Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners

Chun-Fang Huang[a], Donna Erickson[b] and Masato Akagi[a]

[a]Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, 923-1292 Sendai, Japan
[b]Showa University of Music, 808 Sekiguchi, Atsugi, 243-8521 Kanagawa, Japan
akagi@jaist.ac.jp

Language is an important tool in speech communication. Even without the understanding of one language, we can still judge the expressive content of a voice, such as happiness or sadness. It is clear that a voice does not contain only linguistic, but also non-linguistic information. However, sometimes misunderstanding of emotional communication occurs. It is not clear what the common or different characteristics are in non-linguistic information that help or hinder people with different native languages/ culture to make judgments about the expressivity of speech. In order to explore this question, we construct and compare the perceptual model of Japanese expressive speech utterances built for Mandarin and Japanese listeners' perception. This perceptual model uses a concept, called semantic primitives, which are adjectives used by listeners for describing the expressive speech utterances, to connect the acoustic characteristics in voices with the judgments made by humans of expressive speech categories. The comparison of the resulting connections showed that 60% of the adjectives are used commonly by both Mandarin and Japanese listeners. This commonality suggests how expressive speech communication is improved by non-linguistic information.

1 Introduction

Communication is one of the most important activities of a human being. With regard to communication using speech, receivers not only decode linguistic information, but also non-linguistic information. Even without the acquaintance of the language, we can still judge a speaker's expressive speech categories of a voice, such as joy or sadness. However, sometimes misunderstanding of emotional communication occurs. Our question is what are the characteristics in non-linguistic information that help or hinder people with different languages/cultures to make this judgment.

To explore the answer to this question, we built perceptual models of Japanese expressive speech utterances for Mandarin and Japanese listeners. Different from previous research work [1], a special concept, called semantic primitives, are used in the model. These are adjectives that people use to represent the mental feelings toward the expressive contents of speech utterances they hear, e.g., "bright-sounding" or "fast-sounding". By examining the semantic primitives, we can discover how the expressive contents of a voice are perceived by different groups of people, depending on the changes in the acoustic signal. Specifically, we can determine the commonalities and differences in expressive speech perception which help or hinder people with different languages/cultures communicate.

We suggest that semantic primitives are appropriate to use from the following standpoints:

1. Adjectives are a basic tool that we learn from childhood to describe the feeling we have toward different kinds of things, such as food we eat or music we hear.
2. The use of semantic primitives shows how a speech utterance with a particular set of acoustic characteristics can be mentally interpreted by different people. For example, an utterance may be perceived differently by two people as either slow or low, which is then judged as expressing sadness or neutralness, respectively.
3. Different groups of people with different native languages/culture may make similar or dissimilar perceptions and judgments about the expressive contents of a single speech utterance.

The fact that different people use different semantic primitives suggests the following facts: (1) the perception of the expressive contents [1] of a speech utterance is implemented by the changes of acoustic characteristics of the voice, e.g., fundamental frequency, duration, intensity, etc. and (2) the judgment of an expressive speech category is made by the perception of its expressive contents. These two facts are traditionally viewed as two distinct phenomena by previous research. However, we consider these two facts taken together as the key to understanding why the same speech utterance may be judged differently by different people.

These two facts are used to create a three-layered model [4] (see Figure 1), which combines three important elements. In this study, we refer to acoustic characteristics of the voice, e.g., fundamental frequency, duration, intensity, etc., as "acoustic features". These are represented by the bottom-most layer in the model. The middle layer represents the adjectives used by listeners for expressive voices perception. The topmost layer represents the judgment of expressive speech categories made by listeners.

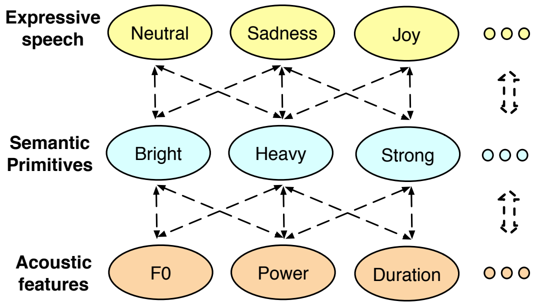


Figure. 1. Conceptual diagram of the perceptual model.

By building a three-layered model for one group of people, we hope to understand the nature of these two facts with regard to expressive speech perception for that group. By comparing models built for two different groups of people, we hope to understand the commonalities/differences of expressive speech perception.

