SMM19, Workshop on Speech, Music and Mind 2019
14 September 2019, Vienna, Austria

Prosodic and voice quality analyses of loud speech: differences of hot anger and far-directed speech

Carlos T. Ishi[1], *Takayuki Kanda*[2]

[1]ATR Hiroshi Ishiguro Labs.
[2]ATR Intelligent Robotics and Communication Labs.
carlos@atr.jp, kanda@atr.jp

Abstract

In this study, we analyzed the differences in acoustic-prosodic and voice quality features of loud speech in two situations: hot anger (aggressive/frenzy speech) and far-directed speech (i.e., speech addressed to a person in a far distance). Analysis results indicated that both types are accompanied by louder power and higher pitch, while differences were observed in the intonation: far-directed voices tend to have large power and high pitch over the whole utterance, while angry speech has more pitch movements in a larger pitch range. Regarding voice quality, both types tend to be tenser (higher vocal effort), but angry speech tends to be more pressed, with local appearance of harsh voices (with irregularities in the vocal fold vibrations).

Index Terms: loud speech, hot anger, prosody, voice quality, paralinguistics.

1. Introduction

Loud speech may appear in different intentional, attitudinal or emotional expressions in speech communication. For example, we naturally raise our voice when speaking to a person in a far distance, for example when greeting or calling, or to a group of people, like in a classroom. It is also quite often to raise the voice when we are extremely angry, the so-called hot anger. However, the way we raise our voices in such different situations is different.

Speakers alter how they produce speech based on the communicative situation, so that changes are made to enhance the information transmission's efficiency. It is also known that in noisy environments, people speak loudly and produce more energy at higher frequencies (the Lombard effect [1]).

The ability to alter speech intensity with changes in listener distance is an important aspect of natural communication. On theory, speech intensity obeys an inverse square law with distance [2]. That is, when the distance between the speakers is doubled, there is a corresponding 6dB reduction in the speech volume due to sound propagation losses. It has been reported that speakers make prosodic, pragmatic, and semantic changes in addition to increasing speech volume to accommodate changes in listener distance [3]. These compensatory changes closely resemble the Lombard effects, which explain that speech intensity is adjusted to compensate for increases in the background noise. Other studies have found that loud speech is also associated with a reduction in speech rate [4].

On the other hand, loud speech may appear in situations where people sometimes behave aggressively and in an offensive way in daily contexts. The understanding of such situations is important for human-computer or human-robot interactions. For example, in stores, there are sometimes complainers who make unreasonable complaints toward store workers. Since it is a stressful situation, people wish a robot to manage such troublesome complaints [5]. In schools, bullying is a serious problem [6], and it is also known that people sometimes bully toward other entities like animals [7]. There is a discussion that such cruelty would turn into interpersonal violence [8]. Offensive behavior is even exhibited toward inanimate entities, like a robot [9],[10]. Commonly to these problems, there would be a benefit if there is a technique to identify such offensive situations.

Direct aggressive speaking sounds hard, hostile, and often comes across as controlling or dominating (from *The Pup Safe Project* [11]). It has no subtlety. A directly aggressive person will do one or more of these things: 1) They raise their voices, get louder as they try to scare you. The aggressive speaker often has threatening body language as well, such as finger pointing and clenched fists, to looming over you; 2) They order to do what they say, demand listening to them, follow their instructions, and insult when you don't comply as they want; 3) They argue like it's a battle to be won. Their way to "win" is to talk over you, attack verbally, and not listening.

In the speech research field, there are numerous studies on acoustic-prosodic features of emotional speech [12-14], but there are no studies contrasting the differences of loud voices in different situations like hot anger and attention drawing.

In this study, we investigated what are the factors involved in the different impressions of loud voices uttered in different situations. For that purpose, we take into account previous studies on acoustic-prosodic and voice quality features [15-20], and analyzed the differences in acoustic-prosodic and voice quality features of loud speech in two situations: hot anger (aggressive/frenzy speech) and far-directed speech (i.e., speech addressed to a person in a far distance).

2. Data

2.1. Data description

In this study, we used a subset of a database collected for analyses of offensive speech [21]. The dataset is composed by short sentences uttered by multiple male and female speakers in different expression types. Headset microphones (DPA 4060) are used to collect audio data.

The offensive-related sentences are uttered in four expression types (the original words in Japanese are in parentheses): "reading out" ("yomiage": read out without emotions/temper), "aggressive" ("bougen": offensive, threatening, aggressive expression), "frenzy" ("kyouran": extreme expression of aggressiveness), and "joking" ("joudan": non-aggressive, non-serious joking/kidding/playful expression).

