EMOTION PERCEPTION ANALYSIS IN WHISPERED SPEECH

Yuya Matsumoto and Kazunori Mano

Graduate School of Engineering and Science, Shibaura Institute of Technology

E-mail: mf14050@shibaura-it.ac.jp, mano@sic.shibaura-it.ac.jp

ABSTRACT

This paper describes the detailed relationship of the emotions and acoustic features in the whispered speech and normal speech utterances. A listening experiment was performed by using emotional speech data of both whispered and normal utterances. The emotion types are divided into four patterns which are emotionless, joy, sorrow, and anger. As a result, it was found that the recognition rate of the whispered speech was usually lower than that of normal speech utterances. In order to clear the problems, acoustic features of intensity, duration time, formant frequencies in whispered speech in the cases of four emotional patterns were examined. In addition, we investigated relationship of the formant frequencies between whispered and normal speech and that of the pitch frequencies.

1. INTRODUCTION

Conversion technologies from whispered speech utterances to normal speech have been developed as studies for speech communication improvement in speech impairments. In addition, the conversion method from whispered speech with emotion to normal speech with emotions is also studied. Since whispered speech does not cause vocal cord vibrations, pitch (=a fundamental frequency) does not exist. However, we can hear any prosodic information of accents and intonations without vibration of vocal cords. In previous researches, it is investigated that the whispered speech is more likely to express information to be equivalent to pitch by controlling formant frequencies [1]. Furthermore, from the whispered speech, we can mostly recognize emotion in the same way as normal speech utterances. Unfortunately, there are a few researches only in discussed about the relation between emotions and acoustic features in whispered speech [2][3].

Whispered speech is a sound only generated by the vocal tract resonance caused by noise of air flow. In other words, whispered speech can be defined as a speech without periodical characteristics. Figure 1 shows two spectrograms of normal speech and whispered speech. Different from the normal speech, the whispered speech does not have pitch harmonics.

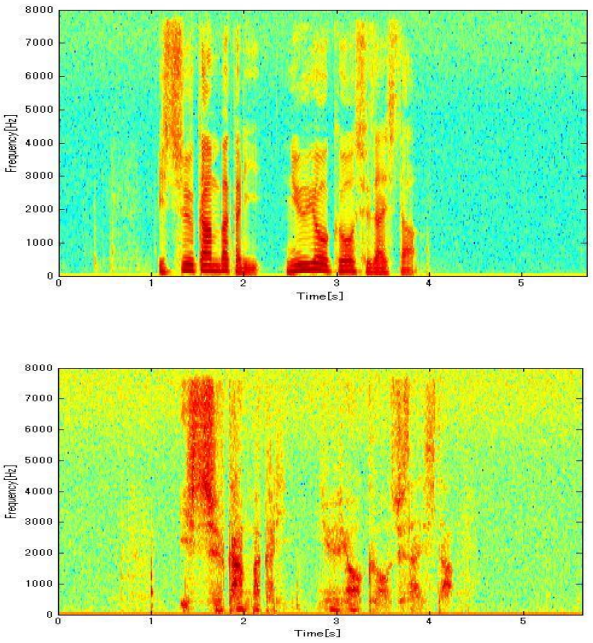

Fig. 1. Spectrograms (top : normal, bottom : whisper)

2. SOUND DATA

Normal speech utterances and whispered speech were recorded in a soundproofing room. WaveSurfer [4] was used as a recording tool. The sampling frequency was 16 kHz. The emotional data consist of 160 samples (two students, 4 emotion patterns, 20 sentences). The emotion patterns are "neutral", "joy", "anger", and "sorrow". The powers (=intensities) were adjusted to be certain average level.

3. LISTENING EXPERIMENT

Listening experiment was conducted to examine how emotion patterns in whispered speech are recognized. In the experiment, recognition rates of both normal speech and whispered speech utterances were examined. Subjects were eight students who have normal hearing ability. They were asked to judge the emotional types of the speech data. The experiment was conducted in the soundproofing room.

3.1 Emotion recognition of the normal speech and whispered speech utterances

Result of the emotion recognition of the normal speech and whispered speech utterances is shown in Table 1. The result shows emotion in whispered speech is to some extent recognized. However, the recognition rate of the whispered speech is lower than that of normal speech utterances. As an opinion of the subjects, although whispered speech has no pitch, something like any movements of fundamental frequency can be heard.

