Perception of French Audio-Visual Prosodic Attitudes

A. Rilliard[1]*, J.C. Martin*[1]*, V. Aubergé*[2] *& T. Shochi*[2]

[1] LIMSI-CNRS, Orsay, France; [2] GIPSA-Lab, Grenoble, France
{rilliard; martin}@limsi.fr ; {auberge; shochi}@gipsa-lab.inpg.fr

Abstract

Experimental studies are required to understand the contribution of audio and visual modalities during affective communication. This paper presents a perception study of the audio-visual expression of six French attitudes. The relative importance of each modality in the decoding of these expressions is analysed, as a first step toward a deeper comprehension of their influence on social affects expression. Two speakers are evaluated, in order to test the influence of speaker's performances on listeners' perception.

1. Introduction

Previous studies on prosody describe it as conveying different levels of information, ranging from linguistic to expressive [17, 12]. The variety of expressive functions has been studied for a long time (cf. [1, 11, 10]). In a similar way, audio-visual speech prosody has more recently been recognized as carrying functions such as the feeling of knowing [22] and the signalling of end of utterances [5].

With respect to the nature of affects, the work of [23] has raised the question of possible differentiation between social affects and emotions. A similar differentiation is used by [16]. Aubergé [3] starts from this distinction between social affects (or attitudes) and emotions to question the existence of different controls underlying the expression of these different affects, i.e. the parameters carrying the expressivity may be organized differently by the subject in timing, shape or range, and this allows the interlocutor to decode emotional state.

Such attitudes or social affects have already been studied by several scientists [24, 14, 4, 9] and also in cross-linguistic contexts [20, 7], but mainly in their acoustic modality only. As the study of the multimodal expression of emotions (in their broad sense) is still a recent field of research [18, 19], only a few works question directly the specific question of social affects, specifically differentiated from emotions [13]. This paper, and the companion paper describing a similar work on Japanese attitudes [21] intend to analyse the specificities and the complementarities of visual and audio modalities in the expression and perception of a specific kind of social affect: attitudes. It also tries to measure the influence of speaker's performance on perception judgements, as many works only rely on the performance of only one speaker.

The first part of the paper describes the construction of the corpus, and explains the choices of expressions; then the experimental setting is tackled. The second part approaches the results analysis and discussion, before some conclusions and perspectives for future works.

2. Corpus & Experimental design

2.1. French Audio-Visual attitudes

Following the work done by Morlec [14] on French prosodic attitudes, based on [11], 6 attitudinal expressions were selected for recording a French audio-visual corpus: *declaration* (DC), *simple question* (QS), *obviousness* (EV), *surprise exclamation* (EX), *doubt-incredulity* (DO), *suspicious irony* (SC). The main differences between the work done by [14] and this one reside in (1) the audio-visual recording of the 6 attitudes, (3) a work on two different speakers, in order to measure the influence of individual performance over recognition scores, and (2) the recording paradigm, designed to set the speaker in a somewhat more natural condition of production of these 6 attitudes: speakers were instructed to produce each of these sentences in order to express one attitude, as an answer to a statement produced by a partner. They had already been trained to produce these attitudes in a preceding session, and had to behave as naturalistic as possible, without any constraints on their expressive strategy.

Two speakers, S1 & S2, both male native French speakers, were recorded in a soundproof room at LIMSI. They were standing in front of a video camera, with an omnidirectional AKG C414B microphone placed 40 cm from their mouth. The microphone was connected to an USBPre sound device connected to a computer outside the room, recording the speech signal at 44,1 kHz, 16bits. A digital DV camera (Canon XM1 3CCD) recorded the speakers' performances. Hands claps between each sentence, recorded both by the camera and the microphone, allow replacing the camera sound by the high quality microphone sound, synchronized thanks to the claps in a post-processing phase. Video clips were encoded using a cinepack codec with a 784 x 576 pixels resolution, either using both MOV and AVI video file formats (respectively for display on Apple or Windows platforms).

The corpus is based on three sentences respectively of 4, 5 and 7-syllable length, without any specific meanings that can bootstrap or forbid one of the 6 attitudes. After the recording and the post-processing, the speakers' performances were judged by both of them, and only the 5-syllable length sentence was kept for the perception test: "*Nicolas revenait.*" [nikola ʁəvnɛ] ("Nicolas was coming back"), played with the 6 attitudes. 6 short videos were thus produced.

