Dimensionality reduction for speech emotion features by multiscale kernels

X. Xu, J. Deng, W. Zheng, L. Zhao, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Xu, X., J. Deng, W. Zheng, L. Zhao, and Björn Schuller. 2015. "Dimensionality reduction for speech emotion features by multiscale kernels." In *INTERSPEECH 2015 - 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 1532–36. ISCA Archive.
https://doi.org/10.21437/Interspeech.2015-335.

Dimensionality Reduction for Speech Emotion Features by Multiscale Kernels

Xinzhou Xu[1,2], *Jun Deng*[1], *Wenming Zheng*[2], *Li Zhao*[2], *Björn Schuller*[3,4]

[1]MISP Group, MMK, Technische Universität München, Germany
[2]School of Information Science and Engineering, Southeast University, P.R. China
[3]Department of Computing, Imperial College London, U.K.
[4]Chair of Complex and Intelligent Systems, University of Passau, Germany
{xinzhou.xu, jun.deng}@tum.de, {wenming_zheng, zhaoli}@seu.edu.cn, schuller@ieee.org

10.21437/Interspeech.2015-335

Abstract

To achieve efficient and compact low-dimensional features for speech emotion recognition, this paper proposes a novel feature reduction method using multiscale kernels in the framework of graph embedding. With Fisher discriminant embedding graph, multiscale Gaussian kernels are used in constructing optimal linear combination of Gram matrices for multiple kernel learning. To evaluate the proposed method, comprehensive experiments, using different public feature sets from the open-source toolbox openSMILE on various corpora, show that the proposed method achieves better performance compared with conventional linear dimensionality reduction methods and single-kernel methods.

Index Terms: speech emotion recognition, dimensionality reduction, multiscale kernels

1. Introduction

In recent years, Speech Emotion Recognition (SER) has turned to be an important subsection in speech processing, affective computing, and pattern recognition [1, 2]. A widely known application of SER is as an increasingly 'necessary' module in Human-Machine Interaction. A typical SER system can be divided into several significant modules, such as speech feature extraction [3–6], feature selection [1, 2], dimensionality reduction [7–9], and classifier designing [3, 10–12].

In this work, we put emphasis on dimensionality reduction. A low-dimensional feature vector describing a speech sample bears a number of advantages such as lower complexity for the learning algorithm and lower bandwidth in distributed processing architectures [13]. Classical dimensionality reduction algorithms, e. g., FDA (Fisher Discriminant Analysis) [14], LLE (Locally Linear Embedding) [15], LPP (Locality Preserving Projections) [16], LDE (Local Discriminant Embedding) [17], GbFA (Graph-based Fisher Analysis) [18], have been proved to be effective methods in image processing and recognition. However, 'original' speech emotion features often include information on other speaker states and traits and the verbal content – partially even more strongly related to this 'side' information. Therefore, most of the existing dimensionality reduction methods are not suitable for the task of SER, due to the strong dependence on labelling information for the features.

This work was supported by Chinese Scholarship Council (CSC), the European Union's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC Starting Grant iHEARu), and No. 645378 (RIA ARIA-VALUSPA), 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.

In order to solve this, we present a discriminant analysis method using multiscale kernels, in the framework of graph embedding [19] with MKL (Multiple Kernel Learning) [20, 21]. Accordingly, we suppose that each sample can be represented by linearly combining several kernels with different scales, ending up its more concise and informative form. Based on kernel theory in graph embedding, different kernels are denoted by different Gram matrices. Further, these Gram matrices can be easily described by various scales when various scaling parameters are tried. To this end, we conduct dimensionality reduction for raw features to construct an optimal speech emotion feature representation, with the help of MKL using a graph embedding optimization form.

2. Theoretical basis

2.1. Notation

Suppose $X = [x_1, x_2, \ldots, x_N] \in \Re^{n\times N}$ and $Y = [y_1, y_2, \ldots, y_N] \in \Re^{d\times N}$ are the sets of N training samples with each of their column standing for one training sample, in the original feature space with the dimensionality of n and in the dimensionality-reduced feature space with the dimensionality of d, respectively. Every column of $\phi(X) = [\phi(x_1), \phi(x_2), \ldots, \phi(x_N)]$ is the RKHS (Reproducing Kernel Hilbert Space) of the corresponding column in X. The Gram matrix $K = \phi^T(X)\phi(X)$. We also assume that any sample (including any training and testing sample) in the original and the reduced dimensionality is represented by column vectors x and y, respectively. x is with its high-dimensional forms $\phi(x)$. For sample x, its kernelized coordinate is $K_x = \phi^T(X)\phi(x)$.

