Modeling Vietnamese Speech Prosody: A Step-by-Step Approach Towards an Expressive Speech Synthesis System

Dang-Khoa Mac and Do-Dat Tran(✉)

International Research Institute MICA, HUST-CNRS/UMI 2954-Grenoble INP, Hanoi Vietnam
{dang-khoa.mac,do-dat.tran}@mica.edu.vn

Abstract. Attempts to add expressivity to synthesized speech is one of the main strategies in speech technologies. This paper summarizes our researches on modeling Vietnamese prosody, with the goal of improving naturalness of synthesized speech in Vietnamese, as well as integrating expressivities (i.e. emotion/attitude). Based on the concept of "rendez-vous" between linguistic levels and prosodic functions, the prosody of utterance is proposed to be decomposed into several components. Therefore, each component is step by step modeled by an independent model: a dynamic linear segment model for tones, a relative registers model for F0 level of syllable, a rule-based approach for phrasing modeling and a F0 stylization modeling for the expressive function. All proposed models were integrated in speech Text-to-speech systems and also were evaluated by perception experiments.

Keywords: Text-to-speech · Vietnamese · Prosody modeling · Tones · Phrasing · Attitude · Expressive speech

1 Introduction

Speech is one of the fundamental human behavior events that simultaneously conveys linguistic information as well as the speaker's affective variability (e.g., mental, intentional, attitudinal, emotional states). Attempts to add expressivity to synthesized speech have existed for more than a decade. For a tonal language like Vietnamese, acoustic parameters implied in the linguistic and affective functions of prosody (typically F0, intensity, timing) also play an important role at the phonemic level for lexical access. Moreover, the Vietnamese tones can imply some voice quality cues that have been shown to be used in the morphology of some attitudes (and emotions) in other languages [1].

The main task of the prosody generator of a TTS system is to provide an acoustic representation of prosody from linguistic information. It is well known that, for a tonal language like Vietnamese, Chinese, the fundamental frequency (F0) contour of a sentence always consists of local tonal components and the intonation of the sentence. Thus the variation of the F0 in the sentences for tonal languages seems more complicated than that of non-tonal languages. The Vietnamese prosodic contour could be generated

X.-L. Li et al. (Eds.): PAKDD 2015, LNCS 9441, pp. 273–287, 2015.
DOI: 10.1007/978-3-319-25660-3_23

automatically by using the Fujisaki model [2, 3] or a linear F0 model combined with relative registers [4]. But there is no model that can generate the prosodic contours of tones combined with expressive prosodic contours.

According to the prosodic model proposed in [6], the intonation is considered as a result of superimposed and independent prototypical gestures belonging to hierarchical linguistic levels: sentence, clause, group, sub-group etc. That concept is called the "rendez-vous" between linguistic levels and prosodic functions of utterance [6, 7]. This theoretical model allows the generation of complex prosodic contours using a superposition process directed by functions. It was applied in the speech synthesis for 3 modalities (declaration, question, surprise), in the automatic generation of 6 expressive prosodic attitudes for French [7] and in the prosody generation of tonal language such as Chinese [8].

Our approach to Vietnamese expressive speech production consists of applying the "rendez-vous" concept above to combine the variation of many prosodic functions such as tone, phrasing, sentences modality and expressivities (attitude/emotion). Each component is modeled by three separated model:

1. a dynamic linear segment model for tones;
2. a relative registers model for F0 height of syllable;
3. a rule-based approach for phrasing modeling and a F0 stylization modeling for the expressive function.

The next sections will present step by step our proposal models for these prosodic components. All proposed models were developed separately and evaluated in the speech different Text-to-speech system.

2 Modeling the Prosodic Contour of Vietnamese Tones in Continuous Speech

In order to build an intonation model for the Vietnamese language, we begun our studies with modeling F0 contours of tones of isolated words.

2.1 Overview

The Vietnamese language has 6 tones as shown in Fig. 1: level (1), falling (2), broken (3), curve (4), rising (5) and drop (6). Tone 5b and 6b correspond to tone 5 and 6 on a syllable ended by a stop consonant. Moreover the Vietnamese tonal system can employ some production of voice quality, within F0. That is the co-occurrence of glottalization during the production of tone 3 and tone 6: tone 3 is accompanied with harsh voice quality due to a glottal stop (or a rapid series of glottal stops) around the middle of the vowel; tone 6 has the same kind of harsh voice quality as tone 3; however, it is distinguished by dropping very sharply and it is almost immediately cut off by a strong glottal stop [9].

