말소리와 음성과학 제2권 제3호 (2010)
pp. 127~134

히스토그램 변환에서 기준분포의 표준편차 변경에 따른 강인한 화자인증 성능 개선

Performance Improvement of Robust Speaker Verification According to Various Standard Deviations of a Reference Distribution in Histogram Transformation

권 철 홍[1)]
Kwon, Chul Hong

ABSTRACT

Additive noise and channel mismatch strongly degrade the performance of speaker verification systems, as they distort the features of speech. In this paper a histogram transformation technique is presented to improve the robustness of text-independent speaker verification systems. The technique transforms the features extracted from speech such that their histogram is conformed to a reference distribution. The effect of different standard deviations for the reference distribution is investigated. Experimental results indicate that, in channel mismatched environments, the proposed technique offers significant improvements over existing techniques. We also verify performance improvement of the proposed method using statistics.

Keywords: robust speaker verification, histogram transformation, statistics.

1. 서론

화자인식(speaker recognition)은 크게 화자식별(speaker identification)과 화자인증(speaker verification)으로 나누어진다[1]. 화자식별은 여러 후보 중에서 발화한 한 명의 화자를 찾는 방법으로 자동 회의록 작성에 응용될 수 있으며, 화자인증은 등록된 화자(claimant, 사용자)와 사칭자(impostor)를 구분하는 기법으로 텔레뱅킹 등에서 본인 인증에 사용될 수 있다. 본 논문에서는 화자식별 보다 응용범위가 광범위한 화자인증을 다룬다. 화자인증은 발화 문장에 따라 문장 종속형과 문장 독립형이 있다. 문장 종속형은 화자가 발성하는 문장이 화자인증 시스템에 알려져 있고, 고정되어 있거나 시스템의 물음에 정해진 대답을 한다. 문장 독립형은 발성 문장이 임의적이어서 화자인증 시스템이 사전에 알 수 없다. 본 논문에서는 문장 독립 화자인증 시스템을 다룬다.

1) 대전대학교 chkwon@dju.ac.kr, 교신저자

접수일자: 2010년 7월 2일
수정일자: 2010년 9월 1일
게재결정: 2010년 9월 27일

화자인증에서 주요 연구과제는 화자간 특성의 차이를 잘 보여주는 특징을 추출하는 방법과 최적의 화자 모델링 방법이다[2]. 근래에 화자인증에서는 MFCC(Mel Frequency Cepstral Coefficients)와 같은 특징 계수와 GMM(Gaussian Mixture Models) 기반 화자 모델링 방식을 널리 사용하고 있다. MFCC를 사용하는 이유는 이 특징 계수가 음성인식뿐만 아니라 화자인식에서도 좋은 성능을 보여 주기 때문이다. GMM은 문장 독립 화자인증 시스템에서 음소 기반 HMM(Hidden Markov Models) 보다 성능이 우수하다는 사실이 알려져 있어[3], 이 분야에서는 주로 GMM으로 화자를 모델링한다. 배경화자를 모델링하는 방법은 UBM(a Universal Background Model) 방식[4]으로, 이는 다수의 화자로부터 음성을 수집하여 하나의 모델을 훈련시키는 방법이다. 본 논문에서는 특징 계수로 MFCC를, 화자와 사칭자 모델은 GMM-UBM 방법을 사용한다.

현재 화자인증 시스템은 조용한 환경에서 고성능 마이크로 수집한 음성 데이터로 훈련하고 인식하였을 경우 충분히 좋은 성능을 보여준다. 그리고 훈련과 인식의 환경이 유사한 경우에도 비교적 좋은 성능을 나타낸다. 그러나 훈련환경과 인식환경이 달라질 경우 시스템의 인식 성능은 크게 저하된다[5]. 이러

한 불일치 원인은 입력 음성에 더해지는 잡음과, 훈련과 인식에서 사용하는 마이크나 전화기 등의 차이 즉 채널 불일치이다. 이 불일치는 잡음과 채널의 종류에 따라 특징 계수 MFCC를 비선형적으로 변형시킨다. 따라서 깨끗한 음성으로 훈련된 화자모델은 변형된 음성을 정확히 모델링하지 못하게 된다.

강인한 화자인증을 위한 성능 개선 방법은 주로 훈련환경과 인식환경의 불일치를 최소화 한다. CMN(Cepstral Mean Normalization)과 MVN(Mean and Variance Normalization) 등이 주로 사용되는데, 이 방법들은 MFCC의 평균을 제거하거나 분산을 정규화 한다[6]. CMN은 MFCC의 평균을 0으로 만들어 확률분포의 첫 번째 모멘트를 등화 시킨다. MVN은 MFCC의 평균뿐만 아니라 분산 즉 확률분포의 두 모멘트를 등화 시킨다.

