한국음향학회지 제30권 제6호 pp. 315~323 (2011)

참조화자로부터 추정된 적응적 혼성 사전분포를 이용한 MAPLR 고속 화자적응

Rapid Speaker Adaptation Based on MAPLR with Adaptive Hybrid Priors Estimated from Reference Speakers

송 영 록, 김 형 순
(Young Rok Song, Hyung Soon Kim)

부산대학교 전자전기공학과
(접수일자: 2010년 12월 30일; 수정일자: 2011년 7월 27일; 채택일자: 2011년 8월 12일)

본 논문은 maximum a *posteriori* linear regression (MAPLR) 기반의 고속 화자적응 성능을 개선하기 위하여 사전분포를 추정하는 두 가지 방식을 제안한다. 일반적으로 MAPLR 방식에서 사용되는 변환행렬의 사전분포는 화자독립모델을 구성하는 훈련 화자들로부터 추정되어 모든 화자들에게 동등하게 적용된다. 본 논문에서는 새로운 화자에게 보다 더 적합한 사전분포를 적용하고자 적응 데이터를 이용하여 새로운 화자의 음향특성과 가까운 참조화자 집단을 선택한 후 참조화자 집단으로부터 사전분포를 추정하는 방법을 제안한다. 또한, 블록 대각 형태의 변환행렬의 사전분포를 추정하는 경우 사전분포의 평균행렬과 공분산행렬을 동일한 훈련 화자들로부터 얻어진 두 가지 형태의 변환행렬집단으로부터 각각 추정하는 방법을 제안한다. 제안된 방법의 성능 평가를 위하여 고립단어 인식실험을 통해 적응 단어의 개수에 따른 단어 인식률을 평가한다. 실험결과, 적응 단어 수가 매우 적을 때 기존의 MAPLR 방식에 비하여 통계적으로 유의미한 성능향상이 얻어짐을 보여준다.
핵심용어: 화자적응, MAPLR, 참조화자, 적응적 사전분포, 혼성 사전분포
투고분야: 음성처리 분야 (2.5)

This paper proposes two methods of estimating prior distribution to improve the performance of rapid speaker adaptation based on maximum a *posteriori* linear regression (MAPLR). In general, prior distribution of the transformation matrix used in MAPLR adaptation is estimated from all of the training speakers who are employed to construct the speaker-independent model, and it is applied identically to all new speakers. In this paper, we propose a method in which prior distribution is estimated from a group of reference speakers, selected using adaptation data, so that the acoustic characteristics of the selected reference speakers may be similar to that of the new speaker. Additionally, in MAPLR adaptation with block-diagonal transformation matrix, we propose a method in which the mean matrix and covariance matrix of prior distribution are estimated from two groups of transformation matrices obtained from the same training speakers, respectively. To evaluate the performance of the proposed methods, we examine word accuracy according to the number of adaptation words in the isolated word recognition task. Experimental results show that, for very limited adaptation data, statistically significant performance improvement is obtained in comparison with the conventional MAPLR adaptation.
Keywords: Speaker adaptation, MAPLR, reference speaker, adaptive priors, hybrid priors
ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

음성인식 기술은 인간의 가장 편리한 의사전달 수단인 음성을 통해 효율적으로 인간과 컴퓨터 사이의 인터페이스 (Human-Computer Interface)를 구현할 수 있는 방법으로서, 최근 스마트폰을 이용한 모바일 음성 검색 서비스의 등장으로 관심이 더욱 증대되고 있다. 음성인식 방식은 일반적으로 화자독립 (speaker independent; SI) 방식과 화자종속 (speaker dependent; SD) 방식으로 나눌 수 있다. 특정 화자의 훈련 데이터가 충분할 경우에는 화

책임저자: 김 형 순 (kimhs@pusan.ac.kr)
609-735 부산광역시 금정구 장전동 산30번지 부산대학교 전자전기공학과
(전화: 051-510-2452; 팩스: 051-516-4279)

자종속 방식이 화자독립 방식보다 높은 인식 성능을 나타내지만, 특정화자로부터 충분한 데이터를 확보하는 것이 현실적으로 매우 어렵기 때문에, 대다수의 응용분야에서 화자독립 방식이 사용된다. 화자적응 (speaker adaptation) 방식은 화자독립 방식을 기반으로 사용자의 제한된 음성 데이터를 활용하여 화자종속 방식에 가까운 성능을 얻고자 하는 방법이다. 일반적으로 적응 데이터가 많을수록 사용자의 특성을 잘 반영할 수 있지만 사용자의 불편이 초래되므로, 소규모의 발화만으로 우수한 성능을 얻을 수 있는 고속 화자적응 방식이 필요하다.

연속확률분포 hidden Markov model (HMM)에 바탕을 둔 화자적응 방식의 대표적인 방법으로는 maximum a *posteriori* (MAP) 방식 [1], maximum likelihood linear regression (MLLR) 방식 [2], 그리고 eigenvoice [3-4]와 cluster adaptive training (CAT) [5]으로 대표되는 화자 군집화 (speaker clustering) 방식 등이 있다. MAP 방식은 새로운 화자의 데이터와 사전 (prior) 정보를 이용해서 HMM 파라미터를 갱신하는 방법이다. 이 방식은 적응 데이터에 나타난 모델들만을 갱신하므로 적응 데이터가 매우 많은 경우 화자종속 시스템의 성능에 근접하지만, 많은 파라미터로 구성된 인식 시스템일수록 적응 속도가 매우 느리다. MLLR 방식은 선형회귀분석을 이용해서 HMM 모델간의 관계를 찾고, 이를 이용하여 적응 데이터에 나타나지 않은 모델도 갱신하는 방법이다. MAP 방식보다 적응 데이터가 비교적 적은 경우에 화자독립 모델을 효과적으로 갱신할 수 있지만, 적응 데이터의 양이 더 증가해도 화자종속 모델로 수렴하지 않는다는 단점이 있다. 이러한 단점을 보완하기 위해 많은 적응 데이터가 주어졌을 때에는 regression class tree를 이용하여 세분화된 변환행렬을 적용할 수도 있다 [6]. 한편, 변환행렬을 추정하기에 충분한 적응 데이터가 주어지지 않았을 때에는 MLLR 방식의 성능을 보장할 수 없는데, maximum a *posteriori* linear regression (MAPLR) 적응방식은 변환행렬에 대한 사전분포를 포함하여 MAP 추정을 함으로써 적응 데이터가 적을 때에 보다 강인하게 변환행렬을 추정할 수 있다 [7]. Eigenvoice 적응방식은 미리 여러 화자로부터 구성한 화자종속 모델들로부터 eigenvoice를 얻어 화자간의 변동을 표현하는 사전정보를 구성한다. 적응 단계에서는 적응 데이터로부터 각각의 eigenvoice의 가중치를 추정하고 eigenvoice들의 가중합으로 화자적응 (speaker adapted; SA) 모델을 구성한다. 특정 화자에 대해 추정해야 하는 파라미터의 수가 매우 적어서 적응 데이터 규모가 매우 적을 경우 다른 방식들에 비해 우수한 성능을 얻을 수 있으나, 적응 데이터가 많아질 경우에는 앞서 언급한 방식들보다 성능이 뒤떨어진다.

