MODELING THE PROSODY OF VIETNAMESE ATTITUDES FOR EXPRESSIVE SPEECH SYNTHESIS

Dang-Khoa Mac [1,2], Eric Castelli [1], Véronique Aubergé [2]

[1] International Research Institute MICA, HUST-CNRS/UMI 2954-Grenoble INP, Hanoi, Vietnam
[2] Laboratory of Informatics of Grenoble (LIG), CNRS, France

ABSTRACT

Attitudes or social affects are strongly implied in interaction processing, and specifically to socio-cultural aspects of language. This paper presents the modeling of attitude to apply in expressive speech synthesis in Vietnamese, an under-resourced tonal language. A prosodic model for Vietnamese attitude is proposed based on the concept of "rendez-vous" between linguistic levels and prosodic functions of utterance. This model is applied to generate the prosody of attitudes in Vietnamese. The perceptual experiment on the synthetic utterances with this model shows that the attitudes are well evaluated.

Index Terms— attitude, tone, prosodic modeling, expressive speech synthesis, Vietnamese

1. INTRODUCTION

During interactions between humans, speech is an important information channel to express mental, intentional, attitudinal and emotional states. According to some theoretical models of affects [1], the affective expression in speech communication may be controlled at different levels of cognitive processing, from the involuntarily controlled expressions of emotion to the intentionally, voluntarily controlled expressions of attitudes. Therefore, attitudes and emotions can be distinguished depends on the nature of the control exerted by the speaker (voluntary vs. involuntary) [2]. Some types of expressivity may be expressed as either an attitude or an emotion. For example, "surprise" can be considered as an attitude when expressed during a voluntary process; otherwise it can be considered as an emotion.

Social affects or attitudes carry the intentions and points of view of the speaker (e.g: surprise, confirmation, etc.), and they can give potency indices about the interaction (e.g authority, politeness) as well as the social context of this interaction (e.g. intimacy, politeness). An utterance without any attitude (e.g., a declaration or "simple" question) can mean that the speaker has no opinion about this utterance or that she/he does not want to or cannot express any attitude [2]. Even if many attitudes are universal in terms of their values or even their prosodic forms, some prosodic implementations and even some attitudes are specific to a given culture or language [6,7]. In any event, the attitudes are built inside each culture and language, and they must be learned by children inside the culture or by second language learners [7].

Attempts to add expressivity to synthesized speech have existed for more than a decade [4]. For a tonal language like Vietnamese, the 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 [7]. The Vietnamese prosodic contour could be generated automatically by using the Fujisaki model or a linear F0 model combined with relative registers [10]. 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 [3], 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 [2, 3]. This theoretical model allows the generation of complex prosodic contours using a superposition process directed by functions [3]. It was applied in the speech synthesis for 3 modalities (declaration, question, surprise) [3], in the automatic generation of 6 expressive prosodic attitudes for French [4] and in the prosody generation of tonal language such as Chinese [5].

Our approach to Vietnamese expressive speech production consists of applying the "rendez-vous" concept above in order to combine the local variation of tones and the global prosodic contours of attitude. As an under-resourced language, one main difficulty with Vietnamese speech processing is the lack of research and data, especially in the expressive speech domain. Therefore, the first part of this paper describes the construction of our corpus for Vietnamese attitudes. The second part presents the prosodic modeling of Vietnamese attitudes, based on the

superposition model proposed in [2]. This model is applied to the generation of prosody of attitudes in Vietnamese and then it is evaluated by a perceptual experiment. This paper ends with some conclusion and perspective for future work.

2. CORPUS OF VIETNAMESE ATTITUDE

Planned as the first corpus of Vietnamese attitudes, our 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 [1]. 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 [4, 6, 7], 16 attitudes have been represented for Vietnamese in our corpus (Table 1).

Table 1: *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

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 minutes of signal after post-processing.

3. MODELING PROSODY OF ATTITUDE IN VIETNAMESE

3.1. Superposition of prosodic contours

In the prosodic model proposed in [2, 3], the main assumption is that the prosody of utterance can be described as the superposition of independent multi prosodic contours which belong to hierarchical linguistic levels: sentence, clause, group, subgroup, etc. Each level is corresponding to a communication function of prosody, as shown in Table 2. The global prosodic contour of a sentence is the result of the additive superposition of all functional contours of all levels. For this preliminary work of integrating the prosody of attitude in expressive speech synthesis for Vietnamese, a tonal language, we concern two linguistic levels:

- syllable level, corresponding to the tonal function of prosody
- sentence level, corresponding to the attitudinal function of prosody

Table 2: *Hierarchical linguistic levels and the prosodic functions corresponding*

Linguistic levels	**Prosodic functions**
Sentence	Modality Attitude
Clause	Syntax
Group	
Syllable	Tone

3.2. Modeling the prosodic contour of Vietnamese tones

The Vietnamese language has 6 tones as shown in Figure 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.

