2008 20th IEEE International Conference on Tools with Artificial Intelligence

Robust Speaker Identification using Greedy Kernel PCA

Min-Seok Kim, IL-Ho Yang and Ha-Jin Yu[1]
School of Computer Science, University of Seoul, Korea
E-mail: {ms, heisco, hjyu}@uos.ac.kr

Abstract

We propose a robust speaker identification system in noisy environments using greedy kernel principal component analysis. We expect that kernel PCA can project important information to some axes and the noise to some other axes in the arbitrary high dimensional space resulting in denoising of the input features. However, it is not easy to use kernel PCA for speaker identification because the storage required for the kernel matrix grows quadratically, and the computational cost grows linearly with the number of training vectors. Therefore, we use greedy kernel PCA which can approximate kernel PCA with small representation error. In the experiments, we compare the accuracy of the greedy kernel PCA with that of the baseline Gaussian mixture models using MFCCs and PCA in noisy environment. As the results, the greedy kernel PCA outperforms conventional methods.

1. Introduction

With the emerging era of ubiquitous computing, speaker recognition have a high potential to be an effective solution of security problems. Speaker identification is a kind of speaker recognition which identify human with their voices, and has many potential applications. However, the level of the accuracy of the current speaker recognition systems has not reached that the users expect. The technology is still developing in various aspects which include extracting good features for speaker discrimination, building good speaker models, and normalizing the scores. Conventionally, melcepstral coefficients and Guassian mixture models[10] are the most basic approaches for features and models respectively.

These days, to cope with the nonlinearity of the speech feature, kernel methods are used in speaker recognition. Kernel methods can transform the features in the input space to a new feature space with much higher dimension. The speech features that cannot be separated linearly in the input space can be separated linearly in the new arbitrary high dimensional feature space.

In this research, we propose a speaker identification system using a kernel method which is expected to model the non-linearity of speech features well. We have been using principal component analysis (PCA) successfully, and extended to kernel PCA, which is used for many pattern recognition tasks such as face recognition. However, we cannot use kernel PCA for speaker identification directly because the storage required for the kernel matrix grows quadratically, and the computational cost to compute eigenvector of matrix grows linearly with the number of training vectors. Therefore, we use greedy kernel PCA which can approximate kernel PCA with small representation error.

In the next section, we introduce kernel principal components analysis. Section 3 describes the greedy KPCA, and section 4 reports experimental evaluation on speaker identification. Finally, section 5 summarizes the conclusions drawn from this study.

2. Feature extraction using kernel PCA

Principal component analysis (PCA)[13] and its nonlinear version kernel PCA (KPCA)[3][6] are well-known techniques for dimensionality reduction. The coordinates in the eigenvector basis corresponding to the maximum variance are called principal components. The projected features onto principal components will retain important information and be denoised by dropping directions with small variances. In KPCA, the d dimensional feature vectors $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_l]^T$, $\mathbf{x}_i \in \Re^d$ (input space) are mapped to a very high dimensional space (feature space) $\Re^\infty$ via a mapping function ϕ, and the PCA is performed in feature space.

$$\phi : \Re^d \rightarrow \Re^\infty . \quad (1)$$

[1] Corresponding author

1082-3409/08 $25.00 © 2008 IEEE
DOI 10.1109/ICTAI.2008.105

The new coordinates for KPCA with centered features $\mathbf{X}_\phi = [\phi(\mathbf{x}_1),...,\phi(\mathbf{x}_l)]^T$ can be calculated by the eigenvalue decomposition problem of kernel matrix $\mathbf{K}$ which is an $l \times l$ matrix whose elements K_{ij} are defined as

$$K_{ij} \equiv \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j). \tag{2}$$

We can employ mercer kernel function [3] to compute dot product in feature space.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{3}$$

Many researches[3][4][5] show that the KPCA is good at analyzing nonlinear structure, however there are disadvantages that the storage of the kernel matrix grow quadratically with the number of feature vectors l, and it is not feasible to compute eigenvectors where l is larger than 5,000 or more because $\mathbf{K}$ is an $l \times l$ matrix. In previous works to apply KPCA to speech processing[4][5], they use randomly selected features to compute KPCA. There are two types of solutions to solve this problem. One is reducing the training set [1][11] and the other is computing KPCA iteratively[12]. In this paper we use the formal method called greedy KPCA (GKPCA)[1]. In section 3, we show the method of computing the reduced kernel matrix $\mathbf{K}$.

