Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 53 (2011) 1198–1209

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Anger recognition in speech using acoustic and linguistic cues

Tim Polzehl [a,b,*], Alexander Schmitt [c,d], Florian Metze [e], Michael Wagner [f]

[a] *Quality and Usability Lab, Technische Universität Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany*
[b] *Quality and Usability Lab, Deutsche Telekom Laboratories, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany*
[c] *Dialogue Systems Group, University of Ulm, Albert-Einstein-Allee 43, D-89081 Ulm, Germany*
[d] *Institute of Information Technology, University of Ulm, Albert-Einstein-Allee 43, D-89081 Ulm, Germany*
[e] *Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*
[f] *National Centre for Biometric Studies, University of Canberra, ACT 2601, Australia*

Available online 12 May 2011

Abstract

The present study elaborates on the exploitation of both linguistic and acoustic feature modeling for anger classification. In terms of acoustic modeling we generate statistics from acoustic audio descriptors, e.g. pitch, loudness, spectral characteristics. Ranking our features we see that loudness and MFCC seem most promising for all databases. For the English database also pitch features are important. In terms of linguistic modeling we apply probabilistic and entropy-based models of words and phrases, e.g. Bag-of-Words (*BOW*), Term Frequency (*TF*), Term Frequency – Inverse Document Frequency (*TF.IDF*) and the Self-Referential Information (*SRI*). SRI clearly outperforms vector space models. Modeling phrases slightly improves the scores. After classification of both acoustic and linguistic information on separated levels we fuse information on decision level adding confidences. We compare the obtained scores on three different databases. Two databases are taken from the IVR customer care domain, another database accounts for a WoZ data collection. All corpora are of realistic speech condition. We observe promising results for the IVR databases while the WoZ database shows lower scores overall. In order to provide comparability between the results we evaluate classification success using the f1 measurement in addition to overall accuracy figures. As a result, acoustic modeling clearly outperforms linguistic modeling. Fusion slightly improves overall scores. With a baseline of approximately 60% accuracy and .40 f1-measurement by constant majority class voting we obtain an accuracy of 75% with respective .70 f1 for the WoZ database. For the IVR databases we obtain approximately 79% accuracy with respective .78 f1 over a baseline of 60% accuracy with respective .38 f1.

Keywords: Emotion detection; Anger classification; Linguistic and prosodic acoustic modeling; IGR ranking; Decision fusion; IVR speech

1. Introduction

Detecting emotions in vocal human-computer interaction (HCI) is gaining increasing attention in speech research. Moreover, classifying human emotions by means of automated speech analysis is achieving a level of performance, which makes effective and reliable deployment possible. Emotion detection in interactive voice response (IVR) systems can be used to monitor quality of service or to adapt emphatic dialog strategies (Yacoub et al., 2003; Shafran et al., 2003).

Anger recognition in particular can deliver useful information to both the customer and the carrier of IVR platforms. It may indicate potentially problematic turns or slots, which could in turn lead to improvements or refinements of the system. It can further serve as a trigger to switch between tailored dialog strategies for emotional conditions to better react to the user's behavior (Metze et al., 2008; Burkhardt et al., 2005a), including the

* Corresponding author at: Quality and Usability Lab, Technische Universität Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany. Tel.: +49 30 8353 58227.

E-mail addresses: tim.polzehl@gmail.com (T. Polzehl), alexander.schmitt@uni-ulm.de (A. Schmitt), fmetze@cs.cmu.edu (F. Metze), michael.wagner@canberra.edu.au (M. Wagner).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.05.002

re-routing of customers to a human operator for assistance when problems occur.

There are many ways, in which a person's emotion can be conveyed. However, in the present voice-based scenario, two factors prevail: the choice of words and acoustic variation. When a speaker expresses an emotion while adhering to an inconspicuous intonation pattern, human listeners can nevertheless perceive the emotional information through the lexical content. On the other hand, words that are not generally emotionally salient can certainly be pronounced in a way which conveys the speaker's emotion in addition to the mere lexical meaning. Consequently, our task is to capture the diverse acoustic and linguistic cues that are present in the speech signal and to analyze their correlation to the speaker's emotion.

