The combined Perception of Socio-affective Prosody: Cultural Differences in Pattern Matching

Takaaki Shochi, Marine Guerry, Albert Rilliard, Erickson Donna, Jean-Luc Rouas

音声研究　第 24 巻　84–96 頁
2020 年

Journal of the Phonetic Society of Japan
Vol. 24, pp. 84–96, 2020

研究論文

The combined Perception of Socio-affective Prosody: Cultural Differences in Pattern Matching

Takaaki Shochi[*,**], Marine Guerry[*], Albert Rilliard[***,****], Donna Erickson[*****] and Jean-Luc Rouas[**]

感情音声のマルチモーダル知覚 —パタンマッチングにおける文化的相違に注目して—

SUMMARY: This study examines cross-cultural differences for the perception by Japanese, French and American judges of audio-visual recordings of a short Japanese utterance produced with nine different social affective expressions. The listeners' task was to create an audio-visual expression that fit a given expressive label, by matching one of the nine video with one of the nine audio stimuli. L1 judges showed a higher rate of correct matching than non-L1 judges; confusions were within semantic categories. Non-L1 judges showed matching patterns similar to L1 ones, with modality-specific differences especially for culturally-related expressions like Japanese politeness and the occidental expression of seduction.

Key words: Multisensory recognition, Pattern matching, Cultural difference, Socio-affective prosody, Automatic combination of synthetic stimuli

1. Introduction

In face-to-face communication, both vocal and visual expressions of affect interact with each other to convey a synergic complex of information (S. Jessen and S. A. Kotz, 2015; A. Rilliard et al., 2009; K. Scherer, 2009; P. Barkhuysen et al., 2010). The study of multimodal perception of affects is relatively recent, preceded by a large number of studies on vocal and visual expressions of "non-emotional" speech (e.g. D. W. Massaro and M. M. Cohen 2000; K. Sekiyama and Y. Tohkura 1991; H. McGurk and J. MacDonald, 1976). The latter study introduced the concept of conflicting cues to speech perception, known as the McGurk Effect, in which visual information sometimes overrode the acoustic information or acoustic information overrode the visual information, sometimes the presentation of incongruent audio-visual phonetic information resulted in the perception of a third distinct phoneme.

Research into which modality, auditory or verbal, more effectively conveys a particular affect or emotion is ongoing. For instance, Q. Summerfield (1992) demonstrated the influence of the visual modality for the recognition of verbal expressions in a noisy environment. F. B. Colavita (1974) discussed the obvious prepotency of the visual information over the auditory one, suggesting that the dominance of visual information over that of hearing for perception of audiovisual stimuli.

Recent work has also shown the importance of the auditory signal in the perception of affects. For example, A. S. Walker and W. Grolnick (1983) were among the first researchers who identified the influence of the auditory modality for the perception of facial expressions of affects. They worked on the audio-visual affective perception (joy and sadness) of infants (3 and 5 months). In their experiment, they presented the facial expressions of the affects with the varied affective voices in order to observe the direction of the head of the newborns as well as the duration of the fixing of their gaze to the presented face. According to the results of their experiment, the 5-month-old babies

[*] CLLE CNRS UMR 5263, Université Bordeaux Montaigne（ボルドーモンテーニュ大学認知科学研究所）
[**] LaBRI, CNRS UMR5800, Université de Bordeaux（ボルドー大学情報処理研究所）
[***] Université Paris-Saclay, CNRS, LIMSI（パリ・サクレー大学情報処理研究所）
[****] Universidade Federal do Rio de Janeiro（リオデジャネイロ連邦大学）
[*****] Haskins Laboratories（ハスキンス研究所）

looked significantly longer at the face of the speaker expressing the same affect by voice rather than the one expressing a different affect. This result indicates the influence of auditory information on the overall perception of affects. In addition to these studies, V. C. Tartter (1980), V. C. Tartter and D. Braun (1994) and V. Aubergé and M. Cathiard (2003) showed that the acoustic signal alone can transmit a "mechanical" smile, without necessarily having visual information, since the configuration of the face (including shape of the lips) linked to fun or joy changes the voice quality corresponding to this affect. J. Ohala (1984) explains that a smile in speech leads to an increase of formant frequencies, which is the leading factor associated with a smile being considered a friendly gesture. As pointed out by P. Noller (1985), multiple factors such as age, gender or choice of affects must be taken into account in order to understand the cognitive process of multimodal perception.

Using congruent and incongruent combinations of audio and visual information is yet another approach to investigating cognitive processing of cross-modal integration of emotional information, a type of emotional McGurk experiment. B. de Gelder et al. (1999) examined both audio and visual modalities on the recognition of emotional expressions, and reported that subjects were faster in recognition of the conveyed information if the facial expression was congruent with the vocal expressivity. Their study, however, used static visual information as visual input and therefore, their results may be different from using dynamic visual movement. S. Takagi et al. (2015) also investigated multisensory perception of affects using dynamic facial expressions to show that the modality dominance between audio and visual information changes for each affect.

A cross-cultural multi-modal emotion perceptual experiment by A. Li et al (2013) examined AV-conflicting stimuli produced by a Chinese female speaker as perceived by Chinese and Japanese listeners. The results showed that listeners tended to use specific modalities for specific emotions, which are often affected by cultural norms of the listener.

E. M. Provost et al. (2015) report on an emotional McGurk dataset at the University of Michigan, which contains emotionally congruent stimuli (emotionally matched faces and voices) as well as emotionally incongruent stimuli (emotionally mismatched faces and voices). They report some fusion effects of incongruent audio and visual stimuli along the dimensions of Valence, Activation, and Dominance. They suggest that the emotional McGurk effect may be better described in terms of its effects on dimensional, rather than categorical, perception.

These studies were carried out with emotional expressions such as anger, joy, sadness, etc. (P. Ekman and W. V. Friesen, 1978) rather than intentional (and voluntarily controlled) social affective expressivities (e.g., irony, contempt, seduction, suspicious, etc.). According to H. Spencer-Oatey (2005), social affects are linked to the speakers' social status, and the intention conveyed in face-to-face interaction. They are supposedly learned during the developmental process in the social environment (V. Aubergé, 2002; T. Shochi et al., 2009). As such, these affects may vary from culture to culture, which can lead to misunderstandings (T. Shochi et al., 2009). The acoustic as well as visual aspects of social affects are described in many languages (A. Abelin, 2004; A. Kleinsmith and N. Bianchi-Berthouze, 2013) and the language-specific aspect of such expressions have been described in a number of studies (A. Rilliard et al., 2009; I. Fónagy, 1982; J. Pierrehumbert and J. B. Hirschberg, 1990; C. Shan et al., 2007; H. Gunes and M. Piccardi, 2009). A. Pavlenko (2005) mentioned the importance of affective meanings during speech communication in her book focusing on the cross-cultural differences and common features of vocal affective expressions.