MAPLR이나 eigenvoice 적응방식과 같이 사전 정보를 이용함으로써 고속 화자적응을 구현하는 방법에서는 적절한 사전 정보를 미리 구성하는 것이 중요하다. 즉, MAPLR 방식에서는 변환행렬의 사전분포를 정의하는 초매개변수 (hyper-parameter)들을 정확히 추정해야 하고, eigenvoice 방식에서는 화자간의 변동을 잘 표현할 수 있는 eigenvoice를 정확히 추정해야 한다. 본 논문에서는 MAPLR 적응방식에서의 사전분포를 기존의 방식과 다르게 적용함으로써 고속 화자적응 성능을 높이고자 한다. 비슷한 의도의 선행연구로서 [14]에서는 화자집단을 군집화하고 화자집단별 사전분포를 구성하는 방법을 제안하였고, [15]에서는 사전분포 확률에 더 큰 가중치를 부여하는 방법을 제안하였으나, 두 방식 모두 성능 개선의 폭은 미미하였다.

기존의 MAPLR 방식에서는 별도로 구성된 변환행렬의 사전분포가 모든 새로운 화자들에 대해 동일하게 적용되는 반면, 본 논문에서는 먼저 새로운 화자와 가까운 참조 화자 (reference speaker)들을 찾고, 충분히 잘 훈련된 각 참조 화자들의 변환행렬들로부터 사전분포를 추정하는 방법을 제안한다. 즉, 모든 훈련 화자들로부터 사전분포를 추정하는 대신 새로운 화자에 가까운 몇몇의 참조 화자들의 정보만을 이용함으로써 새로운 화자에게 보다 더 적합한 사전분포를 추정하고자 하는 것이다. 제안한 방법은 미리 구성한 사전분포들 중 하나만을 선택할 수 있는 이전 방법 [14]에 비해 인식 화자에 따라 선택되는 참조 화자들에 의해 보다 더 적응적으로 사전분포를 추정할 수 있다.

한편, MAPLR 방식에서 변환행렬 형태가 정해지면 사전분포의 경우에도 동일한 형태의 변환행렬을 사용해 왔었다. 즉, 변환행렬이 full 행렬 또는 블록 대각 (block-diagonal) 행렬 형태일 경우 모든 훈련 화자들로부터 full 행렬 또는 블록 대각 형태의 변환행렬들을 얻어 각각의 사전분포를 추정한다. 본 논문에서는 블록 대각 변환행렬 기반의 MAPLR 적응 방식에서, full 행렬 형태의 변환행렬에서 구성된 사전분포와 블록 대각 행렬 형태의 변환행렬에서 구성된 사전분포 파라미터들을 혼합하여 사용함으로써 성능을 개선시키는 방법도 함께 제안한다.

본 논문의 구성은 다음과 같다. 서론에 이어 II장에서는 기존의 MLLR 및 MAPLR 화자적응 방식에 대해서 설명한다. III장에서는 참조 화자를 이용한 사전분포의 추정 방식을 제안하며, IV장에서 블록 대각 변환행렬 기반

의 MAPLR 적응 방식에서의 혼성 사전분포 추정 방식을 제안한다. V장에서 제안된 방식들의 실험 결과를 보여주고, 마지막으로 VI장에서 결론을 맺는다.

II. MLLR 및 MAPLR 화자적응 방식

2.1. MLLR 방식 [2]

MLLR 방식은 HMM 평균벡터를 갱신하는 변환행렬을 구하는 것이 목적이다. 변환된 d차원 평균 벡터 $\hat{\mu}$는 확장된 평균 벡터 ξ에 변환행렬 W를 곱함으로써 얻어질 수 있다. 즉,

$$\hat{\mu} = A\mu + b = W\xi \tag{1}$$

가 된다. 여기서 A는 $d \times d$ 행렬이고 b는 d차원 벡터이며, $W = [b, A]$는 $d \times (d+1)$ 행렬로서 적응 데이터의 우도(likelihood) 함수를 최대로 하는 행렬이고, $\xi = [1, \mu_1, \cdots, \mu_d]^T$이다. 관측벡터의 분포가 가우시안 (Gaussian)이라고 할 때, 상태 s의 확률밀도함수는 다음과 같으며, C_s는 $d \times d$의 공분산 행렬이다.

$$b_s(o_t) = \frac{1}{(2\pi)^{d/2}|C_s|^{1/2}} exp\left\{-\frac{1}{2}(o_t - W_s\xi_s)^T C_s^{-1}(o_t - W_s\xi_s)\right\} \tag{2}$$