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

2.2. Graph embedding and discriminant analysis

The graph embedding framework [19] was proposed to combine subspace and manifold learning. By using graph structures for data, it aims to search for optimal embedding of graphs, together with data mapping types and optimization forms, to discover the internal essence of a data set. The optimization form of the graph embedding framework is shown in Eq. (1) and Eq. (2), with the constraints of penalty and scaling respectively:

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

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

where $W^{(I)}$ and $W^{(P)}$ are the adjacency matrices of the intrinsic graph and the penalty graph, respectively; D is a diagonal matrix to control the weights of the samples, and t is a positive constant value.

Considering the matrix form and the transformation of constraints, we can change the optimization of Eq. (1) to

$$\underset{z}{\text{argmin}} \quad \frac{zL^{(I)}z^T}{zL^{(P)}z^T}, \quad (3)$$

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

For FDA, with $W^{(I)} = W^{(I)FDA} = S^T(SS^T)^{-1}S$ and $W^{(P)} = W^{(P)FDA} = \frac{1}{N}ee^T$, the optimization form in the graph embedding framework can be achieved accordingly. Then, we can draw linear, kernelized, and tensorized forms of FDA in this framework according to Eq. (3).

2.3. Multiple kernel learning FDA in graph embedding

Based on the graph embedding framework of FDA, with the optimization for data mapping $\alpha \in \Re^{N\times 1}$, we can write the kernelized form of FDA, namely KFDA, as

$$\begin{aligned} &\underset{\alpha}{\text{argmin}} \quad \sum_{i=1}^{N}\sum_{j=1}^{N} \| \alpha^T K_{x_i} - \alpha^T K_{x_j} \|^2 W_{ij}^{(I)} \\ &\text{s.t.} \quad \sum_{i=1}^{N}\sum_{j=1}^{N} \| \alpha^T K_{x_i} - \alpha^T K_{x_j} \|^2 W_{ij}^{(P)} = t. \end{aligned} \quad (4)$$

When multiple kernels [21] are required, K_x is written as the linear combination of different kernels, namely $K_x = \sum_{m=1}^{M} \beta^m \phi_m^T(X)\phi_m(x) = \Omega_x\beta$, where the multiple kernel coordinate matrix $\Omega_x = [\phi_1^T(X)\phi_1(x), \phi_2^T(X)\phi_2(x), \ldots, \phi_M^T(X)\phi_M(x)] \in \Re^{N\times M}$ and $\beta \in \Re^{M\times 1}$ is the column vector with corresponding elements β_m for kernel m. The number of kernels is M. Each column of Ω_x is the corresponding coordinate for the sample x.

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 optimization problem. α_i is the ith mapping vector with $i = 1, 2, \ldots, d$. For simplicity, the variables of $W^{(I)FDA}$ and $W^{(P)FDA}$ are represented by $W^{(I)}$ and $W^{(P)}$, respectively. Therefore, the optimization of MKL-FDA is given as

$$\begin{aligned} &\underset{\beta}{\text{argmin}} \quad \sum_{i,j=1}^{N} \| A^T\Omega_{x_i}\beta - A^T\Omega_{x_j}\beta \|^2 W_{ij}^{(I)} \\ &\text{s.t.} \begin{cases} \sum_{i,j=1}^{N} \| A^T\Omega_{x_i}\beta - A^T\Omega_{x_j}\beta \|^2 W_{ij}^{(P)} = t, \\ \beta_m \geq 0, \quad m = 1, 2, \ldots, M. \end{cases} \end{aligned} \quad (5)$$

Then, the bilateral form of MKL-FDA is given by Eq. (6) for solving kernel mappings A, while it is given by Eq. (7) for solving linear weights β of multiple kernels.

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

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

Eq. (6) is approximately changed into the ratio-trace form and can therefore be solved as Generalized Eigenvalue Problem (GEP). Eq. (7) is solvable by Semi-Definite Programming (SDP) relaxation [22]. The solution of Eq. (5) is consequently represented by bilateral steps of Eq. (6) and Eq. (7).

3. Proposed methodology

3.1. Learn multiscale kernels for speech emotion features

To our best knowledge, little research focuses on multiscale representation in SER. In [12], MKL with the optimization form of Support Vector Machines (SVM) is used in SER. However, it requires large computational cost and is only valid in decision making. We develop the method of multiscale kernel learning to show its effectiveness on extracting speech emotion features and then improving the performance of SER.

