Multimodal Indices to Japanese and French Prosodically Expressed Social Affects

Albert Rilliard[1], **Takaaki Shochi**[2], **Jean-Claude Martin**[1], **Donna Erickson**[3], **Véronique Aubergé**[4]

[1] *LIMSI-CNRS, Orsay, France.*
[2] *Kumamoto University, Kumamoto Japan*
[3] *Showa Music University, Kawasaki City, Japan*
[4] *GIPSA-Lab, Grenoble, France.*

Key words

attitudes

audovisual prosody

French

Japanese

perception

Abstract

Whereas several studies have explored the expression of emotions, little is known on how the visual and audio channels are combined during production of what we call the more controlled social affects, for example, "attitudinal" expressions. This article presents a perception study of the audovisual expression of 12 Japanese and 6 French attitudes in order to understand the contribution of audio and visual modalities for affective communication. The relative importance of each modality in the perceptual decoding of the expressions of four speakers is analyzed as a first step towards a deeper comprehension of their influence on the expression of social affects. Then, the audovisual productions of two speakers (one for each language) are acoustically (F0, duration and intensity) and visually (in terms of Action Units) analyzed, in order to match the relation between objective parameters and listeners' perception of these social affects. The most pertinent objective features, either acoustic or visual, are then discussed, in a bilingual perspective: for example, the relative influence of fundamental frequency for attitudinal expression in both languages is discussed, and the importance of a certain aspect of the voice quality dimension in Japanese is underlined.

Acknowledgements: We are deeply grateful to T. Sadanobu from Kobe University, Japan, for his helpful advice and contribution on Japanese attitudes, and to J. Beskow and an anonymous reviewer for their helpful comments and advice on an earlier version of this article. We also thank K. Yoshimoto for his very precious performance, and students at Kanagawa University, Japan, and people at LIMSI and GIPSA-Lab, France, for their participation in the perception tests. This work was supported in part by the CNRS-SHS ATIP grant "La prosodie des affects: des attitudes aux emotions" to the first author, in part by the Japanese Ministry of Education, Science, Sport, and Culture, Grant-in-Aid for Scientific Research, (2007–10):19520371 to the fourth author, and also by SCOPE (071705001) of Ministry of Internal Affairs & Communications (MIC), Japan.

Address for correspondence. Albert Rilliard, LIMSI-CNRS, Bat. 508 – BP 133, 91403 Orsay Cedex, France; e-mail: <albert.rilliard@limsi.fr>

Language and Speech 0023-8309; Vol 52(2/3): 223–243; 103171; DOI:10.1177/0023830909103171
http://las.sagepub.com

1 Introduction

Some theoretical models of affect claim that affective expression may be controlled at different levels of cognitive processing (e.g., Scherer & Ellgring, 2007), from involuntarily to intentionally, deliberately controlled expression of the speaker's attitudes (Daneš, 1994; Ohala, 1996). Our position (Aubergé, 2002) is to distinguish *attitudes* vs. *emotions* as a difference in the nature of the control by the speaker (voluntary vs. involuntary), not the affective value carried by the expressions. Moreover, attitudinal expressions are considered here as affects conventionally encoded in a culture and a language, while emotional expressions may not have been socially constructed to this same extent. According to Damasio (1994), spontaneous emotions involve physiological changes, and are spontaneously expressed inside a "body loop." But emotions can also be simulated, outside this body loop, using the memory of the somatic states associated to a past emotion. These simulation capabilities allow one to express emotion during conversation, and allow actors to act them. It may be that some people are able to detect the imitation, while others are not (Audibert, Aubergé, & Rilliard, 2008). We make the hypothesis that this competence of playing emotions could be used for simulated emotions and that different types of processes exist for authentic vs. simulated expressions. In this view, reproduction of an emotion, sincerely or not, when it occurs outside this "body loop," involves the same type of control as that which produces attitudinal affects that have been specifically built by the language and the culture. Therefore, as all socially encoded tools, such attitudinal expressions must be learned during the child's developmental phase, and they also must be learned by foreign language learners when these attitudes are not shared by L1 and L2 languages (Daneš, 1994).

Recent finding by Paulmann, Schmidt, Pell, and Kotz (2008), based on an event-related potentials (ERP) study, have shown that listeners are able to distinguish neutral prosody from an emotional one (whatever the valence and the label of the emotion) as soon as 200 ms (approximately one syllable) after the stimulus onset. This finding critically differs from the gating experiments held on attitudinal speech (Shochi, 2008) where listeners may decode the attitudinal meaning at various points in time within the utterance, for example, from the beginning of the utterance (in the same way reported by Paulmann et al. for emotions), or gradually throughout the utterance, or even at a specific point in the utterance, either at the middle or end. We therefore propose that attitudes and emotions can be encoded by the producer, using the same dimensions, but—to some extent—may be perceived differently by the receiver (cf. Aubergé & Cathiard, 2003). In other words, the same acoustic and visual dimensions may convey amongst other things both attitudinal and emotional information, but these two kinds of information may be differentiated by the variation of the parameters through time, which are controlled differently. For example the affect labeled as "*surprise*" may be either an attitude or an emotion, according to the control exerted by the speaker on his expression. He may want to voluntarily express some surprise in order to perform an illocutory act (Daneš, 1994), and therefore the prosodic parameters carrying the surprise information will vary according to the linguistic structure of the utterance. Or he may be really surprised and express this surprise involuntarily: the same prosodic dimensions may be used to carry the emotional information, but they may have a different morphology, according to the time course of this emotional surprise. These different expressions of surprise may be recognized as an attitude or an emotion due to the existing correlation amongst the variations of the prosodic dimensions, the linguistic structure of the utterance, and/or the occurrence of a "surprising" event.

Studies on such attitudes, or *social affects*, are important because an attitude is part of the global meaning of speech acts (Daneš, 1994): even if the speaker does not express any attitude by performing a simple declaration, he can still be expressing an attitude, for example, "*the speaker decides not to give information on his attitude*."

But, as claimed by Pavlenko (2005) while advocating the necessity of cross-cultural studies on language and emotions, the role of languages and cultures on the production and perception of affects has to be taken into account: some affects tend towards universality, whereas others seem more specific to one culture (see also Daneš, 1994) for their meaning as well as for their encoding. Therefore, similar to emotions, the study of attitudinal expressions may benefit from a cross-cultural approach (e.g., Shigeno, 1998): attitudinal values can exist or not in one language, and their audiovisual expression in a specific language may not be recognized by foreign speakers—or may be ambiguous in the learner's language (Scherer et al., 2001; Shochi et al., 2006, for a comparison of Japanese, English, and French).

Such a cross-cultural approach recalls the works of Ekman (1999) on facial expression of emotions amongst cultures, which led us to examine what modalities carry the affects to the interlocutor. Attitudinal expressions are linked to language and as such are classically described as one of the basic functions of prosody (Fónagy, 2003; Rossi, Cristo, Hirst, Martin, & Nishinuma, 1981). This expressive function has been studied for a long time with regard to its acoustical aspect (cf. Allerton & Cruttenden, 1978; Danes, 1994; Fónagy, Bérard, & Fónagy, 1984). But more recent works emphasized the multimodal nature of prosody's functions: for example, the one of demarcation (Barkhuysen, Krahmer, & Swerts, 2006), or the feeling of knowing (Swerts & Krahmer, 2005). It seems obvious that in face-to-face interactions, attitudes are perceived within the multimodality of speech (Barkhuysen, Krahmer, & Swerts, 2007a).

Numerous studies of attitudes or social affects have been done (e.g., Bänziger & Scherer, 2005; Campbell, 2005; de Moraes, 2008; Morlec, Bailly, & Aubergé, 2001; van Heuven, Haan, Janse, & van der Torre, 1997) as well as in cross-linguistic contexts (Barkhuysen et al., 2007a; Shochi, Aubergé, & Rilliard, 2007), but mainly in their acoustic modality only. As the study of the multimodal expression of affects (in the broad sense) is still a recent field of research (Scherer, 2003; Scherer & Ellgring, 2007), only a few studies question directly the specific topic of social affects, as differentiated from that of emotions (Granström & House, 2005).

