# 부가 주성분분석을 이용한 미지의 환경에서의 화자식별

유하진(서울시립대)

<차 례>



<Abstract>

## Speaker Identification Using Augmented PCA in Unknown Environments

**Ha-Jin Yu**

The goal of our research is to build a text-independent speaker identification system that can be used in any condition without any additional adaptation process. The performance of speaker recognition systems can be severely degraded in some unknown mismatched microphone and noise conditions. In this paper, we show that PCA(principal component analysis) can improve the performance in the situation. We also propose an augmented PCA process, which augments class discriminative information to the original feature vectors before PCA transformation and selects the best direction for each pair of highly confusable speakers. The proposed method reduced the relative recognition error by 21%.

* Keywords: PCA, Augmented PCA, Speaker recognition, Speech recognition.

# 1. 서 론

유비쿼터스 환경에서는 여러 지역에 있는 사람들이 언제 어디서나 서로를 접근 할 수 있도록 해주는 컴퓨팅 환경이 필요하므로, 인터페이스 자체도 매우 쉽고 편리하여야 한다는 전제 조건이 필요하게 된다. 현재 사람에게 가장 쉽고 편리한 인터페이스는 음성이라고 할 수 있다. 음성은 정보의 전달뿐만 아니라 보안이 필수적인 유비쿼터스 환경에서 중요한 생체인식 도구 중의 하나가 된다. 그런데, 이러한 유비쿼터스 환경은 음성을 처리하는데 있어서 가장 치명적인 환경이 될 수도 있다. 사용자의 위치가 변하면 마이크의 종류가 달라지고, 주변 잡음의 특성도 크게 달라지게 된다. 이러한 요인들은 현재 음성 또는 화자인식[1]의 실용화를 가로막는 가장 중대한 장애요인이 된다. 본 연구에서는 보안을 위한 사용자 인증에서 음성을 사용하는 데 있어서 마이크의 차이와 미지 잡음 하에서의 문제를 해결하는 것을 목표로 하고 있다.

마이크 특성의 차이를 보정하기 위한 연구와 잡음을 처리하기 위한 연구는 수없이 진행되어 왔으며, 많은 우수한 방법들이 소개 되었지만, 대부분의 방법은 서로 다른 마이크의 특성이나 잡음의 특성 또는 잡음 정도에 대한 정보를 필요로 한다[2]. 그런데, 수시로 위치가 변하는 환경에서는 이들 환경의 특성을 파악하거나 적응을 할 만한 여건이 되지 않은 경우가 있을 수 있다. 따라서 본 연구에서는 마이크나 잡음의 특성에 대한 정보가 전혀 없는 상태에서 발성화자를 인식하는 데에 주안점을 두었다. 인식 환경에서의 마이크나 잡음의 특성을 미리 알 수 없으므로, 잡음이 없는 음성으로 모델 학습을 하고, 학습 시와 다른 마이크를 사용하며 잡음이 있는 상태에서 화자인식을 하는 환경을 설정하였다.

마이크의 차이와 잡음에 강인한 특징을 추출하기 위해 먼저 주성분분석(principal component analysis)을 적용하였다. 주성분분석은 음성인식이나 화자인식에서 특징의 차원수를 줄여서 불필요한 정보를 제거하고 모델의 크기나 인식 시간을 줄이는 데 많이 사용되고 있다[3]-[6]. 본 연구에서는 주성분분석의 이러한 특성을 이용하여 화자 특성을 잘 표현하는 특징을 추출하는 데 사용하였다. 실험을 통하여 주성분분석을 통한 특징의 추출이 화자의 특성을 잘 표현함을 확인하였으며, 화자정보를 부가한 주성분 분석 방법을 제안하여 인식 성능을 높일 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 주성분분석 과정과 화자인식에서 가장 일반적으로 사용되는 Gaussian Mixture Model(GMM)을 소개하고 3장에서 제안한 부가주성분 분석과정을 설명한다. 4장에서 실험 환경 및 결과를 제시하고 5장에서는 결론을 맺는다.

## 2. 주성분 분석을 이용한 화자 식별

### 2.1. 주성분분석(Principal Component Analysis)

주성분 분석은 특징 공간을 표현하기 위해 서로 독립적인 축을 구하고, 차원을 축소시켜 저장 공간과 처리시간을 감축하기 위해 주로 사용된다[3]-[6]. 주성분 분석은 다음과 같은 과정을 통해 특징을 변환한다.

