Research Journal of Applied Sciences, Engineering and Technology 7(21): 4535-4540, 2014
DOI:10.19026/rjaset.7.830
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 02, 2014 Accepted: February 15, 2014 Published: June 05, 2014

Research Article

Malay Wordlist Modeling for Articulation Disorder Patient by Using Computerized Speech Diagnosis System

Mohd Nizam Mazenan and Tian-Swee Tan
Department of Biotechnology and Medical Engineering, Faculty of Biosciences and Medical Engineering (FBME), Medical Implant Technology Group (MediTEG), Material Manufacturing Research Alliance (MMRA), Universiti Teknologi Malaysia (UTM), 81310 Skudai Johor, Malaysia

Abstract: The aim of this research was to assess the quality and the impact of using computerized speech diagnosis system to overcome the problem of speech articulation disorder among Malaysian context. The prototype of the system is already been develop in order to help Speech Therapist (ST) or Speech Language Pathologist (SLP) in diagnosis, preventing and treatment for an early stage. Few assessments will be conducted by ST over the patient and mostly the process is still using manual technique whereby the ST relies from their human hearing and their years of experience. This study will surveys the technique and method use by ST at Hospital Sultanah Aminah (HSA) (Speech therapist at Speech Therapy Center) to help patient that suffer from speech disorder especially in articulation disorder cases. Few experiment and result had also been present in this study where the computerized speech diagnosis system is being tested by using real patient voice sample that been collected from HSA and the students from Sekolah Kebangsaan Taman Universiti Satu.

Keywords: Articulation disorder, computerized speech diagnosis system, Hidden Markov Model (HMM)

INTRODUCTION

Speech therapy has many definition and conceptual meaning in it. But in general definition for the speech disorder is a form of therapy or process that has been designed to address language and speech disorder where it concern with disorders of human communication (Tan *et al*., 2007). The definition is always related to the word ST where it is about a person which address the same issue as they occur and will provide preventive care which been designed to stop or avoid such disorder before they start (Tim, 2005). In other hand, speech disorder is refers to a problem of producing sounds, whereas a language disorder refers to a difficulty understanding or putting words together to communicate idea (Bharathi and Shanthi, 2012). Speech disorder not only can happen within specific ages but it also may occur to children and adult.

Based on statistical report by World Health Organization (WHO), speech and language disorder affects at least 3.5% of human population communication skills. Even though it maybe look to be small amount, but in Malaysia it effect 5-10% of children manifest speech and language problems where it is about 10,223 children were reported to have learning disabilities (Sri Raflesia Sdn Bhd, 2008; Jabatan Pendidikan Khas Kementerian Pendidikan Malaysia, 2002; Woo and Teoh, 2010).

The statistic shows the current situation happen regarding to the problem of speech disorder according to health organization. Somehow the speech therapy itself contains a lot of problem especially in the technical factors that related to the therapy assessment, the process of diagnosis, the techniques been use and the speech recognition accuracy issue. For example, the therapy diagnosis assessment in most hospital or speech therapy centers in Malaysia is using too many words as a training data. Some of this word maybe not suitable for specific case of speech disorder (Mohd Nizam, 2013). Other example is where the ST is still using manual technique for the diagnosis which it may lead to time consuming and lack of accuracy (Ooi and Yunus, 2007). The accuracy here is happen when the therapist is using their experience and hearing as a "tool" to detect the speech problem having by the patient (HSA Speech Therapy Center). Because of human hearing can produce major mistake in accuracy, the computerized system should been use in recognition where it can minimal the accuracy problem. Again by using manual technique, the total involvement of ST for each session is too high where the ratio can reach about 1:50 between speech therapist and patient. Furthermore,

Corresponding Author: Tian-Swee Tan, Department of Biotechnology and Medical Engineering, Faculty of Biosciences and Medical Engineering (FBME), Medical Implant Technology Group (MediTEG), Material Manufacturing Research Alliance (MMRA), Universiti Teknologi Malaysia (UTM), 81310 Skudai Johor, Malaysia

even if there are computerized diagnosis or speech recognition system, most of it available in foreign language especially in English language (Ting *et al*., 2003).

