclusion

From this chapter, the data from the database and the methodology that have been used to feature extraction and classification were discussed in detail. In this method, trained with all speakers and 60 speakers for each age for single and multiple frame analysis have been used to determine the identification accuracy. Figure 3.3 shows the flow chart of the overall process in this project.

3.6 Flow Chart of the Proposed Method

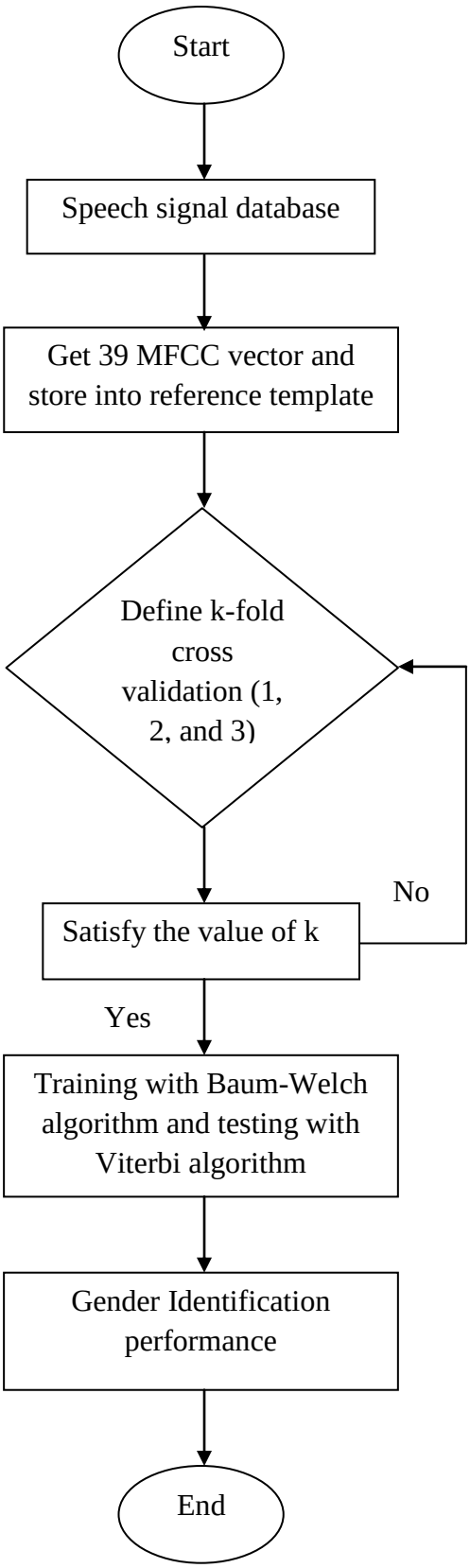

Figure 3.3: Flow chart of the proposed method.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 Introduction

This chapter discusses about the results obtained from the proposed method that was used for the sustained Malay vowel in speech recognition. Furthermore, the accuracy of the classification based on the setting parameters also discussed.

4.2 Experimental Result and Analysis

In this study, the database consists of six Malay vowel; a, e, ə, i, o, u for the 180 male children and 180 female children was used. The three fold cross validation was applied to train 240 speakers and the remaining 120 speakers of speech were used for testing. The Baum-Welch algorithm and the Viterbi algorithm were tested in the HMM training. The number of states for each word is five and was modeled using the left right HMM topologies. Then, the single frame analysis and multiple frame analysis were evaluated using HTK.

4.3 Single Frame Analysis

The classification accuracy of single frame analysis which is 10 ms, 15 ms, 20 ms, 25 ms, 30 ms, 35 ms, 40 ms, 45 ms, and 50 ms with the three fold cross validation was discussed.

Table 4.1 shows the different frame length, accuracy for the three sets of cross validation and the mean accuracy. At 30 ms frame analysis, it shows the highest mean accuracy which is 64.17% compared with the other frame analyses. For the set 1, the accuracy is 65%, set 2 is 63.61% and set 3 is 63.89%. Then, it was followed by 25 ms frame length; 63.94% for mean accuracy. The lowest mean accuracy for this single frame analysis is 60.88% at 20 ms. From the table, the accuracy after 50 ms was slightly decreased which is 62.78%.

