Improvement of Naturalness for an HMM-based Vietnamses Speech Synthesis using the Prosodic information

Conference Paper · September 2013

DOI: 10.1109/RIVF.2013.6719907

Son Thanh Phan
Vietnam Academy of Science and Technology
9 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Tuan Anh Dinh
Oregon Health and Science University
14 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Thang VU
Vietnamese Academy of Science and Technology Institute of Information Technol…
28 PUBLICATIONS **182** CITATIONS
SEE PROFILE

Chi Mai Luong
Vietnam Academy of Science and Technology
63 PUBLICATIONS **250** CITATIONS
SEE PROFILE

Improvement of Naturalness for an HMM-based Vietnamses Speech Synthesis using the Prosodic information

Thanh-Son PHAN, Tu-Cuong DUONG
Faculty of Information Technology
Le Qui Don Technical University
Hanoi, Vietnam
{sonphan.hts, cuongdt60}@gmail.com

Anh-Tuan DINH, Tat-Thang VU, Chi-Mai LUONG
Institute of Information Technology
Vietnam Academy of Science and Technology
Hanoi, Vietnam
{anhtuan, vtthang, lcmai}@ioit.ac.vn

***Abstract*—Natural-sounding synthesized speech is goal of HMM-based Text-to-Speech systems. Besides using context dependent tri-phone units from a large corpus speech database, many prosody features have been used in full-context labels to improve naturalness of HMM-based Vietnamese synthesizer. In the prosodic specification, tone, part-of-speech (POS) and intonation information are considered not as important as positional information. Context-dependent information includes phoneme sequence as well as prosodic information because the naturalness of synthetic speech highly depends on the prosody such as pause, tone, intonation pattern, and segmental duration. In this paper, we propose decision tree questions that use context-dependent tones and investigate the impact of POS and intonation tagging on the naturalness of HMM-based voice. Experimental results show that our proposed method can improve naturalness of a HMM-based Vietnamese TTS through objective evaluation and MOS test.**

Keywords—HTS, HMM, part-of-speech, prosodic information, Vietnamese Speech Synthesis, tri-phone, context-dependent, decision tree-based clustering

I. INTRODUCTION

In a last few decades, speech synthesis systems has been developed and incorporated into several new applications with considerable results. The basic methods for low-level synthesis are the articulatory, formant, concatenate synthesis and statistical parameters synthesis based on hidden Markov models (HMM-based Text-to-Speech Synthesis System-HTS). The HTS has grown in popularity over few years recently; this system can be built without requiring a very large speech corpus for training the hidden Markov models. This technique shows that the resulting HMMs set has the advantage of being small, but till yield the high quality of synthetic speech, which is very important for a synthesis system on PDA, smart-phone or tablet [6].

HTS requires the input signals to be translated into tractable sets of vectors with good properties. In addition, the HTS also offers the attractive ability to be implemented for a new language without requiring the recording of extremely large databases [2], so 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, POS, intonation, and prosodic information that could influence naturalness of the speech to be synthesized.

Using context-dependent HMMs, the system can model the speech spectral, 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 [13], 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 baseline Vietnamese HMM-based speech synthesis system. The labeling text and building decision tree questions with context-dependent information of tone, POS, intonation and prosody are presented in section III. In section IV, experiments and evaluation verify the performance of the proposed method on the quality of synthesized speech, and concluding remarks and our plans for future work are presented in the final section.

II. HMM-BASED SPEECH ANALYSIS AND SYNTHESIS SYSTEM CONFIGURATION

A. HMM model for speech synthesis

In general, speech signals can be synthesized from the feature vectors. In the HTS, the feature vectors include spectral parameters as Mel-Cepstral Coefficients (MCCs, or Mel-Frequency Cepstral Coefficients-MFCCs), duration, and excitation parameters such as the fundamental frequency, F0 (logF0 value).

Feature vector consists of spectral, duration and pitch parameter vectors: spectral parameter vector is 39 coefficients including the MFCC zero-th coefficients, their delta and delta-delta coefficients (12 MFCC coefficients and an energy component). Pitch feature (excitation parameter) vector consists of $\log F_0$, its delta and delta-delta. All of them are modeled simultaneously in a unified framework by using multi-space probability distribution HMMs and multi-dimensional Gaussian distributions [4].

