

## Multiscale kernel locally penalised discriminant analysis exemplified by emotion recognition in speech

Xin Zhou Xu, Jun Deng, Maryna Gavryukova, Zixing Zhang, Li Zhao, Björn Schuller

### Angaben zur Veröffentlichung / Publication details:

Xu, Xin Zhou, Jun Deng, Maryna Gavryukova, Zixing Zhang, Li Zhao, and Björn Schuller. 2016. "Multiscale kernel locally penalised discriminant analysis exemplified by emotion recognition in speech." In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, November 12 - 16, 2016, edited by Yukiko Nakano, 233-37. New York, NY: ACM.  
<https://doi.org/10.1145/2993148.2993184>.

### Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under the following conditions:

#### Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publizieren>



# Multiscale Kernel Locally Penalised Discriminant Analysis Exemplified by Emotion Recognition in Speech

Xinzhou Xu<sup>1,2,3</sup>, Jun Deng<sup>3</sup>, Maryna Gavryukova<sup>3</sup>, Zixing Zhang<sup>3</sup>, Li Zhao<sup>1</sup>, and Björn Schuller<sup>3,4</sup>

<sup>1</sup>School of Information Science and Engineering, Southeast University, China

<sup>2</sup>MISP group, MMK, Technische Universität München, Germany

<sup>3</sup>Chair of Complex and Intelligent Systems, Universität Passau, Germany

<sup>4</sup>Department of Computing, Imperial College London, U.K.

*xinzhou.xu@tum.de, jun.deng@uni-passau.de*

## ABSTRACT

We propose a novel method to learn multiscale kernels with locally penalised discriminant analysis, namely Multiscale-Kernel Locally Penalised Discriminant Analysis (MS-KLPDA). As an exemplary use-case, we apply it to recognise emotions in speech. Specifically, we employ the term of locally penalised discriminant analysis by controlling the weights of marginal sample pairs, while the method learns kernels with multiple scales. Evaluated in a series of experiments on emotional speech corpora, our proposed MS-KLPDA is able to outperform the previous research of Multiscale-Kernel Fisher Discriminant Analysis and some conventional methods in solving speech emotion recognition.

## CCS Concepts

•Information systems → Multimedia information systems; •Human-centered computing → HCI theory, concepts and models; •Theory of computation → Kernel methods;

## Keywords

Locally penalised discriminant analysis; Multiscale kernels; Multiple kernel learning, Speech emotion

## 1. INTRODUCTION

With increasing requirements of advanced intelligent human-computer multimodal interaction, recognising affect and emotions on the basis of spoken signals has shown broad potential such as analysing speaker states [17,18]. The research in the field of Speech Emotion Recognition mainly focuses on exploring suitable feature sets based on prior knowledge [2,13,20], while few works [6,7,21] shed light on mining compact emotional representation from given features.

In order to extract efficient factors from a fixed feature set, common dimensionality reduction methods, including Principal Component Analysis (PCA), Fisher Discriminant Analysis (FDA), Locally Discriminant Embedding (LDE) [5],

Linear Discriminant Projections (LDP) [4], or Locally Linear Embedding (LLE) [14], have been considered, which can be generally transformed into Graph Embedding (GE) frameworks [22,23].

In previous research [21], Multiscale-Kernel Fisher Discriminant Analysis (MS-KFDA) has been proposed to extract efficient features for emotion recognition and general paralinguistics. It has been shown that, learning FDA embedding graphs combined with Gaussian kernels using multiple scaling parameters are beneficial for the task of speech emotion recognition, based on the theory of Multiple Kernel Learning (MKL) [12,19]. However, the method ignored the marginal penalised information, which has been raised in LDE [5] and Marginal Fisher Analysis (MFA) [23]. It may, however, reduce the robustness of the recognition system, since the sample pairs corresponding to outliers hold the same weights as any other pairs.

Inspired by the deficiencies above, we propose a novel algorithm, namely MS-KLPDA (Multiscale-Kernel Locally Penalised Discriminant Analysis), and demonstrate its successful application for recognising emotions in speech. MS-KLPDA employs a locally penalised structure in a penalty embedding graph, which increases the weights of the marginal sample pairs. As in MS-KFDA, the same structure of the intrinsic embedding graph and multiscale Gaussian kernels appears in MS-KLPDA.

