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

# ON LISTENERS' AND SPEAKERS' GENDER-DEPENDENT FEATURES OF AUDITORY IMPRESSIONS OF EMOTIONAL SPEECH IN VARIOUS DEGREES

**Makiko TSURU[a] and Shoichi TAKEDA[*b]**

[a] *Graduate School of Biology-Oriented Science and Technology, Kinki University, and Department of Business Career, Kurume Shin-Ai Women's College, Japan*
[b] *Faculty of Biology-Oriented Science and Technology, Kinki University, Japan*

## ABSTRACT

This paper compares the prosodic features of various types and degrees of emotional expressions in Japanese speech based on the auditory impressions between the two genders of speakers as well as listeners. The speech samples consist 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 are speech rate and F0 parameters. A listening test is conducted using 144-word speech samples uttered by two radio actors and two radio actresses. We use 25 male and 50 female subjects at the ages of 19-21 years old. We then analyze the features of prosodic parameters based on the emotional speech classified according to the auditory impressions of the subjects. Analysis results suggest that prosodic features that identify their emotions and degrees are not only speakers' gender-dependent, but also listeners' gender-dependent.

***Keywords:*** *emotional expression, prosody, auditory impression, listening test, gender*

---

**Corresponding author**: 930 Nishi-Mitani, Kinokawa-shi, Wakayama 649-6493, JAPAN, takeda@gakushikai.jp

## 1. INTRODUCTION

As information communication technology (ICT) advances, there are increasing needs for better human-machine communication tools. 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. The emotion types were "anger", "joy", "sadness", and "gratitude". Speakers were radio actors and actresses, announcers, a Noh-and-Kyogen stage actor, researchers, students, and so on [1]-[3]. We have been taking a minute approach instead of generally investigating various types of typical emotional expression [4]. The degree of emotion was categorized into four categories: "neutral", "light", "medium", and "strong", and the prosodic features of each category have been analyzed [1]-[3].

The quality of speech synthesized based on these 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. We have also been analyzing features of voice quality [5]. 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 [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 Interspeech 2008 [7], 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 our study has been 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 was one of the most significant prosodic-feature parameters of emotional expressions, was known to vary

depending on the gender of speakers [8]. 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 at the ages 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 [8].

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 was 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.

Since then, we have conducted an additional listening test using 25 male subjects at the ages of 20 and 21 years old and have analyzed the prosodic features in the same way.

## 2. EXPERIMENTAL METHOD

### 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 Japanese words that had either of the 3 accent types: flat, mid-high, or head-high. They were the following 3 words: "yamagoya", "naminami", and "imanimo". 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". The total number of words was thus 144 as listed in Table 1.

**Table 1:** Speech samples

(Number of categories)

| Speaker | Emotion | Degree | Accent type | Total |
|---|---|---|---|---|
| (4)<br>Japanese male (2)<br>Japanese female (2) | (3)<br>Anger<br>Joy<br>Sadness | (3)<br>Light<br>Medium<br>Strong | (3)<br>Flat<br>Mid-high<br>Head-high | 108 samples |
| | (1)<br>Neutral | (0) | (3)<br>Flat<br>Mid-high<br>Head-high | 36 samples<br>(12 samples<br>×3 times) |

### 2.2. Prosodic-feature parameters

Prosodic-feature parameters were F0 parameters, i.e., magnitude of accent command (Aa), magnitude of phrase command (Ap), and minimum fundamental frequency (F0min) in Fujisaki's model [9], maximum fundamental frequency (F0max), 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. Experimental conditions

In the listening tests, speech samples were presented to the subjects in random order. There were 16 dummy samples ahead and 144 test samples. We conducted two sessions for the purpose of cancelling the order effect. 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 3 seconds except that the interval after consecutive 10 speech samples was 10 seconds. After a break of 5 minutes, we started the second session.

Seventy-five subjects used a headphone of the same maker and the same sound pressure. Among them, 25 subjects were male university students at the ages of 20 and 21 years old and 50 subjects were female college students at the ages of 19 and 20 years old, both with a normal auditory capacity.

## 3. EXPERIMENTAL RESULTS

We divided the analysis of the experimental results into two groups depending on the gender of the speakers.

### 3.1. Identification rate

To quantify the strength of listeners' auditory impressions, an "identification rate r" was introduced. Table 2 lists the top 5 speech samples in identification rate for each gender combination of speakers and listeners extracted from all types and degrees of emotional speech. The identification rates of speech samples in each of the four quadrants were computed separately.