We thus constructed a three-layer model for two groups of subjects: 10 Mandarin listeners who did not understand Japanese, and 12 Japanese listeners. Both sets of listeners were asked to evaluate the Japanese voice database. The objectives of this study are the following:

1. To find out what semantic primitives are used by Mandarin and Japanese listeners in the perception of

[1] Expressive contents is a dimension of meaning that is separated from the lexical meaning of an utterance

expressive speech.

2. To build a relationship between semantic primitives and perception of expressive speech categories. The comparison of the resulting relationships will show the commonalities and differences of expressive speech categories judgment between Mandarin and Japanese listeners.
3. To analyze and compare the relationship between the changes of acoustic characteristics and the usages of semantic primitives. It is hoped this comparison will show the commonalities and differences of expressive speech perception between Mandarin and Japanese listeners, as affected by the changes in acoustic characteristics.

The remainder of this paper is as follows. Section 2 describes the process of the selection of semantic primitives. Section 3 describes the building of the relationship between expressive speech categories and semantic primitives by the construction of a fuzzy inference system. Section 4 describes the analysis of the relationship between semantic primitives and acoustic features. A general discussion of the overall results is given in Section 5. Section 6 summarizes this paper.

2 Finding semantic primitives

To find out what semantic primitives are used by Mandarin and Japanese listeners in the perception of expressive speech, three experiments are conducted using a speech database with Japanese utterances.

2.1 Experiment 1

Experiment 1 examined listeners' perception of expressive speech utterances.

2.1.1 Method

Stimuli were selected from the database produced and recorded by Fujitsu Laboratory. A professional actress was asked to produce utterances using five expressive speech categories, i.e., *Neutral*, *Joy*, *Cold Anger*, *Sadness*, and *Hot Anger*. In the database, there are 19 different Japanese sentences. Each sentence has one utterance in *Neutral* and two utterances in each of the other categories.

The first group of subjects was 20 Mandarin listeners, 10 males and 10 females, who did not understand Japanese. The second group of subjects was 12 Japanese. They rated the utterances according to how strongly they perceived each of the five expressive speech categories, on a scale of 1 to 5.

2.1.2 Results and discussion

Table 1 lists the results of Exp 1 as a confusion matrix of each intended expressive speech category, which gives the evaluations by Mandarin (a) and Japanese listeners (b), respectively. From the diagonal lines shown in Tables 1, we know that for each category, the highest percentage belongs to the intended category. When comparing Tables 1 (a) and (b), both Mandarin and Japanese listeners have similar patterns of confused categories.

2.2 Experiment 2

Experiment 2 was conducted to construct a perceptual space of utterances in the different expressive speech categories, and the results were analyzed using MDS. The resulting perceptual space was then used in Experiment 3 to select suitable semantic primitives for the perceptual model.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	72%	12%	20%	10%	9%
Joy	10%	83%	1%	0%	1%
Cold Anger	11%	1%	72%	14%c	7%
Sadness	6%	3%	6%	76%	2%
Hot Anger	1%	1%	1%	0%	81%

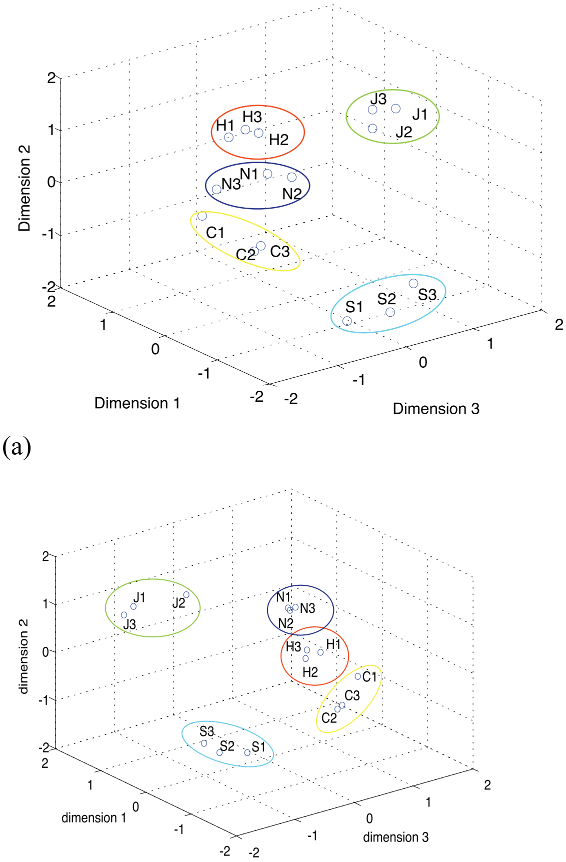
(a)

