KEER2010, PARIS | MARCH 2-4 2010
INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH 2010

SPECTRAL-TILT FEATURES OF EMOTIONAL SPEECH
-RESEARCH ON EMOTIONAL-SPEECH SYNTHESIS BASED ON VOICE-QUALITY CONVERSION-

Shoichi TAKEDA[*a], Yuudai UENO[b], Noboru NAKASAKO[a], Hideo NAKAGAWA[a], Makiko TSURU[c], Risako ISOBE[d] and Shogo KIRYU[d]

[a] *Faculty of Biology-Oriented Science and Technology, Kinki University, Japan*
[b] *Osaka Branch, Chiyoda Integre Co., Ltd., Japan*
[c] *Graduate School of Biology-Oriented Science and Technology, Kinki University, and Department of Business Career, Kurume Shin-Ai Women's College, Japan*
[d] *Graduate School of Engineering, Tokyo City University Graduate Division, Japan*

ABSTRACT

This paper describes the analysis of the voice-quality features of "anger", "joy", etc. depending on the degree of the emotion for expressions in Japanese speech. Among voice-quality features, we turn to the tilt of speech spectra. The analysis results show that these spectral-tilt quantities are emotion-dependent, i.e., the spectral-tilt quantities for "anger" as well as "joy" increase significantly as the degree of each emotion becomes greater. This confirms that the voice quality changes to the one whose higher-frequency band is more emphasized as the degrees of "anger" and "joy" increase.

Keywords: *speech synthesis, emotional expression, voice quality, spectral tilt*

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].

The importance of research on emotional expressions has been widely recognized, and workshops specializing in emotional expressions were held. In the ISCA Workshop [4] held in 2000, for example, a wide variety of research results were reported, ranging from theoretical studies, databases, tools, feature analysis, etc. to applications of speech synthesis and recognition. Other relevant works include Scherer [5], Banse and Scherer [6], Schröder and Grice [7], Juslin and Laukka [8]. Among them, however, reports on Japanese speech synthesis were few.

In the early stage, Nakayama *et al*. analyzed and synthesized Japanese emotional speech [9]. Later, Kitahara and Tohkura [10], Kobayashi and Niimi [11], and some other researchers analyzed rough features of typical emotional expressions such as "joy," "anger," etc. and/or synthesized emotional speech based on these features. These studies, however, gave a mere rough paradigm of emotional expressions such as "joy", "sadness", "anger", etc. They therefore left further studies to give rules to express minute emotional nuances.

We have been taking more minute approach instead of investigating various types of typical emotional expressions roughly. As the first step, we have placed a focus on "anger" expressions since their prosodic features are relatively clear. "Anger" is divided into four degrees: "neutral", "displeasure (weak)", "anger (medium)", and "fury (strong)" and features of each degree have been analyzed. As the next step, we have analyzed the prosodic features of "joy", "sadness", and "gratitude" using the same approach [1]-[3]. Here, "a degree of emotion" refers to the intensity factor of an emotion and it has so far been determined subjectively by the speaker.

There are still few reports on such research in which each emotional expression is divided into several degrees and studied how the features differ depending on the degree of emotion. Examples of such rare studies can be found in Hirose Group's works on "anger", "joy", and "sadness" [12], [13]. One of their reports deals with analysis of Japanese short sentence speech with the above 3 types of emotion uttered by one speaker. Another report is concerned with analysis of 6-mora Japanese word speech with "anger" expression uttered by three speakers. In both studies, they analyzed the features of temporal structures and fundamental frequency. Recently, Bänziger and Scherer investigated the features of emotional expressions in languages other than Japanese in terms of the degree of emotion [14].

In our studies, we have tried to clarify prosodic features of Japanese acted speech comprehensively; not only the features of temporal structures and fundamental frequencies including Fujisaki's model parameters [15], but also those of speech power. Liscombe *et al*. did research on the relationships between subjective ratings and various acoustic features using a huge speech database [16]. This work, however, did not use emotional speech samples whose emotions were divided into several degrees, so the degree-dependent features were unknown. Nor were known the relationships between subjective ratings and features

relating to Fujisaki's model parameters [17], which were indispensible for generating Fujisaki's-model-based F_0 contours for speech synthesis.

