Recognition of Japanese attitudes in Audio-Visual speech

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

[1] GIPSA-Lab, Grenoble France, [2] Showa Music Univ., Kawasaki City Japan, [3]LIMSI, Orsay France
[1]{shochi;auberge}@gipsa-lab.inpg.fr, [2]EricksonDonna2000@gmail.com, [3]{rilliard;martin}@limsi.fr

Abstract

The aim of the present work is to investigate how Japanese listeners recognize 12 audio-visual prosodic attitudes of Japanese. Significant influences of the two speakers and three modalities were observed. Generally the audio-visual condition showed the best recognition score, and interesting behavior for audio and visual modality was observed. Attitudes were regrouped into 3 higher-level perceptual categories for the first speaker: polite expressions, the attitudes of "query", and the expressions of imposition of one's own opinion. The attitude of *kyoshuku* and *surprise* are particularly well recognized by visual information.

1. Introduction

Multimodal expression of affects seems to be controlled at different cognitive processing levels [12], from involuntary controlled expressions to the intentional, deliberate control of the speaker's attitudes [5; 1]. Our position [1] is to distinguish attitudes vs. emotions as a different nature of control by the speaker: voluntary vs. involuntary, not necessarily following the affective value carried by the expressions. In this view, to reproduce, sincerely or not, an emotion outside the body loop described by Damasio [4], is in the same control as to produce attitudinal affects that have been built by the language and the culture, which then must be learned in childhood, or by learners of a second language in the case the attitudes are different from their first language. Our proposal is that attitudes and emotions evolve in acoustic (and perhaps audio-visual) spaces but are identified in terms of speech timing as differences in whether they are deemed "speech" or "under-speech-control" – e.g. the affect value "surprise" may be either an attitudinal expression or an emotion, depending on the control of the expression exerted by the speaker, and its acoustic encoding will be either anchored or not in the linguistic organization of the speech. In this view, study of attitudes is important, because this is a part of the global meaning of speech acts [5]: even if a speaker does not express any attitude by performing a simple declaration, it corresponds to the attitude described by [1] as "the speaker decides not to give information on his attitudes".

In the face-to-face interaction, attitudes are expressed within the multimodality of speech [3]. As attitudes are highly linked to language and culture, the study of attitudinal expression may benefit from a cross-cultural approach [13]. Like all language specifications, and because attitudes are constructed socially for and by the language, some attitudes can be expected to have universal values [5; 1] (e.g. authority or surprise), and possibly a universal prosodic morphology. Attitude values can also exist or not in one language or another, and their realization in a specific language may not be recognized (or may be ambiguous) in the learner's language [11; 14]. This paper presents a study as a continuation of a wider work on the expression of Japanese attitudes in a cross-cultural context [14]. It is a first study of the audio-visual expression of Japanese attitudes, in parallel to a study on audio-visual attitudes of French presented in a companion paper [9].

The multimodal nature of speech prosody has been shown for several functions [15, 2], and has a particular importance in affective communication [12]. Thus, this study tries to investigate the relative contribution of both visual and acoustic cues in the expression of prosodic attitudes, for both Japanese and French languages (this paper is specifically devoted to Japanese). As a first step, the study tries to measure the nature of the quantity of information retrieved by native listeners from each modality on a set of attitudes. A cross-cultural comparison may be envisaged.

After presenting the Japanese corpus and the set of attitudes studied here, the perception experiment and its experimental setting is described. Analyses of the recorded data using various statistical tests are done to arrive at a set of conclusions for follow-up in future work.

2. Method

2.1. Selection of 12 Japanese attitudes

A set of 12 Japanese attitudes which were validated in [14] were used for this experiment. These attitudes were selected according to the literature [6], [7], [10] and Japanese language teaching methods [8] (see [14] for definitions):

Table 1: *list of the 12 attitudes and their abbreviations.*

declaration	DC	exclamation of surprise	SU	evidence	EV
interrogation	IN	sincerity-politeness	SIN	authority	AU
admiration	AD	doubt-incredulity	DO	arrogance	AR
irritation	IR	simple-politeness	PO	kyoshuku	KYO

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 the 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 [10] 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" (p.34).

