말소리와 음성과학 제1권 제1호 (2009)
pp. 69~73

화자식별을 위한 전역 공분산에 기반한 주성분분석

Global Covariance based Principal Component Analysis for Speaker Identification

서 창 우[1] · 임 영 환[2]
Seo, Changwoo · Lim, Younghwan

ABSTRACT

This paper proposes an efficient global covariance-based principal component analysis (GCPCA) for speaker identification. Principal component analysis (PCA) is a feature extraction method which reduces the dimension of the feature vectors and the correlation among the feature vectors by projecting the original feature space into a small subspace through a transformation. However, it requires a larger amount of training data when performing PCA to find the eigenvalue and eigenvector matrix using the full covariance matrix by each speaker. The proposed method first calculates the global covariance matrix using training data of all speakers. It then finds the eigenvalue matrix and the corresponding eigenvector matrix from the global covariance matrix. Compared to conventional PCA and Gaussian mixture model (GMM) methods, the proposed method shows better performance while requiring less storage space and complexity in speaker identification.

Keywords: speaker identification, principal component analysis, global covariance, Gaussian mixture model, eigenvalue, eigenvactor

1. 서 론

전체 공분산 행렬을 이용한 주성분분석(PCA: principal component analysis) 방법은 화자인식(speaker recognition)에서 널리 연구되고 있다. 화자인식은 크게 화자식별(speaker identification)과 화자확인(speaker verification)으로 나눌 수 있다. 화자식별은 발성된 음성 신호가 등록된 화자들 중에서 어떤 화자인지를 찾아내는 방법이고, 화자확인은 발성된 음성 신호가 등록된 화자의 음성과 일치하는지를 판정하는 것으로, 발성한 화자와 등록된 화자와의 확인 과정을 통하여 문턱값보다 유사도가 큰 경우 수락하고 그렇지 않으면 거절하는 것이다.

화자인식에서 성능을 향상시키기 위해서는 높은 차수의 특징벡터가 요구된다. 또한 음성 신호로부터 추출된 특징벡터의 성분은 상관성이 높기 때문에 좋은 근사화를 얻기 위해서 많은 양의 혼합성분을 필요로 한다. 그러나 특징벡터의 차수와 혼합성분 개수의 증가는 또 다른 문제를 발생시킬 수 있다. 예로서, 높은 차수의 특징벡터를 이용하는 인식기는 분류기를 특성화하기 위해서 많은 파라미터와 저장 공간을 요구한다. 이것은 결과적으로 계산량을 증가시키기 때문에 실시간 구현을 더 어렵게 한다[1]-[3]. 무엇보다도, 많은 양의 음성 데이터가 학습을 위해서 요구된다. 그러나 음성인식(speech recognition)과 달리 GMM(Gaussian mixture model)을 기본으로한 화자인식에서는 개인별 녹음된 많은 데이터를 획득하는 것은 사실상 불가능하다[2].

이런 문제점에 대한 해결 방법으로 특징벡터의 차수를 줄이고 신호의 상관성을 제거하기 위한 주성분분석(PCA)[4], [5], 국부 주성분분석(local PCA)[1], 그리고 강인한 주성분분석(robust PCA)[6]을 이용한 방법이 제안되었다. Local PCA는 VQ(vector quantization)에 의한 특징벡터의 공간을 여러 개로 분리한 후 전달함수를 통한 작은 부분공간으로 원래의 특징 공간을 사영(projection)하는 방법이다. Robust PCA는 특징벡터에 이상치가 존재할 경우 M-추정에 의하여 강인한 공분산 행렬을 재추정하여 얻어진 고유벡터로부터 변환 행렬을 구하여 감소된 차원을 갖는 효과적인 방법이다. 그러나 PCA, local PCA, 그리고 robust

1) 숭실대학교 cwseo@ssu.ac.kr, 교신저자
2) 숭실대학교 yhlim@ssu.ac.kr

접수일자: 2009년 2월 3일
수정일자: 2009년 3월 8일
게재결정: 2009년 3월 15일

PCA 방법도 특징벡터의 공분산을 통한 개인별 고유치(eigenvalue)와 고유벡터(eigenvector)를 계산할 때, 많은 양의 학습 데이터를 요구하고 있다.

