10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

An Acoustic Study of Emotional Speech Produced by Italian Learners of Japanese

Motoko Ueyama[1], *Xinyue Li*[2]

[1]Department of Interpreting and Translation, University of Bologna, Italy
[2]Kobe University & ATR Hiroshi Ishiguro Lab, Japan
motoko.ueyama@unibo.it, lixinyue@atr.com

Abstract

Pioneering research on L2 emotional speech provides evidence for crosslinguistic similarities and differences, but there is limited research at the production level. This study examines the production of emotional speech in both L1 and L2 of Italian learners of Japanese and L1 Japanese. We sought to find the acoustic characteristics of emotional speech for this L2-type and investigate transfer effects. We analyzed single-word utterances with five emotions (neutral, happy, angry, sad, and surprised) for six acoustic parameters (F0mean, F0max, F0min, F0range, intensity, and utterance duration). Asymmetric patterns are found for different emotions. For happy and surprised, all speech types show higher pitch and larger pitch range with respect to neutral: i.e., *positive transfer*. In contrast, for angry, we found only larger pitch range in L1 Japanese, but higher F0mean and F0max and larger pitch range in L1 Italian; for sad, lower F0 level and smaller pitch range in L1 Japanese, but higher F0mean, F0max, and F0min and smaller intensity in L1 Italian. For both angry and sad, L2 Japanese shows F0 patterns similar to those of L1 Italian: i.e., *negative transfer*. The findings are discussed in light of valence and arousal features of emotions and other directions for future research.

Index Terms: emotional speech, L2 speech, L2 prosody, L2 Japanese, Italian learners of Japanese, paralinguistics

1. Introduction

In speech communication, speakers produce and perceive not only linguistic information but also nonlinguistic information—that is, information not generally controlled by the speaker, such as the speaker's emotion. Research on emotional speech has increased rapidly in the past decade of this digital age [1], [2].

Emotional speech involves many factors, e.g., a complex interaction between rather complex acoustic and articulatory characteristics of the utterance, as well as the social interaction between speaker and listener [1]. There have been numerous studies on various aspects of emotional speech, including the production of emotional speech.

Earlier research on acoustic properties of emotional speech has shown that multiple parameters are modulated when a speaker's emotion changes. The acoustic cues of emotional speech involve information about F0/pitch; duration at segmental, syllabic, or utterance level; speech rate; amplitude/loudness/power; voice quality; and a combination of all of these [1]–[3]. Different emotions are characterized by certain acoustic patterns. For example, happiness/joy is commonly characterized by high mean pitch, wider pitch range, and high intensity [2]. Sadness is characterized by a decrease in mean pitch, narrower pitch range, and slower speaking rate [4]. Besides such similarities, crosslinguistic differences have also been reported. For example, in [5], the production of hot anger was compared for Japanese and Chinese by analyzing acoustic parameters and electroglottography signals. The results showed that Chinese speakers used higher overall pitch register and tenser voice, while neither of these two patterns was observed in Japanese speakers.

It is easy to predict that such differences may result in L1 transfer to L2 emotional speech. Language transfer is defined as the effects of the learner's L1 on L2 learning [6], and it has been investigated in both segmental and prosodic aspects. However, limited research has been done on L2 emotional speech, especially at the production level. To investigate L1 effects, it is ideal to collect data concerning the same learner's L1 and L2, as well as data from native speakers of the target language as the baseline, using the same experimental setting. These three types of emotional speech were compared in [5]: L1 Chinese, L2 Japanese-L1 Chinese, and L1 Japanese. The results showed similar patterns in L1 Chinese and L2 Japanese-L1 Chinese, which differ greatly from L1 Japanese, indicating L1 transfer effects at the production level. However, not all differences between the learner's native and target languages necessarily result in negative transfer, which is also true for prosodic aspects of L2 production [7]. This is reported in [8], who conducted an analysis of emotional speech produced by learners of Italian as L2 both in Italian and in their own L1s (Spanish, Russian, and Tunisian Arabic). Tunisian learners reproduced intonational contours very similar to L1 in L2 Italian. Both Russian and Spanish learners, however, showed hybrid prosodic patterns that did not adhere to their L1 and diverged from those of native speakers of Italian.

