2009 International Workshop on Nonlinear Circuits and Signal Processing
NCSP'09, Waikiki, Hawaii, March 1-3, 2009

# An emotional speech recognition system based on multi-layer emotional speech perception model

Yuusuke Aoki, Chun-Fang Huang, and Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa, 923–1292 Japan
Phone/FAX:+81–761–51–1699 (Ex. 1391)/+81–761–51–1149
Email: {y-aoki, chuang, akagi}@jaist.ac.jp

**Abstract**

We can communicate using speech from which various information can be perceived. Emotion is an especial element that does not depend on the content of the utterance and is useful in communications that reflects the speaker's intention. Moreover, multiple emotions are usually perceived from one speech utterance, and speech contains various emotions to the various intensity. However, traditional emotion recognition systems directly map the emotional speech to the categories of emotion using acoustic features of speech signals. It can be hardly said that these method imitate the human perception mechanism. To imitate human perception mechanism, in this study, we attempt to construct an emotional speech recognition system based on the multi-layer perception model for emotional speech proposed by Huang and Akagi [1]. We evaluate other emotion recognition systems in terms of Euclid distance and correlation between system outputs and ideal intensity. The results indicated that the multi-layer system shows an internal structure clearly and has the recognition accuracy equivalent to that of the two-layer system. In a sense of imitating the perception mechanism of humans, the constructed system provides a more effective emotion recognition system compared with the conventional methods.

## 1. Introduction

We can communicate using speech from which various information can be perceived. Information included in the speech is roughly divided linguistic information that shows the content that speaker is intended to convey and non-linguistic information that includes individual, emotion and dialect etc. of the utterance. Emotion is an especial element that does not depend on the content of the utterance and is useful in communications that reflects the speaker's intention. Moreover, speech contains various emotions to the various degrees.

In the researches on emotional speech, acoustic features are directly mapped into the emotional category. However, in human's sensory process, emotional perception is performed by perceiving the vague semantic primitives based on acoustic features and by combining these features.

In this study, 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.

## 2. Multi-layer emotional speech perception model [1]

We introduce the elements and connections used in this model. Figure 1 shows the emotion perception multi-layer model. This model was constructed for the vague human perception modeling. This model employs three-layer structure for expressing perception process from acoustic features to emotion. In particular, this model has semantic primitive layer between acoustic feature layer and expressive speech layer. Furthermore, emotion perception is modeled by combination of semantic primitives. This model can judge the change in emotion layer as semantic primitives change.

### 2.1. Composition of this model

In this section, we explain the elements which constitute each layer. Importance of acoustic features and semantic primitives elements were established in modeling process.

#### *2.1.1. Elements of this model*

A total of 16 acoustic features were used: 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

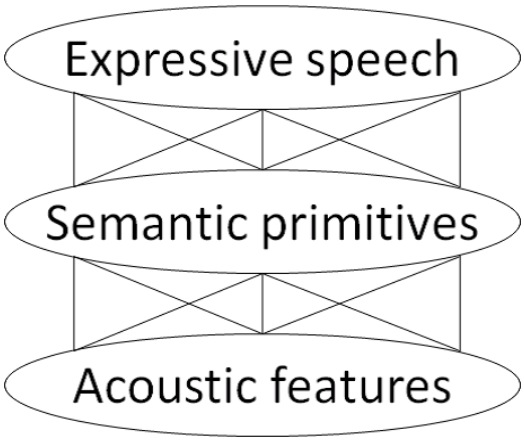


Figure 1: Conceptual diagram of multi-layered perceptual for emotional speech.

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).

Semantic primitives layer uses 17 kinds (bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow) as elements.

Emotional perception layer uses five kinds (Neutral, Joy, Cold-Anger, Sadness, and Hot-Anger) as elements. Neutral is used as a standard of all emotion. Joy, Sadness and Anger are most fundamental emotion. However, we often perceive the difference between Hot-Anger and Cold-Anger. Therefore, Hot-Anger and Cold-Anger are differently processed.

#### *2.1.2. Connections in this model*

This section introduces how to connect the elements in the model.