In contrast to existing related works, the research in this paper contains novelty as follows: Compared to [21], our research adds a locally penalised term in the structure of the penalty embedding graph, which is more suitable for the considered task of emotion recognition in speech. In [12], MKL dimensionality reduction has been proposed by a large amount of kernels, while our research focuses on adopting a kernel with multiple scales and the specifically designed optimisation form for emotion recognition in speech. Compared with [19], the proposed method dispenses utilising the local intrinsic structure to avoid effects from ‘disturbing’ features and tuning in neighbourhood.

The remainder is structured as follows: Section 2 presents the theory of the proposed MS-KLPDA. Then, experiments and results are discussed aimed to evaluate the proposed method in Section 3.

## 2. METHODOLOGY

In previous research using FDA by multiscale kernels [21], the embedding graphs only contain supervised information, which ignores the between-class marginal structure of the penalty graph [5,23]. Here, we propose a novel algorithm

This is the author's version of the work. It is posted here for your personal use. Not for redistribution.

ICMI'16, November 12–16, 2016, Tokyo, Japan  
c 2016 ACM. 978-1-4503-4556-9/16/11...\$15.00  
<http://dx.doi.org/10.1145/2993148.2993184>

m, namely MS-KLPDA, by adopting Locally Penalised Discriminant Analysis (LPDA) as the critical section for optimisation. On the basis of the FDA embedding graphs, the marginal penalty term of LPDA is added in the penalty graph, aiming to penalise local between-class sample pairs.

However, compared with existing methods [5, 23], LPDA avoids to take the neighbouring intrinsic graph into consideration, since the neighbouring information may lead to unfavourable structures in some conditions of computational paralinguistics. In addition, the performance of neighbouring information largely depends on parameters to be tuned. Further, LPDA leads to a convenient control of the marginal sample pairs. This takes advantage of the penalty term by flexibly regulating its weight.

Then, multiscale kernels [21] are learnt by combining the locally penalised embedding graphs and multiple kernel learning, in order to achieve a more optimised value by means of bilateral iterations. Afterwards, the subspace features are fed into the decision maker (i.e., classifiers) to obtain the recognition result. We show the detailed methodology as follows.

## 2.1 Multiscale Kernels

It is assumed that  $N$   $n$ -dimensional training samples  $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{n \times N}$  are given along with their corresponding labels  $Y = [y_1, y_2, \dots, y_N] \in \mathbb{R}^{d \times N}$ . The Reproducing Kernel Hilbert Space (RKHS) of  $X$  is represented as  $\phi(X) = [\phi(x_1), \phi(x_2), \dots, \phi(x_N)]$ , which leads to the Gram matrix as  $K = \phi^T(X)\phi(X)$ . For any sample  $x$  with  $\phi(x)$  in RKHS, the Gram mapping of  $x$  is defined as  $K_x = \phi^T(X)\phi(x)$ .

When multiple kernels [12] are used,  $K_x$  can be written as the linear combination of  $M$  different kernels, namely

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

where the multiple kernel coordinate matrix is

$$\Omega_x = [\phi_1^T(X)\phi_1(x), \phi_2^T(X)\phi_2(x), \dots, \phi_M^T(X)\phi_M(x)], \quad (2)$$

with  $\Omega_x \in \mathbb{R}^{N \times M}$ .  $\beta \in \mathbb{R}^{M \times 1}$  is the column vector with corresponding elements  $\beta_m$  for the kernel  $m$ . The number of kernels is  $M$ . Each column of  $\Omega_x$  is the corresponding coordinate for a sample  $x$ .