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	98%	12%	10%	5%	1%
Joy	0%	87%	0%	0%	0%
Cold Anger	2%	1%	86%	3%	2%
Sadness	0%	0%	4%	92%	0%
Hot Anger	0%	0%	0%	0%	97%

(b)

Table 1. Percentage of ratings of the 5 intended categories. Subjects were Mandarin for (a) and Japanese for (b). The columns are the intended categories and the rows, the categories rated by subjects.

2.2.1 Method



(a)

(b)

Figure 2. The resulting perceptual space of utterances in different categories of expressive speech of Experiment 2. Subjects are Mandarin listeners for (a) and Japanese for (b).

Stimuli were 15 utterances chosen according to the ratings in Exp 1. For each of the five expressive speech categories, three utterances were selected: (1) one that was most confused, (2) one that was least confused, and (3) one that fell in the middle. The subjects were identical to those who participated in Exp 1. Scheffe's method of paired comparison was used. Subjects were asked to rate each of the utterance pairs on a 5-point Liker-type scale (from -2 to 2, including 0, -2 = totally different, 2 = extremely similar) according to how similar they perceived them to be. The pair-wise stimuli were randomly presented to each subject through binaural headphones.

2.2.2 Results and discussion

Figure 2 shows the distribution of utterances in the resulting 3-dimensional perceptual space for Mandarin (a) and Japanese (b) listeners, respectively (STRESS value was 7%). In the figure, one circle represents one utterance and plot symbols like ‘J’ indicate utterances of Joy, etc. The number after each symbol corresponds to the selected utterances (1), (2), and (3) explained in Section 2.2.1 above. As the distribution shows, all categories of expressive speech are separated clearly and the utterances of the same category are close to each other, which means the distribution in the perceptual space can appropriately represent the similarity of the utterances and the position of each emotion. Therefore, this method is reliable for determining the semantic primitives suitable for Exp 3.

2.3 Experiment 3

Experiment 3 was conducted to determine suitable semantic primitives for the perceptual model.

2.3.1 Pre-selection of semantic primitives

In order to determine adjectives related to expressive speech from a large number of possible adjectives applicable to sound, tone, or voice, we carried out this pre-experiment.

Sixty adjectives were selected as candidates for semantic primitives. 46 of these were from the 50 adjectives of Ueda’s work [5] (with English glosses). The 4 adjectives removed from his original list are those that we considered less-related to expressive speech perception because the original adjectives are for music description. An extra 14 adjective considered relative to expressive speech were added based on several informal perception tests.

In the pre-selection of semantic primitives, Japanese subjects listened to 25 utterances (five for each category of expressive speech which were randomly chosen from the expressive speech database) and were asked to circle which adjectives seemed most appropriate to describe each utterance they heard. Finally, from the counts of the 60 adjectives used in the pre-selection, 34 adjectives with their counts larger than the second quartile were selected.

2.3.2 Method

The stimuli and subjects were the same as in Exp 2. The stimuli were randomly presented to each subject. Subjects were asked to rate each of the adjectives on a 4-point scale (0: very appropriate, 3: not very appropriate) when they heard each utterance, indicating how appropriate the adjective is for describing the utterance they heard. In order to clarify which adjectives were more appropriate for describing expressive speech and how each adjective was related to each category of expressive speech, the 34 adjectives were superimposed by the application of a multiple regression analysis into the perceptual space built in Experiment 2. Equation (1) is the regression equation.

$$y = a_1 x_1 + a_2 x_2 + a_3 x_3 \qquad (1)$$

, where x_1, x_2, and x_3 are the positions (x_1, x_2, x_3) of one utterance in the 3 -dimensional perceptual space, and y is the rating of an adjective for a particular utterance. Regression coefficients a_1, a_2 and a_3 were calculated by performing a least squares fit. In addition, the multiple correlation coefficient of each adjective was computed. An example of the resulting diagrams for Japanese listeners is shown in Figure 3.

