Going ba-na-nas: Prosodic analysis of spoken Japanese attitudes

Dominique Fourer, Takaaki Shochi, Jean-Luc Rouas, Jean-Julien Aucouturier, Marine Guerry

Going ba-na-nas: Prosodic analysis of spoken Japanese attitudes

Dominique Fourer[1], *Takaaki Shochi*[2], *Jean-Luc Rouas*[1],
Jean-Julien Aucouturier[3], *Marine Guerry*[4]

[1]LaBRI - CNRS UMR 5800, Univ. Bordeaux 1, France
[2]CLLE-ERSSàB UMR5263 CNRS, Bordeaux, France
[3]IRCAM - CNRS UMR9912 - UPMC, Paris, France
[4]Univ. Bordeaux Montaigne, France

[1]firstname.lastname@labri.fr, [2]Takaaki.Shochi@u-bordeaux3.fr, [3]aucouturier@ircam.fr

Abstract

The aim of this paper is to examine cues for prosodic characterization of attitudes in Japanese. This work is based on previous studies where 16 communicative social affects were defined. The audio signal parameters (fundamental frequency, amplitude and duration) of previously recorded Japanese attitudes, are statistically analyzed. Interesting interactions among the parameters, the gender and the expression of specific attitude (e.g. politeness) were found, and we report on which parameters most significantly characterize each attitude.
Index Terms: speech, prosody, attitude, social affect, emotional speech, Japanese language

1. Introduction

The prosodic expressions of social affects, or attitudes as defined by [1], are a mean used by speakers to drive the illocutionary force of their intended speech acts [2] in face-to-face communication. Such choices are partly linked with the speaker's own proficiency in the spoken language, her/his personality, gender and the communication context which are also constrained at the linguistic level. Thus, each language has specific formulae or conventional prosodic variations for specific interaction contexts. Usually, studies investigating such kinds of prosodic variations rely on stereotypic stimuli [3, 4, 5, 6]. One common difficulty in studies which aim to compare the prosody of social affects is linked to a trade-off between the high sound quality required for acoustic analysis, the need for a neutral lexical content of the studied sentences (ideally identical sentences for all the studied affects), the search for spontaneity of the expressions, and a clear labelling of the communicative goals of the speaker. Most of the cited studies use laboratory corpora. Typically, adhoc sentences are recorded by speakers trying to read a sentence and reproduce a given expressivity. To enhance the spontaneity of these expressions and to facilitate the speaker's task, [7] proposes to place target sentences in affectively loaded texts. Similarly, [6] recorded attitudinally-neutral sentences embedded into dialogues that prepare the speaker to perform an adequate expression for the target sentence. The approach used during this research builds on these works. In order to study the expressive strategies used by speakers of varying linguistic backgrounds, communicative situations have been set-up so they can be plausibly used in different languages. The analysis method used in this paper also shares similar objectives to be applied to any language. We thus decided to use only low-level prosodic descriptors: the values of the fundamental frequency (F_0), the Root Mean Square (RMS) amplitude (computed on vowels) and the syllable duration. Using these features for statistical analysis with Repeated-Measure ANalysis Of VAriance (RM-ANOVA), we are able to determine to which extent each feature may contribute to the differentiation of attitudes. This paper is organized as follows: the framework we used for a speaker express social affects is described in Section 2. The recording set-up procedure and the feature extraction methods are respectively described in Sections 3 and 4. Finally, the results from the statistical analysis are presented in Section 5 and discussed in Section 6.

2. Social contexts for expression of attitudes

In order to immerse subjects in the context, a scenario was set up for each attitude, and the 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, 16 contexts have been selected, corresponding to a set of attitudes used in [8, 9] for different languages. Some of these contexts do not have lexical equivalents in all languages, as the corresponding communication situations have not been conventionalised in that particular culture. It is the case for example of the Japanese notion of *kyoshuku*, described by [10] 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 16 social affects were used in the present corpus: Admiration (ADMI)), Arrogance (ARRO), Authority (AUTH), Contempt (CONT), Doubt (DOUB), Irony (IRON), Irritation (IRRI), Neutral declarative sentence (DECL), Neutral question (QUES), Obviousness (OBVI), Politeness (POLI), Seduction (SEDU), Sincerity (SINC), Surprise (SURP), Uncertainty (UNCE), 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 [11] for details). For all situations, a short neutral target sentence has been used to record the respective prosodic expressions: "A banana". In order to elicite these target sentences in each context, small dialogues were written (cf. [6]), that take place in the prototypical context described above, and that end 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.

