Comparison of Spectral-tilt Features of Emotional Speech Depending on the Degree of Emotions
– Research on Emotional-speech Synthesis based on Voice-quality Conversion –

Yoshiki KABUTA*, Tatsuya TANIGUCHI*, Masashi HATOKO*, Hideaki MATSUI*, Tomomi FUKUMOTO* and Shoichi TAKEDA**

* Graduate School of Biology-Oriented Science and Technology, Kinki University, 930 Nishi-Mitani, Kinokawa-shi, Wakayama, 649-6493, Japan
** Faculty of Biology-Oriented Science and Technology, Kinki University, 930 Nishi-Mitani, Kinokawa-shi, Wakayama, 649-6493, Japan

Abstract: This paper compares the voice quality features of emotions: “hot anger”, “cold anger”, “joy”, “hot sadness (crying type)”, and “cold sadness (dispirited and whispering type)” depending on the degree of the emotions in Japanese speech. Among voice-quality features, we place a focus on a spectral tilt. The analysis results show that the spectral-tilt quantities for “hot-anger” and “joy” speech are degree-dependent, i.e., the spectral-tilt quantities for these emotional speech increase significantly as the degree of the emotions becomes greater. On the other hand, the spectral-tilt quantities for “cold anger” and “cold sadness” speech do not show any significant degree-dependency, and even though those for “hot sadness” speech seem to have specific characteristics, they are speaker-dependent. These results confirm that the voice quality of “hot anger” and “joy” changes to the one whose higher-frequency band is more emphasized as the degree of the emotions increases, but that the voice-quality changes of “cold anger”, “hot sadness”, and “cold sadness” do not show any specific degree-dependency.

Keywords: Hot Anger, Cold Anger, Speech Synthesis, 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 analyzed the prosodic features of various emotional expressions to achieve more natural sounding rule-based synthetic speech [4].

The importance of research on emotional expressions has been widely recognized. Such relevant works include Scherer [12], Banse and Scherer [1], Schröder and Grice [13], Juslin and Laukka [6]. Among them, however, reports on Japanese speech synthesis were few [3].

In early studies, researchers analyzed rough features of typical emotional expressions such as “joy,” “anger,” etc. and/or synthesized emotional speech based on these features [8, 9]. 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 detailed emotional nuances.

We have been taking more detailed approach instead of investigating various types of typical emotional expressions roughly. We have divided such emotions as “anger”, “joy”, and “sadness” into four degrees: “neutral”, “light”, “medium”, and “strong” and features of each degree have been analyzed [4]. 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 was divided into several degrees and it was pursued how the features differed depending on the degree of emotion [5]. Recently, Bänziger and Scherer investigated the features of emotional expressions in languages other than Japanese in terms of the degree of emotion [2].

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 [18] features were investigated.

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 noise levels of speech, we measured the noise levels of the predictive residual signal of speech that expressed several degrees of each emotion and clarified quantitatively that the noise levels differed depending on the type and degree of emotion [16].

We have also done research on the features of the spectral-tilt quantities of speech. So far, many researchers have done work on spectral-tilt features [7, 10].

What is novel in our research is that we have clarified the spectral-tilt features of multiple degrees of emotions in Japanese speech. Our previous paper reported the spectral-tilt features of “anger”, “joy” [17], and “sadness” [14] groups.

This paper reports the analysis results of spectral-tilt features of emotional speech using a new emotional-speech database. In this database, emotion types are divided in more details than those in the previous database: we divide “anger” into two types: “hot anger” and “cold anger”, and “sadness” into two types: “hot sadness” and “cold sadness”.

2. OVERVIEW OF EMOTIONAL SPEECH SYNTHESIS SYSTEM DESIGN

Based on our findings clarified so far, we propose an emotional speech synthesis system as shown in Figure 1, consisting of prosody conversion as well as voice-quality conversion functions.

This system can produce multiple degrees of emotional speech from “neutral” prosodic as well as voice-quality parameters. This system will be applicable to corpus-based speech synthesis including statistical parametric speech synthesis that is being widely used [11]. This time, however, we try to apply our proposed method to our available system, a conventional residual-excited PARCOR synthesis system [15].

