Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Perception of prosodic social affects in Japanese: a free-labeling study

Marine Guerry[1], *Albert Rilliard*[2], *Donna Erickson*[3], *Takaaki Shochi*[1,4]

[1] CLLE-ERRSaB UMR 5263, France
[2] LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
[3] Kanazawa Medical University, Sophia University, Japan
[4] LaBRI UMR 5800, France

marine.guerry@etu.u-bordeaux-montaigne.fr, albert.rilliard@limsi.fr,
ericksondonna2000@gmail.com, takaaki.shochi@labri.fr

Abstract

This paper presents an examination of the variable lexical labels used by listeners to identify 16 social affective expressions in Japanese language, in audiovisual presentations. A free-labeling task allows an open approach to variability in the perception of social affects, that is constrained by pre-defined force-choice paradigms. 27 L1 Japanese listeners participated in the experiment. Subjects were asked to write down one word (noun or adjective) that best describes the intended expression they perceived from the speaker in each stimulus. Results cluster into coherent groups - relative to the expressions intended by the speakers. One Japanese-specific social affect, *kyoshuku* forms one cluster by itself among the 8 main clusters. This result emphasizes its specificity in Japanese culture: this expression was not singularized the same way by L1 French listeners from the same situation. The results also indicate the importance of a separation between assertive and dubitative speech acts in the meaning carried by prosody.

Index Terms: Japanese, multi-modal perception, social affects, free labeling

1. Introduction

The use of prosodic cues during spoken social interactions is well established by studies in many languages. Prosody helps to express various concepts ranging from surprise to irony, or politeness. Most of these works deal with a specific language, for example USA English [1], French [2], Japanese [3], German [4], Brazilian Portuguese [5], Chinese [6], etc. Often targeted to foreign language teaching [7, 2], these approaches focus on the language-specific aspects of prosodic variation, for example in putting forward melodic clichés [8], or concepts that do not exist in another culture (e.g. *kyoshuku* in Japanese – cf. [9, 10]).

Problems arise when one seeks to compare such kinds of prosodic expressions between languages. In perception experiments, translating folk labels raises conceptual shifts that may bias the answers [11]. For the comparison of speakers' production strategies, a similar bias exists if the communication situations where the social affects occur are not similar (i.e. if the role of speakers and listeners, the targeted speech acts, etc. differs to some extend) – as one cannot ascertain an observed prosodic difference originating in two different cultural codes, or in two different interpretations of a folk label under scrutiny.

To bypass these limitations, a recording paradigm has been proposed and applied to different languages [13] (currently recorded in English, Japanese, French, Brazilian Portuguese and German). The naturalness of the prosodic expressions is maximized using scenarios to describe a prototypical interaction (cf. [12]) and performing the scenario with an actual (real person) interlocutor; scenarios are built with a set of controlled interpersonal parameters [13]. These controlled parameters include the hierarchical relationship between speaker and listener, their social distance (cf. [14]), the type of speech act, etc. This results in 16 communication situations, ending with speakers expressing the 16 prosodic expressions on target sentences. The Japanese version is used in the current study.

The perception and decoding of prosodic attitudes has been evaluated in many ways. Uldall [1], for example, uses Osgood's semantic differential technique (based on ratings on a set of scales with opposed adjectives) [15] to measure various semantic aspects conveyed by prosodic contours. Another approach is based on force-choice paradigms, where listeners are asked to recognize the intended attitude in a given set (e.g. [3, 16]). A variant of categorical perception is used by [6] to test the distinctiveness of prosodic expressions along binary scales with two opposed expressions plus a neutral sentence (subjects had to choose among the two ends of the scale or the neutral position). [17] shows that results of categorical perception tests contain information on the main dimensions (valence, arousal) that distinguish these expressions – thus both approaches go in the same direction.