3. Recording procedure

A set of 19 Japanese native speakers (11 females, 8 males) have been recorded. 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 16 attitudes, resulting in a total of 304 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].

4. Feature extraction

We characterized each stimulus with its F_0, amplitude and duration parameters.

4.1. Fundamental frequency

The F_0 parameter measured in Hertz is estimated using the SWIPE algorithm [13] with a 10 ms sampling rate. For statistical analysis (see Section 5), only F_0 values at time indices corresponding to the vocalic phonemes are kept, and averaged to give a mean F_0 value per vowel (3 values per stimulus). We also applied the MultiDimensional Scaling algorithm (MDS) [14] to the complete series of F_0 values normalized by the mean F_0 of the speaker (see Section 5.4). Thus, MDS allows a graphical interpretation of the distance between the attitudes which depends on the computed correlation between the F_0 series.

4.2. Amplitude

The amplitude is estimated using the RMS function where the signal is windowed to result in 20 ms frames with 50% overlap. Thus, the RMS is computed on each frame from the normalized values (in $[-1;+1]$) of the signal samples. As before, for statistical analysis, only RMS values at time indices corresponding to the vocalic phonemes are kept, and averaged to give a mean RMS value per vowel (3 values per stimulus).

4.3. Duration

Additionally, for statistical analysis, each syllable duration measured is computed as the sum of its consonant and vocalic parts, in milliseconds, based on the manual segmentation of the stimuli.

5. Results of the statistical analysis

The statistical significance of prosodic differences is measured between attitudes and gender, separately for the F_0, RMS and duration parameters. For each parameter, each stimulus is considered as a series of three repeated measures, corresponding to the mean parameter value of the stimulus for the three successive vocalic phonemes. For each parameter, we conducted a RM-ANOVA, with attitude (16) $\times$ phoneme (3) as within-subject factors and gender (2) as between-subject factor. For the remainder of this section, $F(a,b)$ denotes the computed statistic which is assumed to follow a Fisher probability density function of parameters a and b. p denotes the p-value which results from the RM-ANOVA and M is used to denote the mean value of the considered parameter.

5.1. Fundamental frequency

The RM-ANOVA on F_0 values reveals a strong statistical interaction of attitude $\times$ time: $F(30,210)=8.12$, $p<0.001$, indicating that the temporal profile of F_0 varies significantly across attitudes. As shown in Figure 2(a), the interaction is most noticeable between F_0 values of the first and second vowel of each utterance: while the majority of attitudes are associated with lower-pitch in the second vowel ("V-shape"), attitudes AUTH, DECL, QUES and (to a lesser extent) SINC are associated with higher-pitched second vowels ("inverted V-shape").

To further characterize this behavior, we extract the normalized F_0 ratio between the second and first vowel of each stimulus (in % increase) - see Figure 2(d). A RM-ANOVA (with attitude (16) as a within-, and gender as a between-subject factor) reveals a main effect of attitude on this ratio: $F(15,225)=13.6$, $p<0.001$. Attitudes AUTH(+26%), DECL (+17%), QUES (+34%) show an increase in F_0 that is more important than the other measures ($p<0.001$, Bonferroni-corrected). Attitude SINC (+7%) also shows an increase in F_0 values more important than for CONT (-14%) ($p=0.01$, Bonferroni-corrected).

Most interestingly, the F_0 differences between attitudes have a marked interaction with gender: $F(30,210)=1.95$, $p=0.003$, indicating that different F_0 patterns are used by male and female to convey the same attitude. Remarkably, this interaction with F_0 is not characterized by the first (attitude $\times$ gender: $F(15,240)=1.36$, $p=0.16$), second (attitude $\times$ gender: $F(15,240)=1.83$, $p=0.03$) or third vowel individually (attitude $\times$ gender: $F(15,240)=0.38$, $p=0.98$), but rather on the difference between values on each vowel.