본 논문에서는 이런 문제를 해결하기 위해서 전역 공분산 행렬(global covariance matrix)을 기본으로 하는 PCA를 이용한 화자식별을 제안한다. 제안된 방법은 먼저 식별에 참여한 모든 화자의 학습 데이터를 이용하여 전역 공분산 행렬을 구한다. 마지막으로, 전역 공분산 행렬을 이용하여 고유치와 대응하는 고유벡터를 구하는 방법이다. 학습 및 테스트 과정에서는 개인별 PCA 계수를 사용하는 대신 전역 공분산 행렬로부터 구해진 고유벡터를 사용해서 새로운 영역으로 사영(projection)한다. 이 경우에 충분치 못한 학습데이터에 의해서 발생할 수 있는 변환 문제는 전역 공분산을 통한 정규화된 고유벡터로부터 해결될 수 있다. 비록 제안된 방법이 전체 학습 데이터에 대한 변환 행렬을 얻기 위한 특별한 단계를 요구하지만, 이들 단계에 대한 계산 비용은 대부분의 시간 소비가 학습의 EM(Expectation-Maximization) 반복 과정에서 일어나기 때문에 무시할 수 있다. 또한 EM 반복 과정에서도 특징벡터의 차원 감소에 의한 학습 시간을 줄일 수 있었다. 제안된 방법의 우수성을 확인하기 위해서 제안된 GCPCA, 일반적인 PCA, 그리고 GMM 방법을 비교 실험의 결과를 통해서 설명하였다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 특징벡터의 차원감소를 위한 PCA를 기술하였다. 3장에서는 전역 공분산 행렬에 기반한 GCPCA 방법을 제안하고, 4장에서는 화자인식에 널리 연구되고 있는 GMM을 설명하였다. 5장과 6장에서는 제안한 방법의 성능을 확인하기 위한 실험 결과와 결론을 기술하였다.

2. 주성분분석(PCA)

PCA는 여러 개의 변수들에 대하여 얻어진 다변량 자료를 분석대상으로 하여 다차원적인 변수들을 축소, 요약하는 차원의 단순화와 함께 일반적으로 서로 상관관계가 있는 반응 변수들 간의 복잡한 구조를 분석하는데 목적이 있다[2], [3]. 따라서 입력된 음성 데이터로부터 추출된 특징벡터들을 상관관계가 없는 새로운 좌표계로 선형변환 시켜 좌표 변환에 의해 새롭게 변형된 성분을 계산한다.

상관성이 밀접한 각 화자 s의 특징벡터 $x_s(t)\in R^L$이고 길이가 T인 학습열이라 하자. PCA는 $x_s(t)$가 가지고 있는 상관성이 높은 특징벡터를 상관성이 없는 새로운 K-차원($K \le L$)의 벡터 $y_s(t)$로 축약시키는 기법이다. 즉, L-차원의 변수를 K-차원으로 축소하여 선형 변환된 $y_s(t)$를 구하고자 하는 것이다. 데이터 집합에서 주성분을 계산하는 일반적인 방법은 공분산 행렬(covariance matrix)의 고유치와 고유벡터를 구한다. 먼저 각 화자 s에 대한 선형변환 행렬 Φ_s를 구하기 위해서는 기준 벡터 γ_s와 공분산 행렬 Σ_s을 구해야 한다.

$$\gamma_s = \frac{1}{T}\sum_{t=1}^{T} x_s(t) \tag{1}$$

$$\Sigma_s = \frac{1}{T}\sum_{t=1}^{T} (x_s(t)-\gamma_s)^{Tr}(x_s(t)-\gamma_s) \tag{2}$$

여기서 Tr은 전치 행렬이다. 식(2)의 공분산 행렬 Σ_s은 다음과 같이 고유치 $\lambda_{s,i}$와 고유벡터 $\phi_{s,i}$로 나타낼 수 있다.

$$\Sigma_s = \sum_{i=1}^{L} \lambda_{s,i}\phi_{s,i}^{Tr}\phi_{s,i} \tag{3}$$

여기서 고유벡터 $\phi_{s,i}$는 변환행렬 Φ_s의 i번째 열벡터(row vector)이다. Φ_s는 $L\times L$인 직교 행렬($\Phi_s^{Tr}\Phi_s = I$)을 이룬다. 위의 설명과 같이 t번째 시퀀스의 특징벡터 $x_s(t)$와 K-차원 주성분 $\Phi_{s,K}$의 관계는 다음과 같다.

