Cultural Differences in Pattern Matching: Multisensory Recognition of Socio-affective Prosody

Takaaki Shochi, Jean-Luc Rouas, Marine Guerry, Donna Erickson

Interspeech 2018
2-6 September 2018, Hyderabad

Cultural differences in pattern matching: multisensory recognition of socio-affective prosody

Takaaki Shochi[1,2], *Jean-luc Rouas*[1], *Marine Guerry*[2], *Donna Erickson*[3]

[1]Univ. Bordeaux, CNRS, LaBRI, UMR 5800, 33405 Talence, France
[2]CLLE-ERRSaB CNRS UMR 5263, Bordeaux, France
[3]Kanazawa Medical University, Japan

takaaki.shochi@labri.fr, jean-luc.rouas@labri.fr, marine.guerry@u-bordeaux-montaigne.fr, EricksonDonna2000@gmail.com

Abstract

This study focuses on the cross-cultural differences in perception of audio visual prosodic recordings of Japanese social affects. The study compares cultural differences of perceptual patterns of 21 Japanese subjects with 20 French subjects who have no knowledge of Japanese language or Japanese social affects. The test material is a semantically affectively neutral utterance expressed in 9 various social affects by 2 Japanese speakers (one male, one female) who were chosen as best performers in our previous recognition experiment. The task was to create a specific audio-visual affect by choosing one video stimuli among 9 choices and one audio stimuli, again among 9 choices. The participants could preview each audio and video stimuli individually and also the combination of chosen stimuli. The results reveal that native subjects can correctly combine auditory and visually expressed social affects, showing some confusion inside semantic categories. Different matching patterns are observed for non-native subjects especially for a type of cultural-specific politeness.

Index Terms: Multisensory recognition, Pattern matching, Cultural difference

1. Introduction

In face to face communication, both vocal and visual expressions of affect interact with each other and convey a synergic complex of information ([1], [2], [3], [4]). In order to understand more precisely about the importance of each modality on the perception of affective information, many studies (for example [5], [4], [6], [7], [8]) have investigated cognitive processing of the crossmodal integration using congruent and incongruent combinations of audio and visual information. For instance, [5] showed in their experiment that both audio and visual modalities have a strong impact on the recognition of emotional expressions. Moreover, the subjects were faster in recognition of the conveyed information if the facial expression is congruent with the vocal expressivity. Their study however used static visual information as visual input and therefore results may be different if using dynamic visual movement. [7] also investigated multisensorial perception of affects using dynamic facial expressions and revealed that the modality dominance between audio and visual information changes for each affect.

Still, these studies were carried out with emotional expressions such as anger, joy, sadness, etc. (see [9]) rather than intentional (and voluntarily controlled) social affective expressivities (cf. irony, contempt, seduction, suspicious, etc.), although some studies used "intended emotion" which is more similar to our "social affect" (ex. [7]). According to [10], social affects are defined by the speakers' social status, and their intention which is conveyed in face to face interaction. These are supposedly learned during the developmental process in the social environment [11]. Therefore, these affects may vary from one culture to another, and this can lead to misunderstandings [2] .

The acoustic as well as visual aspects of social affects are described in many languages ([10], [12], [13]) and the language-specific aspect of such expressions has been stressed in a number of studies ([14], [15], [16], [17], [18]). Indeed, Pavlenko [19] mentioned the importance of affective meanings during speech communication in her book focusing on the cross cultural differences and common features of vocal affective expressions. Despite all those works, we do not know yet which acoustic and visual (facial and body control) parameters allow a listener to distinguish the various affective expressions.

Our current paper investigates the recognition pattern of social affects in audio-visual conditions. Specifically, the purpose of this study is to examine two points: 1) how do native subjects combine auditory cues and speaker's facial expressivity for social affects, and 2) what are some cultural differences between native and non-native subjects of multisensory perceptual patterns of these social affects.

Building on a paradigm targeting cross-cultural recordings [20], the paper presents results comparing the perception of 9 Japanese social affects. These 9 social affects were selected by previous research in linguistics, phonetics and psychology ([21], [22], [23], [24]).