As seen in Figure 2(d), there are significant gender differences in particular in the second-to-first F_0 ratio, both

overall (attitude × gender: $F(15,225)=2.58$, $p=0.001$) and for individual attitudes ADMI (male: +5%, female: -11%, p=0.03), AUTH (male=+10%, female: +37%, $p=0.0008$) and QUES (male=+49%, female=+23%, $p=0.001$; all: Fisher LSD posthocs).

5.2. Amplitude

The RM-ANOVA on RMS values reveals a main effect of gender: $F(1,17)=24.5$, $p=0.00012$, indicating that recordings by female speakers are louder than males; a main effect of time: $F(2,34)=88.51$, $p<0.001$, with first ($M=.00071$) and second vowels ($M=.00066$) both twice louder than the third vowel ($M=.00034$);and a main effect of attitude: $F(15,255)=1.99$, $p=0.015$, showing that some attitudes are generally louder than others, regardless of vowel and gender.

Attitudes ADMI ($M=.00063$), POLI ($M=.00060$), SINC ($M=.00060$) and WOEG ($M=.00059$) are louder than AUTH ($M=.00054$), CONT ($M=.00052$), IRRI ($M=.00052$), QUES ($M=.00054$), SEDU ($M=.00053$), SURP ($M=.00056$) and UNCE ($M=.0052$) at the $p<.05$ level (Fisher LSD test, non Bonferroni-corrected).

There is an interaction of time × gender: $F(2,34)=3.51$, $p=0.04$ - second vowels were more quiet for males than females -, but no interaction of attitude × gender: $F(15,255)=0.88$, $p=0.58$.

Most remarkably, there is a strong interaction of attitude × time: $F(30,510)=6.57$, $p<.001$, showing that, as for F_0, different temporal profiles of amplitudes are used to convey different attitudes (see Figure 2(b)), and, as for F_0, differences are particularly notable in the ratio of RMS at the first and second vowels: while second vowels were quieter than first for the majority of attitudes, they were louder for attitudes AUTH, DECL and QUES.

However, contrary to F_0, there was no interaction of this effect with gender: $F(15,255)=0.33$, $p=0.99$ (attitude × gender interaction, RM-ANOVA on second-to-first RMS ratio). Similar amplitude patterns are used by male and female speakers to convey the same attitudes (see Figure 2(e)).

5.3. Duration

The RM-ANOVA on syllable durations reveals a marginal main effect of gender: $F(1,17)=4.71$, $p=0.04$ - females were slower ($M=137$ ms) speakers than males ($M=117$ ms); a main effect of time: $F(2,34)=95.6$, $p<0.001$, with final syllables twice as long ($M=188$ ms) as the first ($M=99$ ms) and second syllables ($M=98$ ms); and a main effect of attitude: $F(15,255)=13.0$, $p<.001$, showing that some attitudes were generally slower than others, regardless of syllable and gender.

Attitudes AUTH ($M=162$ ms), CONT ($M=166$ ms), DECL ($M=171$ ms), DOUB ($M=155$ ms), QUES ($M=147$ ms) and SURP ($M=165$ ms) were slower than the others ($p<.05$, Bonferroni-corrected). This main effect of attitude was seen both on durations of the first syllable: $F(15,255)=11.09$, $p<.001$, second syllable: $F(15,255)=9.48$, $p<.001$ and third syllable: $F(15,255)=13.08$, $p<.001$.

There was no interaction of time × gender: $F(2,34)=1.57$, $p=0.22$, and no interaction of attitude × gender: $F(15,255)=1.17$, $p=0.29$. However, as mentioned below, there was a strong interaction of attitude × time: $F(30,510)=12.07$, $p<.001$, showing that, as for F_0 and RMS, different patterns of successive syllable durations were used to convey different attitudes (see Figure 2(c)).

