Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 727–742

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Class-specific multiple classifiers scheme to recognize emotions from speech signals☆

A. Milton[a,*], S. Tamil Selvi[b,1]

[a] Department of Electronics and Communication Engineering, St. Xavier's Catholic College of Engineering, Chunkankadai 629003, India
[b] Department of Electronics and Communication Engineering, National Engineering College, Kovilpatti 628503, India

Received 1 October 2012; received in revised form 17 August 2013; accepted 23 August 2013
Available online 4 September 2013

Abstract

Automatic emotion recognition from speech signals is one of the important research areas, which adds value to machine intelligence. Pitch, duration, energy and Mel-frequency cepstral coefficients (MFCC) are the widely used features in the field of speech emotion recognition. A single classifier or a combination of classifiers is used to recognize emotions from the input features. The present work investigates the performance of the features of Autoregressive (AR) parameters, which include gain and reflection coefficients, in addition to the traditional linear prediction coefficients (LPC), to recognize emotions from speech signals. The classification performance of the features of AR parameters is studied using discriminant, *k*-nearest neighbor (KNN), Gaussian mixture model (GMM), back propagation artificial neural network (ANN) and support vector machine (SVM) classifiers and we find that the features of reflection coefficients recognize emotions better than the LPC. To improve the emotion recognition accuracy, we propose a class-specific multiple classifiers scheme, which is designed by multiple parallel classifiers, each of which is optimized to a class. Each classifier for an emotional class is built by a feature identified from a pool of features and a classifier identified from a pool of classifiers that optimize the recognition of the particular emotion. The outputs of the classifiers are combined by a decision level fusion technique. The experimental results show that the proposed scheme improves the emotion recognition accuracy. Further improvement in recognition accuracy is obtained when the scheme is built by including MFCC features in the pool of features.

Keywords: Multiple classifiers; Class specific classification; Classifier fusion; Speech emotion recognition; AR parameters

1. Introduction

Speech is the primary form of communication among human beings. Besides understanding the meaning of speech, the human has the natural ability to estimate the gender, age, speaker and the emotional state of the speaker. The emotional state of a speaker plays an important role because it shapes the real meaning of the spoken language. Emotional states of humans are expressed through changes in speech, facial expression, posture and physiological processes. Recognizing the emotional states by these changes can help us estimate the belief, desire and the likely future behavior of a person (Gratch et al., 2009). Emotion is integral to man's rational and intelligent decisions and

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +91 944 260 2309; fax: +91 465 223 3982.
E-mail addresses: milton@sxcce.edu.in, milton_sxcce@yahoo.com (A. Milton), tamilgopal2004@yahoo.co.in (S. Tamil Selvi).
[1] Tel.: +91 9443721864.

0885-2308/$ – see front matter
http://dx.doi.org/10.1016/j.csl.2013.08.004

expresses feelings and provides feedback (Busso et al., 2009). Anger, boredom, disgust, fear, happiness, sadness and neutral are considered as the seven basic discrete emotions (Yang and Lugger, 2010). The different emotional states of a speaker are associated with different heart rate, skin resistivity, temperature, papillary diameter and muscle activities. These changes result in the production of speech signals that carry the emotional information, making automatic detection of emotions from speech signals possible (Cowie et al., 2001). In the era of increasing human-machine interaction, detection of emotions can make intelligent machines create and understand emotions, like humans. In speech recognition and speaker identification applications, emotions are considered as noise. Therefore the recognition of emotions and its effect on the speech signals can improve the performance of speech and speaker recognition systems. Fear type emotion recognition can be used in audio-based surveillance system (Clavel et al., 2008) in order to gain control of a critical situation. Emotion recognition also finds its application in forensic data analysis and clinical diagnosis.

The four major areas of emotion recognition from speech signals are: acquisition and validation of emotional speech signals, feature extraction, feature selection and classification. In feature extraction and selection, the research community is looking for features which can identify and differentiate one emotion from others. For the features to be universal, it is preferred to be independent of the speaker, language, gender and culture. As the hunt for the best features is not yet complete, some researchers improve the classification accuracy by combining several classifiers (Lee and Narayanan, 2005; Morrison et al., 2007; Albornoz et al., 2011; López-Cózar et al., 2011).

The spectral features MFCC and LPC are the traditional features in automatic speech recognition (Bou-Ghazale and Hansen, 2000). MFCC is also the most investigated spectral feature in automatic speech emotion recognition (Schuller et al., 2011) unlike LPC. But a very few researchers like Nicholson et al. (2000) and Altun and Polat (2009) have used 12th order LPC and 16th order LPC, respectively, as one of the speech features to recognize emotions from speech signals. LPC are the denominator coefficients of an AR model transfer function. AR modeling is a technique which estimates the all-pole transfer function of a system that produces the observed signal. The AR model of speech signal is widely used to estimate pitch, formants, spectra and vocal tract area function. It can accurately model the speech producing system from the vocal cords to the lips as an all-pole linear system. If the number of poles is high enough, the all-pole model can represent the voiced, nasal and fricative of speech signals (Rabiner and Schafer, 2004). Motivated by the performance of LPC in speech recognition applications and its accuracy in modeling the transfer function of speech production system, the present work investigates the emotion classification performance of the features of AR parameters. Instead of concentrating on a single order AR parameters, the paper examines AR parameters of different orders from 3 to 25.

In AR model the combined spectral effect of glottal excitation, vocal tract and radiation are represented by a system function of order P

$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_P(k) Z^{-k}} \tag{1}$$

where G is the gain parameter and $a_p(k)$ are the LPC.

$$G = \sqrt{E_P} \tag{2}$$

where E_P is the minimum prediction error.

The LPC are estimated by a recursive procedure like Levinson–Durbin algorithm in the autocorrelation method (Makhoul, 1975). In the process of estimation of the AR model parameters of order P, we get the reflection coefficients $K_i = a_i(i)$, $i = 1, 2, \ldots, P$ and the prediction error E_P. In the AR model of order P, the vocal tract is considered as an acoustic tube with P sections, each reflection coefficient is an indication of the ratio of the cross sectional area of the consecutive sections of the acoustic tube (Makhoul, 1975). Therefore, reflection coefficients may carry more emotional information than LPC. The relation between the reflections coefficients and the section-wise cross sectional area of the vocal tract motivates us to examine the reflection coefficients. In this paper the collection of LPC, gain parameter and reflection coefficients are referred to as AR parameters.

The classification analysis of the features of the AR parameters with standard classifiers shows that a specific feature vector of the AR parameters, coupled with a specific classifier, recognizes a specific emotion with a high recognition rate. This activates us to propose a classification method that makes use of the specific emotion recognition potential of a specific feature vector of the AR parameters with a specific classifier to improve the overall recognition accuracy. The proposed classification method classifies the emotions at two levels. At the first level, an ensemble of class-specific