단계 1: 모든 데이터의 각 차원에 있는 요소를 각 차원의 평균으로 차감하여 각 차원의 평균이 0이 되도록 한다.

단계 2: 학습 데이터를 이용하여 공분산 행렬을 구한다. 공분산 행렬은 특징벡터의 상관관계와 변이성을 표현한다.

단계 3: 공분산 행렬의 고유벡터(eigenvector)를 구한다. **A**가 n×n 행렬이고, **x**는 n차원 열벡터, λ는 실수 일 때,

$$\mathbf{Ax} = \lambda\mathbf{x}$$

를 만족하는 λ가 고유값(eigenvalue)이고, **x**는 고유벡터이다. 특정 고유값에 대응하는 고유벡터는 무수히 많으므로, 보통 길이가 1인 단위(unit) 고유벡터를 사용한다.

단계 4: 구해진 고유벡터를 모두 모아 변환행렬을 작성한다. 가장 큰 고유값에 해당하는 고유벡터의 방향이 전체 데이터의 분포를 표현하는 가장 중요한 축이 되고, 가장 작은 고유값에 해당하는 고유벡터의 방향이 가장 중요하지 않은 축이 된다. 따라서 일반적으로 가장 중요한 몇 개의 축을 정하여 변환행렬을 만드는데, 본 연구에서는 자료의 차원을 줄이는 것이 목적이 아니므로 모든 축을 사용한다.

단계 5: 변환행렬을 이용하여 모든 데이터를 변환한다.
(변환된 데이터) = (변환행렬)×(입력벡터)

### 2.2. Gaussian mixture model (GMM) [6]

GMM은 텍스트 독립 화자인식에 가장 많이 사용되는 모델링 방법이다. GMM에서는 특징 벡터의 평균과 공분산을 가지고 다차원 가우시안 확률 분포 함수에 의하여 각 화자를 모델링한다. 입력 **x**에 대하여 D차원 확률분포함수(pdf)는 다음과 같이 계산된다.

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

여기서 $\mu_i$ 는 자료의 평균 벡터이고, $\Sigma_i$는 공분산행렬이다.

M개의 혼합(mixture)수를 가진 화자 모델에서 GMM은 다음과 같이 표현된다.

$$p(\mathbf{x}\mid\lambda)=\sum_{i=1}^{M}w_i g_i(\mathbf{x}) \tag{2}$$

$$\sum_{i=1}^{M}w_i=1 \tag{3}$$

여기서 λ는 화자 모델의 파라미터를 나타낸다.

$$\lambda=(w_i,\mu_i,\Sigma_i),\ \text{for } i=1,\ldots,M \tag{4}$$

따라서 파라미터$\lambda$의 화자모델에서 특징벡터 **x**를 관측할 확률은 **x**가 각각의 상태에서 출력될 확률을 그 상태에 있을 확률로 가중하여 합한 것이다. 모델의 학습에는 EM(Expectation-Maximization) 알고리즘을 이용한다. 등록화자로부터 발성된 음성에서 추출된 특징벡터가 주어지면 EM 알고리즘은 반복적으로 모델 파라미터를 다듬어서 학습데이터와 모델파라미터가 잘 정합 되도록 한다. EM 알고리즘은 E단계와 M단계로 나누어진다. E(expectation) 단계는 현재의 모델 파라미터와 관측데이터를 이용하여 숨겨진 구조(hidden structure)를 예측하고, M(maximaization) 단계에서는 예측된 숨겨진 구조를 이용하여 파라미터를 재추정 한다.

화자 식별 시스템에서 최종 결과의 결정은 여러 후보 화자들 중에서 가장 유사도가 높은 화자를 선택하면 된다. S명의 화자의 모델 $\lambda_1, \lambda_2, \cdots, \lambda_s$ 이 있을 때 화자식별은

