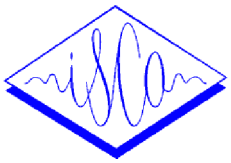

Spoken Language Technologies for
Under-Resourced Languages
(SLTU-2010), Penang, Malaysia

VIETNAMESE MULTIMODAL SOCIAL AFFECTS: HOW PROSODIC ATTITUDES CAN BE RECOGNIZED AND CONFUSED

Dang-Khoa Mac [1,2], *Véronique Aubergé* [2], *Albert Rilliard* [3], *Eric Castelli* [1]

[1] International Research Center MICA, CNRS-UMI 2954, Hanoi, Vietnam
[2] Laboratory of Informatics of Grenoble (LIG), CNRS, France
[3] LIMSI-CNRS, Orsay, France

{dang-khoa.mac, eric.castelli}@mica.edu.vn
veronique.auberge@imag.fr, albert.rilliard@limsi.fr

ABSTRACT

Social affective expression is a main part of face-to-face interaction and it is highly linked to the language through the culture. This paper presents a study on Audio-Visual prosodic attitudes in Vietnamese, an under-resourced tonal language. Based on an audio-visual corpus of 16 attitudes, perception experiments were carried out with Vietnamese and French participants. The result analysis shows the relative contribution of audio, visual, and audio-visual information in attitude perception. It also shows how native and non-native listeners recognize and confuse the attitudes, thus allows us to investigate the cultural specificities and cross-cultural common attitudes in Vietnamese.

Index Terms— Audio-visual corpus, Prosodic social affects, Cross-cultural perception, Vietnamese

1. INTRODUCTION

In speech communication between humans, the expression of mental, intentional, attitudinal, emotional states is a main information channel that is often used by both speaker and listener. Some theoretical models of affect claim that affective expression in speech communication may be controlled at different levels of cognitive processing [1], from the involuntarily controlled expressions of emotion to the intentionally, voluntarily controlled expressions of attitudes. According to [2], attitudinal expressions can be distinguished from emotional expressions by the nature of speaker's control on its expressivity (voluntary vs. involuntary). Some types of expressivity may be expressed as either an attitude or an emotion. For example, "surprise" can be considered an attitude when expressed during a voluntary process; otherwise it can be considered an emotion.

Attitude expression carries the intention and points of view of the speaker (ex: surprise, confirmation, politeness etc.). An utterance without any attitude means that the speaker does not give his opinion in this utterance. Attitudes are constructed for a language and a culture and they need to be learned by children or by second language students. As social affects, attitudinal expressions can vary amongst languages. Some specific attitudes in a language may not be recognized or may be ambiguous in other language. The understanding of this phenomenon may benefit from some cross-cultural studies [3][4].

Vietnamese is a tonal language; therefore the acoustic parameters which are implied in the linguistic and affective functions of prosody play an important role at the phonemic level for lexical access. The Vietnamese tone system has 6 tones: 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. A special feature of the Vietnamese tone system is the co-occurrence of glottalization during the production of tone 3 and tone 6. For example, 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 [5]. These phenomena of voice quality cues also happen in the morphology of some attitudes (and emotions) in other languages [3][6].

This paper presents our study of Vietnamese multimodal social affect in a Vietnamese and French cross-cultural context. Because of the contrast of language characteristics (non-tonal vs. tonal language) and the long geographic and cultural distance (West European vs. East-Asian), French was chosen for this cross-cultural study of Vietnamese social affect. This study was done not only in audio modality but also in visual modality in order to investigate the relative contribution of audio, visual, and audio-visual information in the perception of attitudes for both Vietnamese and French participants.

After presenting the Vietnamese corpus construction and the attitude selection, the perception experiment is described. The experimental results are then presented and analyzed. The results show how the native and non-native listeners can recognize and confuse the Vietnamese

attitudes. After the discussion, this paper ends with some conclusions and perspectives.

