Accent Level Adjustment in Bilingual Thai-English Text-to-speech Synthesis

Chai Wutiwiwatchai, Ausdang Thangthai, Ananlada Chotimongkol,
Chatchawarn Hansakunbuntheung, Nattanun Thatphithakkul

Human Language Technology Laboratory, National Electronics and Computer Technology Center (NECTEC)
112 Pahonyothin Rd., Klongluang, Pathumthani 12120, Thailand

{chai.wutiwiwatchai, ausdang.thangthai, ananlada.chotimongkol, chatchawarn.hansakunbuntheung, nattanun.thatphithakkul}@nectec.or.th

***Abstract*—This paper introduces an *accent level* adjustment mechanism for Thai-English text-to-speech synthesis (TTS). English words often appearing in modern Thai writing can be speech synthesized by either Thai TTS using corresponding Thai phones or by separated English TTS using English phones. As many Thai native listeners may not prefer any of such extreme accent styles, a mechanism that allows selecting accent level preference is proposed. In HMM-based TTS, adjusting the accent level is done by interpolating HMMs of purely Thai and purely English sounds. Solutions for cross-language phone alignment and HMM state mapping are addressed. Evaluations are performed by a listening test on sounds synthesized with varied accent levels. Experimental results show that the proposed method is acceptable by the majority of human listeners.**

I. Introduction

A lot of languages in the world often contain loan words written in either their original script or the local script using transliteration. Text-to-speech synthesis (TTS) hence needs to be able to handle bilingual text or monolingual text containing transliterated loan words. For the Thai-English example, a traditional way is to introduce a letter-to-sound module that could transcribe English words using Thai phones [1]. However, transcribing English words with a phone set of another language often produces non-native accented sounds, which are sometimes not preferred. In some Thai TTS systems, additional phones borrowed from English have been included to improve the system ability to synthesize English sounds. Yet, the mixing of Thai and English phones in transcriptions causes unnatural synthesized voices. A simple way to solve this problem is to use two separated TTS systems one for Thai and another one for English. A bilingual speaker is required to record two sets of training utterances, in Thai and in English. However, given pure English sentences, bilingual speakers often pronounce with a native-like English accent. Voices produced by such Thai-English parallel TTS systems could therefore sound strange from the switching between extreme Thai and English accents. A straightforward way to solve this problem might be to ask the bilingual speakers to read bilingual text instead of two sets of monolingual text. Nevertheless, constructing bilingual text that is phonetically balanced for both languages is highly difficult.

Non-native speakers of English like Thai may not prefer extreme accent styles, neither Thai nor native English. It is thus interesting to provide a mechanism where TTS users could select their preferred accent level, which locates somewhere in between the two extreme accent styles. In Hidden Markov Model (HMM) based TTS [2], HMMs are built for each specific context-dependent phone. One advantage of HMM-based speech synthesis is an ability to do HMM interpolation given two HMMs [3]. This feature allows the mixing of sounds from different HMM sources for several purposes such as for creating non-privacy voices. This ability has also been adopted in dialect modeling [4] and emotional expression and speaking style modeling [5]. This paper applied this ability for cross-language accent level adjustment.

The accent level can basically be adjusted by means of interpolation between HMMs of English phones and HMMs of corresponding Thai phones. Given a loan word, two problems arise when mapping between Thai and English phones: handling missing phones, i.e. phones omitted in one language but appeared in the other language, and aligning a phone cluster possibly defined in one language to a single phone defined in the other language. This paper explains how these problems are solved. Evaluations are based on listener preference tests given by Thai native listeners on sample utterances generated from bilingual text with different HMM interpolation weights.

The next section reviews bilingual Thai-English TTS approaches. Section 3 proposes the accent level adjustment method. Section 4 shows experiments and discussion, and Section 5 concludes this paper.

