# 주성분분석과 선형판별분석의 장점을 이용한 강인한 화자식별

김 민 석, 유 하 진, 김 승 주
서울시립대학교 컴퓨터과학부

# Robust Speaker Identification Exploiting the Advantages of PCA and LDA

Min-Seok Kim, Ha-Jin Yu, Sung-Joo Kim
School of Computer Science, University of Seoul
E-mail: {mskim, hjyu, sung}@venus.uos.ac.kr

## Abstract

The goal of our research is to build a text-independent speaker identification system that can be used in mobile devices without any additional adaptation process. In this paper, we show that exploiting the advantages of both PCA(Principle Component Analysis) and LDA(Linear Discriminant Analysis) can increase the performance in the situation. The proposed method reduced the relative recognition error by 13.5%

## I. 서론

우리는 언제 어디에서나 모바일 기기를 이용하여 네트워크상에 있는 수많은 정보를 이용할 수 있는 시대에 살고 있다. 최근 네트워크 기술이 발전하면서, 기존에 유선으로 연결된 컴퓨터에서만 서비스 가능했던 컨텐츠들이 이제는 모바일 기기에서 무선으로 서비스 되고 있다. 모바일로 서비스되고 있는 컨텐츠 중에는 전자메일과 같이 사용자 보안이 필요한 분야가 있는데, 모바일 기기에서는 일반 컴퓨터에 비해 제한된 입력장치를 이용하여 보안을 수행할 수밖에 없다. 대부분의 모바일 기기에 장착되어 있는 마이크를 사용한 화자인식을 보안에 이용하면 이런 상황에서 좋은 해결책이 될 수 있을 것이다.

모바일 기기에 이용되는 화자 인식 시스템은 언제, 어떤 상황에서 이용될지 예측할 수 없기 때문에 환경 변화와 시차 변화에 강인한 화자 모델을 만들어야 한다. 이 문제를 해결하기 위해 본 연구에서는 주성분 분석과 선형판별 분석의 장점을 결합 하여, MFCC (Mel Frequency Cepstrum Coefficient) 음성 특징에서 환경 변화에 강인한 특징 벡터를 추출한다. 또한 상대적으로 인식기 성능에 영향이 적은 차원을 축소시켜 성능을 유지하면서 낮은 성능의 프로세서와 적은 메모리에서도 수행할 수 있도록 하였다.

기존의 강인한 화자 인식 방법으로는 주성분 분석을 이용한 방법[2-5]과 선형판별 분석[6]을 이용한 방법이 있다. 또한 주성분 분석이 상관관계가 큰 축과 클래스 분포가 나란할 경우 화자 분류에 좋지 않은 축을 찾을 수 있는 문제를 해결하기 위해 클래스 정보를 추가하는 부가 주성분 분석이 제안되었다[5]. 본 연구에서는 Microarray 분류에서 이용되었던 Hybrid PCA-LDA[1]를 화자 인식에 적용하여, 주성분 분석과 선형판별 분석의 장점을 모두 갖는 축을 찾아 화자인식 실험을 하였다.

본 논문의 구성은 다음과 같다. 2 장에서 주성분 분석(PCA: Principal Component Analysis), 선형 판별 분석 (LDA: Linear Discriminant Analysis)과 화자인식에서 가장 일반적으로 이용되는 GMM(Gaussian Mixture Model)을 소개하고 3장에서는 본 논문이 제안한 PCA-LDA 결합에 대하여 기술한다. 4장에서는 실험 환경 및 결과를 제시하고 5장에서 결론을 맺는다.

