

PAPER

Periodicity, spectral and electroglottographic analyses of pressed voice in expressive speech

Carlos Toshinori Ishi* and Jun Arai†

ATR HIL,

2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan

(Received 10 April 2017, Accepted for publication 12 September 2017)

Abstract: Pressed voice is a type of voice quality produced by pressing/straining the vocal folds, which often appears in Japanese conversational speech when expressing paralinguistic information related to emotional or attitudinal behaviors of the speaker. With the aim of clarifying the acoustic and physiological features involved in pressed voice production, we conducted periodicity, spectral and electroglottographic (EGG) analyses on pressed voice segments extracted from spontaneous dialogue speech of several speakers. Periodicity analysis first indicated that pressed voice is usually accompanied by creaky or harsh voices, having irregularities in periodicity, but can also be accompanied by periodic voices with fundamental frequencies in the range of modal phonation. Spectral analysis indicated power is usually reduced in low frequency components of pressed segments. A spectral measure $H1' - A1'$ was then proposed for characterizing pressed voice segments which commonly has few or no harmonicity. $H1' - A1'$ was shown to be effective for identifying most pressed segments, but fails when nasalization occurs. Vocal fold vibratory pattern analysis from the EGG signals revealed that most pressed voice segments (including nasalized vowels) are characterized by glottal pulses with closed intervals longer than open intervals on average, regardless of periodicity.

Keywords: Pressed voice, Voice quality, EGG, Expressive speech, Prosody, “rikimi”

PACS number: 43.72.Ar, 43.72.Kb, 43.70.Gr [doi:10.1250/ast.39.101]

1. INTRODUCTION

Variations in the vibration modes of the vocal folds produce different voice qualities, such as breathy, whispery, creaky (or vocal fry), and harsh voices [1–6]. It is known that changes in voice quality have important roles in the communication of linguistic or paralinguistic information depending on the language, besides intonation-related prosodic features [7–10]. It has been reported that a pressed voice quality category produced by pressing/straining the vocal folds, named “rikimi” in Japanese [11], takes important roles in the expression of emotional or attitudinal behaviors of the speaker [11–14]. The perceptual sensation of pressed voice is of a tense/strident voice quality. Its auditory impression can be better understood by listening to the speech samples in [15].

It is reported that the speaker’s sincerity is displayed by pressed voice primarily in the expression of suffering, admiring, and emphasizing [11]. In [12], it is reported that the speaker’s attitude or emotion is conveyed by pressed

voice in items including emotions like surprise, admiration and disgust in interjections, real feeling expression in adjectives, expressivity in onomatopoeia, hesitation, and modesty. In [14], analysis was conducted to clarify the situations where pressed voice appears. It is reported that the situations where pressed voice appears can be classified in three categories: 1) the subject quotes (reproduces) his own or other’s past utterances; 2) the subject expresses sympathy to the interlocutor; 3) the subject is an expert (i.e., an owner of knowledge), and release (or try to release) knowledge to the interlocutor. This last item has been sub-classified in the following items, according to the speaker’s mental state: a) the subject worries (does not know what to say) with regard to the interlocutor’s opinion or question; b) the subject negates the interlocutor’s opinion, but expressing modesty; c) the subject expresses real feelings, as someone who experienced the dialogue topic.

Given that a variety of paralinguistic information is conveyed by pressed voice, an automatic identification of pressed voice segments from the acoustic features of speech signals, or a correct control of prosodic and voice quality features involved in pressed voice production, are useful for dialogue systems. However, although the

*e-mail: carlos@atr.jp

†e-mail: araijun@hotmail.com

previous studies analyze the communication functions of pressed voice in dialogue speech, a strict definition in terms of acoustic or physiological properties is not provided. There is a categorization of laryngeal voice qualities in modal, breathy, whispery, creaky, harsh and falsetto [1–6], in terms of physiological properties. However, our preliminary analysis indicated that pressed voice does not fit exactly to the descriptions of any of them, but rather may co-exist with some of them [12]. Preliminary results indicated that most of the pressed voice utterances have acoustic features similar to creaky voice (or vocal fry: impulse-like glottal excitation with very low fundamental frequencies, usually accompanied by irregularity in periodicity), or harsh voice (noisy rasping sound, with aperiodic glottal pulses).

The term “pressed voice” has been used as an English translation of “rikimi” in [11,12]. However, we clarify that this term is used in a broader sense, so that it is not strictly equivalent to a “pressed phonation,” in terms of physiological fundamentals. The term “pressed phonation” has been attributed to a “tight voice” or anterior phonation, where the arytenoid cartilages are held together, so that only the anterior ligamental part of the glottis participates in phonation [1,3].

Regarding physiological analysis, electroglottographic (EGG) signals have also been used to analyze the vibration patterns of the vocal folds during pressed voice [12]. However, the dataset was very small (only fifteen conversation passages), and EGG signals were available for only one speaker, who tried to imitate the pressed voice utterances of the audio dataset.

