화자 인식을 위한 GMM기반의 이중 보상 구조

김유진(인하대), 정재호(인하대)

<차 례>

<Abstract>

Double Compensation Framework Based on GMM For Speaker Recognition

Yu-Jin Kim, Jae-Ho Chung

In this paper, we present a single framework based on GMM for speaker recognition. The proposed framework can simultaneously minimize environmental variations on mismatched conditions and adapt the bias free and speaker-dependent characteristics of claimant utterances to the background GMM to create a speaker model.

We compare the closed-set speaker identification for conventional method and the proposed method both on TIMIT and NTIMIT. In the several sets of experiments we show the improved recognition rates on a simulated channel and a telephone channel condition by 7.2% and 27.4% respectively.

* Keywords: speaker recognition, GMM, SBR, channel normalization

1. 서 론

지금까지 HMM과 GMM은 화자간의 차이를 모델링을 하기 위해 가장 효과적인 방법으로 증명되었다. 문장 종속 화자 인식 시스템의 경우 HMM은 패스워드 방식 발성의 음향학적 전이를 매우 효과적으로 활용할 수 있으며, 문장 독립 화자 인식 시스템을 위해서는 GMM이 화자 발성의 음향학적인 분포를 표현함에 있어서 매우 뛰어난 것으로 나타났다[1,2].

특히 GMM을 이용한 화자 확인 시스템인 경우 off-line에서 훈련된 화자독립, 어휘독립 GMM에 화자의 훈련용 발성을 적응시킴으로써 좋은 성능을 나타내는 것으로 나타났다[2]. 일반적으로 전역 모델(world model) 또는 바탕 모델(background model)로 불리는 적응용 GMM은 대용량의 다양한 화자의 음성을 이용하여 훈련되며 화자간 특성을 필터링한 정규화된 음향학적인 분포를 제공하게 된다. 따라서 효과적인 적응 방법을 이용할 경우 적은 양의 화자 발성으로도 화자의 특성만을 효과적으로 모델링할 수 있으며, 특히 훈련용 발성에서 발생되지 않은 음향학적인 특성까지도 모델링할 수 있는 장점을 가진다.

그러나 훈련 환경과 인식 환경이 불일치를 최소화하지 않는다면 이러한 적응 방법은 오히려 치명적인 문제점을 가진다. 적응된 모델은 훈련 환경에 맞추어지므로 상이한 인식 환경의 발성을 인식하지 못하기 때문이다.

이러한 환경 변이를 최소화하기 위한 방법으로서 잡음 성분을 제거함으로써 변이를 정규화시키는 스펙트럼 차감법, CMS 그리고 RASTA 등이 대표적이며 환경 변이에 강인한 음성 특징인 ACWC (Adaptive Component Weighting Cepstrum)가 제안되기도 했다. 또한 환경 변이에 대한 사전 정보를 통해 직접 특징 영역에서 환경 변이를 보상하는 CDCN, SBR 등이 제안되기도 했다[3,4,5,6]. 모델 영역에서는 대표적인 화자 적응 알고리즘인 MLLR과 전통적인 MAP을 이용하여 화자 특징을 적응함과 동시에 환경 변이를 보상해주는 효과를 얻을 수 있으며, Reynolds의 연구는 화자 인식에 적용된 대표적인 예이다. 최근에는 이러한 특징과 모델의 영역에서의 방법들을 동시에 적용하는 연구들이 진행되었는데, Sankar는 HMM기반의 음성인식에 적용된 통계적 정합(Stochastic Matching) 알고리즘을 제안하였다. 또한 통계적 정합과 SBR알고리즘을 HMM기반의 화자 인식에 적용한 연구결과도 보고되었다[7,8].