**Table 2:** The rank of identification rate *r* for all speech

| Speaker | Rank | Male listener (Japanese) | Female listener (Japanese) |
|---|---|---|---|
| | | Degree & emotion (Id. rate *r* %) | Degree & emotion (Id. rate *r* %) |
| Male (Japanese) | 1 | Strong anger (98.0) | Strong anger (94.0) |
| | 2 | Strong anger (96.0) | Strong anger (92.0) |
| | 3 | Strong anger (84.0) | Neutral (84.0) |
| | 4 | Neutral (84.0) | Neutral (80.0) |
| | 5 | Neutral (82.0) | Neutral (80.0) |
| | 5 | Neutral (82.0) | |
| Female (Japanese) | 1 | Neutral (84.0) | Strong joy (82.0) |
| | 2 | Light anger (82.0) | Neutral (79.0) |
| | 3 | Light joy (82.0) | Neutral (76.0) |
| | 4 | Neutral (80.0) | Strong joy (75.0) |
| | 5 | Neutral (80.0) | Light joy (73.0) |

In the case of speech uttered by male speakers, the emotions and degrees that had the top 5 identification rates were "neutral" and "strong anger" perceived by both male and female listeners. In the case of speech uttered by female speakers, on the other hand, neither speech samples perceived as "strong anger" were included in the emotions and degrees that had the top 5 identification rates. The emotions and degrees that had the top 5 identification rates were "neutral", "light anger", and "light joy" for male listeners, and "neutral", "strong joy", and "light joy" for female listeners.

These results suggest that perception of even the same speech sample depends on the gender of listeners.

### 3.2. Listeners' gender-dependent features

We thought that there were specific kinds of prosodic features in the emotional speech which many listeners perceived as the same impression in common. Therefore, we extracted emotional speech of top 5 in identification rate for each type and degree of emotion.

To examine listeners' gender-dependent or gender-independent features, we analyzed several prosodic-feature parameters of speech depending on the types and degrees of emotions. Figures 1-7 show the analysis results, comparing between male and female listeners' impressions that are independent of the speakers' intentions (denoted as L-Male or L-Female). Prosodic feature parameters for emotions that the speakers intended to express are also shown in these figures (denoted as S-Male or S-Female). In the figures, the lengths of bars of S-Mail and S-Female denote the averages of prosodic-feature parameters of speaker-intended emotional speech over 6 samples uttered by male and female speakers, respectively. And the lengths of bars of L-Male and L-Female denote the averages of prosodic-feature parameters of listener-identified emotional speech over 5 samples of highest rank identification rates for each emotion and degree by male and female listeners, respectively. The lengths of error bars denote standard deviations.

3.2.1. $F_0$ features

As shown in Fig. 1, the magnitude of accent command Aa of female speech increased with the increase of the degree of "anger". Aa of male speech has the same tendency (a figure is omitted). Minimum fundamental frequency F0min and maximum fundamental frequency F0max have the same tendency for "anger" speech (figures are omitted). In the case of speech uttered by female speakers, male listeners tended to perceive speech samples as "strong anger" with lower magnitude of accent command Aa than female listeners did. Aa of male speech has the same tendency (a figure is omitted). In the case of "joy" speech uttered by male speakers, as shown in Fig. 2, male listeners were observed to perceive the speech as "strong joy" with approximately half magnitude of accent command Aa compared with female listeners. In the case of "sadness" speech uttered by female speakers, as shown in Fig. 3, male listeners were observed to perceive the speech as "strong sadness" with lower magnitude of accent command Aa compared with female listeners. In the case of "sadness" uttered by male speech, as shown in Fig. 4, male listeners were observed to perceive the speech as "strong sadness" with lower minimum fundamental frequency F0min compared with female listeners.