10.21437/SMM.2019-1

Among the several types of offensive-related utterances (insulting/defame, fooling, offensive, threatening, order, denial), we chose the order/scolding utterances below, which commonly appear in daily troublesome situations, and are thought to be easier for expressing anger. The English translations in parentheses are the equivalent expressions, but it may differ in nuance.

- Order/scolding utterances: “hayaku shi’ro” / “hayaku shiro tsuttendayo” (do it now!), “oriro” (get down!), “ayamare” / “dogeza shiro” (apologize!), “sassato ike” / “acchi ni itte” / “acchi ike” (get out!), “shabere” / “ie” (talk!), “damare” / “shizuka ni shi’ro” (shut up!).

The database also partly includes sentences which are not offensive-related, which are spoken in the following three expression types: “reading out”, “close” (speaking to a person in a close distance, e.g. 1~2 meters), and “far” (speaking to a person in a far distance, e.g. more than 5 meters). The far-directed utterances are expected to be loud, but not necessarily convey a patent attitude or emotion. The non-offensive sentences are:

- Far-directed utterances: “konnichiwa” (hello), “ohayoo” / “ohayoogozaimasu” (good morning), “arigatoo” / “arigatoogozaimasu” (thank you), “otsukaresamadesu” (“have a good night”), “baibai” / “matane” (see you), “ooi” (hey!), “tasukete” (help!), “kajida” (fire!).

For the analysis of the present study, we focused on the utterances which were spoken by loud voices. These include the frenzy utterances (which are closely related to hot anger), and the far-directed utterances (which are loud but not angry).

The database contains utterances from 10 male and 10 female speakers aged 20s to 60s. From those, far-directed utterances are available for 7 male and 4 female speakers.

Some of the speakers self-reported that they do not tend to utter aggressive speech in daily-life, and could not express well the aggressive and frenzy types. These were removed from the subsequent analyses. The IDs and ages of the remaining speakers are F02 (22), F03 (49), F05 (31), F07 (20), F08 (48), F09 (49), F10 (52) for the female speakers, and M02 (66), M03 (61), M05 (22), M06 (49), M07 (22), M08 (44), M09 (20) for the male speakers.

2.2. Data annotation

Considering that some speakers could not well express some of the targeted types, and the degree of expressivity differs for different speakers, perceptual impressions were annotated accounting for several factors. Two research assistants were asked to listen to each utterance and grade their perceptual impressions on the expressivity of the different types.

The following items were annotated in both order/scolding and far-directed utterances.

- Perceived degree of anger (emotional attitude): -3 (very jokey) ~ 0 (not jokey nor angry) ~ 3 (very angry/irritated)
- Perceived sense of distance the utterance is addressed to: -3 (very close: few cm) ~ 0 (1~2m) ~ 3 (very distant: more than 30m).
- Perceived degree of screaming: 0 (not screaming), 1 (shouting), 2 (screaming), 3 (loudly screaming).
- Perceived degree of pressing the larynx: 0 (not pressing) ~ 3 (strongly pressing).

The agreement rates (including exact matching and difference of 1 point) between the perceptual scores of the annotators were 95% for angry/jokey, 96% for the sense of distance, 92% for screaming, and 72% for pressing.

Fig. 1 shows the perceived degrees of distance (close to far), anger (jokey to angry), screaming and pressing, for frenzy and far-directed utterances, for different speakers. Note that far-directed data is absent for the speakers F02, F03, F05 and M02. T-tests showed significant differences between the average scores of frenzy and far-directed utterances for all perceptual factors: $p < 0.001$ for anger, distance and pressing, and $p < 0.05$ for screaming.