Table 4.1: Accuracy for single frame analysis

Frame Length (ms)	**Accuracy (%)**			**Mean Accuracy (%)**
	Set 1	**Set 2**	**Set 3**	
10	64.17	62.08	63.47	63.24
15	64.17	59.72	63.47	62.45
20	60.42	60.28	61.94	60.88
25	65.56	63.61	62.64	63.94
30	65.00	63.61	63.89	64.17
35	65.28	64.72	60.97	63.66
40	64.72	62.64	64.31	63.89
45	66.39	62.92	62.36	63.89
50	64.72	60.42	63.19	62.78

Table 4.2 shows the gender accuracies for the 30 ms frame length since 30 ms gave high accuracy of the single frame analysis. The accuracy for the female is higher than male accuracy. Female accuracy is 67.78% and male accuracy is 60.56%. This is because female speakers have high pitch value than male speakers.

Table 4.2: Gender accuracies in 30 ms frame length

Gender	Accuracy (%)
Male	60.56
Female	67.78

4.4 Single Frame Analysis between Ages

For this part, the single frame analysis was separated the age of the speakers. The result of this part gave six tables of accuracy which is 6 years old until 12 years old and the mean accuracy of these ages for each frame length will be discussed.

Table 4.3 describes the accuracies between 7 years old and 12 years old for the gender identification in the single frame length. The average accuracies of 10 ms to 50 ms were shown. At 40 ms, it shows the highest average accuracy; 61.44% followed by 35 ms which is 60.93%. 10 ms and 15 ms frame length is the lowest average accuracy of the gender identification which is 58.10%. From the table, it shows that 12 years old is higher accuracy between the other ages while 8 and 9 years old is the lowest accuracy when regarding to the gender identification for the children. This situation happened because the children with 7 until 9 years old have some difficulties to identify gender such as disfluencies and breathe noise.

Table 4.3: Accuracies between ages and frame length

Frame Length (ms)	Gender Accuracy between Ages (%)						Average Accuracy (%)
	7 years	8 years	9 years	10 years	11 years	12 years	
10	63.89	50.56	54.44	61.67	58.33	59.72	58.10
15	63.89	50.56	54.44	61.67	58.33	59.72	58.10
20	65.83	49.44	50.28	64.72	57.78	67.50	59.26
25	67.50	48.89	51.67	64.44	60.83	64.45	59.63
30	64.72	51.66	52.22	67.22	59.44	66.67	60.32
35	68.89	52.50	50.28	65.83	57.50	70.56	60.93
40	68.33	55.28	51.67	65.83	58.05	69.45	61.44
45	68.06	53.61	50.56	66.39	55.83	66.94	60.23
50	65.83	53.89	52.22	63.06	57.78	68.89	60.28

Table 4.4 describes the accuracy of gender identification for 40 ms frame length since it gave highest accuracy for 7 years old until 12 years old. Regarding from the table, the accuracy for the female is higher than male. For the female speaker, 12 years old have high accuracy which is 77.78% while 8 years old is the lowest accuracy; 43.33%. For the male speakers, the gender accuracy only reached about 67.22%. This is proved from the previous research, (Meena et al., 2011) male speakers have low frequency compared with the female.

Table 4.4: Gender accuracy for 40 ms frame length

Years	Gender Accuracy (%)	
	Male	Female
7	64.44	72.22
8	67.22	43.33
9	49.44	53.89
10	63.89	67.78
11	55.00	61.11
12	61.11	77.78

4.5 Comparison Accuracy of Single Frame Analysis

Table 4.5 presents the comparison average accuracy of single frame analysis and single frame analysis between ages with 10 ms until 50 ms frame length. From the table, average accuracy of the single frame analysis higher than average accuracy of the single frame analysis between ages. Moreover, average accuracy slightly increased with the increase of frame length. Then, the average accuracy becomes low at 35 ms and onwards. Hence, at 30 ms frame length identified the gender accurately since it is higher than others.

Table 4.5: Comparison average accuracy between single frame analyses

Frame Length (ms)	**Average Accuracy (%)**	
	Single Frame Analysis	**Single Frame Analysis between ages**
10	63.24	58.10
15	62.45	58.10
20	60.88	59.26
25	63.94	59.63
30	64.17	60.32
35	63.66	60.93
40	63.89	61.44
45	63.89	60.23
50	62.78	60.28

Figure 4.1 shows the comparison average accuracy between single frame analyses to see more clearly that the single frame analysis is higher than the accuracy single frame analysis between ages.