Our current paper investigates the cognitive representations of social affect performances in audio-visual modalities. Specifically, the purpose of this study is to examine two points: 1) how do first language (L1) listeners select and combine auditory cues and speaker's facial expressivity for social affects, and 2) what are some potential cultural differences between L1 listeners and language-naïve listeners for the multisensory perceptual patterns of these social affects.

Building on a paradigm targeting cross-cultural recordings (A. Rilliard et al. 2013), the paper presents results comparing the perception of nine Japanese social affects. These nine social affects were selected by previous research in linguistics, phonetics and psychology (A. Wichmann, 2000; T. Shochi et al., 2009; A. Rilliard et al., 2013; T. Sadanobu, 2004; J. A. de Moraes 2008).

This paper is organized as follows: the acquisition of the corpus and the method adopted for automatic combination of synthetic stimuli and their perceptual evaluation are described in Section 2. The results of the perceptual evaluation are presented in Section 3, and discussed in Section 4 before presenting conclusion and perspectives in Section 5.

The experiment aims at enlightening various aspects

of multimodal expression in speech; specifically, the goals are to examine the relative role of modalities, their similarities across dimensions of expressivity, as well as potential cross-cultural variations.

2. Perception Experiment

2.1 Corpus

We tried to avoid the use of labels as an input for speakers' productions of varied social affective behaviors, as their conceptual descriptions have been shown to be language-dependent (e.g. A. Wierzbicka, 1985, 1996; J. Harkins and A. Wierzbicka, 2001). Instead, we proposed communication contexts via short interaction scripts, that explicitly state the communicative aim of the speaker, as well as the social relation between the speaker and the listener (A. Rilliard et al. 2013). These social relations typically focus on the different hierarchical relationships that may exist between the speaker and the listener. A corpus was thus recorded by multiple speakers interacting with the experimenter to produce two target sentences ("a banana" and "Mary was dancing") in different languages (USA English A. Rilliard et al. 2013; French: Guerry et al. 2014; Brazilian Portuguese: A. Rilliard and J. A. Moraes 2017; German: H. Mixdorff et al. 2017), and Japanese (T. Shochi et al. 2015). Sixteen situations were selected, corresponding to sixteen social affects, for which a prototypical interaction script was defined, in order to elicit that affect from the speaker.

These situations were inspired from affects attested in different languages, but do not necessarily correspond to conventionalized expressions in all the studied languages. Hence, seduction is not a conventionalized expression in Japanese, and maybe not in French either; perhaps it could be coined as a "Hollywood seduction".

The Japanese corpus is based on the production of these sixteen social affects, enacted on two target sentences by 19 first language Japanese speakers (10 females) from 20 to 36 year-old (mean age: 22 yrs.), most of whom were students at Waseda University (Tokyo), where the recordings took place. The quality of the recorded productions was evaluated for their quality – i.e. whether the speaker succeeded in expressing the targeted affect. The evaluation was done by 38 first language Japanese speakers, who rated each of the 608 stimuli (audio-visual performances of 16 affects by 2 sentences by 19 speakers) on a 1 to 9 scale (the higher the better). The output of the evaluation showed that one male speaker (hereafter "M", a trained language teacher) outperformed the others with a mean evaluated performance at 7.3 (the three first quartiles were: 7.0, 7.6, 8.0, on a 9 point scale); the second best speaker was a female student (hereafter "F"), whose performances received a mean score of 6.9 on 9 (the three first quartiles were: 6.0, 7.2, 7.8). These two speakers were selected for the current experiment, as best overall performers for these social affects.

Among the sixteen social affects produced by these two speakers, only nine affects on the "banana" sentence, were selected; this was done to reduce the complexity of the experiment. Following previous findings, these nine social affects were subcategorized into 4 categories (see Table 1): Surprise (SURP), a potentially universal affect (T. Shochi et al., 2009; M. Guerry et al., 2014; A. Rilliard et al., 2009); Obviousness (OBVI), Irony (IRON), Contempt (CONT) and Irritation (IRRI), perceptually linked to meanings of imposition (M. Guerry et al., 2014); Politeness (POLI), Sincerity (SINC), and Walking on eggs (WOEG), different strategies of linguistic politeness, (i.e. positive or negative face-protecting devices, following P. Brown and S. Levinson, 1987's conceptualization). Walking-on-eggs was used to denote a situation corresponding, to some extent, to situations where Japanese speakers would express "恐縮 (Kyoshuku)", a Japanese-specific concept defined as "corresponding to 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" (T. Sadanobu 2004, p. 34).

Finally, Seduction (SEDU) transmits the speaker's search for proximity in some vocal phonostyles (P. R. Léon, 1993, for French). E. Hall (1966) noted that the visual cues (especially gaze movement) play the main role for this seductive affect, but recent studies on the seductive voice of French speakers reported that the speakers lowered significantly their fundamental frequency when they spoke to someone they wanted to seduce (A. Aron, 2018)

2.2 Automatic Combination of Synthetic Stimuli

A total of 18 utterances (2 native speakers × 9 affective expressions) were chosen as source signals; each of the auditory and visual recordings were then synchronized (see below for an explanation), such that any possible combination of audio and video presentation could be matched to create one of the nine social affective expressions. That is, since all audio and visual stimuli were synchronized, a subject could match the auditorily-expressed "sincerity" with the visually-expressed "irritation", if they thought this was the best

Table 1 The nine selected social affects

Potentially universal affect:	
Surprise (SURP)	
Cultural Specific affects:	
Obviousness (OBVI)	Imposition of speaker's opinion
Irony (IRON)	
Contempt (CONT)	
Irritation (IRRI)	
Politeness (POLI)	Politeness strategies
Sincerity (SINC)	
Walking on eggs (WOEG)	
Seduction (SEDU)	Voluntary intimacy

match for a particular social affective expression.