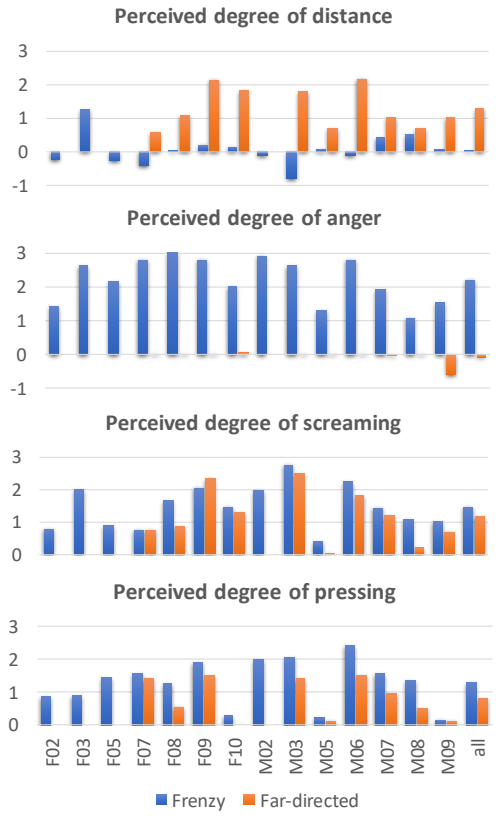

Figure 1: *Perceived degrees of distance (close to far), anger (jokey to angry), screaming and pressing, for frenzy and far-directed utterances, for different speakers. (Note: no far-directed data for speakers F02, F03, F05 and M02.)*

The results in Fig. 1 indicate that the frenzy utterances received high perceptual scores of anger, high scores of screaming, and low scores of sense of distance, for most of speakers. On the other hand, the far-directed utterances received high perceptual scores of sense of distance, high scores of screaming, and low scores of anger. The correlations between screaming and sense of distance were 0.97 for far-directed utterances and -0.02 for frenzy utterances.

From the results above, we can consider the frenzy utterances to be equivalent to angry speech, and contrast them with the far-directed utterances.

Regarding the perceptual scores of pressing, frenzy/angry voices are perceived to be slightly more pressed than far-directed utterances.

3. Prosodic and voice quality analysis

Acoustic analyses were conducted on prosodic and voice quality related features. First, frame-level acoustic features were extracted at 10ms intervals. For the pitch-related parameters, F0 values are estimated by a conventional autocorrelation-based method. All F0 values are then converted to a musical (log) scale before subsequent processing [15]. All utterances were manually segmented, and utterance-level acoustic parameters were extracted for each utterance. Voice quality changes were also manually segmented by an expert (the first author).

3.1. Descriptive analyses of F0 and voice quality features

In this section, we picked up utterance samples from some of the speakers and discuss differences in the expression types of loud voices in two situations: frenzy/angry and far-directed utterances.

Fig. 2 shows examples of pitch contours and harsh/pressed segments for frenzy/angry and far-directed utterances, by three speakers which received high perceptual scores of anger degree.

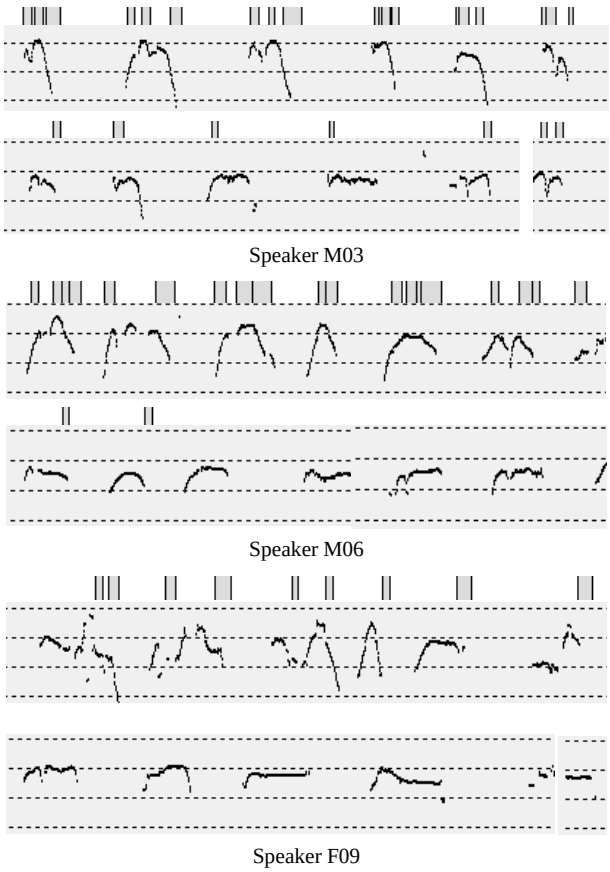

Figure 2: *Examples of pitch contours (octave grid intervals from 55~880 Hz) for frenzy/angry and far-directed utterances. Harsh/pressed segments are shown above the pitch panels. For each speaker, upper panels show frenzy/angry utterances, while bottom panels show far-directed utterances. The duration is about 10 seconds.*

A clear difference observed in Fig. 2 is that there are less up-down pitch movements (the pitch contours are predominantly high and flat) in far-directed voices. Another difference is the presence of harsh/pressed segments (with high energy) in angry voices. Some harsh/pressed segments were also observed in far-directed speech, but these were usually accompanied by low energy in the end of the utterances. Harsh/pressed voices were more frequent in frenzy/angry utterances which received higher scores of anger in Fig. 1.

