Speech Emotion Recognition Based on Robust Discriminative Sparse Regression

Peng Song, Wenming Zheng, *Senior Member, IEEE* , Yanwei Yu, Shifeng Ou

***Abstract*—Speech emotion recognition has recently attracted much interest due to the widespread of multimedia data. It generally involves two basic problems: feature extraction and emotion classification. Most previous algorithms just focus on solving one of these two problems. In this paper, we aim to deal with these two problems in a joint learning framework, and present a novel regression algorithm, namely, robust discriminative sparse regression (RDSR). In RDSR, we propose a sparse regression algorithm to make our model robust to outliers and noises, and introduce a feature selection regularization constraint simultaneously to select the most discriminative and relevant features. In addition, to well predict the labels, we exploit the local and global consistency over labels, and incorporate it into the proposed framework. To solve the objective function of RDSR, we design an efficient alternative optimization algorithm. Finally, experimental results on several public emotion datasets verify the effectiveness and the superiority of our proposed method.**

***Index Terms*—Regression analysis, semi-supervised learning, feature selection, graph Laplacian, speech emotion recognition.**

I. Introduction

AUTOMATIC emotion recognition is an important research field in the area of speech signal processing. It aims at automatically recognizing human emotions from speech signals, and has a wide range of practical applications, e.g., human-computer interaction (HCI), depression and suicide risk assessment, customer satisfaction assessment in call center services [1], [2], [3].

Speech emotion recognition can be formulated as a classic pattern recognition task, which involves two basic problems: feature extraction versus emotion classification. Feature extration aims to extract useful features from speech signals. During the past decades, different types of speech features, have been presented and proven useful for emotion recognition [1], [4], [5]. Overall, the acoustic features can be roughly categorized into two types, i.e., low-level features versus high-level features. The first category refers to the features extracted using time/frequency analysis algorithms, e.g., F0, spectral features, log energy, voice quality, Teager energy operator (TEO), spectral frequency bands of speech formants [4]. The latter one refers to the high-level feature representations which are learned from low-level data using deep learning techniques [6]. One of current research directions focuses on utilizing the combination of different features to obtain a high-dimensional feature set [7], [8]. However, it usually is not good for emotion classification due to the curse of dimensionality and feature redundancy [9], [10]. High dimensionality may also increase the complexity of time and space for subsequent analyzing task. A common strategy to solve this issue is dimensionality reduction. Feature selection, which is designed to select a feature subset from the original high-dimensional feature set, is an efficient technique for dimensionality reduction. Different from other dimensionality reduction algorithms, e.g., principal component analysis (PCA), linear discriminant analysis (LDA), locality preserving projection (LPP) [11], [9], feature selection just selects a discriminative or representative subset of the original feature set, and does not alter its original representation [10], [12]. In speech emotion recognition, different types of features contribute differently [13]. Thus, feature selection can offer better understanding of the contributions of these different features to speech emotion recognition.

Emotion classification is also a critical step for speech emotion recognition. During the last decades, a variety of emotion classification methods, e.g., support vector machine (SVM) [14], Gaussian mixture model (GMM) [15], hidden Markov model (HMM) [16], artificial neural network (ANN) [17], extreme learning machine (ELM) [18], [19], decision tree [19], cooperative learning [20], multiview learning [21], sparse representation [22], and some combinations of these algorithms [1], have been successfully presented. Recently, deep learning techniques have shown extraordinary advantages in speech recognition and many visual classification tasks [23], [24], and also have been applied to speech emotion recognition [25], [26], [6], [27]. However, in reality, they are carried out on the basis of a large number of labeled data, and are not fully applicable for practical environments of small-scale and small samples [28]. Therefore, the study of emotion classification based on traditional machine learning methods, e.g., regression based algorithms, still occupies an important position.

Linear regression based classification has been widely used in supervised learning tasks, e.g., pattern classification and recognition, and has also shown its advantage in processing data with high dimensionality, such as object and face recognition [29], [30]. In [31], [32], Ye and Nie et al. have discussed this property, respectively, and have found that when the dimensionality of data is high and the size of samples is small, the true label matrix is constantly embedded into a

P. Song is with the School of Computer and Control Engineering, Yantai University, Yantai 264005, China. (e-mail: pengsong@ytu.edu.cn)

W. Zheng is with the Key Laboratory of Child Development and Learning Science of Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing 210096, China. (e-mail: wenming_zheng@seu.edu.cn)

Y. Yu is with the Department of Computer Science and Technology, Ocean University of China, Qingdao 266100, China. (e-mail: yuyanwei0530@gmail.com)

S. Ou is with the School of Opto-electronic Information Science and Technology, Yantai University, Yantai 264005, China. (e-mail: ousfeng@126.com)

Corresponding authors: P. Song and W. Zheng.

lower representation of data. The last few years have witnessed the emergence of regression based approaches, which have been proven very efficient for speech emotion recognition [33], [34], [35], [36]. In addition, when high dimensional features are directly used for the classification tasks, the unimportant features are typically included, resulting in performance degradation as well as computational complexity. Therefore, it is necessary to eliminate these unimportant features. Feature selection has been proven to be an efficient tool to address this problem. By using feature selection, the discriminating power of the selected features often outperforms that of the extracted features using conventional subspace learning algorithms, e.g. PCA, LDA and LPP [37].

Motivated by recent progress in regression analysis based classification, feature selection and manifold learning [29], [38], [10], [39], in this work, we present a novel method, called robust discriminative sparse regression (RDSR), which explicitly considers robust regression, feature selection, and structural consistency over labels simultaneously. To make our algorithm robust to noises and outliers, we utilize both labeled and unlabeled data, and introduce a $\ell_{2,1}-$norm on the regression loss function. To select discriminative and relevant features, we perform feature selection by imposing a sparse constraint on the regression matrix. To preserve more discriminative information, we further exploit the structure consistency over labels together.

Finally, we summarize the key contributions of this work in three-folds as follows:

- First, we propose to model the relationship between emotion features and corresponding labels via RDSR. we devise a robust regression algorithm by elegantly performing robust regression and feature selection simultaneously, which is robust to noise and outlier by using the $\ell_{2,1}-$norm. To the best of our best knowledge, RDSR is the first regression framework which explores feature selection and label consistency simultaneously to achieve the desirable learning goal.
- Second, we consider the local and global consistency over labels. By preserving these structural consistencies, our model can preserve more discriminating power.
- Finally, we conduct speech emotion recognition experiments on three widely used speech emotion datasets including EMO-DB, eNTERFACE, and BAUM$-$1s, where the results clearly demonstrate the effectiveness of our method.

