# Categorisation of spoken social affects in Japanese: human vs. machine

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

# CATEGORISATION OF SPOKEN SOCIAL AFFECTS IN JAPANESE: HUMAN VS. MACHINE

Jean-Luc Rouas[1], Takaaki Shochi[1,2], Marine Guerry[2] & Albert Rilliard[3,4]

[1]LaBRI CNRS UMR 5800 Univ. Bordeaux, France [2]CLLE-ERSSàB CNRS UMR 5263, Bordeaux, France [3]LIMSI CNRS Univ. Paris Saclay, Orsay, France [4]Universidade Federal do Rio de Janeiro, Brazil

{jean-luc.rouas, takaaki.shochi, marine.guerry}@labri.fr, albert.rilliard@limsi.fr

## ABSTRACT

In this paper, we investigate the abilities of both human listeners and computers to categorise social affects using only speech. The database used is composed of speech recorded by 19 native Japanese speakers. It is first evaluated perceptually to rank speakers according to their perceived performance. The four best speakers are then selected to be used in a categorisation experiment in nine social affects spread across four broad categories. An automatic classification experiment is then carried out using prosodic features and voice quality related features. The automatic classification system takes advantages of a feature selection algorithm and uses Linear Discriminant Analysis. The results show that the performance obtained by automatic classification using only eight features is comparable to the performance produced by our set of listeners: three out of four broad categories are quite well identified whereas the seduction affect is poorly recognised either by the listeners or the computer.

**Keywords:** Social attitudes; Speech perception; Expressive speech; Prosodic analysis.

## 1. OUTLINE

First, we briefly present the database used in this paper (section 2). Then, a first experiment is carried out which aims at identifying the best performing speakers in the database. After selecting the two best performing speakers for each gender, a perceptual test is carried out using those speakers to assess human performance on the categorisation of nine social affects (section 3). Next, we present the framework and the results for the automatic classification of social affect (section 4). Finally, we discuss the results and draw perspectives in section 5.

## 2. SOCIAL AFFECTS DATABASE

The database used for this experiment is based on the productions of 19 native Japanese speakers uttering the word “banana” at the end of short scripted dialogues ending with 16 different social affects driven by the context using the same paradigm as described in [9].

Most pragmatic situations correspond closely to their English label, with the exception of "walking-on-eggs". This expression denotes the feeling a Japanese speaker would call "Kyoshuku", a 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" ([10], p. 34.).

## 3. PERCEPTUAL EXPERIMENTS

### 3.1. Experimental design

First, an evaluation test was designed in order to evaluate how well each speaker expresses each social affect. This test is carried out with the same protocol as the one used in [9]. 26 subjects (12 females, 14 males, mean age 31), all Japanese native speakers from the Tokyo area listened to 304 stimuli (19 speakers performing 16 social affects) and evaluated each speakers performance. The two best speakers for each gender are thus selected for all further perceptual tests.

Then, in order to investigate the perceptual capacity of listeners to interpret the prosodic expressions of affect (audio alone), we conducted a perceptual test on Japanese native speakers based on a forced choice paradigm derived from the corpus used for the performance test. Accordingly to those results described in [5], we selected for the purpose of this study 9 contexts which can be further regrouped in 4 clusters: The first cluster is composed of Contempt (CONT), Irony (IRON), Irritation (IRRI) and Obviousness (OBVI) and corresponds to expressions of imposition. The second cluster is composed of Politeness (POLI), Seduction (SEDU) and Sincerity (SINC) which are polite and submissive expressions. The third category is composed only of Surprise (SURP) which is a potentially universal affect. The last category contains only "Walking on eggs" which is a dubitative expression. We decided to label these broad categories as respectively Imposition, Politeness, Surprise and Dubitative.

A total of 36 stimuli (9 expressions x 4 speakers) were presented in a random order. 29 listeners, native speakers of the Tokyo dialect (19 F/ 10 M) participated in this test. The subjects listened to each stimulus only once and were asked to answer the perceived affective expressions amongst 9 possible responses. The interface and instruction were displayed in Japanese.