Therefore for an early diagnosis setup, this study will propose computerized technique that use speech recognition that been model by basic Malay language corpus design specifically for articulation disorder. Starting from training the database until recognition of unknown sample data will been done by using HMM technique that also covers the result of accuracy in analysis and diagnosis process.

MATERIALS AND METHODS

Experimental setup: An experiment had been conducted where the goal is to find the accuracy for recognition rate by having a control set of specific target words database which focus more on Malay speech therapy target word. Most of this words can help ST to understand the patients articulatory disability before ST can design an effective therapy strategies that consist set of exercise or training to cure the patient who having speech disorder. But for an early stage, this selected consonant will work on early diagnosis phase for the whole speech recognition system. Based on previous research at HSA, Malay word for therapy are about 108 words which cover most of the Malay articulation function. Table 1 show the total of the simplify target words for Malay speech therapy word design.

In this table, the target word will tackle Malay speech articulation function that concern in solving the problem in speech therapy for speech disorder. There are target words that will cover from Malay consonant, Malay vowel, Malay alveolar and Malay plosives. Malay consonant consists of alphabet (B, D, G, K, L, N, P, R, S, T and Z), Malay vowel consists of alphabet (A, E, I, O, U), Malay alveolar consists of alphabet (D, L, N, R, S, T, Z) and lastly for Malay plosives will cover alphabet (B, D, G, K, P, T).

By using the HMM as the statistical analysis tool in recognition process and Mel-cepstral Frequency Coefficient extraction (MFCC) as feature extraction technique, this experiment is started by capturing the voice sample of the training and target patient. For the sample data collection in this experiment, there are 80 children are involved in Malay target word database for normal speech children. Therefore, each sample needs to speak the word for 6 times to keep the consistency of

Table 1: Target words for Malay speech therapy word design

Consonant	Word target	Total
A	Abjad, angsa, arnab, ayam	4
B	Baju, baldi, bantal, belon, biskut, botol, buku	7
D	Dadu, daun, delima, dodol, dua, duit, duku, durian	8
E	Empat, enam, epal	3
G	Gajah, garfu, gelas, gigi	4
I	Ikan, itik	2
K	Kacang, kad, kambing, kancil, kasut, katak, katil, kek, kelapa, keli, kera, kerang, kijang, kipas, komputer, kopi, kotak, kucing, kunci	19
L	Lampu, langsat, lapan, lembu, lima, limau, lobak, lutut	8
N	Nanas, nangka, nasi, nyamuk	4
O	oren	1
P	Payung, pensel, pinggan, pisang, pisau, puding	6
R	Radio, raga, rambut, rebung, ringgit, roti, rumah	7
S	Sabun, satu, seluar, sembilan, semut, sepuluh, siput, sotong, sudu, syampu	10
T	Tangan, televisyen, telinga, tiga, tikus, tilam, timun, tisu, topi, tuala, tujuh	11
U	Ubat, udang, ular	3
Z	Zip, zoo	2
Total		99

