Available online at www.sciencedirect.com

ScienceDirect

April 2017, 24(2): 1–9
www.sciencedirect.com/science/journal/10058885

The Journal of China Universities of Posts and Telecommunications

http://jcupt.bupt.edu.cn

Noisy speech emotion recognition using sample reconstruction and multiple-kernel learning

Jiang Xiaoqing[1,2], Xia Kewen[1] (✉), Lin Yongliang[1,3], Bai Jianchuan[1]

1. School of Electronic and Information Engineering, Hebei University of Technology, Tianjin 300401, China
2. School of Information Science and Engineering, University of Jinan, Jinan 250022, China
3. Information Center, Tianjin Chengjian University, Tianjin 300384, China

Abstract

Speech emotion recognition (SER) in noisy environment is a vital issue in artificial intelligence (AI). In this paper, the reconstruction of speech samples removes the added noise. Acoustic features extracted from the reconstructed samples are selected to build an optimal feature subset with better emotional recognizability. A multiple-kernel (MK) support vector machine (SVM) classifier solved by semi-definite programming (SDP) is adopted in SER procedure. The proposed method in this paper is demonstrated on Berlin Database of Emotional Speech. Recognition accuracies of the original, noisy, and reconstructed samples classified by both single-kernel (SK) and MK classifiers are compared and analyzed. The experimental results show that the proposed method is effective and robust when noise exists.

Keywords speech emotion recognition, compressed sensing, multiple-kernel learning, feature selection

1 Introduction

Complementarity exists between human's affectivity and logical thinking, so emotional information is significant to understand the real meaning in human's speech. SER is an important research field in the realization of AI [1]. Noise existing in the environment and signal processing systems influences the recognition accuracies and limits the practical applications of SER, such as intelligent customer service systems and adjuvant therapy systems for autism, where accurate recognition of emotions is needed to make a proper response. In this paper, noisy SER is studied using the combination of sample reconstruction based on compressed sensing (CS) theory and multiple-kernel learning (MKL).

In SER two essential aspects influencing the performance of the emotion recognition system are optimal feature set and effective recognition classifier.

The precision and inherent properties of speech features influence the emotional recognizability of the feature set. Noise has negative impact on the extraction of acoustic features, and attempts to cope with the noise in SER started from 2006 [2]. Schuller et al. selected feature subset from a 4k feature set to recognize emotions from noisy speech samples [3]. You et al. proposed enhanced Lipschitz embedding to reduce the influence of noise [4]. Techniques such as switching linear dynamic models and two-stage Wiener filtering etc. were also proposed to handle noisy speech for classification [5]. CS theory proposed by Donoho et al. provides promising methods to noisy speech processing [6–7]. Sparse representation in CS theory has been used in nonparametric classifier. Zhao et al. adopted the enhanced sparse representation classifier to deal with the robust SER [8]. Additionally, as the derived coefficients of noise are not sparse in any transfer domain, it is impossible to reconstruct the noise from measurements. So sparse signals contaminated by noise can be reconstructed with high quality [9]. In this paper, CS theory is utilized in the denoising of noisy speech

Received date: 29-09-2016
Corresponding author: Xia Kewen, E-mail: kwxia@hebut.edu.cn
DOI: 10.1016/S1005-8885(17)60193-6

samples through sample reconstruction. Acoustic features of the reconstructed samples are extracted and selected according to the complementary information in order to constitute robust and optimal feature subset.

SVM is one of the most effective methods in pattern recognition problems. SVM is a kernel method of finding the maximum margin hyperplane in the feature space and its performance depends on the kernel function strongly. So it is necessary to overcome the kernel dependency in designing the effective classifier with SVM. In order to improve the flexibility of kernel function, MKL is proposed and developed to derive the optimal combination of different kernels. Lanckriet et al. proposed MKL with a transduction setting for learning a kernel matrix from data. The method aimed at the optimal combination of predefined base kernels to generate a good target kernel [10]. Jin et al. proposed feature fusion method based on MKL to improve the total SER performance of clean samples. The weights of different kernels corresponding to the global and local features are given through a grid search method [11]. In this paper, MK fusion strategy of Lanckriet is adopted to improve the SVM model in a binary tree structured multi-class classifier, and the fusion coefficients of different kernels are solved by the SDP to find optimal weights of multiple kernels.

