INTONATION ISSUES IN HMM-BASED SPEECH SYNTHESIS FOR VIETNAMESE

NGUYEN Thi Thu Trang[1, 2], *TRAN Do Dat*[1], *Albert RILLIARD*[2], *Christophe D'ALESSANDRO*[2], *PHAM Thi Ngoc Yen*[1]
[1] MICA Institute, HUST - CNRS/UMI2954 - Grenoble INP, Hanoi University of Science and Technology, Hanoi, Vietnam
[2] LIMSI-CNRS (UPR 3251), France
trangntt@soict.hust.edu.vn, {do-dat.tran, ngoc-yen.pham}@mica.edu.vn, {albert.rilliard, cda}@limsi.fr

ABSTRACT
In an HMM-based Text-To-Speech system, contextual features, including phonetic and prosodic factors have a significant influence to the spectrum, F0 and duration of the synthetic voice. This paper proposes prosodic features aiming at improving the naturalness of an HMM-based TTS system (VTed) for a tonal language, Vietnamese. The ToBI (Tones and Break Indices) features are used to learn two crucial prosodic cues i.e. intonation (boundary tones) and pause (break indices), concurrently with another set of features. The result of MOS test showed that the general quality of synthetic voice is rather good, 1.21 point lower than the natural voice. About 55% of the voice trained with ToBI boundary tone feature are perceived as similar to the voice trained without this feature, while a 10% difference in favour of the voice trained without this ToBI feature is observed. This may be linked with F0 contour lowering or raising regardless of lexical tones. This brought two main problems in the synthetic voice: discontinuity in spectrum and F0 or unexpected voice quality. This paper then concluded the need of much more work on intonation modeling that should take into account the Vietnamese tones. A new prosody model can be designed, which may consider the ToBI model, with respect to lexical tones and the syntactic structure of Vietnamese.

Index Terms — *Text-to-speech (TTS), speech synthesis, tonal language, Vietnamese, HMM-based speech synthesis, intonation, ToBI*

1. INTRODUCTION

The hidden Markov model (HMM-) based speech synthesis has recently been demonstrated to be very effective in synthesizing smooth and stable speech. It is most simply described as generating the average of some sets of similarly sounding speech segments [1]. There are many contextual features, including phonetic factors and prosodic factors that a!ect the spectrum, F0 and duration of the synthetic voice [2].

For tonal languages, there are number of works on prosody modeling in speech synthesis for improving the naturalness of the synthetic voice. They did research on improvement of tone intelligibility [3], intonation modeling [4][5] or stress [6]. Vietnamese, a monosyllabic and tonal language has recently been the subject of much linguistic research. Vietnamese tones generation using F0 and power patterns was proposed in [7] while the work in [8] proposed a method for generating F0 contours in syllable-based concatenative speech synthesis. In both work, F0 values were modeled and generated from rules based on corpus analyses. The intonation of the synthetic voice is hence at the "rule-based" level, sometimes discontinuous even based-on non-uniform unit selection [9], compared to statistical parametric synthesis.

In an HMM-based Text-To-Speech (TTS) system, there are many contextual features, including phonetic factors and prosodic factors that affect the spectrum, F0 and duration of the synthetic voice [2]. It is necessary to use a transcription model to generate prosodic labels for both training and synthesis phases. The ToBI (Tones and Break Indices) system [10], which is intended as a standard for the prosodic transcription of American English, is also supposed to be compatible with current work in language processing, explicitly modeling two crucial prosodic cues i.e. intonation (boundary tones) and pause (break indices). This paper experiments the use of ToBI labels for training prosodic features for our HMM-based TTS system for Vietnamese – VTed [11]. Some discussions and future works on intonation issues are given from the evaluation of the obtained results (based on both subjective and objective tests).

The rest of this paper is organized as follows. Section 2 presents the background of this work, including study of Vietnamese phonetics and phonology necessary for a TTS system and the ToBI transcription model. The system architecture and design of Vietnamese features, including how to extract Vietnamese prosodic features from text using ToBI model, are given in Section 3. The implementation and evaluation are presented in the Section 4. The final section gives conclusions and presents future works.

2. BACKGROUND

2.1 Vietnamese phonetics and phonology

Vietnamese, a tonal language, the official language of Vietnam is spoken natively by over seventy-five million people in Vietnam and greater Southeast Asia as well as by some two million overseas, predominantly in France, Australia, and the United States [12]. Both phonetics and prosody are necessary to understand language as a means of communication between people; hence they play important roles in speech processing.