2. PERCEPTION EXPERIMENT

2.1. Attitude selection

Prosodic social affects have been studied in different languages such as English, French, and Japanese [3,4,7]. For these languages, attitudes have been selected thanks to the foreign language didactics' literature. Unfortunately, as Vietnamese is an under-resourced language, there is very little research on Vietnamese expressive speech. We have found only one study [8] dealing with this topic. From this study, we selected 16 attitudes to be examined in Vietnamese speech (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

These 16 attitudes were selected in order to investigate their existence and their realization in Vietnamese. The "*exclamation of surprise*" was divided into three sub-types: "*neutral*", "*negative*" and "*positive*" to verify whether or not they can be distinguished in Vietnamese.

2.2. Corpus construction

The corpus was constituted of 125 skeleton sentences without specific affective meaning in order to be produced naturally in all 16 attitudes. To observe the effects of tone and tonal co-articulation on attitudinal expression, the corpus contains 8 sentences of one-syllable length, which correspond to 8 representations of Vietnamese tones, 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 from 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".

One male speaker, native of Hanoi (standard pronunciation), was chosen to record the corpus. A training phase was carried out in order to ensure that the speaker expressed each attitude as naturally as possible. The corpus was recorded in a sound-proof room. A high quality microphone (AKG C1000S) was placed approximately 40 cm from the speaker's mouth. The microphone was connected to a computer outside the room through an USB sound device. The speech was recorded at 44.1 kHz, 16bits. During the recording, a digital DV camera (Sony DXC990) recorded the speaker's performance. The video clips were encoded with IndeoVideo codec at 784x576 pixels resolution. Vocal fold's vibrations were also measured using an electroglottograph. To control the speaker performance, a specialist in expressive speech and a native Vietnamese speaker observed the recording process from outside the room, through a video system. They could require the speaker to re-produce a stimulus if they thought that it was not performed satisfactorily. The speaker pronounced all 125 sentences in 16 attitudes. The complete corpus contains 2000 stimuli. It corresponds to more than 90 minutes of audio-visual signal after post-processing.

2.3. Experimental protocol

Three skeleton sentences of one-, two- and five-syllable length were chosen from the corpus for the perception experiment. We note that most of Vietnamese words are mono-syllabic or bi-syllabic [8]. As mentioned above, the Vietnamese tone system has certain characteristics that have been shown to be used in the morphology of some attitudes. Therefore the perception of attitude can be affected by tones. In order to limit the complexity of the test, the influence of tone was not investigated in this experiment (it will be studied in another experiment). The three selected sentences include no tone variation: all syllables are based on tone 1 (the level tone). These sentences were then presented in 16 attitudes and in three modalities (audio-only, visual-only and audio-visual). Thus, there were 3*16*3=144 stimuli in the perception test.

Forty listeners participated in this experiment: 20 Vietnamese (10 males and 10 females with a mean age of 25) who speak the same dialect as the speaker; and 20 French (10 males and 10 females with a mean age of 35) who have no experience on Vietnamese language. Both of these Vietnamese and French participants were separated into two groups. The first group listened to the audio-only stimuli first, then watched the video-only stimuli, and finally watched the audio-video stimuli. The second group started with the video-only stimuli, continued with the audio-only stimuli and ended with the audio-video stimuli. For each listener, the stimuli in each modality were chosen randomly in order to counterbalance a possible effect of stimuli presentation order.

The perception tests were carried out in a quiet room, using a high-quality headset (Sennheiser HD 25-13) at a comfortable hearing level. The testing program interface gave the label and the explanation of the 16 attitudes (in the native language of the listener). No listener expressed any difficulty in understanding the concepts of these 16 attitudes. All subjects listened to (and/or watched) each stimulus only once. After each stimulus, they were asked to indicate the perceived attitude among the 16 presented attitudes.

3. EXPERIMENT RESULT

3.1. Attitude recognition

Figure 1 presents listeners' recognition rates of 16 attitudes in three modalities. Globally, most of attitudes were recognized above chance level, and native listeners have higher recognition scores than foreign ones. Some attitudes were well recognized by both Vietnamese and French listeners, such as DEC, EXp, DOU, AUT, IRR, SCO, SED attitudes. The INT, IDS and COL attitudes were well recognized by Vietnamese listeners but were almost not recognized by the French listeners. In case of ADM attitude, the French listeners' recognition rate is higher than that of Vietnamese listeners.

