Comparing Feature Dimension Reduction Algorithms for GMM-SVM based Speech Emotion Recognition

Jianbo Jiang [*†], Zhiyong Wu [*†], Mingxing Xu [†], Jia Jia [†]and Lianhong Cai [*†]
[*]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen Key Laboratory of Information Science and Technology,
Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
[†]Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing, China
E-mail: jjb10@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, {xumx, jjia, clh-dcs}@tsinghua.edu.cn

***Abstract*— How to select effective emotional features are important for improving the performance of automatic speech emotion recognition. Although various feature dimension reduction algorithms were put forward that could help gain the accuracy rate of emotion distinction, but most of them exist various defects, such as high negative impact of the recognition rate, high computational complexity. Regarding this, two dimension reduction algorithms based on PCA (principal component analysis) and KPCA (Kernel-PCA) were comparatively discussed in this paper. The original features extracted from databases were transformed by PCA or KPCA. The weights of these new features over the transforming matrix were calculated and ranked, based on which features were chosen. Experimental results show that feature dimension reduction can make principal contribution to the accuracy of speech emotion recognition, and KPCA slightly outperforms PCA on the hit rate and the remaining dimensions.**

I. INTRODUCTION

Emotion is a significant subject in psychology study which is presented as observable verbal and nonverbal behaviors. Speech is one of observable behaviors expressing emotion, and plays an important role in human-human communication. With the development of computer technology, automatic speech emotion recognition has been one of the latest challenges in the field of human-computer interaction. It has gained great interests in many relevant areas, such as satisfaction evaluation, psychiatric aids and interactive games. It is being further expanded into a broad area of researches, such as human-computer emotional interaction, emotion recognition on visual speeches [1].

To recognize emotion from speech, a variety of acoustic features have been proposed. These features can be categorized into prosodic features, spectral features and voice quality features. Prosodic features consist of statistics derived from the fundamental frequency (f0) and energy contours. Spectral features mainly contain features derived from Mel frequencies, such as Mel-frequency cepstral coefficients (MFCCs). While statistics of jitter, shimmer and harmonic-to-noise ratio (HNR) belong to voice quality features [1][2]. We mainly pay attention to spectral features, or rather, MFCCs. Reference [3] provides MFCCs as emotional features, i.e., 13-dimensional MFCC plus energy, together with their delta and acceleration coefficients, 42 dimensions altogether.

Feature reduction includes feature selection and feature extraction. Feature selection is the process of selecting a subset of relevant features used in model construction, while feature extraction means transforming the input data into a set of features. Feature reduction is also effective in the data analysis process by showing which features are important for prediction, and how these features are related. Ref. [5] provided LDA (Linear Discriminant Analysis) to rank the features and finally got 33-dimensional feature-vector. Ref. [4] proposed a feature selection method based on PCA (Principal Component Analysis). But their biggest drawback is that the dimension of original features is not very high. For example, in [5], original feature was up to over 200 features, and in [4] initial set contained only 85 features. While in this work, the original feature set had a very large dimension of 2688.

In our work, we adopt GMM-SVM (Gaussian mixture model – support vector machine) based system with spectral features for speech emotion recognition. In training the GMM, we use maximum a posteriori (MAP) adaptation method that is widely used in speaker recognition to adapt a universal background model (UBM) to derive the final GMM for each emotion category. The adapted supervector was transformed by PCA or KPCA to calculate and rank the weights of all components. The final supervector features were chosen corresponding to the top largest weights.

The rest of this paper is organized as follows. In Section 2, the GMM-SVM based system for speech emotion recognition is characterized. Then two different feature reduction methods, PCA and KPCA are described in Section 3. Experiments and results are presented in Section 4. Finally, Section 5 gives the conclusions.

II. GMM-SVM BASED EMOTION RECOGNITION

In this work, GMM-SVM based system with spectral features is adopted for emotion recognition of speech. The

This work is partially supported by the National Natural Science Foundation of China (60928005, 60805008, 61375027, 61370023), the Upgrading Plan Project of Shenzhen Key Laboratory (CXB201005250038A) and the Science and Technology R&D Funding of the Shenzhen Municipal.

process of the system is shown in Fig.1, where neutral UBM and the model for each emotion are characterized by GMM.

