ISSN 2320 - 2602
Volume 2, No.1, January 2013
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2013/ijacst01212013.pdf

A study in Vietnamese statistical parametric speech synthesis base on HMM

Son Thanh Phan[1] , Thang Tat Vu[2] , Cuong Tu Duong[3] , Mai Chi Luong[4]
[1]Le Quy Don Technical University, Vietnam, sonphan.hts@gmail.com
[2]Institute of Information Technology, Vietnam Academy of Science and Technology, vtthang@ioit.ac.vn
[3]Le Quy Don Technical University, Vietnam, cuongdt60@gmail.com
[4]Institute of Information Technology, Vietnam Academy of Science and Technology, lcmai@ioit.ac.vn

ABSTRACT

This article describes an approach in Vietnamese speech synthesis, using statistical parameters speech synthesis system based on hidden Markov models (HMMs), that has grown in popularity over the last few years. Spectral, pitch, tone, and phone duration are simultaneously modeled in HMMs and their parameter distributions are clustered independently by using decision tree-based context clustering algorithms. In this system, statistical modeling is applied to learn distributions of context-dependent acoustic vectors extracted from speech signals, each vector containing a suitable parametric representation of one speech frame and Vietnamese phonetic rules to synthesize speech. Several contextual factors such as tone types, syllables, words, phrases, and utterances were determined and are taken into account to generate the spectrum, pitch, and state duration. The resulting system yields significant correctness for a tonal language, and a fair reproduction of the prosody.

Key words : Vietnamese speech synthesis, context dependent, HMM-based, statistical parametric speech synthesis.

1. INTRODUCTION

A text-to-speech (TTS) system converts normal language text into speech using speech synthesis techniques. Speech synthesis is the computer-generated simulation of human speech. Speech synthesis has been developed steadily over the last few decades and it has been incorporated into several new applications with considerable results [1]. The basic methods for low-level synthesis are the articulatory, formant, concatenation synthesis and statistical parameters synthesis based on hidden Markov models. Although many speech synthesis systems can synthesize high quality speech, they still cannot synthesize speech with various voice characteristics such as speaker individualities, speaking styles, emotions, etc. To obtain various voice characteristics in speech synthesis systems based on the selection and concatenation of acoustical units, a large amount of speech data is necessary. However, it is difficult to collect store such speech data. In order to construct speech synthesis systems which can generate various voice characteristics, the HMM-based speech synthesis system (HTS) [1] was proposed.

The statistical parametric speech synthesis system based on HMMs has grown in popularity over few years recently. And speech parameterization and reconstruction is a hot topic at present, mainly because of the great development of this method [1]. HTS requires the input signals to be translated into tractable sets of vectors with good properties. Thus, Mel-frequency Cepstral Coefficients (MFCCs) are widely used for modeling spectral in synthesis and conversion systems [1].

This paper presents a method that extracts MFCCs and F_0 from speech frames, and vice versa, assuming Mel Log Spectral Approximation filter for speech waveforms. The tool has been specifically designed to be integrated into HTS. The implemented method has the following interesting properties:

- It allows extracting high-order MFCCs.
- It does not require excitation parameters other than F_0.
- It achieves considerably high perceptual quality in resynthesize.
- It allows several speech manipulations and modifications.

Since the HTS offers the attractive ability to be implemented for a new language without requiring the recording of extremely large databases, we apply HTS to Vietnamese - a mono-syllabically tonal language. We also constructed a Vietnamese speech database in order to create the synthesis system. The speech waveforms in the database was segmented and annotated with contextual information about tone, syllable, word, phrase, and utterance that could influence the speech to be synthesized [2].

Using context-dependent HMMs, the system can model the speech spectral, excitation as fundamental frequency, and phoneme duration simultaneously. In the system, fundamental frequency and state duration are modeled by multi-space probability distribution HMMs [3] and

multi-dimensional Gaussian distributions [4], respectively. The feature vector of HMMs consists of two streams, i.e., the one for spectral parameter and the other for fundamental frequency, and each phoneme HMM has its state duration densities. The distributions for spectral parameter, fundamental frequency and state duration are clustered independently by using a decision-tree based context clustering technique.

