Analysis of Prosodic Features of Emotional Expressions in Noh Farce ("Kyohgen") Speech according to the Degree of Emotion

Muhd Dzulkhiflee Hamzah1, Shoichi Takeda2, Teruo Muraoka3 & Takumi Ohashi3

1Graduate School of Information Systems, The University of Electro-Communication, Tokyo
2Graduate School of Informatics, Teikyo Heisei University, Chiba
3Graduate School, Musashi Institute of Technology, Tokyo

kifuri74@tlab.is.uec.ac.jp, takeda@gakushikai.jp, t-muraoka@bea.hi-ho.ne.jp, takumi-i@nifty.com

Abstract

We analyzed the prosodic features of “anger,” “joy,” and “sadness” expressions in Noh farce ("Kyohgen") speech depending on the degree of the emotions. The degrees were divided into the following four categories: “neutral,” “low,” “medium” and “high.” A male Noh comedian uttered 6 words that are phonetically similar with one another five times, and the parameters of the prosodic features in the current study were speech rate and fundamental frequency. The analysis results showed the following: (1) Anger: Speech rate decreases when the speaker speaks with “anger.” Maximum fundamental frequencies increase with the increase of the degrees of anger. (2) Joy: Speech rate decreases and maximum fundamental frequencies increase when speaking with “joy.” (3) Sadness: Speech rate decrease with the increase of the degrees of sadness. No conspicuous tendency was found in maximum fundamental frequencies.

1. Introduction

Synthetic speech is being used in various fields, and there is a growing need for synthetic speech of a greater diversity that would include for example read speech and spontaneous conversational speech. Since spontaneous conversational speech is particularly diverse, we must accumulate knowledge through analysis of such speech to enable us to synthesize its various styles.

It is necessary to pay attention not only to linguistic information, but also to para- and non-linguistic information when the target of analysis or synthesis is conversational speech. In other words, it is necessary to analyze the features of various prosodic styles as well as emotional expressions to achieve more natural-sounding rule-based synthetic speech. We have therefore been conducting research on emotional expressions in speech for several years [1-5].

The importance of research on emotional expressions has been widely recognized, and workshops specializing in emotional expressions have been held. In the ISCA Workshop 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. Among them however, reports on Japanese speech synthesis were few.

In the early stage, Nakayama et al. analyzed and synthesized emotional speech [7]. Later, Kitahara and Tohkura [8], Kobayashi and Niimi [9], 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 categories: “neutral,” “displeasure,” “anger,” and “fury.” And features of each category have been analyzed [1, 2]. As the next step, we have recently analyzed the prosodic features of “joy” [5], “sadness” [5], and “gratitude” [3] using the same approach.

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” [10, 11]. 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.

In our studies, we have tried to clarify prosodic features comprehensively; not only the features of temporal structures and fundamental frequencies, but also those of speech power.

This time we pick up speech used in Noh farces. A Noh farce, called “Kyohgen” in Japanese, is a traditional Japanese comedy played in the midst of Noh plays. It was originated about 600 years ago and is still being played. The reason for selecting Noh farce speech is that we think that Noh comedians intentionally emphasize various emotions in plays and that it is easy to extract the essence of the emotions.

This paper reports the analysis results of prosodic features of each emotion mainly by observing the temporal structures and the fundamental frequencies.

2. Measuring method and analysis conditions

The prosodic parameters used in this experiment are (1) temporal structures, and (2) fundamental frequency ($F_0$).

The temporal structure was measured as phoneme duration by viewing a time-scale-enlarged speech waveform, a sound spectrogram, and a spectral differential coefficient in combination.
Table 1: Analysis conditions for pitch extraction and sound-spectrogram production and display.

<table>
<thead>
<tr>
<th>Processing Item</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/D conversion</td>
<td>Sampling frequency 8 (kHz)</td>
</tr>
<tr>
<td></td>
<td>Bit length 16 (bit)</td>
</tr>
<tr>
<td>Pitch extraction</td>
<td>Window length 32 (ms)</td>
</tr>
<tr>
<td></td>
<td>Window type Hamming</td>
</tr>
<tr>
<td></td>
<td>Frame-shift interval 10 (ms)</td>
</tr>
<tr>
<td></td>
<td>Pitch extraction method</td>
</tr>
<tr>
<td></td>
<td>Applying AMDF to low-pass filtered speech and interpolation by sinc function.</td>
</tr>
<tr>
<td>Pitch error correction</td>
<td>Manual correction using graphic dialog system.</td>
</tr>
<tr>
<td>Sound-spectrogram production and display</td>
<td>Window length 16 (ms)</td>
</tr>
<tr>
<td></td>
<td>Window type Hamming</td>
</tr>
<tr>
<td></td>
<td>Frame-shift interval 2.5 (ms)</td>
</tr>
</tbody>
</table>

Table 1 lists the analysis conditions for pitch extraction and sound-spectrogram production and display.

3. Experimental method

The speaker was a professional male Noh comedian in his 50s. He has been playing Noh farces for 56 years. Noh farce or “Kyogen” is a Japanese traditional play, which uses a specific way of expression. Only persons who have been trained for a long time can express it naturally. Since there are only few existing professional Noh comedians with a mastery skill, it is very difficult to find a speaker for this experiment.

As speech materials, we used the following 5-mora words most parts of which consisted of common syllables:

1. For anger and joy, “Migotodazo” (meaning “Excellent!” or “Well done!”) and “Migotodaro” (“I did well!”) in the role of a male, and “Migotodawa” (“Excellent!”) and “Migotodesho” (“I did well!”) in the role of a female.

