Analysis of Production and Perception Characteristics of Non-linguistic Information in Speech and Its Application to Inter-language Communications

Masato AKAGI
JAIST, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: akagi@jaist.ac.jp Tel: +81-761-51-1236

***Abstract*— This paper introduces our ongoing research project concerned with production and perception characteristics of non-linguistic information in speech, and shows our activities for the project. The project aims at constructing universal communication environments beyond languages, nations and cultures based on non-linguistic information. To do this, we are trying to discuss what is essential in production and perception of non-linguistic information, to clarify biological common features among humans independent of languages, nations and cultures, and to apply these common features to man-machine communication as well as human-human communication. In this paper, as the results of our activities in perception of non-linguistic information, we introduce a multi-layer emotional speech perception model and present some results of emotional speech synthesis and recognition using the model.**

I. INTRODUCTION

Our ongoing research project aims at constructing universal communication environments beyond languages, nations and cultures based on non-linguistic information. To globalize and universalize human-human communications in which we can communicate among elders, infants, handicapped persons, etc. and/or machines as well as those in different languages, nations, and cultures (See Fig. 1), we are tackling the following two problems.

Problem 1: Speech production and perception play important roles in human-human communications. Additionally, speech recognition and synthesis systems that mimic human speech perception and production mechanisms also come to play important roles in human-machine communications. Thus, we need global evidence for speech perception and production to obtain knowledge for constructing the models. However, there is little knowledge that could contribute to realize universal communication environments. We have to know what relationships are important between non-linguistic information and speech production and perception, and what is essential in the chain structure of speech perception and production as shown in Fig. 2.

Problem 2: Toward the possibility to communicate with each other beyond languages, nations, and cultures, some biologically common features in speech production and perception, independent of languages, nations, and cultures,are needed, that is;

- ✓ Common organ movements for production,
- ✓ Common features produced by common movements,
- ✓ Common impression and brain activities caused by presenting common acoustic features, and
- ✓ Common behaviors among communicators.

Thus, we have to discuss what is essential in production and perception of non-linguistic information in speech, find out biological common features among humans not depending on languages, nations and cultures, and apply these common features to man-machine communications as well as human-human communications.

In order to solve the problems, in the chain structure of speech perception and production as shown in Fig. 2, our research subjects are as follows;

A) Production --> Acoustic features --> Perception: We clarify what kind of variations (a difference of the tuning, a difference of the vocal cords vibration) occur when generating non-linguistic information, from the vocal cord vibration, the vocal tract shape measurement by the MRI, the bioinstrumentation such as tongue movement measurements and their models. In addition, we examine how these variations influence acoustic features and perception of non-linguistic information through speech analysis and listening experiments using synthesized speech stimuli with controlled non-linguistic information.

B) Perception --> Acoustic features --> Production: We clarify what acoustic features vary in perceiving different sounds and what properties in speech production account for variations of these acoustic features, by tracing the reverse process with A). We elucidate interactions of the production and perception of non-linguistic information in speech through the acoustic features by accomplishing A) and B).

C) Interaction between perception and production: To know how perception and production of the non-linguistic information interact and what routes in the brain are used to interact, we measure brain activities using speech with different variations as stimuli. Thus, brain information processing in production and perception of non-linguistic information in speech becomes clear.

D) Toward global communications: Cooperating with foreign research organizations, we investigate what are common expressions of non-linguistic information and what are the common features in producing and perceiving

09-0105130519©2009 APSIPA. All rights reserved.