The rests of the paper are structured as the followings: Sect. 2 reviews the basic idea of CS in speech signal processing and analyzes the performance of noisy sample reconstruction. Sect. 3 introduces MKL solved by SDP. Acoustic features and feature selection are presented in Sect. 4. The performance evaluation of SER and experimental results are illustrated and analyzed in Sect. 5. Finally Sect. 6 devotes to the conclusions.

2 CS and sample reconstruction of noisy speech

CS combines sampling and compression into one step using the minimum number of measurements with maximum information. CS aims to recover sparse signal with far less than Nyquist-Shannon sampling rate, and the reconstruction can be exact under key concepts such as sparsity and restricted isometry property (RIP) [7,12].

$\boldsymbol{x}=[x(1),x(2),\ldots,x(N)]^{\mathrm{T}}$ is the signal in N-dimensional space $\mathbb{R}^N$, where N is the number of samples. $\boldsymbol{x}$ can be represented by the linear combination of N-dimensional orthogonal basis vector $\{\boldsymbol{\psi}_n \,|\, n=1,2,\ldots,N\}$. Thus $\boldsymbol{x}$ can be represented as:

$$\boldsymbol{x}=\sum_{n=1}^{N}\boldsymbol{\alpha}_n\boldsymbol{\psi}_n=\boldsymbol{\Psi}\boldsymbol{\alpha} \tag{1}$$

In Eq. (1) $\boldsymbol{\Psi}$ is the orthogonal basis matrix, also named representation matrix, $\boldsymbol{\alpha}_n=\langle \boldsymbol{x},\boldsymbol{\psi}_n\rangle$ is projection coefficient, $\boldsymbol{\alpha}$ is the projection coefficients matrix and $\boldsymbol{\alpha}=\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{x}$. It can be said that $\boldsymbol{x}$ and $\boldsymbol{\alpha}$ are the equivalent representations of the same signal with $\boldsymbol{x}$ in time domain while $\boldsymbol{\alpha}$ in $\boldsymbol{\Psi}$ domain. When the signal $\boldsymbol{x}$ only has k non-zero α_n coefficients and $k<<N$, $\boldsymbol{\psi}_n$ is the sparse basis of $\boldsymbol{x}$ and $\boldsymbol{x}$ can be considered k sparse with sparse representation of Eq. (1).

In CS theory, the sensing process can be represented as:

$$\boldsymbol{y}=\boldsymbol{\Phi}\boldsymbol{x} \tag{2}$$

In Eq. (2) $\boldsymbol{\Phi}$ is the $M\times N$ measurement matrix, and $\boldsymbol{y}\in\mathbb{R}^M$ ($M<<N$) is the measurement vector of M-dimensional. Compression is realized because the dimension of measurements $\boldsymbol{y}$ is far less than the dimension of the signal $\boldsymbol{x}$.

With Eq. (1), Eq. (2) can be rewritten as:

$$\boldsymbol{y}=\boldsymbol{\Theta}\boldsymbol{\alpha} \tag{3}$$

where $\boldsymbol{\Theta}=\boldsymbol{\Phi}\boldsymbol{\Psi}$ is $M\times N$ dimensional reconstruction matrix, and $\boldsymbol{\alpha}$ is k sparse vector representing the projection coefficients of $\boldsymbol{x}$ in $\boldsymbol{\Psi}$ domain.

Reconstruction algorithms in CS try to solve Eq. (3), which is an underdetermined equation without a determinant solution. When the signal is sparse and $\boldsymbol{\Theta}$ satisfies the RIP condition, a sparse approximation solution to Eq. (3) can be obtained by minimizing the L_1-norm. RIP of matrix $\boldsymbol{\Theta}$ is defined on isometry constant $\delta_k\in(0,1)$ for a k sparse signal $\boldsymbol{x}$ and satisfies:

$$1-\delta_k\leqslant\frac{\|\boldsymbol{\Theta}\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\leqslant 1+\delta_k \tag{4}$$

It can be loosely said that a matrix obeys the RIP of order k if δ_k is not too close to one. RIP ensures that all subsets of k columns taken from matrix are nearly orthogonal. The equivalent condition of RIP is the incoherence between measurement matrix $\boldsymbol{\Phi}$ and the representation matrix $\boldsymbol{\Psi}$. A variety of reconstruction methods such as greedy algorithms and convex optimization can be used in the solving process of Eq. (3) [13–21].

When CS theory is applied to speech signal processing, the prerequisite is to achieve the sparse representation of speech signals using proper orthogonal basis. The