ACOUSTIC AND EGG ANALYSIS OF PRESSED PHONATION

Carlos Toshinori Ishi, Hiroshi Ishiguro, Norihiro Hagita

Intelligent Robotics and Communication Labs., ATR, Kyoto, Japan
carlos@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

ABSTRACT

Pressed phonation ("rikimi" in Japanese) is a voice quality that carries important paralinguistic information of expressivity in the emotional or attitudinal states of the speaker. Analysis of pressed voice samples extracted from natural conversational speech firstly shows that irregularity in periodicity (such as in vocal fry and harsh voices) is a common but not a strictly determinant feature of pressed voices. Spectral analysis shows that parameters related with spectral slope are effective to identify part of the pressed voice samples, but fail when vowels are nasalized or double-beating occurs within a glottal cycle. Temporal analyses of speech and EGG waveforms indicate that information about the completely closed period can potentially be used for pressed voice identification.

Keywords: pressed phonation, voice quality, EGG, vocal fry, paralinguistic information.

1. INTRODUCTION

It is known that voice quality features due to the vibration mode of the vocal folds (e.g., breathy, whispery, vocal fry or creaky, and harsh voices [1], [10]) take important roles in the communication of paralinguistic (non-verbal) information, besides prosodic features [4], [6], [9]. It has been recently reported that a pressed voice quality called "rikimi" in Japanese [12] takes important roles in the expression of emotional or attitudinal state of the speaker. For example, [12] reports that speaker's sincerity is conveyed by pressed voice, when expressing items like:

- Emotions or attitudes (like surprise, admiration and disgust): "**hee**, sugoine" ("that's really great!"), "**yaa**, anta kandenaine" ("you are not really chewing!"), "**waa**, kimochi warui" ("that's really disgusting!").
- Qualitative adjectives: "**kirei** de,…" ("really beautiful!"), "**umainaa**" ("really delicious!"), "**kimochiwarii**" ("really disgusting!"), "**tsumetaai**" ("really cold!"), "**watashi are kyoufu yawa**" ("I'm really scared!").
- Onomatopoeia: "gibusu o **gaa** tto kitte" ("he cut the gibbs!"
- Hesitation: "**nnn**… mayoimasuwa" ("I'm really uncertain…"), "**fukuwa**… kirarenai" ("I really can't wear it…").
- Politeness, modesty: "**yaa korewa chotto yappari yoku nai to omoimasu**" ("I really think that's not good…"), "**sumimasenga**, chotto yurushite kudasaiyo" ("I'm really sorry, please forgive me…").

Although such pressed voices are easily identified by auditory impression, a clear definition based on acoustic features and the production mechanisms are unclear.

In the present work, aiming an acoustic characterization and automatic detection of pressed voices, acoustic features related to periodicity and spectral properties are investigated in pressed voice segments extracted from natural conversations. Electro-glottographic (EGG) analyses are also conducted to analyze the vibration patterns of the vocal folds during pressed phonation.

2. SPEECH DATA

Most works dealing with voice quality use the stationary portion of specific voice qualities consciously produced by subjects. However, although pressed phonation frequently occurs in natural conversation, most subjects can not produce it consciously. Thus, for the acoustic analysis, we use natural conversational speech data, where pressed phonation was unconsciously produced. Utterances containing pressed phonation were selected from the conversational speech database recorded in the JST/CREST ESP Project [7]. The dataset for acoustic analysis consists of 15 conversational passages (KoB001 ~ 015) including voices of 3 male and 5 female speakers, aging from 10's to 60's [14]. This dataset is the same used in [12] for studying the functional properties of pressed voice in speech communication.

For the EGG analysis, simultaneous speech and EGG waveforms were recorded by one male subject who was able to utter the same utterances of the natural conversational database with voice

qualities similar with the original utterances. Isolated vowels uttered in several voice qualities (e.g., modal phonation, non-pressed fry, pressed fry, periodic pressed voice) were also recorded. The EGG device used in the present experiment is the EG2-PC of Glottal Enterprises.

3. ACOUSTIC ANALYSIS

3.1. Periodicity

Irregularity in periodicity has been reported as one characteristic of pressed voices [12]. In the analysis dataset, we observed that most of pressed voice segments have acoustic features similar with creaky voice or vocal fry (very low fundamental frequencies with discrete glottal pulses, and eventual irregularity in periodicity, found in 11 of the 15 passages). In 4 of the passages, the pressed voice segments have acoustic features similar with harsh voices (noisy rasping sound, with aperiodic glottal pulses). However, pressed voice was also perceived in 3 segments with no particular irregularity in periodicity. This indicates that although pressed phonation is usually accompanied by voice qualities with irregularities in periodicity, it can also be accompanied by periodic phonations.