3. Greedy kernel PCA

Greedy kernel PCA (GKPCA) algorithm [1] can compute KPCA using reduced training set. Let $\mathbf{S}_\phi = [\phi(\mathbf{s}_1),...\phi(\mathbf{s}_m)]^T$ and $\mathbf{X}_\phi = [\phi(\mathbf{x}_1),...\phi(\mathbf{x}_l)]^T$ $(\mathbf{S}_\phi \subset \mathbf{X}_\phi)$ denote the mapped features of $\mathbf{S} = [\mathbf{s}_1,...\mathbf{s}_m]^T$ and $\mathbf{X} = [\mathbf{x}_1,...\mathbf{x}_l]^T$ $(\mathbf{S} \subset \mathbf{X})$ respectively. In GKPCA, the subset $\mathbf{S}_\phi = [\phi(\mathbf{s}_1),...\phi(\mathbf{s}_m)]^T$ is selected from the total training set $\mathbf{X}_\phi = [\phi(\mathbf{x}_1),...\phi(\mathbf{x}_l)]^T$ which minimizes the cost function (defined in equation 5) over the total training set $\mathbf{X}_\phi$ and the size m of $\mathbf{S}_\phi$ should be as small($m << l$) as possible to compute kernel matrix $\mathbf{K}$ and its eigenvectors. Let $I = \{1,...,l\}$ denote indices of the total training set $\mathbf{X}_\phi$ and $J = \{1,...,m\}$ denote indices of m selected subset $\mathbf{S}_\phi$. The reconstructed training set $\widetilde{\mathbf{X}}_\phi = [\phi(\widetilde{\mathbf{x}}_1),...,\phi(\widetilde{\mathbf{x}}_l)]^T$ from subset $\mathbf{S}_\phi$ can be expressed as follows

$$\widetilde{\mathbf{X}}_\phi = \mathbf{S}_\phi \mathbf{B} \tag{4}$$

where $\mathbf{B} = [\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_l]$, $\boldsymbol{\beta}$ is $1 \times m$ (i.e. $\mathbf{B} \in \Re^{m \times l}$) coefficients of subset $\mathbf{S}_\phi$.

The mean square error is the objective function of the reduced set to minimize:

$$\varepsilon_{MS} = \frac{1}{l} \sum_{i \in I} \varepsilon_i^R \tag{5}$$

where ε_i^R is the reconstruction error defined as

$$\varepsilon_i^R = \left\| \phi(\mathbf{x}_i) - \sum_{j \in J} [\boldsymbol{\beta}_i]_j \phi(\mathbf{s}_j) \right\|^2 \tag{6}$$

$\boldsymbol{\beta}_i$ is computed[1] to minimize ε_{MS} as

$$\boldsymbol{\beta}_i = (\mathbf{K}^s)^{-1} \mathbf{k}^s(\mathbf{x}_i), \quad \forall i \in I \tag{7}$$

where $\mathbf{K}^s$ is $m \times m$ kernel matrix of selected subset $\mathbf{S}_\phi$, i.e. the elements $[K^s]_{ij}$ of $\mathbf{K}^s$ is defined as $k(\mathbf{s}_i, \mathbf{s}_j)$, and the vector $\mathbf{k}^s(\mathbf{x}_i) = [k(\mathbf{s}_1, \mathbf{x}_i),...,k(\mathbf{s}_m, \mathbf{x}_i)]$ contains the result of kernel function between subset $\mathbf{S}$ and $\mathbf{x}_i$. Equation 5 is re-expressed using equation 7.

$$\varepsilon_{MS} = \frac{1}{l} \sum_{i \in I} (k(\mathbf{x}_i, \mathbf{x}_i) - 2\mathbf{K}^s \mathbf{k}^s(\mathbf{x}_i) + \mathbf{k}^s(\mathbf{x}_i) \cdot \mathbf{K}^s \mathbf{k}^s(\mathbf{x}_i)) \tag{8}$$

To select $\mathbf{S}_\phi$ efficiently, two adjustments are proposed in [1]. Firstly, we can select $\mathbf{S}_\phi$ gradually using the mean square error $\varepsilon_{MS}^{(t)}$ in t^{th} iteration. The mean square error $\varepsilon_{MS}^{(t)}$ can be upper bounded as the following inequality,

$$\varepsilon_{MS}^{(t)} \le \frac{1}{l}(l-t) \max_{j \in I \setminus J} \left\| \phi(\mathbf{x}_i) - \sum_{j \in J} [\boldsymbol{\beta}_i]_j \phi(\mathbf{s}_j) \right\|^2. \tag{9}$$

Secondly, we can use orthonormalized basis $\mathbf{W} = [\mathbf{w}_1,...,\mathbf{w}_m]$ gradually.

