Robust Speaker Identification Using Ensembles of Kernel Principal Component Analysis

IL-Ho Yang[1] , Min-Seok Kim[2], Byung-Min So[1], Myung-Jae Kim[1] and Ha-Jin Yu[1]

[1] University of Seoul, School of Computer Science,
Seoulsiripdaero 163, Dongdaemun-gu, Seoul, Korea
[2] Advanced Research Institute, LG Electronics Inc., Seoul, Korea
heisco@hanmail.net, minseok3.kim@lge.com, sbm1210@naver.com,
arthmody@naver.com, hjyu@uos.ac.kr

Abstract. In this paper, we propose a new approach to robust speaker identification using KPCA (kernel principal component analysis). This approach uses ensembles of classifiers (speaker identifiers) to reduce KPCA computation. KPCA enhances the features for each classifier. To reduce the processing time and memory requirements, we select a subset of limited number of samples randomly which is used as estimation set for each KPCA basis. The experimental result shows that the proposed approach shows better accuracy than PCA and GKPCA (greedy KPCA).

Keywords: classifier ensemble, greedy kernel PCA, speaker identification.

1 Introduction

The accuracy of the speech and speaker recognition systems can be degraded according to the environmental condition including the channel, noise, etc. Various feature enhancement methods which transform conventional speaker features (such as MFCCs) are used in order to alleviate the problem.

PCA (principal component analysis) [1] is one of the feature enhancement methods which are widely used, but it cannot represent nonlinearly distributed data properly. KPCA (kernel PCA) [2][3] can handle nonlinearly distributed data. But the computational complexity and memory requirement are increased proportionally to the square of the number of samples which are used to estimate the KPCA basis. Therefore, we cannot apply KPCA to speaker identification directly because a large number of features can be extracted from a short utterance.

GKPCA (greedy KPCA) [4] can reduce the computational complexity and memory requirement by using greedy filtering. Greedy filtering selects a subset of the whole feature vectors with minimal representation error. However, in general, the number of subset feature vectors is very smaller than the number of whole features. Therefore, accuracy improvement of GKPCA is limited. We try to overcome this limitation by applying the concept of classifier ensemble.

In [5], a hybrid system for robust speaker identification in adverse environments is proposed. It combines two kinds of classifiers which are trained by different feature-sets. One is based on popular MFCCs and the other on the new parametric feature-set (PFS). In this research, like [5], we combine multiple classifiers which trained by different feature-sets. However, unlike [5], we extract feature-sets using KPCA transforms from original MFCCs feature-set. As in GKPCA, we apply KPCA to subsets of the whole features. At this time, the subsets are selected randomly unlike GKPCA. This process is repeated many times in order to obtain several subsets. Each subset is used to estimate the KPCA basis. Multiple classifiers (speaker identifiers) are trained with these features. Finally, the speaker identification results are combined using majority voting [6].

The remainder of this paper is organized as follows. Section 2 describes the proposed ensemble system. Section 3 and 4 show the experimental results and conclusion respectively.

2 Related Works

2.1 Speaker Identification Using GMM-UBM [7]

GMM (Gaussian mixture model) [8] is well-known modeling method in speaker recognition domain. It represents a speaker model with weighted combination of several Gaussian components (probability density functions). The equation of likelihood function is

$$p\left(\vec{x} \mid \lambda\right) = \sum_{i=1}^{M} w_i p_i\left(\vec{x}\right). \tag{1}$$

Where $\vec{x}$ and λ are input feature vector and model parameter (weights, means and covariances). w_i and p_i represent weight and probability density function of i^{th} Gaussian component respectively. In D-dimensional space, p_i is

$$p\left(\vec{x}\right) = \frac{1}{\left(2\pi\right)^{D/2}\left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\vec{x}-\vec{\mu}_i\right)^{\mathrm{T}}\left(\Sigma_i\right)^{-1}\left(\vec{x}-\vec{\mu}_i\right)\right\}. \tag{2}$$

Where $\vec{\mu}_i$ and Σ_i are mean vector and covariance matrix of i^{th} Gaussian component.

In GMM-UBM method, each speaker dependent model is adapted from UBM (universal background model) which is a speaker independent model to represent general human voice. UBM is represented by a GMM which is trained by EM

(Expectation-Maximization) algorithm [8]. Also, each speaker model is represented by a GMM which is adopted from UBM by MAP adaptation [7].

