강인한 화자 확인을 위한 히스토그램 개선 기법*

최재길, 권철홍(대전대)

<차 례>

<Abstract>

Histogram Enhancement for Robust Speaker Verification

Jae-kil Choi, Chul-hong Kwon

It is well known that when there is an acoustic mismatch between the speech obtained during training and testing, the accuracy of speaker verification systems drastically deteriorates. This paper presents the use of MFCCs' histogram enhancement technique in order to improve the robustness of a speaker verification system. The technique transforms the features extracted from speech within an utterance such that their statistics conform to reference distributions. The reference distributions proposed in this paper are uniform distribution and beta distribution. The transformation modifies the contrast of MFCCs' histogram so that the performance of a speaker verification system is improved both in the clean training and testing environment and in the clean training and noisy testing environment.

* Keywords: Robust speaker verification, Histogram enhancement, Acoustic mismatch.

* 본 연구는 산업자원부의 지역혁신 인력양성 사업의 지원과 대전대학교의 연구 지원으로 수행되었음.

1. 서 론

현재 화자 인식 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 훈련 환경과 인식 환경이 달라질 경우 시스템의 인식 성능을 크게 하락시킨다. 이러한 성능 저하는 훈련과 인식 환경의 불일치 때문이다[1].

본 논문에서는 화자 확인의 성능 향상을 목적으로 음성의 특징으로 사용되는 mel-frequency cepstral coefficients (MFCC)의 분포에 대해 히스토그램(Histogram) 처리 기법을 적용한다. 히스토그램 처리 기법은 이미지 처리 기술에 응용되어 이미지의 밝기 조절과 대비(Contrast) 변조에 효과적으로 사용되고 있다[2]. 음성인식 및 화자인식에서 히스토그램 처리 기법을 사용한 선례로 히스토그램 등화(Equalization) 기법의 적용을 들 수 있다[3][4]. 이는 훈련환경과 인식환경의 불일치에 대한 채널 및 잡음 보상의 의도로 사용되었다.

이 논문에서는 화자 확인 시스템의 성능 개선과 환경 불일치 극복을 위해 발화 음성 단위로 MFCC의 히스토그램을 변환하는 기법을 제안한다. MFCC의 히스토그램 변환은 MFCC의 히스토그램에 나타나는 MFCC 값들 간 차이의 크기 즉 MFCC 값들의 대비를 개선하는 과정으로 히스토그램 개선 기법이라고 명명한다. 여기서 대비라는 말은 화자 확인에서 사용하는 음성 특징의 화자 종속적인 정보(즉, 화자 간의 차이를 보여주는 정보) 간 크기의 차이를 말한다. MFCC 값들의 대비는 MFCC의 히스토그램에 포함된 고유정보이며, 화자 확인 성능의 직접적인 영향 요소로 가정하고 그에 대한 증명을 위해 여러 가지 기준 분포로 MFCC의 히스토그램을 변환하여 성능을 평가하였다. 그리고 히스토그램 개선에 의해 변환된 MFCC가 화자 확인의 성능에 미치는 영향을 분석하였다. 또한, 조용한 환경의 훈련 음성과 가산 잡음 환경의 인식으로 훈련과 인식의 불일치에 대한 화자 확인 성능을 비교 평가하였다.

논문의 구성은 2장에서는 MFCC의 히스토그램을 분석하고, 3장에서는 본 논문에서 제안한 히스토그램 개선 기법에 대해 설명한다. 4장에서는 3장에서 제안한 화자 확인 시스템들을 실험하고, 결과를 분석한다. 그리고 5장에서 결론을 맺는다.

2. MFCC 히스토그램 분석

화자 확인 시스템에서 인식 결과를 결정하는 요소인 스코어(Score)는 입력 특징 벡터 $X=\{x_1,\cdots,x_T\}$에 대해 다음 수식을 따른다[5].

$$\Lambda(X) = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{\overline{hyp}}) \quad (1)$$

여기서 λ_{hyp}는 사용자 모델, $\lambda_{\overline{hyp}}$는 배경화자 모델, $\log p(X|\lambda)$는 유사도(Likelihood), $\Lambda(X)$는 스코어를 나타낸다. 입력 특징벡터 X에 대한 모델 λ의 로그 유사도는 다음 식에 의해 연산된다.

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

여기서 T는 입력 특징벡터의 길이를 나타낸다. 식 (2)를 보면 모델에 대한 로그 유사도 값은 특징 벡터 X의 분포 중 빈도가 높은 평균과 그에 가까운 프레임 값에 의해 결정됨을 알 수 있다. 이것은 식 (1)에서 사용자와 사칭자의 음성 특징 벡터 분포의 평균값의 차이가 크고 분산의 크기가 작을 때 변별력 있는 스코어를 산출할 수 있음을 보여준다.

화자 인식 시스템에서 음성의 특징 벡터로 사용되는 MFCC는 다음과 같은 한계를 갖는다. 발화 음성의 크기가 충분히 클 때, 일반적으로 통계적 MFCC의 분포는 중심극한정리에 의해 가우시안 분포에 가까워진다. 이와 같은 MFCC 분포는 분포의 평균 근처에는 고밀도(높은 빈도수) 값들이 존재하며, 분포의 양 끝에는 상대적으로 저밀도(낮은 빈도수) 값들이 존재한다. 이러한 MFCC의 분포는 서로 다른 발화자의 입력 음성 MFCC 분포의 평균과 그 근처 값들에서 대비가 낮은 즉 저대비 분포라고 말할 수 있다. 일반적인 화자 확인 시스템에서 이러한 MFCC 값들의 저대비는 MFCC 값들 간 특성의 차이가 작으므로 인식성능을 저하시키는 요인이 된다. 따라서 화자 확인 시스템에서 높은 확률의 입력에 해당하는 MFCC 분포의 평균에 해당되는 저대비 특성을 갖는 다수의 값들 간 대비를 개선하면 화자 확인 성능을 향상시킬 수 있다.

