말소리와 음성과학 제4권 제2호 (2012)
pp. 79~86

화자 식별에서의 배경화자데이터를 이용한 히스토그램 등화 기법

Histogram Equalization Using Background Speakers' Utterances for Speaker Identification

김 명 재[1] · 양 일 호[2] · 소 병 민[3] · 김 민 석[4] · 유 하 진[5]
Kim, Myung-Jae · Yang, IL-Ho · So, Byung-Min · Kim, Min-Seok · Yu, Ha-Jin

ABSTRACT

In this paper, we propose a novel approach to improve histogram equalization for speaker identification. Our method collects all speech features of UBM training data to make a reference distribution. The ranks of the feature vectors are calculated in the sorted list of the collection of the UBM training data and the test data. We use the ranks to perform order-based histogram equalization. The proposed method improves the accuracy of the speaker recognition system with short utterances. We use four kinds of speech databases to evaluate the proposed speaker recognition system and compare the system with cepstral mean normalization (CMN), mean and variance normalization (MVN), and histogram equalization (HEQ). Our system reduced the relative error rate by 33.3% from the baseline system.

Keywords: speaker recognition, speaker identification, histogram equalization

1. 서론

화자 인식 시스템은 학습 환경과 인식 환경이 같을 때 좋은 성능을 보여준다. 하지만 실제 상황에서는 학습 환경과 인식 환경이 동일하지 않을 수 있으므로 인식 성능 저하의 요인이 된다. 이러한 학습 환경과 인식 환경의 불일치를 극복하기 위해 사용하는 채널 보상 방법으로 캡스트럼 평균 정규화 (cepstral mean normalization, CMN)[1], 평균-분산 정규화 (mean and variance normalization, MVN)[2] 등이 있다. 이들 채널 보상 방법은 음성 특징에 선형적으로 작용하는 채널 잡음을 효과적으로 제거하기 위해 간단한 선형 변환으로 화자 인식 성능을 개선한다. 그러나 비주기적으로 발생하는 부가 잡음은 음성 특징에 비선형적으로 작용한다[3]. 이러한 비선형적인 특성을 선형으로 근사하여 채널 보상을 하고자 하는 연구들이 진행되었다. 이러한 방법들로 vector Taylor series (VTS)[4]와 static linear approximation (SLA)[5] 등이 제안되었다. 이러한 방법들과 다른 접근 방법인 특징 정규화 방법으로 히스토그램 등화 기법 (histogram equalization, HEQ)[6]이 제안되었다. 히스토그램 등화 기법은 이전에 이미지 프로세싱 분야에서 디지털 이미지의 밝기와 대비를 조절하기 위해 사용되었다[6]. 이후 음성인식 분야에서 채널 보상 방법으로 성공적으로 적용되었다[10][11]. 또한, 히스토그램 등화 기법은 화자 인식 분야에서도 효과적으로 적용되었고, 이를 응용한 feature warping[7], modified segment HEQ[9] 등의 방법이 제안되었다. 이러한 방법들은 부가 잡음에 의해 변형된 특징의 분포를 기준 분포로 비선형 변환한다. 하지만 이러한 방법들은 충분한 길이의 음성 발화가 필요한 단점이 있으며, 음성 발화가 짧은 경우 캡스트럼 평균 정규화, 평균-분산 정규화 등과 같은 채널 보상 방법보다 성능이 떨어지는 문제가 발생할 수 있다. 본 논문은 이러한 문제점을 해결하기 위하여, UBM (universal background model) 학습 데이터의 분포를 이용한다. UBM의 데이터는 다양한 음소와 음원을 포함하는 데이터이기 때문에 세밀한 히스토그램 추정이 가능하다. UBM에서 추정한 분포를 이용하여 히스토그램 등화 기법을 수행하면 짧은 음성 발

1) 서울시립대학교, mj@uos.ac.kr
2) 서울시립대학교, heisco@hanmail.net
3) 서울시립대학교, sbm1210@naver.com
4) LG 전자기술원, minseok3.kim@lge.com
5) 서울시립대학교, hjyu@uos.ac.kr, 교신저자

접수일자: 2012년 3월 13일
수정일자: 2012년 5월 22일
게재결정: 2012년 6월 19일

화의 화자 식별 성능을 높일 수 있다.

본 논문은 2장에서 기존 채널 보상 방법을 설명하고, 3장에서는 제안한 히스토그램 등화 기법에 대해 설명한다. 4장에서는 실험 설계 및 실험 결과를 보이고, 5장에서 결론을 맺는다.