One of the long-term efforts in L2 speech learning research is investigating how L1 influences L2 speech development. More aspects of L2 speech still need to be investigated, including the affective aspect of speech, which is still understudied. This study contributes to remedying this lack by examining the production of emotional speech in L2 Japanese-L1 Italian. The first goal is to find the basic acoustic characteristics of the emotional speech of this L2-type via an analysis of the production by Italian learners of Japanese both in Japanese and in their own L1, as well as by native Japanese speakers as the baseline. The second goal is to determine whether and to what extent the phonetic realization of L2 emotional speech is influenced by transfer phenomena from the learners' L1 background and/or reveals more universal tendencies of the L2 speech-learning process.

10.21437/SpeechProsody.2020-8

2. Method

2.1. Participants

For L1 Japanese and L2 Japanese-L1 Italian, part of our data came from the Kobe Archive of Nonnative Intonation in Japanese (KANI-J Corpus), which is a corpus of L2 Japanese speech under construction that is designed for comparing learners representing various L1 backgrounds (including Italian) for six prosodic aspects (including affective) in addition to a baseline group of native speakers of Tokyo Japanese [9], [10]. All Italian participants were students of the University of Bologna learning Japanese. None of them had visited Japan before. At the time of the recordings, they had completed Japanese lessons for 160–200 hours, taking undergraduate Japanese courses. Their estimated proficiency level was upper-elementary: low enough to observe possible L1 transfer, but high enough to perform production tasks. The Japanese participants were also college students who were native speakers of Tokyo Japanese.

This study first aimed at identifying characteristic acoustic patterns of emotional expressions in each speech type, and we selected relatively expressive speakers as follows. We first selected only female speakers of both groups and asked three native speakers of Japanese to listen to L1 and L2 Japanese data to rate how clearly each speaker differentiated five emotion types (neutral, happy, sad, angry, and surprised) with a five-point scale. Based on the rating results, the top five speakers were selected for each language group. The Italian production of selected Italian learners was also rated with the above procedure by three native Italian speakers to ensure the clarity of their emotional speech in their L1 as well.

2.2. Materials

The emotional speech task of the KANI-J Corpus is structured with a 5 × 5 design (25 stimuli in total):

- Five emotions: neutral, happy, angry, sad, and surprised
- Five one-word utterances: for Japanese, *nani* 'what', *sō* 'that's right', *are* 'that', *Oranda* 'Holland', *Manami-san* 'Manami'

The Italian version of the task was prepared, using five one-word utterances equivalent to the Japanese ones: *cosa* 'what', *sì* 'yes/that's right', *quello* 'that', *Orlanda* 'Holland', *Milena* (personal name).

The participants first read all five stimuli neutrally, without any specific emotion. Subsequently, the remaining 20 stimuli with four emotions (happy, angry, sad, surprised) were presented for each word through mimicked conversation. Each stimulus was presented with a short description of the context of a conversation in Japanese and English for the Japanese task and in Italian for the Italian task. The participants first listened to a pre-recorded audio statement and then pronounced a stimulus in reaction to the statement with the specified emotion type.

2.3. Procedure

2.3.1. Data collection

Most recordings of L1 and L2 Japanese were made using the Online Voice Recorder tool developed by Minematsu and Saitō Lab of the University of Tokyo [11], a computer connected to the Internet, a headset microphone, and headphones or earphones. For some recordings of L1 and L2 Japanese and all recordings of L1 Italian, Microsoft Power Point was used to display the stimuli, with an IC recorder and a pin microphone. In either case, the microphone signal was recorded at a 48kHz sampling rate and an 18-bit quantization rate.

2.3.2. Data analysis

Praat [12] and a script based on [13] were used to measure values of minimum F0, maximum F0, mean F0, F0 range, medians of intensity, and utterance durations. Obtained values were checked manually to ensure the accuracy of the measurements. F0 values were converted into semitones by the formula given in [14]: 12[log(F0 Hz/100)/L2]. Statistical analyses were run using SPSS.

