American Journal of Applied Sciences

Original Research Paper

Statistical Parametric Evaluation on New Corpus Design for Malay Speech Articulation Disorder Early Diagnosis

Mohd Nizam Mazenan, Tan Tian Swee, Tan Hui Ru and Azran Azhim

Medical Implant Technology Group (MediTEG),
Cardiovascular Engineering Center, Material Manufacturing Research Alliance (MMRA),
Faculty of Biosciences and Medical Engineering, Universiti Teknologi Malaysia, Malaysia

Article history
Received: 10-10-2014
Revised: 20-11-2014
Accepted: 14-04-2015

Corresponding Author:
Tan Tian Swee
Medical Implant Technology Group (MediTEG),
Cardiovascular Engineering Center, Material Manufacturing Research Alliance (MMRA),
Faculty of Biosciences and Medical Engineering,
Universiti Teknologi Malaysia, Malaysia
Email: tantswee@biomedical.utm.my

Abstract: Speech-to-Text or always been known as speech recognition plays an important role nowadays especially in medical area specifically in speech impairment. In this study, a Malay language speech-to-Text system was been designed by using Hidden Markov Model (HMM) as a statistical engine with emphasizing the way of Malay speech corpus design specifically for Malay articulation speech disorder. This study also describes and tests the correct number of state to analyze the changes in the performance of current Malay speech recognition in term of recognition accuracy. Statistical parametric representation method was utilized in this study and the Malay corpus database was constructed to be balanced with all the phonetic placed and manner of articulation sample appeared in Malay speech articulation therapy. The results were achieved by conducting few experiments by collecting sample from 80 patient speakers (child and adult) and contain for almost 30,720 sample training data.

Keywords: Speech-to-Text, Hidden Markov Model, Feature Extraction, Malay Corpus

Introduction

Speech is a type of communication method between people (Yong and Swee, 2014b), which people design the content of it to deliver their message. The process of speech sound production is called articulation, it is caused by the movement of lips, tongue, velum and jaws to shape the flow of air into sounds (Thomas and Carmack, 1990). Person who is having difficulties in articulate spoken language correctly are facing articulation disorder problem. According to (Ting *et al*., 2003), the cause of articulation disorder can be organic or functional. Example of organic causes including anatomical, sensory impairment or motor while example of functional causes have several etiologies (Van Riper and Erikson, 1996).

Early screening or diagnosis of articulation disorder and identify the type of the articulation disorder can be beneficial in the next stage which is treatment stage. Currently in Malaysia hospital and speech center, the method of diagnosis for speech disorder is still using the traditional method which is manual diagnosis (Mohd Nizam and Tan, 2012). Manual diagnosis required a lot of involvement of Speech Language Pathologist (SLP) for each session. In current situation, the ratio of speech language pathologist to speech disorder patient is 1:500 in Hospital Sultanah Aminah (HSA) while the ideal ratio number is 1:50 according to World Health Organization (WHO) (Mohd Nizam and Tan, 2012). According to SLP in speech center, Jamilah (2014), a SLP might needs to work in the condition of 1:4 (ratio of SLP to speech disorder patient) and this can reach 1:8 in hospital while the ideal case should be 1:1 ratio. This shows that more effort is needed in this area to decrease the distance between real situation and ideal situation.

Computerized system may comes into this part and assist SLP in early screening diagnosis process to make it more efficient and hope to cut down time consume and raise accuracy. However, the right combination of words need to be designed in the system to address language and speech disorder where it concern with disorders of human communication (Tan *et al.*, 2007) that specifically focus on articulation disorder.

This chapter has introduced the background and importance of the study followed by chapter 2, provides an introduction to the process of voice produced, current research conducted and covered some information about Malay Language. Chapter 3 will introduce the methodology that we used including Malay articulation

corpus design and structure of Hidden Markov Model (HMM) with number of state adjustment and the chapters following are covered for speech recognizer architecture, HMM likelihood evaluation, result and conclusion.

Literature Review

Nowadays, Automatic Speech Recognition (ASR) has gained more and more attention in Malaysia medical field for its possible usage in diagnosis and treatment area especially for speech disorder case. A brief introduction for production of voice, method and Malay language will be presented in this section.

Production of Voice

In order to produce voice, complex planning and coordination of mouth and tongue movements is required (Dronkers, 1996a). It involves muscles along the vocal tract with specific and quick timing. Airflow formed in lungs and flowed through the vocal folds in larynx and vibrated in mouth cavity area including soft palate, hard palate, lips, jaw and tongue is needed in voice production. In detail, when we are planning to speak, our brain will send a signal to larynx muscle for closing the vocal cord (it is remain open for smooth breathing when we are not speaking). After the vocal cord is closed, the flow of air coming up from our lungs when we are speaking encounter the bloackage of vocal cord and vocal cord react open and close repeatedly, which become rapid vibration. This produces the sound waves in the air inside our vocal cord area and we known this as our basic tone of voice (Voice Production, 2009). This human anatomical structure shown in Fig. 1.

Manner of articulation is another important part in the process of speaking after the production of voice. When the moving air pass through vocal cord and reach oral cavity above tongue, the method of people modifying their speech organs and control the flow of air take charge in pronuncing the correct words. Manner of articulation refers to these methods of controlling speech organs such as stop, fricatives, plosives, affricate, nasal consonant and etc. For stop, it is created by trap the air flow in the mouth by different speech organs. For fricative, it is similar to hissing sound, due to a tight contriction made, so the air passing by formed turbulance and creating sound. For nasal consonant, the air has been directed to be released from nasal cavity through nose. Another type of articulation manner is taps and flaps, which is describing the act of make a rapid brush around alveolar ridge area (Hayes, 2011). Thus, human speech production requires complex planning and coordination of mouth and tongue movements (Dronkers, 1996b).