In identification phase, log likelihoods of input feature vector sequence $\vec{X} = \left[\vec{x}_1, \vec{x}_2 ..., \vec{x}_t\right]^T$ are calculated for each speaker model (equation (3)). And then, speaker identification system chooses an identified speaker which has the highest log likelihood.

$$\log p\left(\vec{X} \mid \lambda\right) = \sum_{i=1}^{t} \log p\left(\vec{x}_i \mid \lambda\right). \tag{3}$$

2.2 Feature Enhancement Using GKPCA [4]

PCA (Principal Component Analysis) [1] is used as popular feature enhancement method. It transforms original features to new basis which maximize the variance of the whole features. However, PCA may be ineffective when the feature vectors have nonlinear structure.

KPCA (Kernel Principal Component Analysis) [2][3] is the nonlinear version of PCA. It is more appropriate than PCA when the feature vectors have nonlinear structure. In KPCA, the D-dimensional feature vectors $\vec{X} = \left[\vec{x}_1, \vec{x}_2, ..., \vec{x}_t\right]^T$, $\vec{x}_i \in \Re^D$ in input space are mapped to a very high dimensional space $\Re^\infty$ (feature space) via a mapping function ϕ, and the PCA is performed in feature space.

$$\phi : \Re^D \to \Re^\infty. \tag{4}$$

The new coordinates for KPCA with centered features $\vec{X}_\phi = \left[\phi\left(\vec{x}_1\right), \phi\left(\vec{x}_2\right), ..., \phi\left(\vec{x}_i\right)\right]^T$ can be calculated by the eigenvalue decomposition problem of kernel matrix K which is a t x t matrix whose elements $K_{i,j}$ are defined as

$$K_{i,j} \equiv \phi\left(\vec{x}_i\right) \cdot \phi\left(\vec{x}_j\right). \tag{5}$$

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

$$k\left(\vec{x}_i, \vec{x}_j\right) = \phi\left(\vec{x}_i\right) \cdot \phi\left(\vec{x}_j\right). \tag{6}$$

In researches with toy examples or with a small number of training set, KPCA outperforms PCA. However, the memory requirement and the computational complexity increase quadratically with the number of training data. For this reason, it

is difficult to apply KPCA-based feature extraction method to large-scale problems such as speaker or speech recognition systems.

GKPCA (Greedy Kernel Principal Component Analysis) [4] can compute KPCA using reduced training set. In GKPCA, the subset $\vec{S}_\phi = \left[\phi(\vec{s}_1), \phi(\vec{s}_2), ..., \phi(\vec{s}_n)\right]^{\mathrm{T}}$ ($\vec{S}_\phi \subset \vec{X}_\phi$) is selected from the total training set which minimizes the cost function over the total training set and the size n of the subset should be as small as possible to compute kernel matrix K and its eigenvectors. Fig. 1 shows the process of GKPCA briefly.

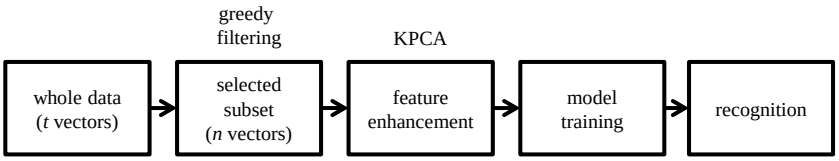

Fig. 1. Brief process of GKPCA.

3 Proposed Method

GKPCA estimates KPCA basis with n samples (speaker feature vectors) which are selected as a subset of the whole t samples ($n << t$). That is, GKPCA uses a small number of n samples. It may not be sufficient because n is very small number in comparison with t.

In this research, we propose an ensemble system using KPCA. The proposed method selects several subsets of the whole data randomly and train multiple classifiers with the enhanced subset features which are transformed using KPCA. Fig. 2 shows the proposed method in case of combining m classifiers.

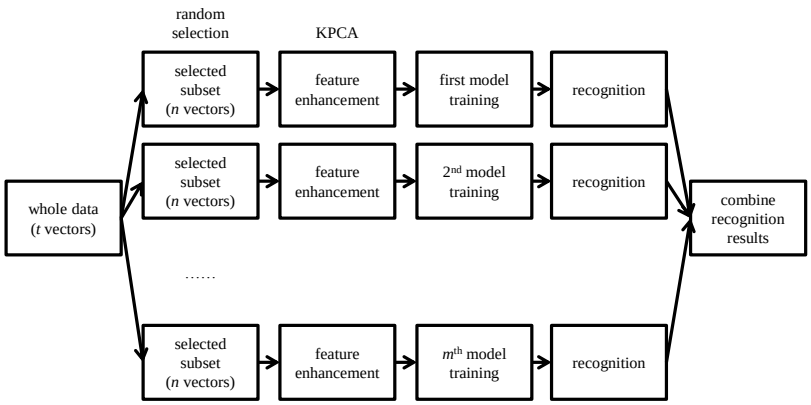

Fig. 2. Combination of m classifiers (proposed method).

