INTERSPEECH 2009 BRIGHTON

Audio-Visual prosody of social attitudes in Vietnamese: building and evaluating a tones balanced corpus

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

10.21437/Interspeech.2009-642

Abstract

This paper presents the building and a first evaluation of a tones balanced Audio-Visual corpus of social affect in Vietnamese language. This under-resourced tonal language has specific glottalization and co-articulation phenomena, for which interactions with attitudes prosody are a very interesting issue. A well-controlled recording methodology was designed to build a large representative audio-visual corpus for 16 attitudes, and one speaker. A perception experiment was carried out to evaluate a speaker's perceived performances and to study the role and integration of the audio, visual, and audio-visual information in the listener's perception of the speaker's attitudes. The results reveal characteristics of Vietnamese prosodic attitudes and allow us to investigate such social affect in Vietnamese language.

Index Terms: Audio-visual corpus, expressive speech, attitudes, perception, Vietnamese

1. Introduction

An increasing number of studies in theoretical as well as applicative fields show that social affect cannot be dissociated from high level cognition [2]. 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 [9]. 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 [13, 14]. The Vietnamese prosodic contour could be generated automatically by using the Fujisaki model [10] or a linear F0 model combined with relative registers [15]. But there is no model that can generate the prosodic contours of tones combined with expressive prosodic contours. The concept of "rendez-vous" between linguistic levels and prosodic functions of utterance [1, 3] allows the generation of complex prosodic contours using a superposition process. This concept was applied to the automatic generation of 6 expressive prosodic attitudes for French [9] as well as for speech synthesis in other languages, like Chinese [6].

Our approach to Vietnamese expressive speech production consists of applying this concept of "rendez-vous" in order to combine the variation of tones and the global prosodic contours of expressive speech. However, 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, our primary task was to construct the first audio-visual expressive speech corpus for Vietnamese.

However, the corpus was not only constructed to be used in speech synthesis, but also to conduct fundamental studies on Vietnamese social attitudes. According to [2], the speaker's attitudes during a verbal interaction are an affect built by the language and the culture. The nature and the prosodic morphology of attitudes have been shown as greatly dependent on social aspects [5, 6, 8, and 13]. The attitudinal expressions are a way for the speaker to give an "opinion" about his own talk, related to his interlocutor. Thus an utterance without any attitude can mean that the speaker expresses an attitude that he does not give his opinion about his talk. Such attitudes are distinguished from emotions by the nature of the speaker control on its expressivity (voluntary vs. involuntary). Some types of expressivity may be expressed both as an attitude or an emotion (e.g. surprise). It can be considered an attitude when expressed during a voluntary process; otherwise it can be considered an emotion [2]. As all expressions constructed for a language and a culture, they can differ between languages [5]. Therefore, some social affects may or may not exist from one language to another, and their realization in a specific language may not be recognized (or may be ambiguous in the learner's language) [13]. Such ambiguities have been shown through audio-visual experiments comparing Japanese, French and English expressions [12, 14]. Therefore, constructing an audio-visual corpus also allows us to study on the expression of Vietnamese attitudes for the first time, and to investigate the relative contribution of visual and acoustic cues.

The first part of this paper describes the characteristics of Vietnamese and the construction of our Vietnamese expressive corpus. The second part presents the perceptual experiment examining the relative contribution of audio and visual modalities in the production and perception of attitudes. The results allow us to answer the question of whether facial indices may have a significant impact on the perception of prosodic attitudes for Vietnamese. This paper ends with some discussions and perspective for future work.

2. Corpus construction

2.1. Vietnamese language and tones system

The Vietnamese language belongs to the "Viet-Muong" group within the "Mon-Khmer" branch of the Austro-Asiatic language group [7]. According to linguists' opinion, Vietnamese is a tonal isolating (monosyllabic, uninflected) language. The Vietnamese tone system consists of 6 tones: level (1), falling (2), broken (3), curve (4), rising (5) and drop

(6). Figure 1 shows examples of tonal prosodic contours in Vietnamese. 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 some tones. 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 is almost immediately cut off by a strong glottal stop [7]. In addition, the height and the shape of a tone can be also altered due to the influence of the neighbour tones. This is the tonal co-articulation phenomena in Vietnamese continuous speech [7]. For example, a tone preceded by a rising tone, such as the tone 3 or 5, will start higher than its normal target value. Otherwise, when it is preceded by a falling tone, such as the tone 2 or 6, it will start at a lower value.