3.2. Spectral tilt measures

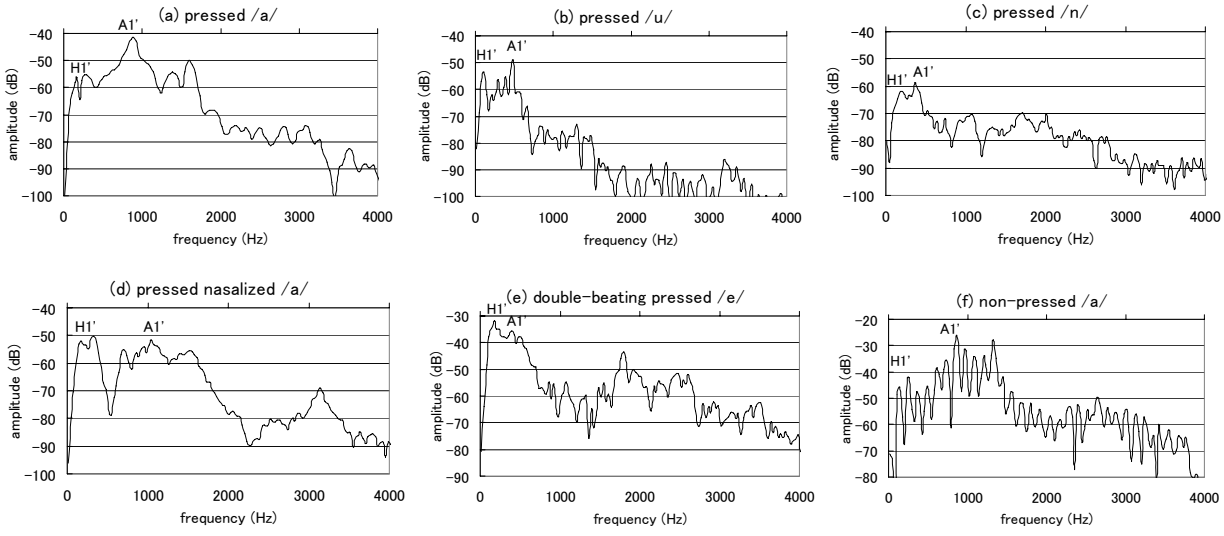
Spectral tilt is a commonly used feature for characterizing voice qualities. [11] reports that spectral tilt is effective for discriminating "tense voice" from "lax voice". As pressed phonation has a tense voice quality, it is thought that spectral tilt can potentially discriminate it from other voice qualities. Classic 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) [5], [11].

A problem of applying such measures for pressed phonation 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 1200 Hz (A1'), where the first formant is likely to occur. For periodic signals, H1' = H1, and A1' = A1.

The measure H1'-A1' was evaluated for the analysis dataset. In 8 passages including vocal fry, 2 passages including harsh voices and 1 periodic segment, H1'-A1' < -10dB was observed in the pressed segments. In almost all non-pressed segments H1'-A1' > -10dB was observed. Fig. 1(a) shows an example of pressed segment, where H1'-A1' is about -15 dB.

Detailed analysis was conducted on pressed segments where spectral tilt measures are problematic, i.e., H1'-A1' > -10dB. In /i/ or /u/ (Fig. 2(b)), where the first formant frequency is low, the difference between the H1' and A1' tends to be smaller than in other vowels. For nasals, a nasal formant appear in the range of 100 to 300 Hz, reducing the difference between H1' and A1' (Fig. 1(c)). Similar behavior happens in nasalized vowels (Fig. 1(d)). A strong amplitude component in the low frequency range was also observed when

Figure 1: Examples of spectrums of pressed and non-pressed intervals: a) pressed interval fry, H1'–A1' < -10dB; b) ~ e) pressed intervals, H1'–A1' > -10dB; f) non-pressed interval, H1'-A1' > -10 dB.



a double-beating pattern occurred in the glottal cycles, causing problems in the H1'-A1' measure (Fig. 1(e)). This double-beating pattern observed in pressed fry voices is investigated in more detail in the EGG analysis of Section 4, in contrast with a double-periodic pattern [3],[8] which is often observed in non-pressed voices.

In non-pressed segments, although almost all samples showed H1'-A1' > -10dB, some samples of vowels /a/ and /o/ showed H1'-A1' < -10 dB (Fig. 1(f)). Although these segments have a more "tense" voice quality than the other non-pressed segments, they are not perceived as pressed voice. However, as such segments are short in duration (less than 200 ms), it is thought that some duration is also necessary for the perception of pressed voice. Detailed analyses are left for future work.

