A two-dimensional framework of multiple kernel subspace learning for recognizing emotion in speech

Xinzhou Xu, Jun Deng, Nicholas Cummins, Zixing Zhang, Chen Wu, Li Zhao, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Xu, Xinzhou, Jun Deng, Nicholas Cummins, Zixing Zhang, Chen Wu, Li Zhao, and Björn Schuller. 2017. "A two-dimensional framework of multiple kernel subspace learning for recognizing emotion in speech." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (7): 1436–49. https://doi.org/10.1109/taslp.2017.2694704.

A Two-Dimensional Framework of Multiple Kernel Subspace Learning for Recognizing Emotion in Speech

Xinzhou Xu, Jun Deng, Nicholas Cummins, *Member, IEEE*, Zixing Zhang, *Member, IEEE*, Chen Wu, Li Zhao, and Björn Schuller, *Senior Member, IEEE*

***Abstract*—As a highly active topic in computational paralinguistics, speech emotion recognition (SER) aims to explore ideal representations for emotional factors in speech. In order to improve the performance of SER, multiple kernel learning (MKL) dimensionality reduction has been utilized to obtain effective information for recognizing emotions. However, the solution of MKL usually provides only one nonnegative mapping direction for multiple kernels; this may lead to loss of valuable information. To address this issue, we propose a two-dimensional framework for multiple kernel subspace learning. This framework provides more linear combinations on the basis of MKL without nonnegative constraints, which preserves more information in the learning procedures. It also leverages both of MKL and two-dimensional subspace learning, combining them into a unified structure. To apply the framework to SER, we also propose an algorithm, namely generalised multiple kernel discriminant analysis (GMKDA), by employing discriminant embedding graphs in this framework. GMKDA takes advantage of the additional mapping directions for multiple kernels in the proposed framework. In order to evaluate the performance of the proposed algorithm a wide range of experiments is carried out on several key emotional corpora. These experimental results demonstrate that the proposed methods can achieve better performance compared with some conventional and subspace learning methods in dealing with SER.**

***Index Terms*—Dimensionality reduction, discriminant analysis, multiple kernel learning (MKL), speech emotion recognition (SER), two-dimensional framework.**

I. Introduction

SPEECH Emotion Recognition (SER), a core area of research within computational paralinguitics [1], focuses on exploiting abstract representations of speech for the classfication or prediction of a range of human affect behaviours [2]–[8]. Emotion recognition from speech has been applied in various real-world cases [9]–[11], including: *Human-Computer Interaction* (HCI) [12], [13], diagnosis and treatment of autism for children [14], [15], and the detection of negative emotions in extreme conditions [6].

The use of functionals, to generate a single high dimensional representation of an utterance from a set of underlying low-level acoustic descriptors, is generally regarded as the default method for capturing paralinguistic information in speech [1]. Previous investigations consistently demonstrate the usefulness of this technique when applied to a range of different SER problems [14], [16], [17]. Popularity notwithstanding, a major drawback of using this feature representation is its non-specificity to the task at hand. Intuitively, an utterance level feature vector should be representative of all categories of information, both linguistic and paralinguistic, present within the particular utterance being modelled.

Therefore for specific applications, such as SER, there is a need to consider the techniques which help minimise the confounding effects of unwanted acoustic information present in the speech signal whilst retaining as much effective information as possible for the task at hand. Subspace learning is considered one such technique to help improve the performance of SER systems [6], [18]. Other such techniques could include segmentation and multi-task learning. However, when compared to subspace learning for SER, segmentation [19] only focuses on the stage prior to feature extraction, whilst multi-task learning [20] requires substantially more (manually annotated) information to successfully recognise the multiple tasks.

Conventional subspace learning algorithms can be employed to carry out dimensionality reduction. Popular examples for

Manuscript received September 18, 2016; revised February 21, 2017; accepted March 20, 2017. Date of publication April 17, 2017; date of current version June 5, 2017. This work was supported in part by the China Scholarship Council, in part by the European Union's Seventh Framework Programme under Grant 338164 (ERC Starting Grant iHEARu), in part by the BMBF IKT2020-Grant under Grant 16SV7213 (EmotAsS), in part by the Horizon 2020 Programme No. 688835 (RIA DE-ENIGMA), in part by the Natural Science Foundation of China under Grants 61673108, 61231002, and 61375028, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20151102. The associate editor coordinating the review of this manuscript and approving it for publication was Sin-Horng Chen. *(Corresponding Author: Xinzhou Xu.)*

X. Xu is with the Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China, also with the Machine Intelligence and Signal Processing group, MMK, Technische Universität München, München 80333, Germany, and also with the Chair of Complex and Intelligent Systems, the University of Passau, Passau 94032, Germany (e-mail: xinzhou.xu@tum.de).

J. Deng, N. Cummins, and Z. Zhang are with the Chair of Complex and Intelligent Systems, the University of Passau, Passau 94032, Germany (e-mail: jun.deng@uni-passau.de; nicholas.cummins@uni-passau,de; zixing.zhang@uni-passau.de).

C. Wu and L. Zhao are with the Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China (e-mail: 230129135@seu.edu.cn; zhaoli@seu.edu.cn).

B. Schuller is with the Department of Computing, Imperial College London, London, U.K., and also with the Chair of Complex and Intelligent Systems, University of Passau, Passau 94032, Germany (e-mail: schuller@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2694704

SER tasks include: *Principal Component Analysis* (PCA), *Linear Discriminant Analysis* (LDA) / *Fisher Discriminant Analysis* (FDA), *Linear Discriminant Projections* (LDP) [21], *Locally Linear Embedding* (LLE) [22], *Isomap* [23], *Locality Preserving Projections* (LPP) [24], *Locally Discriminant Embedding* (LDE) [25] and *Graph-based Fisher Analysis* (GbFA) [26]. Following up on the benefits these, and other such techniques offered to SER, various frameworks have been proposed to combine subspace learning, manifold learning, component analysis, and dimensionality reduction together [27]–[30].

Recent advances in multiple kernel subspace learning techniques indicate the advantages this paradigm brings to dimensionality reduction. Approaches such as the *Multiple Kernel Learning Dimensionality Reduction* (MKL-DR) proposed in [31] learn a unified, low-dimensional, subspace to exploit the information of high-dimensional data representations gained through the use of *Multiple Kernel Learning* (MKL). However, this optimisation combines the multiple kernels by employing only one nonnegative linear mapping direction. Compared with using multiple mapping directions, MKL implicitly uses only one feature when representing a combination of multiple kernels. Since more optimised features often means a more effective information representation for certain tasks, MKL may potentially result in loss of valuable information in MKL-DR, as it does not take other potential mapping directions into consideration.

Fortunately, two-dimensional subspace learning [25], [32]–[36] makes it possible to obtain more mapping directions without nonnegative constraints. The *two-dimensional* trick, as a case of tensorisation, has been employed to solve subspace learning directly for grey-scale image data with two-dimensional features (i. e., *matrix*). Reconstructing the original feature space opens possibilities to accomplish other learning tasks. Along with improving the learning performance, this trick has also been proven to be efficient in preserving the original structure of data and in processing large-size features [25], [34], [36].

It has been shown that, the performance of SER systems benefits from utilising multiscale kernels to describe paralinguistic features [6]. This makes the features represented by two-dimensional forms. In addition, multiple kernels can bring multiple views to reconstruct the original feature space. Potentially, these views provide more possibilities on fitting training models.

However, most of the paralinguistic feature sets are originally represented as *one-dimensional* form (i. e., *vector*) in the application of SER. Therefore, applying current two-dimensional subspace learning methods to a SER task is not straightforward. Nevertheless, MKL dimensionality reduction has provided a possibility to solve this problem on the basis of *Graph Embedding* (GE) frameworks [27], by optimising nonnegative linear combinations of multiple kernels [6], [31].

In this regard, we propose and explore a two-dimensional subspace learning framework based on multiple kernel learning. This framework provides a solution on handling one-dimensional paralinguistic features in SER via a two-dimensional structure. The framework seeks to extend the current MKL methods, by jointly employing two parts (with and without nonnegative constraints) for combining multiple kernels. Specifically, for the application of SER, we further propose a discriminant-based two-dimensional algorithm, namely *Generalised Multiple Kernel Discriminant Analysis* (GMKDA), to process one-dimensional features in SER. The proposed GMKDA makes use of the framework to obtain optimal solutions with the discriminant analysis.

The proposed approach is also compared with some highly related existing works. Kim *et al.* [37] proposed to use *Kernel Fisher Discriminant Analysis* (KFDA) in combination with multiscale kernels for binary classification tasks. However, the authors did not extend their technique to consider the multi-class case. Lin *et al.* [31], [38] proposed to learn one linear mapping in a framework for MKL-DR; again, the authors did not extend the framework to the two-dimensional form. Following Lin *et al.*'s research, we keep the alternative-optimisation way and make use of two-dimensional optimisation steps. Wang *et al.* [39] made improvements on designing more valid local Fisher embedding graphs by maintaining the optimisation structure of [31], while the research in [40]–[42] utilises different optimisation forms based on MKL.

Compared with these works, our research takes advantage of employing multiple linear mapping directions for multiple kernel combinations, by constructing a two-dimensional subspace learning framework for MKL dimensionality reduction [31]. It expands on our previous research into the use of multiscale kernels for SER [6]. Further, this research provides a solution to process a one-dimensional form of features using the two-dimensional scheme, which enables SER to be solved directly by this scheme.

The main contributions of this paper are as follows:

1) A *two-dimensional* framework is proposed to learn an optimal subspace for *one-dimensional* features in SER, by leveraging MKL and additional subspace mapping directions.
2) A novel algorithm, namely GMKDA is proposed based on this framework, taking benefit of a discriminant optimisation object.
3) Concerning the application of SER, the proposed algorithm with multiscale kernels is taken into consideration to achieve better performance compared with conventional methodologies.

The remainder of this paper is organised as follows. In Section II, the theoretical preliminaries are shortly described. Section III introduces both the two-dimensional framework for one-dimensional features, and the algorithm of GMKDA based on this framework, where nonnegative constraints and multiple linear combination directions for multiple kernels are taken into consideration. Afterwards, experimental results on multiple emotional corpora in speech are shown in Section IV. Finally, Section VI offers a succinct conclusion and highlights potential future research directions.

II. Preliminaries

In subspace learning for SER, in order to better describe data, it is expected to learn an optimal mapping $f(\cdot)$ from a certain sample, with features $x \in \Re^{n\times 1}$, to its corresponding new-space

feature vector $y = f(x) \in \Re^{d\times 1}$, where n and d represent the feature numbers of the two spaces respectively.