This paper is organised as follows: the acquisition of the corpus and the method adopted for automatic combination of synthetic stimuli are respectively described in Section 2. Finally, the results from the statistical analysis are presented in Section 3 and conclusion and perspective in Sections 4 and 5.

2. Perception experiment

2.1. Corpus

The database used for this experiment is described in [25]. It consists of 19 native Japanese speakers uttering the word "banana" in 16 different social-affective contexts using carefully designed scenarii. Out of these 16 contexts, we selected for the purpose of this study 9 contexts according to 1) the social power, 2) social politeness strategy and 3) the social proximity (see Table 1).

Most social-affective situations are rather self-explanatory, with the exception of "walking-on-eggs". This category is used to denote a situation corresponding, to some extent, to a situation where Japanese speakers would express "Kyoshuku", a Japanese-specific concept defined as "corresponding to a mix-

10.21437/Interspeech.2018-1795

Table 1: *Selected 9 social affects*

Potentially universal affect:	
Surprise (SURP)	
Cultural specific affects:	
Obviousness (OBVI)	Social hierarchy (power)
Suspicious Irony (IRON)	
Contempt (CONT)	
Irritation (IRRI)	
Politeness (POLI)	Social politeness strategies
Sincerity (SINC)	
Walking on eggs (WOEG)	
Seduction (SEDU)	Social proximity (distance)

ture 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" ([26], p. 34.)

We expect that the "universal" affect "surprise" can be well created by both Japanese and French subjects, but we hope to look at some differences for more cultural-specific affects, particularly for the aforementioned "kyoshuku - walking-on-eggs" affect.

Based on a perceptual evaluation of the 19 Japanese native speakers' performances in each of these situations [20] using ratings from 38 listeners different from the current experiment, the two best performing speakers (1 female and 1 male) were selected for each sentence in each situation.

2.2. Experimental design

The aim of the experiment is to test whether our subjects are able to combine audio stimuli and video stimuli from the different social affective contexts to create a congruent audio-visual stimuli. Thus, the question submitted to the subjects was "Select the audio and video which best expresses the following social affect: XXX", XXX being one of the 9 social affective contexts. The interface should then allow the subjects to 1) look at each of the proposed video stimuli without sound 2) listen to each audio stimuli 3) combine the chosen audio and video stimuli in order to see and listen to their choice simultaneously. These steps could be repeated as long as they wish until they are satisfied with the result.

A screenshot of the interface is displayed on Figure 1. The main difficulty when designing such an interface is to be able to combine the audio and video files (of different lengths) to provide a plausible result even if there is a mismatch between the audio and video stimuli.

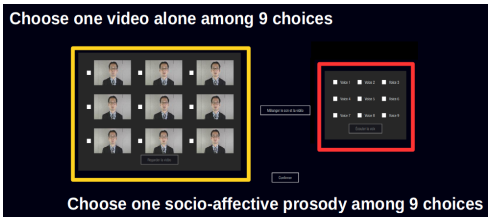

Figure 1: *Interface of the perceptual experiment*

2.3. Automatic combination of synthetic stimuli

The first step to this automatic combination of synthetic stimuli is to separate each recorded audio-visual utterance from our database into audio alone (A) and video alone (V). A manual transcription in phonemes (P) is also needed.

Given a video file from social affect 1, denoted V_1 and an audio file from social affect 2, denoted A_2, after several experiments, we choose to keep the audio files (A_2) as they are and modify the video files (V_1) by removing or duplicating frames at the middle of the phonetic segments from social affect 2 P_2 resulting in a new video file ($\hat{V_{2\to1}}$). This solution was proved more natural after empirical testing. The process for creating the stimuli using this solution is illustrated in Figure 2.

