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

Prosodic and Voice Quality Feature of Japanese Speech Conveying Attitudes: Mandarin Chinese Learners and Japanese Native Speakers

Xinyue Li[1,2], *Carlos Toshinori Ishi*[2], *Ryoko Hayashi*[1]

[1] Kobe University
[2] ATR Hiroshi Ishiguro Labs.
{lixinyue; carlos}@atr.jp, rhayashi@kobe-u.ac.jp

Abstract

To clarify the cross-linguistic differences in attitudinal speech and how L2 learners express attitudinal speech, in the present study Japanese speech representing four classes of attitudes was recorded: friendly/hostile, polite/rude, serious/joking and praising/blaming, elicited from Japanese native speakers and Mandarin Chinese learners of L2 Japanese. Accounting for language transfer, Mandarin Chinese speech was also recorded. Acoustic analyses including F0, duration and voice quality features revealed different patterns of utterances by Japanese native speakers and Mandarin Chinese learners. Analysis of sentence final tones also differentiate native speakers from L2 learners in the production of attitudinal speech. Furthermore, as for the word carrying sentential stress, open quotient-valued voice range profiles based on Electroglottography signals suggest that the attitudinal expression of Mandarin Chinese learners are affected by their mother tongue.

Index Terms: attitude, second language acquisition, EGG, paralinguistic information

1. Introduction

Discussions on paralinguistic and non-linguistic information in speech, including attitudinal speech, have been widely arising not only in phonetic studies but also in psychology, speech recognition, and robotics [1]. Several studies have shown that the precise phonetic realization and production of attitudes are language-dependent [3, 4, 5]. One source of evidence for this claim is the cross-linguistic transfer observed in second language (L2) learners. In particular, L2 learners have been shown to pattern differently from native speakers in the perception of attitudinal speech. For example, Tang [5] recorded Chinese speech conveying six classes of attitudes like dominant/submissive, and asked Chinese listeners, L1 French L2 Chinese listeners, L1 Japanese L2 Chinese listeners, French native listeners and Japanese native listeners without any L2 learning experience of Chinese, to identify the attitudes they listened. The results indicated that all groups of subjects except native Chinese listeners showed perceptual difficulties of attitudes such as sincere/insincere.

In previous research, prosodic features have been closely linked to the attitudinal speech. For instance, in Mandarin Chinese, friendly utterances have a longer syllabic duration and narrower F0 range when compared with hostile utterances, while praising utterances have the shorter syllabic duration and higher F0 mean than blaming utterances [6]. It is also well known that voice quality features are crucial to the identification of paralinguistic information, including attitudinal speech [7, 8], e.g., distinguishing tense voice or lax voice using spectral analysis [7] and breathy voice using an aspiration noise parameter like F1F3syn [8].

Moreover, prosodic analyses of sentence final syllables have been found to be relevant for the discrimination of different paralinguistic types in the previous studies. For example, analysis of the F0 rise of phrase-finals in Japanese question speech conveying paralinguistic information [9] showed that huge F0 rise (long and large F0range) in phrase-final characterizes suspicion/disbelief. In addition, not only the sentence level prosodic and voice quality features are essential to attitudinal speech, but also the acoustic details in the words carrying sentential stress play important roles [6, 7]. For example, wider F0range and longer duration characterizing stressed words have been found to convey more paralinguistic information than unstressed words [13]. Furthermore, it is well known that glottal open quotient [11] (hereinafter, OQ) based on EGG signal is one glottal source measurement of voice quality that is useful to investigate voice quality, such as breathy voice and pressed voice [12], which is crucial to paralinguistic expressions [2, 8]. Hence, when recording the speech material in the present study, the EGG signal was also recorded for all speakers.

At present, however, relatively little is known about the production of attitudinal speech by L2 learners. The present study seeks to fill this gap by documenting what prosodic and voice quality features are crucial to different attitude types and which acoustic cues differentiate native speakers from L2 learners in the production of attitudinal speech, using L1 Mandarin L2 Japanese as a case study.

2. Speech Material

2.1. Recording data

As reported in Gu [6], different from emotions, attitudes can usually be bipolar. The present study defined one pair of attitudes with two opposite poles. Following [6], four pairs of attitudes that are easy to express in daily communication, which are called behavioral attitudes (toward the listener but not toward the content of utterance), were examined.

