Perception of prosodic transformation for Japanese social affects

Dominique Fourer[1], *Takaaki Shochi*[2,3], *Jean-Luc Rouas*[2] *and Albert Rilliard*[4]

[1] IREENA EA4642, Saint-Nazaire, France
[2] LaBRI - CNRS UMR5800, Bordeaux, France
[3] CLLE-ERSSaB UMR5263, Bordeaux, France
[4] LIMSI - CNRS UPR3251, Orsay,France

Abstract

This paper is about the perception of ‘genuine’ social affects versus ‘synthetic’ ones. Our ultimate aim is to create a software for self-teaching language learning that includes a tool where learners will be able to hear their own voice producing the social affect correctly. Towards this goal, we study here how we can construct synthetic stimuli using neutral voices and prosodic parameters, and if such stimuli can be well enough recognized by native listeners. At first, we explain how our corpus is built around contextual scenarios and the recording protocol. Then, we explain how the synthetic stimuli are constructed. These stimuli must comply with several constraints: keeping the original speaker identity, preserving the linguistic content, and of course having the best possible quality. Results from a perception experiment with native speakers of Japanese show that the social affects for natural stimuli are quite well recognized although the results show more variation on the synthetic stimuli, depending on the considered social affect. Some social affects may indeed be expressed quite subtly so that they are difficult to synthesize. An investigation based on statistical analysis is proposed showing where the main difficulties lie.

Index Terms: speech processing, affective prosody, attitudes characterization

1. Introduction

The prosodic expression of social affects plays an important role in face-to-face interaction [1, 2]. As affective prosody can be related to the proficiency in the spoken language, cultural factors or the gender, it conveys various types of information about the emotional state, the mood, the attitude and the personality of a speaker [3].

Nowadays, many efforts are made to characterize the attitudes from the prosodic expression of social affects [4, 5]. This research is motivated by a better understanding of how the linguistic information is coded and distributed in an utterance that could allow improvement in the field of speech recognition and affective speech synthesis.

Our aim is to build a self-teaching language learning software that will allow learners to hear their own voice modified to produce the social affect correctly. This software will be addressed to advanced learners who have a good knowledge of the language but do not have the full gist of the socio-cultural ways to express attitudes, which is often the case if they have not been immersed for a long time with native speakers. The capacity of learners to deal with social affects is addressed in [6].

Towards this aim, we have already recorded native Japanese speakers - the recording procedure is described in section 2. The best performing speakers have been selected according to a listening experiment. These are the speakers we wish to use as ‘targets’ for the prosodic transformation, described in section 3.

The perception evaluation of both kinds of stimuli is described in Section 4 and finally discussed in Section 5.

2. Materials

2.1. Social contexts for expression of attitudes

In order to immerse subjects in the context, a specific scenario was defined for each attitude, and each subject was requested to engage in a short dialogue that would lead to the production of target sentences with the native speaker. For the current experiment, 8 contexts have been selected, corresponding to a set of attitudes used in [7, 8] for different languages.

Some of these contexts do not have lexical equivalents in all languages, as the corresponding communication situations have not been conventionalized in that particular culture. It is the case for example of the Japanese notion of *kyoshuku*, described by [9] as “corresponding to a mixture of suffering ashamedness and embarrassment, which comes from the speaker consciousness of the fact his/her utterance of request imposes a burden to the hearer”. For instance, *Kyoshuku* has no lexical equivalent in English. Meanwhile, “walking on eggs” corresponds to a certain extent to this concept.

The following 8 social affects were used in the present corpus: Admiration (ADMI), Arrogance (ARRO), Doubt (DOUB), Irritation (IRRI), Obviousness (OBVI), Politeness (POLI), Surprise (SURP) and Walking on eggs (WOEG). They are defined by prototypical situations with the social relationship of the two interlocutors specified , as well as the communicative goal of the speaker (see [10] for details). In this paper, the neutral declarative attitude is also used as reference for generating synthetic stimuli, however this expression is not investigated by the perception evaluation.

For all situations, a short neutral target sentence has been used to record the respective prosodic expressions: “banana”. In order to elicit these target sentences in each context, small dialogues were written [11], taking place in the prototypical context described above, and ending with the target sentence.