This paper is structured as follows. First an outline of HMM is given and introduces a brief description for Vietnamese speech synthesis system base on HTS. Then, some experimental results on Vietnamese synthesis and subjective evaluation tests, comparing the quality of synthesized speech with natural speech are here shown. Finally, concluding remarks and our plans for future work are presented.

2. THE HIDDEN MARKOV MODEL

A hidden Markov model $\lambda(A, B, \pi)$ is defined by its parameters: A – state transition probability, B – output probability and π – initial state probability.

Let us have the HMM λ that contains concatenated elementary triphone or monophone HMMs that correspond to the symbols in the word *w*, which has to be synthesized.

The aim of the speech synthesis is to find the most probable sequence of states features vectors $\hat{x}$ from the HMM λ. Figure 1 shows the model in state q_i at time t_i.

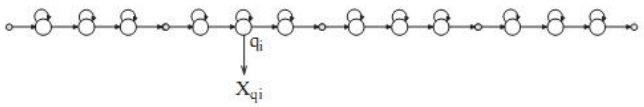

Figure 1: Concatenated HMM chain

X_{qi} is the M-dimensional generated feature vector at the state q_i of the model λ:

$$x_{qi} = (x_1^{(qi)}, x_2^{(qi)}, ..., x_M^{(qi)})^{\mathrm{T}} \quad (1)$$

From model λ we expect to generate a sequence of features vectors $\hat{x} = x_{q1}, x_{q2}, \ldots, x_{qL}$ of length L maximizing the overall likelihood $P(x|\lambda)$ of a HMM:

$$x = \arg\max_x \{P(x|\lambda)\} = \arg\max_x \left\{ \sum_Q P(x|q,\lambda) P(q|\lambda) \right\} \quad (2)$$

where the $Q=q_1, q_2, .., q_L$ is the path through the states of the model λ. The overall likelihood of the model $P(x|\lambda)$ is computed by adding the product of joint output probability $P(x|q,\lambda)$ and state sequence probability $P(q|\lambda)$ over all possible paths Q [11].

3. HMM-BASED SPEECH SYNTHESIS SYSTEM

In general, speech signals can be synthesized from the feature vectors. In the HTS, the feature vectors include spectral parameters as Mel-cepstral coefficients, tone, state duration, and excitation parameters such as the fundamental frequency F_0.

Figure 2 shows the training part of the HMM-based Vietnamese speech synthesis system. In this part, spectral parameters and excitation parameters are extracted from speech database. Then, they are modeled by context-dependent HMMs.

Figure 4 shows the synthesis part of the HMM-based Vietnamese speech synthesis system. In this part, a context-dependent label sequence is obtained and a sentence HMM is constructed by concatenating context dependent HMMs according to the context dependent label sequence. By using parameter generation algorithm [5], spectral and excitation parameters are generated from the sentence HMM. Finally, through a synthesis filter, speech signals in waveforms is synthesized from the generated spectral and excitation parameters [6]. Spectral and excitation parameters are needed for any synthesis filter to generate speech waveforms so both must be modeled by HMMs. Training and synthesis parts of the system are explained with applying them to Vietnamese in the following sections.

2.1. Training part

In the training part, inputs are utterances and their transcriptions at phoneme level, context dependent HMMs are then trained from excitation, spectral parameters together with their dynamic features for each speech unit. Spectral parameters are modeled using continuous distribution HMMs [7], but excitation parameters modeled using Multi-Space probability Distribution HMMs (MSD-HMMs) to overcome the problem of the voiced and unvoiced regions [8]. Also, state duration densities are modeled by single Gaussian distributions [4].

