Abstract:

In terms of human-computer interaction (HCI), synthesized speech has burgeoned at a rapid rate in recent years to fulfill demands for daily speech communication. Natural sounding synthetic speech with only linguistic information is currently used in modern applications such as text to speech systems (TTS), navigation systems, robotic assistants, story teller systems and speech to speech translation systems (S2ST). Information conveyed by speech should be summarized through linguistic information, paralinguistic information as well as nonlinguistic information. Synthesized speech with only linguistic information cannot encompass all of these factors. Therefore, emotional synthesized speech that allows communication of nonlinguistic information is increasingly required. Emotions, ranging from an underlying emotional state to full-blown emotions, contribute substantially to the acoustic manifestation of the spoken language. In order to incorporate emotions to neutral TTS synthesized speech, an emotional voice conversion system which can convert the neutral speech to emotional one is necessary to cope with.

Previous prevalent methods concerned with emotional voice conversion systems mainly used statistical approach such as Gaussian mixture models, deep neural network, or neural network method. They mainly utilized some simple discrete labels such as joy, sad, anger to represent emotion. However, humans have ability to express mild emotion such as a little bit joy or very joy which is a continuum of non-extreme states. This means statistical approaches need large parallel linguistic database of affective voices with a continuum of non-extreme affective states which is a costly and impossible problem. Little attention is paid on controlling the emotional degree in a continuous scale in previous studies.
A rule-based voice conversion technique which can obtain variation tendencies of acoustic features with a limited database is utilized in this research.

When modeling continuous controllable degrees of emotional synthetic speech, two primary problems are firstly needed to be considered. The first one is how to describe emotions and another problem is how to model emotion perception process of human beings.

In the literature, there are many descriptive systems for emotions. The most straightforward description is the utilization of emotion-denoting words or category labels called emotion category. Emotions in daily speech communication are highly diverse. Many human-machine dialogues need machines to express mild and non-extreme emotional states. Therefore, an emotion dimensional approach which satisfies the requirement to express a range from low-intensity to high intensity states is appropriate for representing a continuum of non-extreme emotional states for controlling the degree of emotion. In this research, two dimensions arousal (synonymous to activation and activity) and valence (synonymous to evaluation) are used for emotional speech conversion based on the database we have.

Another problem related to emotion conversion is modeling the process of expression and perception of emotion by human beings. Many researchers based their theory and research on a modified version of the Brunswik's functional lens model of perception. Huang and Akagi proposed a three-layered model for expressive speech perception with emotion (listener attributions) at the top layer, semantic primtives (proximal percepts) at the middle layer, and acoustic feature (distal indicators) at the bottom layer. They assumed that humans perceive emotion not directly from acoustic features, but semantic

primitives, such as fast, bright, and so on also play important roles. The three-layered model had already been applied by many researchers in the emotion recognition area. In this research, we assume that the human production of emotion follows the opposite direction of human perception. This means the encoding process of the speaker is the inverse process of the decoding of the listener. In that case, an inverse three-layered model is employed as the structure between emotion and acoustic feature.

The related acoustic features to each dimension are investigated as applied to emotional speech synthesis. Subjects are asked to evaluate the synthesized speech, the specific acoustic features of which, such as F0, have been replaced by the F0 from the emotional speech but leaving the other acoustic features of the neutral speech. We concluded that both the F0 trajectory and spectral sequence are important to emotion conversion. The power envelope and duration show little influence on the valence axes. In this research, we focused more on the prosody-related features such as duration, F0 and power envelope.

In order to control the emotion degree in dimensional space using the inverse three-layered model, an emotion conversion system was proposed with two inputs (positions in dimensional space and neutral speech) and two steps (rule extraction and rule application). In the first step, the rules between acoustic feature variations of neutral and emotional ones can be extracted using a fuzzy inference system. The inverse three-layered model is set as the structure between emotion dimension and acoustic features with emotion dimension as the bottom layer, semantic primitive layer in the middle and acoustic features layer at the top. The second step is to apply the rule-based voice conversion method to modify the acoustic features of neutral speech to emotional ones following the rules extracted from the first step. It is widely understood that emotion is conveyed by means of a number of prosodic parameters such as voice quality and speech rate as well as fundamental frequency. In this step, some essential prosody features such as duration, F0 contour and power envelope are parameterized by an interpolation method, Fujisaki model and target prediction model. Then the modified acoustic features are synthesized using STRAIGHT. Perceptual evaluation results in V-A space show that the synthesized speech of joyful, sad and cold angry emotion can be perceived well, including the category and the degree, although the perceived degree is decreased compared to the desired values. For hot anger emotion, since the spectral modification was not conducted, the synthesized speech of hot anger is perceived as a joyful emotion.

Commonalities and differences of human perception for perceiving emotions in speech among different languages in dimensional space have been investigated in Han et al., 2016. Results show that human perception for different languages is identical in dimensional space. According to this result, we assume that, given the same direction in dimensional space, we can convert the neutral voices in multiple languages to emotional ones with the same impression of emotion. It means that the emotion conversion system could work for other languages even if it is trained with a databases in one language. We try to convert neutral speech in two different languages, English and Chinese using an emotion conversion system trained with Japanese database. We find that all converted voices can convey the same impression as Japanese voices. On the case, we can make a conclusion that given the same direction in dimensional space, the synthesized speech among multiple language can convey the same impression of emotion. In a word, the Japanese emotion conversion system is compatible to other languages.

In conclusion, this research proposed a method for emotional voice conversion with degree continuously controllable using dimensional representation following human emotion production mechanism. Perception results show that the synthesized stimuli can be perceived with the same tendency as intended in dimension space except hot anger. Neutral voices in other languages were directly inputted into the system without training, and perception results show that the conversion system built in one language is capable for other languages without training. The emotional navigation systems, robotic assistants and S2ST system will bring an intelligent HCI and enormous promotion in human life quality. As this system enables to convert the input neutral speech from any target speaker in any language without training, it can reduce an amount of cost and make the emotional TTS applicable. And this will give a big progress in the field of emotional voice conversion. The emotional navigation systems, robotic assistants and S2ST system will bring an intelligent human computer interface HCI and enormous promotion in human life quality.

Keywords: Emotional voice conversion, rule-based speech synthesis, emotion dimension, three-layered model, Fujisaki F0 model, target prediction model.