A. Notations

This section introduces commonly used notation within this paper. $X = [x_1, x_2, \ldots, x_N] \in \Re^{n\times N}$ will be used to denote the set of N training samples with each column standing for one training sample in the original feature space with the dimensionality of n. $Y = [y_1, y_2, \ldots, y_N] \in \Re^{d\times N}$ represents the set of N training samples with each column standing for one training sample in the dimensionality-reduced feature space with the dimensionality of d. Every column of $\phi(X) = [\phi(x_1), \phi(x_2), \ldots, \phi(x_N)]$ is a Reproducing Kernel Hilbert Space (RKHS) of the corresponding columns in X. $K = \phi^T(X)\phi(X)$ is the Gram matrix. We also assume that, any sample (including any training and testing sample) in the original and reduced dimensionality is represented by column vectors $x \in \Re^{n\times 1}$ and $y \in \Re^{d\times 1}$ respectively. For sample x, its kernelised coordinate is $K_x = \phi^T(X)\phi(x)$, where $\phi(x)$ lies in the RKHS of x.

Each column of $S = [s_1, s_2, \ldots, s_N] \in \Re^{c\times N}$ represents the label information of every corrsponding 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 is the identity matrix and every element of $\mathbf{e} \in \Re^{N\times 1}$ is equal to 1.

B. Graph Embedding Frameworks

Graph embedding frameworks have been proposed to combine subspace and manifold learning together [27]. Embedding graphs, data mapping types and optimisation forms are comprehensively considered in the graph embedding frameworks. Setting $y_i = f(x_i)$, the optimisation of graph embedding frameworks is shown with the constraints of penalty and scaling respectively, as in (1) and (2):

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

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

where $W^{(I)}$ and $W^{(P)}$ denote the adjacency matrices of the intrinsic graph and penalty graph respectively [27]. D is a diagonal matrix to control weights of samples. t is a positive constant value. With one mapping direction $a \in \Re^{n\times 1}$ for sample i in the linear case, $y_i = a^T x_i$, while for the training set with multiple mapping directions $A \in \Re^{n\times d}$, $Y = A^T X$.

As a special case in graph embedding, FDA employs the embedding graphs only including label information. For FDA and Kernel FDA (KFDA), since $N \geq c$, with

$$\begin{cases} W^{(I)} = W^{(I)FDA} = S^T(SS^T)^{-1}S, \\ W^{(P)} = W^{(P)FDA} = \frac{1}{N}\mathbf{e}\mathbf{e}^T, \end{cases} \tag{3}$$

the optimisation of FDA in graph embedding frameworks can be achieved accordingly.

C. Multiple Kernel Learning Dimensionality Reduction

Combining multiple kernels [31], [38] in GE frameworks, K_x is written as the linear combination of different kernels, namely

$$K_x = \sum_{m=1}^{M} \beta_m \phi_m^T(X)\phi_m(x) = \Omega_x \beta, \tag{4}$$

where the multiple kernel coordinate matrix

$$\Omega_x = [\phi_1^T(X)\phi_1(x), \phi_2^T(X)\phi_2(x), \ldots, \phi_M^T(X)\phi_M(x)] \in \Re^{N\times M} \tag{5}$$

and $\beta \in \Re^{M\times 1}$ is the column vector with corresponding elements β_m for kernel m. M represents the number of kernels. Each column of Ω_x is the corresponding coordinate for sample x.

Defining the optimised data mapping as $\alpha \in \Re^{N\times 1}$, we can draw the MKL form as

$$\begin{aligned} \arg\min_{\alpha} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N} \| \alpha^T(K_{x_i} - K_{x_j}) \|^2 W_{ij}^{(I)} \\ \text{s.t.} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N} \| \alpha^T(K_{x_i} - K_{x_j}) \|^2 W_{ij}^{(P)} = t. \end{aligned} \tag{6}$$

By extending the mapping α to $A = [\alpha_1, \alpha_2, \ldots, \alpha_d] \in \Re^{N\times d}$, we obtain multiple mappings by solving the optimisation problem. α_i is the ith mapping vector with $i = 1, 2, \ldots, d$.

Similar as in FDA, Multiple Kernel Learning Fisher Discriminant Analysis (MKL-FDA) utilises the embedding graphs of $W^{(I)FDA}$ and $W^{(P)FDA}$, represented by $W^{(I)}$ and $W^{(P)}$ respectively. The optimisation of MKL is shown as

$$\begin{aligned} \arg\min_{A,\beta} \quad & \sum_{i,j=1}^{N} \| A^T\Omega_{x_i}\beta - A^T\Omega_{x_j}\beta \|^2 W_{ij}^{(I)} \\ \text{s.t.} \quad & \begin{cases} \sum_{i,j=1}^{N} \| A^T\Omega_{x_i}\beta - A^T\Omega_{x_j}\beta \|^2 W_{ij}^{(P)} = t, \\ \beta_m \geq 0, \quad m = 1, 2, \ldots, M. \end{cases} \end{aligned} \tag{7}$$

On solving (7), it has been proposed to use the alternative form, to optimise A and the linear weights β of multiple kernels [31]. To obtain an optimal A, the Generalised Eigenvalue Problem (GEP) is utilised to solve the approximate ratio-trace form, while Semi-Definite Programming (SDP) relaxation provides a solution for the nonconvex problem in optimising β [31], [43].

D. Two-Dimensional Subspace Learning

Two-dimensional subspace learning has been shown in previous research as a special case of tensorised subspace learning [44], e.g., Two-Dimensional PCA (2DPCA) [33], Two-Dimensional LDA (2DLDA) [34], [35], Two-Dimensional LPP (2DLPP) [36], or Two-Dimensional LDE (2DLDE) [25]. By defining $y_i = L^T \Xi_i R$ in (1), the two-dimensional subspace

learning [25] is shown as

$$\arg\min_{L,R} \sum_{i,j=1}^{N} \| L^T \Xi_i R - L^T \Xi_j R \|_F^2 \; W_{ij}^{(I)}$$

$$\text{s.t.} \sum_{i,j=1}^{N} \| L^T \Xi_i R - L^T \Xi_j R \|_F^2 \; W_{ij}^{(P)} = t, \qquad (8)$$

where the two-dimensional samples Ξ_i and $\Xi_j \in \Re^{p\times q}$. p and q are respectively vertical and horizontal sizes of a sample described by two-dimensional features. $L \in \Re^{p\times d^{(L)}}$ and $R \in \Re^{q\times d^{(R)}}$ represent the mapping matrices on the two dimensions respectively. $W^{(I)}$ and $W^{(P)}$ are the same as stated above.

We do not list the case of the scaling constraint, especially for 2DPCA and 2DLPP, since it is similar to the form in GE. In order to deal with the two-dimensional subspace learning, alternative optimisation on L and R by GEP has been utilised with the approximate ratio-trace form [34].

III. Methodology

A. Two-dimensional Subspace Learning With Multiple Kernels

The MKL dimensionality reduction in (7) shows that, the dimensionality-reduced features for a given training or testing sample x can be written as $A^T \Omega_x \beta$, where the newly generated two-dimensional features $\Omega_x \in \Re^{N\times M}$ are constructed by N training samples and M kernels. Thus the linear mapping vectors $A = [a_1, a_2, \ldots, a_d]$ lie in an N-dimensional space.

However, the MKL dimensionality reduction differs from the two-dimensional subspace learning mainly on the aspect of constraints. In detail, the linear combination of multiple kernels contains one nonnegative linear mapping vector β in MKL dimensionality reduction, while in two-dimensional subspace learning there are several mapping vectors without nonnegative constraints. Hence, by replacing the β with a matrix B including several mapping directions, removing the nonnegative constraints, we change the form of MKL into the case of two-dimensional subspace learning:

$$\arg\min_{A,B} \sum_{i,j=1}^{N} \| A^T \Omega_{x_i} B - A^T \Omega_{x_j} B \|_F^2 \; W_{ij}^{(I)}$$

$$\text{s.t.} \sum_{i,j=1}^{N} \| A^T \Omega_{x_i} B - A^T \Omega_{x_j} B \|_F^2 \; W_{ij}^{(P)} = t, \qquad (9)$$

where $B = [b_1, b_2, \ldots, b_{d^{(B)}}]$ and $d^{(B)}$ is the number of linear mapping directions for a multiple kernel combination.

Similar to the solving procedures in [31] and [34], the optimisation of (9) is generally solved by the alternative steps of (10) and (12):

$$\min_B \frac{tr\left(B^T Q^{(I)}(A) B\right)}{tr\left(B^T Q^{(P)}(A) B\right)}, \qquad (10)$$

where $tr(\cdot)$ represents the operator calculating trace, and

$$\begin{cases} Q^{(I)}(A) = \sum_{i,j=1}^{N} (\Omega_{x_i} - \Omega_{x_j})^T A A^T (\Omega_{x_i} - \Omega_{x_j}) W_{ij}^{(I)}, \\ Q^{(P)}(A) = \sum_{i,j=1}^{N} (\Omega_{x_i} - \Omega_{x_j})^T A A^T (\Omega_{x_i} - \Omega_{x_j}) W_{ij}^{(P)}. \end{cases} \qquad (11)$$

$$\min_A \frac{tr\left(A^T Q_0^{(I)}(B) A\right)}{tr\left(A^T Q_0^{(P)}(B) A\right)}, \qquad (12)$$

where

$$\begin{cases} Q_0^{(I)}(B) = \sum_{i,j=1}^{N} (\Omega_{x_i} - \Omega_{x_j}) B B^T (\Omega_{x_i} - \Omega_{x_j})^T W_{ij}^{(I)}, \\ Q_0^{(P)}(B) = \sum_{i,j=1}^{N} (\Omega_{x_i} - \Omega_{x_j}) B B^T (\Omega_{x_i} - \Omega_{x_j})^T W_{ij}^{(P)}. \end{cases} \qquad (13)$$

In essence, the case of (9) indicates a two-dimensional form and it can be solved through the same way as in two-dimensional subspace learning. Therefore, it changes MKL dimensionality reduction on the following two aspects: 1) The nonnegative constraints for kernel weights are removed and thus weights without nonnegative constraints are both considered; 2) Multiple mappings for kernel weights are considered, instead of using only one mapping direction in MKL. When Fisher discriminant embedding graphs are utilised in the form of (9), we denote the corresponding method as 2DMKDA with the relaxed constraints.

B. Proposed Framework

The proposed two-dimensional subspace learning with multiple kernels makes it possible to obtain multiple mapping directions for multiple kernel combinations. However, these mapping directions do not include nonnegative constraints, which have been utilised in MKL to achieve one nonnegative mapping direction. In order to leverage this, we propose a framework presenting extended mapping directions based on MKL dimensionality reduction.

