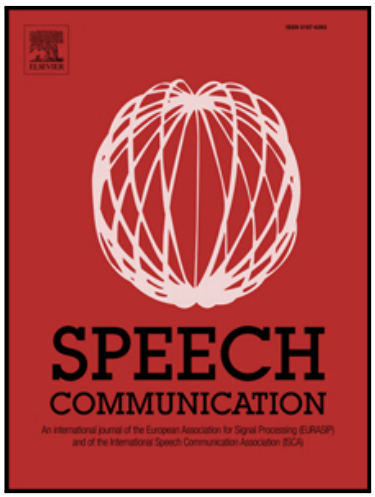

Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space

Yawen Xue, Yasuhiro Hamada, Masato Akagi

PII: S0167-6393(17)30318-7
DOI: 10.1016/j.specom.2018.06.006
Reference: SPECOM 2576

To appear in: *Speech Communication*

Received date: 18 August 2017
Revised date: 31 March 2018
Accepted date: 25 June 2018

Please cite this article as: Yawen Xue, Yasuhiro Hamada, Masato Akagi, Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space, *Speech Communication* (2018), doi: 10.1016/j.specom.2018.06.006

Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space

Yawen Xue*, Yasuhiro Hamada, Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan

Abstract

This paper proposes a rule-based voice conversion system for emotion which is capable of converting neutral speech to emotional speech using dimensional space (arousal and valence) to control the degree of emotion on a continuous scale. We propose an inverse three-layered model with acoustic features as output at the top layer, semantic primitives at the middle layer and emotion dimension as input at the bottom layer; an adaptive-based fuzzy inference system acts as connectors to extract the non-linear rules among the three layers. The rules are applied by modifying the acoustic features of neutral speech to create the different types of emotional speech. The prosody-related acoustic features of F0 and power envelope are parameterized using the Fujisaki model and target prediction model separately. Perceptual evaluation results show that the degree of emotion can be perceived well in the dimensional space of valence and arousal.

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

1. Introduction

In terms of human-computer interaction (HCI), synthesized speech has burgeoned at a rapid rate in recent years to fulfill the demand for daily speech communication. Natural sounding synthetic speech with only linguistic information is currently used in modern applications such as text to speech systems, navigation systems, robotic assistants, story teller systems and speech to speech translation systems. Fujisaki proposed that information conveyed by speech should be summarized through *linguistic information*, which is discrete categorical information explicitly represented by the written language or uniquely inferred from context; but also *paralinguistic information*, discrete and continuous information added by the speaker to modify or supplement the linguistic information, as well as *non-linguistic information*, information not generally controlled by the speaker, such as the speaker's emotion, gender, age, etc [1]. Synthesized speech with only linguistic information cannot encompass all of these factors, thus resulting in unnatural speech sounds. Therefore, affective synthesized speech that allows communication of nonlinguistic information, such as affect and intent, is increasingly required [2] [3] [4]. Affect is not restricted to emotion; for instance [5] [6], there are social affective expressions, such as expression of politeness, sarcasm, irritation, flirtation, etc., which may be more or less controllable. Emotions, ranging from an underlying emotional state to full-blown emotions, contribute substantially to the acoustic manifestation of the spoken language. In order to improve the naturalness of synthetic speech, it is necessary to incorporate the effect of emotion on speech.

Previous methods for emotional voice conversion utilized a categorical approach to express emotional states [7]. One method is the piece-wise linear mapping using a probabilistic model, Gaussian Mixture Models (GMM) [8] [9] [10] [11]. Kawanami [12] first applied GMM for spectrum transformation to emotion voice conversion. Tao [13] tested three different methods for prosody conversion and found that GMM is suitable for a small database while a classification and regression tree model will give better results if a large context-balanced corpus can be obtained. Inanoglu [14] combined a Hidden Markov Model, GMM and F0 segment selection method for transforming F0, duration and short-term spectra in data-driven emotion conversion when large amounts of parallel data are needed. Aihara [15] improved the GMM-based emotional voice conversion for both voice quality and prosody feature conversion.

Former studies [12] [14] [15] [16] [17] [18] [19] considered converting neutral speech to simple categories of emotions such as joy, anger and sad. Tao tried to label the emotion database using four degrees “strong,” “normal,” “weak,” “unlike” to each emotion category [13]. However, daily social emotions conveyed by humans are mild and not purely one emotion or another, but a mixture of emotions, e.g., anger and sad and fearful; they can be described as a continuum of nonextreme states [20] [21]. So synthetic speech with simple categories of emotions is not sufficient. This paper focuses on converting neutral speech

*Corresponding author

Email addresses: xue_yawen@jaist.ac.jp (Yawen Xue), y-hamada@jaist.ac.jp (Yasuhiro Hamada), akagi@jaist.ac.jp (Masato Akagi)