In the present study, in order to clarify the acoustic and physiological properties of the different realizations of pressed voice as a paralinguistic conveyer, we conducted acoustic and electroglottographic analyses on pressed voice utterances extracted from spontaneous dialogue speech of several speakers. The present work is an extended version of our previous studies [12,13].

This paper is organized as follows. In Sect. 2, the speech data and the annotation data used for analysis are described. In Sect. 3, acoustic and EGG analyses are reported. In Sect. 4, discussion about the results is reported, and Sect. 5 concludes the paper.

2. SPEECH AND ANNOTATION DATA

2.1. Speech Data

Most works dealing with voice quality use the stationary portion of specific voice qualities consciously produced by the subjects. However, although pressed voice frequently appears in natural conversation, most subjects cannot produce it in a conscious manner. In the present work, we constructed a dataset of pressed voice utterances extracted from natural conversational speech databases,

where pressed voice is unconsciously produced. In this way, the variations in speaking styles in the realization of pressed voice in spontaneous speech can be analyzed.

The first dataset is constituted by utterances containing pressed voice selected from the Japanese conversational speech database recorded in the JST/CREST ESP Project [16]. This dataset consists of 15 conversational passages (KoB001 ~ 015) including voices of 3 male and 5 female speakers, aging from 10 to 60. This dataset is the same used by [11], for studying the functional properties of pressed voice in speech communication. Only audio signal is available in this database.

The second dataset was collected from our multimodal dialogue speech database [17], which contains spontaneous Japanese dialogues of several speakers. Several multimodal signals (including audio, EGG and motion capture data) are simultaneously recorded for each dialogue partner in sessions of 10 to 15 minutes. For the present analysis, we used the speech and EGG data of 46 recording sessions including 27 different combinations for pairs of dialogue partners. The recording conditions are as follows. The dialogue partners sat in front of each other, separated by a desk, and an approximate distance of 1 meter. Directional microphones (Sanken CS-1) were positioned pointing towards each speaker. (In some of the recordings, headset microphone signals are also available.) The EGG device used in the recordings is the Glottal Enterprises EG2-PC. All waveforms were sampled at 16 kHz, 16 bits. The speakers were instructed to talk freely about any topic. The resulting dialogues were mostly daily conversations, including topics like past happenings, future plans for trips, self-introductions, topics about a common known person, topics regarding family and work, and past experiences.

The 46 recording sessions used for analysis include 10 female speakers (whose ID and ages are FYU (4), FFS (6), FSF (15), FMH (30), FKN (30), FMU (30s), FKI (30s), FYS (30s), FKH (50s), FHT (60s)), and 9 male speakers (MTI (4), MTT (17), MFT (17), MYM (20s), MSR (29), MIT (30), MMS (30s), MSN (38) and MHI (40s)).

2.2. Annotation Data

2.2.1. Identification of pressed voice utterances

Candidates of utterances containing pressed voice were firstly selected by two subjects (research assistants; native speakers of Japanese), by listening to all speech utterances in the dialogue speech database. The only instruction given to the subjects was to judge if an utterance contains pressed voice, i.e., if the speaker is pressing/straining the vocal folds during phonation. 157 utterances were selected in this first stage.

Pressed voice was found in a wide range of ages, regardless of gender. For the second dataset, pressed voice was found in 7 of the 10 female speakers (FYU, FKN,

FMH, FKI, FMU, FKH and FHT) and in 6 of the 9 male speakers (MFT, MTT, MYM, MSR, MIT, and MSN).

Then, in a second stage, four subjects (research assistants; native speakers of Japanese, different from the subjects in the first stage) were asked to rate the degree of perceived pressed voice on a 3-point scale (“2” when pressed voice is clearly perceived, “1” when pressed voice is weakly perceived, and “0” when pressed voice is not perceived). An additional instruction was given to only choose “2” when they perceive that the vocal folds are clearly more pressed than necessary, i.e., the muscle tensions in the vocal folds are clearly stronger relative to the normal phonation. As a result, many “1” and “0” appeared in the annotations for all subjects. The Crombach’s alpha value between the subjective scores was 0.77, indicating acceptable consistency between raters. The scores given by all four subjects were then summed up for each utterance, and used within this paper as a perceived degree of pressed voice quality. The categories {“r1,” “r2,” “r3”} were defined to correspond to the summed score ranges of {2~4, 5~6, 7~8} from the subjective pressed voice degrees. The utterances whose summed scores were smaller than 2, were removed from the analysis, resulting in 123 utterances.

