A Novel Speech Emotion Recognition Method via Incomplete Sparse Least Square Regression

Wenming Zheng, *Member, IEEE*, Minghai Xin, Xiaolan Wang, and Bei Wang

***Abstract*—In this letter, we propose a novel speech emotion recognition method based on least square regression (LSR) model, in which a novel incomplete sparse LSR (ISLSR) model is proposed and utilized to characterize the linear relationship between speech features and the corresponding emotion labels. In training the ISLSR model, both labeled and unlabeled speech data sets are utilized, where the use of unlabeled data set aims to enhance the compatibility of the model such that it is well suitable for the out-of-sample speech data. Another novelty of ISLSR lies in the capability of dealing with feature selection. To evaluate the performance of the proposed method, we conduct experiments on two emotional speech databases. The experimental results on both databases demonstrate that the proposed method achieves better recognition performance in compared with several state-of-the-art methods.**

***Index Terms*—Feature extraction, incomplete sparse least square regression, sparse learning, speech emotion recognition.**

I. Introduction

SPEECH emotion recognition has been a very active research topic in the pattern recognition field. A major goal of emotion recognition from speech is to classify the speech utterances into one of the predefined emotion categories, e.g., anger, joy, sadness, fear, disgust, boredom, neutral [1]. Overall, an automatic speech emotion recognition system can be divided into two major parts, i.e., speech feature extraction versus emotion classification [2]. The main task of the first part is to extract the speech features that are related with the emotions of the speakers, whereas the latter one is to determine the emotion categories based on the extracted speech features. During the last decades, many speech emotion recognition methods had been proposed in the literature [2], among which the regression based approaches had been very popular in recent years [4]. One of the most commonly used approaches of applying regression model to speech emotion recognition is the ordinary least square regression (LSR) model. This method aims to seek a transformation matrix $\mathbf{C}$, such that the difference between emotion label matrix $\mathbf{L}$ and transformed speech feature matrix $\mathbf{CD}$ is minimal. The optimization problem can be formulated as:

$$\arg\min_{\mathbf{C}} \|\mathbf{L}-\mathbf{CD}\|_F^2. \tag{1}$$

In dealing with the speech emotion recognition problem, it is notable that most methods use a two-stage procedure, i.e., first using the training data to train a classification model and then predict the emotion categories of the testing data based on this model. Since the training stage and the testing stage are separated, it would be hard to guarantee that the trained model is well suitable for out-of-sample speech data. This problem occurs especially when there are large dissimilarities between the training speech data and the testing ones.

Another problem one should be noted in the speech emotion recognition is that different speech emotion features contribute differently to the emotion recognition. Some of the features contribute more whereas some of them contribute less. Hence, it would not be a good choice to use all speech features since some of them may contain less useful speech emotion information. In addition, the increase of speech features may also result in the increase of the computational complexity, and the higher dimensionality of the speech feature vector may cause the over-fitting problem [5]. Consequently, it is desired to perform feature selection operation in speech emotion recognition. Feature selection is an active research topic of pattern recognition [19]. There are many feature selection approaches had been proposed during the past decades, among which the approach of sparse learning via ℓ_1 norm penalty has been one of the most popular ones in recent years [6][7].

Based on the above analysis, in this letter we present a novel incomplete sparse LSR (ISLSR) model for speech emotion recognition, in which ISLSR is used to characterize the linear relationship between the speech features and the corresponding emotion labels. In contrast to the traditional LSR model, the ISLSR model requires two speech feature sets, i.e., one labeled feature set and one unlabeled feature set, to train the model parameters. The major reason of using unlabeled feature set is expected to obtain the better compatibility of the model, such that it would be more powerful to deal with the out-of-sample emotion recognition problem. Besides the use of unlabeled feature set, another novelty of ISLSR lies in its ability of dealing with the feature selection problem, where the feature selection is realized by imposing an $\ell_{2,1}$-norm penalty of the regression coefficient matrix on the objective function of ISLSR. To optimize model parameters of ISLSR, we adopt the alternating direction method (ADM) [12] and the inexact