1) Problem Formulation: First, with parameter γ, we define the extended mapping directions for multiple kernels as

$$\bar{\mathbf{B}} = \left[\gamma\beta \;\; (1-\gamma)B\right] \in \Re^{M\times(d^{(B)}+1)}, \qquad (14)$$

where the parameter $\gamma \in [0,1]$ regulates the relationship between nonnegative and other mapping directions for multiple kernel combination. The nonnegative linear combination of multiple kernels, namely $\beta \in \Re^{M\times 1}$, obeys $\beta_m \geq 0$, with $m = 1, 2, \ldots, M$.

Thus, we can obtain the optimal A^*, β^*, and B^* by solving

$$\arg\min_{A,\beta,B} J = \frac{\sum_{i,j=1}^{N} \| A^T (\Omega_{x_i} - \Omega_{x_j}) \bar{\mathbf{B}} \|_F^2 \; W_{ij}^{(I)}}{\sum_{i,j=1}^{N} \| A^T (\Omega_{x_i} - \Omega_{x_j}) \bar{\mathbf{B}} \|_F^2 \; W_{ij}^{(P)}}$$

$$= \frac{\sum_{i,j=1}^{N} \| A^T (\Omega_{x_i} - \Omega_{x_j}) \left[\gamma\beta \;\; (1-\gamma)B\right] \|_F^2 \; W_{ij}^{(I)}}{\sum_{i,j=1}^{N} \| A^T (\Omega_{x_i} - \Omega_{x_j}) \left[\gamma\beta \;\; (1-\gamma)B\right] \|_F^2 \; W_{ij}^{(P)}}$$

$$\text{s.t.} \;\; \beta_m \geq 0, \quad m = 1, 2, \ldots, M, \qquad (15)$$

where $W^{(I)}$ and $W^{(P)}$ stand for the intrinsic and penalty embedding graphs as in Section II. In this framework, selections of the embedding graphs depend on distribution of data. The parameter γ balances the weights between multiple kernel learning and the two-dimensional form. To make it normalised for regulating γ, we set $\beta^T\beta = 1$ and $B^T B = I_{d^{(B)}}$, where $I_{d^{(B)}}$ is a $d^{(B)}$ identity matrix.

The framework of (15) shows that, we can obtain an integral form with the help of MKL and the two-dimensional structure in subspace learning. This framework contains certain designed embedding graphs as in GE frameworks. As shown in the proposed 2DMKDA and previously utilised MKL-FDA [6], the algorithm of GMKDA is proposed by choosing the embedding graphs as in FDA according to (3), which can be found in Section II.

When $\gamma = 0$, GMKDA turns to be 2DMKDA with orthonormalisation constraints, while GMKDA deceases to be MKL-FDA if $\gamma = 1$.

As in [31], [34], [35], we can alternatively optimise $\bar{\mathbf{B}}$ and A here with the objective function (15). The alternative iteration of the solution procedure is presented as follows:

2) Optimising $\bar{\mathbf{B}}$: With fixed A, (15) can be reformulated into

$$\min_{\bar{\mathbf{B}}} \quad \frac{tr\left(\bar{\mathbf{B}}^T Q^{(I)}(A)\bar{\mathbf{B}}\right)}{tr\left(\bar{\mathbf{B}}^T Q^{(P)}(A)\bar{\mathbf{B}}\right)} \quad \text{s.t.} \quad \bar{\mathbf{B}} = \left[\gamma\beta\ (1-\gamma)B\right], \tag{16}$$

where the matrices of $Q^{(I)}(A)$ and $Q^{(P)}(A)$ are shown in (11).

We pose the Iterative Trace Ratio (ITR) algorithm [45], [46] to simplify (16) since it is a trace-ratio problem, which is calculated by iteration of an inner ITR loop. This results in

$$\min_{\bar{\mathbf{B}}} \quad \left[\ tr\left(\bar{\mathbf{B}}^T Q^{(I)}(A)\bar{\mathbf{B}}\right) - \lambda\ tr\left(\bar{\mathbf{B}}^T Q^{(P)}(A)\bar{\mathbf{B}}\right)\ \right], \tag{17}$$

where $\lambda > 0$ represents the iterative value updated in each step of the inner ITR loop. By changing $\bar{\mathbf{B}}$ into $\left[\gamma\beta\ (1-\gamma)B\right]$, (17) can be rewritten as

$$\min_{\beta,B} \quad J_0(\beta, B) = \left[\ \gamma^2\ J_1(\beta) + (1-\gamma)^2\ J_2(B)\ \right], \tag{18}$$

where

$$\begin{cases} J_1(\beta) = \beta^T (Q^{(I)}(A) - \lambda\ Q^{(P)}(A))\beta, \\ J_2(B) = tr\left(B^T (Q^{(I)}(A) - \lambda\ Q^{(P)}(A))B\right), \end{cases} \tag{19}$$

where the values of β and B can be optimised separately with fixed λ. Then, we present the steps of the inner ITR loop to iteratively optimise β, B, and λ with fixed A:

Updating β with fixed λ, A: As in (18), optimising β leads to

$$\begin{aligned} &\min_{\beta} \quad J_1(\beta) = \beta^T (Q^{(I)}(A) - \lambda\ Q^{(P)}(A))\beta \\ &\text{s.t.} \quad \beta^T\beta = 1, \quad \beta_m \geq 0, \quad m = 1, 2, \ldots, M, \end{aligned} \tag{20}$$

which is a Quadratically Constrained Quadratic Programming (QCQP) problem [43] with nonnegative normalisation constraints. This can be relaxed to obtain the SDP form [31].

By adding the auxiliary matrix $\mathbf{P} \in \Re^{M\times M}$ for the relaxation, we can draw the SDP form

$$\begin{aligned} &\min_{\beta,\mathbf{P}} \quad tr\left((Q^{(I)}(A) - \lambda\ Q^{(P)}(A))\mathbf{P}\right) \\ &\text{s.t.} \quad \begin{bmatrix} \mathbf{P} & \beta \\ \beta^T & 1 \end{bmatrix} \succeq 0, \quad tr(\mathbf{P}) = 1, \quad \beta_m \geq 0, \quad m = 1, 2, \ldots, M. \end{aligned} \tag{21}$$

Hence, the optimal β is calculated using (21).

Updating B with fixed λ, A: The optimisation for B depends on minimising $J_2(B)$. Therefore, the optimisation form of (18) is presented with orthonormalisation constraints as

$$\begin{aligned} &\min_{\beta} \quad J_2(B) = tr\left(B^T (Q^{(I)}(A) - \lambda\ Q^{(P)}(A))B\right) \\ &\text{s.t.} \quad B^T B = I_{d^{(B)}}, \end{aligned} \tag{22}$$

which is solved by calculating eigenvalues to obtain an optimal orthonormalised B.

It is worth noticing that, we can utilise parallel computation to calculate β and B since the two procedures are independent to each other.

Updating λ with fixed β, B, A: The calculation of ITR [46], [47] indicates the update criterion of λ as

$$\lambda = \frac{tr\left(\left[\gamma\beta\ (1-\gamma)B\right]^T Q^{(I)}(A)\left[\gamma\beta\ (1-\gamma)B\right]\right)}{tr\left(\left[\gamma\beta\ (1-\gamma)B\right]^T Q^{(P)}(A)\left[\gamma\beta\ (1-\gamma)B\right]\right)}. \tag{23}$$

3) Optimising A: The form of (15) also can be transformed into a form with regard to A. Since $\bar{\mathbf{B}}$ is a constant matrix when β, B, and γ are fixed, the form can be represented as

$$\min_{A} \ \frac{tr\left(A^T C_0^{(I)}(\bar{\mathbf{B}})A\right)}{tr\left(A^T C_0^{(P)}(\bar{\mathbf{B}})A\right)}, \tag{24}$$

where

$$\begin{cases} C_0^{(I)}(\bar{\mathbf{B}}) = \sum_{i,j=1}^{N} (\Omega_{x_i} - \Omega_{x_j})\bar{\mathbf{B}}\bar{\mathbf{B}}^T (\Omega_{x_i} - \Omega_{x_j})^T W_{ij}^{(I)}, \\ C_0^{(P)}(\bar{\mathbf{B}}) = \sum_{i,j=1}^{N} (\Omega_{x_i} - \Omega_{x_j})\bar{\mathbf{B}}\bar{\mathbf{B}}^T (\Omega_{x_i} - \Omega_{x_j})^T W_{ij}^{(P)}. \end{cases} \tag{25}$$

For simplicity, the trace-ratio problem of (24) is then approximately expressed as a ratio-trace form [31] of

$$\min_{A} \ tr\left([A^T C_0^{(P)}(\bar{\mathbf{B}})A]^{-1}[A^T C_0^{(I)}(\bar{\mathbf{B}})A]\right), \tag{26}$$

which is equivalent to solving the GEP of

$$C_0^{(I)}(\bar{\mathbf{B}})\ \alpha_i = \lambda'\ C_0^{(P)}(\bar{\mathbf{B}})\ \alpha_i, \tag{27}$$

by considering $A = [\alpha_1, \alpha_2, \ldots, \alpha_d]$, where $i = 1, 2, \ldots, d$. λ' is the corresponding generalised eigenvalue. It is assumed that, each column of A obeys

$$\frac{\alpha_1^T C_0^{(I)}(\bar{\mathbf{B}})\alpha_1}{\alpha_1^T C_0^{(P)}(\bar{\mathbf{B}})\alpha_1} \leq \frac{\alpha_2^T C_0^{(I)}(\bar{\mathbf{B}})\alpha_2}{\alpha_2^T C_0^{(P)}(\bar{\mathbf{B}})\alpha_2} \leq \cdots \leq \frac{\alpha_d^T C_0^{(I)}(\bar{\mathbf{B}})\alpha_d}{\alpha_d^T C_0^{(P)}(\bar{\mathbf{B}})\alpha_d}.$$

The initial value of A can be selected as $AA^T = I$ according to [31], while the initial value of λ can be a large positive

Algorithm 1: Two-Dimensional Framework for Multiple Kernel Subspace Learning (i.e., GMKDA).