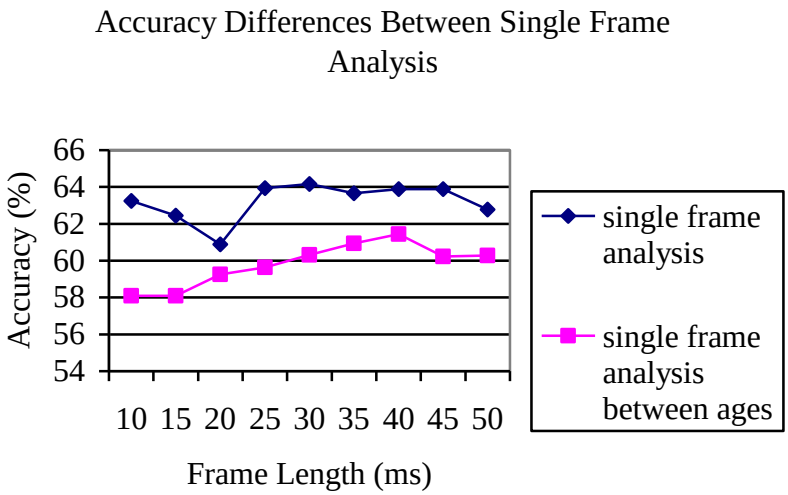

Figure 4.1: Comparison average accuracy between single frame analyses

Figure 4.2 describes the comparison of accuracy between 7 years old and 12 years for gender identification. From the figure, at 10 ms and 15 ms of frame length it showed that the accuracy is very low for 7 years old and 12 years old. Moreover, 12 years old have high accuracy which is 66.67% compared with 7 years old when it reached 30 ms of frame length. 12 years old have high accuracy because the pronunciation of vowel speech is clearer than 7 years old.

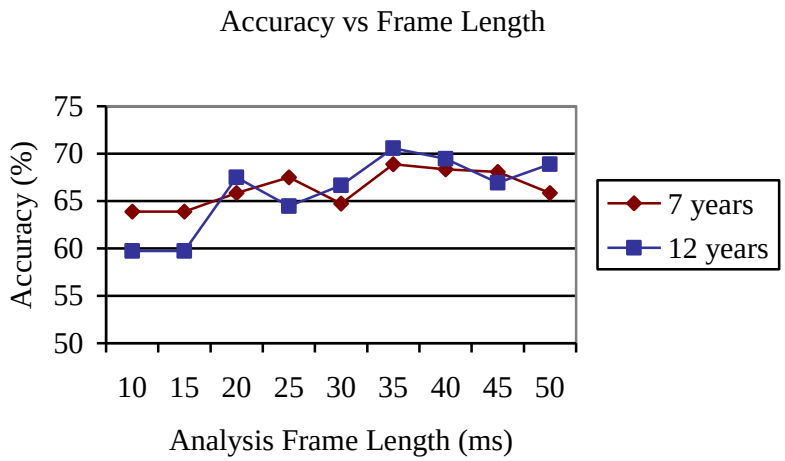

Figure 4.2: Comparison accuracy between 7 years old and 12 years old

4.6 Multiple Frame Analysis

The identification accuracy of multiple frame analysis at frame sizes of 20 ms with a shift of 10 ms and analysis frame length of 10 ms with a shift of 10 ms between 50 ms and 150 ms speech frame analysis was discussed.

Table 4.6 shows the identification accuracy for multiple frame analysis with different speech frame length such as 50 ms, 60 ms, 70 ms, 80 ms, 90 ms, 100 ms, 110 ms, 120 ms, 130 ms, and 140 ms. These speech frame length was divided into two analysis frame length which is 10 ms and 20 ms. Firstly, for the 10 ms analysis frame length, the highest accuracy is at 120 ms speech frame length which is 64.21%. Three set cross validation was obtained from the 120 ms speech frame length. For set 1, the accuracy is 64.31%, set 2 is 61.94% and accuracy for set 3 is 66.39%. The second highest accuracy is at 90 ms speech frame length which is 63.43% followed by 140 ms, 50 ms, 100 ms, 60 ms, 130 ms, 70 ms, 110 ms, and the lowest accuracy is at 80 ms speech frame length.

