Journal of Speech Sciences 6(1): 27-45. 2017.
Available at: http://revistas.iel.unicamp.br/joss

PERCEPTION OF EXPRESSIVE PROSODIC SPEECH ACTS PERFORMED IN USA ENGLISH BY L1 AND L2 SPEAKERS

RILLIARD, Albert [1,3] *
ERICKSON, Donna [2]
DE MORAES, João Antônio [3]
SHOCHI, Takaaki [4]

[1] LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[2] Sophia University, Tokyo, Japan
[3] Federal University of Rio de Janeiro, CNPq, Brazil
[4] CLLE-ERSSàB UMR5263 & LaBRI UMR5800, Bordeaux, France

Abstract: *Attitudes have been described for different languages, with varying labels or contexts of occurrence for same labels. It renders cross-cultural comparison uncertain. A corpus was designed to bypass these limitations. This paper focuses on USA English produced by L1 and L2 speakers. The best performances in 9 attitudes are used in a forced-choice test, in both audio and visual modalities. Results show that 6 categories group the presented attitudes in coherent sets. The cultural origin affects marginally the categorisation of the expressions. An acoustic analysis of the fundamental frequency and intensity allows to test the predictions of two theoretical propositions – the Frequency code and the Effort code. It concludes to a main coherence of cross-language expressivity, and discusses differences. For negative expressions of imposition, L1 speakers follow the Frequency code – and L1 listeners expect this; L2 speakers use the Effort code in the same situations, leading to confusions in the audio-only modality. Differences for seduction and irony are also discussed.*

Keywords: prosody, attitude, cross-cultural comparison, multimodal perception

1 Introduction

Works by Scherer and colleagues on the expression of emotions in voice (28) underlie the comparability of expressive behaviours across cultures (30; 29), and predict acoustical variations from physiological changes induced in speakers by emotional states (2). More particularly, Goudbeek & Scherer (8) show relations between the conceptual dimensions of arousal, valence and potency, and acoustical measures in voice. They notably found a high cross-cultural coherence, and link the dimension of arousal to most of the observed acoustic variance in the voice – and notably to the fundamental frequency (F0) level; the potency dimension is also related to changes in the F0 level, while the valence is related to measures of intensity and spectrum slope. It is interesting to link these findings with the predictions made by Gussenhoven (9) with his "Effort code", which could be related with the arousal dimension of expressions, and also predicts a rise of F0. Meanwhile, the raised F0 level related in Goudbeek & Scherer's study (8) is somehow contradictory to the proposal made by Ohala with his "Frequency code", that links potency to lower and rougher voices (21, 22 – cf. also 9).

Descriptions of the mechanism of vocal folds vibration show that F0 may change under the influence of several factors; one may be the volume of air passing through the glottis, another is related to the tension of the folds (10). Variations in vocal effort are related to the signal energy, to F0, to energy in the higher spectrum and to F1 variations (17, 32). Thus, observed changes in F0 may be respectively due: (i) to changes in vocal effort, when speakers are willing to speak louder; and (ii) to changes in vocal folds settings, when speakers are willing

* Corresponding author: albert.rilliard@limsi.fr

to raise or lower their pitch. These two strategies may be related to Effort code and Frequency code.

Moreover, conventionalization processes have been described, that links the production of affects with cultures and languages. Damasio (6) describes a process that allows an individual to reproduce the effect of external stimuli triggering an emotional reaction, thus explaining the capacities human have to mimic the expressions of emotion, outside the actual feeling. Such a capacity of imitation of recurrent behaviour, linked to experience, is at the basis of the Frequency code's phylogenetic development and symbolic values (23). Such process may thus explain the conventionalization of emotionally triggered expressions down to the linguistic code (16). Wichmann's (33) description of attitudes link them to emotional expressions, for the relation that can be drawn between both types of expressions, but separates clearly both types of variations, showing attitudinal expressions are part of the speaker behaviour in an interaction process, and distinguishing between propositional attitudes that address the meaning of an utterance, from behavioural attitudes that are linked to the speaker's relation with the interlocutor (see 19, 20, for evidence of such a distinction in Brazilian Portuguese).

Such a complex and possibly contradictory picture of voice changes for the expressive variations in voice calls for further work. This paper tries to address two aspects of expressive variations in speech:

- To perceptually measure expressive variations in speech prosody produced by speakers of different linguistic and cultural origins;
- To evaluate the pertinence of the proposed Frequency and Effort codes (23, 9) to predict prosodic changes in given expressive contexts.

This work is based on a corpus of target sentences in USA English, performed by speakers having as their first language (L1) USA English, Japanese or French. After the description of the recorded material and procedure, results of perception tests are detailed, and then discussed in the light of acoustic measures of prosodic changes.

2 Corpus design and recording

One common bias in cross-cultural studies of expressive behaviours (emotions, attitudes, etc.) comes from the use of "folk concepts" (34, 35, 37) to refer to behaviour occurring in complex situations. The translation of words in different languages induces different sets of cognitive references in subjects of different cultural and linguistic origins (e.g. changes of experiential contexts, of moral values). Thus, a "polite" behaviour is not necessary instantiated through the same codes in different cultures (36). To bypass this translation bias, a paradigm has been developed to record a speaker's performance *in situations*, asking them to behave naturally in order to reach a given communication goal (25). The situations and communicative goals are comparable, whatever the languages and cultural origins of the speakers. For each language, sets of speakers were recorded, speaking either their first or second language (L1 or L2). The performances investigated in this paper are those of the speakers recorded for the USA English version of this paradigm, both L1 and L2 speakers.

Sixteen situations corresponding to various attitudes observed in different languages have been designed. Some contexts trigger expressions that are described in all the studied languages. The question in these cases is to observe if such expressions trigger similar expressive strategies across individuals, genders, or cultures. Other contexts did not correspond to any conventionalized behaviour in a given culture. Some languages may lack a lexical entry for a given expressivity. It is the case for the Japanese notion of *kyoshuku*, described as

"*corresponding to a mixture of suffering ashamedness and embarrassment, which comes from the speaker's consciousness of the fact his/her utterance of request imposes a burden to the hearer*" (27, p. 34). The Japanese word *kyoshuku* has no lexical equivalent in e.g. English. Meanwhile, situations exist where a speaker of English may feel/express something akin to *kyoshuku*; will this English speaker then perform something resembling Japanese conventionalized behaviour for *kyoshuku*? These questions are raised for the following 16 social affects, described here with prototypical situations (in each case, A is the recorded speaker, B the interlocutor):

- Admiration (ADMI): A & B are almost the same age and know each other well. Both love French cuisine, and talk about the very delicious food they ate yesterday at a famous French restaurant. The scene is at a coffee shop.
- Arrogance (ARRO): both A & B are from the same university, but A is older and A's father is head of the university and A is a bit of a snob. Both know each other, but are not friends. A organized a social party, and B was not invited to the party, but A is aware of his/her presence during the party. The scene is a party room, and A says to B that only his friends are invited.
- Authority (AUTH): Speaker A is a custom agent; speaker B is a traveller. B is in front of A, requesting permission to enter the country; A needs to impose his authority; the scene is at a custom counter at the airport.
- Contempt (CONT): both A & B are from the same university, but A is older; both know each other, but are not friends. In fact, A really hates B. A organized a social party, and speaker B was not invited, but A is aware of his/her presence. The scene is at a party room
- Doubt (DOUB): A & B are colleagues, same age. A knows that his colleague B didn't go to the baseball game yesterday, but B pretends he went to the game, and A doesn't believe it. The scene is at a coffee shop.
- Irony (IRON): A & B are friends, same age; A is going to Boston to see an important baseball game, and B, who is living in Boston calls A. Unfortunately, the weather in Boston is rainy and A says it's wonderful; the scene is at an airport.
- Irritation (IRRI): A & B are almost the same age and know each other. A is sitting next to B. Suddenly, B starts to smoke, and A is very angry; he wants him/her to stop, expressing his irritation toward speaker B. The scene is at a public place.
- Neutral declarative sentence (DECL): A & B are colleagues, same age; A gives information without any personal perspective; the scene is at a coffee shop.
- Neutral question (QUES): A & B are colleagues, same age. A asks for information, without any personal perspective, awaiting a simple answer. The scene is at a coffee shop.
- Obviousness (OBVI): A & B are colleagues, same age; everyone knows B doesn't like French movies, but A asks B if he likes French movies or not; the scene is at a coffee shop.
- Politeness (POLI): A & B are almost the same age and don't know each other well, but work together professionally. A is sitting next to B; both start social talk. The scene is at a formal party.
- Seduction (SEDU): A loves B and they have an intimate relationship. A gives a compliment to B in a sexually provocative way. The scene is at a clubhouse.
- Sincerity (SINC): B is chief of the section which A belongs to; B is older than A. The chief (B) wants A to take on a big project; A is pleased to be asked to do this, and