The research of MKL provides the possibility of solving multiscale analysis of speech emotion factors. For Gaussian kernels, it is easy to draw the multiscale case by regulating scaling parameters. The kernel transforming between samples x_i and x is shown in Eq. (8), with the parameters $\sigma_m > 0$, where $m = 1, 2, \ldots, M$ and $i = 1, 2, \ldots, N$:

$$(\Omega_{x_i})_m = \phi_m^T(x_i)\phi_m(x) = e^{-\frac{(x_i - x)^2}{\sigma_m^2}}. \quad (8)$$

Kernel methods are originally represented as high-dimensional space by adopting inner product forms in RKHS. However, it can be also assumed that kernel methods bring a dimension-limited feature transformation in graph embedding. This transformation constructs a new feature space for each sample by kernel functions and training samples. Thus, the relationship between a given sample and each training sample leads to the new features. Then, the scales of kernels are mainly determined by the parameters of kernels.

As is shown in Figure 1, for sample x, the original speech emotion features x are transformed into new features $\Omega_x\beta$ by linearly combining multiscale kernels. Then, for the new features of x, the dimensionality-reducted sample can be achieved by using $A^T\Omega_x\beta$ in bilateral ways.

As outlined above, speech emotion features inevitably include much interference resulting from the factors of speakers, text, languages, etc., in spite of state-of-the-art feature acquisition ways. Therefore, in the use of feature reduction methods for SER, supervised information would be helpful to eliminate such interference. Hence, these methods, which are guarded

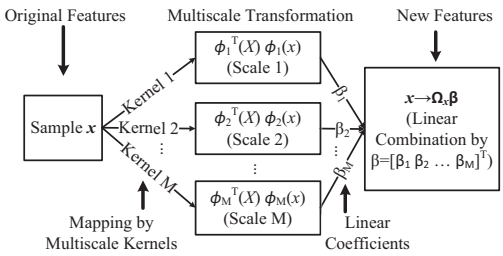


Figure 1: *Schematic diagram of learning multiscale kernels. The original features x are transformed into new features $\Omega_x\beta$ by linearly combining multiscale kernels.*

by supervised information, would seek for new compact feature representations, which in turn fit with the task the SER system is required to solve. Therefore, SER systems would benefit from the novel feature reduction method in combination of the embedding graphs of FDA and multiple kernel learning. In addition, few parameters need to be regulated in FDA.

3.2. Comparisons with related research

Lin et al. [21] proposed dimensionality reduction in the framework of graph embedding using MKL. This is the motivation of our research in SER. We choose the relatively stable FDA structure to better describe speech emotions. Then, multiscale kernels are represented by Gaussian forms with various scaling parameters. In [20] multiscale Gaussian kernels are adopted in KFDA. However, this is attempted only for a binary classification. [23] presents the graph-designing extension of [21], with local-based embedding graphs. In contrast, designing these embedding graphs without processing is not a desirable choice in SER. Compared with [12], our work focuses on extracting appropriate features for SER with 'simple' classifiers.

4. Experimental results

4.1. Corpora and speech emotion features

Three corpora, the Berlin Emotional Speech Database (EMO-DB) [24], the audio parts of eNTERFACE'05 [25] and the Geneva Multimodal Emotion Portrayals (GEMEP) [26] databases, are adopted for evaluation.

EMO-DB includes seven emotions, namely *neutral*, *anger*, *fear*, *joy*, *sadness*, *disgust* and *boredom*. Ten speakers, chosen from 40 ones, were required to say ten German predefined sentences. We use the 494 instances commonly used based on naturalness in our experiments. For the above seven emotions, there are 78, 128, 54, 65, 53, 37, and 79 samples, respectively.

eNTERFACE'05 is a multimodal emotion corpus with both visual and audio parts. It contains examples of the emotions *happiness*, *sadness*, *surprise*, *anger*, *disgust*, and *fear*. The corpus totally includes 42 subjects from 14 nations. The text to be prompted after some story-based emotion elicitation includes five English short sentences. We adopt the commonly chosen 40 speakers in our experiments, with 200 samples per emotion.

GEMEP includes 18 emotions in total, with 10 speakers, including 1 260 multimodal emotion samples. In our experiments, the 12 emotions (*amusement*, *pride*, *joy*, *relief*, *interest*, *pleasure*, *hot anger*, *panic fear*, *despair*, *irritation*, *anxiety*, *sadness*) [27] are adopted, including 1 080 samples, with approximately 90 samples per emotion. It was used as benchmark set in the INTERSPEECH 2013 Computational Paralinguistics Challenge for SER [28] and we use the selection defined in the challenge.