Although there is considerable fluidity and a good deal of conflicting opinion, in general the pronunciation

of educated speakers from the Hanoi area of Vietnam is the most widely accepted as a sort of standard [13]. This section recapitulates the phonological system, phonetics and prosody of the modern Hanoi dialect of Northern Vietnamese (Hanoi Vietnamese), which need considering in design and implementing speech synthesis applications. More detail discussion was presented in [11].

2.1.1 Vietnamese syllable structure

Analysis of syllable structure has a direct bearing on the analysis of the phonemic system: numerous nuclei / vowels; or combination of glide and vowel, and also central for tone system. We adopted the hierarchical structure for Vietnamese syllables (Figure 1). There are 2 main parts of a syllable: An initial consonant and a rhyme. Tone is carried in the rhyme with 3 elements: medial, nucleus and ending. The nucleus and tone are compulsory while others are optional.

SYLLABLE			
Initial (C)	***Rhyme (carrying Tone)***		
	Medial (w)	*Nucleus V (V)*	*Ending (C) / (G)*

Figure 1: ***The hierarchical structure of Vietnamese syllables.***

2.1.2 Vietnamese phonological system

There are totally 19 initial consonants in Hanoi Vietnamese. In this dialect, orthographic ch- and tr- (/c/ and /ʈ/), d-, gi- and r- (/z/ and /ʐ/), x- and s- (/s/ and /ʂ/) are pronounced alike [13] [14] as /c/, /z/, /s/. Hanoi Vietnamese licenses eight segments in coda position: three unreleased voiceless obstruents /p t k/, three nasals /m n ŋ/, and two approximants /j w/. There are nine long vowels /i e ɛ a ɯ ɤ u o ɔ/, four short vowels /ĕ ă ɤ̆ ŏ/ and three falling diphthongs /ie ɯɤ uo/ [15].

2.1.3 Vietnamese tones

Northern Vietnamese speech varieties distinguish six lexical tones: level (1), falling (2), broken (3), curve (4), rising (5), and drop (6) and that some of these tones often involve voice quality contrasts. Figure 2 illustrates Vietnamese tones with a six-tone paradigm (1-4, 5a, 6a) for sonorant-final syllables, and a two-tone paradigm (5b, 6b) for obstruent-final syllables (i.e. the tones of syllables ending in /p/, /t/ or /k/ - checked syllables).

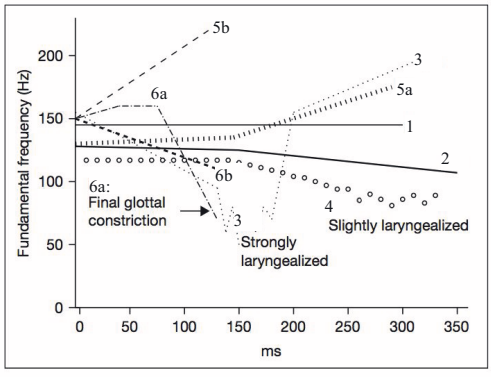

Figure 2: ***Schematic diagram of Hanoi Vietnamese tones*** **[16].**

Phonetically, tones 1, 2 and 5a are produced with modal voice: 1 is High-Level, 2 is Low-Falling, 5a is High-Rising. Tones 3 and 6a are glottalized: 6a has glottal constriction throughout, and is typically falling; 3 has medial glottal constriction and ends on a high fundamental frequency (F0) value. The experiment in [14] warrants the conclusion tones 5b and 6b are not glottalized (both produced in modal voice), either in final or non-final position. The work on oral flow [17] brings out a clear difference between these two sets of rhymes: tone 6a (drop tone in unchecked syllables) has low oral airflow; tone 5b and 6b have relatively high oral airflow, getting close to the range of breathy voice.

2.2 The ToBI transcription model

In ToBI model, the utterances are described by labels structured in tiers: the orthographic tier, the miscellaneous tier (for comments of all kinds), the break tier (which describes the utterance's phrasing) and, of course most importantly, the tone tier.

This section consists of two subsections of a short description of the individual elements of the ToBI tone inventory, which closely follows the example of the ToBI Annotation Conventions by [18]. The first subsection describes phrasal tones and pitch accents while the other one deals with prosodic phrasing.