$$y_s(t) = \Phi_{s,K}^{Tr} x_s(t) \tag{4}$$

K-차원 주성분 벡터의 정보 비율(I)은 다음 식과 같이 구할 수 있다.

$$I = \frac{\sum_{i=1}^{K}\lambda_i}{\sum_{i=1}^{L}\lambda_i} \tag{5}$$

이 정보비율에 따라 고유값이 큰 것부터 K-차원만을 선택하여 $\Phi_{s,K}$를 구하고, 식(4)와 같이 적용할 수 있다[7].

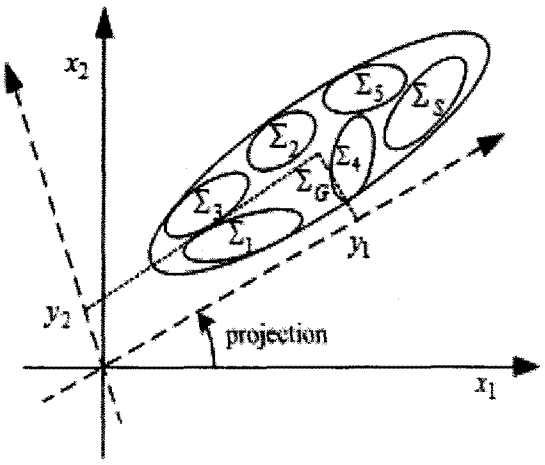

그림 1. 2차원 공간에서의 전역 공분산에 기반한 주성분분석의 사영.
Figure 1. Projection of the global covariance based PCA in 2D space

3. 전역 공분산에 기반한 주성분분석

<그림1>은 2차원 공간에서 전역 공분산에 기반한 주성분분석(GCPCA)을 나타낸 것이다. 여기서 작은 타원은 2장에서 소개된 일반적인 PCA 방법을 위한 개인별 특징벡터에 대한 공분산 $\{\Sigma_1,\Sigma_2,\cdots,\Sigma_S\}$을 나타낸 것이고, 작은 타원을 포함하는 큰 타원은 전체 학습 데이터에서 계산된 전역 공분산행렬 Σ_G을 나타낸 것이다. 변환 좌표계의 최대 분산치에 해당하는 성분이 제1 주성분이 되며 이에 해당하는 기저벡터가 특징벡터로 추출된다. 따라서 전역 공분산 행렬은 학습 데이터가 충분치 못한 경우에도 화자식별에 참가한 모든 학습 데이터를 사용하기 때문에 발성 길이에 상관없이 충분한 데이터를 얻을 수 있다.

각 화자 s의 특징벡터 $x_s(t)\in R^L$는 L-차원 벡터로 길이가 T_s인 학습열의 집합을 $X=\{x_s(t),t=1,\cdots,T_s,s=1,2,\cdots,S\}$이라 가정하자. 전체 학습 데이터에 대한 전역 공분산 행렬을 구하기 위한 기준벡터 γ_G는 다음과 같이 구할 수 있다.

$$\gamma_G=\frac{1}{ST_s}\sum_{s=1}^{S}\sum_{t=1}^{T_s}x_s(t) \tag{6}$$

여기서 S는 식별의 학습에 참여한 화자 수이다. 다음으로 기준벡터 γ_G에 대한 특징벡터의 전역 공분산 행렬 Σ_G은 아래와 같이 계산된다.

$$\Sigma_G=\frac{1}{ST_s}\sum_{s=1}^{S}\sum_{t=1}^{T_s}(x_s(t)-\gamma_G)^{Tr}(x_s(t)-\gamma_G) \tag{7}$$

식(7)로부터 추정된 전역 공분산 행렬은 PCA를 수행하기 위한 고유치 행렬과 대응하는 고유벡터 행렬을 구하기 위해서 사용된다. 전역 공분산 행렬 Σ_G는 식(3)과 같이 고유치 $\lambda_{G,i}$와 고유벡터 $\phi_{G,i}$로 분해(decomposition)할 수 있다.