expresses his enthusiasm, honesty and sincerity for this task. The scene is at B's office.

- Surprise (SURP): A & B are friends, same age. A didn't know that B can sing well. One day, B makes A listen to his beautiful voice. The scene is at friend's home.
- Uncertainty (UNCE): A & B are colleagues, same age. A saw B at the baseball game yesterday, but is not 100% sure if it was really B; the scene is at a coffee shop.
- "Walking on eggs" (WOEG): B is chief of the section which A belongs to; B is older than A. The chief (B) wants A to do a task which is a lot of work, and it seems to A it is impossible to do this, so A tries to reject this request by trying to make sure her/his boss (B) doesn't get angry for refusing. The scene is at B's office.

The situation coined in English "*walking-on-eggs*" corresponds to a certain extent to the Japanese *kyoshuku* situation, adapted to an American English context. Based on the expressivities triggered from these 16 prototypical contexts, dialogues have been written, that end with speaker A uttering either "Mary was dancing." or "A banana"; these two sentences are produced so to carry a similar communication aim, in a similar social context, than those presented in the prototypical situations. Thus, two target sentences are recorded ("Mary was dancing" and "A banana"), in each of the 16 situations; the recording procedure resulting in 32 recorded sentences for each speaker. The recordings took place in sound-treated rooms, with speaker A facing speaker B and a video camera, a microphone placed at one meter from the speaker's mouth, so to avoid significant intensity changes with head position. To capture reliable level of intensity, the microphone was calibrated before each recording session.

For the USA English corpus, 8 L1 speakers have been recorded (5 females, 3 males), plus 11 speakers having English as their L2 (6 from Japan, 5 from France). A total of 608 audio-visual stimuli have thus been obtained. More details of the recording procedure can be found in (25).

3 Perceptual Evaluations

3.1 Evaluation of the speakers' performances

The two target sentences ("banana" and "Mary was dancing") were evaluated for the quality of the speakers' performances, in each of the situations, and for their expressive communicative goal. This first step is seen as an important validation, as the task of speakers may be difficult. To that aim, listeners (having USA English as their L1) were asked to listen to each performance of the 25 speakers (speaking either in their L1 or in their L2 – USA English).

During the experimental procedure, subjects were first informed of the specific context and of the expressive aim of the speaker; they had then to listen to the performance and rate it on a 1 to 9 scale. Different groups of subjects rated each group of speakers – either the speakers with USA English as their L1, or the groups of speakers with Japanese or French as their L1 (speaking in their L2 - USA English). A detailed analysis of the results is given in (26). Figure 1 depicts the relative performance levels, normalized in z-scores for each listener in order to remove any peculiar use of the scale by a given subject. Attitudes are ranked in descending order of performance in the group of L1 speakers of USA English.

The expressions of surprise and doubt received the highest performance scores, whatever the linguistic origin of the speaker; on the other side of the scale, irony received low scores in all language groups. Half the expressions are produced with no significant difference in

performance level between the L1 and the L2 groups. Amongst the other half, the picture is more complex.

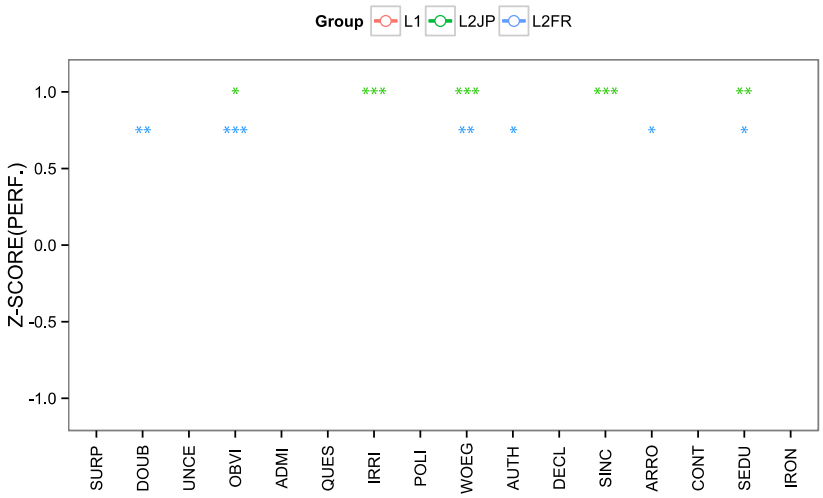

Figure 1: mean z-score of the performance levels received by the speakers in each group of language origin: L1 speakers, L2 from Japanese (JP) and French (FR) origin. Significant differences between the L1 group and either the L1 Japanese or the L2 French groups, for a given attitude, are signalled by stars (*: $p<0.05$; **: $p<0.01$; ***: $p<0.001$).

The speakers of the L1 group outperform L2 speakers from Japan to produce the expressions of *obviousness*, *irritation*, and *seduction*; the L1 speakers outperform L2 speakers from France for the expressions of obviousness, authority and arrogance. On the contrary, the L1 speakers received lower performance scores than the Japanese L2 speakers, while expressing sincerity and "walking-on-eggs"; and the L1 speakers received lower scores than the L2 speaker from France while performing doubt, "walking-on-eggs" and seduction.

3.2 Recognition of the attitudes

The previous test evaluates the performances of the expressive behaviours recorded by 25 speakers. From these files, the two best performances in females and males of each language group (4 recordings for each of the three language groups) have been selected. Amongst these recordings, a subset of 9 expressions was selected to be part of a recognition test. The selected attitudes correspond to the situations of *contempt*, *irony*, *irritation*, *obviousness*, *politeness*, *seduction*, *sincerity*, *surprise*, and *walking-on-eggs*. The selection of these attitudes was made for several reasons:

- Some attitudes received either high or low performance scores and the question of their adequate recognition is thus raised (irony, surprise);
- Some attitudes received different performance scores in different language groups and we would like to know if these ratings are linked to stereotypical representations of the speakers of these origins in the listeners, or to more reliable performances due to cultural training (*irritation*, *obviousness*, *sincerity*, *seduction*, *walking-on-eggs*);
- Expressions of politeness (positive) and contempt (negative) have been added to complete the range of represented expressivities, and add a clear valence opposition, to further study possible variation in their cross-cultural performances (cf. 26).

Subjects had to recognize each expression among the nine possible choices. They were presented with expressions in the audio-only, video-only and audio-visual modalities – in

separate conditions: the presentation order of the audio-only and visual-only conditions is balanced amongst subjects; the audio-visual condition is always given last. Each stimulus is presented once, and subjects give their answers by clicking on the attitude they have perceived, among the nine possible. The presentation order of stimuli is randomized inside each condition of modality, for each subject. All 35 subjects are speakers of USA English.

