Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 1049–1063

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speech-based recognition of self-reported and observed emotion in a dimensional space

Khiet P. Truong [a,*], David A. van Leeuwen [b], Franciska M.G. de Jong [a]

[a] University of Twente, Human Media Interaction, P.O. Box 217, 7500 AE Enschede, The Netherlands
[b] Radboud University Nijmegen, Centre for Language and Speech Technology, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Received 3 April 2010; received in revised form 23 April 2012; accepted 24 April 2012
Available online 3 May 2012

Abstract

The differences between self-reported and observed emotion have only marginally been investigated in the context of speech-based automatic emotion recognition. We address this issue by comparing self-reported emotion ratings to observed emotion ratings and look at how differences between these two types of ratings affect the development and performance of automatic emotion recognizers developed with these ratings. A dimensional approach to emotion modeling is adopted: the ratings are based on continuous arousal and valence scales. We describe the TNO-Gaming Corpus that contains spontaneous vocal and facial expressions elicited via a multiplayer videogame and that includes emotion annotations obtained via self-report and observation by outside observers. Comparisons show that there are discrepancies between self-reported and observed emotion ratings which are also reflected in the performance of the emotion recognizers developed. Using Support Vector Regression in combination with acoustic and textual features, recognizers of arousal and valence are developed that can predict points in a 2-dimensional arousal-valence space. The results of these recognizers show that the self-reported emotion is much harder to recognize than the observed emotion, and that averaging ratings from multiple observers improves performance.

Keywords: Affective computing; Automatic emotion recognition; Emotional speech; Emotion database; Audiovisual database; Emotion perception; Emotion annotation; Emotion elicitation; Videogames; Support Vector Regression

1. Introduction

In recent years, there has been a growing amount of research focusing on the automatic recognition of emotion in several communication modalities, e.g., face, body posture, gesture, speech etc. The ability to automatically recognize emotion in speech opens up many research opportunities and innovative applications. For conversational agents, the assessment of the emotional state in the speech of its human interlocutor is one of the key elements in achieving a humanlike conversation – vocal communication is a very natural way for humans to communicate. Further, with the increasing amount of archived speech and audio data available, the need for useful search queries grows. Searching through speech data by the emotion of the speaker is seen as a novel useful feature. Call centers have also shown interest in automatic emotion recognition systems which can be used for automated quality monitoring of incoming calls of customers. As illustrated with these examples, talking is one of the most natural interaction channels for people and as such, many innovative voice-based applications can be targeted. Hence, we focus here on the vocal modality.

We can identify several major challenges in the affect recognition research community. How to obtain reliable

* Corresponding author.
E-mail addresses: k.p.truong@utwente.nl (K.P. Truong), d.vanleeuwen@let.ru.nl (D.A. van Leeuwen), f.m.g.dejong@utwente.nl (F.M.G. de Jong).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.04.006

emotion annotations of spontaneous emotional behavior is one of these major challenges. The automatic recognition of *non-prototypical* emotions is another one. This paper addresses these two issues by exploring self-reported emotion ratings, i.e., annotation of emotions by the person who has undergone the emotion him/herself, and by adopting continuous arousal and valence dimension to model non-prototypical emotions. For these purposes, spontaneous audiovisual data was collected through a gaming scenario. Using this data, recognizers were trained with acoustic and lexical features in order to recognize scalar values of arousal and valence.

