INCORPORATION OF PHRASE INTONATION TO CONTEXT CLUSTERING FOR AVERAGE VOICE MODELS IN HMM-BASED THAI SPEECH SYNTHESIS

Suphattharachai Chomphan, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
Email: {suphattharachai,takao.kobayashi}@ip.titech.ac.jp

ABSTRACT

This paper describes a novel approach to the context clustering process in a speaker independent HMM-based Thai speech synthesis for improvement of the tone intelligibility of the average voice and also the speaker adapted voice. A couple of phrase intonation features from a generative model including a baseline value of fundamental frequency and a phrase command amplitude are extracted and thereafter exploited in the context clustering process of HMM training stage. In the experiments, subjective evaluations of both average voice and adapted voice in terms of the intelligibility of tone are conducted. The results show that the tone correctness of the synthesized speech is significantly improved.

Index Terms— Phrase intonation, Thai tone, average voice, speech synthesis, hidden Markov models

1. INTRODUCTION

In the present days, the HMM-based speech synthesis is becoming popular among the speech synthesis research area thanks to its ability to generate speech with arbitrary speaker's voice characteristics and various speaking styles. There have been proposed a number of TTS techniques, and state-of-the-art TTS systems based on unit selection and concatenation can generate natural sounding speech. However, to provide various voice characteristics in speech synthesis systems based on the speech unit selection approach, a large amount of speech data is necessary and it is burdensome to obtain enough speech data [1].

For tonal languages such as Thai, Mandarin, Cantonese, and Vietnamese, tone is a very important suprasegmental feature of syllables. The words with the same phoneme sequence may have different meanings if they have different tones [2]. Thus, tone must be carefully taken into account in speech synthesis systems of tonal languages.

As for speaker dependent HMM-based Thai speech synthesis research, we have developed an HMM-based speech synthesizer [3]. In the developed system, a group of contextual factors which affect spectrum, fundamental frequency (F0), and state duration, such as tone type and part of speech are taken into account especially for the purpose of producing natural sounding prosody of the tonal language. We have found that it can provide speech with the better reproduction of prosody over the unit-selection-based Vaja TTS system from NECTEC (National Electronics and Computers Technology Center) [4]. Specifically, a decision tree with a tone-separated structure [3] shows the significant improvement of tone correctness of the synthesized speech. However, some distortion of syllable duration is obviously noticeable when the system is trained with a small amount of data. Some other structures of the decision tree are designed for not only the purpose of maximal correctness of tone but also the purpose of elimination of the syllable duration distortion [3].

In this context, we have attempted to develop a speaker independent HMM-based Thai speech synthesis system. In the system, our speech database contains of quite a large number of speakers with a small amount of data for each speaker (see section 4.1 for details). Although it is desirable that sentence sets of speakers are different from each other to make database rich in phonetic and linguistic contexts, the synthetic speech generated from the average voice model [1] trained using different sentence set for each speaker sounds unnatural compared to the model trained using the same sentence set for all speakers, especially when the amount of training data of each speaker is limited. To treat the problem, the shared decision tree context clustering (STC) [1] is adopted, where every node of the decision tree always has training data from all speakers so that each distribution of the average voice model reflects the statistics of all speakers. Moreover, we also incorporate speaker adaptive training (SAT) [5] into the training procedure of the average voice model to improve the quality of the average model by using the training method in [6].

The naturalness of the synthetic speech generated from the above system is comparable to that of the speaker dependent system, however, the tone correctness of the synthetic speech is degraded considerably. The F0 trajectory of each syllable which indicates a syllable tone is distorted neutrally by the tree based context clustering. This problem does not usually appear in the speaker dependent system, but it is obviously perceived in the speaker independent system when multi-speaker speech database is

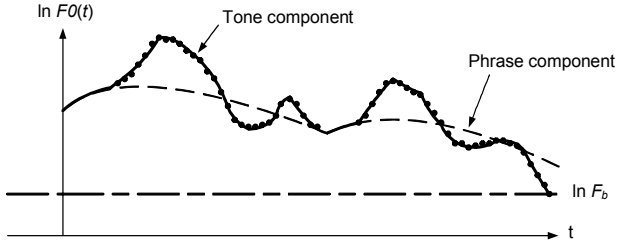

Fig. 1. Representation of F0 contour by generative model

used for training. We therefore propose incorporation of the phrase intonation to the context clustering process in the training stage. The phrase intonation features are extracted by using a generative model (Fujisaki's model) including a baseline value of F0 and a magnitude of phrase command [2]. The results of subjective evaluations of the proposed technique are also discussed in the paper.

