PERCEPTUAL EVALUATION OF SPOKEN JAPANESE ATTITUDES

Takaaki Shochi[1,2], Dominique Fourer[1], Jean-Luc Rouas[1], Marine Guerry[2] and Albert Rilliard[3]

[1]LaBRI - CNRS UMR5800, Bordeaux, France
[2]CLLE-ERSSaB UMR5263, Bordeaux, France
[3]LIMSI - CNRS UPR3251, Orsay,France

[1]firstname.lastname@labri.fr
[2]firstname.lastname@u-bordeaux-montaigne.fr
[3]rilliard@limsi.fr

ABSTRACT

The aim of the present work is to investigate whether Japanese listeners can recognize before the end of an utterance various social affective meanings expressed by a speaker. 28 native Japanese subjects participated in the experiment. For this experiment, one sentence consisting of 3 moras is uttered in 8 different social affects. Each affective expression is produced by 4 native Japanese speakers (2 male, 2 female) who are selected as the best performers by native listeners. These 16 utterances are segmented in three sections (gates) – that correspond to the three moras, adding white noise after a given gate position. The results show a significant effect of gate on recognition scores.

Keywords: social affects, spoken Japanese attitude, affective prosody

1. INTRODUCTION

Recent work on affective prosody describes various types of information including emotional states, mood, attitudes or personality [15]. Among different levels of affective functions of prosody, social affects encoded by cultural factors play an important role in face-to-face interaction [4, 14]. These social affective expressions are supposed to be learnt from childhood in a social environment. [2, 16].

Since the prosody conveying basic emotions is related directly to physiological changes [7], prosodic features of such emotions are distributed throughout the utterance, and are supposed to be perceived immediately by listeners. On the contrary, attitudinal expressions contain culturally conventional prosodic forms, located on a certain temporal part of the utterance [13].

Therefore, the aim of the present work is to investigate how linguistic information is distributed over conventional prosodic forms of utterances, and to identify the part conveying such prototypical prosodic contours which potentially carry more affective information than other parts so that listeners perceive affective information as well as linguistic content by these prosodic indices. Previous research revealed that listeners can predict the entire affective meaning of an utterance from just a part of prosodic information [9, 5, 6]. To investigate such "anticipation" of listeners and especially to identify when they will be able to recognise the intended attitude of the speaker, current work follows the paradigm of gating (i.e. the gradual unveiling) of prosodic contours of a sentence [9].

The gating paradigm is to gradually expose listeners to a speech stimulus from its beginning until its completion. This stimulus may consist of syllables or moras (in the case of Japanese), a word, a phrase or even a complete sentence according to the purpose of the investigation. The division of the stimulus can be such that a portion of the signal is cut at a regular time interval [9, 5, 6], or the end of each syllable [18, 13, 10]. For instance, [3, 13, 10] conducted a perception test using the gating paradigm with six French affective expressions exposing each syllable gradually. The results show a progressive identification (continuous growth rate recognition) for most of the attitudes.

This paper briefly summarizes in Section 2 the corpus of social affective expressions recorded by Japanese native speakers, then describes the methodological protocol for the perceptual experiment based on a gating paradigm. Finally, the results of this perceptual test are analysed in Section 3.

2. EXPERIMENTS

2.1. Corpus

The corpus consisted of recordings of 19 Japanese native speakers (11 females, 8 males) speaking "banana" with 16 different social affective expressions [8].

2.2. Gating experiment

28 Japanese native speakers were recruited among students of Waseda University and Sofia University in Tokyo. The aim of the experiment is to assess if the listeners can recognise, before the end of the utterance, various social affective meanings. According to the gate paradigm described in the introduction, a "gate" position occured for each mora of "banana". All the recordings from our corpus would result in a total of 19 speakers x 16 attitudes x 3 gate positions = 912 stimuli, and would make the perceptual experiment too long and too tiring.

We thus selected the best performing speakers for each attitude (2 males and 2 females) based on an earlier study [8]. This has been done by performing a perceptual categorisation test with 22 listeners (different from the ones used in this experiment). We also reduced the number of attitudes from 16 to 8.

