INTERSPEECH 2009 BRIGHTON

Speaker Adaptation Using a Parallel Phone Set Pronunciation Dictionary for Thai-English Bilingual TTS

Anocha Rugchatjaroen, Nattanun Thatphithakkul, Ananlada Chotimongkol, Ausdang Thangthai, Chai Wutiwiwatchai

Human Language Technology Laboratory,
National Electronics and Computer Technology Center (NECTEC), Thailand
{anocha.rug, nattanun.tha, ananlada.cho, ausdang.tha, chai.wut}@nectec.or.th

10.21437/Interspeech.2009-152

Abstract

This paper develops a bilingual Thai-English TTS system from two monolingual HMM-based TTS systems. An English Nagoya HMM-based TTS system (HTS) provides correct pronunciations of English words but the voice is different from the voice in a Thai HTS system. We apply a CSMAPLR adaptation technique to make the English voice sounds more similar to the Thai voice. To overcome a phone mapping problem normally occurs with a pair of languages that have dissimilar phone sets, we utilize a cross-language pronunciation mapping through a parallel phone set pronunciation dictionary. The results from the subjective listening test show that English words synthesized by our proposed system are more intelligible (with 0.61 higher MOS) than the existing bilingual Thai-English TTS. Moreover, with the proposed adaptation method, the synthesized English words sound more similar to synthesized Thai words.

1. Introduction

Current Thai text-to-speech (TTS) systems are efficient enough and have been applied to many kinds of texts such as, web pages, SMSs, and e-mails. With globalization of the world today, the texts that a TTS system has to read may contain multiple languages within the same document. For example, in Thai texts, which we focus on in this paper, it is quite common to see some English words occur in the same document. For this reason, it is necessary to have a TTS system which can synthesize speech both in Thai and English, i.e., a bilingual TTS system.

A straightforward technique for developing a Thai-English TTS system is to build a speech synthesizer from a bilingual speech corpus of a bilingual speaker similar to the one described in [1]. However, a bilingual speech corpus is quite difficult to create as qualified speakers are limited to only those who can speak both Thai and English fluently. Recently, there was an attempt to make a Thai TTS system able to synthesize English words by transforming an English word into a sequence of Thai phonemes with grapheme-to-phoneme conversion (G2P), then synthesizing this sequence of phonemes with a Thai TTS system similar to the way a Thai word is synthesized [2]. One major problem with this technique is the accuracy of the English-grapheme-to-Thai-phoneme conversion. Only 62% of the generated pronunciations are acceptable.

This paper aims at developing a bilingual Thai-English TTS system that can synthesize English words with higher intelligibility than the existing system. Inaccurate pronunciations from the English-grapheme-to-Thai-phoneme converter reduce the intelligibility of synthesized English words, thus we choose to alleviate this problem with resources from an English TTS system. An English pronunciation dictionary or English G2P would give more accurate pronunciations of English words; however, an English TTS system is required to synthesize these English phonemes. Our bilingual TTS system composes of two monolingual TTS systems, one for Thai and another one for English. Our Thai TTS system is an HMM-based TTS (HTS) system that has been developed according to the approach described in [3]. For an English TTS system, we use the same HTS approach to create an English synthesizer from the CMU ARCTIC databases [4]. When two monolingual TTS systems are used in the same sentence, the voices will be different when switching between languages. Therefore, we apply a speaker adaptation technique to make synthesized English words from the English TTS system sounds more similar to a synthesized voice from the Thai TTS system.

The main problem of cross-language speaker adaptation is the differences in the set of phones used in different languages. The phones that occur only in the target language may not be adequately adapted [5]. One way to solve this problem is to manually create a phone mapping between languages that allow many-to-one or one-to-many mapping, such as in [6]. Liang et al. [5], other the other hand, performed a cross-language mapping at the level of acoustic attributes, represented by HMM states, rather than at the level of phones to avoid the problem of phone mapping. Although the mapping can be identified automatically from similarity distances between HMM states, a bilingual corpus of at least one speaker is necessary in order to obtain an accurate state mapping. This mapping can then be applied to create a bilingual TTS system for any other monolingual speakers.