We then looked closer to the pitch patterns in frenzy/angry utterances. Fig. 3 shows different prosodic realizations of the imperative utterance "sassato ike" ("get out quickly") for different speakers. From the upper panel, waveform, utterance segmentation, perceived harsh/pressed segments, F0 contour (musical scale, grids in octave intervals), phonetic information, power contour (grids in 10dB intervals), and spectrogram (grids in 2kHz intervals).

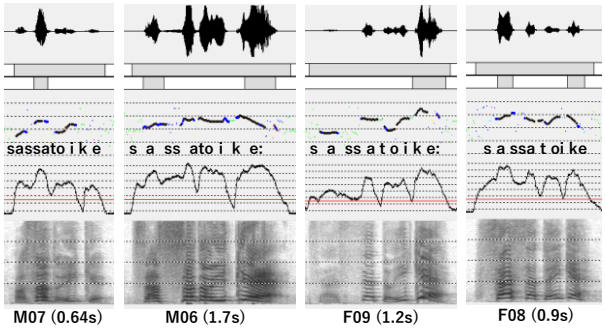

Figure 3: *Samples of frenzy/angry utterances by four speakers. From top: waveform, utterance segment, harsh/pressed voice segments, pitch contour (octave grid intervals from 55~880 Hz), phoneme sequence, power contour (10dB grid intervals), spectrogram (0~8kHz), speaker ID and utterance durations.*

As shown in these examples, some speakers emphasize the word "sassato" ("quickly") (like the speakers M07 and F08), while others emphasize the word "ike" ("go"). Also, the emphasized syllable can differ for the same word. For example, some speakers emphasize the first syllable in "sassato" (like the speaker F08), while others emphasize the second. The emphasized syllable may also differ depending on the dialect background of the speaker. In the collected data, most of speakers are from the Kansai area in Japan.

Regarding voice quality, the examples in Fig. 3 show that harsh/pressed voice qualities frequently appear in hot anger, but not necessarily over the whole utterance. Rather, they usually occur locally, and are related to the prominence of specific words in a phrase. In most of speakers, the last syllable of the phrase is emphasized. In some of the utterances of some speakers, words in the beginning or in the middle of the phrase are emphasized, so that the position of harshness within an utterance seems to depend on accent and prominence.

From the examples in Fig. 3, it can also be observed that some speakers utter fast (M07 and F08), while others utter slowly, in frenzy/angry type (M06 and F09). Thus, speech rate does not seem to be a discriminative factor of anger, and is thought to be a speaker-dependent factor.

3.2. Utterance-level acoustic analysis

Considering the analysis results presented in the previous subsection, acoustic parameters related to f0, power, spectral and aperiodicity features were computed at utterance level.

