A Study of Prosodic Features of Emotional Speech Based on the Auditory Impressions

Makiko Tsuru[1], Shoichi Takeda[2], Noboru Nakasako[2] and Hideo Nakagawa[3]

Abstract

This paper describes the prosodic features of various types and degrees of emotional expressions in Japanese speech based on the auditory impressions. The speech samples consisted of "neutral" speech as well as speech with three types of emotions ("anger", "joy", and "sadness") of three degrees ("light", "medium", and "strong"). Prosodic-feature parameters were speech rate and F_0 parameters including those of Fujisaki's Model. In conversational communication, a speaker's emotion inside his/her mind is not necessarily reflected in his/her utterances, nor is conveyed to the listener in its original form. The purpose of this study is therefore to clarify quantitatively (1) how much the speaker's internal emotion (speaker's intention) is correctly conveyed to the listener, and furthermore, (2) what type of expression can convey the speaker's intension to the listener correctly. We first conducted a listening test to examine how much the speakers' intended emotions agreed with the listeners' auditory impressions, using 144-word speech samples uttered by radio actors and actresses. Subjects were 50 female college students of 19 and 20 years old. The test results showed that the subjects did not necessarily perceive emotional speech as the speakers intended to express. From these results, we learned that it was optimal for emotion communication to use speech that matched the auditory impressions of emotions as a model for synthesis rather than the speaker's intention. We therefore analyzed the features of prosodic parameters based on the emotional speech classified according to the auditory impressions of the subjects. Prior to analysis, we calculated an identification rate for each type and degree of emotion, which is a rate of the number of identifying as a specific type and degree of emotion to the total number of listeners. We selected 5 speech samples whose identification rate ranked the top 5 for each type and degree emotion. Analysis results were summarized as follows: (1) The magnitude of accent command, minimum fundamental frequency, and maximum fundamental frequency increased with the increase of degree of emotion. Contrarily, speech rate decreased with the increase of degree of emotion. (2) The magnitude of accent command was gender-dependent, i.e., that for anger speech uttered by the male speakers increased compared to that for neutral speech, and significant difference was observed from female speech. (3) Minimum fundamental frequency for anger speech uttered by the female speakers increased, and significant difference was observed from male speech. (4) Maximum fundamental frequency for all emotion of male speech increased, and also significant difference was observed from female speech. And (5) Prosodic features that characterized their emotions were speaker's gender-dependent.

1. Introduction

As information communication technology (ICT) advances, there are increasing needs for better human-machine communication tools. According to our experiment, expressive speech is more desirable than non-expressive speech as a means of man-machine dialog. However, the capability of synthesizing expressive speech including emotional speech is currently not high enough to match the needs. We have to explore features from natural speech to achieve a method for a variety of expressive-speech synthesis. Among expressive speech, we have so far placed a focus on emotional speech.

To achieve speech synthesis with rich emotions, we first analyzed the prosodic features of natural emotional speech regarding that prosody might be the most significant factor of emotional expressions in speech. Here, "prosody" is a general concept that consists of voice-pitch height, voice intensity, and a speech rate. The emotion types were "anger", "joy", "sadness", and "gratitude". The degrees of each emotion were "light", "medium", and "strong". Speakers were radio actors and actresses, announcers, a *Noh*-and-*Kyogen* stage actor, researchers, students, and so on [1]-[18].

Emotions in speech are not only characterized by prosodic features but also speaker's voice quality and other

Received 25 November 2008

This study was supported by the Project Research of the School of Biology Oriented Science and Technology No. 06-I-3, 2007-2009.

[1] Program in Electronic Systems and Information Engineering, Graduate School of Biology-Oriented Science and Technology, Kinki University, Wakayama 649-6493, Japan and Department of Business Career, Kurume Shin-Ai Women's College, Fukuoka 839-8508, Japan

[2] Department of Electronic Systems and Information Engineering, Kinki University, Wakayama 649-6493, Japan

[3] Department of Intelligent Systems, Kinki University, Wakayama 649-6493, Japan

features [19]. We have also been analyzing features of voice quality [20]-[22]. This paper, however, places a focus on prosodic features.

Generally speaking, the importance of research on emotional expressions has been widely recognized, and workshops specializing in emotional expressions were held. In the ISCA Workshop [23] 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 Interspeech 2008 [24], reports on emotion and/or expression increased so much that 6 sessions were on emotion and/or expression. However, reports on the relationship between prosodic parameters of Japanese emotional speech samples and their auditory impressions were still few, which aimed at controlling prosodic parameters for emotional speech synthesis.

In our analysis so far, the type and degree of each emotion has been determined by the speakers themselves. In conversational communication, however, a speaker's emotion inside his/her mind is not necessarily reflected in his/her utterances, nor is exactly conveyed to the listener as the speaker intended.

The purpose of this study [25]-[30] is therefore to clarify quantitatively (1) how much the speaker's internal emotion (speaker's intention) is correctly conveyed to the listener, and further, (2) what type of expression is able to convey the speaker's intension to the listener correctly.

We learned in our previous study that the styles of emotional expressions were speakers' gender-dependent: e.g., the features of fundamental frequency, which is one of the most significant prosodic-feature parameters of emotional expressions, is known to vary depending on the gender of speakers [26]. We therefore also took the gender feature into consideration.

We first conducted a listening test to examine how much the speakers' intended emotions agreed with the listeners' auditory impressions, using 144-word speech samples uttered by radio actors and actresses. Subjects were 50 female college students of 19 and 20 years old. The test results showed that the subjects did not necessarily perceive emotional speech as the speakers intended to express.

From these results, we learned that it was optimal for emotion communication to use speech that matched the auditory impression of emotion as a model for synthesis rather than the speaker's intention. We therefore analyzed the features of prosodic parameters based on the emotional speech classified according to the auditory impressions of the subjects. Prior to analysis, we calculated an identification rate for each type and degree of emotion, which is a rate of the number of identifying as a specific type and degree of emotion to the total number of listeners. We selected 5 speech samples whose identification rates ranked the top 5 for each type and degree of emotion.

This paper reports these listening-test results.

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 types of emotions were "anger", "joy", and "sadness". Each word was uttered with the following four degrees of the emotions: "neutral", "light", "medium", and "strong". As listed in Table 1, we used 144 speech samples consisting of all speakers, all accent types, and all emotional types and degrees.

Table 1: Speech samples.

(Number of category)

Speaker	Emotion	Degree	Accent type	Total
(4) male(2)	(3) anger joy sadness	(3) light medium strong	(3) flat mid-high head-high	108
female(2)	(1) neutral	(0)	(3) flat mid-high head-high	36

2.2 Prosodic-feature parameters

Prosodic-feature parameters were F_0 parameters, i.e., magnitude of accent command, magnitude of phrase command, and minimum fundamental frequency ($F_{0\min}$) in Fujisaki's model [31], maximum fundamental frequency ($F_{0\max}$), and speech rate (sk_eve). We did not use the speech power because the distances between the actors /

actresses and the microphone varied largely by their body movements during recording and we could not collect reliable power data.

2.3 Listening test

Speech samples were presented to the subjects in random order. There were 16 dummy samples ahead and 144 test samples. The listening test had two sessions. In the second session, the speech samples were presented to the subjects in the reverse order of those presented in the first session. The interval between two speech samples was three seconds except that the interval after consecutive 10 speech samples was 10 seconds. After a break of 5 minutes, we started the second session.

Fifty subjects used a headphone of the same maker and the same sound pressure. They were female college students of 19 and 20 years old with a normal auditory capacity. Figure 1 shows the answer form.