Statistical Modelling Technique

To access this problem where it's related to recognizing the smallest lexical unit in sound production, the suggested method is by using probabilistic pattern matching technique of Hidden Markov Model (HMM). The ability of statistically modeling the speech variability has been the reason of HMM widely been used in speech recognition field. It is flexible and can be utilized for many other applications especially in medical field that concern on early screening diagnosis of articulation speech disorder which are related to the stochastic modeling tasks. Stochastic task which in simple understanding to deal with the uncertainty and incompleteness of the input variable. The process of HMM will begin by the creating the stochastic models from known utterance. Then the comparison of the probability of that unknown utterance was generated by each model (Paul, 1990). However, at the hidden level, the most common speech recognition system will represent the phonetic information of the underlying speech signal. Some has achieved success by using this methodology but the approach somehow does not explicitly incorporate knowledge of certain aspect of human speech production (Aymen *et al*., 2011).

Another problem arises when developing speech recognition system is regarding the corpus design and pattern matching phase. The detail will be described in more specific in the next chapter of methodology.

Malay Language Overview

Malay language or Bahasa Melayu is the official language in Malaysia, Indonesia and Brunei (Yong and Swee, 2014a). There are about 500 million of Malay speakers existed in the world (Noraini and Kamaruzaman, 2008). With this amount of speakers spreaded in different places at least 3 different countries, it is having several dialect such as Kelantan Malay, Ulu Muar Malay, Langkawi Malay and others (Teoh, 1994). 'Standard Malay' or 'Bahasa Baku' refers to the formal way of Malay language including the language form and usage of it (Sariyan, 1988), is actually based on the Johor-Riau Malay dialect.

According to Noraini and Kamaruzaman (2008), Malay is a phonetic language which written in Roman characters. There are a total of 6 main vowels and 29 of consonants in Standard Malay. In the book of 'The Malay Sound System (1980) has mentioned that there are 9 vocal sound found from the 6 main vowels just now which are /i, e, a, ə, ɥ, o/ and /ɛ, ɜ, ɔ/ of allophones in addition.

In Malay System of Transliteration, Malay language is represent by the roman character with some notes on the pronunciation of malay words (Hugh and Frank, 1894). It is divided into 3 main parts which are vowels, diphthongs and consonants. For example word starting from "A", the Malays pronunce â (long) like the vowel sound a in "calm"; â (medium) is pronunced like the vowel sound "come" and â (short) is pronunced a little more shortly than the vowel sound in the English word "but". Thus, Malay language is quite similar with English pronunciation.

DOI: 10.3844/ajassp.2015.452.462

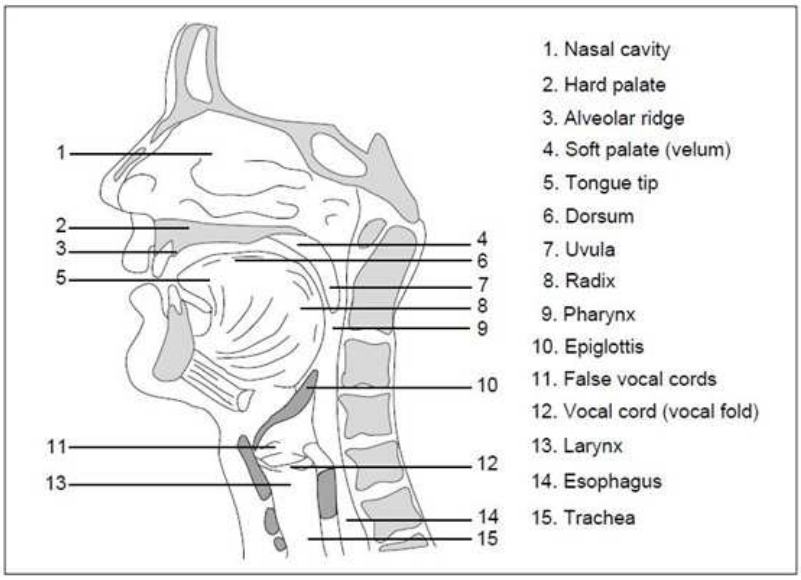

Fig. 1. Human anatomical structures which involved in phonation process (Karjalainen, 2008)

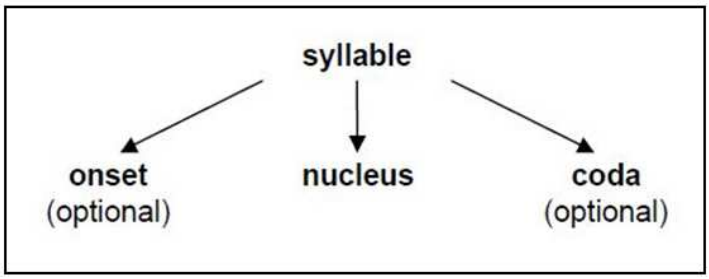

Fig. 2. Illustration of syllable structure (Laver, 1994)