2. PHRASE INTONATION FEATURES

2.1. Phrase Intonation Representation

The tone types in Thai have rather clear manifestations in the F0 contour in the case of isolated syllables as indicated by Seresangtakul [2]. However, the F0 contours are varied considerably in continuous speech due to the influences of such factors as tones of the adjacent syllables, syntactic and pragmatic information of the whole utterance, and the overall speaking rate. The main factors causing these variations are tone coarticulation, tone enhancement/suppression, and phrase intonation [7]. In the conventional HMM-based speech synthesis approach, the speech features including F0 values are modeled statistically. As for speaker dependent system, the intelligibility of tone is degraded significantly when using simple binary tree-based context clustering. This can be alleviated by modifying the tree structure as tone-separated structure [3,4]. Unfortunately, in the case of speaker independent system, the variety of the speaker characteristics and the small amount of training data from one speaker cause the neutralization in the F0 contour. As a result, the intelligibility of tone is considerably degraded in spite of modifying the tree structure. To reduce the effect of the variety of such F0 contours, the phrase intonation as indicated above is thought to be a promising factor. Its relevant features are the baseline value of F0 and the magnitude of phrase command of the generative model as depicted in Fig. 1. We therefore incorporate these features into our existing contextual factors to reduce the variations caused by the phrase intonation factor.

2.2. Feature Extraction using Generative Model

Fujisaki et al. have shown that the F0 contour generally contains a smooth rise-fall pattern in the vicinity of the accented Japanese mora [2]. The F0 contour is treated as a linear superposition of a global phrase and local accent components on a logarithmic scale. The phrase command produces a baseline component, while the accent command produces the accent component of an F0 contour. We use the two parameters of Fujisaki's model as our phrase intonation features including the baseline value of F0 and the magnitude of phrase command. Mathematically, an F0 contour of an utterance generated from an extension of Fujisaki's model for tonal languages has the following expressions [2];

$$\ln F0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi}[G_{pi}(t-T_{0i})] + \sum_{j=1}^{J}\sum_{k=1}^{K(j)} A_{t,jk}[G_{t,jk}(t-T_{1jk}) - G_{t,jk}(t-T_{2jk})], \quad (1)$$

$$G_{pi}(t) = \begin{cases} (\alpha_i^2 t)\exp(-\alpha_i t) & for\ t \geq 0 \\ 0 & for\ t < 0, \end{cases} \quad (2)$$

$$G_{t,jk}(t) = \begin{cases} [1-(1+\beta_{jk}t)\exp(-\beta_{jk}t)] & for\ t \geq 0 \\ 0 & for\ t < 0, \end{cases} \quad (3)$$

where $G_{pi}(t)$ represents the impulse response function of the phrase control mechanism and $G_{t,jk}(t)$ represents the step response function of the tone control mechanism, respectively. The symbols in these equations indicate as follows: F_b is the smallest F0 value in the F0 contour of interest, A_{pi} and $A_{t,jk}$ are the amplitudes of the i-th phrases and of the j-th tone command. T_{0i} is the timing of the i-th phrase command; T_{1jk} and T_{2jk} are the onset and offset of the k-th component of the j-th tone command. α_i and β_{jk} are time constant parameters. I, J, $K(j)$ are the number of phrases, tones and components of the j-th tone contained in the utterance, respectively.

To extract the best representative parameters from the model, the optimization is carried out by minimizing the mean squared error in the ln $F0(t)$ domain through the hill-climbing search in the space of model parameters.