We utilise Gaussian kernels with multiple scales in the proposed MS-KLPDA. The  $m$ th element (corresponding to scale  $m$ ) of  $\Omega_x$  is

$$\Omega_{x_i, m} = \phi_m^T(x_i)\phi_m(x) = e^{-\frac{(x_i - x)^2}{\sigma_m^2}}, \quad (3)$$

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

## 2.2 MS-KLPDA

Suppose for any sample  $x$ , we hope to learn its optimal subspace representation  $A^T x$  by calculating the mappings  $A$ . Then for each pair of training samples  $x_i$  and  $x_j$ , the squared distance of  $A^T K_{x_i}$  and  $A^T K_{x_j}$  can be weighted in order to make the sum of the weighted values minimal or maximal. Explicitly represented by graphs, these weights depends on the relation between training samples. In accordance with GE, the weights are written as the corresponding elements of the adjacency matrices belonging to the intrinsic

(in minimal case) and penalty (in maximal case) embedding graphs, namely  $W^{(I)}$  and  $W^{(P)}$ .

Thus, combining Eq. (1), the optimisation form of MS-KLPDA is shown as

$$\begin{aligned} & \underset{A, \beta}{\operatorname{argmin}} \sum_{i, j=1}^N \|A^T \Omega_{x_i} \beta - A^T \Omega_{x_j} \beta\|^2 W_{ij}^{(I)} \\ & \text{s.t.} \sum_{i, j=1}^N \|A^T \Omega_{x_i} \beta - A^T \Omega_{x_j} \beta\|^2 W_{ij}^{(P)} = \gamma, \\ & \beta_m \geq 0, \quad m = 1, 2, \dots, M, \end{aligned} \quad (4)$$

using the multiple optimised mapping directions  $A = [\alpha_1, \alpha_2, \dots, \alpha_d] \in \mathbb{R}^{N \times d}$ , we obtain multiple mappings by solving the optimisation problem, where  $d$  stands for the dimensionality of the dimensionality-reduced feature space.  $\alpha_i$  is the  $i$ th mapping vector with  $i = 1, 2, \dots, d$  and the  $\gamma > 0$  represents a constant value.

In [21], the embedding graphs only include supervised information, in which the neighbouring between-class sample pairs are not weighted. Consequently, we use the embedding graphs with an additional penalty term, in order to penalise marginal pairs.

The adjacency matrices of the intrinsic and penalty embedding graphs are defined as

$$\begin{cases} W^{(I)} = S^T (S S^T)^{-1} S, \\ W^{(P)} = \frac{1}{N} e e^T + \delta ((e e^T - S^T S) \odot W_{k_0 N N}), \end{cases} \quad (5)$$

where each column of  $S = [s_1, s_2, \dots, s_N] \in \mathbb{R}^{c \times N}$  indicates the label information of every corresponding training sample, where  $c$  is the number of classes.  $S_{ij} = 1$  when sample  $j$  belongs to class  $i$ , otherwise  $S_{ij} = 0$ , where  $i = 1, 2, \dots, c$  and  $j = 1, 2, \dots, N$ . Every element of  $e \in \mathbb{R}^{N \times 1}$  is equal to 1. The operator  $\odot$  represents the element-wise product between two matrices.

$W_{k_0 N N}$  is the  $k_0$  nearest-neighbour adjacency matrix of training samples, where the elements  $(W_{k_0 N N})_{ij} = 1$  (in the current approach) or  $e^{-\frac{\|x_i - x_j\|^2}{t}}$  (for an improved approach), when  $x_i$  is among  $k_0$  nearest-neighbours of  $x_j$  or vice versa, with  $i, j = 1, 2, \dots, N$  and the constant value  $t > 0$ . Otherwise,  $(W_{k_0 N N})_{ij} = 0$ .  $\delta > 0$  denotes the weight of the locally penalised discriminant term.

It is worth noticing that, in Eq. (5) the intrinsic graph and the first term of the penalty graph here are the same as in FDA, while the second term of the penalty graph is similar to the penalty graph in LDE [5] and MFA [23].