In this paper, we propose a new cross-language pronunciation mapping through a parallel phone set pronunciation dictionary. Each entry in this dictionary is an English word and its pronunciation transcriptions, one using a Thai phone set (namely, Thai pronunciation) and another one using an English phone set (namely, English pronunciation). The Thai pronunciations of English words are already exist and were used in our prior TTS system to make it be able to synthesize some English words. This parallel phone set pronunciation dictionary allows us to do cross-language speaker adaption without the need of an explicit mapping between phone sets or a bilingual corpus. The mapping is done automatically through the parallel pronunciations. In our approach, speaker adaptation is done using the following process. We first create adaptation data from the parallel phone set dictionary by synthesizing English words according to their Thai pronunciation transcriptions using our Thai HTS system. Then, we label the synthesized speech with the corresponding English pronunciation transcriptions and use

these labels and the synthesized speech to adapt the voice in the English HTS system toward the voice in the Thai HTS system with a constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation approach [7,8]. In terms of evaluation, we evaluate English words synthesized with our proposed method with two criteria similarity and intelligibility.

2. Proposed system

In this paper, we focus on improving the intelligibility of English words synthesized by a bilingual Thai-English TTS system with resources from a monolingual English TTS system. Figure 1 illustrates our proposed technique. Our speech synthesizer consists of two parts: the first part is a Thai HTS for synthesizing a Thai text and the second part is an English HTS for synthesizing an English text. We apply a speaker adaptation technique to this English HTS to make it sounds more similar to the Thai HTS. The adapted model is denoted as E-TH HTS in Figure 1.

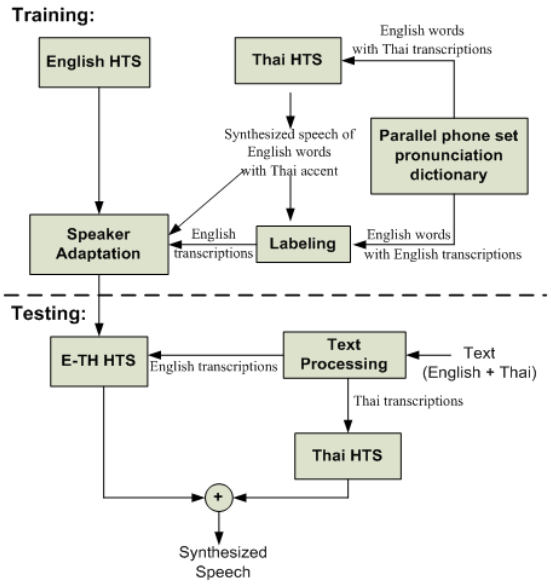


Figure 1: *The proposed system.*

We create speaker adaptation data from a parallel phone set pronunciation dictionary. Each entry in this dictionary is an English word and its pronunciation transcriptions, one using a Thai phone set (namely, Thai pronunciation) and another one using an English phone set (namely, English pronunciation). This parallel phone set dictionary allows us to label speech data with either a Thai pronunciation transcription or an English pronunciation transcription. We use the Thai HTS to generate speech data of a monolingual Thai speaker with the Thai pronunciation in the parallel phone set dictionary. We then label these speech data with the corresponding English pronunciation transcriptions. These English phone labels and the synthesized English words from the Thai HTS are then used to adapt the voice in the English HTS system toward the voice in the Thai HTS. This cross-language pronunciation mapping through the parallel phone set pronunciation dictionary allows us to adapt English phones in the English HTS without having to create an explicit mapping between a Thai phone set and an English phone set. The mapping is done automatically through the parallel Thai-English pronunciations of English words. For speaker adaptation, we use a technique called CSMAPLR. This adaptation technique has been applied successfully for HMM-based speech synthesis [8].

2.1. Pronunciation Mapping

Due to the differences between Thai syllable structure and English syllable structure [2], we found that there are some problems when we try to label an English word which was pronounced according to a Thai pronunciation transcription with an English pronunciation transcription. For instance, the 'C_i V C_f C_f[1]'structure occurs only in English while, in Thai, we only have the 'C_i V' or 'C_i V C_f' structure. Figure 2 shows a spectrogram of a word 'announcement', which is an example of the 'C_i V C_f C_f' structure in English, and its phone labels. When pronouncing this word with Thai pronunciation, only a final sound 'n' is articulated, but not a final sound 't' as illustrated with the first set of labels in Figure 2. Therefore, we have to place labels for two phones when only one phone is articulated. In this case, some part of 'n' will be labeled as 't'. This situation also occurs in other cases where multiple English phones are pronounced with only one Thai phone. For the opposite case where one English phone is pronounced with multiple Thai phones, the boundaries of this English phone cover all of the Thai phones.