2. 기존 채널 보상 방법

2.1 캡스트럼 평균 정규화 (CMN)[1]

음성 신호가 선형 채널 필터를 통과할 때, 선형 채널 필터는 음성 신호에 컨벌루션 잡음을 발생시킨다. 컨벌루션 잡음이 추가된 캡스트럼 특징 벡터는 식 (1)과 같이 나타낼 수 있다.

$$\vec{x} = \vec{s} + \vec{h} \tag{1}$$

$\vec{x}$는 관찰된 캡스트럼 특징 벡터, $\vec{s}$는 음성 신호의 캡스트럼 특징 벡터, $\vec{h}$는 선형 채널 필터의 캡스트럼 특징 벡터이다. 이때, 선형 채널 필터에 의해 발생한 컨벌루션 잡음은 평균을 제거하여 효과적으로 감소시킬 수 있다. 특징 벡터들의 평균 벡터는

$$\vec{\mu} = \frac{1}{T}\sum_{t=1}^{T}\vec{x_t} \tag{2}$$

으로 구한다. 이때 T는 관찰된 캡스트럼 특징 벡터의 전체 길이이다. 채널 보상한 캡스트럼 특징 벡터 $\hat{\vec{x}}$은 다음과 같다.

$$\hat{\vec{x}} = \vec{x} - \vec{\mu} \tag{3}$$

2.2 평균-분산 정규화 (MVN)[2]

선형 채널 필터에 의한 컨벌루션 잡음은 캡스트럼 평균 정규화를 이용하여 효과적으로 제거할 수 있다. 그러나 추가 잡음에 의한 왜곡은 평균뿐만 아니라 분산도 변화시킨다. 따라서 변화된 분산을 보상하기 위하여 분산을 정규화 해야한다. 분산 정규화한 캡스트럼 특징 벡터 $\hat{\vec{x}}$은 다음과 같다.

$$\hat{\vec{x}} = \frac{\vec{x} - \vec{\mu}}{\vec{\sigma}} \tag{4}$$

평균 벡터 $\vec{\mu}$는 식 (2)를 통하여 구할 수 있고, 분산 벡터 $\vec{\sigma^2}$은 다음과 같다.

$$\vec{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(\vec{x_t} - \vec{\mu})^2 \tag{5}$$

2.3 히스토그램 등화 기법 (HEQ)

히스토그램 등화 기법은 디지털 이미지 프로세싱 분야에서 널리 사용되었다[6]. 그 후, 음성 인식과 화자 인식분야에 성공적으로 적용되었다[8][9][10]. 히스토그램 등화 기법의 목적은 학습 데이터와 인식 데이터의 분포를 미리 정의한 기준 분포로 맞추는 것이다. 캡스트럼 평균 정규화 및 평균-분산 정규화는 단지 평균과 분산을 이용하여 특징 벡터를 선형적으로 변환하지만, 히스토그램 등화 기법은 특징 벡터의 누적 분포를 기준 분포의 누적 분포로 비선형적 변환한다.

히스토그램 등화 기법의 수식은 다음과 같다[8].

$$P_{ref}(y)dy = p_x(x)dx \tag{6}$$

P_{ref}는 변환의 기준이 되는 확률 분포이다. 본 논문에서는 평균이 0이고, 분산이 1인 표준 정규 분포를 기준 확률 분포로 사용하며 식은 다음과 같다.

$$P_{ref}(x) = \frac{1}{\sqrt{2\pi}} exp(\frac{-x^2}{2}) \tag{7}$$

x는 1차원 변수이며, 특징 벡터의 한 차원으로 볼 수 있다. $p_x(x)$는 x에 대한 확률 밀도이다. 이때, $y = T(x)$라 하자. $T(x)$는 일가함수(single-valued function)이다. 식 (6)을 $P_{ref}(y)$에 대해 정리하면 다음과 같다.

$$P_{ref}(y) = p_x(x)\frac{dx}{dy} = p_x(G(y))\frac{dG(y)}{dy} \tag{8}$$

$G(y)$는 $T(x)$의 역변환이다.

$p_x(x)$와 $P_{ref}(y)$의 누적 분포 관계를 식 (8)을 이용하여 나타내면 다음과 같다.

$$\begin{aligned} C_x(x) &= \int_{-\infty}^{x} p_x(x')dx' \\ &= \int_{-\infty}^{T(x)} p_x(G(y'))\frac{dG(y)}{dy}dy' \\ &= \int_{-\infty}^{y} P_{ref}(y')dy' \\ &= C_{ref}(y) \\ &= C_{ref}(T(x)) \end{aligned} \tag{9}$$

C_x는 x가 속한 특징 벡터의 확률 분포 함수이며, C_{ref}는 기준 확률 밀도 함수의 누적 분포 함수이다. 식 (9)를 $T(x)$에 대해 다시 정리하면 다음과 같다.

$$T(x) = C_{ref}^{-1}(C_x(x)) \tag{10}$$

C_{ref}^{-1}는 기준 누적 분포 함수의 역함수이다.

표준 정규 분포를 이용한 값의 변환은 누적 분포 함수 (cumulative distribution function, CDF) 표를 이용하면 쉽게 구할 수 있다. <그림 1>은 히스토그램 등화 기법의 누적 분포 변환을 보여준다[9].

