Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 1186–1197
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Formant position based weighted spectral features for emotion recognition

Elif Bozkurt [a], Engin Erzin [a,*], Çiğdem Eroğlu Erdem [b], A.Tanju Erdem [c]

[a] Multimedia, Vision and Graphics Laboratory, College of Engineering, Koç University, 34450 Sariyer, Istanbul, Turkey
[b] Department of Electrical and Electronics Engineering, Bahçeşehir University, 34353 Beşiktaş, Istanbul, Turkey
[c] Department of Electrical and Electronics Engineering, Özyeğin University, 34662 Üsküdar, Istanbul, Turkey

Available online 6 May 2011

Abstract

In this paper, we propose novel spectrally weighted mel-frequency cepstral coefficient (WMFCC) features for emotion recognition from speech. The idea is based on the fact that formant locations carry emotion-related information, and therefore critical spectral bands around formant locations can be emphasized during the calculation of MFCC features. The spectral weighting is derived from the normalized inverse harmonic mean function of the line spectral frequency (LSF) features, which are known to be localized around formant frequencies. The above approach can be considered as an early data fusion of spectral content and formant location information. We also investigate methods for late decision fusion of unimodal classifiers. We evaluate the proposed WMFCC features together with the standard spectral and prosody features using HMM based classifiers on the spontaneous FAU Aibo emotional speech corpus. The results show that unimodal classifiers with the WMFCC features perform significantly better than the classifiers with standard spectral features. Late decision fusion of classifiers provide further significant performance improvements.

Keywords: Emotion recognition; Emotional speech classification; Spectral features; Formant frequency; Line spectral frequency; Decision fusion

1. Introduction

Emotion-sensitive machine intelligence is a basic requirement for more natural human–computer interaction. In this sense, the orientation of research on emotional speech processing shifts from the analysis of acted towards spontaneous speech for advanced real-life applications in human–machine interaction systems (Ververidis and Kotropoulos, 2006; Batliner et al., 2003). The wide use of telecommunication services and multimedia devices will require human-centered designs instead of computer centered ones. Consequently, accurate perception of the user's affective state by computer systems will be crucial for the interaction process (Zeng et al., 2009). Examples of recent emotion-aware systems include call-center applications (Lee and Narayanan, 2005; Neiberg and Elenius, 2008; Morrison et al., 2007), intelligent automobile systems (Schuller et al., 2006) and interactive movie systems (Nakatsu et al., 2000).

Although extensively investigated, automatic emotion recognition from speech remains as an open problem in the field of human–computer interaction. Researchers mostly focus on defining a universal set of features that carry emotional clues and try to develop classifiers that efficiently model these features. Some commonly used speech features for emotion recognition are the mel-frequency cepstral coefficients (MFCC) (Vlasenko et al., 2007; Grimm et al., 2006), the fundamental frequency (F0, pitch), which has been referred as one of the most important features for determining emotion in speech (Nakatsu et al., 2000; Polzin and Waibel, 2000; Lee et al., 2004), and the resonant frequencies of the vocal tract, also known as formants (Nakatsu et al., 2000, 2004).

* Corresponding author.
E-mail addresses: ebozkurt@ku.edu.tr (E. Bozkurt), eerzin@ku.edu.tr (E. Erzin), cigdem.eroglu@bahcesehir.edu.tr (Ç. Eroğlu Erdem), tanju.erdem@ozyegin.edu.tr (A.T. Erdem).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.04.003

The contributions and scope of the paper can be stated under three items: (i) The main contribution is the introduction of novel spectrally weighted mel-frequency cepstral coefficient (WMFCC) features for emotion recognition from speech. Recently, Goudbeek et al. (2009) reported that emotion has a considerable influence on formant positioning. Based on this information, we propose WMFCC features by emphasizing spectral content of the critical spectral bands around formant locations. The spectral weighting is obtained from the normalized inverse harmonic mean function of line spectral frequency (LSF) features. Experimental results demonstrate the superiority of the proposed WMFCC features over traditional MFCC features. (ii) We experimentally evaluate various topologies of hidden Markov model (HMM) classifiers using different spectral and prosody features to gain insight about possible temporal patterns existing in certain feature sets for emotion recognition from speech. (iii) We evaluate the use of decision fusion methods to combine various classifiers with uncorrelated features of emotional speech. It is well-known that in classification systems, data fusion is effective when modalities are correlated, and late fusion is optimal when modalities are uncorrelated (Sargin et al., 2007). Experimental results show that decision fusion of classifiers is beneficial.