As a first step of this automatic combination of synthetic stimuli, each utterance was separated into audio alone (A) and video alone (V), and a manual transcription in phonemes (P) was done. In order to combine a video file from attitude 1, denoted V_1, with an audio file from attitude 2 (A_2), two possibilities were explored: (1) keeping the video file (V_1) as is, but synthesizing the audio file using the time-stretching algorithm from Straight (Kawahara et al., 2008) to change the duration of each mora, such that the audio signal corresponded with the lip motions, thus resulting in a new audio signal ($A_2 \rightarrow {}_1$), or (2) keeping the audio file A_2 as is, but modifying the video file (V_1) by removing or duplicating frames, thus resulting in a new video file ($V_1 \rightarrow {}_2$). Both processes were empirically tested for naturalness; the second solution (2) was chosen, as it seems the eye is more easily fooled than the ear (a similar procedure was used in E. M. Provost et al., 2015). The process for creating the stimuli using this solution is illustrated in Figure 1. Given the duration of each syllable, video frames were duplicated or removed in the middle of the vocalic segments (i.e. no complex processing involved). By this synthesis method, we compiled 162 synthetic audiovisual affective expressions (9 affects × 2 speakers × 9 combination types).

2.3 Perceptual Evaluation Paradigm

The perceptual experiment was taken in an individual session under a JAVA based interface with Bose 5C7N1 high quality headphones (Figure 2). The nine labels were translated into the three languages (with their definitions). On the first page, one simple instruction was given to the subjects: "Select the audio and video which best expresses the following social affect, XXX"; on the next page, the participants were instructed to listen and watch each stimulus as much as they wanted to, and then they were asked to "create" one affective expression by matching the audio and visual files of the 9 affective expressions.

2.4 Subjects

Three groups of judges participated in the experiment: 23 native Japanese subjects (JP), all Tokyo dialect speakers (mean age = 20 y/o), 19 French (FR) and 40 US English subjects (US) without any knowledge of Japanese (FR: mean age = 31 y/o; US: mean age = 22 y/o).

2.5 Processing of Perceptual Evaluation Output

The data processing involved two distinct steps: first a logistic regression process was applied to determine "hits" or "misses" in each modality; and second, a multivariate analysis process evaluated (dis)similarities between labels, by studying the confusions in answers.

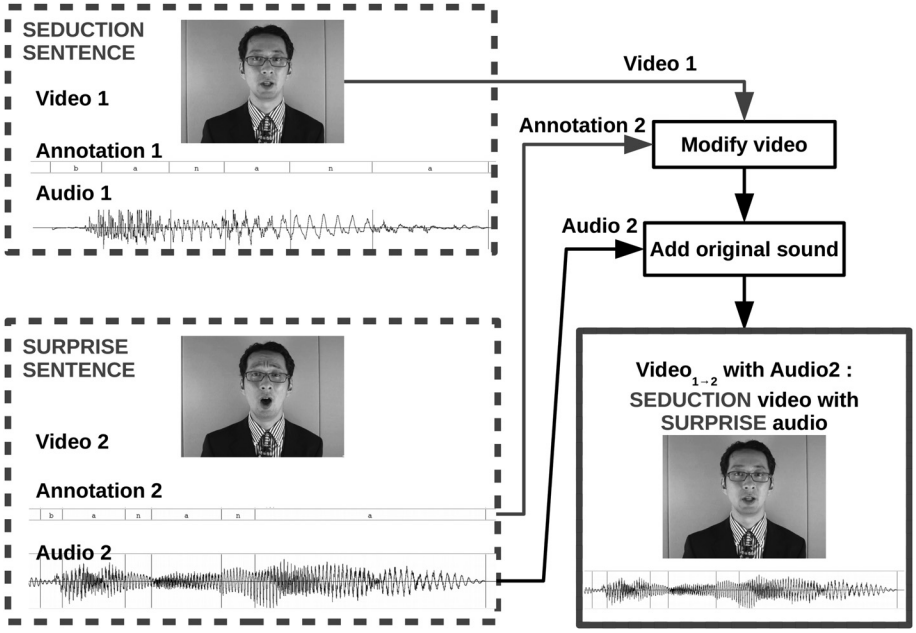

Figure 1　Combining video from seduction *banana* sentence with audio from surprise *banana*

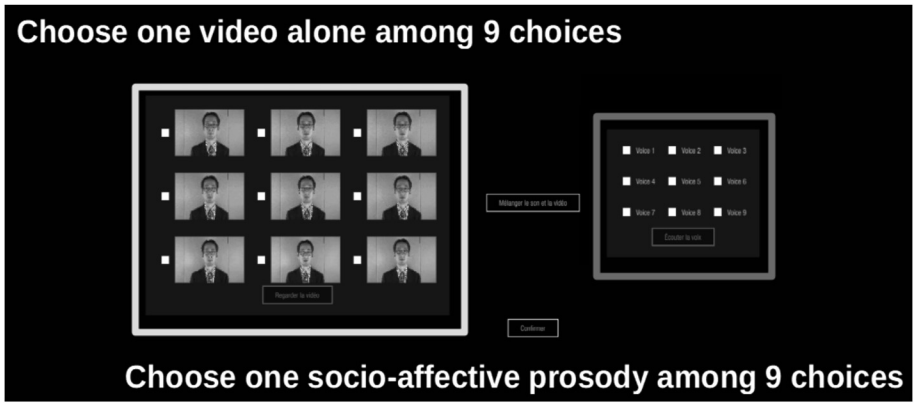

Figure 2　Interface of the perceptual experiment for US judges

As a first step, the answers to the test were analyzed evaluating the accuracy of judges to select the performances corresponding to the target label, according to three controlled factors: (i) the modality of presentation of these performances (either Audio or Visual), (ii) the judges' language Group (JP, FR, US), and (iii) the presented Labels (9 levels). This analysis was done by fitting a logistic regression on the ratio of correct guesses. The second step evaluated the (dis)similarities of audio and visual stimuli selection between labels. This was done by using contingency tables showing the number of times each audio or visual stimulus was selected for each label—for each group of judges—and then applying a multidimensional analysis (Multiple Factor Analysis, henceforth MFA, see F. Husson et al. 2017). This process reduced the dimensionality of the dataset in order to compare different subsets of data.

For building the logistic regression model, the audio and visual stimuli selected by each judge were evaluated as a hit (1) or a miss (0), depending on if they matched the target label or not. This binary outcome was used as the dependent variable for the model, that had the following independent variables: subjects' language Group (3 levels: JP, FR, US), the presented

Label (9 levels), and the Modality of the selected stimuli (2 levels: Audio, Visual). The three independent variables and their double and triple interactions were used to fit a maximal model that was then submitted to a simplification process (thus removing non-significant contributions; see R. H. Baayen 2008; S. T. Gries 2013 for details). The minimal adequate model (i.e. the simplest model able to describe non-random variation in the observed variable) included the three main factors, plus the Group × Label and the Label × Modality interactions. The triple interaction (LRT $\chi^2(16) = 21.8$, $p = 0.15$) and the interaction between Group×Modality (LRT $\chi^2(2) = 4.3$, $p = 0.12$) were found to be non-significant and thus disregarded.