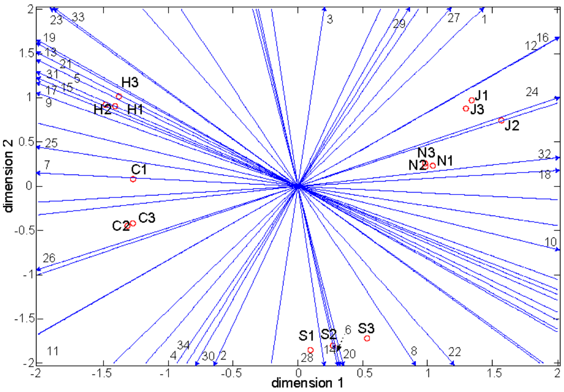


Figure. 3: Direction of adjectives in perceptual space. The figure was plotted with arrow-headed lines in dimension-1 against dimension-2 in Figure. 3.

2.3.3 Results and discussion

In this way, it was possible to find to which category each adjective was related. Semantic primitives were selected according to the following three criteria:

1. The direction of each adjective in the perceptual space: This indicates which category the adjective is most related to.
2. The angle between each pair of adjectives: The smaller the angle is, the more similar the two adjectives.
3. The multiple correlation coefficient of each adjective: When the multiple correlation coefficient of one adjective is higher, it means the adjective is more appropriate for describing expressive speech.

The selected semantic primitives by Mandarin subjects are shown in the top row of Table 3 (a) for Mandarin and (b) for .Japanese. The first ten semantic primitives that are shown in the both tables are identical. These results suggest that semantic primitives are a suitable tool for expressive speech description because those people with different native languages/cultures tend to use the same set of semantic primitives for expressive utterance description. These results are used to build a fuzzy inference system for representing the relationship between expressive speech and semantic primitives. It is explained in the next section.

3 Comparing expressive contents perception by fuzzy inference system

We suggest that the relationship between expressive speech category and semantic primitive represents the way humans use linguistic forms (e.g., words) to describe what they perceive when they hear expressive speech. That expression of the perception is vague, not precise. In this sense, traditional statistical methodology may not be appropriate for solving the problem; rather, fuzzy logic appears to be better suited.

3.1 Construction of a fuzzy inference system

To *precisely* represent the vagueness nature in the relationship between the usage of semantic primitives and the decision of expressive speech categories, the results of the three experiments were then used in training and checking data for the fuzzy inference system (FIS) that maps the relationship between expressive speech categories and semantic primitives.

We built a fuzzy inference system (FIS) for each category of expressive speech. We calculated regression lines that describe the relationship between input (perceptible degrees of semantic primitives) and output (perceptible degrees of expressive speech) of each FIS. Therefore, the slope of the regression line describes the relationship between expressive speech and semantic primitive. The absolute value of the coefficient of the regression line indicates how much the semantic primitives affect the categories of expressive speech. A positive value of slope indicates that the relationship has a positive correlation, and vice versa.

Figure 4 shows one of the results. The solid line is the FIS output and the dotted line is the regression line of the output. The figure depicts a non-linear relationship between *Joy* and the semantic primitive *weak* on the left and another non-linear relationship between *Sadness* and *weak* on the right.

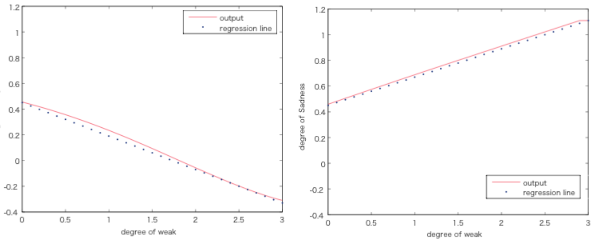


Figure. 4. Slope of regression line. Left graph describes the relationship between *Joy* and *weak*, right graph describes the relationship between *Sadness* and *weak*

3.2 Results

Table 2 lists the five semantic primitives for each expressive speech category. For each semantic primitive, there are three positive correlations (which are the ones that showed the highest correlation values with a positive slope) and two negative ones (which showed the highest correlation values with a negative slope).