Accordingly, the proposed MS-KLPDA can be solved by alternative iteration of Eq. (6) (solving kernel mappings  $A$ ), and Eq. (7) (solving nonnegative linear weights  $\beta$  of multiscale kernels).

$$\begin{aligned} & \underset{A}{\operatorname{argmin}} \operatorname{tr} \left( A^T Q^{(I)}(\beta) A \right) \quad \text{s.t.} \operatorname{tr} \left( A^T Q^{(P)}(\beta) A \right) = \gamma, \\ & Q^{(I)}(\beta) = \sum_{i, j=1}^N (\Omega_{x_i} - \Omega_{x_j}) \beta \beta^T (\Omega_{x_i} - \Omega_{x_j})^T W_{ij}^{(I)}, \quad (6) \\ & Q^{(P)}(\beta) = \sum_{i, j=1}^N (\Omega_{x_i} - \Omega_{x_j}) \beta \beta^T (\Omega_{x_i} - \Omega_{x_j})^T W_{ij}^{(P)}, \end{aligned}$$

which can be approximately changed into the ratio-trace form and therefore it can be calculated as Generalised Eigenvalue Problem (GEP). In order to avoid theoretical minimal

zero values in calculating the GEP, a diagonal matrix with small weights can be added on the Laplacian matrix of  $W^{(I)}$ .

$$\begin{aligned} & \underset{\beta}{\operatorname{argmin}} \quad \beta^T Q^{(I)}(A)\beta \\ & \text{s.t.} \quad \beta^T Q^{(P)}(A)\beta = \gamma, \quad \beta_m \geq 0, \quad m = 1, 2, \dots, M, \\ & Q^{(I)}(A) = \sum_{i,j=1}^N (\Omega_{x_i} - \Omega_{x_j})^T A A^T (\Omega_{x_i} - \Omega_{x_j}) W_{ij}^{(I)}, \quad (7) \\ & Q^{(P)}(A) = \sum_{i,j=1}^N (\Omega_{x_i} - \Omega_{x_j})^T A A^T (\Omega_{x_i} - \Omega_{x_j}) W_{ij}^{(P)}, \end{aligned}$$

which is calculated according to Semi-Definite Programming (SDP) relaxation. The total computational cost is similar as in [12, 21].

### 3. EXPERIMENTS

#### 3.1 Corpora and Features

Two emotional corpora, namely ‘Speech Under Simulated and Actual Stress’ (SUSAS) [11] and the ‘Geneva Multimodal Emotion Portrayals’ (GEMEP) [3] are used in the experiments, in order to show the performance in various conditions.

**SUSAS** is mainly utilised to analyse stress levels containing the four emotional categories of *high stress (hist)*, *medium stress (meds)*, *neutral (neut)*, and *screaming (scre)* in (US) English. On arousal *neut* is low, while the other categories are high. The categories of *neut* and *scre* are both positive on valence – the other two negative. 3 593 utterance samples by seven speaker (three female) in English are contained in the corpus. The numbers of samples in the four emotional categories are 1 202, 1 276, 701, and 414, respectively. In the experiments, the training set of the corpus contains four speakers (two female) with 2 027 samples, while the testing set includes all the remaining speakers with 1 566 samples. **GEMEP** is a French spoken language corpus with 18 detailed emotional categories and 1 260 utterance samples. We chose the 12 categories as used in INTERSPEECH ComParE 2013, namely (*amusement, pride, joy, relief, interest, pleasure, hot anger, panic fear, despair, irritation, anxiety, and sadness*) [9, 18] in the experiments. In total, these are 1 080 samples by ten speakers, which leads to 90 samples per emotion. We divide the corpus into two folds, including five speakers in each fold (three female for the first fold / two female for the second fold). One fold is for training and the other is for testing, and vice versa. This leads to a 2-fold Cross-Validation (CV).

The key information of the corpora used in the experiments is presented in Table 1.

**Table 1: Description of the emotional speech corpora GEMEP and SUSAS used in the experiments, where ‘Tr.’ means training set and ‘Te.’ means testing set.**

Corpus	# Classes	# Speakers	# Samples
SUSAS	Tr.	4 (2 female)	2 027
	Te.	3 (1 female)	1 566
GEMEP	Tr./Te.	5 (2 female)	540
	Te./Tr.	5 (3 female)	540

In the experiments, the open-source paralinguistic tool *openSMILE* [8, 10] is chosen as the feature extractor. We

employ the configuration of the INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE) [18], which also has been used in INTERSPEECH 2014 to 2016 Computational Paralinguistics Challenges [15, 16]. Hence, the feature set with the original dimensionality of 6 373 is obtained for each utterance. This set mainly includes the features of low-level descriptors covering different acoustic characteristics, with various statistical functionals.