We decide to label all English phones according to an English transcription even though some of them are not pronounced by a Thai speaker to make the phone context of the adaptation data similar to the phone context of the source data (English speech corpus), so that the HMM model will get adapted in the right context. Dummy phones, such as 't' in the previous example, are incorrect adaptation data; nevertheless, they do not affect the intelligibility of the adapted model as perceived by Thai listeners much. From a preliminary listening test, we found that the duration of 't' in the adapted model is shorter than the one in the original model with the similar context. However, this doesn't affect Thai listeners because, in the Thai pronunciation, there is no 't'.

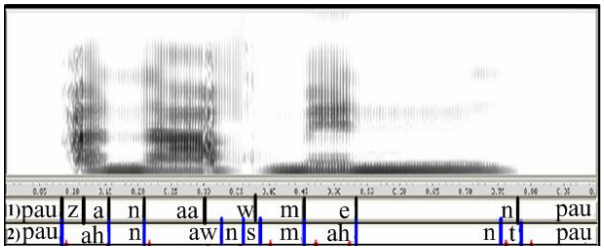


Figure 2: *A synthesized speech of the word "announcement" with 1) the Thai pronunciation transcription and 2) the English pronunciation transcription.*

2.2. Text Processing

To process a text that contains both Thai and English words, we have to first perform a language classification. This is done by simply looking at their character codes. The next step is to find a pronunciation (i.e., a sequence of phonemes) for each word. For a Thai word, we use a Thai pronunciation dictionary. For an English word, we utilize a text processing module of Festival, an English TTS system [10]. Festival uses

[1] C_i stands for initial consonance; V stands for vowel; C_f stands for final consonance.

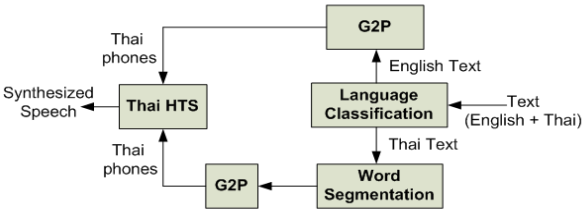
both a dictionary and a G2P conversion to obtain a phoneme sequence of a given word. The phoneme sequences are then passed to the corresponding HTS models, a Thai HTS for a Thai word and a Thai phoneme sequence, and an adapted English HTS (E-TH HTS) for an English word and an English phoneme sequence

3. Experiment

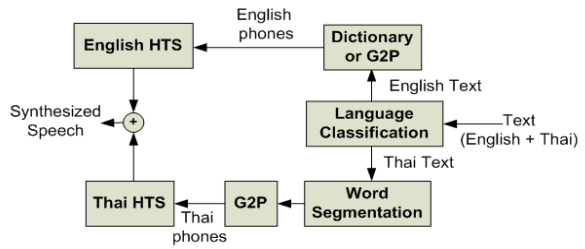
We evaluate the quality of English words synthesized by our proposed system with two evaluation criteria: similarity and intelligibility. We compare our results with the results from two baseline systems, a monolingual Thai TTS system and a combination of monolingual Thai and English TTS systems, described in Section 3.1. Our experimental setting and the results are discussed in Section 3.2.

3.1. Baseline Systems

Figure 3 shows diagrams of two baseline systems. The main difference between these two systems is the pronunciation of an English word. The first baseline system, a monolingual Thai TTS system, uses the English-grapheme-to-Thai-phoneme conversion described in [2] to acquire a Thai phoneme sequence of an English word and then pass it to a Thai synthesizer. The second baseline system, a combination of monolingual Thai and English TTS systems, uses an English text processing module of Festival [10] which utilizes both a CMU pronunciation dictionary and a G2P conversion to obtain an English phoneme sequence of an English word, and then pass it to the English synthesizer. Our proposed system is similar to the second baseline system except that the English synthesizer is adapted toward the voice in the Thai synthesizer.