2.2.1 Transcribing phrasing

Break indices represent a rating for the degree of juncture perceived between each pair of words and between the final word and the silence at the end of the utterance. They are to be marked after all words that have been transcribed in the orthographic tier. All junctures - including those after fragments and filled pauses - must be assigned an explicit break index value; there is no default juncture type. Values for the break index are chosen from the following set:

- 0: for cases of clear phonetic marks of clitic groups.
- 1: most phrase-medial word boundaries.
- 2: a strong disjuncture marked by a pause or virtual pause, but with no tonal marks; or a disjuncture that is weaker than expected at what is tonally a clear intermediate or full intonation phrase boundary.
- 3: intermediate intonation phrase boundary; i.e. marked by a single phrase tone affecting the region from the last pitch accent to the boundary.
- 4: full intonation phrase boundary; i.e. marked by a final boundary tone after the last phrase tone.

2.2.2 Transcribing intonation

The intonation is transcribed as a series of pitch accents and boundary tones each of which can be either low (L), or high (H). Accents are distinguished by appending a star (*), whereas tones are distinguished by appending either a percentage sign (%) or a minus sign (-), denoting boundary and phrase tones, respectively. By tagging individual syllables with these labels, it became possible to identify perceived prominences and major phrase boundaries by * and %, respectively, while the H and L portions of the labels described the shape of the

pitchtrack. The pitchtrack was further described by the use of the "!" diacritic to indicate downstepping, and the inclusion of the HiF0 label to mark the location of the peak F0 value in each major phrase.

Phrasal tones will be assigned at every intermediate or intonation phrase: L- or H- (phrase accent); L% or H% (final boundary tone) and %H (high initial boundary tone). Since intonation phrases are composed of one or more intermediate phrases plus a boundary tone, full intonation phrase boundaries will have two final tones: L-L%, L-H%, H-H% and H-L%. Pitch accent tones will be marked at every accented syllable. Lack of pitch accent assignment for a syllable will be interpreted as meaning that the syllable is NOT accented. The ToBI transcription allows for the five types of pitch accents: H* (peak accent), L* (low accent), L*+H (scooped accent), L+H* (rising peak accent) and H+!H* (a clear step down onto the accented syllable from a high pitch).

3. SYSTEM DESIGN

3.1 System Architecture

Our proposed architecture of an HMM-based TTS system for Vietnamese language is illustrated in Figure 3 [11]. There are three parts to this architecture: Natural language processing (NLP), Training, and Synthesis.

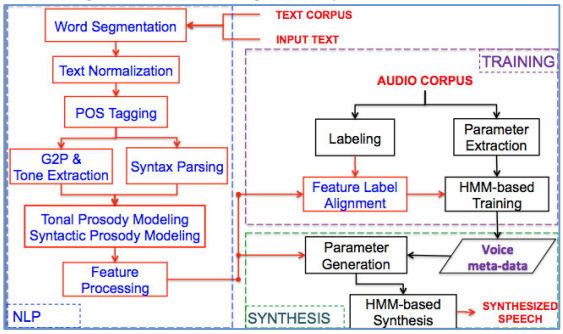

Figure 3: ***Architecture of the HMM-based TTS system for Vietnamese [11].***

There are seven modules in the ***NLP part***, which accepts text as input and finally produces context-based features to both the Training and Synthesis parts. These features include contextual factors in phoneme, syllable, word, phrase and utterance level. The ***Training part*** uses two main inputs to produce a trained voice using HMMs and EM algorithms: (i) Speech parameters including spectral (mel-cepstrum) and excitation parameters, which are extracted from the audio corpus and (ii) Contextual features (extracted from the text corpus) aligned with labels (automatic labeled from the audio corpus). In the ***Synthesis part***, context-based features are used to produce a sequence of speech parameters in such a way that its output probability for the HMM is maximized. High-quality synthesized speech is obtained using these speech parameters and a vocoder.

3.2 Design of Vietnamese features

Contextual features for Vietnamese are chosen in phoneme, syllable, word, phrase and utterance levels, based on Vietnamese phonetic and phonology in section 2. Moreover, we also refer to the work for English [19] and for Vietnamese [20] [21] to build our own feature set. We have more features on functions of phonemes in syllable structure, punctuation and some prosodic features, which will be presented in detail in Section 3.3 (which are formatted in *italic*). There are some slightly different on number of sublevels and relative positions of each level. In this work, following contextual features are taken into account:

