Journal of Computer Science 6 (12): 1474-1478, 2010
ISSN 1549-3636
© 2010 Science Publications

Tone Question of Tree Based Context Clustering for Hidden Markov Model Based Thai Speech Synthesis

Suphattharachai Chomphan
Department of Electrical Engineering, Faculty of Engineering at Si Racha,
Kasetsart University, 199 M.6, Tungsukhla, Si Racha, Chonburi, 20230, Thailand

Abstract: Problem statement: In HMM-based Thai speech synthesis, tone is an important issue that brings about the intelligibility of the synthesized speech. Tone distortion resulted from imbalance of the training data should be appropriately treated. **Approach:** This study described an HMM-based speech synthesis system for Thai language. In the system, spectrum, pitch and state duration are modeled simultaneously in a unified framework of HMM, their parameter distributions are clustered independently by using a decision-tree based context clustering technique. The contextual factors which affect spectrum, pitch and duration, i.e., part of speech, position and number of phones in a syllable, position and number of syllables in a word, position and number of words in a sentence, phone type and tone type, are taken into account for constructing the questions of the decision tree. Since Thai is a tonal language, tone questions play an important role in the context clustering process. **Results:** An experimental result compared F0 contours between those of synthesized speech with and without tone questions; furthermore the size of Thai speech corpus is varied to investigate the synthesized speech quality. **Conclusion:** By using the tone questions in the tree-based context clustering process, the tone distortion is relieved significantly.

Key words: Thai tone, tree-based context clustering, HMM-based speech synthesis

INTRODUCTION

Speech synthesis is one of the key technologies for realizing natural human-computer interaction. For this purpose, Text-To-Speech (TTS) synthesis systems are required to have an ability to generate speech with arbitrary speaker's voice characteristics and various speaking styles. A number of TTS techniques have been proposed and state-of-the-art TTS systems based on unit selection and concatenation can generate natural sounding speech. However, it is still a difficult problem to synthesize speech with various voice characteristics and various speaking styles.

Hidden Markov model (HMM) based TTS system in which each speech synthesis unit is modeled by HMM is proposed in the past decade (Masuko *et al*., 1996; Yoshimura *et al*., 1999; Al-Haddad *et al*., 2008). A distinctive feature of the system is that the speech parameters used in the synthesis stage are generated directly from HMMs by using a parameter generation algorithm (Tokuda *et al*., 2000; Curran *et al*., 2005; Aliwa *et al*., 2010). Since the HMM-based TTS system uses HMMs as the speech units in both modeling and synthesis, the voice characteristics of synthetic speech can be changed by transforming HMM parameters appropriately.

As for Thai speech synthesis, a TTS synthesis system based on unit selection is initially implemented by Luksaneeyanawin in 1991 (Chomphan and Kobayashi, 2008). Subsequently, a TTS synthesis system based on unit selection with TD-PSOLA technique is developed by National Electronics and Computers Technology Center (NECTEC) in 2003 (Hansakunbuntheung *et al*., 2005). Since Thai is a tonal language, this study is proposed to implement a Thai speech synthesis based on HMM which has the ability of synthesizing speech with various voice characteristics and various speaking styles. In the tree-based context clustering stage, the tone question is applied to improve the overall speech quality. An experiment is conducted and it shows a considerable improvement.