3. MFCC 히스토그램 개선 기법

앞에서 설명한 대로 음성의 통계적 MFCC 분포는 화자 확인 시스템에서 화자 간 변별력을 확보하지 못하므로 성능을 저하시킨다. 본 논문에서는 이러한 성능 저하를 피하기 위해 MFCC의 히스토그램을 분포의 평균과 그에 가까운 값들의 대비를 향상시키도록 변환하는 기법인 히스토그램 개선 기법을 제안한다.

3.1 누적 분포 매칭 분포 변환 기법

본 논문에서 제안한 히스토그램 개선 기법(Histogram Enhancement, HEN)에서 MFCC 분포 변환은 히스토그램 등화 기법(Histogram Equalization, HEQ)[4]에서 제시한 누적분포(Cumulative Distribution) 매칭의 분포 변환 기법을 사용한다. HEQ는 훈련과 인식의 환경 불일치에 강한 특징 변환 기법으로 화자 확인에서 Skosan과 Mashao가 제안한 선례가 있다[3].

누적 분포 매칭의 분포 변환 기법은 입력 음성 MFCC 분포의 누적분포 함수(Cumulative Distribution Function, CDF)를 기준 확률분포의 누적분포로 매칭하여 변환한다[3]. 먼저, x를 확률분포 $p_x(x)$를 갖는 랜덤 변수라고 하고, 변환된 값은 $y = T(x)$라고 가정하자. y는 확률분포 $p_x(x)$에 대응되는 기준 확률분포 $p_{ref}(y)$의 유일한 값으로 $p_x(x)$를 $p_{ref}(y)$로 단조 변환한다. 변환 $T(x)$는 구간 dx에서 x를 찾을 확률과 구간 dy에서 y를 찾을 확률을 같게 만든다. 식으로 나타내면 다음과 같다.

$$p_{ref}(y)dy = p_x(x)dx. \tag{3}$$

그리고 변환 $y = T(x)$는 원래의 확률분포 $p_x(x)$에 대해 다음을 만족한다.

$$p_{ref}(y) = p_x(x)\frac{dx}{dy} = p_x(G(y))\frac{dG(y)}{dy}. \tag{4}$$

$G(y) = x$는 $T(x)$의 역함수이다. 위의 수식을 사용하면 누적분포 함수 $C_x(x)$, $C_{ref}(y)$는 $p_x(x)$, $p_{ref}(y)$와 다음과 같은 관계를 갖는다.

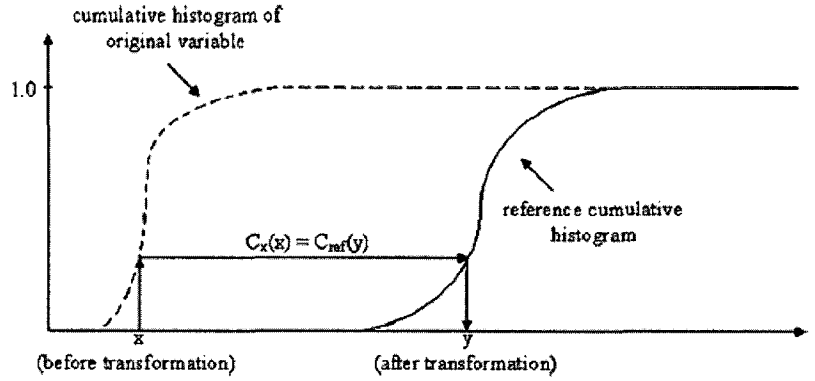
$$\begin{aligned} C_x(x) &= \int_{-\infty}^{x} p_x(x')dx' = \int_{-\infty}^{T(x)} p_x(G(y'))\frac{dG(y')}{dy'}dy' \\ &= \int_{-\infty}^{y} p_{ref}(y')dy' = C_{ref}(y) = C_{ref}(T(x)). \end{aligned} \tag{5}$$

여기서 $T(x)$는 $p_x(x)$를 $p_{ref}(y)$로 변환하며, 위 식을 $T(x)$에 대해 풀면 다음과 같다.

$$T(x) = C_{ref}^{-1}(C_x(x)) \tag{6}$$

여기서 C_{ref}^{-1}는 기준 누적분포의 역함수이다. 이의 실제 구현은 관측 범위가 유한한 수일 때만 가능하다. 결과적으로 누적분포 함수 대신 누적 히스토그램이 사용된다. 그런데 이 변환을 화자 확인의 특징 추출 모듈에 의해 획득한 다차원 특징 벡터에 쉽게 적용하기 힘들다. 그런 이유로, 모든 차원 특징 벡터들이 서로 독립적이라고 가정한다. 이 가정에 따라 각 특징 벡터 요소마다 독립적으로 변환을 적용할 수 있다[3].

HEQ는 입력 음성의 MFCC 누적분포 함수와 기준분포의 누적분포 함수를 매칭하는 방법을 사용한다. 즉 MFCC 원 데이터를 누적분포 함수의 매칭되는 기준분포의 값으로 매핑하여 변환된 MFCC 값을 산출한다. <그림 1>은 누적분포 함수의 매칭을 이용하여 원래의 데이터 x가 y로 변환되는 것을 보여준다[3].

