A Speaker-Adaptive HMM-based Vietnamese Text-to-Speech System

Duy Khanh Ninh
Faculty of Information Technology
The University of Danang – University of Science and Technology
Danang, Vietnam
nkduy@dut.udn.vn

***Abstract*—This paper describes the first attempt in developing a Vietnamese HMM-based Text-to-Speech system using the speaker-adaptive approach. Although speaker-dependent systems have been built widely, no speaker-adaptive system has been developed for Vietnamese so far. We collected speech data from several Vietnamese native speakers and employed state-of-the-art speech analysis, model training and speaker adaptation techniques to develop the system. Besides, we performed perceptual experiments to compare the quality of speaker-adapted (SA) voices built on the average voice model and speaker-dependent (SD) voices built on SD models, and to confirm the effects of contextual features including word boundary (WB) and part-of-speech (POS) on the quality of synthetic speech. Evaluation results show that SA voices have significantly higher naturalness than SD voices when the same limited contextual feature set excluding WB and POS was used. In addition, SA voices trained with limited contextual features excluding WB and POS still have better quality than SD voices trained with full contextual features including WB and POS. These results show the robustness of the speaker-adaptive over the speaker-dependent approach for Vietnamese statistical parametric speech synthesis.**

Keywords—HMM-based speech synthesis, speaker-adaptive approach, average voice model, contextual features, Vietnamese

I. INTRODUCTION

Speech synthesis (or Text-to-Speech) based on statistical models has become the dominant direction in speech synthesis research in the last decade thanks to its high degree of flexibility and natural expressiveness [1]. In this speech synthesis technology, the synthetic speech generated from statistically trained models possesses similar characteristics and speaking style of the recorded voice although a moderate amount of speech data is available. This synthesis framework only requires little language-dependent information, thus has been leveraged to develop speech synthesizers for different languages. Since the statistical models are traditionally hidden Markov models (HMMs), this methodology is often called as HMM-based speech synthesis.

Several HMM-based Text-to-Speech (TTS) systems for Vietnamese have been developed since 2009 [2][3]. Some improvements have been applied to these systems, which are either the incorporation of syntactic and prosodic information to enhance the naturalness of the prosody of generated speech [4][5] or the accurate extraction of pitch contours for glottalized tones to improve the tonal analysis and synthesis [6]. Although the obtained results are promising, all of the above systems were built using the speaker-dependent approach with a moderate amount of training data of one speaker. This traditional approach makes the performance of these systems largely dependent on the voice quality of the selected speaker, and more importantly, has no flexibility in changing the voice characteristics of the systems. Although different voices can be generated by combining a voice conversion method with HMM-based speech synthesis [7], it is not easy to convert prosodic features (e.g., F0 and duration) from one voice to another since these features cover longer time span and are more contextually dependent than spectral ones.

The HMM-based speech synthesis systems using speaker-adaptive approach [8] have proved their robustness in the ability of transforming voice characteristics from the average voice of multiple speakers to the target voice of any other speaker using speaker adaption methods. In this HMM synthesis approach, an average voice model is first trained using speech data from several speakers, then is adapted using speech data of a target speaker. This speech synthesis method can adapt speech parameters (i.e., spectral, excitation, and duration) within a framework based on multi-space distribution hidden semi-Markov models (MSD-HSMMs) [9], an extended version of HMMs for better modeling of F0 and duration parameters of speech. Advanced adaptation algorithms have shown their effectiveness in HMM-based speech synthesis [10]. For tonal languages, an attempt has been made for Thai language and the speaker-adaptive approach was reported to increase the capability of tone reproduction [11].