결과적으로 궁극적인 화자 인식 시스템을 위해서는 화자간의 변이는 최대화시키고 훈련 과정과 인식과정에서의 환경 변이는 최소화켜야 한다. 본 논문에서는 이를 위해 GMM 기반 화자 인식 시스템을 위한 이중 변이 보상 방법을 제안한다. 제안된 방법은 바탕 GMM 모델에 기반하여 특징 영역에서 환경 변이를 보상한 후 모델 영역에서의 MLLR과 MAP을 통해 적응함으로써 환경 변이와 화자간 변이에 대한 상충된 기준을 만족시키고자 한다. 2장에서는 화자 적응을 통한 화자 변

이 보상과 편향 제거를 통한 환경 변이 보상 기법에 대해서 설명하고, 3장에서는 이들 보상 기법이 융합된 바탕 모델에 기반한 보상 구조에 대해서 설명한다. 4장에서는 제안된 구조의 실험 및 결과에 대한 고찰을 제시하고, 마지막으로 5장에서 결론 및 향후 연구 방향으로 논문을 맺는다.

2. GMM을 이용한 화자 및 환경 변이 보상 기법

2.1. 화자 적응을 통한 화자 변이 보상

본 논문에서는 깨끗한 환경의 화자 및 문장 독립 GMM 모델을 off-line에서 훈련시켜 바탕 모델, Λ 로 사용하며 D 차원의 T개의 입력 특징 벡터, $X=\{x_t|t=1,\cdots,T\}$에 대한 확률값은

$$\log p(X|\Lambda)=\frac{1}{T}\sum_{t=1}^{T}\log\sum_{i=1}^{M}\omega_i p_i(x_t) \tag{1}$$

이며 여기서 $\Lambda=\{\ c_i, m_i, \Sigma_i\ \}$, $i=1,\cdots,M$ 이고 단일 가우시안 밀도 함수는

$$p_i(x_t)=\ K_i\exp\left\{-\frac{1}{2}(x_t-m_i)'\Sigma_i^{-1}(x_t-m_i)\right\} \tag{2}$$

로 표현할 수 있다. 이때 $K_i=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}$ 이고 c_i 는 조건 $\sum_i c_i=1$을 만족하는 i번째 혼합 밀도의 가중치, m_i는 평균 벡터 그리고 σ_i 는 공분산 행렬을 나타낸다.

이때 k 번째 화자 모델, $\lambda_k=\{\ \omega_i, \mu_i, \sigma_i\ \}\ i=1,\cdots,L$ 는 MLLR (Maximum Likelihood Linear Regression) 적응 알고리즘을 통해 생성될 수 있다. 새롭게 적응된 평균과 공분산을 구하기 위해 사용되는 변환 행렬은 다음과 같은 식으로 표현된다.

$$\mu_i\ = W m_i \tag{3}$$

$$\sigma_i\ = W\Sigma_i \tag{4}$$

변환 행렬 W는 EM 알고리즘에 의해 계산될 수 있다. 음향학적 유사도에 의해 군집화된 회귀 트리를 사용함으로써 적은 데이터 양으로도 바탕 모델의 전체 음향학적 공간을 빠르게 화자의 음향학적 공간으로 변환할 수 있다[9].

또한 화자 모델은 베이스(Bayes) 학습 또는 MAP (Maximum A Posterior)에 의해서도 구할 수 있다[2].

$$\alpha_i = \frac{n_i}{n_i + r} \tag{5}$$

$$\omega_i = [\alpha_i \dot{\omega}_i + (1-\alpha_i)c_i] \tag{6}$$

$$\mu_i = [\alpha_i \dot{\mu}_i + (1-\alpha_i)m_i] \tag{7}$$

$$\sigma_i^2 = [\alpha_i E(x_i^2) + (1-\alpha_i)(\Sigma_i^2 + m_i^2) + \mu_i^2] \tag{8}$$

