SPOKEN MALAY LANGUAGE INFLUENCE ON AUTOMATIC TRANSCRIPTION AND SEGMENTATION

Husniza Husni[1], Yuhanis Yusof[2], and Siti Sakira Kamaruddin[3]
[1,2,3]*Universiti Utara Malaysia, Malaysia*
[1]*husniza@uum.edu.my,* [2]*yuhanis@uum.edu.my,*[3]*sakira@uum.edu.my*

ABSTRACT. The influence of Malay language into modeling a Malay speech lexicon can be potentially useful for a more accurate transcription and segmentation. The problem arises when trying to discriminate the boundaries between similar sounding phonemes for segmentation, especially in dyslexic children's speech when reading, which have been influenced by the surrounding phonemes (before and after) thus making it harder to distinguish. Hence, this paper explores the need to model spoken Malay into the read speech lexical model that takes into consideration context-dependent model. By modeling spoken Malay language into the lexical model, better transcription can potentially be achieved with regards to the speech data with highly phonetically similar reading errors.

Keywords: Malay automatic speech transcription, Malay phonetic segmentation, Malay language system, accuracy of transcription and labeling

INTRODUCTION

It has been well established that human transcribers always outperform automatic transcriptions (Vasilescu, Yahia, Snoeren, Adda-Decker, & Lamel, 2011, Cucchiarini & Strik, 2003, Lippmann, 1997). However, this advantage comes with a shortfall – that humans tend to make mistake and variability in human transcriptions exists even for the same human transcriber transcribing the same word at different times (Cucchiarini & Strik, 2003). Thus, the need for an automatic approach arises especially when dealing with large vocabulary speech processing. In the effort of enlarging the vocabulary of speech recognition for dyslexic children's reading, the automatic approach seems promising and worth exploring. It is rather difficult and very time consuming to transcribe the children's read speech that contains phonetically similar errors consistently given the variability in human transcriptions and subjective evaluation.

Automatic transcription and segmentation very much depend on automatic speech recognition engine. For dyslexic children's reading, an automatic speech recognition engine has been developed using a carefully modeled context-dependent lexical model to improve the recognition accuracy (Husniza, 2010). Earlier works have shown that this speech recognition engine could recognize dyslexic children's readings, which is full of highly phonetically similar errors, with 75% accuracy (Husniza & Zulikha, 2010a; Husniza & Zulikha, 2010b). However, given the nature of their reading, a high false alarm rate of 15% might be an issue for its implementation to transcribe and segment read speech. Initial findings suggest that it can be used with acceptable agreement to human transcriptions but the findings also suggest that it should be improved for a more accurate segmentation (Husniza, Yuhanis, & Siti Sakira, in press). Thus, this paper aims to investigate the matter further by

looking into the lexical model inspired by the work of Nouza and Silovsky (2010) who suggested that the transcription can be better if the language in the context is taken into consideration.

INITIAL INVESTIGATION

Usually when reading, people tend to read using formal Malay, one that is standardized accordingly to remove the variability in pronunciations. However, we cannot simply ignore the fact that spoken Malay does have its influence in articulating Malay words, especially common words. For children who cannot really read well (or hardly can read), guessing the pronunciation and substituting them with existing knowledge of spoken Malay language that they acquire, can be a fact that needs careful modeling of the lexicon so that accuracy can be improved. Hence, this section discusses on the nature of Malay language and the initial experiment conducted to test the idea.

Malay Language and Dyslexic Children

Malay, although a straight forward language with less complicated spelling system, is still a language that dyslexic children found challenging to read. It is not the language but the nature of their difficulties that challenge them to process text. It can be established that dyslexic children produce phonetically similar reading errors whatever the language might be (Husniza, 2010; Sawyer, Wade, & Kim, 1999). Shaywitz (2003) has explained in depth that the cause of such reading difficulties is due to deactivation of reading pathways in the brain of a dyslexic. The claim has been proven by studies that examined the fMRI images of dyslexic children when reading (Singleton, 2006; Shaywitz, 2003). Apparently, dyslexics (adults and children) activated the wrong parts of the brain for reading creating the difficulties in processing text information. That is why they tend to make phonological errors as the phonological processing of the brain is not activated. Thus, learning to read is a major challenge as reading involves phonological processing to associate phonemes with the corresponding graphemes.