The multivariate analysis process was based on contingency tables built so that for each presented label, organized as the rows of the table, we count the number of times each type of audio and visual stimuli was selected by the judges of each language group. The nine types of selected stimuli build the columns of the table, duplicating them by language group and modality – thus, the complete contingency table consisted of 9 * 2 * 3 = 54 columns and 9 rows, subdivided into six 9 × 9 sub-tables (2 modalities × 3 language groups). During the MFA process, each sub-table was submitted to a correspondence analysis that extracted its main dimensions of variation; these dimensions were then compared and grouped within sub-tables according to their weight on the rows (i.e. the labels presented to judges, which are the common entry to sub-tables), so as to build a main multidimensional analysis. The 5 main dimensions (selected according to an elbow criterion) were kept, and accounted for more than 90% of the total variance. The spread of the presented labels (the rows) onto this 5-dimensional space, which is based on the use of audio and visual stimuli by each group of judges, gives an idea of cognitive similarities and differences among labels and their chosen performances, for each language group (for more on this conceptual approach, see A. K. Romney and C. C. Moore 1998; A. K. Romney et al. 2000). The Euclidean distance between labels on this spread was used as an input for a hierarchical clustering algorithm that allowed grouping the set of nine labels into clusters coherent in terms of their audio and visual selection of performances by judges.

3. Results

3.1 Ratio of Correct Selection of Stimuli

The importance of factors explaining variations in the proportion of correct stimuli selection, according to the logistic model, are summarized in Table 2. It shows that the efficiency of judges to select the corresponding target stimuli depends on (i) their knowledge of the language, with a significant effect of language Group; it also critically depends on (ii) the type of social affect to reconstruct (see Figure 3), and on (ii) the modality – with visual stimuli having a slightly but significant higher outcome than audio ones (0.41 vs. 0.32). Figures 3 and 4 present the scores predicted by the regression model, on the basis of the three independent variables' levels. The significant two-way

Table 2 Analysis of deviance table (type III tests) for the logistic regression model on the ratio of good selection of stimuli reporting, for each independent variable and the two-way interactions, the related likelihood ratio test, degrees of freedom and associated probability.

	LRχ^2	df	p
Group	33.7	2	< 0.05
Label	189.0	8	< 0.05
Modality	4.8	1	< 0.05
Group × Label	63.1	16	< 0.05
Label × Modality	70.4	8	< 0.05

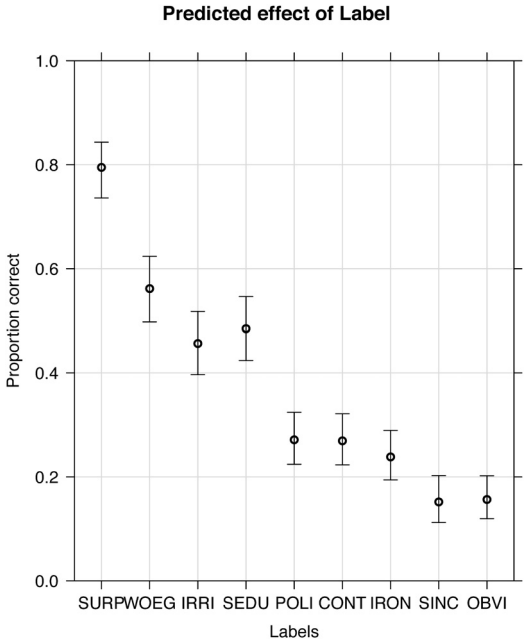

Figure 3 Predicted mean and confidence intervals for the effect of Labels on the proportion of correct guesses of stimuli.

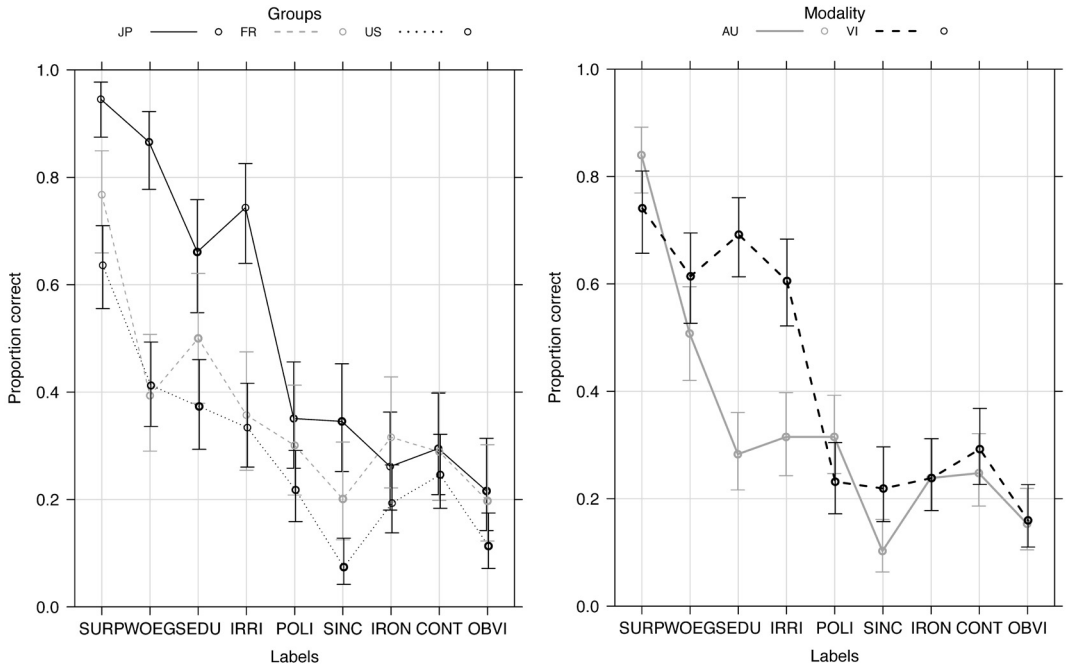

Figure 4 Predicted mean and confidence intervals for the Label x Group interaction (left plot; JP group: plain black line; FR group: grey dotted line; US group: black dashed line) and the Label x Modality interaction (right plot; AU modality: plain grey line; VI modality: dashed black line) on the proportion of correct guesses of stimuli.

interactions between Group & Labels and Label & Modality are presented in Figure 4. Let's recall the other interactions were dismissed for not significantly improving the model (see § 2.5 for details).