The rest of this paper is organized as follows: In Section II, we briefly review the related work and highlight the differences. Section III is dedicated to introducing our proposed robust discriminative sparse regression (RDSR) algorithm for speech emotion recognition. Section IV presents experimental results and comparisons using three real-world datasets. Finally, we provide some concluding remarks and suggestions for future work in Section V.

II. Related work

In this section, we briefly discuss the previous works, i.e., regression methods and feature selection, and also highlight the differences between our work and the existing ones.

A. Regression methods

Regression analysis is a widely used statistical analysis technique [40], and has become a popular tool for many research areas [41], including machine learning [42], [43], [44], computer vision [29], emotion recognition [45], [33], speaker recognition [46].

In [29], Nassem et al. present a linear regression based classification algorithm for face recognition, in which the least-squares regression algorithm is used to the regression coefficients, and then the decision is made by using the minimum distance between the projected vector and the original vector. In [45], Yang et al. apply the linear regression algorithm to detect music emotion variations, and find it exhibits promising prediction accuracy. In [43], Wen et al. present a novel discriminative least square regression algorithm for multi-class problem. By introducing the inter-class sparsity, it can greatly enlarge the inter-class margin and simultaneously reduce the intra-class margin, and thus obtains a better performance. To achieve the goal of semi-supervised classification, Xiang et al. [44] present a local spline regression algorithm, in which the splines developed in Sobolev space is used to map the data points to the class labels. In [33], Zheng et al. propose a incomplete sparse least squares regression (ISLSR) model, for speech emotion recognition, where both labeled and unlabeled data are utilized to enhance the compatibility of the model.

The main limitation of most previous regression algorithms is that they neglect the structural information of labels or do not simultaneously consider feature selection. In reality, it is important to take into account these complementary properties together. Different from the previous regression algorithms, We propose a novel regression algorithm, referred to as RDSR, which considers feature selection and structural consistency over labels together, to achieves more effective and robust regression.

B. Feature selection

Feature selection is one of the most important dimensionality reduction methods. It aims at removing the redundant features, and selecting a subset of features from the high-dimensional feature set [10], [47]. According to design strategies, feature selection methods can be roughly categorized into three groups, i.e, filter methods, wrapper methods and embedded methods [48].

The filter methods first analyze the general characteristics of data, and then evaluate and select the features without involving any learning algorithms. For example, variance and Fisher score might be the two of the most widely used criterias for filter methods due to their good performance [49]. In reality, the filter methods can overcome the overfitting problems to some extent, but may fail to select the most relevant features [50].

The wrapper methods score the features using the predefined learning algorithms [51]. For example, in [52], Maldonado et al. present to select the relevant features according to the performance of SVM classifier. In principal, the wrapper methods can find the most useful features, and often outperform

the filter methods. However, they are prone to the overfitting problem, and have high computation cost.

The embedded methods combine feature selection and the learning algorithms together [53]. Recently, the embedded methods have caught an increasing attention. For example, Wang et al. [54] embed feature selection into a clustering algorithm. In [55], Weston et al. formulate feature selection and pattern classification objectives in a single optimization function. The embedded methods are similar to the wrapper methods, but have less computational cost and are less prone to overfitting.

In practical situations, it is hard to say that after performing feature selection, the selected feature subset is the most suitable one for speech emotion recognition. In light of this, it will benefit from devising a model for incorporating feature selection and emotion classification in a unified fashion. Therefore, in this paper, we present the RDSR method, which is a unified framework of regression analysis, feature selection and label consistency.

III. The proposed RDSR framework

In this section, we present the details of the proposed robust discriminative sparse regression (RDSR) method. First, the sparse linear regression function is developed to predict the mappings between the feature space and the label space. Second, the local and global consistency over labels are learned to make the model be more discriminative to predict the labels. Finally, the overall function of RDSR and optimization algorithm are given.

A. Preliminary

We begin with a brief introduction of some notations used here. Throughout this paper, we use the lowercase characters to denote the vectors, and uppercase characters to denote the matrices. For matrix $A \in R^{n\times m}$, its $i-$th row, $j-$th column are denoted as a^i and a_j, respectively. The Frobenius norm of A is denoted as

$$\|A\|_F = \sqrt{\Sigma_{i=1}^{n}\|a^i\|_2^2} = Tr(A^TA), \tag{1}$$

where $Tr(\cdot)$ demonstrates the trace operation, the superscript ^{T} denotes the transposition of a matrix.

And the $\ell_{2,1}-$norm of A is defined as

$$\|A\|_{2,1} = \Sigma_{i=1}^{n}\|a^i\|_2 = Tr(A^TDA), \tag{2}$$

where $D = [D_{ii}] \in R^{n\times n}$ is a diagonal matrix with $D_{ii} = \frac{1}{2\|a^i\|_2}$. Note that in practical situations, $\|a^i\|_2$ can be zero in theory. According to [56], here we define $D_{ii} = \frac{1}{2\sqrt{\|a^i\|_2^2+\epsilon}}$, where ϵ is a smoothing term with a small value.

B. Problem formulation

Let $X_l = [x_1, x_2, \ldots, x_m] \in R^{N\times m}$ denote a speech feature matrix, in which $x_i \in R^N$ is the feature vector of the $i-$th sample, and $Y_l = [y_1, y_2, \ldots, y_m]^T \in R^{m\times c}$ is the corresponding label matrix. Each column of Y_l indicates a labeling configuration as follows:

$$y_{ij} = \begin{cases} 1 & \text{if } x_i \text{ belongs to the } j-\text{th class;} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

In addition, we introduce a predicted label matrix $F_l = [f_1, f_2, \ldots, f_m]^T \in R^{m\times c}$, whose labels are unknown. Each column of F_l is the predicted label vector of x_i. We assume that there exists a linear mapping between the feature space and label space, i.e.,

$$p_j(x_i) = w_j^T x_i. \tag{4}$$

Denoting $W = [w_1, w_2, \ldots, w_c]$, we can obtain

$$p(X_l) = W^T X_l. \tag{5}$$

As is known to all, least squares regression is one of the most popular methods for classification. However, it is very sensitive to outliers and noises [57]. To solve this problem, first, we introduce the $\ell_{2,1}-$norm on the loss function, which is robust to outliers and noises [57], [58]. Second, we consider both labeled and unlabeled data together to train the model. Therefore, suppose that $X_a \in R^{N\times n}$ is the unlabeled feature matrix, and $F_a \in R^{n\times c}$ is the predicted label matrix accordingly. The objective function is written as

$$\min_{W,F} \|F - X^TW\|_{2,1}, \tag{6}$$

where $F = \left[\begin{smallmatrix} F_l \\ F_a \end{smallmatrix}\right]$ and $X = [X_l, X_a]$.