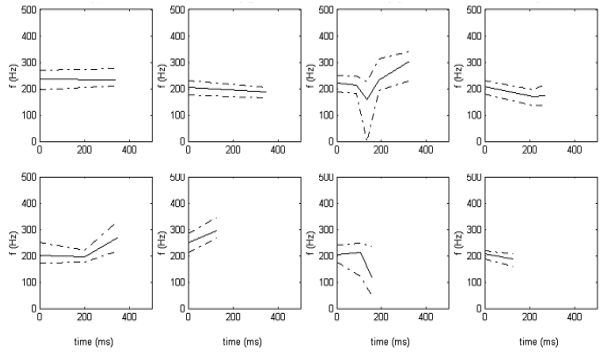

Figure 1: *Examples of contours of 8 Vietnamese tone representations from a female subject [11]. From the left to right, top to bottom: tone 1, 2, 3, 4, 5, 5b, 6, 6b.*

2.2. Attitude selection

Prosodic social affects have been studied in different languages such as English [6], French [9] and Japanese [13]. For these languages, attitudes have been selected thanks to the foreign language didactics' literature. Unfortunately, as mentioned above, there is very little research on the 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).

Because attitudes are highly linked to language and culture, some attitudes may not exist in a given language. For example the occidental concept of "*seduction*" is difficult to translate in Japanese. These 16 attitudes were selected in order to investigate their existence and their realization in the Vietnamese language. The "exclamation of surprise" was divided into three sub-types: "*neutral*", "*negative*" and "*positive*" in order to verify whether or not they can be distinguished in Vietnamese.

Table 1. *Selection of 16 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

2.3. The corpus

In order to create a prosody generation system based on the corpus, the corpus was constructed according to the concept of "rendez-vous" between prosodic contours and linguistic levels [1, 3]. The corpus was constituted from 125 skeleton sentences chosen without specific affective meaning, in order to express them naturally in all 16 attitudes. To observe the effects of tone and tonal co-articulation in attitudinal expression, the corpus contained 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 eight Vietnamese tones. The remainder of the corpus is based on 45 sentences from three to eight syllables in length and systematically varied in their syntactic structure: single word, nominal group, verbal group and a simple structure "subject-verb-object".

One male speaker, a native of Hanoi (standard pronunciation), was chosen to record the speech corpus. A training phase was carried out in order to ensure that the speaker expressed each attitude as naturally as possible. A sample corpus with few sentences had been recorded and it was informally judged by the speaker and eight other native Vietnamese to ensure the naturalness of the speaker's performance.

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 performances. The video clips were encoded with IndeoVideo codec at 784x576 pixels resolution. Vocal folds 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. Perception test

3.1. Experimental protocol

The perception test was intended to evaluate the perceptive relevance of collected attitudinal expressions in Vietnamese and the relative weight of the following factors on the perception of these 16 Vietnamese attitudinal expressions:

- the sentence length (in number of syllables)
- the modality (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-syllables length were chosen from the test corpus. Note that most of Vietnamese words are mono-syllabic or bi-syllabic [8]. In order to limit the complexity of the test, the influence of tones was not investigated in this experiment (the tones influence will be specifically study in a further devoted experiment). Therefore, the three selected sentences include no tone variation: all syllables are based on tone 1 (the level tone). The effect of Vietnamese tones on the attitudinal expression and perception

will be examined later. The three selected sentences were presented in the 16 attitudes and in three modalities (audio-only, visual-only and audio-visual). Thus, there were 3*16*3=144 stimuli for the perception test.

Twenty Vietnamese listeners (10 males and 10 females with a mean age of 25), who speak the same dialect as the speaker, participated in this experiment. They 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. The stimuli in each modality were randomized for each listener 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 also gave the definition of the 16 attitudes. No listener expressed any difficulty understanding the concepts of the 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.

Table 2: *Output of the ANOVA with 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.*

	df	F	p
Att	**15**	**47.804**	**0.000**
Mod	**2**	**45.373**	**0.000**
Ord	**1**	0.022	0.882
Len	2	3.735	0.024
Att*Mod	**30**	**6.096**	**0.000**
Att*Ord	15	1.527	0.087
Att*Len	**30**	**3.542**	**0.000**
Mod*Ord	2	0.749	0.473
Mod*Len	4	1.822	0.122
Ord*Len	2	0.238	0.788
Att*Mod*Ord	30	1.175	0.235
Att*Mod*Len	**60**	**2.104**	**0.000**
Att*Ord*Len	30	0.806	0.763
Mod*Ord*Len	4	0.547	0.701
Att*Mod*Ord*Len	60	0.644	0.985

3.2. Result analyses

3.2.1. Analysis of variance

The mean intensity rating of good answers was chosen as the dependent variable of an ANOVA. In order to measure the effect of each factor listed above, a repeated-measures ANOVAs was calculated, assuming compound symmetry (p>.01). Table 2 shows the results of ANOVA.

