Research Journal of Applied Sciences, Engineering and Technology 7(24): 5278-5282, 2014
DOI:10.19026/rjaset.7.925
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2014 Maxwell Scientific Publication Corp.
Submitted: March 13, 2014 Accepted: April 11, 2014 Published: June 25, 2014

Research Article

Computerized Method for Diagnosing and Analyzing Speech Articulation Disorder for Malay Language Speakers

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

Abstract: This study aims to develop a computerized technique that uses speech recognition as a helping tool in speech therapy diagnosis for early detection. Somehow speech disorder can be divided into few categories which not all type will be fully covered in this research. This study only purposes a solving method for diagnosis of a patient that suffers from articulation disorder. Therefore a set of Malay language vocabulary has already been designed to tackle this issue where it will cover selected Malay consonants as a target words. Ten Malay target words had been choose to test the recognition accuracy of this system and the sample are taken from real patient from Hospital Sultanah Aminah (HSA: Speech therapist at Speech Therapy Center) where the hospital assists in the clinical trial. The result accuracy of the systems will help the Speech Therapist (ST) to give an early diagnosis analysis for the patient before next step can be purposed. An early percentage of correct sample achieved almost 50% in this experiment.

Keywords: Feature extraction technique, Hidden Markov Model (HMM), segmentation, speech articulation disorder, speech recognition, speech therapy

INTRODUCTION

Back to previous years, the speech therapy technique that been done in most hospital, clinic or speech center are using manual technique. This manual technique is often refer as a traditional method which may be considered by many ST to embrace 'traditional' approaches (Powers, 1971). Even nowadays, research been done in Malaysia shows that, this traditional method is still in use widely in Malaysia hospital or speech center. There are no wrong doing by manual technique but from my first observation to the hospital or clinic that still using this technique, the method sometimes lack of accuracy, time consuming (Ooi and Yunus, 2007) and require high number of ST for each session (Saz *et al*., 2009). Speech therapy is a clinical field that focusing on human communication disorder (Tan *et al*., 2007). The common practice of this method can be seen by simply having clients speaking aloud and repeat words over and over again as a way to detect or correct the target words. As of this process going through, the "tool" that only been rely by the ST is their hearing and experience judgment. That would go on for an hour and sometimes it would lead to error elimination during the process (Riper Van, 2007). Compare to computerized technique, human hearing can sometime produce major mistake in accuracy. Therefore, the computerized system should been use in recognition where it can minimal the accuracy problem and the result may be more static. Therefore, this research is proposed the use of computer-based speech articulation diagnostic system for early detection. The computer-based speech therapy system is still new in Malaysia and most of the systems available now are in foreign language such as English (Ting *et al*., 2003).

Few experiments and test has been conducted by using real patient as the target sample for this research. The computerized system will recognize the sample utterance produce by the articulation patient and display a result in a form of % Correct. Other process involve in this experiment was the segmentation of the speech signal and phone mapping by HMM probability tool.

MATERIALS AND METHODS

Experimental setup: The experiment been conducted by having a control set of specific target words that been used in speech therapy, real patient from HSA and a normal quiet room environment that specifically for

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

Table 1: Selected Malay consonants and word design for speech therapy

Vowel/consonant	Selected word
A	Arnab (“rabbit”)
B	Bantal (“pillow”), belon (“balloon”)
M	Manggis (“mangosteen”)
R	Radio (“radio”), raga (“basket”), rambut (“hair”), rebung (“bamboo sprout”), ringgit (“Malaysian money currency”)
T	Tikus (“mouse”)

Table 2: Error rate percentage of manual method

Word	Total sample	Error spoken	Correct	Error rate (%)
Arnab	250	130	120	52
Belon	250	170	80	68
Bantal	250	130	120	52
Manggis	250	140	110	56
Radio	250	200	50	80
Raga	250	160	90	64
Rambut	250	140	110	56
Rebung	250	160	90	64
Ringgit	250	150	100	60
Tikus	250	140	110	56

speech therapy process. For an early stage, selected consonant will be use for early diagnosis phase for the testing part in this speech recognition system. Table 1 shows that selected Malay consonants and word design for the speech therapy.

Based on previous research and experiment, the total word been use is about 99 words. But in this experiment, only 10 selected words had been used as a training set to test the recognition accuracy for articulation pronunciation error. These 10 words are based on therapy training dataset by using hearing as a mechanism. This word is among the words that produce many mistake produce by the patient where the total average of error rate produce is about 60.8%. Table 2 shows the percentage of error rate using manual method, by listening into the total of 250 patient of HSA.

