Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 1172–1185
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Application of speaker- and language identification state-of-the-art techniques for emotion recognition ☆

Marcel Kockmann *, Lukáš Burget, Jan "Honza" Černocký

Brno University of Technology, Speech@FIT, Czech Republic

Available online 1 February 2011

Abstract

This paper describes our efforts of transferring feature extraction and statistical modeling techniques from the fields of speaker and language identification to the related field of emotion recognition. We give detailed insight to our acoustic and prosodic feature extraction and show how to apply Gaussian Mixture Modeling techniques on top of it. We focus on different flavors of Gaussian Mixture Models (GMMs), including more sophisticated approaches like discriminative training using Maximum-Mutual-Information (MMI) criterion and InterSession Variability (ISV) compensation. Both techniques show superior performance in language and speaker identification. Furthermore, we combine multiple system outputs by score-level fusion to exploit the complementary information in diverse systems. Our proposal is evaluated with several experiments on the FAU Aibo Emotion Corpus containing non-acted spontaneous emotional speech. Within the Interspeech 2009 Emotion Challenge we could achieve the best results for the 5-class task of the Open Performance Sub-Challenge with an unweighted average recall of 41.7%. Further additional experiments on the acted Berlin Database of Emotional Speech show the capability of intersession variability compensation for emotion recognition.

Keywords: Emotion recognition; Gaussian mixture models; Maximum-mutual-information; Intersession variability compensation; Score-level fusion

1. Introduction

Spoken emotion recognition is the problem of automatically recognizing the emotional state of a person from their speech. Different moods may change the attributes of the human voice, such as pitch, speaking-rate, and intonation.

In automatic speech processing these properties are usually represented using the appropriate parametrization of speech, so called features. Pattern recognition and machine learning algorithms can then be used to model certain characteristics of emotionally colored speech and recognize emotions in speech utterances. Typically, classifiers like Hidden-Markov-Models (HMMs), Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs) or Neural Networks (NNs) (Bishop, 2006) are used.

While sensing the emotions of an individual from their speech is a relatively new research field in speech processing, a research community has formed in recent years and several methods have been applied successfully (Steidl, 2009; Vlasenko et al., 2007; Seppi et al., 2008; Batliner et al., 2006) and evaluated on special databases containing emotional speech (Ververidis and Kotropoulos, 2003).

Recently the usage of SVMs to directly model large-scale feature vectors has become the standard for emotion recognition (Schuller et al., 2007, 2009). These feature vectors contain diverse kinds of speech parametrization extracted on a per-utterance basis including acoustic, prosodic and voice quality features. Frame based features are usually modeled by HMMs to capture the temporal dynamics of the signal (Schuller et al., 2009).

Using these state-of-the-art techniques, accuracies of over 80% have been reported for emotion classification

☆ This work was partly supported by European project MOBIO (FP7-214324), by Grant Agency of Czech Republic project No. 102/08/0707, and by the Czech Ministry of Education project No. MSM0021630528. Marcel Kockmann is supported by SVOX Deutschland GmbH, Munich, Germany.

* Corresponding author.
E-mail address: kockmann@fit.vutbr.cz (M. Kockmann).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.01.007

tasks on acted non-spontaneous data (Schuller et al., 2006). However, on real life non-acted spontaneous emotionally colored data these accuracies drop drastically (below 40%) (Schuller et al., 2009).

Besides emotion recognition there are many diverse research fields with the goal of extracting certain attributes from speech. These include:

- What is spoken: Automatic Speech Recognition (ASR).
- Who is speaking: Speaker Identification (SID).
- Which language is used: Language Identification (LID).
- Which gender is the speaker: Gender identification (GID).
- What is the age of the speaker: Age Identification (AID).

In many of these fields (like SID, LID and GID) the use of Gaussian Mixture Models has established itself as the standard (Reynolds et al., 2000). HMMs, as used in ASR, are usually outperformed by GMMs (which are actually a HMM containing a single state) on text-independent tasks. Also, best results in all these fields are often obtained using more or less standard acoustic features extracted on a frame-based level, as used in ASR. This is somewhat illogical as features for ASR are optimized to blind out properties like speaker characteristics. Still, these tools seem to provide a good framework for diverse kinds of speech characterization.

As mentioned above, the state-of-the-art for emotion recognition has moved in a different direction. Gaussian mixture modeling of short-time acoustic features has been mostly replaced by Support Vector Machine classification. A similar trend was observed in the field of Speaker Verification as well. However, recent advances in Gaussian Mixture Modeling, like discriminative training or intersession variability compensation, has significantly raised the performance of GMM based systems and currently defines the state-of-the-art (Kinnunen and Li, 2010). This is the main motivation for our work. Our aim is to take basic and newly evolved features and modeling techniques, as used in current LID and SID systems and to apply them to the task of emotion recognition. By doing so we want to provide another view to the problem of emotion recognition. Further enhancement can then be expected by combining both approaches.

Through this paper, we will investigate standard spectral features based on Mel-Frequency-Cepstral-Coefficients (MFCC) (Davis and Mermelstein, 1980) as they are usually used in ASR. There have been many modifications of standard MFCC features to better fit the needs of SID and LID, like longer temporal context and speaker normalization. We will evaluate below some of these techniques for emotion recognition.

Furthermore, prosodic features (incorporating duration, pitch and energy) are often used to enhance the performance of MFCC based systems. Different from spectral features, prosodic features are usually extracted over a longer time span, like on a syllable basis. We examined a prosodic feature extraction method successfully used for GMM based speaker recognition (Kockmann and Burget, 2008).

All these features will be modeled using different flavors of Gaussian Mixture Models. It should be noted, that in all cases we model frame or syllable based features using models without any temporal dependencies. This statistical method of creating a "footprint" has been very successful. We will investigate in detail basic GMM approaches used in speaker and language identification. Furthermore, more sophisticated techniques evolved in the last few years are examined for their applicability in emotion recognition. These include discriminative training of GMMs and intersession variability compensation. Intersession variability for emotion recognition may refer to different acoustic conditions, different speakers or simply the spoken content of the utterance. All these attributes are a nuisance for the task of emotion recognition and we want to "ignore" them during modeling.

To evaluate the performance of the proposed techniques we provide experiments on two independent emotional databases, one containing non-acted spontaneous speech and the other acted non-spontaneous speech. Results on the first database include our submission to the Interspeech 2009 Emotion Challenge (Kockmann et al., 2009) where we could achieve very good results using the techniques described above.

The paper is organized as follows: Section 2 describes the acoustic features we used in our experiments while Section 3 explains the prosodic features used. Section 4 gives detailed information on the Gaussian Mixture Models we used and their training and evaluation procedures. In Sections 5, 6 we present results to evaluate the proposed approaches for emotion recognition. In Section 7 we draw conclusions to our approaches and consider future research.

2. Spectral features

This section will introduce the used MFCC features and the additional techniques applied to make them more suitable for the given task.

2.1. Basic acoustic features

The most widely used features in speech processing are MFCCs (Davis and Mermelstein, 1980). They have been applied successfully for speech recognition as well as for speaker recognition and language identification. We will use them as our basic features for the emotion recognition task. MFCC vectors are generated every 10 ms on a 20 ms frame of speech weighted by a Hamming window. Fast-Fourier-Transform (FFT) output of each speech window is processed by a Mel filter bank with 25 bands. The output is transformed by Discrete Cosine Transform (DCT) and