Connecting subspace learning and extreme learning machine in speech emotion recognition

Xinzhou Xu, Jun Deng, Eduardo Coutinho, Chen Wu, Li Zhao, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Xu, Xinzhou, Jun Deng, Eduardo Coutinho, Chen Wu, Li Zhao, and Björn Schuller. 2019. "Connecting subspace learning and extreme learning machine in speech emotion recognition." *IEEE Transactions on Multimedia* 21 (3): 795–808. https://doi.org/10.1109/tmm.2018.2865834.

Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition

Xinzhou Xu, Jun Deng, Eduardo Coutinho, Chen Wu, Li Zhao, and Björn W. Schuller, *Fellow, IEEE*

***Abstract*—Speech emotion recognition (SER) is a powerful tool for endowing computers with the capacity to process information about the affective states of users in human–machine interactions. Recent research has shown the effectiveness of graph embedding-based subspace learning and extreme learning machine applied to SER, but there are still various drawbacks in these two techniques that limit their application. Regarding subspace learning, the change from linearity to nonlinearity is usually achieved through kernelization, whereas extreme learning machines only take label information into consideration at the output layer. In order to overcome these drawbacks, this paper leverages extreme learning machines for dimensionality reduction and proposes a novel framework to combine spectral regression-based subspace learning and extreme learning machines. The proposed framework contains three stages—data mapping, graph decomposition, and regression. At the data mapping stage, various mapping strategies provide different views of the samples. At the graph decomposition stage, specifically designed embedding graphs provide a possibility to better represent the structure of data through generating virtual coordinates. Finally, at the regression stage, dimension-reduced mappings are achieved by connecting the virtual coordinates and data mapping. Using this framework, we propose several novel dimensionality reduction algorithms, apply them to SER tasks, and compare their performance to relevant state-of-the-art methods. Our results on several paralinguistic corpora show that our proposed techniques lead to significant improvements.**

***Index Terms*—Speech emotion recognition, extreme learning machine, subspace learning, graph embedding, spectral regression.**

I. Introduction

In MANY Machine Learning problems, particularly those involving real world applications, researchers have to deal with high or very high dimensional data, particularly in the context of ‘Big Data’ analytics [1]. Unfortunately, due to computational constraints, dealing with such large feature spaces can be impractical (if not impossible). Therefore, it is essential to explore alternative representations (i.e., lower dimensionality) that are computationally manageable whilst maintaining the relevant information about the original feature spaces. This need is apparent in many human-computer interaction tasks such as Speech Emotion Recognition (SER), which aims to detect emotional information conveyed by the human voice [2]–[5], but many of the typically used acoustic features contain information that may not be relevant for emotion recognition [6], [7]. Thus SER requires new methods for obtaining factors specifically related to emotional representation from large feature spaces [6], [8].

Motivated by this need, in recent years a multitude of techniques for dimensionality reduction have been proposed, and subspace learning has become a central topic in machine learning. Prominent algorithms in the field of pattern recognition are *Graph Embedding* (GE; [9], [10]) and *Spectral Regression* (SR; [11]–[13]). These two classes of algorithms include popular dimensionality reduction techniques, such as, *Principal Component Analysis* (PCA), *Fisher Discriminant Analysis* (FDA), *Linear Discriminant Projections* (LDP) [14], *Locality Preserving Projections* (LPP; [15]), *Locally Discriminant Embedding* (LDE; [16]), and *Graph-based Fisher Analysis* (GbFA; [17]), all of which can be seen as particular cases (i.e., graph structures) of a general GE framework [9], [10], [17]. At their core, both GE and SR utilise embedding graphs to calculate projections for optimal subspaces. However, GE computes these projections directly, whilst SR makes use of regression on spectral coordinates to calculate the projections with the goal of achieving computational efficiency and better performance [12], [13], [18], [19]. The linear and kernelised forms of GE and SR are

This work was supported in part by the China Scholarship Council; in part by the European Union’s Seventh Framework Programme under Grant Agreement 338164 (ERC Starting Grant iHEARu); in part by the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement 645378 (ARIA-VALUSPA) and Grant Agreement 645094 (SEWA); in part by the Natural Science Foundation of China under Grant 61673108, Grant 61231002, and Grant 11701290; in part by the Natural Science Foundation for Jiangsu Higher Education Institutions under Grant 16KJB510031 and Grant 17KJB110012; and in part by the NUPTSF under Grant NY217149 and Grant NY217150. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Martha Larson. *(Corresponding author: Xinzhou Xu.)*

X. Xu is with the College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China, with the Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China, and also with the Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Munich 80290, Germany (e-mail: xinzhou.xu@tum.de).

J. Deng is with the audEERING GmbH, Gilching 82205, Germany (e-mail: jdeng@audeering.com).

E. Coutinho is with the Department of Computing, Imperial College London, London SW7 2AZ, U.K., and also with the Department of Music, University of Liverpool, Liverpool L69 3BX, U.K. (e-mail: e.coutinho@imperial.ac.uk).

C. Wu is with the College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: wuchen@njupt.edu.cn).

L. Zhao is with the Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China (e-mail: zhaoli@seu.edu.cn).

B. W. Schuller is with the Group on Language, Audio & Music, Imperial College London, London SW7 2AZ, U.K., and also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany (e-mail: schuller@ieee.org).

known as *Linear GE* (LGE), *Kernel GE* (KGE) [10], *Linear SR* (LSR; [11], [13]) and *Kernel SR* (KSR; [20]), respectively.

A. Extreme Learning Machines

Recently, *Extreme Learning Machines* (ELMs) have been introduced as an alternative approach [21]–[26] for emotion recognition tasks [27], [28]. ELMs' working principles are similar to those of Single-hidden Layer Feedforward Neural Networks (SLFNs; [21], [24]). However, from an optimisation perspective, the calculation of the weights to the output layer in ELM is related to *Support Vector Machines* (SVMs), *Least Square Support Vector Machines* (LSSVM; [23], [25]), and *Ridge Regression* (RR; [29]–[31]). Compared to SLFNs, a major advantage of ELMs is to generate the input to hidden layer weights directly without training, as well as adopting RR for hidden to output layer weights in order to increase computational speed [23], [25]. Motivated by the appealing characteristics of ELM and SR, various attempts have been made to apply ELM methods to subspace learning. For example, [30], [31] computed a low-dimensional space directly using spectral regression, while *Huang et al.* [32] and *Iosifidis et al.* [33] improved the optimisation procedures of ELMs by adding additional terms for regression.

In spite of the clear benefits of GE, SR, and ELM, there are evident *limitations* associated with the application of these methods in subspace learning. First, the nonlinear extension in GE and SR relies on kernelisation [10], [20], which in essence maps the original feature space to a new space represented by training samples. This process may lead to decreased performance when the training data maps the original features to a poorly represented space. Second, with respect to ELM, the calculation of the decision values at the output layer depends exclusively on the labels [21], [23], whereas the unsupervised relationship between training samples (addressed by most GE based methods, e.g., neighbouring information) is simply ignored. This leads to a limitation related to correctly representing the structure of data according to GE related methods [10], [15]–[17].

B. Overview of This Paper

Inspired by previous work showing that ELM can be adapted for subspace learning by employing different optimisation structures [23], [25], [30]–[33], in this paper we propose a novel framework – *Generalised Spectral Regression* (GSR) – that exploits the relationships between ELM and subspace learning to overcome the above mentioned drawbacks associated with both methods.

We use this framework to design several embedding graphs for SER tasks at the stage of graph decomposition, which takes both of neighbouring and supervised information from training samples into consideration [16], [17], in order to construct suitable graph structures representing the inherent properties of the data. We argue that this makes the system more robust and effective by providing additional unsupervised information. Then, the proposed methods are evaluated and compared to other state-of-the-art methods in multiple SER tasks. The main contributions of this paper are:

- A new framework (GSR) for dimensionality reduction combining ELM and subspace learning;
- A set of novel feature reduction algorithms specifically developed to improve the performance of SER systems;
- A demonstration of the effectiveness of the proposed framework and methods in the context of SER, including a thorough comparison with existent methods.

The remainder of this paper is organised as follows. In Section II, we review the essential background for this work, including the notation used, and the basic principles of ELM, GE frameworks, and SR. In Section III, the proposed GSR framework is presented in detail, and the different GSR strategies developed are described in Section IV. Then, in Section V, we introduce the experimental methodology, and in Section VI we evaluate the performance of the algorithms on multiple SER corpora. Finally, in Section VII we discuss our work, and propose future research directions.

II. Background

In this section, we introduce the basic concepts, variables, and notations used throughout this article.

Let $X = [x_1, x_2, \ldots, x_N] \in \Re^{n \times N}$ and $Y = [y_1, y_2, \ldots, y_N] \in \Re^{d \times N}$ be sets of N labelled training samples in the original feature space with dimensionality n and the lower-dimensional feature space with the dimensionality d, respectively. Each column of $\phi(X) = [\phi(x_1), \phi(x_2), \ldots, \phi(x_N)]$ is the *Reproducing Kernel Hilbert Space* (RKHS) of the corresponding column in X. The Gram matrix K is defined as $\phi^T(X)\phi(X)$. It is assumed that all samples (training and test) in the original and reduced dimensionalities can be represented by column vectors $x \in \Re^{n \times 1}$ and $y \in \Re^{d \times 1}$, respectively. The RKHS of x is defined as $\phi(x)$. For sample x, its kernelised coordinate is $K_x = \phi^T(X)\phi(x)$.

Each column of the label matrix $S = [s_1, s_2, \ldots, s_N] = [\widehat{s}_1, \widehat{s}_2, \ldots, \widehat{s}_c]^T \in \Re^{c \times N}$ represents the labelling information of each training sample, where c is the number of classes. $S_{ij} = 1$ when sample j belongs to class i, otherwise $S_{ij} = 0$, where $i = 1, 2, \ldots, c$ and $j = 1, 2, \ldots, N$. $I \in \Re^{N \times N}$ is the identity matrix. Every element of $e \in \Re^{N \times 1}$ is equal to 1.