II. Bilingual Thai-English TTS

Modern Thai writing often contains the English script as shown in Figure 1. In conventional Thai text-to-speech synthesis, dealing with English word could be done by finding a transcription based on the Thai phone inventory. For example, the word “markets” transcribed as /M AA R K AH T S/ by the CMUDICT is transcribed using Thai SAMPA phones as /m aa k e t/. A lot of sounds appearing in English are distorted when pronounced in Thai e.g. the English phone /AH/ is somewhat different from the Thai phone /e/, and the English phone /S/ is omitted in Thai. This extreme style which pronounces English words using only Thai phones might not be acceptable especially in the current multilingual society.

A straightforward way to improve the English accent in bilingual TTS is by constructing two separated TTS engines one for each language. The system is able to switch between

the two engines according to the input text chunk it found. In order to get the same sound in both languages, a professional bilingual speaker is requested to record phonetically balanced sentences of both languages. A letter-to-sound converter as well as a speech synthesizer of each language will be built from each language corpus using different sets of phones. To build a bilingual Thai-English TTS system in this paper, a bilingual speaker is asked to read 2,712 Thai sentences taken from the TSynC-2 corpus [4] and 1,132 English sentences in the CMU-ARCTIC corpus [5]. The Thai TTS consists of a Thai word segmentation module and a syllable n-gram based letter-to-sound conversion module. The English TTS utilizes the Festival text processor. Speech synthesizers of both languages are based on context-dependent phone HMM [2]. There are 75 phones in Thai and 41 phones in English, details of which are in [4] and [5].

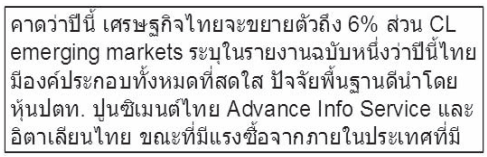

Fig. 1. An example of bilingual text in modern Thai writing

Since the professional speaker usually gives a good English accent when reading English text, bilingual TTS will result in unnatural sounds caused by the switching between Thai and English accents. An ideal way to alleviate the problem is by asking the voice talent to read genuine bilingual text with smoothed accents when switching languages. However, this approach is prohibited by the difficulty of preparing bilingual prompts that are phonetically balanced on both languages.

III. Accent Level Adjustment

Non-native speakers of English like Thai may not prefer an extreme English accent when listening to English words in Thai text. On the other hand, they may also not prefer to listen to English scripts in the purely Thai accent. It is, thus, reasonable to introduce a mechanism where TTS users could select their preferred accent level located somewhere in between the two extreme accent styles.

A. HMM Interpolation

Thanks to HMM-based speech synthesis, two HMMs can be interpolated to form a new HMM, which is likely to produce sounds mixed from those produced by the two original HMMs. The HMM conventionally used in speech modeling consists of Gaussian mixtures in each state. Given N HMMs containing S states with M Gaussian mixtures per state, interpolation is done independently for each Gaussian mixture in each state by

$$\mu_m = \sum_{i=1}^{N} \alpha_i \mu_{mi}, \quad \Sigma_m = \sum_{i=1}^{N} \beta_i \Sigma_{mi} \tag{1}$$

where μ_{mi} and Σ_{mi} are the mean vector and the covariance matrix of the m-th Gaussian mixture of the i-th HMM; μ_m and Σ_m are the interpolated mean vector and covariance matrix of the m-th Gaussian component; α_i and β_i are interpolation weights for the mean and covariance respectively. In the case of HMM-based TTS [2], two sets of Gaussian mixtures are stored in each state, one for the vector of spectral and fundamental frequency (F0) parameters, and the other for the scalar state duration. In interpolation, Eq. (1) is adopted for both Gaussian mixture sets. However, it is noted that in the HMM-based TTS [2], only the mean value of the phone state duration model is used. Therefore, there is no need to deal with the variance of the duration model.

The proposed accent level adjustment is implemented by interpolating HMMs from two extreme accent styles, purely Thai and purely English. It is noted that, the number of states and the number of mixtures per state in the HMMs being interpolated must be the same.