The stimuli, are denoted Gate 1 ("ba"), Gate 2 ("bana") and Gate 3 ("banana"). In order to verify that all stimuli have the same duration, the stimuli were filled with a white noise from the gate position until the target duration of 2 seconds.

This resulted in a total of 96 stimuli (8 attitudes, 4 speakers, 3 gate positions). Subjects were first presented the labels of 8 attitudes with the corresponding situations. During the test, each listener had to listen, and choose the correct label among 8 attitudes (Obviousness, Arrogance, Irritation, Politeness, Admiration, Walking on Eggs, Surprise and Doubt) on the interface created under Livecode software using Bose 5C7NT headphones.

3. RESULTS

3.1. ANOVA

Results collected from the experiment are expressed either as binary recognition scores of the presented attitudes, or as confusion matrices over the eight possible answers, for each presented attitude. Binary recognition scores have been quantified for each presented attitude, and Univariate ANOVA was computed with 3 different fixed factors: Attitude (8 levels), the speaker's Gender (2 levels) and the presented Gate position (3 levels) for 28 listeners' perceptual behaviour. According to the results, the main effect of all three factors (Gate position, Attitude and speaker's gender) were significant ($p.<0.05$). This first result indicates that listeners perceived differently presented stimuli according to the gate position, the type of affective expression and the speaker's gender. Moreover, a significant interaction of 2 factors (Attitude x Gate position) was also observed. However the interaction between Gate position and the speakers' gender was not significant ($p.=0.285$).

3.2. Gate Recognition

According to the results on recognition rate in 3 different gate positions (see Table 1), 3 different types of perceptual behaviours are observed. The first category consists of 3 attitudinal expressions : Doubt (DOUB), Irritation (IRRI), and Admiration (ADMI) showing a continuous growth rate recognition. For instance, the recognition rate for DOUB on gate1 was 52%, 60% in Gate 2, then increases up to 92% when listeners heard the entire utterance. For IRRI, the recognition rate is 39% for Gate 1, 51% for Gate 2, and 73% for Gate 3. Finally, ADMI shows 31% in Gate 1, 54% in Gate 2, then 76% in Gate 3 position.

On the contrary, the attitude of Surprise (SURP) was recognized immediately from the first mora. In fact, a quite high identification rate of this attitude was observed already in Gate 1 position (73%), then this score was stable in Gate 2 (85%) and Gate 3 position (82%).

The third category is composed of the attitudes of Arrogance (ARRO), Obviousness (OBVI), Walking on eggs (WOEG) and Politeness (POLI) indicating a flat shape of recognition indicating some movement in the narrow range of identification rate. For instance, POLI was relatively well recognised even in Gate 1 position (58%), however, this recognition rate was almost the same in Gate 2 (60%) and Gate 3 position (63%). Other attitudes of this category showed a weak identification rate from Gate 1 to Gate 3 position (ex. the range from 22% in Gate 1 to 27% in Gate 3 for OBVI, from 12% in Gate 1 to 33% in Gate 3 for ARRO and from 22% in Gate 1 to 32% in Gate 3 for WOEG). Even if these 3 attitudes were difficult to be perceived as the intended attitude, it does not mean that they are considered unknown attitudes for native listeners.

In order to understand how listeners mix these attitudes with others, the confusion matrix is analysed in the next section.

3.3. Confusion between attitudinal labels

The confusion matrix (Table 1) presents the 8 possible answers (the conceptual attitudes recognised by listeners) as rows, and the presented attitude in each gate (thus 3x8 stimuli, mixing female and male speakers) as columns; it contains the number of times each stimulus was categorised by listeners under a given label. This matrix is analysed thanks