Although speaker-dependent HMM-based TTS systems have been built widely, no average voice based system has been developed for Vietnamese so far. This paper presents the first attempt in developing and evaluating an HMM-based Vietnamese TTS using the speaker-adaptive approach. We have collected speech data from several Vietnamese native speakers and employed state-of-the-art speech analysis, model training and adaptation techniques to develop the system. Besides, we performed perceptual experiments to compare the quality of speaker-adapted voices built on the average voice model and speaker-dependent voices built on speaker-dependent models, and to confirm the effects of word boundary and part-of-speech information on the quality of synthetic speech. These information at word level can be respectively obtained using a word segmenter and a part-of-speech tagger often integrated in a full-featured natural language processing module, which is not always available in the development of a TTS system. For the synthesis of a multi-syllabic language like Vietnamese, the effect of grouping of multiple syllables into one compound word using word boundary information and the effect of part-of-speech tags of the words have not been separately investigated yet, although the combination of part-of-speech tags, pitch accent and phrase-final intonation were reported to noticeably increase the quality of synthesized speech [5]. Thus it is worth to consider carefully their effects on speech quality.

The remaining parts of the paper is organized as follows. Section II reviews speaker-dependent and speaker-adaptive

978-1-7281-3003-3/19/$31.00 ©2019 IEEE

approaches of HMM-based speech synthesis. The development of Vietnamese speech synthesis system using the average voice model is presented in Section III, and the perceptual evaluation results are shown in Section IV. Several conclusions are given in Section V of the paper.

II. SPEAKER-DEPENDENT AND SPEAKER-ADAPTIVE APPROACHES OF HMM-BASED TTS

A. Speaker-Dependent Approach

Fig. 1 shows a convntional speaker-dependent HMM-based TTS system [1]. The statistical models, HMMs, work as the learner in the training part, and as the generator in the synthesis part.

In the training part, spectral parameters (usually, mel-cepstral coefficients) and excitation parameters (usually, logF0) are extracted from speech data of one speaker. These parameter streams are then separately modeled by phoneme HMMs using the maximum likelihood criterion. Since each phoneme exhibits its acoustic realizations differently according to its phonetic and linguistic contexts, these contextual features are embedded into the label for each phoneme, resulting the so-called contextual label. Lists of the contextual features utilized in our Vietnamese systems are given in Sections III.B and IV.A. However, a contextual label only appears very few times (usually only once) in the speech corpus. To overcome this data sparseness problem, a clustering technique based on decision tree is used to cluster acoustically similar instances of a phoneme to construct a context-dependent HMM. This collection of HMMs captures voice characteristics of the training speaker, thus called the speaker-dependent model.

Although sequences of spectral parameters can be modeled by continuous HMMs, those of excitation parameters should be modeled by multi-space probability distribution (MSD) based HMMs [1]. For explicitly modeling the duration of HMM states, hidden semi-Markov models (HSMMs) have been applied to speech synthesis and proved its effectiveness [9]. MSD-HSMM has become the standard model in HMM-based speech synthesis and, thus, the term "MSD-HSMM" is used interchangeably with the term "HMM" in this paper.

In the synthesis part, the text of an input sentence is analyzed and decomposed into a sequence of contextual labels, from which a sequence of corresponding context-dependent HMMs is formulated. Then, these HMMs are used to generate speech parameters with maximal output probabilities. In particular, the HMM state durations, and the sequences of mel-cepstral coefficients and logF0 values are determined, in this order, so that their probabilities are maximized. Finally, a speech synthesis filter is driven by the generated speech parameter vector sequences to synthesize a speech signal.

B. Speaker-Adaptive Approach

Fig. 2 illustrates an HMM-based TTS system using the average voice (or speaker-adaptive) approach [8]. While the synthesis part is similar to that of the speaker-dependent approach, the training and adaptation parts with the aim to build the speaker-adapted model for synthesis are the different points between the two approaches.

In the training part, speech data of multiple speakers is used and the extracted speech parameters are modeled by context-dependent MSD-HSMMs. This collection of MSD-HSMMs captures common voice characteristics among the training speakers, thus called the average voice model. Details of the training techniques for average voice model building is given in Section III.D.

In the adaptation part, the speech data of any other target speaker is utilized for transforming the average voice model toward the voice characteristics of this speaker. Several speaker adaptation methods have been proposed and their effectiveness have been proved [10]. These methods employ linear transformations to transform HMM parameters of the average voice model into those of the speaker-adapted model. Details of the adaptation technique to construct the speaker-adapted model is described in Section III.E.