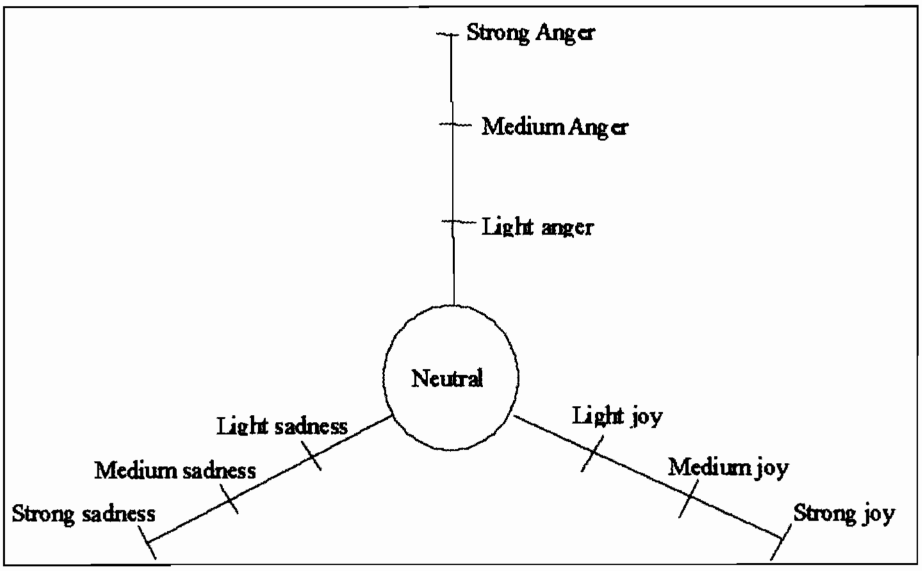

Figure 1: Answer form.

3. Experimental results

We divided the analysis of the experimental results into two groups depending on the gender of the speakers. To confirm the reliability of the subjects' answers, we compared between the first and the corresponding second answers of each subject. We called the rate of the number of the agreed answer pairs to the common number of the presented speech samples to both sessions a "repeatability rate". We used for analysis the subjects' answers whose repeatability rate was not less than 60%.

3.1 Agreement rate

First, as preliminary examination, we investigated how much the speaker's emotion was conveyed to the listener. For this purpose, we calculated the agreement rate for each speech sample. It was defined as the rate of agreement of the listener's receptivity with the speaker's intention in the type and degree of the emotion to the total number of the listeners. The agreement rate r (%) is therefore defined as

$$r = a / N * 100 \qquad (1)$$

where a is the number of the agreed answers and N is the total number of the listeners.

As listed in Table 2, the type and degree of emotion at the highest ranks of agreement rates were the same regardless of the speaker's gender, i.e., among the five highest ranks, the type and degree of emotion were all "neutral". When compared among emotion groups, however, the agreement rates for "neutral" were not necessarily the highest (Fig. 2).

Table 2: High ranks of agreement rates.

Rank	Emotion type	Agreement rate(%) (male)	Emotion type	Agreement rate(%) (female)
1	Neutral	100.0	Neutral	100.0
2	Neutral	100.0	Neutral	100.0
3	Neutral	95.8	Neutral	100.0
4	Neutral	95.8	Neutral	100.0
5	Neutral	95.8	Neutral	95.8

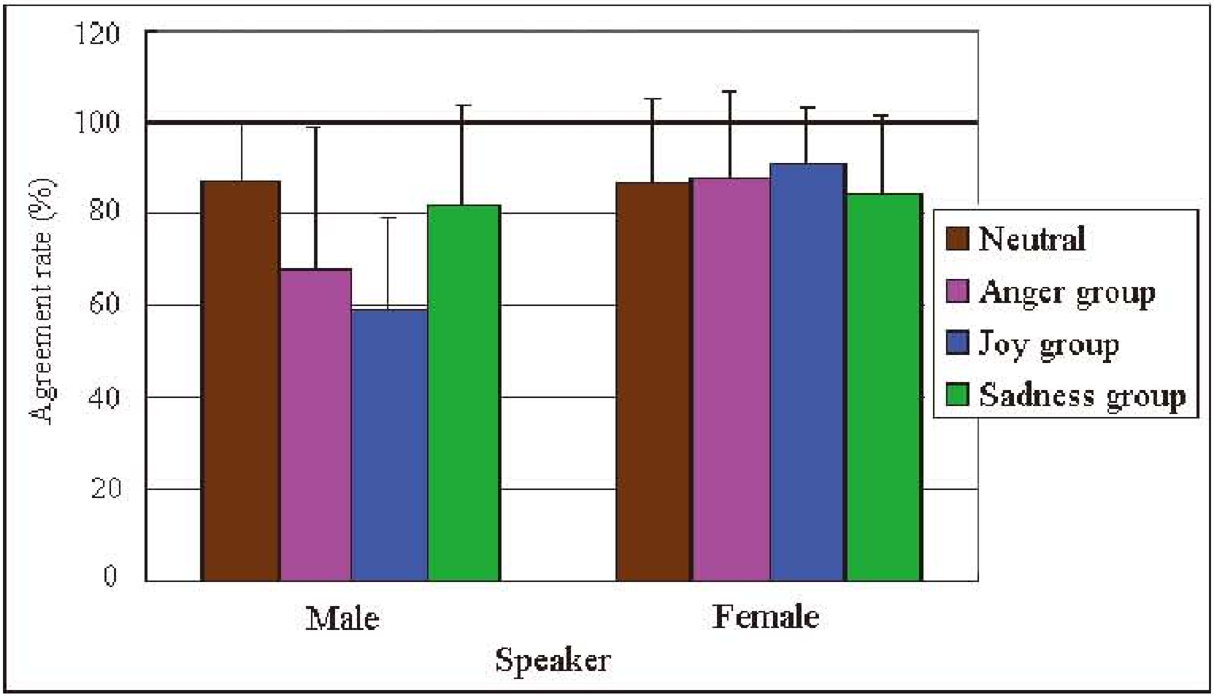

Figure 2: Agreement rate of each emotion group. The height of each bar denotes the mean value and the length of each error bar denotes the standard deviation.

Furthermore, although the average agreement rate of each type of emotion was high as seen in Fig.2, the agreement rate decreased when divided into degrees (see Fig. 3).

Some speech data with specific types and degrees of emotions had particularly strong negative correlations between the agreement rates and the prosodic-feature parameters. For "medium joy" uttered by the male speakers, the correlation coefficient r between the agreement rates and "average speech rate" was -0.9, and for "medium anger" it was also -0.9. For "strong anger" uttered by the female speakers, r between the agreement rates and "minimum fundamental frequencies (F_{0min})" was -0.9, and for "light sadness" it was also -0.9.

According to our observations, however, the emotion that a speaker intended to express was rarely conveyed to a listener accurately. We therefore considered as best for communication of emotions to use the speech that coincided with the auditory impressions of the listeners rather than the speakers' intention as a model of synthesis. We then analyzed the prosodic features of emotions that were classified based on the listeners' auditory impressions. The analysis results will be described in the next subsection.