2.2.2. Segmentation of pressed voice

In order to evaluate the acoustic features and vibratory patterns in pressed voice segments, a more detailed segmentation of voice quality was conducted at a segmental level, for the utterances containing pressed voice. This is because pressed voice is usually manifested by a voice quality change only in part of the voiced portions in a word or a phrase. This type of segmentation and classification requires knowledge about acoustic-phonetics and laryngeal voice qualities. The segmentation was conducted by the first author, which has experience in acoustics and voice quality analysis, based on visual inspection of spectrograms and waveforms, and on auditory impression. The segmentation process resulted in 276 segments. For evaluation purposes, modal (normal) phonation intervals were also segmented around the pressed voice portions. 203 segments were obtained for modal voice.

3. ANALYSIS RESULTS

3.1. Periodicity and Voice Quality Analyses

Irregularity in periodicity has been reported as one characteristic of pressed voice [11]. This section aims on clarifying how these irregularities are manifested in pressed voice segments. The phonation types involved in pressed voice production are also discussed based on periodicity information.

In preliminary analyses, it has been observed that creaky and harsh phonation types often appear in pressed voice utterances. We consider that a strict categorization in

Table 1 Distributions of the perceived degrees of pressed voice, according to the periodicity-based voice quality categories.

	avg. $f_0 < 100$ Hz	avg. $f_0 > 100$ Hz inter-pulse irregularity	avg. $f_0 > 100$ Hz no inter-pulse irregularity	Total (%)
r3	20	15	3	38 (31%)
r2	16	13	11	40 (33%)
r1	30	7	8	45 (36%)
Total (%)	66 (54%)	35 (28%)	22 (18%)	123

creaky and harsh voice qualities is not consistent due to ambiguities in the acoustic and perceptual spaces. Thus, the discrimination of voice qualities related to creaky and harsh phonations were categorized in three groups, based on the periodicity properties, according to the following criterion. The first category is composed by segments with average f_0 smaller than 100 Hz, which is characteristic of creaky phonation. The second category is composed by segments with f_0 larger than 100 Hz and with some pulse-to-pulse irregularity (including double-periodicity with alternations of small and big glottal pulses, and aperiodic variations between adjacent glottal pulses), which is characteristic of harsh phonation. Finally, the third category is composed by segments with f_0 larger than 100 Hz (i.e., in the range of modal phonation), and no evident irregularity in the inter-pulse intervals. This third category includes segments which are ambiguous for the classical classification of voice qualities. This is because f_0 is not as low as in creaky voices, periodicity is not as irregular as in harsh voices, and the auditory impression has a tenser quality compared to normal phonation.

Table 1 shows the distributions of the perceived degrees of pressed voice, according to the periodicity-based voice quality categories.

The distributions in Table 1 show that pressed voice utterances are realized with f_0 lower than 100 Hz (corresponding to creaky voices) in 54% of the utterances, and with f_0 higher than 100 Hz and inter-pulse irregularity (corresponding to harsh voices) in 28% of the utterances. However, it is also shown that 18% of the pressed voice utterances are realized with f_0 higher than 100 Hz, without a clear irregularity in the pulse-to-pulse intervals. Further, it can be observed from the table that the majority of the utterances of both creaky and harsh voices appeared in categories with high degree of pressed voice quality.

The f_0 contours in pressed voice utterances were then analyzed. Figure 1 shows f_0 contours, spectrograms, speech

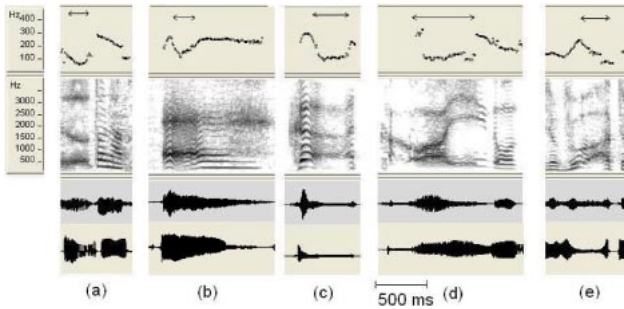


Fig. 1 f_0 contours, spectrograms, speech and EGG waveforms for utterances including pressed voice segments (indicated in the top panels) accompanied by several phonation types. a) pressed creak, $f_0 < 100$ Hz (z[uu]tto, FMU); b) pressed, $f_0 > 100$ Hz (e[e]ee, FYU); c) pressed, $f_0 > 100$ Hz (ga[aa]a, FHT); d) pressed diplophonic (k[awaii], MYM); e) pressed/harsh (b[wa]a, MFT).

and EGG waveforms for pressed voice segments accompanied by different phonation types. In the current subsection, we discuss about the f_0 contours, while the spectrogram and EGG features will be discussed in the following sub-sections.