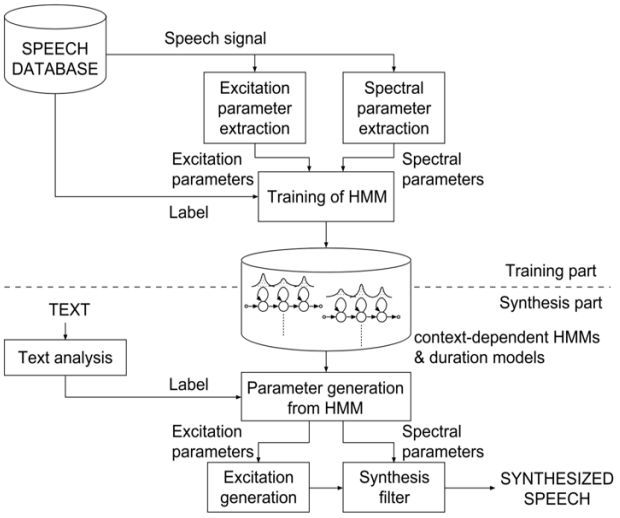

Fig. 1. A typical speaker-dependent HMM-based TTS system [1].

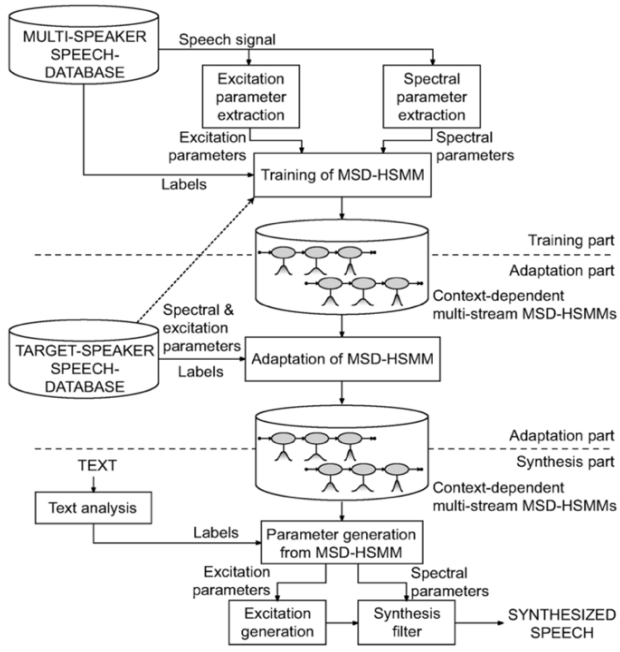

Fig. 2. A typical speaker-adaptive HMM-based TTS system [10].

III. DEVELOPMENT OF A SPEAKER-ADAPTIVE HMM-BASED TTS FOR VIETNAMESE

In this section, we describe the development of our average voice based Vietnamese speech synthesis system.

A. Building Speech Corpus

We collected speech data from eight Vietnamese native speakers, including four males and four females, with standard Northern voices to conduct the experiments. All speakers read the same 1100 phonetically balanced Vietnamese sentences with the text collected from the internet. Speech signals were recorded in a professional studio using a high quality microphone with a sample rate of 44.1 kHz. On average, each speaker produced about one hour of speech signals. Among them, the same 1000 sentences for each of six speakers (three males and three females), were used for average voice model training. For each of the other speakers (one male and one female, referred to as target speakers), 1000 sentences were used as adaptation data and the remaining 100 sentences were used as test data for building and evaluating, respectively, the speaker-adapted voices.