Those semantic primitives that influence the perception of the expressive speech category for both Mandarin and Japanese listeners appear in Table 2 as shaded. The semantic primitives *heavy* and *clear* influence the perception of *Neutral* for both Mandarin and Japanese listeners; specifically, it is positively correlated to *clear*, and negatively correlated to *heavy*. Similarly, *Joy* is correlated with *weak* and *clear* but in opposite directions. *Cold Anger* is positively correlated to *weak*. *Sadness* is positively correlated to *heavy*. *Hot Anger* is positively correlated to *unstable*. These results suggest that for people with different native languages/cultures, the perception of expressive speech categories are affected by a common set of semantic primitives.

Neutral		Joy		Cold Anger		Sadness		Hot Anger	
PF	S	PF	S	PF	S	PF	S	PF	S
dull	-0.181	weak	-0.254	fluent	-0.498	bright	-0.122	muddy	-0.26
heavy	-0.117	low	-0.185	bright	-0.121	smooth	-0.295	heavy	-0.186
bright	0.115	clear	0.178	slow	0.164	heavy	0.179	fluent	0.118
clear	0.234	calm	0.288	weak	0.212	strong	0.181	unstable	0.17
smooth	0.256	smooth	0.44	muddy	0.384	raucous	0.267	hard	0.325

(a)

Neutral		Joy		Cold Anger		Sadness		Hot Anger	
PF	S	PF	S	PF	S	PF	S	PF	S
heavy	-0.329	quiet	-0.039	sharp	-0.079	slow	-0.231	calm	-0.063
weak	-0.181	weak	-0.036	strong	-0.049	monotonous	-0.073	quiet	-0.047
calm	0.103	clear	0.034	quiet	0.044	well-modulated	0.091	sharp	0.103
clear	0.127	unstable	0.063	weak	0.074	fast	0.153	unstable	0.12
monotonous	0.27	bright	0.101	heavy	0.061	heavy	0.197	well-modulated	0.124

(b)

Table 2. The related semantic primitives of each expressive speech category selected by Mandarins (a) and Japanese (b)

4 Comparing expressive speech judgment by acoustic feature analysis

The third objective is to analyze and compare the relationships between the changes of acoustic characteristics and the usage of semantic primitives. Firstly, F0 contour, power envelope, and power spectrum were calculated by STRAIGHT [5]. And then, many of the acoustic features involved with them were measured. We calculated the correlation between the measured acoustic features and the semantic primitives of both Mandarin and Japanese listeners. To consider only those acoustic features significant to the semantic primitive usage, 16 of them were finally chosen and are shown in Table 3.

These 16 acoustic features are the following. Four involved F0--mean value of rising slope (RS), highest pitch (HP), average pitch (AP) and rising slope of the first accentual phrase (RS1st); four involved power envelope--mean value of power range in accentual phrase (PRAP), power range (PWR), rising slope of the first accentual phrase (PRS1st), the ratio between the average power in high frequency portion (over 3 kHz) and the average power (RHT); five involved the power spectrum--first formant frequency (F1), second formant frequency (F2), third formant frequency (F3), spectral tilt (SPTL), spectral balance (SB); and three involved duration total length (TL), consonant length (CL), ratio between consonant length and vowel length (RCV).

Comparison of these two tables brings the following observations:

1. The first 10 semantic primitives that were shared by both Mandarin and Japanese listeners have the same valence (i.e., positive or negative correlation).
2. 6 of the 10 semantic primitives are associated with the same two acoustic features that have the highest correlations: *bright*, *dark*, *low*, *heavy*, and *clear* are associated with average pitch (AP) and highest pitch (HP), and *strong* is associated with power range (PWR) and mean value of power range in accentualphrase (PRAP).

Comparison with Tables 3 (a) (Mandarin listeners) and (b) (which gives the correlation coefficients by Japanese listeners) suggests that F0 plays an important role in the perception of semantic primitives. For easy identification,