The quality of speech synthesized based on only 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.

An example of conventional research on voice-quality analysis and synthesis can be found in Kasuya *et al.*'s study, in which they constructed a speech synthesis system that was capable of synthesizing "suspicious" and "disappointed" expressions [18], [19]. They introduced several paralinguistic features in combination into the system. These features included prosodic and voice-quality features such as F_0 patterns, F_0 range, intensity, duration, glottal-flow waveform and spectrum, turbulent noise, fluctuations, and formant characteristics.

Instead of exploring such sophisticated expressions, we have placed a focus on several basic emotions such as "anger", "joy", etc. with several degrees, as an extension of our study of prosodic features.

Some of the most important factors that affect voice quality of such emotional speech may be noise in speech, specifically in excitation, spectral shapes, i.e., spectral tilt and depth between the peaks and dips of the spectrum, etc.

With regard to the noise levels of speech, we have so far measured the noise levels of the predictive residual signal of speech that expresses several degrees of each emotion and clarified quantitatively that the noise levels differ depending on the type and degree of emotion [20].

This paper places a focus on the features of the spectral-tilt quantities of speech. So far, many researchers have done work on spectral-tilt features [16], [18], [19]. What is novel in our research is that we have clarified the spectral-tilt features of *multiple degrees of emotions in Japanese speech*. As the first step, this paper reports the spectral-tilt features of "anger" and "joy" groups.

2. OVERVIEW OF EMOTIONAL SPEECH SYNTHESIS SYSTEM

We propose a rule-based emotional speech synthesis system as shown in Fig. 1, consisting of prosody conversion as well as voice-quality conversion functions. This system is scheduled to be constructed as an extension of our *residual-excited* PARCOR synthesis system [21], which was confirmed to yield better speech quality than the conventional impulse-/white-noise-excited PARCOR synthesis system did.

Prosody conversion can be achieved using the knowledge of prosodic features of emotional speech [1]-[3].

In voice-quality conversion, excitation conversion is expected to be achieved if we can generate similar excitation of emotional speech artificially so that the noise level can be adjusted to that of emotional speech using the measure called an "N/S ratio" [20]. Spectral

tilt conversion is also expected to be achieved if we can accumulate knowledge of spectral-tilt features, which will be described in the following sections.

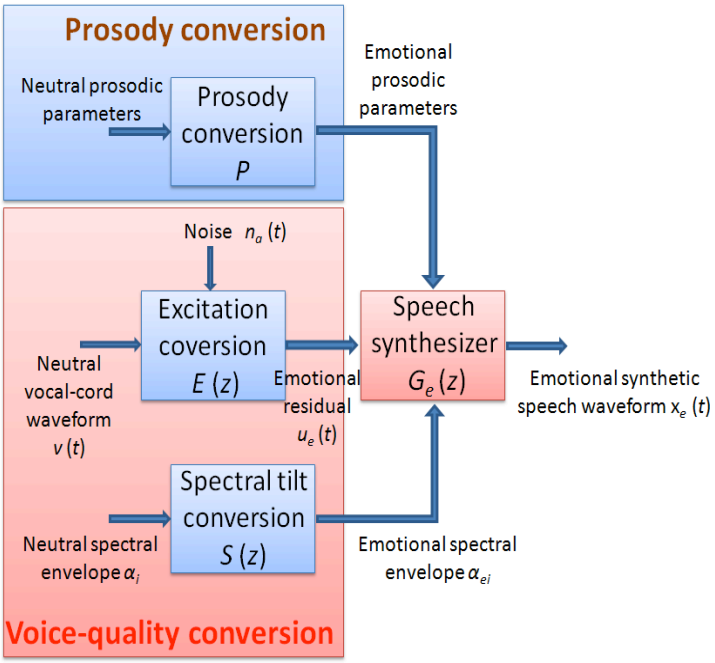

Figure 1: Rule-based emotional speech synthesis system

3. EXPERIMENTAL CONDITIONS

3.1. 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 5 as listed in Table 1. The measuring points here were arbitrarily chosen from quasi-stationary parts of high-accented vowels. The types of emotions were "anger" and "joy". Each word was uttered with the following three degrees of the emotions: "neutral", "medium", and "strong". At this stage, the "degrees" were subjectively defined by the speakers themselves. Standardization of "degrees" in relation to subjective listening tests is being conducted as a separate study [17]. They uttered 5 times a word. The total number of words was thus 500.