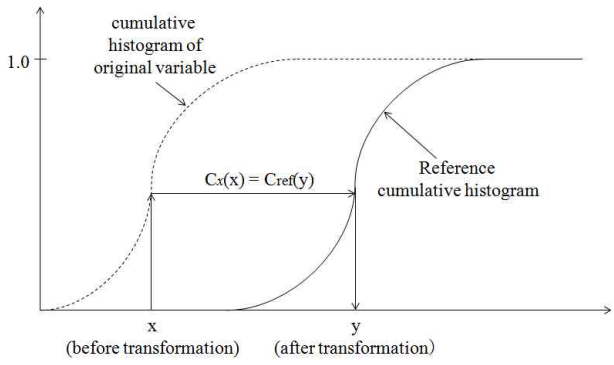

그림 1. 히스토그램 등화 기법을 이용한 누적 분포 변환
Figure 1. The transformation of cumulative distribution using histogram equalization

2.4 히스토그램 등화 기법의 구현

히스토그램 등화 기법의 구현 방법은 순서 기반과 누적 히스토그램 추정 방식이 있다. 순서 기반의 히스토그램 등화 기법의 누적 분포 함수 추정과정은 다음과 같다.

먼저 음성 신호에서 추출한 특정 차수의 특징계수 열 S를 다음과 같이 정의한다.

$$S = \{s_1, s_2, \cdots, s_n, \cdots, s_N\} \tag{11}$$

s_n은 n시점의 특징계수를 의미한다. 이렇게 정의한 S를 계수의 크기에 따라 오름차순으로 정렬하고, 정렬한 순서를 이용하여 S의 서열을 정의한다.

$$s_{D(1)} \le s_{D(2)} \le \cdots \le s_{D(n)} \le \cdots \le s_{D(N)} \tag{12}$$

$D(n)$은 n번째 서열의 인덱스를 의미한다. 오름차순으로 정렬한 S의 누적 분포 Φ_n은 다음과 같이 근사한다.

$$\Phi_n = \frac{R_s(s_n) - 0.5}{N} \tag{13}$$

R_s는 특징계수 S에 대한 s_n의 오름차순 서열을 의미한다. 이렇게 추정한 누적 분포 함수를 이용하여 특징계수의 변환은 다음과 같이 구한다[9].

$$T(s_n) = C_{ref}^{-1}(\Phi_n) \tag{14}$$

누적 히스토그램 추정 방식 기반 히스토그램 등화 기법의 누적 분포 함수 추정 과정은 다음과 같다[9].

1. 음성 신호에서 추출한 특정 차수의 특징계수 열 S의 특징계수 중에서 최대값 $s_{\max}$와 최소값 $s_{\min}$을 결정한다.

2. $[s_{\min}, s_{\max}]$의 범위를 동일한 크기로 M등분한다. 등분한 범위를 B라 하고, B의 원소 범위는 겹치지 않는다. B는 $s_{\min} = b_1 < b_2 < \cdots < b_{M+1} = s_{\max}$ 이며, B_i는 $[b_i, b_{i+1})$의 범위를 갖는다. B_i를 빈 (bin)이라 한다.

3. 이렇게 등분한 B를 가지고 히스토그램을 측정한다. 히스토그램의 측정은 각각 빈에 속하는 특징의 개수를 세는 것으로 한다.

4. 측정한 히스토그램을 정규화하기 위하여 다음 식을 적용한다.

$$p_x(x \in B_i) = \frac{n_i}{N_x} \tag{15}$$

n_i는 B_i의 범위에 속한 특징의 개수이며, N_x는 추출한 특징계수의 전체 개수이다.

5. 정규화한 히스토그램으로 누적 히스토그램을 계산한다.

$$C_x(x : x \in B_i) = \sum_{j=1}^{M} \frac{n_j}{N_x} \tag{16}$$

6. 기준 분포와 계산한 누적 히스토그램을 이용하여, $C_x(x) = C_{ref}(y)$를 만족하는 특징계수 x를 y로 변환한다.

3. 제안한 히스토그램 등화 기법

본 연구에서는 UBM (universal background model) 학습 데이터를 이용한 히스토그램 등화 기법 개선 방법을 제안한다. 제안한 히스토그램 등화 기법은 순서 기반의 히스토그램 추정 방식을 기반으로 하였으며, 누적 분포 함수 추정을 위해 UBM 학습 데이터를 이용한다. 제안한 히스토그램 등화 기법의 추정 과정은 다음과 같다.