A. Extreme Learning Machines

ELM [21]–[24] assumes that L is the number of hidden neurons. $H = [h_1, h_2, \ldots, h_N] \in \Re^{L \times N}$ represents the outputs of the hidden neurons pertaining to the sets of N training samples, with each column representing one training sample in the feature space generated by the input nodes. The label coordinate matrix of extreme learning is $T = 2S^T - ee_c^T$, where each element of $e_c \in \Re^{c \times 1}$ is equal to 1. The hidden neurons' parameters are selected randomly [24], and the activation functions of these hidden neurons can be sigmoidal, hard limit or Gaussian (for more details please refer to [21]–[25]). The typical optimisation function of extreme learning is represented as

$$\min_{\beta} \left(\| \beta \|_F^2 + \ C \| H^T\beta - T \|_F^2 \right), \tag{1}$$

where, $\beta \in \Re^{L \times c}$ represents the matrix of output weights, and $C > 0$ is a constant value controlling the relation between the Frobenius norm of the coefficients β and the linear regression term. The optimal value of β is determined by RR:

$$\beta^* = \left(\frac{I_L}{C} + HH^T \right)^{-1} HT, \tag{2}$$

where I_L is the identity matrix with the dimensionality of L.

In order to reduce the computational cost, we obtain the optimal β as

$$\beta^* = \begin{cases} \left(\frac{I_L}{C} + HH^T\right)^{-1} HT, & L < N, \\ H\left(\frac{I}{C} + H^T H\right)^{-1} T, & L \geq N. \end{cases} \tag{3}$$

B. Graph Embedding Frameworks

GE frameworks [10] aim to find optimal embedding graphs in tandem with data mapping types and optimisation forms, to unveil the internal structure of a given data set. The optimisation forms of GE frameworks with penalty and scaling constraints are shown, respectively, in Eqs. (4) and (5):

$$\min \sum_{i,j=1}^{N} \parallel y_i - y_j \parallel^2 W_{ij}^{(I)} \text{ s.t. } \sum_{i,j=1}^{N} \parallel y_i - y_j \parallel^2 W_{ij}^{(P)} = \mu, \tag{4}$$

$$\min \sum_{i,j=1}^{N} \parallel y_i - y_j \parallel^2 W_{ij}^{(I)} \text{ s.t. } \sum_{i=1}^{N} {y_i}^2 D_{ii} = \mu, \tag{5}$$

where $W^{(I)} \in \Re^{N \times N}$ and $W^{(P)} \in \Re^{N \times N}$ are the adjacency matrices of the intrinsic graph and the penalty graph. D is a diagonal matrix to control weights of samples, and μ is a positive constant value. With one mapping direction $a \in \Re^{n \times 1}$ for sample i, $y_i = a^T x_i$. For the training set with multiple mapping directions $A = [a_1, a_2, \ldots, a_d] \in \Re^{n \times d}$, $Y = A^T X$. The kernelised form of Y can be written as $Y = A_K^T \phi^T(X)\phi(X) = A_K^T K$, where $A_K = [a_{K1}, a_{K2}, \ldots, a_{Kd}] \in \Re^{N \times d}$ is the kernelised mapping. We reformulate the optimisation function (Eq. (4)) to obtain the optimal one-dimensional new features of N samples

$$z^* = \arg\min_z \quad \frac{zL^{(I)} z^T}{zL^{(P)} z^T}, \tag{6}$$

where $z \in \Re^{1 \times N}$. $L^{(I)} = D^{(I)} - W^{(I)}$, with each element of the diagonal degree matrix $D^{(I)} \in \Re^{N \times N}$ as $D_{ii}^{(I)} = \sum_{j=1}^{N} W_{ij}^{(I)}$, where $i = 1, 2, \ldots, N$. Similarly, $L^{(P)} = D^{(P)} - W^{(P)}$ and the diagonal matrix $D^{(P)} \in \Re^{N \times N}$ contains elements $D_{ii}^{(P)} = \sum_{j=1}^{N} W_{ij}^{(P)}$.

The mapping coefficients connecting training and test data can be obtained by reformulating z directly in Eq. 6 and solving the *Generalised Eigenvalue Problem* (GEP; [10], [15]).

For **FDA**, given that $N \geq c$, the adjacency matrices of intrinsic and penalty graphs are represented as $W^{(I)} = W_{FDA}^{(I)} = S^T(SS^T)^{-1}S = \sum_{l=1}^{c}(\widehat{s}_l^T e)^{-1}\widehat{s}_l\widehat{s}_l^T$ and $W^{(P)} = W_{FDA}^{(P)} = \frac{1}{N}ee^T$. Similarly for **LDP**, $W^{(I)} = W_{LDP}^{(I)} = S^T S$ and $W^{(P)} = W_{LDP}^{(P)} = ee^T - S^T S$.

C. Spectral Regression

SR [13], [18], [19] is a two-stage process developed to solve GE problems efficiently that divides the GE solution into spectral graph learning and regression. By setting the new dimensionality of the feature space as one, according to Eq. (6), we can draw a new, optimal one-dimensional feature vector of training samples, written as z^*.

In the linear case, assuming $z = a^T X$, the optimal linear mapping vector a^* can be obtained by a least-square form. However, the least-square solution is often not satisfactory, when the dimensionality of features n is higher than the number of training samples N. Thus we obtain the optimal a as

$$a^* = \arg\min_a \left(\parallel a \parallel^2 + \ C \parallel X^T a - z^{*T} \parallel^2 \right). \tag{7}$$

It has been shown in [18] that when $\frac{1}{C}$ decreases to zero, a^* turns to be the optimal solution for the LS Regression. Eq. (7) also can be solved by RR. In the kernelised form of SR (i.e., KSR), the optimisation method is changed to obtain the optimal kernelised mapping as

$$a_K^* = \arg\min_{a_K} \left(\parallel a_K \parallel^2 + C \parallel K a_K - z^{*T} \parallel^2 \right). \tag{8}$$

III. Proposed Framework

In this section, we describe the proposed GSR framework which consists of three stages: 1) *data mapping*; 2) *graph decomposition*; and 3) *regression*. At the **first** stage, the features of a give sample are projected into a new feature space, and we show that kernel and linear forms in subspace learning [34] follow similar rules as the input-hidden layer procedures in ELMs. By including the data mappings adopted in LSR, KSR, and ELM, different types of data mapping can be employed to generate relevant underlying subspaces. At the **second** stage, we make use of embedding graphs [10] to automatically generate virtual feature vectors for each training sample [35]. This process creates corresponding coordinates of the embedding graphs, which reflect the relationship between each pair of training samples. The goal is to enhance the performance on the target task by better depicting the salient structure in feature space of the data set. Finally, at the **third** stage, in order to construct a connection between the mapping data and the virtual feature vectors, the subspace mapping matrix is learnt by fitting the new feature space to virtual coordinates employing different regression algorithms.

A. Data Mapping

By representing the data mapping of sample x as $f(x)$, we draw the mapping $x \rightarrow f(x)$ for the cases of kernelisation and ELM. Note that for the linear case, the mapping is written as $x \rightarrow x$ directly.

Data mapping in kernelisation

In the linear domain, the lower-dimensional features of N training samples are $Y = A^T X$, assuming that $A = \phi(X)A_K + \phi_\perp(X)A_K'$, where the columns of $\phi_\perp(X)$ represent the basis of the null space of $\{\phi(x_1), \phi(x_2), \ldots, \phi(x_N)\}$, and

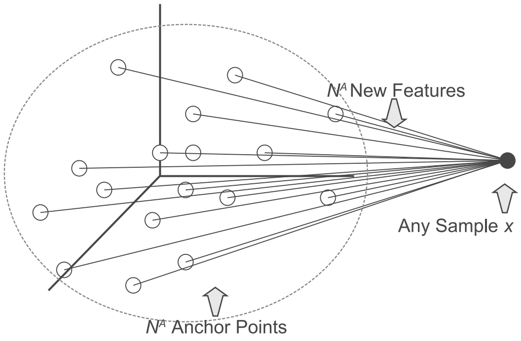


Fig. 1. Generation of N^A new features by N^A anchor points for a certain sample x when the dimensionality of the original feature space is $n = 3$.

A'_K contains their linear coefficients. Then $Y = A_K^T K$, where X is transformed into its RKHS $\phi(X)$.

According to the equation $Y = A_K^T K$, kernel tricks essentially employ a type of data mapping, which maps the original feature space into a new space created from multiple views of all the training samples. The dimensionality of the linear feature mappings in the dimensionality reduction process also changes to fit the dimensionality of the new space.

Further, on the assumption that $\phi_\perp^T(X)\phi(x) = 0$, the *data mapping in kernelisation* for sample x is drawn as

$$x \to f(x) = \phi^T(X)\phi(x) = \mathbf{k}_x, \tag{9}$$

where the column vector $\mathbf{k}_x \in \Re^{N\times 1}$. It can thus be concluded from Eq. (9) that the newly generated features of sample x are written in the form of

$$\mathbf{k}_x = \left[\phi^T(x_1)\phi(x), \phi^T(x_2)\phi(x), \ldots, \phi^T(x_N)\phi(x)\right]^T. \tag{10}$$

Anchor points

In typical kernelisation methods, kernel tricks only present a fixed set of bases in nonlinearisation. Moreover, $\phi_\perp^T(X)\phi(x) = 0$ does not always hold when x is excluded from the columns of X. Hence, it is feasible to solve the problems through changing the columns of $\phi(X)$.

We define a set of N^A anchor points $X^A = [x_1^A, x_2^A, \ldots, x_{N^A}^A]$ to replace X, where usually $N^A > N$ and $x_i^A \in \Re^{n\times 1}$ with $i = 1, 2, \ldots, N^A$. This leads to a set of vectors $\left\{\phi(x_1^A), \phi(x_2^A), \ldots, \phi(x_{N^A}^A)\right\}$ in RKHS, as

$$\phi(X^A) = \left[\phi(x_1^A), \phi(x_2^A), \ldots, \phi(x_{N^A}^A)\right]. \tag{11}$$

We assume that the basis of $\left\{\phi(x_1^A), \phi(x_2^A), \ldots, \phi(x_{N^A}^A)\right\}$ is able to span the space of $\{\phi(x_1), \phi(x_2), \ldots, \phi(x_N)\}$, and $\phi_\perp^T(X^A)\phi(x) = 0$. Therefore, the new features generated by the anchor points are

$$\mathbf{k}_x^A = \phi^T(X^A)\phi(x). \tag{12}$$

Fig. 1 illustrates the generation of new features when anchor points are provided. For a given sample x, its new features can be generated by N^A views of N^A anchor points with the same dimensionality.