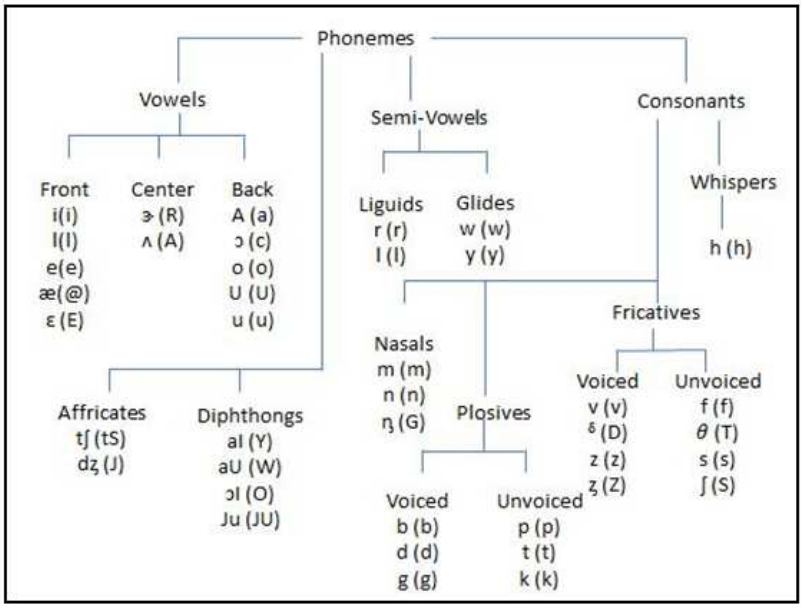

Fig. 3. Types of phonemes (Hina, 2012)

Noraini and Kamaruzaman (2008) said, this language is part of Austronesian lanaguage and it is agglutinative usually, words in Bahasa Melayu are formed by joining syllybles. Syllables are commonly using peak of sonority within an utterance in attempting such definitions (Ladefoged, 1975). A percise definition or structure of a syllybus presented by Laver (1994) is illustrated in Fig. 2.

In the figure shows a syllable formed from a central vowel or vowel-like consonant, which is nucleus. Sometimes, the nucleus will optionally pair up with one or more consonants which formed onset and coda.

The smallest unit of speech is the phoneme (Swee and Salleh, 2008), a family of counds that are close enough in perceptual qualities to be distinguished from other phonemes that is capable of convenying a distinction inmeaning, as the ‘m’ of ‘mat’ and the ‘b’ of ‘bat’ in English (Tischer, 2009). In Malay Language, there are about 24 pure phonemes and 6 borrowed phonemes. Malay Phonemes Features follows the standard of International Phonetic Association (IPA) (Asmah, 1983). Figure 3 shows the possible types of phoneme that can be combined together which presented by (Hina, 2012).

To be clear, the phonology of the speech is also related to diphthongs and vowel combinations. Diphthongs are types of vowels where two vowels sounds are connected in a continuous, gliding motion or sometimes refer as gliding vowels. A total of 3 diphthongs existed in Malay language which is [ai],[au] and [oi] (Raminah and Rahim, 1987). The gliding movement in Malay language itself contains seven combination of vowels such as ‘ai’, ‘au’, ‘ia’, ‘io’, ‘iu’, ‘ua’ and ‘ui’ where it combine in two vowels. However, the vowel combinations of ‘au’ and ‘ai’ are different from diphthong ‘au’ and ‘ai’ in the way they pronounce.

Methodology

Malay Articulation Disorder Corpus

In the process of computerizing the diagnosis process, computer plays a major role in it. However, there is a difficulty in applying computer system in this process which is the current technology of speech signal processing is unable to identify and differentiate the words which is pronounced similarly. For example, ‘Start’ and ‘Stop’ in English is difficult to be differentiated in current speech signal processing technology as they are having the same initial pronunciation of ‘St’. As an addition, both consonant S and T are in the same group of alveolar in International Phonetic Alphabet (IPA) table (Nur Hana, 2007). Looking into the raw speech wave of these two words, it is having similar wave pattern in the initial part (‘St-’ sound) as shown in Fig. 4. As in Malay word case, similar to ‘Start’ and ‘Stop’ words, ‘Lampu’ (Light) and ‘Lembu’ (Cow) are the words which will create confusion in the speech signal processing system of computer due to its similar place and manner of articulation in IPA table, which is plosive bilabial.

(Consonant P and B) (Nur Hana, 2007). Thus, we take this into consideration in the process of simplifying the existed sample word corpus content to be more compatible to computer system. The sample words of the lexicon database were gathered from Hospital Sultanah Aminah by Mohd Nizam and Tan (2012). A total of 128 different Malay words are collected based on the categories of consonants mainly in alveolar and plosives type.

The next factor in simplification process is to make sure each chosen word will have maximum function to test the articulation process of a person, in other word, to takle particular articulation manner of a person. For instance, ‘komputer’(computer), ‘sembilan’(nine) and ‘televisyen’ (television), all these three words can test for 4 places. Take ‘sembilan’(nine) as example, this word can be tested for ‘sem-‘ (S consonant) as initial consonant articulation, ‘-bi-’(B consonant) and ‘-la-’ (L consonant) for middle consonant articulation as well as ‘-an’ (N consonant) for end consonant articulation. All the consonant testing are aiming from alveolar and plosives consonants.

The following Table 1-3 consist the target consonant corpus after simplification for this research. According to (Donald and Katherine, 1996), the most common misarticulated sound is consonant sounds comparing to vowel sounds in English language case. Thus, our corpus is focusing in consonant sound. Total words in here is 64 words.

All of the target words in this corpus are chosen with the consideration of the consonant position for instance initial, middle and end position of a word. In Fig. 5, it is clearly illustrated that consonant R can exist in initial position which is Rumah (House), middle position which is Jari (Finger) and end position which is Motor (Motor).