Our open-source openSMILE toolbox [29] is used for obtaining 'raw' speech emotion features in our experiments. We separately consider the feature sets of the INTERSPEECH 2009 Emotion Challenge (IS09) [4], the 'emobase' configuration [5], and the INTERSPEECH 2010 Paralinguistics Challenge (IS10) [6], with the dimensionality of 384, 988 and 1 582 respectively. These features are obtained by LLDs (Low-Level Descriptors) with subsequential application of statistical functionals. For more details on these features sets, see [4–6]. The choice of the sets is motivated by their gradually increasing, yet reasonable sizes.

4.2. Parameters and system design in experiments

Normalization is conducted on each feature according to the respective training data. Afterwards, MKL with multiscale Gaussian kernels and Fisher discriminant graphs is used in dimensionality reduction. The bilateral way is adopted in solving the optimization of MKL. In the stage of decision, we apply simple nearest-neighbor classifiers to evaluate our method in a transparent way with little influence by the classifier. This is also in line with [30], where deep learning was used for reduction of dimensionality.

We ensure speaker-independent evaluation as follows: On EMO-DB, Leave One Speaker Out (LOSO) cross-validation is used, while on eNTERFACE'05 we use two-fold cross-validation with each fold including 20 speakers. On GEMEP, six out of ten speakers are adopted for training, while the rest are used in testing. In dimensionality reduction we use ten multiscale Gaussian kernels, where the kernel parameters σ_m^2 in Eq. (8) are shown in Table 1. The number of iterations is set as five.

Table 1: *The choices of Gaussian kernel parameters σ_m^2.*

m	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
σ_m^2	$0.001n$	$0.005n$	$0.01n$	$0.03n$	$0.05n$
m	$m=6$	$m=7$	$m=8$	$m=9$	$m=10$
σ_m^2	$0.1n$	$0.3n$	$0.5n$	$0.75n$	n

4.3. Experiments and analysis

We divide the experiments into two sections according to the number of emotion categories. In detail, EMO-DB and eNTERFACE'05 stand for basic emotions, while GEMEP contains more meticulous emotional species.

4.3.1. Experiments on EMO-DB and eNTERFACE'05

The best recognition rates (%) of emotions for different algorithms within low dimensionality (no larger than 70 dimensions) are shown in Table 2, with the rates represented in the form of 'WA (Weighted Accuracy) / UA (Unweighted Accuracy)'. It is noticeable that in eNTERFACE'05 WA and UA are the same, due to the same number of samples for each class in our experiments. 'MS-KFDA' stands for the proposed method. The best three recognition rates for KFDA are represented with the parameters $\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)} = \sigma_i$ $(i = 1, 2, \ldots, M$, with $M = 10)$. In Table 2, some of the linear feature reduction algorithms, including LDA, LPP, LDE, and GbFA, are considered for comparison. In order to raise their low-dimensional performance, SVD (Sigular Value Decomposition) is conducted when solving GEP [14, 16].

One can conclude from Table 2 that, the multiscale-kernel based FDA outperforms KFDA and various linear-mapping-based dimensionality reduction methods. Beyond that, MS-

Table 2: *The best recognition results (%) and their corresponding dimensions of different algorithms on EMO-DB and eNTERFACE'05 considering the feature sets IS09 and emobase ('WA / UA').*

Corpora	EMO-DB		eNTERFACE'05	
Feature Sets	IS09	emobase	IS09	emobase
Baseline	55.2 / 49.8	63.8 / 61.5	42.8	27.5
PCA	56.5 / 52.9	62.5 / 59.5	37.5	32.1
LDA / FDA	74.6 / 72.1	76.0 / 73.4	46.7	38.4
LPP	55.6 / 49.1	63.0 / 59.5	35.7	33.8
LDE	70.1 / 65.6	73.3 / 66.2	49.3	39.8
GbFA	71.6 / 68.1	73.2 / 73.2	48.9	40.8
$\sigma^{(1)}$-KFDA	76.3 / 70.1	81.7 / 75.6	56.8	45.1
$\sigma^{(2)}$-KFDA	75.6 / 69.4	81.1 / 74.4	55.8	44.2
$\sigma^{(3)}$-KFDA	74.5 / 68.7	79.0 / 73.0	55.1	42.0
MS-KFDA	**77.7 / 71.6**	**81.8 / 75.7**	**57.1**	**46.8**

KFDA shows better performance compared with these single-kernel ways in most conditions. The kernel parameters corresponding to the top-3 recognition rates (both WA and UA) are σ_4, σ_5, σ_6 for EMO-DB with IS09, and are σ_6, σ_5, σ_7 for EMO-DB with emobase, while for eNTERFACE'05, the parameters are σ_4, σ_6, σ_5, and σ_6, σ_7, σ_8 with IS09 and emobase respectively. Considering feature sets, the emobase set performs better than IS09 on EMO-DB. However, the rates of emobase fail to outperform those of IS09 on eNTERFACE'05.

