법에 대하여 소개하고, 3장과 4장에서는 본 논문에서 제안한 바이어스 보상을 가진 eigenspace-based MLLR 방법과 multi-stream 기법을 도입하여 이를 일반화한 eigenspace-based MLLR 적응 방식에 대하여 소개한다. 그리고 5장에서는 잡음 환경에서 제안한 방법의 성능 평가를 실시하고, 마지막으로 6장에서 결론을 맺는다.

2. Eigenspace-based MLLR

적응 데이터가 아주 적은 경우 MLLR 방법은 그 성능을 보장하지 못하며, MLLR 방식을 고속화자 적응에서 적용하기 위해 제안된 것이 eigenspace-based MLLR이다[5]. 이는 MLLR 변환행렬을 eigenvoice framework을 이용해서 추정하는 방식이며, 적응 데이터가 아주 적은 경우에도 신뢰성 있게 변환행렬을 추정할 수 있다.

먼저 훈련 DB가 R명의 화자로 구성되어 있다면, 화자독립(speaker independent, SI) 모델을 훈련한 후 화자 각각에 대해 MLLR 화자적응을 적용하여 R개의 변환행렬을 구한다. 그 다음은 r번째 화자의 변환행렬 Wr 의 열벡터(row vector)를 서로 이어 붙여서 식 (1)과 같이 슈퍼벡터(supervector) wr를 구성한다.

$$\mathbf{W}^r = \begin{bmatrix} \mathbf{w}_1^r \\ \mathbf{w}_2^r \\ \vdots \\ \mathbf{w}_D^r \end{bmatrix} \quad \mathbf{w}^r = \begin{bmatrix} \mathbf{w}_1^{rT} \\ \mathbf{w}_2^{rT} \\ \vdots \\ \mathbf{w}_D^{rT} \end{bmatrix} \quad \begin{matrix} ,1 \le r \le R \\ 1 \le d \le D \end{matrix} \tag{1}$$

여기서 $\mathbf{w}_d^r$ 는 Wr의 d 번째 열벡터이며, T는 전치(transpose)를, D는 특징 벡터의 차원을 뜻한다.

두 번째로 R개의 슈퍼벡터에 대해 주성분 분석법(principal component analysis, PCA)를 적용시켜 R개의 eigenvector를 구한다. Eigenvoice framework과 유사하게 새로운 화자의 MLLR 변환행렬은 식 (2)와 같이 구해진 eigenvector들 중 K(<R)개의 eigenvector들의 가중 합으로 표현하며, 이 방법을 eigenspace-based MLLR(ES-MLLR)이라 한다.

$$\hat{\mathbf{w}} = \mathbf{e}(0) + \sum_{j=1}^{K} \alpha(j)\mathbf{e}(j) \tag{2}$$

여기서 e(0) 는 R 개의 슈퍼벡터들의 평균을 의미하며, e(j)는 j-번째 eigenvector를

3. 바이어스 보상을 가진 Eigenspace-based MLLR

본 논문에서는 기존의 MLLR의 단점인 아주 적은 적응 데이터를 사용하는 경우 추정된 변환행렬의 신뢰도가 떨어져 인식성능이 급격히 하락할 수 있는 문제와 ES-MLLR의 단점인 훈련과 인식 환경의 차이가 발생했을 때 이를 극복하지 못하는 문제를 동시에 해결할 수 있는 방법을 제안한다.

ES-MLLR의 경우 변환행렬에 바이어스 항($\hat{\mathbf{b}}_{TRAIN}$)이 있지만 이는 eigenvoice와 마찬가지로 훈련 환경만을 나타내어 줄 뿐이다. 따라서 인식 환경이 훈련 환경과 다른 경우에 바이어스 보상을 적용함으로써 성능 향상을 이룰 수 있다. 식(3)에 바이어스 성분을 추가하여 다음과 같이 확장할 수 있다.

