Study on strained rough voice as a conveyer of rage

Yumiko O. Kato¹, Yoshifumi Hirose¹, Takahiro Kamai¹

¹Intelligent Processing Technology Research Laboratory, Advanced Technology Research Laboratories, Matsushita Electric Industrial Co., Ltd., Kyoto, Japan

{kato.yumiko, hirose.yoshifumi, kamai.takahiro}@jp.panasonic.com

Abstract

It is important to be able to determine anger and its degree for dialog management in an interactive speech interface. We investigated the characteristics of a strained rough voice as a conveyer of a speaker’s degree of anger. In hot anger speech in Japanese, a rough voice with high glottal tension is observed frequently, and the rate of occurrence of the strained rough voice increases according to the degree of anger. In a typical male speaker’s speech sample, amplitude fluctuations observed in a strained rough voice were periodic; the frequency was around between 40–80 Hz. The modulation ratio in rage speech was larger than in other emotional states, suggesting the possibility of determining the speaker’s anger and its degree by detecting strained rough voice.

Index Terms: voice quality, strained rough, anger, Japanese

1. Introduction

In an interactive speech interface, such as an automatic telephone answering system, it is important to perceive the user’s emotion based on their speech to respond to their request more appropriately. If the interactive system fails to recognize the user’s speech correctly, it may request the user to input the speech again, making the user more angry or frustrated. Anger or frustration causes the user’s voice quality or prosody to change, which makes the interactive system more prone to false recognition. To stop this vicious circle, it is necessary to recognize the user’s anger and its degree, and to manage the dialog accordingly. Earlier studies revealed that a harsh voice and tense voice are associated with anger [1, 2, 3, 4]. In native Japanese expressive speech, a strained rough voice is observed in anger and rage speech [5]. In the present study, we investigated phonetic and acoustical characteristics of strained rough voice which are associated with anger and rage. Our objective is to establish a basis for determining the user’s anger degree, which is an important problem of speech interfaces.

2. Phonetic characteristics of strained rough voice

2.1. Strained rough voice

A rough voice is an uneven, bumpy sound related to fluctuating amplitude and a varying fundamental frequency from cycle to cycle. A strained voice is effortful; the hyperfunction of neck muscles and/or the entire larynx might be compressed [6, 7]. Previous studies found that many rough and strained voices exhibited ‘rikimi’ in native Japanese expressive speech. This voice quality is observed in highly tense emotional states, especially rage, which shows explosive hot anger.

2.2. Strained rough voice in various emotional states

To investigate the strained rough voice as a conveyer of rage in Japanese, eight native Japanese professional actors and actresses (four males M1-M4 and four females F1-F4) spoke 50 sentences in a conversational style. They were asked to speak each sentence in three emotional states: neutral, anger, and rage. Speech was recorded in a sound-proof room using a condenser microphone (DPA4061; DPA Microphones). Speech was digitized (EDIROL UA-1000; Roland Corp.) and recorded on a personal computer at 44.1-kHz sampling frequency and 16-bit resolution. An expert labeled phonemes and discriminated accent nuclei and accent phrases. Another expert labeled the strained rough voice based on the GRBAS standard [6], for each vowel and syllabic nasal.

Figure 1 depicts the number of morae produced in the strained rough voice under each emotional state. The occurrence rate of the strained rough voice in the anger and the rage speech was significantly higher than that in the neutral speech (Wilcoxon, p<0.05). No significant difference was found between gender groups in the occurrence rate of the strained rough voice in rage speech. However, under the rage state, two male speakers (M1 and M4) overwhelmingly produced abundant strained rough voices. Under an anger state, a significant difference was apparent between gender groups in the occurrence rate of the strained rough voice.

Figure 1: Bars show numbers of morae uttered with a strained rough voice. Each pattern shows eight speakers’ data. Data are grouped according to the emotional state.
As described above, a specific segmental and prosodic condition exists under which the strained rough voice is easily produced. Among the male speakers, the strained rough voice was produced more during the anger and the rage speech than during neutral speech (Wilcoxon, p<0.05). However, no difference was found in female speakers among the states.