2.2. Perception test

An evaluation test was designed in order to evaluate the relative efficiency of the two modalities to carry the attitudinal information. The factors that have been controlled during this experiment are:

- the 6 attitudes;
- the speaker (S1 or S2);
- the modality (Audio, Visual or Audio-Visual);
- the modalities' presentation order (Audio or Visual first);

Subjects listened to each stimulus only once for each modality, presented in a random order. For each stimulus, they had to select the attitude they perceived in the stimulus as well as its intensity on an open scale ranging from "*hardly perceptible*" to "*very marked*" (encoded on a 1-100 scale, with the 0 score for the 5 not selected attitudes). Subjects had to fill the questionnaire on the PC without any time constraint.

Two groups of subjects passed the experiment. The first group first listened to the audio only stimulus, and then watched the video only stimulus, and finally the audio video stimulus. The second group started with the video only stimulus, continued with the audio only stimulus and finally ended with audio-video stimulus. This enabled to counter-balance a possible effect of the presentation order of the stimuli's modality. During the presentation of one modality, the stimuli corresponding to all attitudes and to the two speakers are randomized – in a different order for each listener.

2.3. Subjects

32 French listeners (17 male and 15 female, mean age = 32) pass the experiment, 16 in each group (Audio only first and Visual only first).

Table 1: *results of the 2 ANOVAs, on the percentage of recognition of each attitude and on the intensity scores. Significant effect (p<.01) are in bold.* Grp *stands for the Group factor,* Spk *for the Speaker,* Mod *for the Modality and* Att *for the Attitudes.*

		% Reco.		**Intensity**	
	df	*F*	*p*	*F*	*p*
Grp	1	0.5	0.508	0.3	0.605
Spk	**1**	**45.3**	**0.000**	**105.5**	**0.000**
Grp:Spk	1	1.9	0.178	7.1	0.012
Mod	**2**	**14.9**	**0.000**	**25.6**	**0.000**
Grp:Mod	2	3.8	0.028	3.7	0.030
Att	**5**	**6.4**	**0.000**	**11.8**	**0.000**
Grp:Att	5	2.2	0.055	0.9	0.466
Spk:Mod	2	1.7	0.199	3.0	0.059
Grp:Spk:Mod	2	1.0	0.364	0.0	0.994
Spk:Att	**5**	**3.8**	**0.003**	**7.5**	**0.000**
Grp:Spk:Att	5	0.7	0.589	0.5	0.764
Mod:Att	**10**	**7.9**	**0.000**	**8.6**	**0.000**
Grp:Mod:Att	10	1.3	0.246	1.4	0.195
Spk:Mod:Att	**10**	**5.4**	**0.000**	**3.6**	**0.000**
Grp:Spk:Mod:Att	10	1.5	0.146	1.7	0.085

3. Results analysis

3.1. Results processing

Results given by listeners are expressed by two measures: as categorical answers (the perceived attitude), and as a relative intensity score given to one category of attitude. Two kinds of results are analyzed: (1) the recognition rate of each attitude, expressed either as the sum of the categorical choice of the attitude by listeners (percentage of good recognitions), or as the mean intensity rating of good answers; and (2) the confusions matrix grouping the categorical answer given by listeners for each presented attitudes – expressed either as categorical recognition rate received by each possible attitude, or as the relative intensity received by each possible attitude compared to the total intensity rating received by the stimuli.

Recognition rates (either categorical or intensity) are analysed by means of two repeated-measure ANOVA analyses (one for each kind of measure). Each ANOVA takes as a dependant variable the recognition rate of each attitude (expressed with percentages or mean intensity), subjects as a random effect, one between-subject factor (the group, or the order of presentation of the Audio and Video modality), and three within-subject factors (the 6 Attitudes, the 2 Speakers and the 3 Modalities). ANOVA analysis mainly aims at measuring the relative importance of the different factors on the listener's behaviour.

Confusion matrices are analysed by means of a correspondence analysis and a cluster analysis [8]. Both of these methods are based on data-reduction techniques that allow a more simple and comprehensive representation of the data. The first one allows a graphic representation of the perception results, in order to analyse the self-recognition of one attitude and their relative dispersion. The second one hierarchically regroups the different stimuli in clusters, the distances of which indicate the perceptive distance between the attitudes (the Ward distance metric is used for clustering), and thus allows for distance judgements.