3. Results

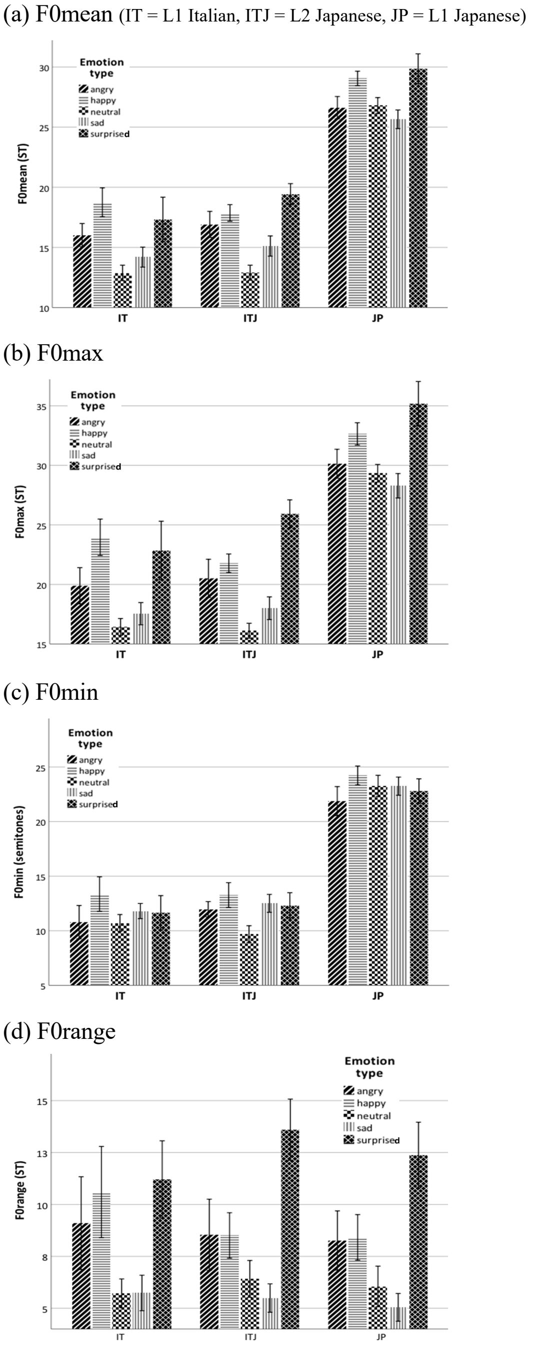
The results of the descriptive statistics were examined first for the individual speakers of each speech type for each acoustic parameter, especially, boxplots (median, upper and lower quartile, minimum and maximum values, and outliers), and mean plots with standard errors for each emotion type to have a general idea of patterns.

3.1. Fundamental frequencies

The F0 results showed common patterns between the speakers of each speech type. We pooled the individual data and performed descriptive statistics for each F0 parameter (see Figure 1 for mean plots). The clear differentiation of means across the five emotions was observed for F0mean, F0max, and F0range in all speech types. To see if those mean differences were statistically significant, a significance test was conducted. Because the requirement of homogeneity of one-way ANOVA was not satisfied, the Kruskal–Wallis H Test, the non-parametric alternative, was performed with emotion type as the independent variable and each F0 parameter as the dependent variable for each speech type (set at 0.05 with alpha). The results showed significant effects for F0mean, F0max, and F0range for all speech types. This indicates that F0 plays a prominent role in emotional speech not only for L1 Japanese (JP) and L1 Italian (IT) but also for L2 Japanese-L1 Italian (ITJ).

The mean plots were further examined to find how different emotions were differentiated in each speech type. Several crosslinguistic differences emerged from the results. First, JP was much higher in F0mean than IT and ITJ. Various cross-linguistic studies indicate language-specific differences with respect to F0 that are socio-culturally motivated [15]; it has been reported that Japanese female speakers use a higher pitch than those of other languages such as American English, Dutch, and Swedish [16]. The F0 analysis of this study shows not only that their pitch level is also higher than one of the Italian female speakers for both L1 and L2.

The second difference was found for angry. In JP, no significant difference was found between neutral vs. angry for F0mean and F0max, although F0range was significantly wider for angry, according to the results of pairwise comparisons (Dunn–Bonferroni test) performed for each parameter ($p < .01$). In IT, however, angry was higher in F0mean and F0max with a wider F0range than neutral. The mean difference for neutral vs. angry was statistically significant for both parameters. The patterns of ITJ were very similar to those of IT. The last difference between the two L1s is observed for sad. In JP, the means of F0mean, F0max, and F0range tend to be lower, although there is no statistical difference between sad and neutral for these parameters. In IT, sad tends to be significantly higher in F0mean and F0max than neutral, although the mean difference for sad vs. neutral is not statistically significant.