During the recordings, each speaker (A) has an active interlocutor (B) who interacts with her/him in order to enhance the naturalness of the communication situation, and to ease the production of realistic expressions. Speakers are indeed not asked to produce an isolated sentence with an identified attitude (e.g. seduction or authority), but rather to immerse in a scenario. For instance, the situation is the following for “walking on eggs”:

- Your boss (Speaker B) has asked you (speaker A) to be

> in charge of setting up a room for a big conference. Your boss is a super compulsive guy who needs to have everything done just right, and gets easily angered if things are not perfect. Your boss walks into the room where the big conference is to be held, and in the wastebasket, there is a half-eaten banana. He is furious.

Currently, these situations have been adapted to three languages: American English, Japanese and French. The present paper focuses on the Japanese results, as performed by native speakers.

2.2. Recording procedure

The present corpus is composed of recordings of 19 Japanese native speakers (11 females, 8 males). Most speakers were recruited amongst university students and were paid for their performance. The recordings took place in a sound-treated room at Waseda University, Japan.

The sound was captured by an *Earthworks QTC1* omnidirectional microphone, placed at one meter from the mouth of the speaker (this distance was chosen to limit the influence of the speaker movements on the sound level). The microphone level was calibrated before each recording session using a *Bruel & Kjaer* acoustical calibrator, thus the sound pressure level can be corrected after recording to a level comparable across all speakers.

The target sentence “banana” was then manually searched for across the recorder corpus, isolated and extracted into individual files. Any speech utterances from speaker B occurring during the expressive gesture of speaker A performing the target sentence were removed from the sound track (none overlapped with their speech). Due to the interactive nature of the recording, some spontaneous changes were observed on the target sentences: typically “banana” sentence with interjections, such as “hmm”, “er”, “oh”, etc., together with the target sentences.

Each speaker recorded one utterance of the word for each of the 8+1 attitudes (with the neutral declaration), resulting in a total of 171 stimuli. These were stored as 16 kHz, 16-bit WAV files. Each stimulus was trimmed to discard the beginning and the ending silence. The wave file of each stimulus was hand-labeled at a phonetic level using the PRAAT software [12].

3. Synthetic stimuli

The synthetic stimuli are obtained by using natural neutral stimuli as source signals which are transformed using the learned characteristics from the target social affect. In order to characterize each attitude, we decided to use the acoustic parameters profile extracted from the best performed stimuli of each gender (with the highest recognition rate during a perceptual listening test [5]). The acoustic parameters extracted from the source signal, chosen as being the neutral declaration attitude [4], are transformed according to the model profile associated to the target attitude. We take care of using as a model the best performed attitude that corresponds to a speaker with the same gender than the source stimulus. In order to obtain realistic stimuli, synthesis constraints have been defined during the transformation process.

3.1. Constraints

As presented in a previous study [4], we assume that each attitude can be described by a set of 3 time-related acoustic parameters: the fundamental frequency (denoted F_0), the amplitude and the duration of each mora.

The constraints on the synthesis of stimuli follow.

- To preserve the speaker identity of each stimulus after transformation, no vocal tract manipulation is allowed. This is the reason why except for the duration parameter, the F_0 and the amplitude profile are normalized to obtain a modulation factor centered around unity.
- To preserve the linguistic content, the formants of the synthetic stimuli have to be almost identical with the source natural ones.
- To obtain realistic synthetic stimuli, the sound quality should be good enough to not be perceived as synthetic. Thus, the pitch-shifting and the time-stretching transformations have to be applied only on the vowels in the signal.

3.2. Parameters estimation

In order to obtain accurate boundaries, each sound excerpt has been segmented and hand-labeled in phonemes using PRAAT [12]. Then the duration of each mora (1 mora per syllable) of the utterance “ba-na-na” is measured in seconds.

The fundamental frequency F_0 measured in Hertz of each stimulus is estimated using the SWIPE algorithm [13] that has been shown to be more robust by its author when it is applied on speech signals compared to other state of the art methods. Each resulting F_0 time series are computed with 10 ms sampling interval.

The amplitude parameter expressed in dB is estimated using the Root Mean Square (RMS) function where the signal is windowed to result in 20 ms frames with 50% overlap. Thus, the RMS amplitude is computed on each frame from the normalized values (in $[-1, +1]$) of the signal samples.

3.3. Synthesis

The parameters transformation results from a combination of pitch-shifting with formant preservation, amplitude modulation and time-stretching, that are successively applied on the source stimulus.