$$\Sigma_G=\sum_{i=1}^{L}\lambda_{G,i}\phi_{G,i}^{Tr}\phi_{G,i} \tag{8}$$

변환 축에서의 중요한 것은 고유치 값의 크기 성분에 의해서 K-차원의 벡터가 선택되기 때문에 고유치 값에 대응하는 고유벡터의 변환 행렬 $\Phi_{G,K}$는 아래와 같다.

$$\Phi_{G,K}=[\phi_{G,1},\phi_{G,2},\cdots,\phi_{G,K}] \tag{9}$$

따라서 GCPCA에 대한 학습과 테스트 과정에서의 각 입력 특징벡터는 다음과 같이 변환된다.

$$y_s(t)=\Phi_{G,K}^{Tr}\,x_s(t) \tag{10}$$

여기서 $y_s(t)$는 GCPCA에 의한 K-차 축소 변환된 특징벡터이다.

4. 화자식별을 위한 GMM

PCA 변환을 통한 $y(t)\in R^K$인 관측열의 길이가 T인 학습벡터열 $Y=\{y(t),t=1,\cdots,T\}$에 대한 가우시안 혼합밀도 $p(y(t)|\theta)$는 다음과 같이 M성분 밀도의 가중화된 합으로 나타낼 수 있다.

$$p(y(t)|\theta)=\sum_{i=1}^{M}w_ib_i(y(t)) \tag{11}$$

여기서,

$$b_i(y(t))=\frac{1}{(2\pi)^{K/2}|\Sigma_i|^{1/2}}$$

$$\cdot\ \exp\left[-\frac{1}{2}(y(t)-\mu_i)^{Tr}\Sigma_i^{-1}(y(t)-\mu_i)\right] \tag{12}$$

여기서 μ_i은 평균벡터(mean vector), Σ_i은 공분산 행렬(covariance matrix)이다. 그리고 w_i는 $\sum_{i=1}^{M}w_i=1$조건을 만족시키는 M차 혼합성분을 위한 가중치벡터(weight vector)를 나타낸다.

길이가 T인 특징벡터 Y가 주어질 때, 화자모델을 위한 GMM은 모든 성분 밀도로부터 가중치 w_i, 평균벡터 μ_i, 그리고 공분산 행렬 Σ_i은 다음과 같이 파라미터화 할 수 있다.

$$\theta=\{w_i,\mu_i,\Sigma_i\},i=1,2,\cdots,M \tag{13}$$

이때, 파라미터 θ에 대한 GMM의 유사도는 다음과 같이 나타낼 수 있다.

$$p(Y|\theta)=\prod_{t=1}^{T}p(y(t)|\theta) \tag{14}$$

식(13)의 파라미터 추정은 EM 알고리즘을 반복적으로 사용하여 다음과 같이 계산할 수 있다[2], [8].

$$\hat{w}_i=\frac{1}{T}\sum_{t=1}^{T}p(i|y(t),\theta) \tag{15}$$

$$\hat{\mu}_i=\frac{\sum_{t=1}^{T}p(i|y(t),\theta)y(t)}{\sum_{t=1}^{T}p(i|y(t),\theta)} \tag{16}$$

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^{T} p(i|y(t),\theta)(y(t)-\mu_i)(y(t)-\mu_i)^{Tr}}{\sum_{t=1}^{T} p(i|y(t),\theta)} \quad (17)$$

음향학적 분류를 위한 사후 확률(a posteriori probability)는 다음과 같이 나타낼 수 있다.

$$p(i|y(t),\theta) = \frac{w_i p(y(t))}{\sum_{m=1}^{M} w_m p(y(t))} \quad (18)$$

화자식별에서 S명의 화자는 각각 GMM의 파라미터 $\theta_1,\theta_2,\cdots,\theta_S$로 나타낼 수 있다. 화자식별의 목적은 다음과 같이 주어진 변환 특징 열로부터 최대 사후 확률을 가지는 화자 $\hat{s}$를 추정하는 것이다.

$$\hat{s} = \max_{1 \le s \le S} \sum_{t=1}^{T} \log p(y(t)|\theta_s) \quad (19)$$