Furthermore, for the 20 ms analysis frame length, the highest accuracy is at 110 ms speech frame length which is 64.26%. The accuracy for set 1 is 65.69%, set 2 is 65% and accuracy for set 3 is 62.08%. The lowest accuracy for this 20 ms analysis frame length is at 130 ms speech frame analysis which is 61.53%. The accuracy for set 1 is 64.58%, set 2 is 61.53% and set 3 is 58.47%. From the table, 20 ms analysis frame length at 110 ms speech frame length is higher than 10 ms analysis frame length at 120 ms speech frame length.

Table 4.6: Classification accuracy of multiple frame analysis

Analysis Frame Length (AFL), ms	Speech Frame Length (SFL), ms	Accuracy (%)			Mean Accuracy (%)
		set 1	set 2	set 3	
10	50	63.47	61.39	61.94	62.27
	60	63.33	59.17	63.33	61.94
	70	62.50	63.06	59.17	61.58
	80	61.67	60.56	60.69	60.97
	90	64.17	61.39	64.72	63.43
	100	61.39	62.22	62.5	62.04
	110	60.97	60.56	63.06	61.53
	120	64.31	61.94	66.39	64.21
	130	63.89	58.89	62.64	61.81
	140	64.86	59.72	64.44	63.01
20	50	63.75	59.86	62.22	61.94
	60	62.92	63.06	63.19	63.06
	70	64.44	62.92	62.78	63.38
	80	63.19	60.56	62.78	62.18
	90	64.58	62.36	62.5	63.15
	100	65.00	63.19	63.47	63.89
	110	65.69	65.00	62.08	64.26
	120	63.47	61.67	60.97	62.04
	130	64.58	61.53	58.47	61.53
	140	65.14	61.81	61.39	62.78

Table 4.7 shows the gender accuracies for 10 ms and 20 ms analysis frame length with 120 ms and 110 ms speech frame length since these frame length gave highest accuracy among the others. Based on the table, the accuracy for the female speakers is higher than the accuracy for male speakers. The accuracy at 120 ms speech frame length for female speaker is 64.44% while accuracy for male is 63.98%. Then, accuracy at 110 ms speech frame length for female is 65.74% while accuracy for male is 62.78%.

Table 4.7: Gender accuracy for 10 ms and 20 ms frame length

AFL (ms)	SFL (ms)	Gender	Accuracy (%)
10	120	Male	63.98
		Female	64.44
20	110	Male	62.78
		Female	65.74

4.7 Multiple Frame Analysis between Ages

This section shows the result of accuracy for each age and the average accuracy for each speech frame length.

Table 4.8 describes the gender accuracy between ages, analysis frame length and speech frame length. For 10 ms analysis frame length, the accuracy is higher at 120 ms speech frame length which is 61.30%. The lowest accuracy at 10 ms analysis frame length is at 70 ms which is 58.43%. Moreover, for 20 ms analysis frame length, the highest accuracy is also at 120 ms speech frame length which is 61.71%. The second higher accuracy is at 110 ms, followed by 90 ms, 100 ms, 140 ms, 130 ms, 50 ms, 80 ms, 70 ms, and the lowest accuracy is at 60 ms speech frame length which is 58.94%.

Table 4.8: Accuracies between ages and frame length

Analysis Frame Length (AFL), ms	Speech Frame Length (SFL), ms	Gender Accuracy between Ages (%)						Average Accuracy (%)
		7 years	8 years	9 years	10 years	11 years	12 years	
10	50	63.61	53.33	46.67	63.34	56.11	64.72	57.96
	60	65.56	50.00	53.06	63.33	56.11	67.22	59.21
	70	65.55	55.00	44.17	62.78	56.94	66.11	58.43
	80	67.78	49.17	48.89	61.39	56.11	67.50	58.47
	90	66.95	54.17	48.61	59.72	55.84	66.94	58.71
	100	68.34	52.78	48.33	65.00	57.50	66.95	59.82
	110	68.33	47.78	48.61	61.94	57.50	68.89	58.84

	120	71.94	53.33	48.33	66.95	59.72	67.50	61.30
	130	70.00	51.94	49.17	61.67	58.89	68.33	60.00
	140	67.50	55.28	51.11	67.50	58.06	67.50	61.16
20	50	68.33	52.78	46.11	64.72	57.78	67.22	59.49
	60	67.78	52.22	53.06	63.33	54.44	62.78	58.94
	70	66.39	50.00	50.56	64.44	58.61	66.39	59.40
	80	70.28	53.06	50.83	62.50	56.39	63.61	59.45
	90	69.44	58.06	46.67	67.22	58.61	67.78	61.30
	100	70.00	55.28	51.11	64.44	56.39	65.55	60.46
	110	70.28	56.39	51.39	68.05	59.45	64.17	61.62
	120	71.94	56.39	50.83	66.39	60.28	64.44	61.71
	130	69.45	54.45	46.67	63.05	58.06	66.67	59.73
	140	66.39	55.00	46.94	65.84	58.61	66.11	59.82