An example words in the corpus, sotong (Squid) can be divided into 3 places to be tested. For initial place, consonant S can be tested and consonant T in the middle position and consonant G in the end position as illustrated in Fig. 6.

Architecture of Speech Recognizer Engine

The core structure of this experiment is based on basics and the state of the art for ASR architecture which consists of front-end processing and back-end processing of the speech sample signaling. Figure 7 shows the ASR architecture in this research. The process of front-end processing is starting by capturing the sound wave of speech sample by using standard microphone at 16 kHz and 16-bit resolution format. Then the process of speech signal processing will be done by converting the speech signal of the waveform format into parametric representation by using FE by considering 12 Mel-Frequency Cepstral Coefficient (MFCC) setup (Davis and Mermelstein, 1980). MFCC had been selected because previous research shows that, MFCC has characteristics of the human auditory system and commonly used in the ASR (Axelsson and Björhäll, 2003).

DOI: 10.3844/ajassp.2015.452.462

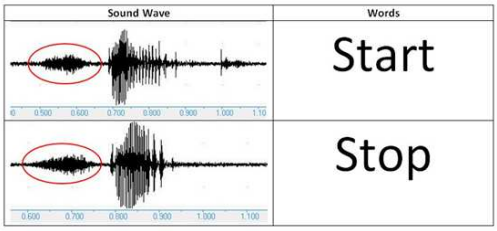

Fig. 4. Comparison between 'Start' and 'Stop' sound wave

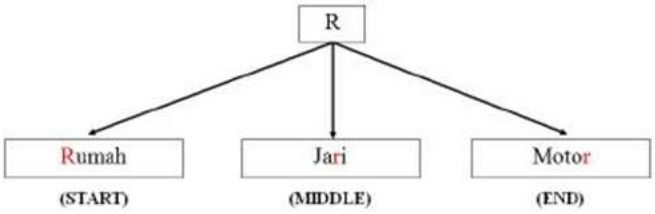

Fig. 5. Consonant arrangement for Malay articulation diagnosis and training (Mohd Nizam and Tan, 2012)

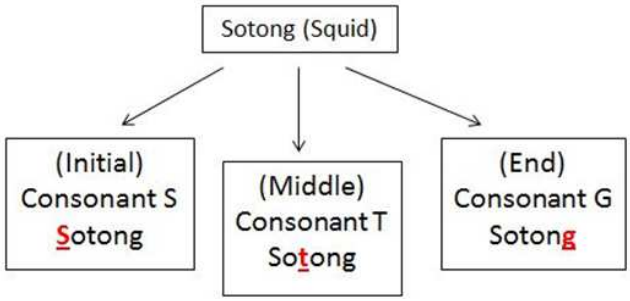

Fig. 6. Example of Sotong (Squid) testing consonant in several parts

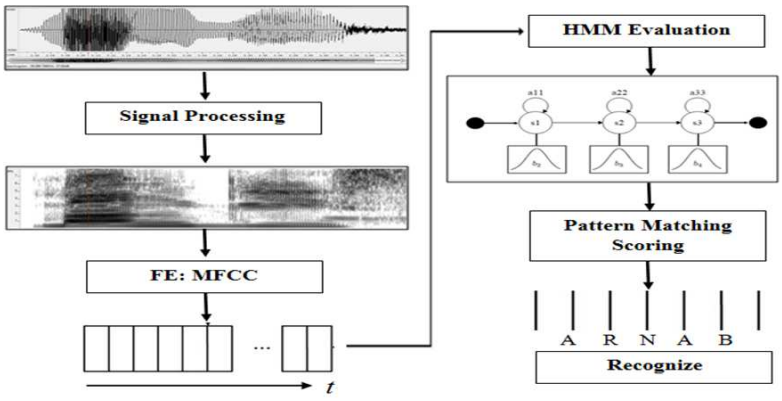

Fig. 7. State of the art speech recognizer architecture

Table 1. Target words for 5 Malay pure alveolar consonant

Consonant	Word target	Total
L	langsat, lapan, lembu, lima, lobak, lutut	6
N	nanas, nangka, nasi, nyamuk	4
R	rambut, rebung, ringgit, roti, rumah	5
S	sabun, seluar, sembilan, sepuluh, siput, sotong, syampu	7
Z	zip, zoo	2
	Total	24

Table 2. Target word for 4 Malay pure Plosives consonant

Consonant	Word target	Total
B	baju, baldi, bantal, biskut, botol	5
G	gajah, garfu, gelas, gigi	4
K	kacang, kad, kambing, katil, kek, kelapa, kerang, kijang, kipas, komputer, kotak, kucing	12
P	payung, pensel, pinggan, pisau, puding	5
	Total	26

Table 3. Target word for 2 Malay Alveolar and Plosives sharing consonant

Consonant	Word target	Total
D	daun, delima, dua, duit, duku, duriam	6
T	tangan, televisyen, telinga, tiga, tikus, tisu, topi, tujuh	8
	Total	14

The next phase reside on back-end processing. As been explain in previous chapter, HMM are dealing with stochastic modeling tasks. The general HMM will covers two stochastic processes that are the transition process between the states and feature vectors generating process by individual states (Zimmermann and Zimmermann Jr, 2002). The acoustic HMM training mechanism will be implemented to generate set of models to represent the observed acoustic vectors of the sample. In typical speech recognition system, The word-based pronounciation dictionary will be created and be used to describe each HMM acoustic model such as phoneme or syllable are mapped to form a word for both training and decoding purpose.

