Extracting MFCC, $F_0$ feature in Vietnamese HMM-based speech synthesis

Son Thanh Phan ¹, Thang Tat Vu ², Mai Chi Luong ³

¹Faculty of Information Technology, Le Quy Don Technical University
²,³Institute of Information Technology, Vietnamese Academy of Science and Technology
¹sonphan.hts@gmail.com
²vtthang@ioit.ac.vn
³lcmai@ioit.ac.vn

Abstract—HMM-based statistical speech synthesis method is not requiring a very large speech corpus for training the system. 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. The method presented in this paper allows accurate MFCC, $F_0$ and tone extraction and high-quality reconstruction of speech signals. Its suitability for high-quality HMM-based speech synthesis is shown through evaluations subjectively.

Keywords—Vietnamese speech synthesis, context-dependent, speech parameterization, statistical parametric speech synthesis

I. INTRODUCTION

The basic methods for low-level synthesis are the articulatory, formant, concatenation synthesis and statistical parameters synthesis based on hidden Markov models (HMM-based Text-to-Speech Synthesis System-HTS). In theory, the most accurate method is articulatory synthesis which models the human speech production system directly, but it is also the most difficult approach. The formant synthesis is based on the modeling of the resonances in the vocal tract and perhaps is the most commonly used during last decades. The concatenate synthesis which is often based on concatenating a large number of pre-recorded samples to generate natural speech is becoming more popular. But its disadvantage is the inflexible, especially when we need the ability of synthesizing speech with various voice characteristics and speaking. One of reasons comes from the fact that, it is impractical to prepare and store a large amount of speech data of arbitrary speakers and speaking styles. 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].
Extracting MFCC, \( F_0 \) feature in Vietnamese HMM-based speech synthesis

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 organized as follows: section II introduces a brief description for Vietnamese speech synthesis system base on HTS. The experimental results on Vietnamese synthesis are presented in section III, the quality of synthesized speech is evaluated in section IV and concluding remarks and our plans for future work are presented in the final section.

II. 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, duration, and excitation parameters such as the fundamental frequency \( F_0 \) (log\( F_0 \) value).

Figure 1 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 2 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.
A. 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.

1) 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-thorder Mel-cepstral analysis, using 40-ms Hamming windows with 8-ms shifts. Output probabilities for the MFCCs correspond to multivariate Gaussian distributions [2].

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].

3) 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 the $n$-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].
Extracting MFCC, $F_0$ feature in Vietnamese HMM-based speech synthesis

4) 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]:

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;
   - 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;

f) 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.
B. 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.

The synthesis part of the HTS is shown in Figure 3. In the synthesis 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.

We used phonetically balanced 310 in 510 sentences (recorded female 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 SPTK tool. 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. 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 180-450 Hz.

For the evaluation, we used remain 200 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.
Extracting MFCC, \( F_0 \) feature in Vietnamese HMM-based speech synthesis

III. EVALUATION

In this section, we aim to evaluate the quality of synthesized speech. The preliminary evaluations show the similarity of spectrogram and pitch contours between synthetic speech signals using MLSA filter, and natural 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 4. (a) Examples of \( F_0 \) and spectrogram extracted from utterance “Tôi sẽ gọi lại” (In English “I will call back”) in natural speech](image)

![Figure 4. (b) Examples of \( F_0 \) and spectrogram extracted from utterance “Tôi sẽ gọi lại” (In English “I will call back”) in generated speech](image)

Figures 4(a) and 4(b) show a comparison of spectrogram and \( F_0 \) patterns between synthesized and original speech signals for a given sentence (utterance “Tôi sẽ gọi lại”, in English: “I will call back”), 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 spectrogram and \( F_0 \) contour are quite close to the natural patterns.

IV. CONCLUSION

In this paper, we present our HMM-based Text-to-Speech system 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.

V. ACKNOWLEDGEMENTS

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.

VI. REFERENCE