Table 4.9 shows the accuracy for gender with 10 ms and 20 ms analysis frame length and 120 ms speech frame length. Based on the table, female speakers have high accuracy compared with a male speaker for each age. This accuracy for female is high because female speakers have high pitch compare to the male speakers (Meena et al., 2011). Furthermore, the result of accuracy is proven by (Chen et al., 2010) since the classification accuracy for the girls; 72% - 96% is higher than boys group which is 69% - 93%.

Table 4.9: Gender accuracy of multiple frame analysis

AFL and SFL	**Years**	**Gender Accuracy (%)**	
		Male	**Female**
AFL - 10 ms SFL - 120 ms	7	67.78	76.11
	8	63.33	43.33
	9	50.00	46.67
	10	62.22	71.67
	11	57.22	62.22
	12	57.22	77.78
AFL - 20 ms SFL - 120 ms	7	67.22	76.67
	8	71.67	41.11
	9	43.89	57.78
	10	66.67	66.11
	11	56.67	63.89
	12	60.56	68.33

4.8 Comparison Accuracy for Multiple Frame Analysis

Table 4.10 shows the comparison of average accuracy for multiple frame analysis with 10 ms and 20 ms analysis frame length at 50 ms until 140 ms speech frame length. At 10 ms of analysis frame length, the average accuracy of multiple frame analysis is higher than the average accuracy of multiple frame analysis between ages. The range of accuracy of multiple frame analysis is around 61% - 64% while the range of accuracy of multiple frame analysis between ages is 58% - 61%. At 20 ms of analysis frame length, the average accuracy of multiple frame analysis is also higher than the average accuracy of multiple frame analysis between ages. The range of average accuracy of multiple frame analysis is 61% - 64% while the range accuracy of multiple frame analysis between ages is 58% - 61%.

Table 4.10: Comparison average accuracy of multiple frame analysis

Analysis Frame Length (ms)	**Speech Frame Length (ms)**	**Average Accuracy (%)**	
		Multiple Frame Analysis	**Multiple Frame Analysis between Ages**
10	50	62.27	57.96
	60	61.94	59.21
	70	61.58	58.43
	80	60.97	58.47
	90	63.43	58.71
	100	62.04	59.82
	110	61.53	58.84
	120	64.21	61.30
	130	61.81	60.00
	140	63.01	61.16
20	50	61.94	59.49
	60	63.06	58.94
	70	63.38	59.40
	80	62.18	59.45
	90	63.15	61.30
	100	63.89	60.46
	110	64.26	61.62
	120	62.04	61.71

	130	61.53	59.73
	140	62.78	59.82

Figure 4.3 describes the comparison of average accuracy of the multiple frame analysis at 10 ms analysis frame length with 50 ms until 140 ms speech frame analysis. From the figure, multiple frame analysis is higher than the multiple frame analysis between ages at 50 ms until 140 ms speech frame length.

