인터넷 전화를 통한 음성의 화자 식별 성능 개선

김 명 재[1] · 김 민 석[2] · 양 일 호[1] · 소 병 민[1] · 유 하 진[1]
서울시립대학교[1], LG전자기술원[2]

The improvement of speaker identification performance on the voice through internet phone

Kim, Myung-Jae[1] · Kim, Minseok[2] · Yang, IL-Ho[1] · So, Byung-Min[1] · Yu, Ha-Jin[1]
University of Seoul[1], LG Electronics Inc.[2]
mj@uos.ac.kr,minseok3.kim@lge.com,hesico@hanmail.net,sbm1210@naver.com,hjyu@uos.ac.kr

요약

인터넷 전화를 위해 사용하는 코덱은 본래의 음성을 왜곡하거나 음성의 일부 성분이 사라지게 하여 화자 인식기의 인식률을 하락시키는 요인으로 작용한다.

본 논문에서는 SILK 코덱을 이용하여 학습 환경과 테스트 환경의 불일치 상황에서 화자식별을 수행하여 하락된 인식률을 개선할 수 있는 채널 보상 방법과 특징 강화 방법을 찾고자 한다.

1. 서론

최근 인터넷 전화의 사용이 증가하면서, 인터넷 전화 환경에서 좋은 성능을 가지는 인식기의 필요성이 증가하였다. 일반적으로 화자 인식 시스템은 동일한 환경에서 학습과 인식을 하는 경우 좋은 인식률을 보이나, 학습과 인식의 환경이 서로 다르게 되면 인식률이 급격하게 하락 할 수 있다.

본 논문에서는 현재 사용자가 가장 많은 스카이프사의 광대역 코덱 SILK[1]를 이용하여 학습 환경과 다른 테스트 음성에 가장 좋은 성능을 갖는 특징 강화 방법을 찾고자 한다.

본 연구에서는 음성 특징의 특징 강화 방법으로 주성분 분석(PCA, principal component analysis)[2], 그리디 커널 주성분 분석(GKPCA, greedy kernel PCA)[3], 커널 다중 판별 분석(KMDA: kernel multimodal discriminant analysis)[4]을 사용하여 비교하였다.

본 논문의 2장에서는 사용한 특징 강화 방법들에 대해 소개한다. 3장에서는 실험방법을 설명하고 결과를 분석한다. 마지막으로 4장에서 결론을 맺는다.

2. 특징강화 방법

2.1 주성분 분석 (PCA)

주성분 분석은 전체 데이터의 분산을 최대화 하는 기저를 찾고, 그 기저로 음성 특징을 사상하는 방법이다. 하지만 음성 특징이 비선형으로 분포되어 있는 경우, 적합한 기저를 찾지 못할 수 있다.

2.2 그리디 커널 주성분 분석 (GKPCA)[3]

커널 주성분 분석(kernel PCA)은 커널 방법을 이용하여 데이터가 비선형으로 분포되어 있는 경우도 처리 가능하게 개선한 방법이다. 커널 방법은 음성 특징을 임의의 고차원 특징 공간으로 확장하여 문제를 해결한다.

그리디 커널 주성분 분석은 기저를 추정 할 때, 특징 모두를 사용하는 것이 아니라 그리디 필터링(greedy filtering)으로 소수의 특징을 선택하여 계산량과 메모리 요구량을 줄이는 방법이다.

본 연구에서는 커널 함수로 가우시안 RBF 커널 함수(1)를 사용하였고, 그리디 필터링으로 선택하는 특징의 수를 100개로 하였다.($\sigma = 21$)

$$k(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right) \qquad (1)$$

2.3 커널 다중 판별 분석 (KMDA)[4]

커널 다중 판별 분석은 특징 공간상에서 서브 클러스터의 중심과 전체 중심 간의 거리를 최대화 하는 축으로 특징을 변환하는 방법이다. 서브 클러스터는 커널 K-means를 이용하여 구한다. 이 때, 커널 K-means의 K

가 서브 클러스터의 개수가 된다. 커널 다중 판별 분석에서도 가우시안 RBF 커널 함수(1)를 사용하였다. 본 연구에서는 커널 다중 판별 분석의 서브 클러스터의 개수를 4개로 하였다.(K=4)

3. 실험 및 결과

3.1 데이터베이스

데이터베이스는 ETRI 중가 마이크 화자 인식용 데이터를 사용하였다. 학습과 테스트에 100명의 문장 발성이 사용되었으며, 학습에는 1주차 데이터의 1번째 발성 10회가 사용되었고, 테스트 데이터에는 2주차 1번째 발성 10회, 3주차 1번째 발성 10회, 4주차 1번째 발성 10회가 사용되었다.