Results are presented either as a proportion of recognition scores for each presented attitude, and as counts of the number of times a stimulus is recognized as one of the nine proposed labels. The complete contingency table is presented in the annex (table A): the factors influencing the variance in recognition proportions (which correspond to the shaded cell of table A) are analysed with an analysis of deviance (5, p. 636); the main dimensions that explain the dispersion of recognition patterns in the contingency table (which is based on the complete table A, in annex) are analysed with a correspondence analysis (1).

The number of recognized vs. unrecognized attitudes are used as the dependant variable to fit a generalized linear model with a binomial error distribution (24). The analysis of deviance took as explicative factors: the presented attitude (9 levels), the modality of presentation (3 levels), the linguistic origin of the speaker who produced the stimulus (3 levels), and the order of presentation of modalities (2 levels). All interactions between the first three factors are also considered.

A stepwise simplification process of the complete model shows that: (i) there is no significant effect of the modalities' presentation order (and no effect of its interactions either); (ii) the triple interaction between the factors attitude, modality and language group, though having a significant effect, did not improve the overall quality of the model, as measured using the AIC criterion (cf. 5, p. 656). The two-way interaction between language group and modality being not significant, it was also removed from the model. The final minimal model thus uses the three remaining main factors (the presented attitude, the modality, the language group) plus the two-way interactions between attitude and modality, and between attitude and language group (cf. table 1).

All the factors kept in the minimal model have highly significant effects on the recognition score. Figure 2 presents the mean proportion of recognition. Attitudes are sorted according to their performance scores (cf. figure 1): it can be observed that performance and recognition are not directly linked – as for example, seduction is amongst the best-recognized expressions. There is, nevertheless, a tendency of decreasing recognition scores with performance. The effect of the modality factor is mostly linked to the highest scores obtained in the bimodal condition. L1 speakers also produce attitudes that are better recognized than those of the L2 speakers – Japanese speakers receiving the lowest scores.

Table 1: output of the analysis of deviance, run on the proportion of recognized attitudes. The factors (the presented attitude, the modality of presentation, the language group) are listed in rows, as well as the two considered interactions. Their associated degrees of freedom (df) and the explained deviance is reported, with the residual deviance and df, and the probability to reject the null hypothesis, based on a chi square distribution.

	df	**deviance**	**resid. df**	**resid. dev**	**p**
NULL	161	870.29			
Attitude	8	436.37	153	433.91	< 0.001
Modality	2	76.24	151	357.67	< 0.001
Group	2	73.54	149	284.14	< 0.001
Attitude:Modality	16	84.76	133	199.38	< 0.001
Attitude:Group	16	40.99	117	158.39	< 0.001

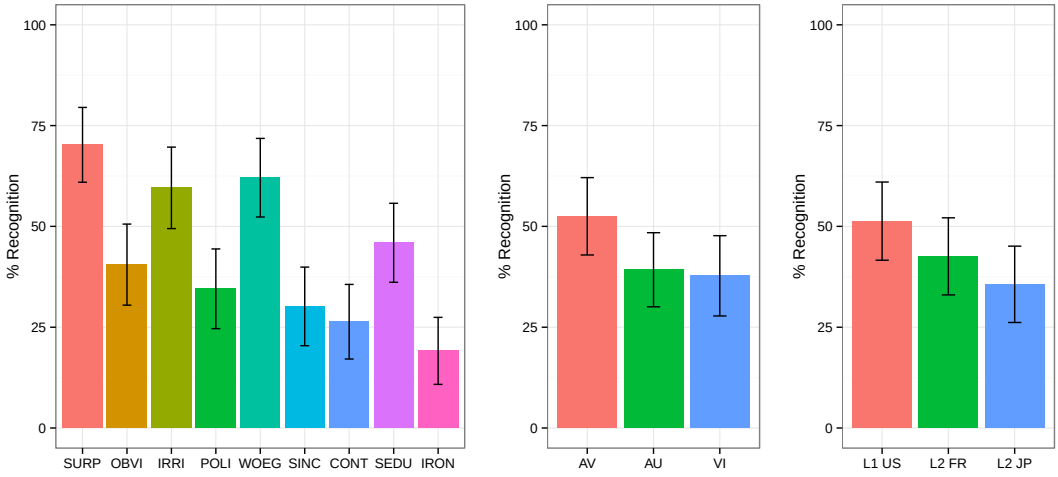

Figure 2: mean proportion of good recognition answers received by each level of the three factors Attitude (left graph), modality (middle) and language group (right graph). Errors bars represent the confidence intervals at 95%.

Figure 3 illustrates the significant effects of the two interaction terms, between (i) attitude and modality and (ii) attitude and language group. The effect of modality on attitudes is specific mostly in the case of seduction, where a dominance of the visual modality is observed over the audio one (and also, to a lesser extent, for irritation, sincerity, contempt and irony); conversely, the audio tend to play the main role in surprise, walking-on-eggs, obviousness and politeness).

The effect of language group on attitudes' recognition is mostly observed through the relative scores of the French vs. Japanese L2 speakers, while L1 speaker always receive the highest scores. French speakers produce attitudes that receive higher recognition scores than Japanese productions in most cases (and interestingly especially for the walking-on-eggs attitude), except for obviousness. For the expressions with the lowest recognition scores, differences between language groups are small.

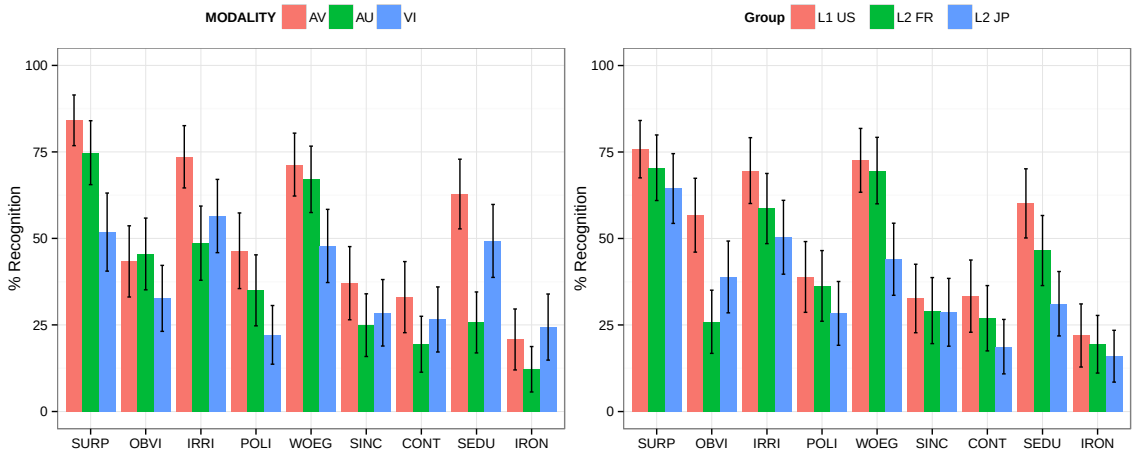

Figure 3: mean proportion of good recognition answers received by each level of the two interactions between attitude and modality (left), and between attitude and language group (right). Errors bars represent he confidence intervals at 95%.

Table 2: output of the CA for the 9 columns, presenting the factor scores (F), the contributions (ct) and the squared cosines (cos) for the first 4 dimensions. Contributions and squared cosines are multiplied by 1000 and rounded for convenience.

