Vietnamese HMM-based Speech Synthesis with prosody information

Anh-Tuan DINH[2],Thanh-Son PHAN[1], Tat-Thang VU[2], Chi-Mai LUONG[2]

[1]Faculty of Information Technology, Le Qui Don Technical University, Hanoi, Vietnam
[2]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

[1]sonphan.hts@gmail.com, [2]{anhtuan,vtthang,lcmai}@ioit.ac.vn

Abstract

Generating natural-sounding synthetic voice is an aim of all text to speech system. To meet the goal, many prosody features have been used in full-context labels of an HMM-based Vietnamese synthesizer. In the prosody specification, POS and Intonation information are considered not as important as positional information. The paper investigates the impact of POS and Intonation tagging on the naturalness of HMM-based voice. It was discovered that, the POS and Intonation tags help reconstruct the duration and emotion in synthesized voice.

Index Terms: Vietnamese speech synthesis, tone characteristics, tonal language, prosody tagging, part-of-speech, hidden Markov models

1. Introduction

In HMM-based speech synthesis systems, the prominent attribute is the ability to generate speech with arbitrary speaker's voice characteristics and various speaking styles without large amount of speech data1].

The naturalness of a Vietnamese TTS system is mainly affected by prosody. Prosody consists of accent, intonation and Vietnamese tones (6 tones). The features with part-of-speech (POS) tagging have close relationships and determine the naturalness and the intelligibility of synthesized voice.

Vietnamese tones consist of level, falling, broken, curve, rising, drop. One syllable can change its meaning when it goes with different tones. So the tonal feature has a strong impact on the intelligibility of synthetic voice. The following Table 1 shows an example about the name, the mark of tone in Vietnamese.

Table 1. *Six tones in Vietnamese*

Name	Tone mark	Example
LEVEL (ngang)	Unmarked	ta – me
FALLING (huyền)	Grave	tà – bad
BROKEN (ngã)	Tilde	tã - napkin
CURVE (hỏi)	Hook above	tả - describe
RISING (sắc)	Acute	tá – dozen
DROP (nặng)	Dot below	tạ - quintal

2. Text to Speech System

A TTS system, showed in Figure 1, is the production of speech from text. It includes the following stages[5]:

- Text tokenization splits the input text stream into smaller units named sentences and tokens. In the phase, written forms of Vietnamese syllables are discovered and tagged. The process is called tokenization.
- Text normalization decodes non-standard word into one or more pronounceable words. Non-standard tokens including numbers, dates, time, abbreviations… are normalized in the phase. The process is also called homograph disambiguation.
- Text parsing investigates lexical, syntactic structures from words which are used for pronunciation and prosodic modeling stages. The stage consists of POS tagging and chunking.

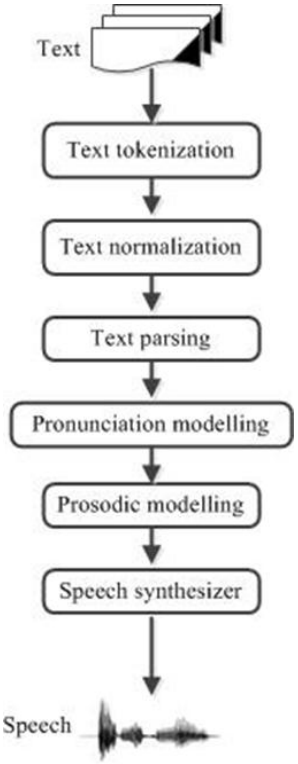

Figure 1: *A TTS system.*

- Pronunciation modeling maps each word to its phonemes. It looks up the words in a lexicon or use grapheme to phoneme (G2P) rules to finish the task. Accents and tones are assigned
- Prosodic modeling predicts the prosody of sentences. Sentence-level stress is identified, the intonation is assigned to sentences which make melody or tune of entire sentences.
- Speech synthesizer generates the speech waveforms from the above information. In Hidden Markov Model-based synthesis, a source filter paradigm is used to model the speech acoustics; information from previous stages are used to make full-context label file of each phoneme in the input sentence. Excitation (fundamental

frequency F_0), spectrum and duration parameters are estimated from recorded speech and modeled by context-dependent HMMs.

Figure 2 is an example of full context model used in HMM-based Speech Synthesis:

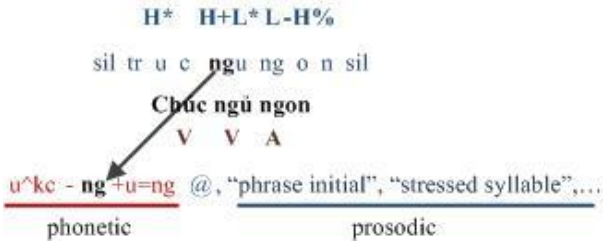

Figure 2: *A HMM full context model.*

Based on the full context model, HMM-based TTS is very flexible easy to add more prosodic information.