Pair 1: *friendly* vs. *hostile*

Pair 2: *polite* vs. *rude*

Pair 3: *serious* vs. *joking*

Pair 4: *praising* vs. *blaming*

Each attitude was designed with two target sentences, including declarative sentence and question sentence, literally neutral. To obtain natural and real attitudinal expression, we provided vignette situations and dialogue scenarios for each sentence. Neutral attitudes serving as baseline were also

10.21437/SpeechProsody.2020-9

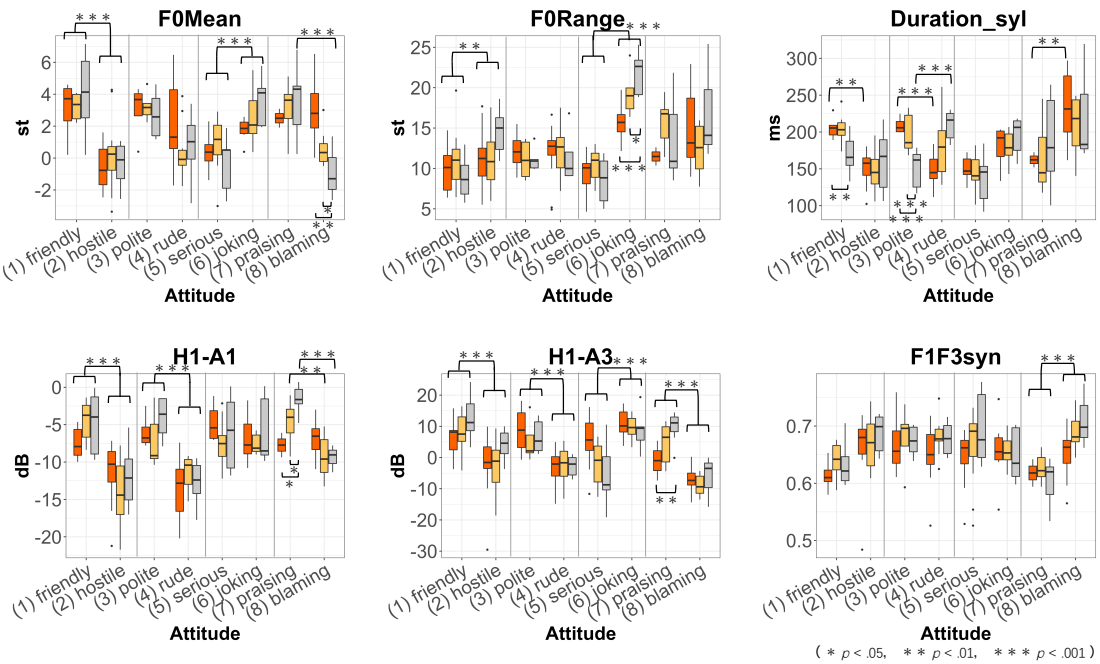

Figure 1: *Distributions of prosodic and voice quality features related to attitudes in all languages*.

(red: Mandarin Chinese by native; yellow: Japanese by Mandarin Chinese learner; grey: Japanese by native)

recorded for each target sentence. Thus, in each of the four pairs of attitudes, three patterns were recorded for each target sentence: the two opposite attitudes and the neutral one.

Eight Japanese native speakers and eight Chinese learners of L2 Japanese speakers were recruited to produce attitudinal speech in Japanese. Each group had 4 males and 4 females, with an overall average age of 29.6 (*SD* = 4.83) across all 16 speakers. All Chinese learners spoke Mandarin as their native language and had passed the highest level ("N1") of the standardized Japanese Language Proficiency Test (JLPT), currently living in Japan. Chinese learners also produced utterances with the same literal meaning in Mandarin Chinese. In total, 576 attitudinal utterances were recorded (384 Japanese tokens and 192 Mandarin Chinese tokens).