Having only six vowels (a, e, ê, i, o, and u), Malay words are made up of combination of consonant, C and vowels, V in a syllable as the basic unit of a word. There are 23 syllable patterns all together that ranges from simple combination to more complex pair of CV. Examples include the word *aku* and *ibu* for V+CV group and *saya* and *kamu* that fall under the CV+CV group. The CV pair can be more complicated (e.g. CV+CVCC+CVC) and sometimes also includes digraphs and diphthongs, and the one that most dyslexic children fail to read correctly – the vowel pair group (*paus*, *saut*, *puas*). In terms of word pronunciation, the sound of each phonemes very much depending upon its surrounding context. Although Malay has a simple spelling system, i.e. the words are written/spelled just like it is uttered (Swan & Smith, 2001), that is not always the case. The word *betul* and *kampung* for example, aren't pronounced with the 'u' sound, instead we say or read the words with the 'o' sound and it is still regarded as correct pronunciation and reading. This common style of pronunciation is adapted to reading task as well as native speakers normally read the two words by substituting the 'u' sound with the 'o' sound and so do the dyslexic children.

Initial Experiment

An experiment was conducted to see how accurate the segmentation is on a few randomly selected speech data of dyslexic children's read speech. The sample test data consists of a collection of 100 speech signals from ten different readers, age between seven to eleven years old. The lexicon involves eight randomly selected simple, common Malay words namely *"abang", "aku", "apa", "betul", "bunga", "baca", "makan", and "umur"*. Of course

having more words would have given a wider range of the findings, however this is an initial experiment to explore how good the transcription and segmentation is given the lexical model for further enhancement and development.

The experiment was a straight forward procedure – the speech signals are fed through the force alignment algorithm that takes the lexical model as the base of its recognition and alignment. The process outputs the text transcription and the phonetic segmentation to be measured against the manual transcription and segmentation of the same speech. Force alignment will use the information contained in the lexical model and the speech recognition engine to search for potential match of an input speech signals (cut into strips of frames of 10 ms each) using the speech recognition engine.

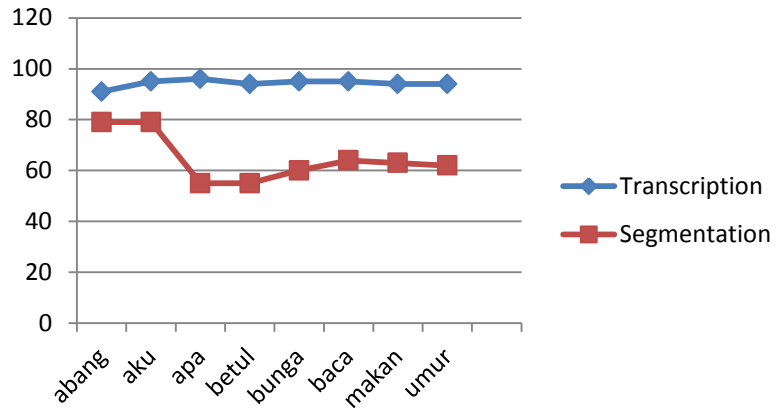

Figure 1. Transcription and segmentation results for the words.

Referring to Figure 1, the transcription scored above 90% similarity with the manual transcription and thus can be considered promising. Consequently, the segmentation however showed variant diversity with most of them scored lower than the acceptable agreement level between 72% to 80% (Cucchiarini & Strik, 2003).

DISCUSSION

Apparently, it is rather interesting to note that the word with less variability in pronunciation, i.e. not really influenced by Malay spoken language, scored the highest segmentation similarity percentage with 79%. These words are "*abang"* and "*aku"*. In spoken Malay, the pronunciations of the two words are phonetically the same as they are spelled. Hence, the lexical models for the words are concrete enough to model the two. The other words, for example "*bunga"*, "*umur"*, and "*betul"* are likely to be articulated with the influenced of spoken Malay. *"Bunga"*, to take as an example, could be uttered by the children without the blending of the 'ng' sound (the digraph) hence producing somewhat different pronunciation. The pronunciation of "*umur"* is also influenced by spoken Malay where the 'r' is somehow silent (for the middle and southern peninsular speakers) or is replaced by another sound similar to the 'q' sound (for the northern speakers). Interestingly, the word "*betul"* is never read nor spoken as it is spelled, but rather the sound of 'u' is replaced with the sound of 'o' quite consistently.

The findings triggers the idea that modeling spoken Malay into the lexical model could help improve the performance as have been discussed by Nouza and Silovsky (2010), who change the spelling of certain words to adapt to the spoken attributes. This means that the lexical model should not only include the context-dependent representation but also

incorporate the variability of spoken Malay into the lexical model. Figure 2 illustrates the idea further.