As for F_0 (but not for RMS), this interaction is significantly related with gender: $F(30,510)=2.43$, $p=<.001$. The gender difference was mainly seen on the duration of the third syllable: $F(15,255)=1.89$, $p=0.02$ (see Figure 2(f)); gender differences were only marginal for the duration of the first syllable: $F(15,255)=1.68$, $p=0.05$, and not significant for the second syllable: $F(15,255)=1.04$, $p=0.4$.

5.4. Multidimensional scaling

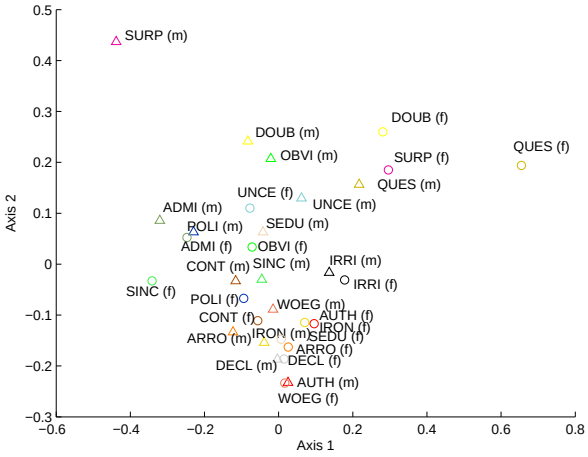
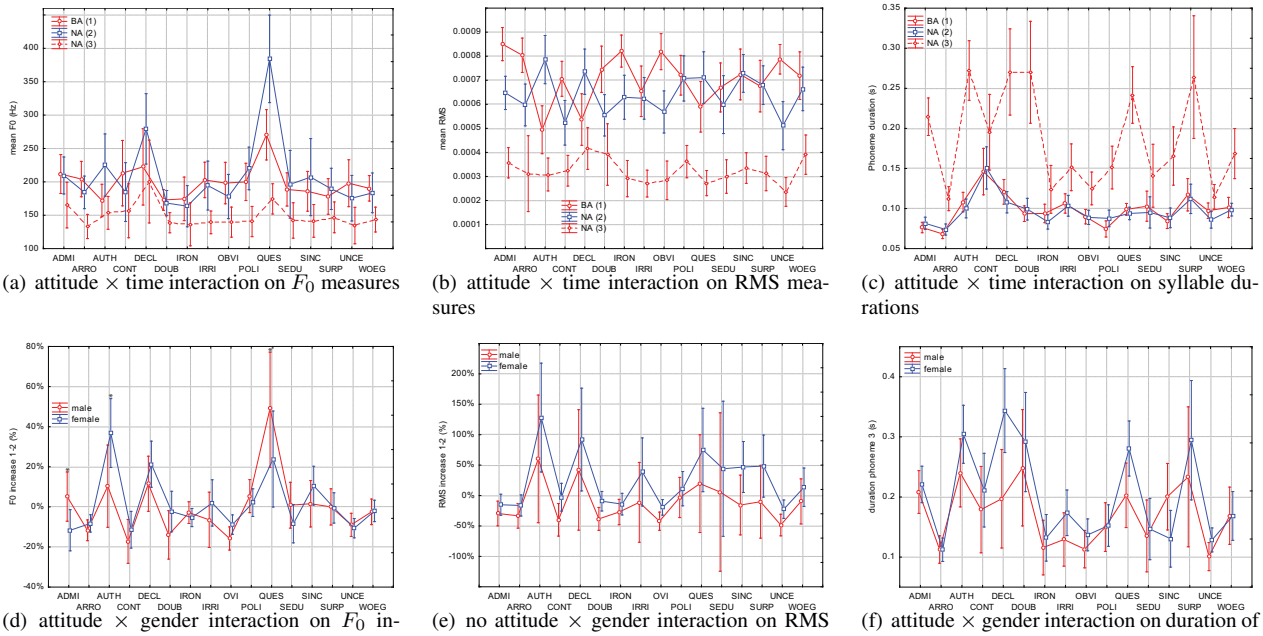


Figure 1: Result of the MDS algorithm based on the correlation distance of the F_0 profiles (circle: female, triangle: male).