As attitudes belong to language, are not the expression of an emotional state, are learned by the speaker, and are able to be produced voluntarily during an interaction, we assume that attitudinal expressions recorded in a lab might not suffer from the same bias as much as do an acted emotional corpus, as has been observed to be different for spontaneous emotional speech (Barkhuysen, Krahmer, & Swerts, 2007b; Wilting, Krahmer, & Swerts, 2006).

Given the particular importance of multimodality in affective communication (Scherer & Ellgring, 2007), this research examines the differences between the production and perception of multimodal expressions of social affect in both Japanese and French. This study aims at investigating: (1) the relative contribution of audio and visual modalities to prosodic attitudes; (2) whether the facial indices may have a significant impact on the perception of prosodic attitudes; and (3) the influence of speaker's performance on the recognition of attitudes. The recognition tests used in this study are described by Shochi, Erickson, Rilliard, Aubergé, and Martin (2008) for Japanese and Rilliard, Martin, Aubergé, & Shochi (2008) for French. We reported the observed results, including acoustical and video analyses of the corpora, in order to compare them to the subjective results. We also discuss the differences and the similarities in audio and visual modalities for both Japanese and French data.

2 Japanese and French corpus

A limited set of audovisual material was recorded in both Japanese and French for this study. The attitudinal expressions were recorded on utterances with an affectively neutral meaning in order to avoid interference between lexicon and prosody (Banse & Scherer, 1996; Mozziconacci, 1998). Lexically expressive sentences may introduce a bias both in the location of prosodic information on the lexically loaded word and in the perception test, as listeners may be influenced by lexical cues, as well as prosodic ones (see Grichkovtsova et al., 2007, for examples of such interaction between prosodic and lexical levels). In order to ease the objective comparison of prosodic parameters, speakers produced each utterance with all the attitudes chosen for one language.

The choice of attitudes for each language is based on didactic foreign language literature (see details for each language in subsequent sections), which provides an empirical field analysis of the culturally-controlled expressions for a language. Our selection of attitudes for the two languages is therefore based on this knowledge, which explains the different sets of attitudes observed for the two languages. Since the literature is intended to help teachers express these attitudes as naturally as possible for the learners, hints for the communication situation as well as for the way to produce the expressions in speech, gesture, and face, are also provided. In addition, the recordings used in this study are based on previous works on prosodic attitudes (Shochi, et al., 2006, for Japanese and Morlec et al., 2001, for French). The main differences between the present work and the two preceding ones are: (1) audovisual recording of the attitudes; (2) recording of two different speakers in order to measure the influence of individual performance over recognition scores; and (3) recording paradigm designed to set the speaker in a somewhat natural condition of production for each attitude. Whereas both Morlec et al. (2001) and Shochi et al. (2006) recorded the audio-only attitude by eliciting them sitting alone in a sound-proof room producing the required sentence with the required attitude, in the present study, the speakers were instructed to produce each sentence in order to express one attitude, as an answer to a statement produced by a partner—that is, we tried to reproduce a basic communication context suitable for eliciting the production of the expression. A training session, in which speakers were asked to behave as naturally as possible, without any constraints on their expressive audiovisual strategy, preceded the actual recordings. Since this work aimed at recording those attitudes specifically prescribed by didactic language teaching materials, we attempted to use trained language teachers, not actors.

2.1
Selection of 12 Japanese attitudes

A set of 12 Japanese attitudes which were validated in Shochi et al. (2006) was used. These attitudes were selected according to the literature (Erickson, Ohashi, Makita, Kajimoto, & Mokhtari, 2003; Maekawa, 1998; Sadanobu, 2004) and Japanese language teaching methods (Mizutani & Mizutani, 1979): *doubt-incredulity* (DO), *obviousness* (EV), *exclamation of surprise* (SU), *authority* (AU), *irritation* (IR), *arrogance* (AR), *sincerity-politeness* (SIN), *admiration* (AD), *kyoshuku* (KYO), *simple-politeness* (PO), *declaration* (DC), and *interrogation* (IN) (see Shochi et al., 2006, for definitions). Some of these attitudes are specific or specifically important for the Japanese culture, especially those linked to the politeness strategy: *simple-politeness*, *sincerity-politeness* and *kyoshuku* vs. *arrogance*. The sincerity-politeness attitude appears when a socially inferior speaker is talking to someone superior to him in Japanese society: the speaker expresses a serious and sincere intention by using this prosodic attitude. The *kyoshuku* attitude (there is no lexical entry to translate this in English) is a typically Japanese cultural attitude. Even if such situations occur in all cultures, the Japanese language has chosen to encode this situation as a prosodic attitude (“attitudineme”). A speaker uses *kyoshuku* when he wants to express a conflicting opinion to an interlocutor considered socially superior aiming to not disturb him but to help him, or when the speaker desires to get a favor from his superior. It is described by Sadanobu (2004, p.34) as “a mixture of suffering ashamedness and embarrassment, [which] comes from the speaker’s consciousness of the fact that his/her utterance of request imposes a burden to the hearer.”

2.2
Selection of six French attitudes

Following the work done by Morlec et al. (2001) on French prosodic attitudes, and based on studies by Fónagy et al. (1984), Calbris and Montredon (1981), Callamand (1973), and Calbris and Porcher (1989), six attitudinal expressions were selected for recording a French audiovisual corpus: *declaration* (DC), *interrogation* (IN), *obviousness* (EV), *surprise exclamation* (SU), *doubt-incredulity* (DO), *suspicious irony* (SC). These attitudes are defined as follows. *Declaration* is used by the speaker to give some simple information, without expressing any point of view. With *interrogation*, the speaker asks for information without expressing any point of view, and waits for a simple answer. *Surprise* is an expression of amazement when something unexpected happens suddenly. The speaker expresses *obviousness* when he says something he feels is self-evident. *Doubt-incredulity* is used to express a feeling of being uncertain about something or of not believing something that has previously been expressed. With *suspicious irony*, the speaker calls into question (and even contradicts) a statement his interlocutor has made. He wants to show he doesn’t agree, that he’s annoyed with what he’s hearing, and condemns what is said or what happened by using an intonation contour which could mean: “Yeah, sure! I believe you . . . (meaning, I don’t).”

2.3
Corpus recording

Two male Japanese native language speakers (SJ1 and SJ2) produced each sentence (detailed later) with the 12 attitudes. SJ1 is a Japanese native language teacher who teaches various attitudes in his class using pragmatic explanations; SJ2 is a naïve native speaker with no teaching experience. Two male French native speakers, SF1 and SF2, the former an experienced teacher, the latter not experienced, recorded the French corpus. All four speakers were recorded in a soundproof room at LIMSI, France. Three sessions were recorded for each speaker (either Japanese or French speaker), separated by pauses and with the possibility for speakers to watch their performance. During the recording, speakers 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 a USBPre-sound device connected to a computer outside the room, recording the speech signal at 44.1 kHz, 16 bits. A digital DV camera (Canon XM1 3CCD) recorded the speakers' performances. Hand claps between each sentence, recorded both by the camera and the microphone, allowed a post-processing, a replacement of the camera sound by the high-quality sound recorded by the microphone, synchronized with the claps. Video clips were encoded with a cinepack codec with a 784 × 576 pixels resolution.

For the Japanese attitudes, one eight-mora sentence was selected from a corpus developed and validated in a previous study (Shochi et al., 2006). This sentence ("*Nagoyade nomimas*" [nagojade nomimas], meaning "He drinks in Nagoya") was constructed on a verb–object syntactic structure. The lexical stress position was located on the first mora. In order to express some attitudes like *doubt* or *surprise*, the vowel [u] may be inserted at phrase final position, and in this case, the lexical stress would be realized at the seventh mora, also. The sentence was constructed in order to avoid any particular affective connotation in any region of Japan.