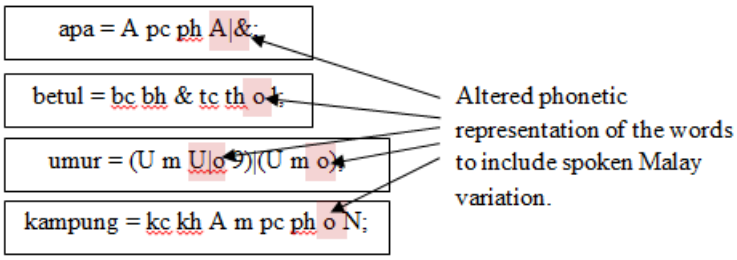

Figure 2. Conceptual design of lexical *apa, betul, umur,* and *kampung*.

Consequently, we will consider the pronunciation model for spoken Malay of more complex words (CV pair) to model the dyslexic children's read speech. In future, the spoken Malay will be modeled into the lexical model allowing the speech recognition engine to adapt to the variability of pronunciation of a word and potentially improve automatic segmentation of the highly phonetically similar speech data.

CONCLUSION

This paper explores the influenced of Malay language, in particular spoken Malay, on the accuracy of automatic transcription and segmentation of dyslexic children's read speech. Although the speech data are all obtained from read speech, the influence of spoken Malay is apparent where early findings suggest that it could causes the segmentation accuracy to fall below the agreement levels. It is therefore can be concluded that a lexical model, which have been modeled with context-dependent pronunciation model, could be enhanced taking into consideration the influence of spoken Malay to leverage the segmentation accuracy at least up to the minimum agreement level. Reaching the agreement level of at least 72% for most of the speech corpus is important to establish a stronger and better automatic transcription and segmentation, especially for children's speech known to be most challenging in the realm of automatic speech processing.

ACKNOWLEDGMENTS

This work is financially supported by the Ministry of Higher Education Malaysia under the Exploratory Research Grant Scheme (ERGS).

REFERENCES

Cucchiarini, C., & Strik, H. (2003). Automatic phonetic transcription: An overview. In *Proceedings of the International Conference of Phonetic Science*, Barcelona, 347-350.

Husniza, H. (2010). *Automatic Speech Recognition Model for Dyslexic Children Reading in Malay*. PhD Thesis, Universiti Utara Malaysia.

Husniza, H., Yuhanis, Y., & Siti Sakira, K. (in press). Evaluation of Automated Phonetic Labeling and Segmentation for Dyslexic Children's Speech. In *Proceedings of World Congress on Engineering*, London.

Husniza, H., & Zulikha, J. (2010a) Improving ASR performance using context-dependent phoneme models. *Journal of Systems and Information Technology (JSIT), 12*(1), 56-69.

Husniza, H., & Zulikha, J. (2010b). Minimizing word error rate in a dyslexic reading-oriented ASR engine using phoneme refinement and alternative pronunciation. In *Proceedings of International Conference on Education and New Learning Technologies EDULEARN'10*, Barcelona, Spain.

Lippmann, N. (1997). Speech recognition by machines and humans, *Speech Communication, 22*(99), 1-15.

Nouza, J., & Silovsky, J. (2010). Adapting lexical and language models for transcription of highly spontaneous spoken Czech, *Text, Speech and Dialogue, Lecture Notes in Computer Science, 6231*, 377-384.

Sawyer, D. J., Wade, S., & Kim, J. K. (1999). Spelling errors as a window on variations in phonological deficits among students with dyslexia. *Annals of Dyslexia, 49,* 137-159.

Shaywitz, S. (2003). *Overcoming Dyslexia: A New and Complete Science-based Program for Reading Problems at Any Levels*, New York: A. A. Knopf Distributed by Random House.

Singleton, C. (2006). *Computer and Dyslexia: Implications for policy and practice*, United Kingdom: Dyslexia Computer Resource Centre.

Swan, M. & Smith, B. (2001). *Learner English: A teacher's guide to interference and other problems*. Cambridge New York: Cambridge University Press.

Vasilescu, I., Yahia, D., Snoeren, N., Adda-Decker, M., & Lamel, L. (2011). Cross-lingual study of ASR errors: On the role of the context in human perception of near-homophones. In *Proceedings of the INTERSPEECH*, Florence, Italy, 1949-1952.