본 논문에서는 음성의 특징 계수로 사용되는 MFCC를 변형 시키는 가산 잡음과 채널 불일치 문제를 다룬다. 화자인증 시스템의 성능 개선과 환경 불일치 극복을 위해 MFCC의 히스토그램을 변환하는 기법을 적용한다. 음성인식 및 화자인증에서 히스토그램 변환 기법을 적용한 연구로 히스토그램 등화(Histogram Equalization, HE) 기법을 들 수 있다[6]-[10]. HE 기법은 특징 계수의 분포를 가우시안 분포로 매핑하는 변환 함수를 이용한다. 즉 가우시안 분포의 누적분포함수(Cumulative Distribution Function, CDF)에 입력 특징 계수의 CDF를 매핑한다. 이 변환 함수는 입력 음성 확률분포의 모든 모멘트를 가우시안 확률분포의 모멘트에 등화 시킨다. 이러한 방식으로 HE 기법은 훈련과 인식환경에서 특징 계수 확률분포의 불일치를 감소시킨다. 이 방식은 가산 잡음과 채널 불일치 환경에서 CMN과 MVN 보다 환경 불일치 문제를 잘 해결할 수 있다고 알려져 있다[6],[7]. HE 기법은 이미지 처리 기술에 응용되어 이미지의 밝기 조절과 대비(contrast) 변조에 효과적으로 사용되고 있다[11]. 이 방식은 나쁜 대비 환경에서 너무 밝거나 너무 어두운 이미지를 보상하는 데 효과적이다.

본 논문에서는 기준분포로 HE 기법에서와 같이 가우시안 분포를 사용하나, 이 분포의 표준편차를 달리하여 표준편차가 인식성능과 관계가 있음을 보인다. 이 관계를 실험을 통하여 밝히고 수학적으로도 규명한다. 또한 통계학을 이용하여 히스토그램 변환 후의 성능이 통계적으로 유의하게 개선되었음을 검증한다.

논문의 구성은 2장에서 MFCC의 히스토그램을 분석하고, 3장에서 HE 기법에 대해 설명하고, 기준분포의 표준편차와 인식성능과의 관계를 밝힌다. 4장에서는 제안한 방법에 따라 화자인증 시스템들을 실험하고, 결과를 분석한다. 그리고 5장에서 결론을 맺는다.

2. MFCC 히스토그램 분석

화자인증 시스템에서 인식 결과는 로그 유사도 차이(Log-Likelihood Ratio, LLR) 즉 스코어에 의해 결정된다. 입력 음성 각 프레임의 D차원 특징 벡터 $\vec{x}$의 수열 $X=\{\overrightarrow{x_1},\cdots,\overrightarrow{x_T}\}$에 대해 스코어는 다음 수식을 따른다[4].

$$\Lambda(X)=\log p(X|\lambda_T)-\log p(X|\lambda_U) \quad (1)$$

여기서 λ_T는 사용자 모델, λ_U는 배경화자인 UBM 모델, $\log p(X|\lambda)$는 유사도(Likelihood), $\Lambda(X)$는 스코어를 나타낸다. 입력 X에 대한 모델 λ의 로그 유사도는 다음 식에 의해 연산된다.

$$\log p(X|\lambda)=\frac{1}{T}\sum_{t=1}^{T}\log p(\overrightarrow{x_t}|\lambda) \quad (2)$$

여기서 T는 입력 특징벡터의 길이를 나타낸다.

화자인증 시스템에서 음성의 특징 벡터로 사용되는 MFCC는 다음과 같은 특성을 갖는다. 발화 음성의 길이가 충분히 클 때 MFCC의 분포는 중심극한정리에 의해 가우시안 분포에 가까워진다. 이와 같은 MFCC 분포는 분포의 평균 근처에는 높은 빈도로 값들이 존재하며, 분포의 양 끝에는 상대적으로 낮은 빈도로 값들이 존재한다. 따라서 식 (2)에서 화자모델에 대한 로그 유사도 값은 MFCC의 분포에서 빈도가 높은 평균과 그에 가까운 값에 의해 결정됨을 알 수 있다. 다른 한편, 식 (1)에서 사용자와 사칭자의 로그 유사도 분포 간에 평균의 차이가 크고 표준편차의 크기가 작을 때 변별력 있는 스코어를 산출할 수 있음을 보여 준다.

<그림1>은 MFCC 1차 계수에 대한 잡음의 영향을 보여 준다. 여기에서 SNR이 5 dB에서 20 dB인 백색 잡음이 음성신호에 더해졌다. 잡음은 음성신호의 확률분포에 크게 영향을 미쳐 그 분포의 평균을 이동시키고 표준편차를 감소시킴을 알 수 있다. 다른 MFCC 차수에서도 유사한 경향이 보인다. 일반적으로 가산 잡음과 채널은 확률분포의 표준편차를 감소시키고 분포의 모양을 변화시킨다[7],[8].

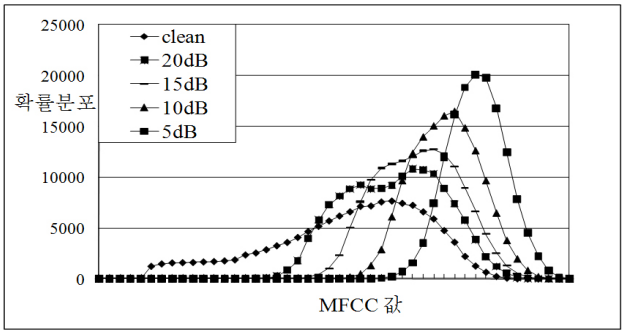


그림 1. 깨끗한 음성과 잡음 음성에서 MFCC 1차 계수의 확률분포

Figure 1. The probability distribution of clean and noisy speech for the first MFCC(clean, 20, 15, 10, 5 dB)

잡음환경에서 화자를 인증하는 경우, 훈련 시에 화자모델은

깨끗한 음성으로 만들고 인식 시에는 잡음음성이 입력된다. <그림1>에서 보듯이 잡음음성은 깨끗한 음성에 비해 MFCC 분포의 표준편차와 모양이 변형되어 두 음성의 분포는 다른 형태를 갖게 된다. 또한, 채널이 존재하는 환경에서는 훈련 시와 인식 시에 사용하는 마이크나 전화기가 달라 MFCC 분포가 다른 모습을 보인다. HE 기법에서는 훈련과 인식 시 입력 음성 모두 미리 정한 기준분포를 갖도록 하여 이와 같은 환경 불일치 문제를 해결한다.