The French attitudinal corpus is based on three sentences of four, five, and seven-syllable length, without any specific affective meanings that can bootstrap or forbid one of the six attitudes The sentences were borrowed from a previous corpus intended to evaluate prosodic expressivity controlled by gesture (d'Alessandro, Rilliard, & Le Beux, 2007), and constructed with the same principle used for the Japanese corpus. After the recording and the post-processing, the speakers' performances were judged by each of the speakers, and only the five-syllable length sentence was kept for the perception test, because of its naturalness and rich audovisual cues: "*Nicolas revenait.*" [nikola ʁəvnɛ] ("Nicolas was coming back"), played with the six attitudes. Six short videos were thus produced for each speaker.

3 Perception test

3.1 Experimental design

An evaluation test was designed in order to evaluate the relative efficiency of the two modalities to carry the attitudinal information for each language separately. Listeners were all native speakers of the language they listened to, that is, French listeners listened to French attitudes whereas Japanese listeners judged Japanese attitudes. The factors controlled during the experiment are:

- the attitudes (6 for French or 12 for Japanese);
- the speakers (2 for each language);
- the modality (Audio, Visual or Audiovisual);

- the modalities' presentation order (audio first, then visual and audovisual; or visual first, then audio and audovisual).

Subjects listened to each stimulus only once for each modality, presented in a random order. For each stimulus, they had to select the perceived attitude 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 used for the five not-selected attitudes). Subjects had to answer to the test on a PC without any time constraint.

For each language, two groups of subjects took the experiment. The first group listened to the audio-only stimulus first, 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 the audio-video stimulus. This enabled us to counterbalance a possible effect of the presentation order of the modalities. During the presentation of one modality, the stimuli corresponding to all attitudes and to the two speakers were randomized in a different order for each listener.

3.2 Subjects

For the Japanese part, 46 Tokyo dialect speakers (mean age = 18.7) participated in this experiment. They were separated into two groups (Audio-only first and Visual-only first) with 28 subjects (4 males and 24 females) for the audio-only condition, and 18 subjects (7 males and 11 females) for the visual-only condition.

For the French part, 32 French listeners (17 males and 15 females, mean age = 32) took the experiment, 16 in each group (7 females and 9 males in the Audio-only-first group and 8 females and males in the Visual-only-first group).

3.3
Statistical processing

Results given by listeners are expressed by two measures: a simple categorical choice (the perceived attitude), and a relative intensity score for the selected 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 confusion matrices, grouping the categorical answers given by listeners for each of the 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. Only the categorical answer (and not intensity scores) will be detailed, as there is no major difference between the two types of measures.

For each language, recognition rates (either categorical or intensity) are analyzed using a repeated-measure ANOVA, which takes as a dependent variable the recognition rate of each attitude (expressed in percentages), *subjects* as a random effect, one between-subject factor (the *group*, that is, the order of presentation of the Audio and Video modality for each group of subjects), and three within-subject factors (the 6 or 12 *Attitudes*, the 2 *Speakers* and the 3 *Modalities*). ANOVA analysis mainly aims at measuring the relative importance of the different factors of the listener's behaviour.

Confusion matrices are analyzed using a correspondence analysis and a cluster analysis (Benzecri, 1973). Both of these methods are based on data-reduction techniques

Table 1

ANOVAs results for Japanese and French expressions, and for categorical and intensity results. Factors used are Grp for the listener group (either audio-only or video-only first); Spk for the two speakers' Mod for the three modalities; Att for 6 or 12 attitudinal expressions. Factors having a significant influence on results ($p < .01$) are marked with a star

	French					*Japanese*				
	% Reco			*Intensity*		*% Reco*			*Intensity*	
	d.f.	*f*	*p*	*f*	*p*	*d.f.*	*f*	*p*	*f*	*p*
Grp	1	0.5	0.508	0.3	0.605	1	0.08	0.779	6.8	0.012
Spk	1	45.3	0.000*	105.5	0.000*	1	163.60	0.000*	211.5	0.000*
Grp:Spk	1	1.9	0.178	7.1	0.012	4	0.02	0.899	0.5	0.481
Mod	2	14.9	0.000*	25.6	0.000*	2	44.8	0.000*	44.0	0.000*
Grp:Mod	2	3.8	0.028	3.7	0.030	2	1.2	0.2999	1.3	0.297
Att	5	6.4	0.000*	11.8	0.000*	11	23.5	0.000*	22.5	0.000*
Grp:Att	5	2.2	0.055	0.9	0.466	11	1.5	0.134	0.8	0.597
Spk:Mod	2	1.7	0.199	3.0	0.059	2	6.8	0.000*	12.9	0.000*
Grp:Spk:Mod	2	1.0	0.364	0.0	0.994	2	0.15	0.856	0.09	0.911
Spk:Att	5	3.8	0.003*	7.5	0.000*	11	11.0	0.000*	17.9	0.000*
Grp:Spk:Att	5	0.7	0.589	0.5	0.764	11	1.4	0.188	2.1	0.018
Mod:Att	10	7.9	0.000*	8.6	0.000*	22	3.7	0.000*	5.0	0.000*
Grp:Mod:Att	10	1.3	0.246	1.4	0.195	22	1.2	0.227	1.0	0.442
Spk:Mod:Att	10	5.4	0.000*	3.6	0.000*	22	7.3	0.000*	8.6	0.000*
Grp:Spk:Mod:Att	10	1.5	0.146	1.7	0.085	22	0.95	0.522	1.0	0.441

that allow a more simple and comprehensive representation of the data. A correspondence analysis allows a graphical representation of the perception results in order to analyze the self-recognition of a particular attitude and its relative dispersion. A cluster analysis hierarchically groups the different stimuli in clusters. The distances between the clusters indicate the perceptive distances between the corresponding attitudes (the Ward distance metric is used for clustering), and thus allow distance judgments.