MATERIALS AND METHODS

HMM-based speech synthesis: A block-diagram of the HMM-based TTS system is shown in Fig. 1. The system consists of two stages including the training stage and the synthesis stage (Tamura *et al*., 2001; Yamagishi *et al*., 2003). In the training stage, mel-cepstral coefficients are extracted at each analysis frame as the static features from the speech database. Then the dynamic features, i.e., delta and delta-delta parameters, are calculated from the static features. Spectral parameters and pitch observations are combined into one observation vector frame-by-frame and speaker dependent phoneme HMMs are trained using the observation vectors. To model variations of spectrum, pitch and duration, phonetic and linguistic contextual factors, such as phoneme identity factors, are taken into account (Yoshimura *et al*., 1999). Spectrum and pitch are modeled by mulyi-stream HMMs and output distributions for spectral and pitch parts are continuous probability distribution and Multi-Space probability Distribution (MSD) (Tokuda *et al*., 1999), respectively. Then, a decision tree based context clustering technique is separately applied to the spectral and pitch parts of context dependent phoneme HMMs (Young *et al*., 1994). Finally state durations are modeled by multi-dimensional Gaussian distributions and the state clustering technique is also applied to the duration distributions (Yoshimura *et al*., 1998). In the synthesis stage, first, an arbitrary given text to be synthesized is transformed into context dependent phoneme label sequence. According to the label sequence, a sentence HMM, which represents the whole text to be synthesized, is constructed by concatenating adapted phoneme HMMs. From the sentence HMM, phoneme durations are determined based on state duration distributions (Yoshimura *et al*., 1998). Then spectral and pitch parameter sequences are generated using the algorithm for speech parameter generation from HMMs with dynamic features (Tokuda *et al*., 1995). Finally by using MLSA filter (Fukuda *et al*., 1992), speech is synthesized from the generated mel-cepstral and pitch parameter sequences.

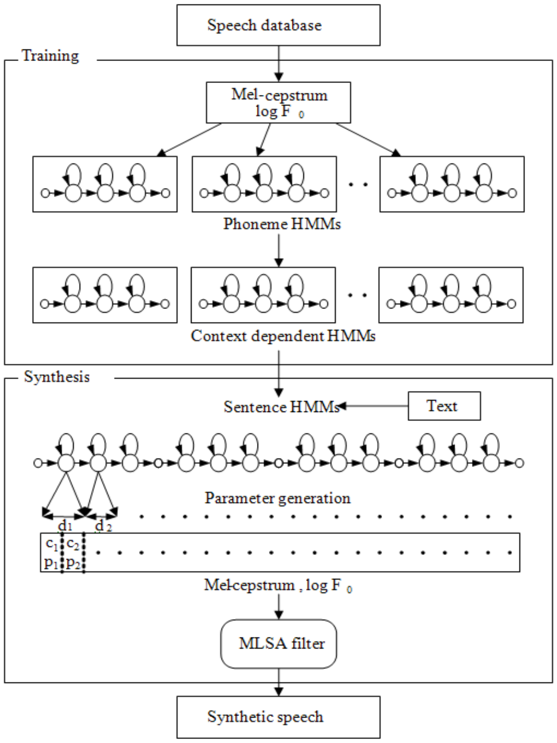

Fig 1: A block diagram of an HMM-based speech synthesis system

Thai Speech Attributes: As for tonal language, such as Thai, a syllable is composed of consonants, vowels and tone (Chompun, 2004). The basic Thai textual syllables can be represented in Fig. 2, where C_i, V, C_f and T denotes an initial consonant, a vowel, a final consonant and a tone respectively.

The significant difference between tonal and toneless language is the syllable tone, where meaning of a syllable changes as the syllable tone changes (Thathong *et al*., 2000; Chompun *et al*., 2001). Table 1 summarizes the number of the Thai characters and phones according to each part of syllables.

In Thai language, four different tone markers are generally used to indicate 5 Thai tones; middle tone (0), low tone (1), falling tone (2), high tone (3) and rising tone (4). For example the syllable “บาน” (to widen) has a middle tone which is pronounced as /ba:n/, meanwhile syllable “บ้าน” (home) has a falling tone which is pronounced as /bâ:n/. Each syllable tone can be characterized by its corresponding fundamental frequency contour which is depicted in Fig. 3 (Chompun, 2004; Chompun *et al*., 2001). Each contour line is constructed by plotting the voice fundamental frequency extracted periodically via the normalized syllable duration.