MLLR 추정치를 구하기 위해 먼저 각각의 상태는 단일 mixture 가우시안 분포를 가지고, 공분산행렬 C_s는 대각행렬이라고 가정한다. 적응 데이터를 $O = [o_1, o_2, \cdots, o_T]$, 현재의 모델을 λ, 갱신된 모델을 $\bar{\lambda}$라고 하고, 모든 가능한 상태열의 집합을 Θ라고 하면, 전체 우도 함수는

$$P(O|\lambda) = \sum_{\theta \in \Theta} P(O, \theta|\lambda) \tag{3}$$

가 된다. 여기서 $P(O, \theta|\lambda)$는 주어진 모델 하에서 상태열 θ를 이용했을 때의 적응 데이터의 우도 함수이다. 식 (3)을 최대로 하는 변환행렬 W_s를 찾기 위해서, 보조함수를 다음과 같이 정의하고

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} P(O, \theta|\lambda) \log P(O, \theta|\bar{\lambda}) \tag{4}$$

expectation and maximization (EM) 알고리즘을 사용하여 변환행렬 W_s를 구할 수 있다. 전역 회귀 (global regression) 형태의 변환행렬, 즉, 모든 상태 s에 대해 동일한 변환행렬을 적용하는 경우, W_s 대신 W로 표현 가능하고, 그 행벡터별로

$$w^{(i)T} = G^{(i)^{-1}} z^{(i)T}, \qquad i = 1, \cdots, d \tag{5}$$

와 같이 적응 데이터와 화자독립 모델로부터 계산되는 $(d+1) \times (d+1)$ 행렬인 $G^{(i)}$와 $(d+1)$차원 벡터인 $z^{(i)}$로부터 추정한다. 여기서,

$$G^{(i)} = \sum_{s=1}^{S} \sum_{t=1}^{T} \frac{1}{\sigma_s^{(i)^2}} \gamma_t(s) \xi_s \xi_s^T \tag{6}$$

$$z^{(i)} = \sum_{s=1}^{S} \sum_{t=1}^{T} \frac{1}{\sigma_s^{(i)^2}} \gamma_t(s) o_t^{(i)} \xi_s^T \tag{7}$$

이다. $\sigma_s^{(i)^2}$는 상태 s의 대각행렬 형태의 공분산행렬 C_s의 i번째 대각성분이고, $\gamma_t(s)$는 o_t가 상태 s에 속할 확률이다. $o_t^{(i)}$는 o_t의 i번째 성분이며, S와 T는 각각 전체 상태 수 및 적응 데이터의 전체 프레임 수를 의미한다.

2.2. MAPLR 방식 [7]

MAPLR 방식은 제한된 적응 데이터가 주어졌을 때 변환행렬에 대한 추정이 제대로 이루어지지 않는 MLLR 방식의 단점을 보완하기 위해 변환행렬의 사전분포를 이용하는 방법이다. 변환행렬의 사전분포가 행렬정규분포 (matrix-variate normal distribution)를 따른다고 가정하면, 사전분포의 수식은

$$p(W) = \frac{1}{(2\pi)^{d(d+1)/2}|\Sigma|^{d/2}|\Psi|^{(d+1)/2}} \times \exp\left\{-\frac{1}{2} tr\left[(W - M)^T \Sigma^{-1} (W - M) \Psi^{-1}\right]\right\} \tag{8}$$

와 같이 표현되며, 평균행렬 $M \in \Re^{d \times (d+1)}$과 두 가지 공분산행렬 $\Sigma \in \Re^{d \times d}$, $\Psi \in \Re^{(d+1) \times (d+1)}$이 변환행렬의 사전분포를 결정한다. Σ가 단위행렬이라고 가정하면, MAP 추정에 기반하여 변환행렬은 행벡터별로

$$w^{(i)T} = \left[G^{(i)} + \Psi^{(i)^{-1}}\right]^{-1} \left[z^{(i)} + m^{(i)} \Psi^{(i)^{-1}}\right]^T, \quad i = 1, \cdots, d \tag{9}$$

와 같이 식 (5)에 사전분포 파라미터들이 추가된 형태로 추정되며, 여기서 $m^{(i)}$는 평균행렬 M의 i번째 행벡터이다. N명의 훈련 화자로부터 사전분포를 추정할 경우, 화

자별 훈련 데이터와 화자독립 모델을 이용하여 변환행렬 집단 $\{\boldsymbol{W}_n\}_{n=1}^{N}$을 생성한 후, 이들로부터 평균행렬과 공분산행렬을 각각

$$\boldsymbol{M} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{W}_n \tag{10}$$

$$\boldsymbol{\Psi}^{(i)} = \frac{1}{N-1}\sum_{n=1}^{N}\left(\boldsymbol{w}_n^{(i)}-\boldsymbol{m}^{(i)}\right)^T\left(\boldsymbol{w}_n^{(i)}-\boldsymbol{m}^{(i)}\right),\quad i=1,\ldots,d \tag{11}$$

와 같이 추정한다. 여기서 $\boldsymbol{w}_n^{(i)}$는 n번째 훈련 화자의 변환행렬 $\boldsymbol{W}_n$의 i번째 행벡터이다.

III. 참조화자로부터의 적응적 사전분포 추정

MAPLR 적응방식에서는 적응 데이터가 적을수록 변환행렬의 사전분포 추정의 정확도가 중요한 역할을 하는데, 일반적으로 N명의 훈련화자로부터 생성된 변환행렬 집단 $\{\boldsymbol{W}_n\}_{n=1}^{N}$을 구성하는 모든 변환행렬들로부터 사전분포를 추정한다. 그러므로 새로운 화자의 특성에 관계없이 모든 새로운 화자에 대해 동일한 사전분포가 적용된다. MAPLR 기반의 고속 화자적응 성능을 보다 더 개선하기 위해서는 사전분포를 적용하는 단계에서도 새로운 화자의 특성을 고려할 필요가 있다. 그림 1에서 보는 바와 같이 많은 화자들의 공통적인 특성보다는 비슷한 특성을 가진 화자들의 특성으로부터 사전분포를 추정하여 새로운 화자에게 보다 더 적합한 사전분포를 적용할 수 있고, 적응 데이터가 적을 때의 성능 개선을 기대할 수 있다. 따라서 본 논문에서는 변환행렬의 사전분포를 미리 전체 훈련화자들로부터 구하는 대신에, 적응 데이터를 이용하여 선택된 참조화자 (reference speaker)들의 변환행렬만으로 사전분포를 추정하여 MAPLR 적응방식에 적용하는 방법을 제안한다.