The training of phoneme HMMs using excitation and spectral parameters simultaneously is enabled in a unified framework by using multi-space probability distribution HMMs and multi-dimensional Gaussian distributions [8]. The simultaneous modeling of F_0 and Mel-cepstral parameter resulted in the set of context-dependent HMMs. Context-dependent clustering of Gaussian distributions was performed independently for spectrum, fundamental frequency and state duration because of the different clustering factor influence.

Spectral Modeling

In this approach the Mel-frequency cepstral coefficients (MFCCs) include tone, state duration parameters and their corresponding delta and delta-delta coefficients are used as spectral parameter. Sequences of Mel-cepstral coefficient vector, which are obtained from speech database using a Mel-cepstral analysis technique, are modeled by continuous density HMMs. The Mel-cepstral analysis technique enables speech to be re-synthesized from the Mel-frequency cepstral coefficients by using the MLSA (Mel Log Spectral Approximation) filter. The MFCCs are extracted through a *24*-th order Mel-cepstral analysis, using 40-ms Hamming windows with 8-ms shifts. Output probabilities for the

MFCCs correspond to multivariate Gaussian distributions [2].

Excitation Modeling

The excitation parameters are composed of logarithmic fundamental frequencies ($\log F_0$) and their corresponding delta and delta-delta coefficients. The variable dimensional parameter sequences such as $\log F_0$ with unvoiced regions properly are modeled by a HMM based on Multi-Space probability Distribution [8].

State Duration Modeling

State duration densities are modeled by single Gaussian distributions [4]. Dimension of state duration densities is equal to the number of state of HMM, and the *n*-th dimension of state duration densities is corresponding to then *n*-th state of HMMs. Here, the topology of HMMs includes left-to-right no-skip states.

There were some proposed techniques for training HMMs using their state duration densities simultaneously. However, these techniques require a large storage and computational load. In this paper, state duration densities are estimated by using state occupancy probabilities which are obtained in the last iteration of embedded re-estimation [4].

a) Phoneme level:

- Two preceding, current, two succeeding phonemes;
- Position in current syllable (forward, backward);

b) Syllable level:

- Tone types of two preceding, current, two succeeding syllables;
- Number of phonemes in preceding, current, succeeding syllables;
- Position in current word (forward, backward);
- Stress-level;
- Distance to {previous, succeeding} stressed syllable;

c) Word level:

- Part-of-speech of {preceding, current, succeeding} words;
- Number of syllables in {preceding, current, succeeding} words;

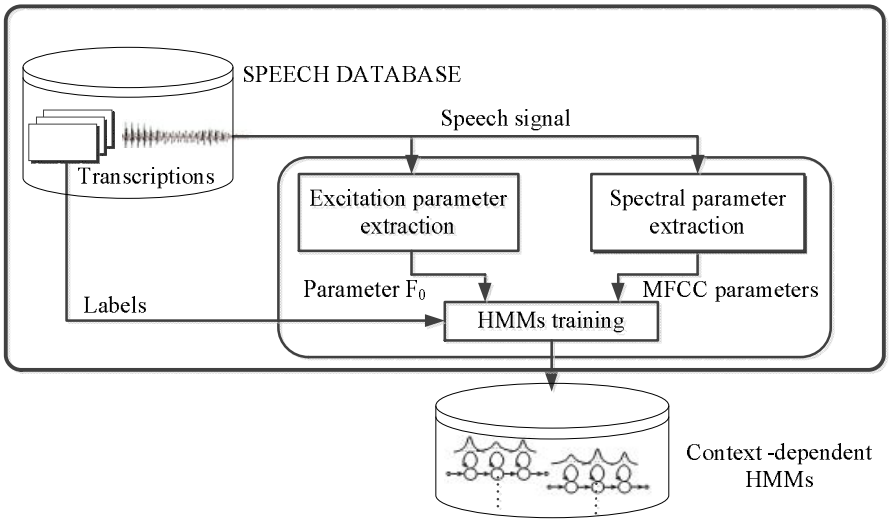

Figure 2: The training part of HMM-based speech synthesis system

Language-dependent Contextual Factors

There are many contextual factors (e.g., phone identity factors, stress-related factors, dialect factors, tone factors) that affect spectrum, pitch and state duration. Note that a context dependent HMM corresponds to a phoneme.