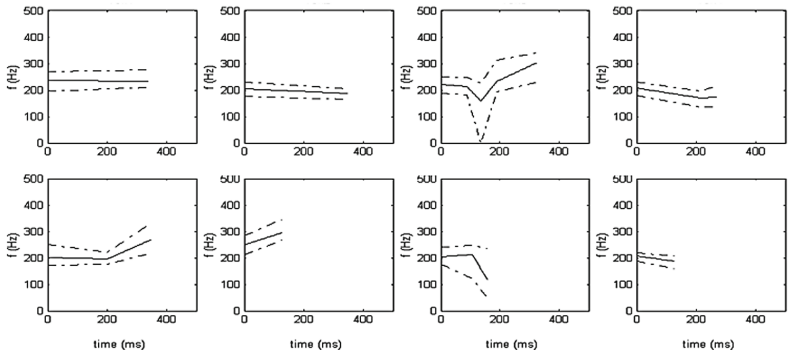


Fig. 1. Examples of contours of 8 Vietnamese tone representations from a female subject [9]. From the left to right, top to bottom: tone 1, 2, 3, 4, 5, 5b, 6, 6b.

In the continuous speech, the F0 contour of the Vietnamese tones with the influence of tonal coarticulation effect can be described by the linear F0 model (as in Fig. 2) combined with relative registers of Vietnamese tone, as proposed in. This method is used in our work to generate the prosodic contour in the syllable level, which correspond to the tonal function of prosody.

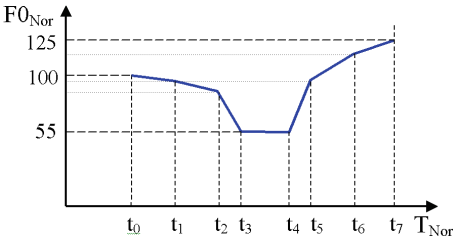


Fig. 2. The normalized linear contour of tone 3

2.2 Linear Segment Model of Vietnamese Tone

2.2.1 Vietnamese Tones Model in Isolated Mode

Several models of generation of F0 contour for Vietnamese tones were presented in the work of [2, 3, 10]. In study of [10] the average F0 contours are used as contour templates for tones. The speech synthesis systems presented in the work of [11, 12] employ the Fujisaki model to generate the F0 contours of 6 tones, but the systems meet difficulties in generating F0 contours of Tones 3 and 6 caused by phenomena of glottalization during their pronunciation. In addition, the three researches – were based on analysis of variation of F0 in voiced syllables, and on the hypothesis that the tone has an effect on the entire syllable. Under these conditions, the patterns of tones concern only voiced syllables. The results of these three studies did not indicate whether these patterns could be used for all types of Vietnamese syllables, particularly for syllables that begin with unvoiced consonants.

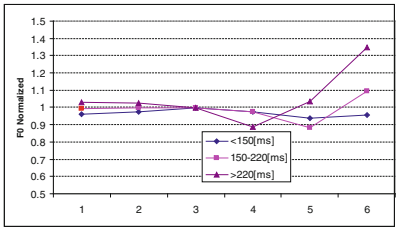


Fig. 3. Three variants of tone 3 in respect to the duration of the syllable

To build the F0 contour patterns for Vietnamese tones, we apply the results of a study [13] about the influence of tone on the syllable. The results showed that the initial consonant of the syllable does not carry the information of the tone, and the Vietnamese tone affects only the final part of the syllable.

Based on the results of [14] about the shapes of F0 contour of the six Vietnamese tones, we got average values of F0 contours of the six Vietnamese tones and then we constructed linear F0 contour models for the Vietnamese tones (Fig. 2 shows a linear contour model for Tone 3). These models only describe the F0 evolution of a final part of the Vietnamese syllable, this is different from studies [2, 10] which model the F0 contour of tones on the whole syllable.

The process of tone modeling composed of three phases:

- Calculating average values of F0 at specific points, the number of points N is dependent on the complexity of tone;
- Normalization of average values, the mean values are divided by the average value of the initial point;

$$F0_i^{Nor} = F0_i / F0_1 \quad i = 1 \ldots N \tag{4.1}$$

- Connecting the points by the lines, the value of F0 of the middle points are calculated by the equation:

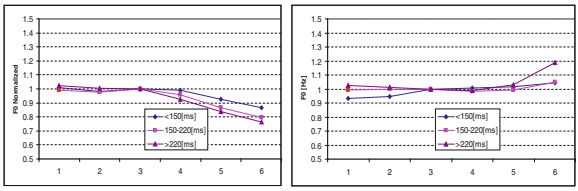
$$F0\,(t) = F0_i^{Nor} + \alpha_i * t \qquad \left(t = t_i - t_{i+1}\right) \tag{4.2}$$

To generate the F0 contour for a tone, the normalized F0 contour is firstly calculated, and then normalized values are multiplied by a specific factor, for example the average value of the fundamental frequency of the speaker.