Table 1: Words for analysis

Word	Measuring point
"yamagoya"	/a/ in "ma"
"gaikokumai"	/a/ in "ma"
"umaoimushi"	/a/ in "ma"
"imanimo"	/i/ in the first "i"
"atonomatsuri"	/a/ in the first "a"

3.2. Analysis Conditions

Analysis conditions are listed in Table 2.

Table 2: Analysis conditions

Processing	Parameter	Value
Digital recording	Sampling frequency Bit length	48 (kHz) 16 (bit)
Down-sampling for speech analysis	Sampling frequency	8 (kHz)
FFT and LPC for spectral analysis	Window length	512 (samples) 64 (ms)
	Window type	Hamming
	Order of LPC	10

4. EMOTION-DEPENDENT SPECTRAL FEATURES

Figure 2 shows the FFT spectrum of "neutral", "fury", and "great joy" speech samples. As designated by the pink ovals in the figure, the powers in the higher band (approximately more than 2000 (Hz)) of the spectra of "fury" and "great joy" increase relative to those in the lower band and the valley of the spectrum is flattened.

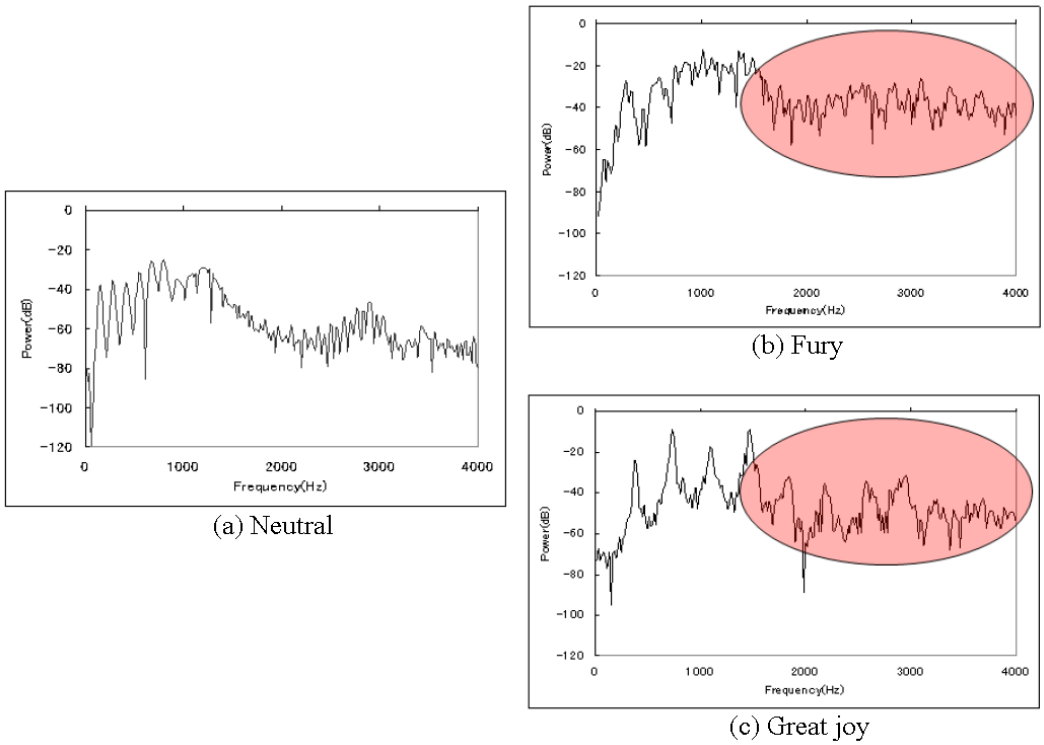

Figure 2: Spectra of "neutral", "fury", and "great joy" speech
Speaker: FY (male), word: "atonomatsuri"

These features can be quantified by introducing a spectral-tilt quantity and a distance quantity between the spectral peak and dip. This paper will only deal with a spectral-tilt quantity.