Post-hoc pairwise comparisons were run on the Group and Label factors and the Group × Label and Label × Modality interactions (Bonferroni correction). The group of Japanese judges shows an overall performance ratio of 0.56 of correct selections, significantly higher than the others, with the French judges having a 0.36 ratio, higher than those of the US at 0.26. Three main groups of labels (Figure 3) can be set, based on their ratio of correct answers, with Surprise having the highest score, at about 0.8, followed by WOEG, IRRI and SEDU around 0.5, while the five others are below 0.3.

The interaction between Group and Label (Figure 4 left) shows differences in performance levels for a given label, linked to cultural/linguistic competence. It is mostly focused on labels with high recognition scores (SURP, WOEG, IRRI, SEDU) that are significantly higher when perceived by L1 speakers of Japanese than by speakers who did not know this language; but note that for SEDU, only the US group has lower scores compared to the JP group. The interaction between Modality and Label (Figure 4 right) shows significant differences in performance ratio for individual labels between modalities in the case of IRRI and SEDU – for which the Visual stimuli received a higher performance ratio than the Audio one. The interaction between Modality and Group is not significant: this means notably the effect of Modality on Labels is consistent across Groups, and that the differences between Groups for four Labels are consistent across Modalities.

3.2 Multivariate Analysis of Performance Construction by Judges

The output of the MFA shows the link between the different Labels and the main dimensions, as presented in Table 3. The first dimension of the MFA is dedicated to the singularity of Surprise (the label with the highest ratio of correct stimuli selection). The second, third and fourth dimensions weight negatively on Seduction – and positively, respectively, on Irritation and Contempt, on Walking-on-Eggs, and on Sincerity and Politeness. The fifth-dimension weights positively on Irony and Obviousness. The first three dimensions, that explain 74% of variance, are linked to answers to the four labels that received high selection ratio (above 50%); it also highlights the confusions between Irritation and

Table 3 Coordinates (D), Contribution (Ct), and squared cosine multiplied by 100 and rounded for convenience (cs) on each dimension (1 to 5) of the MFA, for the nine Labels. The main links between rows (Labels) and columns (dimensions of the MFA) are presented in bold italic fonts, when the row contribution to a column is above the mean, and if the row is well represented by this dimension (squared cosine over 0.20).

	D1	D2	D3	D4	D5	Ct1	Ct2	Ct3	Ct4	Ct5	cs.1	cs.2	cs.3	cs.4	cs.5
Cont	0.6	***2.1***	−0.2	0.2	0.9	1	***13***	0	0	8	4	51	1	0	10
Iron	0.2	0.8	−0.3	0.5	***1.8***	0	2	0	2	***28***	1	11	2	5	***58***
Irri	−0.0	***3.7***	−0.1	−1.5	−1.9	0	***38***	0	12	33	0	69	0	11	19
Obvi	0.1	1.4	−0.2	0.4	***1.2***	0	6	0	1	***13***	0	33	0	2	***23***
Poli	1.2	−1.2	−0.3	***2.0***	−0.8	3	4	0	***22***	6	16	17	1	42	8
Sedu	1.8	***−2.8***	***−2.8***	***−2.2***	0.0	6	***22***	***32***	***29***	0	13	33	33	21	0
Sinc	1.2	−1.0	−0.1	***2.0***	−1.2	3	3	0	***24***	12	15	10	0	46	15
Surp	***−6.4***	−1.3	0.0	−0.1	−0.2	***83***	5	0	0	0	96	4	0	0	0
Woeg	1.4	−1.7	***4.0***	−1.3	0.1	4	8	***67***	10	0	9	13	71	7	0

Table 4 List of the contingency table's columns significantly correlated (p<0.05) to the first five dimensions of the MFA. Columns are separated between positive (higher part of the table) and negative correlations (lower part), for each dimension. For each column, the corresponding expression, its presentation modality and the language group are given.

	D1	D2	D3	D4	D5
+		Irri/V/JP	Woeg/V/JP	Poli/V/JP	Iron/V/JP
		Irri/A/JP	Woeg/A/JP	Poli/V/FR	Iron/A/JP
		Irri/V/FR	Woeg/V/FR	Poli/V/US	Iron/V/FR
		Irri/V/US	Woeg/V/US	Poli/A/US	Iron/A/FR
		Cont/V/JP	Sinc/V/US	Sinc/V/JP	Obvi/V/JP
		Cont/A/JP		Sinc/A/JP	Obvi/A/JP
		Cont/A/FR			Obvi/V/US
		Cont/A/US			
		Sinc/A/FR			
		Sinc/A/US			
−	Surp/V/JP	Sedu/A/JP	Sedu/V/US	Woeg/A/US	
	Surp/A/JP				
	Surp/V/FR				
	Surp/A/FR				
	Surp/V/US				
	Surp/A/US				

Contempt labels. The fourth and fifth dimensions show the confusions patterns between politeness and sincerity (4th axis) and Irony and Obviousness (5th axis), two groups of labels that are not singularized by a specific set of performance (neither audio or visual), but share similar patterns.

Table 4 presents the columns of the contingency table that were found to have a significant correlation (positively or negatively) with each of the five first dimensions of the MFA. These columns are related to the selection by judges of each language group of audio or visual performances.

The first dimension is negatively (note that the sign has no importance but to oppose both ends of a dimension) correlated to all performances of Surprise (i.e. the two modalities were selected by all language groups): this confirms the high Audio and Visual selection ratio observed for this expression in the preceding section. The second dimension is positively related to the performances of Irritation and Contempt, in both modalities for the JP group, but only in the visual modality for FR and US groups. This relates to the significant difference in selection ratio observed for Irritation between the JP group and the others. It also shows that productions of Contempt may be adequate for the expression of Irritation. The low performance of FR and US groups in audio for the Irritation label is further supported by the association to this dimension of the audio performance of Sincerity by these two groups. On the negative side of the second dimension, the JP group associates audio performances of Seduction to this label. The third dimension is positively linked to WOEG performances and labels, in both modalities for JP, but only in the visual mode for FR and US; US judges also select visual performances of Sincerity and oppose this to visual performances of Seduction. The fourth dimension is positively linked to Politeness and Sincerity performances. Politeness is selected mostly for its visual modality (US also selected the audio modality), and only the JP group selected performances of Sincerity. The same dimension is related on its negative side to the Seduction label, and to selection of the WOEG performance by US listeners. The fifth dimension is related for the JP group to audio and visual performances of Irony and Obviousness, while the FR group relates it to AV Irony performance, and US, to visual Obviousness.