Manuscript received November 28, 2013; revised February 05, 2014; accepted February 16, 2014. Date of publication February 27, 2014; date of current version March 18, 2014. This work was supported in part by the National Basic Research Program of China under Grant 2011CB302202, the National Natural Science Foundation of China under Grants 61231002 and 61073137, the Natural Science Foundation of Jiangsu Province under Grant BK20130020, the Ph.D. Program Foundation of Ministry Education of China under Grant 20120092110054, and by the Program for Distinguished Talents of Six Domains in Jiangsu Province of China under Grant 2010-DZ088. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Frederic Bechet.

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

Digital Object Identifier 10.1109/LSP.2014.2308954

augmented Lagrangian multiplier (ALM) approach [13] to design an efficient algorithm in the letter.

The remainder of this letter is organized as follows. In section II and III, we propose the ISLSR model and the speech emotion recognition method, respectively. The experiments are presented in section IV and section V concludes the letter.

II. ISLSR Model

A. Formulation

Let $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_N] \in \mathbb{R}^{m\times N}$ denote a speech feature matrix and $\mathbf{L} = [\mathbf{l}_1, \ldots, \mathbf{l}_N] \in^{c\times N}$ denote the corresponding emotion label matrix, where m, N, and c denote the number of speech features, the number of data samples, and the number of emotion categories, respectively. Each column of $\mathbf{D}$ is a feature vector consisting of emotional speech features, and the entries of $\mathbf{l}_j = [l_{1j}, \ldots, l_{cj}]^T$ take the values of 0 or 1, i.e.,

$$l_{ij} = \begin{cases} 1, & \text{if } \mathbf{d}_j \text{ belongs to the } i-\text{th class;} \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that $\mathbf{D}_t \in \mathbb{R}^{m\times M}$ is another data matrix whose class labels are unknown. Then, the target of our ISLSR model is to estimate the ISLSR model parameters as well as the unknown emotion label matrix, denoted by $\mathbf{L}_t \in \mathbb{R}^{c\times M}$, corresponding to the data matrix $\mathbf{D}_t$ based on the given data matrices $\mathbf{D}$, $\mathbf{D}_t$, and $\mathbf{L}$. The optimization problem of ISLSR can be formulated as:

$$\arg\min_{\mathbf{C},\mathbf{L}_t} \|[\mathbf{L} \quad \mathbf{L}_t] - \mathbf{C}[\mathbf{D} \quad \mathbf{D}_t]\|_F^2, \tag{2}$$

where $\mathbf{C}$ is a regression coefficient matrix.

As to the realization of the feature selection for the speech features, an equivalent way is to make some of the columns of $\mathbf{C}$ be zeros. To this end, we reformulate the above optimization problem in (2) by imposing an $\ell_{2,1}$ norm penalty with respect to $\mathbf{C}$ onto the objective function to shrink the entries of some of the columns to zero, where the $\ell_{2,1}$ norm penalty of $\mathbf{C}$ is defined as the summation of the ℓ_2 norm of columns of $\mathbf{C}$. In this way, only the features corresponding to the non-zero columns of $\mathbf{C}$ can be used for the speech emotion recognition. Then, the ISLSR model can be reformulated by

$$\begin{aligned} &\arg\min_{\mathbf{L}_t,\mathbf{C}} \|[\mathbf{L} \quad \mathbf{L}_t] - \mathbf{C}[\mathbf{D} \quad \mathbf{D}_t]\|_F^2 + \lambda\|\mathbf{C}\|_{2,1}, \\ &\text{s.t.} \mathbf{L}_t^T\mathbf{1} = \mathbf{1}, \mathbf{L}_t \succeq \mathbf{0}, \end{aligned} \tag{3}$$

where $\mathbf{L}_t \succeq \mathbf{0}$ denotes all entries of $\mathbf{L}_t$ are non-negative, $\mathbf{1}$ is a vector with all entries are 1, and λ is a trade-off parameter.

