# Speaker Identification Using Ensembles of Feature Enhancement Methods

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] 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 classifier ensemble of various channel compensation and feature enhancement methods for robust speaker identification on various environments. The proposed ensemble system is constructed with 15 classifiers including three channel compensation methods (including CMS and variance normalization, and without compensation) and five feature enhancement methods (including PCA, kernel PCA, greedy kernel PCA, kernel multimodal discriminant analysis, and without enhancement). Experimental results show that the proposed ensemble system gives the highest average speaker identification rate in various environments (channels, noises, and sessions).

**Keywords:** classifier ensemble, greedy kernel PCA, kernel multimodal component analysis, speaker identification.

## 1 Introduction

Environmental mismatches (channel or noise) of training and test utterances can lower the accuracy of the speaker identification systems. If the system is used at a fixed place, we can avoid the channel mismatch problem by training the speaker models using utterances recorded at the same place. However, there are situations that we cannot know the environment such as telephone lines in advance. The utterances recorded during telephone calls are influenced by the characteristics of both the sender and receiver's devices.

These facts cause the decrease of speaker identification rate at certain domains such as digital forensic investigations where it is difficult to know the channel characteristics of the test utterances. For example, in case of the crimes using telephones, the criminals can call to victims through various channels. We need the suspects' utterances to train the speaker models to identify the unknown voice. At this point, a criminal investigator cannot guarantee the reliability of the speaker

* Corresponding author

identification system if there is channel mismatch. In practice, we may not be able to detect the exact channel, and even if we could detect it, it is hard to record the suspects' voice through the detected channel.

In this research, we use feature enhancement methods to improve the speaker identification rate at various environments such as channels. We transform the original features which are extracted from given utterances to be robust against the environments changes. However it is very hard to select a particular channel compensation or feature enhancement method when we do not know test environment, because each feature enhancement method is optimal only for some specific environments. Therefore, we try to find a way of combining various feature enhancement methods.

We transform MFCC (mel-frequency cepstral coefficients) features by various feature enhancement methods and combine separate classifiers (speaker identification systems) which are trained by the transformed features. We use PCA (principal component analysis) [2], LDA (linear discriminant analysis) [2], GKPCA (greedy kernel principal component analysis) [1] and KMDA (kernel multimodal component analysis) [3] as feature enhancement methods.

The remainder of this paper is organized as follows. Section 2 describes feature enhancement methods used in this research. In section 3, the ensemble system using these feature enhancement methods is proposed. Section 4 and 5 show the experimental results and conclusions respectively.

## 2 Feature Enhancement Methods

MFCC is the most widely used feature in speech and speaker recognition domain. In this research, various feature enhancement methods are used to transform MFCC to robust features. Fig. 1 shows the process of each feature enhancement methods.

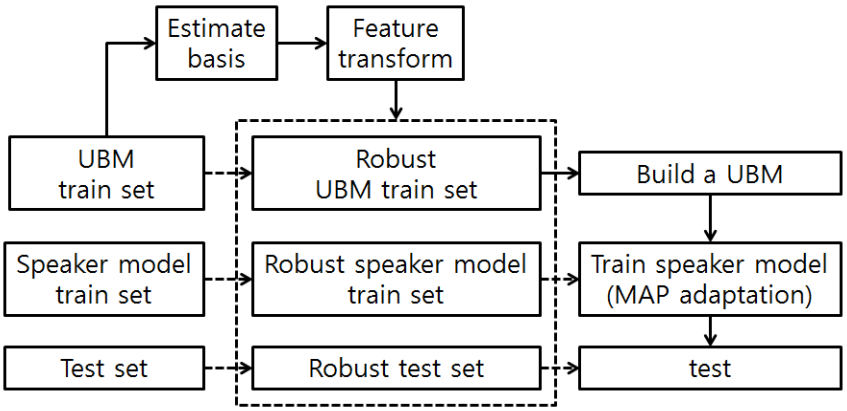


**Fig. 1.** The feature enhancement method.

First, the basis is estimated from the training set for UBM (universal background model) [5]. UBM is a model that represents the characteristics of general persons' voice, and is used as the baseline for speaker verification. The whole MFCC features are transformed into robust features by the estimated basis. Robust features are used for building a UBM, speaker model training and testing. Speaker models are adapted from the UBM using MAP adaptation [5].

