interspeech 2008 incorporating SST 08

Analysis of Voice-Quality Features of Speech that Expresses "Anger", "Joy", and "Sadness" Uttered by Radio Actors and Actresses

Shoichi Takeda[1], *Yuuri Yasuda*[2], *Risako Isobe*[3], *Shogo Kiryu*[3], *Makiko Tsuru*[4]

[1]School of Biology-Oriented Science and Technology, Kinki University, Wakayama, Japan
[2]Osaka Gas Information System Research Institute Co., Ltd., Osaka, Japan
[3]Research Division in Engineering, Musashi Institute of Technology, Tokyo, Japan
[4]Graduate School of Biology-Oriented Science and Technology, Kinki University, Wakayama, Japan
and Department of Business Career, Kurume Shin-Ai Women's College, Fukuoka, Japan

takeda@gakushikai.jp, y_yasuda@aspect-design.info, g0881305@sc.musashi-tech.ac.jp,
kiryu@bme.musashi-tech.ac.jp, tsuru_mkk@kurume-shinai.ac.jp

10.21437/Interspeech.2008-548

Abstract

This paper describes the analysis of the voice-quality features of "anger", "joy", and "sadness" depending on the degree of the emotion for expressions in Japanese speech. The degrees of emotion were "neutral", "light", "medium" and "strong". Among voice-quality features, we turned to the noise level of the glottal-flow waveform. We adopted the AR model and measured the noise levels of the predictive residual signal of speech that expressed each emotion. To measure a relative noise level to the signal level, the "noise-to-signal (N/S) ratio" was introduced. The analysis results showed that the relative noise levels in the residual-waveform spectra were different, i.e., the N/S ratio of each emotion was larger in the order of "anger" > "sadness"≈ "neutral" > "joy" by approximately 4 dB.

Index Terms: speech synthesis, emotional expression, voice quality, noise-to-signal (N/S) ratio

1. Introduction

Owing to recent advancement of speech technology, synthetic speech has remarkably improved its quality and is being used in various fields.

The current synthetic speech applied in practical fields, such as electronic dictionaries, e-mail reading, etc., is, however, mostly non-expressive. It is therefore necessary to develop a technique to synthesize expressive speech if we want to extend its application more widely.

Among expressive speech, we have placed a focus on emotional speech such as "anger", "joy", "sadness", "gratitude", etc. As the first step, we have so far been analyzing the prosodic features of various emotional expressions to achieve more natural-sounding rule-based synthetic speech [1]-[3].

We have been taking a minute approach instead of generally investigating various types of typical emotional expression [4]. The degree of emotion was categorized into four categories: "neutral", "light", "medium", and "strong", and the prosodic features of each category have been analyzed [1]-[3].

The quality of speech synthesized based on these prosodic features, however, did not sufficiently express emotions. We learned that not only prosodic features but also some other features must be used to express emotions. Among such features, voice-quality features were investigated. We observed that "anger" speech was more noisy and, on the other hand, "joy" speech was less noisy than "neutral" speech.

This paper thus discusses the features of glottal-flow waveform based on the AR model in terms of the noise levels of speech. We measured the noise levels of the predictive residual signal of speech that expressed several degrees of each emotion.

2. Experimental conditions

2.1. Speech samples

The speakers were two radio actors and two radio actresses in their 20s and 30s.

As speech samples, we used 4-mora and 6-mora Japanese words that had either of the three accent types: flat, mid-high, or head-high. The number of words was 6. The type of emotion was "anger", "joy", and "sadness". Each word was uttered with the following four degrees of the emotions: "neutral", "light", "medium", and "strong". At this stage, the "degrees" were subjectively defined by the speakers themselves. Standardization of "degrees" in relation to subjective listening tests will be left to future work. They uttered 5 times a word. The total number of words was thus 1260.

2.2. Analysis conditions

Analysis conditions are shown in Table 1. In this experiment the sampling frequency was down-sampled to 8 (kHz) due to the limited performance of the current speech analysis program. In future we will improve the program so that speech analysis can be performed at an arbitrary sampling frequency.

Table 1: *Analysis conditions.*

Processing	Parameter	Value
Digital recording	Sampling frequency	48 (kHz)
	Bit length	16 (bit)
Down-sampling for speech analysis	Sampling frequency	8 (kHz)
DFT for spectral analysis	Window length	256 (samples)
		32 (ms)
	Window type	Hanning
LPC inverse filtering for residual signal extraction	Window length	256 (samples)
		32 (ms)
	Window type	Hanning
	Order of LPC	8

3. Spectral features

Residual signal was extracted from the speech signal by applying an LPC inverse filter to investigate the noise component included in the glottal-flow waveform. The amount of noise component can be quantified by the noise-to-signal (N/S) ratio [5]. Figures 1-10 show the short-time spectra of the residual waveform of the first /i/ in the word "imanimo" uttered by radio actress NN with various types and degrees of emotion.