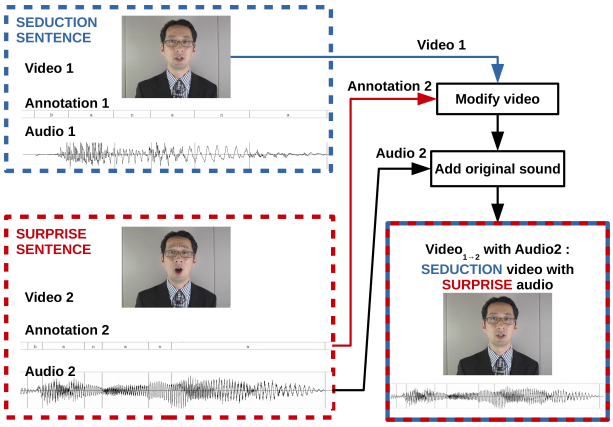

Figure 2: *combining video from seduction sentence with audio from surprise sentence*

A total of 18 sentences (2 native speakers x 9 affective expressions) were chosen as source signals for the synchronisation of audio and video signals in order to create all possible combinations of the selected 9 social affective expressions. For instance, auditory-expressed "sincerity" was synchronised with visual expression of "irritation" etc... By this synthetic method, we compiled 162 synthetic audio visual affective expressions (9 affects x 2 speakers x 9 combination types) which were integrated in the interface.

The interface was implemented in JAVA and allowed to see a preview of the windowed videos when hovering with the mouse and to watch the video alone in full-screen, to listen only to the selected audio, and to display the combined audio-video stimuli in full-screen. The questions were always asked in the native language of the participants, whom were allowed a training session using different speakers before taking the test.

2.4. Subjects

Two groups of listeners participated in the experiment: 21 native Japanese subjects (JP), all Tokyo dialect speakers (mean age= 19.8; 16 females; 5 males) and 20 French subjects without any knowledge of Japanese (FR) (mean age= 32.6; 10 females; 10 males). All the subjects used the same Bose 5C7N1 high quality headphones.

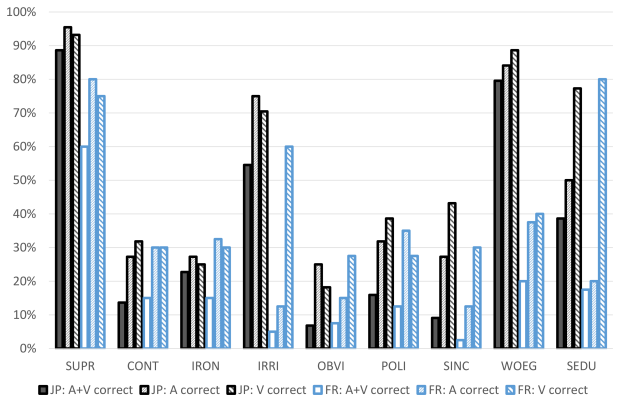

Figure 3: *% of correct answers for 9 affective expressions*

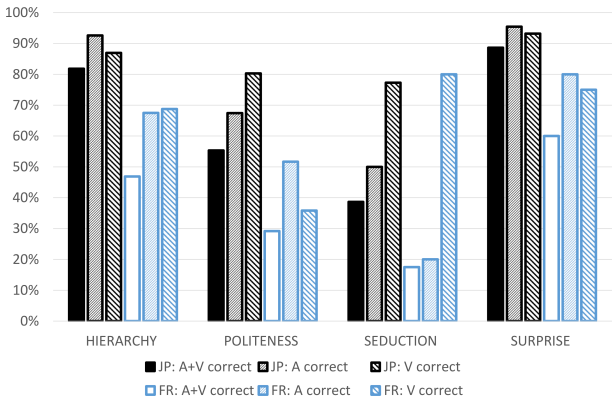

Figure 4: *% of correct answers for 4 classes of social affects*

3. Results

3.1. Correctly produced affects

We first observe the number of correct answers for the 9 affective expressions. Here, a correct answer means that a subject's choice in audio and visual modalities matched the one asked. Given the number of possibilities (9 video x 9 audio = 81 possible combinations) and the quality of the combination procedure, the subjects had some trouble finding both the video and audio corresponding to the affect they were asked to create (see the filled bars on Figure 3). Nevertheless, both Japanese and French subjects managed to create the "surprise" affect which has been hypothesised to be potentially universal.