(a) F0mean (IT = L1 Italian, ITJ = L2 Japanese, JP = L1 Japanese)

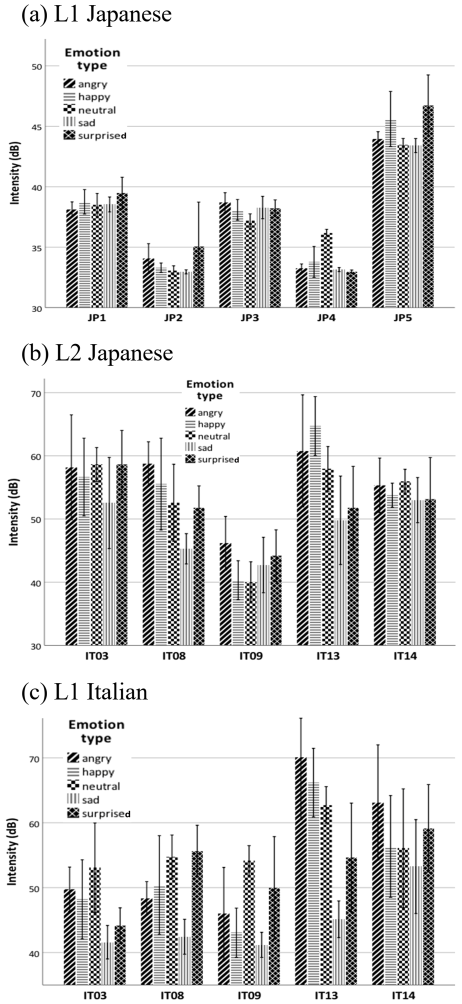
(b) F0max

(c) F0min

(d) F0range

Figure 1 (a-d): *Means and standard errors of the four F0 parameters plotted by group and emotion type.*

The above results show similarities between IT and ITJ as well as differences between those two and JP. For each emotion, a Kruskal–Wallis H test was performed with speech type as the independent variable and each F0 parameter as the dependent variable. There were significant effects for F0mean, F0max, and F0range (p < .001). For these three parameters, pairwise comparisons (Dunn–Bonferroni test) were performed to find which paired means were significantly different. For all parameters, there was no significant difference between IT vs. ITJ, but differences between IT vs. JP and ITJ vs. JP (p < .001) were significant. The results show the grouping of the learners' L1 and L2, which suggests strong transfer effects.


(a) L1 Japanese

(b) L2 Japanese

(c) L1 Italian

Figure 2 (a-c): *Means and standard errors of median intensity plotted by speaker and emotion type.*

For F0range, a mean distribution across the emotions is similar in ITJ and JP (Figure 1-d). Sad is smaller than neutral; surprised shows the largest mean; the means of angry and happy are more or less the same. It is not clear if the similarity of JP and ITJ for F0range is due to learning effects or effects of the use of the same Japanese stimuli.

3.2. Intensity

For intensity, the JP data showed great speaker variations with no particular group pattern. Most JP did not show a clear distinction of means across emotions (Figure 2-a). A Kruskal–Wallis H test was performed with emotion type as the independent variable and intensity as the dependent variable for each speaker. A significant effect was found only for JP4 (p < .01). The speakers of IT showed clearer mean distinctions (Figure 2-c). The same Kruskal–Wallis H test was performed for each IT speaker. There were significant effects for all speakers ($p < .05$) except for IT14. This suggests that intensity is more actively used in IT than JP. One distinctive group pattern of IT is that the mean is the smallest for sad. According to the results of a Dunn–Bonferroni test, the mean difference for sad vs. neutral is significant for those four speakers ($p < .05$). The ITJ patterns were somewhere between JP and IT (Figure 2-b): fewer speakers differentiated emotions with intensity. Two of the five learners—IT08 and IT13—showed significant effects of emotion type on intensity, according to the results of a Kruskal–Wallis H test.

3.3. Utterance duration

For utterance duration, the mean plot of the JP data (Figure 3-a) showed large individual variations and no particular group pattern. In contrast, some group patterns were found in the IT data (Figure 3-c). First, neutral was shorter in duration than the other four emotions. Second, all five speakers showed the greatest mean for happy (IT13) or surprised (all the others). These are considerable tendencies, although the results of Kruskal–Wallis H Test showed significant effects of emotion type on duration only for IT3 and IT8. The ITJ data (Figure 3-b) showed mixed patterns: IT09 and IT13 showed more native-like distributions of means, while the patterns of IT03 and IT14 were more similar to their own L1 patterns. The former and latter may indicate possible learning or transfer effects, respectively.