The detail information on the F0 patterns and results of the perception tests is presented in [15].

2.2.2 Dynamic Model of Vietnamese Tones in Continuous Speech

Figure 3 presents three variants of Tone 3, which depend on the length of the syllable: it is easy to see that these three variants differ greatly from one another. In order to describe more clearly these differences, and also to compare with other tones, we normalized the F0 contours of the three variants (Fig. 4). The F0 value of 3th point was chosen as the reference value because of its stability.



(a) Tone 4 (b) Tone 5

Fig. 4. The normalized F0 contours of (a) tone 4 and (b) tone 5 with the different durations of the syllable

The F0 contours of variants of the five remaining tones were also normalized. Concerning Tones 4 and 5, three F0 contours change in a larger range (Fig. 4), they can be classified into two groups, and each group is represented by an only form.

The same process of tone modeling in isolated mode was carried out for Vietnamese tones in dynamic mode. But in this case, each tone can have more than one model; it depends on the duration of syllable carrying tone.

2.3 Generation of F0 Contour in Continuous Speech

Based on results from on variation of Vietnamese intonation in continuous speech, we believe that it is necessary to consider the four following factors when analyzing and modeling the contour of F0 an utterance in Vietnamese:

- Tones that make up the statement;
- Register for each tone;
- The influence of the phenomena of tonal coarticulation;
- The duration of the syllable;

A method for the generation of the F0 contours for Vietnamese text-to-speech based on the assumption: "*The tonal coarticluation effect takes place from begin to the end of each phrase, and after a pause it is cancelled*".

Suppose that a phrase of N syllables ($S_1\, S_2 \ldots S_N$) will be synthesized. The F0 contour of the phrase is obtained through 3 steps (Fig. 5).

Step 1: The register of all the syllables in the phrase is calculated, based on the phonetic information and position of syllable in the sentence

Step 2: Generation of the tonal contour for each syllable by using tone patterns. The tonal contour is then placed on the register contour.

Step 3: The F0 contour of the phrase is smoothed. For smoothing the discontinuity of the F0 contour, the transition pattern *Exclusive Carry-over* is applied.

And finally the normalized contour of F0 is scaled with a specific factor. For example, in the Fig. 5b, the value of factor equals to the initial F0 value (250 Hz) of the original phrase.

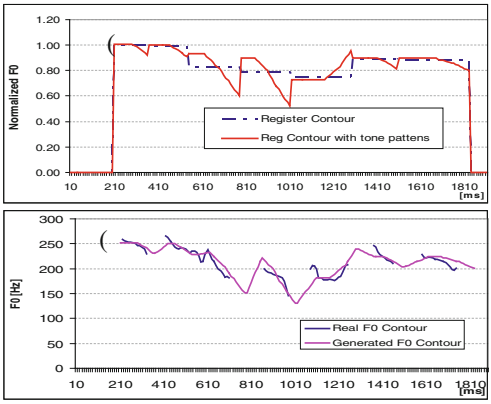


Fig. 5. (a) Generated register contour (dashed line) and the superimposed tone patterns. (b) F0 contour generated by the proposed model and the F0 contour of target speech.

2.4 Evaluation

In order to evaluate the performance of the proposed model on the natural characteristic of synthetic phrase, a perceptual test based on MOS test was performed. Thirty sentences were selected and re-synthesized in different ways. For each phrase, only the contour of F0 is manipulated by applying TD-PSOLA algorithm to have 4 variants. Thus, a speech corpus which contains 5 groups of 30 sentences was built:

- Group 1 contains 30 natural sentences.
- Group 2 includes 30 re-synthetic sentences. Their contour of F0 is generated by applying all of 3 steps the proposed method.
- Group 3 contains 30 re-synthetic sentences whose contour of F0 is generated by applying the first and second steps.
- Group 4 includes 30 re-synthetic sentences whose contour of F0 is generated by applying the second and the third steps. In these sentences, the relative registers of all tones equal 1.
- Group 5 contains 30 re-synthetic sentences whose F0 contour is generated by applying only the tone patterns (step 2). Like the sentences of group 4, the relative registers of all tones equal 1.

Twenty persons participated into the test. The listeners were asked to rate the natural quality of each perceptual sentence on a scale 1–5, where 1 is bad and 5 is completely natural.