2.2. Corpus

We selected one sentence of 8 moras from the corpus developed and validated in previous research [14]. This sentence ("*Nagoyade nomimas.*", [nagojade nomimas], meaning "*He drinks in Nagoya.*") is constructed on a verb-object syntactic structure. The lexical stress position is 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 will be realized at the seventh mora, too. The sentence was constructed in order to have no particular affective connotations in any region of Japan. Two male Japanese native language speakers produce each sentence with all fourteen attitudes. The first speaker is a Japanese native language teacher who teaches various attitudes in his class by pragmatic explanations; the second, a naïve native speaker. A total of 24 utterances (1 sentence x 12 attitudes x 2 speakers) were digitally recorded in a soundproof room at LIMSI. Both speakers were standing in front of the video camera, with an omni-directional AKG C414B microphone placed 40 cm to 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. Hand claps between each sentence, recorded both by the camera and the microphone, allow as post-processing a replacement of the camera sound by the high quality sound recorded by the microphone, synchronized due to the claps. The corpus of 3 modalities (i.e. audio-alone, visual-alone and audio-visual) of 24 utterances was created, thus a total of 72 stimuli (1 sentence x 12 attitudes x 2 speakers x 3 modalities) were used for the following perceptual experiment.

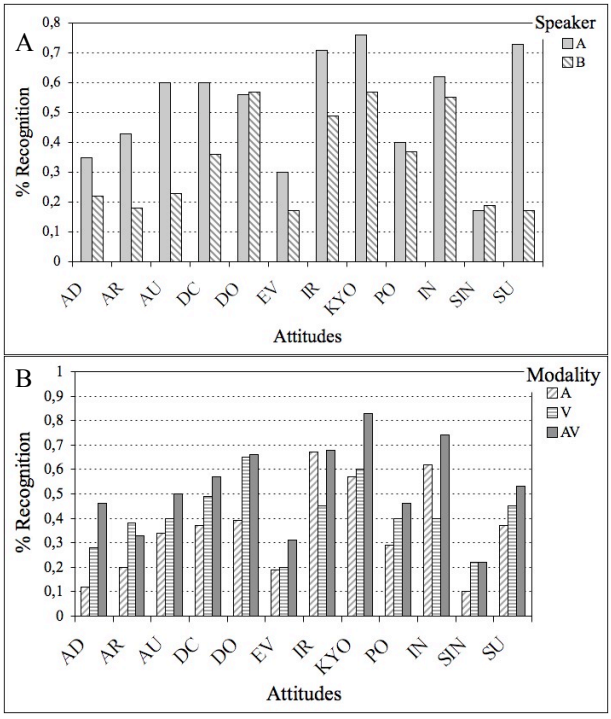

Figure 1: *Effect of speaker (A) or of modality (B) on the recognition of attitudes. X-axis shows attitudes. Y-axis shows recognition rate of 46 Japanese listeners.*

2.3. Experiment protocol

There are three presentation conditions – auditory alone, visual alone, and audio-visual. In the auditory-alone condition, subjects were instructed to listen to what the speaker said and to judge the attitude he expressed. In the visual-alone condition, subjects were required to carefully watch his facial and body movements without sound and to judge the attitude he expressed. In the audio-visual condition, subjects were required to watch the speaker's face and body movements while listening to his voice and to judge the attitude he expressed.