B. Thai-English Phone Alignment

Given sequences of phones from both languages, the first task is to align the phone sequences so that phone-based HMMs of the two languages are correctly matched for interpolation. There are three basic cases possibly occurring in phone alignment as shown in Figure 2. Figure 2 (a) is the simplest case where phones are *one-to-one mapping*. In this first case, HMM interpolation for each phone pair is straight forward.

Figure 2 (b) and (c) show the other two problematic cases. Figure 2 (b) illustrates the *one-to-many mapping* case where one phone in one language is aligned with more than one phone in the other language. *Many-to-one mapping* is considered the same case. This mapping happens in three ways, sharing of the final and the initial consonant in Thai e.g. the English sound /N/ in the word “manner”, consonant clusters e.g. the Thai sound /kr/ in the word “cry”, and syllabic rhymes e.g. the English sound /AY/ in the word “cry”.

In the one-to-many phone mapping case, a single HMM is aligned to multiple HMMs. To make the interpolation technique applicable, the single HMM is modified to have an equivalent number of states to the multiple HMMs. Figure 3 demonstrates a modification approach. Suppose that the phone /AY/ is modeled by a three-state HMM and its corresponding Thai phones are /aa j/ which are represented by two three-state HMMs. In order to lengthen the /AY/ HMM, each HMM state will be replicated according to the duration occupancy of the state. For example, suppose that duration means of the three HMM states of the sound /AY/ are 14, 20, and 6, the duration occupancies of such three states are 35%, 50%, and 15% respectively. The new six-state HMM stretched from the original three-state HMM will contain two copies of the first source state, three copies of the second source state, and one copy of the third source state, corresponding by approximation to the duration occupancy proportion. Finally, the original duration means of source states will be distributed uniformly to their corresponding target states. In this example, duration means of the 6 new states will be 7, 7, 6.7, 6.7, 6.7, and 6.

The last scenario occurs when some phones in one language are omitted in the other language. This *one-to-zero mapping* case often occurs when transcribing English words

by Thai phones. Again, *zero-to-one mapping* is considered the same case. Some English sounds especially final consonant clusters are usually pronounced as a single Thai final consonant. For example, the final English sound /N D/ in the word "sand" shown in Figure 2 (c) is pronounced as only a single Thai sound /n/, while the sound /D/ is omitted. It is noted that in this case, the sound is totally omitted, not combined with its neighboring phone.

Figure 4 shows how to perform HMM interpolation in the one-to-zero phone mapping case. A *zero-duration* HMM, represented by the dash line, is inserted to match the missing phone. The zero-duration HMM has the same emission probabilities as the omitted phone /D/, but its state duration distributions have all zero means. Therefore, interpolating between the /D/ HMM and the zero-duration HMM results in the same emission probabilities with state durations adjusted. Shortening the sound by this proposed method might cause unnatural artifacts, but it might be acceptable if the sound being compressed is relatively short as the example sound /D/ at the end of the syllable "sand".

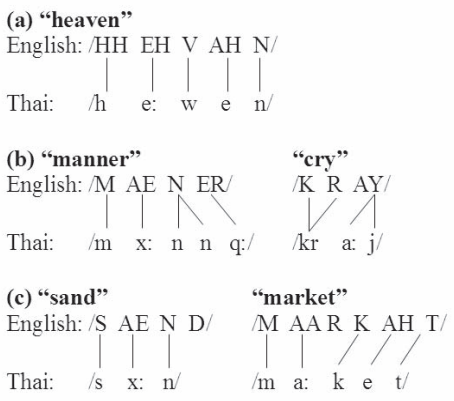

Fig. 2. Examples of phone alignment, (a) one-to-one, (b) one-to-many, and (c) zero-to-one mapping. Phone notations are defined by the CMUDICT for English and by SAMPA for Thai

