Slow Feature Analysis를 이용한 강인한 화자식별

양일호, 김민석, 유하진

서울시립대학교 컴퓨터과학부

Robust Speaker Identification using Slow Feature Analysis

IL-Ho Yang, Min-Seok Kim and Ha-Jin Yu
heisco@hanmail.net, ms@uos.ac.kr, hjyu@uos.ac.kr

요약

본 논문에서는 새로운 특징 추출 방법 중 하나인 SFA(slow feature analysis)[3]를 이용한 강인한 화자 식별 시스템을 제안한다. SFA는 입력 신호에서 느리게 변화하는 특징을 추출하는 방법이다. 휴대폰 환경의 화자 식별 실험에서, MFCCs(mel-frequency cepstral coefficients)와 이것의 delta를 특징으로 실험한 결과에 비해 SFA를 이용한 실험이 높은 성능을 나타내었다.

1. 서론

MFCCs와 가우시안 혼합 모델[1]을 이용한 화자식별시스템은 잡음이 없는 환경에서 높은 성능을 나타낸다. 그러나 잡음환경에서는 그 성능이 급격히 저하된다. 잡음 환경에 강인한 화자 식별 성능을 위해 주성분분석(PCA: principal components analysis)[3]을 이용한 특징 추출 방법이 사용되고 있다[4].

본 연구에서는 주성분분석 대신에 특징 변환의 한 종류인 SFA(slow feature analysis)[2]를 효과적으로 화자식별에 적용하는 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 2장에서는 이론적 배경을 소개한다. 3장과 4장에서 각각 실험 설계와 결과 분석을 서술하고, 5장에서 결론을 맺는다.

2. 이론적 배경

2.1 SFA(slow feature analysis)[2]

SFA는 벡터 형태의 입력 신호로부터 느리게 변화하는 특징을 추출하는 방법이다. I차원의 입력 신호 $\boldsymbol{x}(t)=\{x_1(t),...,x_I(t)\}$를 J차원의 출력 신호 $\boldsymbol{y}(t)=\{y_1(t),...,y_I(t)\}$로 변환할 때, SFA는 다음 수식을 만족하는 변환 $y_j := g_j(\boldsymbol{x}(t))$를 찾는다.

$$\Delta_j := \Delta(y_j) := \langle \dot{y}_j^2 \rangle \text{를 최소화} \qquad (1)$$

즉, SFA는 시간에 따른 데이터 변화(delta, Δ)의 분산을 최소화 한다. quadratic SFA(SFA^2)[2]는 입력 신호를 비선형 확장하여 SFA를 수행하지만, 본 연구에서는 입력 신호를 비선형 확장하지 않고 그대로 사용하는 linear SFA를 이용하였다. linear SFA의 수행 과정은 다음과 같다.

1. 입력 신호의 평균을 0, 분산을 1로 정규화
2. 입력 신호의 시간적 변화(delta) 유도
3. delta의 공분산에 대하여 PCA 수행
4. 고유값이 작은 순으로 고유벡터를 취하여 변환 행렬 구성

2.2 화자인식에의 이용

화자인식에서 사용하는 음성 데이터에서 언어적 정보는 시간에 따라 빠르게 변화하지만, 하나의 발성 속에서 화자 정보는 변하지 않는다. 다시 말해, 학습 데이터 전체에서 볼 때 화자 특성은 느리게 변화하는 특징에 속한다. SFA는 시간적 변화가 느린 순서대로 특징을 추출하므로 이를 이용하여 얻은 특징은 식별 성능 향상에 기여할 수 있을 것이다.

3. 실험 설계

3.1 데이터베이스

본 연구에서는 ETRI 한국어 휴대폰 화자인식용 음성 DB를 이용하였다. 이 중 30명 화자(남성 15명 / 여성

15명)의 4연 숫자 발성 데이터에 대하여 식별 실험을 수행하였다. 모델 학습용으로 화자별 25개 데이터(1번~25번)를 사용하였고, 성능평가를 위해 또 다른 25개 데이터(26번~50번)를 사용하였다(25개*30명=총750개).