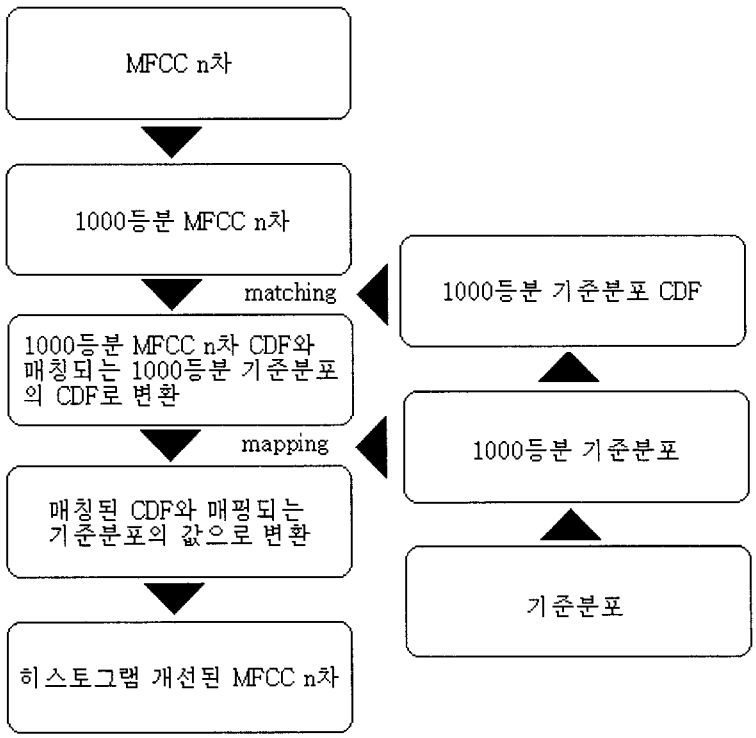


<그림 1> 누적 분포 함수 매칭 과정

3.2. 히스토그램 개선 기법 구현

HEN는 MFCC 값들의 대비를 개선할 목적으로 발화 음성 단위로 MFCC 각 차수의 히스토그램을 변환한다. 다음 1)~6)의 과정은 MFCC 분포를 기준 분포로 변환하는 과정으로 원래의 MFCC 값 x가 y로 변환되는 과정을 나타낸다.

1) 입력 음성에 대해 MFCC 각 차수의 최대값과 최소값을 결정한다. 이를 각각 $x_{\max}$, $x_{\min}$이라 한다.
2) $[x_{\min}, x_{\max}]$의 범위를 M 등분한다. 등분된 각각의 범위를 $B_i=[b_i, b_{i+1}]$라 하면, $x_{\min} = b_1 < b_2 < \cdots < b_{M+1} = x_{\max}$를 만족한다.
3) 이 범위들을 이용하여 입력 음성의 MFCC 각 차수의 히스토그램을 구한다.



<그림 2> 히스토그램 개선 과정

이는 입력에 대한 각 범위마다의 빈도를 구하면 된다.

4) 3)에 의해 구해진 MFCC 히스토그램을 다음 식으로 정규화한다.

$$p_x(x \in B_i) = \frac{n_i}{N_x} \tag{7}$$

여기서 n_i는 범위 B_i에서의 빈도이며, N_x는 입력 음성 MFCC 전체의 빈도이다.

5) 정규화된 MFCC 히스토그램으로 부터 누적 히스토그램을 구한다. 이를 수식으로 표현하면 다음과 같다.

$$C_x(x \in B_i) = \sum_{j=1}^{i} \frac{n_j}{N_x} \tag{8}$$

이는 실제 누적 히스토그램의 이산 함수로 구한 근사치이다.

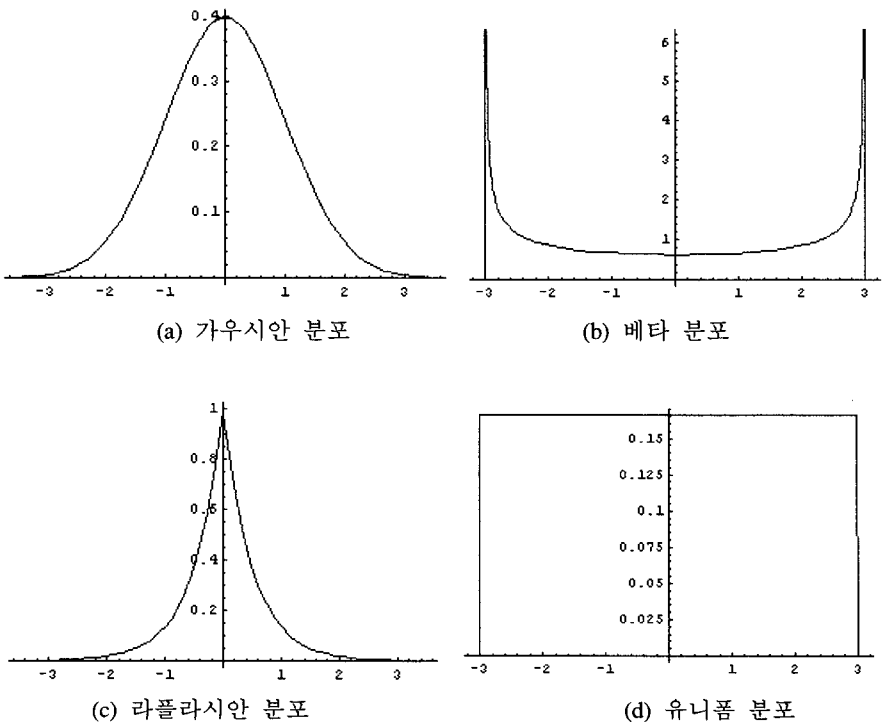
6) 입력 x를 $C_x(x) = C_{ref}(y)$를 만족하는 값 y로 변환한다.

이와 같은 방법으로 MFCC 13차에 HEN을 적용하는 과정을 블록도로 나타내면

<그림 2>와 같다. 여기서, 분포의 등분을 HEQ와 달리 1000 단계로 한 것은 분포를 변환할 때 양자화 잡음의 영향을 최소화하기 위한 것으로, 실제 본 논문에서 실험에 사용된 단문의 경우 약 2.5초(약 250 프레임) 정도로 250등분 한 것과 1000 등분한 것이 화자 확인 성능에 큰 차이는 없다.