Our linguistic approach analyzes the lexical information contained in the spoken word and its correlation to the emotion of anger. The level of anger connotation of a word can be estimated using various concepts. First, we apply the concept of Emotional Salience (Lee and Narayanan, 2005; Lee et al., 2008), which models posterior probabilities of a class given a word and combines this information with the prior probability of a class. This concept can be extended to include contextual information by modeling the salience of not just one word, but word combinations, i.e. n-grams. Further, we compare these models to traditional models from the related field of information retrieval, i.e. models that estimate term frequencies (TF) or words used (Bag-of-Words, BOW) as explained in Section 4.

Our prosodic approach examines expressive patterns that are based on vocal intonation. Applying large-scale feature extraction, we capture these expressions by calculating a number of low-level acoustic and prosodic features, e.g. pitch, loudness, MFCC, spectral information, formants and intensity. We then derive statistics from these features. Mostly, the statistics encompass moments, extrema, linear regression coefficients and ranges of the respective acoustic contours. In order to gain insight into the importance of our features we rank them according to their information-gain ratio. Looking at high-ranked features we report on their distribution and numbers in total, as well as in relation to each other. Only the most promising features are retained in the final feature set for acoustic classification.

In a final step, we fuse information from both linguistic and acoustic classification results to obtain a complex estimate of the emotional state of the user.

We compare our features for three different corpora. One database comprises American English IVR recordings (Schmitt et al., 2010), another contains German IVR recordings (Burkhardt et al., 2009). Both databases account for mostly adult telephony conversations with customer-care hotlines and contain a high number of different speakers. A third database comprises recordings from a Wizard of Oz (WoZ) scenario conducted with a small number of German children (Steidl et al., 2005).

2. Related work and realistic database conditions

When comparing existing studies on anger recognition, one has to be aware of the precise conditions of the underlying database design, as many of the results published hitherto are based on acted speech data. Some of these databases include sets of prearranged sentences. Recordings are usually done in studios, minimizing background noise, recording speakers (one at a time) multiple times until a desired degree of expression is reached. Real life speech does not have any of these settings.

As much as 97% accuracy has been reported for the recognition of angry utterances in a 7 class recognition test performed by humans on the TU Berlin EMO-DB (Burkhardt et al., 2005b), which is based on speech produced by German-speaking professional actors. The lexical content is limited to 10 pre-selected sentences, all of which are conditioned to be interpretable in six different emotional and a neutral-speech contexts. The recordings have wideband quality. Experiments on a subset, which featured high emotion recognition rates and high naturalness votes, both by human listeners, resulted in 92% accuracy when Schuller (2006) classified for the emotions and neutral speech automatically.

Comprising of mostly read sentences, but also some free text passages, a further anger recognition experiment was carried out on the DES database by Enberg and Hansen (1996). The accuracy for classification into 5 classes in a human anger recognition experiment resulted in 75%. All recordings are of wide band quality as well. Classifying this database automatically, Schuller (2006) reported an accuracy of 81%.

When speakers are not acting, namely when there is no professional performance, we need to rely on the impressions of a number of independent listeners. Since no agreed-upon common opinion exists on how a specific emotion 'sounds', it has become standard practice to take into account the opinion of several raters. To obtain a measurement for consistency of such ratings, an inter-labeler agreement measure is often applied. It is defined as the count of labeler agreements, corrected for chance level and divided by the maximum possible count of such labeler agreements. It should be noted that the maximum agreement also depends on the task, as for example the inter-labeler agreement in a gender recognition task is expected to be higher than that in an anger rating task. We assume that low inter-labeler agreement on the different emotion categories in the training and test data would predict a low automatic classification score, since in cases where humans are uncertain about classification, the classifier would likewise have difficulty in differentiating between the classes. Batliner et al. (2000) further analyzes emotion recognition performance degradations when comparing acted speech data, read speech data and spontaneous speech obtained from a WoZ scenario. Performances on acted speech data were much better in all considered experiments.