## II. 강인한 화자 식별 방법

### 2.1 주성분 분석(Principal Component Analysis)

주성분 분석은 다차원 특징 벡터에서 각 차원의 상관관계를 줄이는 독립인 축을 구하고, 그 축으로 특징 벡터를 사상시켜 높은 차원의 정보를 유지하면서 낮은 차원으로 축소시키는 방법이다. 본 연구에서는 여러 클래스로 나뉘어 있는 분포를 나타내기 위하여 다음과 같이 각 차원의 상관관계를 나타내는 $S_\Sigma$ 를 표현한다.[1][8]

$$S_\Sigma = \frac{1}{C}\sum_{j=1}^{C}\frac{1}{N_j}\sum_{i=1}^{N_j}(\overrightarrow{x_i^{(j)}} - \overrightarrow{m})(\overrightarrow{x_i^{(j)}} - \overrightarrow{m})^T \quad (1)$$

여기서 $\overrightarrow{x_i^{(j)}}\{i=1,\ldots,N_j, j=1,\ldots,C\}$ 는 $j$번째 클래스의 $i$ 번째 특징 벡터, $N_j$는 $j$ 클래스의 특징 벡터 개수 $C$ 는 클래스 개수이고 $\overrightarrow{m}$은 전체 평균이다.

주성분 분석을 위한 변환 행렬 $W_{pca}$는 $S_\Sigma$의 고유벡터를 구하고 고유값이 큰 순으로 n개를 선택하여 만든다.

### 2.2 선형판별 분석(Linear Discriminant Analysis)

선형판별 분석은 클래스간 분산( $S_B$ : between-class scatter)과 클래스내 분산( $S_W$ : within-class scatter)의 비율을 최대화하는 축으로 특징 벡터 차원을 축소시키는 방법이다. $S_B$와 $S_W$은 다음과 같이 구한다.[1][8]

$$S_B = \sum_{j=1}^{C} N_j(\overrightarrow{m_j} - \overrightarrow{m})(\overrightarrow{m_j} - \overrightarrow{m})^T \quad (2)$$

$$S_W = \sum_{j=1}^{C}\sum_{i=1}^{N_j}(\overrightarrow{x_i^{(j)}} - \overrightarrow{m_j})(\overrightarrow{x_i^{(j)}} - \overrightarrow{m_j})^T \quad (3)$$

여기서 $\overrightarrow{m_j}$는 $j$번째 클래스의 평균이다.

선형판별 분석을 위한 변환 행렬 $W_{lda}$는 $S_W^{-1}S_B$의 고유벡터를 구하고 고유값이 큰 순으로 n개를 선택하여 만든다.

### 2.3 GMM을 이용한 화자 식별

GMM[7]은 문장 독립 화자 식별 시스템에서 가장 많이 사용되는 모델링 방법이다. 본 연구에서 이용한 GMM 화자 식별 과정은 그림 1과 같다.