#### 3.2 Experimental Setup

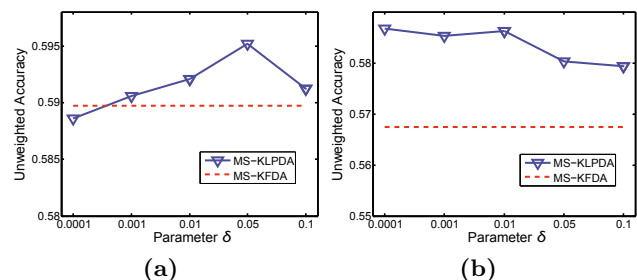
As in [21], the number of scales is set as  $M = 10$ , with the Gaussian scaling parameters  $\sigma_m$  ( $m = 1, 2, \dots, M$ ) (in Eq. (3)) chosen as  $0.001n, 0.005n, 0.01n, 0.03n, 0.05n, 0.1n, 0.3n, 0.5n, 0.75n$ , and  $n$ , respectively. The dimensions  $d$  of the dimensionality-reduced feature space are selected no larger than 7 for SUSAS, and 15 for GEMEP. The neighbouring penalty parameter  $k_0$  is chosen as 30.

At the stage of classification, a  $k$  Nearest-Neighbour (kNN) classifier, which is shown in [21], is adopted to show the basic performance of the proposed methods. Further, we apply generalised Ridge Regression (RR), which was proposed for face recognition in [1], since the kNN classifier requires memory to store all the training samples.

#### 3.3 Results

##### 3.3.1 Experiments on SUSAS

First, we consider the unbalanced SUSAS corpus. Figure 1 shows the changes of Unweighted Accuracy (UA) for MS-KLPDA with the parameters  $\delta$  equal to 0.0001, 0.001, 0.01, 0.05, and 0.1, respectively, compared with MS-KFDA. For the kNN classifier, the best performance can be obtained with  $\delta = 0.05$ , while for RR, it seems more likely to achieve better performance with lower  $\delta$  values. Overall, the proposed MS-KLPDA is capable of achieving better performance given suitable selections of  $\delta$ s.



**Figure 1: Unweighted accuracy of MS-KFDA and MS-KLPDA with various parameters  $\delta$  on SUSAS. (a) with kNN, (b) with RR.**

The UA and Weighted Accuracy (WA) comparison between the proposed MS-KLPDA and conventional methods, including PCA, LDA (Linear Discriminant Analysis) / FDA, LDP, LDE, kNN, RR, and Support Vector Machines (SVM), as well as MS-KFDA, are also shown in Table 2 according to the experiments on SUSAS. One learns from the table that MS-KLPDA with kNN classifier gives best performance among all the evaluated methods.

To better compare the performance of the proposed MS-KLPDA with the ‘conventional’ methods, the results of a one-tail  $z$ -test [24] show that MS-KLPDA is significantly better than SVM at the significance level of 0.05.

**Table 2: Recognition rates by UA and WA (%), of conventional methods and the proposed MS-KLPDA on SUSAS and GEMEP.**

Methods \ Accuracy	SUSAS		GEMEP	
	UA	WA	UA	WA
PCA	54.4	46.6	26.5	26.4
LDA / FDA	55.1	46.3	33.8	33.5
LDP [4]	55.3	46.2	33.1	32.8
LDE [5]	40.3	37.5	35.5	35.4
kNN (Baseline)	53.0	46.5	25.2	25.2
RR (Baseline) [1]	51.1	45.2	33.3	33.2
SVM	56.2	46.6	38.4	38.2
MS-KFDA (with kNN) [21]	59.0	51.5	38.8	38.2
MS-KFDA (with RR)	56.7	54.3	31.0	29.4
MS-KLPDA (with kNN)	<b>59.5</b>	<b>55.4</b>	39.1	38.5
MS-KLPDA (with RR)	58.7	52.7	<b>40.2</b>	<b>40.1</b>

Table 3 shows the comparison of the recalls (%) of each category by the methods of MS-KFDA and the proposed MS-KLPDA. In designing embedding graphs (5), the numbers of neighbouring marginal samples are set as  $k_0 = 5, 30,$  and  $60$  to investigate the tendency of MS-KLPDA. It is concluded from Table 3 that for the states of *high stress* and *medium stress* (with a large sum of samples), the performances keep rising as  $k_0$ s increase. However, for the state of *neutral* (with fewer samples), the recall drops, while it remains stable for *screaming*.

