Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 168–189

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification

William Yang Wang [a,*], Fadi Biadsy [a], Andrew Rosenberg [b], Julia Hirschberg [a]

[a] *Department of Computer Science, Columbia University, United States*
[b] *Computer Science Department, Queens College (CUNY), United States*

Received 2 May 2011; received in revised form 22 March 2012; accepted 23 March 2012
Available online 3 April 2012

Abstract

Traditional studies of speaker state focus primarily upon one-stage classification techniques using standard acoustic features. In this article, we investigate multiple novel features and approaches to two recent tasks in speaker state detection: level-of-interest (LOI) detection and intoxication detection. In the task of LOI prediction, we propose a novel Discriminative TFIDF feature to capture important lexical information and a novel Prosodic Event detection approach using AuToBI; we combine these with acoustic features for this task using a new multilevel multistream prediction feedback and similarity-based hierarchical fusion learning approach. Our experimental results outperform published results of all systems in the 2010 Interspeech Paralinguistic Challenge – Affect Subchallenge. In the intoxication detection task, we evaluate the performance of Prosodic Event-based, phone duration-based, phonotactic, and phonetic-spectral based approaches, finding that a combination of the phonotactic and phonetic-spectral approaches achieve significant improvement over the 2011 Interspeech Speaker State Challenge – Intoxication Subchallenge baseline. We discuss our results using these new features and approaches and their implications for future research.

Keywords: Emotional speech; Paralinguistic; Speaker state

1. Introduction

Although the automatic detection of speaker state has attracted considerable interest in recent years, most studies have focused on the analysis of anger, frustration, and other classic emotions (Litman and Forbes-Riley, 2004; Liscombe et al., 2005; Devillers and Vidrascu, 2006; Ai et al., 2006; Grimm et al., 2007; Gupta and Nitendra., 2007). This focus is motivated primarily by Spoken Dialogue System (SDS) applications, such as call centers and tutoring systems, for which it would be useful to recognize a speaker state such as anger or uncertainty in order to improve the user experience as well as task performance by automatically adapting the system-controlled conversation in real time (Bhatt et al., 2004; Gupta and Nitendra., 2007). The benefit of adapting SDS to the speaker's state is shown by recent work (Forbes-Riley and Litman, 2011) that demonstrates successful deployment of a speaker state classifier in a tutoring system. However, emotional state is not the only important speaker state to recognize. More recently, there have been studies of speaker states that do not map directly to the classic or even derived emotions: studies of charismatic speech

* Corresponding author. Tel.: +1 347 226 1057.
E-mail addresses: yww@cs.cmu.edu (W.Y. Wang), biadsy@google.com (F. Biadsy), andrew@cs.qc.cuny.edu (A. Rosenberg), julia@cs.columbia.edu (J. Hirschberg).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.03.004

(Biadsy et al., 2008), of deceptive speech (Hirschberg et al., 2005), and of medical conditions such as depression or autistic disfunction (Hirschberg et al., 2010) broaden the scope of paralinguistic analysis considerably.

In this paper, we present novel approaches to two of these speaker states: level-of-interest (LOI) and degree of speaker intoxication, both topics in the Interspeech Paralinguistic Challenges. In 2010, the Interspeech Paralinguistic Challenge launched a sub-challenge to detect speaker's LOI (Schuller et al., 2010; Wang and Hirschberg, 2011). Detecting LOI in a topic, product, or person is an important task in many domains. By automatically detecting users' interest in a product or service, for example, it should be easier for sales representatives to identify potential customers. In the political domain, the automatic detection of interest could augment traditional polling activities. Also, understanding the speaker's interest in a conversation might have a significant influence on improving customer service behavior. The 2011 Interspeech Speaker State Challenge launched another sub-challenge: intoxication detection (Schuller et al., 2011; Biadsy et al., 2011), a still more critical task from the point of view of public safety in countries like the United States, where hundreds of thousands of people are the victims of drunk driving every year. A system to detect a person's level of intoxication via minimally invasive means would be able to significantly aid in the enforcement of drunk driving laws, and ultimately to save lives.

We describe our analyses of both data sets from both these Paralinguistic Challenges and compare our features and their performance on each below. In Section 2, we review previous work. In Section 3, we describe our studies of LOI, including the corpus, features and methods we employ. In Section 4, we describe the corpus, features and methods we use for the intoxicated speech studies. We then compare our results on both data sets to understand why different features and methods are better used on different types of data.

2. Related work

2.1. Level-of-interest (LOI)

Schuller et al. (2006) were among the first to study automatic LOI detection from conversational speech. They designed their task as a multiclass classification task, extracting standard acoustic features, such as Mel-Frequency-Cepstral-Coefficients (MFCC), and building a bag-of-words (BoW) vector space model for lexical modeling. When concatenating the bag-of-words feature vector with the acoustic feature vector into a single vector, they achieved good F-measures using a Support Vector Machine (SVM). However, a bag-of-words approach clearly fails to capture contextual information in utterances. For example, the BoW model might not be able to capture negation (e.g. "This product is not bad at all."). In addition, since lexical and acoustic-spectral features are extracted from different domains, a single stage linear combination may not yield optimal results. The 2010 Interspeech Paralinguistic Challenge (Schuller et al., 2010) included an LOI subchallenge, encouraging researchers from many groups to propose new features and methodologies. Each team was given the same conversational speech corpus with annotated LOI, baseline acoustic features, and two baseline results, which were obtained using one a single layer classification. The evaluation metric used for the challenge was primarily the cross correlation (CC) measure (Grimm et al., 2008), with mean linear error (MLE) also taken into consideration. The baseline was built only from acoustic features with Random-Sub-Space meta-learning using unpruned REPTrees, and the CC and MLE for training vs. development sets were 0.604 and 0.118. For the test data, CC and MLE scores of 0.421 and 0.146 were observed.

Participants in this subchallenge included Gajšek et al. (2010), who based their system on the Gaussian Mixture Models as Universal Background Model (GMM-UBM) approach, with relevance MAP (Maximum A-Posteriori) estimation for the acoustic data motivated by the success of GMM-UBM modeling in speaker identification (Reynolds et al., 2000). They achieved CC and MLE of 0.630 and 0.123 in the training vs. development condition, but CC and MLE of only 0.390 and 0.143 in test. This performance difference may have been due to the fact that different subsets of the corpus include different speakers: acoustic features alone may not be robust enough to capture the speaker variation.

Jeon et al. (2010) won the 2010 Subchallenge by including lexical and subjectivity information in the form of term frequency and a subjectivity dictionary. In addition to a linear combination of all lexical and acoustic features, they designed a hierarchical regression framework with multiple levels of combinations. Its first two combiners combine hypotheses from different acoustic classifiers and then use a final stage SVM classifier to combine the overall acoustic posteriors with lexical posteriors to form the final output. They report a result of 0.622 for CC and 0.115 for MLE. On