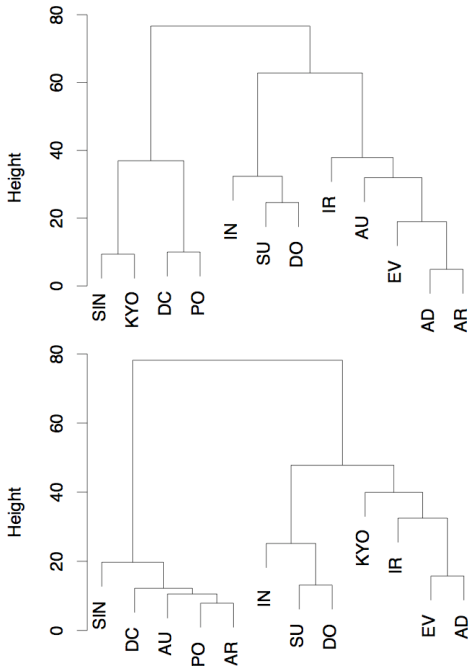

Figure 2: *clustering analysis for 12 attitudes (A: speaker A; B: speaker B) for the audio-only modality.*

The 46 Tokyo dialect speakers who participated in this experiment are separated into 2 groups. Group A consisted of 28 subjects (4 males and 24 females mean age = 18.8) who tested first with audio-only, then visual-only, and finally audio-visual modalities. Group B consisted of 18 subjects (7 males and 11 females mean age = 18.6) who tested first with visual-only, audio-only, and finally audio-visual modalities.

The perception tests were administered in a quiet room, using headphones, and on a Windows-based computer for running the interface program. The test interface gave an explanation of each label, and instructions were read in class. No subject expressed any trouble to understand the concepts referred to by the labels. All subjects listened to (and/or looked at) each stimulus one time only. For each stimulus, they were asked to indicate the perceived attitude among the twelve. Then they were asked to evaluate the intensity of expressiveness of the perceived attitude by moving a cursor located on the bottom of each label on a free scale ranging from "*hardly perceptible*" to "*very marked*" (encoded on a 1-100 scale, with the 0 score for the 11 not selected attitudes). The presentation order was randomized in a different order for each subject. No listener participating in this experiment reported any listening disorder or any visual problem.

3. Results

Results, either the percentage of recognition obtained by each attitude or their mean intensity rating, were analyzed with repeated-measure ANOVAs for testing the relative influence of each factor: the modality's presentation order (i.e. the listeners' group), the 3 modalities, the 12 attitudes and the 2 speakers. Confusions between attitudes were analysed with data-reduction techniques (a correspondence analysis, and a cluster analysis). For more details on the statistical analyses, refer to [9].

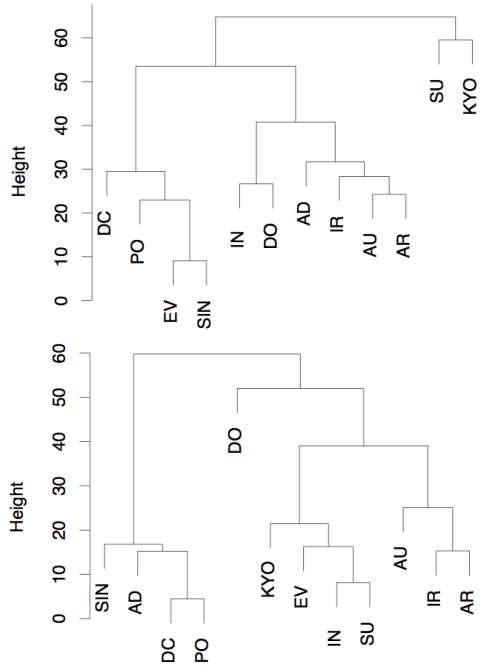

Figure 3: *clustering analysis for 12 attitudes (A: speaker A; B: speaker B) for the video-only modality.*

3.1. Results of the ANOVAs

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 ($p>.01$). First we investigated whether the modality's presentation order influenced the perception of attitude, and evaluated the results of the two groups by ANOVA. The results show no significant influence of the order of presentation on listeners' answers ($F=0.0795$ $p=.78$).