3. 히스토그램 등화 기법과 표준편차 효과

3.1 히스토그램 등화 기법

HE 기법은 입력 MFCC의 CDF와 기준분포의 CDF를 매칭하는 방법을 사용한다. 이런 방법으로 입력 MFCC를 기준분포의 값으로 매핑하여 변환된 MFCC 값을 구한다. <그림2>는 CDF의 매칭을 이용하여 원래의 데이터 x가 y로 변환되는 것을 보여준다[7].

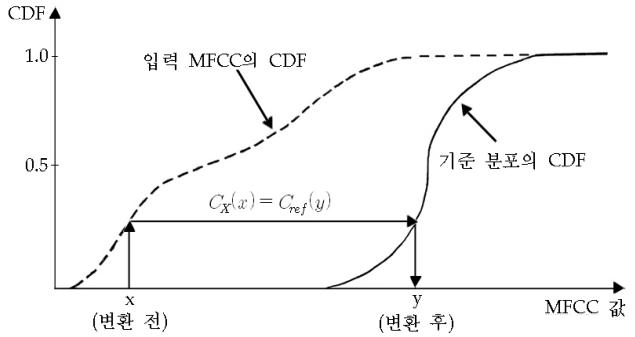


그림 2. CDF 매칭에 의한 MFCC 변환
Figure 2. MFCC transformation by matching CDFs

다음 1) ~ 6)의 과정과 <그림3>은 입력 MFCC의 CDF를 기준분포의 CDF로 매칭하여 원래의 MFCC 값 x를 y로 변환하는 과정을 보여준다.

1) 입력 MFCC 각 차수에 대해 분포를 구하여 각 차수의 최대값 $x_{\max}$과 최소값 $x_{\min}$을 결정한다.
2) $[x_{\min}, x_{\max}]$의 범위를 M등분한다. 등분된 각 구간을 $R_i = [r_i, r_{i+1}]$라고 하자.
3) 각 구간 R_i에서 입력 MFCC 각 차수의 히스토그램을 구한다. 이는 각 구간에서 MFCC의 빈도 n_i를 구하면 된다.
4) 3)에서 구한 MFCC 히스토그램을 다음 식으로 정규화 한다.

$$f_X(x \in R_i) = \frac{n_i}{N_X} \tag{3}$$

여기서 N_X는 입력 음성 MFCC 전체의 빈도이다.

5) 식 (3)의 히스토그램으로부터 누적 히스토그램을 다음과 같이 구한다.

$$C_X(x \in R_i) = \sum_{j=1}^{i} \frac{n_j}{N_X} \tag{4}$$

6) 입력 MFCC x를 $C_X(x) = C_{REF}(y)$를 만족하는 값 y로 변환한다. 여기서 $C_{REF}(y)$는 기준분포의 CDF이고, 기준분포는 가우시안 분포를 사용한다.

본 논문에서 M값을 1000으로 하였다. 분포의 등분을 250 단계로 한 HE 기법과 달리, 1000 단계로 한 것은 분포를 변환할 때 양자화 잡음의 영향을 최소화하기 위한 것이다.

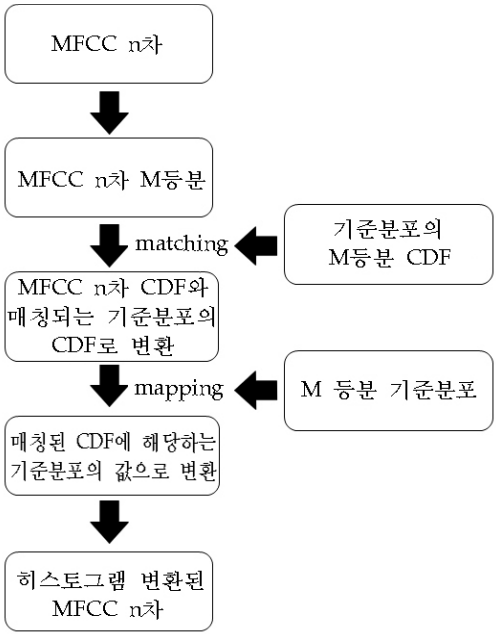


그림 3. 히스토그램 등화 과정
Figure 3. Histogram equalization

3.2 기준분포의 표준편차와 인식성능과의 관계

HE 기법에서 기준분포는 평균이 0이고 표준편차가 1인 가우시안을 사용한다[6],[7]. 본 논문에서는 기준분포의 표준편차를 달리 하여 표준편차가 인식성능에 영향을 미치는 요소임을 제안한다. 이를 규명하기 위해 사용자 입력 음성의 MFCC 값이 사용자 모델의 평균에 가깝고 배경화자 모델의 평균에 멀수록 스코어 값이 커질 수 있음을 밝힌다.

