강인한 VQ-PCA에 기반한 효율적인 화자 식별

Efficient Speaker Identification based on Robust VQ-PCA

이 기 용
Ki-Yong Lee

요 약

본 논문에서는, 효율적인 화자 식별을 위하여 강인한 벡터 양자화 주성분 분석을 제안하였다. 제안된 방법은 화자 식별에서 특징벡터의 학습을 위한 고차원(high dimension) 문제와 이상치(Outlier)에 대한 문제를 해결 하기위하여 제안 되었다. 먼저, 제안된 방법은 M-추정을 이용하여 강인한 벡터 양자화(Vector Quantization : VQ) 에 의한 몇 개의 분리된 영역으로 데이터 공간을 나눈다. 분리된 각 영역에서 공분산 행렬로부터 강인한 주성분 분석(Principal Component Analysis : PCA)이 얻어지게 된다. 마지막으로, 각 영역에서 강인한 PCA에 의하여 줄어든 차원을 갖는 변환된 특징 벡터로부터 화자의 가우시안 혼합 모델(Gaussian Mixture Model : GMM)을 구한다. 제안된 방법은 같은 성능하에서 대각 공분산 행렬을 갖는 전형적인 GMM방법과 비교할 때 더빠른 결과를 얻었으며, 데이터의 저장공간을 줄일 수 있었을 뿐 아니라, 이상치가 존재할 경우에 더욱 강인하였다.

Abstract

In this paper, an efficient speaker identification based on robust vector quantizationprincipal component analysis (VQ-PCA) is proposed to solve the problems from outliers and high dimensionality of training feature vectors in speaker identification. Firstly, the proposed method partitions the data space into several disjoint regions by roust VQ based on M-estimation. Secondly, the robust PCA is obtained from the covariance matrix in each region. Finally, our method obtains the Gaussian Mixture model (GMM) for speaker from the transformed feature vectors with reduced dimension by the robust PCA in each region. Compared to the conventional GMM with diagonal covariance matrix, under the same performance, the proposed method gives faster results with less storage and, moreover, shows robust performance to outliers.

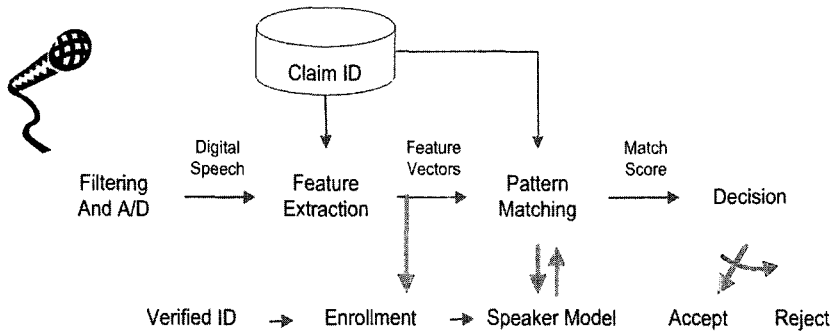
Keyword : Speaker identification, GMM, M-estimation, VQ, PCA

1. 서 론

최근, 화자인식을 위하여 대각 공분산 행렬을 갖는 가우시안 혼합 모델 방법이 점차 많이 사용되고 있다[1]. 화자인식은 화자식별(speaker identification)과 화자확인(speaker verification)으로 나눌 수 있다. 화자식별은 발성된 음성 신호가 등록된 화자들 중에서 가장 유사도가 높은 화자를 선택하여 발성한 화자가 누구인지 결정하는 것이고, 화자확인은 발성된 음성신호가 등록된 화자의 음성과 일치하는지를 판정 하는 것으로, 발성한 화자와 등록된 화자와의 확인과정을 통하여 문턱값(threshold)보다 유사도가 큰 경우 수락(accept)하고, 문턱값보다 유사도가 작은 경우 거절(reject)하는 방법이다. 또한 화자인식을 발성방법에 따라 분류하면, 발성할 단어나 문장이 고정 음운에 기반을 둔 공통적 특징의 개인 차이를 평가하는 문장 종속(text-dependent)형과 발성할 단어나 문장이 정해져 있지 않고 자유롭게 발성 음성 신호로부터 음운 정보를 제거한 화자 정보만을 사용하는 문장 독립(text-independent)형으로 분류된다. 이러한 화자인식 알고리즘은 폰뱅킹, 기밀 정보 지역에 대한 보안 제어, 컴퓨터에 원격 접근등과 같은 서비스에 화자의 신원을 확인하기 위해 음성을 사용한 화자 인식방법이 점차 기대되고 있다. 화자인식을 위한 또 다른 중요한 응용에는 법의학 목적을 위하여 사용된다[2].