We further show the recall (%) of each emotion in Table 3, when using MS-KFDA. The recognition rates of the feature sets of IS09 and emobase are represented as 'IS09 / emobase'.

Table 3: *The recall (%) of the best UA on EMO-DB and eNTERFACE'05 using multiscale kernels, with the form of 'IS09 / emobase'.*

Emotions\Corpora	EMO-DB	eNTERFACE'05
anger	89.8 / 88.3	59.0 / 52.0
boredom	84.8 / 91.1	–
fear	51.9 / 68.5	51.5 / 49.5
disgust	62.2 / 67.6	70.0 / 59.5
joy or *happiness*	60.0 / 61.5	75.0 / 60.5
neutral	84.6 / 91.0	–
sadness	79.3 / 79.3	42.0 / 44.5
surprise	–	45.0 / 15.0

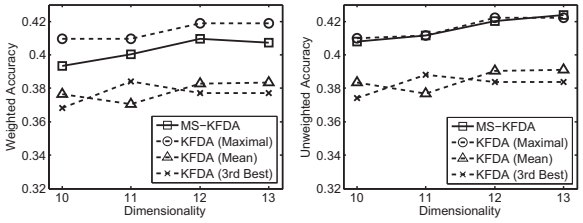
4.3.2. Experiments on GEMEP

Following the description and Table 2 in the former section, the best recognition rates with the dimensionality no larger than 70 are represented in Table 4 for GEMEP.

Table 4: *The best recognition results (%) and their corresponding dimensions of different algorithms on GEMEP using the feature sets IS09, emobase, and IS10 ('WA / UA').*

Feature Sets	IS09	emobase	IS10
Baseline	23.6 / 23.6	30.6 / 30.3	25.9 / 26.0
PCA	27.3 / 27.5	30.3 / 30.4	29.4 / 30.1
LDA / FDA	28.9 / 29.4	32.4 / 33.0	32.4 / 33.2
LPP	24.3 / 24.5	30.6 / 30.3	28.0 / 28.6
LDE	29.6 / 29.6	35.0 / 34.6	35.2 / 36.1
GbFA	29.2 / 30.0	30.6 / 32.5	33.6 / 34.7
$\sigma^{(1)}$-KFDA	35.2 / 36.2	**42.1 / 42.5**	38.2 / 39.5
$\sigma^{(2)}$-KFDA	33.6 / 34.6	41.9 / 42.4	37.7 / 38.9
$\sigma^{(3)}$-KFDA	33.3 / 33.9	38.4 / 39.3	37.3 / 38.3
MS-KFDA	**36.3 / 38.1**	41.0 / 42.4	**39.4 / 40.5**

In Table 4 for GEMEP, the kernel parameters corresponding to the top-3 recognition rates (both WA and UA) are σ_5, σ_4, σ_6 with the IS09 set, while the parameters are σ_4, σ_5, σ_6 when using the feature set IS10. Yet for emobase, the top-3 rates are σ_5, σ_4, σ_6 and σ_4, σ_5, σ_3, respectively for WA and UA. The recognition rates of MS-KFDA outperform the given methods when using the feature sets IS09 and IS10, but the rates fall behind for the emobase set. In order to show the performance of MS-KFDA for the emobase set, in Figure 2 we draw maximal, mean, and 3rd best values of the top-5 recognition rates of KFDA, when separately using the 10 kernel parameters for the dimensionality reaching from 10 to 13, and conduct a comparison with MS-KFDA. According to Figure 2a for WA, MS-KFDA achieves better performance than most KFDA algorithms based on single kernel, though it cannot reach the best result. Figure 2b shows the performance of MS-KFDA can approach the best of KFDA for UA.