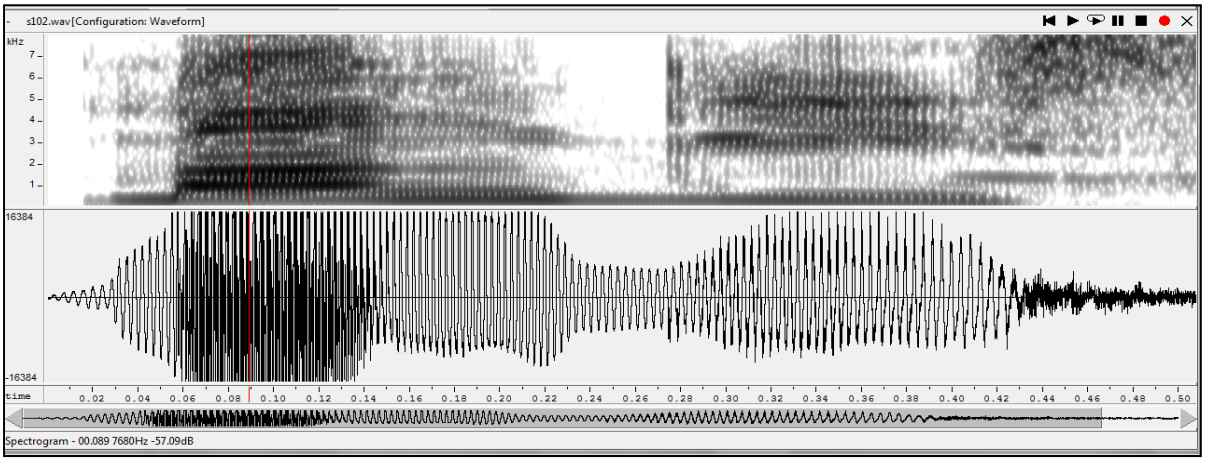

Fig. 1: Sample of normal voice sample in a form of spectrogram and waveform

the wave signal where altogether is about 47520 sample voices has been collected. The training data has approximately 0.0-1500 msec data length. The data sampling rate of the recording was been collected by using GoldWave software at 16 kHz and 16-bit resolution format by using standard vocal microphone. The recording was been done in quiet room environment that specifically for speech therapy to avoid disturbance unknown noise. There are about 15 speech disorder children and 10 normal speech children will be involved in testing the accuracy of the unknown recognition phase. The training data for speech signal sample is illustrated in Fig. 1. Before the training sample been uniformly segmentize, extracting the information by using Feature Extraction (FE) and tying the model, we need to design the dictionary for each of the word that need to be test. Assuming we are using standard Malay phoneme based pronunciation dictionary for our target words, where it has been used to describe the phones HMM acoustic model for the mapping process to form a word for both training and decoding purpose. The list of Malay phoneme will be generated by the -n argument and stored in file monophones1.rtf by using standard HMM command. After the sample data been collected, next step as follows.

Speech signal processing: Speech processing converts the speech waveform into parametric representation. The conversion of the waveform been done by FE where the speech features are used as an inputs to the speech recognition. The FE will construct the combination of variables to get around these problems while still describing the data with sufficient accuracy. In this experiment, we will fully utilize the use of 12 Mel-Frequency Cepstral Coefficient (MFCC) (Davis and Mermelsten, 1980). Axelsson and Bjo¨rhall (2003) explained that MFCC has characteristics of the human auditory system and commonly used in the Automatic Speech Recognition systems (ASR). The front end parameterized an input speech signal into a sequence of MFCC vectors. The process of this front end is shown in Fig. 2.

Pre-emphasize: In the pre-emphasize stage, the speech sample filtering is applied in Eq. (1) to spectrally flatten the signal frames as the high concentration of energy frequencies will be using to evaluate and be used to reduce the effects of the glottal pulses and radiation impedance whereby it can focusing more on the spectral properties of the vocal tract. The value of constant a is generally around 0.9 and 1.0 (Jadhav *et al*., 2013). In our research we used a = 0.97. The pre-emphasize speech signal, $\hat{s}(n)$ will be sent through a high pass filter according to the Eq. (2):

$$H(z) = 1\text{-}az^{-1} \quad (1)$$

$$\hat{s}(n) = s(n) - as(n-1) \quad (2)$$

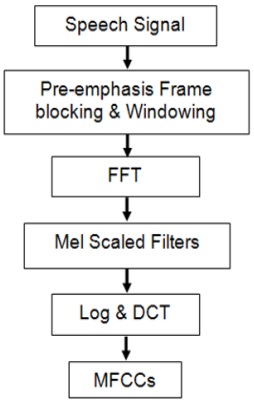

Fig. 2: Process of MFCC front end

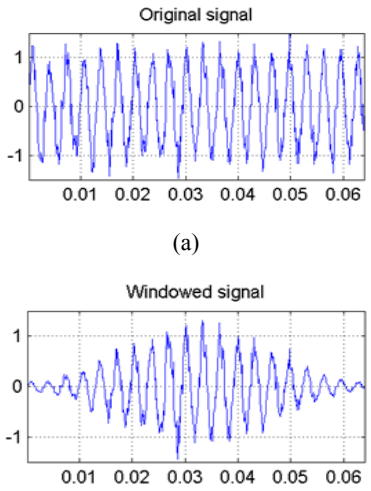

Fig. 3: Effect of multiplying a hamming window by Jang (2011)

