PCA를 이용한 자동차 주행 환경에서의 화자인식

유하진
서울시립대학교 컴퓨터과학부

Speaker Recognition using PCA in Driving Car Environments

Ha-Jin Yu
School of Computer Science, University of Seoul

hjyu@uos.ac.kr

Abstract

The goal of our research is to build a text independent speaker recognition system that can be used in any condition without any additional adaptation process. The performance of speaker recognition systems can be severally degraded in some unknown mismatched microphone and noise conditions. In this paper, we show that PCA(Principal component analysis) without dimension reduction can greatly increase the performance to a level close to matched condition. The error rate is reduced more by the proposed augmented PCA, which augment an axis to the feature vectors of the most confusable pairs of speakers before PCA

Ⅰ. 서론

최근 차량의 증가로 교통체증이 심해져 이동중인 차량에서 보내는 시간이 많아지고 있으므로, 차량 내에서 업무를 처리해 주는 장치의 필요성이 증대되고 있다. 그런데, 주행 중인 차량에서의 기기조작은 안전운행에 심각한 지장을 초래하므로, 운전에 지장을 받지 않고 사용할 수 있는 음성 입력은 필수적이라고 할 수 있다. 특히, 최근에는 이메일이나 금융 정보 등 각종 정보의 송수신에 보안을 위해 화자인식 시스템[1]이 요구되고 있다.

주행 중인 자동차 내에서 음성을 사용하는 데는 아직도 완전히 해결되지 않은 많은 문제점이 있다. 가장 큰 문제점은 주행 중 발생하는 소음이라고 할 수 있다. 자동차의 소음이 특히 어려운 점은 소음의 레벨이나 종류가 일정하지 않고 변한다는 것이다 자동차의 속도에 따라 엔진 등에서 발생하는 소음의 특성이 달라지고, 또한 자동차는 계속적으로 이동하므로 주변에서 발생하는 잡음의 크기나 종류가 위치에 따라 불규칙적으로 변하게 된다. 잡음처리에서 가장 기본적인 방법이라 할 수 있는 스펙트럼 차감법 (spectral subtraction) 등은 일정한 잡음의 성질을 미리 알고 있다고 가정해야 하므로 주행 중인 차량에서와 같이 계속 변하는 잡음에서는 큰 효과를 기대하기 어렵다.

주행 중인 차량 내 음성입력의 또 한 가지 문제점은 사용자의 입과 마이크의 거리를 충분히 가깝게 할 수 없다는 것이다. 음성 입력을 위한 헤드셋은 운전자에게 불편함을 주게 되며, 주행 중 예고 없이 발생하게 되는 상황에서 헤드셋을 착용하는 동작은 안전 운전에 지장을 초래할 수 있다. 또한, 마이크를 입 가까이 고정시키는 것 또한 운전자에게 불편함이나 거부감을 주며, 차량의 구조상 설치가 용이하지 않다. 그러므로 마이크를 고정할 수 있는 위치는 선바이저 등과 같이 사용자의 입과 상당히 떨어져 있어야 하므로, 원거리 입력이 가능한 콘덴서 마이크 종류를 사용해야 하며, 따라서 잡음의 유입이 한 층 더 심해진다.

이와 함께 화자인식에서 큰 영향을 주는 요인 중의 하나는 마이크의 변화이다. 즉, 등록할 때 사용한 마이크와 테스트에서 사용한 마이크의 종류가 다르면 인식 성능이 크게 떨어지게 된다.

본 연구에서는 실용화를 위해 이와 같은 여러 가지 문제점을 고려하여 다음과 같이 목표를 설정하고 있다.

① 잡음의 종류나 레벨에 관한 사전 지식을 사용하지 않는다.

② 마이크의 종류에 대한 사전 지식을 사용하지 않는다.

③ 사용자가 다른 목적을 위한 음성인식 과정에서 입력한 음성을 화자인식에 사용하기 위해, 미리 지정되지 않은 문장을 사용할 수 있는 텍스트독립(text independent)방식으로 화자인식을 한다.