The proposed method uses m times as much features than conventional GKPCA without increasing the time and memory complexity in proportion to the square of the number of training features. Therefore, we expect that it will increase the accuracy of the speaker identification.

We use majority voting scheme to combine the classifiers. This scheme chooses the class which receives the highest number of votes from the classifiers.

$$J = \arg\max_{j=1,C} \sum_{i=1}^{m} d_{i,j} \ . \tag{7}$$

Where $d_{i,j}$ means the result of i^{th} classifier for j^{th} class (0 or 1) and C is number of classes (number of speakers). If the result of i^{th} classifier is j^{th} class, $d_{i,j}$ is 1 and otherwise 0. J^{th} class is the combined result.

4 Experiments and the results

4.1 Database and Speaker Model Training

To evaluate the proposed system in various environments, two types of speech corpora are used: ETRI PC DB and ETRI cellular phone DB. These are consist of Korean speeches which are collected by ETRI (Electronics and Telecommunications Research Institute) speech information research center at Korea. Types of utterances are 2-digit numbers, 4-digit numbers and sentences. Each recording session is repeated 4 times and each session includes 5 recording trials. In ETRI PC DB, whole recording files are saved 16kHz / 16bits mono wave format (headerless linear PCM). The wave files in cellular phone DB are compressed by mu-law (8kHz / 16bits). For the experimental equality, whole recording files of the two DBs are uncompressed and downsampled to 8kHz. The speakers of the corpora are divided into three groups according to the session terms: 'WEEK', 'MONTH' and 'SEASON'. To build a UBM, 'MONTH' speakers' 10 utterances are used as training set (1^{st} month – 1^{st} session). For training speaker models and testing, 'WEEK' speakers' 10 utterances are used (training: 1^{st} week – 1^{st} session, test: 3^{rd} week – 1^{st} session). The number of test speakers for PC DB is one hundred and that for cellular phone DB is 104. The UBM training sets consist of 101 speakers.

The GMM-UBM [7] is used for speaker model training. To build a UBM, GMMs with 256 Gaussian components are trained by EM (expectation-maximization) algorithm [8]. The number of Gaussian components is started from 1 and is doubled, e.g. 1, 2, 4, 8, and 16. The model parameters (weights, means and covariances of Gaussian components) are trained once whenever the number of components is doubled. In the final stage (256 mix), model parameters are trained ten times. The

speaker models are adapted once from UBM using MAP adaptation [7] with each speaker's training set (τ=1).

To evaluate the proposed system in noisy environment, CAR, SUBWAY and RESTAURANT noise in Aurora2 DB is added at SNR 20dB and 10dB. FaNT is used for adding noise.

4.2 Feature Extraction

15 MFCCs (mel-frequency cepstral coefficients) and energy and their derivatives are derived from each utterance (window size = 25ms, shift = 10ms). Silences are removed by energy based method in feature level. The CMVN (cepstral mean and variance normalization) was applied for each utterance.

4.3 Feature Enhancement

Each basis of PCA, GKPCA and proposed method is derived from the UBM training set. Then, the whole features (UBM training set, speaker model train set, test set) are transformed using the methods.

In KPCA, Gaussian kernel function (equation (7)) is used (σ=32).

$$k\left(\vec{x}_i, \vec{x}_j\right) = \exp\left(-\frac{\left\|\vec{x}_i - \vec{x}_j\right\|^2}{2\sigma^2}\right). \quad (8)$$

In greedy filtering and random selection, we select one hundred feature vectors from the UBM training set. This size is maximum number which can process KPCA without out of memory error in our PC (2GB RAM). For the proposed method, we use an ensemble of one hundred classifiers.

4.4 Experimental Results

Table 1 shows the speaker identification rate of the overall experiments which include two types of corpora (PC and cellular phone DB). 'NOISE' represents type and SNR of noises which are added to the original wave data using FaNT. 'CLEAN' means original wave data and others mean noise added data, e.g. CAR20 (CAR noise added at SNR 20dB), RESTAURANT10 (RESTAURANT noise added at SNR 10dB). 'Baseline', 'PCA', 'GKPCA', 'Proposed' mean MFCC features and their enhanced features using PCA, GKPCA and proposed method, respectively.

