MALAY COMPUTER-AIDED SPEECH THERAPY USING AUDIO-FINGERPRINT AND VISUALIZATION

Nor Azan Mat Zin[1], Seyed Yashar Banihashem[2], Salwani Mohd Daud[3], Siti Norul Huda Sheikh Abdullah[1], and Hiroyuki Iida[4]

[1]*Universiti Kebangsaan Malaysia (UKM), Malaysia, azan@ukm.edu.my*
[2]*Universiti Kebangsaan Malaysia (UKM), Malaysia, yasharbanihashem@gmail.com*
[3]*Universiti Teknologi Malaysia (UTM), salwani.kl@utm.my*
[4]*Japan Advanced Institute of Science and Technology (JAIST), iida@jaist.ac.jp*

ABSTRACT. Stuttering affects about one percent of the world's population. Speech therapy for stutterers involves extended interaction between the patient and a skilled speech therapist. Computer aided speech and language therapy (CASLT) help provide therapy for more patients. CASLT has two main phases - assignments of therapy to the stutterer and evaluation of progress. Current research has yet to address the progress evaluation phase. Therefore this research aims to develop an enhanced CASLT which can help user visualize his/her therapy progress. In this paper, we describe a computer aided Malay speech enhancer: a system that visualizes the speech therapy progress of stuttering Malay users. This system includes three main components. The Mel frequency cepstral coefficients, Vector quantization, and Hidden Markov model techniques are used to process the system's audio signal. The student's progress is visualized for him via extraction of his/her audio fingerprint.

Keywords: computerized speech therapy, MFCC, audio fingerprint, vector quantization, decisionfusion, stutter

INTRODUCTION

The role of a child's oral communication skills is vital as it is connected to his/her growth development. It can also be considered as a means to participate in all daily social activities in the family, amongst friends, and in school environments (Michalopoulou, 2009) Stuttering affects about 1% of the world's population including Malaysians, regardless of race, language or culture (Van Borsel et al., 2003). Speech therapy for these individuals generally involves extended interaction between a patient and a skilled speech therapist. As a result, the time and cost of providing this therapy for all stuttering students can make it impractical when large populations of students are involved (Saz et al., 2009). A computer aided speech enhancer is proposed in the current study. The target groups of this study are Malay speaking students who stutter. Hence, this study's proposed system uses the Malay language for analysis.

COMPUTERIZED MALAY SPEECH THERAPY

Prior researches and existing gaps

Assistive technologies play an important role in the lives of disabled people (Banihashem et al. (2013, 2015) and most computer-based assistive technologies can be carried everywhere

by impaired people (Banihashem, 2014). Stutterers must continue to practice speaking at home so that it becomes part of their daily routine. The two main issues i.e. "individual speech therapy cost" and "Practice observation" were the motivators for conducting this study. Providing one-to-one speech therapy is costly for families. It would be a great help if less costly alternatives or supplementary solutions are created and offered to users. Another major problem (Starbuck, 1992) with practicing outside the clinic is that speech therapists cannot ensure that clients are practicing correctly and/or consistently. Clients may practice utterances at home regularly, but if they practice incorrectly, they will not see any progress in their rehabilitation. Lack of progress causes clients to practice less frequently, resulting in a cycle of poor practice, where no improvement is achieved (Oii & Jasmy, 2008). Like most of the non-English languages, the Malay speech recognition (via computer) is still in its infancy (Fook, 2012).

To date, there are only a few published researches available, all presented various techniques to extract audio features only. Very few researches have been done in recent years in the area of Malay Computer-Aided Speech and Language Therapy (CASLT). Ooi and Jasmy (2008) is the most recent research in the area. Unfortunately, most of the Malaysian CASLT researches remain in research library or at the conceptual level. According to Ooi and Jasmy (2008), the use of computer technology in speech therapy and assessment is still new in Malaysia. Tan et al. (2007) analyzed the speech of a stutterer and then gave him therapy assignments. All CASLTs attempt to improve the speech skills of people who stutter.

A comprehensive CASLT can be divided into two main phases. In the first phase, assignments are given to the stutterer to be practiced and accomplished. Most existing CASLTs have already addressed this phase. The next phase, which most CALTSs neglect, is to evaluate the stutterer's progress after they have finished their assignments. In most cases, the second phase, which evaluates the user's progress, is delegated to the speech therapist. If the second phase is accomplished using computers, the stutterer will be less dependent on the therapist. Besides, the stuttering student can be assured that he is practicing correctly. However, in order to achieve better results, after all the assignments and evaluations are accomplished, it is necessary that a speech therapist observe the overall results. In this study, the proposed system tries to address both phases of the CALST. It should be noted that detecting stuttering is different from measuring its intensity and various computerized techniques have been developed to detect stuttering in speech (Czyzewski et al., 2003, Ravikumar et al., 2008, Chee et al., 2009).