Furthermore, native Japanese subjects manage to create the "walking on eggs" affect quite well contrary to the French subjects. The visual expressivity for "irritation" and "seduction" was well selected by both groups.

Globally, the French subjects do not manage to match correctly the audio and the video, with the exception of the "surprise" affect. Visual cues seems however to be perceived correctly by the French on seemingly more "caricatural" affects such as "irritation" and "seduction".

When grouping the social affects in the 4 broad classes defined above, the results are displayed on Figure 4. This figure show that most confusions between affects seem to be made on the same theoretical categories. Although Japanese subjects perform globally better than French ones, "Hierarchy" is well produced by both groups. It is to be noted (though not displayed on the figure) that French subjects used the "walking on eggs" audio stimulus 50% of the time when trying to create Japanese "seduction".

Further analysis and validation of the theoretical clustering are carried out in the next section.

3.2. Correspondence Analysis

In this section, in order to observe the perceptual distance between the real expressivity and all their selected expressivities for each social affect based on the classification made by the subjects, we computed a Correspondence Analysis (CA) using FactoMineR package ([27]) under R software.

First, the percentage of explained variances was computed for 8 dimensions. According to this analysis, the first four dimensions explain 85.6%, so we analysed perceptual points on these dimensions.

Figure 5 shows the CA analysis for JP and FR groups on the

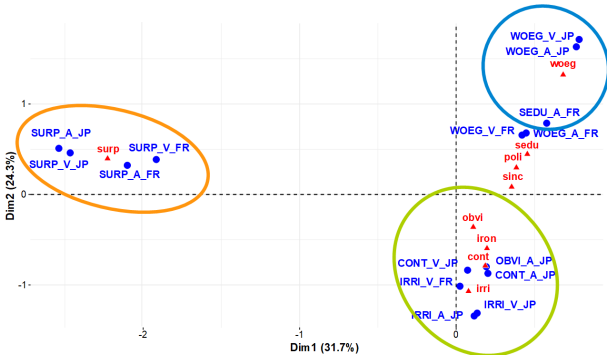

Figure 5: *Perceptual behavior of JP & FR groups in 1st and 2nd dimensions from the CA*

1st and 2nd dimensions. The blue dots on the figures represent the distribution of the perceptual behaviour and the 9 red triangles represent the concept subjects have of the 9 social affects.

On the first dimension, "surprise" is well categorised in Audio and Visual modalities by both French and Japanese groups. On the second dimension, the perceptual point of Japanese subjects for "walking on eggs" is very close to the conceptual point, indicating that this social affect is quite well discriminated. French subjects however tend to select the expressivities linked to "walking on eggs" for "seduction".

On the same dimension, "irritation" is also well discriminated by Japanese subjects. French subjects tend to choose the visual expressivity corresponding to this social affect, but their selected irritated voice included the correct "irritation" affect, but also "contempt", "irony" and "obviousness". This suggests that the irritated voice spoken by the Japanese speakers may be interpreted by French listeners as representing global generic dominant affects.

Figure 6 represents the distribution of perceptual points on the third and fourth dimensions. This figure shows that "seduction" is well discriminated by Japanese subjects in both audio and visual modalities. On the other hand, French subjects select the same expressivity correctly only in visual modality, and they tended to select the auditory expressivity of this affect for "walking on eggs".

Note also that Japanese subjects tend to behave with a similar auditory perception for "sincerity", "politeness" and "seduction".

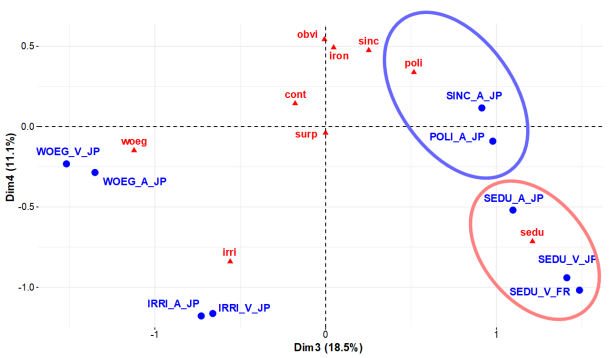

Figure 6: *Perceptual behavior of JP & FR groups in 3rd and 4th dimensions from the CA*

3.3. Clustering Analysis

This matching recognition pattern was also classified by the clustering analysis (ward.D method) using the "stats" package under R software (see figure 7). This statistical analysis allows to establish hierarchical psychometrical categories of social affective expressiveness for the 9 social affects.