테스트에 사용한 데이터는 SILK 코덱을 이용하여 스카이프를 통과한 음성으로 시뮬레이션하여 데이터베이스를 구성하였다.

3.2 모델 학습

모델 학습에는 GMM(gaussian mixture model)[5]을 사용하였다. 혼합 수는 64개를 사용하였으며, 총 10회 반복학습을 하였다.

3.3 특징 추출

학습과 테스트에 사용한 모든 데이터는 표본화 주파수를 8KHz로 하고,에너지를 기반으로 무음구간을 제거하였다. 음성 특징으로는 20차 MFCC와 로그에너지를 사용하였다. 채널 보상으로는 켑스트럼 평균 정규화(CMN)와 켑스트럼 평균-분산 정규화(CMVN)를 사용하였다. 이를 바탕으로 주성분 분석(PCA), 그리디 커널 주성분 분석(GKPCA), 커널 다중 판별 분석(KMDA)으로 특징을 변환하였다.

3.4 실험 결과

실험 결과는 10회 반복 학습을 수행한 것 중 가장 높은 인식률을 기재하였다. WEEK은 주차 데이터를 의미하며 숫자는 음성이 녹음된 시점을 의미한다.

표 1은 PC 환경의 음성 데이터로 학습하고, PC 환경과 스카이프 환경의 음성 데이터로 테스트한 결과이다. 스카이프 환경은 특징 강화 방법을 적용하였다. PC 환경에서는 약 80~87%의 인식률을 보인다. 스카이프 환경에서 인식기의 상대적인 성능 하락 폭은 약 50~107% 이며, 특징 강화 방법을 이용한 상대적인 성능 개선 폭은 커널 다중 판별 분석 기준으로 약 13~25%이다.

표 1. 실험 결과 (PC, 스카이프 환경)

	WEEK 2	WEEK 3	WEEK 4
CMN(PC)	86.8	79.9	82.7
CMVN(PC)	87.2	78.9	85
CMN(skype)	72.6	69.6	68.2
CMVN(skype)	74.3	68.2	72.4
PCA(skype)	77.1	72	75.4
GKPCA(skype)	76.9	72.8	74.5
KMDA(skype)	**79.7**	**73.9**	**76.7**

4. 결론

인터넷 전화를 사용할 때, 화자의 음성을 왜곡시킬 수 있는 여러 가지 요인이 존재한다. 본 연구에서는 여러 가지 요인 중 학습 환경과 테스트 환경의 불일치를 고려하였다. 이를 위해 학습 환경은 PC, 테스트 환경은 스카이프를 선택하였다. 화자 인식기의 인식률 하락요인을 개선하기 위해 켑스트럼 평균 정규화, 켑스트럼 평균-분산 정규화와 같은 채널 보상 방법과 주성분 분석, 그리디 커널 주성분 분석, 커널 다중 판별 분석등과 같은 특징 강화 방법을 수행하였다. 실험 결과 채널 보상으로 켑스트럼 평균 정규화를 수행하고 특징 강화 방법으로는 커널 다중 판별 분석을 수행한 방법이 평균적으로 가장 높은 인식률을 보였다.

참고문헌

1. http://developer.skype.com/silk

2. Duda, R, O., Hart, P. E., Stork, D. G., Pattern Classification, John Wiley & Sons, 2001.

3. Kim, M-S., Yang, I-H., Yu, H-J., "Speaker Identification using Greedy Kernel PCA", Malsori, pp. 105-116, 2008

4. Kim, M-S., Yang I-H., Yu, H-J., "Kernel Multimodal Discriminant Analysis for Speaker Verification", In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4498-4501, 2010

5. Reynold, D. A., Rose, R. C., "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Transaction on Speech Audio Processing Vol. 3, No.1, pp. 72-83, 1995