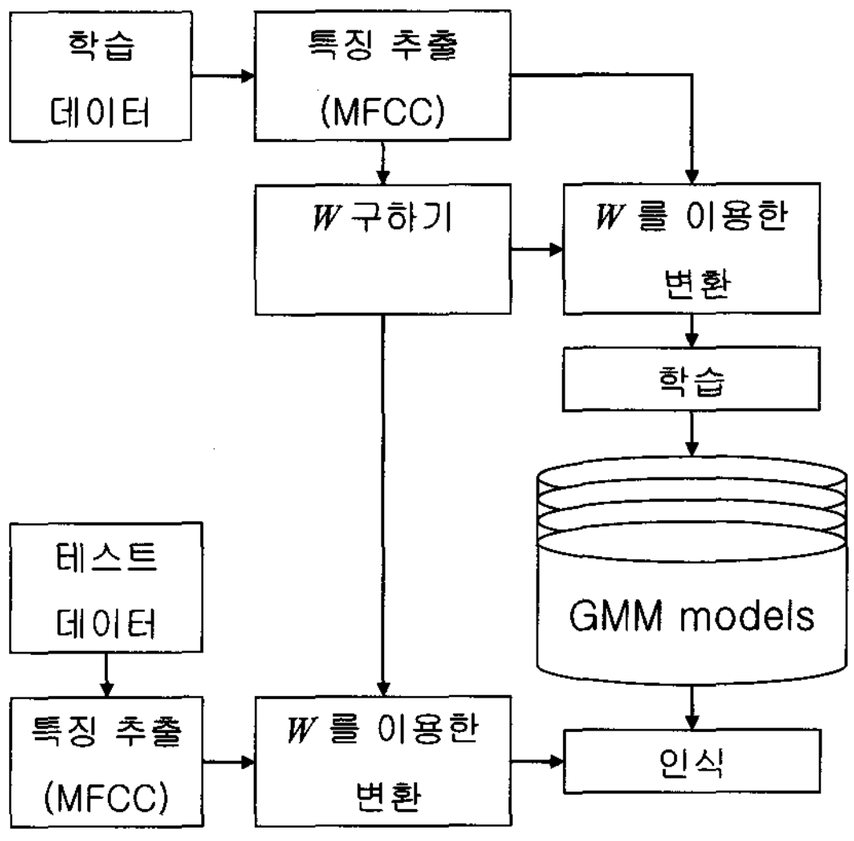


그림 1. 화자식별 과정( $W$는 변환 행렬)

## III. PCA-LDA 결합을 이용한 화자 식별

주성분 분석은 각 차원의 상관관계를 줄이는 것이 목표이고, 선형판별 분석은 $S_B/S_W$ 를 최대화하는 것이 목표이다. 본 절에서는 이 두 가지 특징을 포함하는 변환 행렬을 구하기 위해 다음과 같은 식을 정의한다.[1]

$$W_{opt} = \arg\max_W \frac{|W^T S_1 W|}{|W^T S_2 W|} \quad (4)$$

$$S_1 = [(1-\alpha)\cdot S_B + \alpha \cdot S_\Sigma] \quad (5)$$

$$S_2 = [(1-\beta)\cdot S_W + \beta \cdot I] \quad (6)$$

여기서 $I$ 는 단위행렬이고, α와 β는 각각 $S_B$와 $S_\Sigma$, $I$와 $S_W$의 비율을 결정한다. $W_{opt}$는 본 논문에서 제안한 변환행렬이고, $S_2^{-1}S_1$의 고유벡터를 구하고, 고유값이 큰 순으로 n개를 선택하여 만든다. ( α, β) 쌍이 (0, 0)일 경우 LDA와 동일하고, (1, 1)일 경우 PCA와 동일하다. 특별한 α, β 값에 따른 결과는 표 1[1] 과 같다. α, β가 잘 정의된 경우 $W_{opt}$는 주성분 분석과 선형판별 분석의 장점을 모두 가질 수 있다.

화자인식에 (4)식을 적용하기 위해서 몇 가지 제약조건이 필요하다. 실제 음성 특징 벡터의 분포에서는

$S_B$와 $S_\Sigma$, $I$와 $S_W$의 차이가 너무 크기 때문이다. $S_\Sigma$는 전체 개수로 나누어주는 과정이 있어서 그 크기가 일정 범위 안에 존재하지만, $S_B$는 그런 과정이 없기 때문에 $N_j$이 커지면 커질수록 그 크기는 무한히 증가한다. 또한 $S_W$는 특징벡터가 CMS (Cepstral Mean Subtraction)를 거치기 때문에 그 값이 0에 상당히 가까운 값을 가지게 된다. 화자인식에서는 $S_\Sigma$와 $S_B$, $I$와 $S_W$의 크기를 고려하여 비율이 50%, 50%되는 값을 정했을 때 Trade-off 지점을 구할 수 있다. 행렬의 크기는 다음과 같이 정의 한다.

$$\text{size}(X) = \max_{1 \le i \le N}(|\overrightarrow{x_i}|) \qquad (7)$$
$$X = [x_1, x_2, \ldots x_N]$$

여기서 $|\overrightarrow{x_i}|$는 $\overrightarrow{x_i}$의 크기를 말한다.