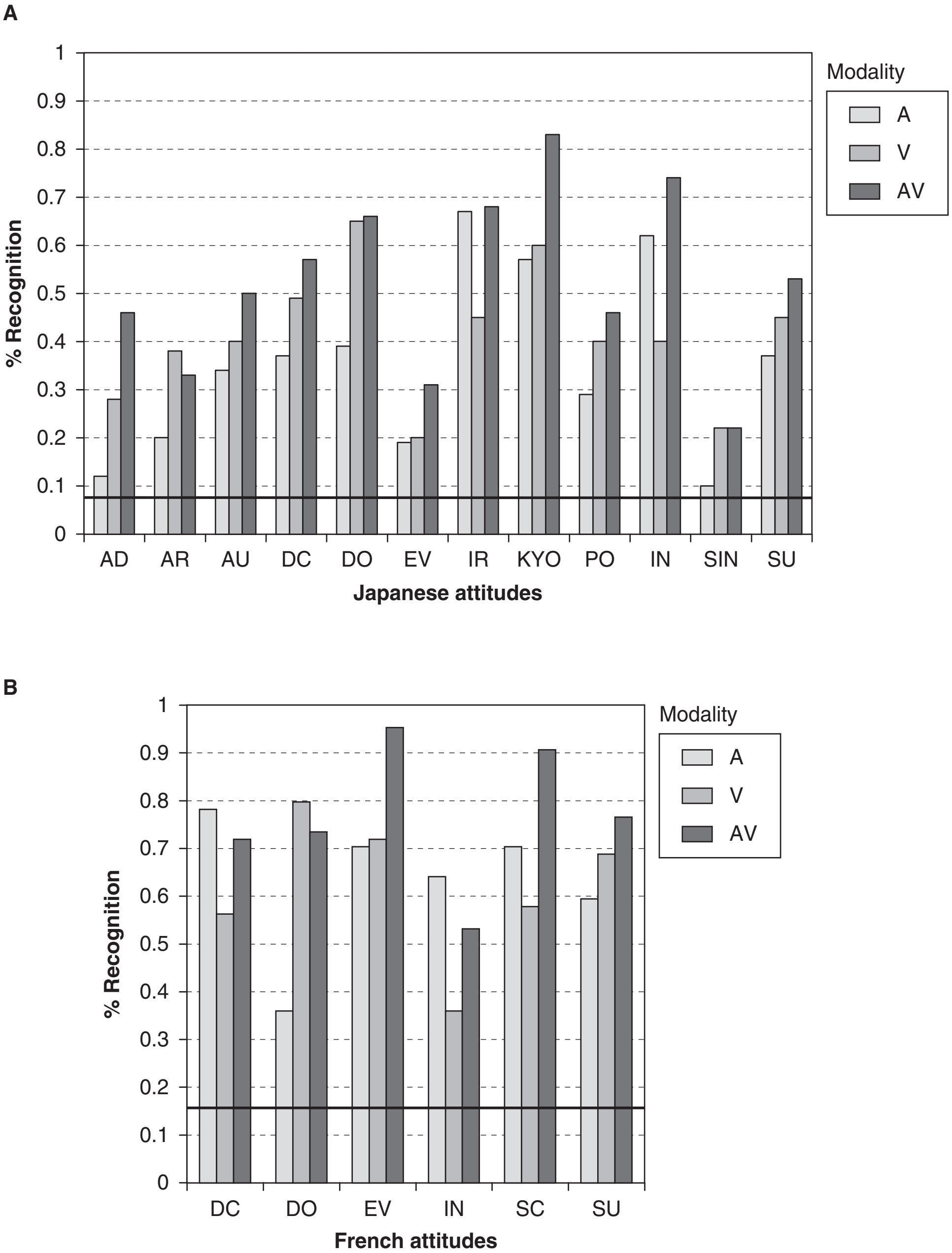
3.4
ANOVA results

Results of the ANOVAs for Japanese and French are summarized in Table 1. For both languages, the effect of the controlled factors are almost identical: all factors show a strong effect on the results, except the *group* factor (corresponding to the different order of presentation of modality for each group), which is never significant at $p < .01$. Interactions between factors (except the group) are all significant, except the interaction between *speaker* and *modality* for French only.

The *attitude* and *modality* factors and their interaction are significant for both languages. Results of the interaction are presented in Figure 1. In their multimodal presentation, all attitudes receive recognition scores higher than chance level, indicating an overall good recognition of the performances even if for some speakers, in some modalities, and for some attitudes, scores may be low.

Figure 1

Percentage of recognition obtained by each attitude, for each modality, in Japanese (top) and French (bottom). The plain horizontal bar on each graph represents the chance level



Amongst the five attitudinal labels represented in both languages (i.e., *declaration*, *doubt*, *obviousness*, *interrogation* and *surprise*), all except *declaration* show a similar influence of the modality on perception results for Japanese and French. Best performances for *doubt* are achieved with visual information; *obviousness* and *surprise*, mainly with audovisual; and *interrogation* mainly with acoustic information. This is a first indication of the coherence of the multimodal data recorded in this corpus. The relative contribution of modalities to the different attitudes will be studied later, together with the analysis of these modalities.

Another factor having a main influence on the perception results is the speaker's performance. For Japanese, the naïve speaker (SJ2) received lower overall recognition scores than the trained one. The strategy of SJ2 regarding the use of each modality is significantly different from the one of SJ1. For French, there is also one speaker (SF1) who systematically received better recognition scores than the other speaker (SF2), even though their strategy towards the two modalities seemed coherent (non significant interaction between the *speaker* and the *modality* for French). Therefore, for the objective analysis, only the data from SJ1 and SF1 will be discussed, in comparison with their perception scores. The reason for this is because we focus on evaluating the possible influences of the visual parameter on prosodic attitude, and not speaker differences. Even if our experiment may show that speaker performance is critical, our data is not sufficient for extracting relevant information concerning this point.

3.5 Analysis of confusion matrices

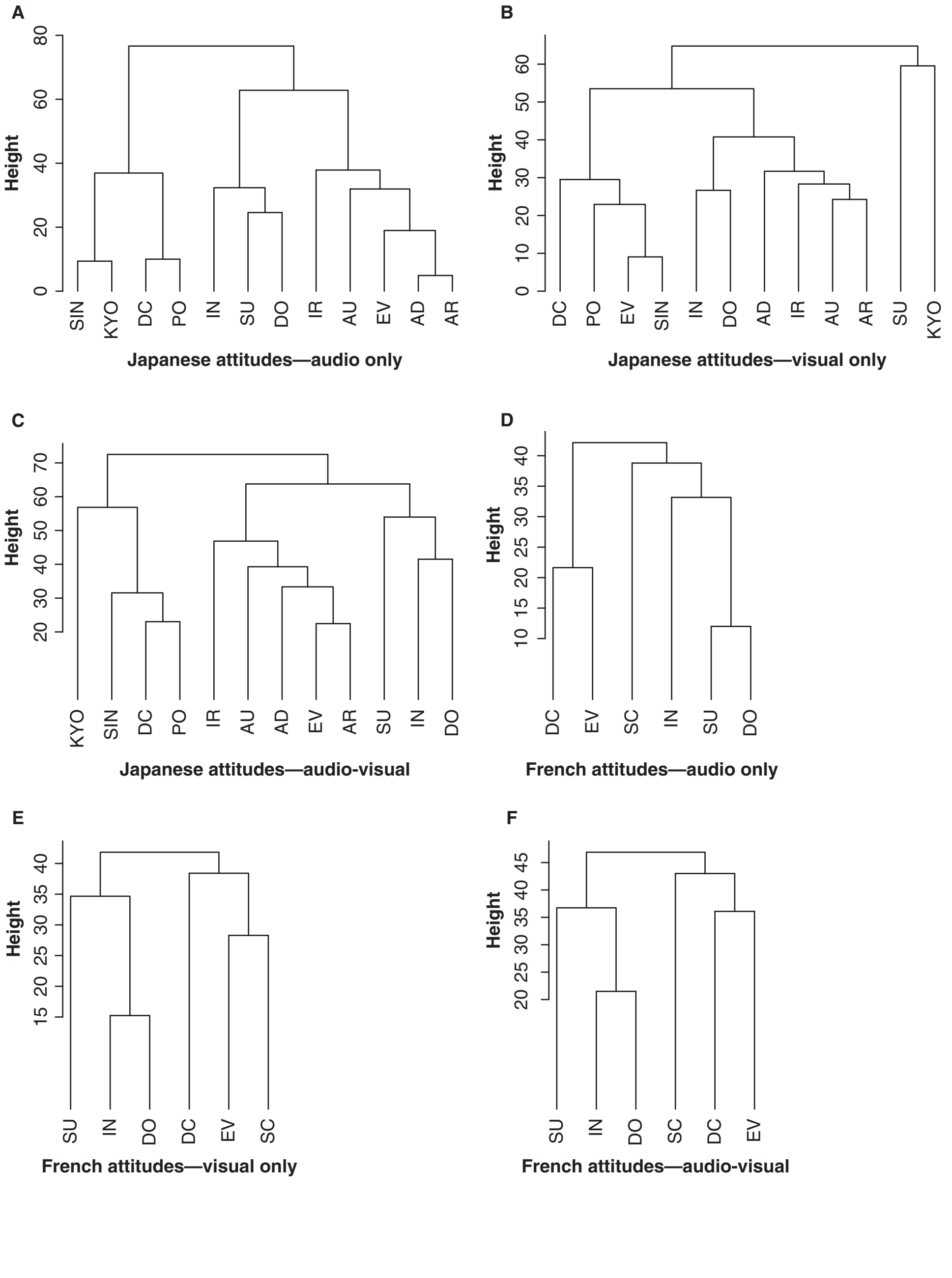
Analysis of the confusion matrices leads to interesting parallels between attitudinal expressions: data reduction techniques allow grouping the attitudinal expression inside a more general cluster, thus enhancing the main perceptive proximities. The results of the clustering analysis are presented here, for both Japanese and French, and for each modality (cf. Figure 2).