1. UBM 학습 데이터 전체에서 추출한 특정 차수의 특징계수 열을 U라 하고, 다음과 같이 정의한다.

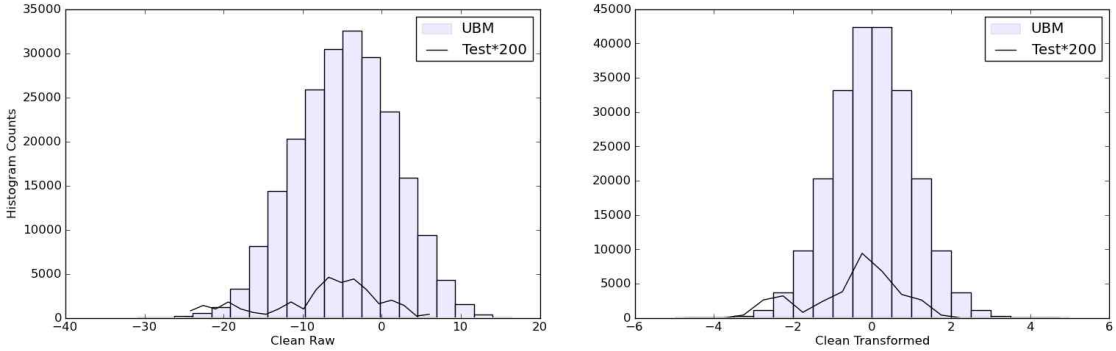

그림 2. UBM 데이터와 테스트 데이터의 특징 분포
Figure 2. Feature distributions of UBM and test data

$$U = \{u_1, u_2, \cdots, u_m, \cdots, u_M\} \tag{17}$$

2. 화자 적응과 화자 식별 실험을 위한 음성 데이터에서 추출한 특정 차수의 특징계수의 열을 S라 하고, 다음과 같이 정의한다.

$$S = \{s_1, s_2, \cdots, s_n, \cdots, s_N\} \tag{18}$$

3. U와 S를 결합하여 새로운 특징계수 열 O를 다음과 같이 정의한다.

$$O = \{o_1, o_2, \cdots, o_k, \cdots, o_K\}, \qquad K = M + N \tag{19}$$

4. O의 특징계수들을 크기의 오름차순으로 정렬한다. $D(k)$는 k번째 서열의 인덱스를 의미한다.

$$o_{D(1)} \le o_{D(2)} \le \cdots \le o_{D(k)} \le \cdots \le o_{D(K)} \tag{20}$$

5. 특징계수 열 O 중 S에 속하는 원소들의 누적 분포 Φ_n을 추정한다. $R_o(s_n)$은 특징계수 열 O에 속하는 특징계수 s_n의 O에 대한 서열을 의미한다.

$$\Phi_n = \frac{R_o(s_n) - 0.5}{K} \tag{21}$$

6. 추정한 누적 분포를 이용하여 특징계수를 기준분포 값으로 변환한다.

$$T(s_n) = C_{ref}^{-1}(\Phi_n) \tag{22}$$

<그림 2>는 UBM 데이터와 테스트 데이터의 분포를 나타낸다. 막대그래프 형식은 UBM 데이터의 분포이고, 선 형식은 테스트 데이터의 분포이다. 테스트 데이터는 UBM 데이터와 비교하면 양이 매우 적기 때문에, 원래의 값보다 200배 크게 나타내었다. <그림 2>의 왼쪽은 변환 전의 데이터이고, 오른쪽은 변환 후의 데이터이다. 변환 전 UBM 데이터는 중심이 -7 근처에 있는데 반해 변환 후 분포의 중심이 0에 있으며 정규 분포를 그린다. 테스트 데이터 역시 분포의 중심이 -5에 있으나 변환 후 분포는 0 근처에 중심이 위치한다.

<그림 3>은 음성 특징 변환을 보여준다. <그림 3>의 순서는 좌측부터 음성의 MFCC 특징, 히스토그램 등화 기법을 적용한 음성 특징, 제안한 방법의 음성 특징을 보여준다. 히스토그램 등화 기법은 한 발성 전체를 기준 분포로 바꾸기 때문에, 한 발성의 분포가 기준 분포인 표준 정규 분포를 나타내는 것을 볼 수 있다. 제안한 방법은 UBM 데이터의 분포를 이용하여 히스토그램 등화기법을 수행하기 때문에, -25~10에 분포하는 원래의 음성 특징의 분포를 표준 정규 분포 범위인 -4~4로 비선형 변환한 것을 볼 수 있다.