	F_1	F_2	F_3	F_4	ct_1	ct_2	ct_3	ct_4	cos_1	cos_2	cos_3	cos_4
CONT	-0.41	0.07	-0.29	0.05	34	1	21	1	226	7	109	3
IRON	-0.07	-0.08	0.17	0.00	1	1	6	0	7	11	44	0
IRRI	-0.76	0.36	-1.03	-0.61	150	37	350	192	256	59	474	166
OBVI	-0.33	0.05	-0.05	0.35	34	1	1	79	176	4	5	208
POLI	-0.27	-0.20	0.41	0.61	17	10	51	180	72	40	170	385
SEDU	-0.27	-0.60	1.33	-1.06	13	70	402	398	22	107	531	336
SINC	-0.22	-0.08	0.26	0.55	12	2	20	139	70	9	96	424
SURP	1.24	1.38	0.24	-0.14	364	486	18	9	437	539	17	6
WOEG	1.39	-1.37	-0.74	-0.09	374	392	131	3	444	430	124	2

To analyse further the recognition results, we then look at the relations between these attitudinal expressions, by analysing the recognition errors. The subjects' answers (selection of one attitude amongst the nine labels) are pooled in a contingency table, for each presented attitude, in each modality, and for each language group (i.e. the significant factors of the previous analysis). This 81 x 9 matrix (81 levels for the three crossed factors, 9 possible answers) is used as the input of a correspondence analysis (CA), based on R's FactoMineR library (12). This analysis allows an observation of the relationships between the stimuli presented, as they are described by L1 listeners on the basis of nine attitudinal labels. Results of the analysis are presented in table 2 by columns (there are too many levels on the rows to present the complete data here). The first four dimensions of the analysis explain more than 85% of the observed variance. Of these four dimensions, plotted on figure 4, one can observe the main distinctions made by listeners, by using the nine labels to rate the stimuli. The four labels that are most distinctively used by listeners are: walking-on-eggs, surprise, seduction and irritation. These labels are also linked to most of the corresponding stimuli. The other five labels show more confusion in their use, and the corresponding stimuli also received more error judgements. To better analyse these confusions, an agglomerative hierarchical clustering was performed on the factor scores obtained from the CA, using the HCPC procedure of R (12).

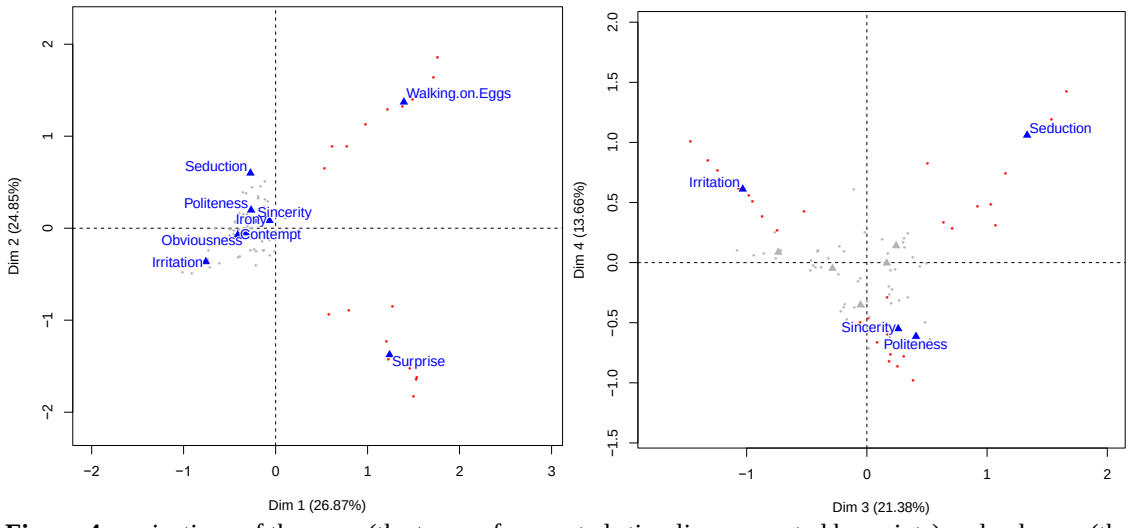

Figure 4: projections of the rows (the types of presented stimuli, represented by points) and columns (the nine attitudinal labels, represented by triangles), on the dimensions 1 & 2 (left graph) and 3&4 (right) of the CA. On the left graph, all the columns' names are shown; on the right graph, only the 4 columns with

the higher squared cosine are named. Individual levels of the rows are not named on these graphs; those with a squared cosine superior to 0.7 are plotted in red, others in grey.

The agglomeration of each type of stimuli presented to listeners is represented as a tree in figure 5. On this tree, one can analyse the similarities and differences perceived between types of stimuli, commencing by the most important clusters. The first four clusters (numbered in order, by cutting the tree from the top) respectively regroup mostly rows corresponding to these types of stimuli: walking-on-eggs, surprise (the first two clusters contain all and only these types of stimuli), seduction and irritation – i.e. the expressions that correspond to the labels most coherently used (cf. figure 4). These four clusters give an "obvious" separation, in light of the preceding analysis. The fifth and sixth clusters group together different kinds of expressions, but are still mostly homogeneous. Adding a seventh cluster bringing no more explanations, we keep this number of 6 clusters. For each of these six clusters (cf. figure 5) obtained amongst the rows of the CA matrix, we look at which columns (i.e. which of the attitudinal labels) are significantly more often used to describe the corresponding set of stimuli: these results are summarized in table 3.