- Phoneme level
  - {Two preceding, current, two succeeding} phoneme
  - The phoneme is onset or coda
  - Number of phonemes {from the beginning, to the end} in the current syllable to the current phoneme.
  - *Break indices of the current phoneme.*
- Syllable level
  - *Tone of {preceding, current, succeeding} syllable*
  - *Position type of the current syllable*
  - Number of phonemes in the {preceding, current, succeeding} syllable
  - Number of syllables {from the beginning, to the end} in the current word to the current syllable.
- Word level
  - *{Previous, next} punctuation in the current sentence*
  - *Part-of-Speech (POS) tags of the {preceding, current, succeeding} word*
  - *Number of words from the {preceding, succeeding} punctuation in the current sentence*
  - Number of {phonemes, syllables} in the {preceding, current, succeeding} word
  - Number of words {from the beginning, to the end} in the current phrase to the current word.
- Phrase level
  - Number of {syllables, words} in the {preceding, current, succeeding} phrase
  - Number of words {from the beginning, to the end} of the current utterance
  - Boundary endtone of the *{preceding, current, succeeding}* phrase.
- Utterance level
  - Number of {words, phrases} in the {preceding, current, succeeding} utterance
  - *Punctuation of the {preceding, current, succeeding} utterance*

3.3 Prosodic features for Vietnamese

3.3.1 Design of Vietnamese prosodic features

This subsection explains in detail for prosodic features presented in the previous subsection. *Tone of a syllable* can be one of 8 values represented for 8 tones: 1-4, 5a, 5b, 6a and 6b. *Position type of a syllable* can be "single" if there is only one syllable in the bearing word; "initial", "middle" or "last" corresponding to it's position in a multi-syllable word. *POS of a word* can be one of the list in the work [22]. *Punctuation* marks in the middle or at the end of sentence are ". , ; : ( ) " ' ? !".

3.3.2 Extraction rules of ToBI features for Vietnamese

In our experiment, intonation phrases are identified by punctuations in the middle of sentences, e.g. " ( ) " ' , ; :". Other break indices are identified by rules in Table 1.

For the intermediate phrase boundary (break indices 2, 3), we need to do more analysis and experiment to have a systematic rules.

Table 1: Phrasing rules

Break indice	*Boundary name*	*Rule*
0	Clitic boundary (within-word)	Between 2 consecutive phonemes in one word
1	Prosodic word boundary	Between 2 consecutive words
2, 3	Intermediate phrase boundary	N/A
4	Intonation phrase boundary	After a punctuation mark in the middle of the sentence
5	*Utterance boundary*	*At end of sentence, not at end of paragraph*
6	*Paragraph boundary*	*At the end of paragraph*

Rules for ToBI boundary tones for phrases in Table 2 are built from the ones for American English [23] with some adaptions for Vietnamese. This study for Vietnamese intonation in [21] discussed some works on the intonation of declarative and interrogative sentences. These works described in a qualitative way for these sentences mode. Declarative sentences are discussed with a falling intonation, F0 declination or "low speech", whereas interrogative sentences are said to be rising contour or "higher pitch". The significant difference between sentence modes would relate to average register, which is situated in the middle of the range for declaratives and towards the periphery for other sentence modes. Other work [22] confirmed that the F0 contour of the last syllable or the one of its second half tends to increase for questions.

Table 2: Intonation rules (boundary tone) for phrases

Position of boundary	*Rule*
End of declarative sentence	L-L%
End of exclamative sentence	L-H%
End of interrogative sentence	H-H%
End of a phrase, terminated by a punctuation mark in the middle of the sentence	H-L%

Three sentence modes are then experimented in this work: Declarative, exclamative and interrogative sentence (Table 2). In declarative mode, sentence-internal boundaries were labeled L-H% and sentence final boundaries were labeled L-L%. Interrogative sentences are transcribed with H-H% while exclamative sentences are labeled with L-H% pattern.

4. IMPLEMENTATION & EVALUATION

4.1 Implementation of VTed

We have built an HMM-based TTS system for Vietnamese, VTed, following the architecture in Figure 3, under the Mary TTS platform.

We adopted previous results [22][24][25] to build the NLP part. The *Prosody Modeling* module is built to extract prosodic features that are presented in the previous section. This enables an automatic training and synthesis process in our system. We used a 5-state left-to-right HMMs with single diagonal Gaussian output distributions. The training was automatically carried out with a corpus using about 92% from the 630 sentences of an existing corpus, *VNSpeechCorpus for speech synthesis*, while the remainding 8% were used in the evaluation phase. These sentences were recorded by a Vietnamese female broadcaster from Hanoi at 48 kHz and 16 bits per sample. The total duration of all sentences is ~37 minutes.

4.2 Subjective evaluations

The subjective evaluations included the assessment of general quality with MOS test with a natural speech reference, and the assessment of intonation model with preference test. Utterances are presented in random order to 18 subjects (9 females) for MOS test and 16 subjects (8 females) for preference test. All subjects are from the North of Vietnam, living for a long time in Hanoi. Participants were 20-35 years old and reported normal hearing and vision. Bars in graphs are presented with the mean values and confident intervals (i.e. interval estimate of a population parameter, calculated from observations, to indicate the reliability of an estimate).