여기서 입력 특징에 대한 바탕 모델 i 번째 혼합 밀도의 누적 확률값 $n = \sum_t p(\Lambda_{b,i}|x_t)$, 갱신된 혼합 밀도 가중치 $\dot{\omega}_i = n_i/T$, 갱신된 평균 벡터 $\dot{\mu}_i = \sum_t x_t \cdot p(\Lambda_{b,i}|x_t)$ 그리고 $E(x_i^2) = \sum_t x_t^2 \cdot p(\Lambda_{b,i}|x_t)$ 이며 $\Lambda_{b,i}$ 는 i 번째 바탕 모델의 혼합 밀도 파라미터이다. 식(5)에서 갱신 속도를 결정하는 상수 r 은 각 혼합 밀도 파라미터에 대해 다른 값을 가질 수 있으며 일반적으로 같은 값에 대해서 혼합 밀도 가중치와 평균벡터에 대해서만 적용하며 공분산 벡터는 갱신하지 않는다. 한편 점유 확률값 n_i 이 문턱 값 이상인 혼합 밀도에 대해서만 고려함으로써 성능 저하 없이 적응 속도 및 인식 속도를 향상시킬 수 있다[2].

2.2. 편향 성분 제거를 통한 환경 변이 보상

Mazin에 의해 제안된 SBR방법은 전화선 채널에 의한 환경 변이에 강인한 음성 인식 시스템을 위해 비편향(bias-free), 어휘 독립 코드북을 음성의 사전 정보로 간주하고 이것과 입력 음성 특징 벡터의 편향을 제거함으로써 환경 변이를 최소화하는 방법이다[5].

SBR 알고리즘은 켑스트럼 영역에서 선형적으로 표현될 수 있는 신호 왜곡을 가정하며 일반적으로 전화선채널에서의 잡음과 왜곡이 대표적인 예이다. 따라서 시간 t 에서의 특징 벡터 x_t 에 대한 왜곡은 첨가된 편향 성분, b 로 다음과 같이 간단히 표현될 수 있다.

$$y_t = x_t + b \tag{9}$$

또한 SBR 알고리즘은 환경 변이를 정규화하기 위해 미지의 편향 성분을 제거함으로써 입력 신호의 깨끗한 환경에서 훈련된 모델에 대한 확률값을 최대화할 수 있다는 가정에 근거하고 있다. 이는 다음과 같은 식으로 표현될 수 있다.

$$p(X|\Lambda) = \prod_{t}^{T} \max_{i} p_i(x_t) \tag{10}$$

여기서 $X=\{x_t|t=1,\cdots,T\}$ 는 이상적인 신호이고 $\Lambda=\{c_i, m_i, \Sigma_i \; i=1,\cdots,M\}$ 는 화자 독립 음소 모델이다. 이때 모델 Λ 가 영차(zeroth order) 마르코프 모델이라면 식 (10)의 우도 함수는 M 개의 가우시안 분포로 구성된 GMM에 대한 우도 함수로 생각할 수 있다. 이때, 각 가우시안 분포 성분은 대용량 화자 독립 음성 데이터에 의한 음향학적 분포를 표현하며 영차 마르코프 모델이다. 따라서 GMM은 화자 특성이 필터링된 음소 모델로 간주할 수 있다.

왜곡된 음성 특징 $Y=\{y_t|t=1,\cdots,T\}$ 에 대해서 편향 성분 $\tilde{b}$ 를 가정하고 식 (9)와의 관계를 적용하면 식 (10)은 다음과 같이 표현된다.

$$p(Y|\tilde{b},\Lambda) = \prod_{t=1}^{T} \max_{i} p_i(y_t - \tilde{b}) \tag{11}$$

이 식을 최대화시키는 편향 성분은 ML (Maximum Likelihood) 편향 성분 추정자, $\bar{b}$,로 생각할 수 있으며 일반적인 EM (Expectation and Maximization) 알고리즘에 의해서 구할 수 있다[10].