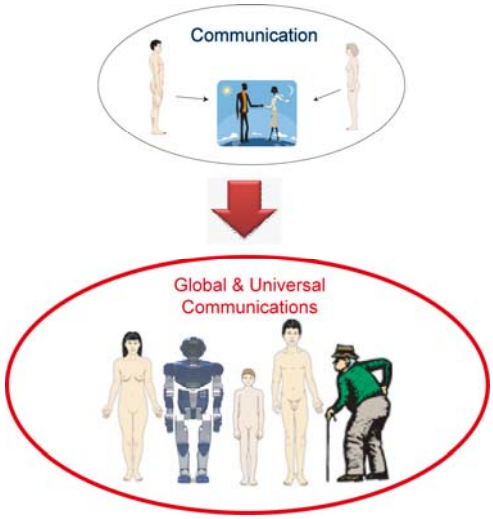


Figure 1. Toward global and universal communications

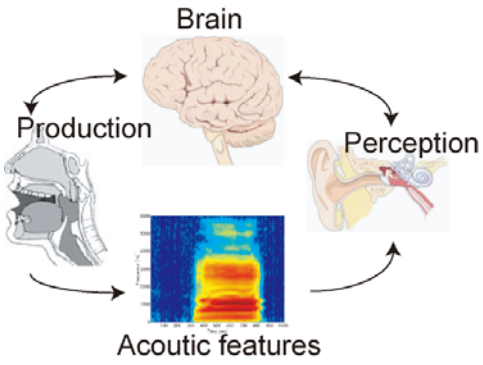


Figure 2. Chain structure of speech perception and production

non-linguistic information independent of languages and cultures.

E) Applications: As applications of the common features, we attempt to recognize and synthesize non-linguistic information in speech, such as adding non-linguistic information in speech (in synthesis) and dialog understanding based on non-linguistic information recognition (in recognition).

In this paper, as a result of our activities in perception of non-linguistic information, we introduce a multi-layer expressive speech perception model and present some application results of the model, that is, emotional speech synthesis and recognition and singing voice synthesis using the model.

II. ACTIVITIES

2-1. Multi-layer expressive speech perception model [1][2]

One of the main purposes of speech synthesis systems is to give linguistic information, which is essentially the contents of a message, to listeners. However, speech also includes non-linguistic information, such as "who speaks" and the "emotional state of the speaker." Incorporating such information into synthesized speech is a challenging research topic because it will enable a speech synthesis system to convey more information.

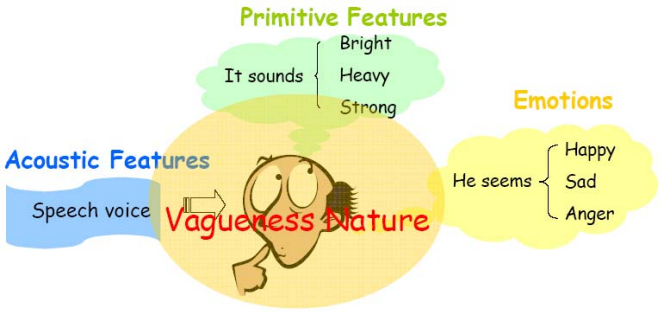


Figure 3. Schematic graph of perception of emotional voices.

Adding non-linguistic information to synthesized speech changes speech quality. It is possible to convey "who speaks" by changing the relevant acoustic features of the synthesized speech so that, for example, male- (female-) like or elder- (child-) like voices can be produced. To convey the "emotional state of the speaker", we can add appropriate acoustic features related to angry or sad voices. The problem with adding non-linguistic information is describing speech individuality or emotions in speech and then using these descriptions to produce signal processing algorithms.

This section shows a model to describe auditory impressions of speech quality and applications of the model to add non-linguistic information into synthesized speech. Two examples of the applications are as follows.

- ✓ Converting a neutral voice to an emotional voice: Synthesis of emotional voices
- ✓ Converting a speaking voice to a singing voice: Synthesis of singing voices

2-1-1. What are "emotional voices" and "singing voices"?

How do we define an angry voice? Do we answer "an angry voice is one where the power of components in the high frequency region is increased by 10 dB over their neutral one"? Although we know this answer is correct in some sense, it does not reflect auditory impressions and not so many people would give such an answer. We would probably answer that an angry voice is loud and shrill.

Since auditory impressions are usually represented by words (adjectives), it is natural to use adjectives to describe both the non-linguistic information conveyed in emotional/singing voices and the signal processing algorithms dealing with physical acoustic features. (Fig. 3) It is necessary to select relevant adjectives to describe auditory impressions to study the relationship between these adjectives and the quality of emotional and singing voices, and to study the relationship between these adjectives and physical acoustic features.

2-1-2. Multi-layer model of auditory impression

We developed multi-layer perception models (Fig. 4) for auditory impressions based on the discussion in 2-1-1 [1][2]. The concepts behind this model are that (1) high-level psychological features such as emotions (Neutral, Sad, Joy, etc.)