Human perceive emotion from semantic primitives by vague judgement. Therefore, we should take vague judgement into consideration when connecting emotion perception and semantic primitives. Fuzzy Inference System (FIS) is able to formulate the mapping from input to output and classify pattern by the rule of IF-THEN form [3]. FIS which includes both symbol processing and numeric processing represents vague experimental knowledge of human according to the IF-THEN form. FIS connected elements between layers

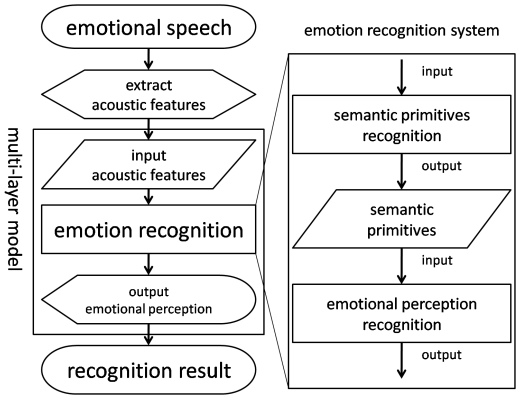


Figure 2: The flow-chart of the emotional recognition system.

based on vague judgement of human.

To connect semantic primitives and acoustic features, we should investigate which acoustic features change when human perceives the changing semantic primitives.

## 3. Emotional speech recognition system

In this study, we construct emotion recognition system based on multi-layer emotional speech perception model [1]. In our recognition system, system output does not decide one emotion category but uses emotion intensity level. Therefore, it is expected that the constructed recognition system can recognize multiple emotion intensity level as humans do.

We connect the layers to satisfy input-output relation requirements at each recognition part. In the proposing recognition system, we connect all relations between layers by Adaptive Neuro-Fuzzy Inference System unlike in the case of Huang and Akagi [1]. We implement semantic primitives recognition part which output semantic primitives intensity by using the input of acoustic features extracted from speech data, and emotional perception part which output emotion intensity by using the input of semantic primitives value with multiple FIS [3]. The recognition system can output multiple emotion intensity after revealing perception process of human using FIS.

The system we constructed is shown in Figure 2.

### 3.1. Extraction of elements

#### *3.1.1. Speech data*

179 utterances of the Fujitsu emotion speech database were used in this research. There are 20 kinds of speech sentences, and there are 9 kinds of emotions in each sentence. However, the data of 1 utterance is missed. These utterances are intended 5 kinds of emotion.

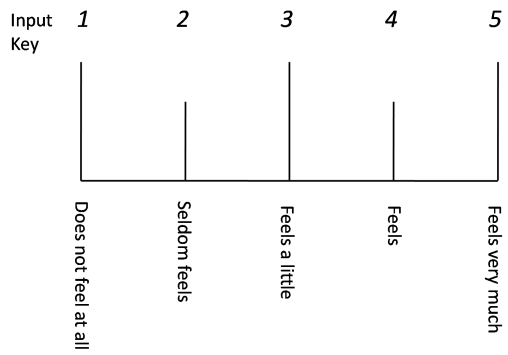


Figure 3: Evaluation measure of the listening test.

### *3.1.2. Acoustic features*

Acoustic features are required as the input of emotion recognition system. In this research, acoustic features which originate from F0, power envelope and power spectrum are extracted by the high quality speech analysis-synthesis system STRAIGHT [2]. Moreover, acoustic features which are related to duration are extracted by segmentation. About acoustic features, the ratios of changed values from Neutral are used. The average of values with the Neutral label is calculated and the value of the ratio on the basis of this is used.

### *3.1.3. Semantic primitives and Expressive speech*

We need values of subjective judgement from speech data as output of semantic primitives recognition part, input and output of emotion recognition part. To extract semantic primitives and expressive speech, we carry out listening test. About each utterance, we extract intensity of semantic primitives and expressive speech by listening test of subjective judgement.

In the listening experiments, intensity of each impression about 17 kinds of semantic primitives and 5 kinds of emotion perception for 179 kinds of utterances used in this research was evaluated by five stages, as shown in Figure 3.

In the experiments, one impression was evaluated in one session. There were 23 sessions in total. The values of semantic primitives and emotional perception were used as the input and output of the recognition system. In this research, 9 native Japanese graduate students participated in the experiments.

## 3.2. Construction of recognition system

In this section, we introduce construction of semantic primitive recognition and emotional perception recognition, as shown in Figure 2. We should also consider connection of weak relation for emotion perception. In this research, the recognition system is implemented using all the elements of this model. We use FIS in all connections.