There is a vast amount of literature available on the modeling of emotional speech (e.g., Williams and Stevens, 1972; Banse and Scherer, 1996) in the speech community. The studies described in this literature usually assume emotion models and descriptions adopted from psychology research. Stemming from Darwin and made popular by researchers such as Ekman and colleagues, the most basic and classical approach to emotion modeling is the use of discrete emotion categories. Ekman (1972) and Ekman and Friesen (1975) applied this approach to the description of facial expressions and proposed six basic emotions ('the big six') that can be assumed universal: happiness, sadness, surprise, fear, anger, and disgust. As an alternative to this theory based on discrete emotions, a dimensional theory of emotion is available which was first described and applied by Wundt (1874/1905) and Schlosberg (1954). In the dimensional approach, emotions are described as points in a multidimensional space. The two main dimensions in this space are the valence dimension (pleasantness ranging from positive to negative) and the arousal dimension (activity ranging from active to passive). Sometimes, a third dimension is used which usually represents the dominance or power dimension. As a third alternative to discrete and dimensional theories of emotion, several researchers (Scherer, 2010) have developed a cognitive approach to emotion. For example, Scherer and colleagues have proposed an appraisal model called the Compoment Process Model. The main assumption here is that an emotion is a reaction (e.g., physiological, feeling) to certain antecedent situations and events that are being evaluated at the cognitive level by the human. In other words, the appraisal (i.e., the evaluation process) of a situation determines how the human is going to react/response to this situation. Componential models emphasize the link between the elicitation of emotion and the response, and as such, these models account for the variability of different emotional responses to the same event that may occur.

One of the attractions of the dimensional approach is that it allows for more flexibility and generality since it provides a way of describing emotions without the use of linguistic descriptors that can be language or culture dependent. Finding category labels to capture every shade of emotion, that frequently occur in everyday daily life, has appeared to be difficult (e.g., Cowie and Cornelius, 2003; Douglas-Cowie et al., 2005). Traditionally, speech-based emotion recognition studies have concentrated on the recognition of discrete emotion categories containing stereotypical emotions. Some of the relevant work include e.g., Batliner et al. (2000), Dellaert et al. (1996), Polzin and Waibel (1998), Petrushin (1999), Devillers et al. (2003), Kwon et al. (2003), Ang et al. (2002), Lee et al. (2002), Liscombe et al. (2003), Nwe et al. (2003), Schuller et al. (2003) and Ververidis and Kotropoulos (2005). Typical emotion categories in these studies are happy, anger, and neutral. Good overviews of these emotion recognition studies can be found in (Cowie et al., 2001; Ververidis and Kotropoulos, 2006). More recently, an increasing number of studies that adopt a dimensional approach to emotion recognition can be observed. Representing (everyday) emotion on a continuous scale could better capture different shades of emotion. Hence, describing emotion by their coordinates in a multidimensional space offers an attractive alternative, especially for computational modeling of emotion. Usually, two dimensions are sufficient to cover the emotions under investigation, where one dimension represents valence and the other dimension represents arousal. Russell (1980) and Schlosberg (1954) have shown that a third dimension, i.e., dominance or power, accounts for only a small proportion of the variance. Hence, the majority of studies have only targeted arousal and valence modeling of emotion. However, one should keep in mind that some information is always lost when mapping to a 2-dimensional emotion space. We give an overview of studies adopting a dimensional approach to emotion recognition in Section 2.

In a slower tempo, progress is also being made in designing procedures for annotation of spontaneous emotion corpora which lead to higher levels of agreement among human labelers and which better reflect the spontaneous nature of the emotion. Emotion annotation is a complex and hard process performed by humans of which the results can have significant impact on the system's performance. Emotion recognition systems need somewhat consistent emotion-labeled data for training and testing. However, it is well-known that the perception of emotion is to a certain extent subjective and person-dependent. In order to deal with this person-dependency and to reach a certain consensus on a specific emotion label, it is common to use several annotators and apply majority voting, i.e., the emotion class with the most 'votes' from the annotators wins (e.g. Batliner et al., 2006). For continuous dimensional annotations, the continuous ratings are usually averaged among the human labelers, see Mower et al. (2009), Truong et al. (2009) and Grimm et al. (2007a). In addition, in order to deal with 'mixed' or 'blended' emotions, which are not uncommon in spontaneous expressive interaction, multi-layered annotation schemes have been proposed (see Devillers et al., 2005). Less attention has been paid in emotion recognition studies to investigate how the annotations from different types of annotators compare to each other. For instance, one could compare annotations from trained emotion labelers to annotations from unexperienced/naïve emotion labelers. Another option is to let the