3.2. ANOVA results

For both recognition percentages and intensity scores, the Mauchly's test of sphericity is not significant (p>.01). Therefore, the repeated-measures ANOVAs were computed, assuming compound symmetry. Results are displayed in table 1 for both ANOVAs.

The first consideration of this analysis is the complete coherence obtained for both kind of measures: all affects receive coherent significance level, whatever the measure. Therefore, in the remaining of the analysis, no difference between the two measures will be made. The order of presentation of the modality (the group factor) is not significant. It only has a small interaction marginally significant with the Modality factor: scores for the Audio only condition are slightly higher when the audio modality is not presented first. The main effects are received by the three main factors: the Speaker, the Modality and the attitudes. Scores for these three factors are presented on the figure 2.

Speaker S1 receives higher recognitions scores than S2, whatever the modality or the expressed attitude. The difference between speakers is even amplified by the intensity ratings – S2's expressions being perceived with a lower activation.

As expected, Audio-Visual stimuli receive the highest ratings. Audio- and Visual-only modalities received comparable mean scores.

Mean recognition rate for each individual attitude is far above the chance level (16,6%), despite important differences between them (cf. figure 2).

Interesting interactions between these three factors can also be noticed. Figure 2A presents the percentage of recognition for each attitude, according to the presentation modality. Audio information appears particularly efficient alone for the expressions of *declaration* (DC) and *simple question* (QS), and video information for *doubt-incredulity* (DO). The Audio-Visual modality receives almost always the best score (or at least nearly the best), for all attitudes, and shows a particularly important synergy between the two modalities for the expressions of *obviousness* (EV), *suspicious irony* (SC) and *surprise exclamation* (EX).

Figure 2B presents the relative performances of the two speakers for each attitude. If S1 scores are almost always higher, he is particularly more efficient for the SC attitude, and surprisingly for DC. For this attitude, speaker S2 moves his eyes up at the beginning of the recording (trying to remember the sentence to produce), and it gives a strong notion of doubt or question, especially on the video-only stimuli: DC is absolutely not recognized for the video modality, whereas it receives good scores for the audio-only one. The difference between the two speakers for the SC expression is mainly due to the relative efficiency of their production.

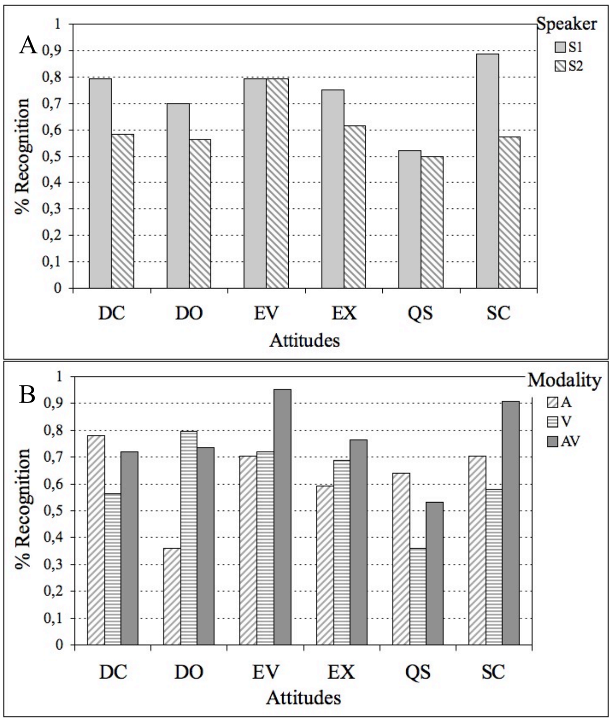

Figure 2: *(A) presents the mean percentage of recognition obtained by each attitude in each of the 3 modalities (Audio-Visual, Audio and Visual). (B) presents the percentages obtained by each attitude for each speaker (S1 and S2).*

3.3. Analysis of confusion matrices

The analysis of the confusion matrices leads to interesting parallel between attitudinal expressions. In order to obtain the main differences and similarities, two data reduction techniques are applied.

