# Maximizing Distance between GMMs for Speaker Verification Using Particle Swarm Optimization

Min-Seok Kim, IL-Ho Yang and Ha-Jin Yu*
*School of Computer Science, University of Seoul, Korea*
*hjyu@uos.ac.kr*

## Abstract

*In this paper, we propose a feature transformation method to maximize the distances between the Gaussian mixture models for speaker verification. The feature transformation matrix is optimized by using particle swarm optimization. We evaluate the transformation using YOHO speech data, and the transformation is applied to some speakers who give poor performance. As the result, the overall equal error rate is reduced to 1.71% from 1.97% of the baseline.*

## 1. Introduction

Speaker verification 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 a good feature for speaker discrimination, building good speaker models, and normalizing the scores. Currently, melcepstral coefficients and Guassian mixture models[1] are the most general approach for features and models respectively. In this research, we aim at improving speaker verification performance by making a better recognition model. We can reduce the error by enlarging the distances between the models. We propose to make the models with larger distances by transforming the feature. The optimized transformation that maximizes the distances between the models can be found by using an optimization technique called particle swarm optimization (PSO). Since PSO can find better solutions incrementally, we can find an optimized feature transformation that enlarges the distances between the models.

In the next section, we will introduce GMMs and the distance measure for GMMs based on Kullback-Leibeler divergence. Section 3 describe the proposed optimization method, and section 4 reports experimental evaluation on speaker verification. Finally, section 5 summarizes the conclusions drawn form this study.

## 2. Distances between the GMMs

### 2.1. Gaussian Mixture Models

Gaussian mixture models (GMMs) [1] is the most prominent approach in speaker verification systems. The GMMs are represented by weighted sum of Gaussian probabilistic density functions of feature vectors extracted from the voice. For a *D*-dimensional feature vector $\mathbf{x}$, the Gaussian mixture density for speaker *s* is defined as

$$p(\mathbf{x} \mid \lambda_s) = \sum_{i=1}^{M} w_i^s g_i^s(\mathbf{x}) \tag{1}$$

$$g_i^s(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} |\mathbf{\Sigma}_i^s|^{1/2}} \times \exp\left\{ -\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i^s)^T \mathbf{\Sigma}_i^{s^{-1}} (\mathbf{x} - \boldsymbol{\mu}_i^s) \right\} \tag{2}$$

$$\lambda_s = \left\{ w_i^s, \boldsymbol{\mu}_i^s, \mathbf{\Sigma}_i^s \right\} i = 1, \ldots, M \tag{3}$$

where $M$ is the number of mixtures and $w_i^s$, $\boldsymbol{\mu}_i^s$, $\mathbf{\Sigma}_i^s$ are weight, mean and covariance of each component respectively. The mixture weight, $w_i^s$, satisfy the constraint $\sum_{i=1}^{M} w_i^s = 1$. $\lambda_s$ is the parameters of speaker *s*'s density model. For efficiency, many applications use diagonal covariance matrix $\mathbf{\Sigma}_i^s$ [5].

### 2.2. Kullback-Leibeler Divergence

Kullback-Leibeler (KL) divergence is the natural way to define a distance measure between probability distributions [2][3]. KL-divergence between two distributions *f* and *g* is defined as

$$KL(f \,\|\, g) = \int f(x) \log \frac{f(x)}{g(x)} dx \,. \tag{4}$$

However, the equation (4) is difficult to apply to GMMs because GMM has two or more distributions. In this paper we use matching based approximations method which is proposed in [2]. We can compute distance between two speaker models $\lambda_A$, $\lambda_B$ using matching function[2]

$$KL(\lambda_A \,\|\, \lambda_B) \approx \sum_{i=1}^{M} \left\{ w_i \min_{j=1}^{M} KL\left(g_i^A \,\|\, g_j^B\right) \right\} \tag{5}$$

and KL-divergence between the Gaussian component $g_1 = N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and $g_2 = N(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ is :

[*] Corresponding author

978-0-7695-3304-9/08 $25.00 © 2008 IEEE
DOI 10.1109/ICNC.2008.820

$$\frac{1}{2}\left(\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}+Tr(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-d\right)\cdot \tag{6}$$