이를 위해 시간 t 에서 왜곡된 특징 벡터 y_t 에 대한 우도함수를 최대화시키는 혼합 밀도 평균 벡터를

$$\begin{aligned} z_t &= \arg\max_{i} p(y_t - \tilde{b}) \\ &= \arg\max_{i} K_i \cdot \exp\left\{-\frac{1}{2}[(y_t - \tilde{b}) - m_i]' \Sigma_i^{-1} [(y_t - \tilde{b}) - m_i]\right\} \end{aligned} \tag{12}$$

라고 정의하면, 식 (11)은 다음과 같이 표현된다.

$$p(Y|\tilde{b},\Lambda) = \prod_{t=1}^{T} K_{z_t} \cdot \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}(y_t - \tilde{b} - m_{z_t})' \Sigma_{z_t}^{-1} (y_t - \tilde{b} - m_{z_t})\right\} \tag{13}$$

이 식을 최대화시키는 $\hat{b}$ 값은 각 시간 t 에서 지수함수 내 이차방정식의 $\hat{b}$

에 대한 편미분에 의해서 구해지며 공분산 행렬이 대각행렬(diagonal)이라면, 다음과 같은 편향 성분 추정자를 얻을 수 있다.

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T}(h_{z_t})^T(y_t - m_{z_t}) \tag{14}$$

여기서 h_{z_t} 는 공분산행렬 $\Sigma_{z_t}^{-1}$ 의 대각 성분 행렬이다. 반복적인 EM 알고리즘에 따라 다시 우도함수를 최대화시키는 편향 성분 $\bar{b}'$ 을 구하기 위한 새로운 최 근접 혼합 밀도 성분은

$$z_t' = \arg\max_i \; p(y_t - \bar{b}) \tag{15}$$

로 표현된다. 이와 같은 최대화 과정과 평균 과정의 반복은 식 (11)의 확률값을 증가시키게 되며 그 증가치가 문턱 값 이하로 수렴될 때, 최대우도에 근거한 최적의 편향 성분 추정자를 구할 수 있다.

이상의 과정은 훈련과 인식과정 모두에서 적용되어 편향 성분이 제거된 특징 벡터에 대해서 처리하게 된다. 단 훈련 과정에서는 적응 알고리즘을 통해 편향 성분이 제거된 데이터를 이용하여 기존의 코드북을 새로운 코드북으로 갱신하여 보다 효과적인 편향 성분 추정을 할 수 있다.

3. 화자 및 환경 변이 보상을 위한 통합된 프레임워크

화자간 변이를 효과적으로 적응시킬 수 있는 보상 방법에서 화자 발성은 기존 바탕 모델의 훈련에 사용된 발성과의 환경 차이를 고려하지 않았으므로 적응된 모델은 화자 종속적인 특징뿐만 아니라 훈련 발성 환경을 그대로 반영하게 된다. 따라서 인식과정에서 또 다른 환경 변이를 포함한 음성이 입력될 경우 기존 모델과의 차이로 인해 성능 저하를 나타낼 수밖에 없다.

결국 화자 고유의 특성은 바탕 모델에 적응하여 보상해 주고 동시에 환경의 변이는 바탕 모델을 근거로 입력 음성을 보상해 줌으로써 불일치를 최소화하는 이중 보상 구조를 가져야 한다. 이러한 각 보상 기법의 효과적인 융합은 다음과 같은 몇 가지 공통점을 근거로 가능하다.

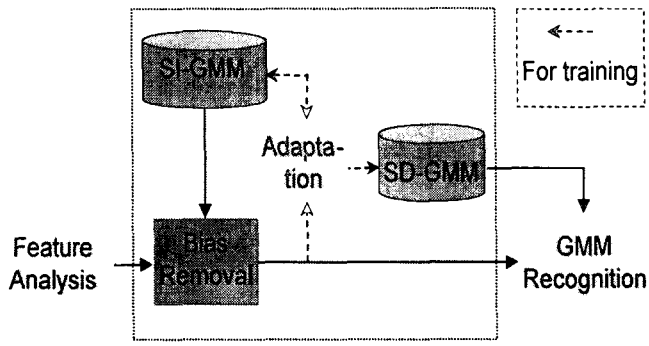
첫째, 깨끗한 환경의 화자 및 문장 독립 기준 모델을 사용한다.

