Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 25 (2011) 556–570
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Spoken emotion recognition using hierarchical classifiers ☆

Enrique M. Albornoz [a,b,*], Diego H. Milone [a,b], Hugo L. Rufiner [a,b,c]

[a] *Centro de I+D en Señales, Sistemas e INteligencia Computacional (SINC(i)), Fac. de Ingeniería y Cs. Hídricas, Univ. Nacional del Litoral, Santa Fe, Argentina*
[b] *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina*
[c] *Laboratorio de Cibernética, Facultad de Ingeniería, Univ. Nacional de Entre Ríos, Entre Ríos, Argentina*

Received 13 April 2010; received in revised form 1 October 2010; accepted 6 October 2010
Available online 26 October 2010

Abstract

The recognition of the emotional state of speakers is a multi-disciplinary research area that has received great interest over the last years. One of the most important goals is to improve the voice-based human–machine interactions. Several works on this domain use the prosodic features or the spectrum characteristics of speech signal, with neural networks, Gaussian mixtures and other standard classifiers. Usually, there is no acoustic interpretation of types of errors in the results. In this paper, the spectral characteristics of emotional signals are used in order to group emotions based on acoustic rather than psychological considerations. Standard classifiers based on Gaussian Mixture Models, Hidden Markov Models and Multilayer Perceptron are tested. These classifiers have been evaluated with different configurations and input features, in order to design a new hierarchical method for emotion classification. The proposed multiple feature hierarchical method for seven emotions, based on spectral and prosodic information, improves the performance over the standard classifiers and the fixed features.

Keywords: Emotion recognition; Spectral information; Hierarchical classifiers; Hidden Markov Model; Multilayer Perceptron

1. Introduction

In human interactions there are many ways in which information is exchanged (speech, body language, facial expressions, etc.). A speech message in which people express ideas or communicate has a lot of information that is interpreted implicitly. This information may be expressed or perceived in the intonation, volume and speed of the voice and in the emotional state of people, among others. The speaker's emotional state is closely related to this information, and this motivates its study. Two antagonistic ideas on the origin of emotions exist. One of these explains emotions from evolutionary psychology and the other as socially constructed (Prinz, 2004). The second theory claims that emotions are generated by society, and they find a different support in each culture. In evolutionary theory, it is widely accepted the "basic" term to define some universal emotions. The most popular set of basic emotions is the *big six*: happiness

☆ This paper has been recommended for acceptance by Dr. Zheng-Hua Tan.

* Corresponding author at: Univ. Nacional del Litoral CC 217, C. Universitaria, Paraje El Pozo, S3000 Santa Fe, Argentina. Tel.: +54 342 457 5233/39/44/45x191; fax: +54 342 457 5224.
E-mail addresses: emalbornoz@fich.unl.edu.ar (E.M. Albornoz), d.milone@ieee.org (D.H. Milone), lrufiner@fich.unl.edu.ar (H.L. Rufiner).

0885-2308/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2010.10.001

(joy), anger, fear, boredom, sadness, disgust and neutral. Ekman et al. (1969) researched it to argue in favour of emotion innateness and universality.

Over the last years the recognition of emotions has become a multi-disciplinary research area that has received great interest. This plays an important role in the improvement of human–machine interaction. Automatic recognition of speaker emotional state aims to achieve a more natural interaction between humans and machines. Also, it could be used to make the computer act according to the actual human emotion. This is useful in various real life applications as systems for real-life emotion detection using a corpus of agent-client spoken dialogues from a medical emergency call centre (Devillers and Vidrascu, 2007), detection of the emotional manifestation of fear in abnormal situations for a security application (Clavel et al., 2008), support of semi-automatic diagnosis of psychiatric diseases (Tacconi et al., 2008) and detection of emotional attitudes from child in spontaneous dialog interactions with computer characters (Yildirim et al., 2011). On the other hand, considering the other part of a communication system, progress was made in the context of speech synthesis too (Murray and Arnott, 2008).

The use of biosignals (such as ECG, EEG, etc.), face and body images are an interesting alternative to detect emotional states (Kim and André, 2008; Schindler et al., 2008; Vinhas et al., 2009). However, methods to record and use these signals are more invasive, complex and impossible in certain real applications. Therefore, the use of speech signals clearly becomes a more feasible option. Most of the previous works on emotion recognition have been based on the analysis of speech prosodic features and spectral information (Dellaert et al., 1996; Noguerias et al., 2001; Borchert and Dusterhoft, 2005; Luengo Gil et al., 2005; Batliner et al., 2011). Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Multilayer Perceptron (MLP) and several other one-level standard techniques have been explored for the classifier (Iliev et al., 2010; Albornoz et al., 2008; Lin and Wei, 2005; El Ayadi et al., 2007; Rong et al., 2007). Good results are obtained by standard classifiers but their performance improvement could have reached a limit. Fusion, combination and ensemble of classifiers could represent a new step towards better emotion recognition systems.

In last years some works using a combination of standard methods have been presented. A fusion scheme where a combination of results at the decision-level based on the outputs of separate classifiers (trained with different types of features) is proposed in Kim (2007). In Truong and van Leeuwen (2007), a similar idea in order to distinguish between laughter and speech is proposed. In this work, two ways to combine classifier outputs are presented: a linear combination of the outputs of independent classifiers and a second-level classifier trained with the outputs from a fixed set of independent classifiers. Two classification methods (stacked generalization and unweighted vote) were applied to emotion recognition of 6 emotional classes in Morrison et al. (2007). These classifiers improved modestly the performance of traditional classification methods, with recognition rates of 73.29% and 72.30%, respectively. In Schuller et al. (2004), a multiple stage classifier with support vector machine (SVM) is presented. Two-class decisions are repetitively made until only one class remains and hardly separable classes are divided at last. Authors built this partition based on expert knowledge or derived it from confusion matrices of a multiclass SVM approach. They reported an accuracy of 81.19% with 7 emotional classes. A two-stage classifier for five emotions is proposed in Fu et al. (2008) and the recognition rate reaches 76.1%. In this work, a SVM to classify five emotions into two groups is used. Then, HMMs are used to classify emotions within each group. In Lee et al. (2009), Bayesian logistic regression and SVM classifiers in a binary decision tree are used. They reported 48.37% of unweighted recall on 5 emotional classes. The order of the classification at each layer of binary classification is motivated by appraisal theory of emotions (Lazarus, 2001). A binary multi-stage classifier guided by the dimensional emotion model is proposed in Xiao et al. (2009). They used six emotion states from the Berlin dataset and reported a classification rate of 68.60%. A true comparison among the results of all the previously mentioned methods is very difficult because they have used different corpus, training/test partitions, etc. Therefore, none of these results can be a baseline for direct comparison with our work and our own baselines are proposed. This will be discussed later.

A simple analysis of the output probability distribution of the HMM states obtained for different emotions is made in Wagner et al. (2007). However, the reasons for success and failure in confusion matrices are not usually analyzed. For example, in Schuller et al. (2004) and Fu et al. (2008) clustering was done based on confusion matrices of standard classifiers, expert knowledge or the goodness of SVM. In the present work, an analysis of spectral features is made in order to characterize emotions and to define groups. Emotions are grouped based on their acoustical features and a hierarchical classifier is designed. The emotions which are acoustically more similar agree with the emotions that are the most difficult to distinguish, as it can be seen in the confusion matrices reported in previous works (Noguerias et al., 2001; Albornoz et al., 2008; Borchert and Dusterhoft, 2005; El Ayadi et al., 2007). The proposed classifier is