5. DEFINITION OF SPECTRAL TILT

Kasuya *et al*. defined a spectral-tilt parameter *TL* as a difference between the peak dB-power level and the dB-power level at 2-3 kHz of the spectrum of the source waveform [18], [19]. In this paper, we defined a spectral-tilt parameter in a different way so that the parameter matched our proposed speech-synthesis system described in Fig. 1. A synthesis method based on this spectral-tilt definition will be reported in a separate paper in future.

Figure 3 shows power spectra and their regression lines obtained by the least squared method. In the figure, the blue curve is the spectrum at /a/ in the word "atonomatsuri" with "neutral" emotion and the red one is the spectrum at the same part of the word with "fury" emotion both uttered by female speaker NN.

Here, a spectral tilt parameter is defined as the primary coefficient of the regression-line function. As seen in this figure, the spectral tilt of "fury" speech is greater than that of "neutral".

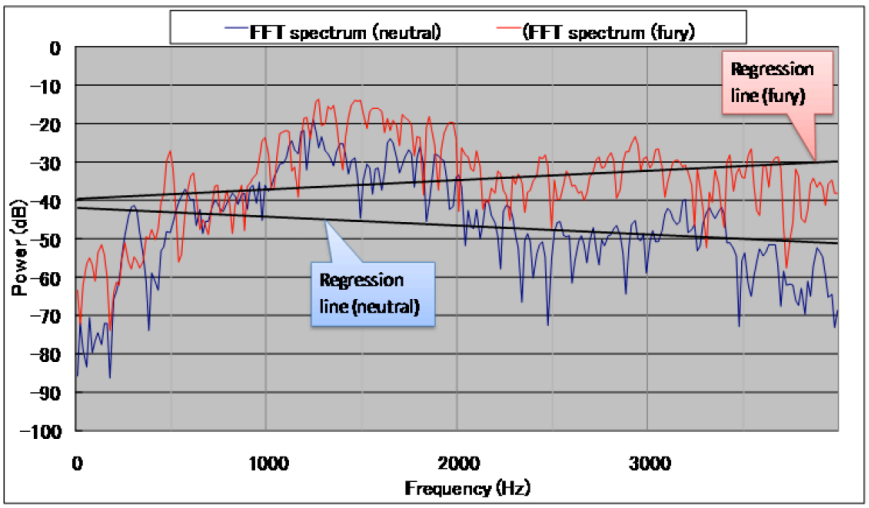

Figure 3: Spectra and their regression lines
Speaker: NN (female)

6. ANALYSIS RESULTS

We conducted spectral-tilt analysis for all speech samples.

Figures 4 and 5 show the results of spectral-tilt analysis. In the figures, the lengths of bars denote mean values and those of error bars denote standard deviations.

Figure 4 shows the spectra-tilt quantities for "anger" speech uttered by two male and two female speakers. From this figure, we knew that as the degree of "anger" increased, the

spectral-tilt quantities tended to increase. This tendency was common to all speakers. The quantities were, however, greater for the female speakers than for the male speakers.

Figure 5 shows the spectra-tilt quantities for "joy" speech uttered by the same speakers. In the case of "joy" also, the tendency was the same as the case of "anger".

