Automatic Speech Recognition System for Malay Speaking Children

Automatic Speech Recognition system

Feisal Dani Rahman, Noraini Mohamed, Mumtaz Begum Mustafa, Siti Salwah Salim
Department of Software Engineering, Faculty of Computer Science & Information Technology,
University of Malaya, Kuala Lumpur, Malaysia
feisaldanirahman@yahoo.com , noraini.binti.mohamed@gmail.com, mumtaz@um.edu.my, salwa@um.edu.my

***Abstract*—Automatic speech recognition or ASR system in short, is the most recent innovation in human computer interaction. An ASR system recognizes human speech and transforms them into outputs such as text or any other machine readable outputs. ASR is increasingly used in various applications such as dictation system, voice or speaker recognition and so on. Despite the advancement in the development of ASR system, not many of such system are developed for children. Children today are increasingly using computers for many daily activities including for education. The lack of ASR system for children causes them to be lagging in behind adult users. One of the reasons for the poor development of ASR system for children is the difficulties of obtaining or creating the speech corpus database of children. Unlike adults, researchers find it difficult to engage children in recording process. This research aims at developing an ASR system for Malay speaking children with the use of a small speech database. The ASR system developed in this research has the ability to recognize words at 76% accuracy.**

Keywords- ASR; Malay; Children speech corpus;HMM

I. INTRODUCTION

Computer is now a technology that is beyond its original design in helping human to solve problem. Nowadays, many developers enhance the ability of computer to perform two way communications with human. The current interaction usually use keyboard and mouse as an input, and the screen monitor and printer become output by computer. However, in recent years it was discovered that it is much easier for human to use a speech as the medium of interaction with computer. Researcher believes that speech is not only a useful tool for human communications, but it also can be used to fetch data by controlling specific machine [1]. Furthermore, computer can also use human speech to identify detailed information of a person, or transcribe the human speech into readable text. The technology that can transcribe spoken language to readable text called "Automatic Speech Recognition" (ASR).

Nowadays, ASR has been widely used in various applications, as it offers reliability, security, hygienist, and easiness in human computer interaction. One example of the application was created by IBM where it allows converting spoken sentences into a letters and words, which could be display on digital paper. Another commercially available ASR is developed by Dragon System and Lernout & Hauspie, where this system can transcribe the speaker voice into readable text. ASR system can also allow users to instruct specific action to be carried out by computer or machine using speech [2].

Despite the benefit offered by ASR system to users of computers, it is unfortunate that there are very few ASR and its related applications design specifically for children [3]. It is a known fact that children nowadays are increasingly using and depending on computers in their daily lives. With little development on ASR system for children, they cannot enjoy the benefits of ASR system like the adults. The main reason for the lack in the development of ASR system for children is the complexity to build a children corpus especially for resource constrains languages such as Malay.

This paper describes the process in developing a speaker independent (SI) ASR system for Malay speaking children. This includes the resource accumulation process, acoustic model creation and testing.

II. CHILREN SPEECHES AND RESOURCES

Building the speech corpus for ASR development is non-trivial especially for children's ASR system [4]. There are many factors that makes building speech corpus of children to be complex. First of all, to store million of vocabulary requires a large data capacity, which could be expensive. Secondly, in order to achieve easy accessibility and availability, it requires hardware with great performance and good memory speed. Finally, to build a high quality speech corpus requires a great deal of time and the need for special recording equipment and studios. The recordings of speech need to be carried out in studio environment with no or little external noises that can distort the recorded speech. Recording process is also time consuming as it need to be carried out several times to obtain good quality recordings.

As the process of building speech corpuses is complex, this problem is worst of for children as it is much harder to get the children to focus during the recording process. There are only number of children speech corpuses that are available for children such as for Filipina, Mandarin, Taiwan, Spanish, and Croatia. Unfortunately, there is none available for Malay children corpus. Table 1 shows a list of some existing available speech corpus of adult and children.

TABLE I. EXISTING AVAILABLE SPEECH CORPUS OF ADULT AND CHILDREN

Corpus	Language	Classification
Pascual, & Guevara [5].	Filipino	• Mode of speech: Isolated • 1.5 hours of recordings • Reading material taken from short fiction stories and various school textbooks.
Das et al. [6]	English	• Isolated and Continuous • Recording hours: NA • The data recorded from 418 children with various reading material, which consist of isolated and continues story book.
Kazemzadeh et al [7]	English	• Isolated • 40 hours of recordings • The speech recorded from 256 children with the variety of background, which are 69% were native speakers of Spanish, 24% were native speakers of English, and 5% were native speakers of both English and Spanish.
Cleuren et al [8]	Dutch	• Continuous • 130 hours • This study aim to investigate the consistency of reading error. The children involve in this research are 400 Dutch speakers, which consist of 274 normal speakers and 126 impairment speakers.
Batliner et al [9]	British, German, Italian and Swedish	• Isolated • 65 hours • The goal of this study is to determine the effects of non-native language and to increase the performance of baselines. The children involve in this recording are 611 children.
Sandler et al, [10]	English	• Isolated