A common feature found in all pressed voice utterances was an f_0 dip in the pressed voice portions. Two transition types could be identified between modal to/from pressed voice segments. One is a fast, but continuous, decrease in f_0 , as shown in the examples of Figs. 1(a)–1(c). This pattern was observed in 17% of the pressed voice utterances. The other type is a sudden (discontinuous) f_0 jump, as in the diplophonic signal (where multiple phonations can be simultaneously perceived [18]) shown in Fig. 1(d). In the remaining 83% of the utterances the f_0 curve was broken, as in the example of Fig. 1(d), or a consistent f_0 curve could not be estimated due to irregularities in the pulse-to-pulse intervals.

3.2. Spectral Measure: $H1' - A1'$

Spectral tilt is a commonly used feature for characterizing voice qualities. In [19], spectral tilt is reported to be effective for discriminating “tense voice” from “lax voice.” As one characteristic of pressed voice is of a tense voice quality, it is expected that spectral tilt can potentially discriminate it from other voice qualities. It can be observed in the spectrograms of Fig. 1 that the power of the frequency components of the lower harmonics is reduced in the pressed voice portions.

Classical measures for spectral tilt are based on the differences between the amplitudes of the first and second harmonics ($H1 - H2$) or between $H1$ and the harmonic closest to the first formant ($H1 - A1$) [7,19]. The $H1 - A1$ measure was considered to be a good alternative for characterizing pressed voice segments, since it reflects the

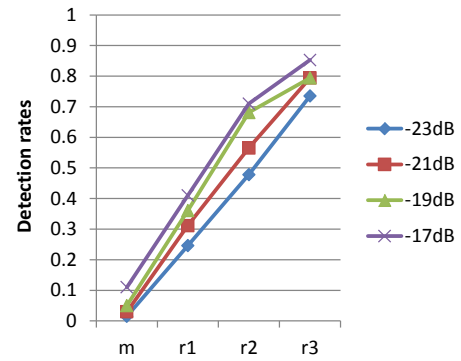


Fig. 2 Discrimination performances of pressed voice segments by using different thresholds for $H1' - A1'$.

effects of loss of energy in the low frequencies, due to high vocal fold tensions.

However, a problem in applying the above measures for pressed voice is that the harmonic structure is disturbed or sometimes inexistent when irregularities in periodicity are present. In such cases, in place of $H1$ and $A1$, we proposed the use of the maximum peak amplitude in the range of 100 to 200 Hz ($H1'$), and the maximum peak amplitude in the range of 200 to 1,200 Hz ($A1'$), where the first formant is likely to occur. For periodic signals, $H1' = H1$, and $A1' = A1$.

For the computation of the power spectrum (prior to the computation of the $H1'$ and $A1'$ values), pre-emphasis processing was introduced in the present study since it is commonly applied before speech analysis to emphasize the formant peaks. A high-pass filter with the transfer function $(1 - 0.96z^{-1})$ was used for implementing the pre-emphasis processing.

Segments with $H1' - A1'$ values smaller than a threshold are judged as pressed voice. Figure 2 shows the discrimination performances of pressed voice segments by using different thresholds for $H1' - A1'$.

Results in Fig. 2 show high detection rates in pressed segments (“r3”) (true positives), low detection rates in modal segments (“m”) (false negatives), and intermediate detection rates for the “r1” and “r2” categories. For example, for a threshold of -21 dB for $H1' - A1'$, a correct detection rate of 80% is achieved in pressed segments, with a false detection rate of 3% in modal segments.

Detailed analysis were also conducted on the pressed segments “r3” which could not be identified as pressed by $H1' - A1'$. Figure 3 shows examples of spectrums of pressed and non-pressed intervals extracted from our dataset. These examples were selected in a way to illustrate the cases where true/false positive/negative detections occur in the pressed segment detection by $H1' - A1'$.

Figures 3(a) and 3(b) show typical examples of pressed creaky and harsh segments, where $H1' - A1'$ is lower than

