EVALUATION OF SPEAKER-DEPENDENT AND AVERAGE-VOICE VIETNAMESE STATISTICAL SPEECH SYNTHESIS SYSTEMS

Duy Khanh Ninh

The University of Danang, University of Science and Technology; nkduy@dut.udn.vn

Abstract - This paper describes the development and evaluation of a Vietnamese statistical speech synthesis system using the average voice approach. Although speaker-dependent systems have been applied extensively, no average voice based system has been developed for Vietnamese so far. We have 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 have 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 is 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 average voice method over the speaker-dependent approach for Vietnamese statistical speech synthesis.

Key words - Vietnamese statistical speech synthesis; hidden Markov model; average voice approach; speaker-dependent approach; contextual features

1. Introduction

Speech synthesis (or Text-to-Speech) based on statistical models has recently drawn much attention in speech synthesis technology owing to its high flexibility and natural-sounding quality [1]. In this speech synthesis method, the statistically trained models can acquire the voice characteristics and speaking style from a moderate amount of speech data of a recorded voice, then generate speech with high intelligibility and smoothness. The whole training and synthesis framework makes use of little language-dependent information. Consequently, it has been applied to building numerous monolingual and multilingual speech synthesizers. Since hidden Markov model (HMM) has been widely used as the statistical model, this speech synthesis method is often called as *HMM-based speech synthesis*.

A couple of HMM-based Text-to-Speech (TTS) systems for Vietnamese have been developed since 2009 [2], [3]. Latest refinements being made to these systems involved in the integration of syntactic information and intonational tags to improve the overall naturalness of generated prosody [4], [5] or the accurate extraction of pitch contours for glottalized tones to enhance the tonal analysis and synthesis [6]. Although the obtained results are promising, all of the above systems are 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 we can synthesize multiple voices by using some voice conversion method combined with HMM-based TTS [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 average voice approach to HMM-based speech synthesis [8], [9] has proved its 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 data from several speakers, then is adapted using a small amount of speech from a target speaker. This speech synthesis method can adapt spectral, excitation, and duration parameters within the same framework of multi-space distribution hidden semi-Markov models (MSD-HSMMs) [10], an extended version of HMMs for better modeling of F0 and duration parameters of speech. Advanced adaptation algorithms have been proposed and their effectiveness in HMM-based speech synthesis has been demonstrated [11], [12]. For tonal languages, an attempt has been made for Thai HMM-based synthesis and the average voice approach was reported to improve the tone correctness of synthesized speech [13].

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 average voice 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 significantly improve the quality of synthesized speech [5]. Thus it is worth carefully considering their effects on speech quality.

The paper is organized as follows. Section 2 reviews speaker-dependent and average voice approaches of HMM-based speech synthesis. The development of Vietnamese speech synthesis system using the average voice method is described in Section 3, and the results of perceptual evaluations are reported in Section 4. Section 5 concludes the paper.

2. Speaker-dependent and average voice approaches of HMM-based speech synthesis

2.1. Speaker-dependent approach

Figure 1 is an overview of a typical speaker-dependent HMM-based speech synthesis system [1]. This system defines a speech synthesis problem in a generative model framework and solves it based on the maximum likelihood criterion. It consists of training and synthesis parts.

