Cross-cultural perception of Vietnamese Audio-Visual prosodic attitudes

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

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

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

Abstract

Prosodic attitudes (social affects) are highly linked to the language through the culture, and are a main part of face to face interaction. Therefore, for description and modeling, as well as for applications like translation, language learning or synthesis, a cross-cultural approach is relevant. This paper presents a cross-perception of 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: firstly, to understand the contribution of audio and visual modalities to affective communication; secondly, to perceptually measure how the native and non-native listeners recognize and confuse the Vietnamese attitudes. The results reveal cultural specificities and cross-cultural common attitudes in Vietnamese.

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

1. Introduction

Speech is a fundamental communication behavior of human that simultaneously conveys linguistic information as well as speaker's affective variability (e.g.: mental, intentional, attitudinal, emotional states). 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. According to [13], for this type of expressivity, voluntary and involuntary process can be discriminated by human.

For simple understanding, attitudinal expression is the way that the speaker gives an "opinion" about his own talk related to his interlocutor. Being social affects, attitudes need to be learned in childhood, or during the time that we learn the second language whose attitudes are different from our first language. As all attitudinal expressions are constructed for a language and a culture, they can differ between languages. Some attitude values can exist in one language but not in another. Some cross-linguistic studies [3,10] show that: an attitude can exist in two languages but with different prosodic forms, or two different social values in two languages have the same prosodic form (false friend).

Vietnamese is an under-resourced language, with the lack of data and research on speech processing, especially in the expressive speech domain. Vietnamese is a tonal language, therefore the acoustic parameters implied in the linguistic and affective functions of prosody play an important role at the phonemic level for lexical access. Figure 1 shows examples of tonal prosodic contours of 6 Vietnamese 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].

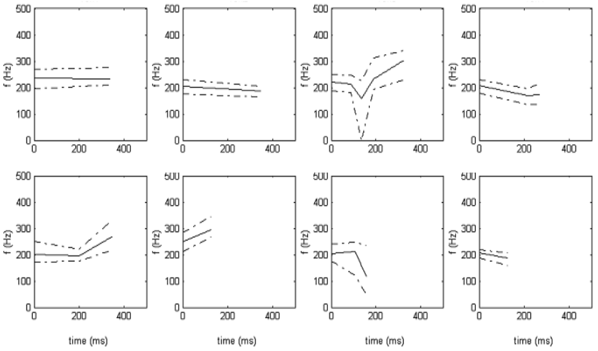

Figure 1: *Examples of contours of 8 Vietnamese tone representations from a female subject [12]. From the left to right, top to bottom: tone 1, 2, 3, 4, 5, 5b, 6, 6b.*

This paper presents a primary study of Vietnamese attitudinal expression in a cross-cultural context. Because of the contrast of language characteristics (non-tonal vs tonal language) and the long distance of geography and culture (West European vs East-Asian), French was chosen as the reference for this cross-cultural study of Vietnamese social affect.

In the face-to-face interaction, attitudes are expressed within the multimodality of speech [4]. Thus 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 attitude perception of both Vietnamese and French participant.

After presenting the construction of Vietnamese attitude corpus, we describe the perceptual experiment with

Vietnamese and French participant. The results show us the role and the integration of audio, visual, and audio-visual information in the listener's perception. This result also shows the difference in attitude recognition, confusion, and perception between the native and non-native speakers. This paper concludes with some discussions and perspectives.

2. Corpus

2.1. Attitude selection

Prosodic social affects have been studied in different languages such as English [8], French [9] and Japanese [3]. For these languages, attitudes have been selected thank to the foreign language didactics' literature. Unfortunately, as an under-resourced language, there is little research on Vietnamese expressive speech. We have found only one study [7] 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

2.2. Corpus construction

The corpus was constituted from 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 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 contained 2000 stimuli. It corresponds to more than 90 minutes of audio-visual signal after post-processing.