표 1. 특별한 경우에 대한 결합식 분석[1]

| (α, β) | PCA-LDA 결합 분석 | 설명 |
|---|---|---|
| (0, 0) | $W_{opt} = \arg\max_W \frac{\lvert W^T S_B W \rvert}{\lvert W^T S_W W \rvert}$ | LDA |
| (0, 1) | $W_{opt} = \arg\max_W \frac{\lvert W^T S_B W \rvert}{\lvert W^T I W \rvert}$ | PCA-LDA 결합 |
| (1, 0) | $W_{opt} = \arg\max_W \frac{\lvert W^T S_\Sigma W \rvert}{\lvert W^T S_W W \rvert}$ | PCA-LDA 결합 |
| (1, 1) | $W_{opt} = \arg\max_W \frac{\lvert W^T S_\Sigma W \rvert}{\lvert W^T I W \rvert}$ | PCA |
| $(\frac{1}{2}, \frac{1}{2})$ | $W_{opt} = \arg\max_W \frac{\lvert W^T [S_B + S_\Sigma] W \rvert}{\lvert W^T [S_W + I] W \rvert}$ | Trade-off |

## Ⅳ. 인식실험 및 결과

### 4.1 실험 환경 [5]

본 연구에서는 음성정보기술지원센터(SITEC)에서 수집한 자동차 화자인증용 음성 DB(CarSpkr01)을 이용하여 실험하였다. 음성은 주행중인 2500CC급 승용차(HYUNDAI GRANDEUR XG, Automatic)에서 수집되었다. 맑은 날씨에 아스팔트 도로를 창문을 닫고 오디오를 끈 상태로 30~60 km/h의 속도로 주행하였다.

음성 수집에 사용된 마이크는 다이나믹 마이크(head-won SHURE SM-10A, Uni-Cardioid), 콘덴서 마이크 (AKG B400-BL, Cardioid), 국산 저가 핸즈프리 마이크 (HYUNDAI Handfree)의 세 종류로 총 8개의 위치에 나누어 장착되었다. 본 연구에서는 다이나믹 마이크로 화자의 입에서 3cm 가량을 유지하며 녹음된 음성(hdw로 표기됨)을 학습에 사용하였고, 선바이저에 장착된 콘덴서 마이크로 녹음된 음성(sv1으로 표기됨)을 테스트에 사용하였다. 이것은 학습과 테스트에서 서로 다른 마이크를 사용하여 인식 성능의 저하를 확인하기 위한 것이다.

본 데이터는 음소가 고루 분포된 문장 및 단어 세트, 4연 숫자음 세트를 총 30명의 화자가 최초발성, 1일 후, 1주일 후, 1개월 후, 2개월 후의 시차를 두고 총 5회 발성하였다. 한 화자 당 발성 수는 250개 이다. 본 연구에서는 최초 발성된 4연 숫자음 세트(화자 당 144개)로 학습하고 최초발성(4320개), 1일 후(4320개), 1주일 후(4320개) 발성된 4연 숫자음 세트를 테스트 데이터로 이용하였다.

화자인식을 위한 특징으로는 15차 MFCC와 에너지, 이의 1차 및 2차 미분을 사용하고, 채널왜곡을 감소시키기 위하여 CMS (Cepstral Mean Subtraction) 방법을 사용한다. 화자 모델은 2.3절에서 설명된 GMM을 사용한다. 혼합수는 100개이고 학습 회수는 5회로 하였다.

### 4.2 제안한 PCA-LDA 결합을 이용한 실험 결과

제안한 방법의 효과를 확인하기 위해서 MFCC를 이용한 Baseline 실험과 기존 널리 이용되는 방법인 PCA와 LDA 실험을 실시하고 PCA-LDA 결합(제안한 방법)과 비교하였다. 본 실험의 size($S_\Sigma$)/size($S_B$)는 6.45e+09 이고 size($I$)/size($S_W$)는 1.37e-08이다. $S_\Sigma$와 $S_B$, $I$와 $S_W$의 크기를 고려한 (α, β) 값은 실험적으로 (1.0e-09, 1 - 1.0e-8)로 정했다. 화자식별 오류율은 표 2와 같다.