For the Japanese speaker, SJ1, the recognition of each attitudinal expression is over the chance level in each modality, with few exceptions (especially the audio-only *admiration*). *Admiration* is perceived as *arrogance* in the audio-only condition, but visual information allows a clear distinction, even if they stay at a global level in the same cluster. Three general clusters emerge from the audio-only and the audio-video analyses. The first cluster groups the politeness expressions plus declaration; whereas the remarkably well-recognized video-only *kyoshuku* is outside this cluster, and the video-only *obviousness* is mixed up with *sincerity-politeness*. The second cluster groups the expression of "query" (*interrogation*, *doubt*, *surprise*), with video-only *surprise*, also recognized without confusion. The third cluster groups *admiration* with attitudes described by Shochi et al. (2008) as "*express[ing] the imposition of the speaker's opinion*" (*arrogance*, *authority*, *obviousness*, *irritation*).

For the French speaker SF1, each attitude is recognized over chance, whatever the modality. The main confusion arises for the audio- and video-only expressions of *doubt*, mixed with *surprise* for the audio-only modality, and with *interrogation* for the video-only modality. At a global level, two main clusters appear, plus the *suspicious irony* expression, always well distinguished from the other attitudes. The first cluster

Figure 2

Hierarchical clustering of attitudinal expressions obtained from recognition scores, for speakers SJ1 and SF1, and for each modality (audio, visual, and audiovisual). The height indicates the relative distances between each attitude or clusters of attitudes



groups the declarative expression of *obviousness* and *declaration*, whereas the second cluster groups the interrogative attitudes of *surprise*, *doubt* and *interrogation*. This configuration remains stable for each modality, except for the visual-only modality, where *declaration* is perfectly recognized but *obviousness* and *suspicious irony* show some confusion.

4 Objective parameters and analysis

4.1 Objective analysis

4.1.1 Extraction of acoustic parameters

Acoustic parameters of prosody were extracted automatically from the hand-labeled signals, with Matlab scripts using the yin algorithm of fundamental frequency (F0) extraction (cf. de Cheveigné & Kawahara, 2002). Three parameters are extracted: F0 (in semitones), duration (in seconds—either syllabic for French or moraic for Japanese), and intensity (in dB). For both F0 and intensity parameters, three values for each vowel were calculated.

For each parameter, we calculated the maximum, minimum, mean, and range, and also the slope of the curve (i.e., the mean value of the first vowel minus the mean value of the last one). These five values for each of the three acoustic parameters are recorded for each attitudinal expression, and used as entries for the principal component analysis.

Due to the complexity and as yet not well-known approach to analyzing voice quality (e.g., d'Alessandro, 2006), the study of voice quality cues is reserved for future work. Moreover, although it is clear that voice quality has a significant impact on perception of Japanese attitudes (Shochi, 2008), it is probably not that useful for French attitudes (Morlec et al., 2001).

4.1.2 Extraction of visual Action Units (AUs)

A researcher with some knowledge about the Facial Action Coding System (FACS) (not a certified FACS coder) from the research team viewed each video and marked appearance of AUs related to the upper face, lower face, and head positions based on appearance changes (intensity was not scored) according to the FACS Manual (Ekman, Friesen, & Hager, 2002). The list of AUs used for this analysis, and their labels, are presented in Table 2.

4.2 Principal Component Analysis (PCA) on audio and video parameters

Using the objective parameters described above as features, separate Principal Components Analyses were run. Results (cf. Figures 3 and 4) present the relative ability of these features to separate each set of attitudes. We compare the results of the PCAs and the results of the clustering analysis (based on perception results).

Table 2

List of visual AU used during the labeling—labels and description

AU's label	*Description of the AU*
AU I+2	Inner + outer brow raiser
AU 4	Brow lowerer
AU 5	Upper lid raiser
AU 6	Cheek raiser
AU 11	Nasolabial deepener
AU 12	Lip corner puller
AU 15	Lip cirber depressor
AU 26	Jaw drop
AU 43	Eyes closed
AU 51	Head turn left
AU 57	Head forward
SH	Shoulder shrug
HN	Head nod

4.2.1 Japanese attitudes

PCAs based on F0 parameters show the following distinctive features. Both a high range of F0 and an important maximum F0 value characterize *doubt* (with the highest range) and *surprise* expressions. Expressions with a lower F0 values are *authority*, *sincerity*, and *kyoshuku*. The general slope of the curve relates to *obviousness* and to a lesser degree to *declaration*. The F0 mean does not have any particular distinctive power.

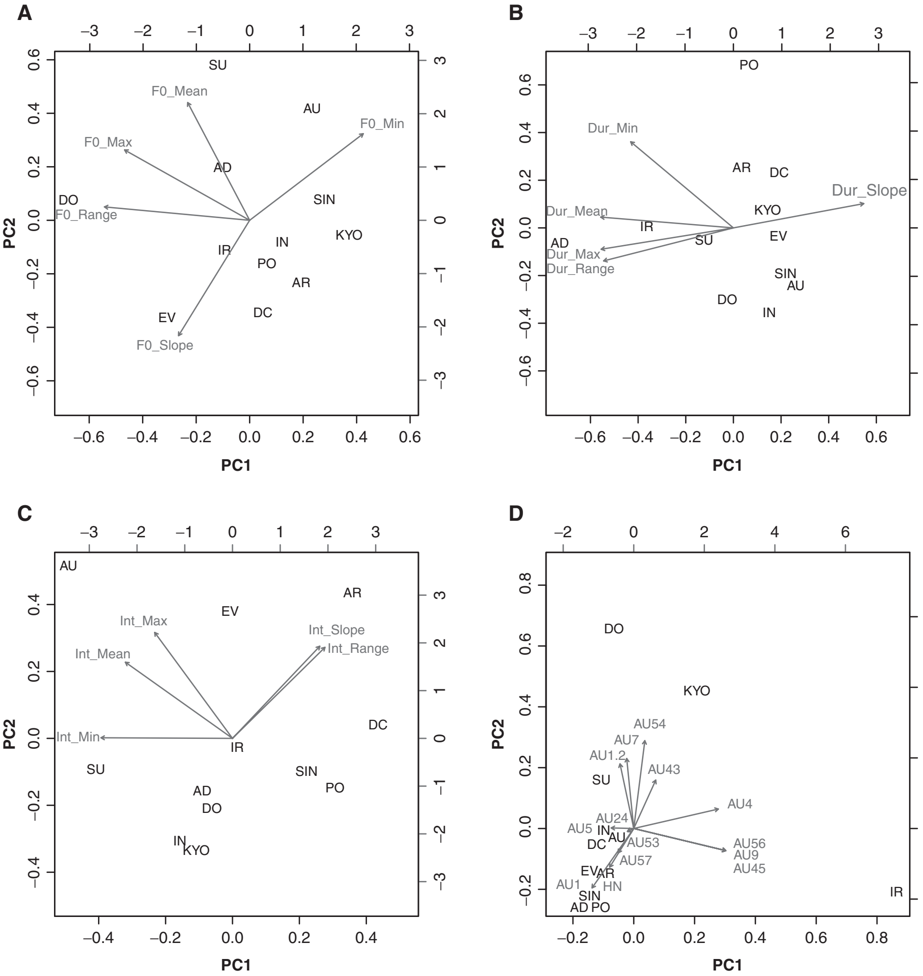
Moraic duration patterns mainly separate *admiration* from the others, with the most important duration range, due to an impressive lengthening of the last syllable, also resulting in the highest maximum of duration (and therefore the most important negative slope). A slightly positive duration slope corresponds systematically to sentences pronounced with the last mora based on a simple [s], without added [u], that is, to *arrogance*, *authority*, *declaration*, *sincerity*, and *kyoshuku*.

The intensity parameter mainly separates the most intense expressions, with important intensity mean and maximum (mainly *authority*, and also *obviousness* and *surprise*) from the others. It also separates expressions with a positive slope (i.e., lowest voice at the end): *arrogance*, *declaration*, and *obviousness*; from the expressions with a rising voice strength: *kyoshuku*, *interrogation*, *surprise*, and *doubt*.

