Acoustic, Electroglottographic and Paralinguistic Analyses of “Rikimi” in Expressive Speech

Carlos T. 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

“Rikimi” is a “pressed-type” voice quality that appears in Japanese conversational speech for expressing paralinguistic information related to emotional or attitudinal behaviors of the speaker. We conducted acoustic, electroglottographic (EGG) and paralinguistic analyses on speech segments including “rikimi”, extracted from spontaneous dialogue speech data. “Rikimi” may be accompanied by several voice qualities (such as creaky or harsh), but vocal fold vibratory pattern analyses based on the EGG signals indicated that a common feature was found in the relation between overall open and closed intervals, in comparison to non-“rikimi” segments. Spectral analyses show that parameters related with spectral tilt are effective to identify part of the “rikimi” segments, but fail when vowels are nasalized. F0 contour analysis showed that a dip occurs during “rikimi” segments, but a change in voice quality is prominently perceived rather than a change in the intonational curve. Linguistic contents are also found to influence the perception of “rikimi” in the conveyance of paralinguistic information.

Index Terms: pressed voice, voice quality, EGG, expressive speech, prosody

1. Introduction

It is known that voice quality features due to different vibration modes of the vocal folds (e.g., breathy, whispery, creaky or vocal fry, and harsh voices [1]-[3]) have important roles in the communication of paralinguistic (non-verbal) information, besides intonation-related prosodic features [4]-[7]. It has been reported that a pressed-type voice quality, called “rikimi” in Japanese [8], takes important roles in the expression of emotional or attitudinal states of the speaker in Japanese [8,9]. For example, it is reported that speaker’s expressivity is conveyed by “rikimi”, in items like: emotions or attitudes (like surprise, admiration and disgust), real feeling expression in adjectives, onomatopoeia, hesitation, and modesty.

The auditory impression of “rikimi” is clearly different from the modal (normal) phonation. However, a clear definition of “rikimi” in terms of acoustic or auditory features does not exist, and its production mechanisms are also unclear. Although there is a categorization of laryngeal voice qualities in modal, breathy, whispery, creaky, harsh and falsetto [1], the “rikimi” quality does not fit exactly to the descriptions of any of them, but rather may co-exist with some of them [9]. Most of “rikimi” segments have been found to have acoustic features similar with creaky voice or vocal fry (impulse-like glottal excitation with very low fundamental frequencies, usually accompanied by irregularity in periodicity). Some “rikimi” segments have been found to have acoustic features similar with harsh voice (noisy rasping sound, with aperiodic glottal pulses). However, “rikimi” has also been found in periodic segments, where the fundamental frequency (F0) does not become as low as in vocal fry, and periodicity is not irregular as in harsh voice [9].

At the present stage, “rikimi” can be defined as a voice produced by “pressing” (“straining”) the vocal folds, having a perceptual sensation of a tense/strident voice quality. The auditory impression can be better understood by listening to the speech samples including “rikimi”, in our webpage [10]. Although the term “pressed voice” has been used as an English translation of “rikimi” [9], we decided to keep the original Japanese term in the present paper, to avoid confusion with a “true pressed voice”, in terms of physiological fundaments. However, along with the analysis, this paper will also speculate the production mechanisms of “rikimi”, i.e. what is “pressed” in the vocal folds.

In our previous work [9], acoustic features related to periodicity and spectral properties were investigated in “rikimi” segments extracted from natural conversations of several speakers. Electroglottographic (EGG) analyses have also been conducted to analyze the vibration patterns of the vocal folds during “rikimi”. However, EGG data was available for only one speaker, who could imitate the “rikimi” of the utterances extracted from natural conversations.

In the present work, aiming at a better acoustic characterization, we used a multi-modal dialogue speech database, and conducted acoustic, EGG and paralinguistic analyses of “rikimi” segments appearing in natural dialogues of several speakers.

2. Analysis Data

Most work dealing with voice quality uses the stationary portion of specific voice qualities consciously produced by subjects. However, although “rikimi” frequently occurs in expressive utterances of natural conversations, most subjects are not able to produce it in a conscious manner. Thus, for the acoustic analysis, we use natural conversational speech data, where “rikimi” is unconsciously produced.