In real-world applications, the emotion features are usually high dimensional and contain redundant features for regression analysis. Thus, it is necessary to select the discriminating and informative features so that the negative influence of the redundant features can be effectively eliminated. To overcome this problem, we introduce the $\ell_{2,1}-$norm based feature selection meanwhile due to its efficacy in recent works [56], [57], where the $\ell_{2,1}-$norm is imposed on the regression matrix W. Therefore, we can obtain the following objective function:

$$\min_{W,F} \|F - X^TW\|_{2,1} + \lambda\|W\|_{2,1}, \tag{7}$$

where $\lambda \geq 0$ is a regularization parameter.

To better predict the emotion labels, we take into account the structural consistency of labels, i.e., local and global consistency over labels [59], [60]. That is, on one hand, the labels should be locally consistent, which means the labels should not change too much between nearby points. On the other hand, the labels are supposed to be globally consistent, which means the predicted labels should not change too much from the groundtruth labels. The two types of label consistency are defined as follows:

1) **The local label consistency:**
Recent studies in spectral theory [61] and manifold learning theory [62] demonstrate that the nearby data share the similar geometric structure. In this paper, first, we construct a $k-$nearest neighbor graph $G = [G_{ij}]$ to encode the geometric information in the labels [39], where each vertex corresponds to a label vector. There are many choices to define the weight matrix G. For simplification and efficiency, as in [39], [63], we use the commonly used $0-1$ weighting strategy, which is defined as follows:

$$G_{ij} = \begin{cases} 1 & \text{if } f_j \in N_p(f_i) \text{ or } f_i \in N_p(f_j) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $N_p(f_i)$ denotes the p nearest neighbors of f_i. Then we use the following term to measure the local consistency:

$$\min_F Tr(F^T L F), \tag{9}$$

where $L = B - G$ is known as graph Laplacian matrix [61], $B = [B_{jj}]$ is a diagonal matrix whose entries are the column sums of G as $B_{jj} = \sum_i G_{ij}$.

2) **The global label consistency:**
To ensure the well prediction of labels, it is reasonable to assume that the predicted labels are consistent with the groundtruth labels. Here we use the following term to measure this global consistency:

$$\min_F Tr\big((F-Y)^T U (F-Y)\big), \tag{10}$$

where U is a diagonal matrix, whose element U_{ii} is a large constant if x_i is labeled and $U_{ii} = 0$ otherwise.

By incorporating the regularization terms (9) and (10) into (7), the proposed robust discriminative sparse regression (RDSR) is formulated, and the objective function can finally be written as:

$$\begin{aligned} \min_{W,F} & \|F - X^T W\|_{2,1} + \lambda \|W\|_{2,1} + \alpha Tr(F^T L F) \\ & + \beta Tr\big((F-Y)^T U (F-Y)\big), \\ \text{s.t.} \quad & F \succeq 0. \end{aligned} \tag{11}$$

where α and β are two trade-off parameters controlling the relative contribution of the corresponding terms, and $F \succeq 0$ denotes all entries of F are non-negative.

In (11), the first regression term $\|F - X^T W\|_{2,1}$ is robust to noises and outliers, the second regularization term $\|W\|_{2,1}$ guarantees W is sparse in rows, which can select the most relevant features to predict the labels, and the third and the final terms are the label consistency regularizer for preserving the local and global consistency over labels, which can make the model have more discriminative power.

C. Optimization algorithm

The objective function in Eq. (11) involves the $\ell_{2,1}-$norm, which is non-smooth and cannnot have a closed-form solution [57]. Hence, it is hard to directly optimize. Consequently, we put forward an iterative optimization algorithm. To facilitate the optimization, we rewrite Eq. (11) as minimizing the equation problem as follows:

$$\begin{aligned} \mathcal{O} = \quad & \|F - X^T W\|_{2,1} + \lambda \|W\|_{2,1} + \alpha Tr(F^T L F) \\ & + \beta Tr\big((F-Y)^T U (F-Y)\big), \quad , \\ \text{s.t.} \quad & F \succeq 0. \end{aligned} \tag{12}$$

The whole alternate procedure of the proposed RDSR is listed as follows:

Fix F Update W

Given fixed F, we solve the regression matrix W, and Eq. (12) reduces to the following sub-problem:

$$\mathcal{O} = \|F - X^T W\|_{2,1} + \lambda \|W\|_{2,1} \tag{13}$$

According to [56], [64], solving (13) is equivalent to solving

$$\mathcal{O} = Tr\big((F - X^T W)^T P (F - X^T W)\big) + \lambda Tr(W^T Q W), \tag{14}$$

where $P = [P_{ii}] \in R^{(m+n)\times(m+n)}$ is a diagonal matrix with

$$P_{ii} = \frac{1}{2\sqrt{\|v^i\|_2^2 + \epsilon}} \tag{15}$$

in which v^i is the $i-$th row of V, where $V = F - X^T W$, and $Q = [Q_{ii}] \in R^{N\times N}$ is also a diagonal matrix with

$$Q_{ii} = \frac{1}{2\sqrt{\|w^i\|_2^2 + \epsilon}} \tag{16}$$

in which w^i is the $i-$th row of W.

Note that P and Q are unknown and depend on W. An iterative algorithm is introduced to solve problem (14). With fixed W, P and Q are obtained by Eqs. (15) and (16), respectively. And with fixed P and Q, W is obtained using Eq. (14). Taking the derivative $\mathcal{O}$ with respect to W, and setting the derivative to zero, we have

$$\begin{aligned} & XP(X^T W - F) + \lambda Q W = 0 \\ \Rightarrow \quad & (XPX^T + \lambda Q) W = XPF \\ \Rightarrow \quad & W = S^{-1} XPF, \end{aligned} \tag{17}$$

where $S = XPX^T + \lambda Q$.

