# 대각공분산 GMM에 최적인 선형변환을 이용한 강인한 화자식별*

김민석(서울시립대), 양일호(서울시립대), 유하진(서울시립대)

<차 례>



<Abstract>

## Robust Speaker Identification Using Linear Transformation Optimized for Diagonal Covariance GMM

Min-Seok Kim, Ha-Jin Yu

We have been building a text-independent speaker recognition system that is robust to unknown channel and noise environments. In this paper, we propose a linear transformation to obtain robust features. The transformation is optimized to maximize the distances between the Gaussian mixtures. We use rotation of the axes, to cope with the problem of scaling the transformation matrix. The proposed transformation is similar to PCA or LDA, but can achieve better result in some special cases where PCA and LDA can not work properly. We use YOHO database to evaluate the proposed method and compare the result with PCA and LDA. The results show that the proposed method outperforms all the baseline, PCA and LDA.

* Keywords: Speaker recognition, Speaker identification, Feature transformation, PCA, LDA.

* 이 논문은 2006년도 서울시립대학교 학술연구조성비에 의하여 연구되었음.

# 1. 서 론

최근 로봇 기술의 발달과 더불어 화자인식기의 수요가 늘고 있다. 로봇에 탑재되는 화자인식기는 일반 가정, 사무실, 공장 등 여러 장소에서 이용되기 때문에 음성이 녹음되는 순간의 주변 잡음 크기 및 종류를 예측할 수 없다. 로봇에 탑재되는 화자인식기를 실용화하기 위해서는 언제 어떤 상황에서도 강인한 인식 성능을 보여야한다. 환경 변화에 강인한 인식기를 만들기 위한 가장 간단한 방법은 다양한 환경에서 데이터를 수집하여 화자 모델을 생성하는 것이다. 그러나 이 방법은 비슷한 환경에서 수집된 음성이 화자 모델에 포함되어 있지 않은 경우에는 인식 성능을 보장하기 어렵다. 또한 다양한 환경에서 동일한 화자의 음성을 수집하기는 쉽지 않다. 제한된 환경에서 데이터를 수집할 수밖에 없는 이런 제약조건 아래서 강인한 화자인식을 위해, 수집된 음성에서 화자의 정보를 많이 포함하고 환경적 영향을 줄이는 화자 특징 변환을 이용한 연구가 진행되었다[1][2].

본 연구에서는 mel-frequency cepstral coefficient (MFCC)로 추출된 특징 벡터에서 환경적 특성보다 화자의 특성을 더 잘 나타내는 특징을 구하기 위해 대각공분산(diagonal covariance) Gaussian mixture model (GMM)에 최적인 선형변환을 제안하고 이를 화자식별에 적용한다.

선형변환을 이용한 기존의 강인한 화자인식 방법으로는 주성분 분석을 이용한 방법[1][2]과 선형판별 분석을 이용한 방법[3], 그리고 주성분 분석과 선형판별 분석의 결합[7]을 이용하는 방법이 있다. 본 논문에서 제안한 GMM에 최적인 선형변환을 이용한 화자식별 방법은 기존 방법과 다르게 변환된 특징 벡터가 대각공분산 GMM으로 모델링된다는 정보를 이용한다. 본 논문에서는 수치적인 최적화 문제를 풀이하는데 이용되는 알고리즘인 particle swarm optimization (PSO)[6]을 통해 GMM에 최적인 선형변환 행렬을 찾는다.

본 논문의 구성은 다음과 같다. 2장에서는 주성분 분석(principal component analysis)과 선형판별 분석(linear discriminant analysis), 화자식별 시스템에 가장 일반적으로 이용되는 GMM 및 최적화에 사용되는 방법인 PSO를 소개하고, 3장에서는 본 논문에서 제안한 대각공분산 GMM에 최적인 선형변환을 이용한 화자식별 방법을 소개한다. 그리고 4장에서는 실험 환경과 주성분 분석, 선형판별 분석, 제안한 방법의 실험결과를 제시하고 마지막으로 5장에서 결론을 맺는다.

# 2. 기존의 방법

## 2.1. 주성분 분석

주성분 분석[4]은 다차원 특징 공간에서 각 차원의 상관관계를 줄이는 독립적인 축을 구하고, 그 축으로 특징 벡터를 사상시켜 낮은 차원으로 축소시키는 방법이다. 주성분 분석에서 각 차원의 상관관계를 나타내는 공분산, $S_\Sigma$은 다음과 같이 구한다.

$$S_\Sigma = \frac{1}{N}\sum_{i=1}^{N}(\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m})^T \tag{1}$$

여기서 $\vec{x}_i$는 $i$번째 특징 벡터, $\vec{m}$은 전체 특징 벡터의 평균, $N$은 전체 특징 벡터의 개수이다. 주성분 분석의 변환 행렬 $W_{PCA}$는 $S_\Sigma$의 고유벡터를 구하고 고유값이 큰 순으로 $d(\leq D)$개를 선택하여 만든다.

$$W_{PCA} = [e_1, e_2, \ldots, e_d], \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \tag{2}$$

여기서 $e$는 고유벡터이고 $\lambda$는 고유값이다. 본 연구에서는 특징 벡터의 차원을 줄이는 것이 목적이 아니므로 모든 고유벡터($d = D$)를 이용한다.