The factors attitude and modality have a significant effect on the perception result and a significant interaction between them. There is also an interaction between attitude and sentence length and an interaction between attitude, modality and sentence length. Figure 2 presents the average intensity for each attitude, in each modality. As expected from the result of [14], for most attitudes, audio and visual information cooperate in completion, since audio-visual scores are better than those using audio- or visual-only information. This synergy is particularly important for EXo, DOU and SED. The visual information is particularly informative for EXp, SCO and POL. The audio information plays an important role for DEC, OBV, AUT, SAR and COL.

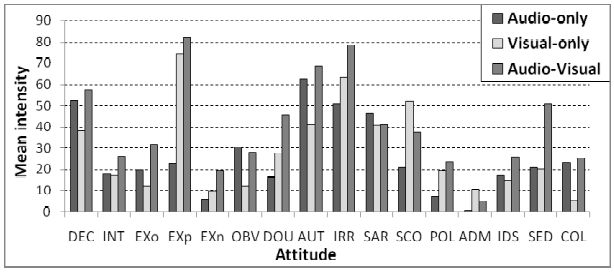

Figure 2: *Mean intensity rating for each attitude in each modality.*

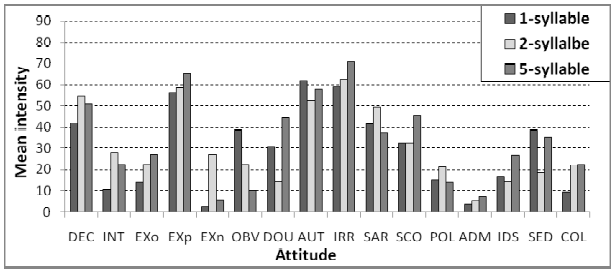

Figure 3: *Mean intensity rating for each attitude in each length.*

Figure 3 shows that the significant interaction between attitude and sentence length can be explained mainly by the low recognition scores obtained by some stimuli in some attitudes, pointing at the worst performances of the test corpus. But the non-significant effect of the length factor in ANOVA result indicates that there is no influence of the three sentences, or the sentence length. The order of presentation of modalities is not significant when measure by the mean intensity rating (p>0.01).

3.2.2. Analysis of confusion matrices

As ANOVA only deals with good answers, they omit a large part of listeners and do not permit a study of the distribution of answers. However, this is a crucial part of the analysis, because it allows the comparison of perceptual proximities between attitudes. This part of the analysis is carried out on the basis of contingency tables, counting the number of each different answer for each presented attitude. This is done separately for each of the three modalities of presentation. This gives for each presented attitude a vector representing its perception by listeners on the possible answers. Then, the Euclidian distance between each pair of vectors is calculated in order to measure the perceptual distances between attitudes. Then, a hierarchical clustering method is applied to group attitudes (using the Ward metric) on the basis of their perceptive proximity (cf. figure 4).

According to the clustering analysis, the 16 attitudes can be separated in wider groups, differing according to the modality considered. The separation threshold was set at 75, almost half the maximum distance obtained between two attitudes. Such an analysis gives 5 clusters of attitudes for the audio modality, that can be labelled under more general kind of expressions: (1) Impolite expressions (SAR, SCO), with the

misunderstood audio-only IDS; (2) assertive expressions (DEC, POL); (3) expressions of imposition of the speaker's view (IRR, OBV, AUT); (4) expression of agreement (ADM, SED, COL) and (5) dubitative expressions (the 3 exclamations, INT, DOU).

