A method for synthesizing emotional speech using three-layered model based on a dimensional approach

Yawen XUE†, Yasuhiro HAMADA‡ and Masato AKAGI†
†School of Information Science, Japan Advanced Institute of Science and Technology,
‡Meiji University
E-mail: †xue_yawen@jaist.ac.jp

Abstract : The purpose of this study is to propose an emotional voice conversion system utilizing the three-layered model for dimensional approach. In this study, we first estimate the values of acoustic features using the three-layered model. The proposed model consists of three layers: acoustic features in the top layer, semantic primitives in the middle layer, and emotion dimension in the bottom layer. Fuzzy Inference System (FIS) is used as estimating tools to connect the three layers. After obtained the synthesized emotional speech, listening tests are carried out to find out whether the converted emotional speech can bring the same impression as intended. Results show that the converted emotional speech can give the same impressions and similar intensity as intended.

Keywords : emotional voice conversion, three-layered model, dimensional approach

1. Introduction

In the field of human-computer-interface (HCI), one of the goals is to improve user experiences by providing genuine human communication. Thus, a speech-to-speech translation (S2ST) system plays a consequential role for converting a spoken utterance from one language into another to enable people who speak different languages to communicate. Conventional S2STs focus on processing linguistic information only, which is deficient in synthesizing affective speech, such as emotional rather than neutral speech. Therefore, a system that can recognize and synthesize emotional speech would be momentous.

Branswikian lens model, Huang and Akagi proposed a three-layered model (acoustic features layer, semantic primitives layer, and emotion layer). For the emotion layer, in this study, a dimensional emotion space spanned by Valence-Activation (V-A) axes, is used to model the human emotions.

The input and output of our system are the dimensional parameter values in V-A space and the corresponding acoustic feature displacements, respectively. An Adaptive-Network-based Fuzzy Inference System (ANFIS) is used to connect the three layers. The related acoustic features of every semantic primitive are selected when synthesizing the emotional speech. Listening tests were carried out to verify whether the synthesis speech can give a position similar to that anticipated. On the basis of the listening test, effectiveness is discussed.

2. Outline of the proposed system

This section outlines the proposed emotional speech conversion system. The system can be divided into two processes. The first block called estimation part, which estimates acoustic feature values for each position on V-A space. The second block converts parameter values according to the estimated acoustic feature values to emotional speech which we called modification part. There are two inputs of the emotional voice conversion system, the position in Valence-Activation space and the neutral speech which you would like to convert. The ultimate goal of this paper is to improve estimation accuracy and enhance the modification of acoustic features. In model creation, the evaluated emotion dimensions, evaluated semantic primitives, and the extracted acoustic features are firstly needed to be acquired by listening tests and some tools. To connect the three layers, we use the fuzzy inference system (FIS). After the system has been built, the parameter values can be estimated when given the intended position in the emotion dimensions in the estimation of acoustic features part. To obtain the emotional converted speech, the parameter values of neutral speech are modified according to the estimated acoustic features using some tools and models.

3. Conclusion

This paper proposed an emotional speech synthesis system using a three-layered model in a dimensional approach. ANFIS was used to connect the three layers for estimating the semantic primitives and acoustic features. The related acoustic features were used for synthesizing the emotional speech by morphing rules. The higher correlation coefficient between the estimated acoustic features and extracted acoustic features shows that the three-layered model estimates acoustic features more accurately than the two-layered model. Results of subjective evaluations revealed that the emotional speech converted by three-layered model using the new modification method, Fujisaki model, can give the intended impression to a much similar degree as than the two-layered model in the emotion dimension, which is a great improvement.