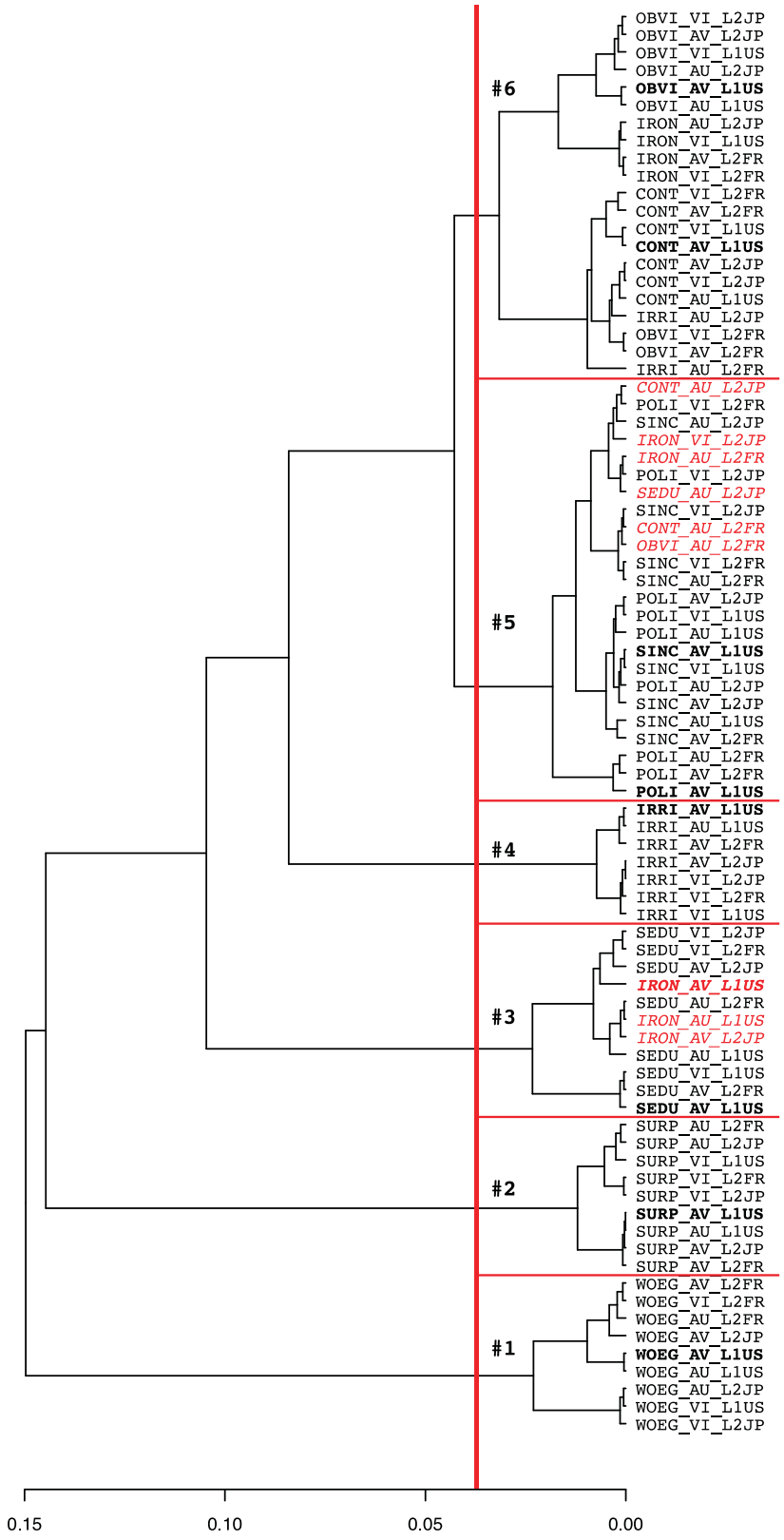

Figure 5: tree representing the hierarchical classification of the type of stimuli presented to listeners, according to the distribution obtained thanks to the CA. The vertical red line indicates the separation into 6 clusters (see text). The audio-visual performances of the L1 speakers are signalled in bold face; the types of stimuli misclassified by listeners are in red italic fonts (see text for details).

Table 3: List of labels that are over-represented in each of the six clusters (named after fig. 4), as compared to their global distribution. The proportion of occurrence of each label in a given cluster and in the whole data (cf. the percentage in cluster and global percentage columns) are compared using a test variable which measures the number of standard deviation separating the two values, the significance of which is tested according to a hypergeometric distribution (15, p. 182).

Cluster	Labels	% in cluster	Global %	V-test	p
#1	Walking-on-Eggs	61.4	9.4	30.1	<0.0001
#2	Surprise	70.5	11.6	31.9	<0.0001
#3	Seduction	47.6	8.8	25.0	<0.0001
#4	Irritation	55.8	12.8	26.4	<0.0001
	Contempt	14.6	9.9	3.6	<0.001
#5	Politeness	26.0	11.9	14.9	<0.0001
	Sincerity	24.3	11.4	13.8	<0.0001
#6	Obviousness	29.5	15.7	13.0	<0.0001
	Irony	13.8	8.4	6.7	<0.0001
	Contempt	14.9	9.9	5.8	<0.0001

The first and second clusters, composed exclusively of stimuli expressing walking-on-eggs and surprise, are linked respectively to the labels of walking-on-eggs and surprise. Thus, these two types of expressions are well produced and recognized as such, and with few confusions with others attitudes, whatever the modality of their presentation and the linguistic origin of the speakers.

Cluster #3 is also linked to the use of a single label: seduction. But it neither contains all types of seductive productions, nor only production of seductions. The audio-only stimuli produced by the Japanese speakers are not recognized as seduction -– they are part of cluster #5 that mostly groups expressions of politeness.

Cluster #3 also contains three types of stimuli supposed to express irony – and interestingly the audio-visual and audio-only versions of irony produced by the L1 speakers, plus the audio-visual irony of L2 Japanese speakers.

Cluster #4 contains only types of stimuli expressing irritation, and these stimuli are predominantly labelled as irritation (thus well recognized), but they also show significant confusions with the label “contempt”. This cluster is only populated with types of stimuli expressing irritation, and with most of them, thus irritated expression may convey a contemptuous meaning. Two types of irritation stimuli (the audio-only performance of both groups of L2 speakers) are not part of this cluster, but can be found in cluster #6, that regroups expressions of obviousness, irony and contempt.

Cluster #5 regroups all the expressions of politeness and sincerity, and they are consistently described by listeners with the labels of politeness and sincerity. This shows the important confusions between both types of expressivities. A few other types of stimuli are part of this cluster. We already mentioned the audio-only seduction by L2 Japanese speakers; one can also find the audio-only expressions of contempt, obviousness and irony by L2 French speakers, and contempt by L2 Japanese plus the visual-only expression of irony by L2 Japanese speakers.

The last cluster #6 contains most of the expressions of obviousness, contempt, and irony. These expressions are described by the labels of obviousness, irony and contempt. It also

contains the expressions of audio-only irritation performed by L2 speakers. This cluster thus mixes negative expressions that lack the distinctiveness of irritation.

4 Acoustic measures

On each vowel of the stimuli presented to listeners for the second experiment, the fundamental frequency and the intensity were measured using Praat (3). Fundamental frequency was measured in semitones, and the mean value of F0 observed for each speaker was subtracted. The A-weighted intensity was measured, because of its proximity to perceived loudness (18); the measured values were also corrected by subtracting the mean intensity measured for each speaker. These normalised values allow an observation of the mean changes induced by attitudes, without taking into account speaker differences. Figure 6 presents the distribution of each attitude, for each group of speaker, according to their mean F0 and intensity values. Two attitudes were added to the set presented to listeners, in order to serve as reference: declaration, that carries a "neutral" prosody; and authority, the description of which varies with theories (cf. introduction). These representations of observed intensity as a function of F0 are inspired by representations of voice range profiles used for example to study the tessitura of singers (14); it is of particular interest to observe the relations between both measures – rise of intensity linking with F0 changes. If this relation is not mandatory, it requires a supplementary control from the speaker to avoid a F0 rise while increasing intensity. Expressions on the upper-right part of this F0 x intensity plane express more strength – e.g. a more active affect (8), or the signalling of a stronger involvement from the speaker (7).

The distributions of these 11 attitudes in the intensity / frequency plane show similarities in the three language groups. Surprise is consistently in the upper right part of the plot, thus performed with a loud and high-pitched voice; this attitude shows the highest pitch in all three groups. Irritation is performed with a loud voice (e.g. more than 10dB higher than declaration for US speakers) – the loudest voice except for surprise performed by L2 French speakers. These two expressions are the most clearly separated from the others.

If one compares the positions of attitudes with reference to declaration, the L1 speakers depart from the two groups of L2 speakers mostly for F0: L2 speakers' declarations are performed with the lowest F0 – while it is in the middle range of L1 speakers' declarations. Declaration is situated in the middle range of intensity, for the three groups. Departing from L2 speakers, L1 speakers do perform some of their expressive attitudes with a lower F0 or at least as low as declaration is. It is typically the case for seduction, and also for authority, contempt obviousness and irony, the four later being also performed with a higher intensity than declaration.