3. Experiment protocol

The perception test was intended to evaluate cross-linguistically the relative efficiency of the following factors on the listener's perception:

- the 16 Vietnamese attitudinal expressions
- the sentence length (in number of syllables)
- the modalities (Audio, Visual, Audio-Visual)
- the presentation order of modalities (Audio first or Visual first)

To examine the influence of sentence length, three skeleton sentences of one, two and five syllable length were chosen from the test corpus. We note that most of Vietnamese words are mono-syllabic or bi-syllabic [7]. 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 (the influence of tone will be studied in another experiment). The three selected sentences include no tone variation: all syllables are based on the 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 attitudes and to indicate the intensity of its expressiveness on a scale ranging from "hardly perceptible" (encoded as 1) to "very marked" (encoded as 100). The score 0 was assigned to the 15 other attitudes.

4. Result analysis

4.1. Effect of factors

In order to measure the effect of the factor listed above, two repeated-measure ANOVAs were calculated (one for each group of listeners), assuming compound symmetry (p>.01). The mean intensity rating of good answers was chosen as the dependent variable of the ANOVA. The within-subjects factors are the 16 attitudes, the sentence length (3 levels), and the modalities (3 levels). One between subjects factor, the presentation order (2 levels), was used. Table 2 shows the results of ANOVAs.

For both of Vietnamese and French subjects, attitude and modality have a significant effect on the perception result. In contrast, sentence length and order of presentation have no influence. However, the interaction between attitude and modality, between sentence length and order of presentation (for French listeners), and between attitude, sentence length and modality have significant effects on subjects' perception.

Table 2: *Output of the ANOVA on the mean intensity rating. Significant effects at the 1% level are set in bold face. Att: attitude; Mod: Modality; Ord: presentation order of modalities; Len: sentence length.*

		Vietnamese		French	
	df	F	p	F	p
Att	**15**	**47.804**	**0.000**	**33.100**	**.000**
Mod	**2**	**45.373**	**0.000**	**74.767**	**.000**
Ord	**1**	.022	0.882	.001	.975
Len	2	3.735	0.024	1.655	.191
Att*Mod	**30**	**6.096**	**0.000**	**9.104**	**.000**
Att*Ord	15	1.527	0.087	2.971	.000
Att*Len	**30**	**3.542**	**0.000**	**3.007**	**.000**
Mod*Ord	2	0.749	0.473	4.955	.007
Mod*Len	4	1.822	0.122	**6.061**	**.000**
Ord*Len	2	.238	0.788	.564	.569
Att*Mod*Ord	30	1.175	0.235	.872	.666
Att*Mod*Len	**60**	**2.104**	**0.000**	**1.721**	**.001**
Att*Ord*Len	30	0.806	0.763	1.138	.277
Mod*Ord*Len	4	0.547	0.701	.913	.455
Att*Mod*Ord*Len	60	0.644	0.985	1.122	.244

The perception results are presented in figure 2 for each attitude, in each modality. Globally, most attitudes were recognized above chance level, and native listeners have higher recognition scores that foreign ones, unless in the case of Admiration, were the average intensity given by French listener is equal or higher than those given by Vietnamese ones.

The modality has a strong effect on attitude perception. As expected, for most attitudes, the average score of audio-visual is better than those of audio-only or visual-only. For Vietnamese listeners, the audio information plays an important role in the case of Declaration, Obviousness, Authority and Colloquial attitudes. For French listeners, audio is mostly informative for Authority and Irritation. The video information is important in case of Positive surprise, Scorn and Politeness for both of Vietnamese and French listeners.

4.2. Clustering analysis

For both Vietnamese and French listeners, confusion matrices were analyzed. By computing the perceptual distances between attitudes from the confusion matrices, a hierarchical clustering method was used to measure the perceptive similarities across these various social affects. Figure 3 shows the hierarchical clustering for Vietnamese and French listeners in each of the 3 modalities. Using a threshold of 75 (about half of the maximum distance obtained between two attitudes), the 16 attitudes can be separated in wider groups who differ according to the considered modality.