Finally, the visual AUs clearly distinguish *irritation* and *kyoshuku* as specific expressions. Then, it groups the questioning attitudes (*interrogation*, *surprise*, and *doubt*), and makes a clear cluster composed of *politeness*, *admiration*, *sincerity*, and *arrogance* around AU 1. *Declaration* and *authority* do not feature any particular facial AU. The somewhat restricted amount of information obtained from the AUs' labeling does not allow a sufficiently clear distinction of Japanese attitudes, even though listeners showed good results. Future work along these lines may require a more detailed analysis of visual information.

Figure 3

Two main dimensions of PCAs for speaker SJ1, showing the relative contribution of the following parameters to each attitude: F0, moraic duration, intensity (mean value, maximum value, minimum value, and slope for each parameter), AUs (as listed in Table 2)

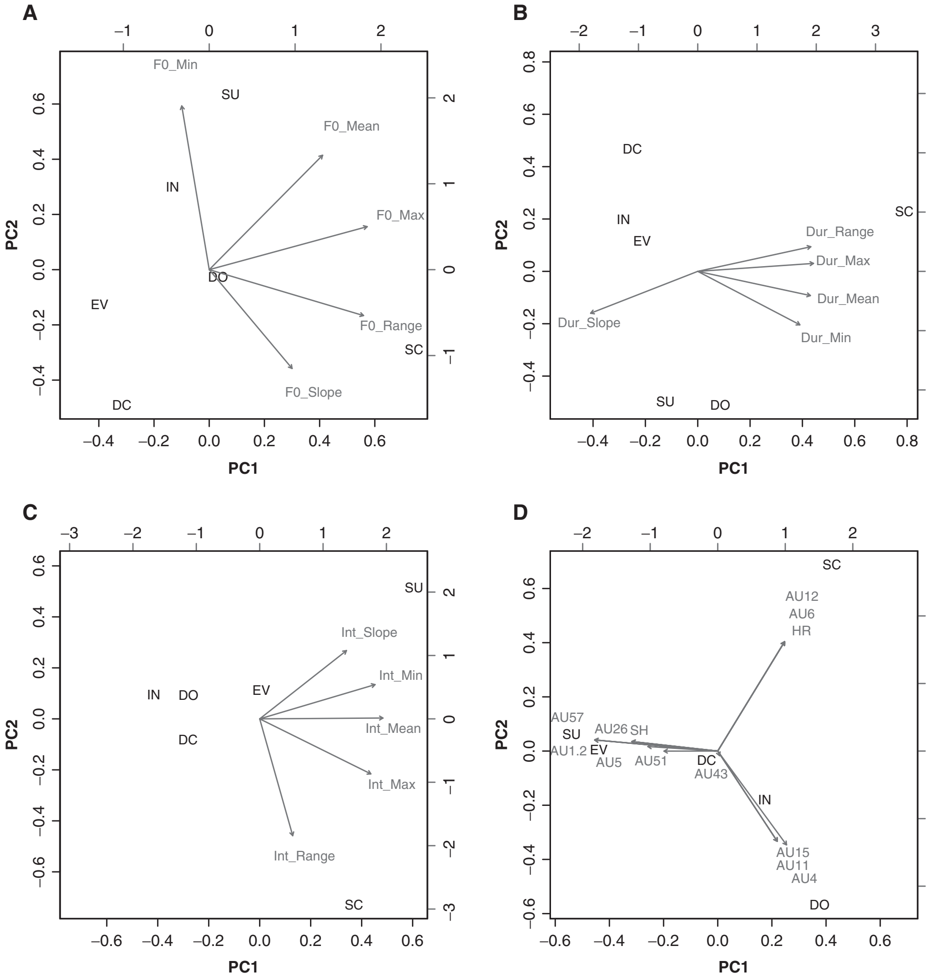


4.2.2 French attitudes

Analysis of the F0 parameters underlines the main influence of F0 in the attitudinal expression of French attitudes, already shown by Morlec et al. (2001). *Suspicious irony* is apart from the other attitudes, with the highest F0 range, an important F0 maximum and a decreasing F0 (positive slope). Declaration is also separated from the other attitudes, with a positive slope and a generally low F0 curve. Interrogative expressions (*surprise*, *doubt*, and *interrogation*) are grouped together, with a negative slope—that is, a rising final F0.

Figure 4

Two main dimensions of PCAs for speaker SF1, showing the relative contributions of the following parameters to each attitude: F0, moraic duration, intensity (mean value, maximum value, minimum value, and slope for each parameter), AUs (as listed in Table 2)



Syllabic duration parameters also separate *suspicious irony* from other, with the highest duration range and the longest syllable (at the end). Other expressions generally show small changes in their durational patterns, except for *surprise* with a positive slope.

Intensity parameters distinguish two expressions: *suspicious irony* and *surprise*. Both attitudes have a high intensity (mean and maximum), but *suspicious irony* has a large range of intensity, whereas intensity for *surprise* is more constant.

Four kinds of facial expressions are distinguished: (1) the *doubt* and *question* expressions; (2) the *surprise* and *obviousness* expressions, that are both expressed with AU 1+2 and AU 57; (3) *suspicious irony*; and (4), finally, *declaration*, that does not display any facial AU.

4.3 Analysis and comparison with subjective results

By comparing the distribution of attitudinal expressions made on the basis of objective and subjective results, it is obvious that results are more coherent for French than for Japanese. That can be explained by the two main drawbacks of our analysis: the absence of voice quality measurement for Japanese, and the fact that our AU labeler may have been less acute on Japanese stimuli, as he is not a speaker of Japanese, and may have been "blind" to some pertinent cues. An automatic detection of the visual parameters which we consider for future research may be more suited to such a study.

For French, F0, duration, and AU, as well as perception results, clearly distinguish *suspicious irony* from others, and oppose interrogative expressions and declarative ones. The main difference between objective and subjective results concerns the facial expressions of *obviousness* and *surprise*, performed with two identical AUs, amongst others. One of the main contributions of facial information is the very clear distinction allowed by these cues of the expressions of *doubt*, *obviousness*, and *suspicious irony*. In preceding works (Morlec et al., 2001), both *doubt* and *suspicious irony* show some confusion (in audio), while for our speaker, *doubt* and *obviousness* are confused in the audio-only condition. It seems then obvious that visual information brings a clear distinction between the more expressive attitudes, while they are less relevant for the two modalities (non-expressive attitudes) of *declaration* and *interrogation*.

The analysis of the Japanese data gives some hints about the main characteristics of attitudinal prosody, but the description of prosodic parameters seems too gross to catch the complexity of all these expressions; specifically it suffers from the size of the corpus, which is too limited to extract statistically relevant data. Two essential measurements may need to be done: a measure of the voice quality dimension, of primary importance for the perception of some Japanese attitudes like *kyoshuku* (cf. Shochi, 2008); and a description of intonation in terms of contours, not only using F0 values averaged on the utterances. However, the results obtained in the audio-only and with the audovisual condition are close to those obtained by Shochi et al. (2006) in an audio-only condition. Furthermore, a set of AUs has been identified for each attitudinal expression. These AUs allow a straightforward identification of the *kyoshuku* expression, and this finding is important in the perspective of a cross-cultural perception of this culturally-built expression—absolutely not recognized in an audio-only condition by occidental listeners (Shochi, 2008). Other information carried by the visual modality allows also a clear distinction between assertive vs. dubitative expressions.