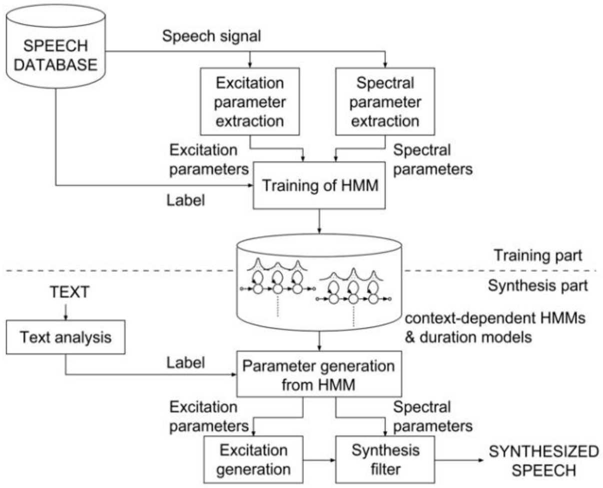


Figure 1. *Speaker-dependent HMM-based speech synthesis*

In the training part, spectrum (i.e., mel-cepstral coefficients and their dynamic features) and excitation (i.e., logF0 and its dynamic features) parameters are extracted from a speech database of one speaker and modeled by phoneme HMMs. 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 in the so-called contextual label. Lists of the contextualfeatures utilized in our Vietnamese systems are given in Sections 3.2 and 4.1. However, a contextual label only appears very few times (usually only once) in the speech corpus. To overcome this data sparseness problem, a decision-tree-based clustering technique 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 mel-cepstral coefficients can be modeled by continuous HMMs, sequences of logF0 cannot be modeled using continuous or discrete HMMs without heuristic assumptions since each logF0 observation can be viewed as consisting of a one-dimensional continuous logF0 value for a voiced frame or a discrete symbol for an unvoiced frame. To model this kind of observation, HMMs based on multi-space probability distributions (MSD) have been proposed [1]. For explicitly modeling the duration of HMM states, hidden semi-Markov models (HSMMs)have been applied to speech synthesis and proved its effectiveness [10]. 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 a sentence to be synthesized is first converted to a contextual-label sequence and then, a sentence HMM is constructed by concatenating the context-dependent HMMs based on the label sequence. Second, the HMM state durations are determined so as to maximize their probabilities. Third, the speech parameter generation algorithm generates the sequences of mel-cepstral coefficients and logF0 values that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated speech parameter vectors by a speech synthesis filter.

2.2. Average voice approach

Figure 2 shows the diagram of an HMM-based speech synthesis systemusing the average voice (or speaker-adaptive) approach [9], [12]. The system consists of three parts: training, adaptation, and synthesis. 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.

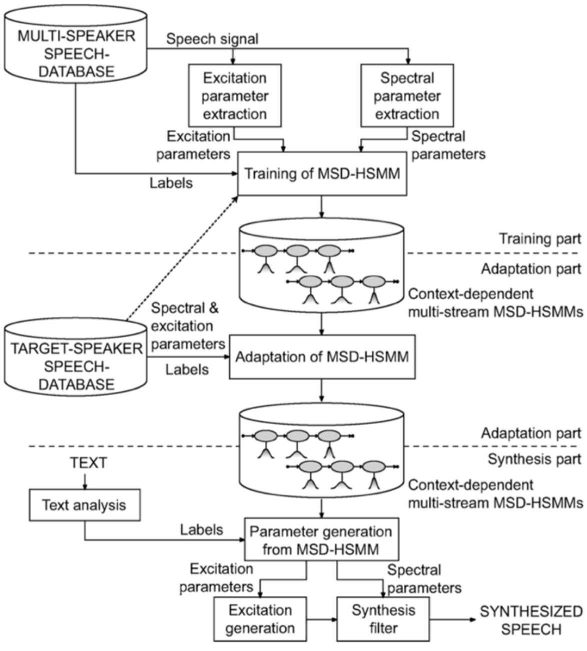


Figure 2. *Average-voice HMM-based speech synthesis*

In the training part, a speech database of multiple speakers is used and 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 to build the average voice model is given in Section 3.4.

In the adaptation part, the speech data of any other target speaker is utilized for adapting the average voice model toward the voice characteristics of this speaker. Several speaker adaptation methods have been proposed and their effectiveness have been proved [11]. 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 modelis described in Section 3.5.

3. Developing a Vietnamese speech synthesis system using average voice approach

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

3.1. 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 sound proof room using a high quality microphone at a sampling 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 training the average voice model. 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 voice based on the average voice model.

3.2. Assigning contextual labels

The speech signals were automatically labelled at phonetic level by using a self-developed phone aligner. These mono-phone labels were then extended to include Vietnamese phonetic and linguistic contexts such as phoneme-level features, syllable-level features, breathe-group-level features, and utterance-level features. The breathe groups in an utterance are separated from each other by either pauses in the speech signal or punctuations in the utterance'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, To BI (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 contextualfeatures used in theaverage 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 utterance.

4) Utterance level

- Number of {syllables, breathe groups} in utterance.

3.3. Extracting speech parameters

The recorded speech signals were down-sampled to 22.05 kHz and windowed by a 25-ms Hamming window with a 5-ms shift. 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 [14]. In particular, each signal frame was parameterized into 39-th order STRAIGHT mel-cepstrum and aperiodicity measures for five frequency sub-bands. In addition to these static features (40 STRAIGHT mel-cepstrum coefficients including the zeroth coefficient, logF0 and 5 aperiodicity measures), their delta and delta-delta features were also used to form acoustic feature vectors for training the average voice model.

