International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:3, 2016

A Two-Stage Adaptation towards Automatic Speech Recognition System for Malay-Speaking Children

Mumtaz Begum Mustafa, Siti Salwah Salim, Feizal Dani Rahman

***Abstract*—**Recently, Automatic Speech Recognition (ASR) systems were used to assist children in language acquisition as it has the ability to detect human speech signal. Despite the benefits offered by the ASR system, there is a lack of ASR systems for Malay-speaking children. One of the contributing factors for this is the lack of continuous speech database for the target users. Though cross-lingual adaptation is a common solution for developing ASR systems for under-resourced language, it is not viable for children as there are very limited speech databases as a source model. In this research, we propose a two-stage adaptation for the development of ASR system for Malay-speaking children using a very limited database. The two stage adaptation comprises the cross-lingual adaptation (first stage) and cross-age adaptation. For the first stage, a well-known speech database that is phonetically rich and balanced, is adapted to the medium-sized Malay adults using supervised MLLR. The second stage adaptation uses the speech acoustic model generated from the first adaptation, and the target database is a small-sized database of the target users. We have measured the performance of the proposed technique using word error rate, and then compare them with the conventional benchmark adaptation. The two stage adaptation proposed in this research has better recognition accuracy as compared to the benchmark adaptation in recognizing children's speech.

***Keywords*—**Automatic speech recognition system, children speech, adaptation, Malay.

I. Introduction

SPEECH is one of the modes of interaction of human that is very effective in information transfer and is now seriously considered for human computer interaction. ASR system is a system or application that transcribes human speech into readable text or other machine readable outputs [1]. ASR system offers many benefits for computer users particularly children for their language acquisition. The ASR system enables the children to have more natural verbal interaction with computers as compared to the traditional means of interaction [2].

Despite the benefits offered by ASR system, the development of this system only focuses on adult with very few developments for children [3]. One major reason for the lack of ASR system for children is the lack in resources for developing one. This is because, ASR systems development requires adequate and good quality resources such as recorded human speech and their relevant transcriptions, which is lacking for children even for the well-resourced languages like English [3].

For developing ASR system for under-resourced languages, [1] have proposed the use of cross-lingual adaptation where the complete resources of one language can be used to develop the ASR system for another language with minimal resources [1]. For example, in [1], a well-resourced language was used to initiate or adapted for the Vietnamese ASR system. Similarly, in [4], a Malay speech synthesis system was developed using English as the source model for cross-lingual adaptation.

While cross-lingual adaptation has been adopted for adult speeches, there is no existing work on ASR systems for children using cross-lingual adaptation. This is due to the lack of freely available speech database of children, even for most of the well-resourced languages. As such, this research looks at the suitability of a cross lingual adaptation for children's ASR system using the resources of other languages.

II. Cross Lingual and Cross Age Adaptation

Developing ASR systems for under-resourced languages are non-trivial as we need to prepare resources that are previously not available (phonological systems, word segmentation, grammatical structure, unwritten language, etc.) [5]. However, acquiring these resources is difficult due to social and cultural aspects, expertise and financial support. As most of the under-resourced languages have no existing corpora, data collection becomes an integral part of the ASR development for these languages [5]. More often than not, the databases of under-resourced languages are small in vocabulary size, poor recording quality and not segmented and labelled. Cross lingual adaptation [1], [5] is a viable solution for a quick development of ASR systems for under-resourced languages. Some of the resources that can be borrowed from another language, including acoustic model, language and pronunciation model (grapheme to phoneme) [1].

Recorded speech databases of children are rare even for some of the well-resourced languages. This is because, developing a speech database for children is not an easy task as it is time consuming for collecting speech data from children. Recording sessions with children is non-trivial, which is different when dealing with adult, where it is much easier to control the recording session with the adults but not with the children. For instance, children are difficult to stay focused on the recording session, and they do may not understand the instruction. On top of that, in some countries, it

Mumtaz Begum Mustafa is with the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. (phone: +603-7967-2500; fax: +603-7957-9249; e-mail: mumtaz@um.edu.my).