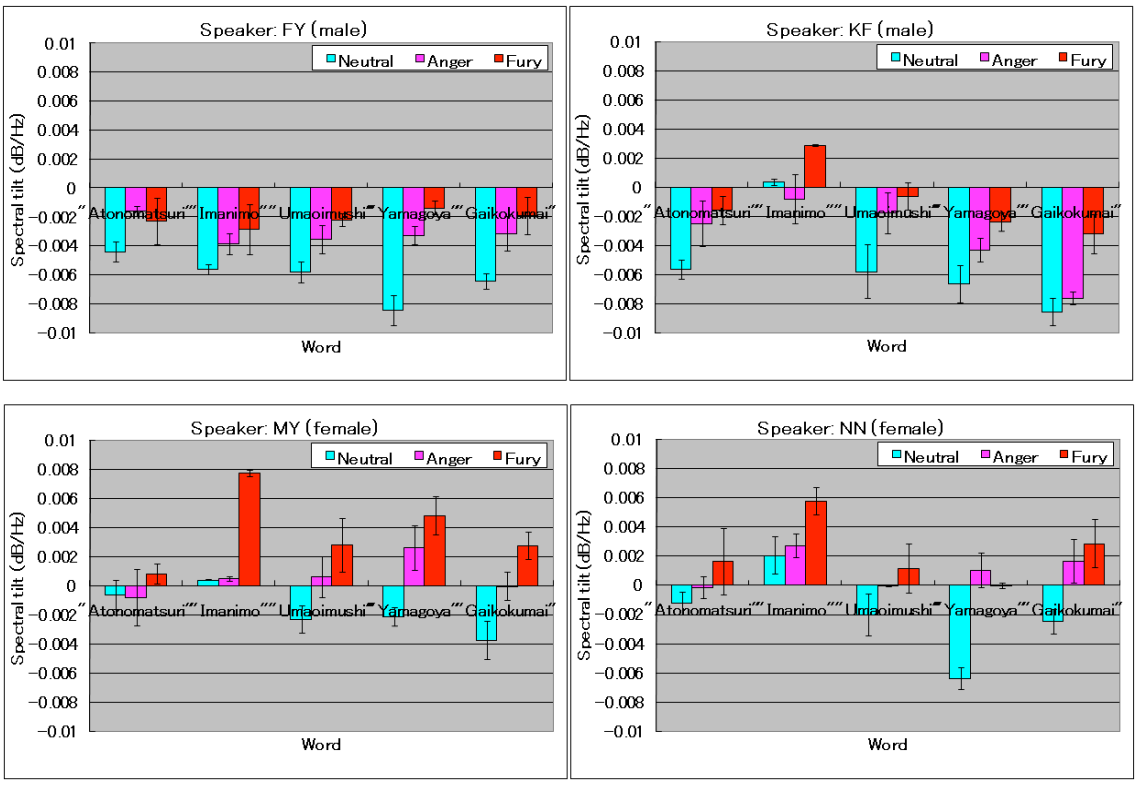

Figure 4: Results of FFT spectral tilt analysis for "anger" speech

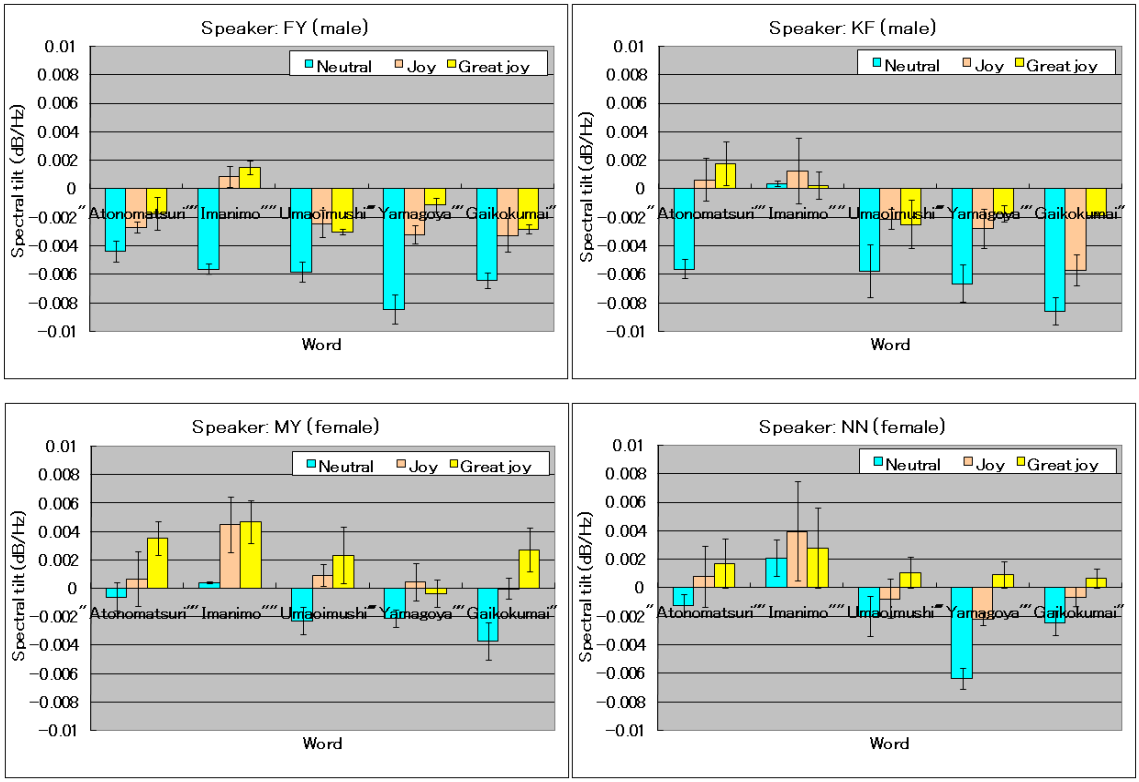

Figure 5: Results of FFT spectral tilt analysis for "joy" speech

These results depicted in Figs. 4 and 5 confirmed that the voice quality changed to the one whose higher-frequency band was more emphasized as the degrees of "anger" and "joy" increased. These results also suggest that a dimensional representation of emotions is useful for more minute expressions of emotions by a speech synthesis system.

One of the potential reasons for finding the same pattern in "anger" and "joy" is presumably that both are realized as high activation/arousal states, which is known to correspond to louder speech with higher vocal effort and should thus be expected to show higher spectral tilt than neutral speech.

7. DISCUSSIONS

7.1. Statistical Tests

Student's *t*-tests were conducted to confirm whether there were significant differences in the spectral-tilt quantities between "neutral" and emotional groups. The test results are listed in Table 3.

From this table, we knew that there were significant differences between "neutral" and "anger" / "fury", and between "neutral" and "joy" / "great joy" in most cases at the 1% level.

Table 3: Results of statistical tests for FFT spectral-tilt quantities between "neutral" and "anger / joy" groups

Speaker (Gender)	Emotion and degree			
	Anger	Fury	Joy	Great joy
FY (male)	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
KF (male)	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
NN (female)	$p > 0.05$	$p < 0.01$	$p < 0.01$	$p < 0.05$
MY (female)	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$

7.2. Phoneme-Dependency of Spectral Tilt

Since a speech power spectral shape depends on the kind of the phoneme, the spectral-tilt quantity is also expected to depend on the kind of the phoneme.