1) Spectral feature parameter model

In this approach the MFCCs include energy component, state duration parameters and their corresponding delta and delta-delta coefficients and tone (partially) are used as spectral parameter. Sequences of Mel-frequency 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 [9].

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 voiced (include tone) and unvoiced regions properly are modeled by HMMs based on Multi-Space probability Distribution [4].

3) State Duration Modeling

State duration densities of phonemes are modeled by single Gaussian distributions [13]. 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 phoneme 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. And the duration of each state is determined by HMM-based speech synthesis system model state durations are modeled as multivariate Gaussian distribution [9].

4) Language-dependent Contextual Factors

There are many contextual factors (e.g., phone identity factors, stress-related factors, dialect factors, tone factors and intonation) 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 [12].

Some contextual information in Vietnamese language was considered as follows [9]:

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

5) HMM-based decision tree 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, POS, ToBI, syllable, word, phrase and utterance. The questions 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. Speech analysising and HMM training stage

Figure 1 shows the training stage of the HMM-based Vietnamese speech synthesis system. In this stage, spectral parameters and excitation parameters are extracted from speech database. Then, they are modeled by context dependent

HMMs. The 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.

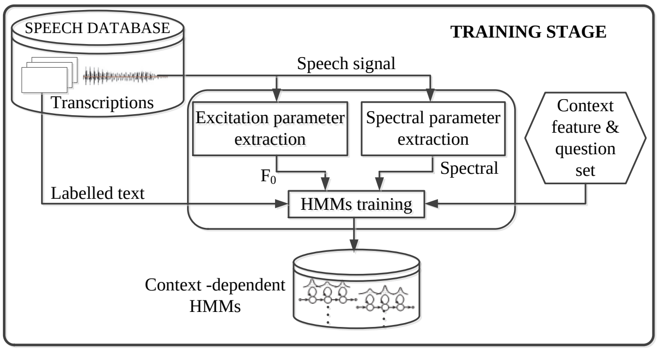

Fig. 1. Block diagram of HMM training

C. HMM-based speech synthesis system

In the synthesis stage, from the set of context-dependent HMMs according to the context label sequence 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.

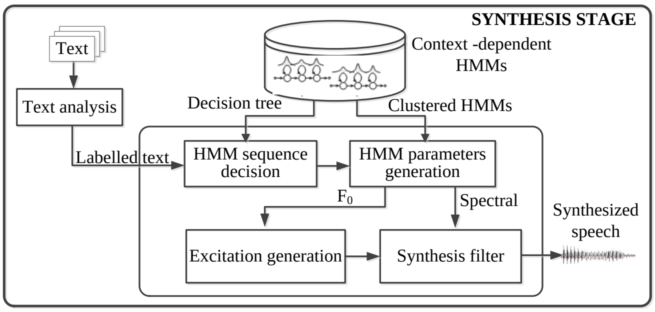

Fig. 2. Block diagram of HMM-based speech synthesis

The synthesis stage of the HTS is shown in Figure 2. In this stage, 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 [14]. 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 the MLSA filter [7].

III. THE IMPROVEMENT OF TONAL AND ACCENTUAL FEATURES IN LABELING

Although the tone on vowels, it plays an important role on entire syllable in Vietnamese. However, tonal features are not as explicit as other features in speech signal. According to Doan Thien Thuat [10], a syllable's structure can be described in Table I. Each syllable could be considered as a combination of Initial, Final and Tone. There are 22 Initials and 155 Finals in Vietnamese [11].

TABLE I. STRUCTURE OF VIETNAMESE SYLLABLE

Tone			
[Initial]	***Final***		
	[Onset]	Nucleus	[Coda]

In the first consonant, we can hear a little of the tone. Tone becomes clearer in rhyme and finished completely at the end of the syllable. The pervading phenomenon determines the non-linear nature of tone. So, with mono syllabic language like Vietnamese, a syllable can't easily be separated into small acoustic parts like European languages.