According to the results, a total of 8 clusters are identified on Table 2.

Table 2: *clusters resulting from the automatic analysis*

Cluster 1	Surprise
Cluster 2	Japanese "walking on eggs"
Cluster 3	French "walking on eggs" and audio "seduction"
Cluster 4	French and Japanese visual "seduction"
Cluster 5	Politeness cluster and Japanese Audio "seduction"
Cluster 6	Japanese "irritation" and French visual "irritation"
Cluster 7	French social hierarchy
Cluster 8	Japanese social hierarchy

Cluster 1 is the "surprise" category. As we observed in the CA analysis, this affect is expressed in a similar way by both French and Japanese subjects. Three other clusters are also shared by both language groups: cluster 4 - "seduction", cluster 5 - politeness expressions including Japanese auditory expressed "seduction" and cluster 6 - "irritation",

On the other hand, there are some clusters which indicate cross cultural differences. For instance, clusters 2 and 3 represent both the "walking of eggs" category. The expressivities of two groups for this Japanese politeness affect are not classified in the same cluster. This confirms the CA analysis result that this affect is expressed quite differently by the two groups. Note also that Cluster 3 includes auditory-expressed "seduction". French subjects tended to use expressivities linked to "walking on eggs" for "seductive expression" as shown in the previous CA analysis. Furthermore, both clusters 7 and 8 are a category of social hierarchy. However, both language groups tended to select different corresponding expressivities.

4. Conclusion

The current work investigated cultural differences of perceptual patterns of native and non-native subjects who have no knowledge of the Japanese language and culture in the recognition of the social affective meanings of utterances extracted from a social interaction database.

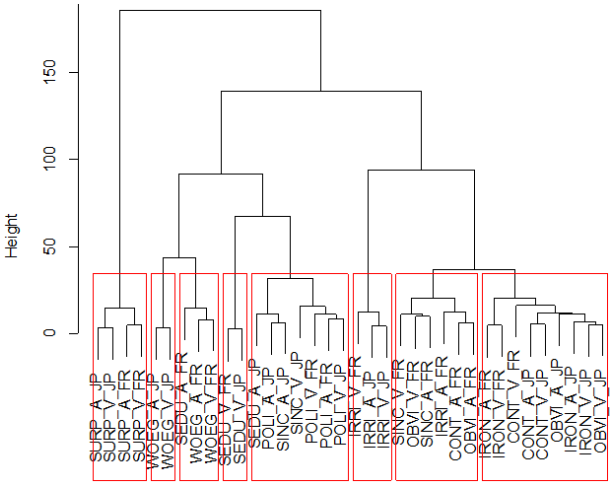

Figure 7: *8 main clusters for both JP & FR groups*

The theoretical broad categories of social affects (potentially universal, social hierarchy, social politeness and social proximity) seem to be confirmed by our experiments, although there exists some confusion between the expressiveness of "seduction" and "politeness".

Statistical analysis of the results indicate that the expression of "surprise" was well discriminated by both Japanese and French groups. This result confirms previous results which revealed that this affective expression is shared among various languages/cultures.

Some cultural differences between the French and Japanese groups exist for a social affect which is linked to social proximity and politeness strategy: "walking on eggs". While Japanese subjects had no trouble selecting the appropriate audio and visual expressivities for this social affect, French subjects did not manage to carry out this task. They rather tended to select the audio expressivity of "walking on eggs" for "seduction". This tendency was previously shown in [2], and our current result confirms this perceptual distortion for this culture-specific social affect. Moreover, all other Japanese politeness expressions are perceived in similar ways by both language groups.