Meanwhile, all these tests are explicitly based on the cognitive ability of subjects to match a set of "folk labels" [11] with what they perceive from the prosodic variations. These tests are thus not suited to study the cross-cultural variations in perception. The present paper presents the results of a free-labeling experiment [18, 19] where listeners had to describe what they understand of the speakers' intended speech act, using any adjective or noun they wished to use. The grouping obtained from these many labels gives a distribution of the presented set of prosodic performances that is *not* dependent on any pre-conception imposed by the experimenters. Comparison of the distributions produced by listeners of various cultural origins thus allows deriving cross-cultural differences in the interpretation of prosodic forms. This paper focuses on the analysis of 16 expressions produced by L1 Japanese speakers, as they are perceived by L1 Japanese listeners. Comparisons are proposed with the distribution of

 doi: 10.21437/SpeechProsody.2016-166

prosodic performance in the same 16 situations produced by L1 French speakers perceived by French listeners.

2. Free-labeling paradigm

2.1. Corpus

The stimuli are extracted from a corpus recorded on the basis of the methodology proposed in [13], where 19 speakers (from Tokyo area, 11 females) have produced two Japanese target sentences (**B**: "*Banana*" -A banana, and **M:** "*Mari wa dansu wo shiteimashita*" -Mary was dancing), in 16 interactional situations that somehow correspond to the following English terms (for details on the situations, cf. [13]): admiration (ADMI), arrogance (ARRO), authority (AUTH), contempt (CONT), doubt (DOUB), irony (IRON), irritation (IRRI), declaration (DECL), question (QUES), obviousness (OBVI), politeness (POLI), seduction (SEDU), sincerity (SINC), surprise (SURP), uncertainty (UNCE), and "walking-on-eggs" (WOEG). The expression "walking-on-eggs" was used to denote a situation corresponding, to some extent, to a situation where Japanese speakers would express "*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 his/her utterance of request imposes a burden to the hearer*" ([9], p. 34).

Based on a perceptive evaluation of the 19 speakers' performances in each of these situations, the two best female and the two best male speakers, for each sentence in each situation, where selected. This results in 128 audio-visual stimuli.

2.2. Experimental paradigm

Japanese listeners (19 females / 8 males) (different from the performance evaluation task) were required to give a label that best describes what they think the speaker expresses, for each of the 128 audio-visually presented stimuli. Stimuli were displayed in a random order. Subjects were asked to write down on a computer device one noun or adjective. There was no constraint about time or number of visualizations.

2.3. Normalization procedure

A total of 799 different labels was recorded. A normalization procedure was then applied. In the case subjects wrote down several words, the first one was kept, as they were instructed to write only one label each time. We also decided to remove all information about tense such as the Japanese particles "*ta*" or "*teiru*". After normalization, a total of 508 different labels remained.

3. Data analysis

A 32x508 contingency matrix was created, that contains the number of times a given label was applied to a given expression for a given sentence by all the subjects (thus mixing the answers received for the four speakers – the 132x508 solution that contains the separated results for each speaker was also tested, but will not be presented here, as it led to a similar global analysis: the detailed differences of idiosyncratic choices cannot be detailed here). This matrix was then used in a correspondence analysis (CA, using R's FactoMineR package [20]) that regroups attitudinal expressions according to their proximity in terms of label usage, reducing the dimensionality of the original matrix: The first 7 dimensions of the analysis were kept, following an elbow criterion. They explain more than 50% of the observed variance among label use; the other dimensions being mostly noise [20]. From analyzing the distribution of the 16 expressions performed with two sentences on the 7 dimensions of the CA, a hierarchical clustering was applied, which led to an 8-cluster solution that optimizes a between-cluster inertia criterion [20]. The dendrogram of figure 1 represents these 8 clusters.