3.2 특징 추출

각 음성데이터에서 16차 특징(MFCCs_0 := 15차 MFCCs + 에너지)을 추출하였다. 여기에 delta, PCA변환결과, SFA변환결과 등을 덧붙여 식별 오류율을 비교하였다. PCA 및 SFA 변환은 MDP(Modular toolkit for Data Processing)[5]를 이용하여 수행하였다. 각 변환은 모델 학습 데이터로부터 추정하였다.

3.3 화자 모델

각 화자별로 가우시안 혼합 모델을 학습하였다(혼합수 32개).

3.4 실험 방법

SFA의 성능 향상 효과를 확인하기 위하여 16차 MFCCs_0특징, PCA변환결과, SFA변환결과만 가지고 실험을 수행하였다. MFCC특징에 대한 SFA변환은 곧 delta에 대한 PCA의 역순 변환이므로, SFA가 delta를 대체할 수 있는지 확인하기 위하여 추가 실험을 진행하였다.

4. 결과 분석

4.1 실험 결과

사용 특징에 따른 식별 실험 결과는 그림 1과 같다.

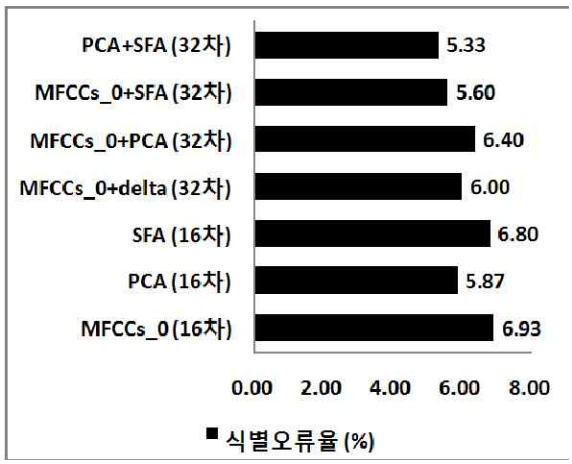

그림 1 사용한 특징에 따른 화자 식별 성능 비교

4.2 결과 분석

MFCCs_0 특징에 대하여 SFA를 적용한 경우 식별 오류율이 6.93%에서 6.80%로 감소하였으나 PCA(5.87%)에 비해서는 오류율이 높았다.

MFCCs_0만 사용한 경우(6.93%)보다 delta를 더하여 실험(6.00%)하였을 때 식별 오류율이 더 낮았다. 이때, delta 대신 MFCCs_0에 대한 SFA변환결과를 더하여 실험(5.60%)하자 식별 오류율이 더욱 감소하는 것을 확인할 수 있었다. 반면에 PCA변환결과로 delta를 대체한 경우는 식별 오류율이 6.40%로 오히려 높아졌다.

또한 MFCCs_0 대신 PCA변환결과를 사용하고, delta 대신 SFA변환결과를 사용할 경우 5.33%로 가장 높은 성능을 보였다(상대 오류율 11.17% 감소).

5. 결론

본 연구에서는 강인한 화자 음성 특징을 얻기 위하여, 시간적으로 느리게 변화하는 특징을 추출하는 SFA를 적용해 보았다. 실험 결과를 통해 SFA는 delta 대신 사용하였을 때 성능 향상 폭이 크며, PCA와 결합하였을 때 더 좋은 성능을 보이는 것을 확인하였다.

향후 연구 주제는 비선형 변환을 수행하는 quadratic SFA의 적용 및 화자 확인 기술에의 접목 등이 있다.

참고문헌

1. Douglas A. Reynolds and Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE transactions on Speech and Audio Processing, 1995.
2. Laurenz Wiskott, Terrence J. Sejnowski, "Slow Feature Analysis: Unsupervised Learning of Invariances," Neural computation, 2002.
3. Richard O. Duda, Peter E. Hart, David G. Stock, "Pattern Classification, Second Edition," John Wiley & Sons, 2001.
4. Z. Wanfeng, Y. Yingchun, W. Zhaohui and Sang Lifeng, "Experimental evaluation of a new speaker identification framework using PCA," IEEE International Conference on Systems, Man and Cybernetics, Vol. 5, pp. 4147-4152, 5-8 October 2003.
5. Zito, T., Wilbert, N., Wiskott, L., Berkes, P. "Modular toolkit for Data Processing (MDP): a Python data processing frame work," Front. Neuroinform, 2008.