Data mapping in ELM

For ELM, the number of anchor points N^A is exactly equal to the number of hidden neurons L. Assuming that the coordinates of the anchor points are the columns of $\Psi = [\psi_1, \psi_2, \ldots, \psi_L] \in \Re^{n\times L}$, the *data mapping of ELM* is

$$x \to f(x) = g(\phi^T(\Psi)\phi(x)) = h_x, \tag{13}$$

where $h_x \in \Re^{L\times 1}$ is the L-dimensional feature vector in the newly generated space. $g(\cdot)$ represents a certain mapping since the output of the activation function is not always nonnegative.

For sample x in ELM, the implicit inner product $\phi^T(\Psi)\phi(x)$ can be represented as an explicit new feature vector h_x. The explicit mapping depends on random anchor points in ELM [21]–[24]. The theory of ELM also indicates that it is possible to select the parameters randomly in the mapping forms. Consequently, the original anchor points and the model parameters can be chosen as random values.

B. Graph Decomposition

We define the embedding graphs as $G^{(I)}$ and $G^{(P)}$. The former represents intrinsic information, whereas the latter represents penalty properties. With this notation, the optimal virtual one-dimensional coordinates are drawn as

$$z^* = \arg\min_z \quad \text{or} \quad \arg\max_z \quad \frac{zG^{(I)}z^T}{zG^{(P)}z^T}, \tag{14}$$

where the selection of minimisation and maximisation, as well as the elements in graphs $G^{(I)}$ and $G^{(P)}$, depends on practical requirements as in [10], [15], [36], [37]. This optimisation process can be adapted to solve the GEP

$$G^{(I)}z^T = \lambda G^{(P)}z^T, \tag{15}$$

where λ is the eigenvalue of the GEP. $G^{(I)}$ and $G^{(P)}$ are the matrices of the intrinsic and penalty graphs, without considering the trivial eigenvectors.

For multiple locally optimal zs in constructing the optimal subspace, we define the zs as the 'virtual coordinates' of training samples. Thus, we get $z_i \in \Re^{1\times N}$, where $i = 1, 2, \ldots, d$, and

$$\left[z_1^T, z_2^T, \ldots, z_d^T\right]^T = Z \in \Re^{d\times N}, \tag{16}$$

which leads to the optimal Z, namely Z^*, by

$$Z^* = \arg\min_Z \quad \text{or} \quad \arg\max_z \quad \frac{tr\left(ZG^{(I)}Z^T\right)}{tr\left(ZG^{(P)}Z^T\right)}, \tag{17}$$

which is a *trace-ratio* problem and it can be solved either by iterative procedures with the orthonormalisation assumption [38], or by the the approximate *ratio-trace* form

$$Z^* = \arg\min_Z \quad \text{or} \quad \arg\max_z$$
$$tr\left(\left[ZG^{(P)}Z^T\right]^{-1}\left[ZG^{(I)}Z^T\right]\right), \tag{18}$$

when $ZG^{(P)}Z^T$ is invertible. This results in solving the GEP of Eq. (15), resorting to the Lagrangian method using *Singular Value Decomposition* (SVD). Theorem 1 shows the effectiveness on using the GEP to solve the optimisation, which can be proved by pre-multiplying z on both sides of Eq. (15).

Theorem 1: The nontrivial eigenvectors corresponding to minimal/maximal eigenvalues in Eq. (15) are the optimal solutions of Eq. (14).

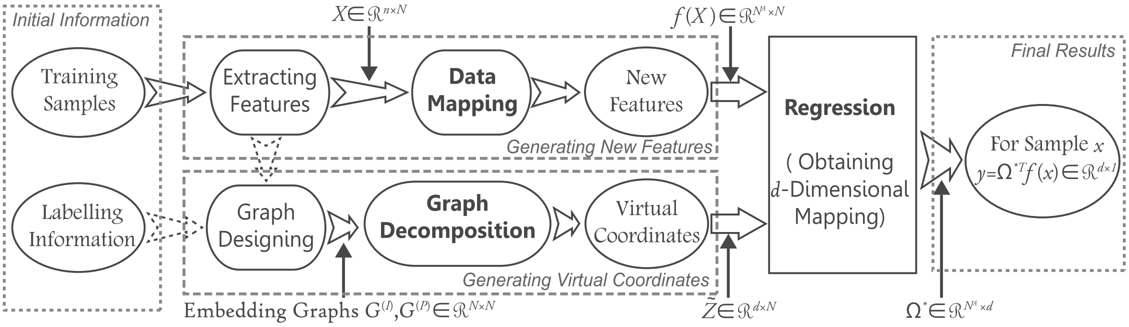

Initial Information
Training Samples
Labelling Information
$X \in \Re^{n\times N}$
Extracting Features
Data Mapping
New Features
Generating New Features
Graph Designing
Graph Decomposition
Virtual Coordinates
Generating Virtual Coordinates
Embedding Graphs $G^{(I)}, G^{(P)} \in \Re^{N\times N}$
$f(X) \in \Re^{N^A\times N}$
$\tilde{Z} \in \Re^{d\times N}$
Regression
(Obtaining d-Dimensional Mapping)
Final Results
For Sample x
$y=\Omega^{*T}f(x) \in \Re^{d\times 1}$
$\Omega^{*} \in \Re^{N^A\times d}$

Fig. 2. Schematic diagram of the proposed GSR framework. The features of training samples X are utilised in data mapping and sometimes in graph design. Given the set of new features and a set of virtual coordinates, we can predict y for sample x in the regression process.

It should be noticed that the newly generated coordinates can be processed before the stage of regression (as in ELM and SR). The processing is defined as

$$Z \rightarrow \widetilde{Z} = [\widetilde{z}_1^T, \widetilde{z}_2^T, \ldots, \widetilde{z}_d^T]^T, \tag{19}$$

with each row corresponding to the respective row in Z. It is also feasible to employ normalisation or orthonormalisation using their corresponding constraints on each column of Z at this stage [39], [40].

Graph decomposition in SR and ELM

In the context of SR, the work presented in [20] shows that the graph-decomposition solution is straightforward to obtain by solving Eq. (17), with $\widetilde{Z} = Z$. According to previous work on ELM [21], [23], [25] and SR [11], [13], [18], [20], $S^T(SS^T)^{-1}S$ and I can be used as the embedding graphs. The lth virtual set of coordinates $z_l^T = \widehat{s}_l$, where the elements of $\widehat{s}_l$ corresponding to the lth class are equal to 1, and all the other elements are equal to 0, with $l = 1, 2, \ldots c$. This solution is presented in Theorem 2, and the proof is shown in Appendix A. For ELM, d is fixed as c, whilst for the SRs (in [11], [13], [18], [20]) d is equal to $c - 1$.

Theorem 2: For SR (in [11], [13], [18], [20]) and ELM, one selection of the embedding graphs is: $G^{(I)} = W_{FDA}^{(I)} = S^T(SS^T)^{-1}S$ and a scaling diagonal matrix $G^{(P)} = I$.

C. Regression

Note that $\omega \in \Re^{N^A\times 1}$ is the mapping direction for dimensionality reduction on samples, and ω_i^* is its optimal value corresponding to the virtual coordinate vector $\widetilde{z}_i$. As in elastic nets [11], [41], the unified regression form of GSR is defined to calculate

$$\omega_i^* = \arg\min_{\omega} \left(\| f^T(X)\omega - \widetilde{z}_i^T \|^2 + \gamma_1 \| \omega \|_1 + \gamma_2 \| \omega \|^2 \right), \tag{20}$$

where $f(X) = [f(x_1), f(x_2), \ldots, f(x_N)]$ is the mapping set of training samples X. $i = 1, 2, \ldots, d$ represents the ith dimension of the virtual coordinates generated at the stage of graph decomposition. γ_1, $\gamma_2 \geq 0$ are the constant weights for the $l1$-norm and $l2$-norm minimisation terms.

Regarding the choices of parameters γ_1 and γ_2, if $\gamma_1 = \gamma_2 = 0$, Eq. (20) becomes an LS Regression problem. When $\gamma_1 = 0$ and $\gamma_2 > 0$, this equation changes into an RR problem, whilst if $\gamma_1 > 0$ and $\gamma_2 = 0$ it becomes a *Least absolute shrinkage and selection operator* (Lasso) problem [42]. These forms are minimised with various norms on the basis of SR.

Finally, $\Omega = [\omega_1, \omega_2, \ldots, \omega_d]$ represents the dimensionality-reduced mapping matrix.

Regression in SR and ELM

In [11] (*Unified Sparse Subspace Learning framework* (USSL)), $\gamma_1 > 0$ and $\gamma_2 = 0$. Eq. (20) can be solved by the *LeastAngleRegression* (LARS) algorithm [11]. In previous SR [13], [18], [20], [35] and ELM [21], [23], [25] research, these parameters were set as $\gamma_2 > 0$ and $\gamma_1 = 0$, and consequently the solution can be written in the form of a column vector:

$$\omega_i^* = \left(\frac{I_{N^A}}{\gamma_2} + f(X)f^T(X) \right)^{-1} f(X)\widetilde{z}_i^T, \tag{21}$$

where I_{N^A} is the N^A-dimensional identity matrix, and z_i is equal to $\widehat{s}_i^T$, where $i = 1, 2, \ldots, c$.

However, for ELM, the decision procedure is similar to what has beenfrequently employed in neural networks. Instead of directly using nearest-neighbour classifiers, ELM implicitly assumes that for each training sample x_j, the regression $f^T(x_j)\alpha_i^* = (\widetilde{z}_i)_j$ always holds for $j = 1, 2, \ldots, N$. With this assumption, the decision process in ELM is equivalent to that of a nearest-neighbour classifier.

D. Generalised Spectral Regression Framework

In this section, we present the complete GSR framework, and demonstrate that several existing methods are particular instances of this framework.

As illustrated in Fig. 2, given the originally extracted features from N training samples X and a given sample x, we can obtain the data mapping $f(X)$ and $f(x)$ for samples X and x, respectively, as described in Section III-A. When designing embedding graphs, both X and their corresponding labels can be employed to construct suitable embedding graphs. The designed embedding graphs are utilised to generate virtual coordinates by solving the GEP in graph decomposition, as described in Section III-B. At the regression stage (see Section III-C), the linear d-dimensional mapping matrix Ω^* is obtained according to Eq. (20), and from it we derive the low-dimensional feature vector $y = \Omega^{*T} f(x)$ for sample x.