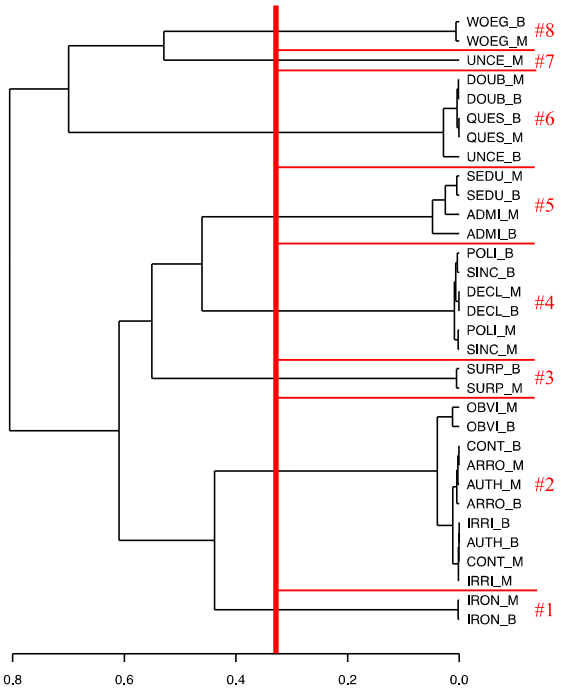

Figure 1: *Dendrogram representing the hierarchical classification of the 16 presented attitudes on each of the 2 sentences, based on the CA's first 7 dimensions. Number indicate the clusters as they are referred to in the text*

3.1. Distribution of the expressions

The tree of figure 1 shows the relative perceptive proximity between the expressions performed on each of the two sentences. The main separation in that tree puts clusters 1 to 5 together, on one hand, and clusters 6 to 8, on the other hand. This corresponds to the first dimension of the CA, and mostly separates assertive expressions from a set of dubitative ones (cf. [21]), along one of the main functions of prosody: the expression of a sentence's mode (declarative, interrogative). Interestingly, the WOEG expression (#8) that corresponds to the Japanese-specific concept of *kyoshuku* is on the dubitative side, while expressions of surprise (#3) are on the assertive one (cf. next section for details).

Among assertive expressions, the main distinction is between clusters 1 and 2 *vs*. clusters 3, 4, and 5. The first set contains mostly negative expressions or impositions; the second set, mostly polite or positive, submissive expressions, plus surprise. Both ironic situations are grouped together, and

form a separate cluster (#1): listeners did identify the prosodic expression elicited by the situation, even though there is no conventional way to express "irony" in Japanese. This differs from the labeling made by the French subjects [22] who mixed ironic prosody on the "Banana" sentence with obviousness, and irony on the "Mary" sentence with the negative cluster #2. French performances of IRON are not singularized as such (note that ironic patterns are reported in this language [16]), maybe because the prosody carries an ironic meaning in its situational context [23] – context that is not presented to the listeners. On the contrary, Japanese speakers may have found a coherent way to behave in these ironic situations. The second negative cluster (#2) groups all the expressions of imposition or negative behaviors (AUTH, IRRI, ARRO, CONT, OBVI). The prosodic expression of obviousness is here grouped with negative expressions, while the French subjects separated it from the cluster of "negative" expressions [22], and used a dedicated cluster for it.

On the dubitative side, and apart from the WOEG cluster, all expressions are grouped but not the expression of UNCE performed with the "Mary" sentence, while the UNCE performed on "Banana" is regrouped with expressions of QUES and DOUB. There is no clear distinction between the expressions of QUES, DOUB and the one for UNCE (cluster 6). This distinction between both UNCE expressions shows the importance of fine details in the situations proposed to speakers for eliciting the attitudes. It may lead to subtle differences, that are perceived and consistently labeled by listeners (cf. next section for the labels). This is illustrated by the two UNCE dialogs with the target M and B sentences:

- "Mary was dancing": speaker A thinks that Mary was dancing at a party, but is not 100% sure – it was very crowded. A and B are colleagues, same age.
    - B: "What was Mary doing when you arrived?"
    - A: "Mary was dancing"
- "Banana": A and B are two friends at a grocery store. They see what is depicted in figure 2.
    - B: "What is this?"
    - A: "A banana"