(a) Monolingual TTS system (ML)



(b) Bi-Monolingual TTS system (BML)

Figure 3: *Baseline system for Bilingual Thai-English TTS.*

All the synthesizers use in this experiment are an HMM-based TTS. A Thai HTS is trained with a Thai speech synthesis corpus 1 (TSynC-1) [11]. This corpus is a single female speaker speech corpus and contains 13 hours of read speech (5,200 utterances) that cover all bi-phone in Thai. A phone set used in TSynC-1 consists of 89 Thai phonemes with 5 variations of tones. An English HTS uses an average voice model trained from 4 speakers of the CMU ARCTIC corpus using speaker independent and speaker adaptive training (SAT). This English HTS model is also an initial model for speaker adaptation. We use the average voice model as one of our baseline systems to serve as a lower bound in terms of similarity. Furthermore, we can identify whether speaker adaptation affects the intelligibility of synthesized speech by comparing this English HTS with the adapted model.

To create an English HTS that sounds more similar to the Thai HTS voice or an adapted English HTS (E-TH HTS), we use a constrained maximum likelihood linear regression (CMLLR) tree-based method with an additional MAP estimation step after MLLR adaptation (CMLLR+MAP). Adaptation data are 8,268 synthesized English words generated by the Thai HTS according to Thai pronunciation transcriptions in the parallel phone set pronunciation dictionary.

3.2. Experiments and Results

We use a subjective listening test to evaluate our proposed bilingual HTS system. Each listening test consists of 20 sentences; each sentence contains a carrier phrase generated from the Thai HTS and an English word synthesized by each bilingual HTS systems described in the previous section. There are three test sets; one for each of the three systems. ML denotes a test set that contains an English word synthesized by the Thai HTS with the English-grapheme-to-Thai-phoneme converter while BML is for the English HTS with an average voice and BML_A is for the adapted English HTS. We also add the forth test set, ML_D, where each English word has a correct Thai pronunciation without any mistakes from G2P conversion.

We evaluate each bilingual HTS system on two aspects: similarity and intelligibility. We asked 10 subjects to listen to the three test sets and rate each sentence along these two aspects. For similarity, the subject were asked how similar the voice that pronounces an English word is to voice that speaks the carrier phrase, which is our target voice, on the scale of 1-5 (5 is very similar). For intelligibility, the subject were asked how well they can understand the pronounced English word on the scale of 1-5 (5 is very well). Mean opinion scores of both measures are reported in the following figure.

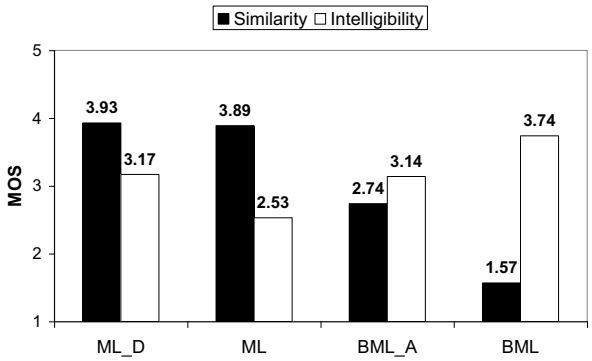


Figure 4: *MOS* of ML_D, ML, BML_A and BML

The result shows that English words synthesized by our proposed system (BML_A) are more intelligible than the ones synthesized with Thai phones generated by a G2P converter (ML). For the case where there is no pronunciation error from G2P (ML_D), the intelligibility is about the same as our system. However, it is difficult to improve the accuracy of the English-grapheme-to-Thai-phoneme converter to the level that is close to the oracle as the current level of accuracy is about 62% .When comparing between ML_D and BML, we found

that when there is no pronunciation error, an English word synthesized with a Thai pronunciation is less intelligible. One reason for this is that it is difficult for a Thai synthesizer to produce good quality speech from uncommon phone sequences occur in a Thai pronunciation of an English word.

When comparing the intelligibility between BML_A and BML, we found that speaker adaptation affects intelligibility. This may come from the fact that the quality of the adapted speech is not as good as the average voice since we adapt the average model with synthesized speech. Nevertheless, this means that it is possible to improve the intelligibility of our proposed system by improving quality of the adapted speech. In terms of similarity, we found that the proposed adaptation method can make the voice of the English HTS sounds more similar to the voice in the Thai HTS. Nevertheless, there is still room for improvement.