In the present work, we used our multimodal dialogue speech database [11], which contains free Japanese conversations of 10 to 15 minutes, recorded for several pairs of subjects, and where several multimodal signals are simultaneously recorded for each dialogue partner. For the present analysis, we used the speech and EGG (electro-glottographic) signals.

The microphones used for the recordings are Sanken CS-1 directional microphones. The EGG device is the EG2-PC of Glottal Enterprises. All waveforms were sampled at 16 kHz, 16 bits. Spectral subtraction was conducted in all waveforms to reduce stationary background noises. (However, the degree of subtraction was controlled in order not to distort much the speech signals.) The dialogue partners sit in front of each other, separated by approximately 1 meter, so that there is an inter-channel leakage in the speech signals. A cross-channel time-frequency binary masking was then conducted in the speech signals, in order to reduce inter-channel leakages. High-pass filtering with a cut-off frequency around 60 Hz was applied to all waveforms, in order to remove DC and undesirable low frequency movements.

46 dialogues including 10 female (whose ID and ages are FYU (4), FFS (6), FSF (15), FMH (30s), FKN (30s), FMU (30s), FKI (30s), FYS (30s), FKH (50s), FHT (60s), and 9 male speakers (MTI (4), MTT (17), MFT (17), MYM (20s), MSR (20s), MIT (30s), MMS (30s), MSN (30s), MHI (40s)) where each speaker participates in three or four dialogues were used for analysis.

Utterances containing "rikimi" were manually selected by three subjects (native speakers of Japanese), by listening to the speech utterances in the dialogue speech database. The tokens where all three subjects agreed were used for the subsequent analysis.

"Rikimi" was found in a wide range of ages, regardless of gender. The numbers of identified segments for each subject are as follows: FYU (1), FFS (0), FSF (0), FMH (21), FKN (21), FMU (4), FKI (3), FYS (2), FKH (12), FHT (7), MTI (0), MTT (5), MFT (3), MYM (5), MSR (13), MIT (4), MMS (0), MSN (1), and MHI (0).

The found "rikimi" segments could be categorized in the following items:

- Interjections: "waa" (I'm *really* impressed), "eee" (I'm *really* hesitated), "eee" (I'm *really* dissatisfied), "hee" (I'm *really* surprised, I'm *really* impressed), "nnn" (I'm *really* embarrassed).
- Adjectives and adverbs: "kawaii" (*really* cute!), "zutto" (for a *really* long time!), "taihen" (*really* serious!), "mottainai" (*really* wasteful!), "itai" (*really* painful!).
- Onomatopoeia: "uwaa"/"bwaa" (a lot, fast), "gaan" (with vigor), "kaaa" (with vigor, energetically), "pyun pyun" (jumping up and down), "nyee" (baby's strident crying sound), "aaa" (excited crying sound expressing strong sadness).
- Utterance quoting: "rikimi" frequently appeared when the speaker quoted his/her past utterance, or an utterance of another person. The expressivity conveyed by quoted utterances with "rikimi" was of excitement or hesitation. "kora!" (Hey!), "moo iikagen ni shii!" ("No more!")

3. Analysis Results

3.1. Vocal fold vibratory patterns: EGG (electro-glottographic) features

Fig. 1 shows speech, EGG and DEGG (derivative of the EGG) waveforms of representative samples of "rikimi" and non-"rikimi" voices (accompanied by several voice qualities) found in our dataset. 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 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 [12]. This way, open intervals can be estimated by the intervals between positive and successive negative peaks in the DEGG waveform, while closed intervals can be estimated by the pulse intervals minus the open intervals.

Fig. 1 (a) shows an example of modal voice (normal phonation) of a female speaker (FMU), for reference. Fig. 1 (b) shows a typical example of a single-pulsed "rikimi" creaky voice of the same speaker, where F0 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 "rikimi" segment of Fig. 1(b).