## 2.2. 선형판별 분석

선형판별 분석[4]은 주성분 분석과 더불어 대표적인 특징 벡터 차원 축소 기법 중 하나이다. 선형판별 분석의 목적은 클래스간 분산($S_B$: between-class scatter)과 클래스내 분산($S_W$: within-class scatter)의 비율($S_B/S_W$)을 최대화하는 것이다. $S_B$와 $S_W$는 다음과 같이 구한다.

$$S_B = \sum_{j=1}^{C} N_j(\vec{m}_j - \vec{m})(\vec{m}_j - \vec{m})^T \tag{3}$$

$$S_W = \sum_{j=1}^{C}\sum_{i=1}^{N_j}(\vec{x}_i^{(j)} - \vec{m}_j)(\vec{x}_i^{(j)} - \vec{m}_j)^T \tag{4}$$

여기서 $\vec{m}$은 전체 특징 벡터의 평균, $\vec{m}_j$는 $j$번째 클래스 특징벡터의 평균, $N_j$는 $j$번째 클래스의 특징 벡터 개수, $C$는 클래스의 개수, $\vec{x}_i^{(j)}$는 $j$번째 클래스의 $i$번째 특징벡터이다. 선형판별 분석을 위한 변환행렬 $W_{LDA}$는 $S_W^{-1}S_B$의 고유벡터를 구하고 고유값이 큰 순으로 $d(\leq D)$개를 선택하여 만든다.

$$W_{LDA} = [e_1, e_2, \ldots, e_d], \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \tag{5}$$

여기서 $e$는 고유벡터이고 $\lambda$는 고유값이다. 본 연구에서는 특징 벡터의 차원을 줄이는 것이 목적이 아니므로 주성분 분석과 마찬가지로 모든 고유벡터($d=D$)를 이용한다.

## 2.3. Gaussian Mixture Model

GMM[5]은 문장 독립 화자 식별 시스템에서 가장 많이 사용되는 모델링 방법이다. GMM으로 모델링된 화자모델은 특징 벡터의 평균과 공분산을 가진 가우시안 확률분포함수(pdf) 혼합(mixture)수 $M$ 개의 선형 결합으로 나타낸다.

$$p(\vec{x}|\lambda)=\sum_{i=1}^{M} w_i g_i(\vec{x}), \qquad \sum_{i=1}^{M} w_i = 1 \tag{6}$$

$$g_i(\vec{x})=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)^T\right\} \tag{7}$$

여기서 $\mu_i$는 $i$번째 pdf의 평균이고, $\Sigma_i$는 $i$번째 pdf의 공분산이다. $\lambda$는 화자 모델 파라미터를 나타낸다.

$$\lambda=\left(w_i,\vec{\mu}_i,\Sigma_i\right), \quad \text{for } i=1,2,...,M \tag{8}$$

모델의 학습에는 expectation-maximization (EM) 알고리즘을 이용한다. 화자식별은 $T$개의 음성 특징 벡터를 $S$명의 식별 대상에 대하여 각각의 유사도를 구하고, 유사도가 가장 큰 $\hat{S}$를 선택하면 된다.

$$\hat{S}=\arg\max_{1\le k\le S}\sum_{t=1}^{T}\log p(\vec{x_t}|\lambda_k) \tag{9}$$

본 연구에서는 계산량, 메모리 절감 및 소용량 데이터를 이용한 학습의 효율성을 위해 대각 공분산(diagonal covariance) GMM을 사용한다.

## 2.4. Particle Swarm Optimization

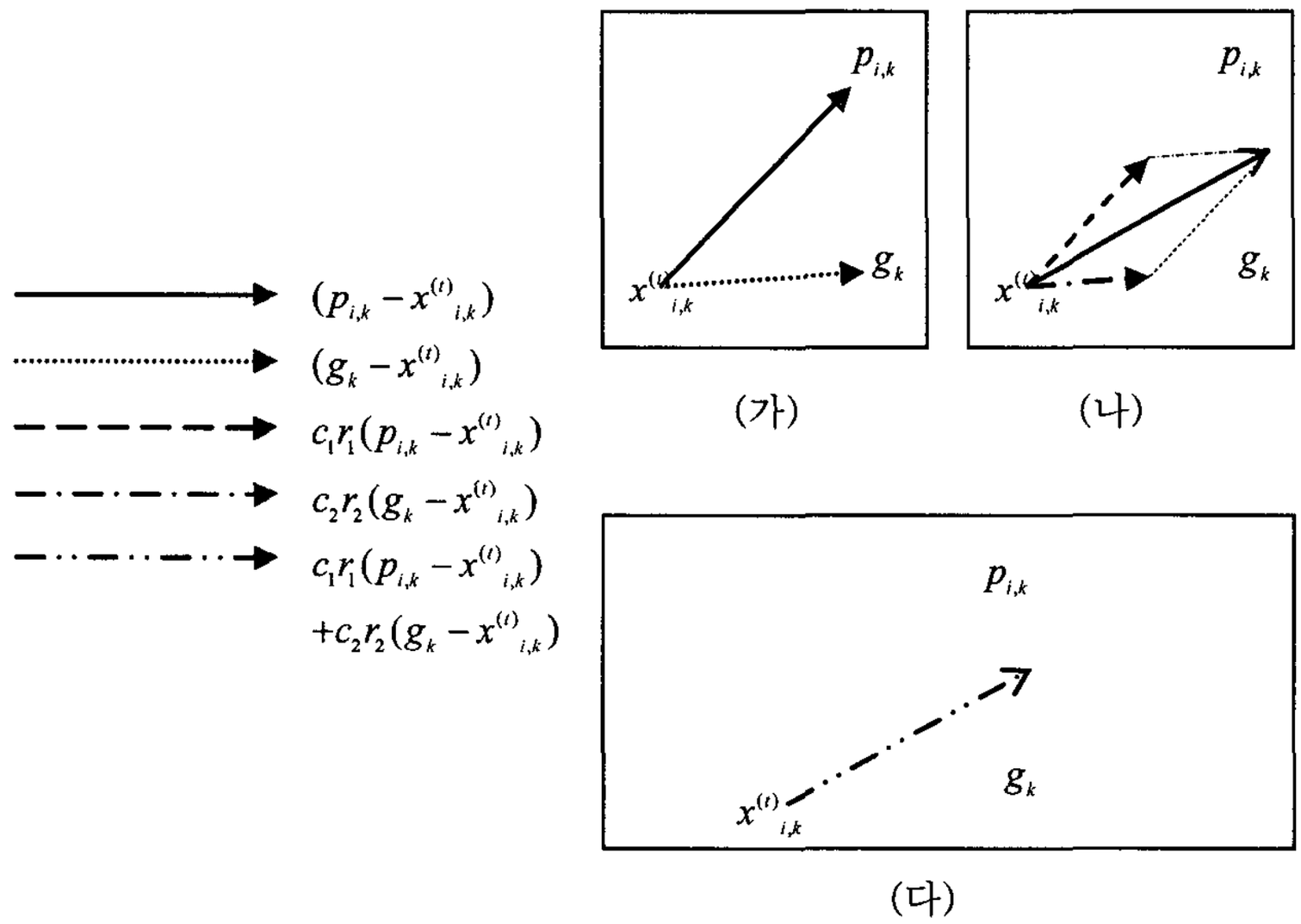
최적화 문제를 해결하는데 많이 이용되고 있는 PSO는 Kennedy와 Eberhart가 1995년에 제안하였다[6]. PSO에서 개체(Particle)는 개체 각각의 이전 세대에서 가장 좋은 값(previous best solution)과 전체 개체의 이전 세대에서 가장 좋은 값(global