제안하는 방법은 다음과 같이 두 단계로 나뉜다.

1) 참조화자 선택 : 적응 데이터로부터 새로운 화자의 특성과 가까운 훈련 화자들을 선택
2) 사전분포 추정 : 참조화자 집단 내의 변환행렬들로부터 사전분포를 추정

제안 방법에서는 1단계에서의 참조화자의 적절한 선택 과정이 성능과 직결되는데, 적응 데이터로부터 참조화자를 선택하는 데에는 두 가지 방법을 고려할 수 있다. 첫 번째로 [10]와 같이 reference speaker weighting (RSW)[1]

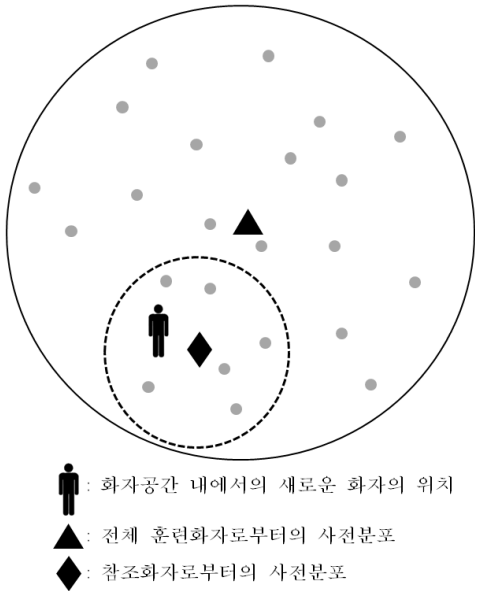

그림 1. 화자공간에서의 사전분포 추정 예시
Fig. 1. Illustration of estimation of prior distribution in speaker space.

[9] 적응방식에서 적응 데이터의 전사 정보를 이용하여 각 화자종속 HMM 모델과의 우도를 계산한 다음, 우도값이 높은 순으로 참조화자를 선택할 수 있다. 아니면 [11]와 같이 각 훈련화자별로 Gaussian mixture model (GMM)을 별도로 구성하여 우도를 계산하고 참조화자를 선택할 수도 있다. 참조화자의 선택에 따라 새로운 화자에게 적응적으로 사전분포가 추정되므로, 본 논문에서는 이 방식을 '참조화자로부터의 적응적 사전분포 추정' 방식이라고 부르기로 한다. 그림 2에 제안한 방법의 순서도를 나타내었다.

그런데 제안된 방식에서는 전체 훈련화자가 아닌 일부 참조화자의 정보만을 이용하여 사전분포를 추정하기 때문에, 사전분포 추정을 위한 변환행렬 샘플 수는 감소하게 되며, 그 결과로 추정을 위한 데이터 불충분의 문제가 발생할 수도 있다. 즉, 전체 N명의 훈련화자 중에서 R명의 참조화자를 선택한다고 가정했을 때 ($R<N$), 공분산행렬의 rank가 d보다 작아져 역행렬이 존재하지 않을 수도 있다. 이로 인하여 식 (9)에서 계산 오류가 발생할 수 있는데, 본 논문에서는 이러한 문제를 예방하기 위해 참조화자가 일정 숫자 이하일 경우 공분산행렬을 full 행렬 대신 블록 대각 (bock diagonal) 행렬 형태로 추정하는

1) Eigenvoice와 공통점을 가지는 고속 화자적응방식으로서, 새로운 화자에 대한 적응모델을 eigenvoice의 가중합이 아닌 미리 선택된 참조화자들의 화자종속모델들의 가중합으로 표현하는 방법이다.

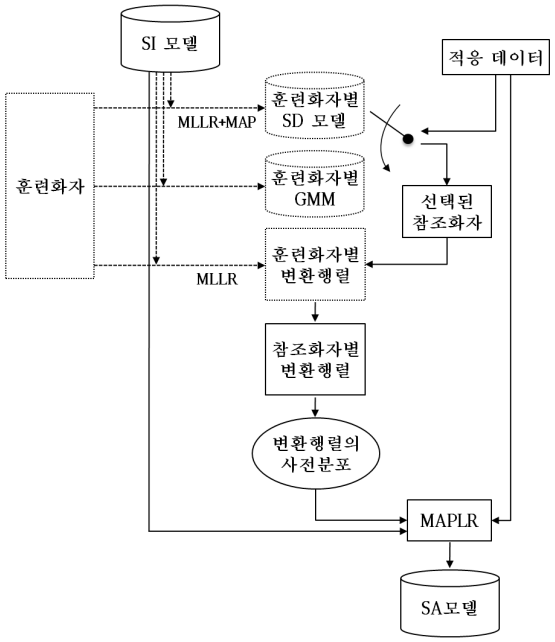

그림 2. 참조화자로부터의 적응적 사전분포 추정
Fig. 2. Estimation of adaptive prior distribution from reference speakers.

방법을 도입한다. 구체적인 예로, p차원의 특징계수와 그 delta, delta-delta 계수를 추출하여 총 d차원의 관측 벡터를 추출하는 경우 (즉, $d=3p$), $(d+1)\times(d+1)$ 행렬 형태인 사전분포의 공분산행렬 $\boldsymbol{\Psi}$은 바이어스 성분에 해당하는 1×1 행렬과 static 계수에 해당하는 $p\times p$ 행렬, delta 계수에 해당하는 $p\times p$ 행렬, delta-delta 계수에 해당하는 $p\times p$ 행렬로 이루어진 블록 대각 행렬 형태로 추정한다. 이와 같이 공분산행렬의 형태를 보다 간략하게 만듦으로써 더 적은 참조화자의 정보만을 선택하여 사전분포에 적용할 수 있다.