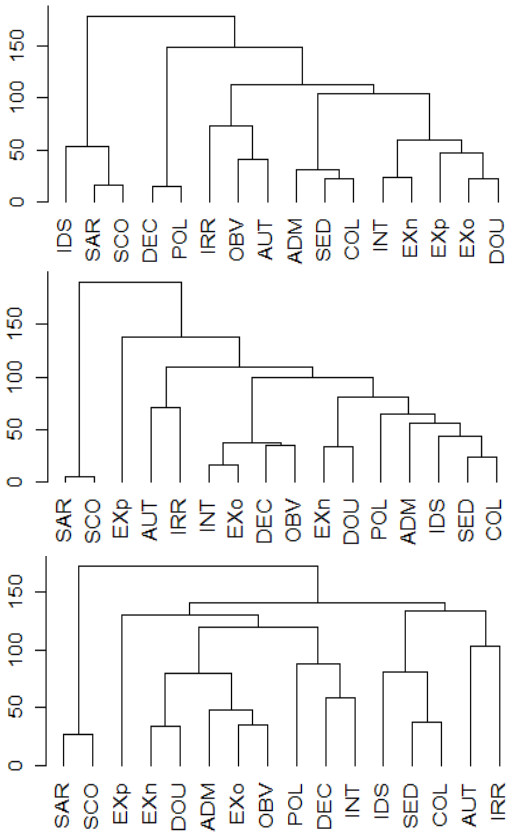

Figure 4: *Clustering analysis for the 16 attitudes in the audio-only (top), visual-only (middle) and audio-visual (bottom) modalities.*

The visual-only clustering gives also 5 wider clusters, plus the EXp, recognized without confusions: (1) the impolite expressions (SAR, SCO), without confusions with IDS; (2) the expression of imposition of the speaker's view (AT, IRR); (3) the neutral dubitative expressions (INT, EXo), with the two expressions of DEC and OBV, misunderstood with this modality; (4) the negative dubitative expressions (EXn, DOU) and (5) all the expressions of agreement (POL, ADM, IDS, SED, COL).

Finally, the clustering of Audio-Visual stimuli gives the most relevant identification, with 3 attitudes recognized without any confusion (EXp, POL, IDS); four main clusters: (1) impolite (SAR, SCO); (2) negative dubitative (EXn, DOU); (3) imposition (AUT, IRR) and (4) agreement (SED, COL). Finally, a cluster regrouping DEC together with INT, and another regrouping ADM, EXo and OBV were found. These two last clusters are below the threshold, but contain attitudinal expressions that are still differentiated.

4. Conclusions and Perspectives

This corpus is the first audio-visual corpus of Vietnamese attitude and was designed to be used for different purposes: prosody modeling, audio-visual expressive speech synthesis, attitudinal speech recognition as well as contrastive social and cultural studies.

General performances of the speaker in the production of these 16 attitudes were quite well evaluated, when compared to similar works in French, English and Japanese [12]. However, some expressions may be performed more adequately (since the listeners did agree with the conceptual existence of this attitude values), or may not be really differentiated by Vietnamese speakers, outside any context of production and/or without lexical cues. The special case of ADM being quite badly recognized calls for further investigation to verify if and how such social affect exists like a conceptual social value, and then can be produced in Vietnamese. In case of Infant Directed Speech (IDS) attitude, because the speaker is a young man and has no children, he designed a situation with his young nephew. Perhaps this kind of attitude not so easily expected to use ecologically such speech, however the good recognition result of this attitude in the audio condition raises interesting questions for the future researches. Such a situation appears to be efficient at least for the visual information, but not at all for the audio cues (SAR, SCO). Future works will also have to explore the importance of the tonal system on the production and the perception of Vietnamese attitudes.

5. Acknowledgements

We are deeply grateful to Christophe Savariaux for his efficient technical contribution.

6. References

[1] Aubergé V., "Developing a structured lexicon for synthesis of prosody", in *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoît, Eds., 1992, 307-321.

[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*, 2002.

[3] Aubergé V. & Bailly, G., "Generation of intonation: a global approach", *Eurospeech 1995*, Madrid, Spain, 1995, 2065-2068.

[4] Chen G.-P., Bailly G., Liu Q.-F. & Wang R.-H., "A superposed prosodic model for Chinese text-to-speech synthesis," in *Int. Conference of Chinese Spoken Language Processing*, 2004.

[5] Danes, F., "Involvement with language and in language", *Journal of pragmatics,* 22, 251-264, 1994.

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

[7] 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, 395-416.

[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.

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

[10] Nguyen, T. D., Mixdorff, H., Luong, C. M., Ngo, H. H., and Vu, K. B., "Fujisaki Model based F0 contours in Vietnamese TTS", in *ICSLP2004*, Korea, 2004, pp. 1429-1432.

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

[12] Rilliard A., Martin J.-C., Aubergé V. & Shochi T., "Perception of French Audio-Visual Prosodic Attitudes", in *Speech Prosody*, Campinas, Brazil, 2008, 685-688.

[13] 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.

[14] 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.

[15] Tran, D. D., Castelli, E., Le, X. H., Serignat, J.-F., and Trinh, V. L., "Linear F0 Contour Model for Vietnamese Tones and Vietnamese Syllable Synthesis with TD-PSOLA", in *2nd Int. Symp. on Tonal Aspects of Language*, Rochelle, France, 2006.