best solution)을 따르는 속도로 탐색 공간을 이동한다. $d$차원 탐색공간에서 $t$시간의 $i$번째 개체가 $X_i^{(t)} = \{x_{i,1}^{(t)}, \ldots, x_{i,d}^{(t)}\}$이고, previous best solution은 $P_i = \{p_{i1}, \cdots, p_{id}\}$, global best solution은 $G = \{g_1, \cdots, g_d\}$일 때 각각의 개체에서 새로운 개체를 생성하는 식은 다음과 같다.

$$\Delta x_{i,k}^{(t+1)} \Leftarrow c_0 (x_{i,k}^{(t)} - x_{i,k}^{(t-1)}) + c_1 r_1 (p_{i,k} - x_{i,k}^{(t)}) + c_2 r_2 (g_k - x_{i,k}^{(t)}) \qquad (10)$$

$$x_{i,k}^{(t+1)} \Leftarrow x_{i,k}^{(t)} + \Delta x_{i,k}^{(t+1)} \qquad \text{for } k = 1, \cdots, d \qquad (11)$$

여기서 $c_0$, $c_1$, $c_2$는 양의 값을 갖는 상수이고, $r_1$과 $r_2$는 0과 1사이의 임의의 실수 값이다. <그림 1>은 PSO의 진행 과정을 보여준다. <그림 1>에서 (가)는 $(p_{i,k} - x_{i,k}^{(t)})$와 $(g_i - x_{i,k}^{(t)})$를 나타내며 (나)는 (가)의 각각에 $c_1 r_1$과 $c_2 r_2$가 가중된 결과를 보여준다. 가중치는 각각 $0 \le c_1 r_1 \le c_1$, $0 \le c_2 r_2 \le c_2$의 범위에 존재한다. (다)는 $c_1 r_1 (p_{i,k} - x_{i,k}^{(t)}) + c_2 r_2 (g_k - x_{i,k}^{(t)})$의 결과를 나타내며, 이전의 세대와의 차이에 $c_0$를 곱한 $c_0 (x_{i,k}^{(t)} - x_{i,k}^{(t-1)})$를 더하여 새로운 속도 $\Delta x_{i,k}^{(t+1)}$를 계산할 수 있다.




<그림 1> Particle swarm optimization 과정

목적함수 $J(\bullet)$를 최적화하는 PSO 알고리즘의 과정은 다음과 같다.

단계 1: $N$개의 개체를 임의의 값으로 초기화한다.
단계 2: 각 개체의 적합도(목적함수 $J(\bullet)$의 결과)를 구한다.
단계 3: 각 개체의 적합도가 previous best solution의 적합도보다 좋다면 previous best solution을 현재의 개체로 교체한다. 또한 현재의 모든 개체의 적합도 중에서 global best solution의 적합도보다 좋은 것이 있다면 global best solution을 그 개체로 교체한다.
단계 4: 식 (10)과 식 (11)을 이용하여 각 개체의 다음 세대를 생성한다.
단계 5: 종료 조건을 만족하면 종료하고, 그렇지 않으면 단계 2로 가서 반복한다.
단계 6: Global best solution을 결과로 취한다.