TABLE I
THE PARAMETERS $(f(\cdot), N^A, G^{(I)}, G^{(P)}, \widetilde{Z}, \gamma_1, \gamma_2)$ OF FREQUENTLY USED SRS AND ELMS IN THE GSR FRAMEWORK

Stages	Data Mapping		Graph Decomposition			Regression	
Parameters	$f(x)$	N^A	$G^{(I)}$	$G^{(P)}$	$Z \to \widetilde{Z}$	γ_1	γ_2
LSR [13], [18]	$x \to x$	n	$S^T(SS^T)^{-1}S$	I	$\widetilde{Z}e = \mathbf{0}_{d\times 1}$, $\widetilde{z}_i\widetilde{z}_j^T = 0$ $(i, j = 1, 2, \dots d)$	$\gamma_1 = 0$	$\gamma_2 \neq 0$
LSR (USSL) [11]	$x \to x$	n				$\gamma_1 \neq 0$	$\gamma_2 = 0$
KSR [20]	$x \to \mathbf{k}_x$	N				$\gamma_1 = 0$	$\gamma_2 \neq 0$
LSR [35] *	$x \to x$	n	$\begin{bmatrix} S^T(SS^T)^{-1}S & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{\bar{N}\times\bar{N}}$	$\begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{\bar{N}\times\bar{N}} + L^{(0)}$	$\widetilde{Z} = Z$	$\gamma_1 = 0$	$\gamma_2 \neq 0$
ELM [21], [23], [25]	$x \to h_x$	L	$S^T(SS^T)^{-1}S$	I	$\widetilde{Z} = 2Z - e_c e^T$	$\gamma_1 = 0$	$\gamma_2 \neq 0$

*$\bar{N} = N + N_{UL}$: *the total number of training samples with* N_{UL} *unlabelled ones.* $L^{(0)}$: *Laplacian matrix of* $W^{(0)}$, *with* $W_{ij}^{(0)}$ *obeying that, when* x_i *and* x_j $(i, j = 1, 2, \dots N)$ *are labelled,* $W_{ij}^{(0)} = 1$ *if the labels are same, while* $W_{ij}^{(0)} = 0$ *if the labels are different; otherwise,* $W^{(0)}$ *is the* k*-nearest neighbours adjacency matrix.*

Table I shows how the proposed GSR can be parameterised to include LSR, KSR, and ELM, through regulating the data mapping form $f(\cdot)$, the number of anchor points N^A, the transformation from Z to $\widetilde{Z}$, as well as the regression parameters $\gamma_1, \gamma_2 \geq 0$ (where $i, j = 1, 2, \dots, d$). Note that, for the methods shown in Table I, we set the optimisation types in graph decomposition of Eq. (14) to maximum. The description of the remaining variables shown in Table I can be found in Section II.

IV. GSR APPLIED TO SPEECH EMOTION RECOGNITION

SER focuses on exploring relevant features and algorithms to infer the emotional state of speakers from paralinguistic information, i.e., the nonverbal aspects of speech (including speech prosody, voice quality, and other quantities estimated directly from the acoustic signal). In a typical framework, a set of (relevant) acoustic descriptors is extracted from spoken utterances, which are then used as features in estimating speakers' emotional states using machine learning methods.

System setup

At the stage of **data mapping**, the anchor points can be random coordinates or training samples. Considering the different data mapping methods, we define three GSRs as:*Random-anchor-points GSR* (RGSR), *Training-sample-anchor-points GSR* (TGSR), and*Linear GSR* (LGSR). In RGSR, the anchor points are randomly chosen (as in ELM), while TGSR utilises training samples as the anchor points (as in KSR).*Linear GSR* (LGSR) represents the data mapping $x \to x$ as in LSR.

At the stage of **graph decomposition**, several pairs of embedding graphs are proposed in the GSR framework, since the embedding graphs of ELM are only related to labelling information. As shown in Fig. 2, together with labelling information, features extracted from training samples are also available for graph design. This leads to the creation of k-nearest neighbour graphs or some other distance-based representations. It should be noticed that we only consider the fully supervised case, where each training sample is labelled by a single emotional class. The embedding graphs ($G^{(I)}$ and $G^{(P)}$) are obtained using pre-existent GE based algorithms – FDA [6], [10], LDP [14], LDE [16], GbFA [17], and *Locally Penalised Discriminant Analysis* (LPDA; [8]). The description of the FDA and LDP embedding graphs can be found in [6], [10], [14] (also introduced in Section II-B). The embedding graphs of LDE and GbFA are shown mathematically in Eqs. (22) and (23).

For LDE,

$$\begin{cases} W^{(I)} = W_{LDE}^{(I)} = S^T S \odot W_{k_1 NN}, \\ W^{(P)} = W_{LDE}^{(P)} = (ee^T - S^T S) \odot W_{k_2 NN}, \end{cases} \tag{22}$$

where the operator '$\odot$' represents the element-wise product between two matrices. $W_{k_1 NN}$ and $W_{k_2 NN}$ are k_1- and k_2- nearest neighbour adjacency matrices of training samples, where the elements $(W_{k_1 NN})_{ij} = 1$ or $e^{-\frac{\|x_i - x_j\|^2}{t}}$, when x_i is among k_1-nearest neighbours of x_j or vice versa, with $i, j = 1, 2, \dots, N$ and the constant value $t > 0$. Otherwise, $(W_{k_1 NN})_{ij} = 0$. $W_{k_2 NN}$ uses k_2-nearest neighbours, similar as in $W_{k_1 NN}$.

For GbFA,

$$\begin{cases} W^{(I)} = W_{GbFA}^{(I)} = S^T S \odot W_{Gram}, \\ W^{(P)} = W_{GbFA}^{(P)} = (ee^T - S^T S) \odot W_{Gram}, \end{cases} \tag{23}$$

where $(W_{Gram})_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$. Thus, nearest neighbour truncation is considered in GbFA.

The embedding graphs for LPDA [8] are given in Eq. (24). The intrinsic embedding graph of LPDA is designed as FDA, so as to remove the impact from the relatively inaccurate neighbouring information in SER. The penalty graph of LPDA aims to penalise neighbouring between-class sample pairs, similar as in LDE and GbFA.

For LPDA,

$$\begin{cases} W^{(I)} = W_{LPDA}^{(I)} = W_{FDA}^{(I)} = S^T(SS^T)^{-1}S, \\ W^{(P)} = W_{LPDA}^{(P)} = \frac{1}{N}ee^T + \delta_0 W_{LDE}^{(P)}, \end{cases} \tag{24}$$

where the constant value $\delta_0 > 0$ controls the relation between scattering (as in PCA) and the labelling penalty of training samples. It is equivalent to FDA when δ_0 deceases to zero.

In accordance with GE frameworks, we use the Laplacian matrices $L^{(I)}$ and $L^{(P)}$ of $W^{(I)}$ and $W^{(P)}$ (respectively) in graph decomposition. This implies setting the optimisation type to minimum for Eq. (14). In order to avoid the theoretically minimal zero value of the term $zG^{(I)}z^T$ for the intrinsic graphs, we add a term δI with a small value $\delta > 0$ in $G^{(I)}$. This leads to $G^{(I)} = \delta I + L^{(I)}$ and $G^{(P)} = L^{(P)}$. We only list each $W^{(I)}$ and $W^{(P)}$, for $G^{(I)}$ and $G^{(P)}$ (respectively) in Eq. (22), (23),

TABLE II
DESCRIPTION OF THE EMOTIONAL CORPORA GEMEP, ABC, VAM, AND eNTERFACE FOR THE AUDIO SECTIONS

Corpus	Language	Sampling Rate (kHz)	# Classes	# Speakers	# Samples	Means of Training
GEMEP [43]	French	44.1	18 (here 12)*	10 (5 female)	1 260 (here 1 080)	Training-Testing
ABC [44]	German	16	6	8 (4 female)	430	2-fold CV
VAM [45]	German	16	4	47 (36 female)	946	5-fold CV
eNTERFACE [46]	English	16	6	42 (here 40, 8 female)**	1 277 (here 1 200)	2-fold CV

We choose 12 classes according to the set in the INTERSPEECH 2013 Computational Paralinguistics Challenge *[47].*
***In order to make the two folds balance, we keep the samples from 40 speakers (each one contains 30 samples) in the experiments.*

and (24). The virtual coordinates relate to the eigenvectors corresponding to minimal eigenvalues [10], [15].

At the **regression** stage, we keep the form of RR employed in the SRs in [13], [18], [20], [35] and the ELMs in [21], [23]–[25]. This leads to the parameter settings as $\gamma_1 = 0$ and $\gamma_2 \neq 0$ in the regression of Eq. (20). The solution of the regression problem can be determined by Eq. (3) or the previously proposed *LSQR* method [13], [18], [35].

The proposed GSRs described above aim at obtaining multiple linear mapping directions reflecting emotional dimensions. However, general feature sets used in SER often include features that are relevant to other linguistic and paralinguistic application, e.g., *Automatic Speech Recognition* (ASR) or *Speaker Identification* (SI) [6]. Since the feature sets contain information that may not be relevant to SER, all the embedding graphs in the proposed methods include supervised information, considering that the target labels play a crucial role in determining the relevant features for the SER task. The new features are used in the procedure of classification or decision making.

Computational complexity

The computational complexity of the proposed methods primarily depends on graph decomposition and regression:

- **Graph decomposition**: The conventional GEP solution using SVD requires the complexity $O\left(N^3\right)$; when using the *fast Monte Carlo algorithm* [48], the complexity is $O\left(r^2N\right)$, where r is the predefined approximate rank in solving SVD.
- **Regression**: When directly solving RR, the maximal complexity is $O\big((\min(N,L))^3 + \min(N,L)LN + \min(N,L)Nd\big)$ for random data mapping, and is $O\left(2N^3 + N^2d\right)$ for kernelisation; when using LSQR, the complexity turns to be $O(NLDl_0)$ for random data mapping, while $O(N^2Dl_0)$ for kernelisation, where l_0 is the iteration number in LSQR.

V. EXPERIMENTAL METHODOLOGY

A. Selected Corpora

In our experiments, we use four corpora that are commonly used in SER tasks. Each corpus is described in the next paragraphs, and Table II summarises the key information.