특징 벡터집합에서 화자 인식의 성능을 향상 시키

* 정 회 원 : 숭실대학교 정보통신 전자공학부 교수
kylee@ssu.ac.kr(제 1저자)



〈그림 1〉 화자 인식 시스템[3]

기 위해서는 고차원이 바람직하다. 뿐만 아니라, 음성 신호의 특징 벡터들의 좋은 근사치값을 구하기 위해서 특징벡터 요소의 상관관계를 표현 하는 많은 수의 혼합성분들이 필요하다[4]. 그러나, 특징벡터의 차원과 혼합성분 개수의 증가는 인식기에서 분류기의 특성을 나타내기 위하여 더 많은 파라미터들이 요구되는 문제들을 야기시킨다. 이러한 것은 더 많은 저장 공간과 계산량의 복잡성을 가져온다. 그 결과, 실시간 구현이 어렵게 되고, 비용도 많이 들게 된다. 또한, 그러한 인식기를 위한 학습을 위해서는 많은 음성데이터가 요구되는데, 이는 현실적으로 불가능하다.

특징벡터의 차원 감소를 위하여, PCA에 기초한 화자인식 방법은 [5], [6]에 제안 되었다. PCA 는 특징벡터의 차원을 감소하고, 변환을 통하여 더 작은 부공간으로 원래의 특징 벡터 공간을 투영시킴으로써 특징 벡터들 사이에 상관관계를 줄이기 위한 특징을 추출하는 방법 중에 하나이다. 그러나, 전형적인 PCA와GMM방법은 관찰되는 값에 이상치가 존재할 경우 매우 민감하다. 따라서, 학습 특징 벡터에 이상치들이 존재할 경우 PCA와 GMM 방법을 이용한 화자인식 시스템은 성능의 저하를 가져 올 수 있다.

본 논문에서는 화자 식별에서 이상치와 고차원 문제를 해결하기 위하여 강인한 VQ-PCA를 이용한 효율적인 GMM 방법을 제안하였다. 먼저, 제안된 방법에서는 특징벡터 공간을 강인한 VQ를 사용하여 여러 개의 분리된 영역으로 나눈다. 강인한 VQ에서는 각 영역에서 정확한 참조 벡터를 얻기 위하여 M-추정에 의한 강인한 K-평균 알고리즘이 적용되었다[7]. 다음으로, 줄어든 차원을 갖은 새로운 특징 벡터를 추출하기 위해, 각 영역에서 특징 벡터의 강인한 공분산 행렬의 고유벡터를 이용하여 구해진 변환 행렬을 사용하여 원래의 n차원 특징 벡터를 L<n 차원 선형 부공간으로 변환시킨다. 마지막으로, 이렇게 변환된 특징 벡터들로부터 대각 공분산 행렬을 갖는 GMM이 얻어진다. 제안된 방법과 전형적인 GMM 과의 실험을 비교하여 제안된 방법이 더 강인하고 효율적임을 보였다.

2. 강인한 VQ-PCA

x_i는 n-차원 특징벡터이고, $X=\{x_1,\ldots,x_T\}$는 화자 데이터의 특징 벡터 공간이라고 하자. 특징벡터 집합 X는 유클리안 거리에 의하여 여러 개의 분리된 영역으로 나누어 진다. 그러면, 양자화기는 각 영역 R^j에 해당하는 K개의 참조 벡터 $r_j, j=1,\ldots,K$들의 집합으로 구성된다. 각 영역 R^j는 모든 X가 다른 참조벡터들보다 r_j에 가장 가까운 곳에 놓여 있다. 특징벡터에 이상치가 존재할 때, 정확한 참조 벡터와 각 영역에서 공분한 행렬을 얻기 위하여 M-추정에 의한 강인한 K-평균 알고리즘이 적용된다. 초기 참조 벡터는 전형적인 K-평균 알고리즘으로부터 구한다[8].