$$\mathrm{Ci(Ci)V^{T}(V)Cf}$$

Fig. 2: Thai syllable structure

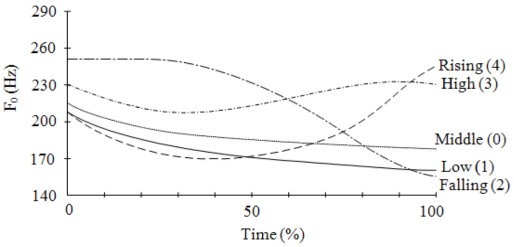

Fig. 3: Fundamental frequency contours of 5 tones in Thai

Table 1: The number of Thai characters and phones

Type	Character	Phone
Initial consonant (Ci)	44	38
Vowel (V)	16	24
Final consonant (Cf)	37	9
Tone (T)	4	5

Questions for tree-based context clustering process: Since tone information is a very crucial factor in Thai language as mentioned above, therefore tone number (0-4) is employed in the context clustering process of HMM-based TTS system. Moreover, the following contextual factors were also taken into account:

- Syllable position in word
- Part of speech
- Number of syllables in word
- Word position in phrase
- Number of syllables in phrase
- Phrase position in sentence
- Number of syllables in sentence
- Number of words in sentence
- Number of phones in syllable
- Phone position in syllable
- Phone type

RESULTS

Experimental conditions: Set of phonetically balanced sentences of Thai speech database from National Electronics and Computer Technology Center (NECTEC) is used for training HMMs. The whole sentence text was collected from Thai part-of-speech tagged corpus, named ORCHID (Hansakunbuntheung *et al*., 2005). The speech in the database is uttered by a professional female speaker with clear articulation and standard Thai accent. The text dependent phoneme labels are extracted based on the phoneme labels and linguistic information included in the database. There are almost 79 phonemes including silence and pause.

Speech signal were sampled at a rate of 16 kHz and windowed by a 25 m sec 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 zeroth coefficient, logarithm of fundamental frequency and their delta and delta-delta coefficients (Tachibana *et al*., 2005).

We used 5-state left-to-right HMMs in which the spectral part of the state is modeled by a single diagonal Gaussian output distribution (Tachibana *et al*., 2006; Yamagishi and Kobayashi, 2005). The number of training utterances is varied as 500, 1000, 1500, 2000 and 2500 sentences.

Subjective evaluations of synthesized speech: First, the naturalness of the synthesized speech generated from 6 approaches; 5 are from the HMM-based system set up by varying number of training utterances and another one is from the unit selection approach with the corpus size of 5200 sentences (Hansakunbuntheung *et al*., 2005), was evaluated by a paired comparison test. The subjects were nine Thai persons.

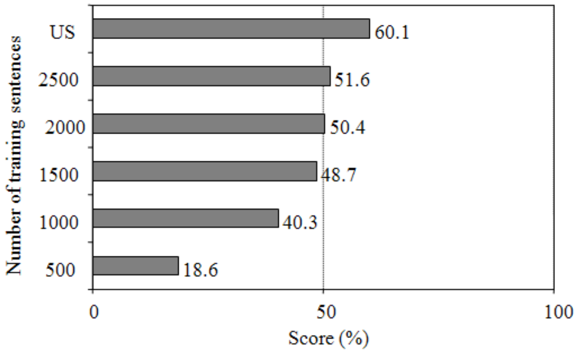

Fig. 4: Evaluation of naturalness of Thai HMM-based system and unit selection (US) approach

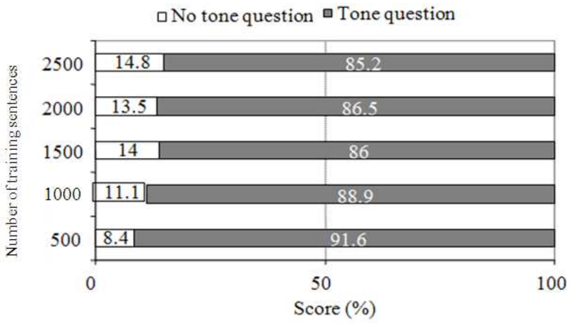

Fig. 5: Evaluation of correction of syllable tone of Thai HMM-based approach

They were presented a pair of speech synthesized from different approaches in random order and then asked which one sounded more natural. For each subject, five test sentences were chosen at random from 25 test sentences which were not contained in the training sentences. The preference scores are shown in Fig. 4.