From these figures, we observed that for "joy" speech, the differences between the peaks and the dips of the harmonic structures were larger than those for "neutral". This meant that the noise levels of the "joy" speech were smaller than those of the "neutral" speech. For "anger" speech, on the other hand, the differences were smaller due to the increase in the noise levels relative to those for "neutral" speech. In addition, the periodicity of the harmonic structures tended to disappear in the range of more than 2 (kHz). For "sadness" speech, there seemed to be nearly the same levels of differences as those for "neutral" speech.

4. Calculation of noise-to-signal ratio

The N/S ratio is an index to quantify the relative amount of noise components included in the spectrum of the signal.

The N/S ratio defined by Muta *et al*. [5] used the smallest value of the signal power over a harmonic peak and valley in the spectrum as the noise component. We used a modified form of the N/S ratio. In our definition, the noise component was the average value of the signal power over a harmonic valley. The advantage of our definition is that the quantity is insensitive to random errors. The N/S ratio is calculated as follows.

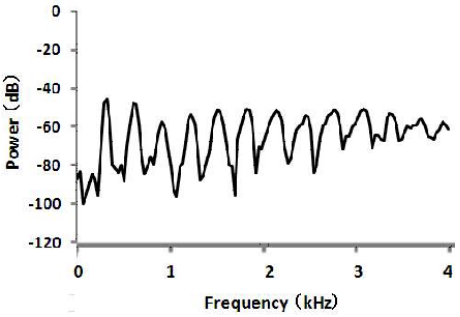

Figure 1: *Spectrum for "neutral" speech.*

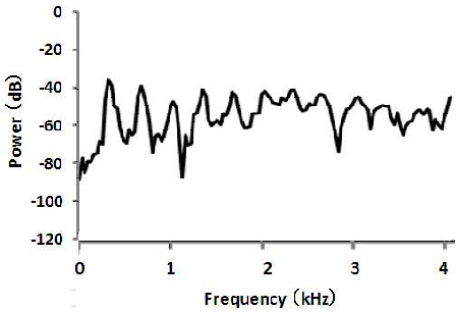

Figure 2: *Spectrum for "displeasure" speech.*

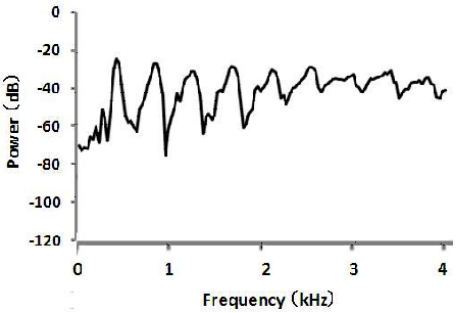

Figure 3: *Spectrum for "anger" speech.*

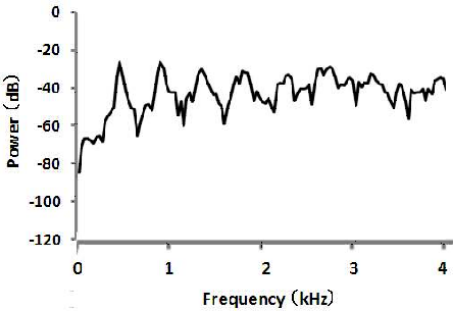

Figure 4: *Spectrum for "fury" speech.*

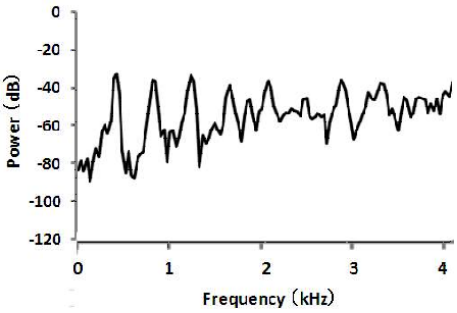

Figure 5: *Spectrum for "slight joy" speech.*

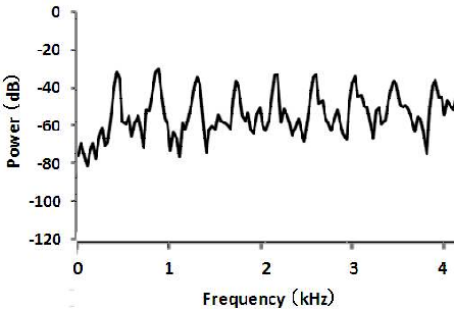

Figure 6: *Spectrum for "joy" speech.*

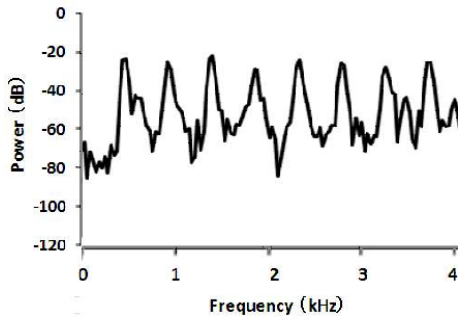

Figure 7: *Spectrum for "great joy" speech.*

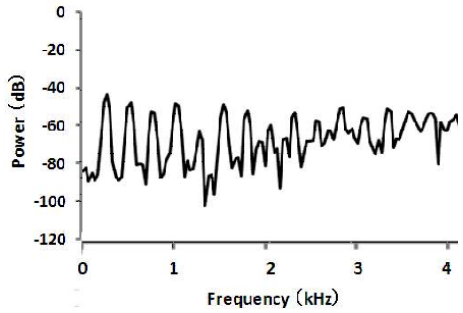

Figure 8: *Spectrum for "slight sadness" speech.*

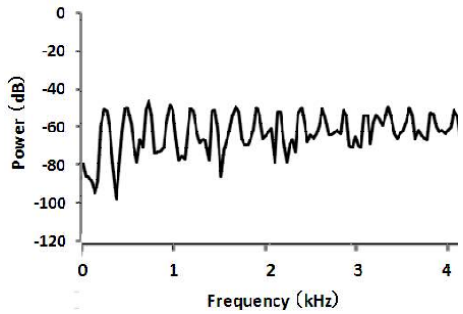

Figure 9: *Spectrum for "sadness" speech.*

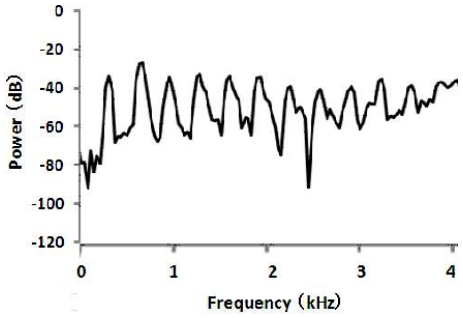

Figure 10: *Spectrum for "deep sadness" speech.*