3.3. 제안 기준 분포

HEN 기법은 음성 특징 분포의 대비를 조절하여 스코어 분포를 최적화하기 위한 것으로 HEQ 기법의 수정을 필요로 한다. HEQ 기법에서는 기준 분포를 평균=0 이고 분산=1인 가우시안을 제안했다[3]. 본 논문의 HEN 기법은 음성 특징 분포의 대비를 조절하기 위하여 기준 분포를 평균=0, 분산=3인 유니폼(Uniform) 분포를 제안한다. 이 분포는 최적의 성능을 낼 수 있는 기준 분포는 아니지만, 음성 특징 분포의 대비 조절에 의한 화자 확인의 성능 차이를 확인할 수 있게 한다. 그리고 음성 특징 분포의 대비를 인식 성능에 대해 비교 평가하기 위해 $\alpha = \beta = 0.5$인 베타(Beta) 분포, 평균=0, 분산=0.5인 라플라시안(Laplacian) 분포의 대조군 기준 분포를 적용하였다. <그림 3>은 실험에서 사용한 기준 분포들을 보여준다.

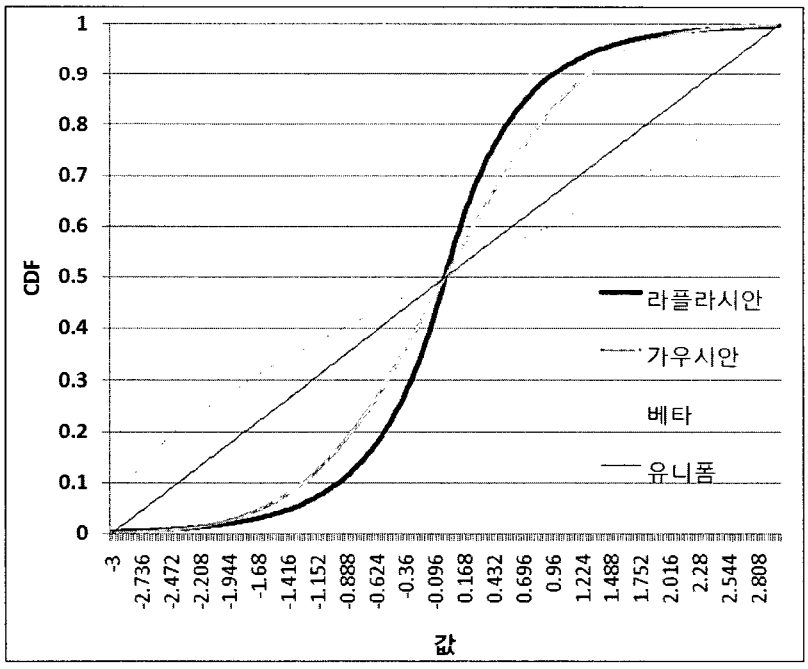


(a) 가우시안 분포 (b) 베타 분포

(c) 라플라시안 분포 (d) 유니폼 분포

<그림 3> 여러 가지 기준 분포

히스토그램 개선 기법은 CDF 매칭에 의한 분포 변환 기법을 사용하는 것으로, MFCC의 분포와 기준 분포의 CDF를 비교하면 MFCC 값들의 대비 변화를 알 수 있다. <그림 4>에서 HEN 시스템의 기준 분포인 유니폼 분포, 베타 분포, 라플라시안 분포와 가우시안 분포의 CDF를 비교하였다. CDF의 기울기는 분포 전체에서 해당 지점에서의 확률밀도를 보여주는 것으로, 기울기가 크다면 높은 확률분포를, 기울기가 작다면 낮은 확률분포를 나타내는 것이다. 즉, CDF에서 기울기가 가우시안보다 커지면 확률분포에서 빈도가 높다는 것을 의미하므로 상대적으로 대비가 낮아지는 것이고, 기울기가 작아지면 확률분포에서 빈도가 낮다는 것을 말하므로 상대적으로 대비가 높아지는 것이다. 유니폼 분포의 경우, 분포의 평균과 그에 가까운 범위에서는 가우시안 분포의 CDF 기울기보다 낮은 기울기로 MFCC 값들의 대비를 고대비화하고, 평균에서 먼 범위에서는 가우시안 분포의 CDF 기울기보다 높은 기울기로 MFCC 값들의 대비는 저대비화한다. 베타 분포는 유니폼 분포 보다 평균 근처에서 더 큰 고대비화와, 먼 범위의 값들에서 저대비화를 한다. 그리고 라플라시안 분포는 가우시안 분포의 평균과 그에 가까운 MFCC 값들에서 상대적으로 저대비화하고, 먼 범위에 있는 값들에서 상대적으로 고대비화한다.



<그림 4> 여러 가지 기준 분포의 CDF

4. 실험 방법 및 결과

4.1 실험 방법

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 PCM, Intel-format으로 저장되었으며 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.

현재 대부분의 문장 독립형 화자 확인 시스템에서는 GMM-UBM 시스템[6]을 사용한다. 본 논문에서도 GMM-UBM 시스템을 사용하였다. 이 시스템에서는 목적 화자 음성으로 훈련되는 화자별 GMM(Gaussian Mixture Model) 이외에 UBM(Universal Background Model)이라는 GMM을 하나 더 필요로 한다. UBM은 음성 특징의 화자 독립 분포를 표현하기 위해 훈련되는 하나의 큰 GMM이다. 그리고 화자인식 시스템에서 화자모델을 만들 때 대부분 훈련 음성 자료를 충분히 얻기 어려운 경우가 많다. 적은 훈련 음성 자료를 효과적으로 이용하는 방법으로 화자적응이 있다. 이는 UBM로 부터 화자적응을 통하여 화자모델을 훈련하여 각각의 화자모델을 생성하는 것이다. 본 논문에서는 화자적응 방법 중 MAP(Maximum A Posteriori) 화자적응 기법[6]을 사용하였다.

GMM은 EM(Expectation-Maximization) 알고리즘으로 2, 4, 8, 16, 32, 64, 128, 256, 512 순으로 mixture를 증가시킨 ML(Maximum Likelihood) 모델이다. 실험 대상 화자모델의 MAP 화자적응 시 각 화자의 0주차(주차 화자의 음성 중 기간별 첫 녹음시점) 6개 단문의 60초로 제한된 5회 발성음성을 사용하였다.

UBM 작성은 월차 화자 음성 중 기간별 첫 녹음 시점인 0개월차 월차 화자 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.