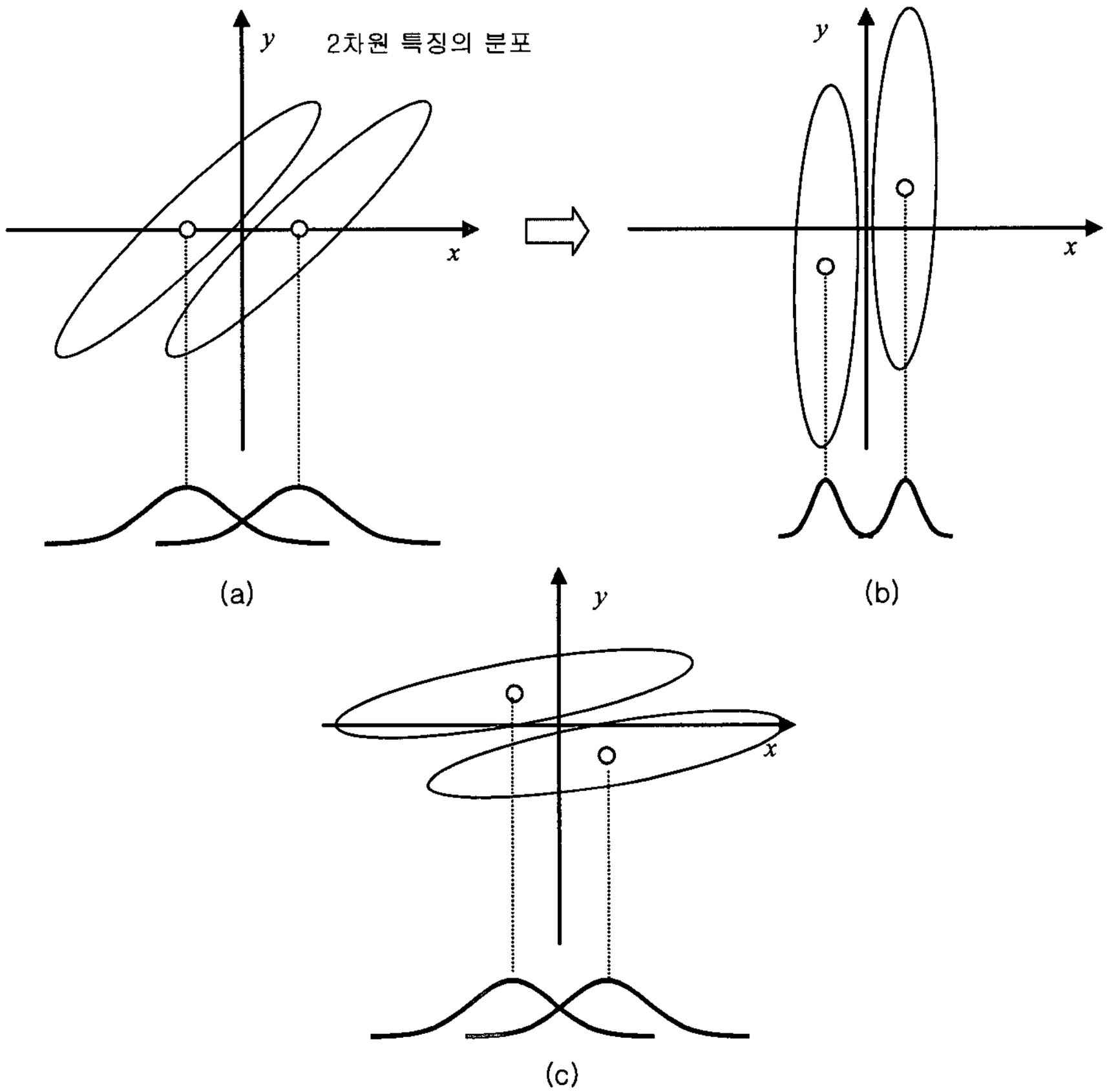
PSO의 가장 큰 장점은 구현하기 쉽다는 것이다. 이 때문에 genetic algorithm[4]과 같은 진화알고리즘을 적용하기 복잡한 문제에 적용할 수 있다[13]-[15]. 또한 PSO는 이전 모든 세대의 기록을 이용하기 때문에 단순히 바로 전 세대의 정보만을 이용하여 최적화를 수행하는 다른 진화알고리즘보다 좋은 성능을 나타낸다[16].

## 3. 대각공분산 GMM에 최적인 선형변환

본 장에서는 화자인식 시스템의 성능을 향상시키기 위하여 제안한 방법을 기술한다. 주성분 분석을 이용하여 구한 $W_{PCA}$는 각 차원의 상관관계를 줄여서 특징 분포를 잘 나타내는 장점을 가지고 있고, 선형판별 분석을 이용하여 구한 $W_{LDA}$는 $S_B/S_W$를 최대화하여 각 화자 특징 분포를 잘 분리하는 장점을 가지고 있다. 본 연구에서는 앞에 설명된 두 가지 방법의 장점과 $W$를 이용한 변환 후에 특징이 대각 공분산 GMM으로 모델링된다는 점을 이용하여 대각 공분산 GMM에 최적인 특징 분포를 갖는 $W_{GMM}$을 찾는 것을 목표로 한다. $W_{GMM}$을 찾기 위하여 목적함수 $J(W)$를 다음과 같이 설정한다.

$$J(W)=\sum_{k=1}^{S-1}\sum_{l=k+1}^{S}\frac{Dist(\lambda_k^W,\lambda_l^W)}{M} \tag{12}$$

$$Dist(\lambda_k^W,\lambda_l^W)=-\sum_{i=1}^{M}\left[\log p(\mu_{i,k}^W|\lambda_l^W)+\log p(\mu_{i,l}^W|\lambda_k^W)\right] \tag{13}$$




<그림 2> 2차원 특징 분포의 회전변환 예: (a) 초기 두 분포의 GMM (b) 제안한 방법을 적용할 경우 (c) PCA를 적용할 경우

$$\lambda_s^W = \left\{ w_{i,s}^W, \mu_{i,s}^W, \Sigma_{i,s}^W \right\} \text{ for } i = 1, \ldots, M \tag{14}$$