Secondly, the subjective evaluation of tone question in the context clustering stage was conducted. The correction of the syllable tone of the synthesized speech generated from 2 systems was evaluated by a paired comparison test. The first system has tone question in the context clustering process, meanwhile another system has no tone question. The preference scores are shown in Fig. 5.

DISCUSSION

It can be seen from Fig. 4 that the more the number of training sentences is increased the more the naturalness of the synthesized speech is obtained. Although, the score of unit selection approach is above of HMM-based approach with 2500 training sentences, the HMM-based approach can be further developed to synthesize the speech with various voice characteristics and various speaking styles as mentioned earlier. It can be said that HMM-based approach is newly constructed for Thai language. Moreover, we expect that the system be further improved in the near future.

From Fig. 5, it can be seen that the score of the system with tone question is considerably superior than that of the system without tone question for every number of training sentences. When increasing the number of training sentences, the percentage score of no tone question case increases. The reason is that the lacking problem of training syllable tones is relieved.

CONCLUSION

In this study, we propose an HMM-based Thai speech synthesis. Thai speech characteristic is investigated and subsequently the conventional HMM-based synthesis system is modified according the tonal attributes of Thai. We found that the number of training sentences affected the naturalness of the synthesized speech while the tone information affected significantly with the output synthesized speech.

ACKNOWLEDGEMENT

The researcher is grateful to NECTEC for providing the TSynC-1 speech database.

REFERENCES

Al-Haddad, S.A.R., S.A. Samad, A. Hussain and K.A. Ishak, 2008. Isolated Malay digit recognition using pattern recognition fusion of dynamic time warping and hidden Markov models. Am. J. Applied Sci., 5: 714-720. DOI: 10.3844/ajassp.2008.714.720

Aliwa, M.B., T.E. El-Tobely, M.M. Fahmy, M.E.S. Nasr and M.H.A. El-Aziz, 2010. A new novel fidelity digital watermarking based on adaptively pixel-most-significant-bit-6 in spatial domain gray scale images and robust. Am. J. Applied Sci., 7: 987-1022. DOI: 10.3844/ajassp.2010.987.1022

Chomphan, S. and T. Kobayashi, 2008. Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis. Speech Commun., 50: 392-404. DOI: 10.1016/j.specom.2007.12.002

Chompun, S., 2004. Fine granularity scalability for MP-CELP based speech coding with HPDR technique. Proceeding of the IEEE Asia-Pacific Conference on Circuits and Systems, Dec. 6-9, IEEE Xplore Press, USA., pp: 197-200. DOI: 10.1109/APCCAS.2004.1412726

Chompun, S., S. Jitapunkul and D. Tancharoen, 2001. Novel technique for tonal language speech compression based on a bitrate scalable MP-CELP coder. Proceeding of the IEEE International Conference on Information Technology: Coding and Computing, IEEE Xplore Press, Apr. 2-4, Las Vegas, NV., USA., pp: 461-464. DOI: 10.1109/ITCC.2001.918839

Curran, K., X. Xi and R. Clarke, 2005. An investigation into the use of the least significant bit substitution technique in digital watermarking. Am. J. Applied Sci., 2: 648-654. DOI: 10.3844/ajassp.2005.648.654

Fukuda, T., K. Tokuda, T. Kobayashi and S. Imai, 1992, An Adaptive algorithm for mel-cepstral analysis of speech. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 23-26, IEEE Xplore Press, San Francisco, CA., USA., pp: 137-140. DOI: 10.1109/ICASSP.1992.225953