There are differences between Japanese and French, but also similarities. In both languages, *declaration* does not use a specific AU but is recognized in the video-only modality. This is a bridge to Danes (1994) assertion, that:

> We must rather assume that any utterance or higher discourse unit has an *emotional value* in its communicative situation, both on the producer's and the receiver's side (though not necessarily for both of them in a given case, or in the same measure). Thus, even the alleged "absence of emotional

> involvement" represents, in fact, an instance of the category of emotional state, and an utterance of this nature may carry—in certain contexts—a high emotional value. (1994, p.258)

Obviousness in both languages is difficult to characterize, and seems to heavily rely on the speaker's performance, whereas *surprise* expressions are performed with a rising F0 at the end of the sentence, and use the AU 1+2, and *surprise* is one of the few attitudes cross-culturally recognized by Japanese, American, and French in Shochi (2008). The dubitative expressions, *interrogation*, *surprise*, and *doubt* seem to behave similarly in Japanese and French: in the audio-only condition, *surprise* and *doubt* are close together, with relative proximity to *interrogation*, while with the visual information (visual-only and audovisual conditions), *interrogation* and *doubt* show a higher proximity and *surprise* receives better discrimination scores, even if it is generally still in the same cluster. The results of the audio-only condition recalls very interestingly those obtained by Shochi (2008), who obtained similar confusions between *doubt* and *surprise*, with a proximity to *interrogation* on Japanese, French, and British English, with native listeners as well as in cross-cultural conditions with Japanese, French, and American English listeners (with some minor differences for the Japanese listeners). This strong perceptual effect obtained on audio-only cues seems to be counterbalanced by the visual cues, which discriminate mainly the *surprise* expression. Finally, in both languages, there is a systematic opposition between the assertive and the dubitative expression, that can be enlighten by the classification of speech acts expressivity proposed by Brandt (2008), from a linguistic point of view. The Japanese language, moreover, adds a dimension of dominance, or of imposition of the speaker, orthogonal to the assertive-dubitative dimension. The French attitude of *suspicious irony* may be seen as one expression of this dimension.

5 Conclusions

Production and perception cues to Japanese and French attitudinal expressions display interesting similarities. First, the two modalities are widely used by all speakers, and are decoded adequately by listeners: almost all attitudes are recognized over the chance level in each single modality, and audovisual performances generally outperform those obtained with individual modalities. The different expressivity and strategy of the speakers also have a strong impact on the subjects' performances. Along these lines, future work would benefit from preparation of more naturalistic situations of interaction, and from more speakers in order to assess a wider range of possible strategies. This would allow us to determine some main coherences, which were not easily seen from the limited set of speakers we had in the current study.

Approaches of attitudinal expressions based on a collection of discrete categories may also raise some issues. It seems promising to restrict such a study to the analysis of one particular dimension that comes out of the results described in this article (such as the politeness–impoliteness dimension for Japanese in comparison with politeness in other languages—Loveday, 1981, such as French). Such a dimension may be more easily tested amongst languages (cf. Sagisaka et al., 2004), and its acoustic and visual correlates could also be extracted. We also intend to investigate the impact of speakers and listeners' personality traits and emotional intelligence on the audiovisual expression and perception of these attitudes as a means of understanding the individual differences that we observe, since some research has shown that personality traits of listeners might affect perception of attitudes (e.g., Cooper, 2002; Matthews, Zeidner, & Roberts, 2002).

Objective analysis based on static cues—either acoustic or visual ones—seems not to be sufficient to understand the performances of listeners: some of the main divergences can be explained, but not all of them. It seems of primary importance to add dynamic and global information to such an objective analysis (Aubergé, Audibert, & Rilliard, 2004) using an automatic labeling tool. For example, rapid eye movement might have a dramatic influence on the subjects' ratings, but are not taken into account in the current analysis. In a similar way, the F0 contour of the sentence was of importance for the expression of French attitudes (Morlec et al., 2001), but is unsatisfactorily described by the mean F0 values, used as descriptors of intonation in this study. The main problem for using such information is the difficulty in our current data reduction technique to deal with contour shapes, not just scales. Future work is required in order to improve the objective analysis, for example, by adding an intensity rating to the AUs, and analysing the dynamics of the audiovisual signs.

Once coherent sets of acoustic and visual cues are at our disposal, we plan to create controlled animated audovisual replays of the attitudinal expressions, using facial animation (Buisine et al., 2006; Martin, Niewiadomski, Devilliers, Buisine, & Pelachaud, 2006) and copy-synthesis approaches, in order to have a better control of the perceptual relevance of the different cues. Virtual characters, the expressivity of which would be controlled with hand-driven interfaces (d'Alessandro et al., 2007; Martin et al., 2007) may allow experimenters to design Wizard-of-Oz scenarios (Dahlbäck, Jönsson, & Ahrenberg, 1993) of man–machine interaction. In this situation, the time-course of the perceptually validated cues, relative to other information such as the syntactic structure of the utterance or the occurrence of an emotionally-loaded event, may be manipulated in order to evaluate if the dynamics of both audio and visual cues can induce the perception of the difference between spontaneous vs. controlled expressive states. Such studies on social affects provide additional knowledge about the audiovisual expression and perception of affects.

References

ALLERTON, D. J., & CRUTTENDEN, A. (1978). Syntactic, illocutionary, thematic and attitudinal factors in the intonation of adverbials. *Journal of Pragmatics*, **2**, 155–188.

AUBERGÉ, V. (2002). A Gestalt morphology of prosody directed by functions: The example of a step by step model developed at ICP. In B. Bel & I. Vincent-Marlien (Eds.), *Proceedings of Speech Prosody 2002* (pp.151–154). Aix-en Provence: Laboratoire Parole et Langage.

AUBERGÉ, V., AUDIBERT, N., & RILLIARD, A. (2004). Acoustic morphology of expressive speech: What about contours? *Proceedings of Speech Prosody 2004* (pp.201–204). Nara, Japan.

AUBERGÉ, V., & CATHIARD. M. (2003). Can we hear the prosody of smile? *Speech Communication*, **40**(1), 87–97.

AUDIBERT, N., AUBERGÉ, V., & RILLIARD, A. (2008). How we are not equally competent for discriminating acted from spontaneous expressive speech. *Proceedings of Speech Prosody 2008* (pp.693–696). Campinhas, Brasil.

BANSE, R., & SCHERER, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, **70**, 614–636.

BÄNZIGER, T., & SCHERER, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, **46**, 252–267.

BARKHUYSEN, P., KRAHMER, E., & SWERTS, M. (2006). How auditory and visual prosody is used in end-of-utterance detection. In *Proceedings of Interspeech 2006 – ICSLP, Ninth International Conference on Spoken Language Processing* (pp.1276–1279). Retrieved February 2009, from http://www.isca-speech.org/archive/interspeech_2006

BARKHUYSEN, P., KRAHMER, E., & SWERTS, M. (2007a). Cross-modal perception of emotional speech. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of PhoneticSciences* (pp.2133–2136). Retrieved February 2009, from http://www.icphs2007.de/conference/Papers/1482/index.html

BARKHUYSEN, P., KRAHMER, E., & SWERTS, M. (2007b). Incremental perception of acted and real emotional speech. In *Proceedings of Interspeech 2007, 8th Annual Conference of the International Speech Communication Association* (pp.1262–1265). Retrieved February 2009, from http://www.isca-speech.org/archive/interspeech_2007

BENZECRI, J. P. (1973). *L'analyse des données*. Paris: Bordas.

BRANDT, P. A. (2008). Thinking and language. A view from cognitive semio-linguistics. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of Speech Prosody 2008* (pp.649–654). Campinhas, Brazil: Editora RG/CNPq.

BUISINE, S., ABRILIAN, S., NIEWIADOMSKI, R., MARTIN, J.-C., DEVILLERS, L., & PELACHAUD, C. (2006). Perception of blended emotions: From video corpus to expressive agent. In J. Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier (Eds.), *Proceedings of 6th International Conference on Intelligent Virtual Agents* (pp.93–106). Berlin/Heidelberg: Springer.

CALBRIS, G., & MONTREDON, J. (1981). *Oh là là. Expression intonative et mimique*. Paris: CLE international.

CALBRIS, G., & PORCHER, L. (1989). *Geste et communication*. Paris: Hatier-Crédif.

CALLAMAND, M. (1973). *L'intonation expressive*. Collection le français dans le monde, B.E.L.C. Paris: Librairies Hachette et Larousse.

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.

COOPER, C. (2002). *Individual differences* (2nd ed.). London: Arnold.