B. Optimization of ISLSR

The ISLSR optimization problem in (3) can be resolved by the ADM method. The detailed procedures of the solution approach can be summarized as follows:

1) Fix $\mathbf{L}_t$ *and optimize* $\mathbf{C}$*:* In this case, the optimization problem of (3) can be simplified as:

$$\arg\min_{\mathbf{C}} \left\|\tilde{\mathbf{L}} - \mathbf{C}\tilde{\mathbf{D}}\right\|_F^2 + \lambda\|\mathbf{C}\|_{2,1}, \tag{4}$$

where $\tilde{\mathbf{L}} = [\mathbf{L} \quad \mathbf{L}_t]$, $\tilde{\mathbf{D}} = [\mathbf{D} \quad \mathbf{D}_t]$. We can solve the optimization problem of (4) using ALM method. For this purpose, we express this optimization problem as:

$$\arg\min_{\mathbf{P},\mathbf{C}} \left\|\tilde{\mathbf{L}} - \mathbf{P}\tilde{\mathbf{D}}\right\|_F^2 + \lambda\|\mathbf{C}\|_{2,1}, \text{s.t.} \mathbf{P} = \mathbf{C}. \tag{5}$$

The corresponding augmented Lagrangian function of (5) can be expressed as:

$$\left\|\tilde{\mathbf{L}} - \mathbf{P}\tilde{\mathbf{D}}\right\|_F^2 + \mathrm{Tr}\left[\mathbf{T}^T(\mathbf{P} - \mathbf{C})\right] + \frac{\mu}{2}\|\mathbf{P} - \mathbf{C}\|_F^2 + \lambda\|\mathbf{C}\|_{2,1},$$

where $\mathbf{T}$ is the Lagrangian multiplier matrix and $\mu > 0$ is a regularization constant. Algorithm 1 provides the pseudo code of solving the optimization of (5).

Algorithm 1 Algorithm for solving (5).

Repeating steps (1) to (3) until convergence:

(1) Fix other parameters and update $\mathbf{P}$:

$$\arg\min_{\mathbf{P}} \left\|\tilde{\mathbf{L}} - \mathbf{P}\tilde{\mathbf{D}}\right\|_F^2 + \mathrm{Tr}\left[\mathbf{T}^T(\mathbf{P} - \mathbf{C})\right] + \frac{\mu}{2}\|\mathbf{P} - \mathbf{C}\|_F^2,$$

which results in

$$\mathbf{P} = \left(2\tilde{\mathbf{L}}\tilde{\mathbf{D}}^T/\mu - \mathbf{T}/\mu + \mathbf{C}\right)\left(2\tilde{\mathbf{D}}\tilde{\mathbf{D}}^T/\mu + \mathbf{I}_m\right)^{-1};$$

(2) Fix other parameters and update $\mathbf{C}$:

$$\arg\min_{\mathbf{C}} \frac{\lambda}{\mu}\|\mathbf{C}\|_{2,1} + \frac{1}{2}\|\mathbf{C} - (\mathbf{P} + \frac{1}{\mu}\mathbf{T})\|_F^2,$$

which results in (see Lemma 4.1 in [14])

$$\mathbf{c}_i = \begin{cases} \frac{\|\mathbf{p}_i + \frac{\mathbf{t}_i}{\mu}\| - \frac{\lambda}{\mu}}{\|\mathbf{p}_i + \frac{\mathbf{t}_i}{\mu}\|}(\mathbf{p}_i + \frac{\mathbf{t}_i}{\mu}), & if \lambda < \|\mathbf{p}_i + \frac{\mathbf{t}_i}{\mu}\|; \\ \mathbf{0}, & \text{otherwise,} \end{cases}$$

where $\mathbf{c}_i$, $\mathbf{t}_i$, and $\mathbf{p}_i$ denote the i-th column of $\mathbf{C}$, $\mathbf{T}$, and $\mathbf{P}$, respectively;