4.2.1 Evaluation of general quality

Subjects were asked to score *"5-Excellent, 4-Good, 3-Fair, 2-Poor and 1-Bad"* their overall impression after listening to an utterance. There were 48 sentences in the test corpus (8% of the whole corpus). For the sake of comparison, this test was also carried out on our previous TTS system adopting non-uniformed unit-selection synthesis (NUS-HoaSung) [9] using the same training corpus (92% of the whole corpus).

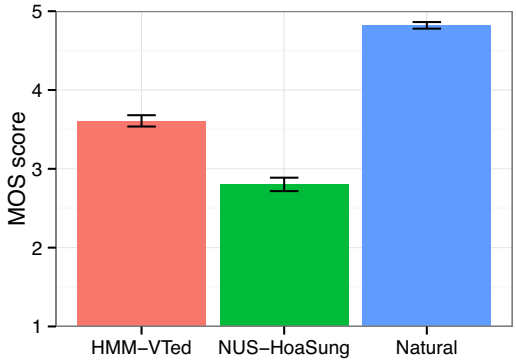

Figure 4: ***Results of quality in general (MOS Test).***

A two-factorial ANOVA was run on the results. The two factors were the TTS system (3 levels) and the Sentence (48 levels). All factors and their interactions have highly significant effect ($p<0.001$); meanwhile the TTS system factor alone explains an important part of the variance (partial $\eta^2=0.63$), while the Utterance factor and the interaction explain only about 0.15 each. A post-hoc Tukey test shows that each TTS system received significantly different mean scores. The experiment results plotted in Figure 4 show that the sound quality of VTed is rather good (0.81 point higher than HoaSung), but still clearly distinguishable from natural speech (1.21 point lower).

4.2.2 Evaluation of prosody model

Preference Test was done with 48 sentences in the MOS Test to test the influence of the ToBI boundary tone feature to the quality of the synthetic voice. There are two synthetic voices for each stimuli: (i) With-ToBI-boundary-tone: The voice synthesized by VTed trained with all proposed prosodic features, including ToBI boundary tone (ii) Without-ToBI-boundary-tone: The voice synthesized by VTed trained with all proposed prosodic features except the ToBI boundary tone.

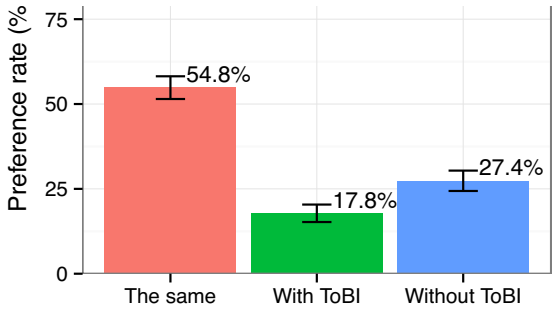

Figure 5: Preference rate of Experiment of ToBI Model.

To help subjects compare both systems, long sound files of both voices were split into 92 utterances with a length ranging between 5 up to 13 syllables. Subjects listened these 92 stimuli, composed of two utterances based on the two synthetic voices, separated by a "beep" sound. The order of the two voices in each pair and the order of utterances are presented randomly to the subjects. For the objective evaluations, we will look at the signal and give some analyses to discover the reasons.

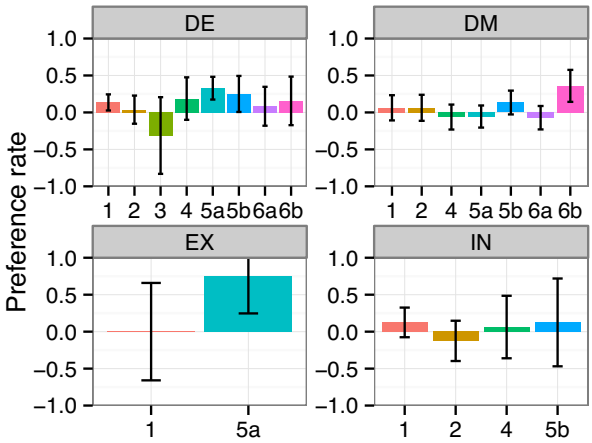

Figure 6: Preference Rate by Tones and Sentence Modes.