The methods that we used to enhance the features are: PCA (principal component analysis) [2], LDA (linear discriminant analysis) [2], GKPCA (greedy kernel principal component analysis) [1] and KMDA (kernel multimodal component analysis) [3]. PCA is an orthogonal transformation of the coordinate system which maximizes the scatter (variance) of the whole data (feature vectors). If PCA is applied to MFCC features, the speaker identification rate can be improved. However, PCA cannot represent nonlinearly distributed data properly. Unlike PCA, LDA maximizes the information for classification. In this research, LDA basis is derived from each speaker's feature in the UBM training set.

KPCA (kernel principal component analysis) [8] can handle nonlinearly structured data using kernel method. The input data with nonlinear structure is mapped to a higher dimensional feature space by kernel method where the nonlinearly related variables can have linear relations. But the computational complexity and memory requirement are increased rapidly depending on the square of the number of samples which are used to estimate the KPCA basis. In speech and speaker recognition tasks, large number of features can be extracted from short utterances, therefore we cannot apply KPCA to speaker identification directly using present computing equipment.

GKPCA [1] use small number of features which represent the whole features. Greedy filtering selects a subset of the whole features with minimal representation error for relaxation of computational complexity and memory requirement. In this research, Gaussian RBF kernel function (equation (1)) is used ($\sigma$=21), and the size of the subset is 100.

$$k(\vec{x},\vec{y}) = \exp\left(-\frac{\|\vec{x}-\vec{y}\|^2}{2\sigma^2}\right) . \quad \textbf{(1)}$$

KMDA [3] maximizes the distances between the center of the whole training data and each center of sub-clusters in high dimensional feature space. In speaker recognition, each speaker's training features are separated into $K$ clusters. Fig. 2 shows an example which has two speakers A and B ($K$ = 2).

If number of speakers is N, then number of sub-clusters is $N \times K$. KMDA is similar to KPCA when $K$ is large. In this research, KMDA basis is derived from each speaker's features in UBM training set ($K$ = 4).

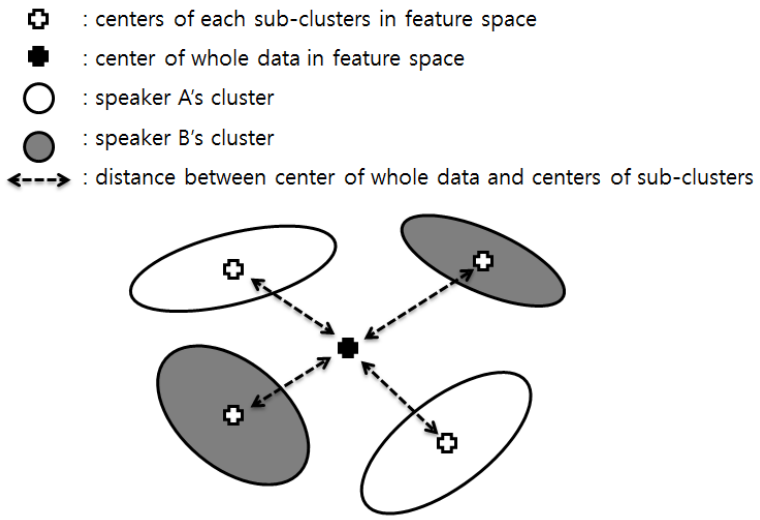


**Fig. 2.** The objective of KMDA.

## 3 The Ensembles of Feature Enhancement Methods

We construct an ensemble system using feature enhancement methods which are described in section 2.