Figure 6 presents the naturalness scores of the five groups. Group 1 which contains the natural sentences has the highest score; this is a predictable result. The next is Group 2 whose sentences are re-synthesized by applying all 3 steps. This is a good result for an intonation model. The lowest score group is belong to Group 5, which applies only the tone patterns. By comparison the score between Group 2 and Group 4, and the score between Group 3 and Group 5, we found that, by applying the relative tone register for

the generation of the F0 contour, the naturalness score of synthetic sentences of Group 2 and 3 (3.84 and 3.72) is significantly higher than that of Group 4 and 5 (3.11 and 3.10).

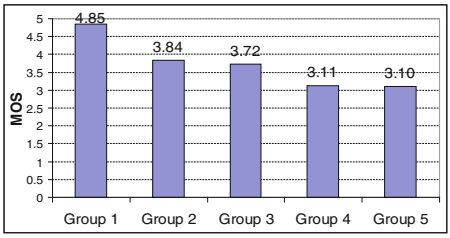


Fig. 6. MOS test result

With respect to the influence of the smoothing F0 contour on the naturalness of re-synthetic sentences, we can see a difference only in the sentences which have a quite natural quality. This is presented clearly when comparing the scores between Group 2 and 3 (3.84 vs. 3.72), and between Group 4 and 5 (3.11 vs. 3.10). Therefore, the obtained results show that, for a tonal language like Vietnamese, the relative tonal register is an important parameter for generation of F0 contour.

3 Vietnamese Phrasing Modeling

Phrasing modeling plays an important role in improving the naturalness for speech synthesis. Many researchers have been working on prosodic structure generation for Chinese [16, 17], pause/break modeling for French [18], German [19], Russian [20] or modeling style specific break [21, 22]. They may use rules or machine learning with lexical information (e.g. POS tagger) or contextual length. However, to the best of our knowledge, there is no such work for Vietnamese, a tonal language. It is believed that there is an interface between syntax and prosodic structure [23–27]. Recently, much effort has been devoted for Vietnamese syntax parsing with some has proven fruitful results [28–30].

3.1 Corpus Preparation

Resources of Vietnamese language is messy and lacking of unified and big-enough corpus, especially for speech processing [31]. To have a preliminary experiment for prosodic phrasing modeling, we adopted the existing corpus, "VNSpeechCorpus for speech synthesis" [32], for analyzing. In this corpus, there are 630 sentences that are recorded by a Vietnamese female broadcaster from Hanoi at 48 kHz and 16 bps (~37 min). Audio files in this corpus are transcribed, time-aligned at the syllable level, and annotated for perceived pauses. Text files are parsed and represented with syntax trees in XML format. These tasks are semi-automatically executed.

3.2 Proposal Syntactic Rules

From the corpus, we discover two types of rules: one between two constituents in phrase structure grammar and the other between two dependents in dependency grammar. Hypotheses are proposed for constituency syntactic rules (Table 1) and for dependency syntactic rules (Table 2) [33]. The IP (Intonation Phrase) boundaries are put for hypotheses either if the left constituent is or contains a clause (HC1, HC2) or both left and right dependents are predicates (HD1) or head elements (HD2). Other cases are made decision based on the syntactic information or number of syllables in the left or right element.

Table 1. Constituency rules and intermediate boundaries

#	Intermediate boundaries	Left constituent	Right constituent
HC1	*IP*	a SBAR or a constituent whose child is a clause	any constituent
HC2		a S >= 6 syllables	any constituent
HC3	*PhP*	any phrase >= 7 syllables	a phrase >=4 syllables
HC4		any phrase	a C following by a constituent >= 5 syllables
HC5		a PP >= 3 syllables	a C or AP/NP/VP
HC6		a S having 3 to 5 syllables or C following by S	any constituent
HC7		a C 'răng' whose parent is a SBAR	any constituent

Table 2. Dependency rules and intermediate boundaries

#	Boundary	Left dependent	Right dependent
HD1	*IP*	a predicate	a predicate
HD2		a H element >= 4 syllables	a head element
HD3		an adjunct >= 3 syllables	any dependent that is a phrase
HD4	*C*	a H element: 2-3 syllables	a head element
HD5		an adjunct having 2-3 syllables	a subject >= 2 syllables

3.3 Evaluation

Based on the proposed model of prosodic phrasing and analysis results on the speech corpus (presented in detail [33]), two new prosodic features are introduced: break levels and syllable position relative to phrase. Table 3 presents our proposed break levels and syllable relative positions used as training features for an HMM-based TTS. The break levels "0", "1", "5" and "6" are easily identified by POS tags or punctuation marks at the end of sentence whereas others ("2", "3", "4") need syntactic rules for prediction. Syllable positions are distinguished for the boundaries above the Word (1) and above the Intonation phrase boundaries ("2").

