Chapter 38
Feature Selection for GUMI Kernel-Based SVM in Speech Emotion Recognition

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

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

ABSTRACT

Speech emotion recognition is the indispensable requirement for efficient human machine interaction. Most modern automatic speech emotion recognition systems use Gaussian mixture models (GMM) and Support Vector Machines (SVM). GMM are known for their performance and scalability in the spectral modeling while SVM are known for their discriminatory power. A GMM-supervector characterizes an emotional style by the GMM parameters (mean vectors, covariance matrices, and mixture weights). GMM-supervector SVM benefits from both GMM and SVM frameworks. In this paper, the GMM-UBM mean interval (GUMI) kernel based on the Bhattacharyya distance is successfully used. CFSSubsetEval combined with Best first algorithm and Greedy stepwise were also utilized on the supervectors space in order to select the most important features. This framework is illustrated using Mel-frequency cepstral (MFCC) coefficients and Perceptual Linear Prediction (PLP) features on two different emotional databases namely the Surrey Audio-Expressed Emotion and the Berlin Emotional speech Database.

INTRODUCTION

Speech is the natural communication form between humans, provides a great deal of information about speaker, language and emotions. This fact has motivated researchers to find a fast and efficient method of natural interaction between man and machine. Presence of emotions makes speech more natural. This has introduced a relatively new research area, namely speech emotion recognition (SER), which is defined as extracting the emotional state of a speaker from his or her speech. This challenging task has several applications in day-to-day life like agent-customer interactions, call-center applications (Herm, 2008),

DOI: 10.4018/978-1-5225-1759-7.ch038

web movies, on- board car driving systems (Hu et al., 2013), medical diagnostic tool and E-tutoring systems (Trabelsi & Bouhlel, 2016a). As in any pattern recognition problem, the performance of emotion recognition from speech depends on label, organization, representation, and evaluation of training data. A significant challenge for emotional research depends on a sense of what emotion is and is in finding appropriate emotional labels. Three labeling methods can be distinguished: (1) categorical approach, (2) dimensional approach, and (3) appraisal-based approach (Cowie & McKeown & Douglas-Cowie, 2012; Hudlicka, 2011). In the first one, emotion is described as a discrete class that differs explicitly and mutually exclusive from one emotion to another. In the second one, emotion is described as a continuous process that will changes dynamically over time, using the multi-dimensional emotion model. However, the appraisal approach, introduces the role of time into the comprehension of emotions (Mortillaro & Meuleman & Scherer, 2012; De Vries, 2015). A critical research challenge in speech emotion recognition systems is to how to encode the spoken emotion by some suitable features (Maji et al., 2015; Saba et al., 2016). This step, called feature extraction, is of a great importance in SER. However, having a large number of potential features increases the complexity of the system and normally results in longer system training times. Therefore, a popular approach is to start with a larger set of features and then removes irrelevant data to reduce dimensionality of the training data and generate a more compact and robust feature set. Another important issue in the evaluation of an emotional speech system is the choice of emotional corpus. The existing emotional databases could be divided into three classes namely: simulated (actor), elicited (induced) and spontaneous (natural) speech databases. For more detailed description, the reader may refer to (Koolagudi& Rao, 2012).

Another important task in speech emotion recognition system is the selection of classifier. Actually, the majority of speaker recognition systems are based on statistical classification methods (Iliou & Anagnostopoulos, 2009) such as Support Vector Machine (SVM) and Gaussian Mixture Model (GMM). The GMM-SVM technique has shown to be the most effective technique for text-independent speaker recognition due to its reliable performance (Trabelsi & Ben Ayed, 2011). In this approach, a big Gaussian Mixture Model (GMM) is trained by EM (expectation-maximization) algorithm from a background training set. This background dataset is called the universal background model (UBM) which is acts as a generic effective model. Then, emotional models are estimated from UBM by MAP (maximum a posteriori) adaptation. The estimated models represent the input for the SVM classifier. The main issue here is how to design a proper SVM kernel between emotional models. In this research, this useful paradigm is utilized to recognize spoken emotional utterances from two different databases. In (Campbell et al., 2006), the authors designed a new kernel function of GMM-SVM based on the Kullback-Leibler (KL) metric. A potential drawback of this kernel is the lack of supplementary covariance information encompassed in the adapted model. In (You et al., 2015), the kernel is modified to include both covariance and weighting information in the kernel by using the Bhattacharyya distance. The kernel, called GMM-UBM mean interval (GUMI) kernel, performs better than the KL kernel in many applications of speech recognition (Bui & Kim, 2015; Trabelsi & Ben Ayed, 2012). For this purpose, the GUMI kernel is used to quantify and recognize the most emotionally salient aspects of emotional speech features. The supervector space has usually very high dimensions (thousands of parameters) and a large redundancy between these dimensions. The supervector contains too many parameters that will affect the SVM model complexity and also lower its performances. Therefore, selection of the pertinent parameters is a crucial step since it might improve the accuracies of the SER. In order to resolve the problem of dimensionality reduction of the supervector space, the paper proposes to utilize feature selection methods of the GMM supervectors. The remainder of paper is structured as follows. Section 2 describes the used speech fea-