The phonemes used for analysis this time were Japanese /a/ and /i/. As seen in Figs. 4 and 5, the spectral-tilt quantities of /i/ for some data were greater than those of /a/. However, the increments of the spectral-tilt quantities of emotional speech from those of "neutral" speech might be less phoneme-dependent. Further analysis for other vowels is needed to clarify phoneme-dependent or independent features.

7.3. Spectral-Tilt Features of "Sadness" Speech

Spectral-tilt features of "sadness" speech are being analyzed. According to the analysis results so far, there seem to be no significant differences between the spectral tilts of

"neutral" speech and those of various degrees of "sadness" speech due to the diversity of "sadness" expressions in Japanese speech. To clarify accurate features, "sadness" needs to be divided into a few categories. We will report these analysis results in the near future.

8. CONCLUSIONS

This paper has described the analysis of voice-quality features of emotional speech in terms of spectral tilt.

From calculation results, we have known that spectral-tilt quantities of "anger" and "joy" group speech are significantly greater than those of "neutral" speech. Furthermore, spectral-tilt quantities of speech uttered by female speakers are greater than those of speech uttered by male speakers. These results confirm that the voice quality changes to the one whose higher-frequency band is more emphasized as the degrees of "anger" and "joy" increase. Furthermore, the results suggest that these tendencies are more emphasized for female speech.

From the above results, it may be concluded that a dimensional representation of emotions will be useful for more minute expressions of emotions by a speech synthesis system.

Future studies will be to analyze spectra-tilt features of other types of emotions such as "sadness" and to synthesize emotional speech using the voice-quality features gained through our research.

9. ACKNOWLEDGEMENTS

The authors express their sincere appreciations to Professor Hideki Kasuya at International University of Health and Welfare for his invaluable comments, and the actors and actresses at Gekidanseinenza Radio Theater for their help in uttering emotional speech.