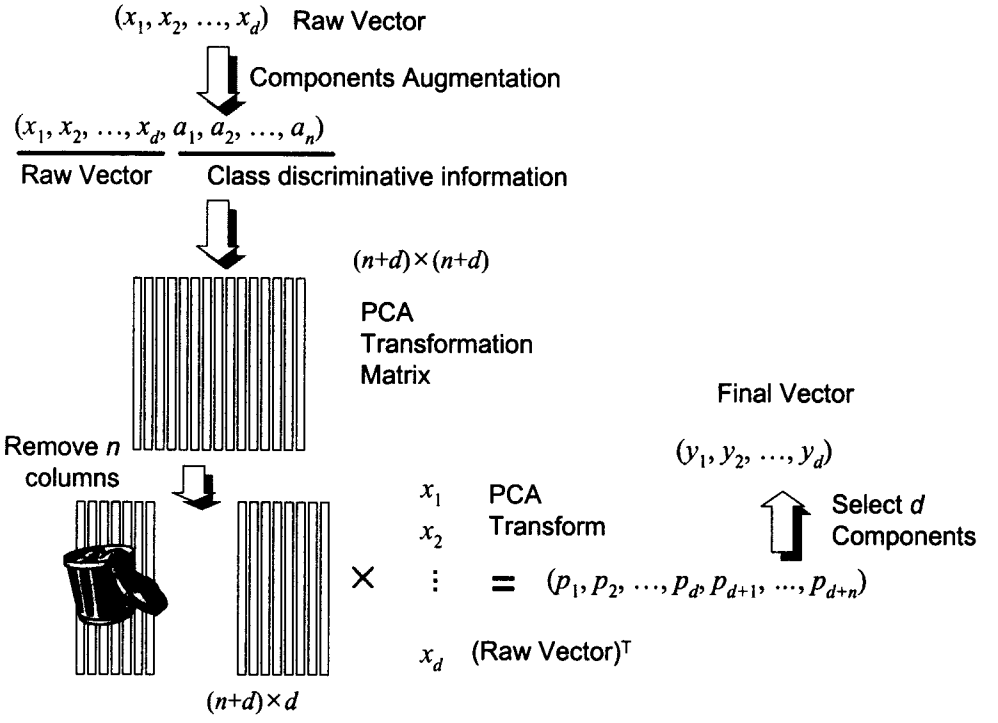
$$\hat{S}=\arg\max_{1\le k\le S}\sum_{t=1}^{T}\log p(\mathbf{x}_t\mid\lambda_k) \tag{5}$$

을 찾는 것이다.

## 3. 제안한 부가주성분분석 (Augmented PCA)

본 연구에서는 불일치 환경에서의 인식 성능을 향상시키기 위하여, 인식 대상 화자 집합에서 혼동되기 쉬운 화자 쌍에 대하여 개별적인 주성분 분석을 수행하였다. 이때, 원래의 특징 벡터 공간의 축에 부가적으로 두 화자를 구분 짓는 축을 추가하여 PCA변환 행렬을 작성하였다. 부가되는 축에는 두 화자를 구분할 수 있도록 상수 r또는 -r을 각각 넣는다. 이렇게 부가정보를 넣는 이유는 잡음이 없는 상태에서 벡터공간에 화자정보를 가진 차원의 축을 추가함으로써 마이크 특성이나 잡음보다 화자의 특성을 잘 표현하기 위한 것이다. 본 실험에서 제안한 부가주

성분분석을 이용한 특징 추출 과정은 다음과 같은 단계를 거친다.




<그림 1> 제안한 부가 주성분 분석 과정

단계 1. 학습데이터의 모든 특징벡터 $x = (x_1, x_2, \ldots, x_d)$에 대하여 다음과 같이 성분을 부가한다.

$$(x_1, x_2, \ldots, x_d, a_1, a_2, \ldots, a_n)$$

여기서, $d$ 는 변환 전 특징 벡터의 차원이고, $n$ 은 분류할 패턴 클래스의 수, 여기서는 화자수가 된다. $x_i$ 는 변환 전 특징 벡터의 $i$번째 원소이고, $a_j$ 는 부가된 원소로서 다음과 같다.

$a_j = r$ 변환할 벡터가 클래스 j 에 속하는 경우

$a_j = -r$ 변환할 벡터가 클래스 j 에 속하지 않는 경우

여기서 $r$은 임의의 상수로, 본 실험에서는 10으로 하였다.

단계 2. 부가된 특징 벡터를 2.2절에서 설명된 과정에 따라 PCA변환하여 변환 행렬 $w'$를 구한다.