참고로, 본 논문에서 제안하는 적응적 사전분포 추정 방식과 RSW 방법 등을 통한 화자군집화 (speaker clustering) 적응방식과의 비교는 다음과 같이 설명될 수 있다. 화자군집화 적응방식에서는 화자공간을 적은 숫자의 파라미터로 표현하고 적응 데이터로부터 이들 파라미터를 추정함으로써 직접 화자적응을 수행하는 것이며, 본 논문의 적응적 방법에서도 동일하게 화자공간을 적은 숫자의 파라미터로 표현하고 적응 데이터로부터 이들 파라미터를 추정하기는 하지만, 이 정보를 MAPLR 적응 방식에서의 사전분포에만 적용한다는 것이 차이점이다. 따라서 전자의 방법의 경우 적응 데이터가 많아지더라도 성능이 빨리 수렴해버리는 문제가 있는 반면에, 후자의 방법에서는 사전분포의 정교성 향상을 통해 적은 양의 적응 데이터에서 성능향상을 얻을 수 있을 뿐 아니라 적응 데이터가 많아질 때에도 MLLR 방식 수준의 성능향상을 보장할 수 있다는 장점이 있다.

IV. 혼성 사전분포 추정

앞서 설명한 바와 같이 MAPLR 적응 방식은 적응 데이터가 충분하지 않은 상태에서 적응 데이터 정보와 사전분포 정보를 함께 이용하여 최적의 변환행렬을 추정한다. 참고로, 식 (9)에서는 적응 데이터로부터 구해지는 $\boldsymbol{G}^{(i)}$와 $\boldsymbol{z}^{(i)}$, 그리고 사전분포 정보인 $\boldsymbol{m}^{(i)}$와 $\boldsymbol{\Psi}^{(i)}$를 함께 이용해서 최종적인 변환행렬을 추정하는 것을 보여준다. 따라서 MAPLR 적응 방식으로 최고의 성능을 얻기 위해서는 적응 데이터 및 사전분포의 양쪽으로부터 최선의 정보를 확보하는 것이 필요하다.

적응 데이터가 충분하지 않을 경우 기존의 MLLR 적응 방식으로 full 행렬 형태의 변환행렬을 안정적으로 추정하기 힘들고, 결과적으로 심각한 성능 저하가 초래된다. MLLR 적응 방식에서 이러한 적응 데이터 부족의 문제를 완화시키는 방법으로 블록 대각 형태의 변환행렬을 사용하는 방법이 사용된다 [2]. 앞 장에서 언급한 것처럼 p차원의 특징벡터와 delta, delta-delta 계수를 추출하여 총 d차원의 특징벡터를 추출하는 경우 (즉, $d=3p$), 식 (1)에서 $d\times(d+1)$ 변환행렬 $\boldsymbol{W}$의 구성성분 중 바이어스 성분에 해당하는 $d\times1$ 행렬 $\boldsymbol{b}$를 제외한 $d\times d$ 행렬 $\boldsymbol{A}$를 3개의 $p\times p$ 부분행렬 (sub-matrix)로 블록 대각 형태의 변환행렬을 구성하게 된다. 블록 대각 형태의 변환행렬은 full 행렬 형태에 비해 추정해야 되는 파라미터 수가 훨씬 적기 때문에 적응 데이터가 충분하지 않을 경우 상대적으로 높은 화자적응 성능을 얻을 수 있다. 물론 적응 데이터가 충분하다면 full 행렬 형태의 변환 행렬이 성능면에서 유리하다.

기존의 MAPLR 방식에서는 변환행렬 형태가 정해지면 당연히 사전분포의 경우에도 동일한 형태의 변환행렬을 사용하는 것을 전제로 한다. 즉, 변환행렬이 full 행렬 또는 블록 대각 행렬 형태일 경우 모든 훈련 화자들로부터 full 행렬 또는 블록 대각 형태의 변환행렬들을 얻어 각각의 사전분포를 추정한다. 본 논문에서는 적응 데이터 규모가 작을 경우에 MAPLR 성능을 보다 향상시키기 위하여, 변환행렬 자체는 불충분한 적응 데이터에 유리한 블록 대각 형태를 사용하되, 사전분포의 파라미터들은 전적으로 블록 대각 행렬들에만 의존하지 않는 방법을 제안한다.

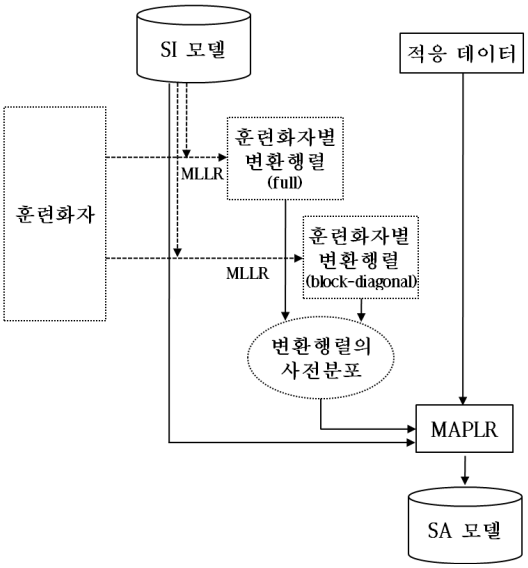

그림 3. 블록 대각 변환행렬을 이용한 MAPLR 방식에서의 혼성 사전분포 추정

Fig. 3. Estimation of hybrid prior distribution in MAPLR with block-diagonal transformation matrix.