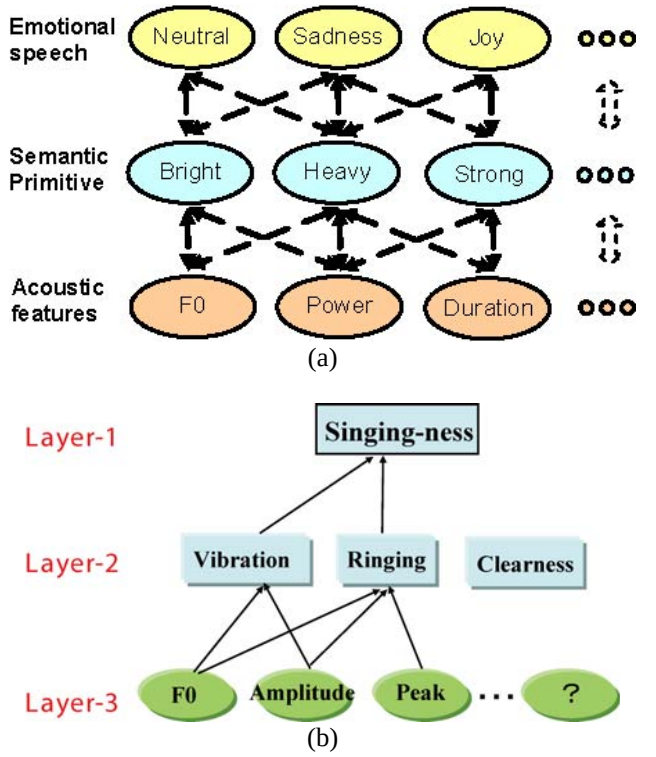


Figure 4. Multi-layer perception models for emotions (a) and singing-ness (b).

or singing-ness are explained by semantic primitives described by relevant adjectives, (2) each semantic primitive is conveyed by certain physical acoustic features, and (3) each high-level psychological feature is related to certain physical acoustic features. Note that "Singing-ness" is taken to be the auditory impression, to which a listener would consider he/she is listening to someone singing rather than talking.

2-1-3. Development of multi-layer model

A. For emotions [1]

To build the three-layer model for emotions, where the emotional speech includes the five categories, Neutral (N), Joy (J), Cold Anger (CA), Sadness (S), and Hot Anger (HA), as shown in Fig. 4(a), three experiments were initially conducted to choose relevant semantic primitives. A fuzzy inference system was then used to build the relationship.

Experiment 1 was conducted to examine utterances in terms of emotion. Stimuli were selected from the database produced by the Fujitsu Laboratory and recorded by a professional actress. Experiment 2 was conducted to construct a perceptual space of utterances in different categories. It was followed by an analysis using a Multi-Dimensional Scaling (MDS). The resulting perceptual space was used in the next experiment to select suitable primitive features for a perceptual model. Experiment 3 was conducted to determine suitable primitive features for the perceptual model. To clarify how each adjective was related to each category, 34 adjectives were superimposed onto the perceptual space using multiple regression analysis. Finally, 17 adjectives were chosen as primitive features. These were bright, dark, high, low, strong,