The modality (Audio only, Visual only and Audio-visual) has a strong effect on attitude perception. As expected, for most attitudes, the average score in audio-visual modality is better than that in audio-only or visual-only modality. For the Vietnamese listeners, the audio information is very important to recognize the DEC, EXo, OBV, AUT and COL attitudes and the visual information play an important role to recognize the EXp, DOU, SCO, POL attitudes. With the French listeners, the audio information is more important to recognize the AUT and IRR attitudes, and the visual information is much more necessary to recognize the DEC, EXp, SCO and ADM attitudes.

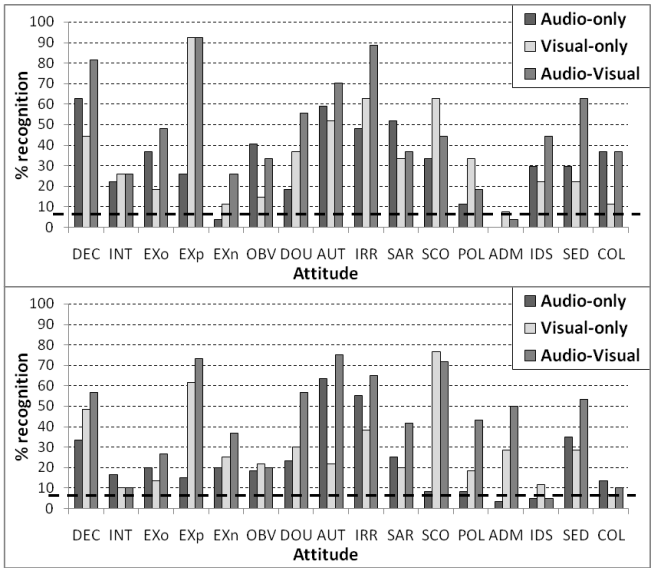

Figure 1: *Recognition rate of each attitude in each modality for Vietnamese listeners (top) and French listeners (bottom). The dash lines: chance level (6.25%)*

3.2. Attitude confusion

From the confusion matrices, confusion graphs were built based on all the confusions higher than twice the chance level (≥ 12.5%). Figure 2 shows the graphic presentation of the confusion among 16 attitudes, in three modalities for Vietnamese and French listeners.

With the audio-only information, the ADM was not recognized by both Vietnamese and French listeners. This attitude was confused with EXo (in case of Vietnamese listeners) and confused with COL and IDS (in case of French listeners). The Vietnamese listeners also did not recognize the EXn and the French listeners did not recognize the IDS. The Vietnamese listeners made a mutual confusion between some pairs/groups of attitudes, such as SAR and SCO; POL and DEC; EXo EXn and DOU. The French listeners have the mutual confusion between AUT and IRR; DOU and EXn; DOU and EXo.

With only the visual information, all attitudes were recognized above the chance level, with both Vietnamese and French listeners. The Vietnamese listeners have the mutual confusion of EXn and DOU; SED and COL; DEC and OBV. The French listeners have the mutual confusion between: EXn and DOU; COL and SED; ADM and EXp. They also strongly confused the SAR with SCO (60%).

As expected, the confusion graph in the case of audio-visual shows less confusion than in case of Audio and Video only. However, several attitudes have the recognition rate below the chance level (ADM for Vietnamese listeners and IDS for French listeners). The Vietnamese listeners confuse between EXn and DOU; SAR and SCO; POL and OBV. The French listeners have also the mutual confusion between SED and COL; EXn and DOU; SAR and SCO.

5. DISCUSSION

According to experimental results, although the mean intensity scores obtained by French listeners are lower than those of Vietnamese, they are fairy coherent with the result of Vietnamese listeners. For both groups of listeners, some attitudes were well recognized: DEC, Exp, DOU, AUT, IRR and SED. It supposes that the concepts and the expressions of these attitudes are similar in the two languages and the two cultures. So they can be seen as cross-cultural social affects (for Vietnamese and French).

Some pairs of attitudes (such as SAR and SCO; EXn and DOU) show a mutual confusion. In the audio channel, this confusion can be explained by the similarity of prosodic characteristic in the expression of these attitudes (the F0 contour, the intensity or the voice quality characteristics). Figure 3 gives an example of the F0 contour of two attitudinal expressions (DEL and POL) with the same sentence. The prosodic forms of these attitudes look nearly similar. Therefore, it is very difficult to distinguish these attitudes with only audio information.