The experiment results plotted in Figure 5 show that subjects perceived the performances of both system as being "The same" in about 55% of the pairs, and the "Without-ToBI-boundary-tone" is preferred in 27%, while the "With-ToBI-boundary-tone" is preferred in 18% of the pairs – thus about a 10% preference for the "Without-ToBI-boundary-tone" voice. To further analyze the factors that may affect the perception of the synthetic voice's intonation – i.e. the sentence mode (DE, EX, IN or DM as in Table 2) and the tone of the last syllable (1-4, 5a, 5b, 6a or 6b) of utterance in each stimulus, a statistical analysis was run on the results, expressed in a three-point scale of preference. Figure 6 illustrates this preference rate for both sentence modes and tones, with this 3-point scale: "Without-ToBI-boundary-tone" is a positive preference mark (+1), while an answer "The-same" is neutral (0), and a preference for the "With-ToBI-boundary-tone" is a negative mark (-1) (this polarity for the scale was chosen after the observed preferences marked by listeners). Since declarative sentences are the most common in our corpus, there provide all tones for both phrase position: middle or end of the sentence; while only tones 1, 5a are presented in exclamative and tones 1, 2, 4, 5b in interrogative sentences.

A two-factorial ANOVA was run on the results to see if there was a difference in the use of this 3-point scale, according to the two factors Sentence modes (4 levels) and Tone of the last syllable (8 levels). The result of this analysis, presented in Table 3, shows that both factors and their interactions have significant effect ($p<0.001$). The two-way ANOVA indicates there is significant interaction between the effects of sentence mode and tone of last syllable on subject's perception.

Table 3: Rules Of Boundary Tone For Phrases

Anova	F value	Significance
Sentence mode	6.5372	0.00021681
Tone of last syllable	4.6968	0.00003188
SentenceMode:Tone	3.0753	0.00070529

If considering the impact of tones of the last syllable, for tones 2, 4 and 6a, most subjects don't find any difference between two voices; whereas the "Without-ToBI-boundary-tone" voice is preferred for tones 1, 5a, 5b and 6b. Conversely, the broken tone 3 is preferred in the "With-ToBI-boundary-tone" version, in spite of its sparseness in the corpus.

4.3 Objective evaluation

To have an explanation for the preference test results, we made some observations of the signal of sentences for wich the voice without ToBI features was preferred. We observed two main problems in the voice with ToBI features: (i) Discontinuity in spectrum and (ii) Unexpected voice quality. This may make subjects feel uncomfortable when hearing these sounds, and lead them to prefer the Without-ToBI voice.

Figure 7 shows an example of the discontinuity in spectrum for the "With-ToBI-boundary-tone" of a phrase "càng nhiều càng tốt" *(as much as possible)*, compared to the voice without ToBI features (at the top). The intonation pattern "L-L%" is applied for this declarative sentence, following the previous rule. However from section 2.1.3 we found that the F0 contour of the syllable bearing the rising tone (Tone 5a,b) normally raises from the beginning to the end of the syllable. But in this case the last syllable "tốt" /tot-5b/ of the phrase synthesized with ToBI model bearing the rising tone (Tone 5) seems to be flat and discontinuous.

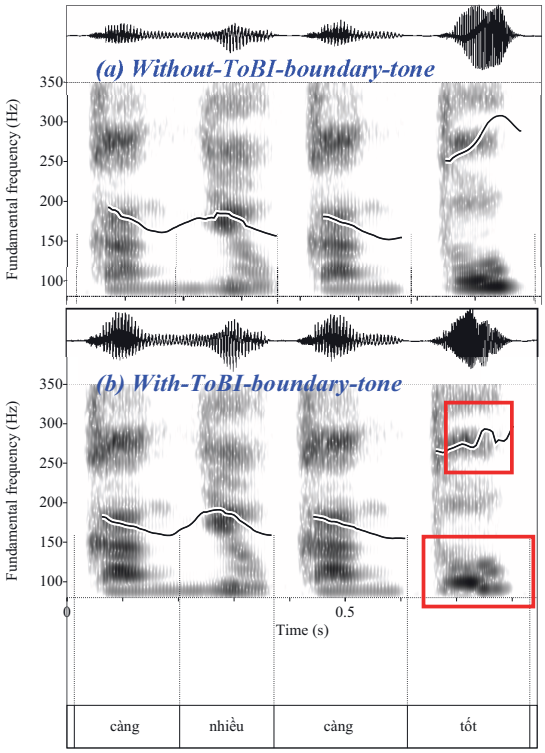

Figure 7: Discontinuity in spectrum and F0 (With ToBI) of "… Càng Nhiều Càng Tốt - /caŋ-1 ŋiew-2 caŋ-1 tot-5b/".