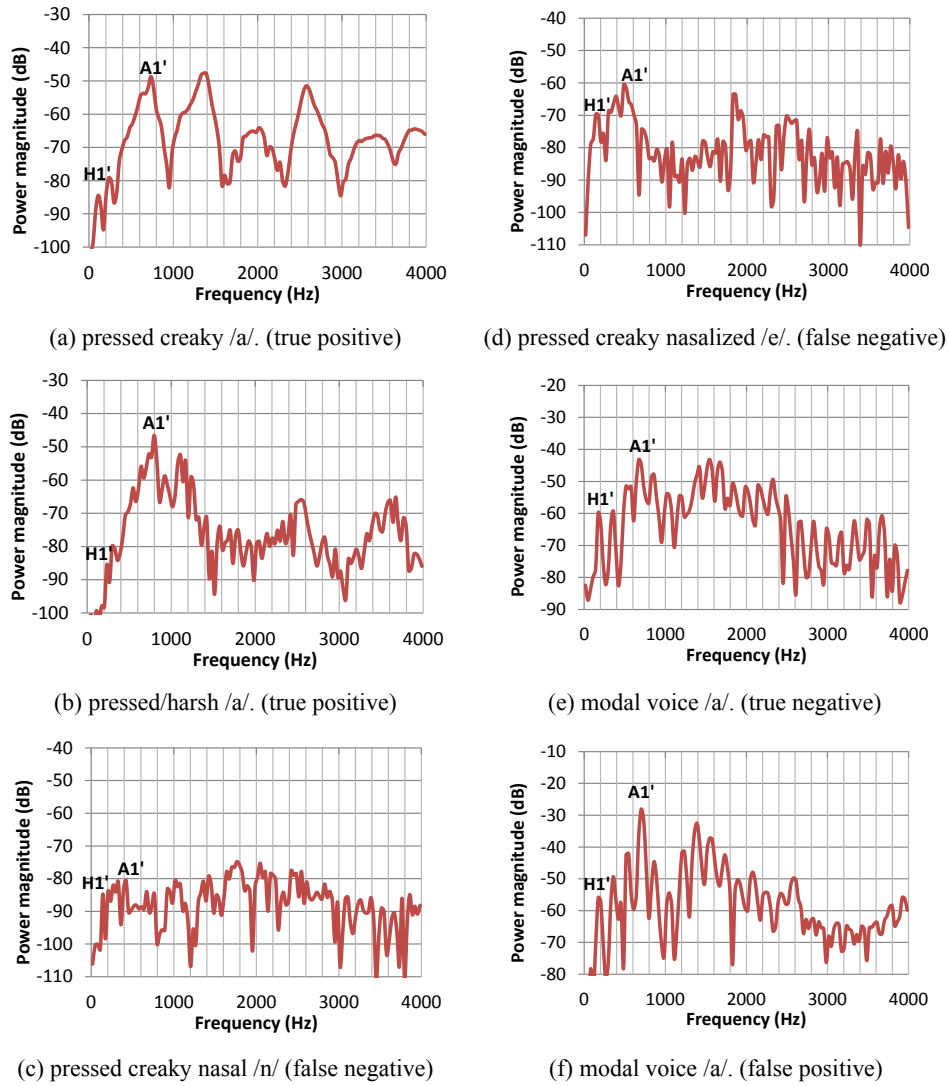


Fig. 3 Examples of spectrums of pressed and non-pressed intervals, with true/false positive/negative detections by $H1' - A1'$.

−20 dB, leading to true positive detections. The problem of $H1' - A1'$ occurs for nasals (Fig. 3(c)) or nasalized vowels (Fig. 3(d)), where a strong spectral component appears in the low frequency bands around 100 to 300 Hz, due to a nasal formant. This causes an increase in $H1' - A1'$ values, and consequent false negative detections of pressed nasal or nasalized segments. Figure 3(e) shows an example of modal segment, where $H1' - A1'$ is larger than −20 dB, leading to a true negative detection. Finally, Fig. 3(f) shows an example of false positive detection in a non-pressed segment. Although these segments have a more “tense” voice quality compared with other modal segments, they are not perceptually judged as a pressed voice. Such tense voice segments have intermediate features between modal and pressed.

How to deal with false negatives in pressed nasalized segments and false positives in non-pressed segments is a topic for future work.

3.3. Vocal Fold Vibratory Patterns: EGG Analysis

Vocal fold vibratory patterns were analyzed based on the EGG signals, for supporting the understanding of the phonation types involved in pressed voice production.

Figure 4 shows speech, EGG and DEGG (derivative of the EGG) waveforms of representative samples of pressed voice found in our dataset. These samples were selected in a way to illustrate the different vibratory patterns including single/multiple glottal pulses in different phonation types (creaky, harsh and periodic) found in pressed voice segments. Modal and non-pressed creaky segments are also included for comparison. The length of the segments is 100 ms for all plots. The amplitudes of the waveforms are re-scaled to allow better visualization.

Regarding the interpretation of the EGG waveforms, the peaks represent high vocal fold contact (or glottal closure), while the valleys represent low vocal fold contact (or glottal opening). The DEGG waveforms are computed