Table 1. Experimental results.

DB	NOISE	Base line	PCA	GK PCA	Pro posed	MAX
PC	CLEAN	**96.50**	96.30	96.30	96.20	96.50
	CAR20	82.90	80.90	85.10	**86.10**	86.10
	SUBWAY20	**70.30**	67.40	67.10	67.80	70.30
	RESTAURANT20	80.20	74.80	82.50	**83.00**	83.00
	CAR10	57.40	40.10	55.00	**58.50**	58.50
	SUBWAY10	35.60	33.10	**35.70**	35.40	35.70
	RESTAURANT10	**63.80**	61.10	60.70	63.40	63.80
CELL PHONE	CLEAN	**79.90**	78.17	78.37	79.62	79.90
	CAR20	45.19	46.63	**47.02**	46.35	47.02
	SUBWAY20	57.69	57.21	59.71	**60.38**	60.38
	RESTAURANT20	53.37	58.56	58.46	**59.42**	59.42
	CAR10	22.21	26.35	28.17	**29.33**	29.33
	SUBWAY10	32.21	38.37	38.27	**42.12**	42.12
	RESTAURANT10	32.02	38.56	39.81	**40.96**	40.96
	AVERAGE	57.81	56.97	59.44	**60.61**	60.61

Our proposed method shows better average accuracy in the overall environments over the other methods. Fig. 3 shows the average speaker identification accuracy according to the number of classifiers.

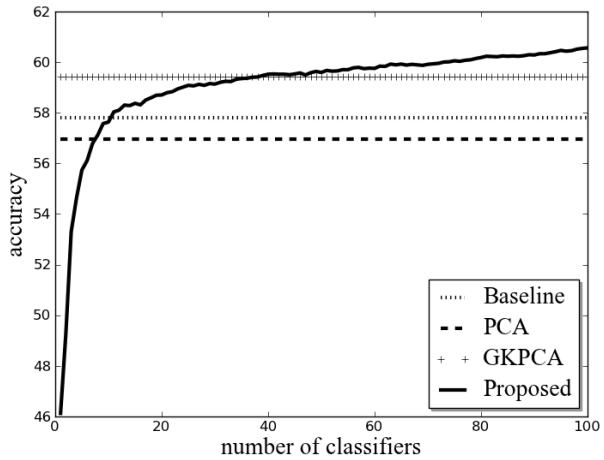

Fig. 3. Average speaker identification accuracy according to the number of classifiers.

If we don't construct classifier ensemble, proposed method shows the lowest identification rate (46.12%). It is a natural result because our method is based on random selection. But, when we combining 11 classifiers, the accuracy of the proposed method is the same as the baseline (58.03%). And, when we combining up

to 39 classifiers, proposed method shows the highest identification rate (59.47% at 39 classifiers and 60.56% at 100 classifiers). These results mean that the limitation of GKPCA can be overcome by the proposed method.

5 Conclusions

We have proposed an ensemble system for speaker identification using kernel PCA. In this research, like GKPCA, a small subset of the whole data is used to estimate each KPCA basis. Unlike GKPCA, each subset for a classifier is selected randomly and multiple classifiers for the ensemble system are trained. As the results, the proposed method shows better average accuracy in various environments.

Acknowledgement

This research was supported by Basic Science Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0024047)

References

1. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification, John Wiley & Sons (2001)
2. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis, Cambridge University Press (2004)
3. Scholkopf, B., Smola, A., Muller, K-R.: Kernel Principal Component Analysis. In Proceedings of International Conference on Artificial Neural Networks, pp. 583--588 (1997)
4. Kim, M-S., Yang, I-H., Yu, H-J.: Robust Speaker Identification using Greedy Kernel PCA. In Proceedings of 20th IEEE International Conference on Tools with Artificial Intelligence, pp. 143--146 (2008)
5. Mashao, Daniel J., Skosan, Marshalleno: Combining classifier decisions for robust speaker identification, Pattern Recognition, vol. 39, pp.147--155 (2006)
6. Polikar, R.: Ensemble based Systems in Decision Making. Circuits and Systems Magazine (2006)
7. Reynolds, D. A., Quatieri, T. F., Dunn, R. B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, Vol. 10, pp. 19--41 (2000)
8. Reynolds, D. A., Rose, R. C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech Audio Processing, Vol. 3, No. 1, pp. 72--83 (1995)