### 3.2. Results

The results of this experiment are shown on Table 1. Although the overall correct identification rate is not very high (33.6%), it is much higher than the chance level (11.1%). The best recognised social affects are Surprise (94.8%) and Irritation (61.6%). The worst recognised affects are Contempt (9.8%), Sincerity (11.2%) and Seduction (13.8%).

**Table 1:** Perceptual categorisation (29 listeners)

| | CONT | IRON | IRRI | OBVI | POLI | SEDU | SINC | SURP | WOEG |
|---|---|---|---|---|---|---|---|---|---|
| CONT | 9.8 | 5.4 | 45.5 | 31.3 | 5.4 | 0.0 | 1.8 | 0.0 | 0.9 |
| IRON | 34.8 | 14.3 | 6.3 | 7.1 | 10.7 | 17.0 | 2.7 | 7.1 | 0.0 |
| IRRI | 1.8 | 2.7 | 61.6 | 15.2 | 12.5 | 0.9 | 2.7 | 2.7 | 0.0 |
| OBVI | 5.4 | 5.4 | 29.5 | 14.3 | 5.4 | 0.0 | 5.4 | 34.8 | 0.0 |
| POLI | 1.8 | 0.0 | 0.0 | 9.8 | 33.0 | 12.5 | 16.1 | 6.3 | 20.5 |
| SEDU | 3.4 | 4.3 | 2.6 | 19.8 | 36.2 | 13.8 | 12.9 | 2.6 | 4.3 |
| SINC | 3.4 | 2.6 | 0.9 | 12.9 | 42.2 | 10.3 | 11.2 | 3.4 | 12.9 |
| SURP | 4.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.9 | 0.0 | 94.8 | 0.0 |
| WOEG | 6.0 | 2.6 | 4.3 | 6.0 | 15.5 | 6.9 | 2.6 | 7.8 | 48.3 |

As can be seen on Table 1, most confusions seem to occur within the theoretical categories. For example, Contempt is confused with either Irritation or Obviousness, while most expressions of Seduction and Sincerity are identified as Politeness. Using the broad categories, the results are shown on Table 2.

**Table 2:** Perceptual categorisation (broad classes)

| | Imposition | Politeness | Surprise | Dubitative |
|---|---|---|---|---|
| Imposition | 72.5 | 16.1 | 11.2 | 0.2 |
| Politeness | 20.6 | 62.8 | 4.1 | 12.5 |
| Surprise | 4.3 | 0.9 | 94.8 | 0.0 |
| Dubitative | 19.0 | 25.0 | 7.8 | 48.3 |

With this clustering, the global correct identification rate is 70%. The most important confusion occurs between the Dubitative and Politeness expressions.

## 4. AUTOMATIC CLASSIFICATION OF SOCIAL AFFECTS

Although many researches are addressing the problem of "emotion" or speaker state automatic recognition (see [11] for example), none to our knowledge try to deal with social affective meaning in speech. There are however experiments aiming at measuring acoustic distances between social affects as for example in [7]. Furthermore, the results of our perceptual experiment lead us to believe that discrimination between social affects may be carried out to a certain extend using automatic classification. We use the same database but the tests are carried out using all the 19 speakers (i.e. not with only the 4 best performing ones as in section 3).

### 4.1. Experimental design

As not much data is available – we have a total of 16 minutes – we devised the experiment as a cross validation one. This means that throughout the experiment, to assure speaker independence of the models, we use all the data from all speakers except the one we test for training the models (i.e. 18 speakers are used for training while 1 speaker is used for the test). This procedure is repeated until all the speakers have been tested.

#### *4.1.1. Preprocessing*

Since phonetic transcriptions of all the excerpts have been done manually, we decided to use them as a basis for our analysis. All the parameters are then computed on the vocalic segments with the exception of duration, which is computed at the syllable level. As our target sentence is /banana/, we then have 3 points of measure, one for each /a/ while duration is computed for /ba/, /na(1)/ and /na(2)/.