Figure 2: I*mage used for eliciting uncertainty with the "banana" sentence.*

3.2. Analysis of labels

To better understand the distribution of the 8 clusters, and what kind of interpretation the listeners made of the proposed prosodic expressions, it is important to have a detailed look at the main labels they used in each case. The HCPC function measured which variables (here the labels) are significantly more represented in each cluster. The details of these labels are given in table 1. Only the labels that are used for more than 10% of the observed occurrences for a cluster are reported.

Table 1: *List of the labels used significantly more often to describe a cluster, as compared to their global distribution – and that also represent more than 10% of the observed labels for this cluster. A tentative English translation is given.*

Cluster	Labels	Translations
#1	嘲笑 バカにする	Ridicule Make fun of
#2	怒り	Anger
#3	驚き	Surprise
#4	嬉しい 喜び	Happy Joy
#5	普通	Normal
#6	疑問 疑い	Question Doubt
#7	不確か 自信がない	Uncertain Unconfident
#8	申し訳ない	I'm sorry

The prosody used by Japanese speakers for both occurrences of ironic expressions (cluster 1) is perceived by the listeners as if the speaker tries to ridicule them. It is not specifically perceived as "ironic" (in its USA English sense), but more as a negative mocking.

Cluster 2, with expressions perceived as negative or imposing, is mostly labeled as "anger", but also with a long list of terms, used frequently for this cluster, that could be translated as (in decreasing order of appearance): *irritation, bothersome, self-important, hopping mad, disappointing, authoritative, bad mood, gloomy, arrogance, emphasis, little anger, unwillingly, lazy, obsessive-compulsive, tough, haughtiness, arrogant, impudent, negligent, fed up, look down, careless, overpowering, arrogance, relevant, pretentious, complaint, abruptness, indifference, disgust, contempt, in a bad mood, dislike, dull, irritation, disagreeableness, decision, not interested, self-confidence*.

This is reminiscent of the results for French, which also include a very long and varied list of terms for a similar negative cluster – using terms that can be compared in meaning and scope. Interestingly, while the French separate obviousness from this negative cluster contrary to the Japanese, there is no occurrence of the "obviousness" label in this list. Expressions of OBVI and IRON do not seem to be something characteristic of normal Japanese conversation. It is nonetheless recognized in a meaningful way, when it occurs.

More than 60% of the answers given to the prosodic expressions of SURP are the label "surprise". This is thus a very well recognized and distinctive set of behaviors. Its grouping among the assertive expressions may be linked with confusions made from other expressions (typically some performances of obviousness and admiration), which are also coined "surprise" (respectively 15 and 29 occurrences). This link between assertive expression misperceived as surprise led the agglomeration algorithm to attach the otherwise separated SURP to assertive; but SURP is not, in itself, perceived as something assertive, just as surprise.

Cluster 4 is perceived as “normal” – and contains the most declarative expressions, that eventually also carry a polite behavior. It’s normal to be polite!

Cluster 5, that groups expressions of ADMI and SEDU, is labeled “happy” and “joy”. This is interesting, as is stresses the positive aspect of these expressions, but also underlies the fact that there is no conventional way to express seduction in Japanese. Listeners resort to more general terms which, interestingly, are close to the one used by French for a similar seduction / admiration cluster (“longing” and “joy”). Note that French also use the term “seduction”, while Japanese don't.

Cluster 6, grouping expression of QUES, DOUB and one form of UNCE, is labeled exactly “question” and “doubt”, while the cluster 7 (which contains the second example of uncertainty) is labeled “uncertain”, “unconfident”. We have here a set of clearly dubitative expression, with a kind of gradient in illocutory strength of the expressions that allows listeners to make some distinctions between them.