n-번째 반복 과정에서, 각 영역에서 참조 벡터와 공분산 행렬의 강인한 추정은 다음식으로 정의 된다.

$$r_j^n = \frac{\sum_{t=1}^{T} c_{t_j}^n x_t}{\sum_{t=1}^{T} c_{t_j}^n} \tag{1}$$

$$\Sigma_j^n = \frac{\sum_{t=1}^{T} c_{t_j}^n (x_t - r_j^n)(x_t - r_j^n)^T}{\sum_{t=1}^{T} c_{t_j}^n}, \quad x_t \in R^l,\ j = 1, \cdots, K \tag{2}$$

여기에서, $c_{t_j}^n$는 모든 t와 j에 대하여, $n = 0$일때 $c_{t_j}^0 = 1$인 가중치 함수이다[9].

$$c_{t_j}^n = \begin{cases} 1 & if \quad d^2(x_t, r_j^{n-1}, \Sigma_j^{n-1}) \le \alpha \\ 0 & otherwise \end{cases} \tag{3}$$

식(3)에서, $d^2(x_t, r_j^{n-1}, \Sigma_j^{n-1})$는 통계적 거리로 $d^2(x_t, r_j^{n-1}, \Sigma_j^{n-1}) = (x_t - r_j^{n-1})(\Sigma_j^{n-1})^{-1}(x_t - r_j^{n-1})^T$이고, α는

$$\alpha = \frac{Median_t (x_t - r_j^0)(\Sigma_j^0)^{-1}(x_t - r_j^0)^T}{0.6745} \tag{4}$$

로 정의되는 문턱치 값이다.

만약 X에 이상치가 존재한다면, 식 (1)과 식(2)에서 이상치의 영향은 각 영역에서 $c_{t_j}^n$에 의해 줄어들게 된다. 그러면, 강인한 PCA가 식(2)의 고유값과 고유벡터들에 의해 얻어지게 된다. 어떤 변형된 축의 중요성은 그것의 고유값의 크기로 결정되므로, 가장 큰 고유값과 연관된 L개의 주성분 벡터가 선택된다[10]. 그러면, 그것들은 원데이타를 최적으로 표현할 수 있는 특징벡터를 변환하는데 사용된다.

학습과 테스트 과정 중에, GMM을 위한 각각의 입력 특징 벡터는

$$y_{t_j} = \Phi_j x_t, \qquad if\ x_t \in R^l \tag{5}$$

로 변환 된다. 여기에서, $\Phi_j = (\phi_1 \phi_2 \cdots \phi_l)_j$는 행이 R^l의 L개의 주성분 고유벡터인 $L \times n$ 가중치 행렬이고, ϕ_i 벡터는 Σ_j의 i-번째 가장 큰 고유값에 해당하는 고유벡터이다. 그리고, 식(2)의 공분산 행렬은 대각 행렬이다.

3. 효율적인 화자 식별을 위한 강인한 VQ-PCA을 갖는 GMM

j번째 영역에서, 상태열 T_j개의 학습 벡터를 $Y_j = \{Y_{t_j=1}, ..., x_T Y_{t_j=T_j}\}$라 하면, 가우시안 성분 밀도는 M_j 성분밀도의 가중된 합으로 표현할 수 있다.

$$p(y_{t_j}|\lambda) = \sum_{i=1}^{M_j} p_{j,i} \frac{1}{(2\pi)^{\frac{L}{2}} |\Sigma_{j,i}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(y_{t_j} - \mu_{j,i})^T \Sigma_{j,i}^{-1}(y_{t_j} - \mu_{j,i})\right\} \tag{6}$$

여기에서, $\mu_{j,i}$는 평균벡터이고, $\Sigma_{j,i}$는 분산행렬이다. 혼합 성분의 가중치는 $\sum_{j=1}^{K}\sum_{i=1}^{M_j} p_{j,i} = 1$를 만족한다.