	OBVI	ARRO	IRRI	POLI	ADMI	WOEG	SURP	DOUB
OBVI (1)	**19.64**	8.04	15.18	20.54	8.04	0.00	19.64	8.93
OBVI (2)	**20.54**	8.04	20.54	18.75	7.14	1.79	12.50	10.71
OBVI (3)	**24.11**	14.29	35.71	3.57	0.89	0.89	10.71	9.82
ARRO (1)	33.93	**10.71**	12.50	28.57	5.36	1.79	3.57	3.57
ARRO (2)	26.79	**27.68**	25.89	10.71	0.00	5.36	0.89	2.68
ARRO (3)	38.39	**29.46**	20.54	7.14	0.00	1.79	1.79	0.89
IRRI (1)	11.61	10.71	**34.82**	16.96	6.25	0.89	11.61	7.14
IRRI (2)	16.07	8.93	**45.54**	18.75	4.46	0.89	4.46	0.89
IRRI (3)	13.39	13.39	**65.18**	7.14	0.89	0.00	0.00	0.00
POLI (1)	21.43	1.79	0.89	**51.79**	5.36	12.50	0.89	5.36
POLI (2)	18.75	0.89	3.57	**53.57**	5.36	15.18	0.89	1.79
POLI (3)	20.54	1.79	0.89	**56.25**	4.46	14.29	0.89	0.89
ADMI (1)	1.79	0.89	4.46	3.57	**27.68**	2.68	50.00	8.93
ADMI (2)	0.89	1.79	1.79	0.89	**48.21**	3.57	39.29	3.57
ADMI (3)	1.79	0.00	0.00	1.79	**67.86**	0.00	28.57	0.00
WOEG (1)	32.14	3.57	8.93	17.86	12.50	**19.64**	1.79	3.57
WOEG (2)	22.32	3.57	1.79	30.36	8.04	**28.57**	3.57	1.79
WOEG (3)	23.21	1.79	0.00	25.89	16.07	**28.57**	2.68	1.79
SURP (1)	2.68	0.00	3.57	3.57	8.93	0.89	**65.18**	15.18
SURP (2)	1.79	0.00	2.68	0.00	7.14	0.00	**75.89**	12.50
SURP (3)	0.00	0.00	1.79	0.00	2.68	0.00	**73.21**	22.32
DOUB (1)	6.25	3.57	7.14	8.04	6.25	4.46	17.86	**46.43**
DOUB (2)	3.57	6.25	8.04	2.68	1.79	0.89	23.21	**53.57**
DOUB (3)	0.89	1.79	4.46	0.00	0.89	0.00	9.82	**82.14**

Table 1: Confusion matrix for 8 attitudes and 3 gate positions.

	F1	F2	F3	Ct1	Ct2	Ct3	cos1	cos2	cos3
ADMI	-0.558	0.995	-0.412	5633	28125	6162	147	466	80
ARRO	0.630	-0.569	-0.501	4711	6041	5992	274	224	173
DOUB	-0.840	-0.916	0.844	15812	29515	32058	299	355	302
IRRI	0.521	-0.704	-0.745	6369	18246	26146	180	328	367
OBVI	0.672	-0.005	-0.001	11615	1	0	768	0	0
POLI	0.711	0.400	0.486	14353	7120	13482	461	146	215
SURP	-1.062	0.206	-0.285	37220	2195	5392	744	28	54
WOEG	0.630	0.719	0.704	4287	8756	10768	202	262	252

Table 2: Output of the CA for the 8 columns of the confusion matrix, presenting the factor scores (F), the contributions (ct) and the squared cosines (cos) for the first 3 dimensions. Contributions and squared cosines are multiplied by 1000 and rounded for convenience.

to a Correspondence Analysis (CA) [1], using R's FactoMineR library [11]. The first three eigenvalues explain more than 80% of the total variance. Table 2 describes the loadings, contribution and squared cosines of the columns. The first dimension of the analysis mostly opposes expressions perceived as OBVI (and POLI to a lesser degree) vs. expressions perceived as SURP. The second dimension opposes expressions perceived as ADMI vs. DOUB and IRRI. The third dimension mostly opposes IRRI to DOUB. To understand the main perceptual mixing between the presented stimuli (here the 24 rows of the matrix), a hierarchical classification of the rows is made, according to their distance calculated on the basis of their position on the dimensions of the CA (cf. [12]). The first five clusters obtained from this classification procedure are depicted in Figure 1. They consisted of: (i) the three gates of ADMI, (ii) the three gates of SURP, (iii) the three gates of DOUB, (iv) all the gates of OBVI, ARRO and IRRI, and (v) all the gates of POLI and WOEG.