In the audio-only condition, the clustering shows five main clusters for Vietnamese listeners and 3 main clusters for French listeners (it is an evidence of the stronger confusions made by non-native listeners). Vietnamese listeners tend to perceive in a similar way the Declaration and Politeness expressions. French listeners mixed up Authority with Irritation, and Seduction with Colloquial.

With visual-only information, results for Vietnamese listeners gives 5 wider clusters, plus the EXp recognized without confusion. Clustering for French listeners shows 4 wider clusters, with two attitudes recognized without any confusion. The elements of each cluster are different from those of the Audio-only condition. For both of Vietnamese and French listeners, the Sarcastic irony and Scorn expressions were regrouped in an “impolite expressions” cluster. Contrary to the case of Audio-only, with the visual information, the Vietnamese listeners mixed up Authority with Irritation. The French listeners show similar confusion patterns than native ones with the visual modality.

As expected (attitudes are devoted to interactive speech), the clustering of Audio-Visual gives the most relevant identification. Vietnamese listeners identify 5 main clusters and recognize 5 attitudes without any confusion (EXp, POL, IDS, AUT and IRR). The clustering of French listeners gives 4 clusters and 4 attitudes well recognized (EXP, ADM, AUT and IRR). For both Vietnamese and French listeners, a cluster regrouping EXn together with DOU, and another regrouping SED and EXo were found. The SAR was grouped together with SCO with Vietnamese listeners. In case of French listeners, the IDS was added in this group. The DEC and INT were in a two-attitude group with Vietnamese listeners, but with French listeners, they are in larger group with 5 attitudes (POL, DEC, EXo, INT, OVB).

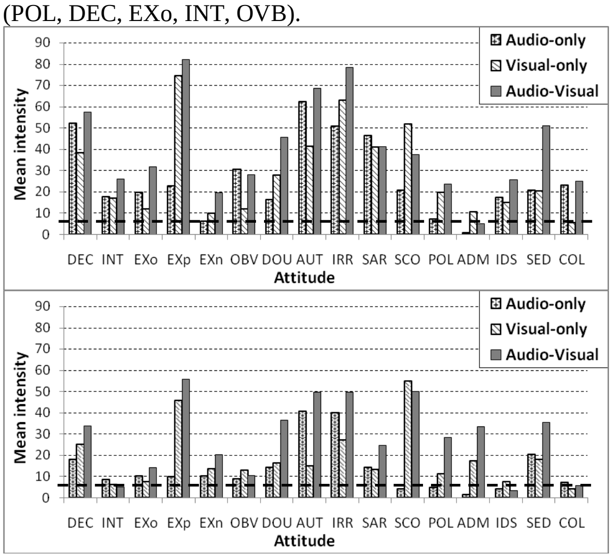

Figure 2: *Mean intensity rating for each attitude in each modality of Vietnamese listeners (top) and French listeners (bottom). The dash line: chance level.*

5. Discussion

According to the results, although the mean intensity scores obtained by French listeners is 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 the similar in two languages and culture. So they can be seen as cross-cultural social affects (for Vietnamese and French).

Otherwise, there are some attitudes recognized quite well by native listeners, which are nearly not recognized by non-natives: INT, IDS and COL. The Vietnamese prosodic performances for these concepts are not shared by French and they need to be learned by foreign language students. Similar findings were already discussed for some Japanese attitudes, which are not recognized in French or English [10].

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 necessary lexical coherency [7]. 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 [11]. However, further investigations need to be performed in order to verify this assumption.

6. Conclusions and Perspectives

This work aims at evaluating cross-linguistically the audio visual social affect of 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 attitude perception: the modality of presentation and the attitudinal expression itself. These results also allows us to investigate the cultural specificities and cross-cultural perception of Vietnamese attitudes, and also raises interesting questions for the 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 prosodic analysis to find out the acoustical and visual parameters that leads 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 for native first, 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.

8. References

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

[2] Aubergé, V., "A Gestalt Morphology of Prosody Directed by Functions: the Example of a Step by Step Model Developed at ICP", in *Speech Prosody*, 151-154, France, 2002.