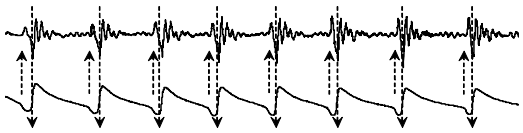
4. EGG ANALYSIS

Fig. 2 shows speech and EGG waveforms of pressed and non-pressed phonations accompanied by several voice qualities. The length of the segments is 200 ms for Fig. 2(a) ~ (f), and 100 ms for Fig. 2(g) ~ (h). Regarding the interpretation of the EGG waveforms, the "mountains" represent maximum vocal fold contact (or maximum glottal closure), while the "valleys" represent minimum vocal fold contact (or maximum glottal opening). Upward/downward arrows indicate approximate instants of glottal opening/closing. These instants were hand-annotated by visual inspection of the waveforms. All waveforms of Fig. 2 are vowel /e/ portions of the same male speaker.

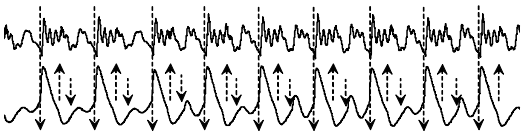
Fig. 2(a) shows a pressed fry segment with single pulses. The interval between the upward and downward arrows corresponds to the open phase of a glottal cycle. It can be noted that the open phase is much shorter than the close phase. On the other hand, the non-pressed fry segment Fig. 2(b) shows the presence of small pulses between the large pulses in the speech waveform. This period-doubled pattern is also evident in the EGG waveform. This indicates that the vocal folds do not close completely after the opening from a completely-closed phase [8].

In the double-beating pressed fry segment of Fig. 2(c), it seems that there are two pulses in a glottal cycle by only looking at the speech waveform. However, it is clear that the glottal pulses are not doubled in the EGG waveform, being more similar with the ones of Fig. 2(a). It is noted that the first "pulse" is due to a positive

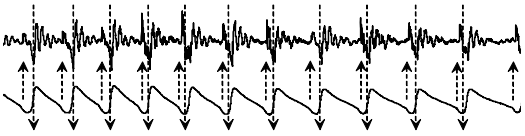
Figure 2: Speech and EGG waveforms for 1 male speaker ((a) ~ (f): 200 ms segments; (g) ~ (h): 100 ms segments). Mountains in the EGG waveforms indicate glottal closure, while valleys indicate glottal openings. Downward arrows indicate negative excitations, upward arrows indicate positive excitations. Small downward arrows indicate weak negative excitations (insufficient glottal closure).



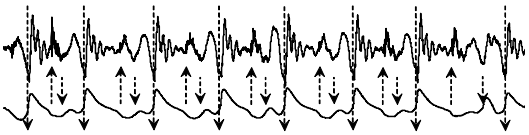
(a) (Single pulse) pressed fry



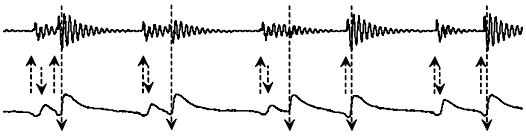
(b) Double-periodic non-pressed fry



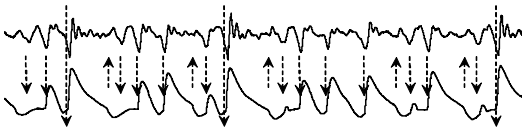
(c) Double-beating pressed fry



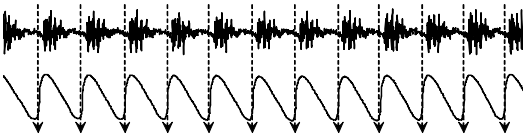
(d) Double-periodic non-pressed fry with strong positive excitations



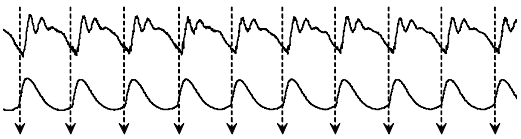
(e) Aperiodic pressed fry



(f) Aperiodic non-pressed fry



(g) Periodic pressed voice



(h) (Periodic) modal voice

excitation due to a sharp glottal opening, while the second "pulse" having a longer closed period is due to the usual negative excitation. It is worth to mention that this pattern is different from a "dicrotic dysphonia" [13] or "dicrotic pattern" [2], where two opening and closing movements occur in rapid succession followed by a relatively longer period of closure.