For the fundamental frequency and the amplitude that are represented by time series, the speaker identity constraint is followed by applying a normalization process to the source and the target stimuli that is obtained by dividing the value of each series by its average. Thus, the resulting parameter denoted $P_{source \to target}$ that corresponds to the transformation of the source attitude parameter P_{source} to the target attitude parameter denoted $P_{target}[n]$, is simply computed as

$$P_{source \to target}[n] = \bar{P}_{source} \cdot \frac{P_{target}[n]}{\bar{P}_{target}} \quad (1)$$

with $n = 1, 2, ..., N$ the sample index and $\bar{P} = \frac{1}{N} \sum_{n=1}^{N} P[n]$ the average of series P.

The syllable duration parameter of the transformed stimulus is defined as being equal to the duration of the target stimulus. This is obtained by a time-stretching operation of ratio $\frac{N_{target}}{N_{source}}$ that is separately applied on each syllable (N_{target} and N_{source} respectively denote the length of the target attitude and the length of the source).

In our implementation, the time-stretching operation is applied after the F_0 and the amplitude parameters transformation using PRAAT. Non formal listening tests show that the resulting synthetic signals obtained with PRAAT provide a sufficiently good quality while verifying all the constraints during the transformation process.

4. Perception evaluation of natural and synthetic stimuli

While the natural stimuli are simply obtained using the recording procedure described in Section 2, the synthetic ones are the result of a transformation applied on the acoustic parameters estimated from stimuli signals as described in Section 3. Both stimuli are used for a recognition test where each listener has to choose a unique answer among a closed choice of 8 attitudes.

4.1. Listeners

The listeners are 21 Japanese native speakers whom have been specifically selected because all of them speak the Tokyo dialect. The mean age of these subjects is 32 years old and most of them are students or staff members from Waseda and Sofia Universities.

4.2. Experimental protocol

A total of 96 stimuli, produced by 6 speakers eliciting the 8 attitudes of each kind (natural and synthetic), have been presented in audio alone condition using a high-quality headphone into a graphical application that has been specifically developed using the livecode software [14].

The stimuli which were presented in a random order were only played one time by each listener before validating its answer. The natural stimuli were merged with synthetic ones without a clue for the listener. The task consisted in selecting the attitude which reflects the most the played stimulus (according to the listener), among a forced choice of 8 attitudes under the graphical interface. At the end of the experiment, subjects were invited to give a feedback about their impressions on the experiment.

4.3. Results

In order to evaluate the listeners' perceptual behavior for both natural and synthetic stimuli, we computed the Kolmogorov-Smirnov and the Pearson's chi-squared tests [15, 16].

These tools can be used to measure the statistical independence between two distributions. Thus, it was used to provide an indication of the "goodness of fit" of the subjects' answers with the presented stimuli, for both kinds of stimuli.

Results reveal a statistically significant association between the presented affect and the subject's answer for both natural and synthetic stimuli ($p < 0.05$).

4.4. Identification rate for both natural and synthetic stimuli

Tables 1 and 2 show the recognition rates of 8 social affective expressions on natural and synthetic sound. According to this table, the recognition rate for synthetic sounds is generally lower than the one which was obtained using naturally expressed prosody.

Indeed, the natural stimuli are globally well recognized by native subjects except OBVI. DOUB as well as IRRI are well discriminated from the other expressions (88% for DOUB, 71% for IRRI). The other social affective expressions show some confusion: ADMI is quite well perceived (65%) although showing some confusion with SURP. SURP which is well discriminated (73%) shows a few confusion with DOUB. The expression of ARRO, which is perceptually located in the category of the expressions of dominance, is perceived rather correctly (44%), but it shows confusion with IRRI and OBVI. Concerning the category of the expressions of politeness (*i.e.* POLI and WOEG), both expressions are well recognized by listeners (56% for POLI, 55% for WOEG). OBVI which obtains a low recognition rate (29%) present some confusion with ARRO, IRRI, POLI.

From another hand, the synthetic stimuli are globally not as well recognized as the natural ones. However, the identification rate for the expressions of politeness: WOEG and POLI, was not very different from the one obtained with the natural stimuli (WOEG: 48%, POLI: 40%). IRRI was also discriminated correctly (40%) even if it shows confusion with ARRO, OBVI and POLI.

4.5. Correspondence Analysis

Figures 1 and 2 display the 2D graphical representation provided by the Correspondence Analysis (CA). The figures result from the projection on the two main axes of inertia, of the presented attitudes and of the subjects answers obtained during the perception evaluation respectively for the natural and the synthetic stimuli.