여기서 $\lambda_s^W$는 $W$에 의하여 변환된 특징 벡터로 모델링된 화자 $S$의 GMM 모델, $Dist(\lambda_k^W, \lambda_l^W)$는 화자 모델 $k$와 $l$의 통계학적인 거리이다. 목적함수 $J(W)$는 결국 각 화자 모델간의 통계학적 거리의 합을 나타낸다. $J(W)$를 최대화하면 본 연구에서 제안한 $W_{GMM}$을 얻을 수 있다. 그러나 $W$가 커질수록 목적함수 $J(W)$ 또한 큰 값을 가지므로 $W$가 무한히 커지는 것을 방지하기 위해 $W$를 좌표축(coordinate axes) 회전 변환으로 설정한다. 이런 과정을 통하여 고유벡터와 같은 성질을 갖는 변환 행렬을 구할 수 있다. <그림 2>는 2차원 특징 공간에서 최적의 대각 공분산

GMM을 위한 회전의 예를 보여준다. 그림에서 타원은 특징의 분포를 보여준다. 변환하기 전에는 <그림 2>의 (a)와 같이 $x$축으로 사상된 두 분포가 대부분 겹쳐지지만, 변환 후에는 <그림 2>의 (b)와 같이 잘 분리된 가우시안 모델을 구하는 것을 볼 수 있다. PCA나 LDA를 사용해도 유사한 결과를 얻을 수 있지만, PCA는 최대 분포의 방향만을 고려하므로 <그림 2>의 (c)와 같이 가우시안 모델을 잘 분리하지 못하는 경우가 있을 수 있고, LDA는 데이터가 유니모달(unimodal) 분포를 가지지 않는 경우에는 분포를 잘 분리하지 못하게 된다. 제안한 방법은 대각 공분산 GMM에 최적화하도록 회전 변환하므로 PCA나 LDA보다 더 좋은 분포를 얻을 수 있다.

$W$는 다음과 같은 방법으로 좌표축을 회전시켜서 구한다. 축의 회전 변환을 나타내기 위해 다음과 같은 행렬 $R$을 정의한다.

$$R=\begin{bmatrix} 0 & \theta_{1,2} & \cdots & \theta_{1,i} & \cdots & \theta_{1,j} & \cdots & \theta_{1,D-1} & \theta_{1,D} \\ 0 & 0 & \cdots & \theta_{2,i} & \cdots & \theta_{2,j} & \cdots & \theta_{2,D-1} & \theta_{2,D} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \theta_{i,j} & \cdots & \theta_{i,D-1} & \theta_{i,D} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & \theta_{j,D-1} & \theta_{j,D} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \theta_{D-1,D} \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{15}$$

여기서 $R$의 $i$행 $j$열 원소는 $i>j$ 일 때 $\theta_{i,j}$이고, 그 이외의 경우에는 0인 행렬이다. $\theta_{i,j}$는 $i$번째 축과 $j$번째 축으로 이루어진 평면을 원점을 중심으로 회전 시키는 각도이다. $R$이 주어졌을 때 $W$를 유도하는 방법은 다음과 같다.

단계 1: $W$와 같은 크기($D\times D$)인 좌표축(D차원 단위행렬과 동일) $A$를 만든다.
단계 2: for $i$ $(1, 2, \cdots, D-1)$
　　for $j$ $(i, i+1, \cdots, D)$
　　$A$에서 $i$축과 $j$축으로 이루어진 평면을 원점을 중심으로 $\theta_{i,j}$만큼 회전
단계 3: $W=A$

여기서 $W$는 $R$을 이용하여 유도할 수 있기 때문에 $J(W)$를 최대화하는 문제는 $J(R)$을 최대화하는 문제로 바꿀 수 있다. $J(R)$을 최대화하는 $R$은 3.1절에 설명한 PSO를 이용하여 다음과 같이 구한다.

단계 1: $N$개의 개체($R_1, R_2, \cdots, R_N$)를 임의의 값으로 초기화한다.

단계 2: 모든 개체($R_1, R_2, \cdots, R_N$)에서 변환행렬($W_1, W_2, \cdots, W_N$)을 유도한다.

단계 3: 변환행렬($W_1, W_2, \cdots, W_N$) 각각을 이용하여 특징을 변환하고 각각의 화자 모델을 만든다.

단계 4: 단계 3에서 얻어진 각각의 화자모델에 대한 적합도를 식 (12)를 이용하여 구한다. 여기서 얻어진 적합도는 곧 각 개체의 적합도가 된다.

단계 5: 각 개체의 적합도가 previous best solution의 적합도보다 좋다면 previous best solution을 현재의 개체로 교체한다. 또한 현재의 모든 개체의 적합도 중에서 global best solution의 적합도보다 좋은 것이 있다면 global best solution을 그 개체로 교체한다.

단계 6: 식 (10)과 식 (11)을 이용하여 각 개체의 다음 세대를 생성한다.

단계 7: 종료 조건을 만족하면 종료하고, 그렇지 않으면 단계 2로 가서 반복한다.