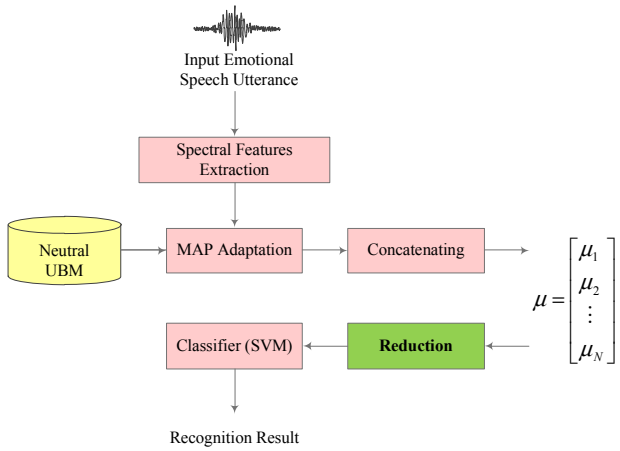

Fig. 1 Architecture of the proposed speech driven talking avatar system based on deep neural network.

The density function of a GMM is defined as following:

$$p(x) = \sum_{i=1}^{M} w_i N(x; \mu_i, \Sigma_i) \tag{1}$$

where $N(;,)$ is the is the Gaussian density function, M is the number of Gaussian mixtures, w_i, μ_i and Σ_i are the weight, mean and covariance matrix of the i-th Gaussian mixture respectively. The supervector of a GMM is defined by concatenating the mean of each Gaussian mixture, which can be thought of as a mapping between an utterance and a high-dimensional vector:

$$\mu = [\mu_1, \mu_2, ..., \mu_M]^{\mathrm{T}} \tag{2}$$

Given an input emotional speech utterance, the spectral features are extracted and used to adapt the GMM from a neutral UBM. The UBM is a GMM that is trained using neutral speeches from a large number of speakers. The MAP adaptation algorithm is used to adapt the GMM from neutral UBM for the input utterance, and during adaptation, only the mean vector μ_i of each Gaussian mixture is adapted. The final GMM supervector is constructed from the adapted GMM as the representation of the input utterance.

The constructed GMM supervector was used to represent emotional space and may have data redundancy. Analysis on the data with such high dimensions would be troublesome to handle. Over-fitting problem may be encountered while the model is trained on a small-scale data as in the situation of emotion recognition. Regarding this, feature reduction is necessary to be investigated. Details of the methods for feature reduction will be elaborated later in the next section.

SVM performs a mapping from an input space to a high-dimensional space through a kernel function. It has been proved to be able to achieve better performance for solving problems in classification, regression and novelty detection than many other classifiers. For simplicity, the linear kernel is selected in the speech emotion recognition system [6].

III. Feature Reduction Algorithm

PCA is a mathematical algorithm that uses an orthogonal transformation to convert a group of possibly correlated features into a group of linearly uncorrelated features, called principal components. Kernel PCA (KPCA) is an extension of PCA using techniques of kernel methods. Details of the algorithms were described in [7].

A. Principal Component Analysis (PCA)

The principle of PCA was shown as follows. After the data set was constructed, the empirical (sample) mean of the distribution was subtracted from the data set. Then the data matrix, $\mathbf{X}^{\mathrm{T}}$, was defined with zero empirical mean, where each of the n rows gives a different repetition in the dataset, and each of the m columns represents a particular kind of datum.

