Recognition of expressive speech based on a multi-layer model for perception

Yuusuke Aoki (0710001)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 5, 2009

Keywords: expressive speech, emotion cognition, multi-layer model, rule-based, fuzzy inference system (FIS).

1 Introduction

We can communicate using speech from which various information can be perceived. Information included in speech is roughly divided into linguistic information that shows the content that speaker intends to convey and non-linguistic information that includes individuality, emotions and dialects, etc.

Recently, speech interface is much in demand. In the future, it is expected that speech interface including non-linguistic information imitates human perception mechanism. Speech applications involving non-linguistic information can reinforce human-human communication and human-machine communication. We focus on emotions within non-linguistic information, because emotion is an special element that does not depend on the content of utterances and reflects the speaker's intention, which is useful in communication.

In the traditional researches on emotion recognition, acoustic features are directly mapped into a certain category. Certainly, phoneme and individuality, etc., are able to map into one category. However, emotions cannot map into one category. Because, multiple emotions are usually perceived

by humans from one speech utterance. Moreover, speech contains various emotions to various intensity. In human's sensory process, emotional perception is performed by perceiving vague semantic primitives based on acoustic features and by combining these semantic primitives. Therefore, it is quite difficult to recognize the emotions in speech by using the simple mapping-based techniques.

To imitate the perception mechanism of humans, we adopt the multi-layer perception model for emotional speech proposed by Huang and Akagi. In this study, we aim to recognize emotions by using this model which imitates the perception mechanism of humans.

2 Emotion recognition system

To recognize emotions by imitating the perception mechanism of humans, we construct an emotion recognition system by using the multi-layer perception model [1][2] and Fuzzy Inference System (FIS) [3]. The multi-layer perception model for emotional speech was constructed for the vague human perception modeling. This model employs a 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. In order to connect these layers, they investigated the elements with strong connection between acoustic features and semantic primitives by correlations and that between semantic primitives and emotional perception by FIS.

FIS which includes both symbol processing and numeric processing represents vague experimental knowledge of human according to the IF-THEN form. FIS is able to express vague judgement of human.

We aspire for the recognition system to imitate perception mechanism of human by using the multi-layer model and FIS. Firstly, we extract acoustic features, semantic primitives and emotional perception results from all utterances of Fujitsu Laboratory database. The acoustic features are extracted using STRAIGHT [4]. The semantic primitives and emotional perception results are obtained through listening tests by subjective as-

sessments. Finally, the multi-layer system is constructed by combining the multiple FISs.

3 Experimental evaluation

In order to investigate performance of the constructed recognition system, we compare our system and other traditional recognition systems. 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. We investigate whether it is able to imitate vague process of human perception by comparing FIS and MRA, and to express sensory process of human perception in term of semantic primitives by comparing the multi-layer model and the two-layer model. Recognition accuracy was measured by Euclidean distance on the absolute scale and the correlation on the relative scale using these systems.

4 Conclusions

In order to imitate human perception mechanism, in this study, we constructed an emotion recognition system based on the multi-layer model proposed by Huang and Akagi.

The evaluation results for FIS and MRA indicate that the recognition systems with FIS are more useful than those with 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. Therefore, we can recognize emotion by imitating human perception mechanism.

References

[1] Chun-Fang Huang, Masato Akagi, "A Multi-Layer fuzzy logical model for emotional speech perception," *Proc. EuroSpeech 2005*, pp. 417–420, Lisbon, Portugal, 2005.

[2] Chun-Fang Huang, Masato Akagi, "A three-layerd model for expressive speech perception," *Speech Commun.*, Vol.50, pp. 810-828, 2008.

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

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