둘째, 기준 모델은 영차(zeroth order) 마르코프 체인 모델인 GMM을 이용한다. 따라서 GMM의 파라미터인 평균, 공분산 등만을 이용하여 간단히 수행된다.

셋째, 각 보상 기법은 훈련과정에서 훈련용 음성을 바탕 모델에 적응시킨 적응 모델을 인식용 모델로 사용한다.

이러한 공통점을 기반으로 다른 목적의 두 기법은 결합될 수 있으며 <그림 1>은 융합된 이중 보상 구조를 보여준다. 구체적인 과정은 다음과 같다.

a. 비편향, 화자 독립 그리고 문장 독립 음성 데이터를 이용하여 바탕 GMM을 훈련한다. 훈련은 일반적인 EM 알고리즘을 사용한다.
b. 입력된 음성에 대해서 바탕 모델을 근거로 식 (14)를 이용하여 편향 성분을 추정하고 입력 음성으로부터 제거하여 갱신한다.
c. 갱신된 음성에 의한 식(10)의 확률값 향상이 문턱값 이하로 수렴될 때까지 (b)과정을 반복한다. 훈련 단계에서는 각 반복 과정에서 갱신된 음성을 MAP(식 3-4)과 MLLR(식 5-7)알고리즘에 의해 바탕 모델에 적응함으로써 갱신한다.
d. 훈련 단계일 경우, 편향 성분이 제거된 음성을 최초의 바탕 모델에 역시 MAP과 MLLR 알고리즘에 의해 적응하여 화자 변이를 보상하여 준다. 인식 단계에서는 편향 성분이 제거된 음성에 대한 화자 모델 확률값을 얻는다.

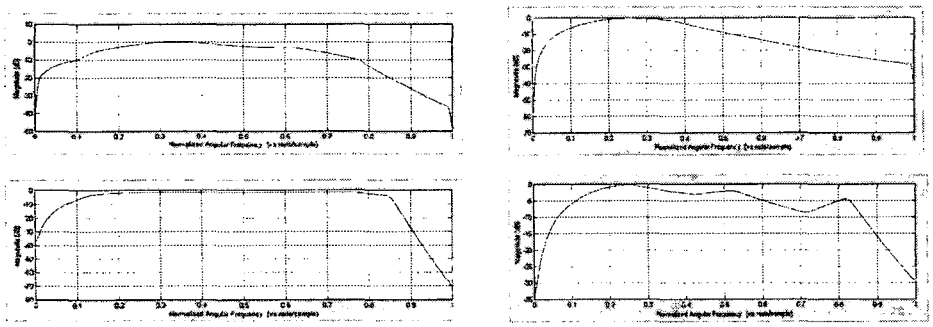


<그림 1> 제안된 GMM기반의 이중 보상 구조

4. 실험 및 결과

제안된 방법을 실험하기 위해 TIMIT 데이터베이스와 이를 실제 전화선 채널을

통과시켜 구축한 NTIMIT 데이터베이스를 이용한 폐집단 화자 식별을 수행하였다. 실험을 위해 각 데이터베이스의 DR3 테스트 부문에서 남자 화자 23명을 선택하였으며 바탕 모델의 훈련을 위해서 TIMIT 데이터베이스의 동일한 지역의 훈련 부문에서 56명의 남자 화자를 선택하였다. 모든 발성은 8Khz로 다운 샘플링하였으며 모의 전화 채널에 대한 실험을 수행하기위해 CMV, CPV, EMV 그리고 EPV의 4가지 모의 전화 채널 필터를 이용하여 화자의 발성을 필터링하였다. 모의 채널들은 각각 통과대역폭, 통과대역 감쇄 그리고 전이구간 등의 특성들이 각각 다르며 201개의 FIR 필터 계수로 표현되어있다[11]. <그림 2>는 각 필터의 주파수 응답을 보여준다.