Some attitudes (INT, IDS and COL) are recognized quite well by native listeners, however they are nearly not recognized by non-natives. Perhaps, the prosodic performances for these concepts of Vietnamese are not shared with French and they need to be learned by foreign students.

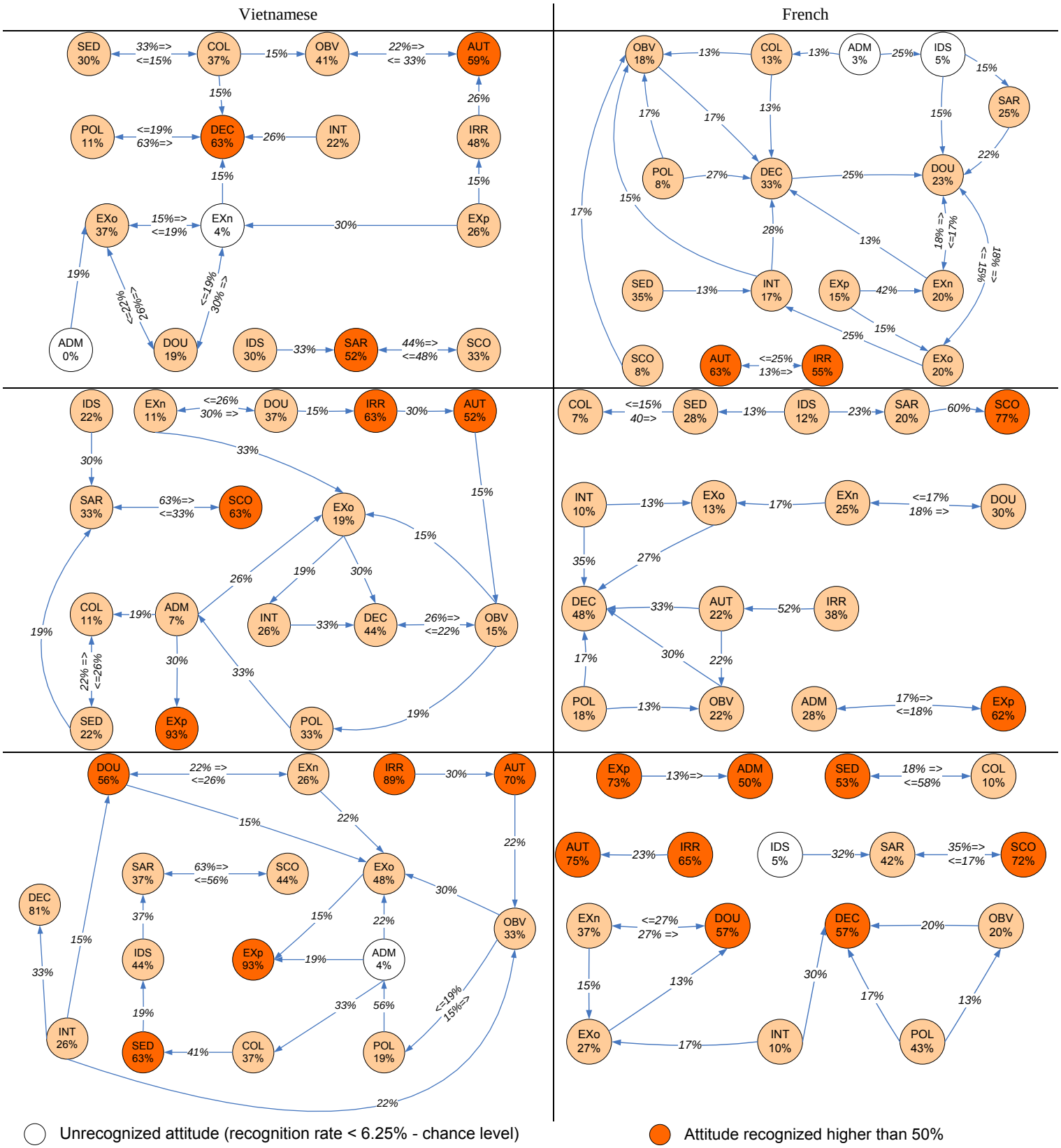

Figure 2*: Confusion graph for the 16 attitudes of Vietnamese and French listeners in Audio only (top), Video only (middle) and Audio visual (bottom)*

Similar conclusions were already discussed for some Japanese attitudes, which are not recognized by French or English [3,7]. An interesting case is the expression of Admiration, which is badly recognized by native listeners but is better recognized by the non-native ones (in visual and audio-visual modalities). Perhaps, for Vietnamese, this attitude cannot occur without lexical coherency [8]. Otherwise, in French, this concept exists and it can be expressed and can be perceived easily by speech prosody or/and gesture of speaker's face.