Fix W Update F

Given fixed W, we compute the label matrix F, and reduce Eq. (12) to the following sub-problem:

$$\begin{aligned} \mathcal{O} = \quad & \|F - X^T W\|_{2,1} + \alpha Tr(F^T L F) \\ & + \beta Tr\big((F-Y)^T U (F-Y)\big), \end{aligned} \tag{18}$$

Then we substitute the expression for W in Eq. (17) into Eq. (18), and can obtain

$$\begin{aligned} \min_F & Tr\big((F - X^T W)^T P (F - X^T W)\big) + \alpha Tr(F^T L F) \\ & + \beta Tr\big((F-Y)^T U (F-Y)\big) \\ \text{s.t.} \quad & F \succeq 0, \end{aligned} \tag{19}$$

Let $\phi = [\phi_{ij}]$ is a Lagrange multiplier matrix, the Lagrange function of Eq. (19) is

$$\begin{aligned} & Tr\big((F - X^T W)^T P (F - X^T W)\big) + \alpha Tr(F^T L F) \\ & + \beta Tr\big((F-Y)^T U (F-Y)\big) + Tr(\phi F^T), \end{aligned} \tag{20}$$

where θ is a regularization parameter to control the orthogonal condition. By using the Karush$-$Kuhn$-$Tucker (KKT) condition $\phi_{ij} F_{ij} = 0$ [9], we get the following equation for F_{ij}:

$$P(F - X^T W) + \alpha L F + \beta U (F - Y) + \phi = 0 \tag{21}$$

This equation will lead to the following updating rules:

$$F_{ij} \longleftarrow F_{ij} \frac{(PX^T W + \beta U Y)_{ij}}{(PF + \alpha L F + \beta U F)_{ij}} \tag{22}$$

Algorithm 1 RDSR: Robust discriminative sparse regression algorithm

Input:

The feature matrix X_l, X_a and label matrix Y_l;

The parameters λ, α and β.

Output:

The regression matrix W.

a). Construct the p nearest neighbor graph G;

b). Initialize the weight matrix U;

repeat

1. $\boldsymbol{W}$ **− step**: Fix F and update W using Eq. (17);
2. $\boldsymbol{F}$ **− step**: Fix W and update F using Eq. (22);

until Converges.

Since the problem in Eq. (11) can be divided into two sub-problems, and each sub-problem is convex w.r.t. one variable, we can solve the sub-problems alternately to obtain a optimal solution, and finally the objective problem of our proposed model can find a local minima [57], [65]. The detailed algorithmic procedure of RDSR is summarized in Algorithm 1. Note that the convergence criterion used in our experiments is that the maximum number of iterations is 20 or $|\mathcal{O}_{t-1} - \mathcal{O}_t|/|\mathcal{O}_{t-1}| \leq 0.001$, where $\mathcal{O}_t$ is the value of the objective function in the $t-$th operation.

TABLE I: Statistics of the datasets.

Datasets	EMO-DB	eNTERFACE	BAUM−1s
Language	German	English	Turkish
Size	494	1170	1222
Classes	Seven	Six	Eight
Modal	Audio	Audio-visual	Audio-visual

IV. Experiments

In this section, we evaluate the RDSR method for speech emotion recognition on three real-world datasets. The followings describe the detail of experiments and results.

A. Data preparation

We conduct experiments on three public emotion datasets, i.e., EMO-DB[1], eNTERFACE[2] and BAUM−1s[3]. The important statistics of these datasets are summarized in Table I.

- The first dataset is EMO-DB [66], which is an open action speech emotion dataset. It contains seven emotion categories: anger, boredom, disgust, fear, happiness, neutral and sadness. Ten professional actors (five mal and five female) are asked to speak ten daily spoken sentences in German. Finally, 494 speech utterances are collected.
- The second dataset is eNTERFACE [67], which is a public audio-visual emotion dataset. It consists of six basic emotions, i.e., anger, disgust, fear, happiness, sadness and surprise. 42 subjects from 14 different nationalities are asked to act these emotions with the pre-defined content in English. Overall, the eNTERFACE contains 1170 recordings.
- The third dataset is BAUM−1s [68], which is another popular audio-visual emotion dataset. It covers eight emotions: anger, boredom, contempt, disgust, fear, happiness, sadness and surprise. 31 Turkish subjects (17 female and 14 male) are employed to simulate these emotions, and each emotional utterance is labeled by five annotators using a majority voting strategy. Finally, 1222 videos are recorded.

It should be noted that, as done in [69], [6], in this work, we aim to recognize six basic emotions, i.e., anger, disgust, fear, happiness, sadness and surprise. Thus, we use total 337 utterances in EMO-DB, 1170 utterances in eNTERFACE and 512 utterances in BAUM−1s for experiments.

B. Experimental setting

For acoustic features, we use the open source openSMILE toolkit [70] as the feature extractor. We use the standard feature set in the Computational Paralinguistics ChallengE (ComParE) of INTERSPEECH 2013 to 2018 [8]. This feature set consists of 65 low level descriptors (LLDs), e.g., MFCC, F0, energy, loudness. 54 statistical functions are applied to 59 LLDs, while 46 statistical functionals are applied to the detla of these 59 LLDs. 39 statistical functionals are employed to apply to the other six LLDs and the corresponding delta values. Moreover, five global temporal statistics are also included into the feature set. Thus, the total feature vector per utterance contains 6373 attributes. In our experiments, we use all these attributes. We concatenate these attributes into a 6373-dimensional feature vector and then normalize these feature vectors.

C. Baseline methods

We compare our proposed RDSR with state-of-the-art methods for speech emotion recognition as shown below:

- Support vector machine (SVM) [71] + feature selection (FS)
- 1-nearest neighbors (NN) [9] + FS
- Sparse representation classifier (SRC) [72] + FS
- Incomplete sparse least square regression (ISLSR) [33]
- Our proposed RDSR without considering label consistency (DSR)
- Deep convolutional neural networks (DCNN) [73]
- Deep convolutional neural networks with a discriminant temporal pyramid matching strategy (DCNN-DTPM) [6]

Specifically, all SVM, NN and SRC algorithms are performed with a generalized Fisher score [49] based feature selection together, and a linear kernel SVM is adopted due to the benefit of less parameters and fast computations [71]. DCNN and DCNN-DTPM are two popular deep learning methods for speech emotion recognition, in which three channels of log Mel-spectrograms are used as the input [6]. In particular, ISLSR is the most closely related method to RDSR, while RDSR differs from ISLSR by introducing the $\ell_{2,1}-$norm

[1]http://emodb.bilderbar.info/docu

[2]http://enterface.net/enterface05/main.php?frame=emotion

[3]http://baum1.bahcesehir.edu.tr

TABLE II: The recognition performance on EMO-DB.

Methods	Recognition rates (%)					
	Anger	Disgust	Fear	Happiness	Sadness	Average
NN+FS	67.18	70.34	71.33	69.54	71.99	73.55
SVM+FS	73.55	80.17	81.31	79.95	80.20	80.69
SRC+FS	70.62	78.83	80.22	78.96	78.13	79.82
ISLSR	82.06	84.89	85.76	87.24	87.03	84.85
DSR	82.31	85.01	85.69	87.44	87.23	84.98
DCNN	82.19	83.85	85.31	87.00	87.21	84.33
DCNN-DTPM	**83.15**	**86.01**	**87.62**	**87.32**	**87.59**	**86.17**
RDSR	**83.26**	**85.12**	**87.60**	**87.55**	**87.96**	**86.19**

TABLE III: The recognition performance on eNTERFACE.