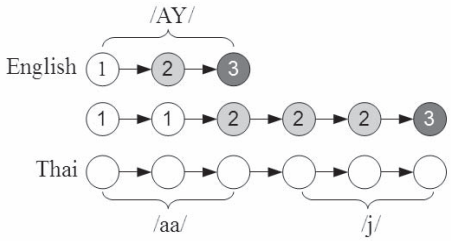

Fig. 3. HMM state extension in one-to-many phone mapping

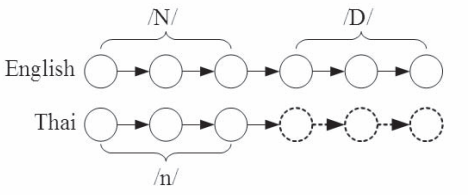

Fig. 4. HMM state extension in one-to-zero phone mapping

It is noted that Thai transcription normally contains syllabic tone marks, 1 to 5 for five Thai tones according to SAMPA. This feature has also been incorporated in HMM decision trees. Interpolation among English and Thai HMMs can be performed regardless of the Thai tone as the F0 parameter included in the model will reflect the change from tonal sounds in Thai to non-tonal sounds in English automatically.

C. Generalization of the Proposed Method

The proposed phone alignment and HMM state mapping methods could possibly be generalized for some other languages. The bilingual scripts do not only happen in Thai, but also appears in many languages such as Japanese and Mandarin Chinese. New English names including abbreviations are mostly written as is. Therefore, the mechanism of accent level adjustment might be able to adopt in the same way. Figure 5 demonstrates two examples of English words transcribed originally in CMU English phones and in Japanese and Mandarin. The alignment shows that there are only three cases similarly to what proposed for bilingual Thai-English. Hence, the proposed HMM interpolation is likely to be applicable.

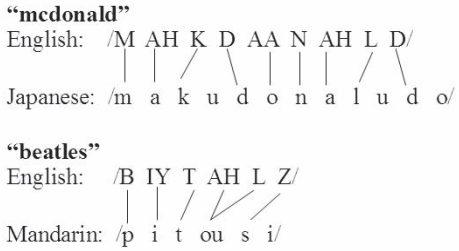

Fig. 5. Examples of phone alignments for bilingual English-Japanese and English-Mandarin cases

IV. Experiments

A. Experimental Setting

The "BSynC" bilingual speech database used in our experiments contains 2,712 Thai and 1,132 English speech utterances, recorded using sentence prompts taken from the Thai TSynC-2 corpus [4] and the CMU-ARCTIC corpus [5]. A bilingual female speaker is a Thai linguist with an English doctoral major. Her English speaking style is not the focus of this paper. The speaker was asked to read the prompts in a quiet room with simultaneous checking of the sound quality during the recording. Table I summarizes key characteristics of the database.

The BSynC database was used to train HMMs of two phone sets, 75 Thai phones and 41 English phones, using the HTS toolkit [2]. Context-dependent HMMs with 5 states per HMM and 1 Gaussian mixtures per state were built based on decision trees separately for each language. 24-order Mel-cepstral coefficients (MCEP) and a fundamental frequency (F0), and their first and second derivatives were used as speech parameters. The phone state duration was modeled separately using a Gaussian Mixture Model (GMM). A diagonal covariance matrix was set for all Gaussian models to

reduce the computational complexity. This bilingual TTS system is able to generate native Thai sounds for Thai scripts and fairly good English accented sounds for English scripts.

The experiments were based on a listening test. To evaluate the capability of the proposed accent level adjustment mechanism, two extreme cases of accent styles for synthesized English words were prepared as follows:

- a purely English accent, where the sound was synthesized using the English phone HMMs, and
- a purely Thai accent, where the sound was synthesized using the Thai phone HMMs corresponding to the Thai-phone transcription of tested English words.

Three more sounds with different accent levels, 75% English and 25% Thai, 50% English and 50% Thai, and 25% English and 75% Thai, were interpolated from the above two extreme accents. For simplicity, the English-to-Thai ratio will be denoted hereafter as "%E-%T", e.g. 75E-25T for 75% English and 25% Thai, 100E-0T for purely English. In all experiments, the interpolation weights α_i and β_i are set equally for simplicity. Each test sentence was speech synthesized with five variations of the accent level described above, resulting in five utterances per test sentence. The utterances were presented to 22 Thai native listeners (13 males, 9 females), aged from 21 to 45 (28 in average). The listeners were asked to select one utterance out of 5 based on their best preference on the accent level of English word sounds. Exact transcriptions both in English and Thai phones as well as their alignments were given by a phonetician.