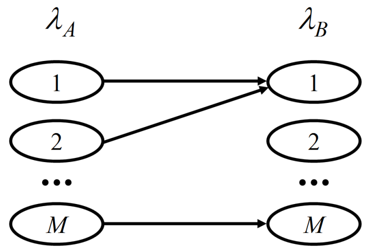


**Figure 1. A matching function between the Gaussian components of two GMMs**

Figure 1 shows a matching function between the Gaussian components of two GMMs (see equation 5). This example illustrates that the first and second components of $\lambda_A$ are the closest to the first component of $\lambda_B$, and also the $M$th component is the closest to the $M$th component of $\lambda_B$. In this paper, we use symmetric version of KL divergence expressed as $KL(\lambda_A,\lambda_B)=KL(\lambda_A \| \lambda_B)+KL(\lambda_B \| \lambda_A)$.

## 3. Maximizing Distance between two GMMs

In universal background model (UBM) based speaker verification system [4] which is the most prominent method, the verification task can be started as a hypothesis test between two hypotheses:

$$\begin{aligned} H_0 &: \ Y \text{ is from the hypothesized speaker } S \\ H_1 &: \ Y \text{ is } \textit{not} \text{ from the hypothesized speaker } S\,. \end{aligned} \tag{7}$$

The optimum test to decide between theses two hypotheses is a likelihood ratio test given by

$$\frac{p(Y\,|\,H_0)}{p(Y\,|\,H_1)}\begin{cases}\geq\theta & \text{accept } H_0\\ <\theta & \text{reject } H_0\end{cases}, \tag{8}$$

where $p(Y\,|\,H_0)$ is the probability density function for the hypothesis $H_0$ which is computed for observed speech segment $Y$ (the likelihood of $Y$ for true speaker model) and $p(Y\,|\,H_1)$ is the likelihood of $Y$ for universal background model.

As mentioned in section 2.1, many applications use diagonal covariance matrix for computational efficiency [5]. However, diagonal covariance matrix implies the assumption that the feature elements are independent. To decorrelate the feature, some methods are applied in feature-space such as principal component analysis (PCA) [6] and linear discriminant analysis (LDA) [7]. However, those methods do not guarantee a good modeling for the recognition. In this paper, we propose a method to maximize the log likelihood ratio between the speaker model and UBM. The transformation is optimized by using particle swarm optimization (PSO) which is proposed by Kennedy and Eberhart [9] and has been used to solve many optimization problems. In PSO, each particle moves in the $d$-dimensional search space with a velocity according to its own previous best solution and its group's previous best solution as follows

$$P^N=P^C+c_0(P^C-P^B)+c_1R_1(P^P-\mathbf{p})+c_2R_2(\mathbf{g}-P^C) \tag{9}$$

where

$P^C$ : current particle
$P^B$ : before particle
$P^N$ : next particle
$\mathbf{p}$ : own previous best particle
$\mathbf{g}$ : group's best particle
$R_1$ : $d$ - dimensional random variables $(0\sim 1)$
$R_2$ : $d$ - dimensional random variables $(0\sim 1)$
$c_0,c_1,c_2$ : positive constants