Comb filtering was applied to the spectrum so that the period of the comb matched that of the harmonic peaks of the spectrum as shown in Fig. 11. Let the sum of the powers in the residual spectrum be S, and the sum of the powers inside the comb filter be C, then the noise component N is $N = S - C$.

The N/S ratio R_{NS} (dB) is defined as

$$R_{NS} = 10 \log_{10}(N/S) \tag{1}$$

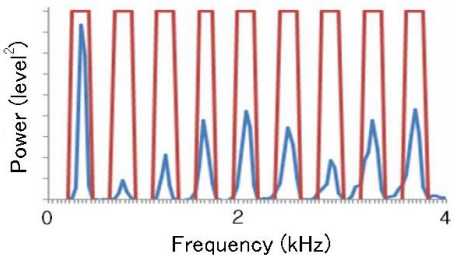

Figure 11: *Comb filtering for N/S ratio calculation.*

5. Results of noise-to-signal ratio calculation

Figures 12 - 14 show several examples of the results of N/S ratio calculation. From these figures, we confirmed that the N/S ratio of each emotion was larger in the order of "anger" $>$ "sadness" $\approx$ "neutral" $>$ "joy" by approximately 4 dB regardless of speakers and words. This meant that the relative noise levels in the residual-waveform spectra were different depending on the type of emotion.

Within the same types of emotion, however, there seemed to be slight differences in the N/S ratios regardless of the type and degree of emotion except for the differences between "slight joy" and the stronger joy. We therefore conducted statistical tests to confirm whether there were significant differences between these data.

Tables 2-5 summarize the statistical test results conducted for all speech samples.

From Table 2, we knew that there were significant differences between "neutral" and "anger", and between "neutral" and "joy" each at the 1% level. There was, however, no significant difference between "neutral" and "sadness". This meant that the noise level of "sad" speech were nearly the same as that of "neutral" speech.

Tables 3-5 list the statistical test results conducted within "anger", "joy", and "sadness" groups, respectively. From Table 3, we knew that even within "anger" group, there were significant differences between any two degrees of "anger" at the 1% level. As for "joy" group (see Table 4), significant differences were observed between "slight joy" and the two other degrees of "joy", but no significant difference was oberved between "joy" and "great joy". As for "sadness" group (see Table 5), however, no significant differences were observed between any two degrees of "sadness". These results meant that the significance of the noise level within each emotion group was different depending on the type of emotion.