<그림 2> 각 채널의 주파수 응답(시계 방향으로 CMV, CPV, EMV 그리고 EPV)

4.1. 특징 추출과 바탕 모델

화자 식별 실험을 위해 12차 LPC 분석을 통한 LPCC를 30ms의 프레임 크기에 대해서 10ms의 간격으로 추출하였다. 또한 강인한 화자 인식을 위한 특징으로 알려진 ACWC를 추출하여 인식 성능을 비교하였다.

바탕 모델은 1024개의 혼합 밀도 성분을 가진 GMM으로 표현하였으며 화자 독립, 문장 독립 모델을 얻기위해 56명 화자에 대한 총 560회의 발성으로 훈련하였다. 단, 바탕 모델의 훈련을 위한 발성은 깨끗한 환경을 가정하기 위해 모의 채널을 통과시키지 않은 TIMIT 데이터베이스의 발성만을 사용하였다. 훈련은 일반적인 EM 알고리즘을 통해 수행되었다.

4.2. 모의 채널에서의 화자 식별

첫 번째 실험에서는 모의 채널을 거친 음성에 대한 화자식별을 수행하였다. 화

자당 10회의 발성 중 5회의 발성은(SX) 묵음을 제외하고 10초 이하의 분량으로 훈련용 발성으로 연결하여 사용하였다. 나머지 5회의 (SA, SI)발성은 역시 묵음을 제외하고 각각 3초 이하의 길이로 인식을 위해 사용하였다.

실험은 화자 변이 보상 방법의 차이에 의한 성능과 대표적인 기존의 환경 변이의 정규화 방법인 CMS와의 성능 비교를 위해 수행되었다. 화자 변이 보상을 위한 MAP 적응 알고리즘의 적응 속도 상수 값은 $r=12$ 이며, MLLR 알고리즘의 회귀트리는 사용하지 않고 바탕 모델의 전체 혼합 밀도 성분에 대해 동일한 변환 행렬 W를 사용하였다. 이는 10초 이하의 비교적 짧은 훈련 발성에 대한 1024개의 혼합 밀도에 대한 회귀 트리의 점유확률이 크지 않으므로 의미가 없다고 판단했기 때문이다. 한편 MLLR과 MAP을 순서대로 사용한 MLLRMAP을 통해서도 화자적응을 시도하였다. 화자 적응 방법을 사용하지 않고 훈련 발성으로 직접 훈련한 방법과의 비교를 위해 46개의 혼합 밀도 성분을 가진 화자 종속 GMM을 생성하였다. 혼합 밀도 성분의 개수는 TIMIT 데이터베이스의 음소 레이블의 개수를 참고하여 결정되었다.

<표 1>은 CMV 모의 채널을 거친 훈련용 발성과 CPV 모의 채널을 거친 인식용 발성에 대한 화자 식별 인식률을 보여주고 있다. 실험 결과는 화자 적응 방법과 채널 정규화 방법에 따라 비교되었다.

<표 1> TIMIT 모의 채널에서의 화자 식별 인식률 (CMV 훈련/ CPV 인식)

	None	MAP	MLLR	MLLRMAP
Clean	81.7	87.8	94.8	95.7
None	18.4	15.8	15.8	16.7
CMS	53.5	59.7	53.5	61.4
DCGMM	-	62.3	63.2	65.8

모의 채널을 거치지 않은 깨끗한 발성과 환경 변이 정규화 방법에 따른 인식률을 비교했을 때, 환경 불일치로 인해 성능 저하가 매우 크다는 것을 알 수 있으며 기존의 CMS방법을 사용했을 경우에도 약 28%의 성능 저하가 나타났다(표의 2열). 화자간 변이를 보상하는 화자 적응에 따른 인식률은 채널 정규화를 시도하지 않은 경우를 제외하고 성능 향상을 보여 주었다(표의 4,5행). 이미 예상함과 같이 환경 변이를 정규화하지 않은 경우 화자간의 차이만 적응되는 것이 아니라 환경 변이 또한 반영되고, 따라서 인식 환경과의 불일치 조건은 더욱 심해지므로 적응 방법과 무관하게 성능이 저하됨을 알 수 있다(표의 3행).

