Jurnal Teknologi

Full Paper

A FAST ADAPTATION TECHNIQUE FOR BUILDING DIALECTAL MALAY SPEECH SYNTHESIS ACOUSTIC MODEL

Yen-Min Jasmina Khaw*, Tien-Ping Tan

School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia

Corresponding author
jasminakhaw87@hotmail.com

Article history
Received
15 May 2015
Received in revised form
1 July 2015
Accepted
11 August 2015

Graphical abstract

Abstract

This paper presents a fast adaptation technique to build a hidden Markov model (HMM) based dialectal speech synthesis acoustic model. Standard Malay is used as a source language whereas Kelantanese Malay is chosen to be target language in this study. Kelantan dialect is a Malay dialect from the northeast of Peninsular Malaysia. One of the most important steps and time consuming in building a HMM acoustic model is the alignment of speech sound. A good alignment will produce a clear and natural synthesize speech. The importance of this study is to propose a quick approach for aligning and building a good dialectal speech synthesis acoustic model by using a different source acoustic model. There are two proposed adaptation approaches in this study to synthesize dialectal Malay sentences using different amount of target speech and a source acoustic model to build the target acoustic model of speech synthesis system. From the results, we found out that the dialectal speech synthesis system built with adaptation approaches are much better in term of speech quality compared to the one without applying adaptation approach.

Keywords: Malay dialect, corpus, dialect adaptation system

1.0 INTRODUCTION

A speech synthesis system is a system that converts text to speech. Speech synthesis technologies have matured and they have been equipped and embedded in many tools, for instance in mobile phones, robotics, telephony system etc. Nevertheless, building a speech synthesis system for an unknown language still requires a lot of effort, especially in the area of phonological analysis of the target language and alignment of speech sounds. Furthermore, it is hard to obtain sufficient amount of resources for unknown language to conduct phonological analysis. Large amount of data collected will lead to higher quality synthesized speech. Some of the languages might not have a written form.

There are many different approaches to speech synthesis: articulatory synthesis, formant synthesis, concatenative synthesis and hidden Markov model (HMM) synthesis [1], [2], [3], [4]. HMM approach is one of the most popular approaches used in many speech synthesis systems. During training part, spectrum and excitation parameters are extracted from speech database and modeled by context dependent HMMs while in the synthesis part, context dependent HMMs are concatenated according to the text to be synthesized. Spectrum and excitation parameters are then generated from the HMM by using a speech parameter generation algorithm [5]. Finally, the excitation generation module and synthesis filter module synthesize speech waveform using the generated excitation and spectrum parameters. Figure 1 shows the overview of HMM-based speech synthesis system.

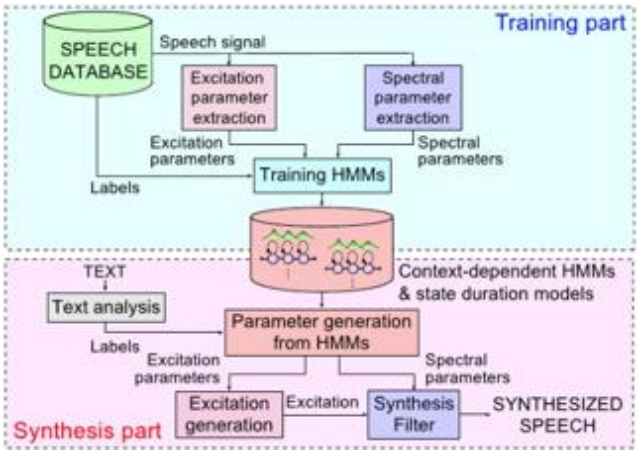

Figure 1 HMM-based speech synthesis system [6]

HMM-based speech synthesis approach can be quickly build with good quality synthesis and the models can be applied different transformations, for example to impersonate another speaker. Thus, HMM-based speech synthesis approach is more flexible over the unit-selection approach. Its voice characteristics, speaking styles, or emotions can easily be modified by transforming HMM parameters using various techniques such as adaptation [7], [8], interpolation [9], [10], eigenvoice [11], or multiple regression [12]. Speaker adaptation is the most successful example. By adapting HMMs with only a small number of utterances, the speech can be synthesized with voice characteristics of a target speaker [7], [8].

To create a speaker dependent HMM acoustic model for speech synthesis, speech alignment need to be conducted at phone-level on the speaker dependent speech synthesis corpus. Alignment can be done manually by human or automatically. Manual alignment of the utterances is expensive and time consuming. For example, to align 1 min of speech, it will take up to 30mins. Figure 2 shows the example of manual alignment for standard Malay speech sound (phones) using Praat. Automatic alignment of phone using automatic speech recognition system (ASR) can also be done, but the ASR requires an acoustic model build from the target speech to create good alignment. From our experiment, we found that using only the speaker's speech to create the acoustic model to perform the alignment does not produce good alignment. On the other hand, when large amount of speech corpus is used to create the acoustic model for ASR, the alignment produced is very good. This can be observed in the synthesized speech produced later. Since there are a lot of data in standard Malay, we attempt to use the existing data to build an acoustic for aligning dialectal speech. In this paper, we attempt to reduce the time needed in building a speech synthesis system for dialectal Malay in the alignment of speech sound, while maintaining the quality of synthesized speech produced usng existing standard Malay corpus. Two adaptation approaches were proposed in this study for aligning collected speech sound. Our study was carried out on Kelantanese Malay. It is one of the Malay dialects from the northeast of Peninsular Malaysia, which is very distinctive without written form.