DAHLBÄCK, N., JÖNSSON, A., & AHRENBERG, L. (1993). Wizard of Oz studies – Why and how. Paper presented at the *Workshop on Intelligent User Interfaces*, Orlando, FL, U.S.A.

d'ALESSANDRO, C. (2006). Voice source parameters and prosodic analysis. In S. Sudhoff, D. Leternova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, et al. (Eds.), *Methods in Empirical Prosody Research* (pp.63–87). Berlin/New York: Walter de Gruyter.

d'ALESSANDRO, C., RILLIARD, A., & Le BEUX, S. (2007). Computerized chironomy: evaluation of hand-controlled intonation reiteration. In *Proceedings of Interspeech 2007* (pp.1270–1273). Retrieved February 2009, from http://www.isca-speech.org/archive/interspeech_2007

DAMASIO, A. R. (1994). *Descartes' error. Emotion, reason, and the human brain*. New York: G. P. Putnam.

DANEŠ, F. (1994). Involvement with language and in language. *Journal of Pragmatics*, **22**, 251–264.

de CHEVEIGNÉ, A., & KAWAHARA, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, **111**, 1917–1930.

de MORAES, J. A. (2008). The pitch accents in Brazilian Portuguese: Analysis by synthesis. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of Speech Prosody 2008* (pp.389–397). Campinhas, Brazil: Editora RG/CNPq.

EKMAN, P. (1999). Facial expressions. In T. Dalgleish & T. Power (Eds.), *The Handbook of cognition and emotion* (pp.301–320). Sussex, U.K.: Wiley.

EKMAN, P., FRIESEN, W. C., & HAGER, J. C. (2002). *Facial action coding system*. Salt Lake City, UT: The Manual on CD ROM., Research Nexus division of Network Information Research Corporation.

ERICKSON, D., OHASHI, S., MAKITA, S., KAJIMOTO, N., & MOKHTARI, P. (2003). Perception of naturally-spoken expressive speech by American English and Japanese listeners. In N. Campbell (Ed.), *Proceedings of CREST International Workshop on Expressive Speech Processing* (pp.31–36). Kobe, Japan: Kobe University.

FÓNAGY, I. (2003). Des fonctions de l'intonation: essai de synthèse. *Flambeau*, **29**, 1–20.

FÓNAGY, I., BÉRARD, E., & FÓNAGY, J. (1984). Clichés mélodiques. *Folia Linguistica*, **17**, 153–185.

GRANSTRÖM, B., & HOUSE, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, **46**, 473–484.

GRICHKOVTSOVA, I., LACHERET, A., MOREL, M., BEAUCOUSIN, V., & TZOURIO-MAZOYER, N. (2007). Affective speech gating. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp.805–808). Retrieved February 2009, from http://www.icphs2007.de/conference/Papers/1539/1539.pdf

LOVEDAY, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates to English and Japanese politeness formulae. *Language and Speech*, **24**(1), 71–89.

MAEKAWA, K. (1998). Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. *Proceedings of the 5th International Conference on Spoken Language Processing* (pp.635–638), Sydney.

MARTIN, J.-C., d'ALESSANDRO, C., JACQUEMIN, C., KATZ, B., MAX, A., POINTAL, L., ET AL. (2007). 3D audiovisual rendering and real-time interactive control of expressivity in a talking head. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *7th International Conference on Intelligent Virtual Agents (IVA'2007)* (pp.17–19). Berlin/Heidelberg: Springer.

MARTIN, J.-C., NIEWIADOMSKI, R., DEVILLERS, L., BUISINE, S., & PELACHAUD, C. (2006). Multimodal complex emotions: Gesture expressivity and blended facial expressions. [Special issue on "Achieving Human-like Qualities in Interactive Virtual and Physical Humanoids," C. Pelachaud & L. Canamero (Eds.)] *Journal of Humanoid Robotics*, **3**(3), 269–291.

MATTHEWS, G., ZEIDNER, M., & ROBERTS, R. D. (2002). *Emotional intelligence—science and myth*. Cambridge, MA: MIT Press.

MIZUTANI, O., & MIZUTANI N. (1979). *Aural comprehension practice in Japanese*. Tokyo: The Japan Times.

MORLEC, Y., BAILLY, G., & AUBERGÉ, V. (2001). Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication*, **33**(4), 357–371.

MOZZICONACCI, S. J. L. (1998). *Speech variability and emotion: Production and perception*. Unpublished Ph.D. thesis, Eindhoven, The Netherlands.

OHALA, J. J. (1996). Ethological theory and the expression of emotion in the voice. In *Proceedings of the 4th International Conference on Spoken Language Processing* (pp.1812–1815). Retrieved February 2009, from http://www.isca-speech.org/archive/icslp_1996

PAULMANN, S., SCHMIDT, P., PELL, M., & KOTZ, S. (2008). Rapid processing of emotional and voice information as evidenced by ERPs. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of Speech Prosody 2008* (pp.205–209). Campinhas, Brazil: Editora RG/CNPq.

PAVLENKO, A. (2005). *Emotions and multilingualism*. Cambridge, U.K.: Cambridge University Press.

RILLIARD, A., MARTIN, J. C., AUBERGÉ, V., & SHOCHI, T. (2008). Perception of French audio-visual prosodic attitudes. In P.A. Barbosa, S. Madureira, & C. Reis (Eds.), *Speech prosody 2008* (pp.685–688). Campinas, Brazil: Editoria RG/CNPq.

ROSSI, M., DI CRISTO, A., HIRST, D., MARTIN, P., & NISHINUMA, Y. (1981). *L'intonation: de l'acoustique à la sémantique*. Paris: Klincksieck.

SADANOBU, T. (2004). A natural history of Japanese pressed voice. *Journal of the Phonetic Society of Japan*, **8**(1), 29–44.

SAGISAKA, Y., YAMASHITA, T., & KOKENAWA, Y. (2004). Speech synthesis with attitude. In *Proceedings of Speech Prosody 2004* (pp.401–404). Nara, Japan.

SCHERER, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, **40**(1–2), 227–256.

SCHERER, K. R., BANSE, R., & WALLBOTT, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural Psychology*, **32**(1), 76–92.

SCHERER, K. R., & ELLGRING, H. (2007). Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion*, **7**(1), 158–171.

SHIGENO, S. (1998). Cultural similarities and differences in the recognition of audiovisual speech stimuli. In *Proceedings of 5th International Conference on Spoken Language Processing* (pp.149–152). Sydney, Australia.

SHOCHI, T. (2008). *Prosodie des affects socioculturels en japonais, français et anglais: à la recherche des vrais et faux-amis pour le parcours de l'apprenant*. Unpublished Ph.D. thesis, University Grenoble 3, France.

SHOCHI, T., AUBERGÉ, V., & RILLIARD, A. (2006). How prosodic attitudes can be false friends: Japanese vs. French social affects. In R. Hoffmann & H. Mixdorff (Eds.), *Speech Prosody 2006 Abstract Book and CD-ROM Proceedings* (pp.692–696). Dresden: TUDpress Verlag der Wissenschaften.

SHOCHI, T., AUBERGÉ, V., & RILLIARD, A. (2007). Cross-listening of Japanese, English and French social affect: About universals, false friends and unknown attitudes. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp.2097–2100). Retrieved February 2009, from http://www.icphs2007.de/conference/Papers/1435/1435.pdf

SHOCHI, T., ERICKSON, D., RILLIARD, A., AUBERGÉ, V., &. MARTIN, J. C. (2008). Recognition of Japanese attitudes in Audio-Visual speech. P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Speech prosody 2008* (pp.689–692). Campinas, Brazil: Editora RG/CNPq.

SWERTS, M., & KRAHMER, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, **53**(1), 81–94.

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. In A. Botinis, G. Kouroupetroglou, & G. Carayiannis (Eds.), *Intonation: Theory, models and applications (Proceedings of an ESCA Workshop, September 18–20, 1997, Athens, Greece)* (pp.317–320). Athens: ESCA and University of Athens Department of Informatics.

WILTING, J., KRAHMER, E., & SWERTS, M. (2006). Real vs. acted emotional speech. In *Proceedings of the Interspeech 2006 – ICSLP, Ninth International Conference on Spoken Language Processing*. Retrieved February 2009, from http://www.isca-speech.org/archive/interspeech_2006