4. 실험 설계 및 결과 분석

4.1 사용한 데이터베이스

본 실험은 UBM (universal background model) 학습을 위해 ETRI (Electronics and Telecommunications Research Institute)에서 배포한 한국어 중가 마이크 화자인식용 음성 데이터베이스를 이용하였다. ETRI 중가 마이크 데이터베이스는 조용한 사무실 환경에서 수집하였으며, 총 250명 (주차 100명, 월차 100명, 3개월차 50명)의 화자가 발성한 2연 숫자, 4연 숫자, 문장 발성으로 구성되어 있다. 250명의 화자는 기간별로 4회 시차 발성하였으며, 시차 발성은 5회 회차 발성으로 구성되어 있다. 본 실험은 월차 100명과 3개월차 50명의 첫 번째 시차 발성 중에서 첫 번째 회차의 10~19번 문장 발성을 이용하여 UBM을 구성하였다. UBM 데이터는 테스트 데이터와 같은 표본화

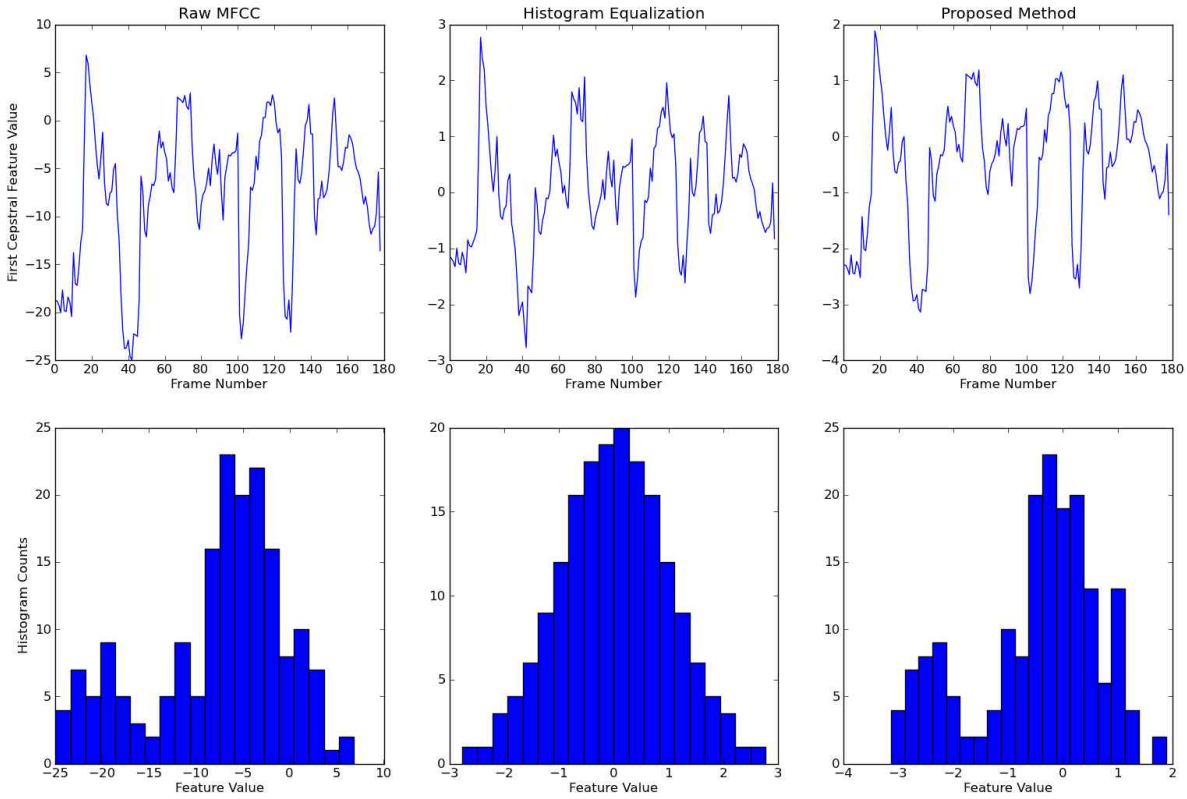

그림 3. 음성 특징의 변환
Figure 3. The transformation of speech features

주파수로 변환하였다.

본 실험의 테스트를 위해 LDC (Linguistic Data Consortium)에서 배포한 YOHO 데이터베이스와 ETRI에서 배포한 한국어 중가 마이크 화자인식용 음성 데이터베이스, 한국어 유선전화 화자인식용 음성 데이터베이스, 한국어 휴대전화 화자인식용 음성 데이터베이스를 이용하였다.

YOHO 데이터베이스는 조용한 사무실 환경에서 138명의 화자가 발성하였으며, 표본화 주파수는 8kHz로 저장되어 있다. 데이터베이스의 구성은 화자 모델의 학습을 위한 ENROLL과 화자 식별을 위한 VERIFY로 구성되어 있다. 화자 모델 적응을 위하여 ENROLL의 4개 시차 발성을 사용하였고, 테스트를 위해 VERIFY의 10개 시차 발성을 사용하였다.

한국어 중가 마이크 화자인식용 음성 데이터베이스를 이용한 실험은 주차 100명의 첫 번째 시차 발성 중에서 첫 번째 회차의 10~19번 문장 발성을 이용하여 화자 모델 적응하였고, 두 번째 시차 발성 중에서 첫 번째 회차의 10~19번 문장 발성을 이용하여 실험을 진행하였다.