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \mathbf{C}_{SI}\hat{\mathbf{w}} + \hat{\mathbf{b}}_{TEST} \\ &= \hat{\mathbf{A}}\boldsymbol{\mu}_{SI} + \hat{\mathbf{b}}_{TRAIN} + \hat{\mathbf{b}}_{TEST} \\ &= \hat{\mathbf{W}}_{ENV}\boldsymbol{\xi}_{SI}\end{aligned} \tag{6}$$

이 때 $\hat{\mathbf{W}}_{ENV}$은 훈련 및 인식 환경 전부를 포함하는 변환 행렬을 뜻한다. 식 (2)를 식 (6)에 대입하면, 환경차이에 대한 보상벡터를 포함한 ES-MLLR에 기반을 둔 적응 모델은 다음과 같이 표현될 수 있다.

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \mathbf{C}_{SI}\left\{\mathbf{e}(0) + \sum_{j=1}^{K}\alpha(j)\mathbf{e}(j)\right\} + \sum_{d=1}^{D}b(d)\mathbf{i}(d) \\ &= \tilde{\mathbf{e}}(0) + \sum_{j=1}^{K}\alpha(j)\tilde{\mathbf{e}}(j) + \sum_{d=1}^{D}b(d)\mathbf{i}(d)\end{aligned} \tag{7}$$

여기서, $\mathbf{i}(d) = [\delta(d-1),\cdots,\delta(d-D)]^T$ 이며, $\delta(x)$는 Kronecker delta 함수를 뜻한다. 식 (7)에서는 단지 eigenvoice 수가 D개만큼 증가하는 형태가 되며, 이들의 eigenvoice의 가중치와 보상벡터의 가중치는 MLED[4]를 통하여 동시에 구할 수 있게 된다.

4. Generalized eigenspace-based MLLR 기반 고속 화자 적응

기존의 eigenvoice 방법에서 적응 데이터 수가 증가하더라도 성능이 향상되지 않은 단점이 있고, 이를 해결하기 위한 방법으로 차원별 eigenvoice 방법이 제안되었으나, 추정 파라미터 수의 증가로 아주 적은 데이터의 경우 MLLR과 같이 성능이 급격히 하락하는 문제점이 있다. 이를 보완하여 적응 데이터 수에 관계없이 고

속 화자적응에서 높은 성능을 얻기 위해 차원 간에 상관성이 높은 몇 개의 sub-stream(또는 multi-stream)으로 나누어서 eigenvoice를 적응시키는 방법인 sub-stream 기반 eigenvoice 방법을 [8]에서 제안하였다. 본 논문에서도 3장에서 제안한 방법에 대해 sub-stream 기반 eigenvoice 적응 방법을 도입하여 식 (7)을 다음과 같이 일반화하였다.

$$\left\langle \hat{\boldsymbol{\mu}}_m^{(s)} \right\rangle_{C_r} = \left\langle \mathbf{e}_m^{(s)}(0) \right\rangle_{C_r} + \sum_{k=1}^{K} w^{(r)}(k) \left\langle \mathbf{e}_m^{(s)}(k) \right\rangle_{C_r} + \left[\sum_{d \in C_r} b^{(r)}(d) \left\langle \mathbf{i}(d) \right\rangle_{C_r} \right]_{\text{opt}} , 1 \le r \le N_{SS} \quad (8)$$

여기서, $\langle \hat{\boldsymbol{\mu}}_m^{(s)} \rangle_{C_r}$는 상태 s와 mixture m에서의 r번째 sub-stream의 평균 벡터를 의미하며, 식(7)에 사용된 용어들에 대한 정의는 [8]과 동일하다. Nss는 sub-stream의 총수이며, {C1, C2,⋯, Cr,⋯, CNss}는 Nss개의 sub-stream 집합을 의미하고 각각의 집합 사이에는 공통인자가 없다. 또한 D1, D2,⋯, Dr,⋯, DNss 는 각각의 sub-stream의 차원을 의미하며 특징벡터 차원(D)와 각 sub-stream 차원 사이의 관계는 다음과 같다.

$$D = \sum_{r=1}^{N_{ss}} D_r \quad (9)$$

그리고 w(r)(k), b(r)(d)는 r번째 sub-stream에 대한 k번째 eigenvoice의 가중치 및 d-차원의 바이어스 벡터의 가중치를 의미한다. 또한 마지막 항에서 "opt"는 option의 의미이다.

