# PROSODIC AND VOICE QUALITY ANALYSES OF OFFENSIVE SPEECH

Carlos Toshinori Ishi and Takayuki Kanda

ATR Hiroshi Ishiguro Lab., ATR Intelligent Robotics and Communication Labs.
carlos@atr.jp, kanda@atr.jp

## ABSTRACT

In this study, differences in the acoustic-prosodic features are analyzed in low-moral or offensive speech. Utterances with the same contents were spoken by multiple speakers with different speaking styles, including reading out, aggressive speech, extremely aggressive (frenzy), and joking styles. Acoustic-prosodic analyses indicated that different speakers use different speaking styles for expressing offensive speech. Clear changes in voice quality, such as tense and harsh voices, were observed for high levels of expressivity of aggressiveness and threatening.

**Keywords**: offensive speech, prosody, voice quality, acoustic features, speaking style.

## 1. INTRODUCTION

People sometimes behave aggressively and in an offensive way in various daily contexts. For example, in stores, there are sometimes complainers who make unreasonable complaints toward store workers. Because it is quite stressful to deal with unreasonable complainers, people may wish a robot to manage such troublesome complaints [1]. In schools, bullying is a serious problem, defined as deliberate, repeated or long-term exposure to negative acts performed by a person or group of persons regarded of higher status or greater strength than the victim [2]. It is also known that people sometimes bully toward other entities like animals [3]. There is a discussion that such cruelty would turn into interpersonal violence [4]. Offensive behaviour is even exhibited toward inanimate entities, like a robot [5],[6]. Commonly to these problems, there would be a benefit if there is a technique to identify offensive utterances, e.g. the supervisor would be alarmed if an offensive utterance occurs, then he/she can better intervene the trouble from such unreasonable complainers and bully children.

Direct aggressive speaking sounds hard, hostile, and often comes across as controlling or dominating (from The Pup Safe Project [7]). A directly aggressive person will do one or more of these things:

- 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.
- They order to do what they say, demand listening to them, follow their instructions, and insult when you don't comply as they want.
- 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 [8-10], but few or no studies clearly focused on aggressive or offensive speech. Part of offensive speech is thought to share features of emotional angry voice. However, offensive speech may express attitudinal behaviours without a specific emotion expression. Further, the same offensive utterance (the same linguistic contents) can be aggressive or playful/joking depending on the accompanying speaking styles.

In this study, we take into account previous studies on acoustic-prosodic and voice quality features [11-15], and aim on clarifying the features involved in the different realizations of offensive speech.

## 2. ANALYSIS DATA

Considering that the same speech contents may convey different impressions depending on the speaking style, we prepared a script of words and sentences that possibly express low-moral offensive speech, and asked multiple male and female speakers to utter those words and sentences in different manners.

### 2.1. Script of low-moral words and sentences

A set of words and sentences was firstly prepared, based on low-moral situations observed in real-world interactions between humans and robots [6], and troublesome situations happened in human daily-life retrieved from the web. The low-moral word and sentences could roughly fall in the following set of categories. The English translations in brackets are the equivalent expressions, but may differ in nuance.

- Insulting/defame utterances: “aho” / “baka” / “doaho” / “bakayarou” / “boke” (fool, idiot, stupid), “gomi” / “kuzu” (garbage, rubbish, thrash), “yakutatazu” (useless), “ikujinashi” (coward), “kuso” (shit), “kono aho/baka” (you fool/idiot!), “omaewa aho/baka ka?” (you’re so fool/idiot!)
- Appearance-related insulting utterances: “hage” (bald), “busu” / “busaiku” (ugly), “chibi” (small), “debu” (fat, chubby), “jijii” (old geezer), “babaa” (old hag), “gaki” (brat).
- Fooling/provocative utterances: “zamaamiro” (Serves you right! You deserved it!), “babaa nani yuuteruno!” (Old hag! What the hell are you saying!),
- Offensive utterances: “konoyarou”/ “temee” / “kisama” / “nanisamada” (You bastard), “shine” / “kiero” (Go to hell!), “nametennoka” (Are you taking the piss?)
- Aggressive utterances: “korosuzo”/ “bukkorosuzo” (“I kill you!”), “tatakuyo”/ “naguruzo” (I punch you!), “keruzo” (I kick you!), “keisatsu yobuzo” (I call the Police)
- Crude comments: “uttooshii” / “urusai” (annoying, irritating)
- Order/scolding utterances: “hayaku shiro” (Do it quickly!), “sassato ike” (Get out of here!), “ayamare” (Apologize to me!), “damare” (Shut up!),

### 2.2. Data collection

We recruited multiple male and female speakers to utter the low-moral words and sentences described in the previous sub-section, in different manners. Headset microphones (DPA 4060) were used to collect audio data.