본 논문에서 화자 확인 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 DB는 훈련과 독립적인(훈련에 사용되지 않은 단문) 3개 단문의 5회 음성을 사용하여 각 화자당 15 단문으로 하였다.

화자 확인 시스템은 GMM-UBM로 화자와 사칭자의 비율을 1:10으로 하였다. <표 1>은 화자 확인에 사용한 화자군의 분류와 음성시료의 개수를 보여주고 있다.

<표 1> 실험에 사용한 화자 및 음성시료의 구성

	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

잡음 보상에 대한 실험을 위해 잡음 환경의 음성을 제작하였다. 화자 확인 인식 시 사용된 테스트 음성에 잡음을 추가하였다. 추가된 잡음은 Noisex-92[7]를 근거로 백색 잡음(white), 차량 잡음(volvo), 군중 잡음(babble)에 5 dB, 10 dB, 15 dB, 20 dB의 SNR을 적용하였다. 잡음 환경의 실험에서 훈련 모델은 SNR이 25 dB 이상의 조용한 사무실 환경에서 수집한 음성 DB로 작성하였다.

사용된 모든 음성은 음성구간 검출을 위해 앞, 뒤의 묵음(전 200ms, 후 200ms)을 제거한 단문의 음성을, 프레임 길이는 25ms, 프레임 주기는 10ms이며 Hamming window을 사용하여 MFCC 13차를 추출하였다. 특징 추출은 HTK ver 3.3[8]을 이용하였다.

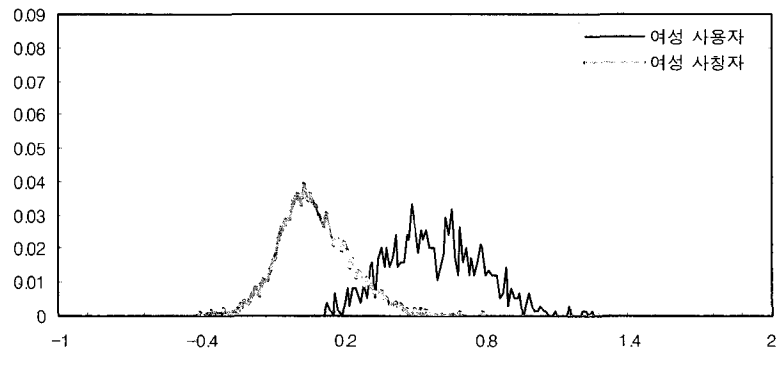
4.2 실험 결과

조용한 훈련 및 테스트 환경의 여성 화자 확인 실험에서, 기본 시스템, HEQ 가우시안, HEN 유니폼, 베타, 라플라시안 시스템의 사용자 및 사칭자의 스코어 분포를 <그림 5>에서 <그림 9>에 각각 나타냈다. 사용자와 사칭자의 스코어 분포에서 두 분포가 겹치는 부분은 에러영역으로 넓이의 1/2은 곧 EER(Equal Error Rate)을 뜻한다. 따라서 두 분포의 평균 차이와 분산의 합은 곧 화자 확인 성능의 지표가 된다. HEN 유니폼과 베타를 적용하였을 때 MFCC의 평균과 그에 가까운 값들의 대비가 커져서 스코어 분포의 평균 차이와 분산의 합이 증가하고, HEN 라플라시안 시스템의 경우 대비가 줄어들면서 스코어 분포의 평균 차이와 분산의 합이 줄어듦을 알 수 있다. HEN을 적용하면서 MFCC의 평균과 그에 가까운 값들의 대비와 스코어 분포의 평균 차이, 분산의 합이 비례하는 성향을 보였다.

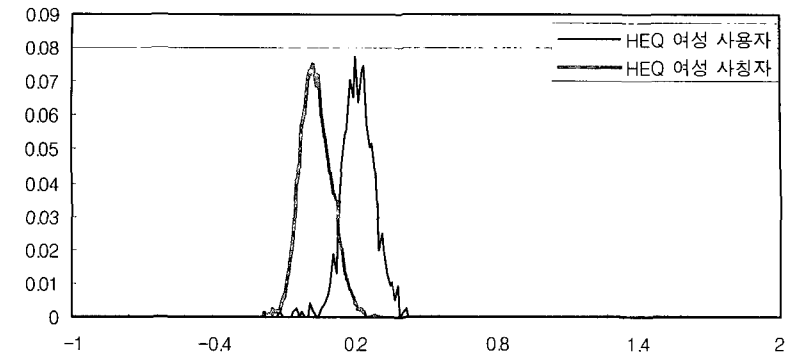
<그림 5>에서 <그림 9>는 스코어 분포의 형태만 보였으며, 스코어 분포의 평균 차이와 분산의 합은 매우 작은 값이므로, HEN을 적용 후 사용자와 사칭자의 스코어 분포 간의 차이가 잘 드러나지 않았다. <표 2>에서는 HEN 적용 후의 시스

템에서 사용자 및 사칭자 스코어 분포의 평균 차이와 분산의 합을 HEQ 시스템과 비교하였다. 이 표는 각 시스템의 스코어 분포의 차이를 명확히 보여주고 있다. HEN 유니폼 시스템은 HEQ 가우시안 시스템보다 남자 26.9%, 여자 30.0% 높은 평균 차이와, 남자 31.7%, 여자 41.5% 높은 분산의 합을 보였다. HEN 베타 시스템은 HEQ 가우시안 시스템과 비교하여 남자 18.8%, 여자 22.6% 높은 평균 차이와, 남자 15.1%, 여자 20.6% 높은 분산의 합을 보였다. 결론적으로 HEN 적용 시 기준 분포에 따라서 사용자와 사칭자의 스코어 분포가 달라졌음을 알 수 있다. 즉, MFCC 값의 대비 변환이 스코어 분포를 변환하여 화자 확인 시스템의 성능에 직접적인 영향을 주는 요인임을 확인할 수 있다. 여기서 HEN 라플라시안 시스템의 경우 MFCC 분포의 평균 값 근처의 대비를 개악한 대조군의 실험으로 성능을 저하시켰다.