Methods	Recognition rates (%)						
	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Average
NN+FS	72.53	48.12	46.86	42.35	67.15	43.11	53.06
SVM+FS	81.13	58.52	56.98	62.78	64.38	57.64	63.49
SRC+FS	76.01	61.32	57.86	53.13	66.90	50.68	61.23
ISLSR	89.99	59.68	61.02	71.22	68.11	65.72	69.36
DSR	90.06	59.73	60.87	71.96	67.84	65.97	69.65
DCNN	89.93	59.79	85.31	60.92	67.65	65.78	69.94
DCNN-DTPM	**90.35**	**60.10**	**61.06**	**72.18**	**68.96**	**66.24**	**71.35**
RDSR	**91.28**	**59.78**	**61.33**	**72.03**	**68.92**	**66.43**	**71.28**

TABLE IV: The recognition performance on BAUM-1s.

Methods	Recognition rates (%)						
	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Average
NN+FS	23.35	22.10	10.32	39.57	29.86	17.59	31.66
SVM+FS	27.14	24.98	13.21	44.63	33.98	19.31	37.12
SRC+FS	25.92	24.23	11.39	42.88	33.75	18.96	35.96
ISLSR	31.16	29.90	13.96	48.93	38.69	22.98	42.17
DSR	31.22	29.86	14.01	48.56	39.28	23.51	42.24
DCNN	31.95	30.76	14.99	49.13	41.02	23.26	43.09
DCNN-DTPM	**32.10**	**30.86**	**15.33**	**49.15**	**41.16**	**23.81**	**43.28**
RDSR	**31.98**	**30.87**	**14.32**	**49.15**	**41.08**	**23.92**	**43.15**

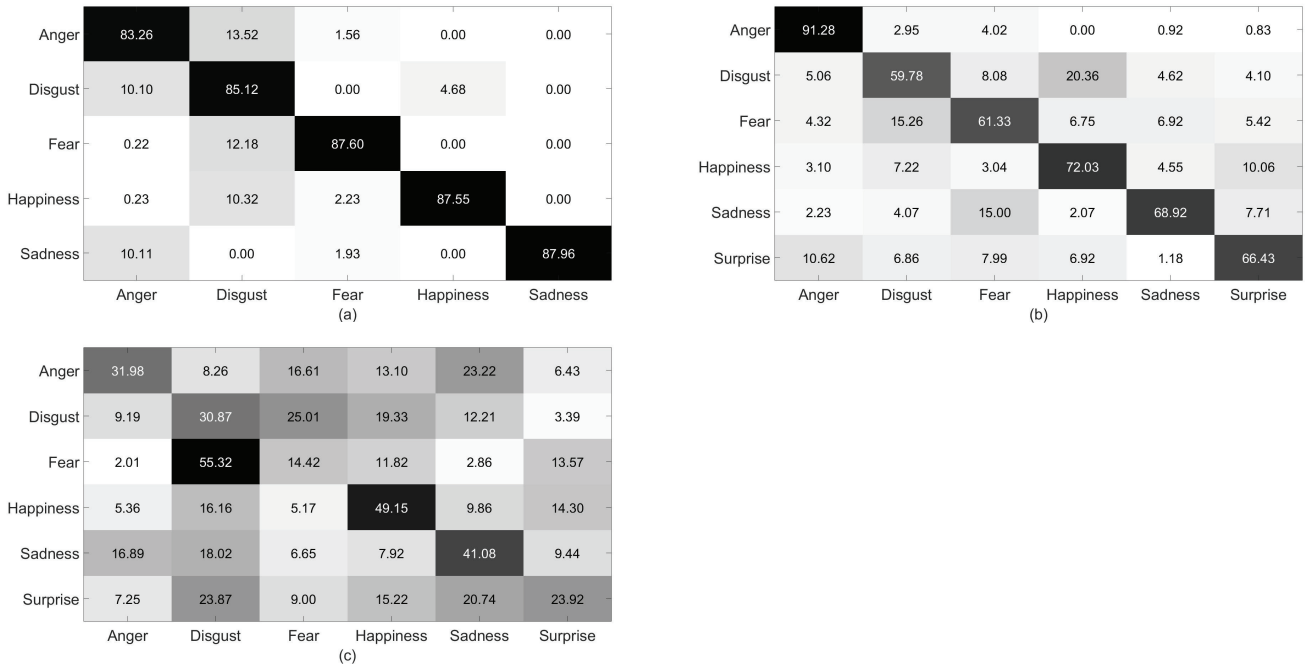


Fig. 1: Confusion matrices of our proposed method under different datasets: (a). the confusion matrix of EMO-DB; (b). the confusion matrix of eNTERFACE; **(c)**. the confusion matrix of BAUM-1s.

sparsity on the loss function and taking into account the label consistency. It should also be noted that DSR is a special case of RDSR with $\alpha = \beta = 0$.

D. Experimental results

In our experiments, we randomly select 50 percent of samples as the training set, 25 percent of samples as the testing set, and the rest as the development set. It should be noted that the training set is labeled, and the testing and development sets are unlabeled. When comparing with the baseline methods, we set the following parameters $\lambda = 1$, $\alpha = 100$ and $\beta = 100$, by searching $\lambda, \alpha, \beta \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. The parameter p is set to 5 in the experiment. The recognition accuracies of RDSR and the seven baseline methods are illustrated in Tables II, III and IV. From these tables, we can make the following observations.

RDSR achieves much better results than the five traditional baseline algorithms with statistical significance. The average recognition rates of RDSR are 86.19%, 71.28% and 43.15%, respectively. Since these results are obtained from different types of public emotion datasets. It can convincingly verify that RDSR is efficient for speech emotion recognition.

Secondly, we notice that DSR performs better than ISLSR, which is a state-of-the-art regression based method for speech emotion recognition. A major limitation of ISLSR is that it uses a ℓ_2-norm minimization the on loss function, which is sensitive to the outliers and noises [57], and DSR avoids this limitation and obtains better recognition results.

Thirdly, we see that all methods perform well on EMO-DB and eNTERFACE datasets, but poorly on BAUM-1s dataset. Note that in BAUM-1s the expressions are spontaneous while in the other two datasets the expressions are acted. This demonstrates that the spontaneous emotions are more difficult to be recognized compared to the acted emotions.

Fourthly, RDSR significantly outperforms DSR, which does not consider the local and global consistency over labels. This validates that the label structural information is very important for emotion label prediction. In addition, we can find that the fear expression in EMO-DB dataset obtains the highest performance gain. This might be attributed to that, the fear expression in EMO-DB dataset is more sensitive to the label structural information compared with other expressions to some extent.