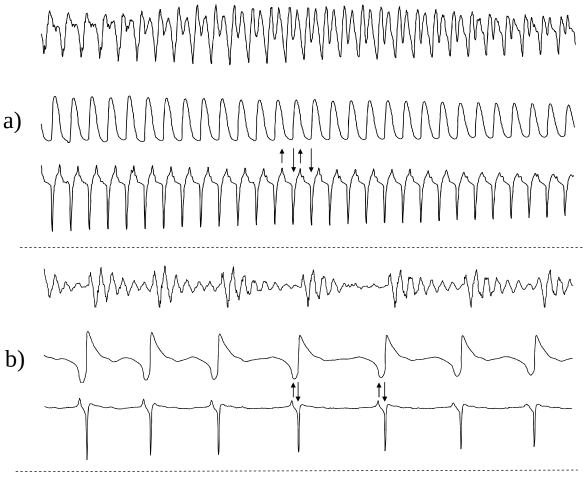
Fig. 1 (c) and (d) show examples of "rikimi" and non-"rikimi" creaky segments of a male speaker (MTT). Both signals have small pulses between large pulses in the EGG waveforms, corresponding to incomplete closures between complete closures. The difference between these signals is that the overall open intervals are smaller than the closed intervals in the "rikimi" creaky segment. Similar behavior was observed in other speakers. Although creaky voice (or vocal fry) has been defined as having a brief glottal excitation pulse followed by a relatively longer closed interval [1], the example of Fig. 1 (d) shows that this definition is not true for non-"rikimi" (lax) creaky segments.

Fig. 1 (e) shows an example of "rikimi" segment with no special irregularity in periodicity, and having F0 ranges above 100 Hz, so that individual pulses cannot be perceived as in creaky voice. Similarly to the previous "rikimi" examples, the open intervals are relatively shorter than the closed intervals.

Figure 1(f) shows an example of a "rikimi" token with a diplophonic quality, where multiple phonations can be simultaneously perceived [13]. It can be noted that larger pulses with stronger negative DEGG peaks occur with larger inter-pulse intervals (lower F0, around 110 Hz), while smaller pulses occur with smaller inter-pulse intervals (higher F0, around 330 Hz).

A common feature found for "rikimi" is that the overall open intervals are much shorter in duration than the overall closed intervals.

However, this feature was not so evident in a "rikimi" token with a rough/noisy voice quality, as in the example shown in Fig. 1(g). Positive peaks are broader and less sharp than the other "rikimi" examples. A correspondence between speech and EGG waveforms is also unclear, so that even though the EGG signal reveals an F0 close to 150 Hz, this F0 component is almost imperceptible in the speech signal. It is questionable if this type of voice should be considered as "pressed" from a production viewpoint.



