Detecting affective states from text based on a multi-component emotion model[☆]

Yingying Gao, Weibin Zhu [*]

Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China

Abstract

A multi-component emotion model is proposed to describe the affective states comprehensively and provide more details about emotion for the application of expressive speech synthesis. Four types of components from different perspectives – *cognitive appraisal*, *psychological feeling*, *physical response* and *utterance manner* are involved. And the interactions among them are also considered, by which the four components constitute a multi-layered structure. Based on the describing model, a detecting method is proposed to extract the affective states from text, as it is the requisite first step for an automatic generation of expressive synthetic speech. The deep stacking network is adopted and integrated with the hypothetic producing process of the four components, by which the intermediate layers of the network become visible and explicable. In addition, the affective states at document level and paragraph level are regarded as contextual features to extend available information for the emotion detection at sentence level. The effectiveness of the proposed method is validated through experiments. At sentence level, a 0.59 *F*-value of the predictions of utterance manner is achieved.

© 2015 Elsevier Ltd. All rights reserved.

Keywords: Multi-component emotion model; Deep stacking network; Visible intermediate layers; Contextual impact; Expressive speech synthesis

1. Introduction

Natural generation of expressive synthetic speech presents the researchers with two main requirements: one is an automatic detection of affective states from the given text, including a proper description of emotion; the other is the effective transformations on speech signal according to the delivered affective messages (Bellegarda, 2011; Roach, 2000; Trilla and Alías, 2013). Most of the studies in expressive speech synthesis (ESS) area are devoted to address the second issue, see Govind and Prasanna (2013) for a comprehensive review, in which the desired emotion is given directly as an additional input along with text. And the emotion is described by the ready-made models borrowed from other disciplines like psychology, which is suggested to be fairly inadequate in the ESS scenario. This work focuses on describing affective states more comprehensively and searching an available text-based detecting method based on

---

the newly proposed description model, as they are the requisite first step for an automatic generation of expressive synthetic speech.

Speech researchers are increasingly paying attention on identifying the affective states from text to direct the subsequent expressive rendering. For example, Trilla and Alías (2013) and Alm et al. (2005) explored the Sentiment Analysis (SA) problem at sentence level to inform a Text-to-Speech (TTS) system about a proper sentiment (*negative, positive* or *neutral*). Wu et al. (2009) adapted the *Pleasure-Arousal-Dominance* (PAD) model to describe the affective information of particular words and then to guide the acoustic variation at different levels. Tao (2003) and Bellegarda (2011) placed emphasis on the generation of a number of typical discrete emotions (*happiness*, *anger*, *sadness*, *fear* and *disgust*). It is implied that the affective states are often simplified to a number of categories or several dimensional values in the previous work. The categorical model supplies an intuitive description of emotion for humans, but it is quite confused for the TTS system to get further guidance about the expressive rendering. The dimensional model supplies two or three primary characteristics which are deemed to be more directly correlated with vocal variations, but it is argued that the projection from the possibly structured multi-dimensional space of emotion to the fairly reduced space with two or three homogeneous dimensions may cause a loss of information (Cowie and Cornelius, 2003).

There are plenty of ways to define or delimit emotion. In a narrow sense, it is restricted to subjective feelings; while in a broader definition, it includes both the emotional states and the emotion-related states (e.g., arousal, attitude), where the latter are suggested to be more pervasive in everyday conversations and probably more important to speech (Cowie and Cornelius, 2003; Scherer, 2003). Therefore, following the broader definition of emotion, this work proposes a multi-component model, in the hope to provide a relatively comprehensive description of emotion and supply more information for the expressive rendering. Four components: *cognitive appraisal*, *psychological feeling*, *physical response* and *utterance manner* are included in the model and distributed in a multi-layered structure. Emotions are represented by different patterns of the four components, and *utterance manne*r will be used as an interface between the affective description and the acoustic signal of emotional speech in the ESS system.

To detect these affective states from textual content, the Deep Stacking Network (DSN) (Deng et al., 2012) is exploited, owning to the multi-layered stacking architecture which suits the structure of the proposed emotion model properly. The DSN is integrated with the hypothetic producing process of the affective components, forming a stacking network with visible intermediate layers. Besides the interactions among different components, the contextual impacts are also taken into account in this work. In the TTS scenario, the analyzing granularity of texts is usually determined as the sentence level, so that the natural expressive variations can be investigated inside the document. In this work, instead of extracting the affective states at the sentence level directly, the emotions of documents and paragraphs are analyzed first, and then sent to the lower levels as a contextual reference. Given that the information provided by a sentence is rather reduced, this method can be regarded as a tactic to extend extra available information. Experiments are implemented to validate the effectiveness of the detecting method and the rationality of the describing model.

The rest of the paper is organized as follows. Section II presents the existing emotion describing models and detecting methods from other related work. Section III describes the details of the proposed multi-component emotion model. Section IV elaborates the detecting method according to the multi-component model. Section V illustrates the verification experiments and discusses the results. The last section draws the conclusions and the future work.

2. Related work

This section outlines some existing emotion describing models in psychology, and the commonly used techniques to detect emotions from text.

2.1. Emotion describing models

There is no generally accepted methodology to describe emotion at present. Extensive theories and approaches have been proposed in psychology. Cowie and Cornelius (2003) divided the affective states into emotional states (subjective feelings) and emotion-related states (e.g., attitudes and physical arousal). Likewise, Ortony et al. (1987) divided these states into physical/bodily states and mental conditions, in which the former corresponds to physical arousal and the latter includes "a significant cognitive component or a significant affective component, some have a significant behavioral component". Cornelius (2000) summarized the explanations of emotion into four major perspectives – *Darwinian*, *Jamesian*, *Cognitive appraisal*, and *Social constructivist*. Similarly, Scherer (2000) summarized these