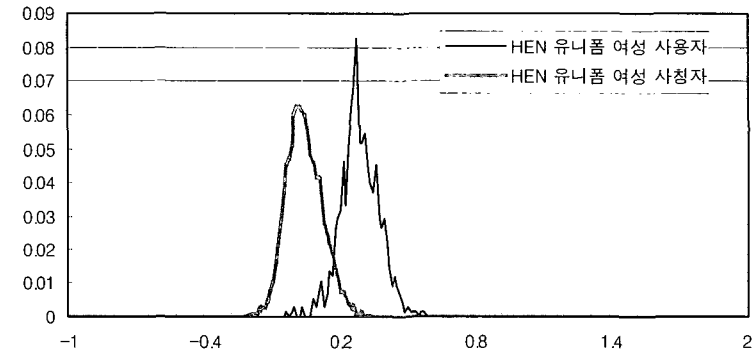
<그림 4>에서 설명한 대로 각 제안 분포의 CDF 기울기에서 알 수 있듯이, 분포의 MFCC 평균과 근처 값들의 대비를 비교하면, 'HEN 베타 > HEN 유니폼 > HEQ 가우시안 > HEN 라플라시안' 순으로 대비의 크기를 나타낼 수 있다. 그리고 MFCC의 평균에서 거리가 먼 값의 MFCC 대비는 이와 반대로 HEN 라플라시안이 가장 고대비, HEN 베타가 가장 저대비이다. 그러나 <표 2>에서 보인 것과 같이 사용자 및 사칭자 스코어 분포의 평균 차이와 분산의 합으로 비교하면, 'HEN 유니폼 > HEN 베타 > HEQ 가우시안 > HEN 라플라시안' 순으로 평균 차이와 분산의 합이 비례해 변화함을 알 수 있다. 따라서 MFCC 분포의 전체 영역에서의 대비는 'HEN 유니폼 > HEN 베타 > HEQ 가우시안 > HEN 라플라시안' 순이라고 말할 수 있다. 화자 확인 시스템은 사용자와 사칭자의 스코어 분포의 평균 차이가 클수록, 각 분포의 분산이 작을수록 성능이 좋아진다. MFCC의 대비를 변환할 경우 사용자와 사칭자 스코어 분포의 평균 차이와 분산의 합이 비례해서 변화하기 때문에 MFCC를 적절한 범위를 넘어선 고대비화는 시스템의 성능을 저하시킬 수 있다.



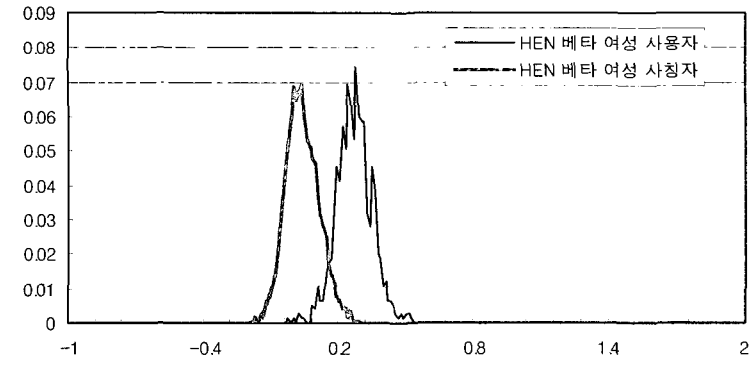
<그림 5> 기본 시스템 여성 사용자, 사칭자 스코어 분포



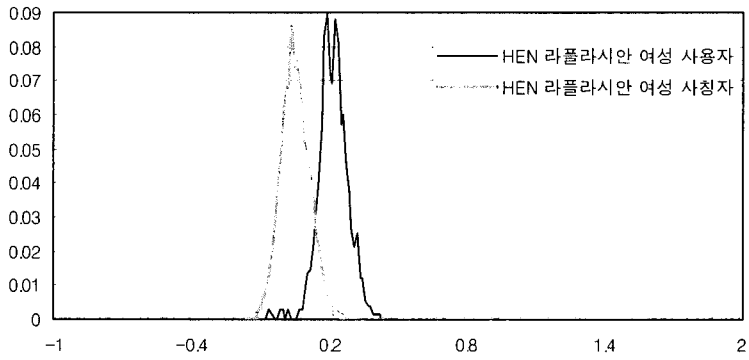
<그림 6> HEQ 여성 사용자, 사칭자 스코어 분포



<그림 7> HEN 유니폼 여성 사용자, 사칭자 스코어 분포



<그림 8> HEN 베타 여성 사용자, 사칭자 스코어 분포



<그림 9> HEN 라플라시안 여성 사용자, 사칭자 스코어 분포

<표 2> 각 시스템의 사용자, 사칭자 스코어 분포 비교

		남자		여자	
		평균	분산	평균	분산
기본 시스템	사용자	0.6231	0.0676	0.5867	0.0386
	사칭자	0.0234	0.0267	0.0823	0.0228
	평균 차이	0.5997		0.5044	
	분산의 합		0.0943		0.0614
HEN 라플라시안 시스템 (A)	사용자	0.2569	0.0043	0.2220	0.0040
	사칭자	0.0504	0.0073	0.0560	0.0038
	평균 차이	0.2065		0.1660	
	분산의 합		0.0115		0.0078
	(A-B)/B	**-11.2%**	**-13.0%**	**-11.9%**	**-17.9%**
HEQ 가우시안 시스템 (B)	사용자	0.2690	0.0053	0.2353	0.0049
	사칭자	0.0364	0.0079	0.0468	0.0046
	평균 차이	0.2326		0.1885	
	분산의 합		0.0132		0.0095
HEN 유니폼 시스템 (C)	사용자	0.3228	0.0076	0.2901	0.0072
	사칭자	0.0278	0.0099	0.0451	0.0063
	평균 차이	0.2951		0.2450	
	분산의 합		0.0174		0.0135
	(C-B)/B	**26.9%**	**31.7%**	**30.0%**	**41.5%**
HEN 베타 시스템 (D)	사용자	0.2989	0.0070	0.2707	0.0062
	사칭자	0.0227	0.0082	0.0396	0.0053
	평균 차이	0.2762		0.231	
	분산의 합		0.0152		0.0115
	(D-B)/B	**18.8%**	**15.1%**	**22.6%**	**20.6%**

주) HEQ 가우시안 시스템을 기준으로 시스템 간 평균 차이와 분산의 합에 대한 상대적인 크기의 비율을 구하여 시스템 간 통계적으로 의미있는 차이가 있음을 보였다. (-) 부호는 HEQ 가우시안 시스템에 비해 평균 차이와 분산의 합이 감소했음을 의미한다.