The result of the MDS algorithm [14] applied on the normalized F_0 profiles associated to the entire phrase "banana" and computed on the corpus is presented in Figure 1. This figure shows the center of mass for each attitude where male and female are separated. The distance between each point is associated to the correlation distance computed between the profiles associated to the attitude of each speaker. For the normalization, each estimated F_0 profile is divided by the averaged F_0 related to the speaker to obtain a modulation function centered on the unity. For the duration, all the estimated profiles are rescaled to a series of 100 frames fitting the mean duration of the corpus which is 531.25 ms

6. Discussion

This study on Japanese spoken language investigated acoustic characteristics of various social affects in order to identify what are the similar and different prosodic patterns for various attitudes. For this study we used the F_0 and the RMS parameters which are objective signal parameters used to describe the prosody of each expressed attitude. Statistical results show that the interaction between attitudes and gender is observed only for the F_0 parameter. It indicates that the gender differences of attitudinal expressivity are more characterized by F_0 rather than RMS or duration. However, we have observed that female expressiveness is not only characterized by a higher pitch but also by a higher RMS and longer syllable durations than males. We also identified that Japanese politeness expressions (ADMI, POLI, SINC, WOEG) are louder than others, including impoliteness expressions (AUTH, CONT and IRRI). Concerning the duration parameter, the Japanese language is well known to be



(a) attitude × time interaction on F_0 measures

(b) attitude × time interaction on RMS measures

(c) attitude × time interaction on syllable durations

(d) attitude × gender interaction on F_0 increase

(e) no attitude × gender interaction on RMS increase

(f) attitude × gender interaction on duration of syllable 3

Figure 2: Differences in F_0 temporal profiles between attitudes 2(a) and gender 2(d). 2(a): mean F_0 value (in Hz) for each vowel of (b)a-(n)a-(n)a for the 16 attitudes. 2(d): difference of mean F_0 between the second and first vowels (in % increase) for the 16 attitudes and for both genders. Error bars in 2(a) and 2(d) correspond to 95% confidence intervals. Asterisks in 2(d) mark significant gender difference ($p<0.05$, Fisher LSD posthoc). Differences in RMS temporal profiles between attitudes 2(b) and gender 2(e). 2(b): mean RMS value for the three vowels of (b)a-(n)a-(n)a for the 16 attitudes. 2(e): difference of mean RMS between the second and first vowels (in % increase) for the 16 attitudes and for both genders. Error bars in 2(b) and 2(e) correspond to 95% confidence intervals. Differences in syllables duration between attitudes 2(c) and gender 2(f). 2(c): duration of each of the three syllables ba-na-na for the 16 attitudes. 2(f): duration of the third syllable for the 16 attitudes and for both genders. Error bars in 2(f) and 2(f) correspond to 95% confidence intervals.