마지막으로 제안된 방법은 모든 화자 적응 방법에서 기존의 CMS에 의한 환경 변이 정규화 방법에 의한 성능보다 우월함을 나타내었으며 특히 MLLRMAP에서

가장 좋은 성능을 보여 주었다(표의 5행).

4.3. NTIMIT에서의 화자 식별

두 번째 실험에서는 실제 전화선 환경에서의 실험을 위해 NTIMIT 데이터베이스의 발성을 이용하여 동일한 실험을 수행하였다. 모든 화자 모델의 훈련과 인식을 위한 발성은 TIMIT을 이용한 실험과 동일하게 선택하였다. 실험 결과는 표 2와 같다.

<표 2> NTIMIT에서의 화자 식별 인식률

	None	MAP	MLLR	MLLRMAP
None	53.9	42.6	29.6	42.6
CMS	46.1	54.8	40.0	50.4
SBRGMM	-	53.0	52.2	58.3

실험 결과는 우선 기존의 채널 정규화 방법인 CMS가 성능 향상에 기여하지 못함을 나타낸다. 이는 음성 성분의 켑스트랄 성분이 평균에 의해 0이 되고 채널 성분만 남는다는 CMS의 간단한 가정에 의해 화자 인식에 유용한 화자 정보가 손상되었기 때문으로 사료된다. 또한 MLLR에 의한 화자 적응은 오히려 성능 저하를 나타내는 것으로 나타났는데 이 역시 비슷한 원인일 것으로 판단된다. 제안된 방법은 MLLRMAP에서 최고의 인식률을 나타내었지만 MAP과 MLLR에 의한 적응 방법에서는 채널 정규화와 화자 적응을 시도하지 않은 방법에 비해 낮은 인식률을 보여주었다. 이것은 편향 성분 추정의 기준이 되는 바탕 모델의 환경이 화자의 발성환경과 큰 차이가 나고 따라서 정확한 추정이 어려웠기 때문으로 사료된다. 결론적으로 바탕 모델에 의한 편향 성분 추정과 정규화는 실제 채널 환경에서는 아직 많은 개선이 필요함을 알 수 있다.

4.4. 강인한 특징을 이용한 NTIMIT에서의 화자 식별

두 번째 실험 결과는 편향 성분 추정 및 정규화 과정에서 알고리즘의 한계로 인해 화자 정보가 상당부분 손실되는 것을 보여주었다. 이를 확인하고 개선하기 위해 환경 변이에 강인한 음성 특징을 사용한 실험을 수행하였다. 사용된 ACWC는 잡음에 민감한 성도 모델의 스펙트랄 곡선의 골(velly)을 정규화하고 봉우리를 강조함으로써 환경 변이에 강인한 성능을 나타내는 것으로 알려졌다[6]. 실험 결과는 <표 3>과 같다.

