On the Differences in Prosodic Features of Emotional Expressions in Japanese Speech according to the Degree of the Emotion

Yasuki Hashizawa[†], *Shoichi Takeda*[†], *Muhd Dzulkhiflee Hamzah*[‡], *& Ghen Ohyama*[*]

† Graduate School of Informatics, Teikyo Heisei University, Chiba
‡ Graduate School of Information Systems, The University of Electro-Communication, Tokyo
*Brain Functions Lab., Kanagawa
twinkle_fairy56@hotmail.com, takeda@gakushikai.jp,
kifuri74@tlab.is.uec.ac.jp, ohyama@bfl.co.jp

Abstract

We analyzed the prosodic 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." Four announcers (two male and two female) uttered 6 words five times, and the parameters of the prosodic features were speech rate and fundamental frequency. The analysis results showed the following. (1) The most significant prosodic feature for expressing "anger" was to enhance the fundamental frequency. (2) The most significant prosodic feature for expressing "joy" was to enhance the fundamental frequency. A significant feature was to emphasize the accent. (3) As a method of expressing "sadness," the female speakers made their pitch low and suppressed accents (i.e., flattened the F_0 pattern). The male speakers did not seem to use prosodic features to express "sadness."

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 of it through analysis 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 [6] 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 of Japanese research, 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, merely gave a rough paradigm of emotional expressions such as "joy," "sadness," "anger," etc. They therefore left it to future studies to give rules on expressing minute emotional nuances.

We have been taking a minute approach instead of generally investigating various types of typical emotional expression. As the first step in our work, we have focused on "anger" expressions since their prosodic features are relatively clear. We categorized the degree of "Anger" into four categories: "neutral," "displeasure," "anger," and "fury." and the features of each category have been analyzed [1, 2]. As the next step, we recently analyzed the prosodic features of "joy" [5], "sadness" [5], and "gratitude" [3] by using the same approach.

There are still few reports on 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 them is an analysis of Japanese short sentence speech with the above types of emotion uttered by one speaker. Another report analyzed 6-mora Japanese word speech with "anger" as the expression uttered by three speakers. Both studies 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.

The analysis results we present here are of prosodic features of emotion that were gotten mainly by observing the temporal structures and the fundamental frequencies including Fujisaki's model parameters [12].

2. Experimental method

The speakers were two male and two female professional announcers in their 60s.

As speech materials, we used 4-mora and 6-mora sense and nonsense words that had either of the three accent types: flat, mid-high, or head-high. Each word was uttered with the following four degrees of emotion: "neutral," "light," "medium" and "strong."

The speakers were requested to utter the nonsense words "Manamana" and "Manamanamana" with the same accent, speech rate, and emotional expression as those for the immediately preceding sense word. They uttered 5 times a word. The total number of words per speaker was thus 360.

The prosodic-feature parameters used were (1) temporal structures (mean speech rate) and (2) fundamental frequency (F_0). The temporal structure was measured as phoneme