This paper is organized as follows. In section 2, background for some research in this study is reviewed. Malay speech corpus preparation is described in Section 3. The proposed approaches are drawn in section 4. Experiment setup is described in section 5 and discussion is in section 6. Finally, section 7 contains conclusion and future work.

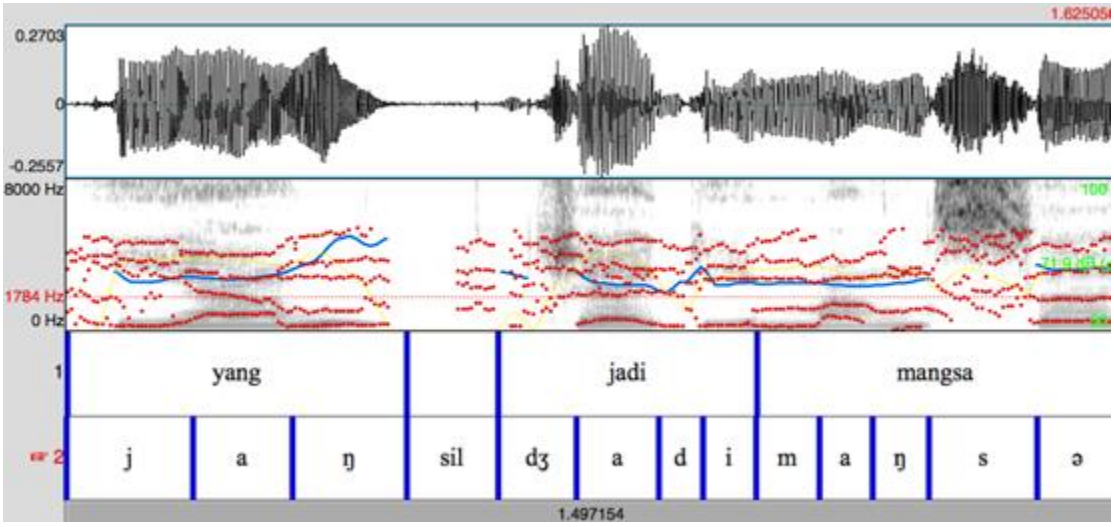

Figure 2 Manual alignment of standard Malay speech sound (phones) using Praat

2.0 BACKGROUND

Malay is a language from the Austronesian family. It is the official language in Malaysia, Indonesia, Singapore, and Brunei. However, Malay spoken in different countries and even within a country itself might vary in terms of pronunciation and vocabulary from one place to another. In Malaysia, the formal Malay language recognized and used is known as standard Malay. Standard Malay is originated from Johor, Riau dialect variety. The prominence of Johor, Riau dialect is due to the influence and importance of empire in the 19 century. Malay dialect is very distinctive, and might not easy to learn as every single pronunciation of the words can bring various meaning. Besides, Malay dialect does not have written form.

There are several Malay dialects found in Malaysia that can be grouped according to the geographical distribution. Asmah (1991) divided the Malay dialects in Peninsular Malaysia to five groups, the North Western group consists Kedah, Perlis, Penang and north Perak, the north eastern group consists of Kelantan dialect, the eastern group consists of Terengganu dialect, the southern group consists of Johor, Melaka, Pahang, Selangor and sounthern Perak and Negeri Sembilan dialect [13]. Each group may be further classified to different subdialects. In this study, Kelantan dialect is studied.

There are some literature studies on phonetic alignment and speech synthesis system. Accurate phonetic alignment is very important, as it will affect the quality of a synthesized utterance from speech synthesis system. Phonetic alignment plays an important role in speech research [14]. It is the starting point for many studies. There are some previous proposed methods in aligning speech recording at phone-level. Techniques borrowed from automatic speech recognition have been successfully applied to such word-level transcription in order to produce time aligned phone transcriptions automatically [15]. The method is referred as forced alignment. The system takes as input speech sound files and matching text files containing word-level transcriptions of the speech. An automatic speech recognition system, usually based on hidden Markov Models (HMM), is applied and a time aligned phone-level transcription is produced [16], [17], [18], [19].

Some acoustic model adaptation techniques in transforming a source speech synthesis hidden Markov model acoustic model to a target acoustic model are proposed in previous studies. Frequency warping is one of the approaches for transforming a source utterance to a target utterance. It is based on a time-varying bilinear function to reduce the weighted spectral distance between the source speaker and the target speaker [20]. Besides, MLLR (Maximum Likelihood Linear Regression) technique is a popular approach used acoustic model adaptation in automatic speech recognition has also been used for transforming the acoustic model of a source speaker to the target speaker [21], [22].