The singular value decomposition of $\mathbf{X}$ is $\mathbf{X}=\mathbf{W}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, where the $m\times m$ matrix $\mathbf{W}$ is the matrix of eigenvectors of the covariance matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, the $m\times n$ matrix $\mathbf{\Sigma}$ is a rectangular diagonal matrix with nonnegative real numbers on the diagonal, and the $n\times n$ matrix $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Then the PCA transformation that preserves dimensionality is given by:

$$\mathbf{Y}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{W} = \mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{V}\mathbf{\Sigma}^{\mathrm{T}} \tag{3}$$

Though $\mathbf{V}$ is not uniquely defined in the usual case when $m<n$-1, $\mathbf{Y}$ will usually still be uniquely defined. Since $\mathbf{W}$ is an orthogonal matrix, each row of $\mathbf{Y}^{\mathrm{T}}$ is simply a linear transformation of the corresponding row of $\mathbf{X}^{\mathrm{T}}$. The n-th column of $\mathbf{Y}^{\mathrm{T}}$ is made up of the "scores" of the cases with respect to the n-th "principal" component, especially, the first column of $\mathbf{Y}^{\mathrm{T}}$ has the scores with respect to the "principal" component, and so on.

B. Kernel Principal Component Analysis (KPCA)

What's different from PCA is that Kernel PCA uses a kernel function to map the d-dimensional data to a higher N-dimensional space. The kernel function is defined as:

$$\begin{aligned} &\Phi : \mathrm{R}^{d} \rightarrow \mathrm{R}^{N} \\ &x \mapsto \xi = \Phi(x) \end{aligned} \tag{4}$$

Similarly, KPCA also operates on zero-centered data as conventional PCA:

$$\sum_{\mu=1}^{N} \Phi(x_\mu) = 0 \tag{5}$$

It operates by diagonalizing the covariance matrix,

$$C = \frac{1}{N}\sum_{\mu=1}^{N} \Phi(x_\mu)\Phi(x_\mu)^{T} \tag{6}$$

In other words, it gives an eigen-decomposition of the covariance matrix:

$$C\nu = \lambda\nu \tag{7}$$

The $n\times n$ kernel matrix $\mathbf{K}$ is defined:

$$\mathbf{K}_{\mu\nu} := (\Phi(x_\mu) \cdot \Phi(x_\nu)) \tag{8}$$

Replacing inner product with kernel function:

$$(v^k \cdot \Phi(x)) = \sum_{i=1}^{M} (\alpha_i)^k \mathbf{K}(x_i, x) \quad (9)$$

If equation (7) does not hold, then the kernel matrix **K** will be revised:

$$\mathbf{K}_{\mu\nu} \to \mathbf{K}_{\mu\nu} - \frac{1}{M}(\sum_{w=1}^{M} \mathbf{K}_{\mu w} + \sum_{w=1}^{M} \mathbf{K}_{w\nu}) + \frac{1}{M^2} \sum_{w,\tau=1}^{M} \mathbf{K}_{w\tau} \quad (10)$$

C. Comparison between PCA and KPCA

KPCA uses a nonlinear kernel function instead of the standard dot product. In fact, KPCA uses a kernel function to map the *d*-dimensional data to a higher *N*-dimensional space, and then performs PCA in the extended space.

PCA and KPCA are essentially different: PCA is based on indicators, while KPCA is based on samples. The advantages of KPCA are: 1) nonlinear principal components afforded better recognition rates than corresponding numbers of linear principal components; and 2) the performance for nonlinear components can be further improved by using more components than possible in the linear case. But if dimension of input space is smaller than the number of examples, KPCA is computationally more expensive than linear PCA.

Otherwise, in linear PCA, we can calculate an "efficient" number of eigenvalues and perform dimension reduction of data by representing the original data as an approximation, projected onto their eigenvectors. However we can't calculate those eigenvectors with KPCA.

IV. Experiments and Results

A. Databases

Two emotional databases are used for the experiments in our work. One is homegrown Chinese Mandarin emotional database named *TV_Movie* Database, which contains 5 kinds of emotions, including four classic emotions (anger, fear, happiness and sadness) and neutral. The database contains 464 voice clips which were interceptions from movies and TV series, with an average length of 3.16s. The data were reconstructed using a sampling rate of 16 kHz with 16-bit resolution, and saved in single channel wav files.

The other is the famous *Berlin* German emotional database [8], which contains about 500 utterances acted by 10 actors in 7 emotional categories (i.e. anger, bored, disgust, fear, joy, sadness and neutral). The data were taken with the sampling rate of 48 kHz and down-sampled to 16 kHz. The average length of the speech recordings of Berlin database is 2.78s.