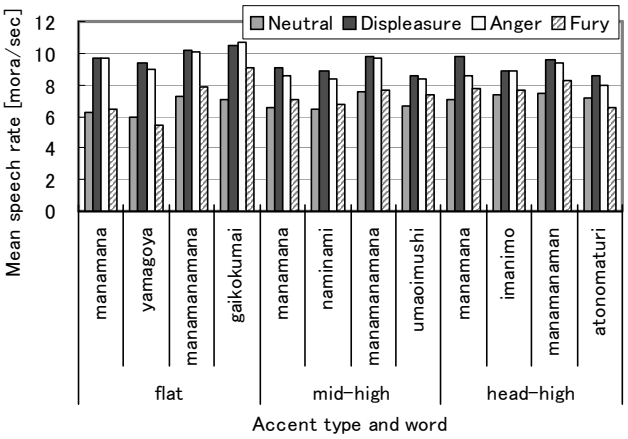

Figure 1: *Mean speech rate (male speaker IK).*

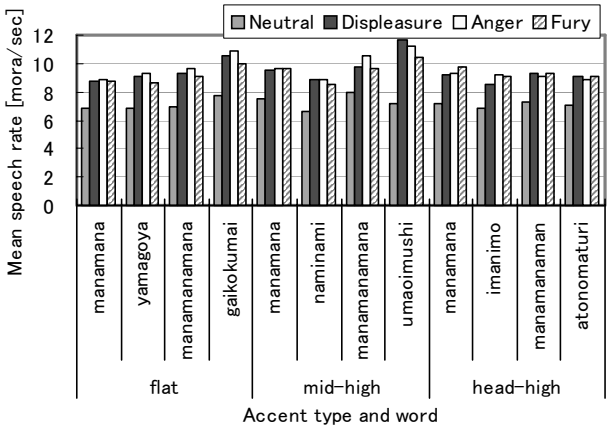

Figure 2: *Mean speech rate (male speaker TA).*

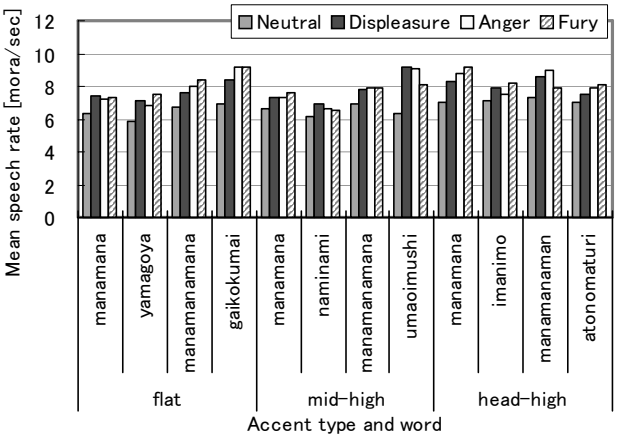

Figure 3: *Mean speech rate (female speaker MT).*

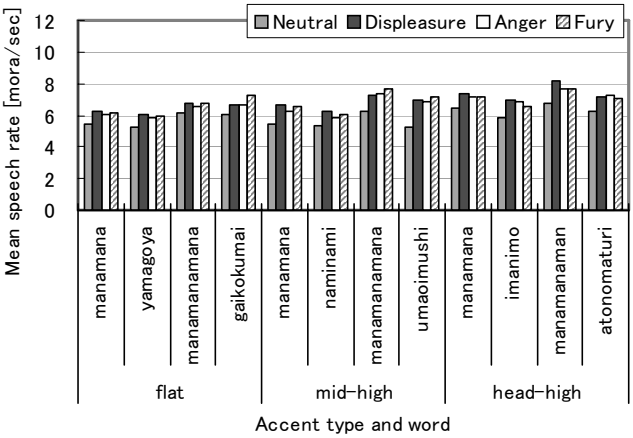

Figure 4: *Mean speech rate (female speaker YK).*

duration by viewing a time-scale-enlarged speech waveform, a sound spectrogram, and a spectral differential coefficient in combination.

3. Experimental results

3.1. Anger

Figures 1-4 show word speech rates for a set of single speech samples uttered by the four speakers. Each word consists of 5 speech samples. For male speaker IK, the mean speech rate tended to increase in the order of "neutral," "displeasure," and "anger," but, conversely, it tended to fall for "fury" regardless of accent type or number of morae. This decreasing tendency was also observed for male speaker TA, but it was word-dependent. For the two female speakers, however, there was an increasing tendency when the utterance had some degree of anger when compared with the "neutral" utterances, and almost no speech rate reduction was observed when the utterance was with "fury."

An increasing tendency was observed in the magnitude of phrase command A_p as the degree of "anger" increased, even though it depended on the accent type, the kind of word, and the number of morae. Henceforth, figures are omitted except for the F_{0max} for "anger."

An increasing tendency was observed in the magnitude of accent command A_a as the degree of "anger" increased regardless of speaker, even though it depended on the accent type, the kind of word, and the number of morae.

Figures 5-8 show experimental results of measuring the maximum fundamental frequency F_{0max}. Figures 5 and 6 show the cases for the male speakers, for which F_{0max} tended to become higher as the degree of anger became larger. This tendency was observed irrespective of accent type, kind of word, or number of morae.

Figures 7 and 8 show the cases for the female speakers. Even though the tendency of increasing F_{0max} was observed in this case as well, it is not as conspicuous as in the case of the male speakers.