3.3.1. Correspondence analysis (CA)

Such analysis extracts the main abstract dimensions that explained most of the variance in the original data. Applied to the confusion matrix for each speaker and for each modality (and for both the percentage of recognition and the intensity scores), it raised the most important divergences between stimuli and also the main proximities. Moreover, as it plots both the rows of the matrices (corresponding to the proposed stimuli) and its columns (i.e. the answered attitudes, corresponding to the concept subject have of each attitude), the proximity between the rows and column dots may indicate a good recognition of the attitude, whereas a great distance indicate lower recognition scores. As the results obtained by recognition percentages and by intensity scores gives completely coherent results, they will not be analysed separately.

The first two dimensions of these analyses explain between 51 to 69% of the variance, for each speaker and each modality; and adding the third one generally explains about 80%. Interestingly, dimensions extracted by the analyses are very comparable (as they grouped or opposed the same attitudes) from one speaker to the other; even if their order sometimes changes[1]. If we compare the first 4 dimensions of each CA, it then gives a very precise sketch of the contribution of French attitudes for each modality.

The main result is the good recognition scores obtained by each attitude, for each speaker and in each modality – unless the DC by S2 in the video-only condition, as already mentioned.

Audio-only information. First dimension opposed affirmative attitudes (DC and EV) to questioning ones (QS, DO, EX). Second one opposed SC to the others. The third and the fourth dimensions opposed the attitudes regrouped by the main opposition of the first dimension: respectively QS vs. EX and DO; and DC vs. EV. For this modality, EX and DO show some perceptive similarities.

Video-only information. The first dimension for the facial information opposes EX to statements expressions. The second dimension opposes SC and EV to DC. The third opposes QS and DO to EV and EX. And finally the fourth opposes SC to EV. Such results show that the four non-linguistically encoded attitudes (i.e. DO, EV, EX and SC) have strong visual cues, whereas expressions of DC and QS are respectively opposed to the first four[2], or have strong confusions with one another (QS is mixed with DO in this modality)

Audio-video information. The first dimension for the audio-video stimuli opposes affirmative to questioning expressions, and the second one opposes SC to the others, as for the audio-only condition.

The third dimension opposes obviousness to declaration, and the fourth opposes EX to DO and QS. The last two dimensions show the main differences existing inside the two main groups drawn by the first dimension, as for the audio condition. But for the AV condition, confusions exist for QS, recognized as DO, as for the visual-only modality.

3.3.2. Cluster analysis

The cluster analysis leads to different conclusions. Its main purpose is to analyse the distances between the stimuli. It has already been said that most attitudes are well recognized, so this analysis will focus on the confusions between attitudes, and on the differences and similarities between speakers by modalities.

Audio-only stimuli. Speakers receive coherent result in modality. DO and EX are closely related, and show a looser connection with QS. This result differs from preceding ones obtained by [2], where DO was mixed up with SC, a confusion that is totally absent of our data. Other attitudes are clearly dissociated.

Video-only stimuli. Visual information clearly differs from one speaker to the others. For S1, QS is mixed up with DO, while all the others expressions are clearly recognized. For S2, the already mentioned problem with DC makes it close to

[1] For the sake of simplicity, the order of dimension will follow the one obtain on the analysis of speaker S1. When dimensions differ between speakers, it will be explicitly mentioned.

[2] We don't refer here to the results obtained by speaker S2 for the visual-only declaration, as they may be biased.

three other attitudes, each of which also showing confusion with another one. QS is mixed up with EX, and SC with DC.

Audio-video stimuli. There is a clear interference between both modalities in these results. For S1, the visually clear expression of EX helps listeners to distinguish it from its acoustically similar DO; whereas the acoustic cues of QS separates it from the more similar visual DO. For S2, audio DC prevails over the problematic eye movements observed in visual modality and the visual difference between DO and EX allows listeners to differentiate them where acoustic cues fail.

4. Conclusions

This work on audio-visual expression of French social affects gives interesting results on several dimensions, and opens up interesting research questions.

By comparing these results to the ones obtained by [2], the importance of the speaker's strategy is clear. Even if most of the data is coherent for the three speakers (the 2 of this study and the one cited above), it is clear that both strategic choices in the use of available acoustic and visual parameters, as well as individual performance to achieve a particular attitude have an influence on the perception results. More works, like [15] may be devoted to data collection on important amounts of social affect expressions, with a specific attention to naturalness.

The comparison of each modality's account to expressivity is coherent with [7] works, which points out the relative importance of multimodality in interaction. Moreover, the relative contribution of audio and visual cues to each attitude is coherent: audio information seems primarily important for *declaration* and *simple question*, both linguistically encoded, whereas the others (and especially *surprise exclamation* and *doubt-incredulity*) seems primarily influenced by visual cues, with strong speaker-dependant changes. Moreover, the synergy between modalities is important for almost each attitude, and especially *obviousness* and *suspicious irony*.