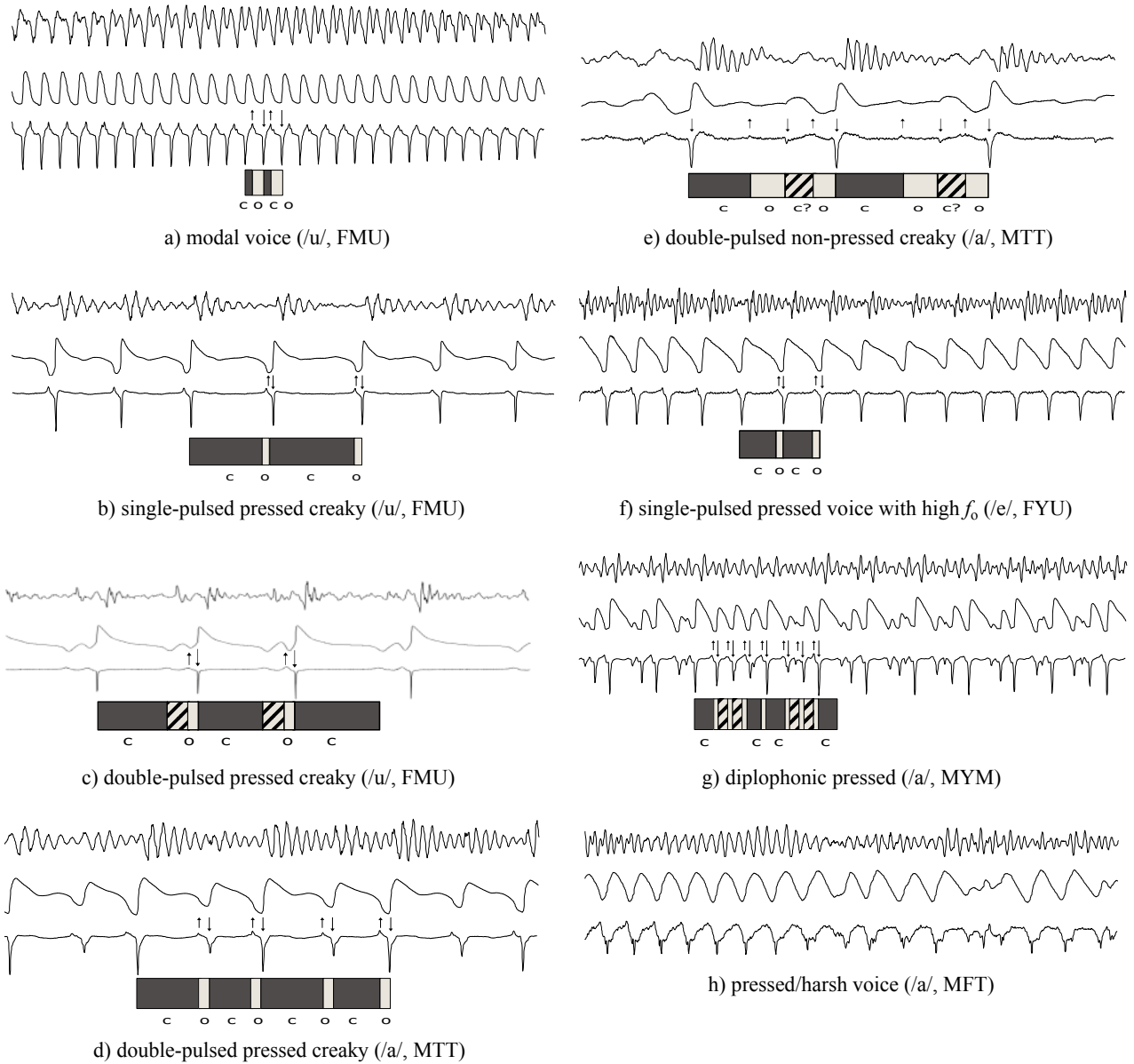


Fig. 4 Speech, EGG and DEGG waveforms, for several types of vibratory patterns in pressed and non-pressed segments. Below the DEGG waveforms, closed intervals (“c”) are indicated by dark gray regions, open intervals (“o”) are indicated by light gray regions, and incomplete closure is indicated by striped regions. Segment lengths are 100 ms.

as the negative of the derivative waveforms of the EGG signals, and provide a better visualization of the approximated instants of opening (positive peaks) and closing (negative peaks) of the vocal folds [20]. This way, open intervals can be visualized as the intervals between positive and successive negative peaks in the DEGG waveform, while closed intervals can be visualized as the pulse cycle intervals minus the open intervals.

Upward/downward arrows indicate approximate instants of glottal opening/closing in Fig. 4. These instants were hand-annotated by visual inspection of the waveforms. The closed and opened intervals are drawn (dark gray for closed intervals and light gray for opened intervals) based on the glottal opening/closed instants.

However, incomplete closures and intervals where glottal opening instant is not clear are shown as striped segments.

Figure 4(a) shows an example of modal voice (normal phonation) of a female speaker (FMU), for reference. Figure 4(b) shows a typical example of a single-pulsed pressed creaky voice of the same speaker (FMU), where f_0 is very low (about 80 Hz). It can be observed in the EGG and DEGG waveforms that the open intervals are much shorter than the closed intervals in the pressed segment of Fig. 4(b). Figure 4(c) shows an example of double-pulsed creaky segment for the same speaker (FMU). Small pulses, corresponding to incomplete closures of the vocal folds, can be observed before bigger and longer pulses with complete closures.