Windowing: After the pre-processing of the speech sample, there are spectral leakages during the selection of frame locations and frame blocking process of non-stationary signals. This is because of the Finite Impulse Response (FIR) that will results ripples in the stop band and pass band of the filter frequency response that is due to Gibb's phenomena. The windowing function will multiplied to the filter impulse response to minimize this effect on each spectral frame. Hamming window will reduce the signal discontinuities where the signal is windowed and multiplying in such a way that the frames overlap (Fig. 3) which w (n) been denoted as Eq. (3):

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \quad 0 \le n \le M-1 \quad (3)$$
$$= 0 \quad \text{otherwise}$$

The overlapping that happen in hamming window will avoid discontinuities on both side lobes of each window.

Mel-frequency cepstrum coefficients computation: The MFCC had been choose as a FE for this experiment prior to the characteristic of this FE which based on human auditory system that focusly on non-uniformly spectral envelop and also widely use in Automatic Speech Recognition (ASR) systems (Axelsson and Bjo¨rhall, 2003; Grigore *et al*., 2011). MFCC will be able to extract important phonetically information of the signals by focusing on lower range of the Mel frequencies. The log mel-scale filter bank is expressed in a form of linear frequency below 1 kHz and a logarithm spacing above 1 kHz that mimic human ear's perceived frequency (Davis and Mermeslsten, 1980). Mel-Scale been described in Eq. (4):

$$mel\,(f) = 2595\ log\,(1 + f/700) \tag{4}$$

where,
f : The real frequency
mel (f) : The perceived frequency

Next, the input training speech sample is transformed by using short-time Fourier Transform method from the time domain to frequency domain that been shown in Eq. (5). $W(n)$ in Eq. (5) is the Hamming window function, that derived the speech signal which it can be treated as stationary and not influenced by other signals within short period of time:

$$\mathrm{Y}(m) = \sum_{\mathrm{n=0}}^{\mathrm{N\text{-}1}} \hat{\mathrm{x}}_{\mathrm{i}}(\mathrm{n}) \bullet \mathrm{e}^{-2\mathrm{jn}\pi\frac{\mathrm{m}}{\mathrm{N}}} \tag{5}$$

The energy spectrum need to be found as denoted exactly in Eq. (6):

$$X(m) = |Y(m)|^2 \tag{6}$$

The energy in each of mel window has to be calculate in mathematical expression below:

$$S[k] = \sum_{j=0}^{\frac{K}{2}-1} W_k(j) X(j) \tag{7}$$

where, 1≤k≤M and M is the number of mel-window in mel scale, which can generally range from 20 to 24. $W_k(j)$ is the triangular weighted function associated with k^{th} mel window in mel scale.

After logarithm and cosine transforms, mel frequency cepstral coefficients can be derived as follow:

$$C[n] = \sum_{k-1}^{M} Log\,(S[k]) \cos[n(k - 0.5\frac{\pi}{M}] \tag{8}$$

where,
c [n] : The n^{th} cepstral vector component for 0≤n≤L
L : The desire order of the MFCC
c [0] : The zero order of MFCC

For the short time energy coefficient is where it incorporated to cepstral coefficients because it contains valuable discriminative information among different phonemic characteristics. Vowels exhibit much larger energy values than fricatives, plosives and silence. Normalized energy is used to normalize speaker loudness variation. Energy can be computed from waveform as:

$$E = \log\left(\sum_{n=0}^{N-1}\left(\hat{x}(n)^2\right)\right) \tag{9}$$

In Eq. (9), E is energy for frame l, which has N discrete time sample in it. As a result, $\hat{x}_1(n)$ has been windowed.