Input:

training samples $X = [x_1, x_2, \ldots, x_N]$, testing sample x;
embedding graphs $W^{(I)}$ and $W^{(P)}$;
iteration number T, parameter γ;

1: normalisation on X and x;
2: calculate two-dimensional features $\{\Omega_{x_1}, \Omega_{x_2}, \ldots, \Omega_{x_N}\}$ and Ω_x, corresponding to X and x;
3: initialise $A^{(0)}A^{(0)^T} = I$, $l = 1$;
4: **while** $l \leq T$ **do**
5: initialise $\lambda^{(0)} > 0$, $l' = 1$;
6: **while** $l' \leq T'$ **do**
7: update $\beta^{(l')}$ using (21), with fixed $A^{(l-1)}$, $\lambda^{(l'-1)}$;
8: update $B^{(l')}$ using (22), with fixed $A^{(l-1)}$, $\lambda^{(l'-1)}$;
9: update $\lambda^{(l')}$ using (23), with fixed $A^{(l-1)}$, $\beta^{(l')}$, $B^{(l')}$;
10: $l' = l' + 1$;
11: **end while**
12: $\bar{\mathbf{B}}^{(l)} = \left[\gamma\beta^{(T')}\ (1-\gamma)B^{(T')}\right]$;
13: update $A^{(l)}$ using (24), with fixed $\bar{\mathbf{B}}^{(l)}$;
14: $l = l + 1$;
15: **end while**
16: optimal $A^* = A^{(T)}$, $\bar{\mathbf{B}}^* = \bar{\mathbf{B}}^{(T)}$;
17: obtain y^* by vectorising on $A^{*T}\Omega\bar{\mathbf{B}}^*$;
18: obtain y by calculating PCA on y^*;

Output:

y, A^*, $\bar{\mathbf{B}}^*$.

value. To avoid the theoretical zero values of the $tr(\cdot)$ terms in (16) and (24), the terms of $t_0 tr(\bar{\mathbf{B}}^T\bar{\mathbf{B}})$ and $t_0 tr(A^T A)$ can be added to the numerators and denominators of the two equations respectively, with a very small value $t_0 > 0$. Note that the vectorised features may include some noise caused by the computational accuracy. We therefore perform PCA following the GMKDA procedure to denoise the features whilst preserving the structure of the data.

To clarify the proposed two-dimensional framework for Multiple Kernel Subspace Learning (i.e., GMKDA) in detail, the key procedures are outlined in Algorithm 1. The maximal numbers of iterations are T (for the outer loop) and T' (for the inner loop). The output values are shown as: the new features y for the testing sample x; the optimal mapping matrices A^* and $\bar{\mathbf{B}}^*$.

In SER, original paralinguistic features are firstly usually obtained by a feature extractor [14], [16], [48], [49]. However, the original feature sets often include factors which are disadvantageous for emotion recognition [13], [50], [51]. Hence, GMKDA is employed in the stage of dimensionality reduction to extract discriminative components from the features for SER.

The whole procedure of GMKDA in paralinguistics (e.g., SER) is shown in Fig. 1. As shown in Fig. 1, for a speech utterance sample x, one-dimensional paralinguistic features are firstly extracted. Then, we reconstruct these features by M kernels respectively, which leads to two-dimensional features with a dimensionality of $N \times M$. Using the two-dimensional features, we put training samples into the iterative procedure to obtain optimal projections β^*, B^*, and A^*.

Though GMKDA is utilised in SER with the embedding graphs of FDA, we can also change the embedding graphs to achieve higher performance if there are embedding graphs better depicting the structure of a certain data set. However, unsuitable embedding graphs may cause a worse result since the optimisation object is misleaded, and vice versa.

C. Theoretical Analysis

1) Convergence: For the loop of optimising A and $\bar{\mathbf{B}}$, the convergence analysis of the proposed method can be drawn according to the existing research [31], [35], [56], since the procedure of the optimisation is presented as a trace-ratio form. For the inner ITR loop, the theoretical convergence can be proven as in Appendix A.

2) Computational Complexity: The computational complexity of the proposed algorithm depends on the number of training samples (N), the number of selected kernels (M), the number of outer-loop iterations (T), and the number of inner-loop iterations (T').

For *2DMKDA*, the computational complexity is given by $O(T\max(N, M)^3)$. When $N >> M$, the complexity turns to be $O(TN^3)$. For the two-dimensional framework for multiple kernel subspace learning (i.e., *GMKDA*), the inner loop costs $O(M^3)$ stem mainly from solving SVD in optimising B, without considering the complexity of utilising the interior point approach in solving SDP, since it is usually not high. Therefore, the complexity of the inner loop is $O(T'M^3)$. Then, obtaining an optimal A using GEP demands $O(N^3)$. Thus, the algorithm requires the upper bound of $O(T\max(N^3, T'M^3))$. When $N >> M$ while T and T' are small, the complexity is $O(N^3)$, which is similar as in [31].

As shown in the analysis, the computational complexity mainly comes from solving eigenvalues or Singular Value Decomposition (SVD). To reduce this complexity, Lanczos method [45] or the Nyström method [57] can be employed in dealing with the large-scale case.

IV. Experimental Setup

A. Corpora

In our experiments, three paralinguistic corpora are utilised to validate the proposed methods in SER. In detail, the corpora are the *GEneva Multimodal Emotion Portrayals* (GEMEP) [14], [54], [55], the *Airplane Behavior Corpus* (ABC) [52], and *eNTERFACE'05* (eNTERFACE) [53] respectively. Only the audio parts of the corpora are taken into consideration to evaluate the performance of the proposed SER system. In our experiments, speaker-independent ways are adopted for fair comparison, which means that, the samples from some speakers are selected in training while the samples from the other speakers are used for testing. As a result, the same speakers cannot appear in both training and testing sets. The corpora of ABC, eNTERFACE, and GEMEP stand for the cases of small size of the sample set, balanced categories, and larger number of emotions,

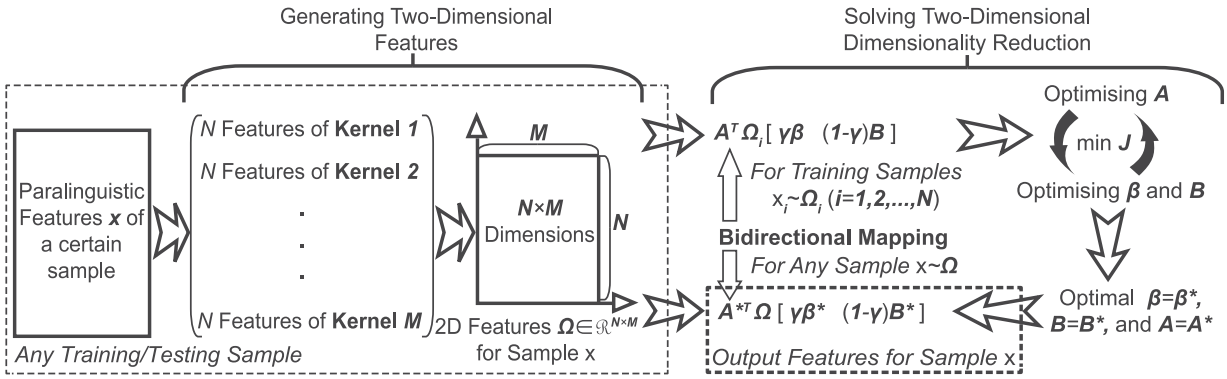


Fig. 1. Schematic diagram of the proposed GMKDA. For any speech signal utterance sample, paralinguistic features are firstly obtained as one-dimensional features by our extractor. Then, the second dimension of features are generated according to multiple kernels. Afterward, the proposed approach is adopted in solving the dimensionality reduction problem to achieve optimal β, B, and A, where γ is between zero and one.

TABLE I
SMALL DESCRIPTION OF THE EMOTIONAL CORPORA ABC, ENTERFACE, AND GEMEP FOR THE AUDIO SECTIONS

Corpus	Language	Sampling Rate	# Classes	# Speakers	# Samples	Evaluation
ABC [52]	German	16 kHz	6	8 (4 female)	430	2-fold CV
eNTERFACE [53]	English	16 kHz	6	42 (here 40, 8 female)	1 277 (here 1 200)	2-fold CV
GEMEP [54], [55]	French	44.1 kHz	18 (here 12)	10 (5 female)	1 260 (here 1 080)	Training-Testing 2-fold CV

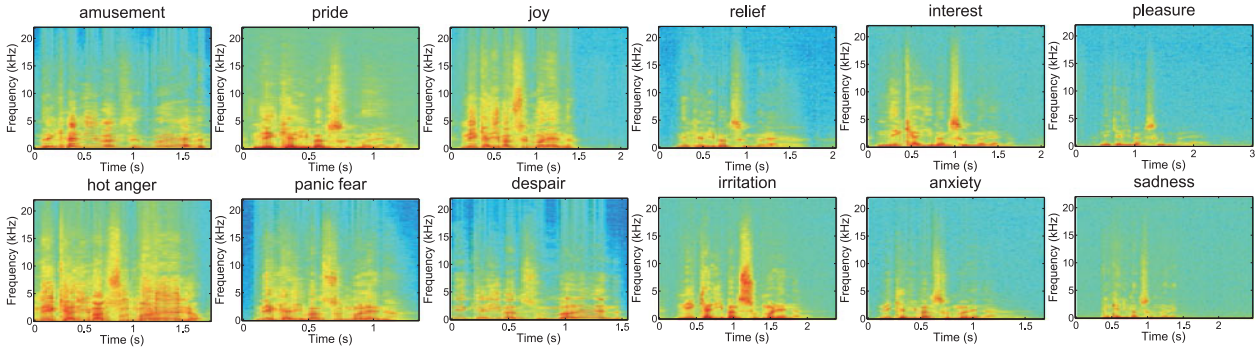


Fig. 2. Speech spectrograms of each used emotion in GEMEP with a certain speaker uttering the same French sentence.

respectively. We depict the audio sections of the three corpora on various aspects as in Table I and follows.

GEMEP is a French-content corpus with 18 speech emotional categories and 1 260 utterance samples. We choose 12 categories of emotions (*amusement*, *pride*, *joy*, *relief*, *interest*, *pleasure*, *hot anger*, *panic fear*, *despair*, *irritation*, *anxiety*, *sadness*) in our experiments as in [58]. Those are totally 1 080 samples by ten speakers belonging to the chosen categories, which leads to 90 samples per emotion. In Fig. 2, we exemplify the speech spectrograms of each emotion in GEMEP when a certain speaker says the same short sentence, in order to show the spectral characteristics of specific emotions in speech. Note that the emotional states corresponding to the upper part of Fig. 2 have higher valence compared to those displayed in the lower images.

We perform two experiments on GEMEP: 1) In order to be in accordance with the challenge settings in [6], [14], the corpus in the experiments is firstly divided into the sets training and testing, which are corresponding to six speakers (three female, three male) and four speakers (two female, two male) respectively. Hence, there are 648 and 432 samples in the sets of training and testing respectively. 2) In order to show results by different experimental ways, we divide the corpus into two folds, including five speakers in each fold (three female and two male for the first fold / two female and three male for the second fold). One fold is for training and the other is for testing, and vice versa. This leads to a two-fold Cross-Validation (CV).