In this study, some of the approaches for fast deployment of dialectal speech synthesis system will be proposed. Since there is a large amount of standard Malay speech corpus, the proposed approaches will try to adapt it using a small amount of dialectal Malay speech corpus. The performance of some different proposed approaches will be then evaluated and compared.

3.0 SPEECH SYNTHESIS CORPUS ACQUISITION

There are some requirements that need to be fulfilled when acquiring a speech synthesis corpus. In term of environment, the recording must be done in a noise free studio. For recording, there are some criteria to be met such as expressiveness control, easy to segment, speaking rate control, prosody structure control, and voice beauty. In term of the content of the speech, the speech corpus should cover as many speech contexts as possible. A considerable amount of speech recordings with carefully selected sentences is very important for developing a good quality of speech synthesis system. One possible source of dialectal speech is from dialog speech. However, dialog speech is less suitable to be used as speech synthesis corpus because of the speed of the discourse, which varies and also the richness in emotion in an uncontrolled recording might not be desirable. Moreover, dialog speech normally does not cover a lot of phonetic context compared to read speech. Therefore, read speech is preferred instead of dialog speech. The challenge in preparing a transcript for recording speech synthesis corpus is the limited amount of text available.

The first step is to acquire some dialog speech. The dialog speech will be transcribed and translated to standard Malay to study dialectal speech, to obtain dialectal word and pronunciation dictionary, and to learn translation rules. The dialog speech is manually transcribed and a parallel corpus is prepared to produce translation rules and unique dialectal vocabularies [23]. The phonology of a language is very important in developing speech synthesis system. The unique pronunciation for each language can be observed through speech corpus. The learned acoustic was then used to develop pronunciation dictionary for the particular language. For the learned translation rules and vocabularies, standard Malay text corpus will be translated into dialectal Malay text since we have a large standard Malay text corpus [24]. Then, G2P tool developed in this study is used to convert dialectal words to their corresponding pronunciation [25]. The dialectal sentences were selected from the translated text corpus based on the information of monophones, triphones and pentaphones [26]. Finally, recording for the selected sentences was carried out. The recording was done in a soundproof room, using

AKG C414XLII microphone, and EMACOP software. The sampling rate is set at 22 kHz.

We have recorded speech from a native speaker of Kelantanese Malay. Around two thousands of sentences (about 4 hours) in text were selected, with phonetically well balanced based on the phoneme distribution. With the translated Malay dialect text and pronunciation dictionary created, sentences were selected such that those with the most number of unique unseen monophones, triphones and pentaphones were selected, so that the context can be evaluated as many as possible. A proper weight was assigned to each phone while selecting the sentences. Sentences which are having maximized number of unseen monophone, triphone and pentaphone, with highest score will be selected first. To ensure that the selected sentences are phonetically well balanced, the phone distribution in the translated Malay dialect corpus was calculated. The following shows the formula for selecting sentences with the most number of unique unseen monophones, triphones and pentaphones from the translated dialectal text.

$$A_v = \sum_{x=1}^{i=A} \left(\frac{P_{vi} + T_{vi} + M_{vi}}{3} \right) \quad (1)$$

where *Av* is the maximum score of monophone, triphone and pentaphone, *Pv* is the score of pentaphone, *Tv* is the score of triphone and *Mv* is the score of monophone, *i* is the iteration and *A* is the size of sentences.

$$P_v = \sum_{x=1}^{i=P} \frac{1}{v} \quad (2)$$

$$T_v = \sum_{x=1}^{i=T} \frac{1}{v} \quad (3)$$

$$M_v = \sum_{x=1}^{i=M} \frac{1}{v} \quad (4)$$

where *P* is the number of unique pentaphone, *T* is the number of unique triphone and *M* is the number of unique monophone.

To have an idea of the phone distribution in the selected sentences compared to the general phone distribution of Kelantanese Malay, a correlation coefficient between these two are calculated. Table 1 shows the correlation coefficient for the selected sentences in Kelantan dialect calculated from the speaker who speak Kelantanese Malay.

Table 1 Correlation coefficient for selected sentences in Kelantanese Malay

Type	Correlation Coefficient
Kelantanese Malay	0.9893

The result shows that the selected sentences have a correlation coefficient of about 0.99 for Kelantanese Malay, which means that it is phonetically well balanced. Figure 3 shows the phone distribution among selected sentences for Kelantanese Malay in the corpus, compared to the general Kelantanese Malay. Notice that our selection strategy has managed to increase the percentage of rare phones in sentences for recording.