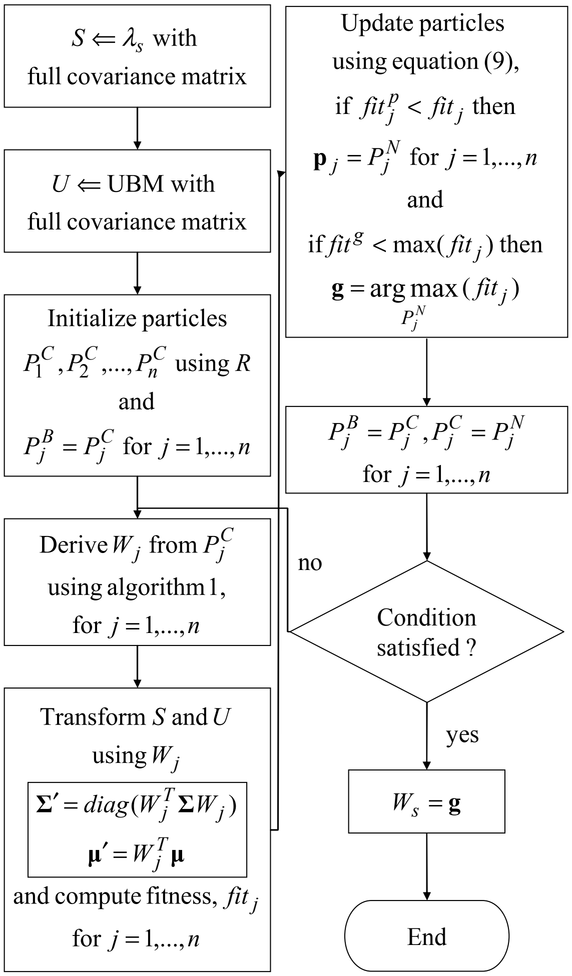
$P^C$, $P^B$, $P^N$, $\mathbf{p}$ and $\mathbf{g}$ are $d$-dimensional vectors, $R_1$ and $R_2$ also $d$-dimensional vectors whose elements are random variables between 0 and 1. Figure 2 shows PSO algorithm with $n$ particles. In our PSO algorithm, the rotation matrix, $W_j$ is derived by rotating coordinate basis using rotation matrix, $R$ as follows:

$$R=\begin{bmatrix} 0 & \theta_{1,2} & \cdots & \theta_{1,i} & \cdots & \theta_{1,j} & \cdots & \theta_{1,d-1} & \theta_{1,d} \\ & 0 & \ddots & \theta_{2,i} & \cdots & \theta_{2,j} & \cdots & \theta_{2,d-1} & \theta_{2,d} \\ & & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ & & & 0 & \ddots & \theta_{i,j} & \cdots & \theta_{i,d-1} & \theta_{i,d} \\ \vdots & & & & \ddots & \vdots & \ddots & \vdots & \vdots \\ & & & & & 0 & \ddots & \theta_{j,d-1} & \theta_{j,d} \\ & & & & & & \ddots & \vdots & \vdots \\ & & & & & & & 0 & \theta_{d-1,d} \\ 0 & & & & \cdots & & & & 0 \end{bmatrix}\cdot \tag{10}$$

The algorithm of rotating coordinate basis is described in algorithm 1. For the fitness of the optimization we use the distance measure $KL(diag(S),diag(U))$ which evaluates the distance between diagonalized speaker model and UBM transformed by $W$. Finally, we can expect that the distance between speaker model and UBM is maximized.

**Algorithm 1. Deriving the transformation matrix using *R***

```
Step 1: Let $W$ be the identity matrix (coordinate basis)
        with $W \in \Re^{d\times d}$.
Step 2:
        for $i$=1,2,…,$d$-1
            for $j$=$i$,$i$+1,…,$d$
                Rotate the plane formed with $i$-th axis and
                $j$-th axis centered at the origin by $\theta_{i,j}$ in $R$
```



$S \Leftarrow \lambda_s$ with full covariance matrix
$U \Leftarrow$ UBM with full covariance matrix
Initialize particles $P_1^C, P_2^C, ..., P_n^C$ using $R$ and $P_j^B = P_j^C$ for $j = 1,...,n$
Derive $W_j$ from $P_j^C$ using algorithm 1, for $j = 1,...,n$
Transform $S$ and $U$ using $W_j$
$\mathbf{\Sigma}' = diag(W_j^T \mathbf{\Sigma} W_j)$
$\boldsymbol{\mu}' = W_j^T \boldsymbol{\mu}$
and compute fitness, $fit_j$ for $j = 1,...,n$
Update particles using equation (9), if $fit_j^p < fit_j$ then $\mathbf{p}_j = P_j^N$ for $j = 1,...,n$ and if $fit^g < \max(fit_j)$ then $\mathbf{g} = \arg\max_{P_j^N} (fit_j)$
$P_j^B = P_j^C, P_j^C = P_j^N$ for $j = 1,...,n$
no
Condition satisfied ?
yes
$W_s = \mathbf{g}$
End