(a) (b)

Figure 2: *Maximal, mean and 3rd best values of 5-top recognition rates of KFDA using 10 kernel parameters, compared with the recognition rates of MS-KFDA, dimensionality 10 to 13 with emobase set. (a) WA, (b) UA.*

In Table 5, the baseline rates, written as 'WA / UA', when using the classifiers Naive Bayes (NB), SVM, and Extreme Learning Machine (ELM) [31] are provided for comparison. Here we use linear and the 'one-against-one' strategy for SVM, and the violation level of Sequential Minimal Optimization (SMO) iterations C_I is set as 0.001. The ELM is set as the linear form, which is similar as ridge regression. It shows that the proposed method can achieve better performance in this condition.

Table 5: *The recognition results (%) ('WA / UA') on GEMEP using the feature sets of IS09, emobase, and IS10, with different algorithms.*

Feature Sets	IS09	emobase	IS10
NB	32.6 / 32.0	32.9 / 32.6	32.9 / 33.5
SVM ($C_I = 0.001$)	31.7 / 32.1	39.1 / 39.4	38.9 / 39.7
ELM (Linear)	32.9 / 33.7	33.3 / 33.7	30.8 / 31.8
MS-KFDA	**36.3 / 38.1**	**41.0 / 42.4**	**39.4 / 40.5**

The experimental results show a satisfying performance for the multiscale-kernel method in dimensionality reduction of speech emotion features. Compared with other dimensionality reduction methods in this paper [14–18], its computational cost depends on iteration times and the solution procedure of SDP.

5. Conclusions

We presented the multiscale-kernel method using MKL based FDA in dimensionality reduction of speech emotion features in this paper. Validated by experiments, the proposed method achieves relatively desirable performance by using multiple speech emotion feature sets on different speech emotion corpora, compared with other dimensionality reduction methods.

The success of this method inspires us to extend this multiple kernel learning framework of a multiscale analysis to a multi-dimensional principal analysis. In addition, Gram matrices with a low-dimensional space can be considered to reduce the computational deviation in the iterations.

6. References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in *Emotion-Oriented Systems*. Springer, 2011, pp. 71–99.

[3] E. Vayrynen, J. Kortelainen, and T. Seppanen, "Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 47–56, 2013.

[4] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2009, pp. 312–315.

[5] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR–introducing the Munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops*. Amsterdam, The Netherlands: IEEE, 2009, pp. 576–581.

[6] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, pp. 2794–2797.

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

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

[9] X. Xu, C. Huang, C. Wu, Q. Wang, and L. Zhao, "Graph learning based speaker independent speech emotion recognition," *Advances in Electrical and Computer Engineering*, vol. 14, no. 2, pp. 17–22, 2014.

[10] A. Sciarrone, A. Delfino, M. Marchese, F. Lavagetto, and I. Bisio, "Gender-driven emotion recognition through speech signals for ambient intelligence applications," *IEEE Transactions on Emerging Topics in Computing*, vol. 1, no. 2, pp. 244–257, 2013.

[11] P. Dumouchel, "Anchor models for emotion recognition from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 280–290, 2013.

[12] Y. Jin, P. Song, W. Zheng, and L. Zhao, "A feature selection and feature fusion combination method for speaker-independent speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, 2014, pp. 4808–4812.

[13] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Distributing recognition in computational paralinguistics," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 406–417, 2014.

[14] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, 2006.

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

[16] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, British Columbia, Canada, 2003, pp. 153–160.

[17] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. Beijing, China: IEEE, 2005, pp. 846–853.

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

[19] 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*, vol. 29, no. 1, pp. 40–51, 2007.

[20] S.-J. Kim, A. Magnani, and S. Boyd, "Optimal kernel selection in Kernel Fisher discriminant analysis," in *International Conference on Machine Learning (ICML)*. Pittsburgh, PA, USA: ACM, 2006, pp. 465–472.

[21] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1147–1160, 2011.

[22] A. d'Aspremont and S. Boyd, "Relaxations and randomized methods for nonconvex QCQPs," *Technical Report, Stanford University*, 2003.

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

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *European Conference on Speech Communication and Technology*, vol. 5, Lisbon, Portugal, 2005, pp. 1517–1520.

[25] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *International Conference on Data Engineering Workshops*. Atlanta, GA, USA: IEEE Computer Society, 2006.

[26] 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, p. 1161, 2012.

[27] 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)*, Lyon, France, 2013, pp. 2044–2048.

[28] 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 *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013.

[29] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *International conference on Multimedia*. Florence, Italy: ACM, 2010, pp. 1459–1462.

[30] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks," in *Proceedings 36th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*. Prague, Czech Republic: IEEE, 2011, pp. 5688–5691.

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