GEMEP: The *GEneva Multimodal Emotion Portrayals* (GEMEP; [43], [47]) is a database of audio and video recordings featuring 10 actors portraying 18 affective states. In this work we use 12 of these states or classes–*amusement*, *pride*, *joy*, *relief*, *interest*, *pleasure*, *hot anger*, *panic fear*, *despair*, *irritation*, *anxiety*, *sadness*, as in [47]. The number of samples per class is 90. The full database was divided into training and test sets, with 648 (6 speakers; 3 female) and 432 (4 speakers; 2 female), respectively.

ABC: The *Airplane Behavior Corpus* (ABC; [44], [49]) contains speech recordings labelled in six emotions: *aggressive*, *cheerful*, *intoxicated*, *nervous*, *neutral*, *tired*. The numbers of samples in each emotional class are 95, 105, 33, 93, 79, 25, respectively, leading to a total number of 430 samples. The samples were obtained from 8 speakers (4 female). For our experiments, the full corpus was divided into two folds, each including the recordings of 4 speakers (2 female). Then, one fold is for training while the other for testing, and vice versa, which is equivalent to a 2-fold CV (Cross-Validation).

VAM: The *"Vera am Mittag"* German audio-visual emotional speech database (VAM; [45], [49]) includes 12 hours of spontaneous and very emotional audio-visual recordings of the German TV show "Vera am Mittag". The corpus includes natural speech utterances in four emotions: *happy* / *excited*, *angry* / *anxious*, *sad* / *bored*, and *relaxed* / *serene*, which are here referred to as *q1* (21 instances), *q2* (50 instances), *q3* (451 instances), and *q4* (424 instances), respectively, based on quadrants in the arousal/valence plane. The corpus includes recordings from a total of 47 different speakers (36 female). For the purposes of our work, the corpus was divided into 5 (speaker-independent) folds. In order to also balance for gender, the first fold contains the samples from 11 speakers, while all other folds contain the recordings of 9 speakers.

eNTERFACE: The *eNTERFACE'05* (eNTERFACE; [46], [49]) database contains speech utterances recorded in office environment expressing six basic emotions (*happiness*, *sadness*, *surprise*, *anger*, *disgust*, *fear*) as defined by Ekman *et al.* [50]. There are a total of 42 different speakers (8 female) in the database. In our work, we employed the samples from 40 speakers (8 female), and 200 samples per emotion category. Half of the samples were used for training while the rest were used for test (and vice versa; 2-fold CV).

B. Features

For our experiments we adopt the official feature set of the *INTERSPEECH 2013 Computational Paralinguistics Challenge* (ComParE; [47]), which includes 6 373 static features of functionals of 65 *Low-Level Descriptor* (LLD) contours (a thorough description of the feature set is given in [51]). All features were extracted with *openSMILE* (version 2.0; [52]), a framework for extracting general-purpose acoustic and prosodic features, which has been successfully applied in a variety of SER and other paralinguistic tasks.

C. Preprocessing

As discussed in Section IV, at the stage of **data mapping**, the anchor points are chosen as training samples for TGSR (KSR), or as random points for the RGSR (ELM) case. The number of anchor points for the TGSR based methods is N, while the number of anchor points L for RGSR based methods is set to 500, 1 000, 3 500, 5 000, and 10 000. In accordance with the min-max normalisation of the training samples in TGSR, we set each attribute of the random anchor points in RGSR ranging between 0 and 1. In addition, the random anchor points are the same for each method when the number of the points is fixed. The random choices of anchor points are repeated 10 times in our experiments. For each experiment, we only consider the best accuracies among the dimensions no larger than 100.

Similar as in existing SR and ELM approaches, the data mapping employs the popular *Gaussian Kernel* form. For sample x, the generated new features, represented as k_x and h_x for TGSR and RGSR respectively, are given as

$$\begin{cases} k_x = \left[e^{-\frac{\|x-x_1\|^2}{t_0}}, e^{-\frac{\|x-x_2\|^2}{t_0}}, \ldots, e^{-\frac{\|x-x_N\|^2}{t_0}} \right]^T, \\ h_x = \left[e^{-\frac{\|x-\psi_1\|^2}{t_0}}, e^{-\frac{\|x-\psi_2\|^2}{t_0}}, \ldots, e^{-\frac{\|x-\psi_L\|^2}{t_0}} \right]^T, \end{cases} \tag{25}$$

where also for fair comparison, the scaling parameter t_0 is set as n (no random selection).

At the stage of **graph decomposition**, the embedding graphs, including LDE, GbFA, and LPDA, as well as FDA and LDP, are used as terms of comparison in our experiments (an introduction to these graph types is given in Sections II-B and IV). The k_1 and k_2 parameters in Eq. (22) for LDE are both set to 30, and the strategy with $(W_{k_1NN})_{ij} = 1$ is adopted. The t parameter in Eq. (23) is chosen as n for GbFA. For LPDA, δ_0 is set to 10^{-4} in Eq. (24).

At the **regression** stage, the weights of γ_1 and γ_2 in Eq. (20) are set to $\gamma_1 = 0$ and $\gamma_2 = 10^{-3}$, respectively, and δ is set to 10^{-6}. At the decision level, a *k Nearest Neighbour classifier* (kNN) is adopted in order to better evaluate the performance compared to other classifiers with complex structures. Thus, kNN can be seen as the baseline. We set $k = 1$ in the experiments.

VI. Results

A. RGSR vs. ELM

This section shows a comparison between our proposed method RGSR and its counterpart ELM. As mentioned above, on the one hand, both ELM and RGSR adopt the same random anchor points at the data mapping stage. On the other hand, it is possible for RGSR to use different embedding graphs for graph decomposition, whereas ELM only employs the fixed embedding graphs as shown in Table I. Hence, in this article, we first conduct a series of experiments with ELM and RGSR on multiple speech emotion corpora using the same sets of random anchor points for all tests. The RGSR algorithms using FDA, LDP, LDE, GbFA, and LPDA embedding graphs are denoted as RGSR-$\{$FDA, LDP, LDE, GbFA, LPDA$\}$, respectively.

In Table III, we show the classification results for the various corpora and different numbers of anchor points ($L \in \{500, 1\,000, 3\,500, 5\,000, 10\,000\}$) on the four corpora. Experiments were run 10 times in order to obtain a distribution of *Unweighted Accuracies* (UAs; indicated as percentages) as the performance measure. As shown in Table III, RGSR-$\{$LDE, GbFA, LPDA$\}$ always achieve better performance than ELM on the four chosen databases. For example, the proposed RGSR-LPDA get a UA of 48.0 % on the ABC database, representing a 4.1 % relative improvement over the ELM system. Note that, RGSR-LPDA performs better than RGSR-LDE, and RGSR-GbFA on GEMEP, ABC, and VAM, except the six-class emotion recognition task on the eNTERFACE corpus. This reflects that the proposed GRSR with LPDA embedding graphs (i. e., GRSR-LPDA) is more robust across different emotional corpora.

In order to verify whether there are statistically significant differences between the performance obtained for each method and number of anchor points, we conducted a three-way *Analysis of Variance* (ANOVA) (factors shown in Table IV) and analysed the main effects. The results reveal significant main effects of *Method* ($F(5, 1167) = 19.068, p < 0.0001$) and L ($F(4, 1167) = 151.327, p < 0.0001$). This indicates that the factors of *Method* and the number of anchor points L have a statistically significant influence on the accuracies for recognising speech emotion. We conducted a post-hoc analysis on the factors to determine the direction of the various effects using *Tukey's Honest Significant Difference* (*Tukey*'s HSD) test. In respect of *Method* (see Table V), the RGSR-$\{$LDE, GbFA, LPDA$\}$ methods result in significantly better performance when compared to the ELM model across the four corpora ($p < 0.0001$), implying that the proposed RGSR algorithms considerably benefit from the graph decomposition.

B. RGSR vs. Conventional Algorithms

Having shown that RGSR algorithms yield better performance than ELMs on the four different corpora, we now focus on the performance comparison between our proposed methods and state-of-the-art supervised learning algorithms that are typically used to build speech emotion recognition systems. To this end, we compared the performance of RGSR-$\{$LDE, GbFA, LPDA$\}$ with standard supervised learning methods, including SVM, *Generalised RR* (noted as RR; [29]), kNN, and *Naive Bayes* (NB). For the SVM tests, we used a 'one-against-one' strategy and the *Sequential Minimal Optimisation* (SMO) for training using a penalty constant $C = 0.001$. For the RR tests, the weight of the $l2$ norm was set as 0.001 (in accordance with the RGSRs), and to achieve a fair comparison, we employed a 1-nearest neighbour classifier (the same used for RGSRs).

The results are depicted in Fig. 3 for each corpus studied: a) GEMEP, b) ABC, c) VAM, and d) eNTERFACE). A one-tailed *z-test* [53] was also conducted to determine the significance of the best algorithm's UA performance (from the 10-time

TABLE III
UAs (%) Including the Means and Standard Deviations in the Ten-Time Repeating Experiments on the GEMEP, ABC, VAM, and eNTERFACE Corpora, Respectively