The only language-dependent requirements within the HTS framework are contextual labels and questions for context clustering. Since Vietnamese is a tonal language, a tone-dependent phone sets and corresponding phonetic and prosodic question set for the decision tree are considered. A tree-based context clustering is designed to have tone correctness which is crucial in Vietnamese speech [9, 10].

Some contextual information in Vietnamese language was considered as follows [2]:

- Position in current phrase;
- Number of content words in current phrase {before, after} current word;
- Distance to {previous, succeeding} content words;
- Interrogative flag for the word;

d) Phrase level:

- Number of {syllables, words} in {preceding, current, succeeding} phrases;
- Position of current phrase in utterance;

e) Utterance level:

- Number of {syllables, words, phrases} in the utterance;

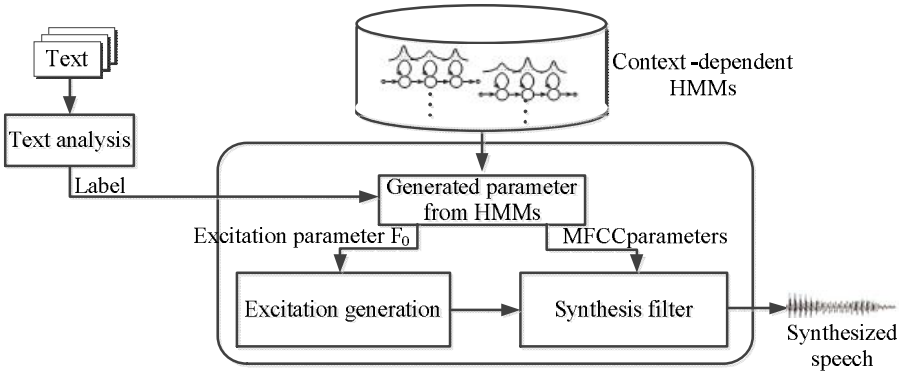

Figure 4: The synthesis part of HMM-based speech synthesis system

Decision tree-based context clustering

In some cases, a speech database does not have enough contextual samples or a given contextual label does not have a corresponding HMM in the trained model set. Therefore, to overcome this problem, a decision tree-based context clustering technique is applied to the distributions of spectrum, fundamental frequency and state duration.

In order to carry out decision tree-based context clustering, some questions were determined to cluster the phonemes. Afterwards, these questions were extended to include all the contextual information, i.e., tone, syllable, word, phrase and utterance. The questions for training part of HTS were derived according to phonetic characteristics of tones, vowels, semi-vowels, diphthongs, and consonants. The classifications for the phonemes and tones were used for making questions and applied to generate the decision trees. The decision trees for context clustering are shown in figure 3.

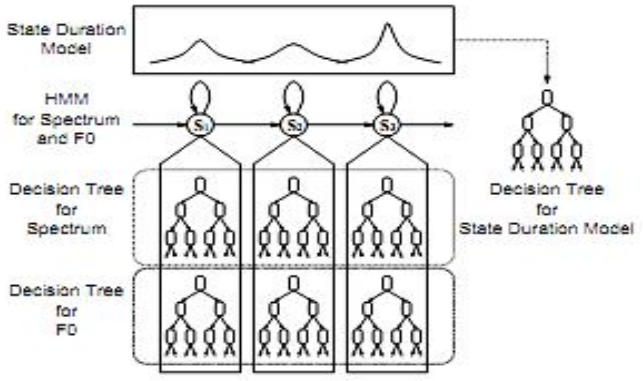

Figure 3: Decision trees for context clustering

2.2. Synthesis part

In the synthesis part, from the set of context-dependent HMMs according to the context label sequence that corresponds to the utterance in the entry text, the speech parameters are generated. The generated excitation parameters and Mel-cepstral parameters are used to generate the waveform of speech signal using the source-filter model. The advantage of this approach is in capturing the acoustical features of context-dependent phones using the speech corpora. Synthesized voiced characteristics can also be changed easily by altering the HMM parameters and the system can be easily ported to a new language.