Figures 4(d) and 4(e) show examples of pressed and non-pressed creaky segments of the same male speaker (MTT). Both signals have small pulses between large pulses in the EGG waveforms, corresponding to incomplete closures between complete closures. The difference between these signals is that the overall open intervals are shorter than the closed intervals in the pressed creaky segment. Similar behavior was observed in other speakers.

Figure 4(f) shows an example of pressed segment with no pulse-to-pulse irregularity in periodicity, and having f_0 ranges above 100 Hz, so that individual pulses cannot be perceived as in creaky voice. Similarly to the previous pressed examples, the open intervals are relatively shorter than the closed intervals.

Figure 4(g) shows an example of a pressed segment with a diplophonic quality (where multiple phonations can be simultaneously perceived). It can be noted that larger pulses with stronger negative DEGG peaks occur with larger inter-pulse intervals (lower f_0 , around 110 Hz), while smaller pulses occur with smaller inter-pulse intervals (higher f_0 , around 330 Hz).

As shown in the examples of Fig. 4, single- and double-pulsed patterns were observed in the glottal pulses of pressed segments. However, a common feature found for most of pressed segments was that the overall open intervals are much shorter in duration than the overall closed intervals, regardless of the periodicity (as shown in the examples of Figs. 4(b), 4(c), 4(d) and 4(f). Nonetheless, this feature was not clear in a pressed voice segment with a harsh voice quality, as in the example shown in Fig. 4(h). Positive peaks in DEGG waveforms are broader and less sharp than the other pressed examples, so that it is difficult to identify a precise glottal opening instant. A correspondence between speech and EGG waveforms is also unclear, due to insufficient closure over the whole glottal pulses. It is difficult to assert if this type of voice can be classified as “pressed” from a production viewpoint. Further, even though the EGG signal reveals an f_0 close to 150 Hz, this f_0 component is almost imperceptible when listening to the speech signal.

4. DISCUSSION

It was shown in Sect. 3.1 that an f_0 dip appears in pressed voice segments, as a consequence of the vocal fold tensions during its production. In our preliminary experiment, it was shown that subjects perceive higher pitch in several pressed voice portions even though the acoustic f_0 is lower than the neighboring modal portions. Further, it was also observed that the pressed voice portion appeared in the accented syllables of the words, where f_0 is raised in normal phonation. However, in pressed voice utterances, f_0 decreased in the accented syllables instead. Therefore, the measured f_0 curve may not be an appropriate representation

of accent/intonation in pressed voice utterances, so that special care has to be taken in the interpretation of f_0 curves of pressed voice segments.

Regarding the vocal fold vibratory analysis, it is worth to mention that there are measures like open quotient (OQ) and speed quotient (SQ) for characterizing glottal waveform shapes [20]. However, there is not a clear consensus on how to deal with the incomplete closures. The work in [21] may provide hints on how to deal with OQ in incomplete closures. This is left for future investigation.

5. CONCLUSIONS

Acoustic and electroglottographic (EGG) analyses were conducted on speech segments including pressed voice, extracted from spontaneous dialogue speech data of several speakers. In accordance with past works, pressed voice was found in a wide range of ages, regardless of gender.

From the periodicity analysis in the present work, it was firstly shown that pressed voice is usually accompanied by creaky or harsh voices, having irregularities in periodicity, but it can also be accompanied by periodic voices with f_0 s in the range of modal phonation. This implies that periodicity is not the main factor involved in pressed voice production. It was also found that the transitions between modal to pressed voice are always accompanied by a continuous or discontinuous f_0 dip.

Spectral analysis indicated reduction of power in low frequency components during pressed voice. The proposed spectral measure $H1' - A1'$ (which is closely related with tense/lax voice properties) was effective for identifying 80% of the pressed voice utterances with high degree of pressed quality, for a 3% insertion error rate in modal segments. However, detailed analyses revealed problems of $H1' - A1'$ for identifying pressed segments when nasalization occurs.

Finally, analysis of vocal fold vibratory patterns from the EGG signals revealed that in most of pressed voice segments (including nasalized segments), the completely closed intervals of the vocal folds are longer than the open intervals on average, regardless of periodicity. Finding acoustic parameters that better reflect the closed interval properties could solve the problems of the current spectral features in nasalized pressed segments. This is a subject for future investigation.

ACKNOWLEDGEMENTS

This research was supported by JST, ERATO, Grant Number JPMJER1401. We thank Yuri Suzuki, Kyoko Nakanishi, Sayaka Taniguchi, Tomoko Honda and Hiroaki Hatano for contributions on data annotation and data analysis. We also thank Ken-Ichi Sakakibara and Mihoko Teshigawara for their valuable comments on voice quality categorization.