Concerning the perceptual results for natural stimuli, all presented affective expressions were generally correctly perceived (*i.e.* the presented stimuli points were very close to the perceptual points). This result confirms the result of the confusion matrix. The stimuli of DOUB were well discriminated by listeners from other presented affective expressions. The listeners showed similar perceptual behavior for the stimuli of SURP and ADMI, but both social affects were also well discriminated from other affective expressions. This figure also shows that two expressions of politeness (WOEG and POLI) were perceived in a similar way. Moreover, other expressions of dominance like OBVI, ARRO and IRRI were perceptually located quite close from each other.

As opposed to natural stimuli, the synthetic ones were much less accurately recognized, as mentioned in the previous section. However, listeners discriminated well WOEG from other affective expressions. Listeners did not confuse this social affect with POLI, which was the case when considering the natural stimuli. The stimuli of POLI were also well perceived, but showed confusion with OBVI. It is important to note that the stimuli of SURP were perceived as DOUB, but the stimuli of DOUB were perceived as ADMI or SURP. Such a perceptual behavior may be explained by the lack of breathiness in the synthetic stimuli of SURP. In addition, the stimuli of dominant expressions (*i.e.* OBVI, ARRO, IRRI) were located close to each other as for natural stimuli. Among these expressions of dominance, IRRI was well perceived, but OBVI and ARRO were both perceived as ARRO.

5. Conclusion

This paper investigated the perception of 'genuine' social affects versus 'synthetic' ones in order to improve the quality of synthetic affective prosody from neutral declarative utterances. The construction of synthetic affective prosody must comply with several constraints: keeping the original speaker identity, preserving the linguistic content, and of course having the best possible quality.

Perception evaluation results with native Japanese speakers, show that the social affects for natural stimuli are quite well recognized although the results show more variation on the synthetic stimuli depending on the considered social affect. For instance, two synthetic expressions of politeness: POLI and

	Recognized attitude							
Reference	Admiration	Arrogance	Doubt	Irritation	Obviousness	Politeness	Surprise	Walking on eggs
Admiration	**65.08**	0.00	0.00	0.00	1.59	0.00	32.54	0.79
Arrogance	1.59	**43.65**	0.79	24.60	19.05	7.94	0.79	1.59
Doubt	2.38	0.79	**88.10**	3.17	0.00	0.00	5.56	0.00
Irritation	0.00	15.08	0.00	**71.43**	9.52	3.97	0.00	0.00
Obviousness	0.79	17.46	5.56	**29.37**	25.40	13.49	7.94	0.00
Politeness	0.79	1.59	1.59	1.59	28.57	**55.56**	3.17	7.14
Surprise	7.14	0.79	17.46	1.59	0.00	0.00	**73.02**	0.00
Walking on eggs	13.49	2.38	0.00	0.79	12.70	15.87	0.00	**54.76**

Table 1: Recognition rate expressed in percents of the total answers for the natural stimuli (overall recognition rate : 59%).

	Recognized attitude							
Reference	Admiration	Arrogance	Doubt	Irritation	Obviousness	Politeness	Surprise	Walking on eggs
Admiration	9.52	6.35	3.17	7.14	**30.16**	24.60	6.35	12.70
Arrogance	0.79	18.25	0.00	13.49	**30.95**	26.19	1.59	8.73
Doubt	3.17	12.70	**30.16**	12.70	19.05	13.49	3.97	4.76
Irritation	0.00	25.40	2.38	**40.48**	15.87	15.87	0.00	0.00
Obviousness	0.79	7.14	6.35	19.05	**35.71**	26.98	3.17	0.79
Politeness	2.38	6.35	0.00	3.17	**43.65**	40.48	0.79	3.17
Surprise	4.76	2.38	**49.21**	7.14	13.49	9.52	12.70	0.79
Walking on eggs	3.17	5.56	2.38	1.59	15.87	23.02	0.00	**48.41**

Table 2: Recognition rate expressed in percents of the total answers for the synthetic stimuli (Overall recognition rate : 29%).