3. FEATURE ARRANGEMENT FOR CONTEXT CLUSTERING

By using the extraction algorithm described in section 2.2, two features are extracted for all training utterances. These features are then prepared to be employed as contextual factors in the context clustering process. As for the first feature of baseline value of F0 (F_b), it ranges from 67.7 Hz to 178.8 Hz, while the second feature of amplitude of phrase command (A_{pi}) ranges from 0.00 to 1.20. Both of them are linearly quantized into 8 classes with an assigned codeword of 0-7. These features are then grouped into two sets (S11, S12) in the phrase level of the following contextual factors as constructed in [4]. Note that our point is to indicate the level of phrase intonation for the current phone, therefore both features have to be used together. As a result, the feature of baseline value of F0 is not classified into the utterance level, although each utterance has its unique value.

- Phoneme level

S1. {preceding, current, succeeding} phonetic type
S2. {preceding, current, succeeding} part of syllable structure

- Syllable level

S3. {preceding, current, succeeding} tone type
S4. the number of phones in {preceding, current, succeeding} syllable
S5. current phone position in current syllable

- Word level

S6. current syllable position in current word
S7. part of speech
S8. the number of syllables in {preceding, current, succeeding} word

- Phrase level

S9. current word position in current phrase
S10. the number of syllables in {preceding, current, succeeding} phrase
S11. codeword of baseline value of F0
S12. codeword of amplitude of phrase command

- Utterance level

S13. current phrase position in current sentence
S14. the number of syllables in current sentence
S15. the number of words in current sentence

In the synthesis stage, the parameter generation algorithm is mostly the same as the previous system [3,4] except for adding the appropriate quantization codewords for these proposed features in the context labels. The mean values of the proposed features for a specific target, which are the a priori knowledge, are statistically expected to be the best representatives. For examples, to generate an average voice, we choose the mean values from all speakers in the training data; 127.4 Hz and 0.40 for F_b and A_{p1}, corresponding to the quantization codewords of 4 and 2, respectively. To generate an adapted voice of a female target speaker, we choose the mean values of the female; 154.6 Hz and 0.38 for F_b and A_{p1}, corresponding to the quantization codewords of 5 and 2, respectively. These statistical figures are from our speech databases as mentioned in section 4.1. Moreover, in the sentence with more than one phrase, the *i*-th successive phrase also needs an associated A_{pi}. The representatives for these amplitudes of phrase command can be obtained by the same statistical method.

4. EXPERIMENTS

4.1. Experimental Conditions

A set of phonetically balanced sentences of Thai speech database named LOTUS from NECTEC [8] was used for training HMMs. Another set of phonetically balanced sentences of Thai speech database named TSynC-1 from NECTEC [9] was used for adaptation and also for a speaker dependent system. The whole sentence text of both databases was collected from Thai part-of-speech tagged ORCHID corpus. In LOTUS, the speech in the database was uttered by 24 female and 24 male speakers with clear articulation and standard Thai accent, while the speech in the TSynC-1 was uttered by a professional female speaker. The phoneme labels included in both databases and the utterance structure from ORCHID were used to construct the context dependent labels with 79 different phonemes including silence and pause.

Speech signal were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were extracted by mel-cepstral analysis. The feature vectors consisted of 25 mel-cepstral coefficients including the zero-th coefficient, logarithm of F0, and their delta and delta-delta coefficients [10].

We used 5-state left-to-right HSMMs [4] in which the spectral part was modeled by a single diagonal Gaussian output distribution. Each context dependent HSMM corresponds to a phoneme-sized speech unit. The average voice model was trained using 35 sentences for each speaker from 24 female and 24 male speaker's speech data. To evaluate the proposed approach, we constructed 4 different training systems including a speaker dependent system using 1,500 training utterances (a reference system), and three speaker independent systems: the systems without using STC and SAT, using SAT and STC, and using SAT and STC with the tone-separated tree structure. In the following results, the entries for "SD", "NONE", "STC+SAT", and "STC+SAT+SEP" correspond to these training systems respectively. In addition, the MLLR-based speaker adaptation [11] with 100 utterances of a female target speaker was used to generate the adapted voice. The proposed technique are applied to all of three speaker independent systems, the intelligibility of tone is subsequently measured as shown in sections 4.3 and 4.4.