B. Experimental Setup

From the recordings from the databases, we extract 13-dimensional MFCCs plus energy, together with their delta and acceleration coefficients, forming 42-dimentional acoustic features. The features are computed every 10ms using the frame length of 25ms, with Hamming windowing and pre-emphasis factor of 0.97. The GMM consists of 64 Gaussian mixtures. The neutral UBM is trained from the neutral speech recordings in the speech database as described in [9]. MAP adaptation is performed to adapt the GMM model for each utterance from UBM. Then the supervectors are generated from the adapted GMM. The supervector has a dimension of 42×64=2688. Then feature reduction method PCA or KPCA is operated over the supervectors.

In the following experiments, 5-fold cross validation is performed for error estimation. More precisely, each of the emotional databases described above is equally divided into 5 disjoint subsets, and the classifiers are trained five times, each time with a different subset held out as a testing set.

The hit rate is calculated for evaluating the experiment, which is defined as the ratio of the number of utterances correctly recognized to the total number of all available utterances.

$$HR = \frac{\#\ of\ correctly\ recognized\ utterances}{\#\ of\ all\ utterances} \quad (11)$$

C. Determining Dimensions after PCA

In this work, we carry speaker-independent experiments on both *TV_Movie* and *Berlin* databases using 5-fold cross validation. PCA is performed in each group. The relationship between dimension of principal component and cumulative contribution is shown in Fig.2.

From the figure, we can find only about 200 components (164 components for *TV_Movie* and 218 components for *Berlin*) can make 90% contribution to the dataset, far less than the dimension of original features (2688). Literally, if we choose the first 20 principal components, over 50% contribution will be achieved. With such a small number of features will greatly reduce the training time and predicting time of the classifier.

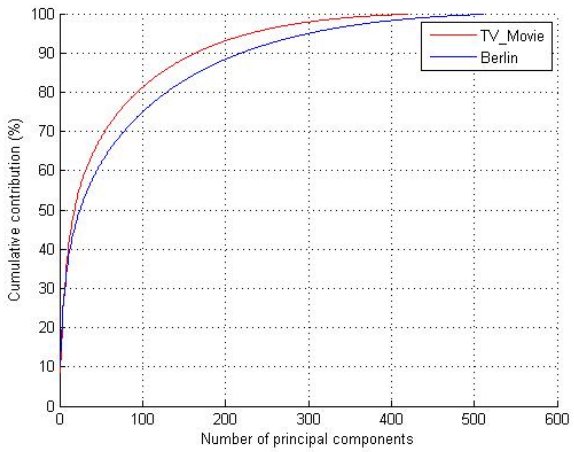

Fig. 2 Relationship between number of principal components and the cumulative contribution. (The red line represents the result on *TV_Movie* database, while the blue line represents the result on *Berlin* database.)

D. Performance of PCA and KPCA

To check the performance using PCA (and KPCA), the following experiment is carried out. In this experiment, PCA and KPCA are both applied to examine the performance. The LDA algorithm is also used for reference. Specific number of

components is chosen according to the cumulative from 100% to 0% with the step of 10% in each database, which could help to accurately determine the inflection point of the curve, and indicate which combinations of principal components are effective. 5-fold cross validation is also performed. The hit rates of the methods (feature reduced by PCA, KPCA or LDA) corresponding to each database are shown in Fig.3.

To quantify the performance of feature reduction, some indices are defined as follows:

- Standard hit rate: 95% of the maximum hit rate with different number of principal components, i.e.

$$HR_{std} = HR_{max} \times 95\% \quad (12)$$

- Minimum dimension:

$$dim_{min} = \min\{dim \mid \forall d(d \geq dim) \Rightarrow (HR_d \geq HR_{std})\} \quad (13)$$

For the above two databases, the standard hit rate and the minimum dimension are tabulated in Table I.