Figure 2 compares the occurrence rate of the strained rough voice in each phone of each mora in rage speech. Each bar shows the number of uttered morae by each speaker; the eight bars show M1-M4 and F1-F4 from the left. Red parts show the number of morae uttered in a strained rough voice. The strained rough voice tended to be produced often in morae with consonants [t], [k], [d], [m] and [n], the palatal and alveolar stops and the nasals or without a consonant (Fig. 2, upper). Regarding vowels contained in the morae, the strained rough voice tended to be produced often in morae with vowels [a], [e], [o], except long vowels ([aː]-[uː]) and vowels following palatalized sounds ([ja]-[jo]) (Fig. 2, lower). When these phones are pronounced, the position and height of the back tongue can easily make the larynx and the pharynx tense. Regarding prosodic condition, the first mora in an intonation phrase were found to be associated with a rather high occurrence rate of the strained rough voice. Rapid control of fundamental frequency and vocal intensity might cause excessive tension in the larynx and pharynx.

2.3. Statistical prediction of the occurrence of a strained rough voice

The strained rough voice is not a timbre of all speech, but rather a local voice quality which appears in a part of speech. As described above, a specific segmental and prosodic condition exists under which the strained rough voice is easily produced.

We made a statistical model that estimates the likelihood for each mora in speech to be uttered with a strained rough voice. Using Hayashi’s quantification method II, discriminant modeling of the occurrence of a strained rough voice of the speech of male speaker M1 was executed using segmental and prosodic conditions as the explanatory variable. Quantification method II is a linear model based on the occurrence of each category to quantify categorical factors such as phones. Figure 3 is a histogram showing the score of each mora derived from applying the quantification method II to the rage speech of male speaker M1. The score from that discriminant model shows the reluctance of the production of the strained rough voice: the strained rough voice had a lower score and the other voices had higher scores. When the appearance of the strained rough voice was estimated by applying the discrimination criterion of −0.20 to the rage speech of M1, the recall ratio was 75%. The score of the point at which the cumulative frequencies of strained rough voice and other qualities showed the same value, was adopted as the discrimination criterion. The cumulative frequency of strained rough voice was calculated from higher to lower, and the cumulative frequency of other qualities was calculated from lower to higher. It is suggested that occurrence of the strained rough voice on each mora can be predicted statistically using segmental and prosodic conditions.

3. Acoustic characteristics of strained rough voice

It might be possible to detect anger if the morae uttered in strained rough voice could automatically be detected by acoustic analysis. Amplitude fluctuation is one salient acoustical characteristic of a rough voice, so we analyzed the frequency and modulation ratio of this amplitude fluctuation to specify the characteristics of a strained rough voice.

![Figure 2: Bars show number of morae included in 50 utterances in rage speech by each included phoneme. Filled parts represent the number of morae uttered with a strained rough voice; blank parts show the number of morae uttered with other qualities. Each bar signifies a speaker: from left, speaker M1, M2, M3, M4, F1, F2, F3, and F4. The upper graph shows the results of grouping the morae according to consonants. The lower graph shows the value of grouping by vowel and syllabic nasal.](image-url)
3.1. Method

We analyzed vowels and syllabic nasals in the speech samples from male speaker M1, whose speech had the most strained rough voice. Figure 4 shows an example waveform and an amplitude envelope of a strained rough voice, along with the procedure of calculating the modulation ratio of amplitude fluctuation. First, the Hilbert envelope was adopted as an accurate amplitude envelope. The Hilbert envelope was calculated from a speech waveform passed through a 1000-step FIR band-pass filter. The pass band of the FIR filter was set as 1.5–2.5 times the fundamental frequency dynamically. Absolute values of the Hilbert transform calculated from the filtered speech waveform, a sinusoidal wave, were calculated as a Hilbert envelope. A sample of the band-pass filtered speech waveform and its envelope is shown in the middle of Fig. 4. The instantaneous frequency of the Hilbert envelope curve was calculated at each sample point. After standardizing by DC component subtraction and smoothing by low pass filtering at 100 Hz of the extracted Hilbert envelope, the instantaneous frequency of each sample point was calculated from the angular velocity of the sinusoidal wave using Hilbert transformation. The standardized envelope of the sample waveform is shown in the lower panel of Fig. 4. The modal class in the histogram of instantaneous frequency values in a vowel or syllabic nasal was adopted as the fluctuation frequency of the vowel or syllabic nasal. The estimated modulation ratio was adopted as the fluctuation depth. First, the third fitting function shown in Fig. 4 was calculated as an estimated envelope without fluctuation. The difference between the value of the extracted Hilbert envelope and the value of the third fitting function at each envelope peak was extracted as modulation. The ratio of the median of peaks in modulation to the median of approximation by third fitting was adopted as the modulation ratio.