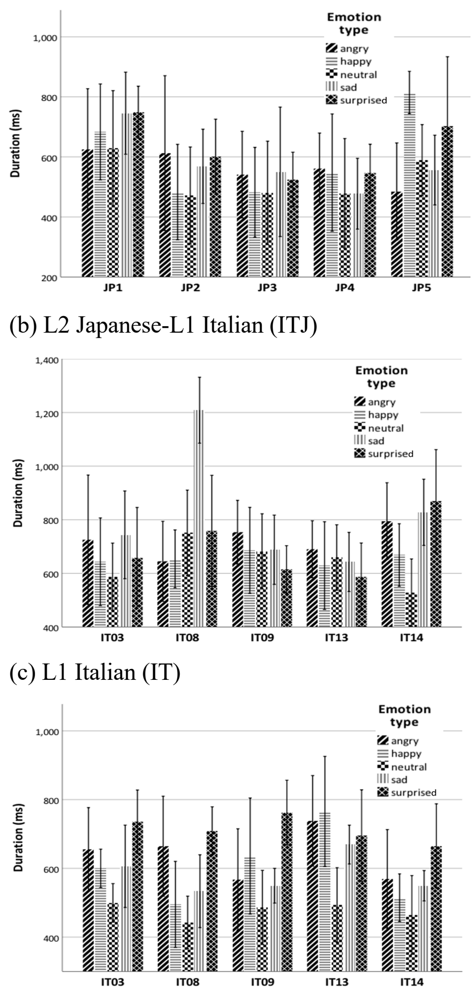
4. Discussion

The results show asymmetric transfer effects for different emotions, primarily for F0 parameters. For happy and surprised, all three speech types are characterized by two F0 modulations: higher pitch and wider pitch range with respect to neutral, which indicates *positive transfer*. In contrast, for angry and sad, different phonetic patterns are observed between L1 Japanese and the other two speech types, which may be due to *negative transfer.* For angry, in L1 Japanese, there is no significant difference between neutral vs. angry for F0mean and F0max, although F0range is significantly larger for angry than for neutral, while L1 Italian is significantly higher in F0mean and F0max with a larger F0range. For sad, L1 Japanese tends to be lower in pitch level and smaller in pitch range with respect to neutral, while L1 Italian is characterized by higher F0mean, F0max, and F0min, with significantly smaller intensity. For both angry and sad, L2 patterns are similar to those of L1 Italian for F0 patterns.

Interestingly, positive transfer is observed for so-called positive emotions (happy, surprised) while negative transfer is seen for negative emotions (angry, sad). In the field of psychology, basic emotions are classified in terms of positive or negative, i.e., values of *valence* that qualify the degree of attractiveness (positive valence) or aversiveness (negative valence) associated with an emotion [18]. [19] found cross-cultural similarities across participants from China, Korea, Canada, the USA in valence values of their perception of basic emotions. In all the cultures, each emotion is perceived as both positive and negative with a general tendency for the dominance of one of the two valance values. Happy and surprised are positive-dominant, while angry and sad are negative-dominant. It seems safe to assume that surprised is positive in this study since the data were elicited in a positive context. In [19], some cross-cultural differences are found as well. Similarities between Koreans and Chinese are stronger than between Americans and Canadians in some cases (e.g., sadness) but not in others (e.g., anger, where the Chinese were more similar to Western cultures than to Koreans). There are cross-cultural differences also in emotional *arousal*, i.e., the perceived intensity of an event from calm/low to excited/high. Such differences can result in negative transfer, as reported in [17] that investigated the emotional speech of L1 Japanese, L1 Chinese and L2 Japnese-L1 Chinese at the production level. They found phonetic evidence for a higher level of emotional arousal in L1 Chinese and L2 Japanese-L1 Chinese than in L1 Japanese, which indicates L1 transfer effects.

Considering all these findings of earlier research, it seems worthwhile to account for emotional valence and arousal systematically for a better understanding of L1 effects on non-native emotional speech, including emotional speech by Italian learners of Japanese. For future development, we are also planning to conduct a more in-depth acoustic analysis of the data analyzed in the present study by adding more phonetic parameters (e.g., voice quality, utterance-final lengthening, or boundary tone movements). A mixed model design should be considered an option for statistical analysis. Last but not least, we hope to eventually extend the study to the relation between production and perception.