한국어 유선전화 화자인식용 음성 데이터베이스는 조용한 환경에서 유선전화를 이용하여 수집하였으며, 표본화 주파수는 8kHz로 저장되어 있다. 총 206명 (주차 84명, 월차 81명, 3개월차 41명)의 화자가 발성한 2연 숫자, 4연 숫자, 문장 발성으로 구성되어 있다. 206명의 화자는 기간별로 4회 시차 발성하였으며, 각 시차 발성은 5회 회차 발성으로 구성되어 있다. 본 실험은 주차 84명의 첫 번째 시차 발성 중에서 첫 번째 회차의 10~19번 문장 발성을 이용하여 화자 모델 적응하였고, 두 번째 시차 발성 중에서 첫 번째 회차의 10~19번 문장 발성을 이용하여 평가하였다.

한국어 휴대전화 화자인식용 음성 데이터베이스는 다양한 환경 (사무실, 집, 거리, 지하철, 백화점, 자동차)에서 수집하였으며, 표본화 주파수는 8kHz로 저장되어 있다. 총 257명 (주차 104명, 월차 102명, 3개월차 51명)의 화자가 발성한 2연 숫자, 4연 숫자, 문장 발성으로 구성되어 있다. 257명의 화자는 기간별로 4회 시차 발성하였으며, 각 시차 발성은 5회 회차 발성으로 구성되어 있다. 본 실험은 주차 104명의 첫 번째 시차 발성 중에서 1~5회차의 10~19번 문장 발성을 이용하여 화자 모델 적응하였고, 두 번째 시차 발성 중에서 첫 번째 회차의 10~19번 문장 발성을 이용하여 테스트하였다.

4.2 음성 특징 추출 및 화자 모델 학습

본 실험은 음성 특징으로 MFCC (mel-frequency cepstral coefficients)를 사용하였다. 음성 신호에 pre-emphasis 필터 (계수 0.97)를 적용하였다. Hamming 창의 크기를 25ms, 이동주기는 10ms로 하여 프레임을 추출하였다. 추출한 프레임에 FFT

(fast Fourier transform) 변환을 통하여 스펙트럼을 얻고, 26차 멜-필터뱅크를 적용하였다. 멜-필터뱅크로부터 얻어진 결과에 로그를 적용한 후, 코사인 변환으로 18차의 MFCC 계수를 추출하였다. 무음 구간은 에너지를 기준으로 제거하였다.

화자 모델 학습에는 GMM-UBM[12] 방법을 사용하였다. UBM은 혼합 수 128개의 가우시안 혼합 모델을 사용하였으며, UBM으로부터 MAP 적응 (τ = 1) 방법으로 화자모델을 구성하였다.

4.3 실험 결과

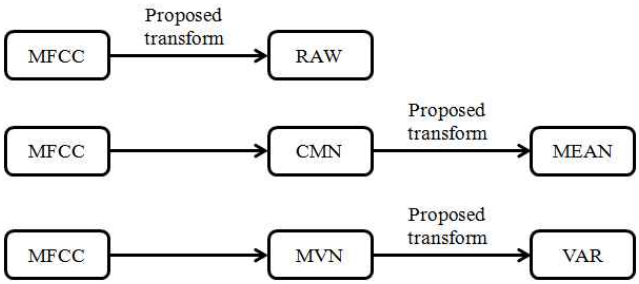

그림 4. 제안한 방법의 변환 과정
Figure 4. The transformation process of proposed method

<그림 4>는 제안한 방법의 특징 변환과정을 보여준다. RAW와 MEAN, VAR 방식은 UBM 학습 데이터와 테스트 데이터의 분포의 중심 위치, 분산 차이에 따른 성능을 비교하고자 수행하였다. RAW는 UBM 데이터와 테스트 데이터의 분포를 그대로 이용한 방법이다. MEAN은 테스트와 UBM 데이터와 테스트 데이터의 평균을 0으로 맞춰준 방법이다. VAR는 UBM 데이터와 테스트 데이터의 평균과 분산을 0과 1로 맞추어준 방법이다.

표 1. UBM 데이터를 이용한 변환과 제안한 방법의 오류율
Table 1. The error rates of the transformation using UBM data and the proposed method.

	UBM			proposed		
	RAW	MEAN	VAR	RAW	MEAN	VAR
YOHO	6.1	6.7	6.6	**6.0**	6.6	6.6
PC	1.6	**1.4**	2.0	1.6	**1.4**	2.0
CELL	28.0	**26.8**	29.0	27.7	26.9	29.0
TEL	**28.5**	33.6	32.2	28.8	33.3	32.3
평균	16.9			16.9		

<표 1>은 UBM 데이터를 이용한 특징 변환과 제안한 방법을 이용한 특징 변화의 화자인식 오류율을 보여준다. 표 1에서 평균 오류율은 UBM 데이터를 이용한 경우와 제안한 방법이 비슷한 성능을 보인다. 자세한 평균 오류율을 살펴보면 UBM 데이터를 이용한 경우는 16.874% 이고, 제안한 방법은 16.854 % 이다. 평균 성능이 좋은 제안한 방법에 대하여 실험을 진행하였다.