There were two sets of test sentences. The first set was for verifying the applicability of the interpolation algorithm applied on the three phone alignment cases, *one-to-one*, *one-to-many*, and *one-to-zero* mapping. One bilingual sentence was designed for each case. A sentence is actually a concatenation of small bilingual phrases appeared in web text. English words appearing in the phrases must be used quite often in Thai scripts, while the number of English words and the number of Thai words in each sentence are kept approximately equal. Table II presents statistics of this first test set.

TABLE I
A SUMMARY OF THE BSYNC BILINGUAL CORPUS

Feature	Thai	English
No. of utterances	2,712	1,132
No. of words	37,952	10,045
No. of phones	115,169	39,153
No. of unique phones	75	41
Length (hrs.)	4.0	1.1

TABLE II
STATISTICS OF THE FIRST TEST SET USED FOR VERIFYING INTERPOLATION RESULTS OF THREE PHONE ALIGNMENT CASES

Phone mapping case	Total no. of words	Total no. of English words	Total no. of case phones
One-to-one	15	6	46
One-to-many	18	6	7
One-to-zero	15	7	9

The second test set was for overall preference testing. Ten bilingual sentences mixed with different cases of phone mapping were prepared, again by extracting from web text. Indeed, one factor that could highly affect the listener preference is how often the English word appeared in Thai scripts. In the extreme case where an English word is usually transliterated into Thai, such word is likely to be included in common Thai dictionary. Thai native listeners may prefer listening to Thai accented sounds of these *common* English words rather than the English accented sounds. On the other hand, for *uncommon* English words, the listeners may prefer more English accent. The second test set then consists of two subsets of bilingual sentences separated by the kind of English words included, common or uncommon. The common English word means an English loan word appearing in the Thai Royal Institute dictionary (http://www.royin.go.th). Table III summarizes the statistics of this second test set.

TABLE III
STATISTICS OF THE SECOND TEST SET USED FOR OVERALL PREFERENCE TESTING

Statistics	Sentences with only common English words	Sentences with only uncommon English words
No. of sentences	5	5
Avg. # words/sent.	19	20
Avg. # Eng words/sent.	6	9

B. Experimental Results and Discussion

Figure 6 reports the listening test result from the first test set. The result clearly shows that the majority of listeners prefer to listen to mixed Thai and English accents rather than the extreme Thai or English accent on every phone alignment case. This result confirms our hypothesis and suggests that each listener should be allowed to adjust the accent level according to his/her own preference. Even though we did not measure the naturalness of interpolated sounds directly using the measure such as MOS, the preference test result could imply that there is no severe degradation in the sound quality from the proposed phone alignment approach and accent level interpolation technique as the interpolated accents are still chosen by many test listeners.

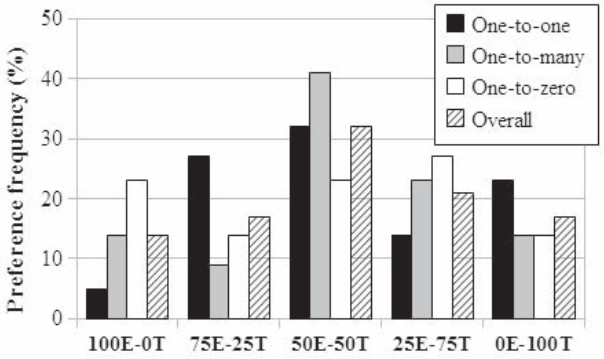

Fig. 6. Preference test results of the first test set

Figure 7 shows comparative results among test utterances containing only *common* and only *uncommon* English words.