#### *4.1.2. Features*

The features we decided to use are mainly coming from the matlab toolbox COVAREP [4] which we modified to add some features and to extract features on selected segments. Out of the 37 computed features, 31 are related to acoustic measurements: fundamental frequency (F0, F0SLOPE, FOVAR), the intensity (NRJ, NRJSLOPE, NRJVAR), duration (DUR), harmonics amplitude (H1, H2, H4), formants amplitude (A1, A2, A3), frequencies (F1, F2, F3, F4) and bandwidth (B1, B2, B3, B4), differences between harmonics amplitude (H1-H2, H2-H4), differences between amplitude of harmonics and formants (H1-A1, H1-A2, H1-A3), cepstral peak prominence (CPP), harmonics to noise ratios on different frequency bands (HNR05, HNR15, HNR25, HNR35). 5 features are glottal features that are computed using inverse filtering (IAIF method): normalised amplitude quotient (NAQ [2, 1]), quasi-open quotient (qOQ), difference between amplitude of harmonics 1 and 2 in the estimated glottal signal (H1H2aiff), parabolic spectral parameter (PSP [3]), PeakSlope [6], maximum dispersion quotient (MDQ [12]).

As these features are computed on each vowel, we thus have three measurements per social affect per speaker. The number of features per social affect per speaker is thus 111. Some of the features are normalised in order to remove the effect of gender whenever possible. A further normalisation is carried out using the "declarative" sentence, which is considered as reference. All values coming from the reference sentence are then subtracted from the values for each social affect.

Given the dynamic nature of speech, incorporating some kind of dynamic measure may help discriminating between the social affects. That is why we decided to compute the differences between the values on successive vowels, for each of the features described above. This results in having, for example for the F0 feature, instead of [$F0_1$, $F0_2$, $F0_3$] a vector containing [$F0_1, F0_2, F0_3, F0_2 - F0_1$, $F0_3 - F0_2$]. Integrating this information adds 74 more features to the original 111 features set, resulting in a total of 185 features.

#### 4.1.3. Features selection

Given this quite important number of features, a feature selection algorithm is used to keep only the most relevant ones. In this work, we decided to use the IRMFSP algorithm, described in [8]. It consists in maximising the relevance of the descriptors subset for the classification task while minimising the redundancy between the selected ones.

This algorithm has the advantage of not trying to combine the different features (as what would occur when using a PCA for instance) and of providing a ranking of the supposed discriminant power of the features, allowing to explore the compromise to be made between number of features and expected classification performance.

### 4.2. Experiments

The classification is carried out using cross validation (leaving one speaker out as described above) and a varying number of features (from 1 to 185 ranked using the IRMFSP algorithm). Thus, the exact process for each step of the cross validation is as follows, until all speakers have been used for testing:

- Select a test speaker (all the other speakers will be used for training).
- Carry out the feature ranking process (IRMFSP) on the training data only.
- For a fixed number of features, train a Linear Discriminant model
- Estimate the class of all the recordings made by the test speaker
- Evaluate the performance of the system for the speaker

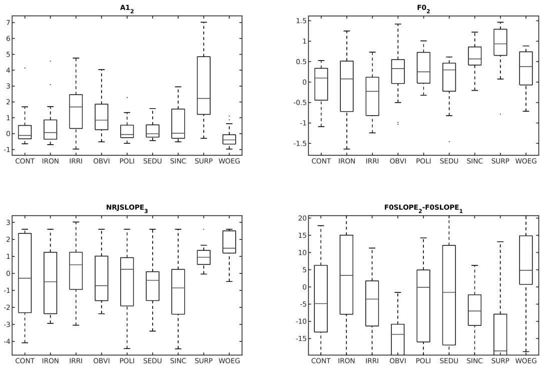
#### 4.2.1. Performance vs. number of features

The first experiment aims at finding the most relevant features. Given the framework described above, we can evaluate the performance of the system, using a varying number of features, starting with the most relevant one according to the IRMFSP algorithm.