In syllable tagging process, contextual features must be considered. There are many contextual factors (ex, phonetic, stress, dialect, tone, intonation) affecting the signal spectral, fundamental frequency and duration. In additional, constructing a decision tree to classify the phonemes based on contextual information. A decision tree-based context clustering and questions set are designed to have tone correctness, POS and ToBI information, which very important in an HMM-based Vietnamese TTS [9].

A. Intonation in Vietnamese

In order to present intonation, we use Tones and Break Indices (ToBI) in intonation transcription phase. ToBI is a framework for developing a widely accepted convention for transcribing the intonation and prosodic structure of spoken sentences in various languages. ToBI framework system is supported in HTS engine. The primitives in a ToBI framework system are two tones, low (L) and high (H). The distinction between the tones is paradigmatic. That is L is lower than H in the same context. Utterances can consist of one or more intonation phrases. The melody of an intonation phrase is separated into a sequence of elements, each made up of either one or two tones. In our works, the elements can be classified into 2 main classes [1].

1) Phrase-final intonation

Intonation tones, mainly phrase-final tones, were analyzed in our work. Boundary tones are associated with the right edge of the prosodic phrase and mark the end of a phrase. It can be established in Vietnamese that, a high boundary tone can change a declarative into an interrogative. To present the boundary tone, 'L-L%', 'L-H%' tags are used. 'L-L%' refers to a low tone; and 'L-H%' describes a high tone. This is a common declarative phrase. The 'L-L%' boundary tone causes the intonation to be low at the end of the prosodic phrase. On the other hand, the effect of 'L-H%' is that first it will drop to a low value and then it will rise towards the end of the prosodic phrase.

2) *Pitch Accent*

Pitch Accent is the falling or rising trends in the top line or baseline of pitch contour. Most noun, verb and adjective in Vietnamese are accented words. An ‘H*’ (high-asterisk) tends to produce a pitch peak while an ‘L*’ (low-asterisk) pitch accent produces a pitch trough. In addition, the two other tag ‘L+H*’ and ‘H+L*’ are also used. ‘L+H*’ rises steeply from a much lower preceding F_0 value while ‘H+L*’ falls from a much higher preceding F_0 value.

It was showed in the experiment that: the intonation tags add valuable contextual information to Vietnamese syllables in training process. Spoken sentences can be distinguished easily between declarative and interrogative utterances. Import information in a speech is strongly highlighted.

B. *Part of Speech*

A POS tag is a linguistic category assigned to a word in a sentence based upon its morphological behavior. Words are classified into POS categories such as noun (N), verb (V), adjective (A), pronoun (P), determine (D), adverb (R), apposition (S), conjunction (C), numeral (M), interjection (I) and residual (X). Words can be ambiguous in their POS categories. The ambiguity normally solved by looking at the context of the word in the sentence.

Automatic POS tagging is processed with Conditional Random Fields. The training of CRFs model is basically to maximize the likelihood between model distribution and empirical distribution. So, CRFs model training is to find the maximum of a log - likelihood function.

Suppose that training data consists of a set of N pairs, each pair includes an observation sequence and a status sequence, D = {(x(i),y(i))} $\forall i = 1..N$. Log-likelihood function:

$$l(\theta) = \sum_{x,y} \tilde{p}(x,y) \log(p(y \mid x, \theta)), \qquad (1)$$

Here, $\theta(\lambda_1, \lambda_2, ..., \mu_1, \mu_2, ,...)$ is the parameter of the model and $\tilde{p}(x,y)$ is concurrent empirical experiment of *x, y* in training set. Replace $p(y \mid x)$ of (1), we have:

$$l(\theta) = \sum_{x,y} \tilde{p}(x,y) \left[\sum_{i=1}^{n+1} \lambda f + \sum_{i=1}^{n} \mu g \right] - \sum_{x} \tilde{p}(x) \log Z, \qquad (2)$$

Here, $\lambda(\lambda_1, \lambda_2, ..., \lambda_n)$ and $\mu(\mu_1, \mu_2, ..., \mu_m)$ are parameter vectors of the model, *f* is a vector of transition attributes, and *g* is a vector of status attributes.