Table 1: Four systems configurations.

| model | multi-layer | two-layer |
|---|---|---|
| using FIS | multi-layerFIS | two-layerFIS |
| using MRA | multi-layerMRA | two-layerMRA |

FIS has the structure of multiple inputs and one output. Therefore, the recognition system needs 17 kinds of semantic primitives FIS and 5 kinds of expressive speech FIS. Each initial FIS are constructed based on input and output. The initial FIS changes to the one with the ideal input-output relation by learning. The initial FIS of 17 kinds for semantic primitives and 5 kinds of emotional perception were constructed using 80% utterances of FUJITSU emotional speech database. Moreover, it is necessary to examine the suitable iteration numbers of FIS so that generation of the recognition error by over-training may not take place in neuro-adaptation study. That number is convergent point of check error of FIS proposed by Lee and Narayanan [4].

In this research, in order to decide the suitable iteration numbers the convergent point of check error of FIS in closed data was investigated based on the research of Lee and Narayanan. From the results, we decided that suitable iteration number of semantic primitives FIS is 120 and that of semantic primitives FIS is 150.

The emotion recognition system was mounted by combining 22 FIS that we constructed.

## 4. Experimental evaluation

For comparison, we adopt a two-layer model for investigating the effectiveness of the multi-layer model and a Multi Regression Analysis (MRA) for investigating the effectiveness of multiple FIS. By combination of these models, we prepared 4 kinds of systems like Table 1. We compared the recognition accuracy of these systems.

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 emotional perception multi-layer model.

As the basis of discussion of recognition result of each recognition system, there are two important points. Namely, whether system output has value close to ideal and whether 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 whether it was being able

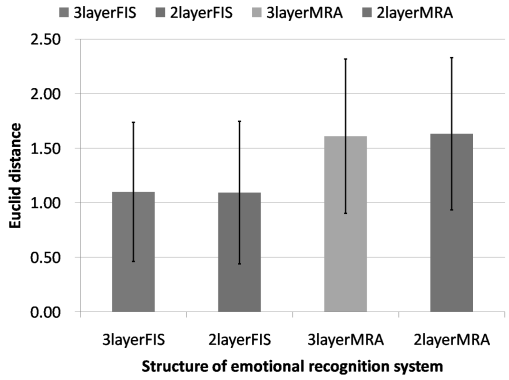


Figure 4: Evaluation results by Euclid distance between system outputs and ideal intensity.

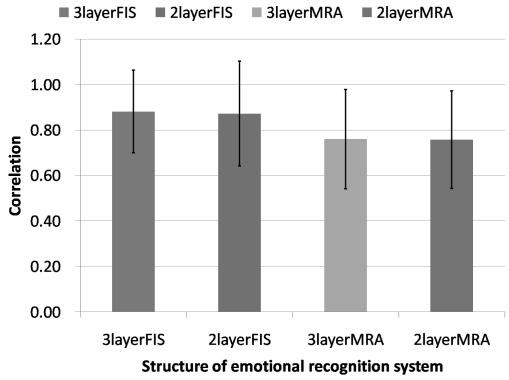


Figure 5: Evaluation results by correlation between system outputs and ideal intensity.

to recognize the relative relation of emotion by correlation.

The results of Euclid distance are shown in Figure 4 and the results of correlation are shown in Figure 5. Euclid distance which is small is better, and correlation which is close to 1 is better.

### 4.1. Evaluation of FIS

The compared 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 of MRA, and correlation of FIS was improved 0.12 from MRA.

### 4.2. Evaluation of the multi-layer model

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.

## 5. Conclusion

In this study, we constructed an emotion recognition system based on the multi-layer model proposed by Huang and Akagi [1], in order to imitate human perception mechanism. The results using FIS are better than those using MRA. Furthermore, the two-layer and multi-layer models can recognize emotion at the almost same accuracy. Since a multi-layer model can also judge the change of semantic primitives, it is better than the two-layer model in imitation of human perception. In a sense of imitating the perception mechanism of human, the constructed system provides a more effective emotion recognition system compared with the conventional methods.

## 6. Acknowledgement

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

## References

[1] C. F. Huang and M. Akagi, “A three-layerd model for expressive speech perception,” *Speech Commun.*, **50**, 810–828, 2008.

[2] H. kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Resturcturing Speech Representations Using a Pitch Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction,” *Speech Commun.*, **27**, 187–207, 1999.

[3] J. S. R. Jang, C. T. Sun, and E. Mizutani, “Neuro-Fuzzy and Soft Computing,” Prentice Hall, 1996.

[4] C. Lee and S. Narayanan, “Emotion recognition using a data-driven fuzzy inference system.” in *Proc. Eurospeech*, pp. 157–160, 2003.

[5] T. Moriyama and S. Ozawa, “Measurement of Human Vocal Emotion Using Fuzzy Control,” System and Computers in Japan, Vol. 32, No. 4, pp. 59–68, 2001.