코덱을 통한 음성과 순수한 음성의 학습량에 따른 화자 식별 성능 비교

허 희 수[1] · 김 민 석[2] · 김 명 재[1] · 양 일 호[1] · 백 록 선[1] · 유 하 진[1]
서울시립대학교[1], LG전자기술원[2]

The comparative analysis of speaker identification performance according to amount of train data on the voice through codec and original voice

Heo, Hee-Soo[1] · Kim,Min-seok[2] · Kim, Myung-Jae[1] · Yang, IL-Ho[1] · Baek, Rock-Seon[1] · Yu, Ha-Jin
University of Seoul[1], LG Electronics Inc.[2]
zhasgone@naver.com, minseok3.kim@lge.com, mj@uos.ac.kr, heisco@hanmail.net, whites86@naver.com, hjyu@uos.ac.kr

요약

최근 스마트폰을 이용하여 음성을 전송하는 서비스가 늘고 있다. 이러한 서비스의 음성 전송을 위해서 사용하는 음성 코덱인 speex[1]는 압축하는 과정에서 음성 품질이 저하된다. 이 때문에 화자 인식기의 인식률이 하락할 수 있다.

본 논문에서는 원래의 음성 데이터와 speex 코덱을 거친 음성 데이터 각각에서 학습 데이터양에 따른 인식률을 비교하는 실험을 하고자 한다.

1.서론

인터넷망을 통해서 음성을 전송하는 경우 대역폭 제한 때문에 한 번에 전송할 수 있는 데이터의 한계가 생긴다. 이 한계를 극복하기 위한 방법으로 코덱을 이용한 손실 압축 방법이 있다. 그런데 손실 압축을 사용하는 경우 음성 데이터의 영구적인 품질이 저하되어 화자 인식기의 성능이 하락할 수 있다. 따라서 코덱을 거친 음성에 대한 화자 인식기의 인식률에 관한 연구가 필요하다.

본 논문은 현재 음성 인식 서비스에서 사용하는 speex 코덱을 이용하여 본래 음성의 품질을 저하시킨 뒤 학습 데이터양에 따른 인식률을 본래 음성의 인식률과 비교하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 화자인식기에서 사용한 가우시안 혼합 모델[2]을 소개한다. 3장에서는 실험방법을 설명하고 결과를 분석한다. 4장에서 결론을 맺는다.

2. 가우시안 혼합 모델

가우시안 혼합 모델은 하나의 가우시안 분포로 표현하기 어려운 복잡한 분포를 다수의 가우시안 분포를 혼합하여 표현하는 방법이다. 이 때, 확률밀도함수는 M개의 가우시안 확률밀도함수의 선형결합으로 정의하며 다음 식과 같이 표현한다.

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (1)$$

여기서 $b_i(\vec{x})$는 혼합 모델의 기본 성분을 이루는 가우시안 확률밀도함수가 된다. p_i는 i번째 성분이 전체 혼합 확률밀도 함수에서 가지는 가중치를 나타낸다. λ는 가중치와 평균, 공분산 행렬로 화자 모델을 정의한다. 가우시안 혼합 모델을 화자 인식기에 적용하기 위해서

S 명의 화자에 대한 화자 모델을 가우시안 혼합 모델을 통해서 정의한다. 특징 벡터 $\vec{x}$를 발생시킬 확률이 가장 높은 화자 모델 λ_s를 찾아서 화자 $\hat{s}$를 선택하는 것으로 화자 인식을 수행하며 다음 식과 같이 표현한다.

$$\hat{s} = \arg max\, p(\vec{x}|\lambda_s), 1 \le s \le S \qquad (3)$$

3. 실험 및 결과

3.1 데이터베이스

본 실험에서는 서울말 낭독체 DB를 사용하였다. 이 중 76명 화자(남성 36명 / 여성 40명)의 발성을 사용해 화자 인식 실험을 수행하였다. 모델 학습을 위해 화자별로 5초에서 600초까지 5초씩 늘려가며 학습에 사용할 발성 데이터를 구성하였다.(600초 * 76명 = 45600초) 실험용으로 화자별로 62개 문장의 발성 데이터를 사용하였다.(62개 * 76명 = 4712개)

3.2 화자 모델

화자 모델에는 가우시안 혼합 모델을 사용하였다. 혼합 수는 256, 512, 1024개를 사용하였으며, 총 15회 반복 학습을 하였다.