ABC consists of the six emotions *aggressive*, *cheerful*, *intoxicated*, *nervous*, *neutral*, *tired*. There are eight German speakers (four female, four male) with totally 430 samples. For the six emotions, the number of samples are 95, 105, 33, 93, 79, and 25 respectively. The corpus is separated into two folds containing 236 and 194 samples respectively, with four speakers (two female, two male) for each fold. Two-fold CV is again adopted in our experiments.

eNTERFACE contains the six basic emotions (*happiness*, *sadness*, *surprise*, *anger*, *disgust*, *fear*) by 42 speakers in English. 40 speakers (8 female) are selected in our experiments to obtain balanced numbers of samples in emotional and speakers' cate-

gories. This results in 1 200 samples in total and 200 samples for each emotion. Then, the corpus is divided into two folds (20 speakers in each fold) with the same size of samples. Similar as above, two-fold CV is adopted in the experiments.

B. Features

The feature extractor adopted in the experiments is our open-source tool *openSMILE* [59], [60]. We use the configuration of the INTERSPEECH 2013 Computational Paralinguistics Challenge (*ComParE*) [14], which also has been used in the INTERSPEECH 2014 to 2017 Computational Paralinguistics Challenges [61], [62]. The feature set of ComParE with the original dimensionality of 6 373 is obtained for each utterance. This set includes the features of 65 Low-Level Descriptors (LLDs) on different acoustic characteristics, with various statistical functionals, as well as some prosodic aspects. 54 functionals are applied to 59 LLDs, while 46 functionals are applied to the delta values of the 59 LLDs. To the other 6 LLDs and the corresponding delta LLDs, 39 functionals are applied. Additionally, 5 global temporal statistics are contained in this set. For full details of the ComParE features, the reader is referred to [63].

C. Preparations

The critical procedures of our proposed GMKDA are shown in Fig. 1. The algorithms of MKL-FDA and 2DMKDA are used for comparison in our experiments. In addition, experiments related to some basic dimensionality reduction methods (i.e., PCA, LDA or FDA, LPP, LLE, Isomap, LDP, GbFA) and classifiers (i.e., *k-Nearest Neighbors* (kNN), *Naive Bayesian classifier* (NB), *generalised Ridge Regression* (denoted as RR) [64], and *Support Vector Machines* (SVM)) are also provided to show the performance of our proposed algorithms. The SVM adopts a 'one-against-one' strategy, with the violation level in Sequential Minimal Optimisation (SMO) set as 0.001. The weight of the $l2$-norm minimisation term in RR is set as 0.001. In initial testing we tested the SVM with both linear and Gaussian kernels, where the Gaussian parameters in the kernelisation are chosen as $0.1n$, n, and $10n$. However, as the kernalised SVM did not perform as well as the linear set-up, the decision was made to proceed with a linear SVM.

Firstly, for each utterance sample, 6 373 features are extracted by *openSMILE*. Then, two-dimensional features are generated by M kernels and N training samples. In our previous research [6], satisfying results can be achieved by using multiscale Gaussian kernels. Therefore, we continue to keep this form and choose $M = 10$. The elements in the two-dimensional features for any sample x, corresponding to kernel m and training sample x_i, are

$$\Omega_{x_i,m} = \phi_m^T(x_i)\phi_m(x) = e^{-\frac{(x_i-x)^2}{\sigma_m^2}}, \tag{28}$$

where $m = 1, 2, \ldots, M$ and $i = 1, 2, \ldots, N$. $\phi_m(x)$ is the column vector in RKHS corresponding to kernel m and sample x. $\sigma_m > 0$ are the scaling parameters of Gaussian kernels.

The Gaussian scaling parameters $\{\sigma_1, \sigma_2, \ldots, \sigma_M\}$ are set as $\{0.001n, 0.005n, 0.01n, 0.03n, 0.05n, 0.1n, 0.3n, 0.5n, 0.75n, n\}$ [6]. Thus, we choose $d^{(B)}$ from 1 to 10, which results in maximal 11 mapping directions for multiple kernels, as in (4) and (14).

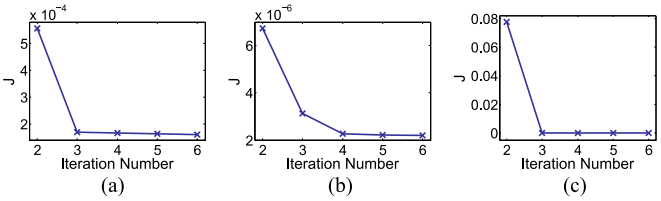


Fig. 3. The J value in each inner-loop iteration exemplified on (a) GEMEP, (b) ABC, (c) eNTERFACE, respectively.

We then perform (two-dimensional) dimensionality reduction on the generated two-dimensional features. The weight γ can be chosen as values between zero and one; results from multiple trials for different values of γ are given in Section V. The number of (outer loop) iterations in MKL-FDA, and the proposed 2DMKDA and GMKDA, is set as $T = 6$; it has previously been shown that the optimisation objects usually converge rapidly in very few iterations [31], [34]. For the inner loop, we set $T' = 10$. We exemplify the cost function $J = \lambda$ in (23) of each inner-loop iteration in Fig. 3.

The maximal dimensionality in all the dimensionality reduction methods is chosen as 100. The reduced dimensionalities are set as 15 for the corpus GEMEP, while as 8 for the corpora ABC and eNTERFACE, due to the mixing of maximal and minimal values of the optimisation problems involved in solving GEP, as well as considering the accuracy in numerical computation.

In the stage of final decision or classification, as adopted in our previous research [6], a kNN classifier is taken into consideration to highlight the basic performance of the proposed methods. We simply choose $k = 1$ as the nearest-neighbour case [65]. In addition, RR is also used in the decision stage, since the kNN classifier requires relatively high space complexity when there exists a pile of training samples [66].

The evaluation metrics here are *Unweighted Accuracy* (UA) (i.e., recall per class added and divided by the number of classes) and *Weighted Accuracy* (WA) [67].

V. Experimental Results

A. Performance Comparisons

The comparisons between GMKDA and other subspace-learning / conventional methods are now presented to completely demonstrate the performance of the proposed methods. Table II shows recognition accuracies (UA and WA) of some common existing subspace learning methods, including PCA, LDA or FDA, LPP, LLE, Isomap, LDP, and GbFA, as well as some conventional methods, including kNN, NB, RR, and SVM, comparing with the results of MKL-FDA, 2DMKDA, and GMKDA (using kNN and RR).

Results gained indicate that our proposed GMKDA paradigm achieves higher performance across all the three emotion corpora, when compared with the subspace-learning and conventional methods, as well as MKL-FDA and 2DMKDA (Table II). In particular, GMKDA obtained UA rates of 42.5%,

TABLE II
SMALL RECOGNITION ACCURACIES (UA AND WA) (%) OF SPEECH EMOTIONS ON THE CORPORA ABC, GEMEP, AND ENTERFACE RESPECTIVELY, USING THE EXISTING MKL-FDA METHOD, AS WELL AS OUR PROPOSED 2DMKDA AND GMKDA

Corpus		GEMEP (Training-Testing)		GEMEP (2-fold CV)		ABC		eNTERFACE
Methods		UA	WA	UA	WA	UA	WA	UA/WA
Subspace Learning Methods	PCA	30.5	29.2	26.5	26.4	40.8	45.8	39.7
	LPP [24]	20.6	20.2	20.2	20.3	32.5	37.3	35.3
	LLE [22]	24.6	24.3	21.7	21.8	35.9	40.0	36.5
	Isomap [23]	28.3	27.3	22.4	22.3	35.2	40.2	33.3
	LDA/FDA	34.6	33.8	33.8	33.5	43.9	52.2	55.2
	LDP [21]	33.7	32.9	33.1	32.8	41.6	50.2	55.1
	LDE [25]	36.5	35.7	35.5	35.4	47.0	56.4	58.7
	GbFA [26]	34.3	33.8	33.4	34.6	41.6	50.2	58.6
Conventional Methods	kNN	28.4	27.8	25.2	25.2	38.0	42.9	39.6
	NB	28.6	27.6	28.0	28.0	33.7	38.4	37.5
	RR [64]	33.0	32.9	33.3	33.2	41.9	49.6	55.7
	SVM	41.2	39.6	38.4	38.2	43.6	51.7	52.2
MKL-FDA [6]	with kNN	41.2	39.1	38.8	38.2	47.7	57.7	59.2
	with RR	28.9	28.5	31.0	29.4	34.9	43.9	58.6
2DMKDA	with kNN	37.8	36.3	35.6	35.0	44.3	53.5	57.2
	with RR	39.9	38.9	40.1	39.9	42.4	50.9	55.8
GMKDA	with kNN	**42.5**	40.7	39.4	39.0	**49.4**	**59.0**	**60.9**
	with RR	42.4	**41.7**	**42.1**	**41.9**	45.8	55.8	60.5

TABLE III
RECOGNITION ACCURACIES (UA AND WA) (%) OF SPEECH EMOTIONS ON THE CORPORA GEMEP (2-FOLD CV), ABC, AND ENTERFACE RESPECTIVELY, USING THE TOP-THREE-PERFORMANCE KERNELS RESPECTIVELY, COMPARED WITH GMKDA (WITH KNN)

Corpus	GEMEP		ABC		eNTERFACE
Methods	UA	WA	UA	WA	UA/WA
KFDA ($\sigma^{(1)}$)	37.6	37.1	46.7	56.3	60.4
KFDA ($\sigma^{(2)}$)	36.7	36.4	46.6	56.1	58.7
KFDA ($\sigma^{(3)}$)	36.4	36.1	46.1	54.9	58.3
GMKDA	**39.4**	**39.0**	**49.4**	**59.0**	**60.9**

42.1%, 49.4%, and 60.9% on the corpora GEMEP (Training-Testing), GEMEP (2-fold CV), ABC, and eNTERFACE respectively, whilst the corresponding UA rates for MKL-FDA were 41.2%, 38.8%, 47.7%, and 59.2% respectively. For the linear methods, results indicate that the supervised methods (LDA / FDA, LDP, LDE, and GbFA, etc.) generally perform better than the unsupervised ways (PCA, LPP, LLE, and Isomap, etc.). For this reason, we keep adopting the Fisher embedding graphs in our proposed methods.

We next present the top-three best recognition accuracies (UA and WA) (%) among the ten kernels [6] for KFDA (MKL-FDA with single-kernel structure) in Table III. It is indicated in Table III that, GMKDA outperforms KFDA by using a combination of multiple kernels. It is worth noting that, although the performance of KFDAs sometimes may be able to approach GMKDA, it is difficult to keep a stable performance as the parameters of the kernels change.

Then, Fig. 4 shows the UAs of SVM, RR, MKL-FDA, and the proposed GMKDA / GMKDA(raw) (respectively using NN and RR) on the corpora, where 'GMKDA(raw)' represents GMKDA without processing of PCA. According to the comparisons of the recognition accuracies, one can see that, the GMKDA / GMKDA(raw) is able to achieve relatively satisfying recognition results compared with MKL-FDA, SVM, and RR. This means that, the proposed framework and GMKDA enhance the performance of multiple kernel learning with multiscale Gaussian kernels in recognising speech emotions.