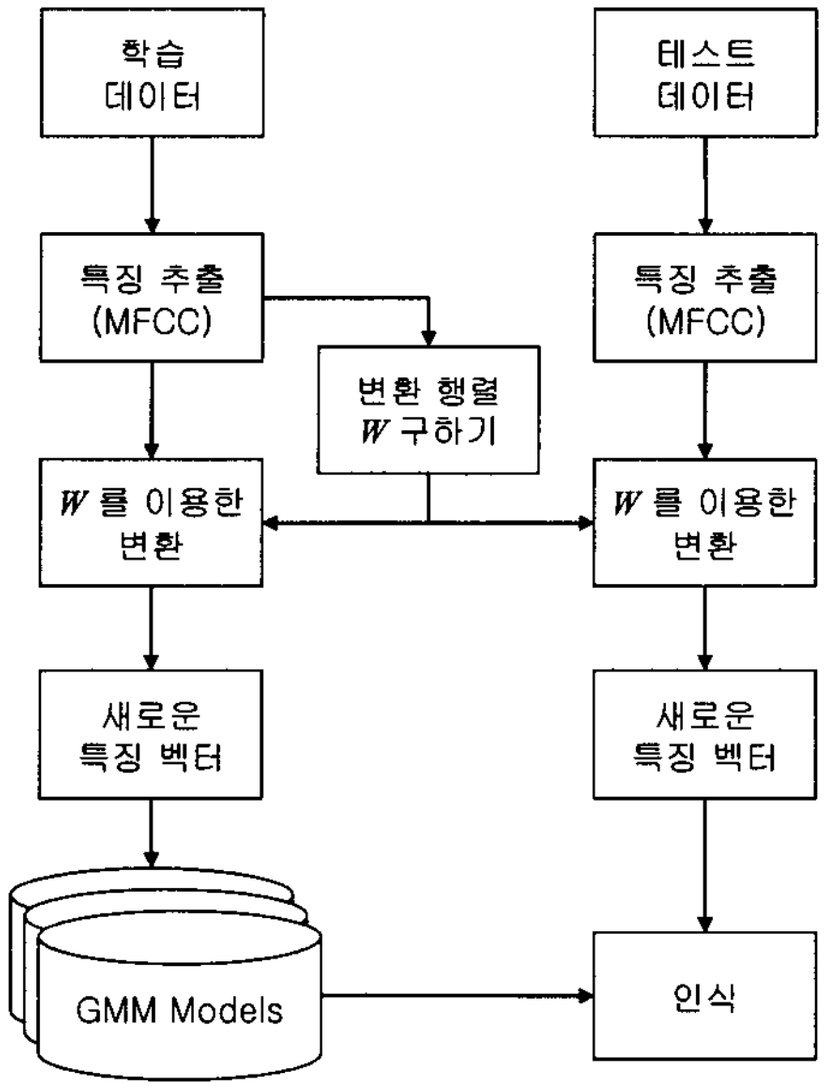
단계 8: Global best solution을 결과로 취한다.

## 4. 실험 및 결과

본 논문에서는 YOHO 음성 데이터베이스[12]의 음성을 이용한 문장독립 화자인식 시스템에서 제안한 방법의 성능을 평가하였다. YOHO 음성 데이터베이스는 화자인식 실험을 위해 만들어진 데이터베이스로 총 138명의 화자로 구성되어 있다. 각각의 화자는 등록용으로 4번, 인식용으로 10번의 세션을 갖도록 녹음하였으며, 녹음은 고급 전화용 마이크를 사용하여 조용한 사무실 환경에서 하였다. 발성 내용은 (35-72-41)과 같이 2자리 숫자를 3개씩 연이어 발음하는 형식으로, 등록 세션에는 24개씩 (한 화자 당 등록 음성 $4 \times 24 = 96$개), 그리고 인식 세션에는 4개씩(한 화자 당 인식 음성 $10 \times 4 = 40$개)의 음성이 있다.

본 논문에서는 138명의 화자 중 50명의 화자를 선택하여 실험하였다. 학습에는 한 개의 등록 세션을 이용하고 (24개의 음성), 인식에는 모든 인식 세션을 이용하였다. 또한 인식 대상 음성에 FaNT(Filtering and Noise Adding Tool)[9][10]를 이용하여 인위적으로 잡음을 삽입하였다. 잡음은 Aurora 2 데이터베이스[11]에서 제공된 잡음 중 babble, train 두 가지를 이용하였고, SNR을 15 dB, 10 dB로 삽입하였다.

특징으로는 20 ms Hamming window의 MFCC 12차와 에너지, 그리고 이의 1, 2차 미분 (39차원)을 이용하였다. 또한 채널 왜곡을 감소시키고 특징 벡터 공간의 불일치를 해결하기 위하여 cepstral mean subtraction (CMS) 방법을 사용하였다. MFCC 특징 벡터에서 앞에서 설명한 선형 변환을 이용하여 새로운 특징 벡터를 추출하여 실험하였다. 화자 모델은 2.3절에서 설명한 GMM[5]을 사용하였다. 혼합




<그림 3> 화자식별 시스템 구성도

(mixture) 수는 64개로 하였고 학습 회수(iteration)는 5회로 하였다. 본 논문에서 이용하는 화자식별 시스템은 <그림 3>과 같다.

본 논문에서는 PCA, LDA, 제안한 방법으로 구한 변환 행렬을 각각 $W_{PCA}$, $W_{LDA}$, $W_{GMM}$이라 부른다. $W_{GMM}$은 각 개체에 대해서 8개의 혼합수로 모델을 생성하여 적합도를 계산하고, 100번째 세대에서의 global best를 선택하고 좌표축을 회전 시켜 구했다. 혼합수가 증가할수록 목적함수의 복잡도가 지수승으로 증가하기 때문에 빠른 계산을 위해서 혼합수를 8개로 한정하였다. 본 연구에서 실험하는 $W_{PCA}$, $W_{LDA}$, $W_{GMM}$과 아무런 변환을 하지 않은 MFCC 특징의 적합도 비는 <표 1>과 같다.