4.2. Influence of Phrase Intonation Features on Clustering Trees

We first investigated how the phrase intonation features affect the clustering trees including mel-cepstrum (mcep), logF0, and duration (dur) trees. The increasing in percentage of the number of existing questions and the dominance score [3,4] for the questions in phrase level is summarized in Fig. 2. It can be seen from both criteria that the logF0 tree is affected most while the mcep tree is affected least. This also reflects the relationship between the phrase intonation features and those three speech features.

Fig. 3 shows the effects of phrase intonation features which cause the changes in the F0 contour levels of some syllables. It can be obviously seen that proposed features can reduce the difference between those of the SD and the STC+SAT+SEP training systems.

4.3. Subjective Evaluations of Average Voice

This section presents how the overall tone correctness of the average voice is improved by embedding the proposed features in section 3 in the different training systems. The tone error

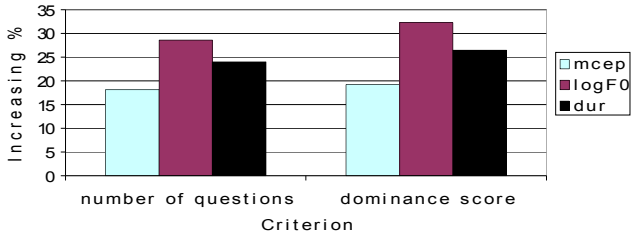

Fig. 2. Increasing percentages of the number of existing questions and the dominance score for the questions in phrase level from different clustering trees

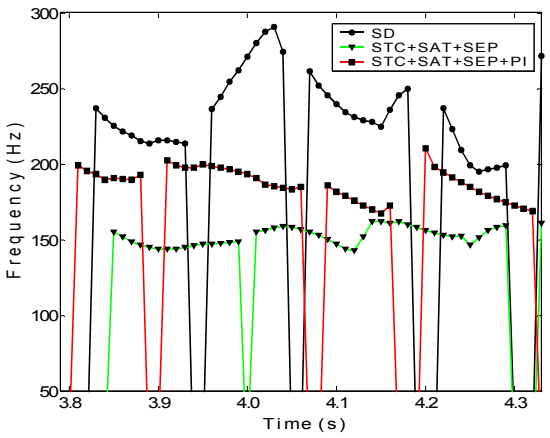

Fig. 3. Examples of generated F0 contours from SD, STC+SAT+SEP, and STC+SAT+SEP+PI (phrase intonation applied) training systems.

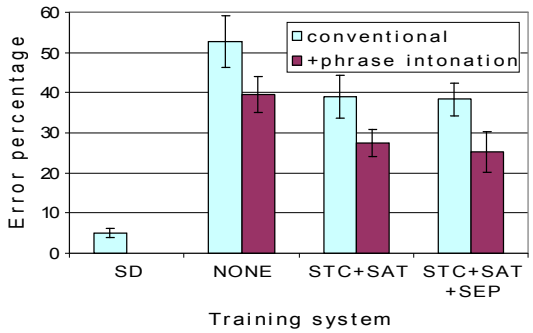

Fig. 4. Tone error percentage of average voice synthesized from different training systems.

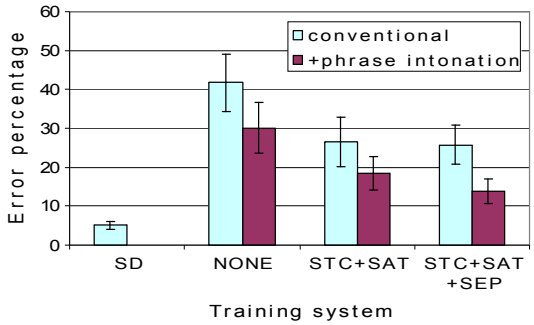

Fig. 5. Tone error percentage of adapted voice synthesized from different training systems.