(3) Update $\mathbf{T}$ and the parameter μ:

$$\mathbf{T} = \mathbf{T} + \mu(\mathbf{P} - \mathbf{C}), \mu = \min(\rho\mu, \mu_{max}),$$

where ρ is a scaled parameter;

2) Fix $\mathbf{C}$ *and optimize* $\mathbf{L}_t$*:* In this case, the optimization problem of (3) can be simplified as:

$$\arg\min_{\mathbf{L}_t} \|\mathbf{L}_t - \mathbf{C}\mathbf{D}_t\|_F^2, \text{s.t.} \mathbf{L}_t^T\mathbf{1} = \mathbf{1}, \mathbf{L}_t \succeq \mathbf{0}. \tag{6}$$

which is equivalent to solving the following ones:

$$\arg\min_{\mathbf{l}_j^{(t)}} \left\|\mathbf{l}_j^{(t)} - \mathbf{C}\mathbf{d}_j^{(t)}\right\|_2^2, \text{s.t.} (\mathbf{l}_j^{(t)})^T\mathbf{1} = 1, \mathbf{l}_j^{(t)} \succeq \mathbf{0}, \tag{7}$$

where $\mathbf{l}_j^{(t)}$ and $\mathbf{d}_j^{(t)}$ denote the j-th column of $\mathbf{L}_t$ and the j-th column of $\mathbf{D}_t$, respectively. The optimization problem of (7) can be solved using quadratic programming algorithm [15].

III. Speech Emotion Recognition Based on ISLSR

A. Emotional Speech Feature Extraction

Basically, the emotional speech features can be classified into different categories according to different criteria. For example, Ayadi *et al.* classified the emotional speech features into the four categories [2]: (1) prosody features, such as pitch, energy, and formants, etc.; (2) spectral features, such as LPC features, MFCC features, etc.; (3) voice quality features; (4) Teager energy operator-based features. Another classification method is to categorize the speech emotion features into the long-term features versus the short-term features [8][9]. The short-term features characterize the speech signals in a short time period, such kind of features include formants, pitch and energy. In contrast to the short-term features, the long-term features mainly characterize the speech voice over the whole utterance via statistical approach. Some of the frequently used long-term features include mean and standard deviation.

In this letter we utilize the OpenSmile toolkit [10] to extract the speech features, where the standard feature set of the Interspeech 2010 Paralinguistic Challenge provided by the organizers [10][11] is adopted. This feature set consists of 1582 features, including 34 basic low-level descriptors (LLD) and their corresponding 34 delta coefficients, 21 functionals applied to the above 68 LLD contours, and 19 functionals applied to the 4 pitch-based LLD and their corresponding 4 delta coefficients. In addition, the number of pitch onsets and the total duration of the utterances are also used as two features and are included in the feature set.

B. Speech Emotion Recognition Based on ISLSR

To assign a label to a testing feature vector, we compare the entry values of the estimated label vector and use the indicator number associated with the largest entry as the emotion label of the testing feature vector. Assume that $\mathbf{d}_t$ is a testing feature vector and $\mathbf{l}_t = [l_{t1}, \cdots, l_{tc}]^T$ is the corresponding label vector estimated by

$$\arg\min_{\mathbf{l}_t} \|\mathbf{l}_t - \mathbf{C}\mathbf{d}_t\|_2^2, \mathrm{s.t.}\mathbf{l}_t^T\mathbf{1} = 1, \mathbf{l}_t \succeq \mathbf{0}.$$

Denote the associated emotion label of the feature vector $\mathbf{d}_t$ by c^*, then c^* can be obtained by the following criterion:

$$c^* = \arg\max_i \{l_{ti}\}. \quad (8)$$

IV. Experiments

In this section, two emotional speech databases will be used to evaluate the performance of ISLSR in speech emotion recognition. The first database is the eNTERFACE audio-visual Emotion Database [16], and the second one is the FAU AIBO emotion corpus [17][18]. In the experiments, we use three state-of-the-art classifiers, i.e., Nearest Neighbor (NN) classifier, support vector machine (SVM) [3], and sparse representation classifier (SRC) [20], as the baseline classifiers. In addition, we also combine each of the three baseline classifiers with the Fisher's discriminant criterion [21] based feature selection operation, which results in three new classifiers, i.e., Fisher+NN, Fisher+SVM, and Fisher+SRC. Then, we evaluate the speech emotion recognition performance of ISLSR compared with the results of the above six classifiers.