L	Methods	GEMEP	ABC	VAM	eNTERFACE
$L = 500$	ELM	35.5 ± 1.5	43.5 ± 1.5	37.9 ± 0.4	51.8 ± 1.4
	RGSR-FDA	35.0 ± 2.1	42.9 ± 1.1	38.5 ± 1.3	48.5 ± 1.2
	RGSR-LDP	35.3 ± 2.2	43.1 ± 1.4	38.5 ± 1.0	48.9 ± 1.1
	RGSR-LDE	35.9 ± 1.6	43.3 ± 1.3	36.6 ± 0.8	49.8 ± 0.5
	RGSR-GbFA	35.6 ± 1.8	42.8 ± 1.2	38.2 ± 1.2	50.8 ± 1.2
	RGSR-LPDA	35.7 ± 1.9	44.3 ± 1.3	$\mathbf{39.2 \pm 1.8}$	49.0 ± 1.3
$L = 1\,000$	ELM	37.9 ± 1.1	43.9 ± 0.9	38.2 ± 0.4	53.9 ± 1.0
	RGSR-FDA	36.9 ± 1.5	43.9 ± 1.2	39.0 ± 1.0	52.0 ± 1.0
	RGSR-LDP	36.8 ± 1.6	44.8 ± 1.2	38.6 ± 0.9	52.5 ± 1.1
	RGSR-LDE	38.1 ± 1.1	44.8 ± 1.0	37.0 ± 0.9	53.3 ± 0.6
	RGSR-GbFA	38.1 ± 1.7	44.2 ± 0.9	39.1 ± 1.5	54.3 ± 1.0
	RGSR-LPDA	38.4 ± 1.3	46.3 ± 1.1	$\mathbf{39.2 \pm 1.5}$	52.3 ± 0.9
$L = 3\,500$	ELM	38.9 ± 0.7	43.9 ± 1.0	37.2 ± 0.4	56.2 ± 1.0
	RGSR-FDA	39.2 ± 0.8	44.8 ± 1.2	38.9 ± 1.3	56.0 ± 0.9
	RGSR-LDP	39.4 ± 0.7	45.3 ± 1.1	38.3 ± 1.1	56.0 ± 0.9
	RGSR-LDE	40.8 ± 1.1	46.0 ± 0.6	36.5 ± 0.3	58.6 ± 0.9
	RGSR-GbFA	40.5 ± 0.8	45.0 ± 1.2	38.5 ± 0.9	59.0 ± 0.8
	RGSR-LPDA	$\mathbf{41.0 \pm 1.0}$	47.8 ± 1.0	38.8 ± 1.4	56.7 ± 0.7
$L = 5\,000$	ELM	38.2 ± 0.9	43.3 ± 1.5	36.8 ± 0.3	56.6 ± 0.6
	RGSR-FDA	39.0 ± 0.8	45.0 ± 1.2	38.7 ± 0.6	56.7 ± 0.8
	RGSR-LDP	39.0 ± 1.0	44.9 ± 1.2	37.5 ± 0.4	56.6 ± 1.0
	RGSR-LDE	40.5 ± 0.7	46.2 ± 1.0	36.6 ± 0.3	59.4 ± 0.8
	RGSR-GbFA	39.9 ± 0.6	44.8 ± 1.1	38.2 ± 0.8	59.5 ± 1.0
	RGSR-LPDA	40.0 ± 0.7	$\mathbf{48.0 \pm 1.0}$	38.0 ± 0.8	57.3 ± 0.8
$L = 10\,000$	ELM	36.5 ± 1.0	42.9 ± 1.0	36.1 ± 0.3	56.4 ± 0.7
	RGSR-FDA	38.2 ± 0.9	44.4 ± 0.5	38.5 ± 1.0	56.4 ± 0.7
	RGSR-LDP	38.1 ± 1.0	44.2 ± 1.0	37.1 ± 0.4	56.5 ± 0.8
	RGSR-LDE	39.8 ± 0.6	46.0 ± 0.6	36.3 ± 0.3	59.8 ± 0.3
	RGSR-GbFA	39.0 ± 0.5	44.6 ± 0.7	38.4 ± 0.6	$\mathbf{60.2 \pm 0.8}$
	RGSR-LPDA	38.8 ± 0.4	47.0 ± 0.6	37.3 ± 0.4	57.3 ± 0.6

TABLE IV
Factors and the Categories (Marked as Categ.) of Each Factor in the Three-Way ANOVA

Factors	# Categ.	Categ.
Method	6	ELM, RGSR-{FDA, LDP, LDE, GbFA, LPDA}
L	5	500, 1 000, 3 500, 5 000, 10 000
Database	4	GEMEP, ABC, VAM, eNTERFACE

TABLE V
Pairwise Comparisons (Mean Difference and Significance) of UAs Between ELM and RGSRs Across the Four Corpora

Method1	Method2	Mean Difference (Method1-Method2)	Significance
ELM	RGSR-FDA	-0.003	$p > 0.05$
	RGSR-LDP	-0.003	$p > 0.05$
	RGSR-LDE	-0.010	$p < 0.0001$*
	RGSR-GbFA	-0.012	$p < 0.0001$*
	RGSR-LPDA	-0.013	$p < 0.0001$*

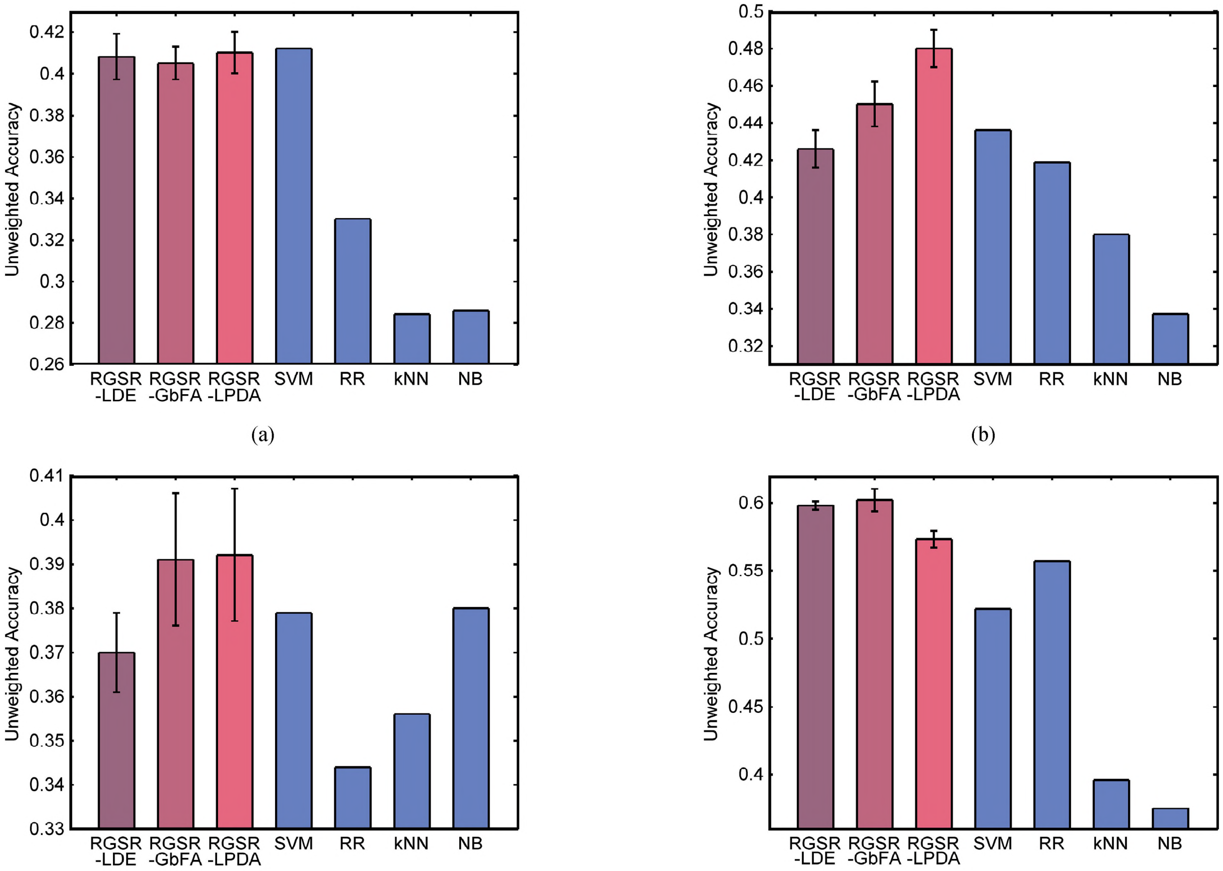
RGSR experiments). As observed, our proposed RGSR-{LDE, GbFA, LPDA}achieve better performance when compared to other conventional methods (e. g., SVM and RR). For example, the average UAs obtained using the RGSR and SVM methods are similar on the GEMEP database, but the RGSR method is significantly ($p < 0.05$) better than the kNN, NB, SVM, and RR methods on the remaining databases.

We also compared the RGSRs with GMKDA [54] method on the GEMEP corpus. GMKDA yielded a UA of 42.5%, while the best UA performance of RGSR-LPDA is 43.3%. In addition, the proposed GMKDA requires a much higher iterative computational complexity than the GSR algorithms.

C. RGSR vs. Conventional Subspace Learning Methods and Spectral Regression

Next, we systematically compared our proposed method with conventional subspace learning and existing SR methods. Specifically, the conventional linear subspace learning methods PCA, LPP [15], LDA, LDE [16] were considered. The recently proposed linear subspace learning methods, LDP [14], GbFA [17], and LPDA [8], were also used for comparison purposes. Furthermore,we tested three kernel subspace learning methods [10] using LDE, GbFA, and LPDA embedding graphs respectively. The SR methods tested were LSR [13], [18] (with $l2$-norm), LSR [11] (with $l1$-norm), KSR [20] (with $l2$-norm), and KSR (with $l1$-norm). As shown in Table I, the γ_2s in $l2$-norm LSR and KSR were set to 0.001, while the γ_1s in $l1$-norm LSR and KSR were set to 0.1. Note that the LSR in [35] was used in semisupervised learning. The experimental results on the four corpora are shown in Table VI.

As it can be observed, the proposed RGSR subspace learning framework always reaches the best results on the four emotional corpora. More specifically, RGSR-LPDA achieves the average UAs of 41.0%, 48.0%, and 39.2% on GEMEP, ABC, and VAM, respectively. RGSR-GbFA achieves the average UA of 60.2% on



(a)

(b)

(c)

(d)

Fig. 3. Column charts of the UAs for our proposed (red) RGSR-{LDE, GbFA, and LPDA }, and other conventional methods (blue) including SVM, RR [29], kNN, and Naive Bayes (NB), on the corpora of (a) GEMEP, (b) ABC, (c) VAM, and (d) eNTERFACE.