We are currently working on the analysis of perception results filtered by the extraversion rating of listeners, and also on the analysis of the acoustic and visual cues to each attitude. Correlations between these cues (and their evolution in time) and perception results will allow a deeper understanding of the modalities' contribution.

Finally, audio-visual replay on a talking head of these affects may allow investigating the indices that lead to naturalness and spontaneity in expressivity. Spontaneity of multimodal expression is still a great challenge, and some improvement of the recording paradigm may be introduced (or spontaneous data collected) in order to acquire more information on this point. Moreover, it may be interesting to specifically design a perception test in order to be able to monitor the recognition speed of listeners, as it has been related to the valence of multimodal stimuli [6].

5. References

[1] Allerton, D.J. & Cruttenden, A. 1978. Syntactic, illocutionary, thematic and attitudinal factors in the intonation of adverbials. *Journal of Pragmatics*, 2, 155-188.

[2] Aubergé, V., Grépillat, T. & Rilliard, A. 1997. Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours. *EuroSpeech*, Rhodos, 871-877.

[3] Aubergé, V. & Cathiard. MA. 2003. Can we hear the prosody of smile? *Speech Communication*, 40 (1), 87-97.

[4] Bänziger, T. & Scherer, K.R. 2005. The role of intonation in emotional expressions. *Speech Communication*, 46, 252-267.

[5] Barkhuysen, P., Krahmer, E. & Swerts, M. 2006. How Auditory and Visual Prosody is Used in End-of-Utterance Detection. Interspeech, Pittsburgh, USA.

[6] Barkhuysen, P., Krahmer, E. & Swerts, M. 2007. Incremental perception of acted and real emotional speech. Interspeech, Antwerp, Belgium, 1262-1265.

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

[8] Benzecri, JP. 1973. L'analyse des données. Paris: Bordas.

[9] Campbell, N. 2005. Getting to the Heart of the Matter; Speech as the Expression of Affect. rather than just Text or Language. *Language Resources and Evaluation*, 39 (1), 111-120.

[10] Danes, F. 1994. Involvement with language and in language. *Journal of Pragmatics*, 22, 251-164.

[11] Fónagy, I., Bérard, E. & Fónagy, J. 1984. Clichés mélodiques. *Folia Linguistica*, 17, 153-185.

[12] Fónagy, I. 2003. Des fonctions de l'intonation : essai de synthèse. *Flambeau*, 29, 1-20.

[13] Granström, B. & House, D. 2005. Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46, 473–484

[14] Morlec, Y., Bailly, G. & Aubergé, V., 2001. Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication*, 33 (4), 357-371

[15] Moroni, V. 1997. *Enquête sur les attitudes du français: définition et interprétation*. Master thesis, Univ. Grenoble 3, France.

[16] Ohala, J.J. 1996. Ethological theory and the expression of emotion in the voice. *ICSLP*, Philadelphia, 1812-1815.

[17] Rossi, M., Di Cristo, A., Hirst, D., Martin, P., Nishinuma, Y. 1981. *L'intonation: de l'acoustique à la sémantique.* Paris: Klincksieck.

[18] Scherer, K.R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2), 227-256.

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

[20] Shochi, T., Aubergé, V. & Rilliard, A. 2007. Cross-listening of Japanese, English and French social affect: about universals, false friends and unknown attitudes. *ICPhS*, Saarbrücken, Germany.

[21] Shochi, T., Erickson, D., Rilliard, A, Aubergé, V. &. Martin, J.C. (submitted). Recognition of Japanese attitudes in Audio-Visual speech. *Speech prosody*, Campinas, Brasil.

[22] Swerts, M. & Krahmer, E. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1), 81-94.

[23] Tomkins, S. S. 1984. Affect theory. In *Approaches to emotion*, K. R. Scherer & P. Ekman (eds.). Hillsdale, N.J.: Erlbaum, 163-196.

[24] van Heuven, V.J., Haan, J., Janse, E., van der Torre, E.J. 1997. Perceptual identification of sentence type and the time distribution of prosodic interrogativity marker in Dutch. *ETRW Workshop on Prosody*, Athens, Greece, 317-320.