$Y = \{Y_1, ..., Y_K\}$가 주어지면, 화자모델을 위한 가우시안 성분 밀도 함수는 성분의 평균벡터, 공분산 행렬, 가중치로 나타낼수 있다.

$$\lambda = \{p_{j,i}, \mu_{j,i}, \Sigma_{j,i}\} \qquad i = 1, \cdots, M_j \ \ \textbf{and} \ \ j = 1, \cdots, K \tag{7}$$

따라서, GMM 유사도는

$$p(Y|\lambda) = \prod_{t_1=1}^{T_1} p(y_{t_1}|\lambda) \cdot \cdots \cdot \prod_{t_K=1}^{T_K} p(y_{t_K}|\lambda) \tag{8}$$

로 쓸 수 있다. 파라미터의 최대 유사도값을 구하기 위하여 EM 알고리즘을 반복 사용하여 얻을 수 있다. 다음의 재추정 식은 모델의 유사도를 단조 증가시킨다.

$$p_{j,i} = \frac{1}{T}\sum_{t_j=1}^{T_j} p\left(j,i\middle|y_{t_j},\lambda\right) \tag{9.a}$$

$$\mu_{j,i} = \frac{\sum_{t_j=1}^{T_j} p\left(j,i\middle|y_{t_j},\lambda\right)y_{t_j}}{\sum_{t_j=1}^{T_j} p\left(j,i\middle|y_{t_j},\lambda\right)} \tag{9.b}$$

$$\Sigma_{j,i}^2 = \frac{\sum_{t_j=1}^{T_j} p\left(j,i\middle|y_{t_j},\lambda\right)\left(y_{t_j}-\mu_{j,i}\right)\left(y_{t_j}-\mu_{j,i}\right)^T}{\sum_{t_j=1}^{T_j} p\left(j,i\middle|y_{t_j},\lambda\right)} \tag{9.c}$$

여기에서, j번째 영역의 i번째 클래스의 사후확률(A posterior probability)은

$$p\left(j,i\middle|y_t,\lambda\right) = \frac{p_{j,i}b_i\left(y_{t_j}\right)}{\sum_{i=1}^{M_j} p_{j,i}b_i\left(y_{t_j}\right)} \tag{10}$$

이다. 여기에서, $L=P, K=1$일 때, 제안된 방법은 [4]방법과 같고, [4]의 방법은 제안된 방법의 특별한 경우로 간주할 수 있다.

화자 식별을 위하여, S명 화자 각각은 GMM의 $\theta_1,\theta_2,\ldots,\theta_s$로 나타내고, 화자의 주성분 벡터를 이용하여 GMM의 최대 사후확률 값을 갖는 화자모델 l를 찾을 수 있다.

$$\hat{s} = \max_{1\le l\le S}\sum_{t=1}^{T}\log p\left(y_t\middle|\theta_l\right) \tag{11}$$

4. 실험결과

우리는 제안된 방법의 검증을 위하여 화자 인식 실험을 하였다. 실험에서, 남자 100명, 여자 100명으로 구성된 200명의 화자가 4개월 동안 4회 방문하여 각 회에 5개의 문장을 발생하였다. 음성 데이터는 12kHz로 샘플링 되었고, 12차 LPC켑스트럼과 13차 델타 켑스트럼으로 사용하였다. 음성 분석 창의 크기는 10ms의 중첩을 가진 20ms를 사용하였다. 첫회에 녹음된 5문장을 화자의 훈련 모델을 위하여 사용하였다. 다른 회에 발성된 3000개의 문장들을 성능 평가를 위하여 사용되었다.

강인한 국부 주성분 분석법 방법과 기존의 전형적인 주성분 분석법의 이상치와 고유값과의 관계를 표 1에 표현하였다. 고유값 비율(Eigenvalue ratio : EVR)은 다음과 같이 나타낼 수 있다.

$$\boldsymbol{EVR}_k = \frac{\sum_{i=1}^{L}\left|\lambda_{i,k}\right|}{\sum_{i=1}^{L}\left|\lambda_{i,k=0}\right|} \tag{12}$$

여기에서, k는 데이터에 포함된 이상치의 양이고, $\lambda_{i,k}$는 Σ_j의 i 번째 고유값이다.