Statistical significance tests using a one-tailed *z-test* [68] suggest that, the GMKDA is significantly better compared with SVM and RR, at the significance levels of 0.05, 0.05, and 0.005, on the corpora of GEMEP, ABC, and eNTERFACE, respectively. Further, on the GEMEP corpus, GMKDA achieved significantly better performances when compared with the conventional subspace learning methods tested at the significance level of 0.05.

In addition to these comparisons, we also conduct experiments using some conventional feature selection methods, namely *Fisher Score*, *Relief-F* [69], and *mRMR* [70], in order to compare GMKDA with other subset-feature based methods. The number of the selected features was varied from $1\,000$ to $6\,000$, with the step size of 500. A linear SVM is chosen as the classifier for feature selection. On the GEMEP, ABC, and eNTERFACE corpora, Relief-F obtains the best UAs of 40.3%, 45.9%, and 53.7%, respectively. Thus, the GMKDA with UAs of 42.1%, 49.4%, and 60.9% respectively outperforms the feature selection methods, where the UA is significantly better on the eNTERFACE corpus at the significance level of 0.005.

B. Parameter Influence

The selection of γ can affect the performance of our proposed recognition system as it is changing the relation between MKL-FDA and 2DMKDA. For GMKDA, $\gamma \in [0, 1]$. As γ changes, GMKDA can decease to MKL-FDA (when $\gamma = 1$) and 2DMKDA (when $\gamma = 0$). In exploring this effect, we firstly show the UA of GMKDA, GMKDA(raw), MKL-FDA, and 2DMKDA, on the corpora GEMEP, ABC, and eNTERFACE

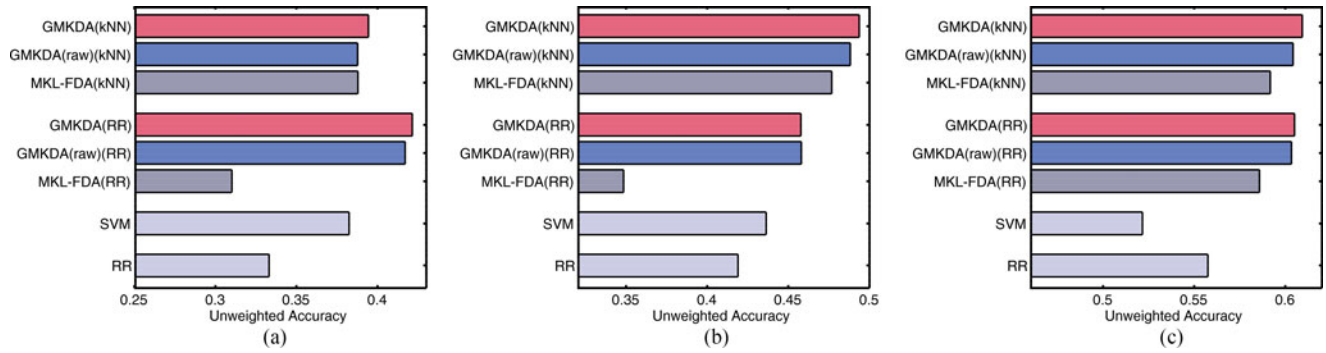


Fig. 4. Row charts of the UAs for the algorithms GMKDA (with kNN and RR), GMKDA(raw) (with kNN and RR), MKL-FDA (with kNN and RR), SVM, and RR, on the corpora (a) GEMEP (2-fold CV), (b) ABC, and (c) eNTERFACE, respectively. GMKDA(kNN): GMKDA using a kNN classifier; GMKDA(raw)(kNN): GMKDA using a kNN classifier without PCA processing; MKL-FDA(kNN): MKL-FDA using a kNN classifier; GMKDA(RR): GMKDA using an RR classifier; GMKDA(raw)(RR): GMKDA using an RR classifier without PCA processing; MKL-FDA(RR): MKL-FDA using an RR classifier.

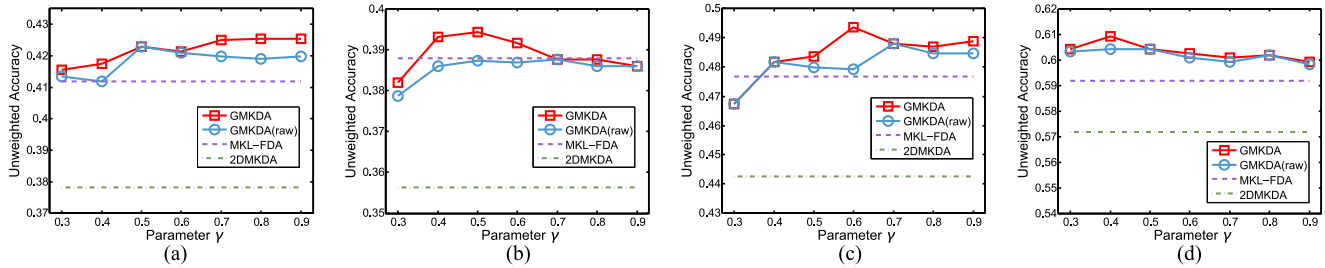


Fig. 5. Line charts of the UAs for the algorithms GMKDA and GMKDA(raw) with various parameters γ, as well as MKL-FDA and 2DMKDA, using the classifiers kNN, on the corpora (a) GEMEP (Training-Testing), (b) GEMEP (2-fold CV), (c) ABC, and (d) eNTERFACE, respectively.

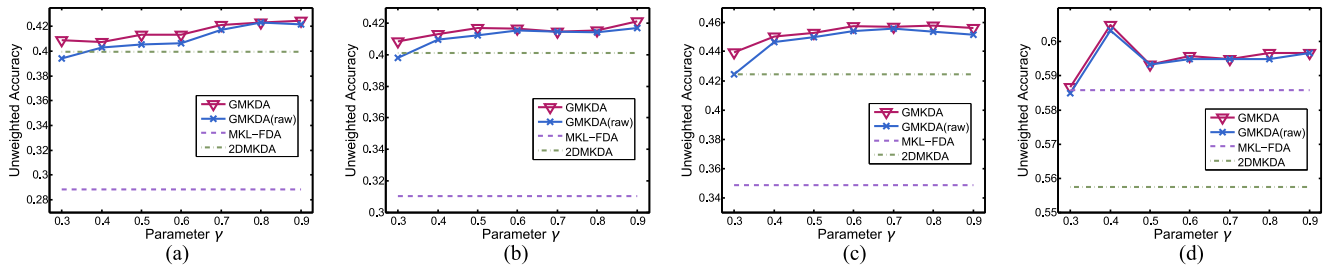


Fig. 6. Line charts of the UAs for the algorithms GMKDA and GMKDA(raw) with various parameters γ, as well as MKL-FDA and 2DMKDA, using the classifier RR, on the corpora (a) GEMEP (Training-Testing), (b) GEMEP (2-fold CV), (c) ABC, and (d) eNTERFACE, respectively.

using kNN in Fig. 5, when different values of γ are adopted from 0.3 to 0.9 in steps of 0.1. Similarly, Fig. 6 investigates the effects of changing γ when using a RR classifier.

Results presented in Fig. 5 indicate that, when kNN is adopted, the selection of the parameter γ affects the UA by a large margin. It can be seen that, the proposed GMKDA was able to outperform both MKL-FDA and 2DMKDA. However, the advantage in the performance of 2DMKDA is not so clear cut when comparing to MKL-FDA. The choice of γ also appears to be database dependent; the fluctuations of UA with the changes of parameter γ keep relatively stable on the copora of GEMEP and ABC. However, the fluctuation on eNTERFACE turns to be better when γ is relatively small in our selections. This may be caused by the problem of computational accuracy in choosing the maximal numbers of reduced dimensionality.

In Fig. 6, the previously utilised MKL-FDA holds low performance on the corpora of GEMEP and ABC with the classifier of RR, while the proposed 2DMKDA can achieve better results. This may result from the fact that 2DMKDA provides more information in the step of regression for RR. Further, depending on choosing suitable parameters γ, the proposed GMKDA is able to outperform 2DMKDA, though there exist some fluctuations as γ changes, probably due to the computational accuracy. However, this trend changes on the eNTERFACE database, where the performance of MKL-FDA exceeds 2DMKDA. One can learn from Figs. 5 and F6 that, the classifier RR shows more stable performance compared with kNN.

Based on the experiments with regard to the parameters, we further show a set of brief results on recognition accuracies (UA and WA) using kNN and RR in Table IV. The γs corresponding to the highest performance are also listed in the two tables. Note that we choose multiple numbers of column vectors for A and $\bar{\mathbf{B}}$ or B, due to the influence from nontrival eigenvectors in solving GEP, and the computational accuracy in the iterative steps.

It can be concluded from Table IV that, when kNN is utilised as the classifier, the proposed 2DMKDA (without nonnegative

TABLE IV
RECOGNITION ACCURACIES (UA AND WA) (%) OF SPEECH EMOTIONS ON THE CORPORA GEMEP, ABC, AND ENTERFACE RESPECTIVELY, USING THE EXISTING MKL-FDA METHOD, AS WELL AS OUR PROPOSED 2DMKDA, GMKDA, AND GMKDA(RAW) METHODS, WITH THE CLASSIFIERS KNN AND RR

Corpus	GEMEP (Training-Testing)		GEMEP (2-fold CV)		ABC		eNTERFACE
Methods	UA	WA	UA	WA	UA	WA	UA/WA
Classifier: kNN							
MKL-FDA	41.2	39.1	38.8	38.2	47.7	57.7	59.2
2DMKDA	37.8	36.3	35.6	35.0	44.3	53.5	57.2
GMKDA(raw)	42.3 ($\gamma = 0.5$)	40.3 ($\gamma = 0.5$)	38.8 ($\gamma = 0.7$)	38.3 ($\gamma = 0.7$)	48.8 ($\gamma = 0.7$)	59.0 ($\gamma = 0.7$)	60.4 ($\gamma = 0.4$)
GMKDA	**42.5** ($\gamma = 0.8$)	**40.7** ($\gamma = 0.6$)	**39.4** ($\gamma = 0.5$)	**39.0** ($\gamma = 0.5$)	**49.4** ($\gamma = 0.6$)	**59.0** ($\gamma = 0.7$)	**60.9** ($\gamma = 0.4$)
Classifier: RR							
MKL-FDA	28.9	28.5	31.0	29.4	34.9	43.9	58.6
2DMKDA	39.9	38.9	40.1	39.9	42.4	50.9	55.8
GMKDA(raw)	42.3 ($\gamma = 0.8$)	41.2 ($\gamma = 0.8$)	41.7 ($\gamma = 0.9$)	41.5 ($\gamma = 0.9$)	45.6 ($\gamma = 0.7$)	55.7 ($\gamma = 0.7$)	60.3 ($\gamma = 0.4$)
GMKDA	**42.4** ($\gamma = 0.9$)	**41.7** ($\gamma = 0.9$)	**42.1** ($\gamma = 0.9$)	**41.9** ($\gamma = 0.9$)	**45.8** ($\gamma = 0.8$)	**55.8** ($\gamma = 0.6$)	**60.5** ($\gamma = 0.4$)

For GMKDA and GMKDA(raw), the γs corresponding to the best results are also presented.