This same target words and sample data from real patient will be use to test the recognition accuracy of computerized technique that was been purposed in this research. The system will use HMM as the statistical analysis tool in recognition process and Mel-cepstral Frequency Coefficient extraction (MFCC) as Feature Extraction (FE) technique for front end parameterized of input speech signal. 12 MFCCs been used as a recognition feature in this experiment because the mel scale of MFCCs is designed based on the human hearing mechanism and commonly used in the Automatic Speech Recognition systems (ASR) (Axelsson and Bjo¨rhall, 2003; Kumar and Mallikarjuna Rao, 2011).

For the voice sample data in this experiment has been collected from normal adult and child with male and female voice patient where the total it’s about 120. Therefore, each person needs to speak the targeted word for 6 times to keep the consistency of the wave signal where altogether is about 7200 sample voices has been collected for the training purpose. The training data has approximately 0.0 to 1500 msec data length. The data sampling rate of the recording was been collected by using GoldWave software at 16000 Hz and 16-bit resolution format by using standard vocal microphone. The recording had been done in a quiet room environment to avoid disturbance of unknown noise that been setup for specifically for speech therapy process. For the testing purpose, the same total sample as the hearing method sample before been used back where it is about 30 speech disorder patient and 20 normal people will be involved in testing the accuracy of the unknown recognition sample to the system. The same word had been use as a testing purpose. Bear in mind that, even though some of the volunteer was been considered as speech disorder patient, not all the target words been pronounce wrongly. It’s all depending on the level of articulation problems that they suffer and this level may be vary from each of the patient. It’s also happen the same thing to the normal people where they might be slightly having a difficulty in pronouncing this target words as the normal speakers is not like anchorman who was trained to be a good and professional speaker.

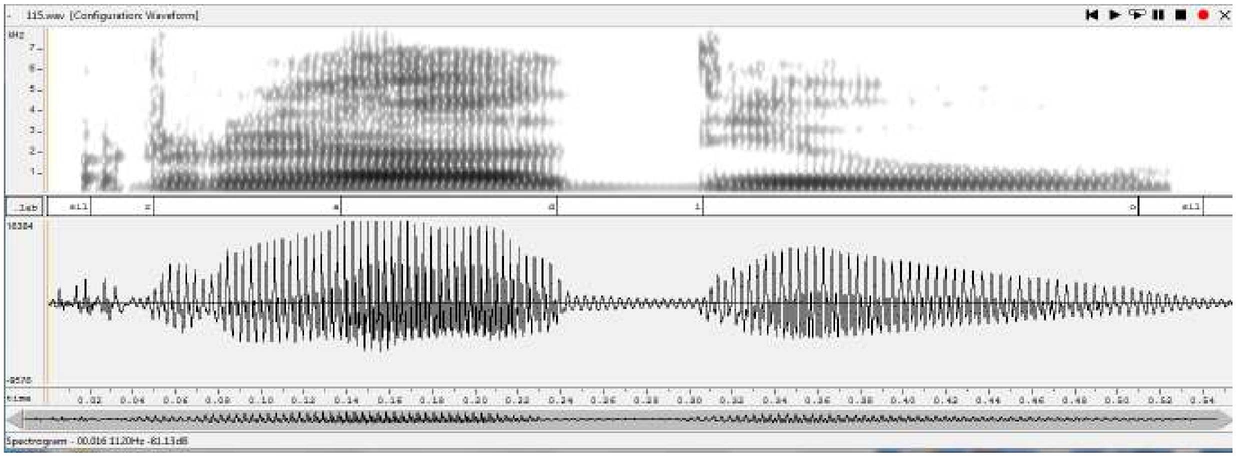

Fig. 1: Training voice sample uniformly segmentized training

Speech data training and testing: The process is begin with training the sample voice in the system. Therefore a word based decoding pronunciation dictionary of the targeted word need to be design. This will describe the phones of HMM acoustic model for the mapping process to form a word for both training and decoding purpose. The process will be continue in speech signal processing where the speech signal in the form of waveform will be converted into parametric representation by using FE that is 12 MFCC setup. Figure 1 shows the sample of waveform that been uniformly segmentize. This sample is from the word "Radio" which produces the highest error spoken rate for about 80% by using traditional hearing method. The transcription for the sample "Radio" above also look very promising which it seems almost segmentize in accurate way. More training data can improve this segmentation as the speech recognition really relies on large training data (Ellis and Morgan, 1999). The most precise way of phonetic segmentation is by manually. But manual segmentation is very costly and requires much time and effort (Ting *et al*., 2007). So in this experiment, it is desirable to have an automatic approach for segmentation, especially when the speech corpus is very large. As the FE will convert the waveform, the speech features will be used as an input to the speech recognition. This step is followed by pre-emphasize stage where the speech sample will be filter applied in Eq. (1) for the signal frames to be spectrally flatten. To focus more on the spectral properties of the vocal tract, high concentration of energy frequencies will be used to evaluate and will reduce the effects of the glottal pulses and radiation impedance. Then, the pre-emphasize speech signal, $\hat{s}$ (n) will be sent through a high pass filter according to the Eq. (2):