Table 1. Emotion recognition result

	NEU-TRAL	JOY	ANGER	SORROW
Normal [%]	96	85	94	100
Whisper [%]	86	78	92	93

3.2 Emotion recognition of pseudo-whispered speech

We generated pseudo-whispered speech which suppresses pitch components of vocal cord vibrations in normal speech utterances. We investigated whether any emotion patterns of this pseudo-whispered speech is recognizable or not. By extracting acoustic features based on the speech analysis-and-synthesis method called STRAIGHT [5], and then re-synthesis a pseudo-whispered speech by suppressing the pitch information, where the spectral envelope components are unchanged from the input normal speech. We conducted listening experiment whether the emotion patterns of pseudo-whispered speech were recognized or not. The condition of the listening experiment is the same as Table 1. Result of the emotion recognition is shown in Table 2.

Table 2. Emotion recognition of pseudo-whispered speech

	NEU-TRAL	JOY	ANGER	SORROW
Recog-nition rate[%]	73	42	70	40

The result of the emotion pattern recognition for the pseudo-whispered speech shows that the rate is clearly lower than that of the whispered speech. The decrease of the recognition rate is remarkable in the cases of "joy" and "sorrow". Since whispered speech has no pitch, it is hypothesized that whispered speech can represent any alternative components which compensate the lack of the pitch information by controlling other acoustic features such as speech intensities, duration times, and formant frequencies. On the other hand, since the pseudo-whispered speech which is generated from normal speech utterance does not compensate the pitch information, the recognition rate may be decreased.

4. ANALYSIS EXPERIMENT

Next, we investigated how the emotional utterances of whispered speech were different from those of normal utterances. Acoustic features of intensity, duration time, and formant frequencies of both normal and whispered speech were analyzed by using WaveSurfer analizing tool.

4.1 Intensity and duration time analysis

Intensity and duration time for every emotion of normal and whispered utterances of shot sentences were analyzed. Table 3 and 4 show the averages of intensity and duration time of all the speech data for each emotional pattern. From these results, remarkable differences between emotional patterns are shown for both normal and whispered speech.

For both normal and whispered speech utterances, the order of the average intensities were "anger", "joy", "neutral", "sorrow". In the case of duration time, it is greater in the order of "sorrow", "neutral", "joy", "anger" for both normal and whispered speech. From these results, much differences between normal and whispered speech are not observed in the order of average intensity and duration time for emotion patterns.

Table 3. Average of intensity

	NEU-TRAL	JOY	ANGER	SORROW
Normal [dB]	42.3	47.2	55.1	37.4
Whisper [dB]	34.8	38.2	50.8	31.4

Table 4. Average of duration time

	NEU-TRAL	JOY	ANGER	SORROW
Normal [s]	1.71	1.69	1.54	2.16
Whisper [s]	2.00	1.93	1.85	2.19

4.2 Relationship between intensity and duration time

To further investigate the relationship of the intensity and duration time for normal and whispered speech, the pair-wise distributions of all the sample data were plotted as in Figure 2. From these results, the ratio of the two kinds of utterances with intensity and duration time are similar. For whispered speech, the power range was generally wider than that of normal speech utterances. Since the recognition of the emotion patterns of whispered speech is more difficult than that of normal speech, it is considered that the intensity of the whispered speech is controlled more precisely than that of normal speech to clarify the emotional differences.

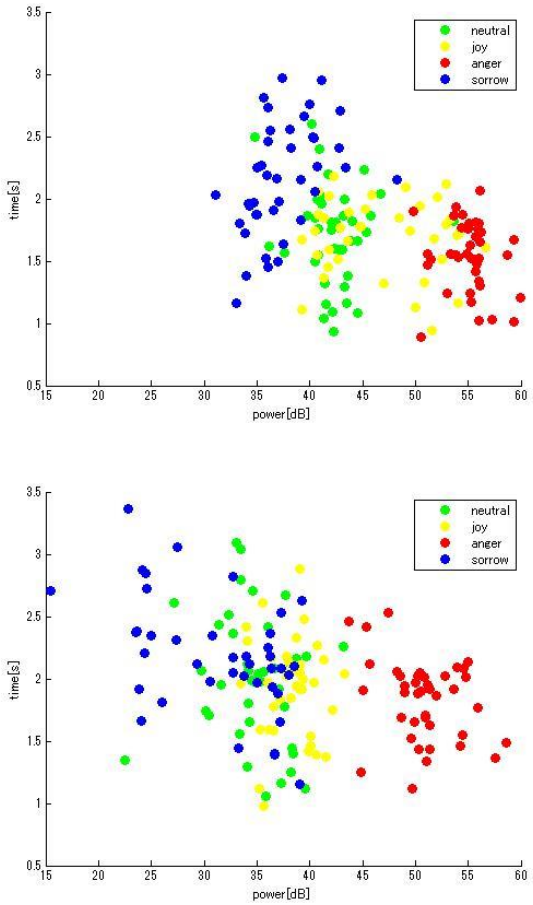

Fig. 2. Relationship of intensity and duration time (top : normal, bottom : whisper)