IV. Experiment and Evaluation

In the experiment, we used phonetically balanced 400 in 510 sentences (recorded female and male voices, Northern dialect) from Vietnamese speech database for training. Speech signals were sampled at 48 kHz, mono channel and coded in PCM 16 bit then the signal is downgraded to 16 kHz in waveform format and windowed by a 40-ms Hamming window with an 8-ms shift. All sentences were segmented at the phonetic level. The phonetic labeling procedure was performed as text-to-phoneme conversion through a forced alignment using a Vietnamese speech recognition engine [12]. During the text processing, the short pause model indicates punctuation marks and the silence model indicates the beginning and the end of the input text.

For the evaluation, we used remain 110 sentences in the speech database, these sentences are used as synthesize data for testing and evaluating. MFCCs and fundamental frequency F_0 was calculated for each utterance using the SPTK tool [15]. Feature vector consists of spectral, tone, duration 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 [8].

A couple of comparisons of synthesized speech qualities, include male and female speech models with only tone and with additional POS, stress and intonation. The information is added to full context model of each phoneme in a semi automatic way.

A. *Objective test*

The objective measurement is described through comparing of waveform, pitch and spectrogram between natural speech and synthesized testing sentences in both cases.

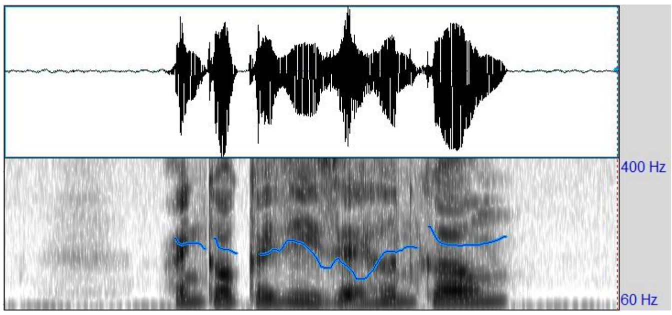

Fig. 3. Examples of wave form, F_0 and spectrogram extracted from utterance “Anh có cái gì rẻ hơn không?” (In English “Do you have anything cheaper?”) in natural speech, male voice

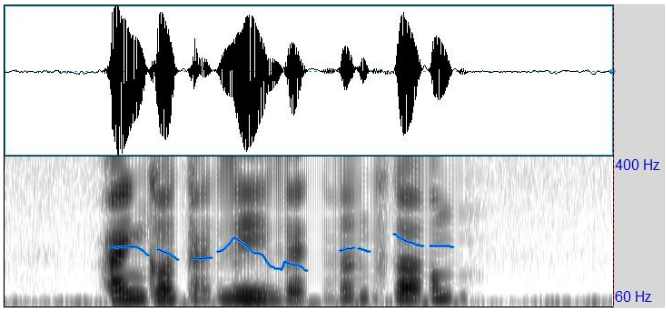

Fig. 4. Examples of wave form, F0 and spectrogram extracted from utterance “Anh có cái gì rẻ hơn không?” (In English “Do you have anything cheaper?”) in synthesized speech without POS and Prosody tagging, male voice

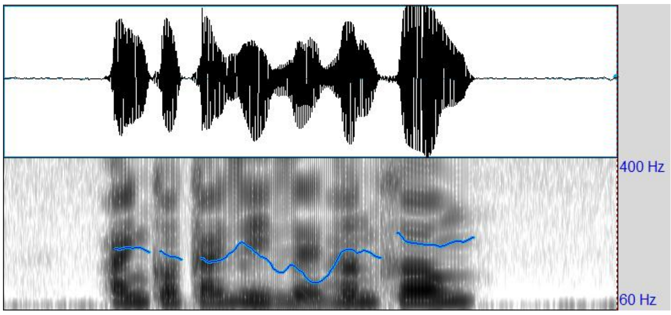

Fig. 5. Examples of wave form, F0 and spectrogram extracted from utterance “Anh có cái gì rẻ hơn không?” (In English “Do you have anything cheaper?”) in synthesized speech with POS and Prosody tagging, male voice