(a) L1 Japanese (JP)



(b) L2 Japanese-L1 Italian (ITJ)

(c) L1 Italian (IT)

Figure 3 (a-c): *Means and standard errors of utterance duration plotted by speaker and emotion type.*

5. References

[1] D. Erickson, “Expressive speech: Production, perception and application to speech synthesis,” *Acoust. Sci. Technol.*, vol. 26, no. 4, pp. 317–325, 2005.

[2] S. Yildirim *et al.*, “An acoustic study of emotions expressed in speech,” in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, pp. 2193–2196.

[3] K. Maekawa, “Production and perception of ‘paralinguistic’ information,” in *Proc. Speech Prosody*, 2004, pp. 367–374.

[4] I.R. Murray and J.L. Arnott, “Toward to simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” in *J. Acoust. Soc. Am.*, vol. 93, no. 2, pp. 1097–1108, 1993.

[5] X. Li *et al*., “EGG analysis of Japanese emotional speech – production by native Japanese speakers and Chinese learners of Japanese,” (in Japanese) in *Proc. Autumn Meeting of the Acoustical Society of Japan,* 2019, pp. 3–4.

[6] L. Selinker, “Interlanguage,” in *Int. Rev. Appl. Ling.*, vol. 10, pp. 209-231, 1972.

[7] M. Ueyama, *Prosodic Transfer: An Acoustic Study of L2 English and L2 Japanese*. Bologna, Italy: Bononia University Press, 2012.

[8] P. Sorianello and A. De Marco, “Sulla realizzazione prosodica delle emozioni in italiano nativo e non nativo,” in *La Fonetica nell'Apprendimento Delle Lingue* (Phonetics and Language Learning), R. Savy and I. Alfano, Eds. Naples, Italy: AISV, pp. 155–177, 2016.

[9] R. Hayashi *et al*., “Development of sound archives for learning Japanese prosody,” (in Japanese) presented at the 2018 Int. Conf. on Japanese Learning (ICJLE) 2018, Venice, Italy, Aug. 4, 2018.

[10] M. Ueyama, A.E. Albin and R. Hayashi, “Development of an L2 Japanese speech corpus for the comparison of prosody across diverse L1 groups,” presented at the 2nd Phonetics and Phonology in Europa (PaPE)”, Lecce, Italy, June 17, 2019.

[11] J. Yue *et al*., “Automatic scoring of shadowing speech based on DNN posteriors and their DTW,” in *Proc. of Interspeech,* 2017, pp. 1422–1426. DOI: 10.21437/Interspeech.2017-728.

[12] P. Boersma and D. Weenink. “Praat: Doing Phonetics by Computer Version 6.1.08.” praat.org. http://www.praat.org/ (accessed Dec. 15, 2019).

[13] A.E. Albin. “Linguistic Analysis of Speech Fundamental Frequency.” rdrr.io https://rdrr.io/github/usagi5886/intonation/ (accessed Dec. 15, 2019).

[14] G. Fant and A. Kruckenberg, “A new approach to intonation analysis and synthesis of Swedish,” in *Fonetik, TMH-QPSR,* vol. 44, no. 1, 2002, pp. 161-164.

[15] B. Andreeva1 *et al*., “Comparison of Pitch Range and Pitch Variation in Slavic and Germanic Languages,” in *Proc. Speech Prosody*, 2014, pp. 776–780.

[16] A. Kemp, *Human Voice: The Story of a Remarkable Talent.* London, U.K.: Bloomsbury, 2011.

[17] X. Li, A. Albin, and R. Hayashi, “A relationship between production and perception of emotional speech by Chinese learners of Japanese,” (in Japanese), in *Proc. 32th Meeting Phonetic Soc. of Japan*, 2018, pp. 1–6.

[18] E. Kensinger, and D. L. Schacter, “Processing emotional pictures and words: effect of valence and arousal,” *Cogn. Affect. Behav. Neurosci. 6*, 110–126, 2006. doi: 10.3758/cabn.6.2.110 (accessed March 9, 2020)

[19] S. An, L. Xi, M. Marks, and Z. Zhang, “Two sides of emotion: exploring positivity and negativity in six basic emotions across cultures,” in *Frontiers in Psychology,* vol. 8, 2017. 10.3389/fpsyg.2017.00610 (accessed March 9, 2020)