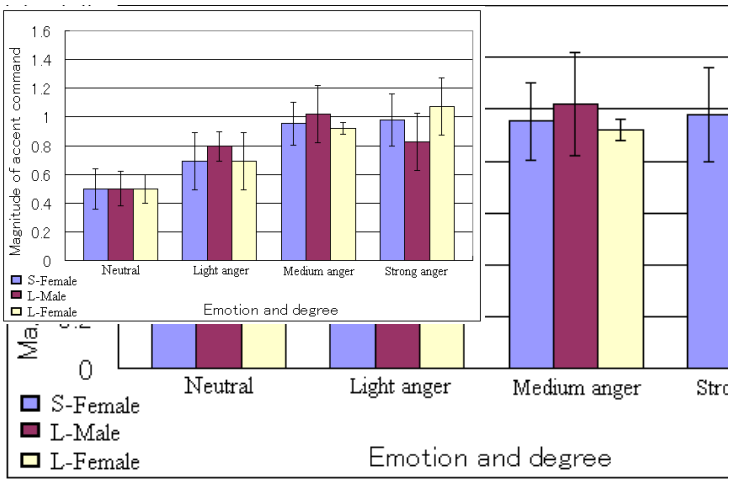


**Figure 1:** Magnitude of accent command $A_a$ for "anger" speech uttered by female speakers

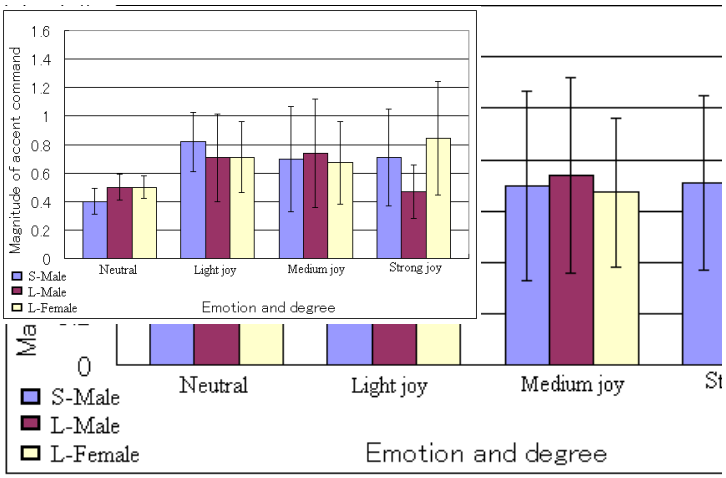


**Figure 2:** Magnitude of accent command $A_a$ for "joy" speech uttered by male speakers

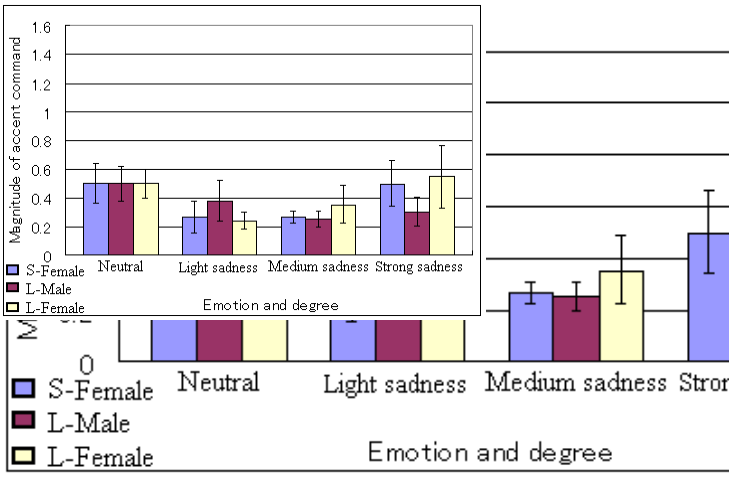


**Figure 3:** Magnitude of accent command $A_a$ for "sadness" speech uttered by female speakers

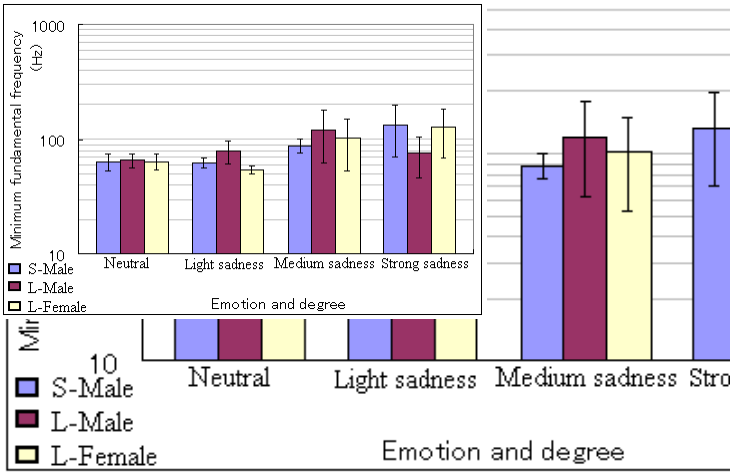


**Figure 4:** Minimum fundamental frequency $F_{0min}$ for "sadness" speech uttered by male speakers

### 3.2.2. Speech-rate features

Figure 5 shows that male listeners tend to perceive the speech as "light anger" with lower speech rate than female listeners do. To the contrary, the male listeners tend to perceive the speech as "strong anger" with higher speech rate than the females listeners do. Next, Figure 6 shows that male listeners tend to perceive the speech as "light joy" with lower speech rate than female listeners do. To the contrary, the male listeners tend to perceive the speech as "strong joy" with higher speech rate than the females listeners do. Figure 7 shows that male listeners tend to perceive the speech as "light sadness" with lower speech rate than female listeners do.