표 2. 화자식별 오류율(%)

| | 차원 | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|---|
| Baseline | 48 | 41.83 | 49.00 | 49.98 | 51.23 | 52.96 |
| LDA(0, 0) | 48 | 30.74 | 42.64 | 44.88 | 44.28 | 39.10 |
| PCA(1, 1) | 48 | 5.05 | 13.52 | 15.44 | 13.54 | 15.67 |
| 제안한방법 | 48 | 1.30 | 1.50 | 1.94 | 1.30 | 0 |
| 제안한방법 | 30 | 1.48 | 1.50 | 1.27 | 1.27 | 0 |
| 제안한방법 | 20 | 2.27 | 2.50 | 2.15 | 2.55 | 0.05 |

여기서 R1은 최초발성, R2는 1일 후, R3는 1주일

후, R4는 1개월 후, R5는 2개월 후 발성이다.

기존 PCA방법과 LDA방법은 시차가 변할수록 식별 오류율이 증가한다. 반면에 제안한 방법에서는 식별 오류율이 크게 변하지 않았다. 뿐만 아니라 차원을 약 2/5로 줄여 20차원으로 축소했을 때에도 PCA방법보다 1주일 후 식별 오류율이 13.29%감소했다. 본 논문에서 제안한 PCA-LDA 장점을 결합한 방법을 이용하였을 때 97%가 넘는 인식율을 얻을 수 있었다.