Prosody conversion can be achieved by using the knowledge of prosodic features of emotional speech [4].

In voice-quality conversion, an excitation conversion method is being developed. Spectral-tilt conversion can be achieved if we apply knowledge of the spectral-tilt features that have been accumulated so far [14, 17].

Among the voice-quality conversion depicted in Figure 1, this paper focuses on spectral-tilt conversion.

3. ANALYSIS CONDITIONS AND SPEECH MATERIALS

Speech samples were recorded at the sampling frequency of 48 [kHz] and down-sampled to 8 [kHz] for analysis. The Hamming window was used for FFT calculations and the window length was 64 [ms].

The speakers were two radio actors and two radio actresses in their 30s and 40s. As speech samples, we used 4-mora Japanese words that had accent types: middle is 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 “hot anger”, “cold anger”, “hot sadness”, “cold sadness” and “joy”. Here, “cold sadness” refers to a dispirited and whispering type of sadness in contrast to “hot sadness” referring to a crying type of sadness, both of which were observed in our both new and old emotional-speech databases. Each word was uttered with the following three degrees of the emotions: “neutral”, “medium”, and “strong”. While the “degrees” had been subjectively defined by the speakers themselves in our old database, the “degree” was clearly standardized emotion-by-emotion by demonstrating the speech samples to the speakers in our new database.

This paper reports the analysis results of spectral-tilt features of emotional speech using a new emotional-speech database. In this database, emotion types are divided in more details than those in the previous database: we divide “anger” into two types: “hot anger” and “cold anger”, and “sadness” into two types: “hot sadness” and “cold sadness”.

<table>
<thead>
<tr>
<th>Table 1: Words for analysis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
</tr>
<tr>
<td>“manamana”</td>
</tr>
<tr>
<td>“minimimi”</td>
</tr>
<tr>
<td>“munumunu”</td>
</tr>
<tr>
<td>“menemene”</td>
</tr>
<tr>
<td>“monomono”</td>
</tr>
</tbody>
</table>
4. ANALYSIS RESULTS

4.1 Male Speaker in His 30s

Figures 2-6 show the analysis results of spectral-tilt quantities for emotional speech uttered by a radio actor in his 30s. In the figures, the height of each bar denotes a mean value of 5-sample spectral-tilt quantities and each error bar denotes a standard deviation.

- **Figure 2**: Analysis results (speaker: radio actor in his 30s; emotion: “hot anger”).
- **Figure 3**: Analysis results (speaker: radio actor in his 30s; emotion: “cold anger”).
- **Figure 4**: Analysis results (speaker: radio actor in his 30s; emotion: “hot sadness”).

4.2 Male Speaker in His 40s

Figures 7-11 show the analysis results of spectral-tilt quantities for emotional speech uttered by a radio actor in his 40s.

- **Figure 5**: Analysis results (speaker: radio actor in his 30s; emotion: “cold sadness”).
- **Figure 6**: Analysis results (speaker: radio actor in his 30s; emotion: “joy”).
- **Figure 7**: Analysis results (speaker: radio actor in his 40s; emotion: “hot anger”).
4.3 Female Speaker in Her 30s
Figures 12-16 show the analysis results of spectral-tilt quantities for emotional speech uttered by a radio actress in her 30s.

4.4 Female Speaker in Her 40s
Figures 17-21 show the analysis results of spectral-tilt quantities for emotional speech uttered by a radio actress in her 40s.
5. SUMMARY OF ANALYSIS RESULTS

The analysis results are summarized as follows.

“hot anger”: speaker-dependent
  type1: neutral ≈ weak < medium < strong
  type2: neutral ≈ weak < medium > strong
“cold anger”: neutral ≈ weak ≈ medium ≈ strong
“hot sadness”: speaker-dependent
  type1: neutral > weak. medium ≈ strong
  type2: neutral ≈ weak < medium
“cold sadness”: weak ≈ medium ≈ strong
“joy”: speaker-dependent
  type1: weak < neutral < medium ≈ strong
  type2: neutral << weak ≈ medium ≈ strong

6. CONCLUSIONS

This paper has analyzed spectral-tilt features of emotions: “hot anger”, “cold anger”, “joy”, “hot sadness”, and “cold sadness” depending on the degree of the emotions in Japanese speech.