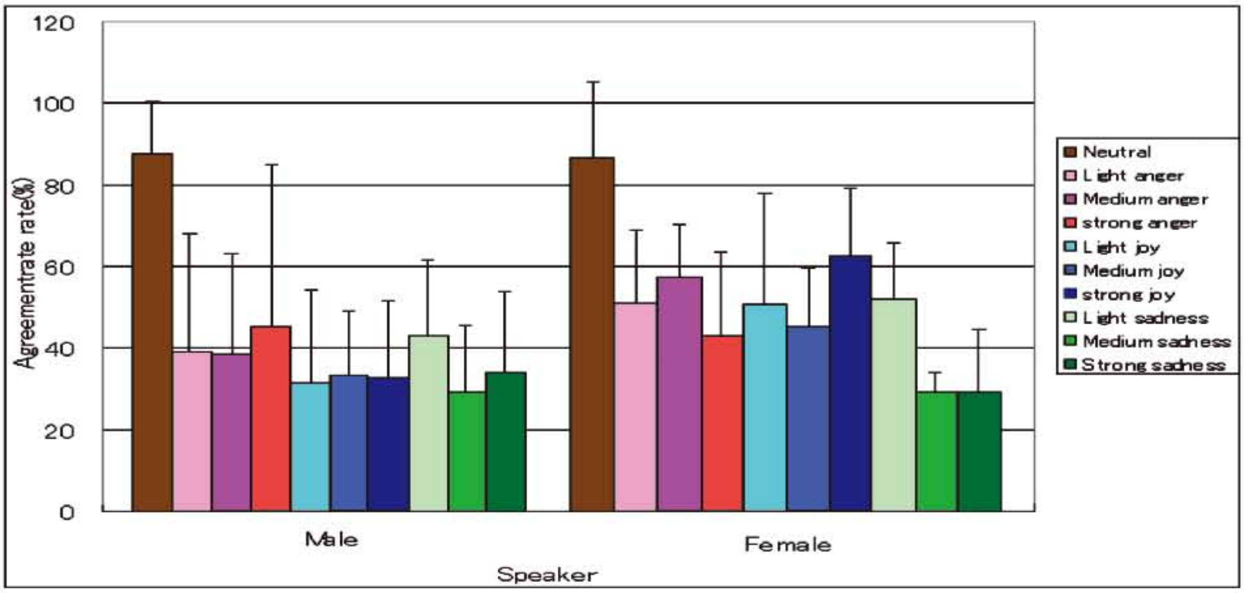

Figure 3: Agreement rate for each emotion type and degree.

3.2 Identification rate

To quantify the strength of listeners' auditory impressions, the quantity called an "identification rate" was introduced. We defined "an identification number" as the number of listeners' identification regardless of the type and degree of emotion that the speakers intended to express. In the same way, we defined "an identification rate" as a rate of the identification number to the total number of listeners.

There were few emotional speech samples of the identification rate more than 50% for each emotion type. However, the listeners must be more likely to catch the same impression for emotional speech whose identification rate was higher. We therefore compared prosodic-feature parameters of the speech with the identification rate for each type and degree of emotion.

For comparison, we extracted emotional speech samples whose identification rates ranked the top 5 for each type and degree of emotion. We then divided each prosodic-feature parameter of these speech samples by the same type of prosodic-feature parameter of speech in "neutral" group for normalization, and named this quotient "an Increase rate per Neutral" (henceforth, "Ir/N"). Then we averaged them by each type and degree of emotion.

Paying attention to the prosodic-feature parameters, we observed the following. In the case of the magnitude of accent command of speech uttered by the male speakers, as shown in Fig. 4, Ir/N was 1.5-2.5 when the subjects perceived the emotion as "anger", approximately 1.5 when they perceived it as "joy", and 0.5-1.5 when they perceived it as "sadness". In the case of the same parameter of speech uttered by the female speakers, Ir/N was 1.5-2.0 when the subjects perceived the emotion as "anger", 1.0-1.5 when they perceived it as "joy", and 0.5-1.0 when they perceived it as "sadness".

In the case of minimum fundamental frequency $F_{0\mathrm{min}}$ of speech uttered by the male speakers, as shown in Fig. 5, Ir/N was 1.5-2.5 when the subjects perceived the emotion as "joy", and 1.0-1.5 when they perceived it as "sadness". In the case of the same parameter of speech uttered by the female speakers, Ir/N was 1.0-1.5 when the subjects perceived the emotion as "anger", approximately 2.0 when they perceived it as "joy", and 1.0-2.0 when they perceived it as "sadness".

In the case of maximum fundamental frequency $F_{0\mathrm{max}}$ of speech uttered by the male speakers, as shown in Fig. 6, Ir/N was 1.5-2.5 when the subjects perceived the emotion as "anger", 1.5-2.5 when they perceived it as "joy", and 1.0-1.5 when they perceived it as "sadness". In the case of the same parameter of speech uttered by the female speakers, Ir/N was 1.0-1.5 when the subjects perceived the emotion as "anger" and "sadness", and 1.5-2.0 when they perceived it as "joy".

In the case of speech rate, the tendency was not similar to that of the F_0 parameters mentioned above. As shown in Fig. 7, Ir/N for speech rate was 0.5-1.0 for all types and degrees of the emotions.

And then, Ir/Ns for the magnitude of accent command, $F_{0\mathrm{min}}$, and $F_{0\mathrm{max}}$ increased as the degree of emotion became stronger. On the other hand, Ir/N for speech rate decreased as the degree of emotion became stronger. These results agreed with our previous findings [16].

Furthermore, we knew from Figs. 4-7 that some types of prosodic-feature-parameter values were different depending on the degrees of emotion for each type of emotion, and that they were also speaker's gender-dependent.

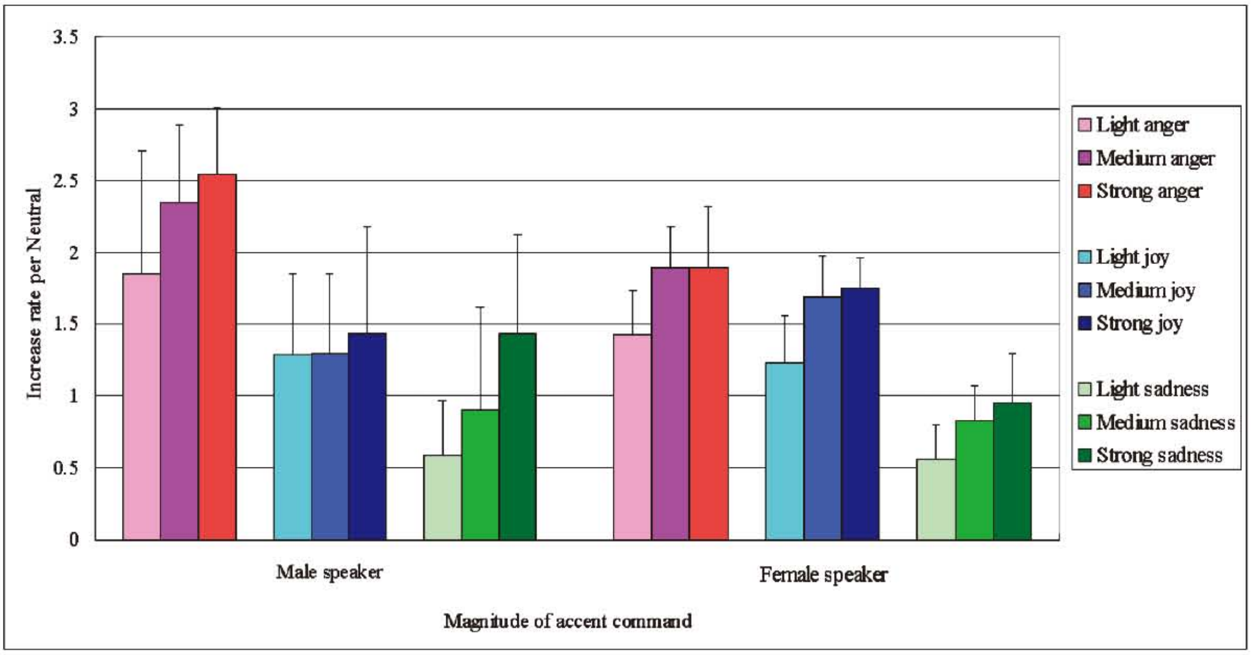

Figure 4: Increase rate per Neutral of magnitude of accent command.