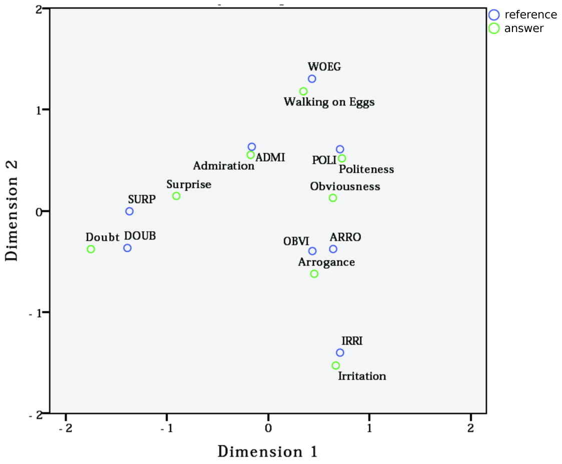

Figure 1: Correspondence analysis of natural stimuli.

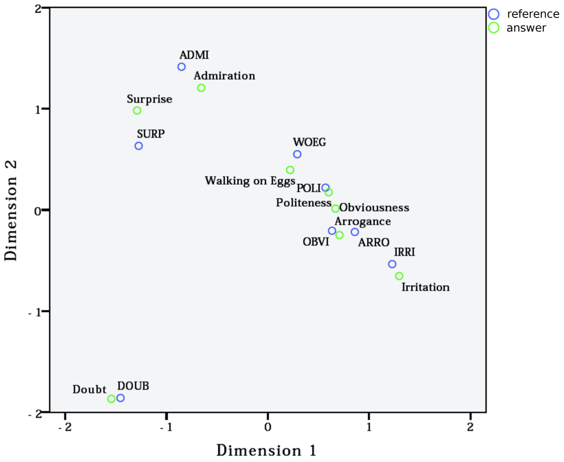

Figure 2: Correspondence analysis of synthetic stimuli.

WOEG are as well recognized by native listeners as for natural stimuli despite the lack of particular voice quality associated to these expressions [7]. However, the synthetic expression of SURP seems to be perceived as doubt. The perceptual behavior for such listeners can be explained by the lack of breathiness in the synthetic stimuli of SURP. In other words, it may indicate that SURP and DOUB have at least some common characteristics on the 3 investigated prosodic parameters. Other synthetic expressions were much less accurately recognized than natural ones showing some important confusions inside similar semantic categories. Therefore, future work will consist in an investigation of other acoustic parameters related to voice quality and in creating a software for self-teaching language learning that includes a tool which will allow learners to hear their own voice producing the social affect correctly.

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

7. References

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

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

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

[4] D. Fourer, T. Shochi, J.-L. Rouas, J.-J. Aucouturier, and M. Guerry, “Going ba-na-nas: Prosodic analysis of spoken japanese attitudes,” in *Proc. Speech Prosody 7*, 2014.

[5] T. Shochi, D. Fourer, A. Rilliard, J.-L. Rouas, and M. Guerry, “Perceptual evaluation of spoken japanese attitudes,” in *International Congress of Phonetic Sciences (ICPhS)*, 2015.

[6] T. Shochi, A. Brousse, M. Guerry, D. Erickson, and A. Rilliard, “Learning effect of social affective prosody in japanese by french learners,” in *Proc. Proc. Speech Prosody 8*, 2016, p. submitting.

[7] T. Shochi, A. Rilliard, V. Aubergé, and D. Erickson, *The role of prosody in affective speech*, ser. Linguistic Insights. Peter Lang, 2009, vol. 97, ch. Intercultural perception of English, French and Japanese social affective prosody, pp. 31–60.

[8] 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, no. 2-3, pp. 223–243, 2009.

[9] T. Sadanobu, “A natural history of Japanese pressed voice,” *Journal of the Phonetic Society of Japan*, vol. 8, no. 1, pp. 29–44, 2004.

[10] A. Rilliard, D. Erickson, T. Shochi, and J. A. D. Moraes, “Social face to face communication - American English attitudinal prosody,” in *Proc. Interspeech*, Aug. 2013.

[11] W. Gu, T. Zhang, and H. Fujisaki, “Prosodic analysis and perception of Mandarin utterances conveying attitudes,” in *Proc. Interspeech*, Aug. 2011, pp. 1069–1072.

[12] P. Boersma and D. Weenink, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[13] A. Camacho and J. G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” *Journal of the Acoustical Society of America*, vol. 124, pp. 1638–1652, 2008.

[14] “Livecode official site,” http://livecode.com/, accessed: April 8, 2016.

[15] G. W. Corder and D. I. Foreman, *Nonparametric Statistics: A Step-by-Step Approach*. Wiley, 2014.

[16] P. E. Greenwood and M. S. Nikulin, *A Guide to Chi-Squared Testing*. John Wiley and Sons, 1996.