Evaluation of Automated Phonetic Labeling and Segmentation for Dyslexic Children's Speech

Husniza Husni, Yuhanis Yusof, and Siti Sakira Kamaruddin

***Abstract*—Phonetic labeling and segmentation have one major outback – they are time consuming, erroneous, and tedious if done manually. Although manual labeling and segmentation are always the best, automated approach is potentially promising as alternative approach for a more efficient process. In an attempt to automatically label and segment dyslexic children's read speech, this paper investigates whether or not the automated approach can be as accurate as compared with the manual one. This is due to the highly phonetically similar reading errors produced when they read that have affected automatic speech recognition (ASR). In this work, experiments were performed using a specifically designed ASR to force-align the read speech and produce the labels and segmentations automatically. The CSLU toolkit's force alignment algorithm has been employed to measure their performances. Selected speech data of dyslexic children's reading in Malay were fed to the algorithm as input and the evaluation resulted in 95% agreement on phonetic labeling and only 65% on segmentation with respect to the manual ones.**

***Index Terms*—automatic phonetic labeling, automatic transcription, speech recognition, dyslexic children reading.**

I. Introduction

MANUAL label and segmentation of speech signals for processing is known to be time consuming, tedious, and costly. Hours are taken by human transcriber to phonetically transcribed and label the speech. Therefore, there has been the need to perform this process automatically. Efforts to perform it automatically have been reported as evidenced in studies such as [1-4]. Most of the studies work with spontaneous speech and with large corpora. Thus the need to perform it automatically becomes more apparent. Another important factor for having an automated approach is that human labeling and segmentation is often error prone due to fatigue or different perception of the speech. Human transcription tends to be influenced by inter-subject and intra-subject variation that requires repeated measurements of the same speech to counter for its differences (which usually differs from each other too!) [4].

Manuscript submitted March 5, 2013. This work is supported by the Ministry of Higher Education Malaysia under the Exploratory Research Grant Scheme (ERGS).

Husniza Husni, Yuhanis Yusof, and Siti Sakira Kamaruddin are with the School of Computing, College of Arts and Sciences, Universiti Utara Malaysia (Husniza, H. is the corresponding author and can be reached by phone: +604-928-7074; fax: +604-928-4753; e-mail: husniza@ uum.edu.my).

Automatic phonetic labeling and segmentation can be performed in two ways: the first approach is by means of phone recognition and the second approach is by forced alignment [4]. Phone recognition involves mapping of speech to phone without relying on existing lexical model and is mainly used for generating or exploring a new acoustic model [5]. In this work, we opt to use the second method, which force aligned the speech using an automatic speech recognizer (ASR) that is built on specific design of lexical model to cater for the varieties of phonetically similar errors produced by dyslexic children when reading in Malay.

Our main objective is to explore whether or not the ASR could perform the task satisfactorily. The automated phonetic labels and segments are useful in the development of ASR for dyslexic children, where manual labeling and segmentation can be removed entirely. Since it could reduce time and cost and alleviate human's error prone labeling and segmentation, the automated ones are beneficial to be used in speech synthesis and in linguistic research.

II. The ASR and the Lexical Model

To achieve the objective, we used an existing ASR, which has been trained on dyslexic children's read speech. The ASR has been trained using lexical model that has been specifically designed for dyslexic children, where it includes selected words in Malay with their four most frequent reading errors namely, vowel substitutions, consonant deletions, nasals, and consonant substitutions [5]. The aforementioned most frequent reading errors were obtained in order to model them into a lexical model for training an ASR. In this case, the speech data were obtained from dyslexic children reading 114 selected Malay words. The words were selected randomly from the standard school syllabus and represent 23 syllable patterns covered by the syllabus. Syllable patterns involve different combination of consonant, *C* and vowel, *V* in a word. Take the word *bunga* for example; it belongs to the pattern of *CV+CV with digraph*. Digraph refers to a single sound made by two successive letters or consonants, in this case the letter 'n' and 'g'. Figure 1 illustrates each of the four most frequent errors made when reading *bunga*, as an example (the word *bunga* in Malay means 'flower' in English).