식 (2)의 유사도는 다음과 같이 표현할 수 있다.

$$p(\overrightarrow{x_t}|\lambda) = \sum_{i=1}^{M} w_i p_i(\overrightarrow{x_t}|\lambda) \tag{5}$$

여기서 w_i는 mixture 가중치이고 M은 mixture 수이다.

$p_i(\vec{x_t}|\lambda)$는 가우시안 확률분포를 가지며 다음 식으로 나타낼 수 있다.

$$p_i(\vec{x_t}|\lambda) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_i|}} exp\left\{-\frac{1}{2}(\vec{x_t}-\vec{\mu_i})'\vec{\Sigma_i}^{-1}(\vec{x_t}-\vec{\mu_i})\right\} \quad (6)$$

여기서 $\vec{\mu_i}$는 MFCC 각 차수의 평균을 나타내는 D차원 벡터이고, $\vec{\Sigma_i}$는 $(D\times D)$ diagonal 공분산 행렬이다.

식의 전개를 단순하게 하기 위해 벡터를 스칼라로 바꾸고 ($D=1$로 가정한다. $D>1$인 경우는 MFCC 각 차수가 서로 독립이므로 쉽게 일반화할 수 있다), mixture 수를 1로 하여 식 (5), (6)에 적용하면 $p(\vec{x_t}|\lambda)$은 다음과 같이 된다.

$$p(x_t|\lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x_t-m)^2}{2\sigma^2}\right\} \quad (7)$$

여기서 m, σ^2은 화자모델 λ의 평균과 분산을 나타낸다. 식 (7)의 양변에 자연로그를 취하면 다음과 같다.

$$\log p(x_t|\lambda) = -\frac{1}{2}log(2\pi\sigma^2) - \frac{(x_t-m)^2}{2\sigma^2} \quad (8)$$

식 (8)을 식 (2)의 우변에 대입하면, 식 (1)의 스코어는 다음과 같이 구할 수 있다.

$$\Lambda(X) = \frac{1}{T}\sum_{t=1}^{T}[\left\{-\frac{1}{2}log(2\pi\sigma_T^2) - \frac{1}{2}\frac{(x_t-m_T)^2}{\sigma_T^2}\right\}$$
$$-\left\{-\frac{1}{2}log(2\pi\sigma_U^2) - \frac{1}{2}\frac{(x_t-m_U)^2}{\sigma_U^2}\right\}] \quad (9)$$
$$= \frac{1}{T}\sum_{t=1}^{T}[-\frac{1}{2}log\frac{\sigma_T^2}{\sigma_U^2} - \frac{1}{2}\left\{\frac{(x_t-m_T)^2}{\sigma_T^2} - \frac{(x_t-m_U)^2}{\sigma_U^2}\right\}] \quad (10)$$

여기서 $m_T, \sigma_T^2, m_U, \sigma_U^2$은 각각 사용자 모델의 평균과 분산, UBM의 평균과 분산을 나타낸다.

4.1 실험방법에서, GMM-UBM 방식에서 UBM 모델에 MAP 적응을 하여 화자모델을 만들 때 분산은 그대로 두고 평균만 갱신한다[4]. 따라서 $\sigma_T^2=\sigma_U^2=\sigma^2$이므로 식 (10)은 다음과 같이 된다.

$$\Lambda(X) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2\sigma^2}\left\{(x_t-m_U)^2 - (x_t-m_T)^2\right\} \quad (11)$$

식 (11)에서 입력 특징 계수 x_t가 사용자 모델의 평균 m_T 근처에 많이 분포하면 둘째 항이 작은 값을 갖게 되고, UBM의 평균 m_U에서 먼 곳에 많이 분포하면 첫째 항이 큰 값을 갖게 되므로 스코어는 커지게 된다. 즉, 입력 특징 계수가 사용자 것이라면 스코어는 크게 될 것이고 사칭자 것이라면 작게 되어, 변별력 있는 스코어를 산출할 수 있음을 알 수 있다.

이와 같이 입력 특징 계수 분포가 평균 근처에서 좁을수록 즉 분포의 표준편차가 작을수록 스코어가 변별력을 갖게 되고 인식 성능이 향상될 수 있음을 예상할 수 있다. 본 논문에서 실험을 통하여 이를 검증하기 위하여 기준분포로 평균이 0인 가우시안 분포를 사용하나, 표준편차가 1 뿐만 아니라 0.25, 0.5, 2, 3 등으로 달리하여 화자인증을 한 뒤 성능을 비교한다.

4. 실험 방법 및 결과

4.1 실험 방법

현재 대부분의 문장 독립 화자인증 시스템에서는 GMM-UBM 방식[4]을 사용하므로, 본 논문에서도 이 방법을 사용한다. GMM은 EM(Expectation-Maximization) 알고리즘으로 2, 4, 8, 16, 32, 64, 128, 256, 512 순으로 mixture를 증가시킨 ML(Maximum Likelihood) 모델이다. 이 시스템에서는 목적 화자 음성으로 훈련되는 화자별 GMM 이외에 UBM이라는 GMM을 하나 더 필요로 한다. UBM은 음성 특징의 화자 독립 분포를 표현하여 배경화자를 대표하는 하나의 GMM이다. 그리고 화자인증 시스템에서 화자모델을 만들 때 대부분 훈련 음성 자료를 충분히 얻기 어려운 경우가 많다. 적은 훈련 음성 자료를 효과적으로 이용하는 방법으로 화자적응이 있다. 이는 UBM으로부터 화자적응을 통하여 화자 모델을 훈련하여 각각의 화자 모델을 생성하는 것이다. 본 논문에서는 화자 적응 방법 중 MAP(Maximum A Posteriori) 화자적응 기법[4]을 사용한다.