The analysis results have shown that the spectral-tilt quantities for “hot-anger” and “joy” speech are degree-dependent, i.e., the spectral-tilt quantities for these emotional speech increase as the degree of the emotions becomes greater, even though their detailed tendencies are speaker-dependent. On the other hand, the spectral-tilt quantities for “cold anger” and “cold sadness” speech do not show any significant degree-dependency, and even though those for “hot sadness” speech seem to have specific characteristics, they are speaker-dependent.

These results confirm that the voice quality of “hot anger” and “joy” changes to the one whose higher-frequency band is more emphasized as the degree of the emotions increases, but that the voice-quality changes of “cold anger”, “hot sadness”, and “cold sadness” do not show any specific degree-dependency.

Future studies will be to analyze the degree-dependency in more details, to determine the voice-quality conversion parameters precisely by analyzing the features of several types and degrees of emotional speech, to synthesize such emotional speech using these conversion parameters, and to evaluate the voice-quality-conversion method by a perception test.

ACKNOWLEDGEMENTS

The authors express their sincere appreciations to 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

Comparison of Spectral-tilt Features of Emotional Speech Depending on the Degree of Emotions


Yoshiki KABUTAR
Yoshiki Kabuta received a B.E. in Electronic Systems and Information Engineering in 2011 from Kinki University. He is currently a graduate student in master’s course at Kinki University and is engaging in research on speech synthesis-by-rule, specifically, on emotional speech synthesis.

Tatsuya TANIGUCHI
Tatsuya Taniguchi received a B.E. in Electronic Systems and Information Engineering in 2012 from Kinki University. From 2011 to 2012, he was engaged in research on speech synthesis-by-rule, specifically, on emotional speech synthesis. In 2012, he joined Forum Engineering Inc. and he is currently with Engineer Division of Forum Engineering Inc.

Masashi HATOKO
Masashi Hatoko received a B.E. in Electronic Systems and Information Engineering in 2012 from Kinki University. He is currently a graduate student in master’s course at Kinki University and is engaging in research on speech synthesis-by-rule, specifically, on emotional speech synthesis and analysis of brain responses when uttering or listening to emotional speech.

Hideaki MATSUI
From 2011 to 2012, Hideaki Matsui was engaged in research on speech synthesis-by-rule, specifically, on emotional speech synthesis.

Tomomi FUKUMOTO
Tomomi Fukumoto received a B.E. in Electronic Systems and Information Engineering in 2012 from Kinki University. From 2011 to 2012, she was engaged in research on digital archives for important cultural properties such as traditional Buddhist buildings and analysis of voice-quality features of emotional speech. In 2012, she joined Nippon Steel & Sumikin Plant Co., Ltd. and she is currently with System Engineering Division of Nippon Steel & Sumikin Plant Co., Ltd.

Shoichi TAKEDA
Shoichi Takeda received an M.E. in Mechanical Engineering for Production in 1974 and a Dr. Eng. degree in Electronic Engineering in 1991 from the University of Tokyo. In 1974, he joined Central Research Laboratory, Hitachi, Ltd., and attended Stanford University on a grant from the Rotary Foundation in 1974-1975. After returning to Japan he was engaged from 1975 to 1991 in research on recognition by artificial vision, speech synthesis-by-rule, etc. From 1991 to 2006, Dr. Takeda was with Teikyo Heisei University. Since 2006, Dr. Takeda has been professor at Kinki University and has been conducting research not only on speech synthesis-by-rule, but also on pathos (kansei) information processing including analysis of emotions by EEG or fNIRS measurements, and on digital archives for important cultural properties such as traditional Buddhist buildings. He is a senior member of the IEEE, and a member of the Japan Society of Kansei Engineering, the New York Academy of Sciences, the Institute of Electronics, Information and Communication Engineers, the Acoustical Society of Japan, etc. He received the Paper Award from the Japan Society of Kansei Engineering in 2006.