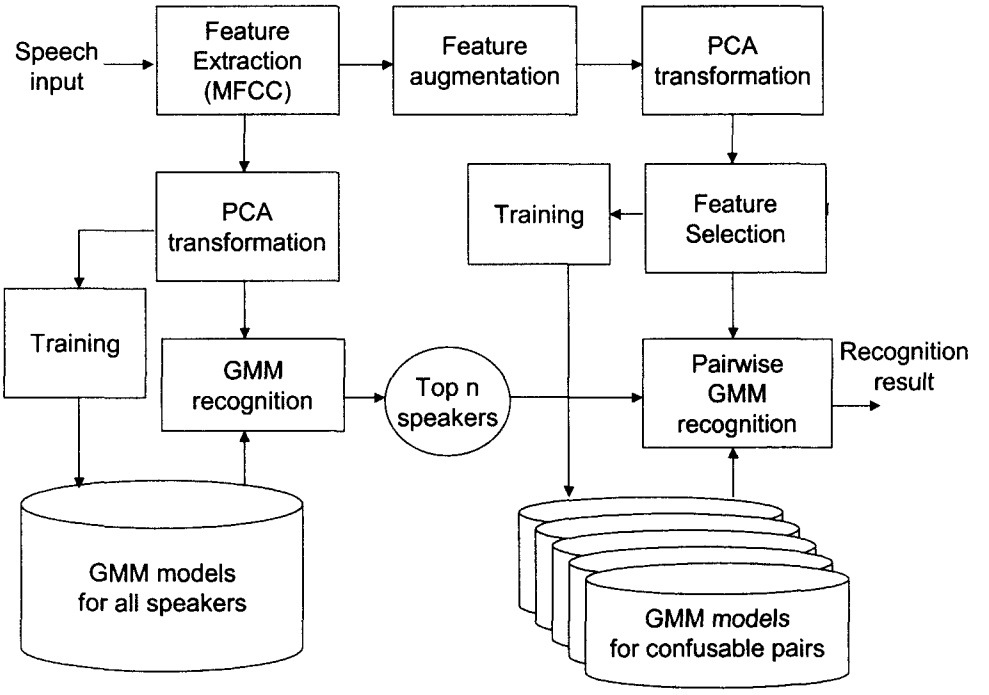
단계 3. 변환 행렬 $w'$에서 부가된 요소에 대응하는 마지막 $n$개의 열을 제거한다. 이것은 단계 1에서 부가되는 요소에 $r$ 또는 $-r$ 대신 모두 0을 대입하여 변환하는 것과 같은 효과를 가진다. 미지의 벡터에 대해서는 클래스 정보를 알

지 못하므로 0을 대입하는 것이다.

단계 4. 학습데이터에 포함된 모든 벡터를 단계 3에서 구한 변환행렬을 이용하여 변환한다. 그 결과 모든 벡터는 $(n+d)$ 차원으로 변환된다.

단계 5. 혼동되기 쉬운 화자의 쌍을 선정하여 두 화자의 변이를 잘 표현하는 축 $d$ 개를 선택한다. 본 실험에서는 각 벡터의 모든 요소에 대하여 화자정보와의 상관계수를 구하여 크기가 큰 $d$개를 선정하였다. 이렇게 하면 최종 결과는 변환 전 벡터와 같은 $d$개의 차원을 가지게 된다.

각각의 화자 쌍에 대하여 단계 5에서 구한 특징벡터로 GMM을 학습시킨다. 인식 테스트 데이터의 특징벡터에 대해서는 마찬가지로 단계 4 와 5를 이용하여 특징을 변환하고 학습된 모델로 인식실험을 수행한다. 즉, 테스트 데이터에 대해서도 단계 3에서 구한 변환행렬과 단계 5에서 선택된 축들을 그대로 이용하여 변환한다. <그림 1>은 제안한 부가주성분분석 과정을 보여주며, <그림2>는 이 방법을 이용한 화자식별 과정을 보여준다. 학습데이터는 PCA 변환 행렬 작성과 GMM학습에 사용되며, 인식데이터는 학습데이터를 이용하여 만들어진 변환 행렬에 의하여 변환된다.