Siti Salwah Salim is a Professor at the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. (email: salwa@um.edu.my).

Feizal Dani Rahman was with the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. (e-mail: feisaldanirahman@yahoo.com).

is difficult to obtain parental consents [6].

For most languages, the ASR systems development focus on adults as it is much easier to obtain recorded speech databases for adults [7]. However, ASR system trained on adult speech yields a very low recognition accuracy in recognizing children speeches. The low recognition accuracy is due to the linguistic and acoustic characteristic differences between adult and children speech features [7]. The recognition accuracy was found to be poorer for young children between the ages of six to ten [7]. To make the ASR system to be more robust in recognizing children speech, speaker adaptive training using Constrained MLLR (C-MLLR) was proposed in [7]. However, the solution proposed in [7] did not consider cross-lingual adaptation between adult and children speech. As such, little is known about the effectiveness of adult to children cross-lingual adaptation.

III. Research Method: Two Stage Adaptation

This research proposes a two stage adaptation (TSA) technique for developing an ASR system for Malay speaking children. The first stage adaptation is the cross lingual adaptation, where the large-sized English TIMIT speech database (phonetically rich and balanced database), is used as a source model to be adapted with the medium-sized Malay adults using supervised MLLR. The adaptation is supervised as both of the databases are labelled accordingly. The purpose of the first adaptation is to develop a language model of Malay that is currently not available.

The second adaptation makes use of the speech acoustic model generated from the first adaptation and adapts it with the small-sized database of Malay speaking children. The proposed solution should enable ASR systems to be developed with very limited database. Fig. 1 depicts the framework of the two stage adaptation for developing the ASR system for Malay speaking children.

In this research, we have used the existing adult database of English (TIMIT) and Malay as source database [4]. As there is no speech database for Malay speaking children, we have developed our own children speech database using very limited resources. The recording uses 150 sentences taken from [4]. The selected sentences range from three to five word length containing 680 word tokens (398 different types), 1,866 syllables and 5.859 phones. Four native Malay children utter these 150 sentences as shown in Table I.

We have conducted the recordings in a noise free environment using Editors Keys Studio Series Portable Vocal Home, Studio Series SL300 USB Studio Microphone and Dual Layer Pop Filter, attached to a portable booth. The microphone is of high quality to reduce external noise such as the breathing sound of the speakers and placed four inches from the mouth. The recording uses lingWAVES 2.50.0042 software at the sampling rate of 16 KHz with 16 bit and 1 (mono) recording channel and saved in .wav format.

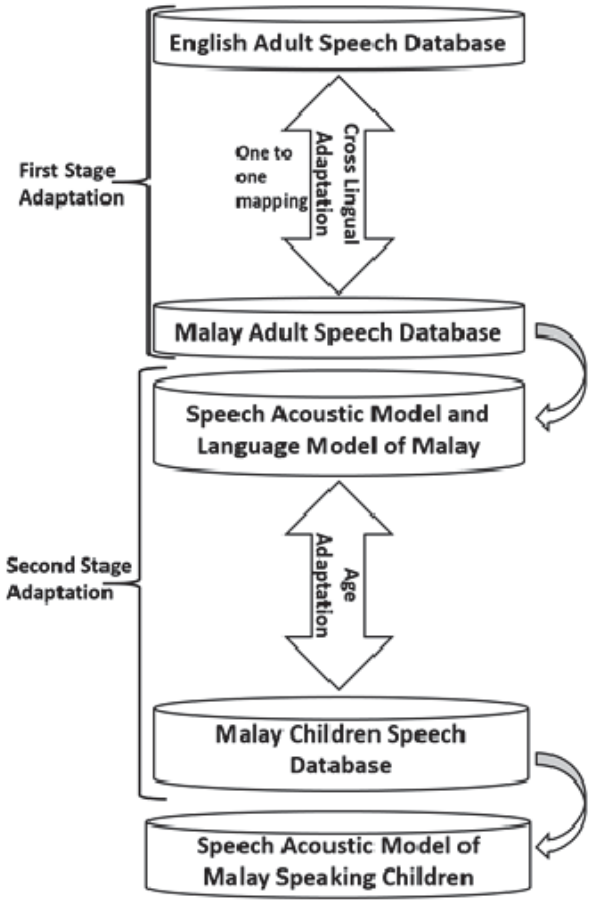

Fig. 1 The Framework of the Two Stage Adaptation for Developing ASR System for Malay Speaking Children