The optimal number of features is found to be 8. Since the IRMFSP feature ranking is computed at each step of the cross validation process, we kept the ranking at each step and computed the median ranking for each feature across all speakers. This way, the eight features that have the best median ranking are: (1) $A1_2$ : amplitude of the first formant on the second vowel, (2) $F0_2$: mean value of the fundamental frequency on the second vowel, (3) $NRJSLOPE_3$: slope of the energy curve on the third vowel, (4) $F4_1$: frequency of the fourth formant on the first vowel, (5) $F0SLOPE_2 - F0SLOPE_1$: difference between the slope of the fundamental frequency on the second and the first vowel, (6) $NRJ_3 - NRJ_2$: difference on the mean value of intensity between the third and the second vowel, (7) $MDQ_1$: maxima dispersion quotient on the first syllable, (8) $F0VAR_3 - F0VAR_2$: difference in the variance of fundamental frequency between the third and second syllable.

Among those selected features, we can observe that the first two, i.e. the most discriminant ones, are measurements made on the middle vowel. On figure 1, it can be seen that the values of the first formant amplitude on the second vowel are higher for Surprise, Obviousness and Irritation, while the fundamental frequency on the second vowel is higher for positive expressions (SURP, POLI, SEDU, SINC) and OBVI and WOEG than for CONT, IRON and IRRI. The approximated slope of energy on the third and last vowel show that ending the sentence with rising energy happens for SURP and WOEG. The difference between the slopes of F0 on the first

**Figure 1:** boxplots of some relevant features (a) $A1_2$ (b) $F0_2$ (c) $NRJslope_3$ (d) $F0slope_3 - F0slope_2$



and second vowel aims at focusing on the contrasts between rising/falling or falling/rising patterns and continuing patterns. In that respect, it seems that the intonation patterns are continuous for most expressions except OBVI, SURP and WOEG.

### 4.2.2. *Results for the best feature set*

The results of the automatic classification experiment are given on Table 3. Overall, the classification achieves a performance of 38.6% of correct identifications. As for the perceptual test, while not being a great performance, this is much higher than chance.

While looking more closely at the results, we can observe that the most easily classified affect is Surprise (78.9%) followed by "Walking on eggs" (57.9%). Some other affects are mildly recognised, such as Irritation (52.6%), Obviousness (47.3%) and Sincerity (47.4%).

**Table 3:** Results of the automatic classification

| | CONT | IRON | IRRI | OBVI | POLI | SEDU | SINC | SURP | WOEG |
|---|---|---|---|---|---|---|---|---|---|
| CONT | 36.8 | 15.8 | 5.3 | 10.5 | 0.0 | 15.8 | 5.3 | 0.0 | 10.5 |
| IRON | 15.8 | 15.8 | 15.8 | 5.3 | 5.3 | 10.5 | 10.5 | 5.3 | 15.8 |
| IRRI | 5.3 | 10.5 | 52.6 | 10.5 | 0.0 | 10.5 | 5.3 | 5.3 | 0.0 |
| OBVI | 5.3 | 0.0 | 10.5 | 47.4 | 5.3 | 10.5 | 10.5 | 10.5 | 0.0 |
| POLI | 5.3 | 10.5 | 0.0 | 5.3 | 5.3 | 10.5 | 26.3 | 5.3 | 31.6 |
| SEDU | 0.0 | 31.6 | 15.8 | 10.5 | 10.5 | 5.3 | 15.8 | 0.0 | 10.5 |
| SINC | 5.3 | 0.0 | 5.3 | 10.5 | 5.3 | 5.3 | 47.4 | 5.3 | 15.8 |
| SURP | 0.0 | 0.0 | 0.0 | 5.3 | 0.0 | 0.0 | 15.8 | 78.9 | 0.0 |
| WOEG | 10.5 | 5.3 | 0.0 | 5.3 | 10.5 | 5.3 | 5.3 | 0.0 | 57.9 |

With the same 8 features, we reproduced the whole experiment using only the 4 theoretical classes, with the same cross-validation procedure. The results are displayed on Table 4. While achieving a overall identification rate of 60%, the system performs poorly for politeness and dubitative categories. Politeness is often confused with imposition while the dubitative expression is confused with both politeness and imposition expressions. The best recognised expressions are the expression of surprise and the expressions of imposition.

