Dimensionality Reduction for Speech Emotion Features by Multiscale Kernels

Xinzhou Xu1,2, Jun Deng1, Wenming Zheng2, Li Zhao2, Björn Schuller3,4
1MISP Group, MMK, Technische Universität München, Germany
2School of Information Science and Engineering, Southeast University, P.R. China
3Department of Computing, Imperial College London, U.K.
4Chair of Complex and Intelligent Systems, University of Passau, Germany
{xinzhou.xu, jun.deng}@tum.de, {wenming.zheng, zhaoli}@seu.edu.cn, schuller@ieee.org

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.

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 \( X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{n \times N} \) and \( Y = [y_1, y_2, \ldots, y_N] \in \mathbb{R}^{d \times N} \) 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 \( S = [s_1, s_2, \ldots, s_N] \in \mathbb{R}^{c \times N} \) 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 \( e \in \mathbb{R}^{N \times 1} \) 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:
\[
\min_{y_i \in \mathbb{R}} \sum_{i,j=1}^{N} ||y_i - y_j||^2 W_{ij}^{(f)} \text{ s.t. } \sum_{i,j=1}^{N} ||y_i - y_j||^2 W_{ij}^{(p)} = t, \quad (1)
\]

\[
\min_{y_i \in \mathbb{R}} \sum_{i,j=1}^{N} ||y_i - y_j||^2 W_{ij}^{(f)} \text{ s.t. } \sum_{i=1}^{N} y_i^2 D_{ii} = t, \quad (2)
\]

where \(W^{(f)}\) 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

\[
\arg\min \quad zL^{(f)}_T z^T \text{ s.t. } zL^{(p)}_T z = t,
\]

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

For FDA, with \(W^{(f)} = W^{(f)}_{FDA} = S^T (SS^T)^{-1} S\) and \(W^{(p)} = W^{(p)}_{FDA} = I\), 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 \(\alpha \in \mathbb{R}^{N \times 1}\), we can write the kernelized form of FDA, namely KFDA, as

\[
\arg\min \quad \sum_{i=1}^{N} \sum_{j=1}^{N} ||\alpha^T K_{x_i} - \alpha^T K_{x_j}||^2 W_{ij}^{(f)} \text{ s.t. } \sum_{i=1}^{N} \sum_{j=1}^{N} ||\alpha^T K_{x_i} - \alpha^T K_{x_j}||^2 W_{ij}^{(p)} = t.
\]

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

By extending the mapping \(\alpha = A^T = [\alpha_1, \alpha_2, \ldots, \alpha_d] \in \mathbb{R}^{N \times d}\), we obtain multiple mappings by solving the optimization problem. \(\alpha_i\) is the \(i\)th mapping vector with \(i = 1, 2, \ldots, d\). For simplicity, the variables of \(W^{(f)}_{FDA}\) and \(W^{(p)}_{FDA}\) are represented by \(W^{(f)}\) and \(W^{(p)}\), respectively. Therefore, the optimization of MKL-FDA is given as

\[
\arg\min_{\beta} \sum_{i,j=1}^{N} ||A^T \Omega_{x_i} \beta - A^T \Omega_{x_j} \beta||^2 W_{ij}^{(f)} \text{ s.t. } \sum_{i,j=1}^{N} ||A^T \Omega_{x_i} \beta - A^T \Omega_{x_j} \beta||^2 W_{ij}^{(p)} = t,
\]

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 \(\beta\) of multiple kernels.

\[
\arg\min_{\beta} tr\left(A^T Q^{(f)}(\beta) A\right) \text{ s.t. } tr\left(A^T Q^{(p)}(\beta) A\right) = t,
\]

\[
Q^{(f)}(\beta) = \sum_{i,j=1}^{N} (\Omega_{x_i} - \Omega_{x_j}) \beta \beta^T (\Omega_{x_i} - \Omega_{x_j})^T W_{ij}^{(f)},
\]

\[
Q^{(p)}(\beta) = \sum_{i,j=1}^{N} (\Omega_{x_i} - \Omega_{x_j}) \beta \beta^T (\Omega_{x_i} - \Omega_{x_j})^T W_{ij}^{(p)},
\]

\[
\arg\min_{\beta} \beta^T Q^{(f)}(\beta)\beta \text{ s.t. } \beta^T Q^{(p)}(\beta)\beta = t, \quad m = 1, 2, \ldots, M,
\]

\[
Q^{(f)}(\beta) = \sum_{i,j=1}^{N} (\Omega_{x_i} - \Omega_{x_j})^T A A^T (\Omega_{x_i} - \Omega_{x_j}) W_{ij}^{(f)}.
\]

(6) and Eq. (7). (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 \(i = 1, 2, \ldots, N\):

\[
(\Omega_{x_i})_m = \phi_m^T(x_i) \phi_m(x) = e^{-\frac{(x_i - x)^2}{\sigma_m^2}}.
\]

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 feature. 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 \(\Omega_{x_i}\) by linearly combining multiscale kernels. Then, for the new features \(x\), the dimensionality-reduced sample can be achieved by using \(A^T \Omega_{x_i}\beta\) 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
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 2015 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 subsequent 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 $\sigma_m^2$ in Eq. (8) are shown in Table 1. The number of iterations is set as five.