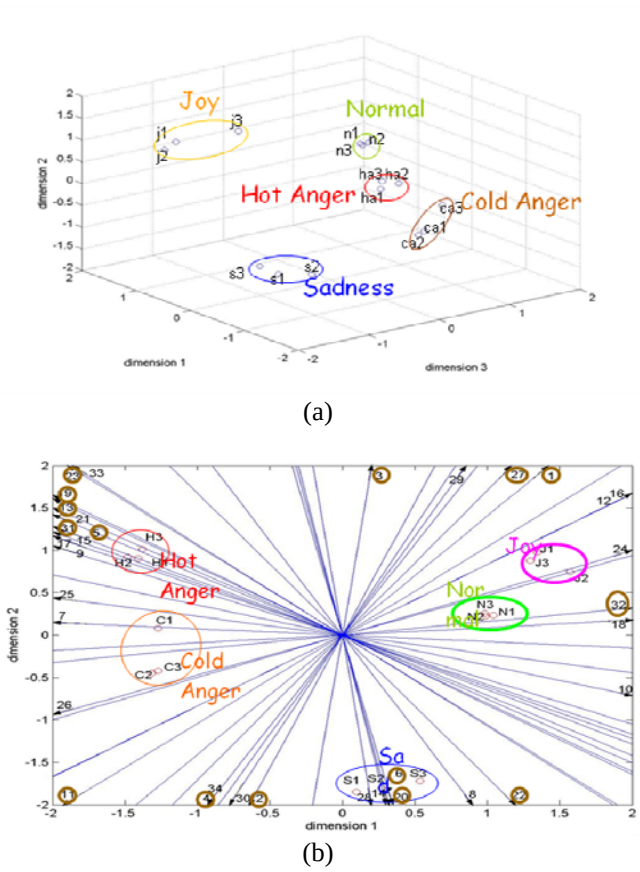


Figure 5. Perceptual space of five emotions. (a) and superimposed arrows.

weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast and slow [1]. Figure 5 (a) shows a perceptual space of the five emotions and 5(b) the 34 adjectives superimposed with arrows onto the perceptual space .

To determine the quantitative relationship between emotion and semantic primitives, we clarified what the relationship meant: the relationship represents how humans use linguistic forms to express what we perceive in speech. However, such expressions are vague, not precise. Therefore, traditional statistical methodology is not appropriate for solving this problem. We used fuzzy logic, which is well suited for building this relationship because 1) fuzzy logic embeds existing structured human knowledge (experience, expertise, heuristics) into workable mathematics; which corresponds to what the model handles, i.e., the perception of emotional speech; 2) fuzzy logic is based on natural language, which corresponds to semantic primitives in linguistic form; and 3) fuzzy logic models nonlinear functions of arbitrary complexity, which corresponds to the nonlinear and complex relationship between emotions and semantic primitives.

Thus, the relationship was determined using the MATLAB Fuzzy Logic Toolbox to process the results of the experiments. For each emotion, a fuzzy inference system (FIS) was constructed, where the input is in perceptible degrees of semantic primitives and the output is given in degrees of emotions.

To evaluate the relationship determined by FIS, we calculated regression lines that describe the relationship between

Table 1. Semantic primitives for each emotion. PF: semantic primitives, S: weighting.

Neutral		Joy		Cold Anger		Sadness		Hot Anger	
PF	S	PF	S	PF	S	PF	S	PF	S
heavy	-0.329	quiet	-0.039	slow	-0.231	sharp	-0.079	calm	-0.063
weak	-0.181	weak	-0.036	monotonous	-0.073	strong	-0.049	quiet	-0.047
clear	0.127	unstable	0.063	fast	0.153	weak	0.065	unstable	0.120
monotonous	0.270	bright	0.101	heavy	0.197	heavy	0.074	well-modulated	0.124
calm	0.103	clear	0.034	well-modulated	0.091	quiet	0.057	sharp	0.103

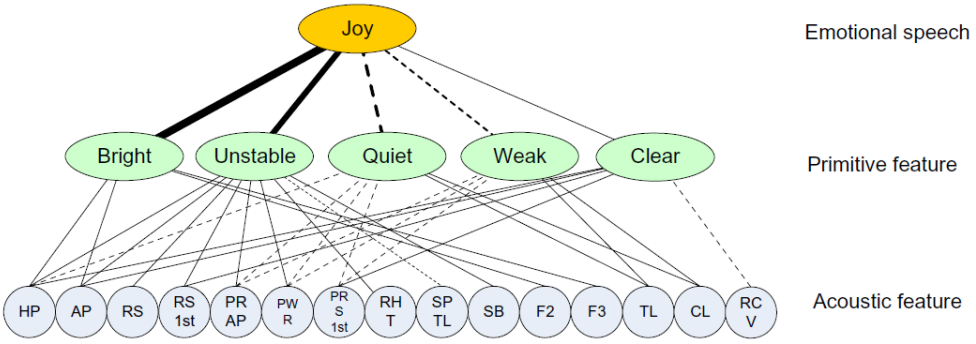


Figure 6. Resultant perceptual model of emotion Joy. The solid lines indicate the relation is a positive correlation, and the dotted ones indicate a negative correlation. The thicker the line is, the higher the correlation.

inputs (perceptible degrees of primitive features) and output (perceptible degrees of emotional speech) of each FIS. The absolute values of the slopes indicate how much the primitive features affect the categories of emotional speech. Table 1 lists the five semantic primitives that have the highest absolute value for each category. Three are positive correlations and two are negative correlations. That is, the category is most characterized by these five primitive features. As Table 1 shows, the relationship determined by FIS corresponds to our reaction when we perceive emotional speech. For example, a joyful voice always sounds bright but not quiet. This matches the FIS result for Joy. Therefore, the relationship determined by FIS is considered acceptable.