	bright	dark	low	heavy	clear	strong	weak	calm	unstable	slow	hard	dull	fluent	smooth	raucous	muddy
AP	0.88	-0.84	-0.88	-0.83	0.72	0.61	-0.59	-0.45	0.47	-0.84	0.21	-0.82	0.77	0.39	0.61	-0.76
HP	0.85	-0.82	-0.87	-0.80	0.68	0.68	-0.63	-0.52	0.56	-0.86	0.29	-0.81	0.74	0.32	0.67	-0.73
RS	0.62	-0.64	-0.67	-0.54	0.46	0.69	-0.67	-0.55	0.57	-0.70	0.49	-0.63	0.48	0.09	0.59	-0.53
RS1st	0.73	-0.75	-0.80	-0.69	0.61	0.66	-0.65	-0.46	0.47	-0.79	0.37	-0.72	0.61	0.17	0.60	-0.68
PWR	0.66	-0.73	-0.78	-0.66	0.50	0.83	-0.76	-0.62	0.64	-0.80	0.58	-0.73	0.57	-0.01	0.77	-0.62
RHT	-0.01	-0.06	-0.12	0.04	-0.19	0.63	-0.35	-0.67	0.76	-0.26	0.66	-0.05	-0.12	-0.53	0.71	0.11
PRS1st	0.74	-0.80	-0.80	-0.74	0.61	0.58	-0.73	-0.34	0.35	-0.75	0.35	-0.80	0.69	0.22	0.47	-0.75
PRAP	0.58	-0.70	-0.73	-0.58	0.45	0.83	-0.80	-0.60	0.61	-0.77	0.67	-0.68	0.47	-0.10	0.70	-0.57
F1	0.63	-0.61	-0.64	-0.61	0.48	0.49	-0.44	-0.33	0.41	-0.58	0.23	-0.58	0.53	0.15	0.48	-0.46
F2	0.48	-0.32	-0.32	-0.41	0.42	-0.05	0.08	-0.02	0.02	-0.30	-0.37	-0.29	0.42	0.57	0.06	-0.29
F3	0.56	-0.42	-0.40	-0.44	0.43	0.22	-0.15	-0.20	0.24	-0.41	-0.10	-0.44	0.48	0.42	0.29	-0.32
SPTL	-0.46	0.45	0.54	0.39	-0.23	-0.67	0.39	0.67	-0.74	0.52	-0.48	0.38	-0.24	0.13	-0.76	0.29
SB	0.39	-0.39	-0.47	-0.34	0.22	0.60	-0.33	-0.63	0.68	-0.51	0.42	-0.34	0.25	-0.10	0.73	-0.26
TL	-0.34	0.50	0.47	0.44	-0.38	-0.36	0.63	0.19	-0.12	0.58	-0.26	0.50	-0.46	-0.12	-0.17	0.55
CL	-0.56	0.65	0.63	0.64	-0.52	-0.41	0.60	0.23	-0.17	0.67	-0.21	0.65	-0.59	-0.19	-0.29	0.69
CL/VL	-0.71	0.74	0.72	0.75	-0.66	-0.40	0.47	0.10	-0.12	0.61	-0.09	0.73	-0.66	-0.32	-0.30	0.71

(a)