mora-timed which means that each mora (taking account of two moras for long vowels, geminate obstruents and nasal /N/) is of similar duration because of the compensation of segmental variation of duration inside the mora-structure [15, 16]. However these results show an important temporal variation among attitudinal expressions. A variation between the two first vowels and the 3rd vowel was especially observed (i.e. 3rd vowel was almost twice as long as 1st and 2nd vowel). It suggests that social affects may change duration adjustment of Japanese rhythmic structure, and it confirms a previous work [17] which finds that the duration of this 3rd vowel is an important factor to make a difference in various attitudinal expressions. Moreover, a correlation between attitude and time based on the F_0 values was observed. Standard Japanese is described as a pitch accent language which is characterized by a rising F_0 at the beginning of the phrase, and an important pitch fall after an accented vowel [18]. The recorded sentence "Banana" has an accentual nucleus on the 1st mora. Thus, we expect to observe F_0 rising on the 1st vowel with a fall at the end of the 1st vowel. However results show that the phrasal initial rising of F_0 until the peak (F_0 max) vary according to attitudes. Our assumption is that the timing of F_0 peak whether it comes on the 1st or the 2nd is correlated with speech rate and RMS. According to our data , it seems that some attitudinal expressions (AUTH, QUES and DECL) where F_0 peak comes later (on the 2nd vowel) are related with slower speech rate and with an RMS peak which comes later (on the 2nd vowel) as opposed to the majority of attitudes which are associated with lower pitch in the second vowel ("V-shape"). MDS analysis for F_0 values of all attitudes identified 3 different categories: polite expressions, impolite expressions and dubitative expressions. The first category consists of 4 impolite expressions (AUTH, ARRO, IRON, IRRI) plus 2 other attitudes (DECL, WOEG). The second category is composed of 4 polite expressions (ADMI, SINC, POLI, SEDU (males only) plus 2 attitudes (OBVI, UNCE). It is important to note that WOEG which is akin to the Japanese polite expression of *Kyosyuku* is located in the category of impolite expressions. These results confirm previous work on Japanese politeness expressions [9, 19]. The expression of WOEG may be differentiated from impolite expressions because it does not have the same voice quality characteristics to the Japanese (polite) expression of *Kyoshuku* [9]. Contrary to the similarity of F_0 values of SEDU for males and females given from RM-ANOVA (see Figure 2(d)), MDS analysis identified that female values for this attitude are quite different from the male ones according to Figure 1, the female speakers' circle is located in the category of impolite expressions, which is different from the male one which is located in the politeness category. Although the pitch of SEDU is similar to impolite expressions, the voice quality of this attitude is softer, and therefore SEDU may not be perceived as impolite expression. For future work, a comparison with non-native speakers will also be done to examine prosodic similarities among different languages.

7. Acknowledgements

Merci à Donna Erickson qui a contribué à améliorer cet article. Merci à Mariko Kondoet à Sylvain Detey pour l'enregistrement du corpus à l'université de Waseda. Cette recherche est financée par le projet ANR PADE et le projet PEPS IDEX/CNRS de l'université de Bordeaux I.

8. References

[1] A. Wichmannn, "The attitunal effects of prosody, and how they relate to emotion," in *Proc. ISCA Workshop on Speech and Emotion*, 2000, pp. 143–148.

[2] I. Fonagy, E. Bérard, and J. Fonagy, "Clichés mélodiques," *Folia Linguistica*, vol. 17, pp. 153–185, 1984.

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

[4] Y. Morlec, G. Bailly, and V. Aubergé, "Generating prosodic attitudes in French: Data, model and evaluation," *Speech Communication*, vol. 33, no. 4, pp. 357–371, 2001.

[5] J. A. de Moraes, "The pitch accents in Brazilian Portuguese: analysis by synthesis," in *Proc. Speech Prosody*, 2008, pp. 389–397.

[6] G. Wentao, T. Zhang, and H. Fujisaki, "Prosodic analysis and perception of Mandarin utterances conveying attitudes," in *Proc. Interspeech*, Aug. 2011, pp. 1069–1072.

[7] I. Grichkovtsova, M. Morel, and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication*, vol. 54, no. 3, pp. 414–429, 2012.

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

[9] 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.

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

[11] 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.

[12] P. Boersma and D. Weenink. (Version 5.3.32 retrieved 17 October 2012 from http://www.praat.org/) Praat: doing phonetics by computer [computer program].

[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] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 5, pp. 1168–1172, 2006.

[15] M. S. Han, "Acoustic manifestations of mora timing in japanese," *Journal of acoustical society of America*, vol. 96, pp. 73–82, 1971.

[16] Y. Sagisaka and Y. Tohkura, "Phoneme duration control for speech synthesis by rule," *IEICE Trans.*, vol. 67, no. 7, pp. 629–636, 1984.

[17] K. Maekawa and T. Kagomiya, "Influence of paralinguistic information on segmental articulation," in *Proc. 6th International Conference on Spoken Language Processing*, Oct. 2000, pp. 349–352.

[18] H. Fujisaki and H. Sudo, "A model for the generation of fundamental frequency contours of japanese word accent," *Journal of acoustical society of Japan*, vol. 27, no. 9, pp. 445–453, 1971.

[19] T. Shochi, V. Aubergé, and A. Rilliard, "How prosodic attitudes can be false friends: Japanese vs. French social affects," in *Proc. Speech Prosody*, May 2006, pp. 692–696.