5. 실험 결과

화자식별 실험에서 제안된 알고리즘의 성능을 검증하기 위해서 제안된 GCPCA, 일반적인 PCA, 그리고 GMM 방법을 사용하여 실험하였다. 실험에 사용된 음성 데이터는 200명의 화자(남자:100, 여자:100)로부터 획득하였고, 개인별 화자의 데이터는 1세션을 주 단위로 하여 3세션 동안 15개(매주 5문장)의 데이터를 구하였다. 처음 2주 동안 수집한 데이터는 학습에 사용하였고, 마지막 주에 수집한 데이터를 테스트에 사용하였다. 수집한 데이터는 모든 화자에 대해서 동일한 고정 문장인 "열려라 참깨"를 사용하였다.

표 1. 제안된 GCPCA, 일반적인 PCA, 그리고 GMM의 파라미터 수.
Table 1. Required number of parameters for the proposed GCPCA, the conventional PCA, and the GMM methods

	Parameters
Proposed GCPCA	$SM(2K+1)+LK$
Conventional PCA	$S(M(2K+1)+LK)$
GMM	$SM(2L+1)$

음성의 분석과정에서 샘플링 주파수는 11.025kHz, 16bit 분해능, 12차 mel-frequency cepstral coefficient(MFCC) 그리고 13차 델타 켑스트럼(delta cepstrum)[9]을 사용하였다. 음성분석 프레임은 16ms이고 50% 중첩을 적용하였다.

<표1>은 식별을 위한 제안된 GCPCA, 일반적인 PCA, 그리고 GMM 방법에서 요구되는 전체 파라미터의 저장 공간을 나타낸 것이다. 비록 제안된 방법에서 변환 행렬을 위한 LK개를 요구하지만, 이것은 일반적인 PCA와 GMM 방법보다 훨씬 작은 파라미터를 요구하고 있다. 예로서, 실험에서 적용한 화자 S=200, 혼합성분 M=16, 변환 전 특징벡터 L=25, 그리고 변환 후 특징벡터 K=20일 때, 제안된 GCPCA, 일반적인 PCA, 그리고 GMM은 각각 131,700, 231,200, 그리고 163,200의 파라미터를 요구한다. 따라서 제안된 방법은 PCA와 일반적인 GMM 보다 평균 43%와 20%의 저장 공간을 줄일 수 있다.

<그림2>은 화자식별의 성능에서 특징벡터의 차수와 혼합성분의 개수를 나타낸 것이다. 실험결과로부터 같은 차수의 특징벡터와 혼합성분을 사용할 때, 제안된 GCPCA(25)은 PCA(25)와 GMM(25)과 비교했을 때 평균 1.13%와 2.09% 높은 식별 율을 보여주었다. 특히 제안된 GCPCA의 특징벡터 차수 K=20이고 M=16일 때, PCA(25)와 GMM(25)보다 파라미터의 저장 용량이 약 50%와 20% 줄어들었지만, 식별 율에서는 오히려 약 0.07%와 1.06% 향상되었다. 그리고 PCA(25)와 GMM(25)을 비교했을 때, PCA(25) 방법이 계산량은 증가되었지만 상관성 제거에 의해서 약 0.95% 향상되었다. 비록 제안된 방법이 전체 학습 데이터에 대한 변환 행렬을 얻기 위한 특별한 단계를 요구하지만, 이들 단계들에 대한 계산 비용은 대부분의 시간 소비가 학습의 EM 반복 과정에서 일어나기 때문에 무시할 수 있다. 따라서 제안된 방법은 화자인식의 특성상 충분한 학습 데이터를 확보할 수 없고 그리고 편리성 측면의 발성 길이가 짧은 경우에 있어서 효과적이라고 할 수 있다.

6. 결 론

본 논문에서는 화자식별에서 전역 공분산 행렬에 기반한 PCA 방법을 제안하였다. 제안된 방법은 먼저 식별에 참여한 모든 화자의 학습 데이터를 이용하여 전역 공분산 행렬을 구한다. 마지막으로, 전역 공분산 행렬을 이용하여 고유치와 대응하는 고유벡터를 구하는 방법이다. 학습과 테스트 과정에서는 개인별 PCA 계수를 사용하는 대신 전역 공분산 행렬로부터 구해진 고유벡터를 사용해서 새로운 영역으로 사영한다. 이 경우 충분치 못한 학습 데이터에 의해서 발생할 수 있는 변환 문제는 전역 공분산 행렬을 통한 정규화된 고유벡터로부터 해결될 수 있다. 제안된 방법의 우수성을 확인하기 위해서 화자식별에서 일반적인 PCA와 GMM 방법을 비교했을 때, 훨씬 작은 계산량으로 우수한 성능을 보였다.