The first three clusters are thus composed of homogeneous expressions, showing that attitudes are already perceived well since the first gate for these expressions; the progression observed through the subsequent gates mostly allows a more characteristic distinction of the type of expression, that are better differentiated from the other clusters – the most specific of each of these three clusters being the attitudes at the third gate (a row in a given cluster is more specific to this cluster when it is farther to the gravity centers of the other clusters).

For the last two clusters, they are composed of several attitudes that show more confusion. Cluster (iv) groups ARRO, IRRI and OBVI; interestingly, the first two attitudes are separated, while OBVI is

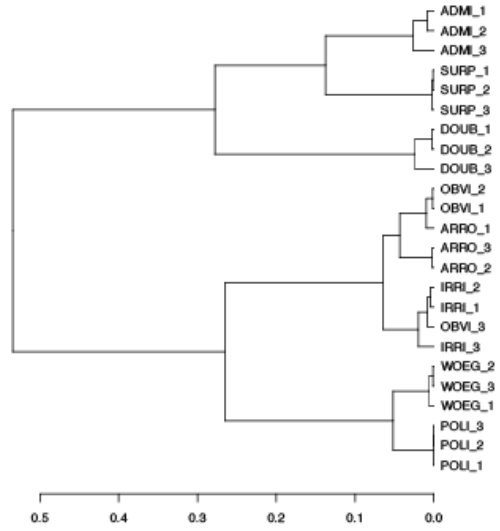

Figure 1: Tree representing the hierarchical classification of CA's rows.

#	Labels	Intern. %	glob. %	p. value	v. test
i	ADMI	46.6	10.4	$<10^{-4}$	19.3
	SURP	40.2	19.0	$<10^{-4}$	9.9
ii	SURP	70.7	19.0	$<10^{-4}$	23.1
	DOUB	17.2	12.9	$<10^{-4}$	2.4
iii	DOUB	61.2	12.9	$<10^{-4}$	23.8
iv	IRRI	30.5	13.5	$<10^{-4}$	20.2
	ARRO	15.0	6.8	$<10^{-4}$	13.2
	OBVI	22.3	14.8	$<10^{-4}$	8.5
	POLI	39.8	16.3	$<10^{-4}$	18.0
	WOEG	19.8	6.2	$<10^{-4}$	15.7
	OBVI	22.6	14.8	$<10^{-4}$	6.4

Table 3: Labels significantly used more often than their average frequency to define each of the five clusters defined by the hierarchical classification of CA's rows.

regrouped with ARRO at the first and second gate, and with IRRI at the third gate. The most specific of this cluster are the third gate stimuli of IRRI, ARRO and OBVI (in decreasing order). Cluster (v) mixes the two expressions expressing different forms of Japanese politeness – even if their acoustic forms are clearly separated (cf. [17]). The most specific in this cluster are the third gate stimuli of both POLI and WOEG.

Table 3 shows which labels are preferentially used by the listeners to describe each of these clusters. Interestingly, even if the two first clusters are composed of homogeneous stimuli, they are described by several labels. Cluster (i), composed of prosodic expressions of ADMI, is described as both ADMI and SURP in close scores. Cluster (ii), composed of prosodic expressions of SURP, is described primarily as SURP, but also as DOUB. The most homogeneous cluster is cluster (iii), composed and labelled as DOUB. These confusions show the cognitive similarities existing between these three prosodic expressions. Cluster (iv), composed of prosodic expressions of IRRI, ARRO and OBVI, is mostly labelled as IRRI, but also as ARRO and OBVI. Finally, cluster (v), composed of expressions of POLI and WOEG, is labelled first as POLI and then WOEG – but it is also labelled as OBVI. This last confusion, as well as the misclassification of the stimuli of OBVI, show the fuzziness of the prosodic expressions and of the concept of OBVI.