percentage is the measured value in this comparison [3,4]. To calculate tone error percentage of our implemented systems, a subjective test was conducted. The 2,289 syllables of 100 synthesized speech utterances were presented to eight native subjects. Then the subjects were requested to decide whether the syllables have the same tones as the given texts or not. The average tone error percentages with 95% confidence interval for different training styles are summarized in Fig. 4. It can be seen that the proposed features can reduce the tone error percentage significantly for all training systems. Applying the proposed features with the STC+SAT+SEP training system achieves the lowest level, in other words, it has the highest tone intelligibility.

4.4. Subjective Evaluations of Adapted Voice

As for evaluating the adapted voice, the experimental conditions were the same as the evaluation test described in the previous section. The average tone error percentages with 95% confidence interval for different training styles are summarized in Fig. 5. The result of adapted voice mostly corresponds to that of the average voice with a lower level of the tone error. The percentage reduction of tone error is between 8 and 12%. Applying the proposed features with STC+SAT+SEP training system achieves the lowest level of 13% which is still little higher than that of the SD training system (5%).

The synthesized speech samples are available on the website http://www.kbys.ip.titech.ac.jp/demo/thai/index.html

5. CONCLUSION

We have described a couple of phrase intonation features to be embedded in the contextual factors for the context clustering process of a speaker independent HMM-based Thai speech synthesis system. These features are extracted based on the parameter optimization of generative model. It is expected to reduce the variation of tone caused by phrase intonation both from intra- and inter-speaker. From the results of subjective tests, the tone intelligibility of the proposed approach is improved in every training system. Our future work will focus on the application of the proposed approach to speaking style.

6. ACKNOWLEDGEMENTS

The authors are grateful to NECTEC for providing us the LOTUS and the TSynC-1 speech databases.

7. REFERENCES

[1] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T, Kobayashi, “A context clustering technique for average voice model in HMM-based speech synthesis,” Proc. ICSLP-2002, pp.133-136, 2002.

[2] P. Seresangtakul and T. Takara, “Analysis and synthesis of pitch contour of Thai tone using Fujisaki's model,” IEICE Trans Inf. & Syst., Vol.E86-D, No.10, pp.2223-2230, 2003.

[3] S. Chomphan and T. Kobayashi, “Design of tree-based context clustering for an HMM-based Thai speech synthesis system,” The 6th ISCA Workshop on Speech Synthesis, pp.160-165, 2007.

[4] S. Chomphan and T. Kobayashi, “Implementation and evaluation of an HMM-based Thai speech synthesis system,” Proc. INTERSPEECH-2007, pp.2849-2852, 2007.

[5] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, “A compact model for speaker adaptive training,” Proc. ICSLP-96, pp.1137-1140, 1996.

[6] J. Yamagishi, T. Masuko, K. Tokuda, and T. Kobayashi, “A training method for average voice model based on shared decision tree context clustering and speaker adaptive training,” Proc. ICASSP-2003, pp.716-719, 2003.

[7] H. Fujisaki, C. Wang, S. Ohno, and W. Gu, “Analysis and synthesis of fundamental frequency contours of standard Chinese using the command-response model,” Speech communication 47, pp.59-70, 2005.

[8] C. Wutiwiwatchai and S. Furui, “Thai speech processing technology: a review,” Speech Communication 49, pp.8-27, 2007.

[9] C. Hansakunbuntheung, A. Rugchatjaroen, and C. Wutiwiwatchai, “Space Reduction of Speech Corpus Based on Quality Perception for Unit Selection Speech Synthesis,” Proc. SNLP-2005, pp.127-132, 2005.

[10] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Speech Synthesis using HMMs with Dynamics Features,” Proc. ICASSP-96, pp.389-392, 1996.

[11] J. Yamagishi, T. Masuko, and T. Kobayashi, “MLLR adaptation for hidden semi-Makov model based speech synthesis,” Proc. ICSLP-2004, pp.1213-1216, 2004.