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

$$s^{\wedge}(n) = s(n) - as(n-1) \tag{2}$$

The next step requires hamming window process to avoid signal discontinuities both side lobes of each window. The windowing will multiply to the filter impulse response to minimize this effect on each spectral frame. Windowing process which w (n) can been denoted as shown in Eq. (3) below:

$$\begin{aligned} \mathrm{w(n)} &= 0.54 - 0.46\cos\left(\frac{2\pi \mathrm{n}}{\mathrm{M-1}}\right) && 0 \le \mathrm{n} \le \mathrm{M-1} \\ &= 0 && \text{otherwise} \end{aligned} \tag{3}$$

Feature extraction computation: MFCC been used in this experiment which provides good performance, better accuracy and minor computational complexity with respect to alternative features (Davis and Mermelsten, 1980). Therefore, MFCC analysis was performed on the sampled speech input. The extraction of important feature for the sample signal will focus 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 logarithm spacing above 1 kHz where the Mel-Scale been describe in (4):

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

Symbol f denotes the real frequency and *mel* (*f*) denotes the perceived frequency. The next stage is the pattern training of the sample where the speech sample is transformed by using short-time Fast Fourier Transform (FFT) with hamming window function. This will make sure the speech signal is treated as stationary and not influenced by other signal within short period of time. Energy spectrum need to be found as denoted in (5) and also the energy in each Mel window has to be calculated as shown in (6):

$$X(m) = |Y(m)|^2 \tag{5}$$

$$S[k] = \sum_{j=0}^{\frac{K}{2}-1} W_k(j) X(j) \tag{6}$$

The different with real ceptrum by the nonlinear, where perceptual motivated frequency scale is being used. Its mimic approximates the behavior of the human auditory system (Tan and Sheikh, 2008).

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

$$C[n] = \sum_{k-1}^{M} Log\,(S[k]) \cos[\,n(k - 0.5\frac{\pi}{M}] \tag{7}$$

where, c [n] is the n^th^ cepstral vector component for $0 \leq n \leq L$ and L is the desire order of the MFCC, c (0) is the zero order of MFCC.

HMM pattern recognition: After FE process completely done, the acoustic HMM training mechanism will be implemented to generate set of models to represent the observed acoustic vectors of the sample. The flow for the training and testing is the most same, where in the testing and recognition part there are additional stage that is correct pattern matching and selection. The stage will then provide the output for the unknown sample recognition. The process of pattern recognition based on HMM begin after speech signal in an acoustic waveform been converted into digital signal by following the FE process that been describe at previous chapter. After that, the signal needs to be model as denoted in Eq. (8) below:

$$P(\vec{O}_t \,|\, (W_{t-1}, W_t, W_{t+1})) \tag{8}$$

Table 3: SENT: % correct of sample recognition

	Syllabus		Words		Phoneme	
Target words	Correct	Wrong	Correct	Wrong	Correct	Wrong
Ringgit		# (arnab)	#		#	
Tikus		# (arnab)	#			# (raga)
Radio	#			# (arnab)		# (bantal)
Raga	#			# (arnab)	#	
Bantal		# (arnab)	#		#	
Rebung		# (raga)	#		#	
Belon		# (arnab)		# (radio)		# (raga)
Manggis	#			# (radio)	#	
Rambut		# (arnab)		# (arnab)		# (arnab)
Tikus		# (arnab)		# (arnab)		# (raga)
Raga	#			# (arnab)	#	
Manggis	#			# (arnab)		# (raga)
Bantal		# (arnab)		# (arnab)	#	
Ringgit		# (arnab)		# (arnab)		# (arnab)
Radio	#		#		#	
Belon		# (arnab)		# (rebung)		# (bantal)
Rebung		# (arnab)	#	# (arnab)	#	
Rambut		# (raga)		# (arnab)		# (raga)
Arnab		# (raga)	#			# (raga)
Arnab	#		#		#	
	35%		40%		50%	