〈표 1〉 제안된 방법과 전형적 PCA 방법에서 이상치와 EVR의 관계

EVR \ K[%]	0	3	5	7	10
RPCA	1.00	1.01	1.01	1.02	1.03
PCA	1.00	1.25	1.46	1.69	2.11

표 1에서 강인한 VQ- PCA 방법이 이상치가 존재할 때 더욱 강인하고, 이상치의 정도가 증가함에 따라서 점점 EVR값이 증가됨을 알 수 있다.

〈표 2〉 제안된 방법과 전형적인 GMM 방법에 필요한 파라메타의 수

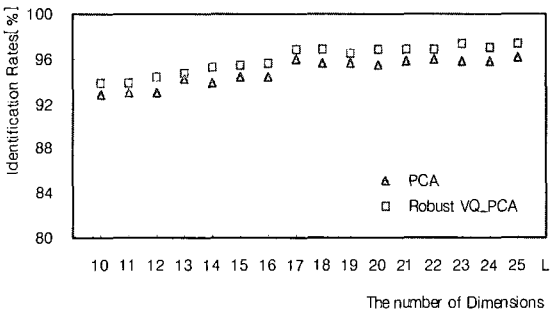
강인한 VQ-PCA를 갖는 GMM	전형적인 GMM
$M_p(2L+1)+Kn(L+1)$	$M_c(2n+1)$

M_p : 강인한 VQ-PCA 를 갖는 GMM에서 전체 혼합 성분 개수

M_c : 전형적인 GMM에서 전체 혼합성분 개수

표 2는 제안된 방법과 기존의 일반적인 GMM방법에서 필요한 파라메타 수를 나타낸 것이다. 제안된 방법에서는 변환행렬을 위한 $K\times L\times n$, VQ를 위한 $K\times n$ 저장공간이 더 필요하지만, 기존의 GMM 방법

의 파라메타 수보다 적게 필요하다. 예를 들어, $M_p = 16, M_c = 64, K = 2, L = 15, n = 25$라 할 때, 제안된 방법과 기존 GMM 방법은 1296개와 3264개의 파라메타를 갖는다.



〈그림 1〉 특징벡터 차원과 식별률과의 관계

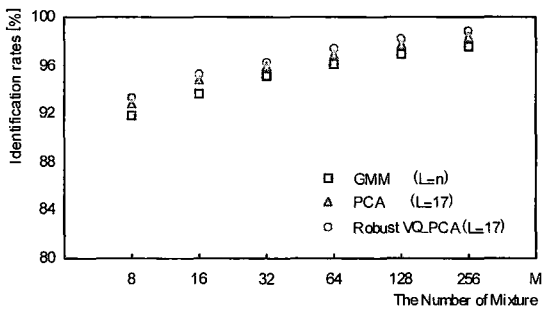
그림 1은 제안된 방법에서 화자식별률과 특징벡터의 차원수와의 관계를 보여준다. 그림에서, 강인한 VQ-PCA와 PCA 방법의 화자식별률은 차원수에 비례하여 증가한다. 강인한 VQ-PCA는 PCA 방법보다 1.1% 정도 우수한 화자식별률을 갖는다. 화자식별 성능은 특징벡터의 수가 17차원에 이른 후에는 성능 변화가 거의 없으므로, 전체 차원을 사용하지 않고, 줄어든 차원을 사용하여 메모리 크기나 계산량을 감소시킬수 있다.

그림 2는 (a) 원 음성과 (b) 이상치가 포함된 음성으로 학습된 화자 모델이 있을 때, 혼합성분의 수와 화자식별 성능사이의 관계를 나타낸 것이다. 여기에서, PCA를 위한 특징벡터의 차원은 L=17이다. 그림 2(a)에서 제안된 강인한 VQ-PCA 방법이 기존의 전형적인 방법(GMM, PCA) 방법과 비교할 때 가장 좋은 화자식별 성능을 보였다. 이 경우에, 전형적인 GMM 방법은 PCA방법과 비교할 때 조차도 더 나쁜 성능을 보였다. 그림 2(b)에서 원음성에 이상치가 포함된 경우에 제안된 강인한 VQ-PCA 방법은 가장 좋은 식별률을 보이지만, PCA 방법은 가장 나쁜 성능을 보였다.