실제로 변환행렬의 사전분포는 훈련용 화자 데이터로부터 구한 화자별 변환행렬들로 추정하며, 훈련 화자별 데이터가 충분한 경우 블록 대각 행렬 형태보다는 full 행렬 형태의 변환 행렬이 화자특성을 더 잘 표현하게 된다. 따라서 사전분포 파리미터 중 평균행렬은 어차피 최종적인 변환행렬이 블록 대각 형태이므로 동일하게 블록 대각 형태의 변환행렬들로부터 추정하더라도, 변환행렬의 각 차원간의 상관관계를 나타내는 공분산행렬은 변환행렬의 신뢰도가 상대적으로 더 높은 full 행렬 형태의 변환행렬 기반으로 추정하는 것을 시도해 볼 만한 방법이다. 따라서 본 논문에서는 블록 대각 행렬 기반의 MAPLR 적응 방식에서 사전분포의 추정을 위한 변환행렬 집단을 full 행렬 및 블록 대각 행렬의 두 가지 형태로 모두 구성한 후, 사전분포 파라미터 중 location 파라미터인 평균행렬은 블록 대각 행렬 형태에서 추정하고, scale 파라미터인 공분산행렬은 full 행렬 형태에서 추정하는 방법을 제안한다. 두 가지 사전분포 파라미터를 각각 다른 변환행렬 집단에서 추정하므로, 본 논문에서는 이를 '혼성(hybrid) 사전분포 추정' 방식이라고 부르기로 한다. 사전분포의 공분산행렬을 full 행렬 형태의 변환행렬로부터 추정하더라도 어차피 최종 변환행렬이 블록 대각 형태이므로 앞 장에서 언급한 바와 같이 공분산행렬 역시 블록 대각 형태로 추정하도록 하며, 이를 통해 변환행렬 샘플 수가 적은 경우에도 상대적으로 안정적인 추정이 가능해진다. 그림 3에 본 장에서 제안한 방식의 순서도를 나타내었다.

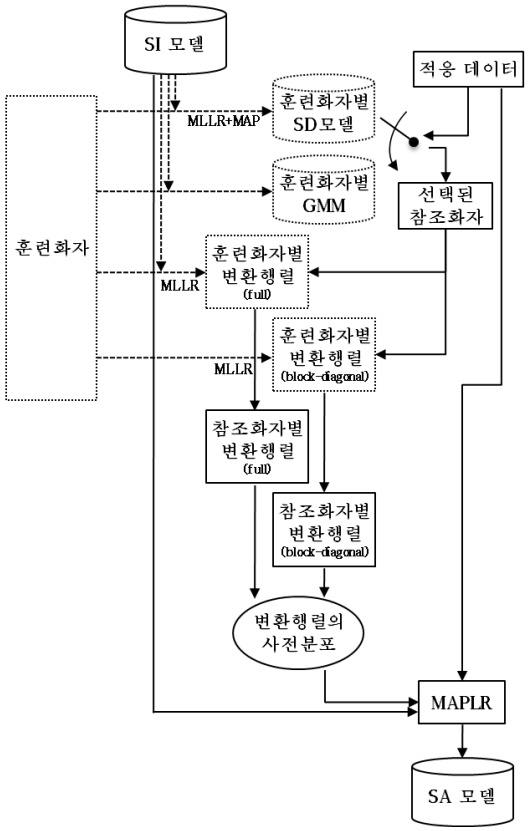

그림 4. 블록 대각 변환행렬을 이용한 MAPLR 방식에서의 참조화자로부터의 적응적 혼성 사전분포 추정

Fig. 4. Estimation of adaptive hybrid prior distribution from reference speakers in MAPLR with block-diagonal transformation matrix.

그림 3에서 보는 바와 같이 제안한 방법에 따라 사전분포를 추정하기 위해서는 두 가지 형태의 변환행렬 집단이 필요하다. 또한 그림 4와 같이 III장에서 제안한 방법과 함께 결합함으로써 고속 화자적응 성능의 추가적인 향상을 도모할 수 있다.

V. 실험 및 결과

5.1. 실험 환경

훈련을 위하여 ETRI에서 구축한 3음소열 최적화 단어(Phonetically Optimized Words, POW) [12] 음성 데이터베이스 전체 (남성화자 40명, 여성화자 40명)를 사용하였다. 본 실험에서 20 ms Hamming 창을 10 ms씩 이동시키면서 얻은 12차의 Mel-Frequency Cepstral Coefficient (MFCC)와 delta, delta-delta 계수들을 추출하여 총 36차의 음성 특징벡터를 사용하였다. 그리고 유성음화 자음 및 묵음을 포함한 46개의 유사음소 단위 (phone-like

unit)를 기본으로 상태 수준에서 트리 기반 군집화 (tree-based clustering)을 적용한 triphone을 기본모델로 사용하였으며 모델 당 상태수는 3개이다. 총 6211개의 군집화된 상태 (node state)를 생성했고, MLLR 변환행렬은 단일 변환행렬만을 사용하는 전역 회귀 (global regression) 형태를 사용하였다. 변환행렬의 사전분포를 추정하기 위해 화자독립 모델의 생성에 사용된 POW DB 각 훈련 화자에 대하여 MLLR을 적용하여 80개의 변환행렬을 full 행렬 형태 및 블록 대각 행렬 형태로 각각 구했다.

또한 참조화자를 선택하는 과정을 구현하기 위해 각 훈련 화자에 대하여 MLLR과 MAP 적응 방식을 적용하여 80개의 화자종속 모델을 구성하였고, 별도로 64개의 mixture를 가지는 화자별 GMM을 구성하였다. 적응 데이터 중 맨 처음 1단어만으로 각 화자종속 모델 또는 GMM과의 우도를 계산하여 참조화자를 선택하도록 하였다.

화자적응 및 성능평가에서는 452 균일 음소 분포 단어 (Phonetically Balanced Words, PBW) [13] 데이터베이스의 남성화자 10명, 여성화자 10명에 대하여 각 화자별로 1개부터 10개까지 단어 수를 늘려가며 supervised mode 화자적응에 사용하였고, 나머지 중 400개의 단어를 성능평가에 사용하였다. 본 논문에서는 상태당 mixture가 1개일 때의 인식성능만을 평가하였으며, 이 때 화자독립 모델을 사용하였을 경우에는 93.58 %의 단어 인식률을 보였다. 적응 데이터의 개수에 따른 기존의 MLLR 및 MAPLR 적응방식의 성능은 다음과 같다.

표 1에서 보는 바와 같이, 적응 데이터가 적을 때, 즉, 1개 또는 5개의 적응 단어가 주어졌을 때 MLLR의 성능은 매우 저조하고, 추정 대상 파라미터 수가 상대적으로 적은 블록 대각 행렬 형태 (표에서 "b-d"로 표기)의 변환행렬이 full 행렬 형태(표에서 "full"로 표기)에 비해 성능면에서 더 우수함을 알 수 있다. 그리고, 사전정보를 포함하는 MAPLR 방식을 적용함으로써, MLLR 방식에 비해 적응 데이터 수가 매우 적을 때의 성능, 즉, 고속 화자적응 성능이 크게 개선되는 것을 확인할 수 있다. 또한, MAPLR 방식에서는 블록 대각 행렬 형태의 변환행렬을 사용하더라도 full 행렬 형태를 사용한 경우의 성능에 필적하는 것을 볼 수 있다.