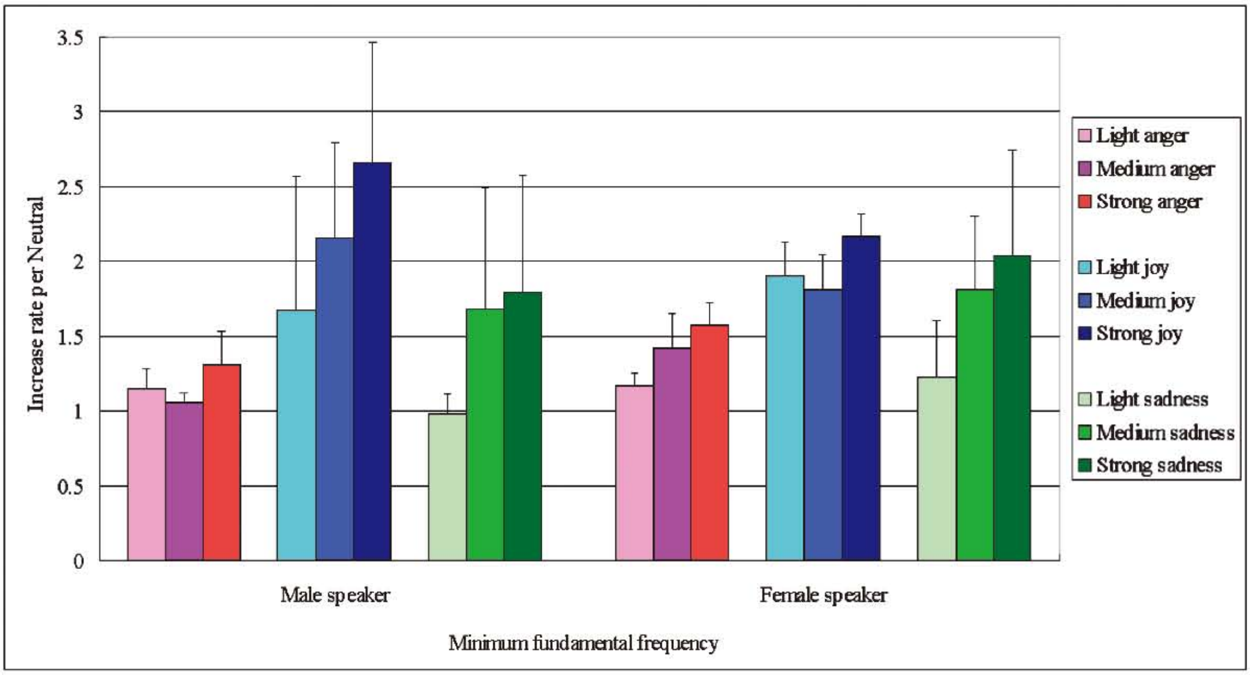

Figure 5: Increase rate per Neutral of Minimum fundamental frequency (F_{0min}).

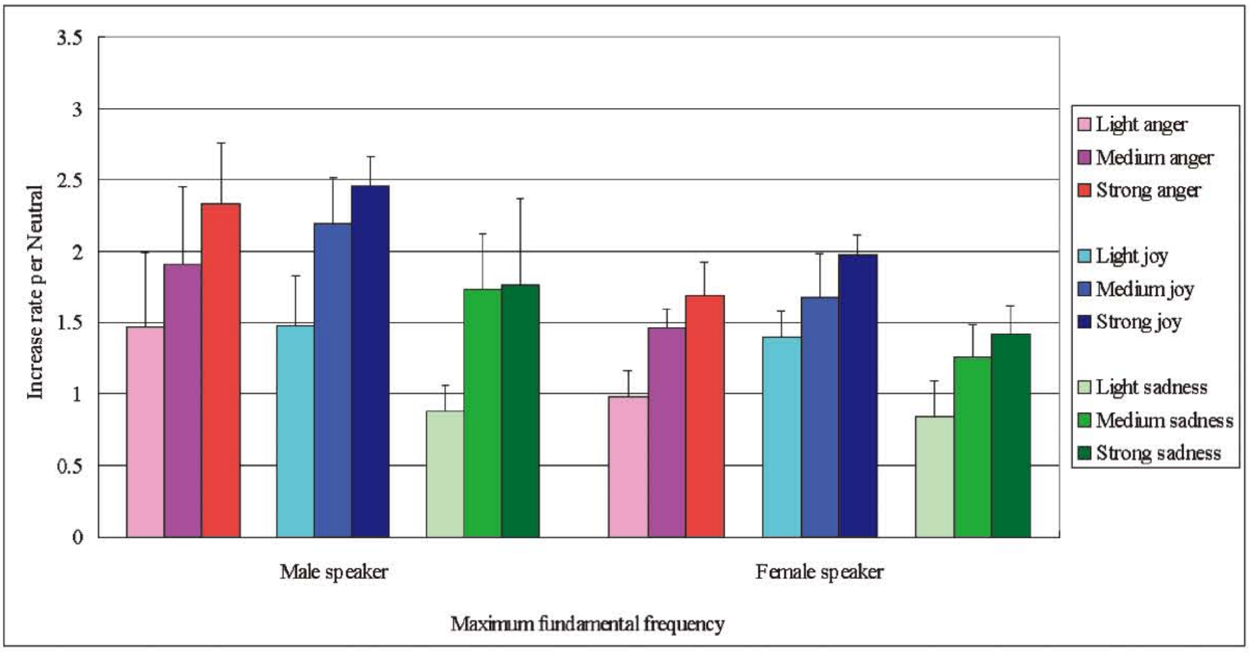

Figure 6: Increase rate per Neutral of Maximum fundamental frequency ($F_{0\max}$).

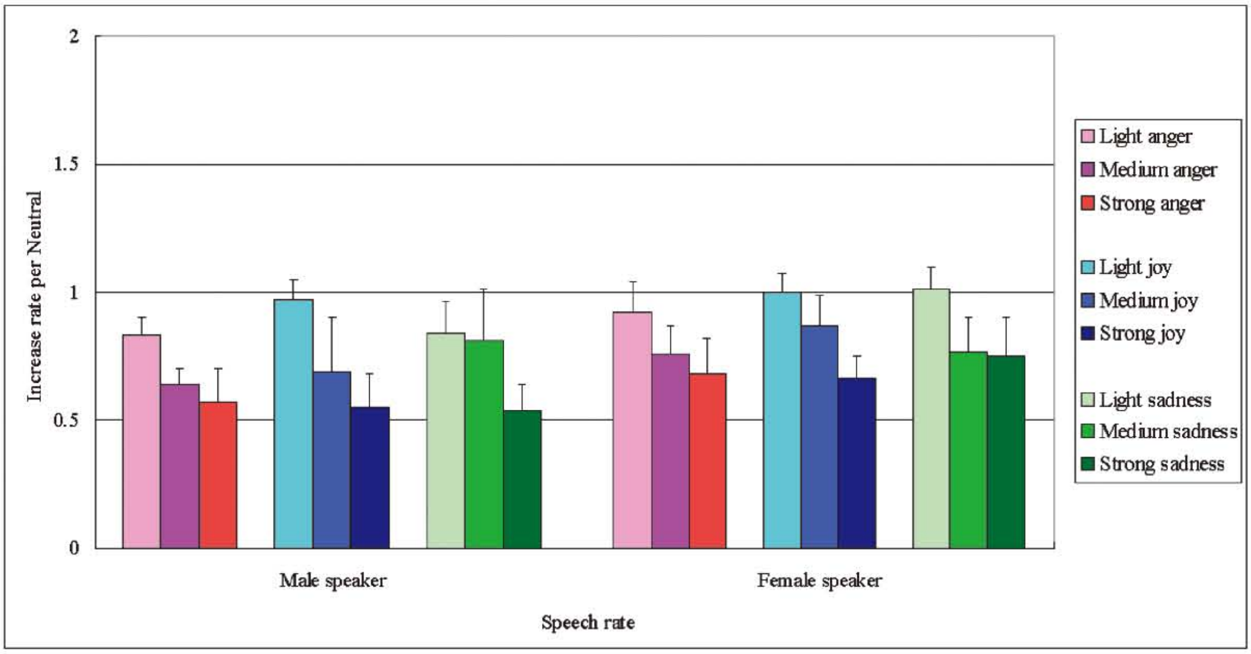

Figure 7: Identification rate of anger emotion (female).