System's Components and Relations

The proposed system's schema is shown in Figure 1. Previously, the audio fingerprint method was used in domains such as radio advertisement detection and music detection systems. The contribution of the current study is that the obtained audio fingerprint will be used in the domain of computer-aided Malay language therapy. There are three types of audio fingerprint algorithms:

Systems that use features based on multiple sub-bands, such as Philips' Robust Hash algorithm, which is reported to be very robust against distortions (Wang, 2003).

Systems that use features based on a single band such as the spectral domain.

Systems using a combination of sub-bands or frames, which are optimized through training. In the current study, the system will use features based on multiple sub-bands to create audio fingerprints.

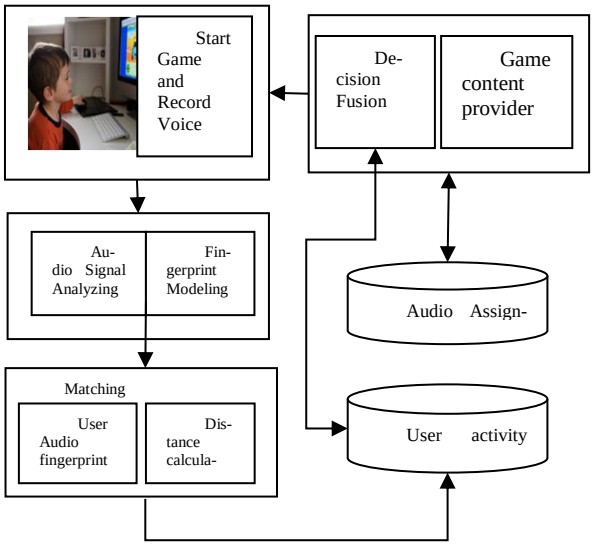

Figure 1. The Schema of a Computer-Aided Speech and Language Therapy (CASLT) System

As indicated in Figure 1, the system has two databases; one for user information and activities and the other for speech therapy exercises. The user should login to load the prepared assignments that they must accomplish. A database of correctly pronounced words is readily at hand. All these words are analyzed and the audio features and fingerprints are extracted for comparison and decision-making. During practice time, the system presents a specific word on the screen and asks the user to listen. In this way, the student will know the correct pronunciation of the word. Then, a red button on the screen will be activated. The user is asked to press this button and record his voice by pronouncing the same word three times. This scenario will be repeated for five different words. After finishing the assignment, the system will calculate the user's audio fingerprint for each of the five words.

The results will be saved in a database. In the next step, the similarity of the user's pronunciation and the pre-recorded words (correctly pronounced) will be calculated via calculation of the energy distance to the centroid point. The decision fusion part will receive the calculated results, which will include five parameters. By analyzing the calculated parameters, the decision fusion will decide the progress of the student. Results from the decision fusion are categorized into three levels i.e. "Not Good", "Good", and "Excellent". To visualize the results, an animated cat will be shown on the screen. The expression on the cat's face depends on the user's results. For example a "Good" result will make the cat smile.

Fingerprint Modeling

Figure 2 outlines the steps for extracting features and modeling fingerprints. Twelve sequential steps are required before the audio fingerprint can be modeled. After the recorded audio signal is normalized, it is then divided into frames. The selected frame length is 20 ms because the speech signal is assumed to be stationary (pseudo-stationary) over a period of about 20–30 ms. Thus, a speech analysis typically requires segmentation of 20 ms long frames (Schafer & Rabiner, 1975).

Sometimes by framing a signal, it is possible to miss some data in the joint point of the frames. To avoid missing the data and to obtain a more consistent data analysis, an overlap-

ping method should be used. In this study, the frame overlapping size is set to 25 percent. To calculate the cepstrum in step six, a non-linear Mel frequency band mapping is used. In step seven, the results of the Mel frequency cepstral coefficients (MFCC) will be used for fingerprint modeling. In the final section, steps 11 and 12, VQ and HMM will be applied to model the fingerprint. The basic idea behind VQ is to divide a large set of data into groups that have approximately similar numbers of data. Usually, each created cluster has a centroid point similar to a k-mean.