From table 1, it can be concluded that the number of speech corpus available for children is extremely limited. The lack of speech corpus for children, dampen the progress of ASR system developed for children. Many researchers believe that ASR system can assist in the area of children's education by teaching them how to read and improved their speech. Moreover, for leisure activities, ASR can help children in two-ways communicating with their toys or games, making it to be more interesting.

As such, in this research, we are proposing to develop an ASR system for Malay speaking children. The development includes the creation of a small vocabulary speech database of Malay speaking children and the speech acoustic models of the Malay speaking children.

III. EXPERIMENTAL SETUP

This section describes the experimental set up and procedures taken to built an ASR system for Malay speaking children.

A. Procedures

1) Building a speech corpus of Malay Speaking Children

The creation of the speech corpus begins with the preparation of text transcriptions or prompt scripts. A total of 390 short Malay sentences were prepared for the purpose of the recording comprises of 1,404 words (987 distinct non repeating words). From the 987 distinct words about 70% or 678 words are daily used words such as 'makan' (eat), 'duduk' (sit), and etc.

The speech data is accumulated through the recordings of selected children from schools around the Klang Valley. Six children aged between nine and twelve years old were involved in the recording process with the consent from their parents. All the children involved in the recording have the ability to speak and understand instructions in Malay.

A total of six recording session were held in recording studious with low noise environment to reduce external noise interference. All speech samples data is recorded using sound recording equipment package by Editors Keys that consists of Editors Keys Studio Series Portable Vocal Home, Studio Series SL300 USB Studio Microphone and Dual Layer Pop Filter.

Each child was seated in front of the recording tools and the microphone was placed four inch from their mouth. They have to face straight at the microphone and the prompt texts were put on display in front of the children on the laptop screen. The lingWAVES 2.5 application software was used to record speech data at the sampling rate of 48000Hz with 16 bit and 1 (Mono) recording channel.

All recorded speeches are then segmented at sentence level using lingWAVES 2.5 software. All segmented speech data is stored in waveform (.WAV) format. The speech corpus built in this research contains 390 speech samples.

2) Building the Speaker Indipendent (SI) Speech acoustic model for Malay speaking children

360 out of 390 recorded utterances are trained using the Malay language phone model. The parameter training is performed using HTK toolkit by applying HMM for building the speech acoustic model of Malay speaking children. In this research, the HMM topology applied is 5-states left-right with no skips and phone level model. The training is performed using HTK tools HCompV for initialisation and HERest for refining existing HMM parameter using Baum-Welch re-estimation and for embedded unit training..

This research applies the HMM model available in HTK toolkit for the development of the ASR system [11]. HTK tools

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors)

run on command-line style interface as to control the simplicity of HTK tools execution as well as the ability to record and document all operations and details of system construction or experimental procedure [11]. As HTK Toolkit is built on Linux/Unix environment, the Cygwin 1.7.9-1 is used to run the HTK Toolkit. Cygwin is a collection of tools that provide Linux-like environment for Windows and contains the Bash Shell and Perl scripting language.

3) Testing using out of vocabulary of children's speech

To test the ability of the ASR system developed in this research in recognizing the out of vocabulary speeches (not used for model training), the remaining 30 recorded utterances from the 390 were used. The ASR recognition result is obtained based on maximum likelihood state sequence in HMM model. The tool provided by HTK toolkit is called HVite and the result comparison analysis tool is called HResults. In order to compute recognition results, the tool used the following formulas:

$$Sentence\ correctness\ (\%) = \frac{Correct\ recognized\ sentences}{total\ \ sentences}\ x\ 100\% \qquad (1)$$

$$Word\ correctness\ (\%) = \frac{Correct\ recognized\ words}{total\ \ words}\ x\ 100\% \qquad (2)$$

IV. Results and Discussion

The performance of ASR system is measured in term of the number of test speeches that are successfully recognized by the system using the speech acoustic model of impaired speeches. For example, the speech input "Guru kelas saya ialah Cikgu Kala" (My class teacher is Miss Kala) is recognized by the ASR system as "Guru kelas saya ialah Cikgu Mala" (My class teacher is Miss Mala). The recognition accuracy is therefore 5/6 which is 83.33%.