표 2. 음성 데이터베이스의 화자 식별 오류율
Table 2. Speaker identification error rates of speech database

DB / 채널 보상		YOHO	PC	TEL	CELL
MFCC		6.27	2.1	30.12	28.08
CMN		6.32	1.8	34.4	28.37
MVN		6.43	1.7	30	28.08
CHEQ		7.16	2.4	32.98	29.42
OHEQ		6.67	1.9	29.64	30
Proposed	RAW	**5.98**	1.6	**28.81**	27.69
	MEAN	6.63	**1.4**	33.33	**26.92**
	VAR	6.59	2	32.26	29.04

<표 2>는 YOHO, 한국어 중가 마이크 화자인식용 데이터베이스 (PC), 한국어 유선전화 화자인식용 음성 데이터베이스 (TEL), 한국어 휴대전화 화자인식용 음성 데이터베이스 (CELL)의 화자 식별 오류율이다. MFCC는 채널 보상을 하지 않은 경우이며, CMN, MVN, CHEQ, OHEQ는 각각 캡스트럼 평균 정규화, 평균-분산 정규화, 누적 기반의 히스토그램 등화 기법, 순서 기반의 히스토그램 등화 기법을 의미한다. 누적 기반의 히스토그램 등화 기법을 위한 빈 (bin) 개수는 1,000개로 하였다. 제안한 방법은 RAW, MEAN, VAR 방식으로 진행하였다. 실험 결과를 살펴보면, 제안한 방법은 다른 채널 보상 방법보다 높은 식별률을 보였다.

표 3. 제안한 방법의 상대적 오류 감소율
Table 3. The reduced relative error of the proposed method

채널 보상 / DB	MFCC	CMN	MVN	CHEQ	OHEQ
YOHO	4.6	5.4	7.0	16.5	10.3
PC	33.3	22.2	17.6	41.7	26.3
CELL	4.1	5.1	4.1	8.5	10.2
TEL	4.3	16.3	4.0	12.6	2.8
평균	11.6	12.2	8.2	19.8	12.4

<표 3>은 다른 채널 보상방법들과 제안한 방법의 상대적 오류 감소율이다. 제안한 방법은 성능이 좋게 나온 방법을 기준으로 다른 채널 보상방법과 비교하였다. 상대적으로 화자 식별률이 많이 떨어진 히스토그램 등화 기법들과 비교하여, 제안한 방법의 성능이 많이 개선되었다.

4.4 결과 분석

기존의 히스토그램 등화 기법을 수행하는 경우와 UBM 데이터를 이용하여 히스토그램 등화 기법을 수행하는 경우를 살펴보자. 실제 테스트 데이터의 정렬된 값이 [-17.42, -14.32,

-12.19, -10.11, …] 라 할 때, 이 테스트 데이터의 서열은 [1, 2, 3, 4, …] 로 정의된다. 하지만 UBM 데이터를 사용하여 서열을 정의하면, [12, 69, 77, 96, …] 로 정의된다. 또 다른 테스트 데이터의 정렬된 값이 [-3.14, -2.99, -2.08, -1.77, …] 로 정의될 때, 이 데이터의 서열은 [1, 2, 3, 4, …] 로 정의되고, UBM 데이터를 사용한 서열은 [301, 348, 391, 423, …] 로 정의된다. 일반적인 히스토그램 등화 기법을 이용하여 비선형 변환하면, 위의 두 데이터는 서로 다름에도 불구하고 비슷한 데이터로 변환 되어 화자 인식 성능을 하락 시킬 수 있다. 하지만 제안한 방법을 이용하면, 서로 다른 데이터로 변환되므로 일반적인 히스토그램 등화 기법에서 일어나는 성능 하락을 방지할 수 있다.

각 분포를 선형적으로 변화시켰을 때, 평균적으로 성능이 가장 높은 방법은 RAW 방법이었다. 일반적으로 학습과 인식의 채널이 같은 경우 CMN과 MVN은 평균과 분산정보를 잃게 되어 인식성능이 하락하는데 이와 비슷한 경향을 보인다. 이는 제안한 방법을 수행할 때, 분포에 변환을 수행하지 않고, 원래의 분포로 수행하는 것이 평균적으로 좋은 성능을 얻을 수 있을 것으로 보인다. 추후 소음 환경에서 MEAN, VAR의 경향성과 CMN, MVN의 경향성의 상관관계를 보이는 실험이 필요하다.