만약 식 (8)에서 바이어스 보상항을 고려하지 않는 경우, Nss = 1이면 일반적인 ES-MLLR가 되고, Nss = D이면 차원별 eigenvoice와 동일한 형태의 ES-MLLR이 되며, 또한 Nss가 1과 D사이에서 임의로 선택된다면 sub-stream 기반의 eigenvoice 방식과 동일한 ES-MLLR 방법이 됨을 알 수 있다. 그리고 바이어스 보상 항이 있는 경우, Nss = 1이면 바이어스 보상을 가진 ES-MLLR이 되고, 또한 Nss 수가 증가하면 sub-stream 기반의 ES-MLLR에 대해 바이어스 보상이 동시에 적용된 형태가 된다.

Sub-stream 기반 eigenvoice 적응 방법에서, 적응데이터 수가 적을 때는 sub-stream 수가 적은 경우가 성능 향상에 유리하고 적응데이터 수가 증가함에 따라 sub-stream 수도 함께 증가하는 것이 유리하다. 따라서, sub-stream 수는 적응데이터 수에 따라 자동적으로 정해지는 방법이 필요하며, 본 논문에서도 자동적으로 sub-stream을 나누기 위해 [8]에서 사용한 통계적 군집분석(clustering analysis) 방법

먼저 SI 모델을 사용한 경우 SNR이 낮아질수록 성능이 급격히 감소하는 것을 알 수 있다. 이에 대해 기존의 ES-MLLR 방법을 사용한 경우 SNR이 낮아지더라도 높은 성능 향상을 보였으며, 또한 본 논문에서 제안한 바이어스 보상을 가진 ES-MLLR의 경우 성능 향상이 기존의 방법에 비해 SNR 10dB 환경에서 적응 단어 수가 20개인 경우 33%의 인식 오류 감소율을 보였다.

본 논문에서 제안한 sub-stream 기반 ES-MLLR에서 sub-stream을 나누기 위해 [8]에서 사용한 문턱치 중 TH2를 사용하여 실험을 수행하였다. 또한 [8]과 같이 적응 데이터만을 사용하여 구한 공분산 행렬의 경우에는 신뢰성이 떨어지므로 본 논문에서는 군집화시 사용할 상관 행렬을 구할 때 MAP 적응을 사용하여 미리 구해놓은 훈련 DB의 공분산 행렬과 적응 데이터의 공분산 행렬을 가중합하는 방법을 취하였다. 그리고 통계적 군집을 위한 결합(linkage) 방법으로는 평균결합법(average linkage)을 사용하였다[8].

이상의 sub-stream 기반 eigenspace-based MLLR 실험 모두 적응 데이터 수가 1개인 경우에서는 성능을 제대로 얻지 못하였다. 따라서 적응 데이터가 1개인 경우에는 식(8)에서 Nss = 1이 되도록 sub-stream의 수에 제약조건을 부여하였다.

<표 1> ES-MLLR과 제안한 방법의 성능 비교

Adaptation Scheme	SNR	Number of Adaptation Words							
		0	1	5	10	20	30	40	50
ES-MLLR	Clean	95.78	97.28	98.00	97.93	97.98	98.08	98.10	98.08
ES-MLLR &EC			97.63	98.45	98.68	98.58	98.68	98.60	98.58
MES-MLLR			97.28	98.45	98.88	98.88	98.83	98.83	98.88
MES-MLLR &EC			97.63	98.43	98.68	98.88	98.88	98.83	98.80
ES-MLLR	20dB	93.23	96.70	97.53	97.60	97.73	97.78	97.83	97.73
ES-MLLR &EC			96.68	97.98	98.10	98.18	98.20	98.25	98.23
MES-MLLR			96.70	98.05	98.38	98.75	98.70	98.50	98.58
MES-MLLR &EC			96.68	97.93	98.63	98.78	98.73	98.58	98.73
ES-MLLR	10dB	80.18	92.50	93.75	94.05	94.25	94.35	94.38	94.30
ES-MLLR &EC			93.30	95.65	95.73	96.13	95.88	96.03	95.90
MES-MLLR			92.50	96.70	97.78	97.93	98.10	97.98	98.18
MES-MLLR &EC			93.30	96.85	98.05	97.80	97.98	97.88	98.05