5.2. 실험 결과 및 검토

본 논문의 실험 조건에서는 full 행렬 형태의 변환행렬을 이용하여 참조화자들로부터 사전분포를 추정하는 경우, 공분산행렬의 역행렬의 계산을 위해 적어도 40명의 훈련 데이터가 필요했다. 또한, 참조화자가 40명 이상으로 증가할수록 대체적으로 성능이 감소하는 경향을 보였다. 그러므로 참조화자의 선택에서 사용되는 두 가지 방법의 성능 비교를 위해 full 행렬 형태의 변환행렬에 대해 40명 및 50명의 참조화자를 선택해서 MAPLR 적응 실험을 하였으며, 그 결과를 표 2에 나타내었다. 표에서 볼 수 있듯이 GMM을 이용하여 참조화자를 선택하는 방법과 훈련화자별 화자종속 HMM 모델을 이용하여 참조화자를 선택하는 방법의 MAPLR 화자적응 성능은 거의 비슷하게 나타났다.

따라서, 본 논문의 이후의 실험에서는 별도의 GMM을 구성하지 않고 훈련화자별 화자종속 모델을 기반으로 참조화자를 선택하는 방법을 사용하였다. 표 3에 III장에서 제안한 참조화자로부터의 적응적 사전분포를 적용한 MAPLR 적응방식의 성능을 나타내었다.

표에서 "Ref. spk."는 참조화자의 수이며, "full"은 기존의 방식과 마찬가지로 full 행렬 형태로 사전분포의 공분산행렬을 추정 및 적용하였음을 의미하고, "b-d"는 더 적은 수의 참조화자로부터 사전정보를 추정하기 위해 사전분포의 공분산행렬을 블록 대각화하여 추정 및 적용하였음을 의미한다. 본 논문의 실험 조건에서 블록 대각 행렬 방식으로 참조화자들로부터 사전분포를 추정하는 경우, 공분산행렬의 역행렬의 계산을 위해 적어도 15명 이

표 1. 기존의 MLLR 및 MAPLR 적응방식의 단어인식률 (%)
Table 1. Word accuracy of conventional MLLR and MAPLR adaptation methods.

Method		No. of adaptation words			
		1	3	5	10
Baseline		93.58			
MLLR	full	0.00	11.25	75.84	96.84
	b-d	10.57	87.53	96.33	97.61
MAPLR	full	97.16	97.49	97.61	97.89
	b-d	97.09	97.41	97.54	97.86

표 2. 참조 화자 선택 과정에서 GMM과 HMM을 이용했을 때의 MAPLR 성능 비교 (%)
Table 2. Performance comparison of MAPLR adaptation method using GMM and HMM in reference speaker selection.

Method	Ref. spk.	No. of adaptation words			
		1	3	5	10
GMM	40	97.33	97.59	97.64	97.95
HMM		97.34	97.58	97.69	97.89
GMM	50	97.22	97.47	97.62	97.86
HMM		97.29	97.50	97.62	97.89

표 3. 참조화자로부터 추정된 적응적 사전분포를 적용한 MAPLR 적응방식의 단어인식률 (%)
Table 3. Word accuracy of MAPLR using adaptive prior distribution estimated from reference speakers.

Method	Ref. spk.	Prior cov.	No. of adaptation words			
			1	3	5	10
MAPLR	-	full	97.16	97.49	97.61	97.89
Adaptive prior	40	full	97.33*	97.59	97.64	97.95
	20	b-d	97.39*	97.51	97.62	97.97
	40	b-d	97.34*	97.58	97.69	97.89

*p-value < 0.01.

상의 훈련 데이터가 필요했으며, 참조화자가 40명 이상으로 증가할수록 대체적으로 성능이 감소하는 경향을 보였기 때문에, 표에서 참조화자가 20명 및 40명일 때의 성능을 표시하였다. 적응 데이터가 5단어 이상일 때의 성능 개선은 미미하지만 특히 1단어일 때 기존의 MAPLR방식에 비해 성능이 개선되었고, 20명의 참조화자로부터 블록 대각 방식으로 사전분포의 공분산행렬을 추정했을 때 1단어에서의 성능이 가장 크게 개선되었다. 개선된 성능의 통계적 유의미성 [8]을 확인한 결과, 적응 단어 수가 1개일 경우에만 1 %의 유의수준에서 통계적으로 유의미한 성능향상이 얻어졌다. (표에서 *로 표시된 수치가 p-value가 0.01 미만인 성능 향상임을 나타낸다.)

다음으로 IV장에서 제안하였던 혼성 사전분포 추정 방식을 III장에서의 방법과 함께 사용하여 실험을 하였고, 그 결과는 다음과 같다.

먼저, 혼성 사전분포만을 적용한 경우, 즉 80개의 full 행렬 형태의 변환행렬 집단과 블록 대각 행렬 형태의 변환행렬 집단에서 각각 공분산행렬과 평균행렬을 추정하여 MAPLR 적응을 수행하였을 때, 모든 적응 데이터 구간에서 기존의 MAPLR 방식에 비해 통계적으로 유의미한 성능 개선을 보이는 것을 확인하였다. 또한 참조화자에 의한 적응적 사전분포 방식을 함께 적용할 경우 대체적으로 20명의 참조화자를 적용했을 때 성능이 가장 높은 경향을 보였다. 적응 데이터가 1단어일 때 혼성 사전분포만을 적용한 경우에 비해 부가적인 성능 개선을 보였는데 1 %의 유의수준에서 통계적으로 유의미하였으며, 적응 데이터가 3단어일 때에는 20명의 참조화자의 경우에 한해서 1 %의 유의수준에서 통계적으로 유의미한 추가 성능향상이 얻어졌다.