B. Assigning Contextual Labels

The speech signals were automatically labelled at phonetic level by using a self-developed phone aligner. These monophone labels were then extended to include Vietnamese phonetic and linguistic contexts such as phoneme-level features, syllable-level features, breathe-group-level features, and sentence-level features. The breathe groups in an sentence are separated each other by either pauses in the speech signal or punctuations in the sentence's transcription. Due to the lack of a complex natural language processing module, the contextual labels do not cover word-level features such as part-of-speech and prosodic features such as prosodic phrasing, ToBI (Tones and Break Indices), and phrase-final intonation like other Vietnamese systems [4][5]. Instead, the phonological features of a phoneme (e.g., voiced/unvoiced, long/short vowel, fricative/plosive/labial consonant, etc) and the vowel identity of a syllable were added to the contextual labels as they have some effect on segmental characteristics, particularly on speech spectrum.

Details of the Vietnamese contextual features used in the average voice based system are as follows:

1) Phoneme level:

- Two preceding, current, two succeeding phonemes
- Position in current syllable (forward, backward)
- Phonological features of current phoneme

2) Syllable level:

- Tone types of two preceding, current, two succeeding syllables
- Number of phonemes in {preceding, current, succeeding} syllables
- Position in current breath group
- Name of the vowel of current syllable

3) Breathe-group level:

- Number of syllables in {preceding, current, succeeding} breathe group
- Position of current breathe group in sentence

4) Sentence level:

- Number of {syllables, breathe groups} in sentence

C. Extracting Speech Parameters

The recorded speech signals were down-sampled to 22.05 kHz. We used Hamming windows with the length of 25 ms and the window shift of 5 ms for speech analysis. In Vietnamese speech synthesis, pitch (or F0) is an important parameter since it represents not only utterance's intonation but also syllabic tones. For glottalized tones such as Broken and Drop tones ("Thanh ngã" and "Thanh nặng" in Vietnamese) and for some creaky voices, particularly those of Northern Vietnamese speakers, it is difficult to extract complete and accurate F0 contours from speech signal due to large variations of the signal's degree of periodicity. Thus the F0 extraction method proposed in [6] was employed in our system to alleviate this problem. Besides, we used the high-quality speech vocoding method STRAIGHT to extract spectral and aperiodicity measurements from speech signals as described in the Nitech-HTS 2005 system [12]. The acoustic feature vectors for model training include static features (40 mel-cepstrum coefficients, logF0 and 5 aperiodicity components) and their first and second order dynamic features.

D. Training Average Voice Model

We utilized 5-state MSD-HSMMs for modeling the acoustic features and phoneme duration. Gaussian probability density function (pdf) was employed as the state output and state duration distributions. We used 6000 sentences of six training speakers for average voice model training. First, monophone MSD-HSMMs were trained and turned into context-dependent models based on contextual labels, and the model parameters were re-estimated. After that, these context-dependent models were clustered by a context clustering technique in which decision trees were shared among training speakers [8]. We finally re-trained the clustered MSD-HSMMs using the speaker adaptive training (SAT) technique [8], which further maximizes the likelihood of the training data of multiple speakers. The resulting average voice model consists of robustly trained Gaussian pdfs at leaf nodes of the trees.

E. Building Speaker-Adapted Models

The average voice model was then modified towards the two target speakers based on their adaptation data. Among the adaption methods investigated in [10], the structural maximum a posteriori linear regression (SMAPLR) adaptation method was chosen to alleviate the processing load when the size of adaptation data is relatively large. The shared decision trees formed during the average voice model training phase was utilized for estimating multiple transformation matrice during the speaker adaptation. The transformed Gaussian pdfs form the speaker-adapted model for each target speaker.

F. Synthesizing Speech Waveform

The sentence's text is analyzed and a sequence of context-dependent MSD-HSMMs is formed. Then, the state durations of these models were determined and the acoustic parameter trajectories were generated based on the maximal output probability (MOP) criterion [1]. Finally, a mixed excitation signal was generated, then drove a synthesis filter to synthesize the speech signal in a way similar to the system described in [12] to ensure the high quality of synthesized speech while maintaining a low synthesis time.