HEN을 제안하면서 화자 확인 시스템의 스코어 분포가 음성 특징 대비의 종속 변수라고 가정하였다. 이를 음성 특징의 대비 즉, MFCC의 대비를 변환하여 화자 확인 시스템의 인식 성능이 향상됨을 실험으로 보인다.

<표 3>에 조용한 훈련 및 테스트 환경에서 각 시스템의 성능을 보여주는 EER을 정리하였다. 실험 결과의 EER을 비교하면, HEN 베타 시스템은 기본 시스템의 성능 보다 남자 6.39%, 여자 13.92%의 에러를 감소시켰다. 그리고 HEQ 가우시안 시스템 보다 남자의 경우 22.68%, 여자의 경우 15.72%의 에러를 감소시켰다. HEN 유니폼 시스템은 기본 시스템의 성능 보다 남자 2.18%, 여자 12.31%의 에러를 감소시켰다. 그리고 HEQ 가우시안 시스템 보다 남자의 경우 19.2%, 여자의 경우 14.15%의 에러를 감소시켰다.

<표 3> 각 시스템의 화자 확인 EER 비교

	남자	여자
기본 시스템	6.41%	7.47%
HEQ 가우시안 시스템	7.76%	7.63%
HEN 라플라시안 시스템	9.08%	8.40%
HEN 유니폼 시스템	6.27%	6.55%
HEN 베타 시스템	6.00%	6.43%

잡음 환경에 대한 실험의 화자 확인 결과는 <표 4>에서 <표 7>과 같다. HEN 베타 시스템은 HEQ 가우시안 시스템 대비 백색 잡음 환경 5 dB~20 dB에서 평균적으로 남자 약 3.0%의 에러율이 증가하였고, 여자 약 0.72%의 에러율을 개선하였다. 차량 잡음 환경 5 dB~20 dB에서 평균적으로 남자는 약 10.32%의 에러율을 개선한 반면, 여자 약 4.45%의 에러율을 증가시켜 성능을 저하시켰다. 군중 잡음 환경 5 dB~20 dB에서 평균적으로 남자 약 7.33%, 여자 약 1.68%의 에러율을 감소시켜 성능을 개선하였다. 그리고 <표 7>의 세 잡음의 평균 EER을 보면 남자는 약 1.8%의 에러율을 개선한 반면, 여자는 거의 같은 성능을 보여 주었다.

HEN 유니폼 시스템은 HEQ 가우시안 시스템 대비 백색 잡음 환경에서 평균적으로 남자 약 4.46%, 여자 약 4.2%의 에러율을 개선하였다. 차량 잡음 환경에서 평균적으로 남자는 약 12.48%의 에러율을 개선한 반면, 여자 약 6.67%의 에러율을 증가시켜 성능을 저하시켰다. 군중 잡음 환경에서 평균적으로 남자 약 8.78%, 여자 약 0.84%의 에러율을 감소시켜 성능을 개선하였다. 그리고 <표 7>의 세 잡음의 평균 EER을 보면 남자 약 7.5%, 여자 1.6%의 에러율을 개선하였다.

<표 4> 백색 잡음 환경 화자 확인 EER 비교

	남자(%)					여자(%)				
	5dB	10dB	15dB	20dB	평균	5dB	10dB	15dB	20dB	평균
기본 시스템	45.20	41.60	37.49	32.75	39.26	46.93	43.33	37.60	30.53	39.60
HEQ 가우시안 시스템	34.27	28.00	22.00	16.96	25.31	38.40	34.93	30.40	23.87	31.90
HEN 라플라시안 시스템	37.36	32.00	25.35	18.96	28.42	38.48	34.60	29.73	24.13	31.74
HEN 유니폼 시스템	32.57	26.53	21.47	16.13	24.18	38.53	33.87	28.13	21.73	30.57
HEN 베타 시스템	34.67	28.53	23.07	18.00	26.07	39.07	34.93	29.47	23.20	31.67

<표 5> 차량 잡음 환경 화자 확인 EER 비교

	남자(%)					여자(%)				
	5dB	10dB	15dB	20dB	평균	5dB	10dB	15dB	20dB	평균
기본 시스템	18.53	14.67	11.60	10.00	13.70	31.07	26.93	22.73	19.20	24.98
HEQ 가우시안 시스템	8.53	8.05	7.47	7.33	7.85	10.80	9.20	8.13	7.87	9.00
HEN 라플라시안 시스템	9.87	9.07	8.55	8.27	8.94	12.93	11.00	10.00	8.93	10.72
HEN 유니폼 시스템	8.00	6.67	6.55	6.27	6.87	11.33	10.07	9.07	8.00	9.62
HEN 베타 시스템	8.13	7.13	6.57	6.31	7.04	11.73	9.33	8.80	7.73	9.40

<표 6> 군중 잡음 환경 화자 확인 EER 비교

	남자(%)					여자(%)				
	5dB	10dB	15dB	20dB	평균	5dB	10dB	15dB	20dB	평균
기본 시스템	28.67	19.92	14.0	11.20	18.45	34.29	28.27	22.53	18.40	25.87
HEQ 가우시안 시스템	12.40	10.07	8.67	7.60	9.68	16.0	12.53	10.28	8.80	11.90
HEN 라플라시안 시스템	13.53	10.87	9.47	8.53	10.60	18.93	15.24	12.40	10.67	14.31
HEN 유니폼 시스템	11.33	9.20	7.87	6.93	8.83	16.53	12.53	9.73	8.41	11.80
HEN 베타 시스템	12.27	9.07	7.89	6.67	8.97	16.13	12.40	10.00	8.27	11.70