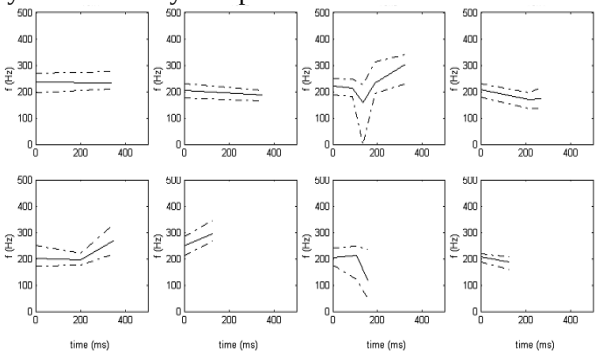

Figure 1: *Examples of contours of 8 Vietnamese tone representations from a female subject [8]. From the left to right, top to bottom: tone 1, 2, 3, 4, 5, 5b, 6, 6b.*

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 [8].

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 Figure 2) combined with relative registers of Vietnamese tone, as proposed in [9, 10]. This method is used in our work to generate the prosodic contour in the syllable level, which correspond to the tonal function of prosody.

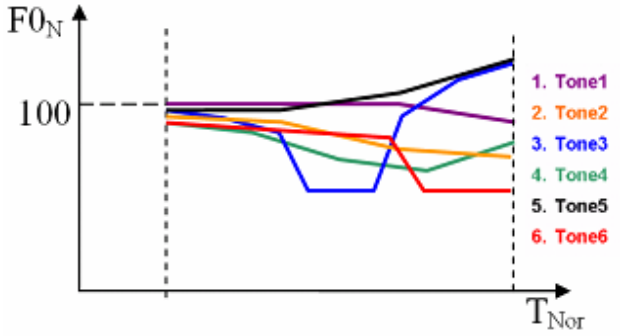

Figure 2: *Normalized linear F0 contour models for 6 Vietnamese tones [4]*

3.3. Modeling the prosodic contour of attitudes

The prosodic contour of attitude represents the attitudinal function of prosody and it corresponds to the sentence level. According to [3], 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).

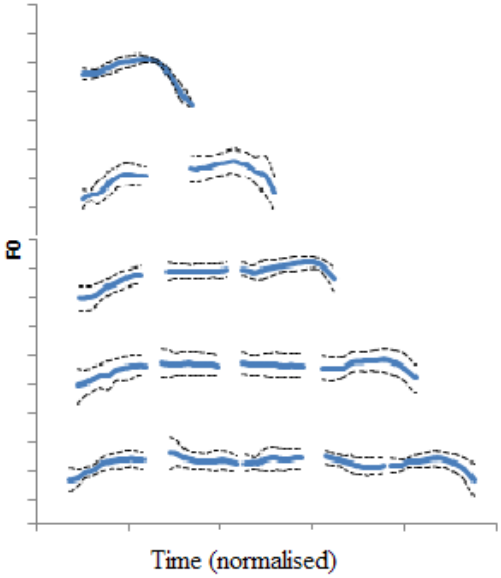

Figure 3: *Mean and deviation of F0 contours of 1 to 5 syllables sentences uttered with the attitude Authority.*

Figure 3 shows an example of the mean F0 contours of neutral-tone sentences with the length from 1 to 5 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.

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 Figure 4. 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.

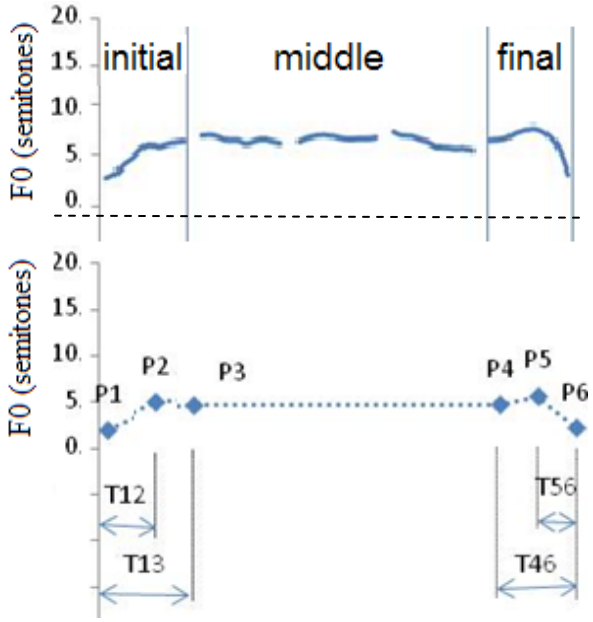

Figure 4: *An example of stylization F0 contour of attitude*

4. PERCEPTUAL EVALUATION

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

4.1. Experimental method

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 [11]), 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 without 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 with two methods:

- re-synthesis with TD-PSOLA technique in Praat environment
- generation with the speech synthesis system developed by the Institute MICA [10]

With both of methods, 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. The testing program interface gives the label and the explanation of the 4 attitudes. All subjects listened to each stimulus only one time. After each stimulus, they were asked to indicate the perceived attitude among the 4 attitudes and to indicate the intensity of its expressiveness (or the confidence about this choice) on a scale ranging from "hardly perceptible" (encoded as 1) to "very marked" (encoded as 100).