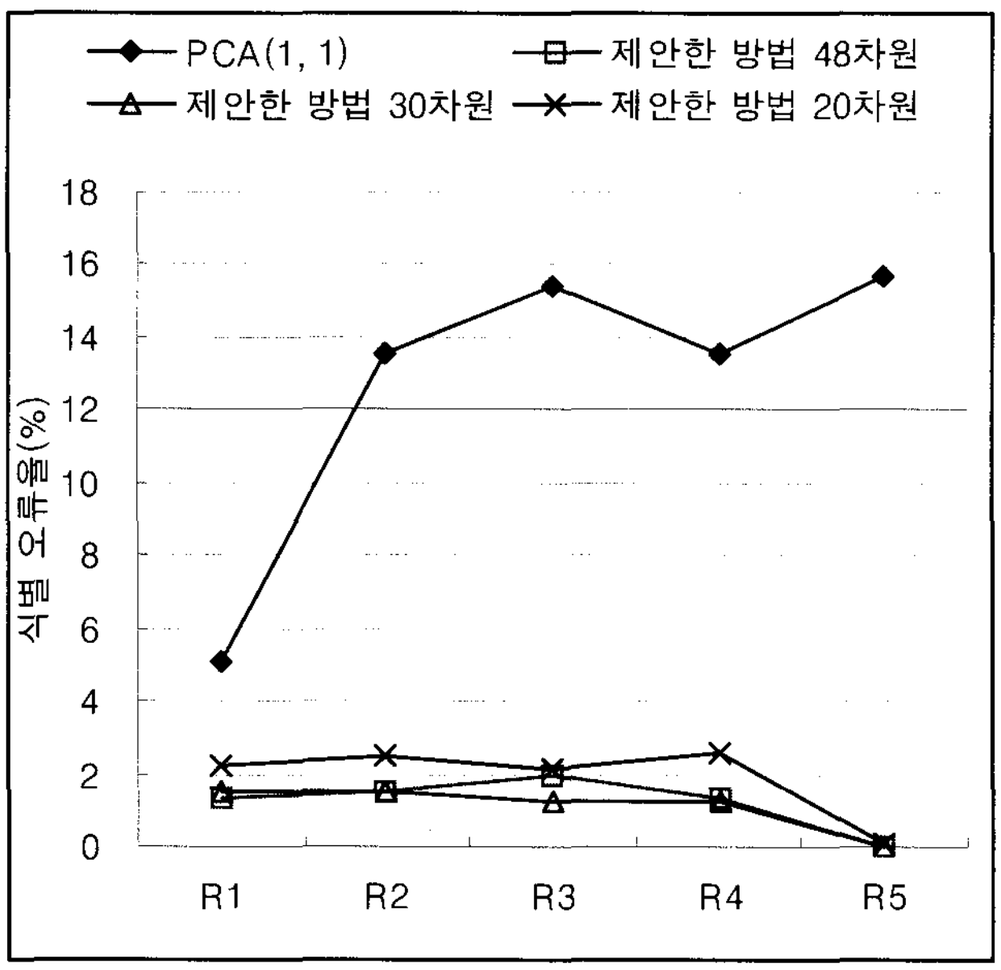


그림 2. 화자 식별 오류율 그래프

## V. 결론

본 논문에서는 모바일 기기에서의 환경 불일치 상황과 시차 변화에 따른 인식률 저하를 해결하기 위해서 기존 주성분 분석과 선형판별 분석의 장점을 결합한 방법을 제안 하고, 차원을 축소시키면서 실험을 수행하였다. 다이나믹 마이크로 녹음된 잡음이 거의 없는 음성을 이용하여 화자 모델을 만들고, 주행 중인 차량에서 콘덴서 마이크로 녹음된 음성을 테스트에 이용하였다. 1주일 후 식별 오류율은 기존 방법인 주성분 분석은 15.44%, 선형판별 분석은 44.88%였고, 제안한 방법은 1.94%였다. 제안한 방법은 주성분 분석에 비해서 식별 오류율이 13.5% 감소했다. 시차 변화에 따른 실험에서는 주성분 분석 방법이 최초발성에 비해서 1주일 후 식별 오류율이 10.39% 증가한 반면에 제안한 방법에서는 0.64%만 증가하여서 시차 변화에 강인한 것을 알 수 있었다. 또한 차원을 축소한 실험에서 48차원에 비해서 20차원에서의 1주일 후 식별 오류율이 0.21%증가 하였으나 주성분 분석에 비해서 13.29% 감소하여서 높은 성능을 나타내는 것을 알 수 있었다.

향후 계획으로는 화자인식기의 성능을 최적화 시키는 ( α, β)를 찾는 알고리즘 연구, 다른 음성 DB에서의 성능평가 등이 있다.

## 참고문헌


[1] Yijuan Lu, Qi Tian, Maribel Sanchez, Yufeng Wang, "Hybrid PCA and LDA Analysis of Microarray Gene Expression Data," IEEE Symposium on Computational Intelligence in Bio informatics and Computational Biology, Nov. 2005

[2] Zhang Wanfeng, Yang Yingchun, Wu Zhaohui and Sang Lifeng, "experimental evaluation of a nre speaker identification freamework using PCA," IEEE International Conference on Systems, Man and Cybernetics, Vol 5, pp. 4147 - 4152, Oct. 2003

[3] Peliv Ding, Limig Zhang, "Speaker Recognition using Pricipal Component Analysis," Proceeding of ICONIP 2001, 8th International Conference on Neural Information Processing, Shanghai China, November 14-18, 2001.

[4] 이윤정, 서창우, 강상기, 이기용, "화자식별을 위한 강인한 주성분 분석 가우시안 혼합 모델," 한국음향학회지 제22권, 제7호, pp. 519-527, 2003

[5] 유하진, "부가 주성분분석을 이용한 미지의 환경에서의 화자식별," 대한음성학회지:말소리 제54호, pp. 73-83, 2005.

[6] Q Jin, A Waibel, "Application of LDA to speaker recognition," Proc. ICSLP-00, Beijing, China, October 2000.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Processing, vol. 3, no. 1, pp. 72-83, 1995.

[8] Richard O. Duda, Peter E. Hart and David G. Stork, "Pattern Classification 2nd edition"