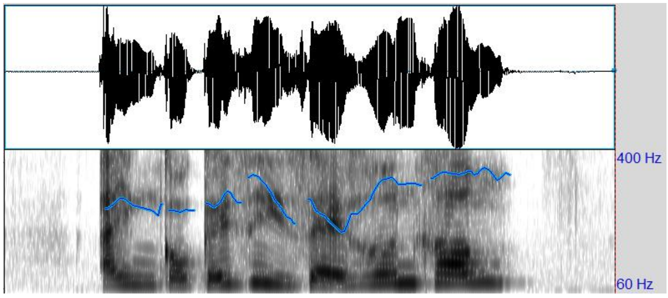

Fig. 6. Examples of wave form, F_0 and spectrogram extracted from utterance “Anh có cái gì rẻ hơn không?” (In English “Do you have anything cheaper?”) in natural speech with POS and Prosody tagging, female voice

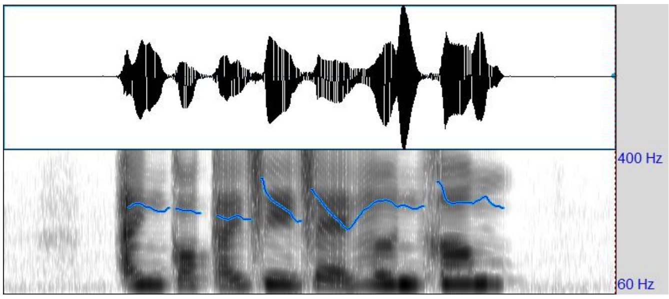

Fig. 7. Examples of wave form, F_0 and spectrogram extracted from utterance “Anh có cái gì rẻ hơn không?” (In English “Do you have anything cheaper?”) in synthesized speech without POS and Prosody tagging, female voice

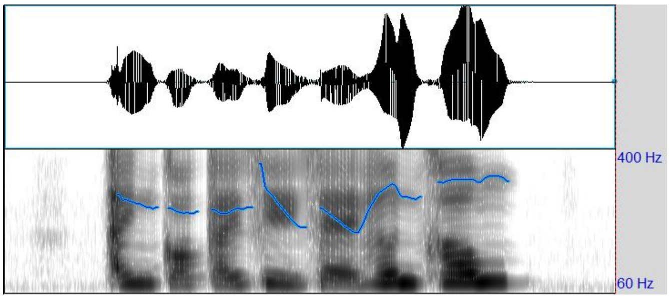

Fig. 8. Examples of wave form, F_0 and spectrogram extracted from utterance “Anh có cái gì rẻ hơn không?” (In English “Do you have anything cheaper?”) in synthesized speech with POS and Prosody tagging, female voice

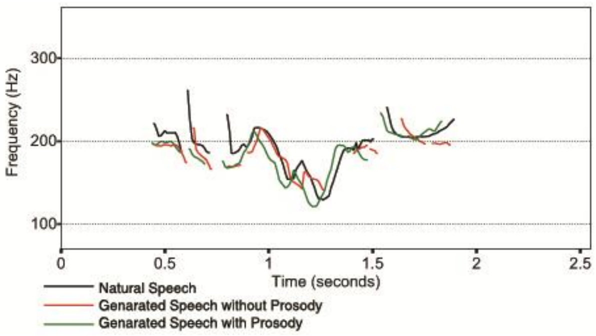

Fig. 9. Comparison F_0 contour of Natural Speech and Generated Speeches, male voice

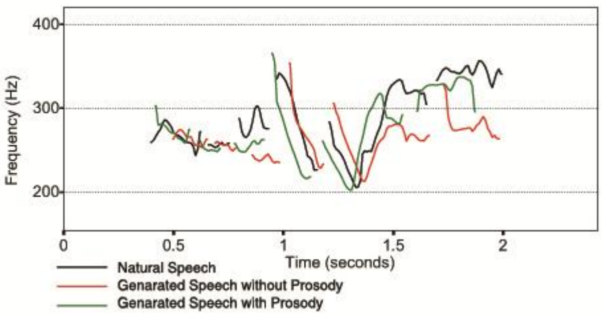

Fig. 10. Comparison F_0 contour of Natural Speech and Generated Speeches, female voice

B. MOS test

As a further subjective evaluation, MOS tests were used to measure the quality of synthesized speech signals in comparison with natural ones. The rated levels were: bad (1), poor (2), fair (3), good (4), and excellent (5). In this test, fifty sentences were randomly selected. With three types of audio, (1) natural speech signals, (2) the synthetic speech signals without POS, accent and intonation, and (3) the synthetic speech signals with POS, accent and intonation, the number of listeners were 50 people. The speech signals were played in random order in the tests.