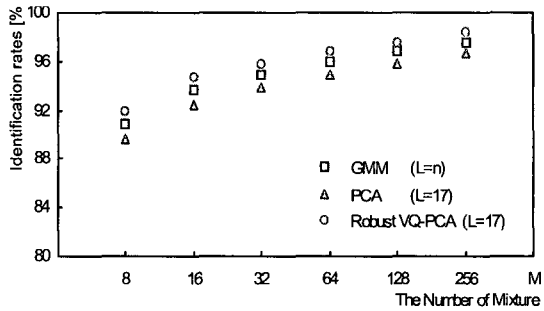
전형적인 PCA 방법의 식별률은 학습데이터에 이상치가 존재할 때 심각하게 저하된다. 이 결과는 화자식별을 위한 PCA 변환 행렬이 정확하지 않은 것으로부터 알 수 있다. 이상치에 의해 손상된 학습 데이터가 사용되었기 때문이다. 그러나 제안된 강인한 VQ-PCA방법의 식별률은 이상치가 존재하거나 존재하지 않은 경우 모두 일정하게 성능을 유지하였다.

5. 결 론

본 논문에서는 화자식별에서 이상치와 고차원 문제를 해결하기 위하여 M-추정을 이용하여 강인한 국부 PCA방법에 기초한 GMM방법을 제안하였다. 먼저, 음성 신호의 고차원 특징 벡터를 강인한 K-평균 알고리즘을 갖는 PCA에 의해 저차원 부공간으로 변환시킨다. 그러면, 변환된 특징벡터를 이용하여 화자식별을 위하여 GMM에 적용된다. 같은 성능을 가질 때 16개의 혼합 성분을 갖는 제안된 방법은 64개의



(a) 원음성



(b)이상치가 포함된 음성

〈그림 2〉 혼합 성분수와 화자 식별 성능사이의 관계

혼합 성분을 갖는 전형적인 대각 GMM 방법과 비교할 때, 더 적은 저장공간을 필요로 하고 더 빠른 실험결과를 얻을 수 있었다. 기존의 GMM 방법은 28개의 혼합성분이 필요하지만, 제안한 방법은 16개의 혼합성분이 필요하였다. 더군다나, 제안된 방법은 학습데이터에 이상치가 존재할 때 더욱 강인하였다.

감사의 글

본 논문은 2004학년도 숭실대학교 교내학술연구비 지원에 의하여 수행되었습니다.

참 고 문 헌

[1] Reynolds, D.A. and Rose, R.C., "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Tr. SAP., vol.3, no. 1,pp.72-82, 1995.

[2] Furui,S., "Recent advances in speaker recognition", Pattern Recognition Letters 18, pp. 859-872, 1997.

[3] Campbell, J.P., "Speaker Recognition:A Tutorial", Proceedings of the IEEE, vol.85, no.9, 1997.

[4] Liu, L. and He, J., "On the use of orthogonal GMM in speaker recognition", Proc. ICASSP, pp.845-849, 1999.

[5] Seo, C., Lee, K.Y.and Lee, J., "GMM based on local PCA for speaker identification", Electronic Letters, vol.37, no.24, pp.1486-1488, 2001.

[6] Ariki, Y., Tagashira, S. and Nishijima, M., "Speaker recognition and speaker normalization by projection to speaker subspace", IEEE ICASSP 96, pp. 319-322, 1996.

[7] Croux, C. and Haesbroeck, G., "Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies", Biometrika 87, no.3, pp. 603-618, 2000.

[8] Gersho, A. and Gray, R.M. "Vector quantization and signal compression", Kluwer Academic.

[9] Huber, P., Robust Statistics, New York : Wiley, 1981.

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

◐ 저 자 소 개 ◑

이 기 용

1983년 숭실대학교 전자공학과 졸업(학사)
1985년 서울대학교 대학원 전자공학과 졸업(석사)
1991년 서울대학교 대학원 전자공학과 졸업(박사)
1991년~1997년 8월 창원대학교 전자공학과 교수
1997년 9월~현재 : 숭실대학교 정보통신 전자공학부 교수
관심분야 : 음성신호처리, 음성향상, 화자인식, 신경망, 적응 신호처리 etc.
E-mail : kylee@ssu.ac.kr