<표 3> ACWC음성 특징을 이용한 NTIMIT에서의 화자 식별 인식률

	None	MAP	MLLR	MLLRMAP
None	53.0	36.5	24.6	38.3
CMS	47.0	45.2	40.8	47.8
DCGMM	-	60.9	55.7	59.1

실험의 결과는 제안된 방법에 있어서 두 번째 실험의 결과에 비해 향상된 결과를 나타내었으나 기존의 CMS 방법에서는 뚜렷한 성능 향상을 나타내지 못했다. 이는 CMS가 환경 변이에 강인한 특징의 경우에도 화자 정보를 상당 부분 왜곡시키므로 적절하게 화자 적응의 효과를 나타내지 못하는 것으로 사료된다. 반면 제안된 방법은 적절하게 화자간의 차이점을 손상시키지 않고 환경 변이를 제거함으로써 화자 적응에서 향상된 성능을 보여주는 것으로 사료된다. 결론적으로 실험결과에서 제안된 방법의 인식률은 기존의 CMS에 의한 방법에 비해 최고 27.4% 향상되는 것으로 나타났다.

5. 결 론

본 논문에서는 화자 인식 알고리즘을 위한 이중 보상 구조를 제안하였으며, 특히 화자 적응을 위한 보상 과정과 HMM에 적용되었던 편향 제거 알고리즘이 동일한 GMM기반에서 효과적으로 융합될 수 있음을 보였다. 이를 위해 화자의 특성을 적응하기 위해 사용했던 바탕 모델을 기준으로 편향 성분을 추정하고 제거하였으며 소량의 데이터에 의한 훈련 과정을 고려하여 적응 알고리즘을 통해 편향 제거 과정의 모델 훈련을 수행하였다.

깨끗한 환경의 TIMIT과 전화선 환경의 NTIMIT을 이용한 폐집단 화자 식별 실험에서, 제안된 방법은 기존의 채널 정규화 알고리즘인 CMS와 비교하여 우월한 성능을 보여주었다.

앞으로 다수의 화자 음성을 이용한 화자 확인 실험을 통해서 제안된 방법을 검증하고 향상시킬 것이며 가산 잡음을 고려한 연구를 진행할 예정이다.

참 고 문 헌

[1] A. E. Rosenberg, F. K. Soong, "Sub-Word Unit Talker Verification Using Hidden Makov Models", *Proceedings of ICASSP-90*, pp.269-272, 1990.

[2] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, pp.19-41, 2000.

[3] R. J. Mammone, X. Zahng, R. P. Ramachandran, "Robust Speaker Recognition", *IEEE signal processing magazine*, pp.58-71, 1996.

[4] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition", Ph.D. thesis, Kluwer Academic Publishers.

[5] M. G. Rahim, B-H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", *IEEE Trans. On SAP*, Vol. 4, No. 1, pp.19-30, 1996.

[6] K. T. Assaleh, R. J. Mammone, "New LP-Derived Features for Speaker Identification", *IEEE trans. on Speech and Audio Processing*, Vol. 2, No. 4, 1994.

[7] A. Sankar, C-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. On SAP*, Vol. 4, No. 3, pp.190-202, 1996.

[8] L. Docio-Fernandez, C. Garcia-Mateo, "Application of Several Channel and Noise Compensation Techniques for Speaker Recognition", *Proceeding of Eurospeech97*, pp. 1115-1118, 1997.

[9] C. J. Leggetter, "Improved Acoustic Modelling For HMMs Using Linear Transformations", Ph.D. thesis, University of Cambridge, 1995.

[10] S. M. Kay, *Fundamentals of Statistical Signal Processing Vol. I: Estimation Theory*, Prentice Hall, 1993.

[11] J. Kupin, "A wireline Simulator[Software]", *CCR-P*, 1993.

접수일자: 2003년 2월 11일
수정일자: 2003년 3월 7일
게재결정: 2003년 3월 8일

▶ 김유진(Yu-Jin Kim)
주소: 402-751 인천시 남구 용현동 253
소속: 인하대학교 전자공학과
전화: 032)860-7420
FAX: 032)868-3654
E-mail: egkim@ieee.org

▶ 정재호(Jae-Ho Chung)
주소: 402-751 인천시 남구 용현동 253
소속: 인하대학교 전자공학과
전화: 032)860-7420
FAX: 032)868-3654
E-mail: jhchung@inha.ac.kr