**Table 3: Recalls (%) of MS-KFDA and MS-KLPDA (with the numbers of neighbours  $k_0 = 5, 30,$  and  $60$ ) on SUSAS.**

Methods \ Species	<i>hist</i>	<i>meds</i>	<i>neut</i>	<i>scre</i>
MS-KFDA ( $k_0 = 0$ )	43.2	66.2	39.5	98.6
MS-KLPDA ( $k_0 = 5$ )	44.8	68.8	38.3	98.6
MS-KLPDA ( $k_0 = 30$ )	47.1	75.4	31.0	97.8
MS-KLPDA ( $k_0 = 60$ )	48.8	85.6	21.9	97.1

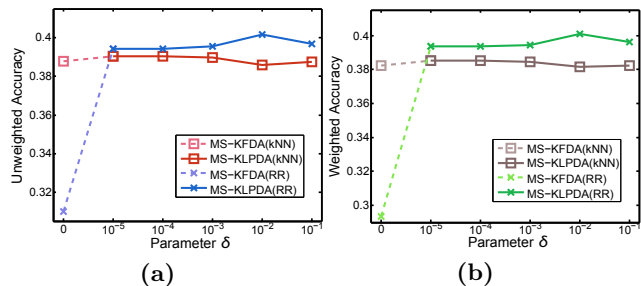
### 3.3.2 Experiments on GEMEP

Next, for further comparison, we show the experiments on GEMEP with 12 emotional classes, which demonstrate the case of more fine-grained emotional modelling.

Similar as in the experiments on SUSAS, the ‘conventional’ methods of subspace learning (PCA, LDA/ FDA, LDP, LDE), kNN, RR, and SVM are chosen for comparison in the right part of Table 2. The proposed MS-KLPDA with RR is able to achieve the best performance among these methods, while the MS-KLPDA with kNN is also competitive.

The recognition rates of the proposed MS-KLPDA on GEMEP, represented by UA and WA, are presented in Figure 2, with the parameters  $\delta$  equal to  $0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2},$  and  $10^{-1}$ , respectively, using kNN and RR. Note that the MS-KLPDA with  $\delta = 0$  represents MS-KFDA marked by dotted lines in Figure 2. It can be drawn from Figure 2 that, the proposed MS-KLPDA methods generally outperform MS-KFDA. The best performance is observed for the parameter  $\delta = 10^{-2}$  using RR.

Let us now focus on the comparison between the MS-KLPDA and MS-KFDA in the right part of Table 2, which shows that, MS-KLPDA with RR provides a relatively good increase on the basis of MS-KFDA compared with the methods using kNN on GEMEP.



**Figure 2: Recognition rates (UA and WA) of MS-KFDA and MS-KLPDA with various parameters  $\delta$ s using kNN and RR on GEMEP. (a) UA, (b) WA.**

**Table 4: Recognition rates by UA and WA (%), of the top-three-performance KLPDA and MS-KLPDA variants on GEMEP.**

Classifier Methods	with kNN		with RR	
	UA	WA	UA	WA
KLPDA ( $\sigma^{(1)}$ )	37.8	37.4	<b>40.9</b>	<b>40.7</b>
KLPDA ( $\sigma^{(2)}$ )	37.5	37.0	40.1	40.0
KLPDA ( $\sigma^{(3)}$ )	36.9	36.6	39.1	39.0
MS-KLPDA	<b>39.1</b>	<b>38.5</b>	40.2	40.1

To show the performance of learning multiscale kernels, we present the comparison between MS-KLPDA and KLPDA (with single Gaussian kernels) in Table 4, where the UA and WA of the top-three-performance KLPDA variants are listed, denoted with the scaling parameters  $\sigma^{(1)}, \sigma^{(2)},$  and  $\sigma^{(3)}$ , respectively. According to the table, for the kNN classifier, MS-KLPDA provides ‘dominant’ performance, while for RR, the multiscale approach is still competitive though it fails to outperform the best single-kernel way.