REFERENCES

- [1] J. Catford, *Fundamental Problems in Phonetics* (Edinburgh University Press, Edinburgh, 1977), pp. 98–105.
- [2] J. Laver, “Phonatory settings,” in *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge, 1980), pp. 93–135.
- [3] M. J. Ball, J. Esling and G. J. Dickson, “The transcription of voice quality,” in *Voice Quality Measurement*, R. D. Kent and M. J. Ball, Eds. (Singular Publishing Group, San Diego, 2000), pp. 49–58.
- [4] J. Kreimann and B. Gerratt, “Measuring vocal quality,” in *Voice Quality Measurement*, R. D. Kent and M. J. Ball, Eds. (Singular Publishing Group, San Diego, 2000), pp. 73–101.
- [5] B. R. Gerratt and J. Kreiman, “Toward a taxonomy of nonmodal phonation,” *J. Phon.*, **29**, 365–381 (2001).
- [6] J. H. Esling, “Voice quality,” in *Encyclopedia of Language and Linguistics*, 2nd ed., Vol. 13, K. Brown, Ed. (Elsevier, Oxford, 2006), pp. 470–474.
- [7] M. Gordon and P. Ladefoged, “Phonation types: A cross-linguistic overview,” *J. Phon.*, **29**, 383–406 (2001).
- [8] C. Gobl and A. Ní Chasaide, “The role of voice quality in communicating emotion, mood and attitude,” *Speech Commun.*, **40**, 189–212 (2003).
- [9] G. Klasmeyer and W. F. Sendlmeier, “Voice and emotional states,” in *Voice Quality Measurement* (Singular Thomson Learning, San Diego, 2000), pp. 339–358.
- [10] C. T. Ishi, H. Ishiguro and N. Hagita, “Automatic extraction of paralinguistic information using prosodic features related to *F0*, duration and voice quality,” *Speech Commun.*, **50**, 531–543 (2008).
- [11] T. Sadanobu, “A natural history of Japanese pressed voice,” *J. Phon., Soc. Jpn.*, **8**, 29–44 (2004).
- [12] C. T. Ishi, H. Ishiguro and N. Hagita, “Acoustic and EGG analysis of pressed phonation,” *Proc. Int. Conf. Phonetic Sciences (ICPhS 2007)*, pp. 2057–2060 (2007).
- [13] C. Ishi, H. Ishiguro and N. Hagita, “Acoustic, electroglottographic and paralinguistic analyses of “rikimi” in expressive speech,” *Proc. Speech Prosody 2010 (SP2010)*, ID 100139, pp. 1–4 (2010).
- [14] J. Arai, C. Ishi and N. Hagita, “The situations where “rikimi” is used,” *Proc. Int. Conf. Japanese Language Education (ICJLE2010)* (CD-ROM), pp. 1–10 (2010) (in Japanese).
- [15] <http://www.irc.atr.jp/~carlos/pressed/> (accessed 2018-02-06).
- [16] N. Campbell, “Speech & expression; The value of a longitudinal corpus,” *Proc. LREC 2004*, pp. 183–186 (2004).
- [17] C. T. Ishi, H. Ishiguro and N. Hagita, “Analysis of inter- and intra-speaker variability of head motions during spoken dialogue,” *Proc. AVSP 2008*, pp. 37–42 (2008).
- [18] S. Kiritani, “High-speed digital image recording for observing vocal fold vibration,” in *Voice Quality Measurement*, R. D. Kent and M. J. Ball, Eds. (Singular Publishing Group, San Diego, 2000), pp. 269–283.
- [19] I. Maddieson and P. Ladefoged, ““Tense” and “lax” in four minority languages of China,” *J. Phon.*, **13**, 433–454 (1985).
- [20] D. G. Childers and C. K. Lee, “Voice quality factors: Analysis, synthesis, and perception,” *J. Acoust. Soc. Am.*, **90**, 2394–2410 (1991).
- [21] H. Yokonishi, H. Imagawa, K.-I. Sakakibara, A. Yamauchi, T. Nito, T. Yamasoba and N. Tayama, “Relationship of various open quotients with acoustic property, phonation types, fundamental frequency, and intensity,” *J. Voice*, **30**, 145–157 (2016).

Carlos T. Ishi received the Ph.D. degree in engineering from The University of Tokyo in 2001. He worked at the JST/CREST Expressive Speech Processing Project from 2002 to 2004 at ATR. He joined ATR Intelligent Robotics and Communication Labs, since 2005, and is currently the group leader of the Dept. of Sound Environment Intelligence at ATR Hiroshi Ishiguro Labs.

Jun Arai received the B.E. degree from Bunkyo University in 1999, and the M.A. and Ph.D. degrees from The University of Kobe in 2008 and 2011. He taught Japanese language in several universities at China, Sri Lanka and Czech Republic, during 1999–2006 and 2013–2017. Recently he joined The Japan Foundation Japanese-Language Institute, Kansai.