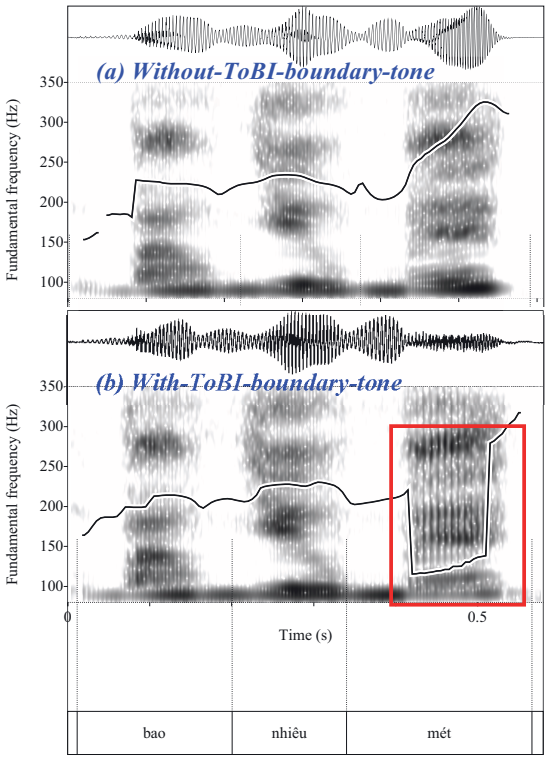

Figure 8: Wave form, Strectrogram and F0 Contour of "bao nhiêu mét - /baw-1 ŋiew-1 mɛt-5b/".

An example of the unexpected voice quality of the "With-ToBI-boundary-tone" is illustrated in the Figure 8. The sentence "Nhà này rộng bao nhiêu mét? *(How many meters is the house?)*" was applied the intonation pattern "H-H%" for an interrogative sentence. The F0 contour of the last syllable "mét" /mɛt-5b/ bearing the rising tone (originally high register) is traditionally raised, but in this case we found a phenomena of glottalization in the syllable. The F0 contour of this syllable looks like the F0 contour of the broken tone 3; meanwhile the F0 contour of this syllable for the Without-ToBI-boundary-tone maintains the traditional form of the tone 5b.

5. CONCLUSIONS AND DISCUSSIONS

This paper presented the design of prosodic features for training and synthesis an HMM-based TTS system for Vietnamese, VTed. The result of a MOS test showed that the general quality of synthetic voice is rather good, 1.21 point lower than the natural voice. A preference test was carried out to evaluate the effect of the ToBI boundary tone feature to VTed. About 55% of the voice trained with ToBI boundary tone are perceived as similar to another model without this features, while a 10% difference in favour of the without ToBI one is observed. Explanations may come from the fact that the F0 contour happens to be lowered or raised regardless of the tone of the last syllable. It raises two main problems in the synthetic voice: discontinuity in the spectrum and F0 and inadequate voice quality (e.g. unexpected glottalization).

These results showed the need for more efforts in intonation modeling for Vietnamese, which should take care of the lexical tones and other prosody cues of Vietnamese. In our work, there are also a lack of the break indices 2 and 3, which identify the intermediate phrase boundaries. We are now working on the interface between Vietnamese prosodic hierarchy and syntax to discover the systematic rules of break indices on syntax. In the generation model of F0 contour for questions from declarative sentences in [26], the whole contour of the declarative sentence is raised by *alpha* (normalized register ratio), and the contour of the last syllable is then raised by *beta* (increasing slope). However, this work was done in a limited corpus disregarding tone types of the last syllable and some types of simple questions. In conclusion, the prosodic system of Vietnamese seems to be too complex for being described by ToBI model, and it will need to take into account not only prosodic parameters but also the syllabic structure, the tone features and the syntactic structure.

6. REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] T. Yoshimura, "Simultaneous Modeling of Phonetic and Prosodic Parameters, and Characteristic Conversion for HMM-Based Text-To-Speech

Systems," Nagoya Institute of Technology, Japan, 2002.

[3] S. Chomphan and T. Kobayashi, "Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis," *Speech Communication*, vol. 50, no. 5, pp. 392–404, 2008.

[4] C.-C. Y. Hung-Yan Gu, "An HMM Based Pitch-Contour Generation Method for Mandarin Speech Synthesis.," *J. Inf. Sci. Eng.*, vol. 27, pp. 1561–1580, 2011.

[5] Y. Wang, J. Jia, and L. Cai, "Analysis of Chinese Interrogative Intonation and its Synthesis in HMM-Based Synthesis System," in *2011 International Conference on Internet Computing Information Services (ICICIS)*, 2011, pp. 343–346.