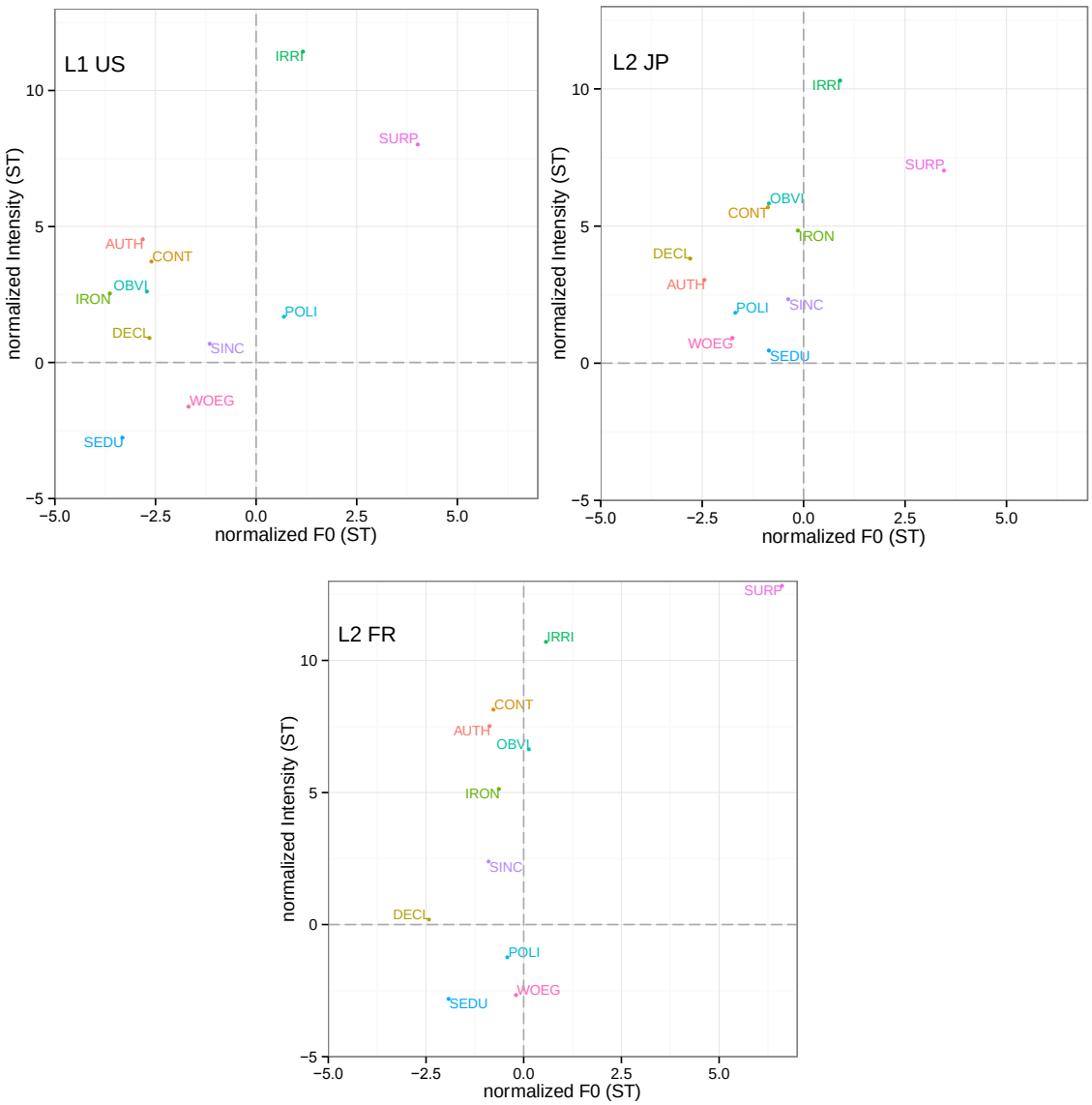

Figure 6: scatterplots of each attitude, according to their mean F0 and intensity values, for each group of speakers (L1 US, L2 Japanese and L2 French speakers).

The three polite expressions (politeness, sincerity and walking-on-eggs) are performed with a higher pitch than declaration – for all three groups. Their use of intensity does differ for sincerity: this attitude is performed with a higher mean intensity than politeness by L2 French speakers (+3.6dB), which is not the case for the other speakers. The expression of walking-on-eggs is performed with a decrease of voice toward the lower-left part of the graph, and especially with a low intensity for all three groups of speakers.

Using a lower pitch for expression of dominance (here typically *authority* and *contempt*) is reminiscent of the predictions of the Frequency code (23) that is accordingly observed in L1 speakers, but the figure for L2 speakers is more complex. These expressions in L1 speakers in addition are performed with a higher intensity. They are thus more pronounced (i.e. more intense) and rather low in voice (i.e. performed with a low F0); this may give an effect of a "bigger" voice, typical of what underlies the Frequency code (i.e. the size of the animal producing the voice). On the contrary, the L2 Japanese speakers produced authority with a lower intensity, and a small rise of pitch, while they produced contempt with a mean rise of intensity of 1.8 dB and a mean rise of F0 of 1.9ST; these two expressions are thus separated for

this group of speakers. L2 French speakers mostly resort to intensity change, keeping F0 rather constant, unless for surprise or seduction. For authority and contempt, an increased intensity over declaration of more than 7dB is observed.

5 Discussion & Conclusion

The figure of attitudinal expressions in USA English shows many similarities across the three groups of speakers (either L1 or L2 speaker, from Japan or France). On the recognition side, the classification of stimuli (obtained from the confusions scores) shows mostly groupings by attitudes, whatever the linguistic origins of the speakers – or the modality. This is in support of the findings of Scherer and colleagues on emotional expressions: the main part of the expressivity may be shared across cultures, and decoded adequately. Note that the speakers recorded for this corpus all have in common the same language (English) – strongly linked to the dominant American culture, which has a strong presence in these three countries. This may explain part of the observed convergence. The perceived convergence is particularly strong for the expressions of surprise and walking-on-eggs. For the former, this result reproduces already documented similarities (31), and is certainly related to the acoustic specificity of the expression (highest F0 and high intensity). It is a newer finding that the expression of walking-on-eggs (an expression modelled from the Japanese-specific expression of *kyoshuku*) could be perceived similarly for expressions produced by Japanese and non-Japanese speakers. This should be investigated in more detail in order to know if the expressive strategy of Japanese speakers differs in their L1; nevertheless, the three groups of culturally-varied speakers produce this expression with a coherent pattern of both low intensity and F0; (26) also report important lengthening for these productions.

Divergences remain that are worth being observed and discussed further. Amongst the attitudes that have been selected because of their cultural specificity, the expression of seduction is typical of the American culture, but is not described as a conventional expression in Japan (at least maybe not for males). This expression received some of the highest recognitions scores, and is regrouped in one coherent cluster – except for the audio-only production by L2 Japanese speakers. Both L1 and L2 French speakers produce seduction with the lowest observed intensity and the lowest F0 (apart from declaration by French); on the contrary, L2 Japanese seduction, if produced with the lowest intensity, uses F0 values closer to the mean. This relatively high F0 value (especially high with regard to the intensity level) brings these seductive productions closer to what is perceived by listeners as expressions of politeness. These L2 Japanese audio-only stimuli are actually grouped with politeness and sincerity, and also acoustically similar. Thus, this difference may be linked to a cultural difference, with the L2 Japanese subject having either no conventional prototypes, or resorting to other conventions for expressing social proximity or anointment (such as e.g. politeness, infant-directed-speech, cf. 11), that may lack the sexually-explicit aspect of the American convention. Meanwhile, this difference is disambiguated in the bimodal presentation: the visual behaviour of these speakers is explicit enough to allow good recognition.

The expressions of politeness and sincerity are perceptually grouped, and show important confusions. They are also produced with a similar expressive strategy: a raised F0, with a relatively low intensity. It should be noted that each group of speakers express acoustically the more marked nature of sincerity over politeness in a different way: with a lowering of F0 for L1 speakers – and conversely with a rise of F0 for the L2 Japanese speakers, and with a rise of

intensity for the L2 French speakers. These varying strategies may explain part of the observed confusions between both attitudes.