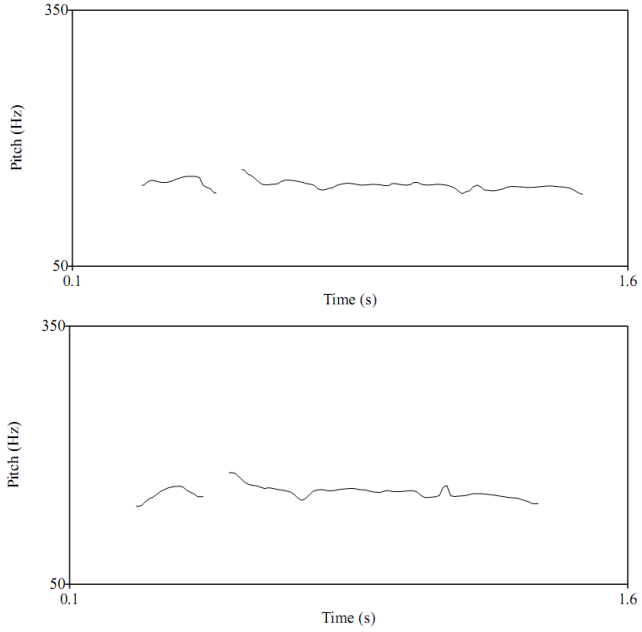

Figure 3: *F0 contour of 5-syllable length sentence in 2 attitudinal expressions: DEC (top) and POL (bottom)*

6. CONCLUSIONS AND PERSPECTIVES

Using the cross-cultural perception of audio and visual social affect in Vietnamese, the speaker's performance for 16 Vietnamese attitudes was quite well evaluated by native and non-native listeners. Experimental results reveal the influential factors on the attitudinal perception: the modality of presentation and the attitudinal expression itself. These results allow us to investigate the cultural specificities and the cross-cultural perception of Vietnamese attitudes, and also raise interesting questions for future researches as well as for educational purposes – mostly in the field of foreign language teaching.

However, the results need to be further validated by a deeper prosodic analysis to find out the acoustical and visual parameters that lead to the perception of these social affects. Other perception experiments including variations of Vietnamese tones are scheduled in order to explore the importance of such a tonal system on the perception of attitudes not only for native, but also for foreign speaker without any linguistic knowledge of a tonal language: will they be able to separate tonal from attitudinal information?

7. ACKNOWLEDGEMENTS

We are deeply grateful to Christophe Savariaux for his efficient technical contribution, as well as to the subjects of our experiments. This study was done in the framework of the KC.03.15/06-10 project.

8. REFERENCES

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

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

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

[4] Shochi, T., Aubergé, V. & Rilliard, A. (2007). Cross-Listening of Japanese, English and French social affect: about universals, false friends and unknown attitudes. Proceedings of ICPhS, Saarbrucken, Germany, 2007, pp. 2097-2100

[5] Do T.D., Tran T.H. & Boulakia G., "Intonation in Vietnamese", in Intonation systems: A survey of 22 languages, D. Hirst and A. Di Cristo, Eds.: Cambridge University Press, 1998, pp. 395-416.

[6] Shochi, T., Erickson, D., Rilliard, A., and Aubergé, V., "Recognition of Japanese attitudes in Audio-Visual speech", in Speech Prosody, Campinas, Bresil, 2008, pp. 689-692.

[7] Shochi, T., Rilliard, A., Aubergé, V. & Erickson, D. "Intercultural Perception of English, French and Japanese Social Affective Prosody", in The role of prosody in Affective Speech, ed. S. Hancil, Linguistic Insights 97, Peter Lang AG, Bern, 2009, pp.31-59.

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