Next, we examined the effects of degrees of emotions on the prosodic parameters. Figures 8 and 9 show bubble charts of three types of prosodic-feature parameters (magnitude of accent command, minimum fundamental frequency $F_{0\min}$, and maximum fundamental frequency $F_{0\max}$) of speech uttered by male and female speakers, respectively. These bubble charts consist of an abscissa, an ordinate, and a size of a bubble on the coordinates. The abscissa indicates Ir/N of minimum fundamental frequency, the ordinate indicates Ir/N of maximum fundamental frequency, and the size of a bubble indicates Ir/N of magnitude of accent command for each type and degree of emotion.

As seen in Fig. 8, in the case of “anger” speech uttered by male speakers, Ir/Ns of $F_{0\min}$ were almost unchanged and Ir/Ns of $F_{0\max}$ increased with increase of the degree of emotion. In the case of “joy” speech, Ir/Ns of $F_{0\min}$ and $F_{0\max}$ increased with increase of the degree. In the case of “light sadness”, Ir/Ns of $F_{0\min}$ and $F_{0\max}$ were almost the same as those for "neutral". In the case of "medium" and "strong sadness", on the other hand, Ir/Ns of F_{0min} and F_{0max} increased with increase of the degree of emotion.

As seen in fig. 9, in the case of "anger" speech uttered by female speakers, Ir/Ns of both F_{0min} and F_{0max} increased with increase of the degree of emotion. In the case of "joy" speech, Ir/N of F_{0max} increased more rapidly than that of F_{0min} with increase of the degree. Ir/N of magnitude of accent command increased with increase of the degree. In the case of "sadness" speech, on the other hand, Ir/N of F_{0min} increased more rapidly than that of F_{0max} with increase of the degree.

Speakers' gender-specific features were observed in the ranges of F_{0min} and F_{0max} variations in terms of Ir/N according to the degrees of emotions: i.e., the ranges generally tended to expand more widely for male speech than those for female speech. This tendency was particularly conspicuous for "joy". This fact suggested that emotions of female speakers were easier to be perceived with less voice-pitch-height changes than emotions of male speakers.

Common to both genders, there were inconspicuous differences among "light joy", "medium sadness", and "strong sadness". This suggested that there was a limitation in emotional expressions only using prosodic features. We have also been exploring other features [20]-[22].

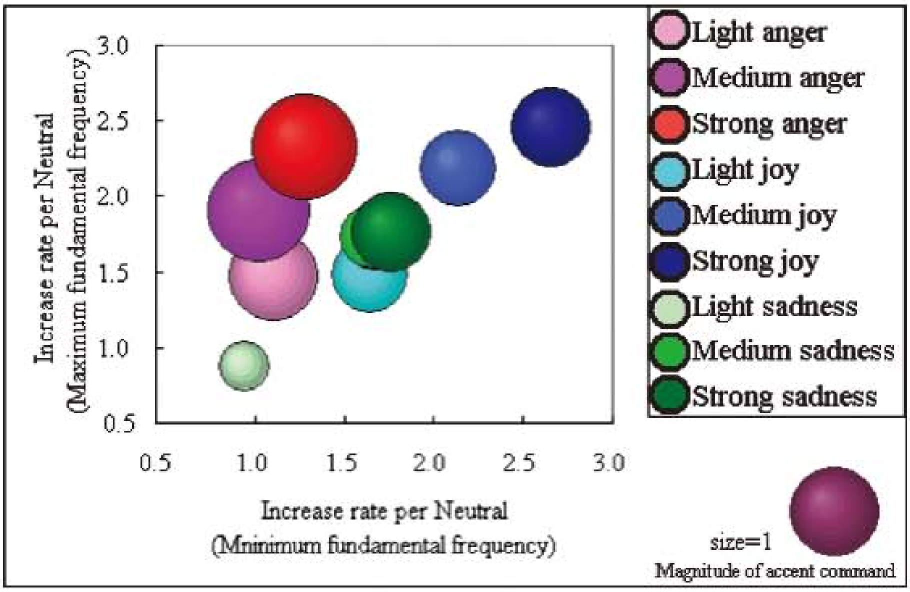

Figure 8: Bubble chart of Ir/N of prosodic features (male)

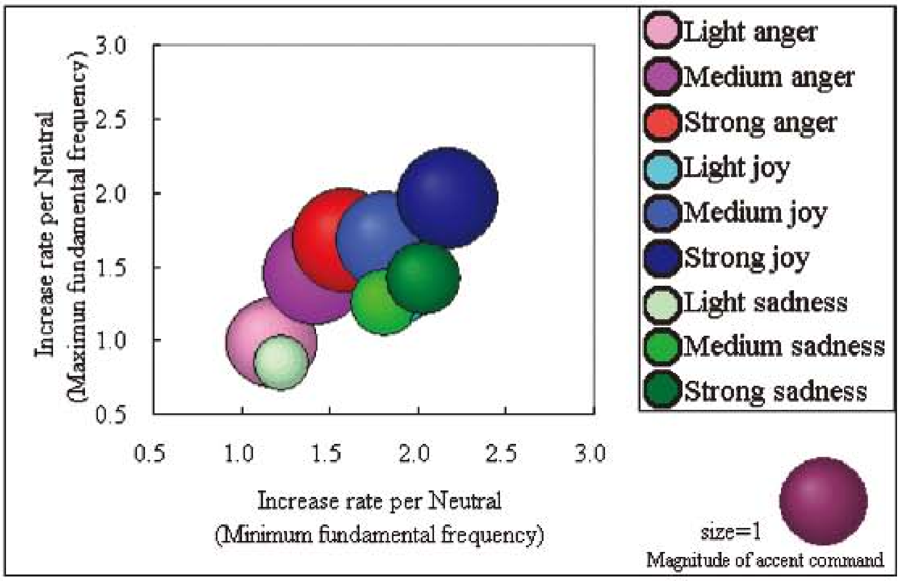

Figure 9: Bubble chart of Ir/N of prosodic features (female)

3.3 Difference of prosodic-features parameters between speakers' genders

In this subsection, we describe the difference of prosodic-features parameters between the genders of speakers for each type and degree of emotion. We compared between prosodic features of male speakers and those of female speakers statistically. As for $F_{0\mathrm{min}}$ and $F_{0\mathrm{max}}$ as height of voice pitch, since they have difference between male and female speech, these F_0 parameters must be normalized to eliminate this difference. We therefore adopted the ratio of these F_0 parameters for emotional speech to those for neutral speech for statistical analysis.

As shown in Fig. 10, in the magnitude of accent command of "anger" speech, there was significant difference between speakers' genders ($p < 0.05$). We understood that the listeners perceived as "anger" for male speech with larger magnitude of accent command than for female speech. This showed that male speakers could not communicate their anger emotion with the same intensity of accent as that for female speakers.

As shown in Fig. 11, there was significant difference between $F_{0\mathrm{min}}$ for anger speech uttered by male speakers and that uttered by female speakers ($p < 0.01$). Figure 12 shows that, in the case of $F_{0\mathrm{max}}$, there was significant difference between both genders for all emotions. According to Fig. 6, male speakers needed to increase $F_{0\mathrm{max}}$ for listeners to perceive their emotions, which was higher than that of speech uttered by female speakers.

Through this listening test, we learned that the listeners were able to perceive emotions of female speakers with a less increase in a voice height and accent level than those of male speakers.

