Title	Rule-based emotional voice conversion utilizing three-layered model for dimensional approach
Author(s)	薛，雅文
Citation	
Issue Date	2015-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/12926
Rights	
Description	Supervisor: Masato Akagi, School of Information Science, Master

Rule-based emotional voice conversion utilizing three-layered model for dimensional approach

Yawen Xue (1310201)

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

September 24, 2015

Keywords: Emotional voice conversion, three-layered model, dimensional approach, fuzzy inference system.

In the field of human-computer-interface (HCI), improving user experiences by providing genuine human communication with computer is one of the motivations. A speech-to-speech translation (S2ST) system plays a consequential role for converting a spoken utterance from one language into another using computer technology to enable people who speak different languages to communicate with each other. 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. This research consider the emotional speech synthesis part which use the conversion method to get the emotional speech.

To construct a system for synthesizing emotional speech, several studies have already obtained some achievements. Most methods are based on a concatenative approach, like unit selection, or a statistical parametric approach, like the Hidden Markov Model (HMM) with the Gaussian Mixture Model (GMM). Both methods can synthesize emotional speech with good quality when the emotion is present in a category such as happy, sad, or angry. However, they can only synthesize the emotional speech with the average emotion intensity (not strong or weak emotions) in the emotion category, and both need a huge database for training, although it is difficult to collect many human responses when listening to emotional speech. In human speech communication, people sometimes strengthen or weaken emotional expressions depending on the situation. Thus, a small number of discrete categories is not sufficient to mimic the emotional speech in daily life.

Therefore, some researchers proposed a multi-dimension approach to express emotion on a continuous-valued scale instead of categorical methods. The advantage of the dimensional approach is not only that they can represent the emotion in all degrees but also providing a standard criterion for all research because the value of position in multi-dimensional is done by human experiment in the same dimension. And by using the rule-based synthesis method, tendencies of the variations in multi-dimensional space can be acquired using a small database. With the tendencies of variation, the synthesized speech can convey all degrees of an emotion.

The purpose of this study is to propose a emotional conversion system utilizing three-layered model for dimensional approach. The ultimate goal of our work is to improve the conventional emotional speech conversion system based on dimensional method in order to precisely predict the acoustic features as well as to synthesize affective speech much similar to what intended in dimensional space.

To improve the accuracy of estimated acoustic features, a three-layered model is adopted. According to the Branswikian lens model, people's emotion perception is multi-layered. Human beings do not perceive emotion directly from the acoustic features, so semantic primitives such as bright, high, strong, and so on are also of great importance. In these circumstances, this paper basically utilized the three-layered model proposed by Huang and Akagi (acoustic features layer, semantic primitives layer, and emotion layer). For the emotion layer, in this study, a dimensional emotion space is used to model the human emotions. The Valence-Activation (V-A) axes in the two-dimensional emotion space can describe the strength, such as very or slightly happy, which gives a more flexible interpretation of emotional states. An Adaptive-Network-based Fuzzy Inference System (ANFIS) connects the three layers and estimates corresponding values of dimensional axes.

The modification of estimated acoustic features for emotional speech synthesis is revised following the work of Hamada, especially for the fundamental frequency (F0) related acoustic features. The Fujisaki model is adopted to extract the trajectory of the F0 contour from which we can obtain the F0 related acoustic features all at once. This work was done by Hamada and we adopt this method in three-layered model.

In this study, two problems are anticipated to solve which are improving the conventional emotional speech conversion system based on dimensional method in order to precisely predict the acoustic features as well as to synthesize affective speech much similar to what intended in dimensional space. The higher correlation coefficient shows that the three-layered model estimates acoustic features more accurately than the previous two-layered model. Results of subjective evaluations revealed that the emotional speech converted by three-layered model using the new modification method can give the intended impression to a much similar degree as than the previous two-layered model in the emotion dimension. And the naturalness of the converted speech achieve an average level by subject evaluations. Above all, a conclusion can be made that an emotional conversion system utilizing three-layered model in dimensional approach can achieve better quality converted emotional speech than previous method.