본 논문의 실험에서 사용한 음성 DB는, ETRI 음성정보 연구센터에서 구축한 한국어 화자인식용 영리용 음성 DB로, SNR 25 dB 이상 확보 가능한 조용한 사무실 PC 환경에서 중가의 마이크(모델명: Sennheiser MD425)를 사용하여 수집하였고, 16 kHz, 16 bit, Linear PCM으로 저장되었다. 250명(기간별: 주차 100명, 월차 100명, 3개월차 50명)의 화자가 발성한 2연 숫자, 4연 숫자, 문장으로 구성되어 있다. 문장 음성의 발성목록은 개인정보와 관련된 10개의 질문과 3어절 이내로 구성된 단문 10개로 구성되며, 한 화자당 동일한 목록을 5회 발성하고, 녹음 간격에 따라 주차/월차/3개월차로 구분하여 4회 반복한 것이다.

UBM 작성은 월차화자 음성 중 기간별 첫 녹음 시점인 0개월차 월차화자 100명으로 남자 50명과 여자 50명으로 구성하였다. 훈련 DB의 환경적 데이터가 균형이 맞으므로 남녀 화자 모두의 음성을 사용하여 하나의 UBM을 작성하였다. UBM 훈련 시 각 화자당 6단문의 5회 발성 음성으로, 화자당 약 72초 정도로 총 2시간 분량을 사용하였다. Mixture 수는 512로 하였다.

UBM에 실험 대상 화자모델의 MAP 화자적응 시 각 화자의 0주차(주차 화자의 음성 중 기간별 첫 녹음시점) 6개 단문의 60초로 제한된 5회 발성음성을 사용하였다. MAP 화자적응 시 모델의 평균만 갱신하였다[4].

본 논문에서 깨끗한 환경에서 화자인증 실험의 테스트 음성 DB는 주차화자 100명을 대상으로 하였는데, 남자 50명과 여자 50명으로 구성하였다. 사용된 테스트 음성은 훈련 시점과 1주일 차이의 음성인 주차화자의 1주차 음성이다. 이 테스트 음성 DB는 훈련과 독립적인(훈련에 사용되지 않은 단문) 3개 단문의 5회 음성을 사용하여 각 화자당 15 단문으로 하였다.

본 논문에서 화자와 사칭자의 비율을 1:10으로 하였으므로, 각 화자당 사칭자 문장 수는 150개 이다. <표1>은 화자인증에 사용한 화자군의 분류와 음성시료의 개수를 보여주고 있다.

표 1. 실험에 사용한 화자 및 음성시료의 구성
Table 1. Speaker and speech DB for experiments

	남자	여자	전체
화자 수	50 명	50 명	100 명
사용자 테스트 음성시료	750 단문	750 단문	1,500 단문
사칭자 테스트 음성시료	7,500 단문	7,500 단문	15,000 단문

채널 불일치 실험을 위해 다양한 종류의 전화기에서 수집한 음성 DB를 사용하였다. 이 음성 DB도 ETRI에서 수집한 것으로, 훈련 음성과 테스트 음성은 서로 다른 종류의 전화기로 수집한 것이다.

사용된 모든 음성은 음성 구간을 검출하여 앞, 뒤의 묵음을 제거한 음성을, 프레임 길이는 25ms, 프레임 주기는 10ms이며 Hamming window을 사용하여 HTK ver 3.3[12]을 이용하여 MFCC를 추출하였다. MFCC 12차에 에너지를 더하고 여기에 delta와 delta-delta를 추가하여 MFCC 39차를 사용하였다.

4.2 기준분포의 표준편차에 따른 유사도 분포 비교

채널 불일치 환경의 남성에 대한 화자인증 실험에서, 표준편차에 따른 사용자 및 사칭자의 로그 유사도 분포가 <그림4>에 보인다. <그림4>에서 실선과 점선은 각각 사용자와 사칭자의 로그 유사도 분포를 나타낸다. <그림4 (a)>의 기본 시스템은 히스토그램 변환 기법을 적용하지 않은 화자인증 시스템이다. 사용자와 사칭자의 로그 유사도 분포에서 두 분포가 겹치는 부분은 에러 영역으로 넓이의 1/2은 곧 EER(Equal Error Rate)을 뜻한다. 따라서 두 분포의 평균 차이와 표준편차는 곧 화자인증 성능의 지표가 된다.