Table 1: The choices of Gaussian kernel parameters $\sigma_m^2$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$m = 1$</th>
<th>$m = 2$</th>
<th>$m = 3$</th>
<th>$m = 4$</th>
<th>$m = 5$</th>
<th>$m = 6$</th>
<th>$m = 7$</th>
<th>$m = 8$</th>
<th>$m = 9$</th>
<th>$m = 10$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_m^2$</td>
<td>0.004n</td>
<td>0.005n</td>
<td>0.01n</td>
<td>0.03n</td>
<td>0.05n</td>
<td>0.06n</td>
<td>0.1n</td>
<td>0.3n</td>
<td>0.5n</td>
<td>0.75n</td>
<td>n</td>
</tr>
</tbody>
</table>

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 dimensionality) 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 $\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)} = \sigma_i (i = 1, 2, \ldots, M, \text{ with } M = 10)$. 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 (Singular 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-
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 $\sigma_5$, $\sigma_6$, $\sigma_8$ for EMO-DB with IS09, and are $\sigma_6$, $\sigma_5$, $\sigma_7$ for EMO-DB with emobase, while for eNTERFACE'05, the parameters are $\sigma_5$, $\sigma_4$, $\sigma_6$ 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 WA on EMO-DB and eNTERFACE'05 using multiscale kernels, with the form of 'IS09 / emobase'.

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Corpora</th>
<th>EMO-DB</th>
<th>eNTERFACE'05</th>
</tr>
</thead>
<tbody>
<tr>
<td>happy</td>
<td>IS09</td>
<td>9.3 / 8.8</td>
<td>35 / 37.2</td>
</tr>
<tr>
<td>sad</td>
<td>IS09</td>
<td>9.3 / 8.8</td>
<td>35 / 37.2</td>
</tr>
<tr>
<td>disgust</td>
<td>IS09</td>
<td>5.1 / 4.9</td>
<td>70 / 59.5</td>
</tr>
<tr>
<td>joy</td>
<td>IS09</td>
<td>60 / 61.5</td>
<td>75 / 60.5</td>
</tr>
<tr>
<td>neutral</td>
<td>IS09</td>
<td>84 / 91.0</td>
<td>42 / 44.5</td>
</tr>
<tr>
<td>sadness</td>
<td>IS09</td>
<td>79 / 79.3</td>
<td>45 / 15.0</td>
</tr>
<tr>
<td>surprise</td>
<td>IS09</td>
<td>79 / 79.3</td>
<td>45 / 15.0</td>
</tr>
</tbody>
</table>

### 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’).

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Feature Sets</th>
<th>IS09</th>
<th>emobase</th>
<th>IS10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>Baseline</td>
<td>27.3 / 27.5</td>
<td>30.3 / 30.4</td>
<td>29.4 / 30.1</td>
</tr>
<tr>
<td>LDA / FDA</td>
<td>28.9 / 29.4</td>
<td>32.4 / 33.0</td>
<td>32.4 / 33.2</td>
<td></td>
</tr>
<tr>
<td>LPP</td>
<td>24.3 / 24.5</td>
<td>30.8 / 30.3</td>
<td>28.0 / 28.6</td>
<td></td>
</tr>
<tr>
<td>MS-KFDA</td>
<td>36 / 38.1</td>
<td>41.0 / 42.4</td>
<td>39.4 / 40.5</td>
<td></td>
</tr>
</tbody>
</table>

In Table 4 for GEMEP, the kernel parameters corresponding to the top-3 recognition rates (both WA and UA) are $\sigma_5$, $\sigma_4$, $\sigma_6$ with the IS09 set, while the parameters are $\sigma_4$, $\sigma_5$, $\sigma_6$ when using the feature set IS10. Yet for emobase, the top-3 rates are $\sigma_5$, $\sigma_4$, $\sigma_6$ and $\sigma_4$, $\sigma_5$, $\sigma_6$, 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.

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.

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Feature Sets</th>
<th>IS09</th>
<th>emobase</th>
<th>IS10</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>Baseline</td>
<td>27.3 / 27.5</td>
<td>32.5 / 32.0</td>
<td>34.7 / 34.5</td>
</tr>
<tr>
<td>ELM</td>
<td>Linear</td>
<td>32.9 / 33.7</td>
<td>33.3 / 33.7</td>
<td>30.8 / 31.8</td>
</tr>
<tr>
<td>MS-KFDA</td>
<td>36.3 / 38.1</td>
<td>41.0 / 42.4</td>
<td>39.4 / 40.5</td>
<td></td>
</tr>
</tbody>
</table>

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