The words, grouped in different syllable patterns, were used as stimuli to obtain dyslexic children's read speech. The syllable patterns ranges from easy to slightly more complex combination of *C* and *V*. Since reading is a problem for these children, it is noticed that even simple syllable patterns such as *V+CV* and *CV+CV* with common, everyday

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

words (e.g. *aku* and *saya*) were challenging that they could read and make mistakes, phonetically.

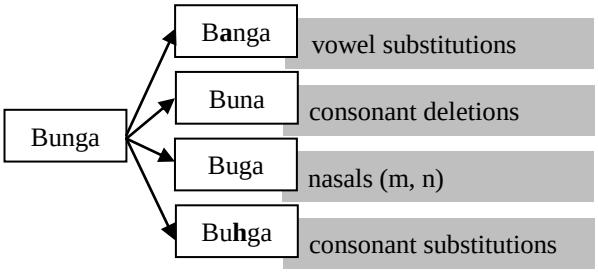

Fig. 1. The errors modeled in the lexical model for dyslexic children's reading. Note that the first example is the case of vowel substitution where the letter 'u' is replaced with letter 'a'. The second error is when 'g' is deleted from the word; the third error occurred when 'n', which is a nasal is omitted; and the last error is when 'n' is confused and replaced with 'h'.

Dyslexic children suffer a condition called dyslexia, a specific learning difficulties that mostly impedes reading abilities among children, as well as adults. Dyslexia is mainly caused by the problem in processing information (text, print, and anything related) in brain. Apparently, as have been proven by fMRI images, dyslexics are using different parts of the brain to process information when reading [6]. Hence, creating somewhat unique difficulties for them to read since reading involve the ability to correctly associate a phoneme with its corresponding grapheme. Broken link of this association results in producing incorrect sound for a particular grapheme (e.g. confusing the sound that the letter 'b' and 'p' make). Similar looking letters make it even more difficult for them to recognize and associate.

Due to their reading difficulties, mainly because of phonological deficits [7-11], dyslexic children's readings are full of mistakes. These mistakes, phonological in origin, remain a challenge to ASR as they tend to reduce its accuracy for it is difficult to recognize and differentiate between the phonetically similar sounds. Thus, the reading errors are modeled into the lexical model as phoneme refinement and adaptation to the original words. The lexical model of the word *bunga* for example, is constructed using WorldBet [12], which also includes the errors and represented as the following:

```
bunga = bc bh U N|n A|&;
```

where, `N` is the phonetic symbol of the sound of 'ng' in *bunga* and `n` is the phonetic symbol of the letter 'n'. The symbol `A` phonetically represents the letter 'a' while `&` symbol represents the letter 'e'. In this example, the errors included are classified as nasals and consonant deletion, where it involves miss-pronunciation of 'ng' in *bunga* or complete omission of the letter 'g'. Whereas the `A|&` simply denotes the pronunciation adaptation as both sounds, if spoken/read, are normally considered to be correct.

Modeling the errors into the lexical model has proved to significantly increase the accuracy of the ASR by lowering it to 25% from 30% [13]. In addition, context-dependent phonetics model is also used in the design of the lexical model, given the lexical a better representation of the actual production of speech hence, reducing the WER [13, 14].

III. FORCE ALIGNMENT AND EVALUATION METHOD

A. Force Alignment

With the existing ASR, force alignment is performed on the read speech for the purpose of evaluation. Force alignment is an automated approach to labeling and segmenting speech. The Viterbi algorithm is used to force align the speech in order to produce the output, which in the form of phonetic labeling and segmentation. This algorithm works by finding the most probable path or sequence through hidden states in order to look for the best solution.

The speech samples were fed to the ASR. Here, the Viterbi algorithm search for the most probable solution and outputs the maximum likelihood that a particular state is representing the input fed through it. The states represents all the phonemes involved or modeled in the ASR, hence the lexical model is used as one of the input to supply all the required sequence of potential words on the list. Supposed a speech input is fed to ASR to force align and retrieve its phoneme sequence that make up the word *baca*, for example. Of course, the expected output would be the phoneme sequence as the following: `bc bh A tS A`, which makes up the word.