Lastly, note that in our experiments, we have also compared our proposed RDSR with two state-of-the-art deep learning algorithms, i.e., DCNN and DCNN-DTPM. From the tables, we observe that, RDSR outperforms DCNN in all cases, and can obtain similar performance to DCNN-DTPM. These results demonstrate the effectiveness of RDSR.

To further investigate the recognition performance of RDSR, we present the confusion matrices corresponding to Tables II-IV. Figs. 1 (a), (b) and (c) show the confusion matrices of experiments on EMO-DB, eNTERFACE and BAUM-1s, respectively, from which we can find that the mostly confused expressions are disgust and fear. From these figures, we can also observe that the emotions in BAUM-1s are more likely to be confused with each other, which coincides with the findings in [6], where the authors point out that the confusion is due to spontaneous emotions.

E. Effectiveness verification

In this section, we further verify the effectiveness of RDSR by inspecting the feature selection and label consistency. Here we consider three cases of RDSR as follows:

- RDSR$_1$: It can be viewed as a special case of RDSR without considering feature selection, with $\lambda = 0$ in Eq. (11).
- RDSR$_2$: It can be viewed as a special case of RDSR without considering the local structural consistency over labels, with $\alpha = 0$ in Eq. (11).
- RDSR$_3$: It can be viewed as a special case of RDSR without considering the global structural consistency over labels, with $\beta = 0$ in Eq. (11).

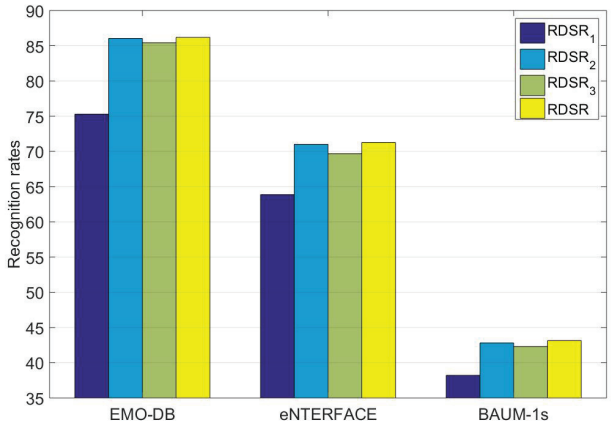


Fig. 2: Effectiveness verification: recognition accuracies (%) of our proposed RDSR under different settings.

We run RDSR$_1$, RDSR$_2$, RDSR$_3$ and RDSR on all datasets using the optimal parameters. Fig. 2 shows the recognition accuracies for each method. We can have the following observations: First, without feature selection, the RDSR$_1$ performs the worst. This indicates that feature selection can help much for RDSR model. Second, RDSR$_2$ can obtain better performance than RDSR$_3$ in all cases. The reason might be that the global label consistency is more important than the local label consistency for our model. Finally, RDSR$_2$ and RDSR$_3$ perform worse than RDSR. This proves that only considering the local or global consistency over labels is not enough for effective recognition.

F. Parameter sensitivity analysis

We conduct empirical parameter sensitivity analysis using all datasets, which validates that RDSR can obtain optimal recognition performance under a wide range of parameter values.

Feature selection regularization λ: We run RDSR with varying values of λ. Theoretically, λ controls of the degree of feature selection. When $\lambda \to 0$, feature selection will not be performed. When $\lambda \to \infty$, the optimization problem will be ill defined. We plot the recognition accuracies w.r.t. different values of λ in Fig. 3 (a), and choose λ as 1.

Label consistency regularization α and β: We run RDSR with varying values of α and β. Theoretically, α and β control the weight of label consistency regularization, and the larger values of α and β make the structural consistency over labels more important in RDSR. We plot the recognition accuracies w.r.t. different values of α and β in Fig.3 (b) and (c), respectively, and choose α and β as 100.

Number of nearest neighbors p: We also run RDSR with varying values of p. Initially, the larger value of p will result in a dense-neighbor graph. Thus, p should not be neither too large or too small to ensure a optimal performance. We plot the recognition accuracies w.r.t. different values of p in Fig.3 (d), and we choose p as 5 for our experiments.

G. Convergence analysis

Since the optimization of RDSR is an iterative algorithm, we need to check the convergence property by conducting the corresponding experiments on the EMO-DB, eNTERFACE and BAUM-1s datasets. Fig. 4 shows the convergence performance of the values of the objective function. From these figures, we can find that the values of objective function monotonically decrease when the iteration round increases for these three datasets, demonstrating that our proposed method is efficient and can converge quickly. Here, we also give the time cost of Algorithm 1 on these three datasets. By using a computer which has an Intel Core E5-2660 of 2.20GHz and 48GB RAM, the execution time on EMO-DB, eNTERFACE and BAUM-1s datasets is around 8.23s, 14.25s and 13.86s, respectively.

V. Conclusions and Future Work

In this paper, we have proposed a novel algorithm, called robust discriminative sparse regression (RDSR), for speech emotion recognition. In details, by utilizing both labeled and unlabeled data, we develop a joint learning framework by taking into account regression and feature selection together, and thus our algorithm can empirically have classification power. Moreover, the label consistency is considered, in which the local and global structural consistency over labels are incorporated into our model to make it be more discriminative. Experimental results on several emotion benchmarks demonstrate that RDSR is efficient for speech emotion recognition problem, and can significantly outperform state-of-the-art methods.

In our future work, we plan to develop a transferred version of RDSR, such that our model can be suitable for practical cross-corpus generalizability of speech emotion recognition. In addition, we would like to involve subspace learning algorithms into RDSR model to improve the recognition performance. Moreover, deep learning architectures, e.g., CNN, LSTM, have been successfully employed for feature learning in speech emotion recognition [6], [26]. Thus, it is worth trying to integrate the deep features into our model to further boost the recognition performance.

VI. Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grants 61703360, 61921004 and 61773331, and the Fundamental Research Funds for the Central Universities under Grant CDLS-2019-01.