### 3.1 Proposed Ensemble System

Fig. 3 shows an overview of the proposed ensemble system. First, the whole features are transformed using the new bases which are derived from the UBM training set using PCA, LDA, GKPCA and KMDA. In training and testing, each enhancement step is processed in parallel as described in section 2 (Fig. 2). UBM is trained using the transformed features, and speaker models are adapted from UBM using MAP adaptation. Finally, each speaker identification result is combined using various ensemble methods.

### 3.2 Combining the Classifiers

We use two kinds of combining schemes. First, majority voting [4] chooses the class which receives the highest number of votes from the classifiers. Each classifier can vote for only one class. Equation (2) shows how to calculate the majority voting when the ensemble is constructed with T classifiers and the number of classes ($\omega$) is C.

$$J = \arg\max_{j=1,C} \sum_{t=1}^{T} d_{t,j} \quad . \tag{2}$$

Where $d_{t,j}$ means the result of $t$-th classifier for $j$-th class $\omega_j$ (0 or 1). If the result of $t$-th classifier is $\omega_j$, $d_{t,j}$ is 1 and otherwise 0. $\omega_J$ is the combined result.

Another combining scheme, the Borda count determines the winner by giving each candidate a certain number of points corresponding to the position in which it is ranked according to log likelihood by each voter. In this research, each 5-best classes get from 5 to 1 points from each classifier. The equation of Borda count is the same as (2), except that $d_{t,j}$ is the score from $t$-th classifier for $j$-th class.

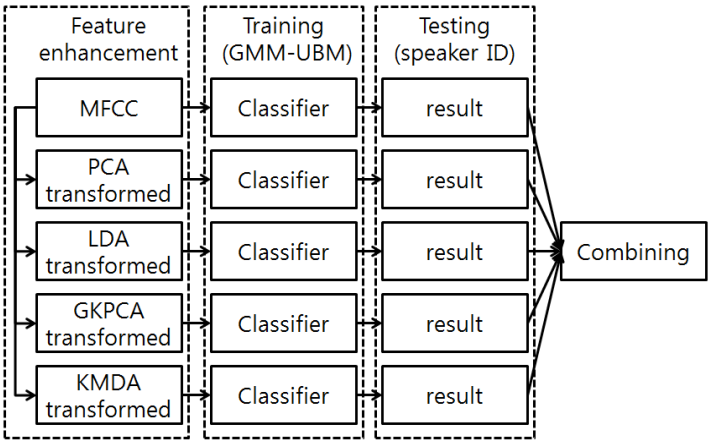


**Fig. 3.** The ensemble of feature enhancement methods.

## 4 Experimental Results

### 4.1 Database

To evaluate the proposed system in various environments, three types of speech corpora are used: ETRI PC DB, ETRI phone DB, and ETRI cellular phone DB. The speakers of the corpora are divided into three groups according to the session terms: 'WEEK', 'MONTH' and 'SEASON'. Experiments are performed for each database in parallel. To build a UBM, 'MONTH' speakers' 10 utterances are used as training set (first month – first try). For training speaker models and testing, 'WEEK' speakers' 10 utterances are used. (training: first week – first try, test with the same session: first

week – third try, test with different session: third week – first try). The scripts for the 10 utterances in UBM training set, speaker model training set and test set are the same (text-dependent). In PC DB, the number of speakers is 100. In phone and cellular phone DB, UBM training set consists of 101 speakers and speaker model training set consists of 104 speakers.

### 4.2 Speaker Model Training

The GMM-UBM is used for speaker model training. To build a UBM, GMMs with 32 Gaussian components are trained. The number of Gaussian components is started from one and increased double. In each of 1, 2, 4, 8 and 16 components, model parameters are trained once. For the model with 32 components, model parameters are trained 10 times. The speaker models are adapted up to three times from UBM using MAP adaptation with each speaker's training set ($\tau$=1).