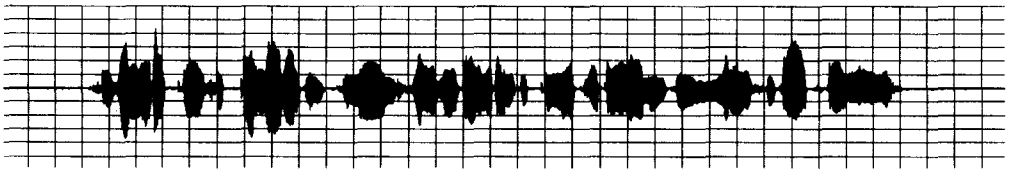

<그림 2> 제안한 부가주성분분석을 이용한 화자식별 과정

# 4. 실험 및 결과

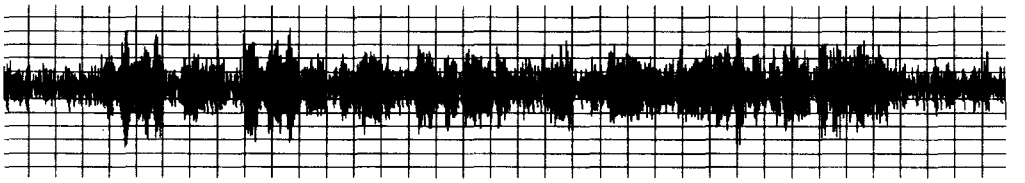
## 4.1. 실험 환경

본 연구에서는 음성정보기술산업지원센터(SITEC)에서 수집한 자동차 화자인증용 음성 DB(CarSpkr01)를 이용하여 실험하였다. 음성은 주행 중인 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-worn SHURE SM-10A, Uni-Cardioid), 콘덴서 마이크(AKG B400-BL, Cardioid), 국산 저가 핸즈프리 마이크(HYUNDAI Handsfree)의 세 종류로 총 8개의 위치에 나누어 장착되었다. 본 연구에서는 다이나믹 마이크로 화자의 입에서 3 cm 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. <그림3>에서 보는 바와 같이 테스트에 사용된 음성 (b)는 입과 비교적 떨어져 있고, 콘덴서 마이크를 사용하였으므로 잡음이 상당히 포함되어 있다. 이에 반하여, 학습에 사용된 음성 (a)는 다이나믹 마이크로 입에 가까이 대고 녹음하여 잡음이 거의 없는 것을 알 수 있다.



(a) 다이나믹 마이크로 녹음된 학습데이터의 예



(b) 동일한 음성의 콘덴서 마이크 입력

<그림 3> 학습과 인식에 사용한 음성 데이터의 비교

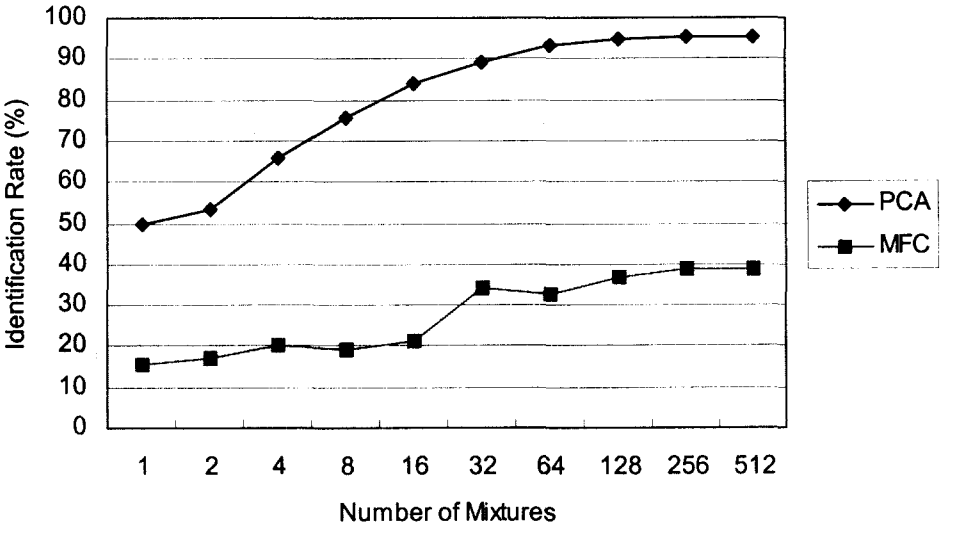
본 데이터는 음소가 고루 분포된 문장 및 단어 세트, 4연 숫자음 세트를 총 30명의 화자가 최초발성, 1일 후, 1주일 후, 1개월 후, 2개월 후의 시차를 두고 총 5회 발성하였다. 한 화자 당 총 발성 수는 250개 이다. 본 연구에서는 최초 발성된 모든 텍스트의 음성을 학습데이터로 사용하고 1주일 후에 발성된 모든 텍스트의

음성을 테스트 데이터로 사용한다.