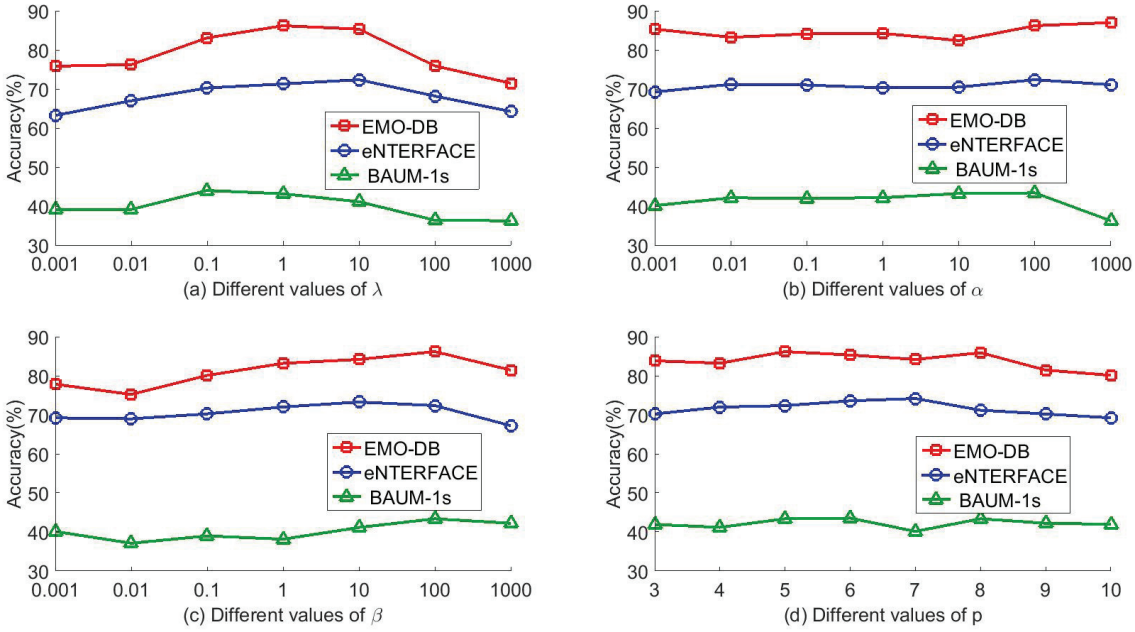
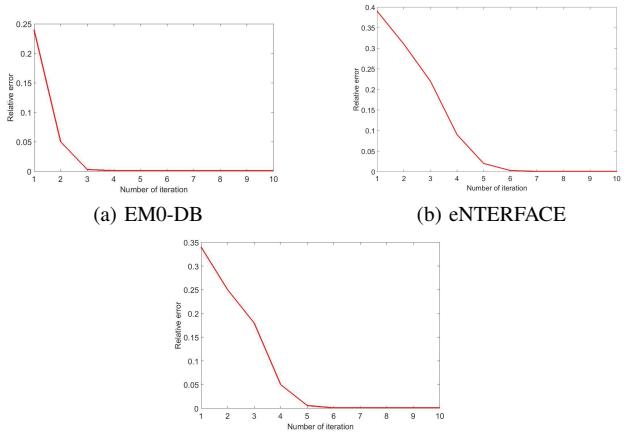


Fig. 3: Parameter sensitivity study for RDSR: (a) feature selection λ; (b) local label consistency regularization α; (c) global label consistency regularization β; (d) number of nearest neighbors p.



(a) EM0-DB (b) eNTERFACE

(c) Baum-1s

Fig. 4: Convergence curve of the relative errors of objective function values of RDSR under different settings.

REFERENCES

[1] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Iintelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[2] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[3] Peng Song, Wenming Zheng, Shifeng Ou, Xinran Zhang, Yun Jin, Jinglei Liu, and Yanwei Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.

[4] Chenjian Wu, Chengwei Huang, and Hong Chen, "Text-independent speech emotion recognition using frequency adaptive features," *Multimedia Tools and Applications*, pp. 1–11, 2018.

[5] Shaoling Jing, Xia Mao, and Lijiang Chen, "Prominence features: Effective emotional features for speech emotion recognition," *Digital Signal Processing*, vol. 72, pp. 216–231, 2018.

[6] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

[7] Diana Torres-Boza, Meshia Cédric Oveneke, Fengna Wang, Dongmei Jiang, Werner Verhelst, and Hichem Sahli, "Hierarchical sparse coding framework for speech emotion recognition," *Speech Communication*, vol. 99, pp. 80–89, 2018.

[8] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, 2013.

[9] Christopher Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[10] Jie Gui, Zhenan Sun, Shuiwang Ji, Dacheng Tao, and Tieniu Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 7, pp. 1490–1507, 2017.

[11] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Iintelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[12] Adrian Barbu, Yiyuan She, Liangjing Ding, and Gary Gramajo, "Feature selection with annealing for computer vision and big data learning," *IEEE Transactions on Pattern Analysis and Machine Iintelligence*, vol. 39, no. 2, pp. 272–286, 2017.

[13] Peng Song and Wenming Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2018.

[14] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004, vol. 1, pp. I–577.

[15] Hao Hu, Ming-Xing Xu, and Wei Wu, "Gmm supervector based svm with spectral features for speech emotion recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007, vol. 4, pp. IV–413.

[16] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[17] Joy Nicholson, Kazuhiko Takahashi, and Ryohei Nakatsu, "Emotion recognition in speech using neural networks," *Neural computing & applications*, vol. 9, no. 4, pp. 290–296, 2000.

[18] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *Proc. INTERSPEECH*, pp. 223–227, 2014.

[19] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[20] Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 115–126, 2015.

[21] Mehrdad J Gangeh, Pouria Fewzee, Ali Ghodsi, Mohamed S Kamel, and Fakhri Karray, "Multiview supervised dictionary learning in speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1056–1068, 2014.

[22] Xiaoming Zhao, Shiqing Zhang, and Bicheng Lei, "Robust emotion recognition in noisy speech via sparse representation," *Neural Computing and Applications*, vol. 24, no. 7-8, pp. 1539–1553, 2014.

[23] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[24] Qingchen Zhang, Laurence T Yang, Zhikui Chen, and Peng Li, "A survey on deep learning for big data," *Information Fusion*, vol. 42, pp. 146–157, 2018.

[25] Zakaria Aldeneh and Emily Mower Provost, "Using regional saliency for speech emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2741–2745.

[26] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[27] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak, "Emotion identification from raw speech signals using DNNs," *Proc. Interspeech 2018*, pp. 3097–3101, 2018.

[28] Marie Tahon and Laurence Devillers, "Towards a small set of robust acoustic features for emotion recognition: challenges," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 1, pp. 16–28, 2016.

[29] Imran Naseem, Roberto Togneri, and Mohammed Bennamoun, "Linear regression for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2106–2112, 2010.

[30] Liang Chen, "Dual linear regression based classification for face cluster recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2673–2680.

[31] Jieping Ye, "Least squares linear discriminant analysis," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 1087–1093.

[32] Feiping Nie, Zinan Zeng, Ivor W Tsang, Dong Xu, and Changshui Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1796–1808, 2011.

[33] Wenming Zheng, Minghai Xin, Xiaolan Wang, and Bei Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569–572, 2014.