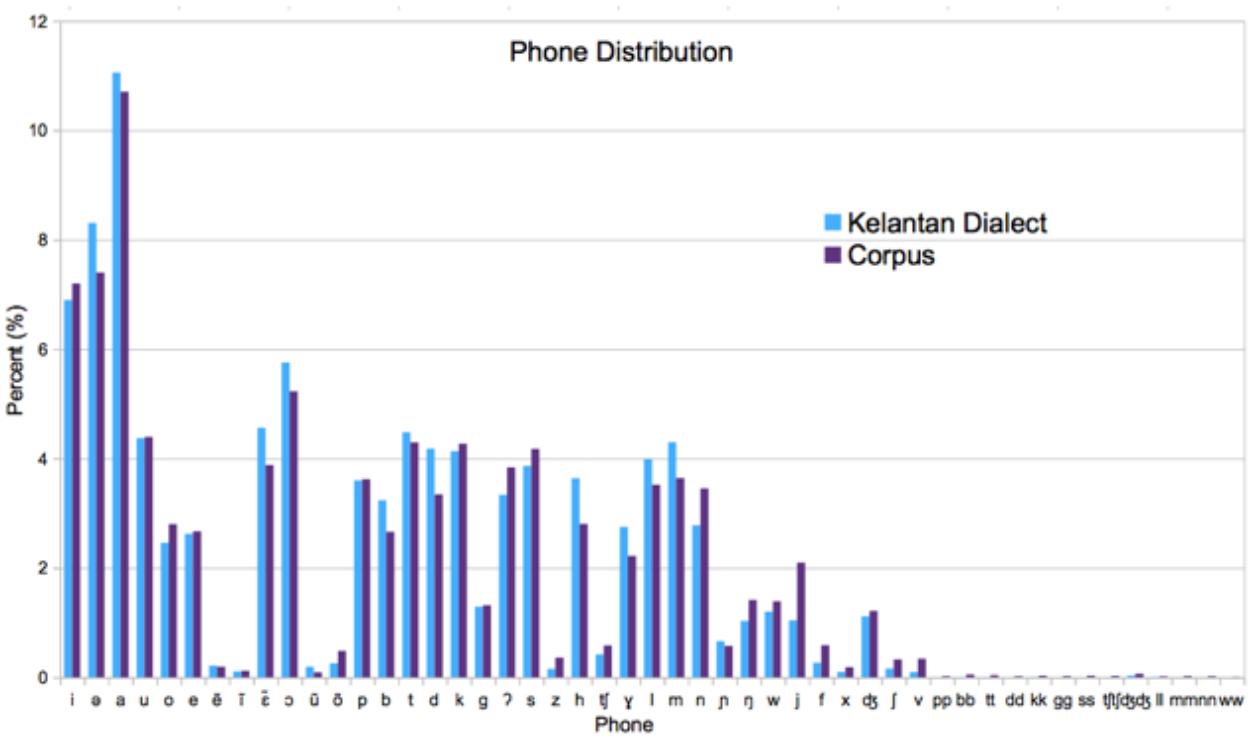

Figure 3 Phone distributions for Kelantanese Malay

4.0 PROPOSED SPEECH SOUND ALIGNMENT FOR DIALECTAL SPEECH SYNTHESIS

4.1 Forced Alignment

After acquiring a speech synthesis corpus, the speech sounds were aligned. Forced alignment is one of the alignment approaches that have been widely used for automatic phone segmentation in speech recognition. Viterbi algorithm [27] in the automatic speech recognition system is applied for performing the phonetic alignment task. The ASR system will need an acoustic model and a pronunciation dictionary to align the recorded audio to the word transcription. The speech signal is analyzed as a successive set of frames. The alignment of frames with phonemes is determined via the Viterbi algorithm, which finds the most likely sequence of hidden states given the observed data and the acoustic model represented by the hidden Markov models (HMMs). The acoustic features used for training HMMs are normally cepstral coefficients such as MFCCs [28] and PLPs [29]. A common practice involves training single Gaussian HMMs first and then doubling the Gaussian mixtures in HMMs.

4.2 Proposed Adaptation Technique for Forced Alignment

From our previous study, it showed that the synthesized speech is not clear and the quality is quite bad if the phone alignment of speech synthesis corpus is not done properly. This is because the alignment produced will be used to model the speech sounds or phones (HMM-based speech synthesis acoustic model). Thus, for automatic alignment to work, the acoustic model used in the alignment must be robustly trained using large speech corpus. Our study shows that using only the speech synthesis corpus to create the acoustic model for the phonetic alignment does not produce good quality synthesized speech. This observation is similar to the finding in the field of automatic speech recognition, where an acoustic model created from a large speech corpus and many speakers is better in decoding the speech of a speaker than using a small speaker dependent speech corpus alone. However, acquiring large speech corpus is expensive. Thus, in this study we will use existing available speech resources (e.g. Standard Malay). Dialectal Malay (e.g. Kelantanese Malay speech synthesis corpus) will be used to adapt the standard Malay acoustic model, and then used for phonetic alignment. Two proposed adaptation approaches are investigated.

4.2.1 Initialise Target Language Acoustic Model Using Source Language Acoustic Model for Forced Alignment

The first approach is to initialize the target language acoustic model using a source language acoustic model for forced alignment. The idea is to initialise the dialectal Malay acoustic model using standard

Malay acoustic model and then adapted the model by using the dialectal Malay speech. The adapted acoustic model is then used to align the dialectal Malay utterances in the speech synthesis corpus. The number of phonemes in standard Malay might be different compared to dialectal Malay. Therefore, there would be unique dialectal Malay phonemes, which do not exist, in standard Malay. To overcome this problem, similar phones of standard Malay and dialectal Malay are mapped together, while unique phones in dialectal Malay are mapped to a phone, which is perceptional closest. Dialectal speech is then used to adapt the standard Malay acoustic model where dialectal Malay acoustic model was created. Finally, forced alignment for dialectal Malay speech can be carried out in order to build a dialectal speech.