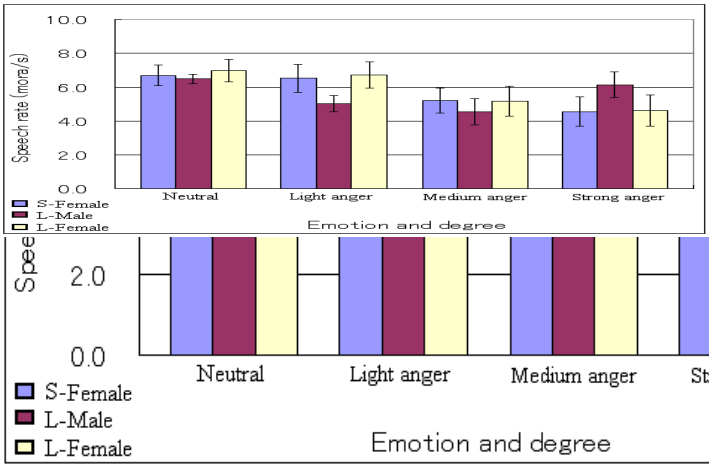


**Figure 5:** Speech rate for "anger" speech uttered by female speakers

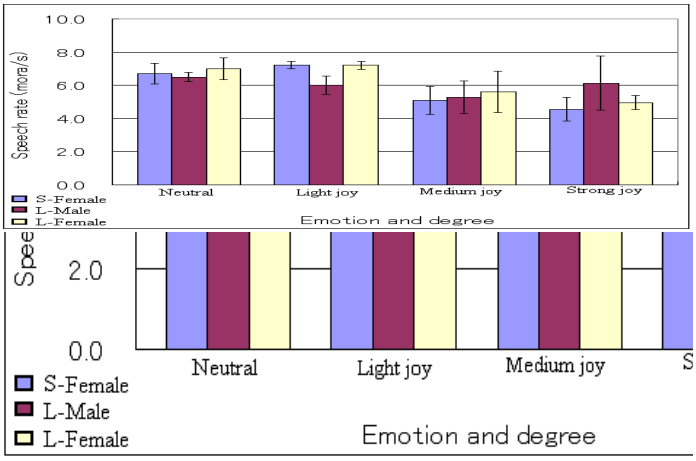


**Figure 6:** Speech rate for "joy" speech uttered by female speakers

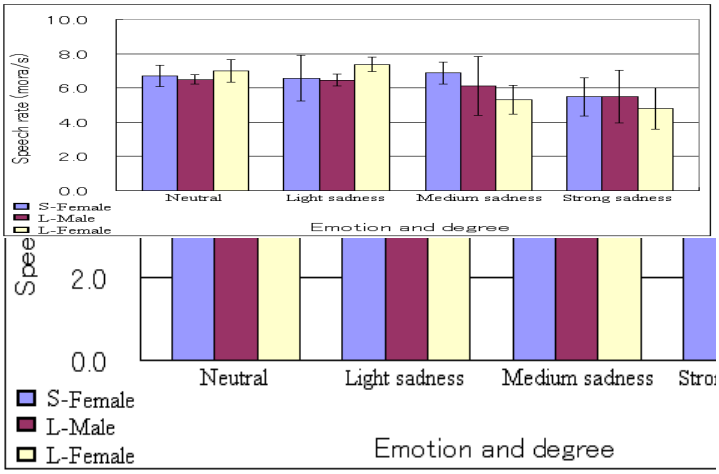


**Figure 7:** Speech rate for "sadness" speech uttered by female speakers

All these results described in subsubsections 3.2.1 and 3.2.2 suggest that male listeners perceive different impressions from those female listeners do. Additional listening tests, however, are needed using sufficient number of speech samples to generalize these results, and we will leave them as a future work.

#### 3.2.3. Other prosodic features

No conspicuous listeners' gender-dependent features were observed in other prosodic feature parameters. In other words, these prosodic parameters were found to be listeners' gender-independent.

## 4. CONCLUSIONS

Listening tests have been conducted to investigate whether or not there are listeners' and speakers' gender-dependent differences in the prosodic features of speech samples that have the same auditory impression on the type and degree.

The test results have suggested that there are listeners' as well as speakers' gender-dependent differences in the prosodic features to identify the type and degree of emotion.

We will investigate gender-dependent features of emotional speech uttered by more number of speakers to generalize the knowledge obtained through this study, and clarify effective prosodic-feature parameters for emotional speech syntheses based on the results of this study in future.

### 4.1. Acknowledgments

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

[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, *Proc. ICSP2001*, 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, *Proc. 2nd Int. Conf. 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, *Proc. 2nd Int. Conf. Speech Prosody*, Nara, Japan, pp. 651–654, 2004.

[4] 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.

[5] 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.

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

[7] *Proc. Interspeech 2008*, Brisbane, Australia, 2008.

[8] 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.

[9] Fujisaki, H. and Hirose, K., Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *Journal of the Acoustical Society of Japan (E)*, 5(4), pp.233-242, 1984.