3.2. Results

Figure 5 shows a scatter diagram of morae based on the fluctuation frequency and the modulation ratio. The mean value of fluctuation frequency in all states was 58.60 Hz; the standard deviation was 16.88 Hz. The mean of the modulation ratio in all states was 0.178; the standard deviation was 0.123. The means of fluctuation frequencies in emotional states—neutral, in anger, and in rage—were, respectively, 51.35 Hz (SD: 17.79 Hz), 55.70 Hz (14.07 Hz), and 59.57 Hz (17.45 Hz). By multiple comparison, a significant difference was found between the mean values of fluctuation frequency in anger speech and rage speech (p<0.01). The mean of the modulation ratio in each emotional state—neutral, in anger, and in rage—were, respectively, 0.10 (0.05), 0.10 (0.06), and 0.20 (0.13). The modulation ratios in rage were significantly different from those of the other two emotional states (p<0.01). The correlation coefficient between the fluctuation frequency and modulation ratio was significant (p<0.01) but small, 0.19.

The coefficient of correlation between the fundamental frequency and the fluctuation frequency was significant (p<0.01), but very small, 0.11.

4. Discussion

In native Japanese expressive speech, the occurrence of strained rough voice is prone to increase according to the degree of anger. Using segmental and prosodic parameters, the reluctance of the occurrence of strained rough voice can
The amplitude fluctuation of a strained rough voice is a problem to be studied. We analyzed the amplitude fluctuation of a strained rough voice to identify its acoustical characteristics. The modulation ratio of a strained rough voice in rage speech was significantly larger than the ratios of other emotional states. Differences of fluctuation frequency among emotional states were not so large, but the fluctuation frequency in rage was significantly larger than that in anger. Furthermore, a slight correlation was observed between the fluctuation frequency and modulation ratio, which may increase according to the anger degree. It is necessary to investigate the correspondence of the perceptual anger degree to the modulation ratio and fluctuation frequency to clarify details.

Correlation between the fluctuation frequency and fundamental frequency was small, which suggests that amplitude fluctuation of the strained rough voice is not associated with the fundamental frequency, there would be a certain level of correlation if the amplitude fluctuation resulted from diplophonia. Amplitude fluctuation is considered to result from periodic fluctuation of tissues around the larynx according to high glottal tension.

The average of the amplitude fluctuation frequency of the strained rough voice was about 60 Hz. For 80% of morae uttered with a strained rough voice, the fluctuation frequency was 40–80 Hz (Fig. 4). This fluctuation band of frequency overlaps that of roughness perception. Fluctuation around 20–150 Hz does not cause beat perception; this is the perceived timbre of roughness [9, 10]. The perception of roughness in the speech sound might be recognized as the quality of a strained rough voice. A strained rough voice appears to be one way of effectively transmitting para-linguistic or non-linguistic information by controlling acoustic characteristics corresponding to perceptual characteristics.

The results suggest that a strained rough voice is detectable by extracting amplitude fluctuation. It is necessary to study the varying fundamental frequency cycle to cycle for additional specificity.

5. Conclusion

We investigated phonetic and acoustic characteristics of a strained rough voice as a conveyer of the degree of anger. Such a voice is characterized by amplitude fluctuation and high glottal tension. The occurrence rate of a strained rough voice increased according to the anger degree. Furthermore, the reluctance to utter a strained rough voice can be estimated using a statistical prediction model. Expressive speech samples of a typical male speaker containing neutral, anger and rage states were analyzed acoustically. Amplitude fluctuations observed in the strained rough voice were periodic: their frequency was around 40–80 Hz. The modulation ratio in rage speech was larger than that of other emotional states, which suggests that it may be possible to determine the speaker’s anger by detecting a strained rough voice. A future study is needed to determine the anger degree using the reluctance estimation and detection of the strained rough voice.

6. Acknowledgements

We sincerely thank Prof. Kasuya at the International University of Health and Welfare and Prof. Kawahara at Wakayama University for their helpful suggestions and discussion.

7. References