4.2.2 Adapting Target Language Pronunciation Modeling Based on Source Language Phoneset for Forced Alignment

In the second proposed approach, the pronunciation dictionary for dialectal Malay was prepared based on the phoneset of standard Malay. Similar phones of standard Malay and dialectal Malay are mapped together in the phoneset. For unique phones that exist in dialectal Malay only, they are mapped by perception similarity. The dialectal pronunciation dictionary created was then use to adapt the standard Malay acoustic model using dialectal speech. With the dialectal acoustic model created, forced alignment of the dialectal speech is conducted. The aligned speech sound was then used to build a dialectal speech synthesis system.

5.0 EXPERIMENT

5.1 Experiment Setup

Automatic phonetic alignment was carried out by forced aligning the utterances using an automatic speech recognizer, Sphinx3 from CMU, was applied in this experiment. Standard Malay was used as a source language whereas dialectal Kelantanese Malay as target language in this study. About 4 hours of speaker dependent dialectal Malay speech synthesis corpus described in Section 3 was used in this experiment. On the other hand, to create standard Malay acoustic model, it was trained through automatic speech recognizer using MASS speech resources [30] that contains about 140 hours of speech, and our pronunciation dictionary. The aligned Kelantanese Malay utterances were then used to train acoustic model for the hidden Markov model (HMM) based speech synthesis system. In this study, there are two fast adaptation approaches being studied for aligning dialectal Malay speech described in the following subsections. Figure 4, 5 and 6 show the baseline approach and the two proposed approaches for forced alignment.

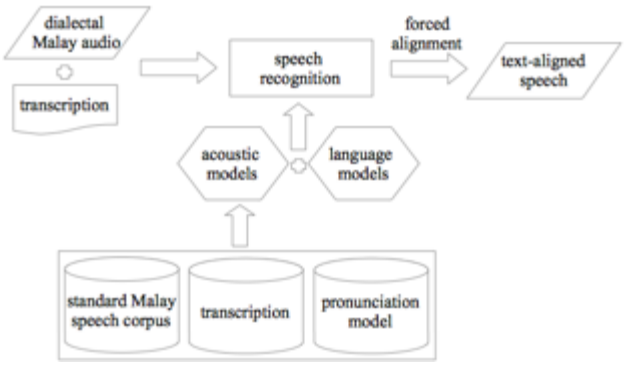

Figure 4 Baseline approach for forced alignment

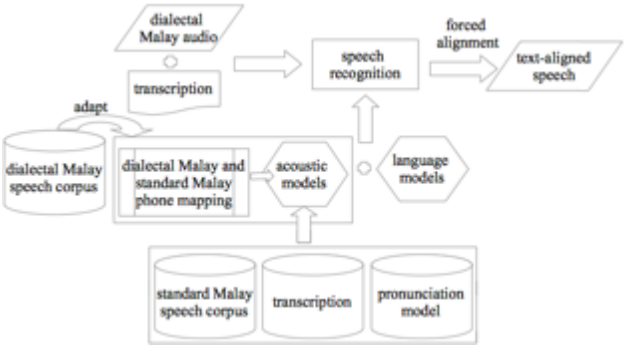

Figure 5 Initialise target language acoustic model using source language acoustic model for forced alignment

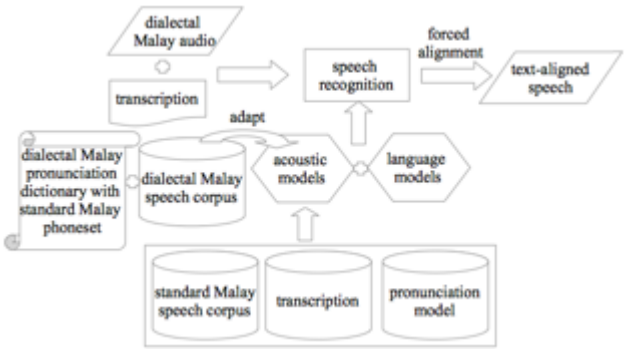

Figure 6 Adapting target language pronunciation modeling based on source language phoneset for forced alignment

Table 2, 3 and 4 describe the phoneset for standard Malay and Kelantanese Malay [25].