5. 결론

본 논문에서는 UBM 데이터를 이용한 히스토그램 등화 기법을 제안하였다. 테스트 데이터의 길이가 충분하지 않은 경우, 기존의 히스토그램 등화 기법은 다른 채널 보상 방법보다 성능이 떨어지는 경향을 보이나, UBM 데이터를 이용하여 테스트 데이터의 누적 분포를 추정하면 다른 채널 보상 방법보다 성능이 향상되었다.

UBM 데이터로 변환 함수를 만들어 특징을 변환하였을 때, 녹음 상태가 좋은 YOHO 데이터베이스에서는 성능이 떨어지는 현상을 보였고, ETRI PC 데이터베이스는 성능차이를 보이지 않았다. 하지만 녹음 상태가 좋지 않은 ETRI 유선전화, 휴대전화 데이터베이스에서는 성능이 오른 것을 볼 수 있다. 또한, UBM 데이터로 변환 함수를 만든 방법과 제안한 방법에 따른 오류 차이는 최대 4개이고, 평균 성능은 동일하므로 성능차이는 미미하다.

전체적인 성능을 살펴보면, 실험에 사용한 데이터베이스는 학습 환경과 인식 환경이 동일한 데이터베이스이므로 대체로 MFCC의 성능이 다른 채널 보상 방법보다 높은 경향을 보인다. 하지만 제안한 방법을 사용하면 다른 채널 보상 방법뿐 아니라 MFCC를 사용했을 때보다 더 높은 성능을 보여준다. 따라서 실제 환경에 제안한 방법을 사용하면 더 높은 성능 향상이 기대된다.

감사의 글

이 논문은 2011년도 서울시립대학교 교내학술연구비에 의하여 연구되었습니다.

참고문헌

[1] Atal. B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *Journal of the Acoustical Society of America*, vol. 55, No. 6, 1304-1312.

[2] Viikki. O. and Laurila. K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition, *Speech Communication*, Vol 25, 133-147.

[3] Huang. X., Acero. A. and Hon. H. (2001). *Spoken Language Language Processing: A Guide to Theory, Algorithm, and System Development*, Upper Saddle River NJ: Prentice-Hall.

[4] Moreno. P. J., Raj. B. and Stern. R. M. (1996). A vector Taylor series approach for environment independent speech recognition, *Proc. ICASSP*, 733-736.

[5] Kim. N. S. (2008). Statistical linear approximation for environment compensation, *IEEE Signal Processing Letters*, Vol. 5, No. 1, 8-10.

[6] Gonzalez. R. C. and Wintz. P. (1987). *Digital Image Processing*, Reading MA: Addision-Wesley.

[7] Pelecanos. J. and Sridharan. S. (2001). Feature warping for robust speaker verification, *A Speaker Odyssey - The speaker recognition workshop*, 213-218.

[8] Skosan. M. and Mashao. D. (2004). Matching feature distributions for robust speaker verification, *Proc. PRASA*, 42-47.

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

[10] Segura. J. C., Benitez. C., A. Torre. de la, Rubio. A. J. and Ramirez. J. (2006). Cpestral domain segmental nonlinear feature transformations for robust speech recognition, *IEEE Signal Processing Letters*, Vol. 11, 517-520.

[11] Torre. A. de la, Peinado. A. M., Segura. J. C., Perez-Cordoba. J. L., Benitez. M. C. and Rubio. A. J. (2005). Histogram equalization of speech representation for robust speech recognition, *IEEE Trans. Speech and Audio Processing*, Vol. 13, 355-366.

[12] Reynolds. D. A., Quatieri. T. F. and Dunn. R. B. (2000). Speaker verification using adapted gaussian mixture models, *Digital Signal Processing*, Vol. 10, 19-41.

• **김명재 (Kim, Myung-Jae)**
서울시립대학교 컴퓨터과학부
서울시 동대문구 서울시립대로 163
Tel: 02-2210-5322 Fax: 02-2210-5275
Email: mj@uos.ac.kr
관심분야: 화자인식, 음성인식

• **양일호 (Yang, IL-Ho)**
서울시립대학교 컴퓨터과학부
서울시 동대문구 서울시립대로 163
Tel: 02-2210-5322 Fax: 02-2210-5275
Email: heisco@hanmail.net
관심분야: 화자인식, 음성인식

• **소병민 (So, Byung-Min)**
서울시립대학교 컴퓨터과학부
서울시 동대문구 서울시립대로 163
Tel: 02-2210-5322 Fax: 02-2210-5275
Email: sbm1210@naver.com
관심분야: 화자인식, 음성인식

• **김민석 (Kim, Min-Seok)**
LG전자 전자기술원
서울시 서초구 양재동 221
Email: miseok3.kim@lge.com
관심분야: 화자인식, 음성인식

• **유하진 (Yu, Ha-Jin), 교신저자**
서울시립대학교 컴퓨터과학부
서울시 동대문구 서울시립대로 163
Tel: 02-2210-5613 Fax: 02-2210-5275
Email: hjyu@uos.ac.kr
관심분야: 화자인식, 음성인식