Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language xxx (2014) xxx–xxx

ELSEVIER

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech☆

Houwei Cao [a,*], Ragini Verma [a], Ani Nenkova [b]

[a] *Department of Radiology, Section of Biomedical Image Analysis, University of Pennsylvania, 3600 Market Street, Suite 380, Philadelphia, PA 19104, United States*

[b] *Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut Street, Philadelphia, PA 19104, United States*

Received 17 April 2013; received in revised form 15 October 2013; accepted 27 January 2014

Abstract

We introduce a ranking approach for emotion recognition which naturally incorporates information about the general expressivity of speakers. We demonstrate that our approach leads to substantial gains in accuracy compared to conventional approaches. We train ranking SVMs for individual emotions, treating the data from each speaker as a separate query, and combine the predictions from all rankers to perform multi-class prediction. The ranking method provides two natural benefits. It captures speaker specific information even in speaker-independent training/testing conditions. It also incorporates the intuition that each utterance can express a mix of possible emotion and that considering the degree to which each emotion is expressed can be productively exploited to identify the dominant emotion. We compare the performance of the rankers and their combination to standard SVM classification approaches on two publicly available datasets of acted emotional speech, Berlin and LDC, as well as on spontaneous emotional data from the FAU Aibo dataset. On acted data, ranking approaches exhibit significantly better performance compared to SVM classification both in distinguishing a specific emotion from all others and in multi-class prediction. On the spontaneous data, which contains mostly neutral utterances with a relatively small portion of less intense emotional utterances, ranking-based classifiers again achieve much higher precision in identifying emotional utterances than conventional SVM classifiers. In addition, we discuss the complementarity of conventional SVM and ranking-based classifiers. On all three datasets we find dramatically higher accuracy for the test items on whose prediction the two methods agree compared to the accuracy of individual methods. Furthermore on the spontaneous data the ranking and standard classification are complementary and we obtain marked improvement when we combine the two classifiers by late-stage fusion.

Keywords: Emotion classification; Ranking models; Spontaneous speech; Acted speech; Speaker-sensitive

1. Introduction

Research on emotion recognition from cues expressed in human voice has a long-standing tradition (Cowie et al., 2000; Ververidis and Kotropoulos, 2006). The urgency for developing accurate methods for emotion recognition has

☆ This paper has been recommended for acceptance by S. Narayanan.

* Corresponding author. Tel.: +1 5302676972.
E-mail address: Houwei.Cao@uphs.upenn.edu (H. Cao).

0885-2308/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2014.01.003

become even greater with the wide-spread use of interactive voice systems in call centers (Petrushin, 1999; Lee et al., 2002; Yu et al., 2004), car navigation systems (Fernandez and Picard, 2003), education (Litman and Forbes-Riley, 2004) and human–robot interaction (Steidl, 2009). There is also increasing interest in incorporating emotion in search over audio content, which has motivated work on emotion prediction in talk shows (Grimm et al., 2007b) and movies (Giannakopoulos et al., 2009).

The traditional paradigm of emotion recognition in speech is to extract acoustic features from the speech signal, then train classifiers on these representations, which when applied to a new utterance are able to determine its emotion content. A variety of pattern recognition methods have been explored for automatic emotion recognition such as gaussian mixture models (Luengo et al., 2005; Vlasenko et al., 2007; Vondra and Vich, 2009), hidden Markov models (Nwe et al., 2003; Shafran et al., 2003; Meng et al., 2007), neural network (Nicholson et al., 2000) and support vector machines (Kwon et al., 2003; Tabatabaei et al., 2007; Bitouk et al., 2010; Lee et al., 2011), regression (Grimm et al., 2007b). All of these seemingly diverse methods are designed to predict the emotion of a single test utterance in isolation. In many practical applications, however, emotion analysis is to be performed on a recording of complete conversations. In recordings of meetings, a user who was not present at a meeting may want to view only parts of the discussion in which participants expressed emotions. Similarly in telephone and broadcast conversations or political debates and speeches a user may want to identify parts where emotion was expressed. In all these scenarios multiple utterances from the same speaker are available and an emotion detection system could make use of this information. In such case, the state-of-the-art classification methods produce a classification score for each test utterance that does not take into account any potentially beneficial information from other utterances present in the test set. Deciding for example if an utterance expresses anger may be easier if the decision is made with respect to a larger set of utterances conveying a variety of emotions.

Ranking approaches to the task of emotion recognition offer a way of sorting all utterances in a given sample of speech from the same speaker with respect to the degree with which they convey a particular emotion. The benefit from modeling the extent to which all possible emotions are expressed in the utterance has been documented on work on emotion profiles (Mower et al., 2011), where each utterance is characterized by the distance from the hyperplane for several binary emotion classifiers. The benefit from incorporating speaker information has been confirmed by a number of studies which show that emotion recognition is higher when the same speaker is present in both training and testing. No prior work however has shown how to exploit user-specific information when prediction is done on speakers not previously seen in training. Ranking approaches seamlessly incorporate both of these desirable properties.

In this paper we show that ranking approaches lead to considerable and consistent improvement of prediction accuracy compared to conventional classification on both acted and natural spontaneous emotional speech. We use ranking SVM for our analysis (Joachims, 2002) and rely on a standard large set of acoustic features which we describe in Section 4. First we carry out experiments on two publicly available datasets of acted emotional speech which we introduce in Section 2, and report results from speaker independent analysis using leave-one-subject-out evaluation paradigm in Section 5. In Section 6 we further evaluated the proposed ranking models on the more challenging task of spontaneous emotion recognition. We also discuss the complementarity of conventional SVM classifiers and ranking-based models and further investigate the combination of the two classifiers in Section 7.

2. Corpora of emotional speech

In this study we experiment with ranking approaches on three publicly available datasets: the Berlin emotional speech database of German emotional speech (Burkhardt et al., 2005), the LDC emotional speech database of English emotional utterances (Linguistic Data Consortium, 2002), and the FAU Aibo emotional database (Steidl, 2009). The Berlin and LDC datasets consist of acted utterances, rendered by actors to convey some target emotions. The FAU Aibo database contains realistic spontaneous emotional speech from recording of children interacting with an Aibo robot.

2.1. Berlin emotional speech database

The Berlin dataset contains Recordings of 10 native German actors (5 female/5 male), expressing in German each of the following seven emotions: *anger, disgust, fear, happy, neutral, sadness, boredom*. Each actor was asked to speak one of the 10 pre-selected sentences which were chosen to maximize the number of vowels. In our experiments, we used 454 emotional utterances corresponding to the six basic emotions (Cowie et al., 2000), which are represented in