### 4.3 Feature Extraction and Enhancement

G.712 channel simulation is applied to the test utterances. CAR noise in Aurora2 DB is added at SNR 20dB. FaNT is used for channel simulation and adding noise. The sampling rate of all DB is set to 8kHz. 20 MFCC coefficients and log energy are extracted from the given speech. Silences are removed by energy based method in feature level. CMS (cepstral mean subtraction) and variance normalization are used as channel compensation methods. After channel compensation, features are transformed by PCA, LDA, GKPCA and KMDA.

### 4.4 Experimental Results

Table 1 shows the average speaker identification rate in various 24 type of environments which include three types of corpuses (PC, phone and cellular phone DB), two types of channel mismatch (same channel and different channel simulated by G.712), two types of noises (clean and noisy with CAR noise 20db) and two types of sessions (same and different session).

'MFCC', 'CMS' and 'VAR' in Table 1 and Fig. 4 mean MFCC features and their compensated features using CMS and variance normalization respectively. '[original feature]-[feature enhancement method]' notation shows the result using a single feature enhancement method. 'VOTE' and 'BORDA' mean combining classifiers with majority voting and Borda count, respectively. '[channel compensation method]-[combining scheme]' notation shows the results with the ensemble of feature enhancement methods (five types of classifiers in the same channel compensation method). 'TOTAL-[combining scheme]' notation shows the results of the ensembles of channel compensation and feature enhancement methods (15 types of classifiers with the combinations of three channel compensation methods and five feature enhancement methods).

**Table 1.** Average speaker identification rates.

| Feature | Same channel | Different channel | Clean | Noisy | Same session | Different session | Total average |
|---|---|---|---|---|---|---|---|
| MFCC | 58.33 | 39.97 | 67.17 | 31.14 | 57.46 | 40.84 | 49.15 |
| CMS | 55.95 | 53.69 | 71.61 | 38.03 | 63.81 | 45.83 | 54.82 |
| VAR | 62.97 | 60.69 | 73.36 | 50.30 | 71.33 | 52.32 | 61.83 |
| MFCC-PCA | 57.45 | 29.77 | 61.78 | 25.45 | 50.15 | 37.08 | 43.61 |
| CMS-PCA | 58.12 | 56.57 | 78.51 | 36.18 | 64.66 | 50.03 | 57.35 |
| VAR-PCA | 65.61 | 62.98 | 76.32 | 52.26 | 73.97 | 54.61 | 64.29 |
| MFCC-GKPCA | 58.19 | 34.39 | 63.69 | 28.88 | 54.12 | 38.45 | 46.29 |
| CMS-GKPCA | 59.55 | 57.68 | 78.00 | 39.23 | 66.93 | 50.30 | 58.61 |
| VAR-GKPCA | 65.46 | 63.51 | 76.64 | 52.33 | 74.20 | 54.78 | 64.49 |
| MFCC-KMDA | 58.09 | 37.49 | 66.17 | 29.41 | 55.76 | 39.81 | 47.79 |
| CMS-KMDA | 60.45 | 58.99 | 78.83 | 40.61 | 68.29 | 51.15 | 59.72 |
| VAR-KMDA | 66.26 | 64.71 | 77.37 | 53.59 | 75.58 | 55.38 | 65.48 |
| MFCC-LDA | 56.05 | 27.95 | 59.31 | 24.70 | 48.51 | 35.50 | 42.00 |
| CMS-LDA | 59.86 | 57.35 | 77.27 | 39.94 | 67.22 | 50.00 | 58.61 |
| VAR-LDA | 65.21 | 63.46 | 77.57 | 51.10 | 73.73 | 54.94 | 64.33 |
| MFCC-VOTE | 59.46 | 35.60 | 64.96 | 30.10 | 55.25 | 39.80 | 47.53 |
| MFCC-BORDA | 61.27 | 59.62 | 79.26 | 41.63 | 68.76 | 52.13 | 60.44 |
| CMS-VOTE | 67.69 | 66.18 | 78.79 | 55.09 | 76.63 | 57.25 | 66.94 |
| CMS-BORDA | 59.09 | 33.97 | 63.96 | 29.10 | 53.89 | 39.16 | 46.53 |
| VAR-VOTE | 60.43 | 59.38 | 78.59 | 41.22 | 68.06 | 51.75 | 59.90 |
| VAR-BORDA | 67.59 | 65.73 | 78.35 | 54.97 | 76.48 | 56.84 | 66.66 |
| TOTAL-VOTE | 69.10 | 67.21 | **82.01** | 54.29 | 77.51 | 58.80 | 68.15 |
| TOTAL-BORDA | **70.51** | **67.49** | 80.72 | **57.28** | **78.20** | **59.81** | **69.00** |

