Abstract

In human-human communication, speech is the most direct way for communication. Speech contains a lot of information of speaker, such as emotion, gender, age, native and level of education. Emotions play a vital role in speech understanding. By using appropriate emotions, the same textual information can be used to convey different meanings. Emotions in speech can be used not only to express intentions, but also to understand intended information by combining potential emotions, voice information, and other linguistic factors.

Although acoustic features of emotional speech have been investigated, it is still difficult to model emotions by using only these acoustic features. Many studies have shown that the speech production organs features, such as glottal waveforms and vocal tract shapes are important. However, the properties of glottal source and vocal tract to expressive emotions via acoustic features have not been investigated deeply yet. Thus, this study focuses on investigating the effects of glottal source and vocal tract cues on emotional speech perception, particularly for vowel, since vowel plays more important role in emotional speech.

The research aims to investigate the effects of glottal source and vocal tract cues on perception of emotional vowels, especially after removing known effects of dominant prosody features (e.g., pitch, intensity, and duration). Thus, (1) an analysis-by-synthesis method is firstly developed for estimating glottal source waveform and vocal tract shape of emotional vowel. (2) the glottal source waveform and vocal tract shape are estimated from Japanese vowel /a/, and the spectral tilt and character of vocal tract shape are consistent with previous results. (3) F_0/pitch (fundamental frequency), intensity (E_e-related), duration of source related features (prosody features), spectral tilt of glottal source waveform, and F_1 (first formant frequency) are discussed, which in a controlled way of modifying the estimated glottal source waveform and vocal tract shape and utilized for establishing an analysis-by-synthesis method for resynthesizing the emotional vowels. Then, Japanese natives with normal hearing participate in the evaluation of perceptual rating emotions in the valence and arousal space. The results show that the glottal source information plays an essential role in perception of emotions in vowels, whereas the vocal tract information contributed to the valence and arousal perception after neutralizing the F_0, intensity, and duration cues effects.

This study investigates emotional vowels from the point of view of speech production. The results contribute to further understanding the emotional speech production mechanism, also can enlighten many emotional speech fields, such as emotional speech recognition, synthesis and conversion. Moreover, an accurate estimation method of glottal source waveform and vocal tract shape is proposed for vowels in this study. It can be used in many speech signal processing fields, for example, speech analysis, speech synthesis, voice pathology detection,

speaker recognition, and speech recognition.

Keywords: Emotional vowel production, emotional vowel perception, glottal source waveform, vocal tract, ARX-LF model, valence and arousal