3.4. Training average voice model

As in state-of-the-art HMM-based synthesis systems, we utilized 5-state context-dependent multi-stream left-to-right MSD-HSMMs without skip paths [10] to simultaneously model the above acoustic features and phoneme duration. Each state had a single Gaussian probability density function (pdf) with a diagonal covariance matrix as the state output pdf and had a single Gaussian pdf with a scalar variance as the state duration pdf. The average voice model was trained using 6000 sentences of six training speakers. We first trained speaker-independent mono-phone MSD-HSMMs using mono-phone labels. These were converted into context-dependent MSD-HSMMs, and the model parameters were re-estimated again. Then, shared-decision-tree-based context clustering [8] using the minimum description length criterion was applied to the MSD-HSMMs, and the model parameters of the MSD-HSMMs at each leaf node of the decision trees were tied. Note that a decision tree was constructed independently for each combination of state index and acoustic parameter (mel-cepstrum, logF0, aperiodicity) or duration. We then re-estimated the clustered MSD-HSMMs using the speaker adaptive training (SAT) technique described in [9] to normalize the influence of speaker differences among the training speakers. The estimated Gaussian pdfs at leaf nodes of the shared decision trees form the *average voice model*.

3.5. Building speaker-adapted models

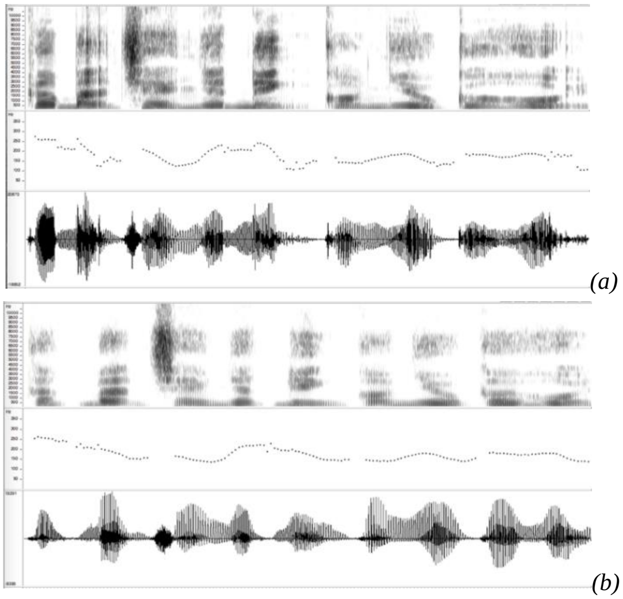
We then adapted the average voice model to the two target speakers based on their adaptation data. In order to reduce the computational time needed for the large amount of adaptation data, we used the structural maximum a posterior linear regression (SMAPLR) adaptation method proposed in [15], [11] for adapting the average voice model. In the speaker adaptation and speaker adaptive training, the estimation of multiple transforms was based on the shared decision trees constructed in the training stage of the average voice model. The tuning parameters for the adaptation algorithm and the thresholds to control the number of transforms were manually adjusted. The transformation matrices for the output Gaussian pdfsin the SMAPLR algorithm were diagonal triblocks corresponding to the static, delta, and delta-delta coefficients. The transformed Gaussian pdfs form the *speaker-adapted model* for each target speaker.

3.6. Synthesizing speech waveform

The text of a sentence to be synthesized was first converted to a contextual label sequence, and then the sentence MSD-HSMM was constructed based on the label sequence. Second, the state durations were determined so as to maximize their probabilities based on the state duration pdfs. Third, the speech parameter generation algorithm considering global variance (GV) [16] was used to generate each component of the acoustic feature vector from a sequence of Gaussian pdfs of the context-dependent HSMMs. Note that the GV pdf of each component of the acoustic feature vector was directly estimated from the adaptation data of each target speaker beforehand. Finally, an excitation signal was generated using mixed excitation (pulse plus a noise component band-filtered according to the five aperiodicity parameters) and pitch-synchronous overlap add [17]. This signal was used to excite a mel-logarithmic spectrum approximation (MLSA) filter corresponding to the STRAIGHT mel-cepstral coefficients and thus generate the speech waveform. These vocoding components are similar to those of the Nitech-HTS 2005 system [14], which reduce computational cost at synthesis time while maintaining the high quality of synthesized speech.