A feature vector is usually consists of 12 MFCC and energy coefficient can be length 39 by taking account the first order derivative of cepstral coefficient (delta coefficients) and second order derivative of cepstral coefficient (Delta-Delta coefficients) and acceleration of speech and energy coefficient. Based on previous research by Stemmer *et al*. (2001), this first and second order derivative and normalized energy are added to mel frequency cepstral coefficients that consists of the differences of features between predecessor and successor frames whereby it will capture the dynamic changes of the signals. Those experiments will add these delta coefficients to improve recognition accuracy performance as follows:

- First-order derivative of MFCC:

$$\Delta C_t[n] = C_{t+1}[n] - C_{t-1}[n],\ 0 \le n \le L \tag{10}$$

- Second-order derivative of MFCC:

$$\Delta\Delta C_t[n] = \Delta\, C_{t+1}[n] - \Delta C_{t-1}[n],\ 0 \le n \le L \tag{11}$$

- First and second order differenced energy coefficient:

$$\Delta E_t[n] = E_{t+1}[n] - E_{t-1}[n]$$
$$\Delta\Delta E_t[n] = \Delta E_{t+1}[n] - \Delta\, E_{t-1}[n] \tag{12}$$

A typical feature vector, yt is composed by those MFCC acoustic vector and its first and second order differences which can be characterized by:

$$y_t = \{E_t, C_t, \Delta E_t, \Delta C_t, \Delta\Delta E_t, \Delta\Delta C_t\} \tag{13}$$

RESULTS AND DISCUSSION

The speech signal in this experiment will be process to evaluate the matching point between target words with unknown speech signal that suppose to be

Table 2: Malay word recognition evaluation

Word	Total sample	Correct	Error	Accuracy rate (%)
Abjad "alphabet"	200	179	21	89.50
Empat "four"	180	170	10	94.44
Itik "duck"	190	177	13	93.16
Oren "orange"	200	183	17	91.50
Ular "snake"	200	184	16	92.00

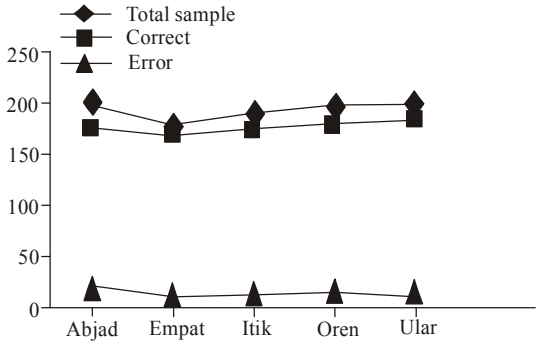

Fig. 4: Recognition evaluation of five Malay words

correspond with the training model. The detection of words accuracy rely on start and end points where the subsequent processing of the data need to be kept to a minimum (Jadhav *et al*., 2013). The detection of this start and end point is based on analyzing its energy profile. When the energy of voice signal rise at the setup threshold value, it will marks the presence of voice input and vice versa.

The experiment consists of training speech sample taken from 80 children age from 8 to 13 years old, male and female. Each of samples will require speaking the word for 6 times where altogether is about 47520 sample voices has been collected. For the earlier output, only few results will be shown by taking only 5 sample results of recognition evaluation that shown in Table 2. As been stated above, there are about 15 speech disorder children and 10 normal speech children will be involved in testing the accuracy of the unknown recognition phase (Fig. 4).

Based on the result, it shows that the word "Empat" ("four") achieve higher accuracy of recognition by the rate of 94.44% compare to the other words. For the lowest recognition evaluation accuracy goes to word "Abjad" ("alphabet") where it capture only about 89.5%. After we see and analyze the result, an early hypothesis can be made by assuming that, the word "Empat" can get the best accuracy in recognition is because there are no similar language model that close to that word among the entire training model. Therefore the probability distribution of this sample word utterance may be far and not close to each other. The hypothesis for the lowest recognition accuracy which is from the word "Abjad", can be assume as the target word is close to other training word that may sound the same. The syllable of "Ab" and "jad" is near to the pronunciation of word "kad" and syllable "nab" from the word "Arnab". The total recognition evaluation was in an excellent result.