Table 3. Break levels as training features

Break level	Syllable position	Prosodic hierarchy	Rule
0	0	Within word	Between 2 consecutive phonemes in one word
1	0	Word	Between 2 consecutive words
2	1	Clitic group	HD4, HD5
3	1	Phonological phrase	HC3, HC4, HC5, HC6, HC7
4	1	Intonation phrase	After a punctuation mark in the middle of the sentence or HC1, HC2, HD1, HD2, HD3
5	2	Utterance boundary	After punctuation marks at end of sentence, not of paragraph
6	2	Paragraph boundary	At the end of paragraph

For evaluations, the MOS test was carried out with two versions of TTS system and a natural speech reference, presented in random order. 19 subjects (8 females) participated in the tests. 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. In the test corpus, 40 sentences are chosen so that each sentence covers only one hypothesis for ease of analysis. There are three to four examples designed for each hypothesis.

Subjects were asked to score "5-Excellent, *4-Good, 3-Fair, 2-Poor and 1-Bad*" for the naturalness after listening to an utterance. The experimental results illustrated in Fig. 7 show an increase of 0.35 on a 5 point MOS scale, for the new prosodic informed system (3.96/5) compared to the previous TTS system (3.61/5).

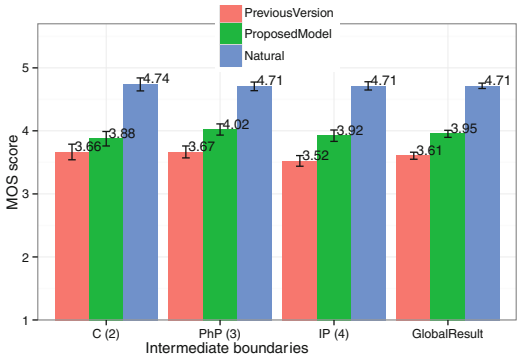


Fig. 7. Results of naturalness using MOS test.

4 Modeling Prosody of Attitude in Vietnamese

As mention above, our approach to Vietnamese expressive speech production consists of applying the "rendez-vous" concept in order to combine the local variation of tones and the global prosodic contours of attitude.

4.1 Vietnamese Attitude Corpus

Our first work is the construction of first corpus for Vietnamese attitudes. This corpus was not only constructed to be used in speech synthesis, but also to conduct fundamental studies on Vietnamese social affects. In the face-to-face interaction, attitudes are expressed within the multimodality of speech such as speech, face, gestures, etc. Thus this corpus was done not only in audio modality but also in visual modality, in order to investigate the relative contribution of audio and visual information in the generation and perception of Vietnamese attitude.

Based on research on attitudes in Vietnamese and other languages [34, 35], 16 attitudes have been represented for Vietnamese in our corpus (Table 4). To observe the effects of tone and tonal co-articulation on attitudinal expression, the corpus contains 8 sentences of one-syllable length, corresponding to the 8 types of Vietnamese tone, and 72 sentences of two-syllable length, which correspond to all combinations of two tones among the 8 Vietnamese tones. The remainder of the corpus is based on 45 sentences of 3- to 8-syllable length and systematically varied in their syntactic structure: single word, nominal group, verbal group and a simple structure "subject-verb-object". That means that the corpus is built from 125 sentences without specific affective meaning produced with all the 16 attitudes and balanced in terms of tone position. These sentences were recorded (both audio and video, but only audio is focused in this paper) by one male speaker native of the Hanoi dialect (standard pronunciation). The whole corpus thus contained 2000 sentences corresponding to more than 90 min of signal after post-processing.

Table 4. 16 selected Vietnamese attitudes, with their abbreviations

Declaration	DEC	Irritation	IRR
Interrogation	INT	Sarcastic irony	SAR
Exclamation of neutral surprise	EXo	Scorn	SCO
Exclamation of positive surprise	EXp	Politeness	POL
Exclamation of negative surprise	EXn	Admiration	ADM
Obviousness	OBV	Infant-directed speech	IDS
Doubt-incredulity	DOU	Seduction	SED
Authority	AUT	Colloquial	COL

4.2 Modeling

The prosodic contour of attitude represents the attitudinal function of prosody and it corresponds to the sentence level. According to [6], the forms of these contours are independent of others linguistic factors (syntax, tone) and depend only on the type of attitude. Therefore, we propose that the form of the prosodic contour of attitudes can be obtained by the mean value of prosodic contour of the neutral-tone sentences (all syllable produced with tone 1).