A large cluster of mostly negative expressions is observed in the perception data that is described by listeners with the labels of obviousness, irony and contempt. These three expressions, if they differ widely in their definition and communicative aims, share common features – and particularly a dominance feature, and may be perceived as a negative feature. This dominance feature is shared with the expression of authority. The Frequency code would predict a lowering of F0 for dominant expressions, whereas the Effort code predicts a higher F0 for authority because the speaker has to increase his/her implication on the speech act. On the production side, we observe differences between the L1 and L2 speakers: L1 speakers produce stimuli with increased intensity and correspondingly, a lowered F0 (with regard to declaration); they also show similar mean values for all the expressions in this cluster. L2 French speakers also group these four expressions, albeit producing them with even higher intensity and higher F0. The L2 Japanese speakers do not group authority with the other three expressions: these are produced with a higher F0, with little change in intensity. The strategy of both L2 groups is more typical of the Effort code: speakers may thus seek to acoustically mark the speech act's strength rather than the dominance or the negative valence – which may be expressed in the visual modality. These strategic choices may explain the perceptual confusions of the audio-only stimuli of contempt, irony and obviousness that are grouped with the cluster of polite expression. If L1 listeners expect F0 changes predictable under the Frequency code, they may misperceive these expressions – but only in the audio-only version. The visual performance of L2 speakers is not ambiguous for L1 subjects. There may thus be two competing prosodic strategies observed here, one following the Frequency code (used by L1 speakers, and expected by L1 listeners), while the other follow the Effort code – and the observations of (8) – and is used to mark the speech act's strength by both groups of L2 speakers. It will be interesting to verify if these L2 speakers also use the same strategy in their L1 – and if listeners from Japan and France would also mix these expressions of obviousness, contempt and irony with a wider cluster of politeness.

The relative position of these expressions as compared to irritation may give an idea of which strategy is used: L2 speakers tend to produce expression much closer to irritation than do L1 speakers. Irritation is typically produced under the Effort code, and expresses discontent about something. L2 Japanese speakers do not change their expression of authority in this direction (authority sticks close to declaration), which is typical of the importance of the given social position of the speaker in this culture (11), while in more egalitarian societies, one has to express one's authority, which is not necessarily given by context or situation. L1 and L2 French speakers, however, don't show the same strategic choices for this expression: they respectively use the Frequency code and the Effort code.

Notice also that the "contempt" label is used to describe both the irritation cluster and the "obviousness-irony-contempt" one. Let us recall the irritation cluster has a double label but is only composed of irritation stimuli: this shows the conceptual similarities between the two labels, along with the typicality of the expressive behaviour. The second cluster is also interesting as it links the concept of obviousness with negative expression, and typically with irony: such a display of obviousness may be interpreted as a (negative) sarcastic form of ironic remark.

The case of irony is complex – as is this expression. It has been described as a highly varying phenomenon (4), which most importantly expresses a contrast with some other aspects of the discourse. One problem of our methodology regarding this specific attitude lies precisely in that contrastive nature: as we present to listeners extracts brought out of their production

contexts, the changes that are characteristic of the ironic nature may be misunderstood. This may explain the confusion by listeners of irony, as performed by L1 speakers, with seduction. L1 speakers' productions of irony show a very low F0, as seduction (even intensity is different). This similarity, in absence of contextual information, may lead to misinterpretation. The same explanation could stand for audio-only irony by L2 French and visual-only irony by L2 Japanese speakers: without cues to decode the mismatch between some specific cues and the context (as the context is not given), irony is misperceived in some cases. It would be interesting to analyse further the cues that contrast in these cases to test if they can be accurately decoded in an adequate presentation.

6 Acknowledgements

This work was supported by the ANR grant PADE, JSPS Grants A #25240026 and A #23320087. The authors warmly thank the speakers as well as Mariko Kondo and Sylvain Detey from Waseda University for their help, and Caroline Smith for her support of the second author as a Guest Researcher at the Linguistics Department at the University of New Mexico; we also warmly thank the listeners at various universities in the U.S.

REFERENCES

1. Abdi H, Béra M. Correspondence analysis. In: Alhajj R, Rokne J, editors. Encyclopedia of Social Networks and Mining. New York: Springer Verlag; 2014. p. 275-284.
2. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology. 1996:70:614–636.
3. Boersma P, Weenink D. Praat: doing phonetics by computer (version 5.3.32) [computer program]. retrieved October 17, 2012.
4. Bryant GA. Verbal irony in the wild. Pragmatics & Cognition. 2011:19(2):291-309.
5. Crawley MJ. The R book. John Wiley & Sons; 2012.
6. Damasio, AR. Emotion in the perspective of an integrated nervous system. Brain research reviews. 1998:26(2):83-86.
7. Daneš F. Involvement with language and in language. Journal of pragmatics. 1994:22(3-4):251-264.
8. Goudbeek M, Scherer K. Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. The Journal of the Acoustical Society of America. 2010:128(3):1322-1336.
9. Gussenhoven C. The phonology of tone and intonation. Cambridge: Cambridge University Press; 2004.
10. Henrich, N, d'Alessandro, C, Doval, B, Castellengo, M. Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. Journal of the Acoustical Society of America. 2005:117(3):1417-1430.
11. Hill B, Ide S, Ikuta S, Kawasaki A, Ogino T. Universals of linguistic politeness: Quantitative Evidence from Japanese and American English. Journal of Pragmatics. 1986:10:347-371.
12. Husson F, Josse J, Le S, Mazet J. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. R package version 1.27. http://CRAN.R-project.org/package=FactoMineR. 2014.
13. Kerbrat-Orecchioni, C. (2005). Politeness in France: How to buy bread politely. In: Hickey L, Stewart M, editors. Politeness in Europe. Clevedon: Multilingual Matters; p. 29-44.
14. Lamesch S, Doval B, Castellengo M. Towards a more informative Voice Range Profile: the role of laryngeal vibratory mechanisms on vowels dynamic range. Journal of Voice. 2012:26(5):672.e9-672.e18.
15. Lebart L, Morineau A, Piron M. Statistique exploratoire multidimensionnelle. Paris: Dunod; 2000.
16. Léon, P. Précis de phonostylistique: parole et expressivité. Paris: Nathan; 1993.
17. Liénard, JS, Di Benedetto, MG. Effect of vocal effort on spectral properties of vowels. The Journal of the Acoustical Society of America. 1999:106(1):411-422.
18. Liénard J-S, Barras C. Fine-grain voice strength estimation from vowel spectral cues. In: Proceedings of Interspeech; 2013 Aug 25-29; Lyon. ISCA; 2013, 128-132.