<그림4>에 로그 유사도 분포의 모양을 보였으나 그 차이가 분명히 드러나지 않으므로, t-검정[13]을 시행하여 두 로그 유사도 분포의 차이가 통계적으로 유의한가를 검증하고, 어느 경우에 두 분포의 차이가 가장 큰 가를 확인한다. <표2>에서 표준편차 값에 상관없이 p 값이 모두 0.05보다 작으므로 95% 신뢰수준에서 두 분포가 통계적으로 유의하게 차이가 있음을 알 수 있다. 그런데, 표준편차 값에 따라 t 값에 차이가 있다. 표준편차가 작을수록 t 값이 크므로 두 분포의 차이가 더 커짐을 알 수 있다. 이와 같이 표준편차를 작게 하면 사용자와 사칭자의 로그 유사도 분포의 변별력이 커짐을 알 수 있다. 즉, 히스토그램 변환에서 기준분포의 표준편차는 로그 유사도 분포를 변환하여 화자인증 시스템의 성능에 영향을 주는 요인임을 확인할 수 있다.

4.3 화자인증 실험결과

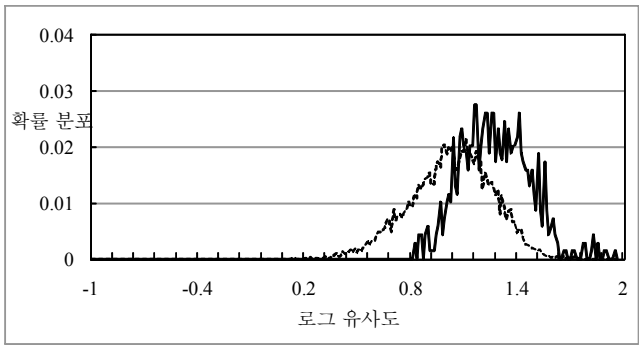
<표3>에 조용한 훈련 및 테스트 환경에서 각 시스템의 성능을 보여주는 EER을 정리하였다. 일반적으로 채널 보상 방법을 깨끗한 음성에 적용하면 음성에 왜곡이 발생하므로 성능이 나빠지는 경향이 있는데[8],[14], 본 실험에서도 비슷한 경향을 보였다. 남자의 경우 모든 표준편차에서 기본 시스템 대비 성능이 나빠졌으나, 여자의 경우 표준편차가 3.0인 경우를 제외하고 성능이 개선되었다. 평균적으로, 표준편차가 0.25와 0.5인 경우 기본 시스템 보다 성능이 약간 개선되었고, 나머지 경우는 저하되었다. <표3>에서 주목할 점은, 표준편차에 따른 인식률을 보면 표준편차가 작을수록 인식성능이 좋다는 것이다.

표 3. 깨끗한 환경에서 기준분포의 표준편차에 따른 화자인증 EER(%) 비교
Table 3. Comparison of EERs according to standard deviations of the reference distribution in clean environments

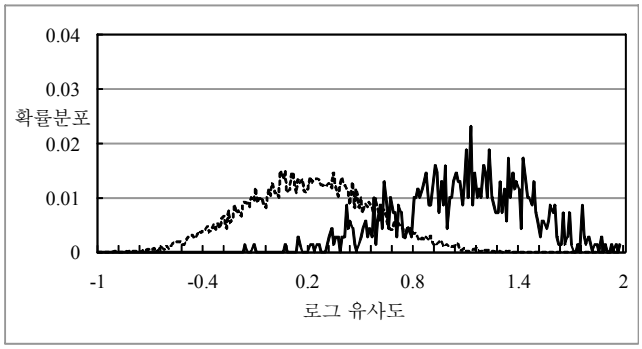
표준편차	남자	여자	평균
기본 시스템	3.62	5.49	4.56
1.0	4.27	4.93	4.60
0.25	4.15	4.40	4.28
0.5	4.18	4.62	4.40
2.0	4.33	5.18	4.78
3.0	5.86	6.94	6.4

채널 불일치 환경에 대한 화자인증 실험결과는 <표4>와 같다. 실험결과의 EER을 비교하면, 표준편차가 0.25와 0.5인 경우 1.0인 경우 보다 평균적으로 각각 13.1%와 7.4%의 에러가 감소하였다. 표준편차가 2.0와 3.0인 경우 1.0인 경우 보다 평균적으로 각각 8.4%와 19.4%의 에러가 증가하였다. 이 환경에서도 기준분포의 표준편차가 작을수록 인식성능이 좋음을 알 수 있다.

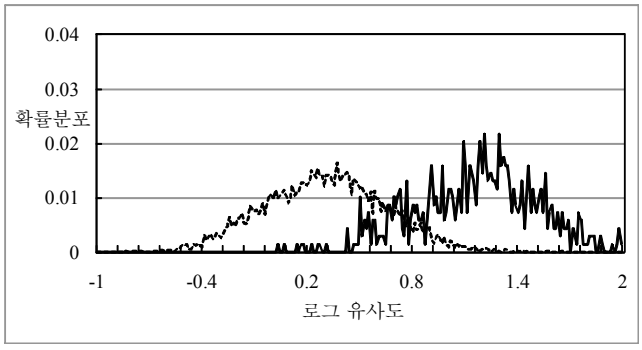
이러한 성능 개선의 정도가 통계적으로 유의한가를 보기 위하여 t-검정을 시행하였다. <표4>에 보이는 EER 값은 화자인증 실험한 여러 화자의 평균값으로, 실제로는 화자마다 성능이 다르다. 시행한 t-검정에서 각 화자의 인식률을 입력 데이터로 삼아, 표준편차가 1.0인 경우를 기준으로 표준편차가 0.25, 0.5,