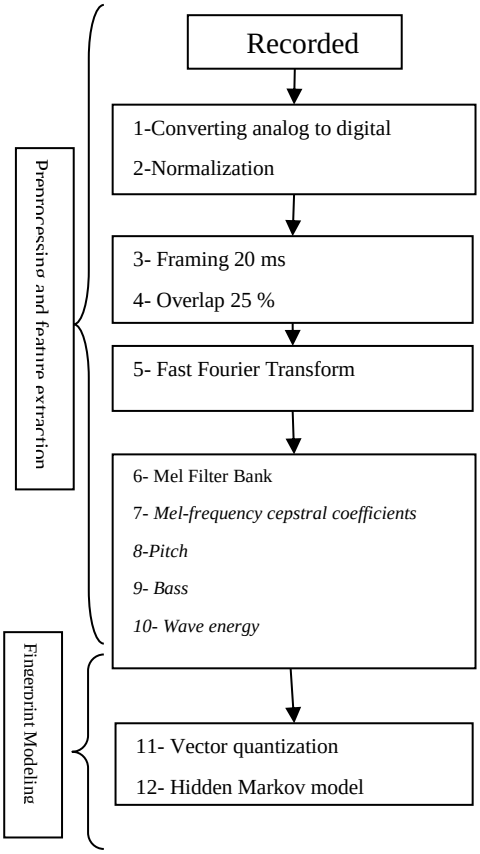

Figure 2. Audio Fingerprint Extracting Steps

The data in VQ are presented via index of their distance to the centroid point. In addition, the centroid point plays an important role for judging similarity. Usually, HMM is used to carry out comparison steps between two speech patterns. One of the reasons HMM is such a popular technique in speech recognition research is its ability to model the time distribution of speech signals (Solera-Ureña, 2007).

Decision Fusion

The decision fusion diagram is shown in Figure 3. It takes five input parameters to make a decision regarding the similarity level and content of the next assignment.

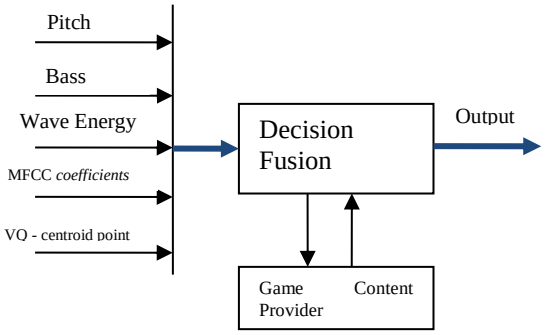

Figure 3. Decision Fusion Module

To implement VQ, the Linde-Buzo-Gray algorithm is used. Figure 4 indicates the sample code of prototyping using Matlab. In the code, "res" is a VQ codebook with the number of required centroids; "CW" is correctly pronounced as word vector and "Centro" is the number of centroid points.

```
function res = vqlindouzogray(CW, Centro)
eq = 0.01;
res = mean(CW, 2);
dataper = 10000;
for i = 1:log2(Centro)
    r = [res*(1+eq), res*(1-eq)];
    while (1 == 1)
        a = distance(CW, res);
        [m,th] = min(a, [], 2);
        tm = 0;
        for j = 1:2^i
res(:, j) = mean(CW (:, find(th == j)), 2);
x = distance (CW (:, find(th == j)), res(:,
j));
for b = 1:length(x)
                tm = tm + x(b);
```

Figure 4. The Methods Of Implementing VQ For A System

Results

The system prototype is not yet completed. An example of results, obtained from prototyped parts of the system is shown in Figure 5.

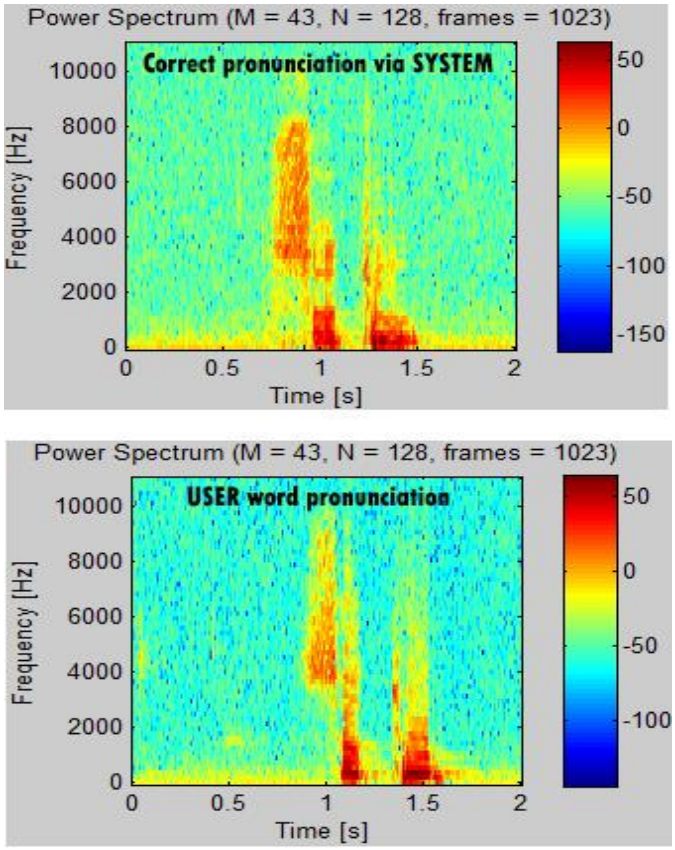

Figure 5. An Example of the Power Spectrum for the Word "SATU" Analyzed via the System Prototype