4. Conclusions

Our bilingual Thai-English TTS combines two monolingual HMM-based TTS systems, one for Thai and another one for English. The use of an English HTS system is to improve the intelligibility of synthesized English words as an English pronunciation dictionary or English G2P would give more accurate pronunciations of English words. The results from the subjective listening test show that the English HTS (BML) achieved the highest intelligibility score compared to the system that uses the English-grapheme-to-Thai-phoneme converter (ML) even when the generated Thai pronunciations are correct (ML_D).

To make the voices from both HTS systems sound more similar, we adapt the voice in the English HTS toward the voice in the Thai HTS with CSMAPLR adaptation. The adaptation data are English words synthesized by the Thai HTS with Thai pronunciation transcriptions, but are labeled with their corresponding English pronunciation transcriptions in the dictionary. This proposed cross-language pronunciation mapping through the parallel phone set pronunciation dictionary allows us to adapt English phones in the English HTS without having to create an explicit mapping between a Thai phone set and an English phone set. The labeling is done automatically using a phone boundary specification module in a phone recognizer. We also adapt the English mono-phone model toward the synthesized adaptation data with MLLR and MAP and found that its labeling accuracy is much better than the non-adapted model and the resulted phone boundaries and close to those obtained manually.

Our proposed system (BML_A) achieved a higher intelligibility score than the existing bilingual Thai-English TTS (ML). Moreover, with the proposed adaptation method, our system produces English words that sound more similar to synthesized Thai words. However, we found that speaker adaptation affects intelligibility due to the quality of adaptation data. Nevertheless, this means that it is possible to improve the intelligibility of our proposed system by improving quality of the adapted speech. In future, we plan to create a speech corpus that is parallel to the CMU ARCTIC but in Thai accent to improve the adaptation performance.

5. Acknowledgements

The authors would like to appreciate all testers who spend their valuable time to score our synthesized speech.

6. References

[1] Liang, H., Qian, Y., and Soong, F. K., “An HMM-Based Bilingual (Mandarin-English) TTS.” ISCA Workshop on Speech Synthesis (SSW-6), 137-142, 2007.

[2] Thangthai, A., Wutiwiwatchai, C., Ragchatjaroen, A., and Saychum, S., “A Learning Method for Thai Phonetization of English Words.”, Interspeech, 1777-1780, 2007.

[3] Chomphan, S. and Kobayashi, T., “Implementation and Evaluation of an HMM-Based Thai Speech Synthesis System.” Interspeech, 2849-2852, 2007.

[4] Kominek, J. and Black, A.W., “The CMU ARCTIC Speech Databases.”, ISCA Workshop on Speech Synthesis (SSW-5), 223-224, 2004.

[5] Liang, H., Qian, Y., Soong, F. K. and Liu, G., “A Cross-Language State Mapping Approach to Bilingual (Mandarin-English) TTS”, Acoustics, Speech, and Signal Processing (ICASSP), 4641 – 4644, 2008.

[6] Y.-J. Wu, S. King and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis", Proc. of ISCSLP 2008, Kunming, China, Dec.2008.

[7] Nakano, Y., Tachibana M., Yamagishi, J., and Kobayashi, T., “Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis.” Interspeech, 2286-2289, 2006.

[8] Yamagishi, J., Kobayashi, T., Nakano Y., Ogata, K. And Isogai, J., “Analysis of Speaker Adaptation Algorithm for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm.” IEEE Tran. Audio, Speech, and Language Proc., 17(1):66-83, 2009.

[9] Zhao, Y., Wang, L., Chu, M., Soong, F.K., and Cao, Z., “Refining Phoneme Segmentations Using Speaker-Adaptive Context Dependent Boundary Models.”, Interspeech, 2557-2560, 2005.

[10] Black, A.W., Taylor, P., and Caley, R., "Festival Speech Synthesis System," University of Edinburgh. Online: http://www.cstr.ed.ac.uk/projects/festival/

[11] Hansakunbuntheung, C., Tesprasit, V., and Sornlertlamvanich V., “Thai Tagged Speech Corpus for Speech Synthesis.”, The Oriental COCOSDA, 97-104, 2003.