본 논문에서는 제안된 방식의 기본석인 성능 특성을 파악하기 위한 방편으로서 에너지 파라미터를 함께 사용하지 않은 12차 MFCC와 그 delta, delta-delta 계수들

표 4. 블록 대각 변환행렬을 이용한 MAPLR 방식에서 참조화자로부터의 적응적 혼성 사전분포를 적용했을 때의 단어인식률 (%)
Table 4. Word accuracy of MAPLR with block-diagonal transformation matrix using adaptive hybrid prior distribution from reference speakers.

Method	Ref. spk.	No. of adaptation words			
		1	3	5	10
MAPLR	-	97.09	97.41	97.54	97.86
Hybrid prior	-	97.30*	97.62*	97.83*	97.97*
Hybrid prior & adaptive prior	20	97.51*	97.74*	97.86	98.04
	30	97.56*	97.64	97.80	97.99
	40	97.47*	97.59	97.78	98.00

*p-value < 0.01.

을 이용하여 단일 mixture 가우시안 확률분포를 가지는 HMM에 대해 고립단어 인식 실험만을 수행한 결과를 나타내었다. 앞으로 보다 심층적인 성능 평가를 위해서는 확장된 특징 파라미터 및 복수 mixture의 확률모델을 이용하여 연속음성인식에 적용하는 추가 실험이 진행될 필요가 있다고 판단된다.

VI. 결 론

본 논문에서는 MAPLR 화자적응 방식에서 사용되는 사전분포를 온라인상에서 추정하거나 새로운 방식으로 추정함으로써 고속 화자적응 성능을 높이고자 하였다. 첫번째로, 참조화자로부터 추정된 적응적 사전분포를 적용하는 방법을 제안하였다. 많은 화자들의 공통적인 특성보다는 비슷한 특성을 가진 화자들의 특성으로부터 사전분포를 추정하여 새로운 화자에게 보다 더 적합한 사전분포를 적용하고자 하였다. 고립단어 인식실험 결과, 적응 데이터가 매우 적을 때의 성능은 통계적으로 유의미하게 개선되는 것을 확인하였다. 또한 블록 대각 행렬 기반의 MAPLR 방식에서 혼성 사전분포를 추정하고 적용하는 방법을 제안하였다. 훈련화자와 화자독립 모델에서 얻어진 변환행렬 집단을 두 가지 형태로 구성하여 변환행렬의 사전분포 파라미터를 각각 독립적으로 추정하여 혼합할 수 있으며, 실험을 통해 적응 데이터의 개수와 무관하게 기존의 MAPLR 방식보다 통계적으로 유의미한 성능 개선을 보이는 것을 확인하였다. 또한 참조화자에 의한 적응적 사전분포 방식과 함께 적용함으로써 적응 데이터가 매우 적을 경우 추가적인 성능 개선을 얻었다. 본 논문에서 제안한 방법은 MAPLR 화자적응 방식의 고속 화자적응 성능을 더욱 개선함으로써 매우 적은 적응 데이

터만이 주어졌을 때 유용한 화자적응 방식이 될 것으로 사료된다.

감사의 글

이 논문은 부산대학교 자유과제 학술연구비 (2년)에 의하여 연구되었으며, 지원에 감사드립니다.

참 고 문 헌

1. J. L. Gauvain and C. H. Lee "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.
2. C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 1, pp. 171–185, 1995.
3. R. Kuhn, P. Nguyen, J. C. Jungua, L. Goldwasser, N. Niedzielski, S. Finche, K. Field and M. Contolini, "Eigenvoices for speaker adaptation," in *Proc. ICSLP*, pp. 1771–1774, 1998.
4. R. Kuhn, J. C. Jungua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.
5. M. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 417–428, Jul. 2000.
6. M. J. F. Gales, *The Generation and Use of Regression Class Trees for MLLR Adaptation*, Cambridge University, Cambridge, U. K., Tech. Rep. CUED/F–INFENG/TR263, 1996.
7. W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix variate priors," in *Proc. Eurospeech*, vol. 1, pp. 1–4, 1999.
8. L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 532–535, 1989.
9. T. J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communication*, vol. 31, pp. 15–33, 2000.
10. B. Mak, T.–C. Lai, and R. Hsiao, "Improving reference speaker weighting adaptation by the use of maximum–likelihood reference speakers," in *Proc. ICASSP*, pp. 229–232, 2006.
11. C. Huang, T. Chen and E. Chang, "Transformation and combination of hidden Markov models for speaker selection training" in *Proc. ICSLP*, pp. 1001–1004, 2004.
12. Y. Lim and Y. Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," in *Proc. ICASSP*, pp.89–91, 1995.
13. 이용주, 김봉완, 김종진, 양옥렬, 임선영, "음성 DB용 PBW에 관한 검토," *제12회 음성통신 신호처리 워크샵 논문집*, pp. 310–314, 1995.
14. 송영록, 김형순, "화자적응에서의 복수의 사전분포의 유용성", 제 *27회 음성통신 및 신호처리 학술대회 논문집*, pp. 136–137, 2010.
15. 송영록, 김형순, "MAPLR 기반의 화자적응에서의 weighted prior 적용", *2010 한국음성학회 가을 학술대회 발표논문집*, pp. 136–137, 2010.

저자 약력

•송 영 록 (Young Rok Song)

2009년 2월: 부산대학교 전자전기공학부 (학사)
2011년 2월: 부산대학교 전자전기공학과 (석사)
2011년 ~ 현재: LG전자
※ 주관심 분야 : 음성인식, 화자적응

•김 형 순 (Hyung Soon Kim)

1983년 2월: 서울대학교 전자공학과 (학사)
1984년 2월: 한국과학기술원 전기및전자공학과 (박사과정 조기진학)
1989년 2월: 한국과학기술원 전기및전자공학과 (박사)
1987년 ~ 1992년: 디지콤정보통신연구소 선임연구원
1992년 ~ 현재: 부산대학교 전자전기공학부 교수
※ 주관심 분야 : 음성인식, 음성합성, 음성신호처리