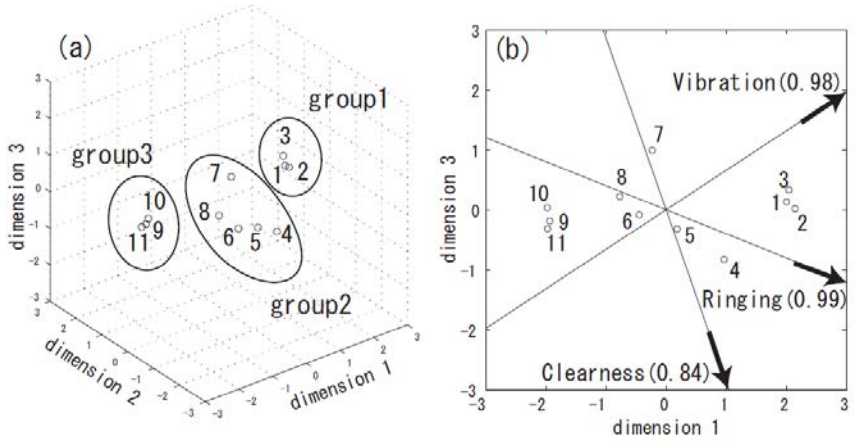
The relationship between the semantic primitives and acoustic features was determined by analyzing acoustic features in speech signals in terms of the fundamental frequency (F0) contour, power envelope, power spectrum, and duration. The F0 contour, power envelope, and power spectrum were calculated using STRAIGHT [3]. We also measured acoustic features on the basis of two aspects – accentual phrase and overall utterance – because most people do not speak continuously. For example, in Japanese the sentence /a ta ra shi i ku ru ma o ka i ma shi ta/ (“I bought a new car” in English) was always spoken with pauses in this way / a ta ra shi i ku ru ma o ka i ma shi ta/, forming 3 accentual phases, indicated by extra spaces. By comparing utterances that are the same sentence but spoken in different categories, the variation of the F0 contour and power envelope in both accentual phrases and overall utterances was evident. Taking this factor into account, some acoustic features were measured in each accentual phrase of an utterance.

Eight acoustic features were measured from the F0 contour, eight from the power envelope, three from the power spectrum, and seven from the duration. Correlation coefficient values between acoustic features and those semantic primitives that have at least one correlation coefficient over 0.6 are considered significant. There are 16 acoustic features/ Four are for F0: mean value of rising slope (RS), average pitch (AP), highest pitch (HP), and rising slope of the first accentual phrase (RS1st). Four are for the power envelope: mean value of power range in accentual phrase (PRAP), power range (PWR), rising slope of the first accentual phrase (RS1st), and the ratio between the average power in the high frequency portion (over 3 kHz) and the average power (RHT). Five are for the power spectrum: the first, second, and third formants (F1, F2, and F3); spectral tilt (SPTL); and spectral balance (SB). Three are for duration: total (TL) and consonant (CL) lengths, and the ratio between consonant and vowel lengths (RCV). Figure 6 is a resultant perceptual model of emotion Joy.

B. For singing-ness [2]

To determine acoustic features for transforming a speaking voice into a singing voice, a three-layer model was proposed (Fig 4 (b)). This model shows the relationship amongst “singing-ness” (1st layer) and semantic primitives (2nd layer) and acoustic features (3rd layer). The model is able to not only analyze acoustic features affecting “singing-ness” but also identify the semantic primitives that constitute “singing-ness” by investigating the relationship between the layers.

The singing and speaking voices used in our experiment were selected in the following way: (1) selected 80 voices of data containing the vowel /a/ from several singing and speak



(a) (b)

Figure 7. Perceptual space of "singing-ness". (a) 3-D, (b) 2-D.

ing voice sources [4], (2) performed listening tests to arrange the data in order according to "singing-ness", and (3) chose 11 voices from the 80 voices based on the result of the listening test.