위와 같은 목표 달성을 위한 실험을 위해 SITEC에서 구축한 음성자료를 사용하였다. 본 논문에서는 주성분분석(Principal component analysis, PCA)을 사용하여 실험한 초기 결과를 보고하고자 한다. 주성분분석은 주로 특징벡터의 차원을 줄여서 데이터 처리량을 줄이고 모델 크기를 줄여서 인식시간을 단축하는데 많이 사용되어져 왔다[2-5]. 본 연구에서는 PCA가 마이크의 불일치와 잡음에 강인한 효과가 있음을 보인다. 또한, 특징의 차원을 줄이는 대신에, 제안한 방법으로 오히려 정보를 추가함으로서 추가적인 성능의 향상을 얻을 수 있음을 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 본 실험에서 목표로 하는 환경을 조성하기 위한 실험 환경을 설명하고, 3장에서는 주성분 분석과 GMM을 이용한 화자식별 방법과 실험 결과를 기술한다. 4장에서는 본 연구에서 인식 성능을 더욱 향상시키기 위하여 제안한 부가 주성분분석의 간단한 설명과 실험결과를 보이며, 5장에서 본 논문의 내용을 요약하고 결론을 맺는다.

II. 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. 또한, 테스트에 사용된 음성은 입과 비교적 떨어져 있고, 콘덴서 마이크를 사용하였으므로 잡음이 상당히 포함되어 있다. 이에 반하여, 학습에 사용된 음성은 다이나믹 마이크로 입에 가까이 대고 녹음되었으므로 잡음이 거의 없는 것을 알 수 있다.

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 249개 이다. 본 연구에서는 최초 발성된 모든 텍스트의 음성을 학습데이터로 사용하고 1주일 후에 발성된 모든 텍스트의 음성을 테스트 데이터로 사용하였다.

화자인식을 위한 특징으로는 MFCC (Mel-frequency cepstral coefficients)와 이의 1차 및 2차 미분을 사용하고, 잡음을 감소시키기 위하여 CMS (Cepstral mean subtraction) 방법을 사용한다.

화자 모델은 GMM(Gaussian mixture model)[6]을 사용한다. 혼합(mixture)수에 따른 성능의 변화를 살펴보기 위하여 혼합수를 1, 2, 4, 8, , 512 과 같이 두 배씩 증가시키면서 실험하였다. 각 혼합수 별 반복 학습(iteration) 회수는 5회로 하였다.

III. 주성분 분석을 이용한 화자 식별

1. 주성분 분석 (PCA)

주성분 분석은 특징 공간을 표현하기 위해 서로 독립적인 축을 구하고, 차원을 축소시켜 저장 공간과 처리 시간을 감축하기 위해 주로 사용된다[2-5] 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. 특정 행렬에 대한 고유벡터들은 서로 직교(orthogonal; 내적값이 0)한다.

단계 5: 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 가

장 중요한 몇 개의 축을 정하여 변환행렬을 만드는데, 본 연구에서는 자료의 차원을 줄이는 것이 목적이 아니므로 모든 축을 사용하였다.

단계 6. 변환행렬을 이용하여 모든 데이터를 변환한다.

(변환된 데이터) = (변환행렬)×(입력벡터)

2. GMM(Gaussian mixture model) [6]

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.

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

여기서 μ_i 는 자료의 평균 벡터이고, Σ_i는 공분산행렬이다.

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

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

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

여기서 λ는 화자 모델의 파라미터를 나타낸다.

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

따라서 파라미터λ의 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 EM알고리즘은 반복적으로 모델 파라미터를 다듬어서 학습데이터와 모델파라미터가 잘 정합되도록 한다. EM알고리즘은 E단계와 M단계로 나누어진다. E(expectation) 단계는 현재의 모델 파라미터와 관측데이터 (observations)를 이용하여 숨겨진 구조(hidden structure)를 예측하고, M(maximaization)단계에서는 예측된 숨겨진 구조를 이용하여 파라미터를 재추정한다.

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 있을때 화자 식별은

$$p(\vec{x}_t\mid\lambda_k) = \sum_{i=1}^{T} w_i g_i(\vec{x}_t)$$

을 최대로 하는 k를 찾는 것이다.

3. 실험 결과

일반적인 멜켑스트럼 계수를 특징으로 사용하여 GMM으로 학습하고 인식하였을 경우와 여기에 주성분 분석을 적용하였을 경우의 결과를 그림 1에서 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.