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

[4] Barkhuysen, P., Krahmer, E. & Swerts, M. "Cross-modal perception of emotional speech", in *ICPhS*, 2133-2136, Saarbruecken, Germany, 2007.

[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, 395-416, 1998.

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

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

[8] Diaféria, M.-L., "Les Attitudes de l'Anglais : Premiers Indices Prosodiques", Master thesis, INP Grenoble, France 2002.

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

[10] Shochi, T., Aubergé, V. & Rilliard, A. "Cross-Listening of Japanese, English and French social affect: about universals, false friends and unknown attitudes", in *ICPhS*, 2097-2100, Saarbrucken, Germany, 2007.

[11] Rilliard, A., Shochi, T., Martin, J.C., Erickson, D. and Aubergé, V. "Multimodal Indices To Japanese And French Prosodically Expressed Social Affects", Language and Speech 52(2&3), 223-243, 2009.

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

[13] Audibert, N., Aubergé, V. and Rilliard, A. "Acted vs. spontaneous expressive speech: perception with inter-individual variability", 2nd International Workshop on Corpora for Research on Emotion and Affect, 23-27, Marrakech, Morrocco, 2008.

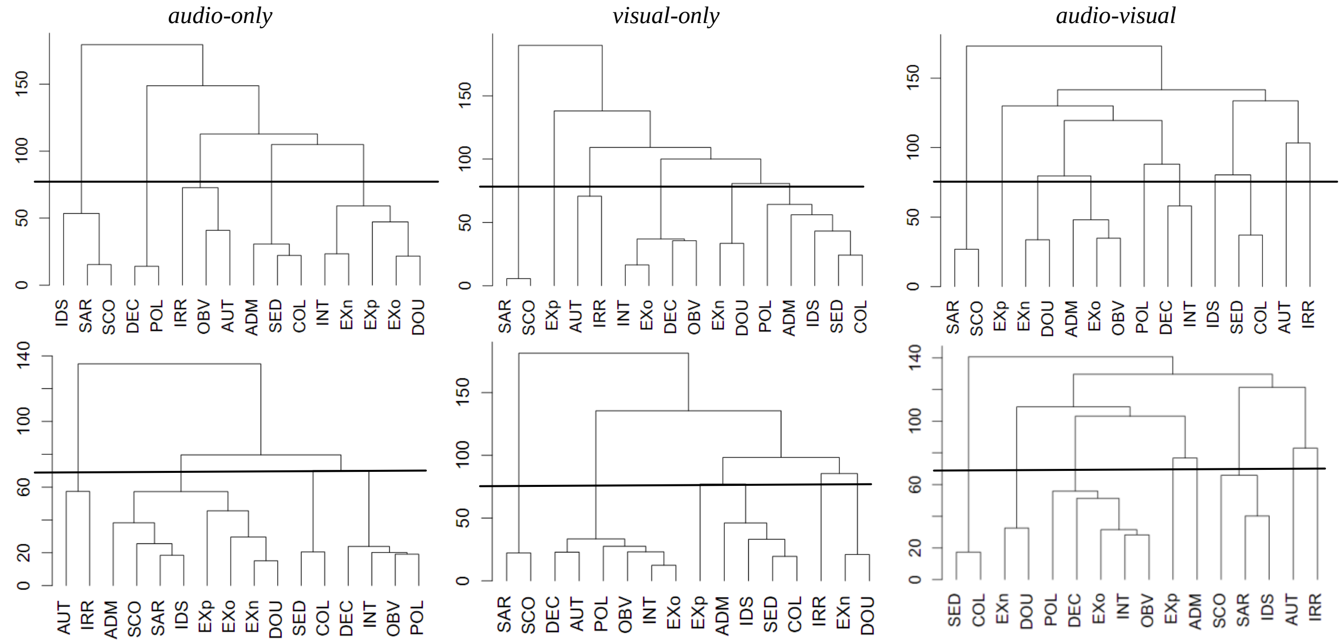

Figure 3*: Clustering analysis for the 16 attitudes of Vietnamese listeners (top) and French listeners (bottom)*