<표 7> 백색, 차량, 군중 잡음 화자 확인 평균 EER 비교

	남자(%)					여자(%)				
	5dB	10dB	15dB	20dB	평균	5dB	10dB	15dB	20dB	평균
기본 시스템	30.80	25.40	21.03	17.98	23.80	37.43	32.84	27.62	22.71	30.15
HEQ 가우시안 시스템	18.40	15.37	12.71	10.63	14.28	21.73	18.89	16.27	13.51	17.60
HEN 라플라시안 시스템	20.25	17.31	14.46	11.92	15.99	23.45	20.28	17.38	14.58	18.92
HEN 유니폼 시스템	17.30	14.13	11.96	9.78	13.29	22.13	18.82	15.64	12.71	17.33
HEN 베타 시스템	18.36	14.91	12.51	10.33	14.03	22.31	18.89	16.09	13.07	17.59

조용한 환경에서의 화자 확인 실험결과와 가산 잡음 환경에서의 시스템 성능의 추이가 다름을 <표 3>에서 <표 6>의 결과로 부터 확인할 수 있다. 조용한 환경에서는 HEN 베타 시스템이 가장 좋은 성능을 보인 반면, 가산 잡음 환경에서는 평균적으로 HEN 유니폼 시스템이 가장 좋은 성능을 보였다. 입력 음성의 가산 잡음은 MFCC 분포의 분산을 변형시키는 것으로 알려져 있다[3]. HEN 시스템은 분포의 비선형 변환이므로, 이러한 원인으로 시스템의 성능 왜곡이 있을 수 있다. <표 4>에서 <표 6>의 실험결과는 HEN 유니폼 시스템이 HEN 베타 시스템 보다 가산잡음에 의한 분포의 왜곡에 유연한 시스템임을 보여준다. 결론적으로 조용한 환경과 잡음 환경 모든 측면에서 HEN 유니폼 시스템이 HEN을 적용할 때 가장 좋은 기준 분포라고 말할 수 있다.

지금까지의 실험 결과는 음성의 채널 변화가 없는 조용한 환경 및 잡음 환경의 화자 확인 시스템에서 얻어진 것이다. HEN은 HEQ의 장점인 채널 독립적인 특징 분포로 변환하는 방식은 같지만, 채널 변화가 있는 화자 확인 시스템에서의 성능을 평가하지는 못했다. 따라서 이 논문에서의 실험결과와 동등 비교 평가를 할 수 없다. 그러나 HEQ의 방식과 동일하게 특징 분포를 변환하였다는 점에서 채널 보상의 측면에서도 좋은 결과가 나올 것으로 생각된다. 이 추측을 확인할 수 있는 실험이 뒤따라야 하겠다.

5. 결 론

본 논문은 화자 확인에서 음성 특징의 파라미터인 MFCC의 대비라는 용어를 정의하였다. 여기서 대비라는 말은 화자 확인에서 사용하는 음성 특징의 화자 종

속적인 정보(화자 간의 차이를 보여주는 정보) 간 크기의 차이를 뜻한다. 본 논문에서는 화자인식 성능 향상을 목적으로 MFCC의 대비를 조절하기 위해 HEN 기법을 제안하였다. 또한, 이 대비가 화자 확인의 성능지표인 사용자와 사칭자 스코어 분포의 종속변수임을 실험적으로 보였다.

조용한 훈련 및 테스트 환경에서 HEN 유니폼 시스템은 기본 화자 확인 시스템과 HEQ 가우시안 시스템 보다 인식 성능을 개선하였다. 그리고 백색, 차량, 군중 잡음 환경에서, 차량잡음 환경의 여자 화자 확인을 제외한 모든 실험에서 HEQ 가우시안 시스템 보다 HEN 유니폼 시스템이 더 나은 성능을 보여 환경의 변화에도 더 강인한 시스템임을 보였다.

본 논문에서 화자 확인의 성능이 MFCC의 대비에 종속변수임을 보였다. 그러나 본 논문은 MFCC의 대비에 대해 정형화된 척도를 정의하지 못하였다. 그리고 MFCC의 대비와 스코어 분포의 평균 차이, 분산에 대한 비례관계의 수학적 정의를 밝히지 못했다. 따라서 이후의 연구에서는 MFCC의 대비에 대한 정형화된 척도와, MFCC의 대비와 스코어 분포간의 수학적 정의를 도출하여, 화자 확인의 성능을 최적화 할 수 있는 기준 분포에 대한 연구를 진행할 예정이다.

참 고 문 헌

[1] 유하진, "화자인식 기술 및 국내외시장 동향", *대한음성학회 2004 봄 학술대회 발표논문집*, pp. 91-97, 2004.

[2] R. C. Gonzalez, P. Wintz, *Digital Image Processing*, Addison-Wesley, 1987.

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

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

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

[6] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, Vol. 10, Nos. 1-3, pp. 19-41, 2000.

[7] Noisex-92, http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

[8] S. Young, *The HTK Book*, Cambridge University Engineering Department, 2001.

접수일자: 2007년 8월 15일
게재결정: 2007년 9월 20일

▶ 최재길(Jae-Kil Choi)
주소: 300-716 대전광역시 동구 용운동 96-3 대전대학교
소속: 대전대학교 정보통신공학과 BMW 연구실
전화: 042) 280-2567
E-mail: u2u2u2u2u2@nate.com

▶ 권철홍(Chul-Hong Kwon) : 교신저자
주소: 300-716 대전광역시 동구 용운동 96-3 대전대학교
소속: 대전대학교 정보통신공학과 BMW 연구실
전화: 042) 280-2555
E-mail: chkwon@dju.ac.kr