This study was partly supported by the Project Research of the School of Biology Oriented Science and Technology, Kinki University No. 06-I-3, 2007-2009, and Grant-in-Aid for Scientific Research (C) "Interdisciplinary Research in Physiology and Acoustics on Active Characteristics of *Kansei* Evoked by Speech and Music Stimuli" from Japan Society for the Promotion of Science (No. 21500209), 2009-2012.

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, *In the Proceedings of International Conference on Speech Prosody 2001*, Taejon, Korea, pp.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, *In the Proceedings of 2nd International Conference on Speech Prosody*, Nara, Japan, pp.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, *In the Proceedings of 2nd International Conference on Speech Prosody*, Nara, Japan, pp.651-654, 2004.

[4] *The Proceedings of ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, 2000.

[5] Scherer, K. R., Vocal Affect Expression: A Review and a Model for Future Research, *Psychological Bulletin*, Vol. 99, No. 2, pp.143-165, 1986.

[6] Banse, R. and Scherer, K. R., Acoustic Profiles in Vocal Emotion Expression, *Journal of Personality and Social Psychology*, Vol. 70, No. 3, pp.614-636, 1996.

[7] Schröder, M. and Grice, M., Expressing vocal effort in concatenative synthesis, *In the Proceedings of 15th ICPhS*, Barcelona, pp.2589- 2592, 2003.

[8] Juslin, P. N. and Laukka, P., Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?, *Psychological Bulletin*, Vol. 129, No. 5, pp.770–814, 2003.

[9] Nakayama, T., Ichikawa, A., Nakada, K., and Miura, T., Control rules of sound-source characteristics for speech synthesis, *Tech. Rep. Speech*, 1969.

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

[11] Kobayashi Y. and Niimi, Y., On a prosodic information control method that reflects emotions in speech, *In the Proc. Fall Meet. Acoust. Soc. Jpn. 2-8-7*, pp.233-234, 1993.

[12] Kawanami, H. and Hirose, K., Considerations on the prosodic features of utterances with attitudes and emotions, *Tech. Rep. IEICE*, SP97-67, pp.73-80, 1997.

[13] Hirose, K., Minematsu, N., and Kawanami, H., Analytical and perceptual study on the role of acoustic features in realizing emotional speech, *In the Proc. ICSLP2000*, Beijing, China, pp.369-372, 2000.

[14] Bänziger, T. and Scherer, K., Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus, *In Affective Computing and Intelligent Interaction*, pp. 476-487, 2007.

[15] Fujisaki, H. and Hirose, K., Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *J. Acoust. Soc. Jpn (E)*, 5(4), pp.233-242, 1984.

[16] Liscombe, J., Venditti, J., and Hirschberg, J., Classifying Subjective Ratings of Emotional Speech Using Acoustic Features, *In the Proceedings of Eurospeech 2003*, Geneva, pp.725-728, 2003.

[17] Tsuru, M., Takeda, S., Nakasako, N., and Nakagawa, H., A Study of Prosodic Features of Emotional Speech Based on the Auditory Impressions, *Memoirs of the School of Biology-Oriented Science and Technology of Kinki University*, 23, pp.1-14, 2009.

[18] Kasuya, H., Maekawa, K., and Kiritani, S., Joint Estimation of Voice Source and Vocal Tract Parameters as Applied to the Study of Voice Source Dynamics, *In the Proc. ICPhS99*, San Francisco, California, pp.2505-2512, 1999.

[19] Kasuya, H., Yoshizawa, M., and Maekawa, K., Roles of Voice Source Dynamics as a Converter of Paralinguistic Features, *In the Proc. ICSLP2000*, Paper #1283, Beijing, China, 2000.

[20] Takeda, S., Yasuda, Y., Isobe, R., Kiryu, S., and Tsuru, M., Analysis of Voice-Quality Features of Speech that Expresses "Anger", "Joy", and "Sadness" Uttered by Radio Actors and Actresses, *Interspeech 2008*, Brisbane, Australia, pp.2114-2117, 2008.

[21] Takeda, S,, Asakawa, Y., and Ichikawa, A., A Comparison of Synthetic Speech Quality Generated by the Residual- and Multipulse-Excited Analysis-Synthesis Methods, *Electronics and Communications in Japan*, Part 3, 74(4), pp.97-105, 1991.