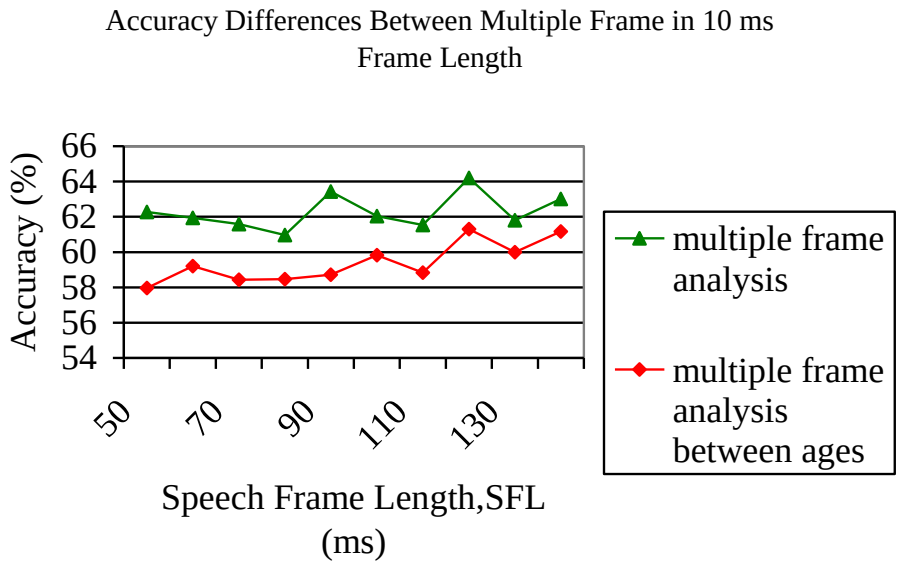

Figure 4.3: Comparison average accuracy of multiple frame analysis at 10 ms frame length

Figure 4.4 shows the comparison average accuracy for multiple frame analysis at 20 ms with the speech frame length. Based on the figure, multiple frame analysis shows higher accuracy compared with the multiple frame analysis between ages at a given speech frame length.

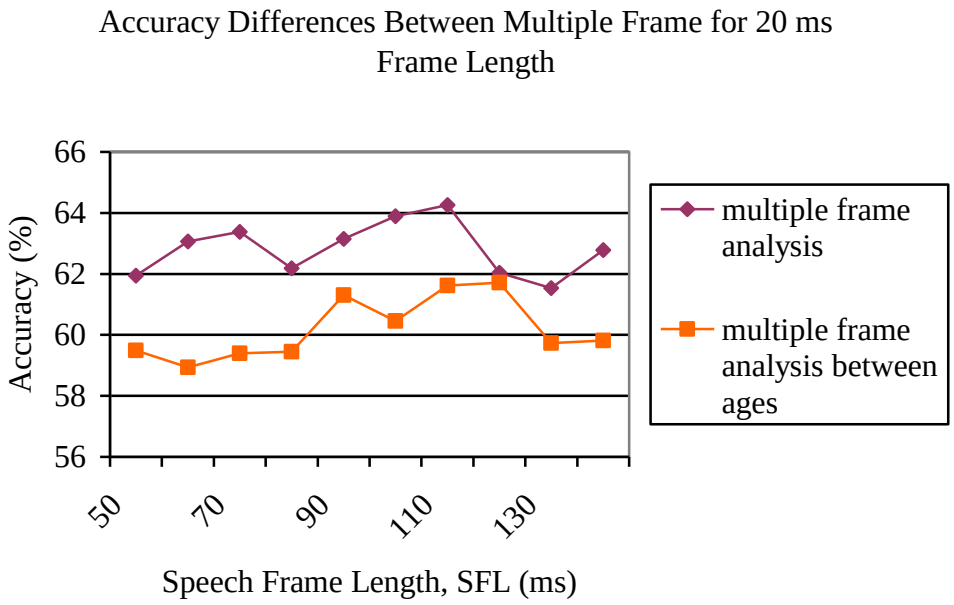

Figure 4.4: Comparison average accuracy of multiple frame analysis at 20 ms AFL

CHAPTER 5

CONCLUSION

5.1 Summary

In this project, 39-dimensional vector of MFCC was employed to the sustained Malay vowel speech database as a proposed method for the feature extraction. The original speech signal is firstly filter which emphasizes high frequency and it increased the energy of the signal. Then, segment speech samples of small frames and the hamming windowing was used to consider next block in the feature extraction process chain. The Fast Fourier Transform is to convert each frame of the samples from time domain to frequency domain and mel spectrum was obtained from the mel filter bank. Finally, Discrete Cosine Transform convert log mel spectrum to time domain and it results the coefficients of the MFCC.

The features have been obtained from the MFCC was used to get the gender identification accuracy using Hidden Markov model toolkit (HTK). Three fold cross validation was used in this study to get more accurate result. In this study, two frame analyses were evaluated. Firstly, single frame analysis with 10 ms, 15 ms, 20 ms, 25 ms, 30 ms, 35 ms, 40 ms, 45 ms, and 50 ms was used and secondary, multiple frame

analysis, analysis frame sizes of 20 ms with a shift of 10 ms and analysis frame length of 10 ms with a shift of 10 ms was used for speech length between 50 ms and 150 ms.