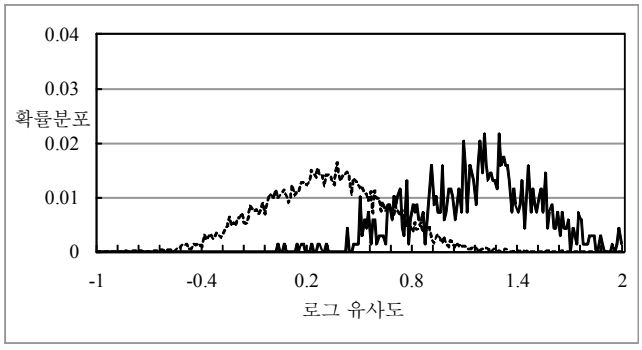
(a) 기본 시스템
(a) A basic system



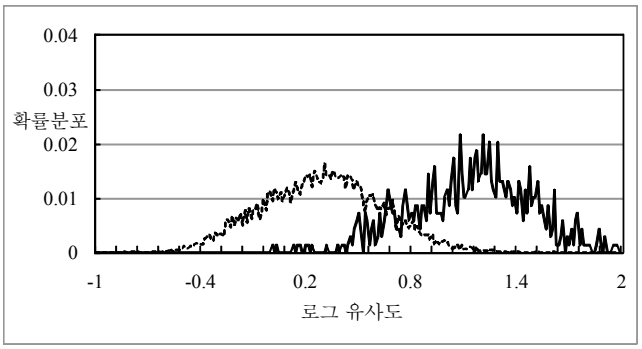
(b) 표준편차가 1.0인 경우
(b) in case of standard deviation = 1.0



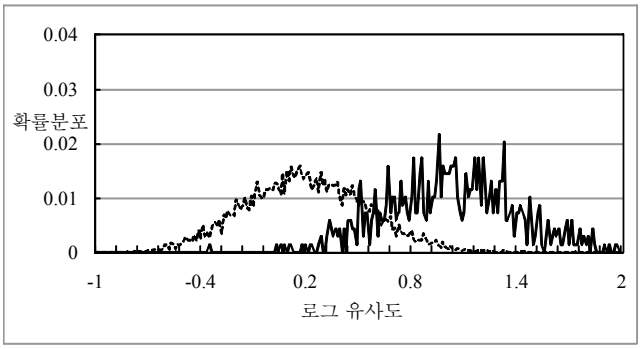
(c) 표준편차가 0.25인 경우
(c) in case of standard deviation = 0.25



(d) 표준편차가 0.5인 경우
(d) in case of standard deviation = 0.5



(e) 표준편차가 2.0인 경우
(e) in case of standard deviation = 2.0



(f) 표준편차가 3.0인 경우
(f) in case of standard deviation = 3.0

그림 4. 남성 사용자와 사칭자 로그 유사도 분포
Figure 4. Log-likelihood distributions for male target speakers and impostors

2.0, 3.0인 경우와 비교하여 인식률의 평균차이, t 값, p 값을 구하였다. <표5>를 보면 p 값이 모두 0.05 보다 작으므로 95% 신뢰 수준에서 성능의 차이가 통계적으로 유의하다는 것을 알 수 있다.

표 2. 남성 사용자와 사칭자 로그 유사도 차이 t-검정 결과
Table 2. T-test results for log-likelihood differences of male target speakers and impostors

표준편차	기본 시스템	1.0	0.25	0.5	2.0	3.0
사용자 평균	1.311	1.105	0.903	1.143	1.029	0.986
사칭자 평균	0.899	0.223	0.535	0.305	0.210	0.229
사용자 표준편차	0.191	0.264	0.149	0.244	0.366	0.361
사칭자 표준편차	0.594	0.356	0.147	0.338	0.340	0.357
평균차이	0.412	0.882	0.368	0.838	0.819	0.757
t 값	40.41	60.87	62.64	61.97	56.34	53.05
p 값	0.000	0.000	0.000	0.000	0.000	0.000

표 4. 채널 불일치 환경에서 기준분포의 표준편차에 따른 화자인증 EER(%) 비교
Table 4. Comparison of EERs according to standard deviations of the reference distribution in channel mismatched environments

표준편차	남자	여자	평균
기본 시스템	28.11	27.98	28.05
1.0	12.62	14.93	13.77
0.25	11.24	12.70	11.97
0.5	11.88	13.64	12.76
2.0	13.55	16.31	14.93
3.0	15.22	17.69	16.46

표 5. 채널 불일치 환경에서 화자인증 실험결과 성능 검증
Table 5. Performance verification of experimental results in channel mismatched environments

표준편차	인식률의 평균차이	t 값	p 값
1.0 vs 0.25	1.80	5.094	0.000
1.0 vs 0.5	1.01	2.647	0.010
1.0 vs 2.0	-1.16	-2.867	0.005
1.0 vs 3.0	-2.69	-4.559	0.000

지금까지의 실험 결과는 잡음이 없는 조용한 환경 및 채널 불일치 환경의 화자인증 시스템에서 얻어진 것이다. HE 기법은 채널 불일치 환경뿐만 아니라 잡음 환경에서도 좋은 성능을 보여 준다[6]-[8]. 본 논문에서 사용한 알고리즘은 HE 기법에 기반 하므로 잡음 환경에서도 좋은 결과가 나올 것으로 생각된다. 이 추측을 확인할 수 있는 실험이 뒤따라야 하겠다.