2.2. Perceptual experiment

Four Japanese native evaluators performed a perceptual experiment to examine if the intended attitudes can be correctly recognized. The evaluators were asked to answer two questions:

a. perceived attitude score, e.g., for *friendly*/*hostile*: *very hostile* (-3), *hostile* (-2), *slightly hostile* (-1), *neither* (0), *slightly friendly* (1), *friendly* (2), *very friendly* (3)

b. the word carrying sentential stress, if any

Firstly, Cronbach's alpha coefficient for each pair of attitudes was examined to describe the reliability of the perceived attitude scores in the perceptual experiment. The value of *friendly*/*hostile* was .866, *polite*/*rude* was .863, *serious*/*joking* was .881, and *praising*/*blaming* was .948. Hence, the values of all attitudes are higher than .8, and we regard internal consistency among all evaluators of the perceptual experiment as acceptable results. We then calculated the differences of perceived attitude scores and excluded the utterances if the differences were above two points (e.g., between *friendly* and *very friendly*, there is 1 point difference). Furthermore, for the remaining utterances, if polarity of the scores were judged reversed (i.e., *friendly* (+) utterance was misjudged as *hostile* (-)), the corresponding utterances were also excluded. Consequently, 502 well conveyed attitudinal tokens (*friendly*/*hostile*: 118 tokens, *polite*/*rude*: 120 tokens, *serious*/*joking*: 137 tokens, *praising*/*blaming*: 127 tokens) were used for the subsequent acoustic analysis.

3. Acoustic Analysis

The following six measurements were extracted from each token to describe the prosodic and voice quality features of Japanese and Mandarin Chinese speech: F0mean, F0range, mean syllabic duration, H1-A1, H1-A3, and F1F3syn. To control between-speaker differences in F0 usage, we normalized the F0 values for each speaker by converting the F0 measurements into semitones and subtracting each speaker's mean value. H1-A1 (difference between the spectral amplitudes of the first harmonic and the first formant, in dB) and H1-A3 (difference between the amplitudes of the first harmonic and the third formant, in dB) were measured for the spectral analysis. And F1F3syn (synchronization of the waveform amplitude envelopes of the first and third formant frequency bands) reflecting aspiration noise components was also measured to differentiate breathy voice from non-breathy voice. F1F3syn values closed to 0 indicate breathiness, while F1F3syn values closed to 1 indicate non-breathiness.

The results of acoustic analyses across each target sentence are grouped by attitude and language, as shown in Figure 1 (grey: Japanese by native speakers, yellow: Japanese by Mandarin Chinese learners, red: Mandarin Chinese). Two-way ANOVA (language × attitude) between each attitude within each pair of attitudes was conducted. Significant differences between the two opposite attitudes are shown at the upper side, while significant differences between the three language groups are shown at the lower side of each graph in Figure 1. The quantitative results of acoustic analyses are summarized

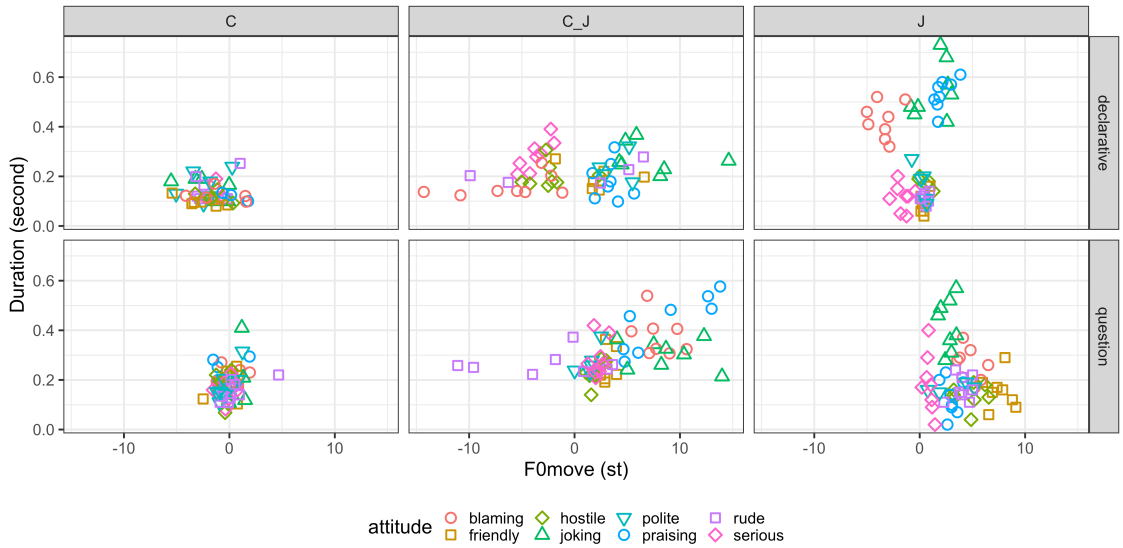

Figure 2: *Distributions of sentence final tones related to attitudes in all languages*.