**Table 4:** Automatic classification in broad theoretical classes

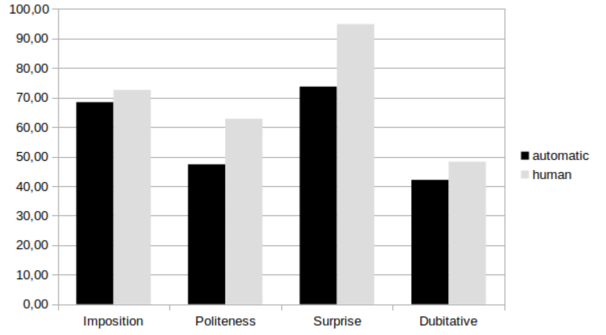
| | Imposition | Politeness | Surprise | Dubitative |
|---|---|---|---|---|
| Imposition | 68.4 | 19.7 | 3.9 | 7.9 |
| Politeness | 38.6 | 47.4 | 3.5 | 10.5 |
| Surprise | 5.3 | 21.1 | 73.7 | 0.0 |
| Dubitative | 26.3 | 31.6 | 0.0 | 42.1 |

## 5. DISCUSSION AND PERSPECTIVES

Unfortunately, we were only able to evaluate the four best speakers in the perceptual experiment. This is due to the fact that we need to keep the experiment simple to avoid cognitive overload and that we need to replicate the experiment with many listeners to assess their global behaviour. Concerning the automatic classification experiment, the design is of course different: we do not need many trials because the classification produces the same result each time, but we need to have as many speakers as possible to assess the generalisation of the approach. In that study, we can therefore consider the machine as a particular listener which is used to evaluate all the speakers.

Considering only the broad classes of affects, when looking at the confusion matrix for the perceptual test (Table 2) and the for automatic classification (Table 4), we can observe a rather similar behaviour: Seduction is in both case poorly identified, while the other classes of affect are mostly correctly classified. As a graphic way of comparison, Figure 2 presents the performance obtained separately for each social affect for the automatic classification system and for a categorisation perceptual experiment.

**Figure 2:** % correct for broad classes of social affects by human and machine



These results show the similar behaviour between human perception and automatic classification of the broad class of social affects. We will need to confirm these results using more data, particularly by testing complete utterances rather than a single word. In the future, we will also reproduce the same experiment using different languages such as French and English.

## 6. REFERENCES

[1] Airas, M., Alku, P. 2007. Comparison of multiple voice source parameters in different phonation types. *Proceedings of Interspeech 2007*. Citeseer.

[2] Alku, P., Bäckström, T., Vilkman, E. 2002. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America* 112(2), 701–710.

[3] Alku, P., Strik, H., Vilkman, E. 1997. Parabolic spectral parameter - a new method for quantification of the glottal flow. *Speech Communication* 22(1), 67–79.

[4] Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S. 2014. Covarep - a collaborative voice analysis repository for speech technologies. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE 960–964.

[5] Guerry, M., Rilliard, A., Erickson, D., Shochi, T. 2016. Perception of prosodic social affects in japanese: a free-labeling study. *SPEECH PROSODY, 8th* 811–815.

[6] Kane, J., Gobl, C. 2011. Identifying regions of non-modal phonation using features of the wavelet transform. *INTERSPEECH* 177–180.

[7] Mixdorff, H., Hï¿½nemann, A., Rilliard, A., Lee, T., Ma, M. K. 2017. Audio-visual expressions of attitude: How many different attitudes can perceivers decode? *Speech Communication* 95, 114 – 126.

[8] Peeters, G. 2003. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. *Audio Engineering Society Convention 115*. Audio Engineering Society.

[9] Rilliard, A., Erickson, D., Shochi, T., De Moraes, J. A. 2013. Social face to face communication - american english attitudinal prosody. *Proc. of Interspeech 2013*.

[10] Sadanobu, T. 2004. A natural history of japanese pressed voice. *Journal of the Phonetic Society of Japan* 8(1), 29–44.

[11] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM 3–10.

[12] Yanushevskaya, I., Ní Chasaide, A., Gobl, C. 2015. The relationship between voice source parameters and the maxima dispsersion quotient (mdq). *Interspeech 2015*.