4.3 Formant frequency analysis

Formant frequencies are the frequencies of the spectral peaks of the spectral envelopes. Those frequencies are called ‘first’, ’second’, 'third' formants from the lower frequencies. In speech recognition, typical methods of sound processing often focused on formant frequencies which characterize phonemic information. Although, in emotional expression researches, formant frequencies have not been necessarily studied, in whispered speech processing, since formant frequencies may play an important role to analyze and synthesis emotional expressions.

In this research, we focus on first formant frequencies because they have mainly higher intensities than other formats. Figure 3 shows the distribution of the first formant frequencies for each speech data. The formant values are the averages of the frequencies over all the speech frames for every emotional normal and whispered speech utterances. The first formant frequencies of the normal speech utterance are distributed between 300 to 700 Hz, while, in the case of whispered speech, the frequencies are between 600 to 1100 Hz. The distribution of the frequencies of whispered speech is wider than that of the normal one. From this result, whispered speech may change first formant frequencies to represent any pitch-related information. In addition, among emotional patterns, the average formant frequencies change from lower to higher frequencies in accordance with the patterns, , it was found that be greater both of utterance in the order of "sorrow", "neutral", "joy", and "anger". The first formant frequency distributions of each emotional whispered speech are very clearly separated compared with those of the normal speech utterances.

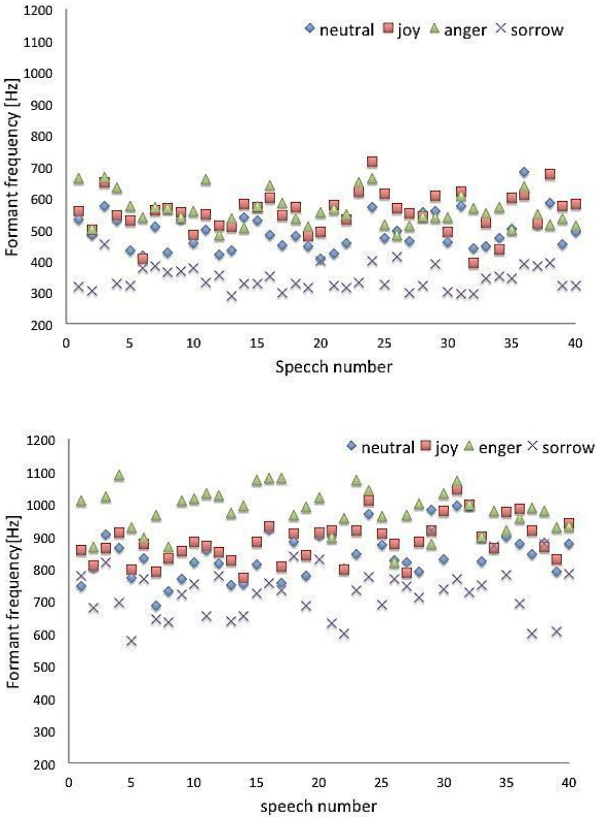

Fig. 3. Formant frequency (top : normal, bottom : whisper)

4.4 Formant frequency distributions in each phonemic sound

In the whispered speech, it is hypothesized that the first formant conveys any prosodic information related to pitch information to compensate the lack of vocal cord vibrations. We examined the relation of the formant frequency and pitch. In the experiment, in order to analyze the formant frequencies as a function of each utterance at certain pitch frequencies in normal speech, formant frequencies at each phoneme were extracted in vowels.

Firstly, labeling of phonemes are provided to every speech utterances by using WaveSurfer tool. Secondly, in the case of normal speech utterance, formant frequencies were associated with the corresponding pitch frequencies. In the case of whispered speech, formant frequencies were associated with the pitch frequencies of the corresponding normal speech utterances.

Figure 4 shows the relationship between the phoneme-wise pitch and formant frequencies in vowels. As a result, the distribution of the formant frequencies of whispered

speech is more widely spread than that of the normal speech utterances. Table 5 shows the standard deviation of these formant frequencies.