The force alignment is performed using CSLU Toolkit [14]. The toolkit's force alignment is a straight forward process where the algorithm is executed by giving these files as inputs: the trained ASR, its specification files, lexicon file, and a speech file of which we like to automatically label and segment the phonemes. The output of this process is a file that stores the phonetic labels and segmentation of the speech as shown in Fig. 2. Since it is an automatic approach to obtain the desired phonetic labels and segmentation, which reduce time exponentially, the question is whether or not the output, i.e. the phonetic labels and segmentation are as accurate as the ones transcribed by a human transcriber? How are we going to measure it before we can actually use it further?

```
MillisecondsPerFrame: 1.000
END OF HEADER
0 1670 .pau
1670 1710 m
1710 1950 A
1950 1960 l
1960 2230 U
2230 2290 m
2290 2450 A
2450 2470 tc
2470 2530 t
2530 2790 .pau
```

Fig. 2. The output of force alignment. This is an output file that contains the phonetic labels for phonemes within the word *maklumat* (means *information*). Notice that there are three columns – the first column is the start time and the second column is the end time of a phoneme. This represents a segment in the input speech file. The third column is the automatic phonetic label generated.

B. Evaluation of Automatic Approach

Automatic phonetic labeling needs to be measured for its accuracy and normally, it is measured in reference to human labeling. Thus, human phonetic labels, although previously described as potentially error prone and tedious, are still being used as the benchmark data [4, 16]. This is due to the widely accepted assumption in the field that human labels are always better than the automated ones. Nevertheless,

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

researchers have also identified issues regarding human labeling as reference to measure against the automated ones [4]:

- Variation in human transcriptions
- Lack of reference data for evaluation (especially when it comes with Malay and dyslexic speech data)

What has been and still currently a practice is that researchers tend to use arbitrary human transcribers [4]. There have been effort to use more than one transcribers (9 human transcribers) whose tasks was to judge whether or not a phone was present in a speech file but it can be concluded that the results suggested variations where the agreement was below than 53% [4]. So, even with human transcribers, exact and accurate transcription is not always guaranteed. Sometimes, when it comes to highly phonetically similar errors, even the same human transcriber transcribe differently. Thus, we ask the questions – How can we evaluate the automatic labeling and segmentation? How do we know whether the result achieved is satisfactory?

To avoid the aforementioned issues, the reference phonetic labeling and segmentation used are the ones that have been agreed by at least two human transcribers, as have been performed by some researchers [4, 17]. In this case, only the ones with the same transcriptions are considered, though varied slightly in terms of time aligned phoneme segmentations.

To evaluate, a Java program was developed that takes manual and automatic phonetic labeling and segmentation as input and outputs a similarity percentage. It measures two different similarities – one is the similarity between manual and automated phonetic symbols; another is the similarity of their segmentation boundaries (start time and end time of each phoneme in a particular word) with respect to time. The justification behind separating the two is because we want to see how much the lexical model affected in generating automated phonetic symbols given the nature of the read speech that is highly with phonetically similar errors. However, this does not mean that evaluating the similarities of the segmentation for each phoneme in a word is less important. In transcription task, both are equally important.

For the purpose of evaluation, 101 speech data were selected, which phoneme labeling and segmentation have been manually created. These data were force aligned to obtain their automated phonetic labeling and segmentation. Thus, evaluation needs to be performed to measure whether or not they are satisfactory. The following section discusses the evaluation results.

IV. Results and Discussion

The results showed a promising percentage of automated phonetic labeling, however not so much on segmentation with respect to the reference. The acceptance percentage of human transcriber is between 76-84% [2]. The resulting percentage from the evaluation in average is 95% for the phonetic labels and only 65% in average for the segmentation of the phonetic labels. Table 1 presents a snippets of the results obtained consisting the words *abang* (older brother), *aku* (I or me), *apa* (what), *baca* (read), *betul* (correct), *bunga* (flower), *makan* (eat), and *umur* (age).