Four speaking styles shown below were requested to express (the original words in Japanese are in brackets):

- “Reading out” (“yomiage”): read out without emotions/temper.
- “Aggressive” (“bougen”): offensive, threatening, aggressive expression.
- “Frenzy” (“kyouran”): extreme expression of aggressiveness; furious, mad, hysteric expression, losing one’s temper.
- “Joking” (“joudan”): non-aggressive, non-serious joking/kidding/playful expression.

We collected data from 10 male and 10 female speakers aged 20s to 60s. Some of the speakers self-reported that they seldom utter aggressive speech in daily-life, and could not express well the aggressive and frenzy styles. On the other hand, others could not express well the joking style. Then, in order to check how appropriate each speaker could express the targeted styles, we asked two annotators (research assistants) to listen to the speech utterances and grade their perceptual impressions on the expressivity of the different styles. A perceived degree of aggressiveness was graded from -3 (very jokey) to 3 (very aggressive, seriously aggressive), and a perceived degree of threatening was graded from -3 (very gentle) to 3 (very scary).

Based on the perceptual scores of the annotators, four male and four female speakers were selected for the subsequent analyses, who could better express the different situations, in comparison to the other speakers. Fig. 1 shows the average scores of the perceived degree of aggressiveness and threatening for different styles, for the eight selected speakers. The IDs and ages of the selected speakers are F05 (31), F06 (35), F07 (20), F09 (49) for the female speakers, and M02 (66), M03 (61), M06 (49), M09 (20) for the male speakers. The speakers removed from the analysis received scores closer to zero. The correlation coefficients between the two annotators for the perceptual scores of the selected speakers were 0.95 for aggressiveness and 0.92 for the threating degrees.

It can be observed in Fig. 1 that the collected joking speech utterances were graded between slightly jokey (-1) to jokey (-2), and slightly gentle (-1), on average. The aggressive speech utterances were graded around slightly aggressive (1) to aggressive (2) and around slightly scary (1) to scary (2). The frenzy speech utterances were graded as very aggressive (3) and very scary (3). Among the selected speakers, F05 and M09 were graded to be less aggressive and less scary than the others.

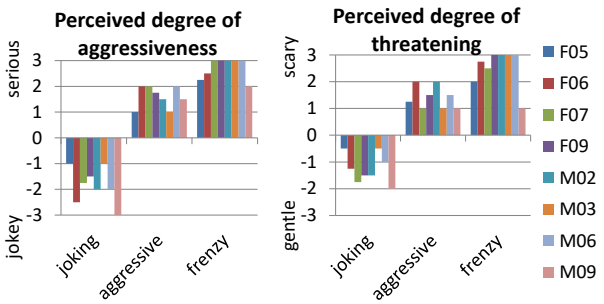


**Figure 1:** Perceived degree of aggressiveness (serious to jokey) and threatening (scary to gentle), for different styles.

## 3. PROSODIC AND VOICE QUALITY ANALYSIS

### 3.1. Acoustic analysis

Acoustic analyses were conducted on prosodic and voice quality related features. First, frame-level acoustic features were extracted each 10ms frames.

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 [11]. All utterances were manually segmented, and utterance-level acoustic parameters were extracted for each utterance. Fig. 2 shows the distributions (average and standard deviations) of the six utterance-level acoustic parameters described below, for four speakers.

- "f0max" is the maximum f0 in the utterance, in semitone units. In Fig. 2, the F0 values are normalized (subtracted) by the F0 value of the reading out style. The mean F0 was also calculated, but the results are omitted from the figure since similar trends have been found.
- "power" is the maximum power value in the utterance, in dB. In Fig. 2, the power values are normalized (subtracted) by the power value of the reading out style.
- "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 vocal tension and pressed voices [12].
- "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 [13].
- "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 [14]. The values in Fig. 2 are scaled by 10 times to allow better visualization.
- "breathiness" is the total length of vocalic segments detected as breathy, in seconds. Breathy segments are detected by the method proposed in [15].