3. The improvement of tonal and accentual features

It is thought that tone lies on vowel; however, tone plays an important role on all over a 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 2:

Table 2. *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) affecting the signal spectral, fundamental frequency and duration. In additional, constructing a decision tree to classify the phonemes based on contextual information. The construction of the decision is very important in HMM-based Vietnamese TTS system[6].

Some contextual information include tone, accent, part-of-speech, was considered as follows[9]:

- Phoneme level:
  - Two preceding, current, two succeeding phonemes
  - Position in current syllable (forward, backward)
- 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
- 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
- Phrase level:
  - Number of {syllables, words} in {preceding, current, succeeding} phrases
  - Position of current phrase in utterance
- Utterance level:
  - Number of {syllables, words, phrases} in the utterance

4. 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[3]:

4.1. 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.

4.2. 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.

5. 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))} $\forall i = 1 \ldots N$. Log-likelihood function:

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

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:

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

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.

6. 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. 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[11]. 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. 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 logF_0, its delta and delta-delta[7].

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.

6.1. Objective test

The objective measurement is described through comparing of pitch contour between natural speech and synthesized testing sentences in both cases and showed in Figure 3 and Figure 4.

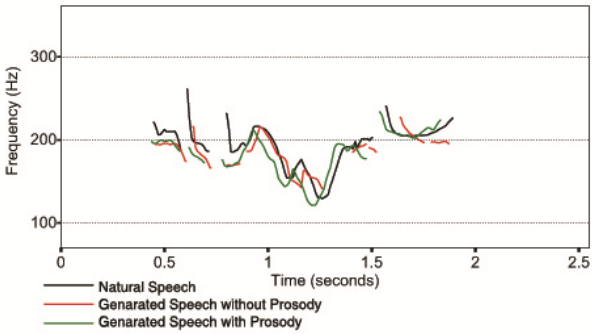

Figure 3: *ComparisonF_0 contour extracted from utterance "Anh có cái gì rẻ hơn không?"of Natural Speech and Generated Speeches, male voice*

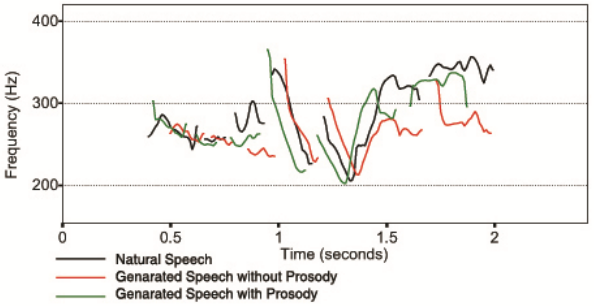

Figure 4: *ComparisonF_0 contour extracted from utterance "Anh có cái gì rẻ hơn không?"of Natural Speech and Generated Speeches, female voice*

6.2. 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, a hundred 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 3 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 3. *Resultsof the MOS test*

Speech	Mean Opinion Score
Natural	**4.53**
Without POS, Accent, Intonation	**3.15**
With POS, Accent, Intonation	**3.89**

7. Conclusions

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. 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.

8. 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.

9. References

[1] J.Yamagishi, K.Ogata, Y.Nakano, J.Isogai, T.Kobayashi,"HSMM-Based Model adaptation algorithms for Average-Voice-Based speech synthesis", 77–80, ICASSP 2006.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura,"Speech parameter generation algorithms for HMM-based speech synthesis", 1315–1318, Proc. ICASSP,June 2000.

[3] H. Mixdorff, H. B. Nguyen, H. Fujisaki, C. M. Luong,"Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese", 177-180, Proc. EUROSPEECH, Geneva, 2003.

[4] Phu Ngoc Le, Eliathamby Ambikairajah, Eric H.C. Choi,"Improvement of Vietnamese Tone Classification using FM and MFCC Features", 01-04, Computing and Communication Technologies RIVF 2009.

[5]Schlunz, GI, Barnard, E and Van Huyssteen, GB,"Part-of-speech effects on text-to-speech synthesis", 257-262, 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), Stellenbosch, South Africa, 22-23 November 2010.

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

[7] Son Thanh Phan, Thang Tat Vu, and Mai Chi Luong,"Extracting MFCC, F_0 feature in Vietnamese HMM-based speech synthesis, International Journal of Electronics and Computer Science Engineering", 2(1):46-52, Jan 2013.

[8]Tang-Ho Le, Anh-Viet Nguyen, Hao Vinh Truong, Hien Van Bui, and Dung Le,"A Study on Vietnamese Prosody", 63-73, New Challenges for Intelligent Information and Database SystemsStudies in Computational Intelligence Volume 351, 2011.

[9] Thang Tat Vu, Mai Chi Luong, Satoshi Nakamura,"An HMM-based Vietnamese Speech Synthesis System", 116 - 121, Proc. Oriental COCOSDA, 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", 1689-1692, Proc. INTERSPEECH, 2005.