Each word was uttered with the following four degrees of emotions: “neutral,” “low,” “medium” and “high.”

The speaker uttered 5 times a word and “neutral” words were used as references for all the types of emotion. The number of emotion types was therefore 10 and the number of roles were two (male and female), so the total number of speech samples was 100.

4. Temporal features

Mean word speech rate was measured as the feature of temporal structures. The purpose of this experiment is to investigate how this feature changes depending on the degrees of anger, joy, and sadness.

4.1. Results

Figure 1 compares the statistical quantities of word speech rate between different degrees of (a) anger, (b) joy, and (c) sadness uttered by the male Noh comedian. Each word consists of 5 speech samples.

The results for anger are shown in Figure 1(a). The mean speech rate tended to decrease in the order of “neutral,” “low,” and “medium,” but tended to increase for “high.” The statistical tests confirm significant differences at the 1% level between “neutral” and “low,” “low” and “medium,” and “medium” and “high.” This result is different from our
previous studies on the speech uttered by announcers and radio actors, which reported the opposite tendency when the degree of anger becomes higher. The peculiar style of speech in Noh farce might be the reason for this difference.

In the case of joy, there is a slightly increasing tendency from “neutral” to “low,” but decreasing tendency when the degrees of joy become larger (see Figure 1(b)). A statistical test shows a significant difference at the 1% level between all the degrees of joy, but there is not any difference between “neutral” and “low.”

Figure 1(c) shows the result on sadness. The mean speech rate tended to increase from “neutral” to “low,” but tended to decrease when the degree of sadness becomes larger. There is a significant difference at the 1% level between all the degrees of sadness (except for the difference at the 5% level between “neutral” and “low”).

5. $F_0$ features

The parameter that is thought to attract auditory attention most may be the “maximum fundamental frequency $F_{0\text{max}}$” in a word or sentence, especially when expressing the emotion such as anger and joy. Other candidates of feature parameters may be those of Fujisaki’s Model [12] such as magnitude of phrase and accent, minimum fundamental frequency, etc.

The purpose of this experiment is to clarify how the maximum fundamental frequencies change with the degrees of (1) anger, (2) joy, and (3) sadness.
5.1. Results

The features of fundamental frequency measured in our study are (1) magnitude of phrase command, (2) magnitude of accent command, and (3) maximum $F_0$. But the results in our analysis on phrase and accent commands show that there is no significant difference between the degrees of emotions used in our study.

Figure 2 shows a comparison of statistical quantities of maximum $F_0$ between different degrees of (a) anger, (b) joy, and (c) sadness. As shown in Figure 2(a) and 2(b), maximum $F_0$ tended to increase as the degrees of anger and joy becomes larger. No tendency was found in maximum fundamental frequencies. The statistical test results are shown in Table 2.

Table 2: Summary of statistical test results done over the degrees of anger, and joy.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>“neutral” and “Low”</th>
<th>“Low” and “Medium”</th>
<th>“Medium” and “High”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>p = 0.011</td>
<td>p = 0.033</td>
<td>p = 0.82</td>
</tr>
<tr>
<td>Joy</td>
<td>p = 0.117</td>
<td>p &lt; 5.13E-5</td>
<td>p &lt; 2.07E-4</td>
</tr>
<tr>
<td>Sadness</td>
<td>p = 0.209</td>
<td>p = 0.127</td>
<td>p = 0.476</td>
</tr>
</tbody>
</table>

Figure 3 shows comparison of $F_0$ contours for different degrees of joy: (a) low, (b) medium, and (c) high. Measured $F_0$ values (symbols “-x”) and their best approximations obtained by Fujisaki’s model are both shown. The dotted curves show the phrase commands, and the stepwise waveforms show the accent commands. Quantitative analyses will be carried out in future.

6. Conclusions

The degrees of (1) anger, (2) joy, and (3) sadness have been divided into the following four categories: “neutral,” “low,” “medium,” and “high.” We used a set of speech uttered by a male Noh comedian and conducted an analysis to see how the speech rate, and fundamental frequency change in accordance with the degrees of the emotions. Analysis results are as follows:

(1) Anger: Speech rate decreases when the speaker speaks with “anger.” In Noh farce, the actor might speak faster to express “anger” at the highest degree. Maximum fundamental frequencies increase with the increase of the degrees of anger.

(2) Joy: Speech rate decreases and maximum fundamental frequencies increase when speaking with “joy.”

(3) Sadness: Speech rate decrease with the increase of the degrees of sadness. No conspicuous tendency was found in maximum fundamental frequencies.

Future studies will be to analyze more data from other Noh comedians, to carry out a perceptual rating, to compare the results with those obtained from the analyses for the announcers’ and radio actors’/actresses’ speech, and to analyze power features, spectral features, etc.

7. Acknowledgments

The authors would like to express their thanks to Noh comedian (and also instructor of Teikyo Heisei University) Mr. Juro Zenchiku for their help in uttering emotional Noh farce speech, to the helpful people at Hitachi's Central Research Laboratory for giving permission to use their analysis programs, and to Ms. Maiko Yamada and Ms. Yuko Takahashi of Musashi Institute of Technology for their help in producing Noh farce speech database.

This research was partly supported by Grant-in-Aid from Teikyo Heisei University as well as Grant-in-Aid for Scientific Research on Priority Areas(2) “Diversity of Prosody and its Quantitative Description” from the Ministry of Education, Culture, Sports, Science and Technology, Japan (No. 12132206).

8. References