where, the P is the probability of the given observation model to evaluate the probability of different HMM in generating the same observation sequences of. The W is the sequence of $\vec{O}_t$ words. This step is actually to highlight important and represent enough speech signal information. Next step is pattern matching that using equation shown below:

$$[W_t^i, P(\vec{O}_t, \vec{O}_{t-1}, ... | W_t^i)] \qquad (9)$$

Pattern matching is been done in a sequence of individual observation for the important feature in the sample speech signal data. Search algorithm will be applied in the process to uncover the word sequence of $W = W_1 W_2, ..., W_m$. That has the maximum posterior probability (Chen, 1998; Juang, 1985) which been derived as follow equation:

$$P(W_t^i | O_t) = \frac{P(O_t | W_t^i) P(W_t^i)}{P(O_t)} \qquad (10)$$

where,

$P(W_t^i)$ = The language model

$P(O_t | W_t^i)$ = The acoustic model probability

The last step requires the recognition and comparison of the pattern by using (11):

$$P(S | O) = \arg\max|_T \prod_i P(W_t^i | (\vec{O}_t, \vec{O}_{t-1}, ...)) \qquad (11)$$

RESULTS AND DISCUSSION

This experiment was carried out to evaluate and test the performance of computerized technique in diagnosing the voice sample from articulation disorder patient. The accuracy result will be used to compare with the traditional method which currently been using at HSA. The effect of the performance is differentiate by few parameter of the level setting that is phoneme, syllable and word based. Table 3 shows the SENT: % Correct of the recognized target words based on training sample. Based on the result, the highest score of recognition accuracy was at the phoneme level setting where the % correct achieved about 50%. This look very promising as the accuracy of recognition can be improve by using correct energy setting of FE, increase target samples for a training purpose and multiply the number of re-estimation at the training level. HMM pattern recognition also needs to be study in more specific. Correct algorithm can be implement to get better result. An early hypothesis based on the analyzed result shows that, phoneme based is appropriate form for the recognition of speech articulation disorder case where it will focus more on single word correction rather than sentence correction.

CONCLUSION

In this research study, the used of computerized technique is better compare to traditional approach that using hearing and experience judgment as a tool of diagnosing the patient that suffer from speech disorder. The computerized technique produce recognition result that more stable, accurate and consistent. Even though the total accuracy for computerized recognition is still at lower rate, but an improvement can be make as been explained at the previous chapter. Therefore, more research study and experiments need to be continue as the early system gives an average results, which open much more room for improvement.

ACKNOWLEDGMENT

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

Chen, B., 1998. Search Algorithms for Speech Recognition (PowerPoint Slides). Retrieved from: http://berlin.csie.ntnu.edu.tw/Courses/SpeechRecognition/Lectures2011/SP2011F_Lecture 10_Search Algorithms.pdf.

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.

Ellis, D. and N. Morgan, 1999. Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2: 1013-1016.

Juang, B.H., 1985. Hidden Markov Models. Encyclopedia of Telecommunications.

Kumar, S. and P. Mallikarjuna Rao, 2011. Design of an automatic speaker recognition system using MFCC, vector quantization and LBG algorithm. Int. J. Comput. Sci. Eng., 3: 2942-2954.

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.

Powers, M.H., 1971. Clinical Educational Procedures in Functional Disorders of Articulation. In: Travis, L.E. (Ed.), Handbook of Speech Pathology and Audiology. Prentice-Hall, Englewood Cliffs, N.J.

Riper Van, C., 2007. Personal Correspondence to Wayne Secord. In: Secord *et al*. (Eds.), 2nd Edn., Eliciting Sounds: Techniques and Strategies for Clinicians. Thomas Delmar Learning, Clifton Park, NY, pp: 8.

Saz, O., S.C. Yin, E. Lleida, R. Rose, C. Vaquero and W.R. Rodríguez, 2009. Tools and technologies for computer-aided speech and language therapy. J. Speech Commun., 51(10): 948-967.

Tan, T.S. and H.S.S. Sheikh, 2008. Corpus-based Malay text-to-speech synthesis system. Proceeding of the 14th Asia-Pacific Conference on Communications (APCC, 2008). Tokyo, Oct. 14-16, pp: 1-5.

Tan, T.S., 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 IEEE International Conference on Intelligent and Advanced Systems (ICIAS '07), Nov. 25-28, pp: 330-334.

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 Information, Communications and Signal Processing, 3: 1423-1426.

Ting, C.M., S.H. Salleh, T.S. Tan and A.K. Ariff, 2007. Automatic phonetic segmentation of Malay speech database. Proceeding of the IEEE 6th International Conference on Information, Communications and Signal Processing, pp: 1-4.