The result of the test shows that the recognition accuracy at sentence level is 71.51%. At the word level, the ASR system recognition accuracy is 76.70%, which is based on a total of 160 words made available in the 30 test data. The word error rate (WER) of the ASR system speech acoustic model is 23.30%, which comprises of 3 deletion (1.84%), 30 substitutions (18.39%) and 5 insertion (3.07%). Among the six children, the highest recognition accuracy is 85.96% while the lowest is at 67.44% with standard deviation of 8.05%.

We have analyzed the ASR system's sentence correctness result by the speakers' age group, 9 – 10 years old and 11 – 12 years old. There are three speakers in age group 9 – 10 years old and three speakers in age group 11 – 12 years old. This analysis demonstrates that the ASR system recognize speeches from older speakers better compared to younger speakers. The reason for better recognition performance for older speakers is that as the children grow, they improved their articulation skill due to body and physiological change [3].

We have also analyzed the recognition accuracy of ASR system in term of gender. From this analysis, it can be said that speeches from female speakers are recognized by the ASR system slightly better than speeches from male speakers [12], explained that for children, female speakers' physical vocal tract grows and stabilizes earlier than male speakers, and by the age of twelve years old, the female speakers have speech frequencies similar to adult female speakers. Therefore, the recognition for female speakers' speeches can be recognized by ASR system better than male speakers' speeches [12]. In addition, female speakers produce consistent pronunciation compared to male speakers [13].

V. Conclusion

This research aims at developing an ASR system for Malay speaking children. The speech corpus developed in this research comprises the voices of six children uttering a total of 390 sentences. This speech corpus is small compared to adult speech corpus due to the difficulty in engaging children in the recording process. This is because children are easily distracted and seldom has the patience to handle repeated recording.

Nevertheless, we have managed to build a speaker independent speech acoustic model using the speeches of these six children using the HTK toolkit. The testing of the ASR system found that it can recognize of up to 76% of test words accurately. The best ASR system built for adults can recognizes speech with 70%-90% accuracy. As such the ASR system developed in this research can recognizes speech input at an acceptable level despite the use of small speech database.

The future direction of this research is to further enhance the recognition accuracy of the ASR system for Malay speaking children by performing adaptation on adult speaker's speech acoustic model. This can increase the recognition accuracy of the ASR system with the use of very limited speech corpus of children.

Acknowledgment

This research is supported by UM High Impact Research Grant UM-MOHE UM.C/HIR/MOHE/FCSIT/05 from the Ministry of Higher Education Malaysia.

References

[1] F. Jelinek, A Real-Time, Isolated-Word, Speech Recognition System For Dictation Transcription. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP 1985, pp. 858-861.

[2] Q.H. Ngo and W. Winiwarter, Building an English-Vietnamese Bilingual Corpus for Machine Translation. Asian Language Processing (IALP) 2012, pp.157 – 160.

[3] M. Gerosa, D. Giuliani and F. Brugnara, Acoustic variability and automatic recognition of children's speech, Speech Communication vol 49(10-11 pp. 847-860, 2007.

[4] M. Gerosa, D. Giuliani, S. Narayanan and A. Potamianos, A Review of ASR Technologies for Children's Speech. ICMI-MLMI 2009 Workshop on Child, Computer and Interaction.

[5] R.M. Pascual and R.C.L Guevara, Developing a children's Filipino speech corpus for application in automatic detection of reading miscues and disfluencies, TENCON 2012, pp. 1 - 6. doi: 10.1109/TENCON.2012.6412235

[6] S. Das, D. Nix and M. Picheny, Improvements in children's speech recognition performance. *Acoustics, Speech and Signal Processing, 1,* 1998 pp.433 - 436 .

[7] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P.Price, E. Andersen, S. Narayanan and A. Alwan,). TBALL data collection: The making of a young children's speech corpus,2005 pp.1581-1584.

[8] L. Cleuren, J. Duchateau, P. Ghesquière and H.V. Hamme, Children’s oral reading corpus (CHOREC): Description and assessment of annotator agreement 2008.

[9] A. Batliner, M. Blomberg, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl and M. Wong, The PF_STAR children's speech corpus 2005.

[10] U. Sandler andY. Sonnenblick, A system for recognition and translation of the speech of handicapped individuals. *Electrotechnical Conference, 1998, pp.*16 – 19

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, and G. Moore, The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department , 2006

[12] J.E. Huber, E.T. Stathopoulos, G.M. Curione, T.A. Ash, and K. Johnson, Formants of Children, Women, and Men: The Effects Of Vocal Intensity Variation. The Journal of the Acoustical Society of America, vol 106 (3 Pt 1), pp 1532-42, 1999).

[13] M. Adda-decker and L. Lamel, Do Speech Recognizers Prefer Female Speakers? In Proceedings of INTERSPEECH 2005, pp. 2205-2208.