Abstract

Analyses and synthesis of emotional sounds is an exciting research direction. Moreover, a sophisticated emotional sound synthesis system can significantly improve experiences of human-computer communication. Although humans can express subtle emotional changes in their voices, most researches on emotional speech synthesis focus on the categorical approach to emotional states expressions, such as synthesizing speech to joy, sadness or anger. Besides categorical approach, some studies tried to control speech emotion continuously as humans do. As the study of Y. Xue has constructed an emotional speech conversion system using a rule-based approach and a three-layer model, following the emotional perception and production of human being. However, the system has rooms to improve in continuous emotion control, especially on Valence scale.

Whether categorical or continuous emotional speech synthesis, it is necessary to face a common problem, which researchers can only get categorical emotional voice data in the most case, and it is impossible for asking a human actor to record emotional voices data in a regular gradient variation, whether respecting physical acoustic features or emotional perception. Therefore, categorical data determines that many studies focus on categorical approaches. Even study as Xue's emotion conversion system, which focuses on continuous emotional speech synthesis, is trained by categorical data. Consequently, discontinuous training data had distorted the mapping rules between acoustic features and emotional impression to a certain extent. Also, limited and discontinuous training data makes studies as Xue's system fail to clarify the correspondence between some important acoustic features variations and emotion impression.

For the purpose to obtain emotional voices continuously spanned on the V-A space, discuss what acoustic features are important to emotional impressions and how those features relate to emotion perception in a more detailed way, this study has two sub-goals: (i) Obtaining emotional speech samples continuously spanned on the V-A space by morphing techniques and collect the impressions of synthesized voices. (ii) Examining how acoustic features related to perceptions of emotional speech. Therefore, this study interpolates voices from pairs of typical emotions with a morphing method, collects emotion scores on Arousal-Valence space by a listening test, and analyzes which acoustic features significantly influence emotion perception and how those features vary changes emotion impression.

Analyses based on acoustic features and evaluation scores show that Arousal perception can be stably described by merely using fundamental frequency (F0). Power related features have a significant influence on Arousal perception, however, limited on sad-related voices. Comparing to Arousal, this research found that F0 and formants significantly influence Valence perception simultaneously, and how acoustic features correspond to Valence perception vary with different morphing references. Considering the correspondence and significances vary across different acoustic features for different morphing groups, this study proposed an assumption that how acoustic features relate to Valence perception depends on different areas of V-A space, and it is necessary to manipulate formants-related features in order to obtain high quality of Valence control in synthesized emotional voices.

Keywords: morphing voices, emotional speeches, acoustic features, relation analysis.