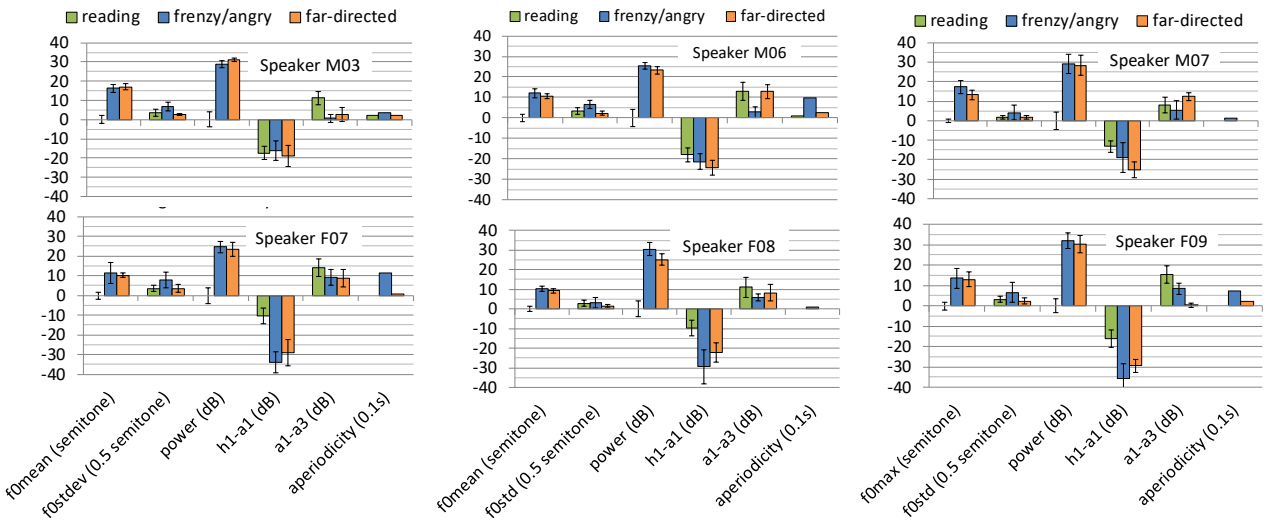

Figure 4: *Distributions of six acoustic-prosodic features for different types, for four speakers (M03, M06, F07, F09). The units of the vertical axis are shown in parentheses for each feature in the horizontal axis. Mean and standard deviations are shown for f0, power and spectral features, while total durations are shown for aperiodicity.*

- "*f0mean*" and "*f0std*" are the mean and the standard deviation F0 in the utterance. In Fig. 4, the *f0mean* values are normalized (subtracted) by the average *f0mean* of the reading out type. The max F0 values were also calculated, but the results are omitted from the figure since similar trends have been found with f0mean. The *f0std* values in Fig. 4 are doubled for allowing better visualization.
- "*power*" is the maximum power value in the utterance, in dB. In Fig. 4, the power values are normalized (subtracted) by the power value of the reading out type.
- "*h1-a1*" is the difference of the power of the first harmonic and the power around the first formant, specifically in the range between 200 to 1200 Hz, and is given in dB. This parameter is related to the vocal tension, and is correlated with pressed voice [17]. Large negative values indicate tenser voice quality.
- "*a1-a3*" is the difference of the power around the first formant (200 to 1200 Hz) and the power around the third formant (1500 to 4000 Hz), and is also given in dB. This parameter provides an estimate of the spectral tilt, and is also related to the vocal tension [18]. Smaller values indicate tenser voice quality.
- "*aperiodicity*" is the total length of vocalic segments detected as aperiodic (i.e., when auto-correlation peaks are lower than 0.5 in the F0 estimation), in seconds. In order to reflect aperiodicity caused by harsh voices, the aperiodic segments are disregarded if vocal fry is detected [19]. The values in Fig. 4 are scaled by 10 times to allow better visualization.

Fig. 4 shows the distributions for six utterance-level acoustic parameters described above, for four selected speakers (3 male and 3 female) which showed higher perceived scores of anger and distance (in Fig. 1). Pairwise t-tests (Welch t-tests) were conducted for checking statistical significances between different conditions.

It can be observed in Fig. 4 that both frenzy/angry and far-directed speech have higher pitch (about 1 octave (12 semitones) higher), stronger power (about 20~30dB higher) in comparison to reading out speech (t-test, $p < 0.001$ for *f0mean* and *power*). Tenser voice quality (lower *h1-a1* and *a1-a3* values) was observed in all speakers, female speakers having lower *h1-a1* values, while male speakers having lower *a1-a3* values (t-test, $p < 0.001$). The far-directed utterances by speaker M06 showed no differences with reading out type for *a1-a3* values, but lower *h1-a1* values (t-test, $p < 0.001$).

Differences between frenzy and far-directed speech can be observed in *f0std* and *aperiodicity*. Frenzy/angry speech shows higher f0 variation (larger *f0std* values, $p < 0.001$) and presence of aperiodicity due to harshness. Harsh voices were observed with high frequency in the frenzy/angry speech of four speakers (F07, F09, M02 and M06).

4. Conclusions and future directions

We analyzed the differences in expression types of frenzy/angry and far-directed utterances, accounting for five acoustic-prosodic and voice quality features.

Prosodic analysis results indicated that both expression types are accompanied by louder power and higher pitch, by around 20 to 30 dB and 1 octave higher in comparison to reading out type, while differences were observed in the intonation patterns: far-directed voices tend to have large power and high pitch over the whole utterance, while angry speech has more up-down pitch movements resulting in a larger pitch range. Regarding voice quality, both expression types tend to be tenser (higher vocal effort), but angry speech tends to be more pressed, with local appearance of harsh voices (with irregularities in the vocal fold vibrations).