<표 1> 각 $W$ 별 baseline과의 적합도 비 ($I$=단위행렬)

| 변환행렬 | 목적 함수 | 실제 값 | 적합도 비 | |
|---|---|---|---|---|
| 없음 | $J(I)$ | 169842 | $J(I)/J(I)$ | 1.0000 |
| $W_{PCA}$ | $J(W_{PCA})$ | 151441 | $J(W_{PCA})/J(I)$ | 0.8917 |
| $W_{LDA}$ | $J(W_{LDA})$ | 74745 | $J(W_{LDA})/J(I)$ | 0.4401 |
| $W_{GMM}$ | $J(W_{GMM})$ | 222817 | $J(W_{GMM})/J(I)$ | 1.3119 |

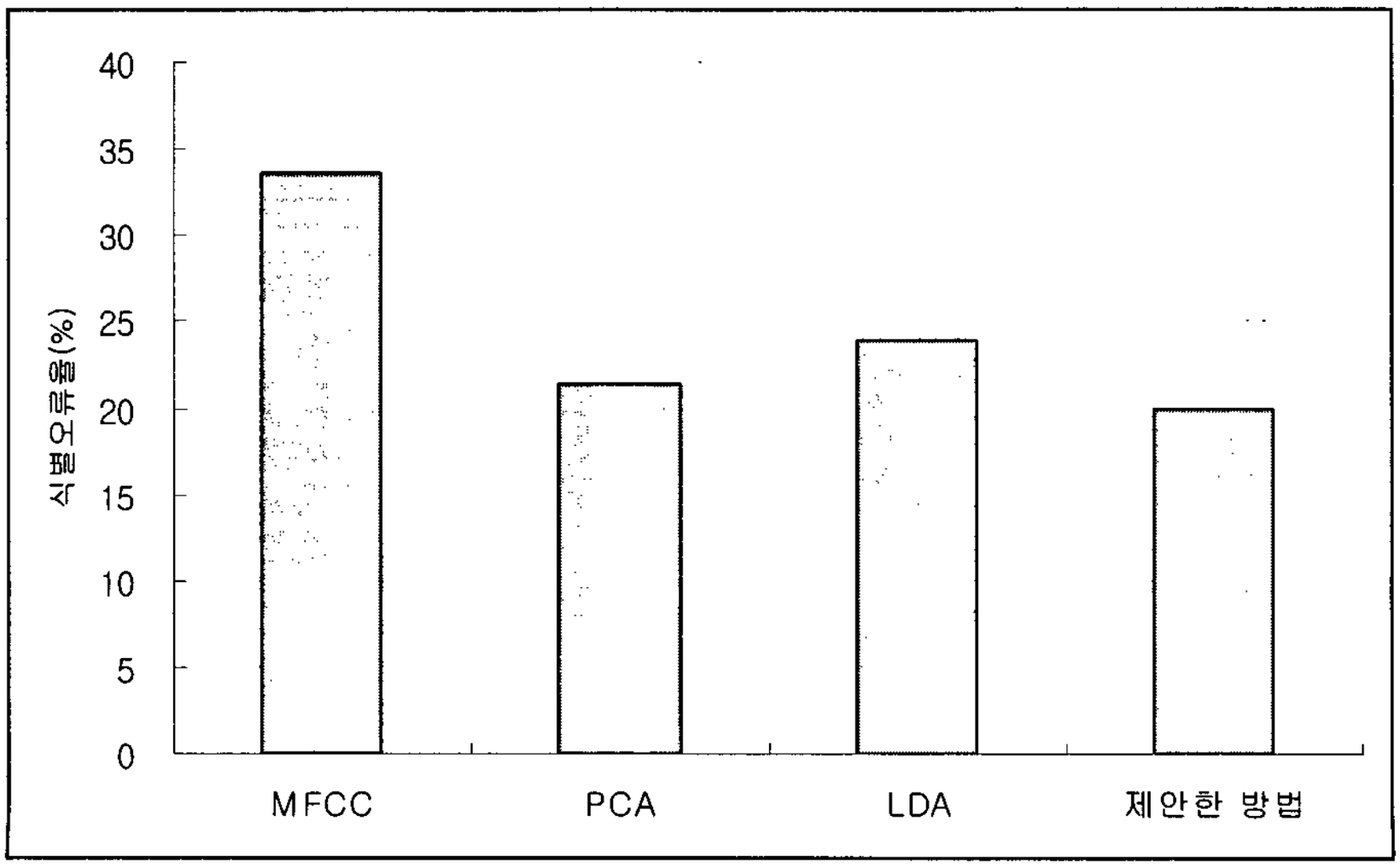
<표 2>와 <그림 4>는 MFCC 특징 벡터를 이용한 Baseline 실험과 강인한 화자 인식을 위한 변환 행렬 $W_{PCA}$, $W_{LDA}$, $W_{GMM}$을 적용한 화자식별 결과이다. 기존의 선형변환 방법인 PCA와 LDA는 각각 MFCC 특징을 이용한 실험보다 12.3%, 9.7%의 오류 감소 결과를 얻을 수 있었다.

Babble 잡음이 15 dB로 삽입된 환경인 경우 제안한 방법이 MFCC와 PCA 특징보다 각각 0.80%, 0.45% 낮은 성능을 나타내었다. 모든 환경에 대한 평균은 기존의 방법보다 높은 성능을 얻을 수 있었다. 평균 화자식별 오류율은 "MFCC > LDA > PCA > 제안한 방법" 순이며, 제안한 방법을 이용한 화자식별 실험은 평균적으로 MFCC, LDA, PCA보다 각각 13.72%, 4.02%, 1.39%의 오류가 감소하였다.

<표 1>과 <표 2>에서 PCA(분산 최대화), LDA($S_B/S_W$ 최대화), 제안한 방법(모델간 거리 최대화) 각각의 목적이 인식률 향상에 도움이 되었지만, 제안한 방법이 이들 중 가장 좋다는 것을 알 수 있다.