Pairwise t-tests (Welch t-tests) were conducted for checking statistical significances between different conditions within a speaker and within an acoustic parameter.

From the results in Fig. 2, it can be firstly observed that power becomes higher in aggressive and frenzy styles by about 10 to 25dB compared with the reading out style, for all speakers (Welch t-tests, $p<0.01$).

Regarding F0, a gradual increase can also be observed for aggressive and frenzy styles. Half of the speakers showed F0 higher than an octave (12 semitones) in frenzy style (Welch t-tests, $p<0.01$).

Regarding the vocal tension-related spectral features "h1-a1" and "a1-a3", it can be observed that both parameters decrease from reading out to frenzy styles. This indicates that there is not only a louder or higher voice, but also a tenser voice quality in aggressive and frenzy styles. In speakers M03 and M06, there is not a big change in the "h1-a1" parameter, but the changes in the spectral tilt parametrized by "a1-a3" are significant (Welch t-tests, $p<0.01$). The spectrum becomes flatter ("a1-a3" values closer to 0dB) in frenzy styles, which have tenser voice qualities.

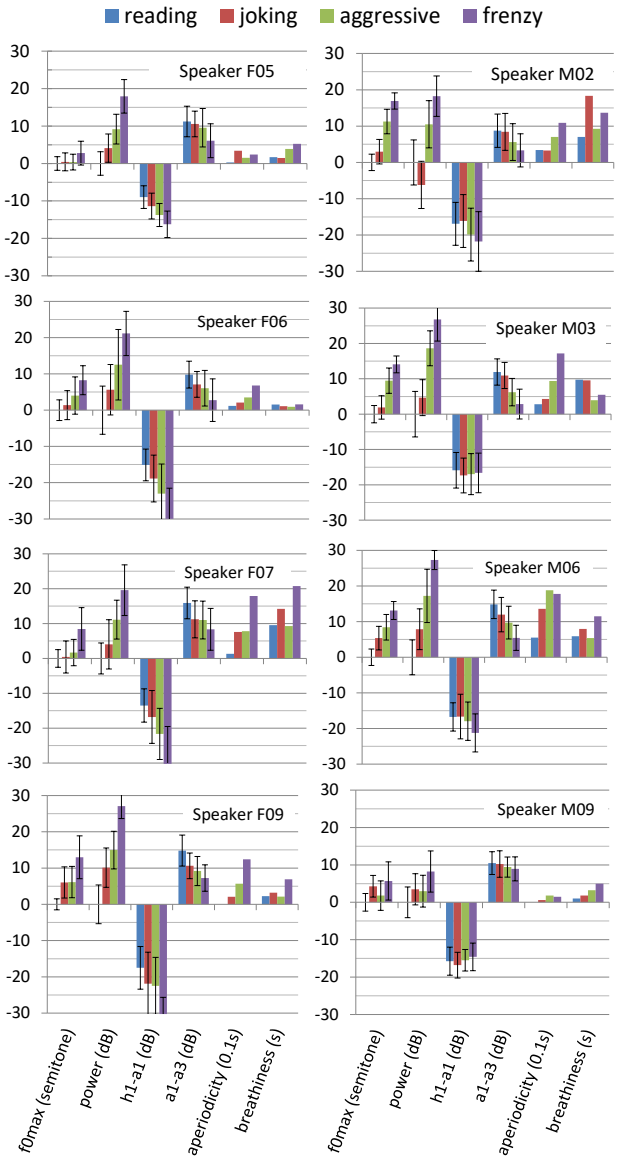


**Figure 2:** Distributions of six acoustic-prosodic features for different styles, for each speaker. The units of the vertical axis are shown in brackets for each feature in the horizontal axis Mean and standard deviations are shown for f0, power and spectral features, while total durations are shown for aperiodicity and breathiness parameters.

Finally, regarding the voice quality parameters, presence of aperiodicity, breathiness and pressed/tense voices, which are associated with harsh voices, are observed in four speakers (F07, F09, M02 and M06).

### 3.2. Speaker differences

The analysis results above showed common features in the speaking styles of the different offensive speech expressions. However, a close look on the data revealed that some of the speakers expressed offensive speech in different ways.