4.2. Results

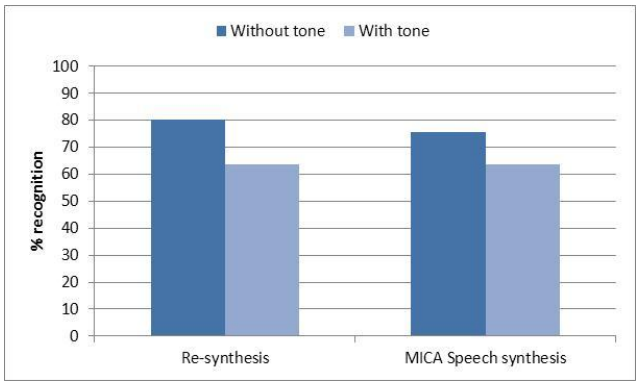

Figure 5: *The recognition rate (%) of synthetic utterances generated by re-synthesis method and by MICA speech synthesis system.*

Figure 5 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.

Figure 6 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%).

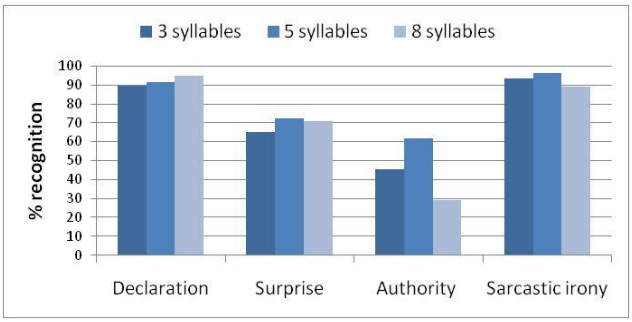

Figure 6: *The recognition rate (%) of synthetic utterances with four attitudes.*

5. CONCLUSIONS AND PERSPECTIVES

This paper presents our preliminary attempt of modeling the prosody of attitude for speech synthesis in Vietnamese, a tonal and under-resourced language. 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.

12. REFERENCES

[1] K.R. Scherer, and H. Ellgring, “Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns?”, Emotion, 7(1), 2007, pp. 158-171

[2] V. Aubergé, "A Gestalt Morphology of Prosody Directed by Functions: the Example of a Step by Step Model Developed at ICP", Speech Prosody, 2002.

[3] Aubergé, V. “Developing a structured lexicon for synthesis of prosody”. Talking Machines: Theories, Models and Designs. G. Bailly and C. Benoît, 1992, pp 307-321

[4] Morlec, Y., Bailly, G., & Aubergé, V., "Generating the prosody of attitudes", in ETRW Workshop on Prosody, Athens, Greece, 251-254, 1997

[5] Chen, G.-P., G. Bailly, et al. “A superposed prosodic model for Chinese text-to-speech synthesis”. International Conference of Chinese Spoken Language Processing, 2004

[6] Le T.X., "Etude contrastive de l’intonation expressive en français et en vietnamien", PhD thesis of Linguistic and Phonetic, Université Paris 3, 1989

[7] Shochi, T., Aubergé, V., and Rilliard, A., "How prosodic attitudes can be false friends: Japanese vs. French social affects", in Speech Prosody, Dresden, 2006, pp. 692-696

[8] Pham, T. N. Y., Castelli, E., and Nguyen, Q. C., "Gabarits des tons vietnamiens", in JEP, Nancy, France, 2002, pp. 23-26

[9] Tran, D. D., E. Castelli, et al. « Linear F0 Contour Model for Vietnamese Tones and Vietnamese Syllable Synthesis with TD-PSOLA ». Second International Symposium on Tonal Aspects of Language, Rochelle, France, 2006

[10] Tran, D. D. « Synthèse de la parole à partir du texte en langue vietnamienne », PhD Thesis, INP Grenoble, 2007

[11] Mac D.K., Auberge V., Rilliard A. & Castelli E. “Audio-Visual prosody of social attitudes in Vietnamese: building and evaluating a tones balanced corpus”, INTERSPEECH 2009, Brighton, UK, pp 2263—2266