CONCLUSION

In this study, a HMM has been use as the statistical analysis tool in recognition process and MFCC as feature extraction technique for front end parameterized of input speech signal. Both technique are among the best technique for ASR and still been using until today. During the recognition evaluation test, the recognition result from this experiment shows that the approximation of the matching signal from training model and unknown utterance is very well. Even the lowest present is still in 80% above range which we can consider that the recognition of an early phase of the speech recognition system is very promising. Therefore, more specific experiment needs to be conduct by improving the language model and the utilization of the recognition and segmentation technique.

ACKNOWLEDGMENT

The authors gratefully acknowledge the research grant provided by Research Management Centre (RMC), sponsored by Ministry of Higher Education (MOHE), Malaysia. Vot: Q.J130000.2545.04H41 and GUP Universiti Teknologi Malaysia, Johor Bahru, Malaysia.

REFERENCES

Axelsson, A. and E. Bjo¨rhall, 2003. Real time speech driven face animation. M.S. Thesis, The Image Coding Group, Department of Electrical Engineering, Linko¨ping University, Linko¨ping.

Bharathi, C.R. and Dr. V. Shanthi, 2012. Disorder speech clustering for clinical data using fuzzy C-means clustering and comparison with SVM classification. Indian J. Comut. Sci. Eng., 3: 713-719.

Davis, S. and P. Mermelsten, 1980. Comparison of parametric representation for monosyllabic word recognition in continuous spoken sentence. IEEE T. Acoust. Speech, 28: 357-366.

Grigore, O., C. Grigore and V. Velican, 2011. Impaiered speech evaluation using mel-cepstrum analysis. Int. J. Circ. Syst. Signal. Process., 5: 70-77.

Jabatan Pendidikan Khas Kementerian Pendidikan Malaysia, 2002. Maklumat Pendidikan Khas Kuala Lumpur. Kementerian Pendidikan Malaysia.

Jadhav, S., S. Kava, S. Khandare, S. Marawar and S. Upadhya, 2013. Voice activated calculator. Int. J. Emerg. Technol. Adv. Eng., 3: 504-507.

Jang, J.S.R., 2011. Audio Signal Processing and Recognition. In: Roger Jang's Homepage. Retrieved from: http://www.cs.nthu.edu.tw/~jang (Accessed on: April 5, 2011).

Ooi, C.A. and J. Yunus, 2007. Computer-based system to assess efficacy of stuttering therapy techniques. Proceeding of the International Conference on Biomedical Engineering, 15: 374-377.

Sri Raflesia Sdn Bhd, 2008. Learning Support and Intervention Services. Web. Southampton. Retrieved from: www.srirafelsia.com (Accessed on: April 8, 2013).

Stemmer, G., C. Hacker, E. Noth and H. Niemann, 2001. Multiple time resolutions for derivatives of Mel-frequency cepstral coefficients. Proceeding of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '01), pp: 37-40.

Tan, T.S., H. Liboh, A.K. Ariff, C.M. Ting and S.H. Salleh, 2007. Application of Malay speech technology in Malay speech therapy assistance tools. Proceeding of the International Conference on Intelligent and Advanced Systems, pp: 330-334.

Tim, P., 2005. Research Methods in Communication Disorders. Whurr Publishers, Ltd., London.

Ting, H.N., J. Yunus, S. Vandort and L.C. Wong, 2003. Computer-based Malay articulation training for Malay plosives at Isolated, Syllable and Word Level. Proceeding of the International Conference on Infornation, Communication and Signal Process, 3: 1423-1426.

Woo, P.J. and H.J. Teoh, 2010. An investigation of cognitive and behavioural problems in children with attention deficit hyperactive disorder and speech delay. Malays. J. Psychiatr., 16: 50-58.