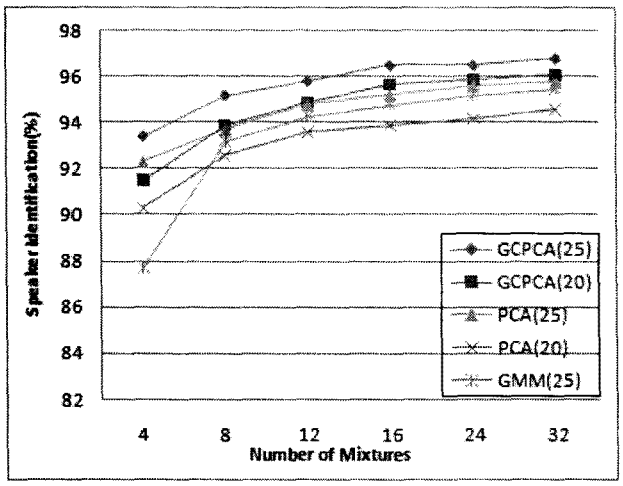

그림 2. 제안된 GCPCA(25)/(20), 일반적인 PCA(25)/(20), 그리고 GMM(25)을 이용할 때의 화자식별 성능.

Figure 2. Speaker identification performance using proposed GCPCA(25, 20), PCA(25, 20), and GMM(25)

감사의 글

본 논문은 2009년도 숭실대학교의 교내연구비 지원으로 이루어졌습니다.

참 고 문 헌

[1] C. Seo, K.Y. Lee, and J. Lee, "GMM based on Local PCA for Speaker Identification", *Electronics Letters*, Vol. 37, No. 24, pp. 1486-1488, 2001.

[2] D. Reynolds, and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. on SAP*, Vol. 3, No. 1, pp. 72-82, 1995.

[3] L. Liu, and J. He, "On the use of orthogonal GMM in speaker recognition", *International Conference on ASSP*, pp. 845-849, 1999.

[4] Y. Ariki, S. Tagashira, and M. Nishijima, "Speaker recognition and speaker normalization by projection to speaker subspace", *International Conference on ASSP*, pp. 319-322, 1996.

[5] N. Kambhatla, and T.K. Leen, "Dimension reduction by local PCA", *Neural Computation*, Vol. 9, pp. 1493-1503, 1997.

[6] Y. Lee, C. Seo, S. Kang, and K.Y. Lee, "RPCA-GMM for Speaker Identification", *Journal of ASK*, Vol. 22, No. 7, pp. 519-527, 2003.
(이윤정, 서창우, 강상기, 이기용, "화자식별을 위한 강인한 주성분 분석 가우시안 혼합 모델", 한국음향학회, Vol. 22, No. 7, pp. 519-527, 2003.)

[7] B.N. Flury, "Common Principal Components in k Groups", JASA, Vol. 79, No. 388, pp. 892-898, Dec, 1984.

[8] A. Dempster, N. Laird, and D. Doubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Stat. Sco.*, Vol. 29, pp. 1-38, 1977.

[9] S. Young, *The HTK Book, Cambridge University*, 2001.

• **서창우 (Seo, Changwoo)** 교신저자
숭실대학교 글로벌 미디어학부
서울시 동작구 상도동 511번지
Tel: 02-826-9872 Fax: 02-826-9872
Email: cwseo@ssu.ac.kr
관심분야: 음성신호처리, 멀티미디어, 모바일 시스템
2008~현재 글로벌 미디어학부 연구교수

• **임영환 (Lim, Younghwan)**
숭실대학교 글로벌 미디어학부
서울시 동작구 상도동 511번지
Tel: 02-820-0685 Fax: 02-820-0685
Email: yhlim@ssu.ac.kr
관심분야: 멀티미디어, 모바일 시스템
1996~현재 글로벌 미디어학부 교수