5. 결론

본 논문은 잡음과 채널 불일치 환경에 강인한 문장 독립 화자인증 시스템의 성능 개선을 위하여 HE 기법에 기반한 히스토그램 변환 방식을 제안하였다.

잡음 음성이나 채널 음성의 MFCC 분포가 깨끗한 음성의 분포와 다르다는 사실을 보이고, HE 기법을 적용하여 훈련과 인식 시 입력 음성 모두 가우시안 분포를 갖도록 하여 이와 같은 환경 불일치 문제를 감소시켰다.

MFCC가 분포의 평균 근처에서 빈도가 높고, 로그 유사도 값이 이 영역의 MFCC 값에 의해 결정됨을 보이고, 수식을 전개하여 MFCC가 사용자 모델의 평균에 가깝고 UBM 분포의 평균에 먼 값들이 많을수록 변별력 있는 스코어를 산출할 수 있음을 밝혔다. 이와 같은 사실을 화자인증 성능 개선에 적용하기 위해 HE 기법에서 기준분포로 이용한 가우시안 분포의 표준편차를 달리하면서 실험하였다. 실험 결과 표준편차가 작을수록 채널 불일치 환경에서 화자인증 시스템의 성능이 개선됨을 알 수 있다.

그동안 음성인식이나 화자인식 분야에서는 단순히 실험결과 성능만을 제시했으나 본 논문에서는 통계학을 적용하여 성능을 검증하였다. 통계학을 이용하여 성능을 검증한 점도 이 논문이 기여한 바라고 말할 수 있다.

감사의 글

이 논문은 교육과학기술부 2007년도 한국학술진흥재단(No. KRF-2007-D00741 (I00101))과 한국한의학연구원 기관고유 사업인 체질건강수준 표준개발과제 (K10070)의 지원으로 수행되었음.

참고문헌

[1] Yu, H.J. (2004). “An overview and market review of speaker recognition technology”, *Proc. 2004 Spring Conf. on the Korean Society of Phonetic Sciences and Speech Technology*, pp. 91-97. (유하진, (2004). “화자인식 기술 및 국내외시장 동향”, 대한음성학회 2004 봄 학술대회 발표논문집, pp. 91-97.)

[2] Auckenthaler, R., Carey, M., Lloyd-Thomas, H. (2000). “Score normalization for text-independent speaker verification systems”, *Digital Signal Processing*, Vol. 10, pp. 42-54.

[3] Auckenthaler, R., Parris, E., Carey, M. (1999). “Improving a GMM speaker verification system by phonetic weighting”, *Proc. International Conf. on Acoustics Speech Signal Proc.,* Vol. 1, pp. 313-316.

[4] Reynolds, D.A., Quatieri, T.F., Dunn, R.B. (2000), “Speaker verification using adapted Gaussian mixture models”, *Digital Signal Processing*, Vol. 10, pp. 19-41.

[5] Campbell, J.P. (1997). “Speaker recognition: a tutorial”, *Proceedings of the IEEE*, Vol. 85, No. 9, pp. 1437-1462.

[6] de la Torre, A., Peinado, A.M., Segura, J.C., Perez-Cordoba, J.L., Benitez, M.C., Rubio, A.J. (2005). “Histogram equalization of speech representation for robust speech recognition”, *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 3, pp. 355-366.

[7] Skosan, M., Mashao, D. (2006). “Modified segmental histogram equalization for robust speaker verification”, *Pattern Recognition Letters,* Vol. 27, No. 5, pp. 479-486.

[8] Pelecanos, J., Sridharan, S. (2001). “Feature warping for robust speaker verification”, *Proc. Odyssey 2001*, Vol. 1, pp. 213-218.

[9] Xiang, B., Chaudhari, U.V., Navratil, J., Raamaswamy, G.N., Gopinath, R.A. (2002). “Short-time Gaussianization for robust

speaker verification", *Proc. International Conf. on Acoustics Speech Signal Proc.*, Vol. 1, pp. 681-684.

[10] Seo, Y.J., Kim, H.R., Lee, Y.K. (2006). "Robust speech recognition based on class histogram equalization", *Malsori*, Vol. 60, pp. 145-164.
(서영주, 김회린, 이윤근, (2006). "클래스 히스토그램 등화 기법에 의한 강인한 음성인식", 한국음성학회 말소리, 제60호, pp. 145-164.)

[11] Grundland, M., Dodgson, N.A. (2005). "Color histogram specification by histogram warping", *Proc. SPIE*, Vol. 5667, pp. 610-621.

[12] Young, S. (2001). *The HTK Book*, Cambridge University Engineering Department.

[13] Sung, T.J. (2007). *Understanding and application of modern basic statistics*, Kyoyookbook.
(성태제, (2007). 현대 기초통계학의 이해와 적용, 교육과학사.)

[14] Reynolds, D. (1995). "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication*, Vol. 17, No. 1-2, pp. 91-108.

• **권철홍 (Kwon, Chul Hong),** 교신저자
대전대학교 정보통신공학과
대전시 동구 용운동 96-3
Tel: 042-280-2555 Fax: 042-280-2559
Email: chkwon@dju.ac.kr
관심분야: 화자인증, 음성공학
1997~현재 정보통신공학과 교수