## 4. CONCLUSIONS

A novel algorithm of multiscale-kernel locally penalised discriminant analysis was proposed in this paper. The key idea of this algorithm is adopting locally penalised embedding graphs by learning multiscale kernels in the framework of graph embedding. The extensive experimental results on emotional speech corpora show a performance improvement of the proposed MS-KLPDA compared with MS-KFDA and conventional alternative algorithms.

In our future work, more suitable embedding graphs are expected to be designed in order to better describe the structures of training samples. In addition, a wider scope of scaling parameters for kernels can be utilised in optimisation to obtain a more optimal object over the iterations.

## 5. ACKNOWLEDGMENTS

This work was supported by China Scholarship Council (CSC), the EC’s 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), the EU’s Horizon 2020 Programme through the Innovative Action No. 644632 (MixedEmotions), the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant agreement #16SV7213 (EmotAsS), the Natural Science Foundation of China under Grants No. 61231002 and No. 61273266, and the Doctoral Fund of the Ministry of Education of China under Grant No. 20110092130004.

## 6. REFERENCES

- [1] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7, Minneapolis, MN, 2007. IEEE.
- [2] J. P. Arias, C. Busso, and N. B. Yoma. Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech & Language*, 28(1):278–294, 2014.
- [3] T. Bänziger, M. Mortillaro, and K. R. Scherer. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5):1161, 2012.
- [4] H. Cai, K. Mikolajczyk, and J. Matas. Learning linear discriminant projections for dimensionality reduction of image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):338–352, 2011.
- [5] H.-T. Chen, H.-W. Chang, and T.-L. Liu. Local discriminant embedding and its variants. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 846–853, San Diego, CA, 2005. IEEE.
- [6] J. Deng, Z. Zhang, E. Marchi, and B. Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 511–516, Geneva, Switzerland, 2013. IEEE.
- [7] H. P. Espinosa, C. A. R. García, and L. V. Pineda. Features selection for primitives estimation on emotional speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5138–5141, Dallas, TX, 2010. IEEE.
- [8] F. Eyben and B. Schuller. openSMILE:) The Munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records*, 6(4):4–13, 2015.
- [9] F. Eyben, F. Weninger, and B. Schuller. Affect recognition in real-life acoustic conditions—a new perspective on feature selection. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2044–2048, Lyon, France, 2013. ISCA.
- [10] F. Eyben, M. Wöllmer, and B. Schuller. OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proc. International conference on Multimedia*, pages 1459–1462, Florence, Italy, 2010. ACM.
- [11] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom. Getting started with SUSAS: A speech under simulated and actual stress database. In *Proc. European Conference on Speech Communication and Technology*, volume 97, pages 1743–1746, Rhodes, Greece, 1997. ISCA.
- [12] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Multiple kernel learning for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1147–1160, 2011.
- [13] S. Ntalampiras and N. Fakotakis. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *IEEE Transactions on Affective Computing*, 3(1):116–125, 2012.
- [14] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
- [15] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang. The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 427–431, Singapore, 2014. ISCA.
- [16] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger. The INTERSPEECH 2015 computational paralinguistics challenge: Degree of nativeness, parkinson’s & eating condition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 478–482, Dresden, Germany, 2015. ISCA.
- [17] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet, et al. The INTERSPEECH 2012 speaker trait challenge. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, Oregon, 2012. ISCA. no pagination.
- [18] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013. ISCA. no pagination.
- [19] Z. Wang and X. Sun. Multiple kernel local Fisher discriminant analysis for face recognition. *Signal Processing*, 93(6):1496–1509, 2013.
- [20] C.-H. Wu and W.-B. Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21, 2011.
- [21] X. Xu, J. Deng, W. Zheng, L. Zhao, and B. Schuller. Dimensionality reduction for speech emotion features by multiscale kernels. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1532–1536, Dresden, Germany, 2015. ISCA.
- [22] S. Yan, D. Xu, B. Zhang, and H. Zhang. Graph embedding: A general framework for dimensionality reduction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 830–837, San Diego, CA, 2005. IEEE.
- [23] 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 Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.
- [24] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):115–126, 2015.