G. Evaluating Objectively Speaker-Adapted Voices

A simple text analyzer was implemented to extract the contextual labels described in Section III.B from the input text. The remaining modules described from Section III.C to Section III.F were carried out by our scripts, which were modified from the HTS toolkit, from which two speaker-adapted voices were built. Fig. 3 and Fig. 4 show the waveforms and extracted spectrograms and F0 contours of natural and synthesized speech signals, respectively, of an utterance in the test data for the target male speaker. Fig. 5 and Fig. 6 show the same information for the target female speaker. It can be observed that the spectral and F0 features of the synthesized speech is quite similar to those of the natural speech for both of the two speakers.

IV. PERCEPTUAL EXPERIMENTS

We conducted several perceptual experiments to compare the quality of speaker-adapted (SA) voices built using the speaker-adaptive approach and speaker-dependent (SD) voices built using the speaker-dependent approach. Besides, we would like to confirm the effects of word boundary (WB) and part-of-speech (POS) on the quality of SD voices when they are added into contextual labels.

A. Experimental Conditions

In parallel with building two SA voices of the two target speakers (one male and one female) as presented in the preceding section, we built six SD voices of themselves (three voices for the male, three voices for the female) using the SD approach. The adaptation data of 1000 sentences of each target speaker was used as the training data to train SD models, and the remaining 100 sentences was used for testing. These models and the resulting voices, trained with different techniques and contextual features, are summarized in Table I.

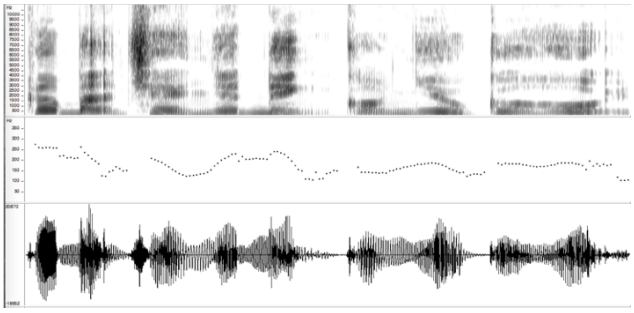

Fig. 3. Spectrogram, F0 contour, and waveform of the male's natural speech for the utterance "Các bạn trẻ nhất định có nhiều cơ hội" (in English "Young people surely have many opportunities").

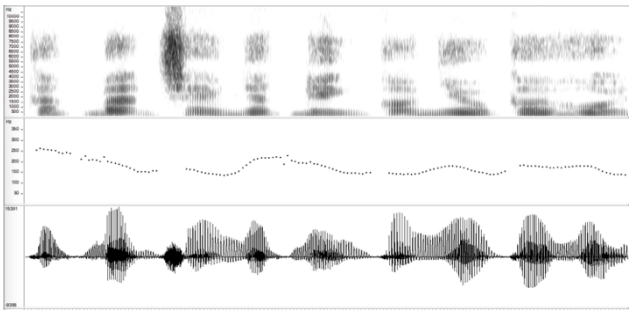

Fig. 4. Spectrogram, F0 contour, and waveform of the male's synthesized speech for the utterance "Các bạn trẻ nhất định có nhiều cơ hội" (in English "Young people surely have many opportunities").

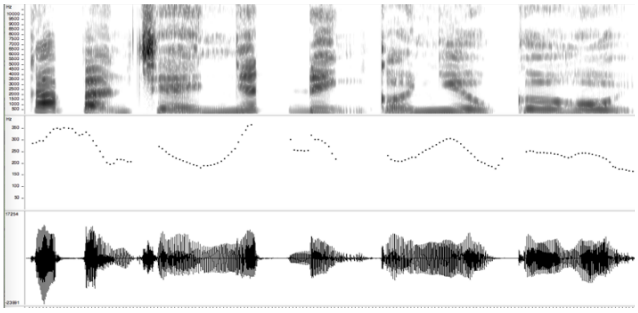

Fig. 5. Spectrogram, F0 contour, and waveform of the female's natural speech for the utterance "Các bạn trẻ nhất định có nhiều cơ hội" (in English "Young people surely have many opportunities").