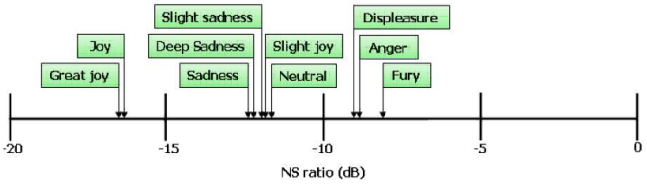

Figure 12: *N/S ratio for the word "atonomatsuri" uttered by radio actor FK.*

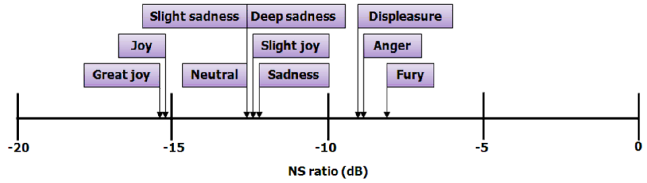

Figure 13: *N/S ratio for the word "imanimo" uttered by radio actor FK.*

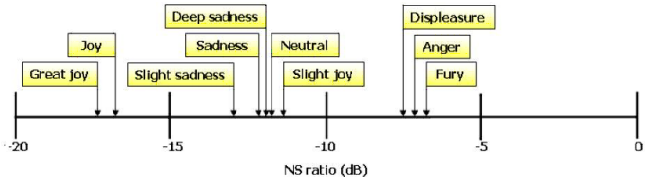

Figure 14: *N/S ratio for the word "atonomatsuri" uttered by radio actress NN.*

6. Conclusions

This paper has described the analysis of voice-quality features in terms of the noise level. From observation of the spectra of the residual signal, we have known that the noise levels are different depending on the types of emotion. The noise-to-signal (N/S) ratio has been introduced to quantify the relative noise level in emotional speech.

From calculation results, we have known that the N/S ratio of each emotion is significantly larger in the order of "anger" > "sadness" ≈ "neutral" > "joy" by approximately 4 dB. The significance of the N/S ratio within each emotion group is different depending on the type of emotion.

Future studies will be to explore other voice-quality features such as spectral tilt and other spectral shape features, to examine robustness of these feature parameters, to clarify perceptual relevance, and finally, to synthesize emotional speech using the voice-quality features gained through our research.

7. Acknowledgments

The authors would like to express their thanks to the actors and actresses at Gekidanseinenza radio theater for their help in uttering emotional speech.

This research was partly supported by Grant-in-Aid from Kinki University.

Table 2: *Statistical test results conducted between "neutral" and emotion groups.*

	Anger group	Joy group	Sadness group
Neutral	$p < 0.01$	$p < 0.01$	$p > 0.05$

Table 3: *Statistical test results conducted between different degrees of "anger".*

	Displeasure	Anger
Anger	$p < 0.01$	
Fury	$p < 0.01$	$p < 0.01$

Table 4: *Statistical test results conducted between different degrees of "joy".*

	Slight joy	Joy
Joy	$p < 0.01$	
Great joy	$p < 0.01$	$p > 0.05$

Table 5: *Statistical test results between different degrees of "sadness".*

	Slight sadness	Sadness
Sadness	$p > 0.05$	
Deep sadness	$p > 0.05$	$p > 0.05$

8. References

[1] Takeda, S., Ohyama, G., and Tochitani, A., "Japanese project research on "Diversity of Prosody and its Quantitative Description" and an example: analysis of "anger" expressions in Japanese speech", Proc. ICSP2001, Taejon, Korea, 423–428, 2001.

[2] Hashizawa, Y., Takeda, S., Muhd Dzulkhiflee Hamzah, and Ohyama, G., "On the Differences in Prosodic Features of Emotional Expressions in Japanese Speech according to the Degree of the Emotion", Proc. 2nd Int. Conf. Speech Prosody, Nara, Japan, 655–658, 2004.

[3] Muhd Dzulkhiflee Hamzah, Takeda, S, Muraoka, T., and Ohashi, T., "Analysis of Prosodic Features of Emotional Expressions in Noh Farce ("Kyohgen") Speech according to the Degree of Emotion", Proc. 2nd Int. Conf. Speech Prosody, Nara, Japan, 651–654, 2004.

[4] Kitahara, Y. and Tohkura, Y., "Prosodic control to express emotions for man-machine speech interaction", IEICE Trans. Fundamentals, E75-A(2):155–163, 1992.

[5] Muta, H., Baer, T., Wagatsuma, K., and Muraoka, T., "A pitch-synchronous analysis of hoarseness in running speech", J. Acoust. Soc. Am., 84(4):1292–1301, 1989.