"Singing-ness" consists of several semantic primitives. Therefore, the relationship between the first layer and the second layer was investigated using MDS. Since the stress in 3-D analysis was initially less than 10 %, 3-D analysis was used for MDS. Figure 7(a) shows scatter plots of the voices in 3-D space. The 11 voices are divided into 3 groups: group 1 contains voices that exhibit "more singing-ness", group 3 contains voices that exhibit "less singing-ness", and group 2 contains voices that exhibit a "medium singing-ness" between groups 1 and 3. The same grouping is shown in 2-D plane in Fig 7(b).

To reveal semantic primitives affecting "singing-ness", some experiments were done by the following steps: (1) selecting some semantic primitives that constitute "singing-ness", (2) calculating the psychological distance in each selected semantic primitives using Scheffe's paired comparison (5 grade evaluation), and (3) calculating the direction of each semantic primitives using multiple regression analysis in the space of "singing-ness". In step (1), "vibration", "ringing", and "clearness" were selected as the candidates to be semantic primitives for "singing-ness". In step (3), multiple correlation coefficients of adjectives in the 3-D were 0.99 for "vibration", 0.99 for "ringing", and 0.84 for "clearness". Since these values are high and vectors of each semantic primitive indicate different directions to each other, "singing-ness" is strongly associated with "vibration", "ringing" and "clearness".

The acoustic features affecting each semantic primitive were studied. To investigate acoustic features affecting "vibration", we focused on periodic fluctuation. There is a periodic fluctuation that is called Vibrato in the F0 contour of a singing-voice, and this characteristic affects singing-voice perception. Therefore, fluctuations in the amplitude envelope and formants were analyzed using STRAIGHT [3]. From the results, the following characteristics were found to affect "vibration."

- ✓ 4-6 Hz modulation in the F0 and amplitude envelope
- ✓ Frequency and amplitude of formants fluctuated with the same modulation frequency.
- ✓ Intervals and phases of formant deviation correspond to those of F0 deviation.

To investigate how much these characteristics affect "vibration", some synthesized voices were generated by adding these features to speaking-voice data, and listening experiments were carried out using the synthesized voices. The result shows that "vibration" increases by adding each characteristic.

Sundberg reported that there was a remarkable peak in the spectrum at around 3 kHz, and this characteristic is peculiar to a singing-voice [5]. Therefore, the spectral envelope and aperiodicity index were analyzed using STRAIGHT [3]. The results showed there are 2 types of variation in a spectral sequence creating the impression of more "ringing". Type 1 is a remarkable peak at around 3 kHz that was reported as the singing-formant [5]. Type 2 is a dip in the aperiodicity index at around 3 kHz. Such a remarkable peak indicates a strong harmonic component.

2-1-4. Testing the models

To test the resultant models, we synthesized emotional and singing voices based on the models and determined whether acoustic impressions such as semantic primitives, each emotion or singing-ness in the higher-level layers occurred by morphing acoustic features from bottom to top [1].

A. For emotions

Semantic-primitive rules were developed for morphing a neutral utterance into utterances that convey semantic primitives, based on the analysis of acoustic features that were done when building the model. There is one rule for each semantic primitive. One rule has 16 parameters that control the 16 acoustic features. The morphed utterances were then evaluated by subjects. The experimental results showed that subjects can perceive semantic primitives from the morphed utterances. That suggests the semantic-primitive rules are effective.

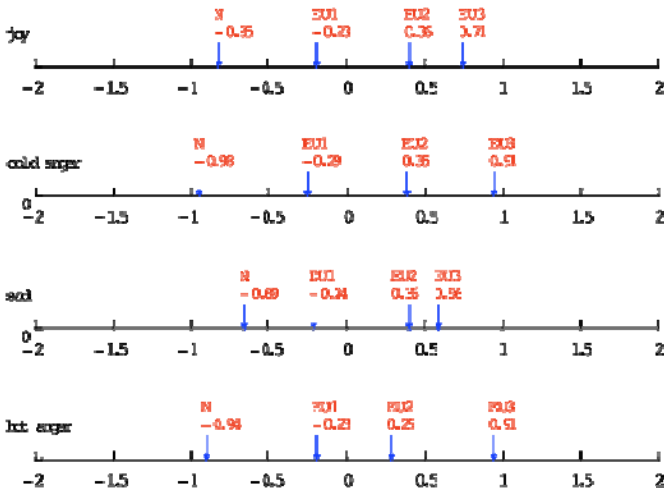


Figure. 8. Experimental result of intensity rule testing. The intended intensity levels are N < EU1 < EU2 < EU3. N is the neutral utterance, and EUn represent emotional speech utterances created by the combination of semantic primitives.