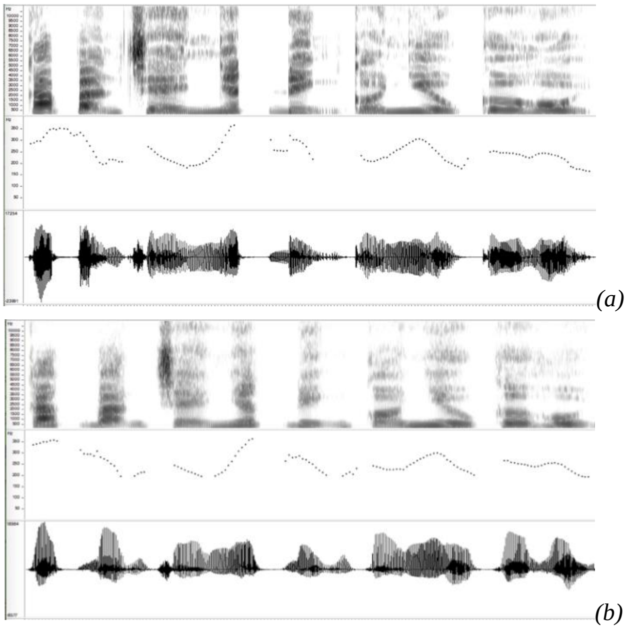
3.7. Evaluating speaker-adapted voices

A simple text analyzer was implemented to extract the contextual labels described in Section 3.2 from the input text. The remaining modules described from Section 3.3 to Section 3.6 were carried out by a modified version of the HTS toolkit version 2.3 [18], from which two speaker-adapted voices were built. Figure 3 and Figure 4 show the waveforms and corresponding spectrograms and F0 contours of natural and synthesized speech of a sentence in the test data of the target male and female speakers, respectively. It can be observed that the spectral and F0 features of the synthesized speech is quite similar to those of the natural speech.


(a)

(b)

Figure 3. *Spectrogram, F0 contour, and waveform of natural speech (a) and synthesized speech (b) of the sentence "Các bạn trẻ nhất định có nhiều cơ hội" of the male speaker*


(a)

(b)

Figure 4. *Spectrogram, F0 contour, and waveform of natural speech (a) and synthesized speech (b) of the sentence "Các bạn trẻ nhất định có nhiều cơ hội" of the female speaker*

4. Perceptual experiments

We carried out several perceptual experiments to compare the quality of speaker-adapted (SA) voices built using the average voice approach and speaker-dependent (SD) voices built using the SD 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.

4.1. 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 Table1.

***Table 1.** 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 1, the limited set consists of contextual features listed in Section 3.2 (i.e., WB and POS are excluded). The WB features are those that 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} breathegroup;
- Number of words in the utterance.

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 JVn Text Pro [19], respectively.

4.2. Experimental results

We conducted 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 were asked to listen to 20 sentences randomly selected from the test set, each of which was either natural speech or synthesized using four configurations in Table 1. The speech signals were played in random order. The 5-point MOS scale includes: bad (1), poor (2), fair (3), good (4), and excellent (5). This is a standard test widely used in speech synthesis research [20].

Figure5 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 slightly improve 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.

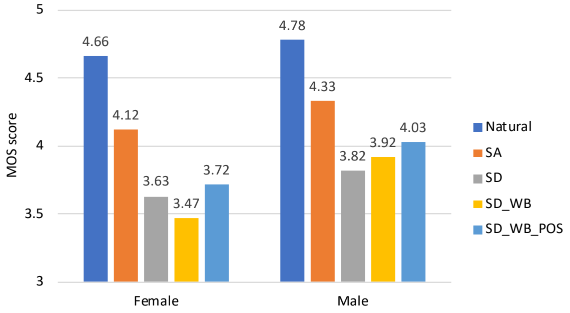


***Figure 5.** MOS score of different voices of two speakers*

5. Conclusion

This paper presents the first attempt in developing and evaluating an HMM-based Vietnamese statistical speech synthesis system using the average voice 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 average voice method 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.

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.

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 VietnameseSpeech 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, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “A training method of average voice model for HMM-based speech synthesis”, *IEICE Transactions on Fundamentals*, vol. E86-A, no.8, pp.1956–1963, Aug. 2003.

[9] 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.

[10] 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.

[11] 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.

[12] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, “Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[13] 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.

[14] 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.

[15] O. Shiohan, T. Myrvoll, and C. Lee, “Structural maximum a posteriori linear regression for fast HMM adaptation”, *Computer Speech and Language,* vol. 16, no. 3, pp. 5–24, 2002.

[16] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis”, *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[17] E. Moulines and F. Charpentier, “Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones”, *Speech Communication*, vol.9, pp.453–467, 1990.

[18] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, “The HMM-based speech synthesis system version 2.0”, *Proc. 6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, pp.294–299, Aug. 2007.

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

[20] T. Dutoit, “An Introduction to Text-to-Speech Synthesis”, Springer, 1997.

(The Board of Editors received the paper on 15/3/2019, its review was completed on 23/8/2019)