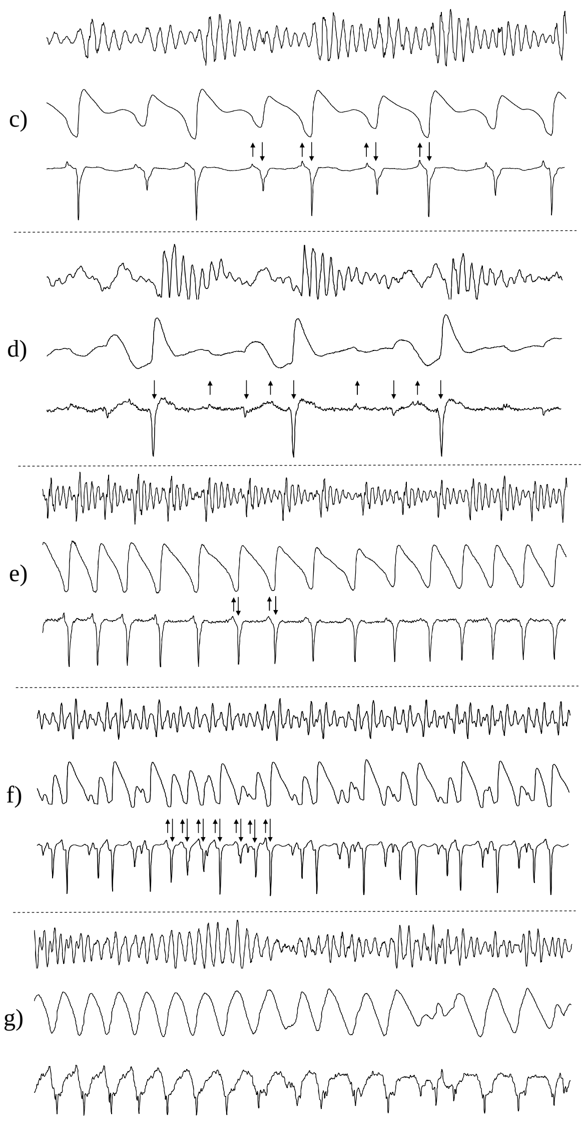


Figure 1: *Speech, EGG and DEGG waveforms, for several types of vibratory patterns in "rikimi" and non-"rikimi" segments. Segment lengths are 100 ms. a) modal voice (ky[ua], FMU) b) "rikimi" creak (z[uu]tto, FMU); c) "rikimi" creak (w[a]a, MTT); d) non-"rikimi" creaky (katt[ara]na, MTT); e) "rikimi" (e[e]e, FYU); f) "rikimi" diplophonic (k[awaii], MYM); g) "rikimi" harsh (b[wa]a, MFT).*

It is worth to mention that although there are measures like open quotient (OQ) and speed quotient (SQ) for characterizing glottal waveforms [12], we didn't compute these measures because it is not clear how to deal with the incomplete closures.

3.2. Periodicity and voice qualities

Irregularity in periodicity has been reported as one characteristic of "rikimi" [8]. In our previous work [9], we stated that "rikimi" was usually accompanied by voice qualities having irregularities in periodicity (creaky and harsh voices), but also appeared with periodic pulses with F0s in the range of modal phonation.

Fig. 2 shows F0 contours, spectrograms, speech and EGG waveforms for speech segments containing "rikimi". We can observe that the transitions between modal and "rikimi" usually occur accompanied with a fast but continuous (gradual) decrease/increase in F0, as shown in the examples of Fig. 2 (b), (c) and (e), but may also occur with an F0 jump in diplophonic signals, as in the example in Fig. 2(d).

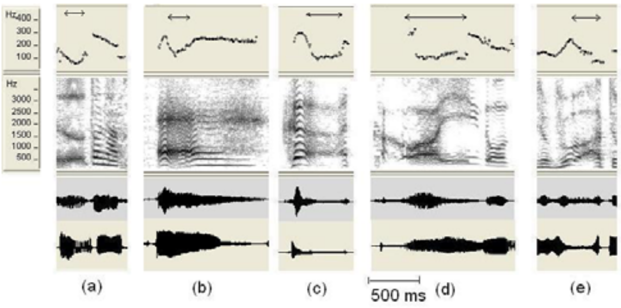


Figure 2: *F0 contours, spectrograms, speech and EGG waveforms for utterances including "rikimi" segments (indicated in the top panels). a) "rikimi" creak (z[uu]tto, FMU); b) "rikimi" (e[e]ee, FYU); c) "rikimi" (ga[aa]a, FHT); d) "rikimi" diplophonic (k[awaii], MYM); e) "rikimi" harsh (b[wa]a, MFT)*

This F0 dip arises as a consequence of the vocal fold tensions during "rikimi", but an interesting feature is that a change in voice quality is more prominently perceived rather than a change in the intonational curve. Thus, even though the F0 curve is continuous without a pulse-to-pulse irregularity, it can be said that there is irregularity in periodicity, in the sense that the F0 curve is not perceived as an intonational curve.

3.3. Spectral features

Spectral tilt is a commonly used feature for characterizing voice qualities. In [14], spectral tilt is reported to be effective for discriminating "tense voice" from "lax voice". As "rikimi" has a tense voice quality, it is thought that spectral tilt can potentially discriminate it from other voice qualities. In fact, the spectrograms in Fig. 2 show that the power of the frequency components of the lower harmonics is reduced in all "rikimi"segments.

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) [4], [14].

A problem of applying such measures for "rikimi" 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 have proposed in our previous work [9] 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 be present. For periodic signals, H1' = H1, and A1' = A1.

A threshold of 10dB (H1'–A1' < 10dB) was effective to identify almost all "rikimi" segments of the present dataset. However, misdetections were found in nasalized segments, where the nasal formant increases the power of lower frequency components, decreasing the H1'–A1' measure. Also, insertion errors were found in non-"rikimi" segments mainly in /o/, where the first formant is high.

Although it is clear that a reduction of the lower frequency components occurs in "rikimi" segments, these spectral tilt related parameters are not enough for automatic detection purposes.

3.4. Power

It was observed that the power of the speech signal in "rikimi" segments is smaller than that in modal voice segments, by about 10 dB on average, ranging from 3 to 20 dB. This decrease in power is mainly due to the reduction of power components of the lower harmonics, as stated in the previous sub-section.