Hansakunbuntheung, C., A. Rugchatjaroen and C. Wutiwiwatchai, 2005. Space reduction of speech corpus based on quality perception for unit selection speech synthesis. Proceeding of the International Symposium on Natural Language Processing, Dec. 13-15, NECTEC, Chiang Rai, Thailand, pp: 127-132. http://www.hlt.nectec.or.th/publications.php

Masuko, T., K. Tokuda, T. Kobayashi and S. Imai, 1996. Speech synthesis using HMMS with dynamics features. Proceeding of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing, May 7-10, IEEE Xplore Press, Atlanta, GA., USA., pp: 389-392. DOI: 10.1109/ICASSP.1996.541114

Tachibana, M., J. Yamagishi, T. Masuko and T. Kobayashi, 2006. A style adaptation technique for speech synthesis using HSMM and suprasegmental features. IEICE Trans. Inform. Syst., E89-D: 1092-1099.

Tachibana, M., J. Yamagishi, T. Masuko and T. Kobayashi, 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Trans. Inform. Syst., E88-D: 2484-2491.

Tamura, M., T. Masuko, K. Tokuda and T. Kobayashi, 2001. Text-to-speech synthesis with arbitrary speaker’s voice from average voice. Proceeding of the 7th European Conference on Speech Communication and Technology, Sept. 3-7, IEEE, Aalborg, Denmark, pp: 345-348.

Thathong, U., S. Jitapunkul and V. Ahkuputra, 2000. Classification of Thai consonants naming using Thai tone. Proceeding of the International Conference on Spoken Language Processing, Oct. 16-20, ISCA, Beijing, China, pp: 47-50. http://www.isca-speech.org/archive/icslp_2000/i00_3047.html

Tokuda, K., T. Kobayashi and S. Imai, 1995. Speech parameter generation from HMM using dynamics features. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 9-12, IEEE Xplore Press, Detroit, MI., USA., pp: 660-663. DOI: 10.1109/ICASSP.1995.479684

Tokuda, K., T. Masuko, N. Miyazaki and T. Kobayashi, 1999. Proceedings of the 1999 IEEE International Conference on Acoustics, Speech and Signal Processing: Hidden Markov Models based on Multi-space Probability Distribution for Pitch Pattern Modeling, Mar. 15-19, IEEE Xplore Press, Phoenix, AZ., USA., pp: 229-232. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=758104

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, 2000. Speech parameter generation algorithm for HMM-based speech synthesis. Proceeding of the 2000 IEEE International Conference on Acoustics, Speech and Signal Processing, June 5-9, IEEE Xplore Press, Istanbul, Turkey, pp: 1315-1318. DOI: 10.1109/ICASSP.2000.861820

Yamagishi, J. and T. Kobayashi, 2005. Adaptive training for hidden semi-Markov model. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 23-23, IEEE Xplore Press, USA., pp: 365-368. DOI: 10.1109/ICASSP.2005.1415126

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proceeding of the 6th European Conference on Speech Communication and Technology, Sept. 5-9, IEEE, Budapest, Hungary, pp: 2347-2350.

Yamagishi, J., T. Masuko, K. Tokuda and T. Kobayashi, 2003. A Training method for average voice model based on shared decision tree context clustering and speaker adaptive training. Proceeding of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 6-10, IEEE Xplore Press, Hong Kong, China, pp: 716-719. DOI: 10.1109/ICASSP.2003.1198881

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1998. Duration modeling for HMM-based speech synthesis. Proceeding of the International Conference on Spoken Language Processing, Nov. 30-Dec. 4, IEEE, Sydney, Australia, pp: 29-32. http://www.shlrc.mq.edu.au/proceedings/icslp98/PDF/AUTHOR/SL980939.PDF

Young, S.J., J.J. Odell and P.C. Woodland, 1994. Tree-based state tying for high accuracy acoustic modeling. Proceeding of the ARPA Human Language Technology Workshop, Mar. 6-8, ACM Press, Princeton, New Jersey, pp: 307-312. DOI: 10.3115/1075812.1075885