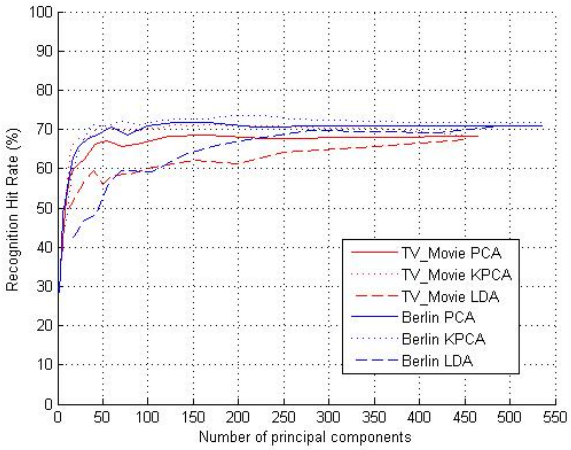

Fig. 3 Hit rates on utterances with principal components using PCA, KPCA or LDA on *TV_Movie* and *Berlin* databases. (The red lines represent the results on *TV_Movie* database, while the blue lines represent the results on *Berlin* database; the solid lines indicate PCA, the dotted lines are for KPCA, while the dashed lines stand for LDA.)

TABLE I
THE STANDARD HIT RATE AND THE MINIMUM DIMENSION UNDER VARIOUS CIRCUMSTANCES

Database	Method	Standard Hit Rate	Minimum Dimension
TV_Movie	PCA	65.3%	41
	KPCA	67.5%	23
Berlin	PCA	68.1%	44
	KPCA	69.9%	38

From the figure and the table, we can find by both feature reduction method PCA and KPCA, the hit rates are obviously higher than the results of LDA in contrast. If the number of principal components is greater than a specific value not exceeding 50, the hit rate is acceptable. As for theoretical analysis, nonlinear principal components afforded better hit rates (i.e. recognition rates) than corresponding numbers of linear principal components, when the number of the components is the same, the corresponding point on the dotted line (i.e. KPCA) is above the one on the solid line (i.e. PCA). From the numerical sense, KPCA is 2.2% higher than PCA on average on *TV_Movie* database, while 1.8% higher than PCA on average on *Berlin* database. Another advantage of KPCA over PCA is, since the adapted GMM has a high degree of non-linearity, features can be reduced to smaller dimensions. For example, on *TV_Movie* database, 41 dimensions are required for PCA to reach the required standard, while only 23 dimensions are needed for KPCA to meet the same standard. Finally, by both PCA and KPCA, can the original 2688 dimension of features be reduced to dozens of principal components, which indicates the two algorithms are both efficient and correctness.

V. CONCLUSIONS

This paper comparatively discusses two feature reduction algorithms based on principal component analysis (PCA) and Kernel-PCA (KPCA) respectively. In detail, the original features extracted from databases are transformed by PCA or KPCA; the principal components are chosen based on the ranking of the weights of these new features over the transforming matrix. Experimental results shows that, components selected by the method can make principal contribution to the hit rate of speech emotion recognition. Both the two algorithms are efficient and correctness. Feature reduction methods can find out meaningful hidden low-dimensional structures to avoid the influence of excessive redundant dimension on complexity of the classifier. In addition, KPCA outperforms PCA a bit in two indices: the hit rate and the remaining dimension.

REFERENCES

[1] A. Tawari, "Speech emotion analysis: Exploring the role of context," *IEEE Trans. Multimedia*, 12(6): 502-509, 2010.

[2] I. Luengo, E. Navas, I. Hernaez and J. Sanchez, "Automatic emotion recognition using prosodic parameters," in *Proc. Interspeech*, 493-496, 2005.

[3] H. Hu, M.X. Xu and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition," in *Proc. ICASSP*, 413-416, 2007.

[4] X.H. Luo, D.L. Yang, M.X. Xu and L. Xu, "PCA based feature selection algorithm on speaker-independent speech emotion recognition," *Computer Sciences*, 38, 2011 (in Chinese).

[5] B. Schuller, G. Rigoll and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," in *Proc. ICASSP*, 2004.

[6] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, 2(3), 2011.

[7] L.Xu and X.B. Zhang, "Fast kernel feature extraction method and its application," *Computer Engineering*, 35, 2009 (in Chinese).

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, 2005.

[9] M. Dai, D. Yang and M.X. Xu, "Research on the composition of UBM training set in speech emotion recognition," in *Proc. NCMMSC*, 2011 (in Chinese).