$$\phi(\widetilde{\mathbf{x}}) = \mathbf{W}\phi(\mathbf{z}) \tag{10}$$

where $\phi(\mathbf{z})$ is a new representation of $\phi(\mathbf{x})$. Using the orthonormal condition that $\mathbf{W}^T\mathbf{W}$ is an identity matrix, the optimal $\mathbf{z}$ minimizing the reconstruction error ε^R is computed as

$$\phi(\mathbf{z}) = \mathbf{W}^T \phi(\mathbf{x}) . \quad (11)$$

As the result of equation 10, the reconstruction error ε^R in equation 6 is computed as

$$\varepsilon^R = \left\| \phi(\mathbf{x}) - \phi(\tilde{\mathbf{x}}) \right\|^2 = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}) - \phi(\mathbf{z}) \cdot \phi(\mathbf{z}) . \quad (12)$$

The detailed algorithm can be found in [1] which selects $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_m]^T$ instead of $\mathbf{S}_\phi$ using Gram-Schmidt orthogonalization process. Finally, the kernel matrix can be computed by using $\mathbf{Z}$.

$$\mathbf{K} \approx \mathbf{Z}^T \mathbf{Z} . \quad (13)$$

The reduced kernel matrix $\mathbf{K}$ is an $m \times m$ matrix where $m << l$. Therefore, we can compute KPCA in feasible time.

4. Experiments and results

We used the YOHO database which consists of 138 speakers prompted to read combination lock phrases, for example, "67 34 85." The YOHO database has 'enroll' and 'verify' mode. The 'enroll' consists of 4 sessions with 24 utterances in each session, and the 'verify' consists of 10 sessions with 4 utterances in each session. We used 50 speakers, labeled from 101 to 154, for speaker identification task. We use only the first session in 'enroll' for training and all sessions in 'verify' for testing. In the verification data, we added noises ('babble', 'restaurant' in Aurora2 database[9]) with SNR 15dB and 20 dB artificially using FaNT[7][8].

Each frame was represented by 12 mel-frequency cepstrum coefficients, the log energy, and their first- and second-order time derivatives (delta and delta-delta features), for a resulting 39-dimensional feature vector, with 20ms window and 10ms shift from the audio recording. The frames of data corresponding to silence were removed from the utterances using energy threshold. To build the speaker models, we use GMMs with 128 mixtures.

Figure 1 shows the flow of the robust speaker identification system. Firstly, we extract speaker feature vectors using MFCCs, and derive the new coordinates using GKPCA from clean speech for training. Then we project MFCCs onto new coordinates and apply the robust feature vectors to training models and to speaker identification task.

We used radial basis kernel function with $\sigma^2 = 2$ to compute equation 3.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2} \frac{\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2}{\sigma^2} \right) \quad (14)$$

To compute kernel function effectively, the feature vectors are normalized to be in the range -1 to 1.

We compared the system using GKPCA with the baseline system using MFCCs feature vectors and PCA. Figure 2, 3 and table 1 show the error rate of speaker identification results. As the results, robust feature vectors derived by GKPCA show 25.85% and 13.90% error rate when babble noise and restaurant noise are added respectively with SNR 15dB, and 7.45%, 3.25% respectively with SNR 20dB. The average relative error rate reduction over baseline system is 16.8%.

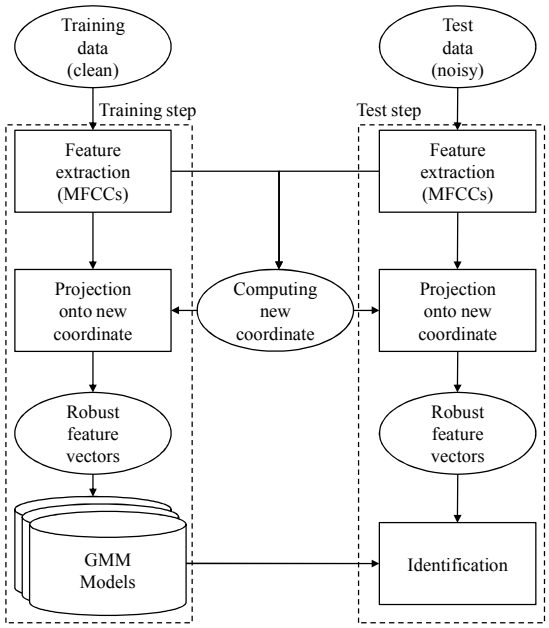

Figure 1. The flow of the robust speaker identification system

Table 1. Error rate of speaker identification

		MFCCs	PCA	GKPCA
SNR15dB	Babble	27.65	33.00	25.85
	Restaurant	18.25	16.00	13.90
SNR20dB	Babble	8.25	8.25	7.45
	Restaurant	6.50	4.00	3.25
average		15.16	15.31	12.61