(C: Mandarin Chinese by native; C_J: Japanese by Mandarin Chinese learner; J: Japanese by native)

below.

For *friendly*/*hostile*, *friendly* has higher F0mean ($p < .001$) and narrower F0range than *hostile* ($p = .007$). Regarding voice quality features, H1-A1 and H1-A3 in *friendly* are higher than those in *hostile* (*p*s < .001). Moreover, *hostile* is faster than *polite* in Mandarin Chinese ($p = .008$). When it comes to *friendly*, Japanese utterances by natives are faster than those of Mandarin Chinese utterances ($p = .002$). For *polite*/*rude*, *polite* has higher H1-A1 and H1-A3 than *rude* (*p*s < .001). Regarding the results for syllable duration, *rude* is slower than *polite* in Japanese by native speakers ($p < .001$). Conversely, *rude* is faster than *polite* in Mandarin Chinese by native ($p < .001$). In addition, *polite* from Japanese by native is faster than Japanese by Mandarin Chinese and Mandarin Chinese by native (*p*s < .001). For *serious*/*joking*, *serious* has lower F0mean and narrower F0range than *joking* (*p*s < .001). Regarding voice quality, H1-A3 is lower in *serious* than in *joking* ($p < .001$). Moreover, F0range of *joking by Japanese* native is larger than Japanese by Mandarin Chinese learners and Mandarin Chinese utterances ($p = .041$, $p < .001$). For *praising*/*blaming*, H1-A3 in *praising* is higher than in *blaming*, and F1F3syn in *praising* is lower than in *blaming* (*p*s < .001). More specifically, *praising* has higher F0mean than *blaming* in Japanese by native speakers ($p < .001$). In Mandarin Chinese speech, *praising* is faster than *blaming* ($p = .002$). In addition, Japanese native's *praising* is more lax than Mandarin Chinese utterance as indicated by the differences in H1-A3 ($p = .003$). And Japanese native's *blaming* has lower F0mean than the others ($p = .016$, $p = .003$).

4. Analysis of Sentence Final Tones

Following [8, 10], the present study utilized F0move (the difference between two equally long segments broken from the last syllable of each target sentence) to quantify the amount and direction (positive for rising and negative for falling) of F0 movement within the last syllable. Figure 2 shows the distribution of F0move and duration in the last syllable of target sentences (question sentence and declarative sentence) for each attitude and language. The results in Figure 2 indicate that Japanese attitudinal speech can be classified by sentence final tone-types, but not for Mandarin Chinese speech.

For Japanese declarative sentences by native speakers, we can observe that *praising* shows small rising tones (positive F0move) and relatively longer duration (around .5s); *joking* shows long duration and rising tones partially; *blaming* shows falling tones (negative F0move) and intermediately long duration; *Serious* shows slightly falling tones but short duration; the other attitudes tend to be flat tones (F0move around 0). On the other hand, Japanese declarative sentences by Mandarin Chinese learners show sharper rising or falling tones than Japanese natives in the last syllable. Deep rising tone in short duration (around .2s) is found in *joking*; *praising* and *polite* also show rising tones and comparatively short duration; while deep falling tones (some F0move values are under -5st) and short duration are observed in *blaming*; *serious* and *hostile* also show falling tones; both rising and falling tones are found in *friendly* and *rude*.

For Japanese question sentences by native speakers, rising tones within 10st and flat tones are observed in all attitudes. Comparatively deep rising tones are found in *friendly* (all above 5st); while *serious* show flat tones. Conversely, for Japanese question sentences by Mandarin Chinese learners, falling tones are found in *rude*, and larger rising tones (large F0move and long duration) are found in *praising*, *joking* and *blaming*.

5. Analysis of OQ in Stressed Words

In the present study, besides the Japanese listeners in the perceptual experiment described in Section 2, four L1-Mandarin Chinese of L2 Japanese learners judged the words carrying sentential stress of Mandarin Chinese utterances. The stressed words were deemed when more than 3 listeners agreed. As a result, we leveraged Japanese words that clearly conveyed the attitudes. For *friendly*/*hostile*: "raishuu (next week)"; for *polite*/*rude*: "tantousha (the person in charge)"; for *serious*/*joking*: "ashita (tomorrow)"; and for *praising*/*blaming*: "sugoi (wonderful)" were selected from one Japanese female and one Mandarin Chinese learner female. Mandarin Chinese words having the same meaning were also analyzed. Based on

EGG signals, OQ within the stressed word was calculated. Then, OQ-valued voice range profile (VRP) [14] of the words carrying sentential stress is utilized to clarify the dynamic properties of voice quality among all attitudes and languages, as shown in Figures 3 to 5. X-axis and y-axis represent F0 (in 2 semitone intervals, normalized by subtracting each speaker's mean value) and power (in 5dB intervals), color represents OQ (the smaller OQ, the deeper red color, indicating the tenser voice; the larger OQ, the lighter green color, indicating the laxer voice).