Figure 8 shows an example of the mean F0 contours of neutral-tone sentences with the length from 1 to 8 syllables. In observation the mean value of F0 contours, duration and intensity of all attitudes, we found that for each attitude, the F0 contour remains a common form when the number of syllables in the utterance increases. This common form can be divided into three parts: initial, middle and final part. The initial and final parts cover typically one or two syllables. The difference between F0 contours of 16 attitudes are mainly represented in these two parts. For all attitudes, the middle part is stable and can be simply represented by a line connecting the initial and the final part. For the duration and intensity, the differences between 16 attitudes are also mainly characterized by the duration and the mean intensity of the first and the last syllable.

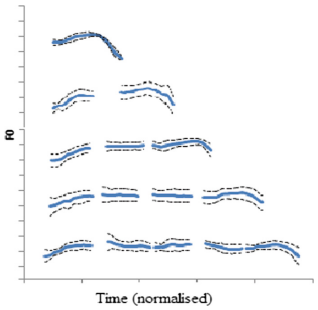


Fig. 8. Mean and deviation of F0 contours of 1–5 syllables sentences uttered with the attitude authority.

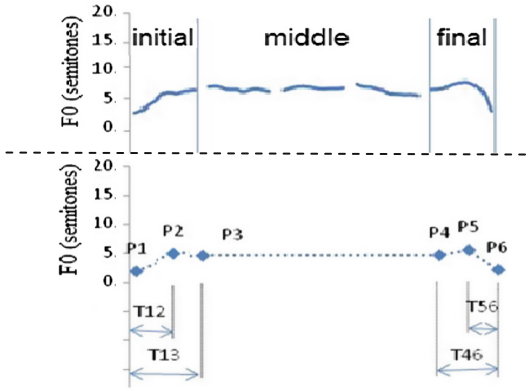


Fig. 9. An example of stylization F0 contour of attitude

The description given above enables us to stylize the prosody of attitudes when the number of syllables in the utterance increases:

- The F0 contour is stylized by 6 points as in Fig. 9. The mean values of 6 point and the relative distance between them represent the common form of F0 contour for each attitude.

- The duration and intensity of each attitude is characterized by the mean value of the first and the last syllable.

4.3 Perceptual Evaluation

An experiment was designed to perceptually evaluate the predicted prosody of attitudes, generated with the proposal model.

As mentioned above, in the face-to-face interaction, attitudes are expressed within the multimodality: audio and visual information. In this experiment, we aim to evaluate our prosodic model on the attitudes which are well transferred by the audio information. Using the result of the perception test on 16 attitudes with both of audio and visual modality (presented in [36]), we chose 4 attitudes well recognized with audio information for this experiment, they are: *Declaration, Exclamation of neutral surprise, Authority* and *Sarcastic irony*.

Four sentences (with tone and non-tone) from 3 to 8 syllables are used for this experiment. Using these sentences, the synthetic utterances corresponding to 4 selected attitudes above are generated generation with the speech synthesis system developed by the Institute MICA [5].

The prosody synthetic utterances (with 4 attitudes) are predicted by using the proposed model. The 32 synthetic utterances (4 attitudes, 4 sentences, 2 methods) above are then used in a perceptual test in order to examine whether the listeners can indicate the attitudes of synthetic utterances or not. Twenty Vietnamese listeners participated in this experiment. All subjects listened to each stimulus only one time.

Figure 10 presents the mean recognition rates of synthetic utterances (with tone and without tone) generated by re-synthesis method and by the MICA speech synthesis system. Overall, for both type of synthetic utterances and both type of sentence, the recognition rates are over 60 %. The sentences without tone are better recognized than the sentence with tone. That means that the local perturbation by tones increases the complexity of the global cues of prosody of the sentence. The perception result on the utterances generated by re-synthesis method is slightly better than on the utterances generated by MICA speech synthesis system.

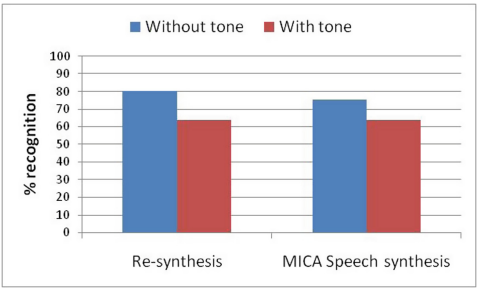


Fig. 10. The recognition rate (%) of synthetic utterances generated by re-synthesis method and by MICA speech synthesis system.

Figure 11 shows the recognition rates for 4 attitudes with difference lengths of sentence. Except in the case of Authority, the length of sentence shows no affect on the perception of attitude. The attitudes Declaration and Sarcastic irony have very good result (recognition rate >90 %). The attitude Authority has the lowest recognition rate (from 30 to 60 %).