Table II shows the mean of opinion scores which were given by all the subjects. The MOS result implied that the quality of natural speech is from good to excellence, and the quality of synthesis speech is from fair to good.

TABLE II. RESULTS OF THE MOS TEST

Speech	Mean Opinion Score
Natural	**4.57**
Without POS, Accent, Intonation	**3.26**
With POS, Accent, Intonation	**3.92**

V. Conclusion

The experimental results, shown that POS and prosody information do contribute to the naturalness (specifically in terms of pitch) of a TTS voice when it forms part of a small phoneme identity-based feature set in the full context HTS labels. However, the same effect, even an improvement, can be accomplished by including segmental counting (phrase) and positional information in segment instead of the POS tags in the HTS labels-and no extra resources are used. The experiments were limited by Northern dialect corpus. It would be prudent to test the effects on the other dialects, especially the South Central dialect.

The proposed tonal features can improve the tone intelligibility for generated speech. In addition, beside tonal features, the proposed POS and prosody features give the better improvement of the synthesized speech quality. These results confirm that the tone correctness and prosody of the synthesized speech is significantly improved and more naturalness when using most of the extracted speech features. Future work includes the improvement of text processing automatically and work on the contextual information.

Acknowledgment

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. Mixdorff, H. B. Nguyen, H. Fujisaki, C. M. Luong, “Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese,” Proc. EUROSPEECH, Geneva, pp.177-180, 2003.

[2] H. Zen, K. Tokuda, A. W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol.51, no.11, pp.1039-1064, 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, Phoenix, Arizona, USA, pp. 229-232, 1999.

[4] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi,“Multi-space probability distribution HMM,” IEICE Vol.E85-D,No.3, pp.455-464, March 2002.

[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] Phan Thanh Son, Vu Tat Thang, “HMM-based Speech Synthesis for Vietnamese language,” International Conference on Science and Technology, 45th Anniversary of Electric Power University, Hanoi, Vietnam, pp 338-346, November 2011.

[7] S. Imai, “Cepstral analysis synthesis on the mel frequency scale,” Proc. of ICASSP, pp.93–96, 1983.

[8] Son Thanh Phan, Thang Tat Vu, Cuong Tu Duong, and Mai Chi Luong, “A study in Vietnamese statistical parametric speech synthesis base on HMM,” IJACST, Vol. 2, No. 1, pp. 01-06, Jan 2013.

[9] Thang Tat Vu, Mai Chi Luong, Satoshi Nakamura, An HMM-based Vietnamese speech synthesis system, Proc. Oriental COCOSDA, Urumqi, China, pp. 108-113, 2009.

[10] T.T. Doan, “Vietnamese Acoustic,” Vietnamese National Editions, Second edition, 2003.

[11] T.T Vu, D.T. Nguyen, M.C. Luong, J.P Hosom, “Vietnamese large vocabulary continuous speech recognition,” Proc. INTERSPEECH, pp. 1689-1692, 2005.

[12] T.T Vu, T.K. Nguyen, H.S. Le, C.M. Luong, “Vietnamese tone recognition based on MLP neural network,” Proc. Oriental COCOSDA, Kyoto, Japan, pp. 116-121, 2008.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Duration modeling in HMM-based speech synthesis system, Proc. of ICSLP,” Vol. 2, Sydney, Australia, pp. 29-32, 1998.

[14] 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.

[15] Department of Computer Science, Nagoya Institute of Technology, Speech Signal Processing Toolkit, SPTK 3.6. Reference manual, http://sourceforge.net/projects/sp-tk/, Japan, 12- 2003. [Updated 25-12-2012].