In this part, an arbitrarily given text to be synthesized is converted to a context-based label sequence. Then, according to the label sequence, a sentence HMM is constructed by concatenating context dependent HMMs. State durations of the sentence HMM are determined so as to maximize the likelihood of the state duration densities [6]. According to the obtained state durations, a sequence of Mel-cepstral coefficients and pitch values including voiced/unvoiced decisions is generated from the sentence HMM by using the speech parameter generation algorithm [5]. Finally, speech is synthesized directly from the generated Mel-cepstral coefficients and pitch values by using the MLSA filter.

4. EXPERIMENTS

We used phonetically balanced 400 in 510 sentences (recorded male voice) from Vietnamese speech database for training. Speech signals were sampled at 16 kHz, and stored in a 16-bit PCM encoded waveform format and windowed by a 40-ms Hamming window with an 8-ms shift. MFCCs and fundamental frequency F_0 was calculated for each utterance using the Snack Sound ToolKit (Tksnack) tool on Ubuntu. Feature vector consists of spectral, tone and pitch parameter vectors: spectral parameter vector consists of 39 Mel-frequency cepstral coefficients including the zero-th coefficient, their delta and delta-delta coefficients (12 MFCC coefficients and an energy component). Pitch feature vector consists of $\log F_0$, its delta and delta-delta. We used 5-state left-to-right HMMs with single diagonal Gaussian output distributions, number of iterations of embedded training, expectation-maximization (EM) algorithm with 20 iterations is used to generate speech parameter, limit for F_0 extraction in 80-350 Hz.

For the evaluation, we used remain 110 sentences in the speech database, these sentences are used as synthesize data. Context-dependent labels were automatically generated from texts using a Vietnamese text analyzer. Context-dependent HMMs were trained for each of the spectral, F_0, and periodic

components using a decision-tree based context clustering technique.

5. EVALUATION

In this section, we aim to evaluate the quality of synthesized speech. The preliminary evaluations show the similarity of spectrogram and pitch contours of natural speech signals with synthetic speech signals using MLSA filter, and by NHMTTS software ones.

Since the means of state duration models are used in speech generation, the duration of a generated utterance can be different from that of the original. In this experiment, a sequence of states, which are obtained by force-aligning the original feature observations with the spectral and pitch models, is used for speech parameter generation. Therefore, we can make a comparison between synthesized and original speech signals while isolating duration differences.