For this experiment, two sets of pronunciation dictionary were created which are training pronunciation dictionary and the decoding pronunciation dictionary. In order to develop accurate and robust model, few thousand of utterance sample must be involved in HMM acoustic training (Young *et al*., 2006). The training samples for this experiment consists of different Malay accent (chinese and Malay speaker), gender, normal and disorder patient are taken into account during the process of data collection. the word-based pronunciation dictionary been created for the use of acoustic training as shown in Fig. 8 with the total of 23 monophone been used.

HMM State Likelihood Evaluation

The experiment been done in this research is by applying the general approach of identifying FE by using HMM that provides statistical framework for modeling speech patterns which most widely use technique in speech recognizer (Rabiner, 1989). First, the sequence of feature vectors from FE process of MFCC been taken as a realization of concatenation elementary process describe by HMM. This HMM models will be observed through stochastic process that produces the time set of observations. HMM speech recognizer will identify unknown speech by estimating likelihood of each phoneme at the frames of the speech signal. Searching procedure will determine the highest likelihood of phoneme sequence that only been correspond to the words in vocabulary.

Isolated word recognition in this experiment will assume the spoken word of the speech utterance will be represent by a sequence of speech vectors observation of O, denoted in Equation 1 below:

$$O = o_1 o_2, \ldots\ldots, o_T \tag{1}$$

o_T is the speech vector observed at time t. The isolated word recognition occurrence can be denoted as in Equation 2:

$$\arg \max_i \{P(w_i \mid O)\} \tag{2}$$

where, w_i is the i^{th} of vocabulary word. For the probability is not compute directly, but will be compute using Bayes' Rule as:

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)} \tag{3}$$

ARNAB	a r n a b	LEMBU	l e m b u
BAJU	b a j u	LIMA	l i m a
BALDI	b a l d i	LOBAK	l o b a k
BANTAL	b a n t a l	LUTUT	l u t u t
BELON	b e l o n	MANGGIS	m a n g g i s
BISKUT	b i s k u t	NANAS	n a n a s
BOTOL	b o t o l	NANGKA	n a n g k a
DAUN	d a u n	NASI	n a s i
DELIMA	d e l i m a	NYAMUK	n y a m u k
DUA	d u a	PAYUNG	p a y u n g
DUIT	d u i t	PENSEL	p e n s e l
DUKU	d u k u	PINGGAN	p i n g g a n
DURIAN	d u r i a n	PISAU	p i s a u
GAJAH	g a j a h	PUDING	p u d i n g
GARFU	g a r f u	RADIO	r a d i o
GELAS	g e l a s	RAGA	r a g a
GIGI	g i g i	RAMBUT	r a m b u t
KACANG	k a c a n g	REBUNG	r e b u n g
KAD	k a d	RINGGIT	r i n g g i t
KAMBING	k a m b i n g	ROTI	r o t i
KATIL	k a t i l	RUMAH	r u m a h
KEK	k e k	SABUN	s a b u n
KELAPA	k e l a p a	SELUAR	s e l u a r
KERANG	k e r a n g	SEMBILAN	s e m b i l a n
KIJANG	k i j a n g	SEPULUH	s e p u l u h
KIPAS	k i p a s	SIPUT	s i p u t
KOMPUTER	k o m p u t e r	SOTONG	s o t o n g
KOTAK	k o t a k	SYAMPU	s y a m p u
KUCING	k u c i n g	TANGAN	t a n g a n
LANGSAT	l a n g s a t	TELEVISYEN	t e l e v i s y e n
LAPAN	l a p a n	TELINGA	t e l i n g a
		TIGA	t i g a
		TIKUS	t i k u s
		TISU	t i s u
		TOPI	t o p i
		TUJUH	t u j u h
		ZIP	z i p
		ZOO	z o o

Fig. 8. Word-based training pronunciation dictionary

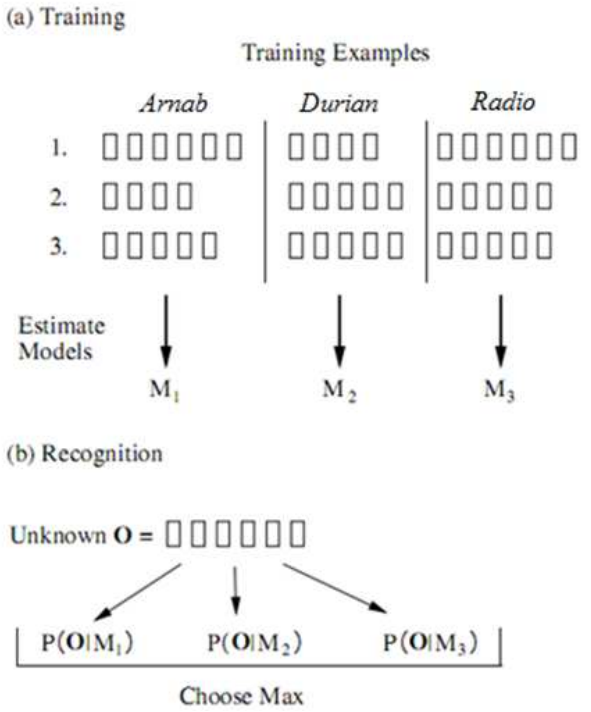

Fig. 9. HMM recognition of isolated word

$P(w_i)$, is the set of prior probabilities where the most probable spoken word depends on the likelihood of $P(O|w_i)$. For the HMM based speech recognition, the O will be correspond to each word that been generated by Markov Model as denoted in Equation 3. A Markov Model is a finite state machine which change state once every time unit and each time t that a state j is entered. Then the speech vector of o_T is generated from the probability density denoted as $bi(Ot)$. To generate the sequence of o_1 to o_6 the six model will move through the state sequence X = 1,2,2,3,4,4,5,6 which the entry and exit states are non-emitting.