TABLE I
The Demographic of the Children Speakers

Speaker	Gender	Age
M10	Male	10
F10	Female	10
M13	Male	13
F13	Female	13

Based on Fig. 1, the proposed two-stage adaptation begins with cross-lingual adaptation with the use of TIMIT English database and the Malay adult speech database (containing recorded speech of 1,000 utterances of one male and one female speech each). The adaptation makes use of one-to-one mapping of Malay phonemes with the English phonemes as proposed in [4]. The outcome of this adaptation is the language model of Malay. The cross lingual of Malay and English is possible due to the similarity of pronunciation of consonants and vowels [4].

The adapted model from the first stage adaptation is used for the second stage adaptation, where the acoustic and language model of Malay speaking adult is adapted with 100 utterances uttered by each child, making a total of 400 utterances for adaptation. Adaptation is performed using

Maximum Likelihood Linear Regression (MLLR) for both stages [1], [4]. The outcome of the second stage adaptation is the speech acoustic model for Malay speaking children.

IV. Experiments

In this research, we have compared the proposed TSA technique is compared against the benchmark adaptation (single adaptation). We have opted two types of benchmark adaptation. The first one (B1), is the adaptation of the English adult database (TIMIT) with the Malay children's speech database using a one-to-one mapping [4]. The second benchmark adaptation (B2) is where the Malay adult speech database [4] is adapted with the children's speech database.

All adaptation make use of 100 recorded utterances of each child (a total of 400 utterances), and the remaining 50 for testing. The training makes use of MFCC as the feature vector and HMM as the classifier. In speaker adaptation, the regression class tree is used to generate the specific number of transformations dynamically. The adaptation is performed in two steps, which are the global adaptation, performed initially, and the second step, where the global adaptation is transformed to produce a better frame or alignment.

The proposed approach and the two benchmark adaptations performances are evaluated using the recognition accuracy measures at word level. 50 utterances of each child that were not used for training were tested for each of the experiments (a total of 200 utterances). Both training and testing were performed using the HTK toolkit, where the recognition accuracy automatically and objectively generated by the toolkit. Table II shows the breakdown of the training and test data used in this research.

TABLE II
The Breakdown of Training and Test Data from the Children's Database

	Training	Testing
Number of utterances	100	50
Number of words	480	200
Number of syllables	1,298	568
Number of phones	4,114	1,745

V. Result and Discussion

Table III shows the recognition accuracy of each of the experiments in recognizing the 50 utterances of each child as stated in Section IV. From Table III, it was found that the average recognition accuracy of the two-stage adaptation (74.0%) was higher than both the B1 (62.9%) and B2 (59.5%). The recognition accuracy of the proposed method was 18% better than the recognition accuracy of the benchmarks. From Table III, when we compare speaker to speaker, the recognition accuracy percentage differences of speaker F13 between the proposed adaptation and the benchmark was the highest (at 21.8% higher), while speaker M10 was the lowest (14.7%). This suggests that the use of two stage adaptation that generates the language model of Malay during the first adaptation run and a speech acoustic model of children from the second adaptation run is effective in recognizing children's speeches, especially the female speakers. However, the recognition accuracy of this research is not as high as other works [7], as the size of the speech database is very small (the recording time for each child is less than 10 minutes).

The recognition accuracy of the benchmark adaptation B1 that uses TIMIT as the source data is better than B2 that uses the Malay adult speech database. The performance of B1 is better as the TIMIT database is much larger at 630 speakers as compared to the Malay adult database (2 speakers). The large number of speakers of TIMIT also means that the database covers more speech variation and therefore is a suitable source model for adaptation. Another reason for the good performance of B1 is because the TIMIT database was better labeled than the Malay adult speech database due to limited resources.