A. Brief Reviews of Emotional Speech Databases

The eNTERFACE database is a public, audiovisual, and acted emotion database. It contains 6 archetypal emotions defined by Ekman, i.e., happiness, sadness, surprise, anger, disgust and fear, which are acted by 46 subjects with the pre-defined spoken content in English. We discard the video samples of 3 subjects whose emotions are not clearly recognized. Moreover, only the audio data of 1287 video samples from 43 subjects are used in the experiments. It should be noted that there is a long silent segment at both ends of each video in the database. Hence, we first remove the silent segment from each speech utterance in order to improve the emotion classification performance. After the pre-processing procedure, we extract 1582 features from each utterance by using the OpenSmile toolkit and concatenate them into a 1582-dimensional feature vector and then normalize them.

The FAU Aibo Emotion Corpus was recorded in real conditions with children's spontaneous, emotionally coloured speech. The children were led to play and interact with the Sony's pet robot Aibo which was actually controlled by a human operator to elicit the children's natural speech. The data was recorded at two different schools from 51 children. The recordings were labeled by 5 human labellers on the word level belonging to 11 emotions (e.g. joyful, surprised, motherese, neutral, bored, emphatic, helpless, touchy, reprimanding, angry, other). However, some words were not labeled due to the lacks of majority voting (MV). We also extract 1582 features from each utterance using openSmile toolkit and concatenate them into a 1582-dimensional feature vector and then normalize the feature vectors.

B. Experimental Results

We use subject independent cross-validation strategy to compare the recognition performance among the aforementioned seven speech emotion recognition methods. Specifically, for each of the two speech databases, we partition all the subjects into five subsets with approximately equal size. Then, three of five subsets are selected to constitute the training data set, and one is selected as the testing data set, and the last subset is used as the development data set. For the training data set, the emotion class labels are known, whereas for testing data set or development data set, the emotion class label information are unknown. The target of our experiments is to predict the emotion class labels of testing data samples based on both the training data set and the development data set. It should be noted that the development data set can not be used for the other classifiers except for ISLSR because the class labels of samples in the development data set are unknown.

The experiments are repeated 20 trials such that they cover all the possible cases of training data set, development data set, and testing data set. After the experiments, we use the the average recall and the average precision over the 20 trials to evaluate the emotion classification performance. Table I shows the experimental results of the various classification methods on the eNTERFACE database and Table II shows the results on the

TABLE I
THE AVERAGE RECOGNITION ACCURACY (%) OF VARIOUS METHODS ON THE ENTERFACE DATABASE

Methods	Overall number of features	Experimental Results							
		Anger	Disgust	Fear	Happy	Sad	Surprise	Average Recall	Average Precision
NN	1582	72.22	47.84	46.50	40.81	67.23	42.97	52.94	52.97
Fisher+NN	1500	72.40	48.31	46.83	42.01	67.51	42.86	53.32	53.34
SVM	1582	79.85	58.27	57.93	62.27	64.22	58.07	63.44	63.43
Fisher+SVM	1500	80.41	57.59	56.93	62.51	64.06	57.19	63.12	63.11
SRC	1582	75.68	60.97	57.66	52.41	66.53	50.12	60.56	60.60
Fisher+SRC	1500	75.84	62.06	57.60	52.92	66.06	50.40	60.81	60.85
ISLSR	648	90.09	59.51	60.61	71.80	68.40	65.59	**69.33**	**69.32**