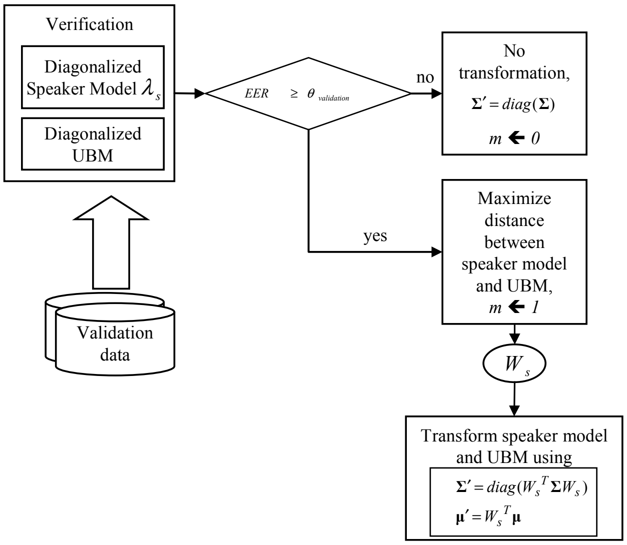
**Figure 2. Our PSO algorithm to maximize the distance between speaker model and UBM**

## 4. Experimental Results

We used the YOHO database which consists of 138 speakers prompted to read combination lock phrases, for example, "67 34 85." The features were derived using 12th order MFCC analysis from the audio recording and deltas computed making up a twenty four dimensional feature vector. We used 20 speakers, labeled from 101 to 120, for training and testing (as *true speaker*) and 18 speakers, labeled from 121 to 140, as *imposter*. Furthermore the rest of the speakers are used for universal background model. The frames of data corresponding to silence were removed from the utterances using energy threshold. The YOHO database has 'enroll' and 'verify' mode. The 'enroll' consists of 4 sessions with 24 utterances each session. And the 'verify' consists of 10 sessions with 4 utterances each session. Table 1 shows the settings for the training and test data.

**Table 1. Database setup**

| | |
|---|---|
| Training data for speaker model | *true speaker*'s 8 utterances in first session of 'enroll' mode |
| Training data for UBM | All utterances in 'enroll' mode of speakers who are designated for UBM |
| Validation data | 1. *true speaker*'s 24 utterances in second session of 'enroll' mode<br>2. *imposter*s' 180 utterances in second session of 'enroll' mode (10 utterances each *imposter*) |
| Test data | 1. *true speaker*'s 88 utterances (third and fourth sessions of 'enroll' and all 'verify' session, 24+24+40)<br>2. *imposter*s' 180 utterances in 'verify' (10 utterances each *imposter*) |



Verification
Diagonalized Speaker Model $\lambda_s$
Diagonalized UBM
Validation data
$EER \geq \theta_{validation}$
no
No transformation,
$\mathbf{\Sigma}' = diag(\mathbf{\Sigma})$
$m \leftarrow 0$
yes
Maximize distance between speaker model and UBM,
$m \leftarrow 1$
$W_s$
Transform speaker model and UBM using
$\mathbf{\Sigma}' = diag(W_s^T \mathbf{\Sigma} W_s)$
$\boldsymbol{\mu}' = W_s^T \boldsymbol{\mu}$

**Figure 3. The proposed speaker verification system (training step)**

In validation test, total EER (equal error rate) is 3.14% and labeled 102, 104, 111, 112 speakers' EERs are upper than 4.0% (in this paper, $\theta_{validation}$ is 4.0). So these speakers are transformed to maximize distance between speaker model and UBM. Through our proposed algorithm, the total EER of validation test becomes 2.73%. In verification test, total EER of baseline that is not transformed and proposed algorithm are 1.97%, 1.71% respectively. Table 2 and figure 5 show the result of speaker verification. In addition, we experiment using linear discriminant analysis (LDA) instead of our proposed algorithm. But the EER of speaker verification is 2.21% and poorer than baseline and our proposed algorithm.