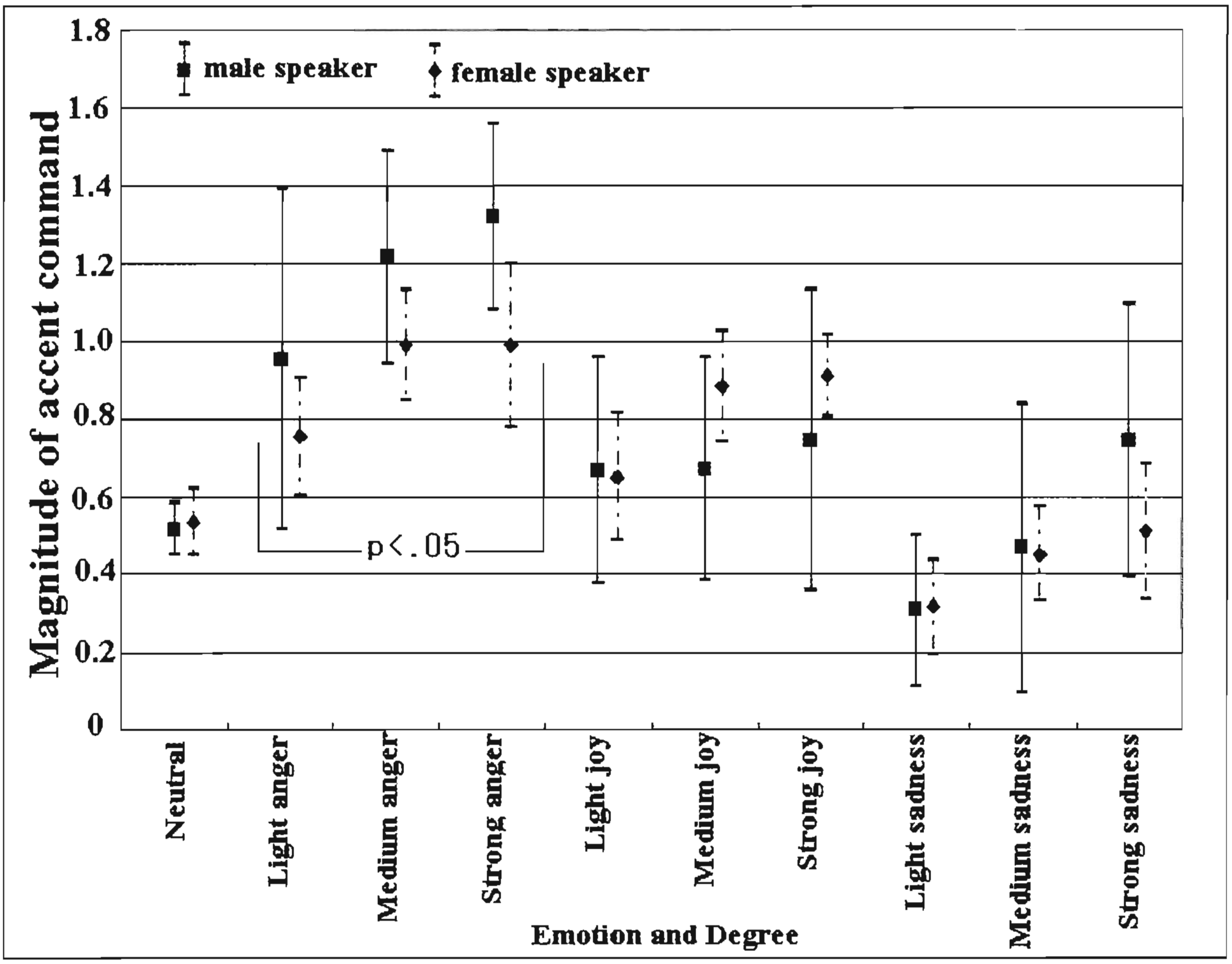

Figure 10: Comparison of magnitude of accent command. The plot in the center of each error bar denotes the mean value and the length of each error bar denotes the standard deviation.

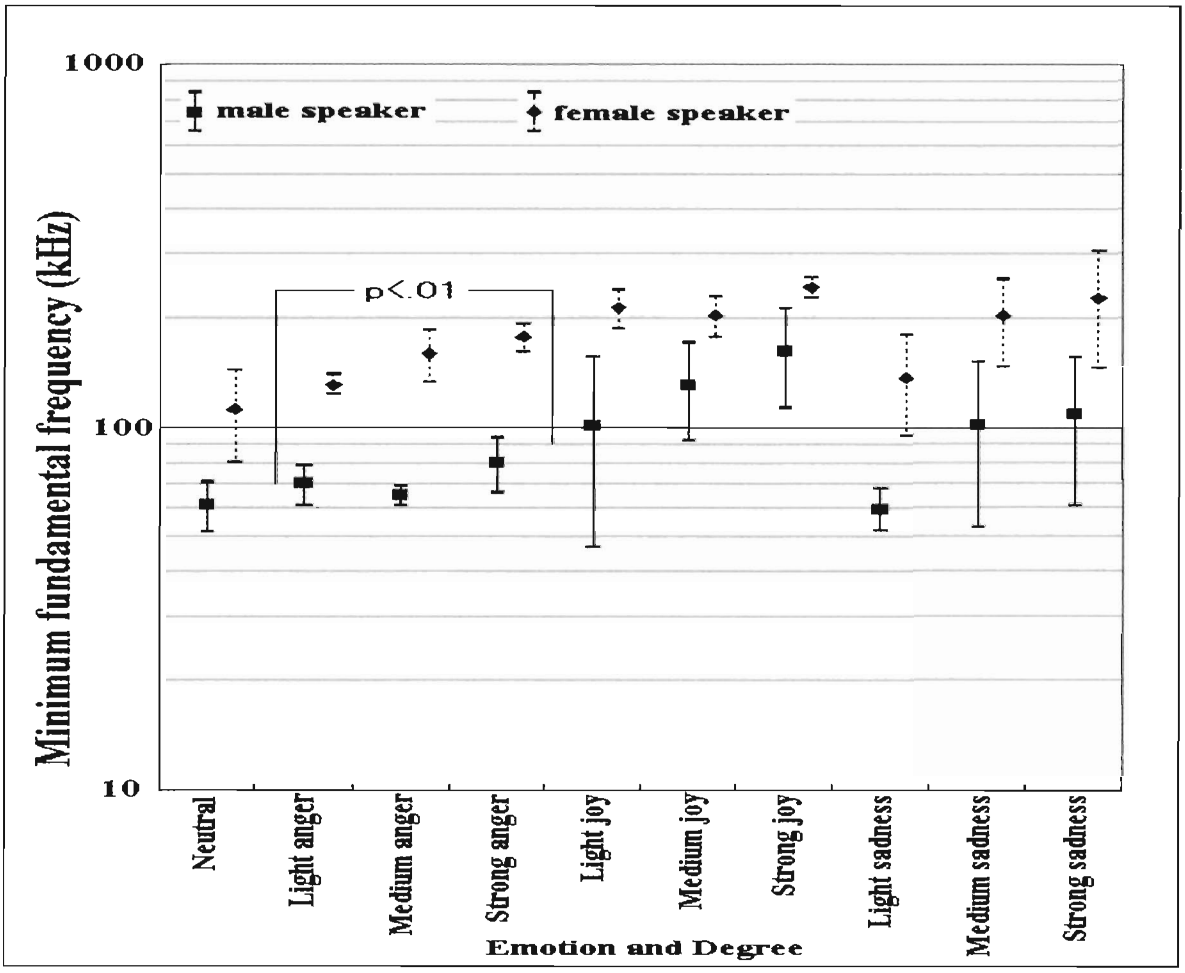

Figure 11: Comparison of minimum fundamental frequency.