High classification accuracy was obtained from the single frame analysis and multiple frame analysis. For the single frame analysis, 30 ms of frame length gives the highest accuracy among the others which is 64.17%. This frame length; 30 ms achieves better accuracy for the gender identification since the low and the high frame length gives poor results. So, the classification accuracy depends on the low and the high frame length. Moreover, 40 ms frame length gives the highest accuracy when it considers the children age which is 61.44%.

For the multiple frame analysis, there are two analysis frame length was evaluated which is 10 ms and 20 ms. At 10 ms, 120 ms of speech frame length give highest classification accuracy which is 64.21% and for the 20 ms of analysis frame length, 110 ms speech frame length gives higher accuracy, 64.26%. Furthermore, the identification accuracy between ages was obtained. For both analysis frame length, 120 ms of speech frame length gives the highest accuracy which is 61.30% and 61.71%.

For the single frame analysis, the accuracy of female children was 67.78% while accuracy for male children is 60.56%. For the multiple frame analysis, the accuracy for female children is 65.74% and male children are 62.78%. In this study, female speakers give highest identification accuracy since female have high pitch or fundamental frequency compare with the male speakers, which is have low pitch. Moreover, vowels are most important in study gender identification since it is easy to identify in the speech signal and their features can distinguish the genders. The objective of this project is achieved. The limitation of this study is children have some difficulties to identify

gender due to the disfluencies of speech, false pronunciation and breathe noise. Then, MFCC also has some limitations such as MFCC captures speech information at a very short time scale, increase in computation complexity and has a problem of over training that can give poor results.

5.2 Recommendation of Future Project

For the recommendation, there are some other techniques that can be use to get better results, which is:

- The proposed method can be applied to another type of speech identification problems such as to identify the age of the speakers and the Malay vowel recognition.
- The other method of feature extraction and classifier can be used for the identification accuracy. For example, the adaptation techniques like Vocal Tract Length Normalization (VTLN) and Linear Predictive Coding (LPC) is for feature extraction method. For the classifier, Support Vector Machine (SVM), Fuzzy Logic and Neural Network (NN) can be used.

REFERENCES

Abdulla, W., & Kasabov, N. (1999). The Concepts of Hidden Markov Model in Speech Recognition. *99*(9).

Arica, N., & Yarman-Vural, F. T. (1999, June 20-23). *A New HMM Topology for Shape Recognition.* Paper presented at the Proc. of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, Turkey.

Bell, T. E. (2006). The Quiet Genius: Andrew J. Viterbi. *The Bent of Tau Beta Pi*, 17-21.

Blunsom, P. (2004). Hidden Markov Models. 1-7.

Bouman, P. C. A. (2009). Speech Processing (Part 1). In D. Williamson (Ed.).

Chen, G., Feng, X., Shue, Y.-L., & Alwan, A. (2010). *On Using Voice Source Measures in Automatic Gender Classification of Children's Speech*. Paper presented at the Eleventh Annual Conference of the International Speech Communication Association.

Churbanov, A., & Winters-Hilt, S. (2008). Implementing EM and Viterbi algorithms for Hidden Markov Model in linear memory. *BMC Bioinformatics, 9*(224), 1-23.

Deiv, D. S., Gaurav, & Bhattacharya, M. (2011). Automatic Gender Identification for Hindi Speech Recognition. *International Journal of Computer Applications, 31*(5), 1-8.

DeMarco, A., & Cox, S. J. (2008). An Accurate and Robust Gender Identification Algorithm. *Journal of Neuroscience Methods, 172*(1), 122-130.

E.Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A Maximization Technique Occuring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Analysis of Mathematical Statistics, 41*(1), 164-171.

Elmezain, M., Al-Hamdi, A., Rashid, O., & Michaelis, B. (2009). *Posture and Gesture Recognition for Human Computer Interaction*. Retrieved 19 November 2012.

Fink, G. A. (2008). *Markov Models for Pattern Recognition: From Theory to Applications*. Retrieved 24 November 2012.

Gaikwad, S., Gawali, B., & S.C., M. (2012). Gender Identification Using SVM with Combination of MFCC. *Advances in Computational Research, 4*(1), 69-73.

Gaurav, Deiv, D. S., Sharma, G. K., & Bhattacharya, M. (2012). Development of Application Specific Continuous Speech Recognition System in Hindi. *Journal of Signal and Information Processing, 3,* 394-401.

Gelfer, M. P., & Mikos, V. A. (2005). The Relative Contributions of Speaking Fundamental Frequency and Formant Frequencies to Gender Identification Based on Isolated Vowels. *Journal of Voice, 19*(4), 544-554.

