Preliminary evaluation of angry voice in automatic speech recognition

Rikio Ueno, Mitsunori Mizumachi, Yoshihisa Nakatoh

Kyushu Institute of Technology
1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka 804-8550, Japan

Rikio Ueno: m349405r@tobata.isc.kyutech.ac.jp

Abstract

Speech recognition has been introduced as an interface for the various devices; especially operator assistance in call center operations is needed. But when speech recognition is introduced into the call center operations, the recognition performance may deteriorate because the voices of customers include emotion (angry). Previous study reported that the recognition performance of “angry voice” tend to deteriorate than that of “calm voice”. The acoustic features of “angry voice” are different from those of “calm voice”, for example, loud power and high voice. In this study, to explore what factors make the recognition performance deteriorating, we record the parallel speech corpus of “calm voice” and “angry voice” in Japanese, carry out the recognition experiments. And, we compare speech pitch, speech power and spectral envelope between speaker of little deteriorating of speech recognition rate and speaker of deteriorating of speech recognition rate for five vowels. In the results, about speaker deteriorating recognition rate, speech pitch of /i/ increased (about 5dB) and speech power of /u/ increased (about 40Hz) on “angry voice” of “incorrectly words”. Particularly, it was confirmed that the spectral envelopes of /i/ and /u/ on “angry voice” were changed the form.

Key words speech recognition, emotional speech, calm voice, angry voice, spectrum, speech analysis, formant,

1. Introduction

Auto speech recognition (ASR) has been introduced as an interface to the various devices. Also, the needs of business support, for example call center, are increasing. However, ASR technique isn't able to support the call center work of all. Especially, problem occurs in the response operations of

complaint because it contains the feeling of anger to the voice emitted by the customers. The use conditions of ASR assume usual speech, and including feelings affect the performance. In this study, we carry out speech recognition experiments about two pattern voices, which record “calm voice” and “angry voice”. And, comparison “angry voice” and “calm voice” associate with acoustic features and the recognition performance.

2. Evaluation method

Previous study reported that the recognition performance of “angry voice” tended to deteriorate than the “calm voice”^[1]. It also reported that acoustic features of “angry voice” were different from those of “calm voice”, for example, increase power and F0^[2]. However, what acoustic features have a bad influence on ASR isn't known. Therefore, we study the relationship between acoustic features and recognition performance of “angry voice”. And, two pattern voices, which record “calm voice” and “angry voice” experiment on recognition engine in Japanese. Based on the speech recognition rate, divide into two groups, speaker of little deteriorate in the speech recognition rate (group A), and speaker of deteriorate in the speech recognition rate (group B). After all, we divide three words, which are “calm voice” of “correctly words” (Words A), “angry voice” of “correctly words” (Words B) and “angry voice” of “incorrectly words” (Word C) in each vowels to compare “calm voice” with “angry voice” to each group about acoustic features, which the entire power comparison, fundamental frequency (F0), and spectrum envelope. Comparison is performed for each vowel in the word in relation to the above-mentioned acoustic features.

3. Recognition experiment

3.1 The experimental conditions

We select the 50 words, which assume to be used complaint of call center operation. We record two patterns emotional voices. These emotional voices are “calm voice” which usually speak, and “angry voice”, which emit to feel angry. These voices are recorded by 18 adult male speakers. The sound wave records in 44.1kHz and 16 bits sampling frequency, after all the recorded voice down-sampled to 11 kHz using the recognition experiment. And, the ASR system for the recognition experiment is our original speech recognition engine, COMPATS^[3,4]. The specifications of the ASR engine are shown in Table 1. In our ASR system, a text-based word lexicon provides convenience in changing the word vocabulary for recognition. The size of the vocabulary is about 543, which consist of Tohoku University and Panasonic isolated spoken word database, and 50 words, which assume to be used complaint of call center operation.

Sampling freq.	10kHz
Window length	30[ms] Hamming
Window shift	15[ms]
Parameters in word matching	LPC-mel Cepstral parameters (7 dimensions), Normalized residual power, Delta log power
Sound unit	CV, VC
Acoustic models	5 - 10 states, Single mixture, single covariance and gender-independent

3.2 Experimental Results

Figure 1 shows the word recognition rate is for 18 speakers. As a result, in all speakers, the recognition performance of “angry voice” was less than the “calm voice”. Especially, in 6 speakers (which consist of No.1, 3, 5, 8, 14, 18), “angry voice” was more 15% deteriorate than “calm voice”. In 4 speakers (which consist of No.4, 7, 12, 15), “angry voice” was more 10% deteriorate than “calm voice”. And, in 8 speakers reduction frequency rate was not confirmed the others speakers.