19. de Moraes, JA, Rilliard, A. Illocution, attitudes and prosody: A multimodal analysis. In: Raso, T, Ribeiro De Mello, H, editors. Spoken Corpora and Linguistic Studies. Amsterdam: John Benjamins Publishing Company; 2014. p. 233-270.
20. de Moraes, JA, Rilliard, A. Prosody and Emotion. In: Armstrong, ME, Henriksen, N, del Mar Vanrell, M, editors. Interdisciplinary approaches to intonational grammar in Ibero-Romance. Amsterdam: John Benjamins Publisher; 2016. p. 135-152.
21. Ohala JJ. Cross-language use of pitch: an ethological view. Phonetica. 1983:40(1):1-18.
22. Ohala JJ. An ethological perspective on common cross - language utilization of f0 of voice. Phonetica. 1984:41(1):1-16.
23. Ohala JJ. The frequency code underlies the sound symbolic use of voice pitch. In: Hinton L, Nichols J, Ohala JJ, editors. Sound symbolism. Cambridge: Cambridge University Press; 1994. p. 325-347.
24. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. URL http://www.R-project.org/. 2014.
25. Rilliard A, Erickson D, Shochi T, de Moraes JA. Social face to face communication – American English attitudinal prosody. In: Proceedings of Interspeech; 2013 Aug 25-29; Lyon. ISCA; 2013, 1648-1652.
26. Rilliard A, Erickson D, Shochi T, de Moraes JA. US English attitudinal prosody performances in L1 and L2 speakers. In: Proceedings of Speech Prosody; 2014 May 20-23; Dublin. ISCA; 2014, 895-899.
27. Sadanobu T. A natural history of Japanese pressed voice. Journal of the Phonetic Society of Japan. 2004:8(1):29-44.
28. Scherer KR. Vocal affect expression as symptom, symbol, and appeal. In: Papousek H, Jürgens U, Papousek M, editors. Nonverbal vocal communication: Comparative and developmental approaches. Cambridge: Cambridge University Press; 1992. p. 43-60.
29. Scherer KR, Wranik T, Sangsue J, Tran V, Scherer U. Emotions in everyday life: Probability of occurrence, risk factors, appraisal and reaction patterns. Social Science Information. 2004:43(4):499-570.
30. Scherer KR, Wallbott HG. Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology. 1994:66:310-328.
31. Shochi T, Rilliard A, Aubergé V, Erickson D. Intercultural perception of English, French and Japanese social affective prosody. In: Hancil S, editor. The role of prosody in affective speech. Bern: Germany; 2009. Linguistic Insights. 97; p. 189-220.
32. Traunmüller, H, Eriksson, A. Acoustic effects of variation in vocal effort by men, women, and children. The Journal of the Acoustical Society of America, 2000:107(6): 3438-3451.
33. Wichmann, A. The attitudinal effects of prosody, and how they relate to emotion. In: Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion; 2000 September 5-7; Newcastle. ISCA; 2000, 143-148.
34. Wierzbicka A. A semantic metalanguage for a cross-cultural comparison of speech acts and speech genres. Language in Society 1985:14(4):491-513.
35. Wierzbicka A. Defining Emotion Concepts. Cognitive Science. 1992:6:539-581.
36. Wierzbicka A. Japanese cultural scripts: Cultural psychology and "cultural grammar". Ethos. 1996:24(3):527-555.
37. Wierzbicka A. Empirical Universals of Language as a Basis for the Study of Other Human Universals and as a Tool for Exploring Cross-Cultural Differences. Ethos. 2005:33(2):256-291.

ANNEX

Table A contains the number of responses each type of stimuli used in experiment 2 (grouped by expressed attitude, modality of presentation, and language of the speaker; see section 3.2 for details) received from the listener, for the 9 possible answer categories.

Table A: contingency table showing the number of answers given by listeners to stimuli expressing a given attitude (ATT), in a given presentation modality (MOD), by a speaker of a given language (SPK) (along the lines), for each of the 9 possible answers (along the columns; see text for details); cells in grey shade indicates correct guess (scores used in the deviance analysis), other cells indicates the confusions made by listeners.

Presented stimuli			**Responses**								
ATT	**MOD**	**SPK**	**CONT**	**IRON**	**IRRI**	**OBVI**	**POLI**	**SEDU**	**SINC**	**SURP**	**WOEG**
		USA	13	8	11	9	2	0	2	0	0
	AU	Fr	9	1	8	10	7	0	11	0	1
		Jp	5	4	6	11	5	2	6	6	3
		USA	17	4	9	12	2	0	1	0	0
CONT	AV	Fr	18	2	18	5	2	0	2	0	0
		Jp	11	8	10	7	2	0	1	6	3
		USA	15	1	9	10	3	1	3	1	2
	VI	Fr	11	5	19	4	2	0	4	0	2
		Jp	11	10	10	8	0	0	2	5	2
		USA	4	4	2	5	6	14	6	0	3
	AU	Fr	3	3	1	7	10	4	7	6	7
		Jp	5	10	1	12	3	4	8	3	2
		USA	2	13	0	5	5	16	3	0	0
IRON	AV	Fr	3	13	1	10	4	8	7	1	1
		Jp	6	3	2	8	7	10	7	1	4
		USA	6	12	0	9	7	4	5	1	0
	VI	Fr	4	12	0	14	4	9	4	1	0
		Jp	5	10	1	8	8	2	6	8	0
		USA	3	1	34	3	1	0	3	0	0
	AU	Fr	6	1	17	9	1	11	5	0	0
		Jp	5	8	17	6	0	2	2	4	1
		USA	3	0	35	6	0	0	0	1	0
IRRI	AV	Fr	8	0	42	0	0	0	0	0	0
		Jp	5	3	26	8	0	0	1	2	0
		USA	3	2	25	7	2	0	5	1	0
	VI	Fr	8	1	29	7	0	0	0	3	2
		Jp	5	4	25	9	0	0	1	1	0
		USA	3	4	5	29	4	0	2	0	0
	AU	Fr	6	1	8	13	6	0	9	2	0
		Jp	4	8	4	22	0	2	1	6	1
		USA	0	2	2	31	6	0	6	0	0
OBVI	AV	Fr	6	4	12	12	5	0	2	4	0
		Jp	9	6	5	18	5	0	3	1	1
		USA	4	2	5	20	8	1	2	5	0
	VI	Fr	7	4	9	10	6	0	2	7	0
		Jp	6	6	4	16	5	1	5	5	0

Presented stimuli			Responses								
ATT	**MOD**	**SPK**	**CONT**	**IRON**	**IRRI**	**OBVI**	**POLI**	**SEDU**	**SINC**	**SURP**	**WOEG**
		USA	0	1	7	8	17	2	11	2	0
	AU	Fr	1	1	1	10	19	5	4	3	1
		Jp	5	2	0	10	13	1	14	2	0
		USA	3	2	1	5	26	0	11	0	0
POLI	AV	Fr	2	4	1	6	24	4	4	0	0
		Jp	1	5	1	7	15	0	13	1	4
		USA	4	5	3	7	13	0	13	0	3
	VI	Fr	4	4	5	8	6	4	10	4	0
		Jp	3	3	3	9	12	1	7	3	6
		USA	8	3	4	3	1	18	2	1	6
	AU	Fr	6	3	0	5	7	15	6	3	3
		Jp	6	6	0	8	8	3	5	5	5
		USA	0	6	0	1	2	34	3	0	0
SEDU	AV	Fr	2	2	0	3	7	32	2	0	0
		Jp	0	4	0	5	5	22	4	5	1
		USA	1	4	0	2	4	31	4	0	0
	VI	Fr	2	4	2	2	12	20	5	1	0
		Jp	0	3	0	0	13	18	8	2	2
		USA	8	2	2	4	11	5	12	0	5
	AU	Fr	10	1	4	10	7	2	13	0	1
		Jp	6	2	0	12	6	0	10	7	0
		USA	2	3	2	7	16	0	18	0	1
SINC	AV	Fr	7	6	2	4	10	1	17	0	1
		Jp	2	1	1	7	9	1	17	5	0
		USA	2	1	4	9	15	0	18	0	0
	VI	Fr	13	2	5	8	5	1	12	1	1
		Jp	7	2	4	8	8	0	10	2	2
		USA	0	2	0	2	0	0	2	43	2
	AU	Fr	0	1	4	1	2	0	2	32	2
		Jp	3	3	0	4	0	0	2	30	3
		USA	0	3	0	1	0	1	1	43	2
SURP	AV	Fr	1	0	1	1	0	0	2	39	0
		Jp	0	3	0	0	1	0	3	36	2
		USA	1	4	3	2	1	1	0	30	9
	VI	Fr	2	4	1	4	3	1	5	22	2
		Jp	5	3	2	5	3	1	5	21	0
		USA	0	1	0	1	2	0	1	2	38
	AU	Fr	1	0	0	2	2	2	2	3	33
		Jp	4	3	1	3	2	6	6	3	22
		USA	0	0	1	0	2	0	1	0	41
WOEG	AV	Fr	0	7	0	4	0	0	1	2	31
		Jp	1	9	0	4	4	2	2	1	27
		USA	3	2	3	3	5	4	5	1	19
	VI	Fr	3	4	2	4	1	0	0	1	30
		Jp	5	6	2	5	2	5	5	3	17