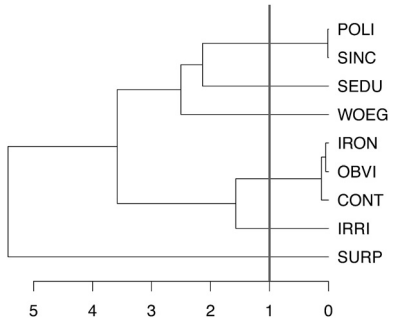

Figure 5 Dendrogram presenting the Euclidean distance between labels calculated from their position along the first five dimensions of the MFA, hierarchically grouped following Ward's minimum variance criterion. The vertical line indicates the place where the tree was cut for the six cluster solutions.

A hierarchical clustering algorithm was applied to the distribution of labels across the five main dimensions. The obtained dendrogram (see Figure 5) shows the proximities and oppositions between the nine presented labels. Surprise is separated from the others that are grouped under two main types: a set of relationship expressions (SEDU, WOEG, POLI, SINC), and a set of expressions of imposition (IRRI, CONT, OBVI, IRON). Within each of these two groups, Irritation is separated from the other impositions, while Seduction and WOEG are separated from other relationship expressions. By cutting the dendrogram at this level, one obtains a six-cluster solution.

The six clusters are linked to the following types of performances (i.e. these performances are selected significantly more often by judges within the clusters than they are globally). **Cluster #1** regroups performances selected for the label "Surprise" (Label|Modality|Group, in decreasing order of association): Surp|A|JP, Surp|V|JP, Surp|V|FR, Surp|A|FR, Surp|A|US, Surp|V|US. **Cluster #2** regroups performances linked to "Irritation": Irri|A|JP, Irri|V|JP, Irri|V|US, Irri|V|FR, Cont|A|US, Cont|A|FR. **Cluster #3** regroups performances linked to "Contempt, Irony, Obviousness": Iron|A|JP, Iron|V|FR, Iron|A|FR, Obvi|A|JP, Iron|V|JP, Obvi|V|JP, Obvi|A|US, Cont|A|JP, Cont|V|JP, Iron|A|US **Cluster #4** regroups performances linked to "Politeness, Sincerity": Poli|V|JP, Sinc|V|JP, Poli|A|US, Sinc|A|JP, Poli|V|FR, Poli|A|FR, Poli|V|US, Poli|A|JP, Sedu|A|US. **Cluster #5** performances linked to "WOEG": Woeg|A|JP, Woeg|V|JP, Woeg|V|US, Woeg|V|FR, Obvi|V|FR. **Cluster #6** regroups performances linked to "Seduction": Sedu|V|JP, Sedu|V|FR, Sedu|V|US, Woeg|A|FR, Sedu|A|FR.

We note some variation in terms of association between labels and performances when we compare the cluster analysis and the associations along the MFA dimensions. This may be explained by the fact clusters consider the five dimensions, not a single one.

4. Discussion

The current work investigated cross-cultural differences of multisensory perception of Japanese socio-affective prosody extracted from a social interaction dataset, as judged by three groups of judges: Japanese and also French and U.S., the latter two groups having no knowledge of Japanese.

According to the logistic regression, the ratio of correct guesses was highest for the L1 judges (JP) than for the two groups of language-naïve judges (FR and US) in three expressions (WOEG, IRRI, SEDU). Hence, only the expression of Surprise has high levels of bimodal matching in all language groups. This result supports previous findings about the similarity of surprise across these languages and cultures (T. Shochi et al. 2009). Concerning SEDU and IRRI, the judges showed a significant decrease in audio performances, compared to the high levels of visual modality. This is the strongest mismatch between modalities in terms of performances.

The MFA showed relationships between labels and audio or visual stimuli, including those that are not the targeted performances. The analysis of its third dimension showed that, for the WOEG label which is an expression of politeness (as it aims at taking care of the interlocutor's negative face, following P. Brown and S. Levinson, 1987, conceptualization of linguistic politeness) typical of Japanese culture, the performances selected by L1 judges are the correct ones in both modalities. Conversely, the two non-L1 groups showed a different spread in their choices of performance: the visual performances are more specific (i.e. the distribution across categories of answers is reduced to a shorter list) than the audio ones: visual WOEG performances are associated to that dimension, while the audio performances selected by listeners are spread across the possible categories and do not show significant association to the dimension. The MFA revealed also that the US group selected the Sincerity (another politeness strategy) visual performance for this label; both visual performances share similar nodding patterns. The perceptual difference observed for this affect between L1 and non-L1 groups may be linked

to the specific voice quality features that are associated with this expression; typically performed in a pressed voice (T. Sadanobu 2004) as a means to convey a sense of suffering for imposing on a "higher-up". Acoustic analysis of such performances includes tenseness, breathy voice or harsh phonation as possible cues to Japanese's *Kyoshuku* (T. Shochi, 2008). In occidental cultures, it is not customary to use such a voice quality to express politeness; hence the variety of associations that were observed in the audio modality, and thus the relative lack of consistency. This fact had been already experimentally shown in T. Shochi et al. (2009); our findings here confirm the perceptual distortion for this culture-specific strategy of social politeness for the audio modality, while the specificities of the audio modality are perfectly interpreted within the context of the visual cues (A. Rilliard et al. 2009). The complexity of the association process which is underlined here shows that the conventionality of the WOEG audio cues in the Japanese culture is not shared by Japanese-naïve judges, even if they may interpret them correctly: it is not an obvious choice for them.

The two other expressions of politeness, POLI and SINC, received low ratios of performance selection, even by the JP group. This may be explained by their strong expressive similarities, which led subjects to select both types of behavior without discriminating between them – hence the cluster they form in the MFA output. For these affects, the visual performance of POLI and the audio performance of SINC tended to be chosen as best "polite" prototypes.

Another cluster is based on CONT, IRON and OBVI, a cluster which is strongly related in the dendrogram of Figure 5 to the IRRI expression: all are social affects conveying an imposition trait. The three expressions composing the cluster received low selection ratio for all groups, and show important intra-cluster confusions. These three labels are reconstructed by judges using productions of Irony and Obviousness for all groups, and also of CONT by the Japanese group, while the US and FR groups linked the contemptuous performances for the Irritation cluster. This confirms the similarity in terms of performances for these two expressions across cultures (M. Guerry et al. 2014; A. Rilliard and J. A. Moraes 2017), plus the negative trait that may be associated to Irony in the Japanese culture (see also M. Guerry et al. 2016) : it is here linked to Contempt by the Japanese group, but not by the two other groups, who link Contemptuous performances with Irritation.