Future investigations include discrimination of the different types of loud voice from acoustic and linguistic information. The combination with facial expressions and body movements are also future challenges for discrimination of audio-visual aggressive attitudes.

5. Acknowledgements

This work was supported by JST CREST Grant JPMJCR17A2, and JST ERATO Grant JPMJER1401. We thank Taeko Murase, Megumi Taniguchi, Miki Okuno and Yuka Nakayama for contributions in the experiment conduction and data annotation.

6. References

[1] Junqua, J.C. The Lombard reflex and its role on human listeners and automatic speech recognizers. *J. Acoust. Soc. Am. 93*, 510-524 (1993)

[2] Zahorik, P. & Kelly, J.W. (2007). Accurate vocal compensation for sound intensity loss with increasing distance in natural environments. *Journal of the Acoustical Society of America 122*(5), 143-150, 2007.

[3] Michael, D.D., Siegel, G.M., & Pick, H.L. (1995). Effects of distance on vocal intensity. *Journal of Speech and Hearing Research, 38*, 1176-1183, 1995.

[4] Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *Journal of the Acoustical Society of America 85*(1), 295-312, 1989.

[5] Hayashi, K., Shiomi, M., Kanda, T., & Hagita, N. "Are Robots Appropriate for Troublesome and Communicative Tasks in a City Environment?," *IEEE Trans. on Autonomous Mental Development*, 4(2), 150-160, 2012.

[6] Olweus, D., Bullying at school: What we know and what we can do, *Oxford: Blackwell Publishers*, 1993

[7] Arluke, A., "Animal Abuse as Dirty Play", *Symbolic Interaction*, vol.25, pp.405-430, 2002.

[8] Miller, C. "Childhood animal cruelty and interpersonal violence", *Clinical Psych. Review*, vol.21, pp.735-749, 2001.

[9] Salvini, P., Ciaravella, G., Yu, W., Ferri, G., Manzi, A., Mazzolai, B., Laschi, C., Oh, S.-R., Dario, P., "How safe are service robots in urban environments? Bullying a Robot", Proc. *IEEE Int. Symp. on Robot and Human Interactive Communication* (RO-MAN), pp.1-7, 2010.

[10] Brscić, D., Kidokoro, H., Suehiro, Y., & Kanda, T. (2015). "Escaping from Children's Abuse of Social Robots," Proc. of *the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (HRI2015).

[11] http://pupsafeproject.org/social/aggressive-speech/

[12] Scherer, K.R., Johnstone, T., and Klasmeyer, G. 2003. Vocal expression of emotion. In *Handbook of the Affective Sciences*, R. J. Davidson, K. R. Scherer, and H. Goldsmith, Eds. Oxford, UK: Oxford University Press, ch. 23, 433–456.

[13] Schuller, B., Batliner, A., Steidl, S., Seppi, D. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 53(9–10), Nov.–Dec. 2011, 1062-1087.

[14] Weninger, F., Wollmer, M., Schuller, B. 2015. Emotion recognition in naturalistic speech and language – a survey. In *Emotion Recognition: A Pattern Analysis Approach*, A. Konar, A. Chakraborty. Eds., Publisher: John Wiley & Sons, Inc., ch 10, 237-268.

[15] Ishi, C.T., Ishiguro, H., Hagita, N. (2008). Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication* 50(6), 531-543, June 2008.

[16] Ishi, C.T., Arai, J., Hagita, N. (2017). "Prosodic analysis of attention-drawing speech," Proc. Interspeech 2017, 909-913.

[17] Ishi, C.T., Arai, J., (2018). Periodicity, spectral and electroglottographic analyses of pressed voice in expressive speech. Acoustical Science and Technology, Vol. 39, No. 2, 101-108.

[18] Ishi, C.T. (2004). "A New Acoustic Measure for Aspiration Noise Detection," Proceedings of The 8th International Conference of Speech and Language Processing 2004 (ICSLP 2004), Vol. II, 941-944.

[19] Ishi, C.T., Sakakibara, K-I., Ishiguro, H., Hagita, N. (2008). A method for automatic detection of vocal fry. IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, No. 1, 47-56, Jan. 2008.

[20] Ishi, C.T., Ishiguro, H., Hagita, N. (2010). Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. EURASIP Journal on Audio, Speech, and Music Processing 2010, ID 528193, 1-12 Jan. 2010.

[21] Ishi, C.T., Kanda, T. (2019). "Prosodic and voice quality analyses of offensive speech," accepted to *ICPhS2019*.