TABLE VI

UAs (%) Obtained by Subspace Learning (SL) Methods (Including Linear SL, and Kernel SL), and Existing Spectral Regression (SR) Methods (Including Linear SR (LSR) and Kernel SR (KSR)), and Our Proposed RGSR Methods, Including RGSR-{LDE, GbFA, LPDA} on the Four Speech Emotion Corpora

%UA	Corpus	GEMEP	ABC	VAM	eNTERFACE
Linear SL	PCA	30.5	40.8	36.1	39.7
	LPP [15]	20.6	32.5	34.0	35.3
	LDA	34.6	43.9	34.7	55.2
	LDP [14]	33.7	41.6	35.2	55.1
	LDE [16]	36.5	47.0	35.8	58.7
	GbFA [17]	34.3	41.6	35.3	58.6
	LPDA [8]	37.1	46.4	35.3	56.0
Kernel SL	Kernel LDE [16]	37.0	46.2	34.1	59.2
	Kernel GbFA [17]	34.9	44.5	35.2	57.5
	Kernel LPDA [8]	37.1	46.4	35.2	56.3
Linear SR	$l2$-norm LSR [13], [18]	32.2	40.8	34.0	55.1
	$l1$-norm LSR [11]	32.7	38.5	34.7	50.8
Kernel SR	$l2$-norm KSR [20]	37.8	43.6	37.3	55.1
	$l1$-norm KSR	37.1	42.0	36.4	49.6
Our Methods	**RGSR-LDE**	40.8 ± 1.1	46.2 ± 1.0	37.0 ± 0.9	59.8 ± 0.3
	RGSR-GbFA	40.5 ± 0.8	45.0 ± 1.2	39.1 ± 1.5	$\mathbf{60.2} \pm 0.8$
	RGSR-LPDA	$\mathbf{41.0} \pm 1.0$	$\mathbf{48.0} \pm 1.0$	$\mathbf{39.2} \pm 1.5$	57.3 ± 0.6

the eNTERFACE corpus. The one-tailed *z-tests* reveal that our proposed RGSR method significantly outperforms the widely used subspace learning methods, PCA and LDA, and the adopted linear spectral regression methods, $l1$-norm LSR and $l2$-norm LSR. Further, RGSR consistently surpasses the performance of other kernel subspace learning and kernel spectral regression methods on the four speech emotion recognition benchmarks by a large margin. These results suggest that our proposed RGSR

TABLE VII
PAIRWISE COMPARISONS (MEAN DIFFERENCE AND SIGNIFICANCE) OF UAS FOR THE RGSRS BETWEEN THE NUMBER OF ANCHOR POINTS $L = 3\,500$, $L = 5\,000$, AND THE OTHER LS ACROSS THE FOUR CORPORA

$L1$	$L2$	Mean Difference ($L1$-$L2$)	Significance
$L = 3\,500$	$L = 500$	0.035	$p < 0.0001$*
	$L = 1\,000$	0.017	$p < 0.0001$*
	$L = 10\,000$	0.006	$p < 0.005$*
$L = 5\,000$	$L = 500$	0.033	$p < 0.0001$*
	$L = 1\,000$	0.016	$p < 0.0001$*
	$L = 10\,000$	0.005	$p < 0.05$*

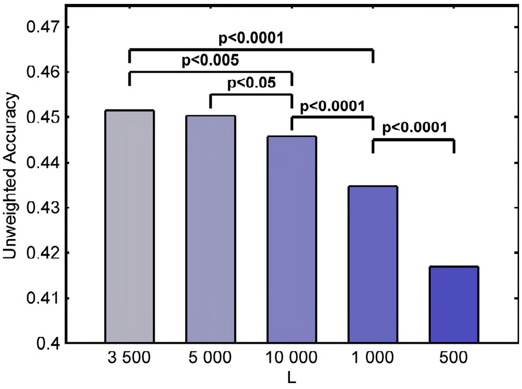


Fig. 4. The column bar chart of the averaging UAs across the four corpora when using different number of anchor points Ls, with the corresponding significance level between Ls.

is an efficient method for supervised dimensionality reduction in SER.

D. *Influence of Anchor Points*

As discussed above (cf. Section III), one major advantage of our proposed method over previous spectral regression counterparts is that anchor points are randomly generated, leading to the new features in RKHS. Here, we investigate the influence of anchor points.

Following the three-way ANOVA analysis conducted in Section VI-A, we first look into the effect of the number of anchor points L on our proposed method. The results are depicted in Table VII and Fig. 4. As shown in Table VII, '$L = 3\ 500$' and '$L = 5\ 000$' lead to a significantly better performance than '$L = 500$', '$L = 1\ 000$', and '$L = 10\,000$' ($p < 0.05$). Furthermore, as can be seen in Fig. 4, the tests using more than $3\,500$ anchor points yield significant performance improvements when compared with the ones using a lower number of anchor points ($p < 0.0001$). This indicates that the number of anchor points plays a significant role in our proposed algorithm. Further, our proposed method highly benefits from a large number of anchor points, yet, too many anchor points may degenerate the performance.

Now, we show the importance of randomly generated anchor points in our introduced GSR framework. We compare our proposed RGSR relying on random anchor points with the corresponding GSR methods using training sample anchor points, which are referred as to TGSR. Linear GSR (i.e., the data mapping $x \rightarrow x$) methods, which do not need the help of anchor points, are also reported. Table VIII shows the highest UAs yielded by these various methods. Note that the UAs of RGSRs correspond to the best performance of RGSR-$\{$LDE, GbFA, LPDA$\}$, respectively.

TABLE VIII
UAS OBTAINED BY LINEAR GSR (LGSR; WITHOUT USING ANCHOR POINTS), TGSR (WITH ANCHOR POINTS FROM TRAINING SAMPLES), AND THE RGSR (WITH RANDOM ANCHOR POINTS), WITH THE EMBEDDING GRAPHS LDE, GBFA, AND LPDA. 'EG' STANDS FOR THE CORRESPONDING EMBEDDING GRAPHS ON THE FOUR CORPORA

EG	Methods	GEMEP	ABC	VAM	eNTERFACE
LDE	LGSR-	36.2	46.6	34.7	59.0
	TGSR-	38.0	**47.5**	38.2	57.8
	RGSR-	**42.8**	**47.5**	**38.9**	**60.3**
GbFA	LGSR-	35.1	43.2	35.5	58.2
	TGSR-	40.3	**46.8**	40.1	59.1
	RGSR-	**41.7**	46.0	**42.5**	**61.3**
LPDA	LGSR-	36.4	47.3	35.1	55.7
	TGSR-	40.3	48.3	41.3	58.2
	RGSR-	**43.3**	**49.4**	**42.3**	**58.7**

As seen in Table VIII, the GSR algorithms using anchor points (i.e., RGSR and TGSR) outperform the linear GSR. Furthermore, the proposed RGSR algorithms often achieve the best performance on the four speech emotion databases, which demonstrates the importance of random anchor points.

VII. CONCLUSIONS AND OUTLOOK

In this article, we proposed the *Generalised Spectral Regression* framework that exploits the combination of Extreme Leaning Machines (ELMs) and subspace learning to overcome the drawbacks of ELM and Graph Embedding (GE) based spectral regression. The GSR framework consists of three stages, namely *data mapping*, *graph decomposition*, and *regression*. In *data mapping*, samples with original features are mapped into new spaces by using anchor points. In *graph decomposition*, designed embedding graphs reflecting the intrinsic structure of data are decomposed to obtain virtual coordinates. In *regression*, combining the virtual coordinates and data mapping, dimension-reduced mappings are calculated employing different regression types.

Using the GSR framework, we designed multiple embedding graphs to specifically represent the relations between data in the application of *Speech Emotion Recognition* (SER). Extensive experiments on four speech emotional corpora demonstrate the effectiveness and practicality of the proposed approaches when compared with related existing methods including ELM and subspace learning approaches.

Despite the excellent results using our proposed approach, there are various aspects on which this work could be improved. In *data mapping*, optimising the selections of anchor points or data mapping types is a meaningful research direction, since there exist significant differences in performance in the experiments. In the *graph decomposition* phase, it still remains unknown which type of embedding graphs can be more beneficial for SER tasks. Exploring more specific embedding graphs may further improve the system performance. In the *regression* stage, we only considered *Least-Square Regression* with $l2$-norm minimisation. Other approaches such as deep learning could also be explored in future work.

Appendix A
Proof of Theorem 2

According to Section II-A, we change the z^T in both the left and the right parts of Eq. (15) into $\widehat{s_l}$, where the label $l = 1, 2, \ldots, c$. It can be proved that

$$G^{(I)}\widehat{s_l} = S^T(SS^T)^{-1}(S\widehat{s_l}) = I_N\widehat{s_l}. \tag{26}$$

With the left part of Eq. (15) equal to $\lambda I_N \widehat{s_l}$, $\lambda = 1$ can be drawn. According to the SVD of

$$W_{FDA}^{(I)} = S^T(SS^T)^{-1}S = U\Lambda U^T, \tag{27}$$

where $U^TU = I$, $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_N)$, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$. Accordingly, the c eigenvalues $\lambda_1 = \lambda_2 = \ldots = \lambda_c = 1$, while the others are equal to zero.

Thus, $\lambda = 1$ are the c largest eigenvalues, which are corresponding to the maximal values of $zG^{(I)}z^T$. This is equivalent to solving

$$\max \quad \left(zG^{(I)}z^T\right) \Rightarrow \min \quad \left(zL^{(I)}z^T\right), \tag{28}$$

where $L^{(I)}$ is the Laplacian matrix of $S^T(SS^T)^{-1}S$.

The transformation for virtual coordinates is then changed into $2\widehat{s_l} - e$. Therefore, Theorem 2 is proved.

Acknowledgment

The authors would like to thank Zixing Zhang for the valuable help.

References

[1] B. Schuller, “Speech analysis in the big data era,” in *Text, Speech, and Dialogue*. New York, NY, USA: Springer, 2015, pp. 3–11.

[2] A. Tawari and M. M. Trivedi, “Speech emotion analysis: Exploring the role of context,” *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 502–509, Oct. 2010.

[3] I. Luengo, E. Navas, and I. Hernáez, “Feature analysis and evaluation for automatic emotion identification in speech,” *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 490–501, Oct. 2010.

[4] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[5] P. Song, S. Ou, W. Zheng, Y. Jin, and L. Zhao, “Speech emotion recognition using transfer non-negative matrix factorization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 5180–5184.

[6] X. Xu, J. Deng, W. Zheng, L. Zhao, and B. Schuller, “Dimensionality reduction for speech emotion features by multiscale kernels,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, 2015, pp. 1532–1536.

[7] S. Mariooryad and C. Busso, “Compensating for speaker or lexical variabilities in speech for emotion recognition,” *Speech Commun.*, vol. 57, pp. 1–12, 2014.

[8] X. Xu *et al.*, “Multiscale kernel locally penalised discriminant analysis exemplified by emotion recognition in speech,” in *Proc. ACM Int. Conf. Multimodal Interact.*, Tokyo, Japan, 2016, pp. 233–237.

[9] S. Yan, D. Xu, B. Zhang, and H. Zhang, “Graph embedding: A general framework for dimensionality reduction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, 2005, vol. 2, pp. 830–837.

[10] S. Yan *et al.*, “Graph embedding and extensions: A general framework for dimensionality reduction,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[11] D. Cai, X. He, and J. Han, “Spectral regression: A unified approach for sparse subspace learning,” in *Proc. Int. Conf. Data Mining*, Omaha, NE, USA, 2007, pp. 73–82.