Testing of the second relationship, between emotional speech and semantic primitives, was also carried out. The results showed that the combination of semantic primitives works as a way of generating emotional speech and that a different intensity of semantic primitives can give different perceptions of the intensity of emotional speech, which corresponds to the resultant model (See Fig. 8). The demonstration of this relationship completes the construction of our proposed model for the perception of emotional speech.

B. For singing-ness

The relationships between the first layer (singing-ness) and the third layer (acoustic features) were investigated. The results showed that "singing-ness" of a voice increases by adding each acoustic feature to the speaking-voice. In particular, the effect of acoustic features affecting "vibration" on "singing-ness" is larger than that of acoustic features affecting "ringing." Therefore, to control the spectral sequence for transforming speaking-voice to singing-voice, the following procedures are required: (1) adding amplitude and frequency modulations (AM and FM) to formants corresponding to the vibrato in F0, and (2) emphasizing peaks of spectral envelopes or dips of aperiodicity indexes at around 3 kHz. The modulation frequencies of AM and FM are 5 Hz, which is the same as that of the vibrato in F0, and peak amplitude is increased by 18 dB. In order to evaluate "singing-ness" of a synthesized singing-voice, Scheffe's paired comparison was carried out. The result showed in Fig. 9 that the "singing-ness" of the synthesized singing-voice is almost the same as that of a real singing-voice.

C. Emotional recognition system

We constructed an emotional speech recognition system by using this model. The state-of-the-art emotion recognition systems directly map the emotional speech to the categories of emotion using acoustic features of speech signals. For the emotion perception of humans, however, emotion is generally

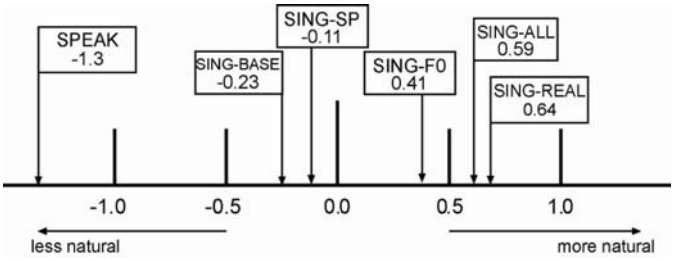
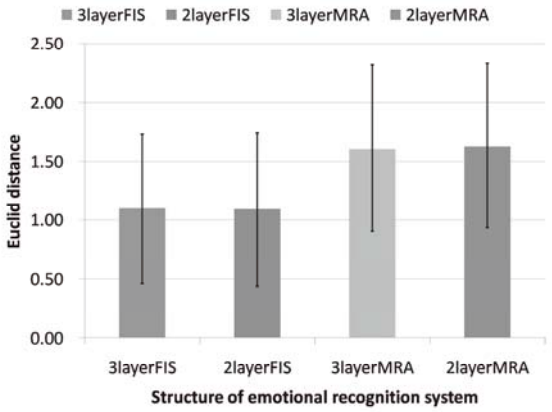
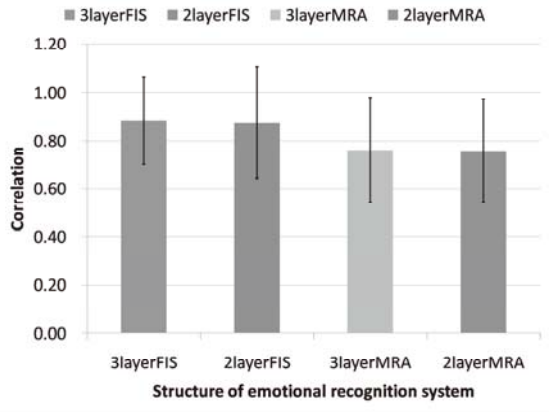


Figure 9. Result of Scheffe's paired comparison. SPEAK: speaking voice, SING-REAL: singing voice, SING-BASE: adding melody, SING-SP: adding spectral features, SING-F0: adding F0 features and SING-ALL: adding all of features.