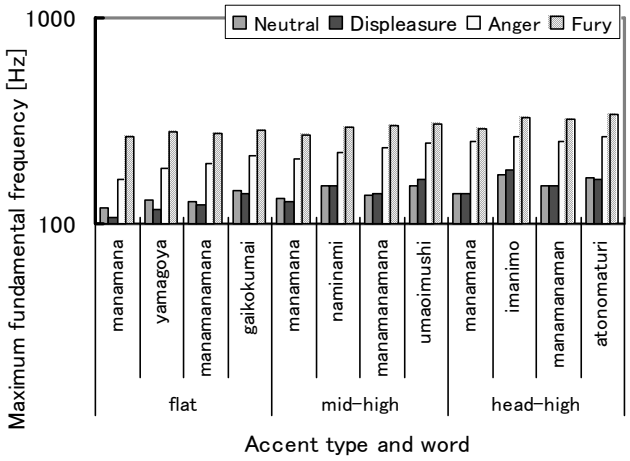

Figure 5: *Maximum fundamental frequency (male speaker IK).*

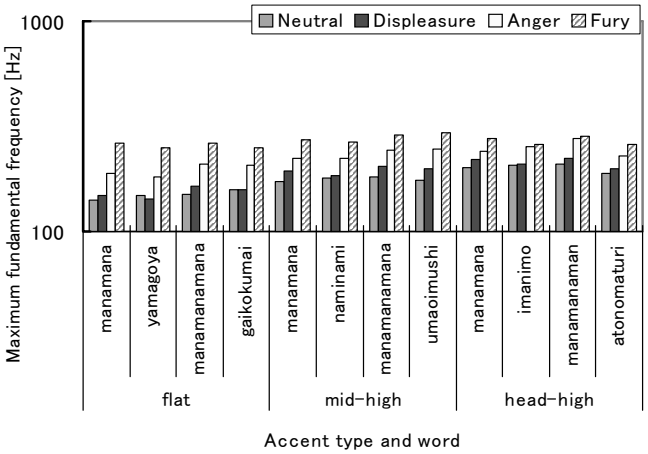

Figure 6: *Maximum fundamental frequency (male speaker TA).*

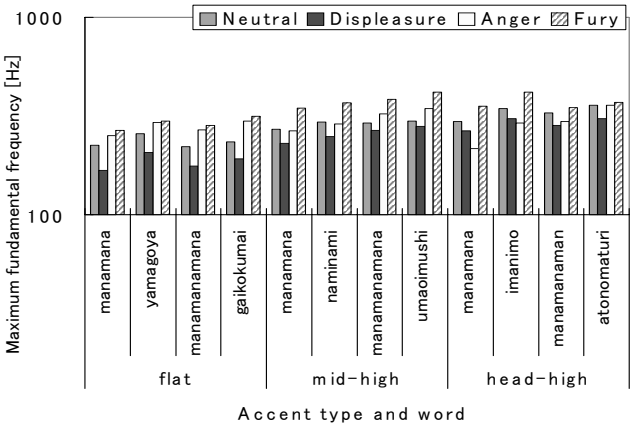

Figure 7: *Maximum fundamental frequency (female speaker MT).*

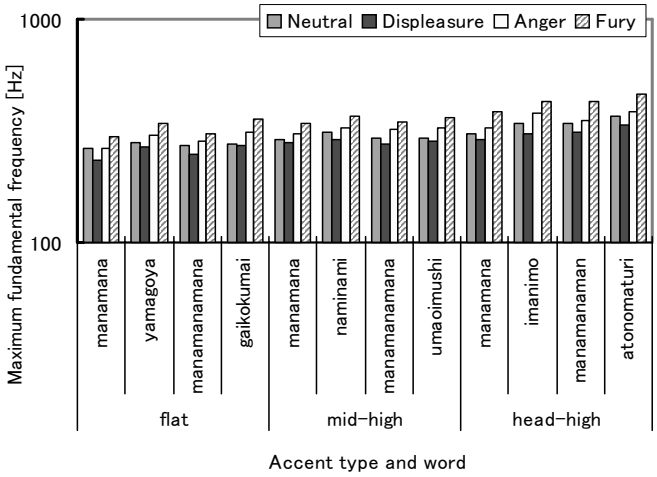

Figure 8: *Maximum fundamental frequency (female speaker YK).*

3.2. Joy

Neither the mean speech rate nor A_p showed any conspicuous tendencies.

An increasing tendency was observed in A_a as the degree of "joy" increased regardless of speaker. This was commonly observed and did not depend on accent type, kind of word, or number of morae. This tendency was more conspicuous for the male speakers than for the female speakers.

An increasing tendency was observed in F_{0max} as the degree of "joy" increased for all the speakers. This was commonly observed and did not depend on accent type, kind of word, or number of morae.

3.3. Sadness

No conspicuous tendencies were observed in the mean speech rate.

A decreasing tendency was observed both in A_p and A_a as the degree of "sadness" increased. This was observed only for the female speakers and it did not depend on accent type, kind of word, or number of morae.