화자인식을 위한 특징으로는 20차 MFCC(Mel-frequency cepstral coefficients)와 에너지, 이의 1차 및 2차 미분을 사용하고, 채널왜곡을 감소시키기 위하여 CMS (Cepstral mean subtraction) 방법을 사용한다. 화자 모델은 2.2절에서 설명된 GMM (Gaussian mixture model)[6]을 사용한다. 혼합(mixture)수에 따른 성능의 변화를 살펴보기 위하여 혼합수를 1, 2, 4, 8, ..., 512와 같이 두 배씩 증가시키면서 실험하였다. 각 혼합수 별 반복 학습(iteration) 회수는 5회로 하였다.

### 4.2. PCA를 이용한 실험 결과

일반적인 멜켑스트럼 계수를 특징으로 사용하여 GMM으로 학습하고 인식하였을 경우와 여기에 주성분 분석을 적용하였을 경우의 결과를 <그림 4>에서 비교할 수 있다. 주성분 분석을 사용하지 않았을 경우에는 인식률이 40% 이하로, 사용이 불가능할 정도의 성능을 보이며, 혼합수가 증가하여도 인식률이 일정하게 증가하지 않는다. 반면, 주성분 분석을 사용할 경우에는 인식률이 혼합수에 따라 규칙적으로 증가하는 것을 볼 수 있으며, 혼합수 512개 에서는 95% 이상의 성능을 얻을 수 있었다.