We evaluate the proposed WMFCC features and the combined classifiers on the spontaneous emotional speech corpus FAU Aibo (Steidl, 2009), which is an elicited corpus with clearly defined testing and training partitions ensuring speaker-independence and different room acoustics as in real-life. We achieve significant performance improvements over the best scoring emotion recognition systems in the Interspeech 2009 Emotion Challenge (Schuller et al., 2009) with the proposed WMFCC features and the decision fusion of classifiers. Furthermore, we observe evidence of temporal formant patterns in discriminating emotion related classes of speech signal.

The remainder of this paper is structured as follows. Section 2 defines the components of the proposed emotion recognition system. The employed spectral and prosody features together with the proposed WMFCC features are presented in Section 2.1. Section 2.2 defines the HMM based classifier for emotion recognition, and Section 2.3 presents the decision fusion method for HMM based classifiers. Experiments to assess the performance of the proposed system are discussed in Section 3. Finally, the concluding remarks are presented in Section 4.

2. Proposed system

A block diagram of the proposed automatic speech driven emotion recognition system is given in Fig. 1. This system consists of three main blocks: feature extraction, classification and late fusion of classifiers. The feature extraction block computes prosodic and spectral features including the proposed WMFCC features. The classification block includes HMM based classifiers. HMM based classifiers with several states are capable of modeling temporal clusters, where each state can represent a different distribution of observations. We target to capture emotion related patterns in syntactically meaningful chunks of speech segments using HMM based classifiers. Syntacti-

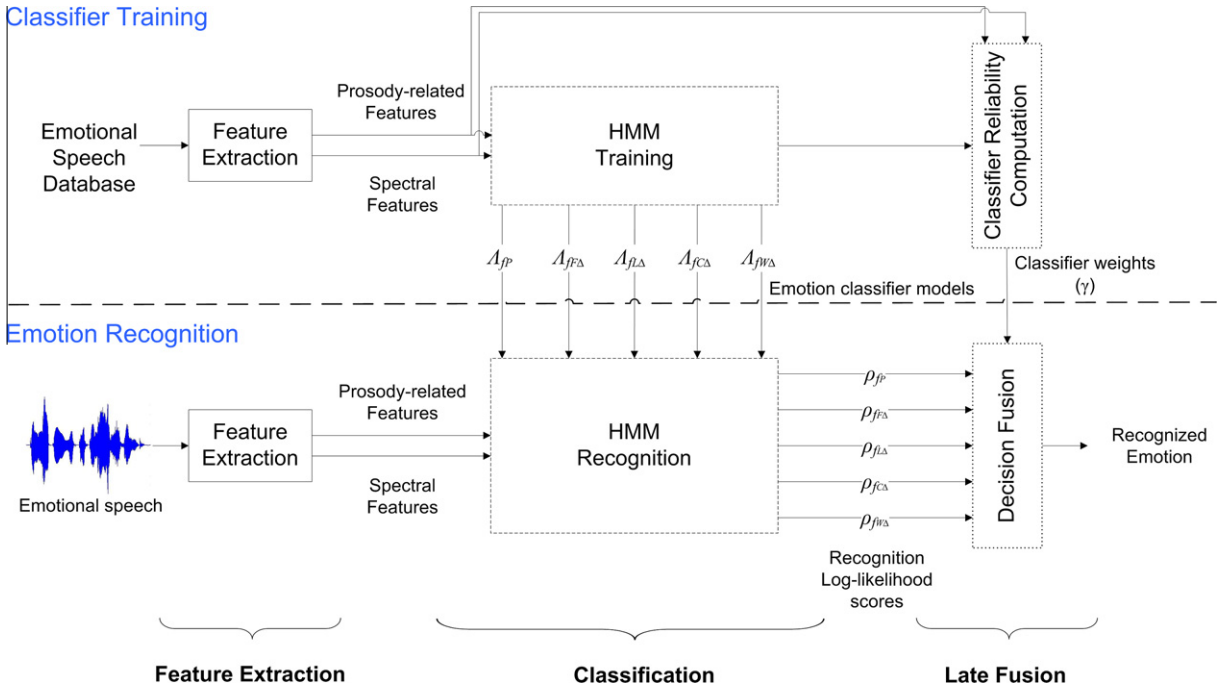

Fig. 1. Overview of the proposed emotion recognition system. The system is composed of classifier training and emotion recognition parts. Each spectral and prosody-related feature sequence, f, is used to train hidden Markov model sets, Λ_f, for all emotion classes. The highest log-likelihood scores, ρ_f, are evaluated through Viterbi decoding to be used in the decision fusion.