It is interesting to note that "rikimi" is an effective strategy for expressing strong emotions and attitudes, by using less power than in normal phonations.

3.5. Perception of "rikimi": Effects of duration, position within the utterance, and linguistic contents

In many segments which were not identified as "rikimi", acoustic and EGG features similar to that of "rikimi" were observed. Nonetheless, most of these segments were short in duration. Thus, it is thought that a minimum duration is necessary for the perception of "rikimi" in continuous speech.

The average length of "rikimi" segments was about 400 ms, and the minimum duration found in the present data was about 150 ms. However, many segments not classified as "rikimi" and having durations larger than 150 ms were found as having "rikimi" features.

We then conducted a detailed analysis, and firstly found that, many segments not classified as "rikimi" were at the boundary of phrases, where pitch is lowered, appearing mostly in male speakers. In these cases, it was also noted that most of the boundary words were particles, auxiliary verbs, or anastrophes (subject-verb inversions). Segments with "rikimi" features were also found in disfluent utterances, but these were not classified as "rikimi", probably because they don't convey specific expressivity.

Thus, it is thought that linguistic information (position in the utterance, and morpho-syntactic content) also affects the perception of "rikimi" as a paralinguistic function. The paralinguistic functions found in Section 2 show that expressivity is conveyed in interjections, adjectives, adverbs, and onomatopoeia. In the case of utterance quoting, "rikimi" may appear over whole utterances, but usually including some of the above items.

It was also noted that "rikimi" tend to appear in the accent nucleus syllables of the words.

4. Discussion

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

We also observed that in almost all "rikimi" segments, the EGG waveform amplitudes are smaller than in the neighboring modal segments, as in the examples of Fig. 2. This probably indicates that the vocal fold vibrations are changing from full-glottal to anterior phonation, where only the anterior part of the glottis (ligamental glottis) is vibrating, so that the variations in the glottal impedance (measured by the EGG signals) become smaller in "rikimi" segments. We observed that the amplitude of the EGG waveforms don't change much between modal and non-"rikimi" creaky voice, so that the full-length glottis would be vibrating. High-speed images would help to verify these hypotheses.

5. Conclusion

"Rikimi" was found in a wide range of ages, regardless of gender, for expressing a variety of emotional or attitudinal behaviors of the speaker.

EGG analysis revealed that "rikimi" may occur along with several other voice qualities, like creaky, harsh and diplophonic, but a common feature was that the glottal open intervals are relatively smaller than the closed intervals.

Periodicity analysis revealed that the transitions between modal to "rikimi" are often accompanied by a fast and gradual F0 lowering, but such F0 curves are not perceived as intonational curves.

Spectrographic analysis revealed a reduction of energy in the lower frequency components. Spectral tilt measures were partly effective for identifying "rikimi", but have problems when nasalization occurs.

The segmental duration, position within the utterance and linguistic contents were found to potentially influence the perception of "rikimi" as an expressivity conveyer.

Future works are to investigate more robust acoustic features that better reflect the vibratory pattern features of "rikimi", and evaluate a subsequent paralinguistic information extraction.

6. Acknowledgements

This work was partly supported by the Ministry of Internal Affairs and Communications (MIC), and the Ministry of Education, Culture, Sports, Science and Technology (MEXT). The authors thank Yuri Suzuki and Jun Arai for helping with the analysis.

7. References

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

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

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

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

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

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

[7] 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.

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

[9] Ishi, C.T., Ishiguro, H., and Hagita, N., 2007. Acoustic and EGG analysis of pressed phonation, Proc. of International Conference on Phonetic Sciences (ICPhS'2007), 2057-2060.

[10] http://www.irc.atr.jp/~carlos/vocalfry/ visited 17-Apr-09

[11] Ishi, C.T., Ishiguro, H., and Hagita, N., 2008. Analysis of inter- and intra-speaker variability of head motions during spoken dialogue, Proc. of AVSP' 2008, 37-42.

[12] Childers, D.G., Lee, C.K., 1991. Voice quality factors: Analysis, synthesis, and perception. J. Acoust. Soc. Am. 90(5), Nov. 1991., 2394-2410.

[13] 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.

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