TABLE I
THE SNIPPET OF THE EVALUATION RESULTS

Word	Phonetic Label	Phoneme Segmentation
abang	91.0%	79.0%
aku	95.0%	79.0%
apa	96.0%	55.0%
baca	94.0%	64.0%
betul	95.0%	55.0%
bunga	95.0%	60.0%
makan	94.0%	63.0%
umur	94.0%	62.0%

From Table 1, it can be concluded that the phonetic labeling performs better than that of phonetic segmentation. The results indicated that the phoneme segmentation differ 35% from the manual transcriptions (i.e. 65% agreement) thus resides lower than the usual acceptance rate between human transcribers. To clearly visualize how they differ, Fig. 3 illustrates a sample of comparison between manual and automated phonetic labeling and segmentation.

Referring to Fig. 3, the manual transcription transcribed the speech signal into a sequence of phonemes `A bc bh A N` (correct transcription). However, in this case, the automated transcription transcribed it into a sequence of `A pc ph A N`, which is read as *apang*, which carries a completely different meaning (*apang* is not a frequently used word in Malay). According to this example, the first phoneme segmentation represented by the phonetic symbol `A`, slightly differs in terms of the duration. The manual transcription segment is slightly longer whereas the automated transcription is a bit shorter by a few milliseconds. Fig. 4 compares the same example in the view of their phoneme files created as outputs of both method of transcription.

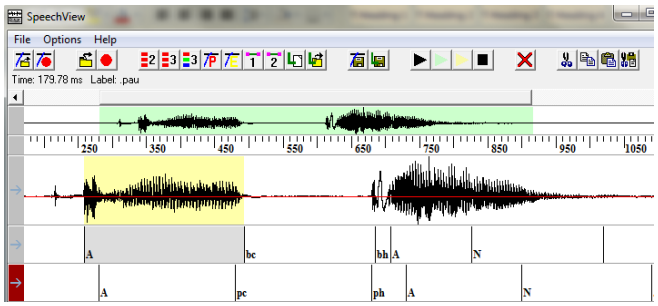

Fig. 3. The manual and automated phonetic labeling and segmentation for the word *abang*. The manual ones are presented in the third row while the automated ones are presented in the fourth row after the emphasized speech signal row (in the second row). The first row presents the same speech signal but only in smaller scale.

```
MillisecondsPerFrame: 1.0
END OF HEADER
259.134277 492.863251 A
492.863251 684.673401 bc
684.673401 707.538208 bh
707.538208 826.943237 A
826.943237 1018.753418 N
```

```
MillisecondsPerFrame: 1.000
END OF HEADER
0 280 .pau
280 480 A
480 680 pc
680 730 ph
730 900 A
900 1090 N
1090 1490 .pau
```

Fig. 4 Manual (left) and automated (right) phonetic transcriptions. Notice the duration of symbol `A` in both transcriptions – the start and end points are given in the first and second columns. The start and end points for both methods of transcriptions slightly differ by approximately 20.87 ms at start and 12.86 ms at end time. The `.pau` denotes the pause (which in this case refers to no significant wave signal detected).

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

This is the real challenge when dealing with dyslexic children's read speech – difficulties to differentiate between highly phonetically similar errors, in this case discriminating between the letter 'b' and 'p' (denoted by phonetic symbols `bc bh` and `pc ph` respectively). Even human transcribers make mistakes when dealing with these sorts of reading errors while transcribing since the nature of articulation to produce such sound is somewhat similar.

Even though the agreement of phonetic symbols is promising with 95% as mentioned earlier, the segmentation is lower than the acceptance rate. However, this is as expected. Deciding the boundary for each phoneme is no trivial task that even human transcribers' transcriptions could differ. As shown in Fig. 3 and Fig. 4, the segmentation difference is not that large but our assumption is that it can affect the accuracy of automated transcriptions generated.

Nevertheless, this result shows existing ASR can be used to force align the speech and obtain the corresponding phonetic labels with their segmentations. Even though the ASR's WER is 25% with a slightly higher FAR, the automated phonetic labeling and segmentation seem to be independent of the accuracy of the ASR. Thus, this finding conforms to the claim made in [18] that lower WER does not always guarantee better and satisfactory rate of transcriptions.