Table 2 Number of vowel, consonant and diphthong in standard Malay and Kelantanese Malay

Category	Standard Malay	Kelantanese Malay
Vowel	6	12
Consonant	25	37
Diphthong	3	0

Table 3 List of vowels, consonants and diphthongs in standard Malay

Vowels	Consonants				Diphthongs
/a/	/p/	/h/	/m/	/l/	/aw/
/e/	/b/	/f/	/n/	/r/	/aj/
/i/	/t/	/v/	/ŋ/	/z/	/oj/
/o/	/d/	/ʃ/	/ɲ/	/ʔ/	
/u/	/k/	/x/	/w/		
/ə/	/g/	/ɣ/	/j/		
	/s/	/tʃ/	/dʒ/		

Table 4 List of vowels and consonants in Kelantanese Malay

Vowels		Consonants					
/i/	/ẽ/	/p/	/s/	/m/	/f/	/bb/	/tʃtʃ/
/ə/	/ĩ/	/b/	/z/	/n/	/x/	/tt/	/dʒdʒ/
/a/	/ɛ̃/	/t/	/h/	/ɲ/	/dʒ/	/dd/	/ll/
/u/	/ɔ/	/d/	/tʃ/	/ŋ/	/ʃ/	/kk/	/mm/
/o/	/ũ/	/k/	/ɣ/	/w/	/v/	/gg/	/nn/
/e/	/õ/	/g/	/ll/	/j/	/pp/	/ss/	/ww/
		/ʔ/					

5.2 Baseline: Dialectal Synthesis System without using a Source Acoustic Model

In this approach, a pronunciation dictionary for Kelantanese Malay is prepared where all the phoneset used is included as shown in Table 4. Table 2 shows the number of vowel, consonant and diphthong of Kelantanese Malay, which is slightly different compared to standard Malay. Phonology of standard Malay is shown in Table 3.

A Kelantanese Malay acoustic model was then trained through automatic speech recognizer using Kelantanese Malay speech corpus only without any adaptation technique carried out. Forced alignment was then conducted to align the dialectal speech using the acoustic model and then used to train the acoustic model for HMM-based speech synthesis system. Finally, the dialectal Malay utterances were synthesized.

5.3 Proposed Adaptation Technique for Forced Alignment

As there is only about 4 hours of dialectal Malay speech collected in this research, some adaptation approaches were proposed in order to build a better quality of dialectal speech synthesis acoustic model. Two approaches for building dialect adaptation system were proposed in this study. The following sub section will describe each proposed approach in details using Kelantanese Malay.

5.3.1 Initialise Target Language Acoustic Model Using Source Language Acoustic Model for Forced Alignment (Approach A)

First, a Kelantanese Malay acoustic model was initialised using standard Malay acoustic model. Since there is more number of phoneme used in Kelantanese Malay, additional phones were initialised with the closest standard Malay phone in perception. Finally, MLLR [31] and MAP [32] acoustic model adaptation algorithms, which are part of the Sphinx speech recognition package was applied using Kelantanese Malay speech synthesis corpus on the Kelantanese Malay acoustic model that we initialized earlier to create a new/adapted Kelantanese Malay acoustic model. Table 5 shows the Kelantanese Malay phones that matched with closest standard Malay phones.

Table 5 Mapping of Kelantanese Malay phones to the standard Malay phone

Kelantanese Malay Phoneme	Standard Malay Phoneme	Kelantanese Malay Phoneme	Standard Malay Phoneme
/ẽ/	/e/	/ss/	/s/
/ɛ̃/	/e/	/tʃtʃ/	/tʃ/
/ɔ/	/o/	/dʒdʒ/	/dʒ/
/pp/	/p/	/ll/	/l/
/bb/	/b/	/mm/	/m/
/tt/	/t/	/nn/	/n/
/kk/	/k/	/ww/	/w/
/gg/	/g/		

With the adapted Kelantanese Malay acoustic model, forced alignment was carried out to align recorded Kelantanese Malay speech synthesis corpus. The acoustic model for the HMM-based speech synthesis system was then trained using align speech and synthesized Kelantanese Malay utterances were obtained.

5.3.2 Adapting Target Language Pronunciation Modeling Based On Source Language Phoneset for Forced Alignment (Approach B)

Standard Malay acoustic model was adapted to sounds of dialectal Malay by creating a phoneset map and creating dialectal pronunciation dictionary with standard Malay phoneset, which shown in Table 6. The standard Malay acoustic model was then adapted with MLLR [31] and MAP [32]. Some examples are shown in Table 7.

Table 6 Phoneset mapping between standard Malay and Kelantanese Malay

Kelantanese Malay Phoneme	Standard Malay Phoneme	Kelantanese Malay Phoneme	Standard Malay Phoneme
/ẽ/	/e/	/dd/	/d/
/ĩ/	/i ŋ/	/kk/	/k/
/ɛ̃/	/e/	/gg/	/g/
/ɔ/	/o/	/ss/	/s/
/ũ/	/u n/	/tʃtʃ/	/tʃ/
/õ/	/o m/ or /on/ or /o ŋ/	/ll/	/l/
/pp/	/p/	/mm/	/m/
/bb/	/b/	/nn/	/n/
/tt/	/t/	/ww/	/w/

Table 7 Pronunciation of Kelantanese Words based on standard Malay Phoneset

Malay Word	Meaning	Kelantanese Malay Pronunciation	Pronunciation after Phoneme Mapping
dalam	inside	/d a l **ɛ̃**/	/d a l **e**/
tahun	year	/t a h **ũ**/	/t a h **u n**/
gila	crazy	/g i l **ɔ**/	/g i l **o**/

The recorded Kelantanese speech was then aligned using the adapted acoustic model by forced alignment. After that, the aligned speech was used to train acoustic model for the HMM-based speech synthesis system and finally Kelantanese Malay utterances were synthesized.