5. Perspectives

For future work, we expect to increase the number of subjects in order to analyse the speaker's gender effect for their perceptual behaviours. Comparison of French perceptual behaviours with those of American English subjects will also be explored in order to identify cultural specific and universal affects. Moreover, application of this knowledge to E-learning of foreign language as well as to emotion recognition psychotherapy are expected to be some of the outcomes of this research into cultural similarities/differences of audio-visual affective expressivities.

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 (ANR-10-IDEX-03-02), Cluster of excellence CPU. We are deeply indebted to A. Rilliard (Limsi-CNRS), S. Detey (Waseda University), and students from Bordeaux and students from Waseda University (Japan).

7. References

[1] S. Jessen and S. A. Kotz, "Affect differentially modulates brain activation in uni-and multisensory body-voice perception," *Neuropsychologia*, vol. 66, pp. 134–143, 2015.

[2] T. Shochi, A. Rilliard, V. Auberge, and D. Erickson, "Intercultural perception of english, french and japanese social affective prosody," *The Role of Prosody in Affective Speech*, pp. 32–59, 2009.

[3] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1, pp. 227–256, 2003.

[4] P. Barkhuysen, E. Krahmer, and M. Swerts, "Crossmodal and incremental perception of audiovisual cues to emotional speech," *Language and speech*, vol. 53, no. 1, pp. 3–30, 2010.

[5] B. de Gelder, J. Vroomen, and G. Pourtois, "Seeing cries and hearing smiles: Crossmodal perception of emotional expressions," *Advances in psychology*, vol. 129, pp. 425–438, 1999.

[6] O. Collignon, S. Girard, F. Gosselin, S. Roy, D. Saint-Amour, M. Lassonde, and F. Lepore, "Audio-visual integration of emotion expression," *Brain research*, vol. 1242, pp. 126–135, 2008.

[7] S. Takagi, S. Hiramatsu, K.-i. Tabei, and A. Tanaka, "Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality," *Frontiers in integrative neuroscience*, vol. 9, 2015.

[8] B. de Gelder, K. B. Böcker, J. Tuomainen, M. Hensen, and J. Vroomen, "The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses," *Neuroscience letters*, vol. 260, no. 2, pp. 133–136, 1999.

[9] P. Ekman and W. V. Friesen, *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[10] F. Daneš, "Involvement with language and in language," *Journal of pragmatics*, vol. 22, no. 3, pp. 251–264, 1994.

[11] V. Aubergé, "Prosodie et émotion," in *Actes des deuxiemes assises nationales du GdR I3*, 2002, pp. 263–273.

[12] Å. Abelin, "Cross-cultural multimodal interpretation of emotional expressions _ an experimental study of spanish and swedish," in *Speech Prosody 2004*, 2004.

[13] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[14] 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.

[15] I. Fónagy, *La vive voix, Essais de psycho-phonétique*. Paris: Payot, 1982.

[16] J. Pierrehumbert and J. B. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," *Intentions in communication*, pp. 271–311, 1990.

[17] C. Shan, S. Gong, and P. W. McOwan, "Beyond facial expressions: Learning human emotion from body gestures." in *BMVC*, 2007, pp. 1–10.

[18] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 64–84, 2009.

[19] A. Pavlenko, *Emotion and Multilinguism*. Cambridge: Cambridge University Press, 2005.

[20] 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.

[21] A. Wichmann, "The attitudinal effects of prosody, and how they relate to emotion," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[22] W. Gu, T. Zhang, and H. Fujisaki, "Prosodic analysis and perception of mandarin utterances conveying attitudes," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[23] T. Shochi, A. Rilliard, V. Aubergé, and D. Erickson, "The role of prosody in affective speech, volume linguistic insights 97, chapitre intercultural perception of english, french and japanese social affective prosody," *Peter Lang*, pp. 31–59, 2009.

[24] H. Fujisaki, "A model for the generation of fundamental frequency contours of japanese word accent," *Journal of the Acoustic Society of Japan*, vol. 27, pp. 445–453, 1971.

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

[26] T. Sadanobu, "A natural history of japanese pressed voice," *Journal of the Phonetic Society of Japan*, vol. 8, pp. 29–44, 2004.

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org