[6] A. Li, S. Pan, and J. Tao, "HMM-based speech synthesis with a flexible Mandarin stress adaptation model," in *2010 IEEE 10th International Conference on Signal Processing (ICSP)*, 2010, pp. 625–628.

[7] T. T. Do and T. Takara, "Precise tone generation for Vietnamese text-to-speech system," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, 2003, vol. 1, pp. I–504–I–507 vol.1.

[8] D. D. Tran and E. Castelli, "Generation of F0 contours for Vietnamese speech synthesis," in *Proceedings of the third International Conference on Communications and Electronics (ICCE)*, Nha Trang, Vietnam, 2010, pp. 158–162.

[9] V. T. Do, D. D. Tran, and T. T. T. Nguyen, "Non-uniform unit selection in Vietnamese speech synthesis," in *Proceedings of the Second Symposium on Information and Communication Technology*, Hanoi, Vietnam, 2011, pp. 165–171.

[10] K. Silverman, M. Beckman, and Pierrehumbert, "TOBI: A standard scheme for labeling prosody," in *Proceedings of the Second International Conference on Spoken Language Processing*, 1992.

[11] T. T. T. NGUYEN, C. ALESSANDRO, A. RILLIARD, and D. D. TRAN, "HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation," *INTERSPEECH 2013*, Lyon, France, Aug-2013.

[12] J. P. Kirby, "Vietnamese (Hanoi Vietnamese)," *Journal of the International Phonetic Association*, vol. 41, no. 03, pp. 381–392, 2011.

[13] L. C. Thompson, *A Vietnamese Reference Grammar*. University of Hawaii Press, 1987.

[14] A.-G. Haudricourt, "'The origin of the peculiarities of the Vietnamese alphabet' (Translated by Alexis Michaud)," *Mon-Khmer Studies*, vol. 39, pp. 89–104, 2010.

[15] Đoàn T. T., *Ngữ Âm Tiếng Việt (Vietnamese phonetics)*. Đại Học và Trung Học Chuyện Nghiệp, 1999.

[16] A. Michaud, "Final consonants and glottalization: new perspectives from Hanoi Vietnamese," *Phonetica*, vol. 61, no. 2–3, pp. 119–146, 2004.

[17] A. Michaud, T. Vu-Ngoc, A. Amelot, and B. Roubeau, "Nasal release, nasal finals and tonal contrasts in Hanoi Vietnamese: an aerodynamic experiment," *Mon-Khmer Studies*, vol. 36, pp. pp. 121–137, 2006.

[18] M. E. Beckman and J. Hirschberg, "The ToBI Annotation Conventions," Ohio State University, 1994.

[19] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings of the 2002 IEEE Workshop on Speech Synthesis*, California, USA, 2002, pp. 227–230.

[20] T. S. Phan, T. T. Vu, T. C. Duong, and C. M. Luong, "A study in Vietnamese statistical parametric speech synthesis base on HMM," *International Journal of Advances in Computer Science and Technology*, vol. 2, pp. 1–6, 2012.

[21] L. He, J. Yang, L. Zuo, and L. Kui, "A trainable Vietnamese speech synthesis system based on HMM," in *Proceedings of the International Conference on Electric Information and Control Engineering (ICEICE)*, Wuhan, China, 2011, pp. 3910–3913.

[22] H. P. Le, A. Roussanaly, T. M. H. Nguyen, and M. Rossignol, "An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts," in *Traitement Automatique des Langues Naturelles - TALN 2010*, Montreal, Canada, 2010.

[23] M. Jilka, G. Möhler, and G. Dogil, "Rules for the Generation of ToBI-based American English Intonation," *Speech Communication*, vol. 28, pp. 83–108, Jun. 1999.

[24] H. P. Le, T. M. H. Nguyen, A. Roussanaly, and T. V. Ho, *A Hybrid Approach to Word Segmentation of Vietnamese Texts*, vol. 5196. Springer-Verlag Berlin, Heidelberg ©2008, 2008.

[25] T. T. T. Nguyen, T. T. Pham, and D. D. Tran, "A method for Vietnamese text normalization to improve the quality of speech synthesis," in *Proceedings of the 2010 Symposium on Information and Communication Technology*, Hanoi, Vietnam, 2010, pp. 78–85.

[26] A.-T. Le, D.-D. Tran, and T.-T. T. Nguyen, "A Model of F0 Contour for Vietnamese Questions, Applied in Speech Synthesis," in *Proceedings of the Second Symposium on Information and Communication Technology*, Hanoi, Vietnam, 2011, pp. 172–178.