The variety of selected performances for the Seduction label illustrates potential cultural differences regarding this expression. While the L1 judges select audio seduction performances, US and FR judges seem to build their selections on the basis of visual seduction performance, mixed with audio WOEG. This is interesting as the low-pitched and breathy voice typical of WOEG, which is not used by these groups for "polite" situations (see above), is described as typical of an occidental "charming" voice in the literature (e.g. P. R. Léon 1993; see also D. Erickson et al. 2020).

5. Conclusion & Perspective

At a more global level of analysis (cf. Figure 5 dendrogram), the distinction between expressions of relationship vs. imposition is interesting in that is shows the use of voice and facial features in the management of interpersonal relationships. The first set of expressions aims at continuing the interaction, protecting face, so that social interaction continues; on the opposite side, these dominant expressions impose the speaker's view on the listener and work to cut short the interaction. This opposed dimension of interaction may be nicknamed "buzz-in" or "buzz-out" expressions, respectively. The fact the visual modality was found to be more reliable across cultures, while the audio performances were more subject to change, supports the theoretical description of social affects, or prosodic attitudes, as constructs that may be divided in two broad sets – one expressing propositional meanings, and the other conveying social or behavioral expressions (A. Wichmann 2000). Ongoing research suggests that audio and visual modalities control the organization of the code, with propositional expressions, as part of the linguistic meaning being organized mainly by audio cues, while social interactions rely more on visual ones (J. A. de Moraes and A. Rilliard 2014).

Future plans are to study gender effects on perceptual behaviors, which in this study was not possible due to the limited number of participants. Other applications of this research about cultural similarities/differences of audio-visual affective expressivities are important, as they relate to second language learning/teaching, as well as to the field of emotion recognition psychotherapy.

Acknowledgments

This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of ANR-2010-JCJC-1902-01 and of the "Investments for the fu-

ture" Programme IdEx Bordeaux (ANR-10-IDEX-03-02), Cluster of excellence CPU. "This research was also partly supported by the Major Program of the National Social Science Fund of China (13&ZD189). We are deeply indebted to S. Detey (Waseda University), and students from Bordeaux and students from Waseda University (Japan).

References

Abelin, A. (2004) "Cross-cultural multimodal interpretation of emotional expressions—an experimental study of Spanish and Swedish." *Proceeding of Speech Prosody 2004*.

Aron, A. (2018) "L'abaissement de la fréquence fondamentale comme pratique de seduction." *Proceeding of XXXIIe Journées d'Études sur la Parole*, Aix-en-Provence, France.

Aubergé, V. (2002) "Prosodie et émotion." *Actes des deuxiemes assises nationales du GdR I3*, 263–273.

Auberge, V. and M. Cathiard (2003) "Can we hear the prosody of smile?" *Speech Communication* 40 (1), 87–97.

Baayen, R. H. (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Barkhuysen, P., E. Krahmer and M. Swerts (2010) "Cross modal and incremental perception of audiovisual cues to emotional speech." *Language and speech* 53(1), 3–30.

Brown, P. and S. C. Levinson (1987) *Politeness: Some universals in language usage* 4. Cambridge university press.

Colavita, F. B. (1974) "Human sensory dominance." *Perception & Psychophysics* 16 (2), 409–412.

Collignon, O., S. Girard, F. Gosselin, S. Roy, D. Saint-Amour, M. Lassonde and F. Lepore (2008) "Audio-visual integration of emotion expression." *Brain research* 1242, 126–135.

Daneš, F. (1994) "Involvement with language and in language." *Journal of pragmatics* 22 (3), 251–264.

De Gelder, B., J. Vroomen and G. Pourtois (1999) "Seeing cries and hearing smiles: Crossmodal perception of emotional expressions." *Advances in psychology* 129, 425–438.

De Gelder, B., K. B. Böcker, J. Tuomainen, M. Hensen and J. Vroomen (1999) "The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses." *Neuroscience letters* 260 (2), 133–136.

Ekman, P. and W. V. Friesen (1978) *Manual for the facial action coding system*. Consulting Psychologists Press.

Erickson, D., S. Kawahara, A. Rilliard, R. Hayashi, T. Sadanobu, Y. Li and K. Obert (2020) "Cross cultural differences in arousal and valence perceptions of voice quality." *Proc. 10th International Conference on Speech Prosody 2020,* 720–724.

Erickson, D., A. Rilliard, J. de Moraes and T. Shochi (2018). "Personality judgments based on speaker's social affective expressions." In Qiang Fang, Jianwu Dang, Pascal Perrier, Jianguo Wei, Longbiao Wang and Nan Yan (eds.). *Studies on Speech Production*, Lecture Notes in Artificial Intelligence 10733, 3–13, Springer International Publishing.

Fónagy, I. (1982) *La vive voix, Essais de psycho-phonétique*. Paris: Payot.

Fourer, D., T. Shochi, J.-L. Rouas, J.-J. Aucouturier and M. Guerry (2014) "Going ba-na-nas: Prosodic analysis of spoken Japanese attitudes." *Proceedings of Speech Prosody 2014*.

Fujisaki, H. (1971) "A model for the generation of fundamental frequency contours of Japanese word accent." *Journal of the Acoustic Society of Japan* 27, 445–453.

Gries, S. T. (2013) *Statistics for linguistics with R: A practical introduction (2nd edition)*. Walter de Gruyter.

Gu, W., T. Zhang and H. Fujisaki (2011) "Prosodic analysis and perception of Mandarin utterances conveying attitudes." *Twelfth Annual Conference of the International Speech Communication Association*.

Guerry, M., A. Rilliard, D. Erickson and T. Shochi (2016) "Perception of prosodic social affects in Japanese: A free-labeling study." *Proc. Speech prosody*, 811–815.

Guerry, M., T. Shochi, A. Rilliard and D. Erickson (2014) "Perception of prosodic social affects in French: A free-labeling study." *18th International Congress of Phonetic Sciences (ICPhS 2015)*, Glasgow, UK.

Gunes, H. and M. Piccardi (2009) "Automatic temporal segment detection and affect recognition from face and body display." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39 (1), 64–84.

Hall, E. (1966) *The hidden dimension*. New York: Doubleday.

Harkins, J. and A. Wierzbicka (2001). *Emotions in crosslinguistic perspective* (Vol. 17). Mouton de Gruyter.

Husson, F., S. Lê and J. Pagès (2017) *Exploratory multivariate analysis by example using R*. CRC press.

Jessen, S. and S. A. Kotz (2015) "Affect differentially modulates brain activation in uni-and multisensory body-voice perception." *Neuropsychologia* 66, 134–143.