4. CONCLUSION AND PERSPECTIVE

The purpose of this work was to investigate whether Japanese listeners can recognise before the end of utterances different social affective expressions produced by native speakers. A gating paradigm was used to that aim. The results showed a significant effect of gate of the presented attitude and of the speaker's gender on recognition scores. This indicates that listeners change their perception of various social affects with regard to the temporal evolution of the prosodic signal. According to the observation of listeners' perception for 8 attitudinal expressions on 3 different gate positions, 3 types of perceptual categories were observed. The first category shows a clear positive correlation between the evolution of gate position and the increase of global recognition score, indicating that important acoustic cues associated with the perception of the intended attitudes were distributed globally from the beginning to the end of utterances. The immediate recognition of the attitude of SURP from the first gate was probably due to the particular voice quality (i.e. breathy, but tense voice) and an important pitch amplitude of this attitude compared to others. The second cluster, composed of dominant attitudes (ARRO and OBVI), has a low recognition rate in all 3 gate positions, due to the mutual confusions between these attitudes, and with IRRI. Similarly, a cluster composed of politeness expressions (WOEG and POLI), shows little progression of the recognition rate, and confusion within the expressions of politeness. In future work, we will analyse correlation between perceptual results and acoustic features in order to determine the most important acoustic cues for the identification of various social affects.

5. ACKNOWLEDGEMENT

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) and ANR

PADE. The authors warmly thank M. Kondo and S. Detey from Waseda University, J. Moore from Sophia University and D. Erickson from Kanazawa Medical University for their help. Thank also all the speakers and listeners from Waseda University and Sophia University, Japan for their participation.

6. REFERENCES

[1] Abdi, H., Béra, M. 2014. Correspondence analysis. In: *Encyclopedia of Social Networks and Mining*. Springer Verlag 275–284.

[2] Aubergé, V. 2002. A gestalt morphology of prosody directed by functions: the example of a step by step model developed at icp. *Speech Prosody 2002, International Conference*.

[3] Aubergé, V., Grepillat, T., Rilliard, A. 1997. Can we perceive attitudes before the end of sentences? the gating paradigm for prosodic contours. *Fifth European Conference on Speech Communication and Technology*.

[4] Campbell, N. 2004. Perception of affect in speech-towards an automatic processing of paralinguistic information in spoken conversation. *INTERSPEECH*.

[5] Cotton, S., Grosjean, F. 1984. The gating paradigm: A comparison of successive and individual presentation formats. *Perception & Psychophysics* 35(1), 41–48.

[6] Dupoux, E. 1989. *Identification des mots parlés: détection de phonèmes et unité prélexicale*. PhD thesis Paris EHESS.

[7] Fónagy, I. 1983. *La vive voix: essais de psychophonétique*. Langages et sociétés. Payot.

[8] Fourer, D., Shochi, T., Rouas, J.-L., Aucouturier, J.-J., Guerry, M. 2014. Going ba-na-nas: Prosodic analysis of spoken japanese attitudes. *Speech Prosody 2014* 4.

[9] Grosjean, F. 1985. The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics* 38(4), 299–310.

[10] Grépillat, T. 1996. Perçoit-on, par l'intonation, l'attitude d'un locuteur avant la fin de l'énoncé ? Master's thesis Université Stendhal Grenoble III.

[11] Husson, F., Josse, J., Le, S., Mazet, J. 2010. Factominer: Multivariate exploratory data analysis and data mining with r. version 1.14.

[12] Husson, F., Josse, J., Pages, J. 2010. Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data. *Applied Mathematics Department*.

[13] Morlec, Y., Bailly, G., Aubergé, V. 1999. Training an application-dependent prosodic model corpus, model and evaluation. *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*.

[14] Pavlenko, A. 2007. *Emotions and multilingualism*. Cambridge University Press.

[15] Pierrehumbert, J., Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. *Intentions in communication*.

[16] Scherer, K. R., Banse, R., Wallbott, H. G. 2001. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural psychology* 32(1), 76–92.

[17] Shochi, T., Rilliard, A., Aubergé, V., Erickson, D. 2009. *The role of prosody in affective speech* volume Linguistic Insights 97 chapter Intercultural perception of English, French and Japanese social affective prosody, 31–59. Peter Lang.

[18] Thorsen, N. G. 1980. A study of the perception of sentence intonation – evidence from danish. *The Journal of the Acoustical Society of America* 67(3), 1014–1030.