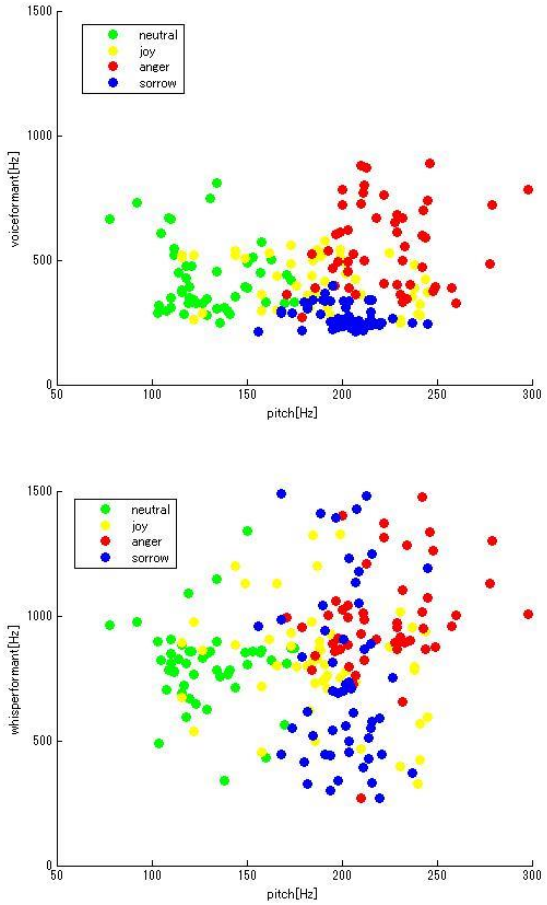

Fig. 4. formant frequencies as a function of pitch frequencies of normal speech utterances. (top: normal(=voiced) utterances, bottom: whispered utterances).

Table 5. Standard deviations of the first formant frequencies

	NEUTRAL	JOY	ANGER	SORROW
NORMAL	140	99	167	45
WHISPER	166	221	208	343

As results of the standard deviation in Table 5, the first formant frequencies of the whispered speech are greater than those of the normal speech utterances. The differences are 26 for "neutral", 122 for "joy", 41 for "anger", and 298 for "sorrow". It is suggested that reason, there is no pitch in whispered speech, is considered to have made the kind of pitch by movement of the formant frequency.

In particular, the differences of the standard deviations between normal and whispered speech utterances were significant for "joy" and "sorrow". Since the recognition rates of "joy" and "sorrow" emotion were low comparing with other emotions for whispered speech, some compensation mechanisms may occur.

5. CONCLUSION

In this study, we investigated the difference between emotional and acoustic features in normal speech utterance and whispered speech. Firstly, in the listening experiment, the emotional characteristics were able to be recognized in not only normal speech utterance but also whispered speech. Then, we analyzed intensity and duration time and formant frequencies in both speeches. As results, the ratio of speech intensity and duration time in both normal and whispered speech were similar. However, whispered speech shows clear differences for each emotional pattern due to avoid confusing the emotional categories. In the comparison of the distributions of the first formant frequencies of each vowel, whispered speech has larger standard deviations than those of normal speech utterances. This result suggests some compensation mechanisms which represent pitch-related prosodic information by using formant frequencies.

REFERENCES

[1] M. Sugito, M. Higasikawa, A. Sakakura, H. Takahashi, “Perceptual, acoustical, and physiological study of Japanese word accent in whispered speech," IEICE Technical Report., SP91-1, pp.1-8, 1991.

[2] G. Chenghui, Z. Heming, T. Zhai, Y. Zongyue, G. Xiaojiang, “Feature analysis on emotional Chainese whispered speech," Proc. ICINA, Vol.2, pp. V2-137 – V2-141, 2010.

[3] Y. Jin, Y. Zhao, C. Huang, L. Zhao, Y. Jin, “Study on the emotion recognition of whispered speech," Proc. GCIS'09, Vol.3, pp.242-246, 2009.

[4] WaveSurfer, http://www.speech.kth.se/wavesurfer/

[5] H. Kawahara, “STRAIGHT, a very high quality VOCODER: Insights from auditory scene analysis," Acoustical Society of Japan, Vol.54,No.7,pp521-526,1998.

Yuya Matsumoto received B.E. from Shibaura Institute of Technology in 2015. He is admitted to Graduate School of Shibaura Institute of Technology in 2015. His current research theme is speech processing.

Kazunori Mano received the B.E., M.E. and Dr. Eng. degrees in electrical engineering from Waseda University, Japan, in 1982, 1984 and 1987, respectively. From 1987 to 2008、 he engaged in research on speech coding and synthesis at NTT labs. Since 2008, he has been a Professor, Department of Electronic Information Systems, Shibaura Institute of Technology.