Giegerich, H. J. (1992). *English Phonology: An introduction*. Retrieved 24 November 2012.

Gomathy, M., Meena, K., & Subramaniam, K. R. (2011). Classification of speech signal based on gender: a hybrid approach using neuro-fuzzy systems. *International Journal Speech Technology, 14*(4), 377-391.

Gotoh, Y., Hochberg, M. M., & Silverman, H. F. (1998). Efficient Training Algorithms for HMM's Using Incremental Estimation. *IEEE Transactions on Speech and Audio Processing, 6*(6), 539-548.

Gurgen, F. S., Fan, T., & Vonwiller, J. (2006). *On The Analysis of Phoneme-based Features for Gender Identification with Neural Networks.* University of Sydney NSW.

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012). *Speech Recognition using MFCC*. Paper presented at the International Conference on Computer Graphics, Simulation and Modeling.

Kouemou, G. L. (2011). *Hidden Markov Models, Theory and Applications*.

Meena, K., Subramaniam, K., & Gomathy, M. (2011). Gender Classification in Speech Recognition using Fuzzy Logic and Neural Network. *The International Arab Journal of Information Technology*.

Paul, D. B. (1990). Speech Recognition Using Hidden Markov Models. *The Lincoln Laboratory Journal, 3*(1), 41-62.

Phoophuangpairoj, R., Phongsuphap, S., & Tangwongsan, S. (2009). *Gender Identification from Thai Speech Signal Using a Neural Network.* Mahidol University, Nakhonpathom, Thailand.

Raymond C. Vasko, J., El-Jaroudi, A., & Boston, J. R. (1996, 7-10 May). *An Algorithm To Determine Hidden Markov Model Topology.* Paper presented at the Acoustics, Speech, and Signal Processing.

Rosdi, F. (2008). *Isolated Malay Speech Recognition Using Hidden Markov Models.* Unpublished Signal Processing, University of Malaya, Kuala Lumpur.

Shu, H., Hetherington, I. L., & Glass, J. (2003). Baum-Welch Training for Segment-based Speech Recognition. *Computer Science and Artificial Intelligence Laboratory*, 43-48.

Siraj, Yaacob, Y., S. A. M., F., S., P., P. M., & Nazri, A. (2010). *Improved Malay Vowel Feature Extraction Method Based on First and Second Formants*. Paper presented at the Second International Conference on Computational Intelligence, Modelling and Simulation.

Tabrizi, G. T., Setayeshi, S., & Kakhki, M. M. (2011). HMM-Based Recognition and Adaptation of Persian Children's Speech. *Contemporary Engineering Sciences, 4*(5), 221 – 228.

Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International Journal of Emerging Technologies, 1*(1), 19-22.

Vaseghi, P. S. V. (2006). Chapter 13 Speech Processing. In *Advanced Digital Signal Processing and Noise Reduction*. London: J. Wiley.

Wang, X. (1994). Durationally Constrained Training of HMM without Explicit State Durational PDF. *Institute of Phonetic Sciences, 18*, 111-130.

Wang, X., Xia, M., Cai, H., Gao, Y., & Cattani, C. (2012). Hidden Markov Models Based Dynamic Hand Gesture Recognition. *Mathematical Problems in Engineering*, 1-11.

Wiggers, I. P., & Rothkrantz, D. d. L. J. M. (2003). *Automatic Speech Recognition Using Hidden Markov Models.* Data and Knowledge System Group.

Young, S. (1994). *The HTK Hidden Markov Model Toolkit: Design and Philosophy*. Cambridge: Cambridge University Engineering Department.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., et al. (2009). *The HTK Book (for HTK Version 3.4)*: Cambridge University Engineering Department

Yusof, S. A. M., P.M.Raj, & Yaacob, S. (2007, June 17-19). *Speech Recognition Application based on Malaysian spoken vowels using Autoregressive Model of*

the Vocal Tract. Paper presented at the Proceeding of the International Conference on Electrical Engineering and Informatics, Institute Technology Bandung, Indonesia.

Zeng, Y.-M., Wu, Z.-Y., Falk, T., & Chan, W.-Y. (2006, 13-16 August). *Robust GMM based Gender Classification Using Pitch and RASTA-PLP Parameters of Speech.* Paper presented at the Fifth International Conference on Machine Learning and Cybernetics.