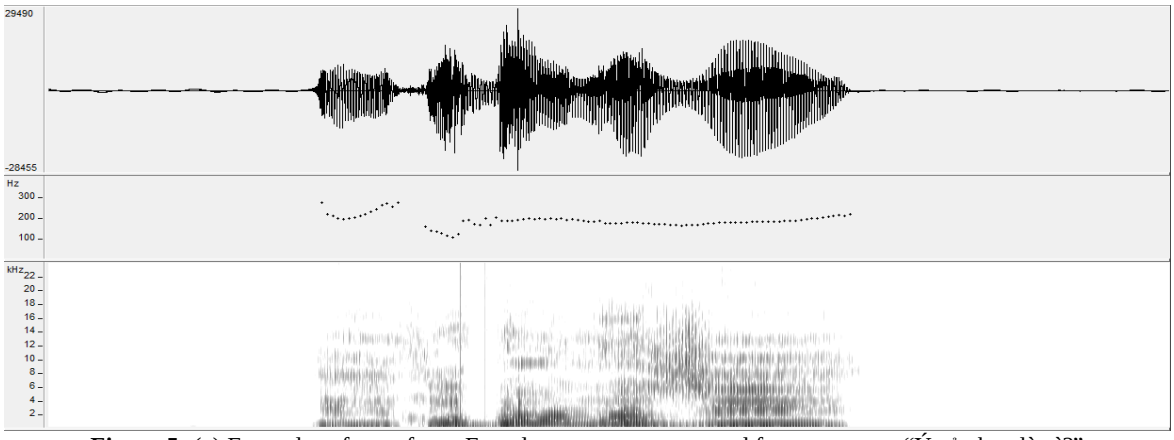

Figure 5: (a) Examples of waveform, F_0 and spectrogram extracted from utterance "Ý của bạn là gì?" (In English "What do you mean?") in natural speech

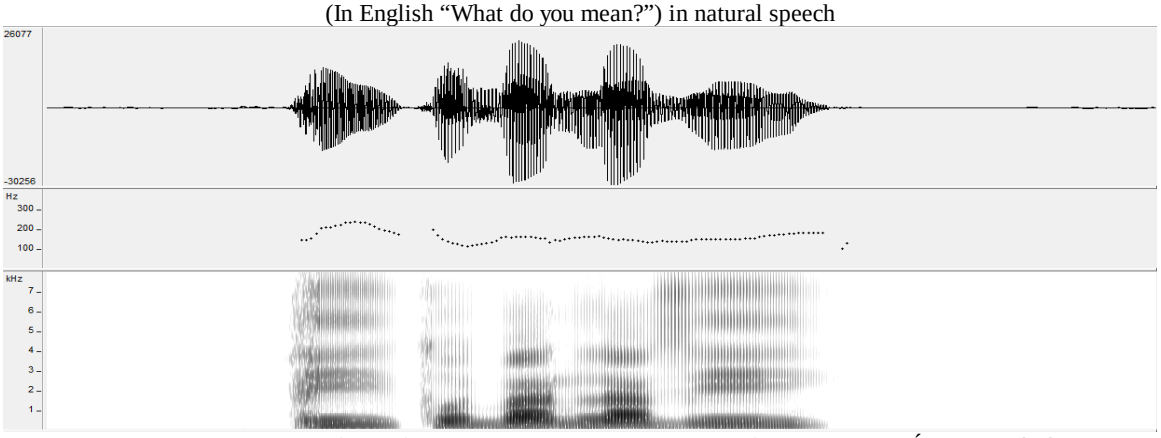

Figure 5: (b) Examples of waveform, F_0 and spectrogram extracted from utterance "Ý của bạn là gì?" (In English "What do you mean?") in generated speech by our system

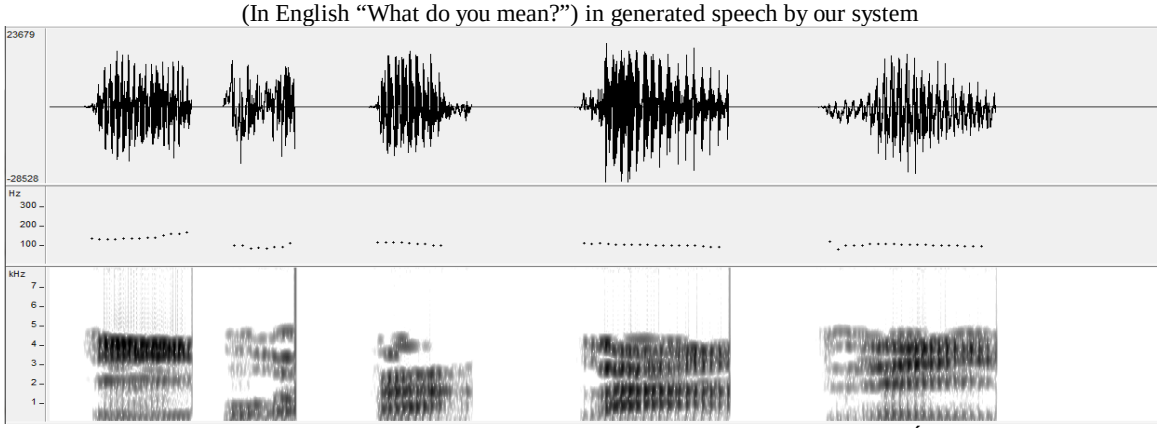

Figure 5: (c) Examples of waveform, F_0 and spectrogram extracted from utterance "Ý của bạn là gì?" (In English "What do you mean?") in generated speech by NHMTTS software