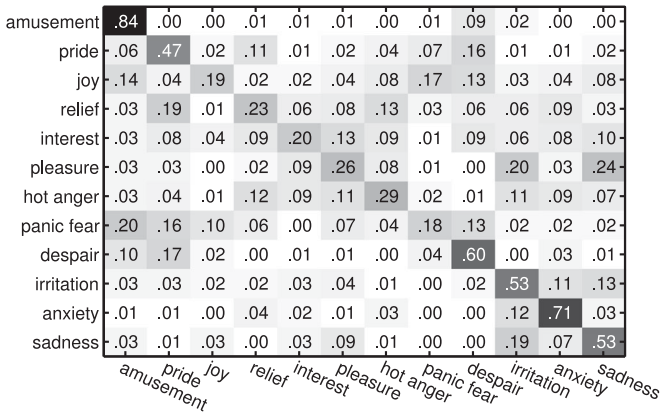


Fig. 7. Recall matrix when using GMKDA with RR on GEMEP.

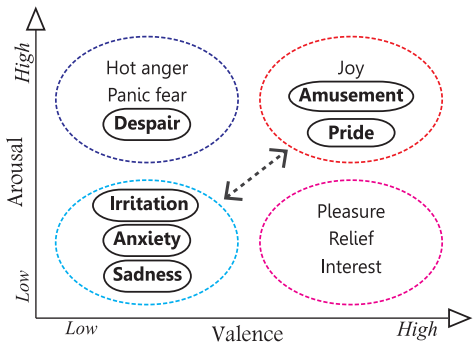


Fig. 8. The analysis of emotions on the dimensions of valence and arousal, on GEMEP when using GMKDA with RR.

constraints) fails to outperform MKL-FDA (with nonnegative constraints). Even though the relaxation in 2DMKDA does not result in an improvement, our proposed GMKDA is still able to achieve better performance compared with MKL-FDA, by balancing the linear combinations with and without nonnegative constraints. Table IV also indicates that when using RR, the proposed GMKDA outperforms the systems MKL-FDA (with $\gamma = 1$) and 2DMKDA (with $\gamma = 0$).

C. Emotional Analysis

Finally, we investigate the emotion analysis on the GEMEP database with 12 emotional classes. The recall matrix using the proposed GMKDA with RR on the GEMEP database is presented in Fig. 7. It is learnt from Fig. 7 that compared with the remaining emotions, the emotional classes *amusement*, *pride*, *despair*, *irritation*, *anxiety*, and *sadness* are relatively easier to recognise. These emotions achieve the recalls of 84%, 47%, 60%, 53%, 71%, and 53% respectively, while for MKL-FDA, the recalls corresponding to the emotions are 59%, 41%, 38%, 50%, 62%, and 40% respectively. This indicates that the proposed GMKDA achieves higher accuracies on recognising these emotions. Thus further improvements, i.e., methods fusion, could be employed on the basis of the analysis.

Further, employing the general psychological classification on the dimensions of *valence* and *arousal* [55], the emotions of GEMEP are analysed in Fig. 8. As shown in the figure, most of the easily recognised emotions possess high *valence* and *arousal*, or low *valence* and *arousal*. In addition, for the remaining emotional classes, most of their easily-confused classes do not lie in their 'opposite' parts in the *valence-arousal* space, which indicates that it is less likely to confuse high-*valence* / low-*arousal* and low-*valence* / high-*arousal* emotions. Comparing the recalls of MKL-FDA and GMKDA, it is more likely to correctly recognise these emotional pairs. In addition, we specifically investigated the emotional state of 'despair', which has a relative high recall as a low-*valence* / high-*arousal* case. Results presented in Fig. 7, indicate that this emotional state is easy to be classified as 'pride', which also has high arousal. This partially supports a commonly reported finding for emotion recognition in speech, that classifying valence is more difficult than classifying arousal.

VI. CONCLUSION

In this paper, a two-dimensional multiple kernel subspace learning framework is proposed to be applied in speech emotion recognition. By extending the optimisation of multiple kernel learning subspace learning, this framework leverages both multilple kernel learning and two-dimensional subspace learning

and combines them into a unified structure. On the basis of the framework, we develop the algorithm GMKDA by using Fisher discriminant embedding graphs. The proposed GMKDA benefits from jointly employing the nonnegative linear combination and the mapping directions without nonnegative constraints for multiple kernels.

This gives rise to learning more optimised information representations effective for recognising emotions. The experimental results gained on the emotional corpora when using multiscale kernels indicate that, the proposed GMKDA achieves improved performance when compared with both the previously proposed MKL-FDA and conventional methods.

An initial set of experimental results indicates that the proposed GMKDA method generally outperforms conventional methods on the emotional corpora. A secondary set of experiments focused on comparing the GMKDA, MKL-FDA, and 2DMKDA methodologies. These results indicate that, the regulation of the weight of the nonnegative kernel combination, γ, allows our proposed method to obtain better results when compared with MKL-FDA and 2DMKDA, both of which are special cases of GMKDA. Further analysis on emotions exemplified on GEMEP indicates the detailed performance of the proposed method for further applications.

Future work includes the following aspects. First, although the proposed framework is theoretically reasonable, the calculation in an application cannot be guaranteed to be convergent, since the solution in optimising the mapping directions A is usually not optimal. Thus, constraints on A should be added in the optimisation. Second, the current research only employs multiscale kernels in generating two dimensional features. Hence, in our future research, a more desirable generation of two-dimensional features will be investigated in order to obtain better performance. Third, the embedding graphs in GMKDA only contain discriminant information, without considering neighbouring relationship of training samples. Accordingly, it is promising to work on constructing embedding graphs with a relatively valid representation for paralinguistic application. In addition, following our research, tensorised extensions on this two-dimensional framework are expected to be conducted in more applications.

APPENDIX A
PROOF OF THE THEORETICAL CONVERGENCE OF THE INNER ITERATIVE TRACE RATIO LOOP

For the first time of the inner-loop iteration, we have

$$J_0(\beta^{(1)}, B^{(1)}, \lambda^{(0)}) = tr(G_0(\beta^{(1)}, B^{(1)}, \lambda^{(0)})) \leq 0, \quad (29)$$

where

$$G_0(\beta^{(1)}, B^{(1)}, \lambda^{(0)}) = \begin{bmatrix} \gamma\beta^{(1)^T} \\ (1-\gamma)B^{(1)^T} \end{bmatrix} \times (Q^{(I)}(A) - \lambda^{(0)} Q^{(P)}(A)) \left[\gamma\beta^{(1)}(1-\gamma)B^{(1)}\right] \quad (30)$$

is negative semi-definite with a large positive initial value $\lambda^{(0)}$.

Then, it can be drawn according to (29) that,

$$\lambda^{(0)} \geq \frac{tr\left(\left[\gamma\beta^{(1)}\ (1-\gamma)B^{(1)}\right]^T Q^{(I)}(A)\left[\gamma\beta^{(1)}\ (1-\gamma)B^{(1)}\right]\right)}{tr\left(\left[\gamma\beta^{(1)}\ (1-\gamma)B^{(1)}\right]^T Q^{(P)}(A)\left[\gamma\beta^{(1)}\ (1-\gamma)B^{(1)}\right]\right)}, \quad (31)$$

since $Q^{(P)}(A)$ is positive semi-definite. With the right part of (31) equal to $\lambda^{(1)}$, we have $\lambda^{(0)} \geq \lambda^{(1)}$.

For $l' \geq 2$, it is obtained that,

$$J_0(\beta^{(l')}, B^{(l')}, \lambda^{(l'-1)}) \leq J_0(\beta^{(l'-1)}, B^{(l'-1)}, \lambda^{(l'-1)}) = 0, \quad (32)$$

which results in

$$\lambda^{(l'-1)} \geq \lambda^{(l')} = \frac{tr\left(\left[\gamma\beta^{(l')}(1-\gamma)B^{(l')}\right]^T Q^{(I)}(A)\left[\gamma\beta^{(l')}\ (1-\gamma)B^{(l')}\right]\right)}{tr\left(\left[\gamma\beta^{(l')}(1-\gamma)B^{(l')}\right]^T Q^{(P)}(A)\left[\gamma\beta^{(l')}(1-\gamma)B^{(l')}\right]\right)}. \quad (33)$$

Combining (31) and (33), we draw that, λ is monotonically non-increasing in calculating the inner-loop iteration, which indicates that the objective function of (16) is monotonically non-increasing in each step of the iteration. In addition, $\lim_{l'\to\infty} \| [\gamma\beta^{(l'+1)}(1-\gamma)B^{(l'+1)}] - [\gamma\beta^{(l')}(1-\gamma)B^{(l')}] \| = 0$ is similarly drawn according to [46]. ■

ACKNOWLEDGMENT

The authors would like to thank H. Sagha, W. Zheng, and the anonymous reviewers for their valuable help.

REFERENCES

[1] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Hoboken, NJ, USA: Wiley, 2013.

[2] A. Batliner *et al.*, "The automatic recognition of emotions in speech," in *Emotion-Oriented Systems*. New York, NY, USA: Springer, 2011, pp. 71–99.

[3] M. J. Gangeh, P. Fewzee, A. Ghodsi, M. S. Kamel, and F. Karray, "Multiview supervised dictionary learning in speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 6, pp. 1056–1068, Jun. 2014.

[4] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.

[5] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011.

[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: ISCA, 2015, pp. 1532–1536.

[7] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, no. 9, pp. 1062–1087, 2011.

[8] A. Sciarrone, A. Delfino, M. Marchese, F. Lavagetto, and I. Bisio, "Gender-driven emotion recognition through speech signals for ambient intelligence applications," *IEEE Trans. Emerg. Topics Comput.*, vol. 1, no. 2, pp. 244–257, Dec. 2013.

[9] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization," in *Proc. Afeka-AVIOS Speech Process. Conf.*, Tel Aviv, Israel: ACLP, 2011, pp. 157–163.

[10] J. Deng, Z. Zhang, and B. Schuller, "Linked source and target domain subspace feature transfer learning–exemplified by speech emotion recognition," in *Proc. Int. Conf. Pattern Recognit.*, Stockholm, Sweden, pp. 761–766.