The transition probabilities and output probabilities can be express in Equation 4 below:

$$P(O,X|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)\ldots \quad (4)$$

where, the joint probability that O is generated by the model M that moving through the state sequence of X. only observation O is known and X as the underlying state sequence is hidden.

The likelihood can be express by considering the most likely state sequence denoted in following Equation 5 below:

$$P(O|M) = \max_x \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(O_t) a_{x(t)x(t+1)} \right\} \quad (5)$$

where, $x(O)$ is constrained to be the model entry state and $x(t+1)$ is constrained to be the model of exit state. The process can be display in Fig. 9.

Based on Fig. 9; the training set are the sample that corresponding to particular model that can be determine automatically by HMM re-estimation procedure. This will provide enough number of representative samples of each word that can be collected. Then HMM will be constructed to implicitly models all O(t) of the sources variability inherent in real speech by training HMM model for each vocabulary word. The recognition phase of the unknown word is where the likelihood of each model generating that word is calculated to find the most likely model identifies the word.

Results

The experiment setup been divided into 2 phase that is training and testing. The sample been divided into 2 categories of children and adult. The acoustic training data been selected from best speaker and patient speaker for giving the rich input acquisition for the training database. The total of children involved was about 42 and adult was 38 people. Explained in previous chapter, the selected word for this new corpus design been simplified from 128 word into 64 word. Therefore, each target sample need to speak the word for 6 times to keep the consistency of the wave signal which will sum up training sampling into 80×64×6 where altogether is about 30,720 voice sample has been done for training set. The main concern is to test the unknown utterance accordingly into places and manner of articulation to test for the anomalies in the unknown sample. Table 4 shows the list of Malay phoneme according to place and manner of articulation which is a common guideline among the linguistic researcher (Celce-Murcia and McIntosh, 1979; Michailovsky, 1994).

Two main categories in Table 4 which are place of articulation and manner of articulation where place of articulation is the point of contact where an obstruction occurs in the vocal tract between an articulatory gesture and manner of articulation is the configuration and interaction of the speech organs when making a speech sound. From all the elements in the Table 4, we are focusing on two priority group that is alveolar (place) and plosives (manner) because these two groups have the highest possibilities to be articulated wrongly (Jamilah, 2014).

For the overall words in the corpus had been categorized into the categories of testing places as been mentioned in Table 5. The highest word to be tested is in the 3 places with the total percentage for almost 50% from the corpus. The less words can be tested is on the 4 places as the corpus design don't have much words on this categories which for about only 6.25% can be tested.

Evaluating the recognizer accuracy is consists of recognition results where it's been evaluated by string alignment. This is the process when the reference transcription is aligned with the recognizer's transcription by using dynamic programming (Stein *et al*., 2001). Then the differences are counted. There are three different types of error; substitutions, deletions and insertions. With the total number of phonemes in reference transcription and the number of this three type error, the two informative values of accuracy and correctness can be compute. The formula as shown below:

$$Accuracy = \frac{N - S - D - I}{N} \quad (6)$$

$$Correctness = \frac{N - S - D}{N} \quad (7)$$

Where:

N = The total number of phonemes in the reference transcriptions
S = The number of substitution errors
D = The number of deletion errors
I = The number of insertion errors

The accuracy is where the recognizer has inserted excess phonemes and correctness is the proportion of recognized phonemes that actually correct.

Besides that, the intelligibility of the recognizer been also computed by using the formula of Word Error Rate (WER) denoted in Equation 8. The elements of *S*, *D* and *I* was gathered from the transcription of the speaker to be computed. C is the number of correct words:

$$WER = \frac{S + D + I}{S + D + C} \quad (8)$$

Based on previous research, the indicator shows that, the lowest WER means better speech recognition accuracy for the recognizer.

For the overall results which been showed in Table 6 had achieved almost 55% result of % Correct for sentence-level accuracy based on the total number of label files which are identical to the transcription files. The second line is the word accuracy based on the Dynamic Programming-based string alignment Procedure (DP) matches between the label files and the transcriptions. The results had achieved 75.89% of % correctness with total of accuracy had achieved for about 66.96%. The result of error type for D, S and I been showed above. The result been calculated by using Equation 6 and 7 respectively.

The example informative values of each type of error cases been illustrated as in Fig. 10 above. The example took case of 3 sample reference and recognition sample for the testing data. The illustration shows how the informative values been pointed to the string alignment to the number of deletions, substitutions and insertions before accuracy and correctness can be computed. The WER for baseline state had achieved the highest percentage of correctness which about 55% compare to other number of states. Somehow the WER for baseline is the best for each different state that is about 0.33 with total accuracy is the highest compared to other setting for almost 66.96% which shown in Table 7.

Figure 11 shows the result been plot by number of states over % of accuracy. It's clearly indicated that the increasing number of state might not increase the percentage of accuracy. Several test been done from 80 different speaker with more than hundred sample utterance taken from the database.