V. CONCLUSION

Phonetic labeling and segmentation have been a challenge to researchers, particularly in linguistics and in speech recognition as well as speech synthesis. Due to the time consuming, tedious, and erroneous process of manual labeling and segmentation, automated approach has been used as alternative. However, manual phonetic labeling and segmentation still is being regarded as the best and thus being a reference to the automated ones. Hence, the aim of this paper is to explore the performance of the automated approach, using force alignment, in terms of producing automated phonetic labeling and segmentation for dyslexic children's read speech. For that, CSLU toolkit's force alignment algorithm is used with an existing ASR trained on a lexical model for dyslexic children's read speech. To evaluate, the speech data are fed to the algorithm as inputs. Results have shown that the automated phonetic labeling generated are 95% in agreements with the manual ones. However, the automated segmentation of phonemes differs 35% from the manual ones. The results show that the automatic approach could potentially be used to automatically transcribe dyslexic children's speech with certain tolerance on its discrepancy on the segmentation boundaries.

REFERENCES

[1] M. Garcia & R. J Gonzales, "Automatic phonetic transcription by phonological derivation," in *Proc. of the 10th International Conference on Computational Processing of the Portuguese Language*, 2012, pp. 350–361.

[2] S. Chang, L. Shastri, & S. Greenberg, "Automatic phonetic transcription of spontaneous speech (American English)," *Proc. of International Conference on Speech and Language Processing*, 2000.

[3] C. Van Bael, W. Strik, H. van dan Heuvel, "Application-oriented validation of Phonetic Transcriptions: Preliminary results," *Proc. of 15th International Congress on Phonetic Sciences*, pp. 1161–1164, 2003.

[4] C. Cucchiarini & H. Strik, "Automatic Phonetic Transcription: A Review," in *Proc. 15th International Congress on Phonetic Sciences*, Barcelona, 2003, pp. 347–350.

[5] H. Husniza & J. Zulikha, "Dyslexic children's reading pattern as input for ASR: Data, analysis, and pronunciation model," *Journal of Information and Communication Technology (JICT)*, vol. 8, pp. 1–13, 2009.

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

[7] J. Frost, "Phonemic awareness, spontaneous writing, and reading and spelling development from a preventive perspective," *Reading and Writing: An Interdisciplinary Journal*, vol. 14, pp. 487 – 513, 2001.

[8] I. Lundberg, "The computer as a tool of remediation in the education of students with reading disabilities: A theory-based approach," Learning Disability Quarterly, vol. 18, no. 2, pp. 88–99, 1995.

[9] S. E. Shaywitz, " Dyslexia," *Scientific American*, pp. 98-104, 1996.

[10] M. J. Snowling, *Dyslexia* (2nd ed.), UK: Blackwell Publishers, 2000.

[11] J. Ziegler, "Do differences in brain activation challenge the universal theories of dyslexia?" *Brain and Language*, vol. 98, pp. 341–343.

[12] J. L. Hieronymus, *ASCII phonetic symbols for world's languages: Worldbet* (Tech. Rep.). NJ, USA: AT&T Bell Labs, 1993.

[13] H. Husniza & J. Zulikha, "Improving ASR performance using context-dependent phoneme models," *Journal of Systems and Information Technology (JSIT)*, vol. 12, no. 1, pp. 56–69, 2010.

[14] H. Husniza & J. Zulikha, "Pronunciation variations and context-dependent model to improve ASR performance for dyslexic children's read speech," in *Proc. of International Conference on Computing and Informatics (ICOCI 2009)*, Kuala Lumpur, Malaysia.

[15] S. Sutton, R. A. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon et al., "Universal speech tools: The CSLU toolkit," *Proc. of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, vol. 7,pp. 3221-3224, 1998.

[16] S. Chang, L. Shastri, & S. Greenberg, "Automatic phonetic transcription of spontaneous speech (American English)," *Proc. of INTERSPEECH* 2000, pp. 330–333, 2000.

[17] C. Cucchiarini & D. Binnenpoorte, "Validation and improvement of automatic phonetic transcription," *Proc. of International Conference on Spoken Language Processing ICSLP 2002*, pp. 313–316, 2002.

[18] J. M. Kessens & H. Strik, "Lower WERs do not guarantee better transcriptions," *Proc. of EUROSPEECH 2001*, pp. 1721–1724, 2001.

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)