3.3 특징 추출

학습과 테스트에 사용한 모든 데이터는 표본화 주파수를 16kHz로 하고, 에너지를 기반으로 무음 구간을 제거하였다. 음성 특징은 25ms의 윈도우 크기로 10ms씩 중첩하여 이동하면서 12차 MFCC와 0번째 캡스트럼 계수를 사용하여, 총 13차 특징을 추출하여 사용하였다.

3.4 실험 결과

그림 1은 512개 혼합 수에 대한 원본 음성 데이터와 코덱을 거친 음성 데이터의 학습량에 따른 인식률을 비교한 것이다.

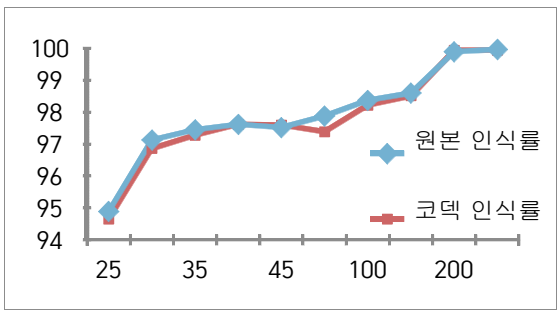

그림 1. 혼합 수 512개에서 학습량의 따른 원본 음성과 코덱을 거친 음성의 인식률 비교

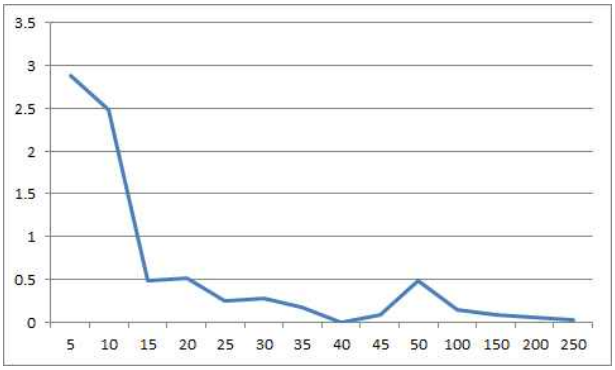

그림 2. 혼합 수 512개에서 학습량에 따른 원본 음성과 코덱을 거친 음성의 인식률 차

그림 2는 512개의 혼합 수에서 학습량에 따른 원본 데이터와 코덱을 거친 음성 데이터의 인식률의 차를 나타낸 그래프이다.

혼합 수	원본 인식률	코덱 인식률	차
256	99.94	99.94	0
512	99.96	99.94	0.02
1024	99.98	99.92	0.06

표 1. 학습 데이터가 250초인 경우 각 혼합 수별 원본과 코덱을 거친 음성 데이터의 인식률과 그 차

표 1은 학습 데이터가 250초인 경우 256, 512, 1024개의 혼합 수에 대한 원본 음성 데이터와 코덱을 거친 음성 데이터의 인식률과 그 차를 보여준다.

4. 결론

speex 코덱을 사용해서 음성 데이터를 압축할 때, 데이터가 손실되는 화자 인식기의 인식률을 하락시킬 수 있다. 본 연구에서는 왜곡된 음성에 대한 화자 인식기의 인식률을 본래 음성의 인식률과 비교하기 위해서 두 데이터 그룹으로 실험을 수행하였다. 실험 결과 두 그룹 간의 인식률 차는 학습량이 증가할수록 적어지는 경향을 보였으며 50초 이상의 충분한 학습이 이루어지면 인식률의 차가 0.6% 이하로 나타났다.

참고문헌

1. http://speex.org/docs/manual/speex-manual/

2. 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.