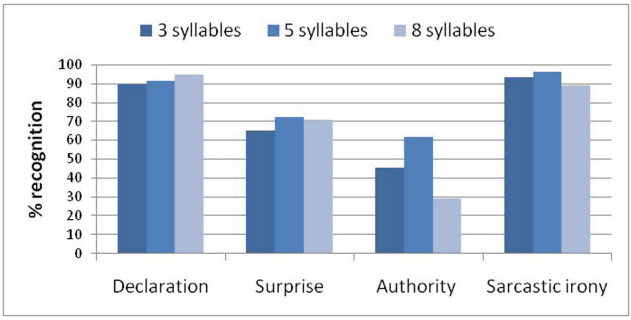


Fig. 11. The recognition rate (%) of synthetic utterances with four attitudes.

5 Conclusions and Perspectives

This paper presents our preliminary attempt of modeling the completed prosody of Vietnamese speech. Based on the concept of superposition the prosodic contour, a prosodic model was proposed to encode the attitudinal function of prosody for Vietnamese attitudes. This model was applied in generation the prosody of attitudes in Vietnamese. The predicted prosody of attitudes using this model was well recognized in the perception experiment. This result shows us the ability of applying the proposed model in generation the prosody of attitude for the tonal language such as Vietnamese. With this result, the hypothesis of global prosodic contours encoding speaker attitudes is also verified.

However, this work concerns only with the three basic parameters of prosody (F0, duration, intensity). The future work will also have to analyze the role of voice quality in the production and perception of attitudes, in order to characterize the voice quality of attitudes and to be applied in expressive speech synthesis for Vietnamese.

Acknowledgment. We would like to thank Mrs. NGUYEN Thi Thu Trang for her contributions in the frame work of the paper and of the research group.

References

1. Scherer, K.R., Ellgring, H.: Multimodal expression of emotion: affect programs or componential appraisal patterns? Emotion **7**(1), 158 (2007)
2. Nguyen, D.T., Luong, C.M., Vu, B.K., Mixdorff, H., Ngo, H.H.: Fujisaki model based F0 contours in vietnamese TTS. In: INTERSPEECH (2004)

3. Fujisaki, H., Gu, W.: Phonological representation of tone systems of some tone languages based on the command-response model for F0 contour generation. In: Tonal Aspects of Languages (2006)
4. Do Dat, T., Castelli, E., Hung, L.X., Serignat, J.-F., Van Loan, T.: Linear F0 contour model for Vietnamese tones and Vietnamese syllable synthesis with TD-PSOLA. In: Second International Symposium on Tonal Aspects of Languages (2006)
5. Trần, Đ.Đ.: Synthèse de la parole à partir du texte en langue Vietnamienne. INPG, Grenoble (2007)
6. Aubergé, V.: A gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP. In: International Conference on Speech Prosody 2002 (2002)
7. Morlec, Y., Bailly, G., Aubergé, V.: Generating the prosody of attitudes. In: Intonation: Theory, Models and Applications (1997)
8. Chen, G.-P., Bailly, G., Liu, Q.-F., Wang, R.-H.: A superposed prosodic model for Chinese text-to-speech synthesis. In: 2004 International Symposium on Chinese Spoken Language Processing, pp. 177–180 (2004)
9. Yên, P.T.N., Castelli, E., Cuong, N.Q.: Gabarits des tons vietnamiens. In: JEP 2002, Journées d'Etude Sur Parole XXIV, Nancy, France, pp 23–26 (2002)
10. Do, T.T., Takara, T.: Vietnamese text-to-speech system with precise tone generation. Acoust. Sci. Technol. **25**(5), 347–353 (2004)
11. Mixdorff, H., Nguyen, B.H., Fujisaki, H., Luong, C.M.: Quantitative analysis and synthesis of syllabic tones in Vietnamese. In: EuroSpeech2003, Geneva, pp. 177–180 (2003)
12. Fujisakia, H., Gu, W.: Phonological representation of tone systems of some tone languages based on the command-response model for F0 contour generation. In: TAL2006, pp. 59–62 (2006)
13. Trần, Đ.Đ., Castelli, E., Serignat, J.-F., Trinh, V.L., Le, X.H.: Influence of F0 on Vietnamese syllable perception. Presented at the Interspeech 2005, Lisbon, Portugal, pp. 1697–1700 (2005)
14. Nguyen, Q.C.: Reconnaissance de la parole en langue Vietnamienne. Ph.D. thesis, INP-Grenoble, Grenoble, France (2002)
15. Trần, Đ.Đ., Castelli, E., Lê, X.H., Segrinat, J.F., Văn Loan, T.: Linear F0 contour model for Vietnamese tones and vietnamese syllable synthesis with TD-PSOLA. In: TAL2006, France, pp. 103–107 (2006)
16. Chou, F.-C., Tseng, C.Y., Lee, L.-S.: Automatic generation of prosodic structure for high quality Mandarin speech synthesis. In: ICSLP (1996)
17. Tao, J., Dong, H., Zhao, S.: Rule learning based Chinese prosodic phrase prediction. In: 2003 International Conference on Natural Language Processing and Knowledge Engineering. Proceedings, pp. 425–432 (2003)
18. Doukhan, D., Rilliard, A., Rosset, S., d' Alessandro, C.: Modelling pause duration as a function of contextual length. In: INTERSPEECH (2012)
19. Apel, J., Neubarth, F., Pirker, H., Trost, H.: Have a break! Modelling pauses in German speech. In: KONVENS (2004)
20. Chistikov, P., Khomitsevich, O.: Improving prosodic break detection in a Russian TTS system. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS, vol. 8113, pp. 181–188. Springer, Heidelberg (2013)
21. Jokisch, O., Kruschke, H., Hoffmann, R.: Prosodic reading style simulation for text-to-speech synthesis. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 426–432. Springer, Heidelberg (2005)
22. Parlikar, A.: Style-Specific Phrasing in Speech Synthesis. Carnegie Mellon University, Pittsburgh (2013)