Figures 5(a), 5(b) and 5(c) show a comparison of waveform graph, spectrogram and F_0 patterns between original speech signal with speech signals are synthesized by our HTS and by NHMTTS software (author Nguyen Huu Minh) for a given sentence (utterance "*Ý của bạn là gì?*", in English: "*What do you mean?*"), which is not included in the training database but was uttered by the speaker who recorded the database. It can be noticed that the generated waveform, spectrogram and F_0 contour base on HMM are quite close to the natural patterns.

6. CONCLUSION

This paper presented a description of the HMM-based speech synthesis technique implemented for Vietnamese language, in which spectral, tone, state duration and fundamental frequency are modeled simultaneously in a unified framework of HMM. Contextual information and questions for decision tree-based context clustering were designed whereas a tone-dependent phone set is employed in training HMMs with phonetic and prosodic question set in corresponding decision trees. The evaluation results show that our system can generate highly intelligible speech with naturalness and can be understood. Overall, our system yields fair reproductions of prosody.

As a result, it might be possible to synthesize speech with various voice characteristics, e.g., emotion expression, by applying speaker adaptation or speaker interpolation technique. Future work will be directed towards investigation of contextual factors and conditions of the context clustering, improvement of text processing, and evaluation of synthetic speech. Synthesizing speech with various voice characteristics by applying speaker adaptation and speaker interpolation techniques is also our future work.

ACKNOWLEDGEMENT

This work was partially supported by ICT National Project KC.01.03/11-15 "Development of Vietnamese – English and English – Vietnamese Speech Translation on specific domain". Authors would like to thank all staff members of Department of Pattern Recognition and Knowledge Engineering, Institute of Information Technology (IOIT) - Vietnam Academy of Science and Technology (VAST) for their support to complete this work.

REFERENCES

1. H. Zen, K. Tokuda, A. W. Black. **Statistical parametric speech synthesis**, *Speech Communication*, Vol.51, no.11, pp.1039-1064, 2009.
2. Thang Tat Vu, Mai Chi Luong, Satoshi Nakamura. **An HMM-based Vietnamese Speech Synthesis System**, *Proc. Oriental COCOSDA*, 2009.
3. K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi. **Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling**, *Proc. of ICASSP*, 1999.
4. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura. **Duration Modeling in HMM-based Speech Synthesis System**, *Proc. of ICSLP*, Vol.2, pp.29–32, 1998.
5. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. **Speech parameter generation algorithms for HMM-based speech synthesis**, *Proc.ICASSP 2000*, pp.1315–1318, June 2000.
6. T. Yoshimura. **Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems**, Doctoral Dissertation, Nagoya Institute of Technology, January 2002.
7. K. Tokuda, H. Zen, and A. Black. **An HMM-based speech synthesis system applied to English**, *in IEEE Speech Synthesis Workshop*, 2002.
8. K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi. **Multi-space probability distribution HMM**, *IEICE* Vol.E85-D,NO.3 March 2002.
9. T.T Vu, T.K. Nguyen, H.S. Le, C.M. Luong. **Vietnamese tone recognition based on MLP neural network**, *Proc. Oriental COCOSDA*, 2008.
10. H. Mixdorff, H. B. Nguyen, H. Fujisaki, C. M. Luong. **Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese**, *Proc. EUROSPEECH*, pp.177-180, Geneva, 2003.
11. L. R. Rabiner. **A tutorial on hidden Markov models and selected applications in speech recognition**, *Proc. IEEE*, Vol. 77, No. 2, pp. 257–286, 1989.