[11] E. Marchi, A. Batliner, B. Schuller, S. Fridenzon, S. Tal, and O. Golan, "Speech, emotion, age, language, task, and typicality: Trying to disentangle performance and feature relevance," in *Proc. Int. Conf. Priv. Secur. Risk Trust Int. Conf. Soc. Comput*., Amsterdam, The Netherlands: IEEE, 2012, pp. 961–968.

[12] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[13] 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.

[14] 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: ISCA, 2013, pp. 148–152.

[15] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting autism, emotions and social signals using adaboost," in *Proc. Annu. Conf. Int. Speech Commun. Assoc*., Lyon, France: ISCA, 2013, pp. 220–224.

[16] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc*., Brighton, U.K.: ISCA, 2009, pp. 312–315.

[17] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2015.

[18] P. Fewzee and F. Karray, "Dimensionality reduction for emotional speech recognition," in *Proc. Int. Conf. Priv., Secur. Risk Trust Int. Conf. Soc. Comput.*, Amsterdam, The Netherlands: IEEE, 2012, pp. 532–537.

[19] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. IV, Honolulu, HY, USA: IEEE, 2007, pp. 941–944.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "A multi-task approach to continuous five-dimensional affect sensing in natural speech," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 1, 2012, Art. no. 6.

[21] 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.

[22] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[23] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[24] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst*. Vancouver and Whistler, BC, Canada: MIT Press, 2003, pp. 153–160.

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

[26] 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.

[27] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "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.

[28] F. De la Torre, "A least-squares framework for component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1041–1055, Jun. 2012.

[29] S. Wang, S. Yan, J. Yang, C.-G. Zhou, and X. Fu, "A general exponential framework for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 920–930, Feb. 2014.

[30] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[31] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.

[32] H. Zhang, Q. J. Wu, T. W. Chow, and M. Zhao, "A two-dimensional neighborhood preserving projection for appearance-based face recognition," *Pattern Recognit.*, vol. 45, no. 5, pp. 1866–1876, 2012.

[33] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

[34] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. Adv. Neural Inf. Process. Syst*. Vancouver, BC, Canada: MIT Press, 2004, pp. 1569–1576.

[35] J. Yang, D. Zhang, X. Yong, and J.-y. Yang, "Two-dimensional discriminant transform for face recognition," *Pattern Recognit.*, vol. 38, no. 7, pp. 1125–1129, 2005.

[36] D. Hu, G. Feng, and Z. Zhou, "Two-dimensional locality preserving projections (2DLPP) with its application to palmprint recognition," *Pattern Recognit.*, vol. 40, no. 1, pp. 339–342, 2007.

[37] S.-J. Kim, A. Magnani, and S. Boyd, "Optimal kernel selection in Kernel Fisher discriminant analysis," in *Proc. Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA: ACM, 2006, pp. 465–472.

[38] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Dimensionality reduction for data in multiple feature representations," in *Proc. Adv. Neural Inf. Process. Syst*. Vancouver, BC, Canada: MIT Press, 2009, pp. 961–968.

[39] Z. Wang and X. Sun, "Multiple kernel local Fisher discriminant analysis for face recognition," *Signal Process.*, vol. 93, no. 6, pp. 1496–1509, 2013.

[40] M. Uzair, A. Mahmood, and A. Mian, "Sparse kernel learning for image set classification," in *Computer Vision–ACCV 2014*. Berlin, Germany: Springer, 2014, pp. 617–631.

[41] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, "Multiple kernel sparse representations for supervised and unsupervised learning," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2905–2915, Jul. 2014.

[42] A. Nazarpour and P. Adibi, "Two-stage multiple kernel learning for supervised dimensionality reduction," *Pattern Recognit.*, vol. 48, no. 5, pp. 1854–1862, 2015.

[43] A. d'Aspremont and S. Boyd, "Relaxations and randomized methods for nonconvex QCQPs," EE392o Class Notes, Stanford Univ., Stanford, CA, USA, 2003.

[44] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "A survey of multilinear subspace learning for tensor data," *Pattern Recognit.*, vol. 44, no. 7, pp. 1540–1551, 2011.

[45] 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.

[46] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Minneapolis, MN, USA: IEEE, 2007, pp. 1–8.

[47] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 729–735, Apr. 2009.

[48] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc*., Makuhari, Japan: ISCA, 2010, pp. 2794–2797.

[49] B. Schuller *et al.*, "The interspeech 2012 speaker trait challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc*. Portland, OR, USA: ISCA, 2012.

[50] 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.

[51] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarrone, "Gender-driven emotion recognition through speech signals for ambient intelligence applications," *IEEE Trans. Emerging Top. Comput.*, vol. 1, no. 2, pp. 244–257, Dec. 2013.

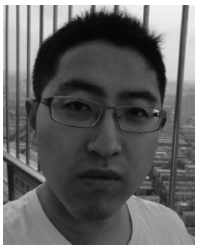
[52] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, Honolulu, Hawaii: IEEE, 2007, pp. II–733.

[53] 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: IEEE, 2006.

[54] 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.

[55] T. Bänziger and K. R. Scherer, "Introducing the Geneva multimodal emotion portrayal (GEMEP) corpus," in *Blueprint for Affective Computing: A Sourcebook*, A Sourcebook, K. R. Scherer, T. Bänziger, and E. B. Roesch, Eds. Oxford, U.K.: Oxford Univ. Press, 2010, pp. 271–294.

[56] Y.-L. Chen and C.-T. Hsu, "Multilinear graph embedding: Representation and regularization for images," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 741–754, Feb. 2014.
[57] M. Li, J. T.-Y. Kwok, and B. Lü, "Making large-scale Nyström approximation possible," in *Proc. Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 631–638.
[58] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-life acoustic conditions—A new perspective on feature selection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Lyon, France: ISCA, 2013, pp. 2044–2048.
[59] 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: ACM, 2013, pp. 835–838.
[60] F. Eyben and B. Schuller, "openSMILE:) The Munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Rec.*, vol. 6, no. 4, pp. 4–13, 2015.
[61] B. Schuller *et al.*, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proc. Annu. Conf. Int. Speech Commun. Assoc*., Singapore: ISCA, 2014, pp. 427–431.
[62] B. Schuller *et al.*, "The INTERSPEECH 2015 computational paralinguistics challenge: Degree of nativeness, Parkinson's & eating condition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc*., Dresden, Germany: ISCA, Sep. 2015, pp. 478–482.
[63] F. Eyben, *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction*. New York, NY, USA: Springer, 2016.
[64] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA: IEEE, 2007, pp. 1–7.
[65] X. Liu, S. Yan, and H. Jin, "Projective nonnegative graph embedding," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1126–1137, May 2010.
[66] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, pp. 249–268, 2007.
[67] 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. Underst.*, Merano, Italy: IEEE, 2009, pp. 552–557.
[68] S. L. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Min. Knowl. Discovery*, vol. 1, no. 3, pp. 317–328, 1997.
[69] H. Liu and H. Motoda, *Computational Methods of Feature Selection*. Boca Raton, FL, USA: CRC Press, 2007.
[70] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.



Xinzhou Xu received the Bachelor's degree from Nanjing University of Posts and Telecommunications, Nanjing, China, and the Master's degree from Southeast University, Nanjing, China, in 2009 and 2012, respectively. He is currently working toward the Ph.D. degree in Southeast University, Nanjing, China. He is also with the Machine Intelligence & Signal Processing group, MMK, Technische Universität München, München, Germany, and with the Chair of Complex and Intelligent Systems in the University of Passau, Germany. 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, Haerbin, China, in 2009, the Master's degree in information and communication engineering from Harbin Institute of Technology, Heilongjiang, China, in 2011, and the Doctoral degree for his study on feature transfer learning for speech emotion recognition, in electrical engineering and information technology from Technische Universität München (TUM), Munich, Germany, in 2016. He is currently a Postdoctoral Researcher at the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany. His research interest focuses on machine learning methods such as transfer learning and deep learning with an application preference to affective computing.

Nicholas Cummins (M'13) received the undergraduate degree at the University of New South Wales (UNSW), Sydney, Australia, in 2011 and the Ph.D. degree in electrical engineering from UNSW in February 2016. His Ph.D. investigated whether the voice can be used as an objective marker in the diagnosis and monitoring of clinical depression. He is a Postdoctoral Researcher at the Chair of Complex and Intelligent Systems, Universität Passau, Passau, Germany, where he is involved in the EU-FP7 starting grant project iHEARu. His current research includes areas of behavioural signal processing with a focus on the automatic analysis and understanding of speaker characteristics. He has published regularly in the field of depression detection since 2011; these papers have attracted significant attention and citations. He led the 2nd ranked team in the AVEC2013 Depression Challenge and was a member of the 2nd ranked team in the AVEC2014 Depression Challenge. Previous to starting his degree, he worked for eight years as an electrician in Australia, the U.K., and Ireland.

Zixing Zhang (M'15) received the Ph.D. degree in engineering from the Institute for Human-Machine Communication at Technische Universität München (TUM), Germany, 2015. Before that, he earned his master degree in physical electronics from Beijing University of Posts and Telecommunications (BUPT), China, 2010. Currently, he is a postdoctoral researcher at the University of Passau, Germany. Until now, he has authored more than thirty publications in peer-reviewed journals and conference proceedings. His research interests mainly lie in deep learning, semi-supervised learning, active learning, and multi-task learning, in the applications of computational paralinguistics (e.g., emotion recognition) and robust automatic speech recognition.

Chen Wu received the bachelor's degree from Nanjing University of Aeronautics and Astronautics, Nanjing, China, and the master's degree from Southeast University, Nanjing, China, in 2009 and 2012, respectively. He is currently working toward the Ph.D. degree in Southeast University. His research interests include signal processing and pattern recognition.

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

Björn Schuller (M'05–SM'15) received his diploma in 1999, his doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, and his habilitation and Adjunct Teaching Professorship in the subject area of Signal Processing and Machine Intelligence in 2012, all in electrical engineering and information technology from TUM in Munich/Germany. He is Reader in Machine Learning in the Department of Computing at the Imperial College London/UK, Full Professor and head of the Chair of Complex and Intelligent Systems at the University of Passau/Germany where he previously headed the Chair of Sensor Systems in 2013, and an Associate of the Swiss Center for Affective Sciences at the University of Geneva. Dr. Schuller is president of the Association for the Advancement of Affective Computing (AAAC), elected member of the IEEE Speech and Language Processing Technical Committee, and member of the ACM, IEEE and ISCA and (co-)authored 5 books and more than 600 publications in peer reviewed books, journals, and conference proceedings leading to more than 13 000 citations (h-index = 56).