(a)



(b)

Figure 10. Evaluation results by (a) Euclid distance between system outputs and ideal intensity, and (b) correlation between system outputs and ideal intensity.

mapped into vague semantic primitives and further recognized based on the combination of these primitives. Moreover, multiple emotions are usually perceived by humans for one speech utterance. Therefore, it is quite difficult to recognize the emotions in speech by using the mapping based techniques. To solve this problem, we attempt to construct an emotional speech recognition system which imitates the human perception mechanism by adding semantic primitives between acoustic features and emotional perception. This

constructed system is also able to recognize the multiple emotions in speech due to the use of semantic primitives.

The constructed system is composed of multiple FISs. In order to build up each FIS, we extract the acoustic features, the semantic primitives and the emotional perception from all utterances of the Fujitsu Laboratory database. The acoustic features are extracted using STRAIGHT [3]. The semantic primitives and emotional perception are obtained through listening tests by subjective assessments. Finally, the multi-layer system is combined with the multiple FIS.

To evaluate the effectiveness of the multi-layer and multiple FIS model, a two-layer model is constructed for comparative evaluation of the multi-layer model and a recognition system using Multiple Regression Analysis (MRA) is constructed for comparative evaluation of the system using FIS. Moreover, the recognition system which combines the multi-layer model and FIS is further compared with the system which combines the two-layer model and MRA. The two-layer recognition system using FIS is based on the system proposed by Moriyama and Ozawa [5]. Acoustic features and emotional perception were made the same as the emotional perception multi-layer model.

As the basis of discussion of the recognition results of each recognition system, there are two important points: Namely, whether the system output has values close to the ideal ones and whether it resembles the interaction of intensity of each emotion by listening tests. It was checked whether the output of a system would be absolutely close to an evaluation value according to the Euclid distance. It was checked also whether it was able to recognize the relative relation of emotion by correlation. The results of Euclid distance are shown in Fig. 10(a) and the results of correlation are shown in Fig. 10(b). Euclid distance which is small is better, correlation which is close to 1 is better.

The results for vagueness of human perception, i.e., for FIS and MRA, indicate that the recognition systems with FIS are more useful than those with MRA, because the Euclid distance of FIS was suppressed by 0.68 using MRA, and correlation of FIS was improved 0.12 using MRA.

The comparison results for imitation of human perception, i.e., for multi-layer and two-layer, indicate that significant differences were not observed between the recognition systems based on the multi-layer model and those based on the two-layer model even at the significance level of 0.01. These results indicate that the multi-layer system shows an internal structure clearly, and has the recognition accuracy equivalent to the two-layer system. In terms of imitating the perception mechanism of humans, the constructed system provides a more effective emotion recognition system compared with conventional methods.

III. SUMMARY

We introduced some activities in the perception parts of our ongoing research project. The contents we showed are the multi-layer model for expressive speech perception and its application to expressive speech synthesis and recognition. These are results simulated from new ideas, and their evaluations are not yet complete. We plan to illustrate the effectiveness of the model in the future with many examples of applications.

ACKNOWLEDGMENTS

This study was supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan. This work was done with Dr. Takeshi Saitou and Dr. Chun-Fang Huang.

REFERENCES

[1] Huang, C-F. and Akagi, M. "A three-layered model for expressive speech perception," *Speech Communication* 50, 810-828 (2008).

[2] Saitou, T., et al., “Analysis of acoustic features affecting “singing-ness” and its application to singing-voice synthesis from speaking-voice,” *Proc. ICSLP2004*, CD-ROM (2004).

[3] Kawahara, H., et al., “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*, 27, 187-207 (1999).

[4] Nakayama, I., “Comparative studies on vocal expression in Japanese traditional and western classical-style singing, using a common verse,” *Proc. ICA*, Kyoto, Mo4. C1. 1. (2004).

[5] Sundberg, J., “The Science of Singing Voice,” *Northern Illinois University Press* (1987).

[6] Moriyama, T. and Ozawa, S. “Measurement of human vocal emotion using fuzzy control,” *System and Computers in Japan*, 32, 4, 59–68 (2001).