It is noted again that common English words mean words appearing in the Thai Royal Institute dictionary. It is clear that Thai native listeners tend to prefer more Thai-accented sounds for English words commonly used in Thai. For uncommon English words, mixed accent and more English-accented sounds are preferred. Figure 8 interestingly shows that the listeners' field of work highly influences their accent level preference. Younger people like Bachelor students tend to prefer more Thai-style accents. Linguists tend to prefer a mixed accent, whereas Engineers seem to prefer either of the two extreme accents. While the English skills of listeners have not been clearly examined, the relationship between their skills and preference points is of interest. This issue will be explored in the future work.

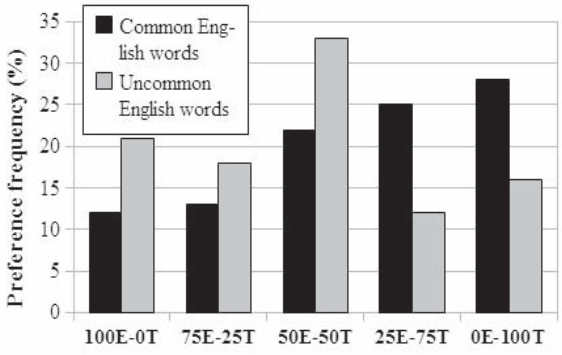

Fig. 7. Preference test results on the second test set containing either common or uncommon English words

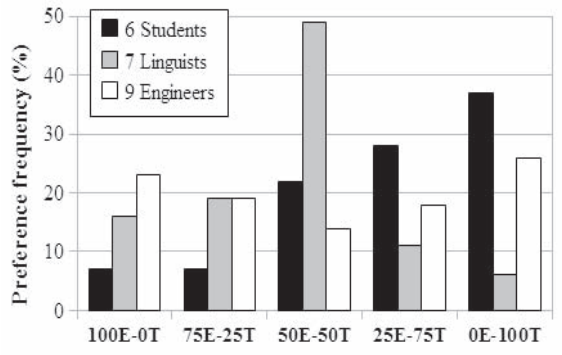

Fig. 8. Preference test results on the second test set distributed by the listener field of work

V. CONCLUSION

An *accent level* adjustment mechanism is introduced for bilingual text-to-speech synthesis (TTS). Having HMMs trained by speech utterances of each language, accent level adjustment is performed by doing cross-language phone alignment, HMM state mapping, and HMM interpolation. The unequal number of phones in the scripts of two languages for a given word is mainly caused by different definitions of phones, phone clusters, and the phone omitting nature in some languages. The one-to-many phone mapping case is handled by duplicating HMM states, i.e. duplicating state emission and duration distributions. The one-to-zero phone mapping case is handled by copying the omitted HMM states but setting state durations to be zero. Mixed-accent sounds generated by the interpolated HMM were best preferred by many test listeners. The result verified the applicability of the proposed accent level adjustment method and implied that the sound quality and naturalness were not severely degraded after interpolating. Future works are to carefully analyze whether there is a systematic way that non-native English speakers reduce from the native English accent to their local accent. Such phonological change aspects have been investigated in the Austrian German and Viennese dialect [6], for example. Having such knowledge will help improving the HMM state alignment and interpolation.

REFERENCES

[1] Thangthai, A., Wutiwiwatchai, C., Ragchatjaroen, A., Saychum, S., "A learning method for Thai phonetization of English words", In Proc. INTERSPEECH 2007, pp. 1777-1780, 2007.

[2] Tokuda, K., T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features", In Proc. ICASSP 1995, pp. 660-663, 1995.

[3] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T., "Speaker interpolation for HMM-based speech synthesis system", The Journal of the Acoustical Society of Japan (E), vol.21, no.4, pp. 199-206, Jul. 2000.

[4] Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., Strom, V., "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis", Speech Communication 52 (2010), pp. 164-179.

[5] Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T., "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing", IEICE Trans. Inf. Syst. E88-D (11), 2484–2491.

[6] Wutiwiwatchai, C., Saychum, S. and Rugchatjaroen, A., "An intensive design of a Thai speech synthesis corpus", in Proc. SNLP 2007, pp. 201–206, 2007.

[7] Kominek, J. and Black, A., "The CMU ARCTIC speech databases for speech synthesis research," Tech. Rep. CMULTI-03-177, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.