	Bright	dark	low	heavy	clear	strong	weak	calm	unstable	slow	high	well-modulated	monotonous	noisy	quiet	sharp	fast
AP	0.71	-0.88	-0.91	-0.78	0.76	0.33	-0.54	-0.66	0.60	-0.62	0.87	0.41	-0.10	0.52	-0.70	0.34	0.35
HP	0.69	-0.88	-0.89	-0.73	0.74	0.42	-0.56	-0.72	0.67	-0.62	0.90	0.50	-0.18	0.60	-0.73	0.44	0.42
RS	0.44	-0.64	-0.60	-0.40	0.44	0.56	-0.54	-0.74	0.67	-0.56	0.70	0.54	-0.32	0.63	-0.67	0.59	0.50
RS1st	0.50	-0.79	-0.78	-0.61	0.66	0.45	-0.58	-0.67	0.60	-0.51	0.77	0.42	-0.10	0.57	-0.72	0.47	0.24
PWR	0.43	-0.74	-0.65	-0.41	0.57	0.70	-0.66	-0.80	0.78	-0.57	0.74	0.59	-0.27	0.76	-0.79	0.69	0.38
RHT	-0.10	-0.05	0.00	0.24	-0.10	0.68	-0.14	-0.55	0.67	-0.16	0.29	0.52	-0.41	0.72	-0.29	0.68	0.36
PRS1st	0.48	-0.80	-0.70	-0.56	0.64	0.45	-0.78	-0.61	0.51	-0.64	0.64	0.27	-0.01	0.44	-0.78	0.42	0.37
PRAP	0.31	-0.67	-0.56	-0.30	0.47	0.73	-0.67	-0.77	0.73	-0.55	0.62	0.55	-0.26	0.76	-0.78	0.73	0.31
F1	0.41	-0.64	-0.60	-0.49	0.47	0.25	-0.39	-0.49	0.52	-0.29	0.59	0.17	0.10	0.43	-0.52	0.29	0.30
F2	0.60	-0.41	-0.56	-0.66	0.44	-0.31	0.07	-0.11	0.07	-0.06	0.50	0.08	0.05	-0.03	-0.09	-0.27	0.11
F3	0.60	-0.47	-0.54	-0.55	0.49	0.01	-0.15	-0.33	0.33	-0.10	0.61	0.33	-0.16	0.23	-0.29	0.02	0.27
SPTL	-0.29	0.49	0.53	0.30	-0.32	-0.48	0.17	0.62	-0.71	0.24	-0.65	-0.49	0.21	-0.72	0.42	-0.53	-0.23
SB	0.27	-0.44	-0.48	-0.28	0.28	0.49	-0.16	-0.55	0.66	-0.31	0.63	0.55	-0.29	0.68	-0.39	0.51	0.20
TL	-0.26	0.42	0.30	0.21	-0.28	-0.41	0.69	0.52	-0.28	0.80	-0.25	-0.19	0.19	-0.22	0.63	-0.39	-0.59
CL	-0.36	0.64	0.53	0.47	-0.44	-0.34	0.71	0.50	-0.32	0.59	-0.39	-0.10	-0.04	-0.29	0.71	-0.31	-0.37
CL/VL	-0.41	0.78	0.71	0.66	-0.66	-0.14	0.58	0.29	-0.23	0.28	-0.47	0.02	-0.32	-0.27	0.58	-0.12	0.00

(b)

Table 3. Correlation coefficients between semantic primitives and acoustic features, Mandarin for (a) and Japanese (b)

the cells of acoustic features that are also significant to the perception of semantic primitives by both Mandarin and Japanese listeners are shadowed in Table 3.

5 General discussion

Comparison of the results from the research work with Mandarin listeners with the results of Japanese listeners suggests the following: (1) People who are from different native-languages/cultures show a certain similarity in the way they use semantic primitives. These results suggest that the two language speakers use a set of the same semantic primitives. The findings discussed in the last section also suggest the possibility of some type of universality of acoustic cues associated with semantic primitives. (2) However, there are some semantic primitives used differently between the languages: Mandarin listeners associate *Neutral* with *dull*, *bright*, and *smooth* while Japanese listeners associate it with *weak*, *calm* and *monotonous*; Mandarin listeners associate *Joy* with *low*, *calm*, and *smooth*, while Japanese listeners associate it with *quiet*, *unstable* and *bright*, etc. These differences may account for why communication of emotions between people of different languages and cultures may go relatively smoothly for the most part yet may suddenly fall apart - a phenomena sometimes experienced in cross-linguistic communication situations.

6. Summary

To explore the question of what common/different characteristics in non-linguistic information that help or hinder people with different native languages/cultures background in making judgments about the expressivity of speech, the usage of semantic primitives by Mandarin and Japanese listeners in the perception of expressive speech was investigated by three experiments. The comparison of expressive contents perception between Mandarin and Japanese listeners is conducted by the building of fuzzy inference system (FIS). The comparison of expressive speech judgment between Mandarin and Japanese listeners is conducted by acoustic feature analysis. The three-layered approach helps reveal the commonalities and differences in the perception of expressive speech utterances. The results showed that Mandarin and Japanese listeners tend to use a set of the same semantic primitives for describing expressive contents of a speech utterance. However, the results also showed that there are some different semantic primitive usages, and this may account for the misunderstanding of expressive contents between people with different native languages/cultures background. More research is needed to explore the differences further.

Acknowledgments

This study was supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan.

References

[1] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis", *Acoust. Sci. & Tech*, 26, 317-325, (2005)

[2] C. F. Huang and M. Akagi, "A Multi-Layer Fuzzy Logical Model for Emotional Speech Perception," *In Proc. Interspeech, 2005*

[3] Ueda, K., 1988. "Should we assume a hierarchical structure for adjectives describing timbre?" *Acoustical Science and Technology* 44(2), 102-107 (in Japanese)

[4] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigne, Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds, Speech Communication,187-207. (1999)