Results show a significant influence of the speaker on the recognition scores, expressed either as categorical choice ($F=163.6$, $p<.0001$) or as intensity ratings ($F=211.5$, $p<.0001$). Figure 1A presents listeners' recognition rate for the two speakers, and for each attitude. This graph shows that the attitudes expressed by speaker A, who is the Japanese native language teacher, were globally better recognized than speaker B's expressions (mean scores for the 2 speakers: recognition A=.52, B=.34; intensity: A=35.3%; B=19.4%). Especially the perception of AU and SU was quite different.

The modality also influenced the listeners' perception: a significant effect was observed with both categorical choices ($F=44.8$, $p<.01$) and intensity ratings ($F=44$, $p<.01$). Figure 1B presents listeners' recognition rates of the three modalities and for each attitude. The audio-visual condition shows the best recognition score for all the attitudes except AR. The recognition scores of AR, DO and SIN in visual alone condition are mostly the same as in the audio-visual condition, indicating a predominance of the visual information over the auditory one for these attitudes. However, the attitude of IR seems mostly recognized by auditory information rather than visual information. For most attitudes (AD, AU, DC, EV, KYO, PO, IN, SU), audio and visual information cooperate, as audio-visual scores are better than audio or visual only information. Auditory and visual information complemented each other.

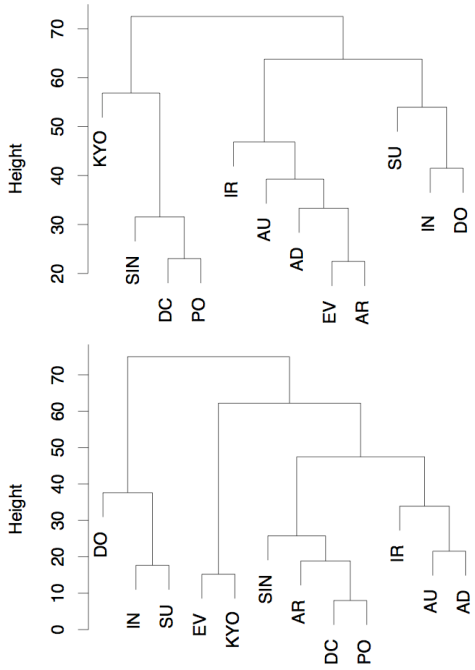

Figure 4: *clustering analysis for 12 attitudes (A: speaker A; B: speaker B) for the audio-video modality.*

3.2. Confusion analysis by data reduction techniques

Similarities between listeners' recognition between the 12 attitudes will be detailed (for each modality and each speaker) hereafter using the clustering analysis.

Audio alone modality

For speaker A, the results regroup in 3 main categories (cf. fig. 2A). The first group consists of SIN, KYO, PO and DC. Listeners tend to perceive in a similar way all the politeness expressions, plus DC. The second group contains IN, DO and SU, and therefore constitutes a more global group of query expressions. The third group consists of IR, AU, EV, AD and AR. All these attitudes except admiration express the imposition of the speaker's opinion.

Similarly, there are 3 categories observed for speaker B (cf. fig. 2B), but the components of each category are not the same. The first group consists of SIN, PO, DC, AU and AR. Contrary to the first group for speaker A, AU and AR are included instead of KYO. The second group contains IN, DO and SU, which are the same components seen with speaker A: all these attitudes express a query. The third group consists of KYO, IR, EV and AR. As already suggested above, the

performance of untrained speaker B was less robust than that of speaker A; consequently, this could cause confusions for subjects, even between expressions that are supposed to be very different.

Visual alone modality

For speaker A, the perceptual categories of attitudes in the visual modality (cf. fig. 3A) are the same as those in the audio modality, except KYO and SU in that each behave very specifically in the visual modality.

Results for speaker B (fig. 3B) show a radical change of perceptual behavior in the visual modality. The first group consists of 2 polite expressions plus DC and AD. However, KYO is added to the second group, with EV, IN and SU, and mainly seems to group the most confused expressions for this modality and this speaker. A third group consists of AU, IR, and AR, all expression of dominance. DO does not show any confusion.