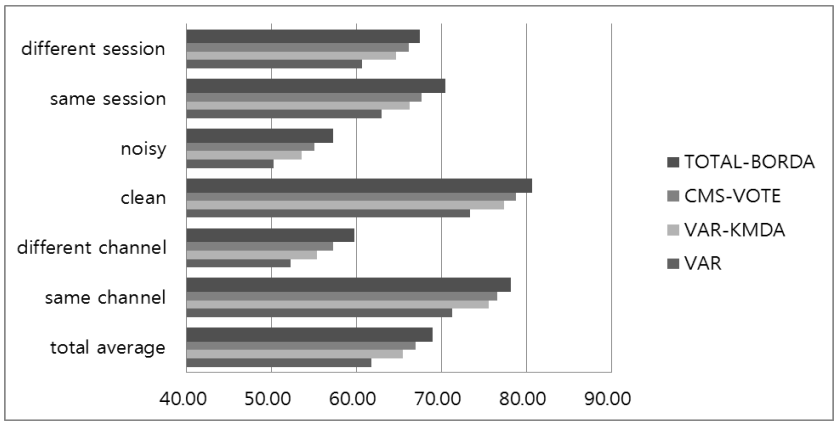


**Fig. 4.** Speaker identification rates of major algorithms.

Fig. 4 shows major results in Table 1. If we do not construct classifier ensemble, variance normalization and KMDA ('VAR-KMDA') shows the highest speaker identification rate. Otherwise, if we construct ensembles with five types of feature

enhancement methods, the speaker identification rate increases to higher than 'VAR-KMDA' in some cases. Experimental results show the highest speaker identification rate when we construct an ensemble with 15 classifiers (three types of channel compensation methods and five types of feature enhancement methods) and combine them using Borda count ('TOTAL-BORDA').

## 5 Conclusion

In certain domains such as digital forensic investigation, the recording environments of train and test speech may be different, and the users of speaker identification systems may not have the prior knowledge about such environmental mismatch. In this case, we cannot guarantee the reliability of the speaker identification system. In this research, we tried to find a way of combining various feature enhancement methods to improve speaker identification rate at various environments. The proposed ensemble system is constructed with several classifiers. Each classifier is trained using enhanced features which are transformed with different channel compensation and feature enhancement methods. In the experimental results, the proposed method shows the highest average speaker identification rate 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. Kim, M-S., Yang, I-H., Yu, H-J.: Speaker Identification using Greedy Kernel PCA. Malsori, pp. 105--116 (2008)
2. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification, John Wiley & Sons (2001)
3. Kim, M-S., Yang, I-H., Yu, H-J.: Kernel Multimodal Discriminant Analysis for Speaker Verification. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4498--4501 (2010)
4. Polikar, R.: Ensemble based Systems in Decision Making. Circuits and Systems Magazine (2006)
5. 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)
6. 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)
7. Scholkopf, B., Smola, A., Muller, K-R.: Kernel Principal Component Analysis. In Proceedings of International Conference on Artificial Neural Networks, pp. 583--588 (1997)
8. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis, Cambridge University Press (2004)