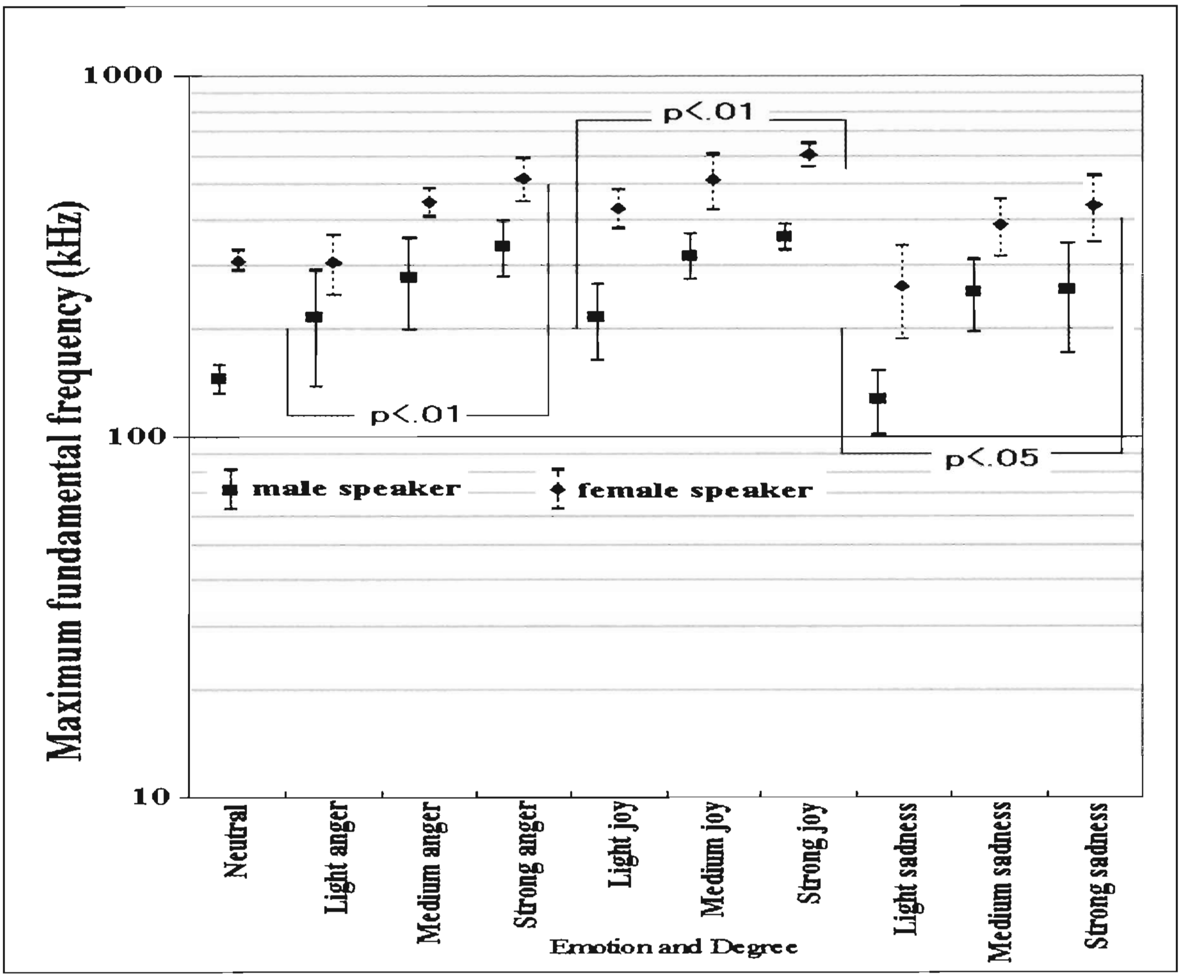

Figure 12: Comparison of maximum fundamental frequency.

4. Conclusions

In this study, we have placed a focus on subjects' auditory impressions on various types and degrees of emotional speech uttered by radio actors and actresses to determine optimal prosodic parameters for speech synthesis.

Listening test results have shown that the subjects do not necessarily perceive emotional speech as the speakers intend to express. From these results, we have analyzed the features of prosodic parameters based on the emotional speech classified according to the auditory impressions of the subjects. Prior to analysis, we have calculated an identification rate for each type and degree of emotion, which is a rate of the number of identifying as a specific type and degree of emotion to the total number of listeners. We have selected 5 speech samples whose identification rates rank the top 5 for each type and degree of emotion.

Analysis results are summarized as follows: (1) The magnitude of accent command, minimum fundamental frequency, and maximum fundamental frequency increase with increase of degree of emotion. Contrarily, speech rate decrease with increase of degree of emotion. (2) The magnitude of accent command is gender-dependent, i.e., that for anger speech uttered by the male speakers increases compared to that for neutral speech, and significant difference has been observed from female speech. (3) Minimum fundamental frequency for anger speech uttered by the female speakers increases, and significant difference has been observed from male speech. (4) Maximum fundamental frequency for all emotion of male speech increases, and also significant difference has been observed from female speech. And (5) Prosodic features that characterize their emotions are speaker's gender-dependent.

Because this listening test has been conducted for female subjects, a listening test for male subjects must be conducted for future work to explore gender-dependent or gender-independent features.

5. Acknowledgements

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

6. References

[1] Takeda, S., Sato, M., and Yagishita, E., (1997) Analysis of prosodic features of "Anger" expressions in drama conversations, Proc. Spring Meet. Acoust. Soc. Jpn., 1-7-2, pp.203-204. (in Japanese)

[2] Takeda, S., Ishizuka, F., and Hiramatsu, M., (2000) Power features of "Anger" expressions in pseudo-conversational speech, Proc. Autumn Meet. Acoust. Soc. Jpn., 2-1-9, pp.191-192. (in Japanese)

[3] Takeda, S., (2001) Analysis of features of "Anger" expressions in Japanese speech, Report of Scientific Research on Priority Areas (B) Realization of advanced spoken language information processing from prosodic features, pp.35-41. (in Japanese)

[4] Takeda, S., Nishizawa, Y., and Ohyama, G, (2001) Some considerations of prosodic features of "Anger" expressions, Tech. Rep. IEICE, SP2000-164, pp.33-40. (in Japanese)

[5] Takeda, S., Ohyama, G., and 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), pp.423-428.

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

[7] Tochitani, A., Takeda, S., Koshiba, Y., Munakata, M., Ohyama, G., and Kato, S., (2002) Differences in prosodic features of "joy" and "Sorrow" expressions in spoken Japanese depending on the degree of emotion, Proc. Spring Meet. Acoust. Soc. Jpn., 1-P-6, pp.363-364. (in Japanese)

[8] M. Dzulkhiflee Hamzah, Takeda, S., and Tochitani, A., (2002) Difference in prosodic features of "Gratitude" expressions in spoken Japanese depending on the degree of emotion, Proc. Autumn Meet. Acoust. Soc. Jpn., 1-10-19, pp.267-268. (in Japanese)

[9] Takeda, S., Tochitani, A., Doshita, M., M. Dzulkhiflee Hamzah, Aoyama, S., and Ohyama, G., (2002) Comparison of prosodic features of emotional expressions in spoken Japanese depending on the degree of emotion, Proc. Autumn Meet. Acoust. Soc. Jpn., 1-10-20, pp.269-270. (in Japanese)

[10] Takeda, S. and M. 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, pp.441-442. (in Japanese)

[11] Doushita, M. and Takeda, S., (2003) Differences in prosodic features of "Sorrow" expressions in spoken Japanese uttered by a radio actor depending on the degree of emotion, Proc. Spring Meet. Acoust. Soc. Jpn., 2-Q-35, pp.443-444. (in Japanese)

[12] Takahashi, T., Takeda, S., and M. Dzulkhiflee Hamzah, (2003) Difference in prosodic features of "Sorrow" expressions in spoken Japanese depending on the degree of emotion, Proc. Spring Meet. Acoust. Soc. Jpn., 2-Q-9, pp.391-392. (in Japanese)

[13] M. Dzulkhiflee Hamzah, Takeda, S., and 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. Autumn Meet. Acoust. Soc. Jpn., 2-Q-28, pp.367-368. (in Japanese)

[14] Hashizawa, Y., Takeda, S., M. Dzulkhiflee Hamzah, and Ohyama, G., (2004) On the Differences in Prosodic Features of Emotional Expressions in Japanese Speech according to the Degree of the Emotion, Proceedings of the 2nd International Conference on Speech Prosody, Nara (Japan), pp.655-658.