Kawahara, H., M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno (2008) "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation." *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, 3933-3936.

Kleinsmith, A. and N. Bianchi-Berthouze (2013) "Affective body expression perception and recognition: A survey." *IEEE Transactions on Affective Computing* 4 (1), 15–33.

Léon, P. R. (1993). *Précis de phonostylistique. Parole et expressivité*. Paris: Nathan.

Li, A., Q. Fang and J. Dang (2013) "Emotional McGurk Effect? A Cross-Cultural Investigation on Emotion Expression under Vocal and Facial Conflict." *CCL 2013*

Computer Science Psychology.

Massaro, D. W. and M. M. Cohen (2000) "Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception." *Journal of the Acoustical Society of America* 108 (2), 784–789.

McGurk H. and J. MacDonald (1976) "Hearing lips and seeing voices." *Nature* 264 (5588), 746–748.

Mixdorff, H, A. Hönemann, A. Rilliard, T. Lee and M. K. H. Ma (2017) "Audio-visual expressions of attitude: How many different attitudes can perceivers decode?" *Speech Communication* 95, 114–126.

de Moraes, J. A. (2008) "The pitch accents in Brazilian Portuguese: Analysis by synthesis." *Proc. Speech Prosody* 389–397.

de Moraes, J. A. and A. Rilliard (2014) "Illocution, attitudes and prosody: A multimodal analysis." In T. Raso and H. Mello (eds.) *Spoken Corpora and Linguistic Studies*, (studies in corpus Linguistics 61), 233–270 Amsterdam: John Benjamins.

Noller, P. (1985) *Nonverbal communication and marital interaction*. New York: Pergamon.

Ohala, J. J. 1984. "An ethological perspective on common cross-language utilization of F0 of voice." *Phonetica* 41, 1–16.

Pavlenko, A. (2005) *Emotion and Multilingualism*. Cambridge: Cambridge University Press.

Pierrehumbert, J. and J. B. Hirschberg (1990) "The meaning of intonational contours in the interpretation of discourse." *Intentions in communication* 271–311.

Provost, E. M., Y. Shangguan and C. Busso (2015) "UMEME: University of Michigan Emotional McGurk Effect Data Set." *IEEE Transactions on Affective Computing*.

R Core Team (2017) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. https://www.R-project.org

Rilliard, A., D. Erickson, T. Shochi and J. A. D. Moraes (2013) "Social face to face communication—American English attitudinal prosody." *Proc Annual Conference of the International Speech Communication Association (INTERSPEECH 2013),* Lyon, France, 1648–1652.

Rilliard, A. and J. A. Moraes (2017) "Social affective variations in Brazilian Portuguese: A perceptual and acoustic analysis." *Revista de Estudos da Linguagem* 25(3), 1043–1074.

Rilliard, A., J. A. Moraes, D. Erickson and T. Shochi (2014) "Cross-cultural perception of some Japanese politeness and impoliteness expressions." In F. Baider and G. Cislariu (eds.) *Linguistic approaches to emotion in context*, 251–276, Amsterdam: John benjamins.

Rilliard, A., T. Shochi, J.-C. Martin, D. Erickson and V. Aubergé (2009) "Multimodal indices to Japanese and French prosodically expressed social affects." *Language and speech* 52 (2-3), 223–243.

Romney, A. K. and C. C. Moore (1998) "Toward a theory of culture as shared cognitive structures." *Ethos* 26(3), 314–337.

Romney, A. K., C. C. Moore, W. H. Batchelder and T. L. Hsia (2000) "Statistical methods for characterizing similarities and differences between semantic structures." *Proceedings of the National Academy of Sciences* 97(1), 518–523.

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

Scherer, K. R. (2009) "Emotions are emergent processes: they require a dynamic computational architecture." *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1535, 3459–3474.

Sekiyama, K. and Y. Tohkura (1991) "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility." *Journal of the Acoustical Society of America* 90 (4, Pt 1), 1797–1805.

Shan, C., S. Gong and P. W. McOwan (2007) "Beyond facial expressions: Learning human emotion from body gestures." *BMVC* 1–10.

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.* Gipsa-Lab, Univ. de Grenoble 3.

Shochi, T., D. Erickson, K. Sekiyama, A. Rilliard and V. Aubergé (2009) "Japanese children's acquisition of prosodic politeness expressions." *Proceeding of Interspeech 2009*, Brighton, UK, 1743–1746.

Shochi, T., D. Fourer, J.-L. Rouas, M. Guerry and A. Rilliard (2015) "Perceptual Evaluation of Spoken Japanese Attitudes." *Proc. 18th International Congress of Phonetic Science*, Glasgow, Scotland UK.

Shochi, T., A. Rilliard, V. Auberge and D. Erickson (2009) "Intercultural perception of English, French and Japanese social affective prosody." *The Role of Prosody in Affective Speech*, 32–59. Peter Lang.

Spencer-Oatey, H. (2005) "(Im)Politeness, face and perceptions of rapport: unpackaging their bases and interrelationships." *Journal of politeness research* 1(1), 95–119.

Summerfield, Q. (1992) "Lipreading and audio-visual speech perception." In V. Bruce, A. Cowey, A. W. Ellis and D. I. Perrett (eds.) *Processing the facial image*, 71–78. Clarendon Press/Oxford University Press.

Takagi, S., S. Hiramatsu, K. Tabei and A. Tanaka (2015) "Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality." *Frontiers in integrative neuroscience* 9 (1).

Tartter, V. C. (1980) "Happy talk: Perceptual and acoustic effects of smiling on speech." *Perception and Psychophysics* 27, 24–27.

Tartter, V. C. and D. Braun (1994) "Hearing smiles and frowns in normal and whisper registers." *Journal of the Acoustical Society of America* 96 (4), 2101–2107.

Venables, W. N. and B. D. Ripley (2002) *Modern applied statistics with S-PLUS (4th edition).* Springer.

Walker-Andrews, A. S. and W. Grolnick (1983) "Discrimi-

nation of vocal expressions by young infants." *Infant Behavior and Development 6* (4), 491–498.

Wichmann, A. (2000) "The attitudinal effects of prosody, and how they relate to emotion." *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Wierzbicka, A. (1985) "Different cultures, different languages, different speech acts: Polish vs. English." *Journal of Pragmatics* 9(2–3), 145–178.

Wierzbicka, A. (1996) "Japanese cultural scripts: Cultural psychology and "cultural grammar"." *Ethos*, 24(3), 527–555.

(Received Jul. 31, 2020, Accepted Nov. 25, 2020,
e-Published Dec. 30, 2020)