6.0 DISCUSSION

There are two experiments conducted in this study. Fifteen synthesized sentences in each approach were randomly chosen for evaluation. Twenty listeners participated in the perception test conducted. The first experiment was carried out to evaluate the synthesized utterances in term of naturalness, ease of listening and articulation for each proposed approach. The following Table 8 shows the scale for Mean Score Option (MOS) of experiment 1. Listeners were asked to rate the three aforementioned quality dimensions for each sentence by grading on a scale of 1 to 5 for each dimension. The standard deviation (std) is then calculated.

Table 8 Scale labels for MOS evaluation: experiment 1

Attributes	Naturalness	Ease of Listening	Articulation
1	Unnatural	No meaning understood	Bad
2	Inadequately natural	Effort required	Not very clear
3	Adequately natural	Moderate effort	Fairly clear
4	Near natural	No appreciable effort required	Clear enough
5	Natural	No effort required	Very clear

Table 9 shows the evaluation results in experiment 1. From the results, we noticed that the baseline approach without adaptation data is having lower quality of synthesized speech compare to adaptation approaches of approach A and approach B. The overall score for approach A of "naturalness", "ease of listening" and "articulation" which lies 4 is particularly high, considering that 4 corresponded to "near natural", "no appreciable effort required" and "clear enough". It is worth noting that the "ease of listening" and the "articulation" received remarkably high grades, illustrating that the synthetic speech contains minimal number of distractions that would otherwise demand more effort from the listener for perceiving the transmitted message. For approach B, the "naturalness", "ease of listening" and "articulation" shows scale near 4 where approach A is slightly better than approach B.

Table 9 Evaluation result: Experiment 1

Proposed Dialect Adaptation System	Naturalness		Ease of Listening		Articulation	
	MOS	STD	MOS	STD	MOS	STD
Approach A	4.24	0.26	4.30	0.31	4.20	0.20
Approach B	3.92	0.42	3.90	0.39	3.82	0.40

Figure 7 shows the scale labels for MOS evaluation in experiment 1.

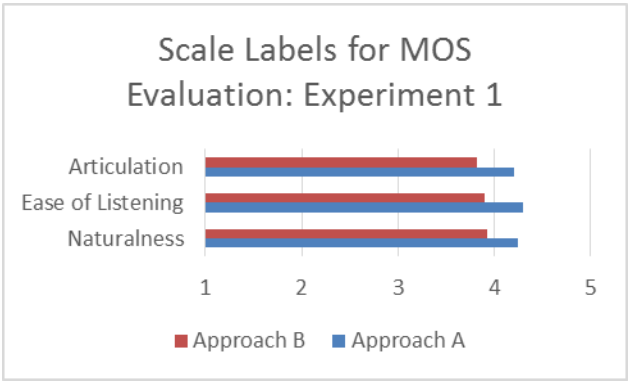

Figure 7 Scale labels for MOS evaluation: Experiment 1

In second experiment, the overall performance of the synthesized utterances for each proposed approach was evaluated. Table 10 shows the scale for MOS of experiment 2. Listeners were asked to rate each sentence by grading on a scale of 1 to 5 for overall performance.

Table 10 Scale labels for MOS evaluation: Experiment 2

	1	2	3	4	5
Overall Performance	Bad	Poor	Fair	Good	Excellent

The evaluation results of experiment 2 are shown in Table 11. The overall performance of approach A has a score of 4, which is particularly high, considering that 4 corresponded to "good". For approach B, the overall performance scale lies between 3 and 4. From the result of perception test carried out, approach A and approach B are having higher mark compare to the approach that without conducting adaptation for building dialectal speech synthesis system.

Table 11 Evaluation result: Experiment 2

Overall Performance	Proposed Dialect Adaptation System	
	Approach A	Approach B
MOS	4.04	3.57
STD	0.25	0.37

Figure 8 shows the overall performance result for experiment 2.

Figure 8 Overall Performance: Experiment 1

From the experiments conducted, we found out that an acoustic model created from a large speech corpus is better in decoding the speech of a speaker than using a small speaker dependent speech corpus alone. Therefore, the proposed adaptation approaches in this study were reliable.

7.0 CONCLUSION

In this paper, two adaptation approaches were proposed to build dialect speech synthesis system quickly with quality concerned as collecting a large corpus of speech is very time consuming. With this, Kelantanese Malay was collected in this study. For future works, other dialectal Malay such as Sarawak dialect and Kedah dialect can be conducted. Building dialectal synthesis system will be interesting for those who wish to learn a particular dialect.

Acknowledgement

This project is supported by the research university grant 1001/PKOMP/817068 from Universiti Sains Malaysia.

References

[1] Huang, X. D., Acero, A. and Hon, H-W. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR, New Jersey.

[2] Baeza-Yates, R. and Ribeiro-Neto, B. 1999. *Modern Information Retrieval.* Addision-Wesley.

[3] Rank, E. and Pirker, H. 1998. *Generating Emotional Speech with a Concatenative Synthesizer, ICSLP'98*. 671-674.

[4] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. 1999. *Simultaneous Modeling Of Spectrum, Pitch and Duration In HMM-Based Speech Synthesis, Eurospeech.* 2347-2350.