[15] M. Dzulkhiflee Hamzah, Takeda, S., Muraoka, T. and Ohashi, T., (2004) Analysis of Prosodic Features of Emotional Expressions in Noh Farce ("Kyohgen") Speech according to the Degree of Emotion, Proceedings of the 2nd International Conference on Speech Prosody, Nara (Japan), pp.651-654.

[16] Hashizawa, Y., Takeda S., M. Dzulkhiflee Hamzah, and Ohyama, G., (2005) Comparison of prosodic features of emotional speech uttered by announcers with those uttered by radio actors/actresses according to the degree of emotion, Proc. Spring Meet. Acoust. Soc. Jpn., 2-1-2, pp.207-208. (in Japanese)

[17] Sotoda, M., Kiryu, S., Takeda, S., and Muraoka, T., (2006) Prosodic features of Kyogen speech with "anger", "joy", and "sadness" compared according to the degree of emotion -the case of the increased number of speech samples- , Proc. Autumn Meet. Acoust. Soc. Jpn., 1-6-11, pp.175-176. (in Japanese)

[18] Yamada, M., Ohashi, T., Muraoka, T., and Takeda, S., (2004) Prosodical Analysis of "Kyogen" Speech, Proc. Spring Meet. Acoust. Soc. Jpn., 2-6-7, pp.419-420.

[19] Ishii, C. T., Ishiguro, H., and Hagita, N., (2005) Using prosodic and voice quality features for paralinguistic information extraction in dialog speech, SIG-Challenge-05, pp.71-76. (in Japanese)

[20] Takeda, S., Yasuda, Y., Isobe, R., Kiryu, S., and Tsuru, M., (2008) Analysis of voice-quality features of speech that expresses "anger", "joy", and "sadness" uttered by radio actors and actresses, Proc. Spring Meet. Acoust. Soc. Jpn., 3-Q-33, pp.447-448.

[21] Isobe, R., Kiryu, S., Takeda, S., Yasuda, Y., and Tsuru, M,, (2008) Anger speech synthesis based on glottal-flow information, Proc. Autumn Meet. Acoust. Soc. Jpn., 1-Q-12, pp.347-348.

[22] S. Takeda, Y. Yasuda, R. Isobe, S. Kiryu, and M. Tsuru, (2008) 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.

[23] Proc. ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, (2000) Belfast (Ireland).

[24] Proc. Interspeech 2008, (2008) Brisbane (Australia).

[25] Tsuru, M. and Takeda, S., (2004) A Study of Prosodic Features of Emotional Speech -A study by a listening test-, Kurume Shin-Ai Women's College bulletin 27, pp.67-73. (in Japanese)

[26] Tsuru, M. and Takeda, S., (2007) PCA-based comparison between prosodic features and auditory impressions of “anger” speech according to the degree of emotion uttered by announcers, Kurume Shin-Ai Women's College bulletin 30, pp.65-70. (in Japanese)

[27] Tsuru, M. and Takeda, S., (2005) On auditory impressions compared with prosodic features of emotional speech according to the degree of emotion uttered by announcers, Proc. Autumn Meet. Acoust. Soc. Jpn., 2-6-17, pp.295-296. (in Japanese)

[28] Tsuru, M. and Takeda, S., (2007) PCA-based comparison between prosodic features and auditory impressions of emotional speech according to the degree of emotion uttered by announcers, Proc. Spring Meet. Acoust. Soc. Jpn., 3-8-7, pp.237-238. (in Japanese)

[29] Tsuru, M. and Takeda, S., (2008) On auditory impressions compared with prosodic features of emotional speech according to the degree of emotions uttered by radio actors and actresses, Proc. Spring Meet. Acoust. Soc. Jpn., 3-Q-32, pp.445-446. (in Japanese)

[30] Tsuru, M. and Takeda, S., (2008) Difference in auditory impressions on emotional speech depending on the subjects’ gender, Proc. Autumn Meet. Acoust. Soc. Jpn., 1-4-14, pp.267-268. (in Japanese)

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

和文抄録

被験者の聴覚的印象に基づく感情音声の韻律的特徴の検討

靍真紀子[1]，武田昌一[2]，中迫　昇[2]，中川秀夫[3]

表情豊かな感情音声の合成を目的として，これまで「怒り」，「喜び」，「悲しみ」などの感情を表現した自然音声の音響的特徴を「韻律（声の大きさ，高さ，発話速度の総称）」という視点から解析してきた。筆者らの研究の特徴は，より微妙なニュアンスの特徴まで調べることを目的に，それぞれの感情を，感情を含まない「平常」のほか，「弱い」，「中程度」，「強い」の 4 段階の度合に分類して解析を行い，それぞれの特徴を体系的に明らかにしてきたことである。

会話によるコミュニケーションにおいて，話者の心の内面にある感情がそのまま発声に反映され，更にそのまま聞き手に伝わるとは限らない。

本研究の目的は，(1)このような話者の内面の感情（話者の意図）がどの程度正しく聞き手に伝わるのか，(2)どのような表現のときに話者の意図した感情が正しく聞き手に伝わりやすいのか，を定量的に明らかにすることである。

そこでまず，発話の意図がどの程度聞き手の聴覚的印象と一致しているかを調べるために声優が発声した 144 単語を対象として聴取実験を行った。今回は，19 歳と 20 歳の 50 名の女性短大生を被験者とした。聴取実験の結果，話者が発声した感情音声は，被験者には話者の意図通りに受容されていないことが多いことがわかった。

この結果より，話者の意図よりも聴覚的な印象に適合した音声を合成のモデルとして用いるのが正しい感情コミュニケーションのために最適であると考え，表出された音声の特徴に着目して，被験者の印象に基づく感情音声の分類を行い韻律パラメータの特徴を解析した。被験者がそれぞれの感情と度合に受容した割合の高い 5 音声の特徴に着目して解析した結果，以下のことが明らかになった。

（1）アクセント指令の大きさ，最低基本周波数，および最高基本周波数は，感情の度合の増大とともに増大し，逆に平均発話速度は減少している。

（2）アクセント指令の大きさには話者の性差が見られ，男性話者の「怒り」音声は平常音声に比べ増加し，その増加率は女性話者の音声と有意差が見られた。

（3）最低基本周波数では，女性話者の「怒り」音声の増加率は大きく，男性話者との間に有意差が見られた。

（4）最高基本周波数では，すべての感情の男性話者の増加率が大きく，女性話者との間に有意差が見られた。

（5）同じ感情を特定づける韻律的特徴が，男性話者と女性話者の感情音声の間に違いがあることが分かった。

[1] 近畿大学大学院生物理工学研究科　電子システム情報工学専攻，〒649-6493 和歌山県紀の川市西三谷 930／久留米信愛女学院短期大学ビジネスキャリア学科，〒839-8508 福岡県久留米市御井町 2278-1

[2] 近畿大学生物理工学部　電子システム情報工学科，〒649-6493 和歌山県紀の川市西三谷 930

[3] 近畿大学生物理工学部　知能システム工学科，〒649-6493 和歌山県紀の川市西三谷 930