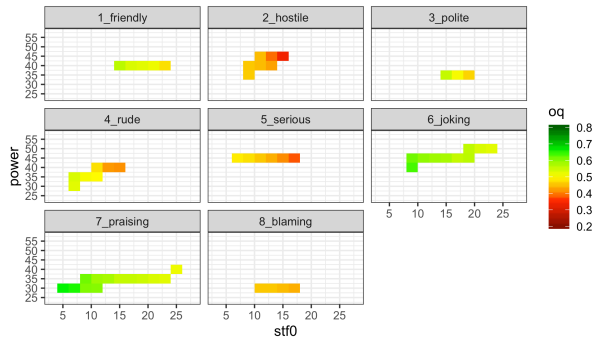

Figure 3: *Japanese attitudinal speech by Japanese native speaker: OQ-valued range profile.*

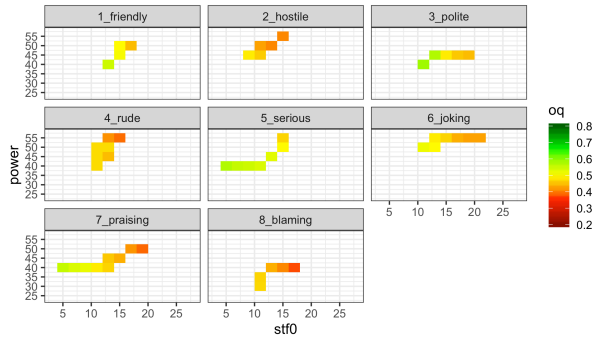

Figure 4: *Japanese attitudinal speech by Mandarin Chinese learner: OQ-valued range profile.*

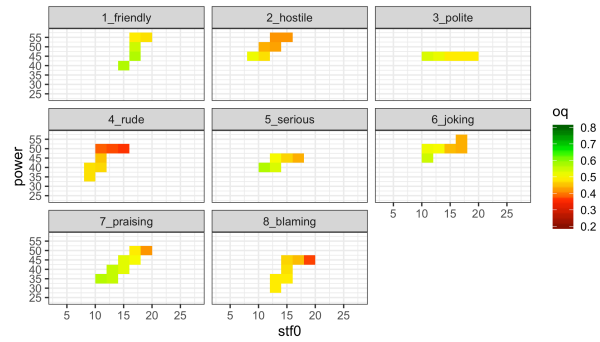

Figure 5: *Chinese attitudinal speech by Mandarin Chinese speaker: OQ-valued range profile.*

Overall, OQ decreases when F0 and power become bigger in all languages and attitudes. This is consistent with the results of [15]. For Japanese produced by native speakers in *joking* and *praising*, lax voice has been found (OQ: the minimum value is above .55, presented in green and light yellow color), contrary to that of Mandarin Chinese and Mandarin Chinese utterances, which have smaller OQ (wider orange color zone), suggesting a tenser voice. And in *serious*, OQ in whole utterance from Japanese native is smaller than that of Mandarin Chinese learners and Chinese speech. For Figure 4 and Figure 5, OQ of *rude* shows extremely small values (the red color), but not observed in Japanese by native speaker.

6. Discussion and Conclusion

First, based on the acoustic analysis, we can observe that prosodic features, such as F0mean and F0range are distinctive for *friendly*/*hostile* and *serious*/*joking*, while voice quality features are distinctive in all attitude pairs. The results of global F0 and duration features are consistent with the results of [6], where Mandarin Chinese produced by native speakers were discussed, suggesting that Japanese and Chinese attitudinal speech have some prosodic features in common. Specifically, *rude* is faster than *polite*, *serious* has lower F0mean and narrower F0range than *joking,* and *praising* is faster than *blaming*. Furthermore, the results of voice quality showed that *friendly, polite, joking and praising* utterances are closer to lax voice. Conversely, *hostile, rude, serious* and *blaming* utterance are closer to tense voice. In addition, *praising* is breathier than *blaming*.