Fig.1 recognition rate of all speakers

4. Analysis the acoustic features between “angry voice” and “calm voice”

Based on the results of the above-mentioned, we compare with “angry voice” and “calm voice” about the power, fundamental frequency, and the spectral envelope. We focus on the five vowels because the speech features are appeared well. In order to perform the comparison for each five vowels, clipping the vowel portion of the word, calculating the average value of a plurality word feature quantity and plurality of frames on the same vowel. The calculating of average value performs to the deteriorating in the speech recognition rate. Also, we separate the words of them from the words of recognition error and those of correctly word.

4.1 Comparison of the average power of each vowel

The results show Figure 2. The vertical axis represents the size of the power. The horizontal axis is the type of vowels. For each vowel shows the standard deviation of the average power about “Words A”, “Words B” and “Words C” in group B. As a result, an increase of the power was seen comparing the “calm voice” and the “angry voice.” the tendency was seen features of “angry voice”. On the other hand, the vowel of /u/ was different slight features from others vowels about comparing with “Words B” and “Words C”. However, there was no significant difference between “Words B” and “Words C”.

Fig.2 each phoneme of average power on group B

4.2 Comparison of F0 of each vowels

We compare to the spectral envelope for each vowel. Figure 3 shows the results. The vertical axis is the fundamental frequency. The horizontal axis represents the type of vowels. As a result, when I compared “angry voice” with “calm voice”, fundamental frequency of “angry voice” was more than “calm voice” about 40-80Hz. Although some slight difference in the vowel of / i /, there was no significant difference in the total vowels about the comparison between “Words B” and “Words C”.

Fig.3 each phoneme of average F0 on group B

4.3 Comparison of the spectral envelope of each vowel

In the section, we compare the spectral envelope for each vowel. Especially, we focus the speaker No.8 in group B and speaker No.17 in group A. Figure 4-8 show a comparison of the spectral envelope of the speaker No.8 about vowels /a/, /i/, /u/, /e/ and /o/. And figure 9-10 show a

comparison of the spectral envelope of the speaker No.17 about vowels /i/ and /u/. The vertical axis show the power of voice, the horizontal axis is frequency. In the spectral envelopes of speaker No.8, which are /a/, /e/, and /o/, these were similar curve between “Words A”, “Words B” and “Words C”. However, in the spectral envelopes, which were /i/ and /u/, these changed a large shape of spectral envelopes. About vowels /i/, the deformations of the formants were remarkable between “Words A”, “Words B” and “Words C”. About vowels /u/, first-formant and second formant changed toward the high-frequency in “Words C” compared with “Words A” and “Words B”. And in the spectral envelopes of speaker No.17, the above-mentioned changes didn’t occur. In summary, vowels /i/ and /u/ were susceptible to feelings of anger, so the deformations of the formants were remarkable compared to other vowels. Especially, vowels /i/ of the changes was more remarkable than vowels /u/. And the changes didn’t occur in group A. In other words, it was supposed that cause of the deterioration of recognition performance about “angry voice” was caused by the deformation of the spectrum /i/ and /u/.

Fig.4 Spectrum envelopes of vowel /a/ in speaker No.8

Fig.5 Spectrum envelopes of vowel /i/ in speaker No.8

Fig.6 Spectrum envelopes of vowel /u/ in speaker No.8

Fig.10 Spectrum envelopes of vowel /u/ in speaker No.17

Fig.7 Spectrum envelopes of vowel /e/ in speaker No.8

Fig.8 Spectrum envelopes of vowel /o/ in speaker No.8

Fig.9 Spectrum envelopes of vowel /i/ in speaker No.17

5. Conclusion

As a result, the spectral envelopes, which were /i/ and /u/ about “angry voice”, changed a large shape from “calm voice” in comparing correctly words with incorrectly words about “angry voice”. Especially, the spectral envelopes of group B changed more remarkable than those of group A. In the future, we are going to improve the speech recognition rate about “angry voice”.

REFERENCES

- [1] N. Kadotani, H. Aso and M. Suzuki, “An investigation on discrimination among emotion expressions contained in speech,” IEICE Transactions on Information and System SP2001-141, pp43-50, Mar, 2002
- [2] I. Hiradate and A. Masato, “Analyses of acoustic features of “anger” emotional speech,” IEICE Transactions on Information and System (NLC2000-34), SP2000-82, pp43-48, Dec, 2000
- [3] T. Kimura, H. Kuwano, A. Ishida, T. Watanabe and S. Hiraoka, “Compact-Size Speaker Independent Speech Recognizer for Large Vocabulary Using “COMPATS” Method,” Proc. ICSLP 94, pp. 1379-1382 (1994).
- [4] T. Kimura, M. Endo, S. Hiraoka and K. Niyada, “Speaker independent word recognition using continuous matching of parameters in time-spectral form based on statistical measure,” Proc. ICSLP 92, pp. 169-172 (1992).