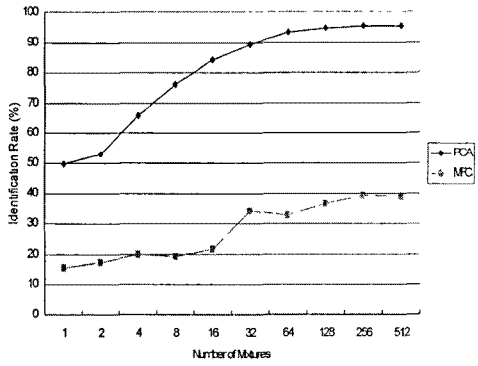

그림 1 마이크 불일치 및 잡음 환경에서의 주성분분석(PCA)의 효과

Ⅳ. 제안한 부가주성분분석 (Augmented PCA)

본 연구에서는 불일치 환경에서의 인식 성능을 향상시키기 위하여, 인식 대상 화자 집합에서 가장 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 a와 $-a$를 각각 넣는다. 이 상수 a는 모든 학습 데이터의 특징벡터 내 모든 원소의 최대값으로 하였다. 그리고 이 변환 행렬을 학습 데이터에 적용하여 두 화자만을 구분하는 GMM을 학습시킨다. 테스트 시에는 상위 n개의 인식 결과에 대하여 두 화자만을 구분하는 후처리를 함으로써 인식 오류를 줄일 수 있었다. 그림 2에서, 본 실험에서 가장 혼동되는 두 화자만의 인식 오류가 제안한 방법(AUGPCA로 표시)에 의하여 감소되는 것을 볼 수 있다. 전체적으로 혼합수가 128개 일 때 인식 오류가 5.4%에서 4.3% 로 줄어 21%

의 상대 오류율 감소를 얻을 수 있었다.

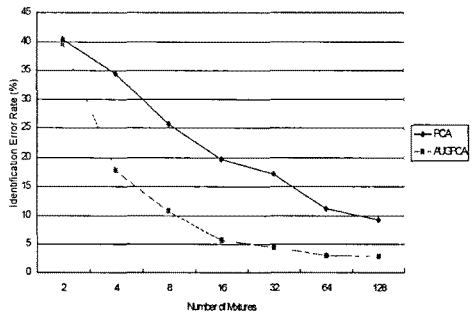

그림 2 가장 혼동되기 쉬운 두 화자에 대한 부가주성분분석(AUGPCA)의 효과

V. 결론

본 연구에서는 유비쿼터스 환경에서와 같이 언제 어디서나 화자식별을 수행하기 위하여 학습시와 다른 마이크와 잡음 환경에 대하여 사전 정보가 없는 상태로 추가적인 적응과정을 거치지 않고 인식하는 것을 목표로 한다. 음성정보기술산업지원센터(SITEC)에서 수집한 자동차 화자인증용 음성 DB를 이용한 실험결과, 주성분분석(PCA)를 사용한 경우에 불일치 환경에서도 높은 인식 성능을 얻을 수 있었다. 또한, 제안한 부가주성분분석(augmented PCA)를 사용하여 상대 인식 오류를 추가로 21% 줄일 수 있었다. 향후 연구 과제로는 LDA(linear discriminant analysis)등 유사한 방법에 대한 비교와 이를 이용한 성능 향상이 있을 수 있다.

감사의 글

본 연구의 실험에 큰 도움을 준 서울시립대학교 컴퓨터과학부 3학년 신연식 학생에게 감사를 표합니다.

참고문헌

[1] Joseph P Campbell, JR, "Speaker Recognition: A Tutorial," Proceedings of the IEEE, Vol 85, No 9, September 1997, pp. 1437-1462

[2] Shang-nien Tsai, Lin-shan Lee, "Improved robust features for speech recognition by integrating time-frequency principal components (TFPC) and histogram equalization (HEQ)," 2003 IEEE Workshop onAutomatic Speech Recognition and Understanding, 2003. ASRU '03, 30 Nov -3 Dec. 2003 pp 297 - 302

[3] Zhang Wanfeng, Yang Yingchun, Wu Zhaohui and Sang Lifeng, "Experimental evaluation of a new speaker identification framework using PCA," IEEE International Conference on Systems, Man and Cybernetics, 2003, Volume 5 , 5-8 Oct 2003, pp. 4147 - 4152

[4] Peilv Ding, Liming Zhang, "Speaker Recognition using Principal Component Analysis," Proceedings of ICONIP 2001, 8th International Conference on Neural Information Processing, Shanghai China, November 14-18, 2001

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

[6] 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