TABLE III
The Recognition Accuracy of Each Experiment

Speaker	Two-Stage Adaptation	Benchmark 1	Benchmark 2
M10	62.5	54.5	50.5
F10	63.0	53.5	51.0
M13	86.5	73.5	69.5
F13	84.0	69.0	67.0
Average	**74.0**	**62.9**	**59.5**

When comparing the average recognition accuracy of each speaker, speaker M13 is the highest, while speakers M10 and F10 were the lowest as depicted in Fig. 2. From Fig. 2, it can be observed that the speeches from both the 13 years old speakers have better recognition accuracy than the ten years old. This is due to the closeness of the speech properties of these children with the adults as stated in [7]. The recognition accuracy of the speeches by the children aged 10 were low due to dissimilarity of their speeches with the adult speech. We also found that the recognition accuracy of the speeches by M13 speaker was the highest and this probably due to the presence of more male speakers' speeches for the TIMIT database.

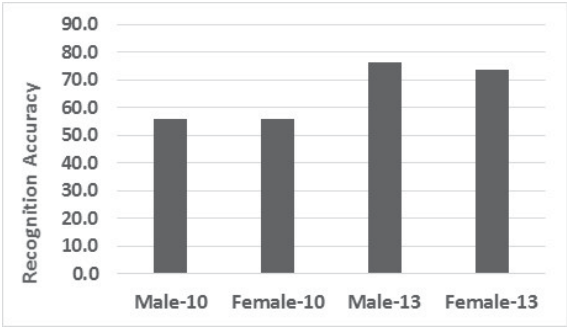

Fig. 2 The Average Recognition Accuracy for each Speaker

V. Conclusions

In this research, we have proposed a two-stage adaptation technique for developing an ASR system for Malay speaking children. The proposed approach was aimed to improve the recognition accuracy of ASR systems for children where the

language is under-resourced. The proposed technique allows us to develop a speech acoustic model for Malay speaking children. We found that the performance of the proposed method was compared with two benchmark adaptation methods using the word level recognition accuracy.

The findings show that the proposed adaptation can recognize the children's speech better than the two benchmark adaptations. The findings of the experiments also reveal that the speech generated by older children was better recognized than younger children due to the similarity with the adults. The recognition accuracy of younger children is lower due to much dissimilarity with the adult's speeches. Some of the dissimilarities that could lead to lower recognition accuracy of the young children's speech include variability in speech properties between the adults and children, and pronunciation error in children's speech.

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 and University of Malaya Research Grant (UMRG) Grant no.: RG284-14AFR.

REFERENCES

[1] Le, V-B., Besacier, L. (2009). Automatic Speech Recognition for Under-Resourced Languages: Application to Vietnamese Language. Audio, Speech, and Language Processing, IEEE Transactions, 17(8), 1471-1482. doi: 10.1109/TASL.2009.2021723.

[2] Warschauer, M. (2013). Comparing face-to-face and electronic discussion in the second language classroom. CALICO journal, 13(2-3), 7-26.

[3] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., & Rose, R. (2007). Automatic speech recognition and speech variability: A review. Speech Communication, 49(10), 763-786.

[4] Mustafa, M. B., Don, Z. M., Ainon, R. N., Zainuddin, R., & Knowles, G. (2014). Developing an HMM-Based Speech Synthesis System for Malay: A Comparison of Iterative and Isolated Unit Training. IEICE TRANSACTIONS on Information and Systems, 97(5), 1273-1282.

[5] Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56, 85-100.

[6] Plowman, L. (1999). Using video for observing interaction in the classroom. Edinburgh: Scottish Council for Research in Education.

[7] Gerosa, M., Giuliani, D., & Brugnara, F. (2009). Towards age-independent acoustic modeling. Speech Communication, 51(6), 499-509.