[5] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura. 2000. Speech Parameter Generation Algorithms For HMM-Based Speech Synthesis. *Proc. of ICASSP 2000*. 3: 1315-1318, June 2000.

[6] Tokuda, H. Zen, A. W. Black. 2002. An HMM-based Speech Synthesis System Applied to English. *IEEE Workshop on Speech Synthesis.* 227-230.

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamur. 2000. Speech Parameter Generation Algorithms For HMM-Based Speech Synthesis. *Proc. of ICASSP 2000.* 3: 1315-1318, June 2000.

[8] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi. 2001. Adaptation of Pitch and Spectrum for HMM-based Speech Synthesis using MLLR. In *Proc. ICASSP, 2001*. 805-808.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 1997. Speaker interpolation in HMM-based Speech Synthesis System. In *Proc. Eurospeech, 1997*. 2523-2526.

[10] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi. 2005. Speech Synthesis with Various Emotional Expressions and Speaking Styles by Style Interpolationand Morphing. *IEICE Trans. Inf. & Syst*. E88-D(11): 2484-2491.

[11] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 2002. Eigenvoices for HMM-based Speech Synthesis. In *Proc. ICSLP, 2002*. 1269-1272.

[12] T. Nose, J. Yamagishi, and T. Kobayashi. 2006. A Style Control Technique For Speech Synthesis Using Multiple Regression HSMM. In *Proc. Interspeech, 2006*. 1324-1327.

[13] Asmah Haji Omar. 1991. *Aspek Bahasa dan Kajiannya*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

[14] Sergio, P. and Luis Oliveira, C. 2003. DTW-based Phonetic Alignment Using Multiple Acoustic Features, EUROSPEECH 2003 – GENEVA.

[15] Brugnara, F., Falavigna, D. and Omologo, M. 1993. Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. *Speech Communication.* 12(4): 357-370.

[16] Sjolander, K. 2003. An HMM-based System For Automatic Segmentation and Alignment Of Speech, Umea University, Department of Philosophy and Linguistics PHONUM. 9: 93-96.

[17] Jakovljević, N., Mišković, D., Pekar, D., Sečujski, M. and Delić, V. 2012. Automatic Phonetic Segmentation for a Speech Corpus of Hebrew. *INFOTEH-JAHORINA*. 11.

[18] Mizera, P. and Pollak, P. 2013. Accuracy of HMM-Based Phonetic Segmentation Using Monophone or Triphone Acoustic Model.

[19] Yuan, J., Ryant, N. and Liberman, M., Stolcke, V. Mitra, and W. Wang. 2013. Automatic Phonetic Segmentation using Boundary Models, in INTERSPEECH. 2306-2310.

[20] Gao, W. and Cao, Q. 2014. Frequency Warping for Speaker Adaptation in HMM-based Speech Synthesis. *Journal of Information Science and Engeering.* 30: 1149-1166.

[21] Tamura, M., Masuko, T., Tokuda, K. and Kobayashi, T. 1998. Speaker Adaptation for HMM-Based Speech Synthesis System using MLLR.

[22] C. J. Leggetter and P. C. Woodland. 1995. Maximum Likeliood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*. 171-185.

[23] Khaw, J-Y. M. and Tan T. P. 2014. Hybrid Approach for Aligning Parallel Sentences for Languages without a Writteen Form using Standard Malay and Malay Dialect, Asian Language Processing (IALP). 170-174.

[24] Tao, J., Liu, F., Zhang, M. and Jia, H. 2008. Design of Speech Corpus for Mandarin Text to Speech.
[25] Khaw, J-Y. M. and Tan, T. P. 2014. Grapheme To Phoneme for Kelantan Dialect. Cocosda'14, Phuket, Thailand. 206-211.
[26] Khaw, J-Y. M. and Tan, T. P. 2014. Preparation of MaDiTS Corpus for Malay Dialect Translation and Speech Synthesis System. *Proceeding of the 2nd International Workshop on Speech, Language and Audio in Multimedia (SLAM 2014), Penang, Malaysia.* 53-57.
[27] Wightman, C. and Talkin, D. 1997. The Aligner: Text to Speech Alignment Using Markov Models. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (ed.). *Progress in Speech Synthesis*. Springer Verlag, New York. 313-323.
[28] Davis, S. & Mermelstein, P. 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing.* ASSP-28: 355–366. 2.2.1.
[29] Hermansky, H. 1990. Perceptual Linear Predictive Analysis of Speech. *The Journal of the Acoustical Society of America.* 87: 1738-1752. 1.1, 2.2.1.
[30] Tan, T. P., Xiao, X., Tang, E. K, Chng, E. S. and Li, H. 2009. Mass: A Malay Language LVCSR Corpus Resource, Cocosda'09, Beijing. 10-13.
[31] Goronzy, S. and Kompe, R. 1998. Speaker Adaptation of HMMs using MLLR. *Proceedings of SRF*.
[32] Kompe, R. and Goronzy, S. 1998. MAP Adaptation of an HMM Speech Recognizer. *Proceedings of SRF*.