<표 1>에서 보는 바와 같이, sub-stream ES-MLLR과 바이어스 보상 방법을 함께 사용함으로써, SNR 10dB 환경에서 ES-MLLR에 비해 적응 단어수가 10개인 경우 67%의 인식 오류 감소율을 보였으며, 또한 바이어스 보상을 가진 ES-MLLR에

비해서도 54%의 인식 오류 감소율을 보여 잡음 환경에서도 상당한 성능 향상을 얻었다. 따라서, 제안된 바이어스 보상을 가진 sub-stream 기반 ES-MLLR 방법이 잡음 환경에서도 효과적으로 적용 가능한 고속 화자적응 방법임을 알 수 있다.

6. 결 론

본 논문에서는 기존의 eigenspace-based MLLR 화자 적응 방법이 훈련 및 인식 환경에 차이가 나는 경우 그 차이를 보상하지 못하는 단점을 해결하기 위해, 바이어스 보상 방법을 도입하여 잡음 환경에서 화자 및 환경 적응을 동시에 적용하였다. 특히 적응 데이터가 아주 적은 경우에도 높은 성능향상을 얻을 수 있었으며, 바이어스 보상을 동시에 적용한 eigenspace-based MLLR 적응 방식을 사용하여 기존 eigenspace-based MLLR 적응방식에 비하여 적응 데이터를 50개까지 사용했을 경우 18~33%의 인식 오류 감소율을 얻을 수 있었다. 또한 본 논문에서는 sub-stream 기반의 eigenvoice 방법을 제안한 바이어스 보상을 가진 eigenspace-based MLLR 방법과 통합시켜 깨끗한 환경뿐만 아니라 잡음 환경에서도 추가적인 성능 향상을 얻었으며, 적응 데이터를 50개까지 사용했을 경우 50~67%의 인식 오류 감소율을 얻었다.

참 고 문 헌

[1] A. Sanker, "A maximum-likelihood approach to stochastic matching for robust speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 3, pp. 190 -202, 1996.

[2] C. H. Lee, C. H. Lin and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models", *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 806-814, 1991.

[3] 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.

[4] 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*, vol. 5, pp. 1771-1774, 1998.

[5] K. T. Chen, W. W. Liau, H. M. Wang and L. S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression", in *Proc. ICSLP*, Beijing, pp. 742-745. 2000.

[6] J. S. Park, H. J. Song and H. S. Kim, "Performance improvement of rapid speaker

adaptation based on eigenvoice and bias compensation", in *Proc. Eurospeech*, 2003.

[7] H. J. Song and H. S. Kim, "Simultaneous estimation of weights of eigenvoices and bias compensation vector for rapid speaker adaptation", in *Proc. ICSLP*, Jeju, 2004.

[8] 송화전, 이종석, 김형순, "Sub-stream 기반의 eigenvoice를 이용한 고속 화자적응", *말소리*, 제 55호, pp. 93-102, 2005.

[9] J. T. Chien, "Quasi-Bayes linear regression for sequential learning of hidden Markov model", *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 268-278, 2002.

[10] K. T. Chen, H. M. Wang, "Eigenspace-based maximum a posteriori linear regression for rapid speaker adaptation", in *Proc. ICASSP*, Salt Lake City, vol. 1, pp. 317-320, 2001.

[11] Y. Lim and Y. Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database", In *Proc. ICASSP*, vol. 1, pp. 89-91, 1995.

[12] 이용주, 김봉완 외, "음성 DB용 PBW에 관한 검토", *제 12회 음성통신 신호처리 워크샵 논문집*, pp. 310-314, 1995.

[13] ITU recommendation p.56, "Objective measurement of active speech level", 1993.

접수일자: 2006년 4월 11일
게재결정: 2006년 4월 24일

▶ 송화전(Hwa Jeon Song) : 교신저자
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 컴퓨터 및 정보통신연구소
소속: 부산대학교 컴퓨터 및 정보통신연구소
전화: 051) 510-1704
E-mail: hwajeon@pusan.ac.kr

▶ 김형순(Hyung Soon Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
E-mail: kimhs@pusan.ac.kr