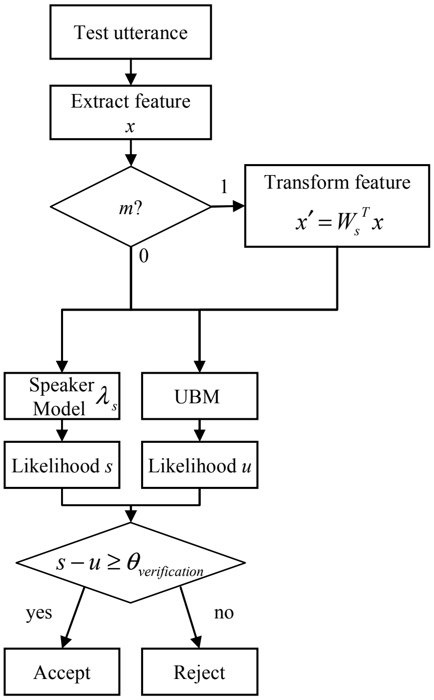


**Figure 4. The proposed speaker verification system (test step)**

**Table 2. Equal error rate of speaker verification**

| | EER (%) |
|---|---|
| Baseline | 1.97 |
| Proposed algorithm | 1.71 |

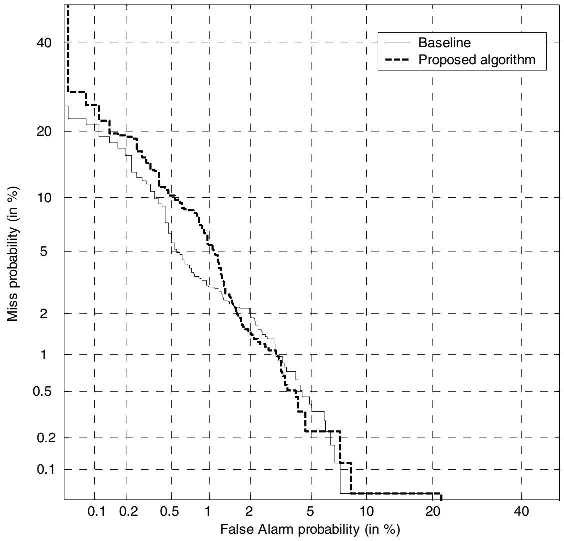


**Figure 5. DET curves for baseline and proposed algorithm**

## 5. Conclusions

We proposed a feature transformation method based on particle swarm optimization for speaker verification. The optimized feature space is sought by rotating the axes of the base feature space to make the distances of the recognition models large enough for discrimination. The distances between the Gaussian mixture models are evaluated based on Kullback-Leibeler divergence. The optimal degree of axes rotation is found by using particle swarm optimization which can find the better solution incrementally. The proposed method is evaluated using YOHO database, and the feature transformation is applied to the speakers who show lower performances compared to others. As the result, the overall equal error rate is reduced to 1.71% from 1.97% of the baseline. For further research, we are considering to find a deterministic method for the optimized feature transformation.

## 6. References

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

[2] J. Goldberger and H. Aronowitz, “A distance measure Between GMMs Based on the Unscented Transform and its Application to Speaker Recognition,” In Proc. Interspeech 2005, pp. 1985-1988, 2005.

[3] S. Kullback, *Information theory and statistics*, Dover Publications, New York, 1968.

[4] D. Reynolds, T. Quatieri and R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” Digital Signal Processing, vol. 10, no. 1-3, pp. 93-112, 2000.

[5] X. Zhou, Z. Yao and B. Dai, “Improved covariance modeling for GMM in speaker identification,” INTERSPEECH-2005, pp. 3113-3116, 2005.

[6] Z. Wanfeng, Y. Yingchun, W. Zhaohui and Sang Lifeng, “Experimental evaluation of a new speaker identification framework using PCA,” IEEE International Conference on Systems, Man and Cybernetics, Vol. 5, pp. 4147-4152, 5-8 October 2003.

[7] Q Jin, A Waibel, “Application of LDA to speaker recognition,” Proc. ICSLP-00, vol. 2, pp. 250-253, Beijing, China, October 2000.