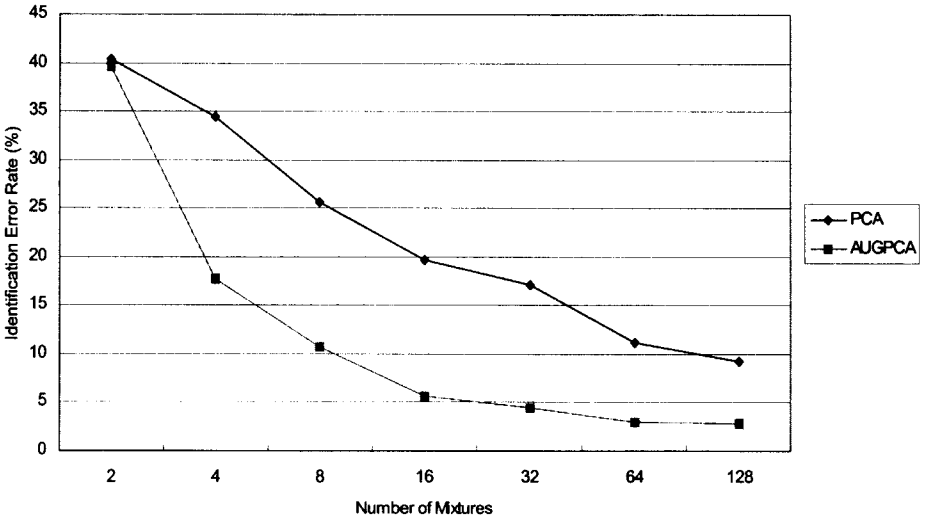



<그림 4> 미이크 불일치 및 잡음 환경에서의 주성분분석(PCA)의 효과

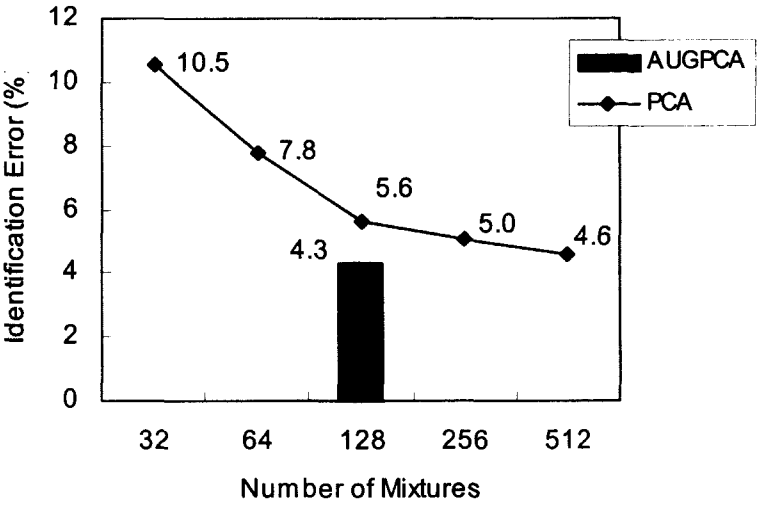
### 4.3. 제안한 부가 PCA를 이용한 실험 결과

제안한 방법의 효과를 확인하기 위하여 기존의 PCA를 적용한 상위 5개의 인식 결과에 대하여 제안한 부가주성분분석을 이용하여 두 화자만을 구분하는 후처리를 하였다. 모든 화자를 대상으로 한 부가PCA변환은 성능의 향상을 가져오지 않으므로 사용하지 않았다. <그림 5>에서, 본 실험에서 가장 혼동되는 두 화자만의 인식 오류가 제안한 방법(AUGPCA로 표시)에 의하여 감소되는 것을 볼 수 있다. <그림 6>에서 보는 바와 같이 전체적으로 혼합(mixture)수가 128개 일 때 인식 오류가 5.4%에서 4.3% 로 줄어 21%의 상대 오류율 감소를 얻을 수 있었다.

기존의 주성분분석 방법만을 사용한 인식과정에서 혼합수를 두 배인 256개로 늘려도 오류는 5.03%로 10% 정도의 상대 오류만 감소하는 것을 알 수 있었다. 이 결과로 볼 때 제안한 방법은 단순히 혼합수를 늘이는 것에 비하여 높은 성능 향상을 얻을 수 있음을 알 수 있었다.




<그림 5> 가장 혼동되기 쉬운 두 화자에 대한 부가주성분분석(AUGPCA)의 효과




<그림 6> 제안한 부가주성분분석을 이용한 전체 화자 식별 오류의 감소

## 5. 결 론

본 논문에서는 언제 어디서나 정보를 주고받을 수 있는 유비쿼터스 환경에서의 화자 식별을 목표로 하여, 잡음이 없는 상태에서 다이나믹 마이크로 녹음한 음성으로 학습한 모델을 주행중인 차량에서 콘덴서 마이크로 녹음한 음성에 대하여 테스트 하는 환경을 설정하였다. 먼저 마이크의 특성이나 잡음에 대한 적응 과정을 거치지 않은 상태에서 주성분분석을 이용한 특징 변환에 의하여 인식률이 40% 이하에서 95% 이상으로 향상되는 것을 확인할 수 있었다. 또한, 특징벡터에 화자 정보를 부가하여 주성분분석을 함으로써 화자식별에 효과적인 축으로의 사상을 유도하는 부가주성분분석을 제안하였고, 화자의 쌍에 대하여 제안한 방법을 적용하여 GMM을 학습하는 방법을 사용한 실험결과 21%의 상대오류 감소를 얻을 수 있었다.

향후 계획으로는 부가 주성분 분석의 수학적인 분석, 부가하는 상수의 크기 변화에 대한 영향 조사, 인식에 유효한 축의 선정 방법 비교, MFCC계수의 초기 차수 변형, 선형 판별분석(LDA: Linear Discriminant Analysis)과의 비교 등이 있다.

## 감사의 글

실험에 큰 도움을 준 서울시립대학교 컴퓨터과학부 3학년 신연식 군에게 감사의 뜻을 전합니다.

## 참 고 문 헌

[1] J. P. Campbell, Jr., "Speaker recognition: a tutorial", *in Proc. of the IEEE*, Vol. 85, No. 9, pp. 1437-1462, Sep. 1997.

[2] A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recognition, Kluwer Academic Publishers, Boston, 1993.

[3] S.-N. Tsai and L.-S. Lee, "Improved robust features for speech recognition by integrating time-frequency principal components (TFPC) and histogram equalization (HEQ)", *in Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 297 - 302, Nov.-Dec. 2003.

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

[5] P. Ding and L. Zhang, "Speaker recognition using principal component analysis", *in Proc. of International Conference on Neural Information Processing*, Shanghai, China, Nov. 2001.

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

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

접수일자: 2005년 5월 25일
게재결정: 2005년 6월 14일

▶ 유하진(Ha-Jin Yu)
주소: 서울시 동대문구 전농동 90번지 서울시립대학교
소속: 서울시립대학교 컴퓨터과학부
전화: 02)2210-5613
FAX: 02)2210-5275
E-mail: hjyu@venus.uos.ac.kr