Second, analyses of sentence final tone-type revealed different patterns between native speakers and L2 learners in the production of attitudinal expression, e.g., for Japanese question sentences by Mandarin Chinese learners, falling tones are found in *rude*, and larger rising tones are found in *praising*, *joking* and *blaming*, in comparison to native speakers. These results indicated the final tone-type intonation instruction for L2 learning is also important for attitude expression. It is also interesting to note that the tone patterns were quite different from the phrase final tones in Chinese sentences, suggesting that F0 movement in the last syllable do not contribute much to the attitudinal expression of Mandarin Chinese. In addition, a large overlap in attitudes indicates analyses should be conducted for further investigation.

Third, analysis of EGG in stressed words showed that OQ decreases when F0 and power become larger, fitting with [15], where OQ is reported to be strongly related to F0 and power in speech, sung sentence and shouting utterance etc. from eighteen singers. And the results of OQ-valued VRP indicated that Mandarin Chinese learners tend to produce *joking* and *praising* with tenser voice in Japanese and Chinese, whereas the same cannot be said for Japanese speakers, suggesting that Chinese learners of Japanese transfer their native glottal source cues when they produce attitudinal speech in Japanese. This is a piece of evidence that suggests that L2 transfer exists in the production of speech conveying attitudes.

Future works include the investigation of the acoustic differences between correctly judged Japanese utterances and misjudged Japanese utterances from Mandarin Chinese learners of L2 Japanese, to contribute to the speech communication instruction for the L2 learners.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 17H02352 and JST ERATO Grant Number JPMJER1401, Japan.

8. References

[1] H. Fujisaki, "Prosody, models, and spontaneous speech," In: Y. Sagisaka, N. Campbell, and N. Higuchi, (eds.), *Computing Prosody*, pp. 27–42, New York: Springer-Verlag, 1996.

[2] H. Mori, K. Maekawa, and H. Kasuya, *What Does Speech Convey? Speech Science of Emotion, Paralinguistic Information, and Speaker Individuality*, Tokyo: Corona Publishing, 2014.

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

[4] D. Erickson, "Expressive Speech: Production, perception and application to speech synthesis." *Acoust. Sci. & Tech*. vol. 26, 317–325, 2005.

[5] P. Tang, "Cross-linguistic perceptual experiment and acoustic analysis of Chinese attitudinal speech," MD Thesis, Nanjing Normal University. (In Chinese with English abstract)

[6] W. Gu, T. Zhang, and H. Fujisaki, "Prosodic Analysis and Perception of Mandarin Utterances Conveying Attitudes," *Proceedings of INTERSPEECH*, pp. 1069–1072, 2011.

[7] C. Menezes and K. Maekawa, "Paralinguistic effects on voice quality: a study in Japanese," *Proceedings of Speech Prosody*, pp. 656–659, 2006.

[8] C. T. Ishi, H. Ishiguro, and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality," *Speech Communication*, vol. 50, no. 6, pp.531–543, June 2008.

[9] K. Maekawa, "Production and perception of 'Paralinguistic' information," *Proceedings of Speech Prosody*, pp. 367–374, 2004.

[10] C. T. Ishi, "Perceptually-related F0 parameters for automatic classification of phrase final tones," *IEICE Trans. Inf. & Syst.*, Vol. E88-D, No. 3, 481-488, 2005.

[11] R. Timcke, H. von Leden, and P. Moore, "Laryngeal vibrations: Measurements of the glottis wave," *AMA Archives of Otolaryngology*, vol. 68, no. 1, pp. 1–19, 1958.

[12] D. Klatt, and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers,'' *J. Acoust. Soc. Am*, vol. 87, pp. 820–857, 1990.

[13] S. Zhang, P. C. Ching, and F. Kong, "Acoustic analysis of emotional speech in Mandarin Chinese," *Proceedings of ISCSLP*, pp. 57–66, 2006.

[14] K. Wakasa, H. Terasawa, H. Kawahara, and K. Sakakibara, "Comparison between operatic singing and choral sing by physiological and acoustic feature quantity analysis," *Proceedings of Acoust. Soc. Ja*, pp. 1121–1124, 2019.

[15] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *J. Acoust. Soc. Am*, vol. 117, no. 3, pp. 1417–1430, 2005.