TABLE II
THE AVERAGE RECOGNITION ACCURACY (%) OF VARIOUS METHODS ON THE FAU AIBO EMOTION CORPUS

Methods	Overall number of features	Experimental Results					
		Subsuming angry touchy, reprimanding	Emphatic	Neutral	Motherese	Average Recall	Average Precision
NN	1582	44.27	45.72	61.00	44.64	48.91	52.82
Fisher+NN	1500	44.69	46.21	61.37	44.44	49.18	53.17
SVM	1582	46.85	49.89	69.69	41.97	52.10	57.75
Fisher+SVM	1500	46.85	50.13	69.57	41.94	52.12	57.80
SRC	1582	36.37	39.22	90.10	24.86	47.64	60.03
Fisher+SRC	1500	36.65	38.91	90.15	24.72	47.61	60.02
ISLSR	487	65.98	54.29	57.40	64.35	**60.50**	**60.25**

FAU Aibo emotion corpus. In both Table I and II, we also show the overall number of features selected during the experiments.

From Table I, we can see that the proposed ISLSR method achieves much better results than the other classifiers in terms of the average recall and the average precision, in which the average recall and the average precision of ISLSR are 69.33% and 69.32%, respectively. The better recognition results of ISLSR can also be observed in Table II, from which we can see that the ISLSR method achieves the average recall and the average precision are 60.50% and 60.25%, respectively. These results are competitive to the best ones obtained by the other classifiers.

V. CONCLUSIONS

In this letter, we have proposed a novel ISLSR method for speech emotion recognition. Different from most of the traditional speech emotion recognition methods such as SVM or SRC, the proposed method makes use of both labeled and unlabeled data samples to enhance the compatibility of the model, such that it is more powerful to predict the emotion class labels of the testing data samples. To evaluate the recognition performance of the proposed method compared with the state of the art classifiers, we conducted experiments on two speech emotion databases. The experimental results on both databases demonstrate that the proposed method achieves better or competitive recognition performance in contrast to the other classifies. A major reason of the better classification performance of ISLSR is most likely due to the use of development data set during model training, which enhances the compatibility of the recognition model for out-of-sample data.

REFERENCES

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech*, 2005, pp. 4–8.
[2] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Patt. Recognit.*, vol. 44, pp. 572–587, 2011.
[3] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Disc.*, vol. 2, no. 2, pp. 121–167, 1998.
[4] W. Zheng and X. Zhou, "Speech emotion recognition based on kernel reduced-rank regression," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2012, pp. 1972–1976.
[5] C. M. Bishop, *Pattern Recogntion and Machine Learning*. Berlin, Germany: Springer, 2006.
[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.
[7] J. Yang, L. Zhang, Y. Xu, and J.-Y. Yang, "Beyond sparsity: The role of L_1-optimizer in pattern classification," *Patt. Recognit.*, vol. 45, pp. 1104–1118, 2012.
[8] Y. Li and Y. Zhao, "Recognizing emotions in speech using short-term and long-term features," in *Proc. Fifth Int. Conf. Spoken Language Processing*, 1998, pp. 2255–2258.
[9] D.-N. Jiang and L.-H. Cai, "Speech emotion recognition with the combination of statistical features and temporal features," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2004, pp. 1967–1970.
[10] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.
[11] B. Schuller, S. Steidl, and A. Batliner *et al.*, "The Interspeech 2010 paralinguistic challenge," in *Proc. Interspeech*, 2010, pp. 2794–2797.
[12] Z. Qin and d. Goldfarb, "Structured sparsity via alternating direction methods," *J. Mach. Learn. Res.*, vol. 13, pp. 1333–1367, 2012.
[13] Z. Lin, M. Chen, L. Wu, and Y. Ma, The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices UIUC, Tech Rep. UILU-ENG-09-2215, 2009.
[14] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 35, no. 1, pp. 172–184, 2013.
[15] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical Mathematics*. New York, NY, USA: Springer, 2000.
[16] O. Martin *et al.*, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd IEEE Int. Conf. Data Engineering Workshops*, 2006.
[17] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.
[18] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin, Germany: Logos Verlag.
[19] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 19, no. 2, pp. 154–158, 1997.
[20] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.
[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Cassification*. New York, NY, USA: Wiley, 2001.