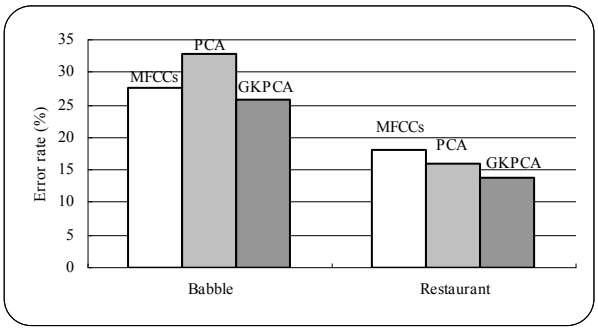

Figure 2. Error rate of speaker identification with SNR 15dB

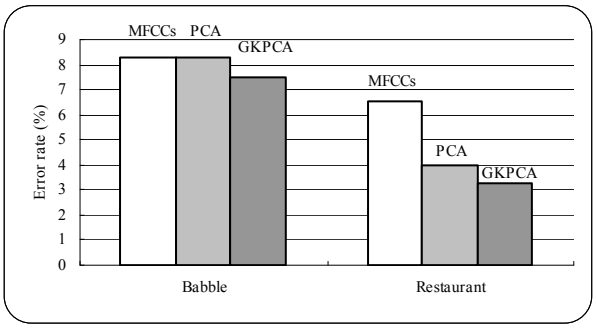

Figure 3. Error rate of speaker identification with SNR 20dB

5. Conclusions

We have proposed a noise-robust speaker identification system using greedy kernel principal component analysis (GKPCA). The basic assumption in this research is that speech features are not linearly separable, hence we expect that the performance of speaker identification will improve if we transform the features into an arbitrary high dimensional space where the speech features can be separated linearly. Kernel principal component analysis is one of the methods which can find efficient axis system in the arbitrary high dimensional space, but it is not practical to use it because we should compute eigenvectors of the $l \times l$ matrix where l is the number of feature vectors. Therefore, we applied greedy kernel PCA which can approximate kernel PCA by selecting relevant feature vectors incrementally. As the result, the average relative error rate reduction over baseline system is 16.8%. The proposed system can be effective where the training data are very limited because KPCA cannot make full use of the large amount of training data.

6. References

[1] V. Franc. *Optimization Algorithms for Kernel Methods*, PhD thesis, Centre for Machine Perception, Czech Technical University, 2005.

[2] T. Tangkuampien and D. Suter, "Human motion de-noising via greedy kernel principal component analysis filtering", *In Int. Conf. on Pattern Recognition*, 2006, pp. 457–460.

[3] B. Schölkopf, A. Smola, and K.Müller, "Kernel principal component analysis", *In Int. Conf. on Artificial Neural Networks*, 1997, pp. 583–588.

[4] T. Takiguchi and Y. Ariki, "Robust Feature Extraction using Kernel PCA", *In Int. Conf. on Acoustics, Speech and Signal Processing*, 2006, pp. 509-512.

[5] A. Lima, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda and T. Kitamura, "On the Use of Kernel PCA for Feature Extraction in Speech Recognition", *In eurospeech 2003*, 2003, pp. 2625-2628.

[6] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York, 2004.

[7] H.-G. Hirsch, "Fant-fltering and noise adding tool", http://dnt.kr.hs-niederrhein.de/download.html.

[8] X. Domont, M. Heckmann, F Joublin and C. Goerick, "Hierarchical Spectro-temporal Features for Robust Speech Recognition", *In Int. Conf. on Acoustics, Speech and Signal Processing*, 2008, pp 4417-4420.

[9] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *in ASR 2000*, 2000, pp. 181–188.

[10] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication*, vol. 17, 1995, pp. 91–108.

[11] B. Scholkopf, P. Knirsch, A. J. Smola, and C. Burges, "Fast approximation of support vector kernel expansions and an interpretation of clustering as approximation in feature spaces", *In Proc. DAGM Symp. Mustererkennung*, 1998, pp. 124–132.

[12] T.-J. Chin, D. Suter, "Incremental kernel Principal component analysis", *IEEE Transactions on Image Processing*, Vol. 16, No. 6, 2007, pp. 1662-1674.

[13] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification (2nd Edition)*, Wiley-Interscience, 2000

Acknowledgement

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD) (KRF-2007-321-A00155)