23. Selkirk, E.O.: On Prosodic Structure and Its Relation to Syntactic Structure. Indiana University Linguistics Club, Bloomington (1980)
24. Selkirk, E.: The syntax-phonology interface. In: Goldsmith, J., Riggle, J., Yu, A.C.L. (eds.) The Handbook of Phonological Theory, pp. 435–484. Wiley, New York (2011)
25. Nespor, M., Vogel, I.: Prosodic structure above the word. In: Cutler, D.A., Ladd, D.D.R. (eds.) Prosody: Models and Measurements, pp. 123–140. Springer, Berlin Heidelberg (1983)
26. Hayes, B.: The prosodic hierarchy in meter. Phon. Phonol. **1**, 201–260 (1989)
27. Dehé, N., Feldhausen, I., Ishihara, S.: The prosody–syntax interface: focus, phrasing, language evolution. Lingua **121**(13), 1863–1869 (2011)
28. Viet, H.A., Thu, D.T.P., Thang, H.Q.: Vietnamese parsing applying the PCFG model. In: Proceedings of the Second Asia Pacific International Conference on Information Science and Technology, Vietnam (2007)
29. Nguyen, P.-T., Vu, X.-L., Nguyen, T.-M.-H., Nguyen, V.-H., Le, H.-P.: Building a large syntactically-annotated corpus of Vietnamese. In: Proceedings of the Third Linguistic Annotation Workshop, Suntec, Singapore, pp. 182–185 (2009)
30. Le, A.-C., Nguyen, P.-T., Vuong, H.-T., Pham, M.-T., Ho, T.-B.: An experimental study on lexicalized statistical parsing for Vietnamese. In: Proceedings of the 2009 International Conference on Knowledge and Systems Engineering, Hanoi, Vietnam, pp. 162–167 (2009)
31. Le, V.-B., Besacier, L.: Automatic speech recognition for under-resourced languages: application to Vietnamese language. IEEE Trans. Audio Speech Lang. Process. **17**(8), 1471–1482 (2009)
32. Tran, D.D., Castelli, E.: Generation of F0 contours for Vietnamese speech synthesis. In: Proceedings of the third International Conference on Communications and Electronics (ICCE), Nha Trang, Vietnam, pp. 158–162 (2010)
33. Trang, N.T.T., Rilliard, A., Trần, Đ.Đ., D'Alessandro, C.: Prosodic phrasing modeling for Vietnamese TTS using syntactic information. In: INTERSPEECH 2014, Singapore, pp. 2332–2336 (2014)
34. Le Thi, X.: Etude contrastive de l'intonation expressive en français et en vietnamien. Ph.D. thesis, Université Paris 3, Paris, France (1989)
35. Shochi, T., Aubergé, V., Rilliard, A.: How prosodic attitudes can be false friends: Japanese vs. French social affects. In: Speech Prosody, Dresden, pp. 692–696 (2006)
36. Mac, D.-K., Aubergé, V., Rilliard, A., Castelli, E.: Audio-visual prosody of social attitudes in Vietnamese: building and evaluating a tones balanced corpus. In: Tenth Annual Conference of the International Speech Communication Association (2009)