[12] D. Cai, “Compressed spectral regression for efficient nonlinear dimensionality reduction,” in *Proc. Int. Joint Conf. Artif. Intell.*, Buenos Aires, Argentina, 2015, pp. 3359–3365.

[13] D. Cai, X. He, and J. Han, “SRDA: An efficient algorithm for large-scale discriminant analysis,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 1–12, Jan. 2008.

[14] H. Cai, K. Mikolajczyk, and J. Matas, “Learning linear discriminant projections for dimensionality reduction of image descriptors,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 338–352, Feb. 2011.

[15] X. He and P. Niyogi, “Locality preserving projections,” in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2003, pp. 153–160.

[16] H.-T. Chen, H.-W. Chang, and T.-L. Liu, “Local discriminant embedding and its variants,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, 2005, vol. 2, pp. 846–853.

[17] Y. Cui and L. Fan, “A novel supervised dimensionality reduction algorithm: Graph-based Fisher analysis,” *Pattern Recognit.*, vol. 45, no. 4, pp. 1471–1481, 2012.

[18] D. Cai, X. He, and J. Han, “Spectral regression for efficient regularized subspace learning,” in *Proc. Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[19] D. Cai, X. He, and J. Han, “Spectral regression for dimensionality reduction,” Comput. Sci. Dept., Univ. Illinois Urbana-Champaign, Champaign, IL, USA, Tech. Rep. 2856, 2007.

[20] D. Cai, X. He, and J. Han, “Efficient kernel discriminant analysis via spectral regression,” in *Proc. Int. Conf. Data Mining*, Omaha, NE, USA, 2007, pp. 427–432.

[21] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[22] G.-B. Huang, Z. Bai, L. L. C. Kasun, and C. M. Vong, “Local receptive fields based extreme learning machine,” *IEEE Comput. Intell. Mag.*, vol. 10, no. 2, pp. 18–29, May 2015.

[23] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, “Extreme learning machine for regression and multiclass classification,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[24] G.-B. Huang, L. Chen, and C.-K. Siew, “Universal approximation using incremental constructive feedforward networks with random hidden nodes,” *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006.

[25] G.-B. Huang, X. Ding, and H. Zhou, “Optimization method based extreme learning machine for classification,” *Neurocomputing*, vol. 74, no. 1, pp. 155–163, 2010.

[26] G. Huang, G.-B. Huang, S. Song, and K. You, “Trends in extreme learning machines: A review,” *Neural Netw.*, vol. 61, pp. 32–48, 2015.

[27] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, and P. Oittinen, “Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments,” *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2085–2098, Dec. 2014.

[28] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Singapore, 2014, pp. 223–227.

[29] S. An, W. Liu, and S. Venkatesh, “Face recognition using kernel ridge regression,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, 2007, pp. 1–7.

[30] B. Liu, S. Xia, F. Meng, and Y. Zhou, “Extreme spectral regression for efficient regularized subspace learning,” *Neurocomputing*, vol. 149, pp. 171–179, 2015.

[31] P. Liu, Y. Huang, L. Meng, S. Gong, and G. Zhang, “Two-stage extreme learning machine for high-dimensional data,” *Int. J. Mach. Learn. Cybern.*, vol. 7, pp. 1–8, 2014.

[32] G. Huang, S. Song, J. N. Gupta, and C. Wu, “Semi-supervised and unsupervised extreme learning machines,” *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2405–2417, Dec. 2014.

[33] A. Iosifidis, A. Tefas, and I. Pitas, “Graph embedded extreme learning machine,” *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 311–324, Jan. 2016.

[34] J. Yang, A. F. Frangi, J. Yang, D. Zhang, and Z. Jin, “KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.

[35] D. Cai, X. He, and J. Han, “Spectral regression: A unified subspace learning framework for content-based image retrieval,” in *Proc. Int. Conf. Multimedia*, Augsburg, Germany, 2007, pp. 403–412.

[36] M. Belkin and P. Niyogi, “Laplacian eigenmaps and spectral techniques for embedding and clustering,” in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2001, vol. 14, pp. 585–591.

[37] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, “Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 2705–2712.

[38] T. T. Ngo, M. Bellalij, and Y. Saad, “The trace ratio optimization problem for dimensionality reduction,” *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 5, pp. 2950–2971, 2010.

[39] J. Ye, “Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems,” *J. Mach. Learn. Res.*, vol. 6, pp. 483–502, 2005.

[40] F. Song, S. Liu, and J. Yang, “Orthogonalized Fisher discriminant,” *Pattern Recognit.*, vol. 38, no. 2, pp. 311–313, 2005.

[41] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.

[42] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 58, pp. 267–288, 1996.

[43] T. Bänziger, M. Mortillaro, and K. R. Scherer, “Introducing the Geneva multimodal expression corpus for experimental research on emotion perception,” *Emotion*, vol. 12, no. 5, pp. 1161–1179, 2012.

[44] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, “Audiovisual behavior modeling by combined feature spaces,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, USA, 2007, vol. 2, pp. II-733–II-736.

[45] M. Grimm, K. Kroschel, and S. Narayanan, “The Vera am Mittag German audio-visual emotional speech database,” in *Proc. IEEE Int. Conf. Multimedia Expo*, Hannover, Germany, 2008, pp. 865–868.

[46] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The eNTERFACE’05 audio-visual emotion database,” in *Proc. Int. Conf. Data Eng. Workshops*, Atlanta, GA, USA, 2006.

[47] B. Schuller *et al.*, “The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Lyon, France, 2013, pp. 148–152.

[48] P. Drineas, R. Kannan, and M. W. Mahoney, “Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix,” *SIAM J. Comput.*, vol. 36, no. 1, pp. 158–183, 2006.

[49] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, “Acoustic emotion recognition: A benchmark comparison of performances,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, Merano, Italy, 2009, pp. 552–557.

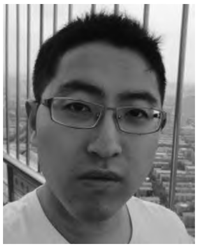
[50] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1977.

[51] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, “On the acoustics of emotion in audio: What speech, music and sound have in common,” *Frontiers Psychol.*, vol. 4, May 2013, Art. no. 292.

[52] F. Eyben, F. Weninger, F. Groß, and B. Schuller, “Recent developments in openSMILE, the Munich open-source multimedia feature extractor,” in *Proc. Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 835–838.

[53] S. L. Salzberg, “On comparing classifiers: Pitfalls to avoid and a recommended approach,” *Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 317–328, 1997.

[54] X. Xu *et al.*, “A two-dimensional framework of multiple kernel subspace learning for recognizing emotion in speech,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1436–1449, Jul. 2017.



Xinzhou Xu received the bachelor’s degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009, and the master’s and Ph.D. degrees from Southeast University, Nanjing, China, in 2012 and 2017, respectively. He is currently a Lecturer with the College of Internet of Things, Nanjing University of Posts and Telecommunications. Previously, he was with the Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Munich, Germany (from 2014 to 2016), and with the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany (from 2015 to 2016). His research interests include spoken signal processing, pattern recognition, machine learning, and affective computing.

Jun Deng received the bachelor’s degree in electronic and information engineering from Harbin Engineering University, Harbin, China, in 2009, the master’s degree in information and communication engineering from the Harbin Institute of Technology, Harbin, China, in 2011, and the Doctoral degree for the study on Feature Transfer Learning for Speech Emotion Recognition in electrical engineering and information technology from Technische Universität München, Munich, Germany, in 2016. From 2015 to 2017, he was a Postdoctoral Researcher with the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany. He is currently a Leader Researcher with the audEERING GmbH, Gilching, Germany, where he focuses on conducting fundamental research toward design and development of cutting edge technology for a wide range of affective computing applications. His current research focuses on machine learning methods such as transfer learning and deep learning with an application preference to affective computing.

Eduardo Coutinho received the Diploma in electrical engineering and computer sciences from the University of Porto, Porto, Portugal, in 2003, and the Doctoral degree in affective sciences from the University of Plymouth, Plymouth, U.K., in 2008. He is currently a Lecturer in music psychology with the University of Liverpool, Liverpool, U.K., and a Senior Researcher in affective computing with the Department of Computing, Imperial College London, London, U.K. He works in the interdisciplinary fields of music psychology and affective computing, where his expertise is in the study of emotional expression, perception and induction through music, and the automatic recognition of emotion in music and speech. He pioneered research on the analysis of emotional dynamics in music, and made significant contributions to the field of music emotion recognition, setting the new standard approach for recognition of emotional dynamics in music. Dr. Coutinho was the recipient of the Knowledge Transfer Award from the National Center of Competence in Research in Affective Sciences in 2013, and the Young Investigator Award from the International Neural Network Society in 2014.

Chen Wu received the bachelor’s degree from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009, and the master’s and Ph.D. degrees from Southeast University, Nanjing, China, in 2012 and 2015, respectively. He is currently a Lecturer with the College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include signal processing and pattern recognition.

Li Zhao received the bachelor's degree from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1982, the master's degree from Soochow University, Suzhou, China, in 1988, and the Ph.D. degree from the Kyoto Institute of Technology, Kyoto, Japan, in 1998. He is currently a Professor with the School of Information Science and Engineering, Southeast University, Nanjing, China. His research interests include spoken signal processing and affective computing.

Björn W. Schuller (M'05–SM'15–F'18) received the Diploma in 1999, the Doctoral degree for the study on Automatic Speech and Emotion Recognition in 2006, and the Habilitation and Adjunct Teaching Professorship in the subject area of Signal Processing and Machine Intelligence in 2012, all from the Technische Universität München, Munich, Germany, all in electrical engineering and information technology. He is currently a Reader in machine learning with the Department of Computing, Imperial College London, London, U.K., the Full Professor and Head of the Chair of Embedded Intelligence for Health Care and Wellbeing, Augsburg University, Augsburg, Germany, and Centre Digitisation.Bavaria, Garching, Germany, and an Associate of the Swiss Center for Affective Sciences with the University of Geneva, Geneva, Switzerland. He (co)authored 5 books and more than 600 publications in peer reviewed books, journals, and conference proceedings leading to more than 18 000 citations (h-index = 65). Prof. Schuller is the President-Emeritus of the Association for the Advancement of Affective Computing, elected member of the IEEE Speech and Language Processing Technical Committee, and member of the ACM and ISCA.