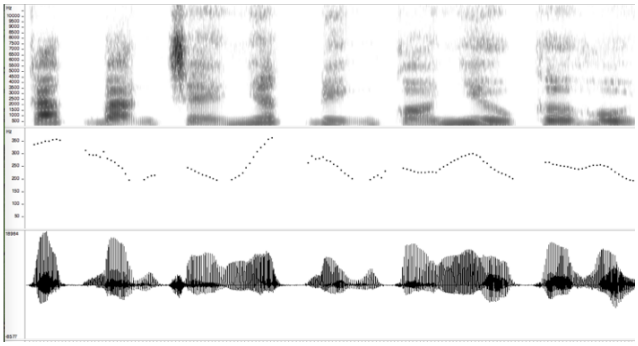

Fig. 6. Spectrogram, F0 contour, and waveform of the female's synthesized speech for the utterance "Các bạn trẻ nhất định có nhiều cơ hội" (in English "Young people surely have many opportunities").

TABLE I. SUMMARY OF MODELS TRAINED WITH DIFFERENT TECHNIQUES AND CONTEXTUAL FEATURES

Model/Voice	Training technique	Contextual features
SA	SAT + SMAPLR adaptation	Limited set
SD	SD training	Limited set
SD_WB	SD training	Limited set + WB features
SD_WB_POS	SD training	Limited set + WB and POS features

In Table I, the limited set consists of contextual features listed in Section III.B (i.e., WB and POS are excluded). The WB features are those can be added to contextual labels once word boundaries were determined, including:

- Position of syllable in current word (forward, backward)
- Number of syllables in {preceding, current, succeeding} words
- Position of word in current breathe group
- Number of words in {preceding, current, succeeding} breathe group
- Number of words in the sentence

Meanwhile, the POS features added to contextual labels include: POS of {preceding, current, succeeding} words.

To extract the WB and POS information, we employed the word segmenter and the POS tagger integrated in JVnTextPro [13], respectively.

B. Experimental Results

We performed MOS (Mean Opinion Score) tests to evaluate the synthetic voices of the two target speakers. Their

natural voices were also rated for reference. Nine native subjects from Northern Vietnam were told to listen to 20 utterances randomly selected from the test set, each of which was either natural speech or synthesized using four configurations in Table I. The speech signals were introduced to the listener in random order.

Fig. 7 shows the MOS score averaged by all the test subjects. It can be seen that the natural voices were rated from good to excellent, while the synthetic voices were rated from fair to good. Among the synthetic voices, while the SA voices achieve an average score ranging from 4.0 to 4.5 points on the MOS scale, the SD voices range between 3.5 and 4.0 points. The evaluation results show that SA voices have significantly higher naturalness than SD voices (about 0.5 points on MOS scale) when being trained with the same limited contextual feature set excluding WB and POS. In addition, SA voices trained with limited contextual features excluding WB and POS still have better quality than SD voices trained with full contextual features including WB and POS (from 0.3 to 0.4 points on MOS scale). Considering SD voices only, the introduction of both WB and POS related features into contextual labels helps improve slightly the naturalness of synthetic speech, from 0.1 to 0.2 points on MOS scale. These results are consistent among two target speakers. Sources of quality difference among the voices reported by the listening subjects are mostly on two aspects: the naturalness of pitch and the reduction of sound artifacts. However, the introduction of only WB related contextual features exhibits effectiveness on the male SD voice (MOS score increases from 3.82 to 3.92), while has negative effect on the female SD voice (MOS score decreases from 3.63 to 3.47). It is due to the fact that the female speaker has a fast speaking rate while the male voice has normal rate. Thus the female voice is adversely affected by the syllable-grouping effect introduced by WB features.