The power spectrum for word "SATU", which means "one" in the Malay language is presented in Figure 5. As indicated in the figure, there is an overall similarity between pronunciations of the system and the stuttering student (user). However, they are still not exactly similar, so the system calculates these differences via the decision fusion part. The result will be presented to notify the user of his progress in speech therapy. The system should be precisely tuned to analyze both the pronunciation similarity and differences at the same time. By considering this system analysis, speech therapists can provide more effective exercises to the stuttering student.

CONCLUSION

Once the complete system is prototyped, it will be evaluated with more stuttering students. The vocabulary domain will also be increased based on the suggestions of the speech therapist. A portable version of this system would prove useful for stuttering students. In the future, it might be possible to develop an Android or an iOS version of this system.

ACKNOWLEDGMENTS

The authors would like to thank the Ministry of Education, Malaysia, for funding this project under the fundamental research grant FRGS/1/2014/ICT05/UKM/02/1.

REFERENCES

Michalopoulou. (2009). Oral language in preschool education. *Theoretical approaches and didactical applications*. Thessaloniki, Greece: Epikentro Publisher.

Van Borsel, J., Achten,E., Santens,P., Lahorte,P., and Voet, T. (2003). fMRI of developmental stuttering: a pilot study. *Brain and language*, 85, 369-376.

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

Banihashem, Seyed Yashar, Nor Azan Mat Zin, Noor Faezah Mohd Yatim, Norlinah Mohamed Ibrahim. (2013). Real Time Break Point Detection Technique (RBPD) in Computer Mouse Trajectory. *TELKOMNIKA*, 11(5), 2710-2715.

Banihashem, S. Y., Shishehchi, S., Zin, N. A. M. and Yatim, N. F. M. (2011). Accessible targets for motion impaired users with Hidden Click Zone technique. *Proceedings of the International Conference on Pattern Analysis and Intelligent Robotics (ICPAIR)*, 188-191.

Banihashem, S. Y., N. A. Mat Zin, N. F. Mohd Yatim, and N. Mohamed Ibrahim, (2014). Improving mouse controlling and movement for people with Parkinson's disease and involuntary tremor using Adaptive Path Smoothing technique via B-Spline. *Assistive Technology: The Official Journal of RESNA*, 26(2), 96-104. DOI: 10.1080/10400435.2013.845271.

Starbuck, H. B. (1992). *Therapy for stutterers*. Stuttering Foundation of America. Memphis, Tennessee,.

Oii. C. A and Jasmy Yunus (2006). Computer-based System to Assess Efficacy of Stuttering Therapy Techniques. *Proceedings of The 3rd Kuala Lumpur International Conference on Biomedical Engineering*, 374-377.

Fook, C., Hariharan, M., Yaacob, S. and Adom, A. (2012). A review: Malay speech recognition and audio visual speech recognition. *Proceedings of the International Conference on Biomedical Engineering (ICoBE)*, 479-484.

Tan, T.-S., Ariff, A., Ting, C.-M.and Salleh, S.-H. (2007). Application of Malay speech technology in Malay speech therapy assistance tools. *Proceedings of the International Conference on Intelligent and Advanced Systems*, *ICIAS*. 330-334.

Czyzewski, A., Kaczmarek, A. and Kostek, B. (2003). Intelligent processing of stuttered speech. *Journal of Intelligent Information Systems*, 21, 143-171.

Ravikumar, K., Reddy, B., Rajagopal, R. and Nagaraj, H. (2008). Automatic Detection of Syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies. *Proceedings of World Academy of Science: Engineering & Technology*, 48, 270-273.

Chee, L. S., Ai, O. C., Hariharan, M. and Yaacob, S. (2009). MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA. *Proceedings of IEEE Student Conference on Research and Development (SCOReD)*, 146-149.

Wang, A. (2003). An Industrial Strength Audio Search Algorithm. (2003). *ISMIR*, 7-13.

Schafer R. W. and Rabiner, L. R. (1975). Digital representations of speech signals. *Proceedings of the IEEE*, 63, 662-667.

Solera-Ureña, R., Padrell-Sendra, J., Martín-Iglesias, D., Gallardo-Antolín, A., Peláez-Moreno, C.and Díaz-de-María, F. (2007). SVMs for automatic speech recognition: a survey. *Progress in nonlinear speech processing*, 190-216, ed: Springer