For example, in the frenzy style, two different expressions were observed. One is a shouting style (M02, M03, M06, F06, F09), having powerful, high-pitched, tense, harsh voice; the other is a hysteric style (M02, F07, F09) having presence of pressed falsetto, and presence of inhaled falsetto after the offensive speech utterance (F07).

Another feature of aggressive and frenzy styles is the trill of "r" consonants. In standard Japanese, "r" consonants are not trilled. The "r" trilling in Japanese usually appears in threating expressions mainly by male speakers. Trilling was observed in some of the aggressive and frenzy utterances of speakers M03 and M06.

Different expressions were also observed in the joking style. The acoustic features of joking utterances in Fig. 2 showed intermediate values between reading out and aggressive styles. This trend was different in the speakers F07 and M02, which showed a softer and breathier voice quality. In some of the speakers, lengthening and rising tones in the last syllable of the utterance were observed (F05, M02, M03). Some speakers expressed the joking style by including laughing speech (F06, F09, M06).

For the task of discrimination of offensive speech, the features above should also be taken into account.

Finally, two of the speakers (M09 and F05) showed smaller changes among the acoustic features of different speaking styles in Fig. 2. Speaker M09 was the only one showing smaller degree of power increase over different styles, while speakers F05 and M09 showed smaller degree of increase in F0 in frenzy style. Further, M09 did not show significant differences in the vocal tension-related spectral parameters. These results are in correspondence to the lower perceived degree of aggressiveness and threatening for these two speakers as shown in Fig. 1.

## 4. DISCUSSION

We stated in the introduction that attitudinal offensive speech is related, but not equal to emotional angry voice. The samples of shouting style in frenzy speech (with high aggressiveness scores and high threatening scores) are thought to correspond to a "hot anger" emotion expression. On the other hand, part of the aggressive speech (with lower aggressiveness scores) was not particularly felt as angry. For the identification of offensive attitudes, and for the clarification of the differences between offensive attitudes and emotional expressions, the combination of linguistic and prosodic features should be taken into account.

The results for A1-A3 and breathiness might look contradictory from the "breathy voice" definition, which is produced with low vocal tension. However, the parameter "breathiness" includes both breathy and whispery voices, so that the lower A1-A3 values in frenzy style correspond to whispery voices.

## 5. CONCLUSIONS AND FUTURE DIRECTIONS

We collected offensive speech data of male and female speakers in four different manners (reading out, aggressive, frenzy and joking) and analysed the differences in speaking style accounting for several acoustic-prosodic and voice quality features.

Aggressive utterances were generally found to be expressed louder, with higher pitch, and with a tenser voice quality. Two distinct styles were found in the expression of frenzy utterances, a shouting style with powerful, high-pitched, tense/pressed and harsh voice, and a hysteric style, with presence of pressed falsetto voice qualities. The joking utterances were often expressed by a softer voice quality and with acoustic features intermediate between reading out and aggressive utterances.

Among the analysed acoustic-prosodic and voice quality features, six were found to well represent the different styles in offensive speech expression: maximum or mean f0 in the utterance, power of the utterance, spectral features related to vocal tension ("h1-a1" and "a1-a3"), and vocal fold vibration-related features ("aperiodicity" and "breathiness").

More detailed analysis at utterance and syllable levels, including durational differences, and consideration of linguistic information are topics for future investigation. Detection of laughing speech and falsetto inhalation are also remaining topics for acoustic analysis. Finally, the combination with facial expressions and body movements are also future challenges for detection of audio-visual offensive attitudes.

## 6. ACKNOWLEDGMENT

This work was supported by JST CREST Grant Number JPMJCR17A2, Japan. We thank Taeko Murase, Miki Okuno and Megumi Taniguchi for contributions in the experiments and data annotation.

## 7. REFERENCES


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

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

[3] Arluke, A. 2002. Animal Abuse as Dirty Play. *Symbolic Interaction* 25, 405-430.

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

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

[6] 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*).

[7] What is aggressive speech? - The Pup Safe Project. http://pupsafeproject.org/social/aggressive-speech/, last access Dec. 2018.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Ishi, C.T., Arai, J. 2018. Periodicity, spectral and electroglottographic analyses of pressed voice in expressive speech. *Acoustical Science and Technology* 39(2), 101-108.

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

[14] 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* 16(1), 47-56.

[15] 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.