Finally, the expression of WOEG, corresponding to *kyoshuku*, is mostly labeled “I’m sorry”, which sounds an accurate contraction of [9] definition. Listeners also use a few times (16) the *kyoshuku* word to denote these performances. The typically Japanese expression is thus well recognized. Interestingly, it also shares similarities (if not grouped under the criterion chosen in this analysis) with UNCE, the closest expression in the dengrogram tree. This proximity is also denoted by French listeners (on French performance for the same situation), that group together UNCE and WOEG, of course without naming it in such a specific way. So, culture-specific concepts may share similarities at a higher level of organization of these expressive concepts.

4. Conclusions

Presented here are the results of a free-labeling experiment that allow L1 Japanese listeners to coin what they understood from behaviors of several speakers, in predefined constrained communication situations.

A multidimensional analysis allowed a description of the distribution of the labels, and gives a grouping of the 16 expressive behaviors into 8 clusters – each of which is labeled consistently by a rather limited number of terms.

This dimensional analysis of “prosodic meaning” is built on numerous works about dimensions of meaning [15, 1, 24, 25], that have shown the importance of the *valence*, *activation* and *dominance* dimensions in characterizing many spaces of meaning. One important finding in this study is to uncover the assertive/interrogative dimension [21] as one of the primary ones to separate the 16 prosodic expressions. This assertion/interrogation dimension is indeed very classical for prosodic meaning, but has not been related as an important “dimension of meaning”. This may be due to several factors – and notably the fact that most of the works on the dimensions of meaning are based on isolated written words, and not communicative expressions. Moreover, this dimension is also robust to cross-cultural changes: it is the main dimension found for French using the same 16 expressive situations and paradigm [22], and also one of the main dimensions reported in work on prosodic attitudes based on forced choice paradigms in the past, with other sets of expressions (e.g. [26]).

The free-labeling paradigm allows an accurate and precise analysis of what subjects perceive from speakers’ performances. This is a broad description given in these four pages; much more in-depth analysis is possible. Such analysis would look further at the complete list of labels denoting a given cluster, or at the relationships between languages, by comparing the labels used for comparable clusters, or for culture-specific expression.

In this way, the cross-cultural recording that is used in this study allows to compare the production strategies of various speaker groups, for the same communication situation. If one sticks with the walking-on-eggs situation, one can see that both USA English speakers and French (who do not have it in their prototypical set of culturally-encoded expressions) mostly resort to hesitations (introducing schwa, pauses, etc.) in this situation. On the contrary, Japanese speakers mostly use irregular phonation (breath or creak). The French listeners did group WOEG with UNCE, while the Japanese did singularize WOEG as “I’m sorry” – but they also put this rather close to their uncertainty cluster. There is thus still room to explore cross-cultural similarities.

Our next work will focus on USA English perception of English performances – and then, cross-cultural tests shall be run to compare the perception of behaviors from other culture: what do French listeners perceive from the Japanese WOEG performances? What do Japanese think of the USA English irony and seduction?

5. Acknowledgements

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 the "Investments for the future" Programme IdEx Bordeaux - CPU (ANR-10-IDEX-03-02), the ANR PADE Grant and the Social affEcts Discrimination Using Combined acousTic and visual informatiON project (SEDUCTION); also, Japan Society for the Promotion of Science (JSPS), Grants-in-Aid for Scientific Research (A) #25240026. We thank Japanese listeners from Waseda University who participated in the perceptual experiment.

6. References

[1] Uldall, E., "Attitudinal meanings conveyed by intonation contours", Language and Speech, 3(4), 223-234, 1960.

[2] Martins-Baltar M., "De l'énoncé à l'énonciation: une approche des fonctions intonatives", Paris, Didier, 1977.

[3] Fujisaki, H., & Hirose, K., "Analysis and perception of intonation expressing paralinguistic information in spoken Japanese", in ESCA Workshop on Prosody, 1993.