The non-pressed fry segment in Fig. 2(d) shows a speech waveform similar with the double-periodic pattern of Fig. 2(b). However, the speech waveform shapes between the large and small pulses are less similar in Fig. 2(d), than in Fig. 2(b). The EGG waveforms show that this is due to a stronger positive excitation in the small pulses of Fig. 2(d). Nonetheless, the EGG waveforms reveal that both are double-periodic patterns.

Fig. 2(e) and (f) show aperiodic pressed and non-pressed fry segments. In both segments, small pulses (insufficient closures) eventually appear between the complete closure segments. However, the pulse shapes are similar to the pressed and non-pressed segments of Fig. 2(a) and (b), respectively.

Pressed phonation accompanied by fundamental frequencies with similar range to modal phonation is shown in Fig. 2(g), and a non-pressed modal segment is shown in Fig. 2(h). The F0 of these examples are ranged between 100 and 120 Hz. The positive excitations are not strong compared with the previously presented fry segments. But from the EGG waveforms, it can be observed that the widths of the EGG pulses are wider in the pressed segment, which means that the closed phase is longer than the open phase.

Finally, although the EGG analysis above presented results only for the vowel /e/, similar waveforms were obtained for the other vowels. It is also worth to mention that although measures like open quotient (OQ) and speed quotient (SQ) are often used to characterize glottal waveform shapes, we avoided to use such measures since it is unclear if the incomplete closures should be included in the open phases or be treated as closed phases.

5. CONCLUSION

The main findings of the present work are as follows: 1) pressed voice is usually accompanied by vocal fry or harsh voices, having irregularities in periodicity, but it can also be accompanied by periodic voices; 2) spectral tilt measures (which are closely related with tense/lax voice properties) are partly effective for identifying pressed voices, but have problems when nasalization or double-beating in the glottal excitation occur; 3) single glottal pulses tend to occur in pressed fry, while double-periodic pulses often occur in non-pressed fry; 4) EGG analysis showed that regardless the periodicity, the open period is shorter than the completely closed period in pressed voice, while vibrations with incomplete closure often occurs in non-pressed fry.

From the above results, we can infer that the voice quality which is perceived as "pressed voice" or "rikimi" in Japanese, is a phonation type where, regardless irregularities in periodicity, the completely closed intervals of the vocal folds are predominant to the open intervals plus eventual incompletely closed intervals.

Future works are to investigate acoustic features that correspond to the EGG features, aiming for automatic detection of pressed phonation, and subsequent paralinguistic information extraction.

6. ACKNOWLEDGEMENTS

This work was partly supported by the Ministry of Internal Affairs and Communications.

7. REFERENCES

[1] Catford, J., 1977. *Fundamental Problems in Phonetics*, Edinburgh: Edinburgh Univ. Press, 98-105.

[2] Cavallo, S.A., Baken, R.J., Shaiman, S., 1984. Frequency perturbation characteristics of pulse register phonation. *J. Commun. Desord.* 17, 231-243.

[3] Gerratt, B. R., Kreiman, J., 2001. Toward a taxonomy of nonmodal phonation. *J. of Phonetics* 29, 365-381.

[4] Gobl, C.; Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40, 189-212.

[5] Gordon, M., Ladefoged, P., 2001. Phonation types: a cross- linguistic overview. *J. of Phonetics* 29, 383-406.

[6] Ishi, C.T., Ishiguro, H., Hagita, N., 2006. Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction. *CD-ROM Proc. Speech Prosody 2006*.

[7] JST/CREST ESP Project. http://feast.atr.jp/ visited 28-Feb-07

[8] Kiritani, S., 2000. High-speed digital image recording for observing vocal fold vibration. In: Kent, R.D., Ball, M.J., (eds), *Voice Quality Measurement*. San Diego: Singular Publishing Group, 269-283.

[9] Klasmeyer, G.; Sendlmeier, W. F., 2000. Voice and Emotional States. In *Voice Quality Measurement*, Singular Thomson Learning. 339-358.

[10] Laver, J., 1980. Phonatory settings. In: *The Phonetic Description of Voice Quality*, Cambridge: Cambridge Univ. Press, 93 - 135.

[11] Maddieson, I., Ladefoged, P., 1985. "Tense" and "lax" in four minority languages of China. *J. Phonetics* 13, 433-454.

[12] Sadanobu, T., 2004. A Natural History of Japanese Pressed Voice, *J. of Phonetic Society of Japan*, Vol. 8 (1): 29-44.

[13] Zemlin, W.R., 1998. Speech and Hearing Science – Anatomy and Physiology, Allyn and Bacon, 169.

[14] http://www.irc.atr.jp/~carlos/vocalfry/ visited 28-Feb-07