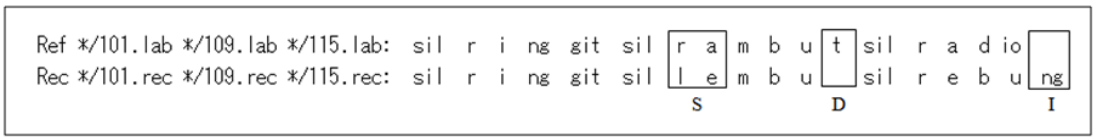

Fig. 10. Informative values for 3 type of error

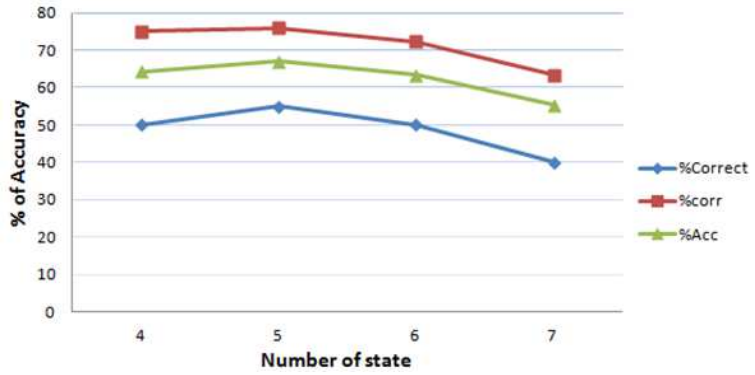

Fig. 11. Recognition accuracy and correctness with the output probability densities approximated by a mixture of state 4,5, 6 and 7

Table 4. List of malay phoneme according to place and manner of articulation

Place of articulation/ manner of articulation	Labial	Labial-dental	Alveolar	Palato-alveolar	Palatal	Velar	Uvular	Glottis
Nasal	m	-	n	ɳ		ŋ		
Plosive	p,b	-	t, d			k, g	q	k
Affricate	-			tʃ (c) dʒ(j)				
Fricative	-	f, v	s, z	sy		X(kh)y(gh)		h
Semi-vowels	-	-			y	w		
Tap	-	-	l					
Trill	-	-	r					

Table 5. Total words and percentage which can be tested

Category	Total words	Percentage
Words which can test for 4 places	4	6.250
Words which can test for 3 places	30	50.00
Words which can test for 2 places	20	31.25
Words which can test for 1 places	10	15.60

Table 6. Overall Results of % SENT and WORD

	HTK results analysis overall results
SENT:	%Correct=55.00 [H=11, S=9, N=20]
WORD:	%Corr=75.89, Acc=66.96 [H=85, D=5, S=22, I=10, N=112]

Table 7. WER for different number of state

# state	% correct	% corr	% Acc	WER
4	50.0	75.00	64.29	0.36
Baseline 5	55.0	75.89	66.96	0.33
6	50.0	72.32	63.39	0.37
7	40.0	63.39	55.36	0.45

Conclusion

In this study, the main concern is the proposed of HMM as statistical modeling technique for speech recognition system for diagnosisng patient that suffer from articulation disorder. Its emphasize on the design of Malay speech corpus that balanced with all the phonetic placed and manner of articulation sample appeared in Malay speech articulation therapy environment. The architecture of speech recognition engine had been also been describe with few discussion on HMM state likelihood evaluation. The 64 word corpus design been tested with few changes in the HMM setting and also the changes of the probability densities approximated by a mixture of different state setting. The output show, the baseline 5 state is the best setting which produces WER for about 0.33 and result accuracy achieved for about 75.89%. In short, phonetic balanced database could provide a good recognizer speech database with correct HMM setting.

Future Work

The process of preparing and designing the training corpus involve a lot of work. By joining the process of designing with expert such as speech language pathologist, it can shorten the process in the future. Few techniques of segmentation the training utterance and adjusting the FE setting might also the best way to improve the recognition accuracy especially for isolated phoneme recognition. It might be the future interest to improve this project.

Acknowledgement

The authors gratefully acknowledge the Ainuddin Wahid scholarship provided by Universiti Teknologi Malaysia and university research grant provided by Research Management Centre and sponsored by Ministry of Education, Malaysia. Vot: Q.J130000.2545.04H41 and Flagship Vot: 10H93 (Research University Grant) GUP Universiti Teknologi Malaysia, Johor Bahru, Malaysia.

Funding Information

This article was funded by the Ministry of Higher Education, Malaysia. Vot: Q.Jl30000.2545.04H4l, GUP Universiti Teknologi Malaysia. Johor Bahru, Malaysia and Ainuddin Wahid Scholarship provided by Universiti Teknologi Malaysia and University Research Grant.

Author's Contributions

Mohd Nizam Mazenan: Designed the research plan, organized the study, experimental setup and data analysis. Also work with writing the manuscript.

Tan Tian Swee: Computer analysis with the test data, enhance the mathematical model used and contributed to the writing of the manuscript.

Tan Hui Ru: Participated in the corpus design, data collection and experiments. Also contributed to the writing of the manuscript.

Azran Azhim: Participated in all experiments, setup and preparations. Also work with meU1od used and writing the manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all other authors have read and approved the manuscript and no ethical issues involved.

References

Asmah, H.O., 1983. The Malay Peoples of Malaysia and Their Languages. 1st Edn., Dewan Bahasa dan Pustaka, Kuala Lumpur, pp: 682.

Axelsson, A. and E. Björhäll, 2003. Real time speech driven face animation. MSc Thesis, Institutionen för systemteknik.

Aymen, M., A. Abdelaziz, S. Halim and H. Maaref, 2011. Hidden Markov Models for automatic speech recognition. Proceedings of the International Conference on Communications, Computing and Control Applications (CCA' 11).

Van Riper, C. and R.L. Erickson, 1996. Speech Correction: An Introduction to Speech Pathology and Audiology. 9th Edn., Allyn and Bacon, Boston, ISBN-10: 0138251428, pp: 532.