<표 2> 전체 화자식별 오류율 (%)

| 특징 종류 | 잡음 종류 | | | | 평균 |
|---|---|---|---|---|---|
| | Babble | | Train | | |
| | 15 dB | 10 dB | 15 dB | 10 dB | |
| MFCC | 7.60 | 34.50 | 31.45 | 61.00 | 33.64 |
| PCA | 7.95 | 30.40 | 11.15 | 35.75 | 21.31 |
| LDA | 11.92 | 31.65 | 15.35 | 36.85 | 23.94 |
| 제안한 방법 | 8.40 | 29.35 | 10.10 | 31.85 | 19.92 |




<그림 4> 평균 화자식별 오류율

## 5. 결 론

본 논문에서는 어떤 상황에서 수행될지 알 수 없는 화자 인식기의 성능 향상을 위하여 대각 공분산 GMM에 최적인 선형변환을 이용한 강인한 화자식별 방법을 제안하였다. 실험에서 학습에는 깨끗한 음성(clean speech)을 사용하고, 식별에는 잡음이 포함된 음성(noisy speech)을 이용하였다. 주성분 분석과 선형판별 분석을 이용한 화자식별 오류율은 기존 MFCC 특징을 이용한 실험보다 각각 12.3%, 9.7% 감소하였다. 본 논문에서 제안한 방법은 MFCC 특징, 주성분 분석, 선형판별 분석보다 각각 13.72%, 1.39%, 4.02% 화자식별 오류율이 감소하는 결과를 얻을 수 있었다.

제안한 방법이 우수한 이유는 실제 화자식별을 위한 대각 공분산 가우시안 혼합 모델(GMM)에 최적인 특징 공간으로 특징을 변환시키기 때문이다. 다시 말해 제안한 방법을 이용하여 변환된 특징은 식별 대상 화자들 간의 거리를 최대화하면서 가우시안 혼합 모델로 모델링하기 좋은 분포를 갖는다.

실험을 통해 PCA(분산 최대화), LDA($S_B/S_W$ 최대화), 제안한 방법(모델간 거리 최대화) 각각의 목적이 인식률 향상에 도움이 되었지만, 제안한 방법이 이들 중 가장 좋다는 것을 알 수 있었다.

향후 계획은 본 논문에서 제안한 대각 공분산 GMM에 최적인 선형변환 행렬을 다양한 최적화 알고리즘을 이용하여 구하는 것과 결정적(deterministic) 방법으로 구하는 방법에 대해서 연구하는 것이다.

## 참 고 문 헌

[1] 유하진, "부가 주성분분석을 이용한 미지의 환경에서의 화자식별", *말소리*, 제54호, pp. 73-83, 2005.

[2] Z. Wanfeng, Y. Yingchun, W. Zhaohui, S. Lifeng, "Experimental evaluation of a new speaker identification framework using PCA", *Proc. IEEE International Conference on Systems, Man and Cybernetics*, Vol. 5, pp. 4147-4152, 2003.

[3] Q. Jin, A. Waibel, "Application of LDA to speaker recognition", *Proc. ICSLP*, Vol. 2, pp. 250-253, 2000.

[4] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd Ed., Wiley-Interscience, 2000.

[5] D. A. Reynolds, R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72-83, 1995.

[6] J. Kennedy, R. C. Eberhart, "Particle swarm optimization", *Proc. IEEE International Conference on Neural Networks*, pp. 1942-1948, 1995.

[7] M.-S. Kim, H.-J. Yu, K.-C. Kwak, S.-Y. Chi, "Robust text-independent speaker identification using hybrid PCA&LDA", *Lecture Notes in Artificial Intelligence*, Vol. 4293, pp. 1067-1074, 2006.

[8] D. Pearce, H. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under conditions", *Proc. ICSLP*, Vol. 4, pp. 29-32, 2000.

[9] M. Cernak, "Unit selection speech synthesis in noise", *Proc. ICASSP*, Vol. 1, pp. 761-764, 2006.

[10] G. Hirsch, "Fant - filtering and noise adding tool", available at http://dnt.kr.hsnr.de/download.html, 2005.

[11] H. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *Proc. ISCA ITRW on Automatic Speech Recognition: Challenges for the Next Millennium*, pp. 181-188, 2000.

[12] J. P. Campbell, Jr., "Testing with the YOHO CD-ROM voice verification corpus", *Proc. ICASSP*, pp. 341-344, 1995.

[13] J. Kennedy, R. Eberhart, *Swarm Intelligence*, Morgan Kaufmann, 2001.

[14] R. Eberhart, Y. Shi, "Comparison between genetic algorithms and particle swarm optimization", *Proc. Seventh Annual Conference on Evolutionary Programming*, pp 611-619, 1998.

[15] L. Diaz, T. Milligan, *Antenna Engineering Using Particle Optics: Practical CAD Techniques and Software (Artech House Antenna and Propagation Library)*, Artech House Publishers, 1996.

[16] C. Veenman, M. Reinders, E. Backer, "A cellular coevolutionary algorithm for image segmentation", *IEEE Transactions on Image Processing*, Vol. 24, No. 9, pp. 1273-1280, 2002.

접수일자: 2007년 5월 17일
게재결정: 2008년 2월 28일

▶ 김민석(Min-Seok Kim)
주소: 130-743 서울 동대문구 전농동 90
소속: 서울시립대학교 컴퓨터과학부 박사과정
전화: 02) 2210-5322
E-mail: ms@uos.ac.kr

▶ 양일호(Il-Ho Yang)
주소: 130-743 서울 동대문구 전농동 90
소속: 서울시립대학교 컴퓨터과학부 석사과정
전화: 02) 2210-5322
E-mail: heisco@hanmail.net

▶ 유하진(Ha-Jin Yu) : 교신저자
주소: 130-743 서울 동대문구 전농동 90
소속: 서울시립대학교 컴퓨터과학부
전화: 02) 2210-5613
E-mail: hjyu@uos.ac.kr