[34] Yaxin Sun and Guihua Wen, "Ensemble softmax regression model for speech emotion recognition," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 8305–8328, 2017.

[35] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007, vol. 4, pp. IV–1085.

[36] Xinzhou Xu, Jun Deng, Eduardo Coutinho, Chen Wu, Li Zhao, and Bjorn W Schuller, "Connecting subspace learning and extreme learning machine in speech emotion recognition," *IEEE Transactions on Multimedia*, 2018.

[37] Yu Zhang, Jianxin Wu, and Jianfei Cai, "Compact representation for image classification: To choose or to compress?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 907–914.

[38] Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 156–171, 2017.

[39] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[40] Shiming Xiang, Feiping Nie, Gaofeng Meng, Chunhong Pan, and Changshui Zhang, "Discriminative least squares regression for multi-class classification and feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 11, pp. 1738–1754, 2012.

[41] Ashish Sen and Muni Srivastava, *Regression analysis: theory, methods, and applications*, Springer Science & Business Media, 2012.

[42] Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, III, "The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984.

[43] Jie Wen, Yong Xu, Zuoyong Li, Zhongli Ma, and Yuanrong Xu, "Inter-class sparsity based discriminative least square regression," *Neural Networks*, vol. 102, pp. 36–47, 2018.

[44] Shiming Xiang, Feiping Nie, and Changshui Zhang, "Semi-supervised classification via local spline regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2039–2053, 2010.

[45] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen, "A regression approach to music emotion recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 448–457, 2008.

[46] Xiao-Lei Zhang, "Linear regression for speaker verification," *arXiv*, pp. 1–10, 02 2018.

[47] Ronghua Shang, Yang Meng, Chiyang Liu, Licheng Jiao, Amir M Ghalamzan Esfahani, and Rustam Stolkin, "Unsupervised feature selection based on kernel fisher discriminant analysis and regression learning," *Machine Learning*, pp. 1–28, 2018.

[48] Sotiris Kotsiantis, "Feature selection for machine learning classification problems: a recent overview," *Artificial Intelligence Review*, vol. 42, no. 1, pp. 157–176, 2011.

[49] Quanquan Gu, Zhenhui Li, and Jiawei Han, "Generalized fisher score for feature selection," in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2011, pp. 266–273.

[50] Xiaofeng Zhu, Xuelong Li, Shichao Zhang, Chunhua Ju, and Xindong Wu, "Robust joint graph sparse coding for unsupervised spectral feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 6, pp. 1263–1275, 2017.

[51] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[52] Sebastián Maldonado and Richard Weber, "A wrapper method for feature selection using support vector machines," *Information Sciences*, vol. 179, no. 13, pp. 2208–2217, 2009.

[53] Deng Cai, Chiyuan Zhang, and Xiaofei He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 333–342.

[54] Suhang Wang, Jiliang Tang, and Huan Liu, "Embedded unsupervised feature selection," in *Twenty-ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 470–476.

[55] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping, "Use of the zero norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, 2003.

[56] Ran He, Tieniu Tan, Liang Wang, and Wei-Shi Zheng, "$l_{2,1}$ regularized correntropy for robust feature selection," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2504–2511.

[57] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *Proc. NIPS*, 2010, pp. 1813–1821.

[58] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou, "L2, 1-norm regularized discriminative feature selection for unsupervised," in *AAAI*, 2011, pp. 1589–1594.

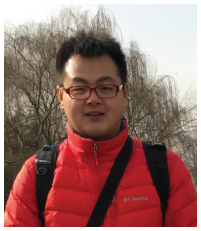
[59] Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems*, 2004, pp. 321–328.

[60] Lei Zhang and David Zhang, "Visual understanding via multi-feature shared learning with global consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 247–259, 2016.

[61] Fan RK Chung, *Spectral graph theory*, vol. 92, American Mathematical Soc., 1997.
[62] H Sebastian Seung and Daniel D Lee, "The manifold ways of perception," *Science*, vol. 290, no. 5500, pp. 2268–2269, 2000.
[63] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai, "Graph regularized sparse coding for image representation," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1327–1336, 2011.
[64] Feiping Nie, Wei Zhu, and Xuelong Li, "Unsupervised feature selection with structured graph optimization," in *Thirtieth AAAI conference on artificial intelligence*, 2016.
[65] Hong Tao, Chenping Hou, Feiping Nie, Yuanyuan Jiao, and Dongyun Yi, "Effective discriminative feature selection with nontrivial solution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 796–808, 2016.
[66] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of german emotional speech.," in *Proc. INTERSPEECH*. ISCA, 2005, vol. 5, pp. 1517–1520.
[67] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, "The enterface'05 audio-visual emotion database," in *Proc. 22nd International Conference on Data Engineering Workshops*, 2006, pp. 8–8.
[68] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem, "Baum-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2017.
[69] Yongjin Wang and Ling Guan, "Recognizing human emotional state from audiovisual signals," *IEEE transactions on multimedia*, vol. 10, no. 5, pp. 936–946, 2008.
[70] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.
[71] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.
[72] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
[73] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

Peng Song is currently an associate professor with the school of computer and control engineering, Yantai University, China. He received the B.S. degree in EE from Shandong University of Science and Technology, China in 2006, the M.E. and P.h.D degrees in EE both from Southeast University, China in 2009 and 2014, respectively. From 2007 to 2008, he was a research intern at Microsoft Research Asia. From 2009 to 2011, he worked as a software engineer at Motorola. His current main research interests include affective computing, speech signal processing and machine learning.

Wenming Zheng (SM'18) received the B.S. degree in computer science from Fuzhou University, Fuzhou, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, China, in 2004. Since 2004, he has been with the Research Center for Learning Science, Southeast University, where he is currently a Professor with the School of Biological Science and Medical Engineering and the Key Laboratory of Child Development and Learning Science of the Ministry of Education. His current research interests include affective computing, pattern recognition, machine learning, and computer vision.



Yanwei Yu received his Ph.D. degree in Computer Science from University of Science and Technology Beijing, China in 2014. From 2012 to 2013, he was a visiting scholar in the Worcester Polytechnic Institute, Massachusetts. From 2016 to 2018, he was a postdoctoral researcher at the College of Information Sciences and Technology, Pennsylvania State University, Pennsylvania. He is currently an associate professor at the Department of Computer Science and Technology, Ocean University of China. His research interests include data mining, machine learning and distributed computing.

Shifeng Ou received the Ph.D. degree in communication and information system from Jilin University, China in 2008. Currently, he is a professor at Yantai University, China. His main research interests include speech signal processing and blind source separation.