[4] Kohler, K. J., "Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions", From Traditional Phonology to Modern Speech Processing, 205-214, 2004.

[5] de Moraes, J. A., "The pitch accents in Brazilian Portuguese: Analysis by synthesis", in 4th International Conference on Speech Prosody, Campinas, Brazil, 389–397, 2008.

[6] Gu, W., Zhang, T. & Fujisaki, H., "Prosodic Analysis and Perception of Mandarin Utterances Conveying Attitudes", in Proceedings of Interspeech, Firenze, Italy, 1069-1072, 2011.

[7] Delattre, P., "Comparing the prosodic features of English, German, Spanish and French", International Review of Applied Linguistics in Language Teaching, 1(1), 193-210, 1963.

[8] Fónagy, I., Bérard, E. and Fónagy, J., "Clichés mélodiques", Folia Linguistica 17: 153–185, 1984.

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

[10] Shochi, T., Rilliard, A., Aubergé, V. & Erickson, D., "Intercultural perception of English, French and Japanese social affective prosody", S. Hancil (ed.), The role of prosody in affective speech, Linguistic Insights 97, Bern: Peter Lang, AG, Bern, 31-59, 2009.

[11] A. Wierzbicka, "Different cultures, different languages, different speech acts: Polish vs. English", Journal of Pragmatics, vol. 9, no.2-3, 145–178, 1985.

[12] Bänziger, T., Mortillaro, M., & Scherer, K. R., "Introducing the Geneva multimodal expression corpus for experimental research on emotion perception", Emotion, 12(5), 1161-1179, 2012.

[13] A. Rilliard, D. Erickson, T. Shochi, and J. A. Moraes, "Social face to face communication - American English attitudinal prosody", in Proceedings of Interspeech, Lyon, France, 2013.

[14] Spencer-Oatey, H., "Reconsidering power and distance", Journal of pragmatics, 26(1), 1-24, 1996.

[15] Osgood, C. E., Suci, G. J., & Tannenbaum, P. H., "The measurement of meaning", Urbana: University of Illinois Press, 1957.

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

[17] Amir, N., Rubinstein, R., Shlomov, A., & Diamond, G., "Comparing categorical and dimensional ratings of emotional speech", in 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Vol. 5, 1-5, 2015.

[18] S. C. Widen and J. A. Russell, "A closer look at preschoolers' freely produced labels for facial expressions", Developmental Psychology, Vol.39, 114–128, 2003.

[19] Y. Greenberg, N. Shibuya, M. Tsuzaki, H. Kato, and Y. Sagisaka, "Analysis on paralinguistic control in perceptual impression space using multiple dimensional scaling", Speech Communication 51, 585–592, 2009.

[20] Husson, F., Le, S. and Pages, J., *Exploratory Multivariate Analysis by Example Using R*, Chapman and Hall, 2010.

[21] Brandt, P. A., "Thinking and language. A view from cognitive semio-linguistics", in 4th International Conference on Speech Prosody, Campinas, Brazil, 649–654, 2008.

[22] M. Guerry, T. Shochi, A. Rilliard, and D. Erickson, "Perception of prosodic social affects in french: a free-labeling study", in Proceedings of 18th of International Congress of Phonetic Sciences, Glasgow, Scotland, 2015.

[23] Bryant, G. A., "Prosodic contrasts in ironic speech", Discourse Processes, 47(7), 545-566, 2010.

[24] Osgood, C. E., May, W. H., & Miron, M. S. *Cross-cultural universals of affective meaning*, University of Illinois Press, 1975.

[25] Romney, A. K., & Moore, C. C., "Toward a theory of culture as shared cognitive structures", Ethos, 26(3), 314-337, 1998.

[26] A. Rilliard, T. Shochi, J.-C. Martin, D. Erickson, and V. Aubergé, "Multimodal indices to Japanese and French prosodically expressed social affects", Language and Speech, vol. 52, 223–243, 2009.