V. CONCLUSION

This paper presents the first attempt in developing and evaluating an HMM-based Vietnamese speech synthesis system using the speaker-adaptive approach. Details of the system development process from speech data collection to speech synthesis have been described. Built upon a large collection of speakers and speech data, our average-voice based system achieves an average score higher than 4.0 points on the MOS scale. Besides, the effects of WB and POS related contextual features on the quality of speech synthesized from HMMs have been investigated. Perceptual experiments show the robustness of the speaker-adaptive over the speaker-dependent approach. Specifically, SA voices built from the average voice model possess remarkably higher naturalness than SD voices (around 0.5 points on MOS scale) when being trained the same limited contextual feature set excluding WB and POS. In addition, SA voices trained with limited contextual features excluding WB and POS still gain better quality than SD voices trained with full contextual features including WB and POS. These results suggest that the use of the average voice model can compensate for the lack of WB and POS information given by a full-featured natural language processing module in building a synthetic voice.

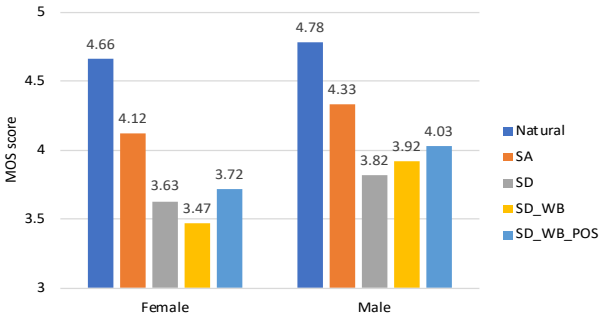

Fig. 7. Average MOS score of different voices of two target speakers.

ACKNOWLEDGMENT

This research is funded by Funds for Science and Technology Development of the University of Danang under grant number B2016-DNA-38-TT and the University of Danang - University of Science and Technology under grant number T2017-02-93. The author thanks the subjects whose participation in speech recording and speech quality evaluation made this study possible.

REFERENCES

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi and K. Oura, "Speech synthesis based on Hidden Markov Models," Proceedings of the IEEE, vol. 101, no. 5, pp. 1234–1252, May 2013.

[2] T. T. Vu, M. C. Luong, and S. Nakamura, "An HMM-based Vietnamese speech synthesis system," Proc. Oriental COCOSDA, Urumqi, China, pp. 116–121, Aug. 2009.

[3] T. T. T. Nguyen, C. Alessandro, A. Rilliard, and D. D. Tran, "HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation," Proc. INTERSPEECH, Lyon, France, pp. 2311–2315, Aug. 2013.

[4] T. T. T. Nguyen, A. Rilliard, D. D. Tran, and C. Alessandro, "Prosodic phrasing modeling for Vietnamese TTS using syntactic information", Proc. INTERSPEECH, Singapore, pp. 2332–2336, Sept. 2014.

[5] T. S. Phan, T. C. Duong, A. T. Dinh, T. T. Vu, C. M. Luong, "Improvement of naturalness for an HMM-based Vietnamese speech synthesis using the prosodic information", Proc. IEEE RIVF, Vietnam, pp. 276–281, 2013.

[6] D. K. Ninh and Y. Yamashita, "F0 parameterization of glottalized tones in HMM-based speech synthesis for Hanoi Vietnamese", IEICE Transactions on Information and Systems, vol.E98-D, no.12, pp. 2280–2289, 2015.

[7] T.-N. Phung, "HMM-based speech synthesis with multiple individual voices using exemplar-based voice conversion", International Journal of Computer Science and Network Security, vol.17, no.5, pp. 192–196, May 2017.

[8] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Transactions on Information and Systems, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[9] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," IEICE Transactions on Information and Systems, vol. E90-D, no. 5, pp. 825–834, May 2007.

[10] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," IEEE Transactions on Audio, Speech and Language Processing, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[11] S. Chomphan, T. Kobayashi, "Tone correctness improvement in speaker-independent average-voice-based Thai speech synthesis", Speech Communication, vol. 51, no. 4, pp. 330–343, 2009.

[12] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," IEICE Transactions on Information and Systems, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[13] C.-T. Nguyen, X.-H. Phan, and T.-T. Nguyen, "JVnTextPro: A Java-based Vietnamese Text Processing Tool," http://jvntextpro.sourceforge.net/, 2010.