Celce-Murcia, M. and L. McIntosh, 1979. Teaching English as a Second or Foreign Language. 1st Edn., Newbury House Publishers, Rowley, ISBN-10: 088377125X, pp: 389.

Davis, S. and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Proc., 28: 357-366. DOI: 10.1109/TASSP.1980.1163420

Donald, S. and S.H. Katherine, 1996. An experimental approach to the problem of articulation in aphasia. Cortex, 2: 277-292. DOI: 10.1016/S0010-9452(66)80008-4

Dronkers, N.F., 1996a. A new brain region for coordinating speech articulation. Nature, 384: 159-161. PMID: 8906789

Dronkers, N.F., 1996b. A new brain region for coordinating speech articulation. Nature, 384: 159-61. PMID: 8906789

Hayes, B., 2011. Introductory Phonology. 1st Edn., John Wiley and Sons, Hoboken, ISBN-10: 1444360132, pp: 336.

Hina, H., 2012. Phoneme and feature theory. University of the Punjab.

Hugh, C. and A.S. Frank, 1894. A Dictionary of the Malay Language. 1st Edn., Authors at the Government's printing Office, Taiping.

Jamilah, S., 2014. Personal Interview regarding process diagnosis and treatment for articulation disorder.

Karjalainen, M., 2008. Kommunikaatioakustiikka. 2nd Edn., Teknillinen korkeakoulu, Espoo, ISBN-10: 9512297493, pp: 255.

Ladefoged, P., 1975. A course in Phonetics. 1st Edn., Harcourt Brace Jovanovitch, New York.

Laver, J., 1994. Principles of Phonetics. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 052145655X, pp: 707.

Michailovsky, B., 1994. Manner Vs place of articulation in the Kiranti initial stops. Proceedings of the 26th International Conference on Sino-Tibetan Languages and Linguistics, (TLL' 94), Osaka, Japan, pp: 766-772.

Mohd Nizam, M. and T.S. Tan, 2012. Malay alveolar vocabulary design for malay speech therapy system. Universiti Teknologi Malaysia.

Noraini, S. and J. Kamaruzaman, 2008. Acoustic Pronunciation Variations Modeling for Standard Malay Speech Recognition. Comput. Inform. Sci. DOI: 10.5539/cis.v1n4p112

Nur Hana, S., 2007. Study on phonetic context of Malay syllables towards the development of Malay speech synthesizer. Universiti Sains Malaysia.

Paul, D.B., 1990. Speech recognition using hidden markov models. Lincoln Laboratory J., 3: 41-62.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 77: 257-286. DOI: 10.1109/5.18626

Raminah, S. and S. Rahim, 1987. Kajian bahasa untuk pelatih Maktab Perguruan. Universiti Utara Malaysia.

Sariyan, A., 1988. Isu-isu Bahasa Malaysia. 1st Edn., Penerbit Fajar Bakti, Petaling Jaya, ISBN-10: 9679331903, pp: 263.

Stein, C., T.H. Cormen, R.L. Rivest and C.E. Leiserson, 2001. Introduction to Algorithms. 1st Edn., MIT Press, Cambridge, ISBN-10: 0262032937, pp: 1180.

Swee, T.T. and S.H.S. Salleh, 2008. Corpus-based malay text-to-speech synthesis system. Proceedings of the 14th Asia-Pacific Conference on Communications, Oct. 14-16, IEEE Xplore Press, Tokyo, pp: 1-5.

Teoh, B.S., 1994. The Sound System of Malay Revisited. 1st Edn., Dewan Bahasa dan Pustaka, Ministry of Education, Malaysia, Kuala Lumpur, ISBN-10: 9836241841, pp: 150.

Thomas, P.J. and F.R. Carmack, 1990. Speech and Language: Detecting and Correcting Special Needs. 1st Edn., Allyn and Bacon, Boston, ISBN-10: 0205123643, pp: 175.

Tan, T.S., Helbin-Liboh, A.K. Ariff, C.M. Ting, 2007. Application of Malay speech technology in Malay speech therapy assistance tools. Proceedings of the International Conference on Intelligent and Advanced Systems, Nov. 25-28, IEEE Xplore Press, Kuala Lumpur, pp: 330-334. DOI: 10.1109/ICIAS.2007.4658401

Ting, H., J. Yunus, S. Vandort and L. Wong, 2003. Computer-based Malay articulation training for Malay plosives at isolated, syllable and word level. Proceedings of the 4th Pacific Rim Conference on Information, Communications and Signal Processing, Dec. 15-18, IEEE Xplore Press, pp: 1423-1426. DOI: 10.1109/ICICS.2003.1292700

Tischer, S., 2009. U.S. Patent Application l 2/357,456. Filed January 22, 2009 for Methods. Systems and Products for Synthesizing Speech.

Voice Production, 2009. Voice care for teachers program. Department of Education and Early Childhood Development, Voice Production.

Yong, L.C. and T.T. Swee, 2014a. Low footprint high intelligibility Malay speech synthesizer based on statistical data. J. Comput. Sci., 10: 316-324. DOI: 10.3844/jcssp.2014.316.324

Yong, L.C. and T.T. Swee, 2014b. Low footprint high intelligibility Malay speech synthesizer based on statistical data. J. Comput. Sci., 10: 316-324. DOI: 10.3844/jcssp.2014.316.324

Young, S., G. Evermann, M. Gales, T. Hain and D. Kershaw *et al*., 2006. The HTK Book (for HTK Version 3.4), Cambridge University Engineering.

Zimmermann, J. and J. Zimmermann Jr, 2002. Stochastic speaker recognition model.