Audio-visual modality

According to the 3 perceptual categories shown in fig. 4, the attitudes belonging to each category are almost the same as in the audio-alone modality (fig. 2), for speaker A. However, the results of speaker B in fig. 4 are quite different from both audio-alone and visual-alone except for the category of "query" observed in audio-alone. In addition, KYO and EV form one independent category.

Such predominance of the audio modality over the visual one for the confusion of attitudinal expressions is coherent with the results observed with French (see [9]).

4. Conclusion

The aim of the present work is to investigate how Japanese listeners recognize audio-visual attitudinal expressions, and the completion or redundancy of the two modalities to each attitude represented in this work. The results show no significant difference due to presentation order, whereas the speaker performance, the modality and the expressed attitudes did have a significant effect on listeners' recognition scores. Generally the audio-visual modality showed the best recognition score for all attitudes. For some attitudes, the one modality alone seems to carry all the needed information, with the additional modality not adding extra information. For most attitudes, auditory and visual information complemented each other.

The listeners generally group the set of 12 attitudes for the first speaker into 3 general perceptual categories. The first group consists of the polite expressions, the second, the attitudes of "query", and the third, the expressions of imposition of one's own opinion. The attitude of KYO and SU are particularly well recognized by visual information for speaker A. Worst recognition scores and more important confusions shown on speaker B's results may be due to several factors: As a naïve speaker, he may have experienced more difficulty performing the 12 attitudes, and may have chosen different strategies than speaker A. He may also have expressed these attitudes in a more subtle and less intense way, leading to stimuli harder to be recognized out of context – and this is reflected by the difference between the mean intensity ratings obtained from speakers A and B.

In order to better understand some of the anomalies in the results presented here, future work will involve acoustic and visual analysis of the productions of each of the speakers. Also analysis-resynthesis of these audio-visual attitudes are planned in order to extract the most pertinent parameters for perception.

5. Acknowledgements

We are deeply grateful to T. Sadanobu from Kobe University for his helpful advice and contribution on Japanese attitudes. We also thank students at Kanagawa University for their participation in the perception tests. This work was performed as collaboration between the Graduate School of Cultural Studies and Humane Science of Kobe University and the GIPSA-Lab of Grenoble under the auspices of the "College Doctoral Franco-Japonais". It was supported in part by the Japanese Ministry of Education, Science, Sport, and Culture, Grant-in-Aid for Scientific Research, (2007-2010):19520371 to the second author, and also by SCOPE (071705001) of Ministry of Internal Affairs & Communications (MIC), Japan.

6. References

[1] Aubergé V. 2002. A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP. *Speech Prosody*, Aix-en-Provence, France, 151-155.

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

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

[4] Damasio, A.R. 1994. *Descartes' error. Emotion, reason, and the human brain*. New-York: G.P. Putnam.

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

[6] Erickson, D., Ohashi, S., Makita, S., Kajimoto, N., Mokhtari, P. 2003. Perception of naturally-spoken expressive speech by American English and Japanese listeners. CREST International Workshop on Expressive Speech Processing, 31-36.

[7] Maekawa, K., 1998. Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. ICSLP, 635-638.

[8] Mizutani O., Mizutani N., 1979. Aural Comprehension Practice in Japanese. The Japan Times.

[9] Rilliard, A., Martin, J.C., Aubergé, V. & Shochi, T. (submitted). Perception of French Audio-Visual Prosodic Attitudes. Speech Prosody, Campinas, Brasil.

[10] Sadanobu, T. 2004. A natural history of Japanese pressed voice. J. of the Phonetic Society of Japan 8 (1), 29-44.

[11] 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.

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

[13] Shigeno, S. 1998. Cultural similarities and differences in the recognition of audio-visual speech stimuli. *ICSLP98*,

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

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