Chapter 15
Comparison of Several Acoustic Modeling Techniques for Speech Emotion Recognition

Imen Trabelsi
Sciences and Technologies of Image and Telecommunications (SETIT), University of Sfax, Tunisia

Med Salim Bouhlel
Sciences and Technologies of Image and Telecommunications (SETIT), University of Sfax, Tunisia

ABSTRACT

Automatic Speech Emotion Recognition (SER) is a current research topic in the field of Human Computer Interaction (HCI) with a wide range of applications. The purpose of speech emotion recognition system is to automatically classify speaker's utterances into different emotional states such as disgust, boredom, sadness, neutral, and happiness. The speech samples in this paper are from the Berlin emotional database. Mel Frequency cepstrum coefficients (MFCC), Linear prediction coefficients (LPC), linear prediction cepstrum coefficients (LPCC), Perceptual Linear Prediction (PLP) and Relative Spectral Perceptual Linear Prediction (Rasta-PLP) features are used to characterize the emotional utterances using a combination between Gaussian mixture models (GMM) and Support Vector Machines (SVM) based on the Kullback-Leibler Divergence Kernel. In this study, the effect of feature type and its dimension are comparatively investigated. The best results are obtained with 12-coefficient MFCC. Utilizing the proposed features a recognition rate of 84% has been achieved which is close to the performance of humans on this database.

INTRODUCTION

Human emotion recognition is one of the major challenges in Human-Computer interactions (Grunberg, 2012) due to its wide range of applications and complex tasks: agent-customer interactions communication, speech driven facial animation, E-tutoring applications etc.

DOI: 10.4018/978-1-7998-2460-2.ch015

The classification of emotions has been researched from two fundamental viewpoints: one that emotions are basic, distinct and fundamentally different constructs e.g. fear and anger; or two, that emotions can be characterized on a dimensional space (Cowie & McKeown & Douglas-Cowie, 2012; Hudlicka, 2011; Gunes, 2010). The speech emotion recognition systems (SER) use several types of databases of acted, simulated or real emotions. A wide range of pattern recognition methods exists. They include two powerful paradigms in machine learning: generative and discriminative methods. A generative method is a full probabilistic model for all variables, such as those based on Bayes decision theory and related techniques of density estimation. A Discriminative model learns the conditional probability distribution, such as nearest-neighbor classification, support vector machines.

Integrating generative machines learning models such as Gaussian Mixture Model (GMM) and discriminative machines learning models such as Support Vector Machines (SVM) in a hybrid system has shown great success. Favorable properties of SVM such the non-linear kernels and the inherent class-discriminative model structure represent an attractive way to enhancing GMM. Thereby, a combination between these two powerful models based on the Kullback-Leibler Divergence Kernel is presented in this paper. The Speech signal contains a large number of parameters that reflect the emotional characteristics, and the different parameters result in changes in emotion. Thus, the most important challenge in speech emotion recognition is how to determine the best feature parameters, which can express mostly the emotional characteristics of speech.A lot of research has been done in the speech parameterization area, resulting in many different feature methods. These methods can be categorized into three broad categories (1) spectral features, (2) prosodic features and (3) high-level features. These speech features can be also divided into two categories: utterance level features and frames level features (Trabelsi & Bouhlel, 2016). In this paper, frames level features are explored, in particularly the spectral features such as, the Mel Frequency cepstrum coefficients (MFCC), Linear prediction coefficients (LPC), linear prediction cepstrum coefficients (LPCC), Perceptual Linear Prediction (PLP) and Relative Spectral Perceptual Linear Prediction (Rasta-PLP) and Formants.

The paper is organized as follows. Section 2 represents the related literature. Section 3 describes the proposed method. Section 4 discusses the results, and finally Section 6 concludes the work.

Related Literature

The progress made in the area of speech emotion recognition by various group researchers so far is briefed in this section. SER using various utterance level statistical features was described by the authors (Trabelsi & bouhlel, 2016) in their paper. An extensive comparative study about the significance of the different spectral and prosodic features for SER tasks can be found in (Koolagudi & Rao, 2012). A review on various pattern recognition techniques of speech emotion recognition, features and emotional databases were explained in the paper, along with some directions for future research on SER (Iliou & Anagnostopoulos, 2009). The GMM-UBM mean interval (GUMI) kernel based on the Bhattacharyya distance combined with different feature extraction methods was successfully used on the Surrey Audio-Expressed Emotion and the Berlin Emotional speech Databases (Trabelsi & Bouhlel, 2016a). In (Zhang & Zhao, 2013), a new type of kernel, called the enhanced kernel isometric mapping, is applied for speech emotion recognition in human-robot interaction. Other types of SVM kernels were employed in (Sikka, 2013; Yang, 2012). The concept of high aroused, low aroused and neutral emotions was emphasized using hierarchical GMM model (Trabelsi & Bouhlel, 2018). Spectral features such as LPC, LPCC, Mel