Like A_p and A_a, a decreasing tendency was also observed in F_{0max}.

3.4. Summary of prosodic features

Table 1 summarizes the prosodic features we have obtained. These results are consistent with the previous findings with a smaller number of data [1, 2, 5].

4. Conclusions

We investigated the differences in the prosodic features according to the degree of emotion of "joy," "sadness," and "anger" for expressions in spoken Japanese words. Using 4- and 6-mora words that four announcers (two male and two female) uttered as speech materials, we analyzed the features of the mean speech rate, the magnitude of phrase command, the magnitude of accent command, and the maximum fundamental frequency. As a result, the following conclusions were obtained:

(1) The most significant prosodic feature for expressing "anger" is to enhance the fundamental frequency. However, the effective way to achieve more pronounced "anger" may be to combine several prosodic and non-prosodic features.

Table 1: *Summary of the prosodic features of various emotions*

Emotion	Gender	Temporal structure	Fundamental frequency			Note
		Mean speech rate	Magnitude of phrase command A_p	Magnitude of accent command A_a	Maximum fundamental frequency F_{0max}	
Anger	Male	Increase (anger) Reduction (fury)	Increase	Increase	Increase	
	Female	Increase				
Joy	Male	No tendency	No tendency	Increase	Increase	
	Female					
Sadness	Male	No tendency	No tendency	No tendency	No tendency	Speaker-dependent
	Female		Reduction	Reduction	Reduction	

(2) The most significant prosodic feature for expressing "joy" is to enhance the fundamental frequency. Another significant feature is to emphasize the accent. The utterance speed is not used. In other words, "joy" is expressed by making a pitch high and emphasizing an accent.

(3) As a method of expressing "sadness," the female speakers made their pitch low and suppress an accent (i.e., flatten the F_0 pattern). The utterance speed was not used. The male speakers did not seem to use prosodic features to express "sadness."

In the future, we will analyze a larger number of speech samples with more speakers, and we will analyze their power features, spectral features, etc.

5. Acknowledgments

The authors would like to express their thanks to the helpful people at Hitachi's Central Research Laboratory for giving permission to use their analysis programs, to the announcers at NHK for their help in uttering emotional speech, and to Mr. Naoki Saito, Mr. Tsutomu Kobayashi, Mr. Kazuto Ito, Mr. Masahiro Nashizawa, Mr. Yusuke Katagishi and other students of Teikyo Heisei University for their help in analyzing the data.

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).

6. References

[1] Takeda, S.; Ohyama, G.; Tochitani, A., 2001. 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.

[2] Takeda, S.; Ohyama, G.; Tochitani, A.; Nishizawa, Y., 2002. Analysis of prosodic features of "anger" expressions in Japanese speech. *J. Acoust. Soc. Jpn.* 58(9), 561-568. (in Japanese)

[3] Takeda, S.; Muhd Dzulkhiflee Hamzah, 2003. Comparison of prosodic features of "gratitude" expressions in spoken Japanese uttered by radio actor and actress depending on the degree of emotion. *Proc. Spring Meet. Acoust. Soc. Jpn.* 2-Q-34, 441-442. (in Japanese)

[4] Tochitani, A.; Takeda, S., 2003. Differences in the prosodic features of ATM guidance speech between with emotions and without them. *Proc. Spring Meet. Acoust. Soc. Jpn.* 2-Q-28, 429-430. (in Japanese)

[5] Muhd Dzulkhiflee Hamzah; Takeda, S.; Ohyama, G., 2003. Comparison of prosodic features of "joy" and "sorrow" expression in spoken Japanese uttered by a radio actor depending on the degree of emotion. *Proc. Fall Meet. Acoust. Soc. Jpn.* 2-Q-28, 367-368. (in Japanese)

[6] 2000. Proc. ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research.

[7] Nakayama, T.; Ichikawa, A.; Nakada, K.; Miura, T., 1969. Control rules of sound-source characteristics for speech synthesis. *Tech. Rep. Speech.* (in Japanese)

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

[9] Kobayashi Y.; Niimi, Y., 1993. On a prosodic-information control method that reflects emotions in speech. *Proc. Fall Meet. Acoust. Soc. Jpn.* 2-8-7, 233-234. (in Japanese)

[10] Kawanami, H.; Hirose, K., 1997. Considerations on the prosodic features of utterances with attitudes and emotions. *Tech. Rep. IEICE* SP97-67, 73-80. (in Japanese)

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

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