한국음향학회지 제22권 제1호 pp. 48~53 (2003)

차원별 Eigenvoice와 화자적응 모드 선택에 기반한 고속화자적응 성능 향상

Performance Improvement of Fast Speaker Adaptation Based on Dimensional Eigenvoice and Adaptation Mode Selection

송 화 전*, 이 윤 근**, 김 형 순*
(Hwa-Jeon Song*, Yun-Keun Lee**, Hyung-Soon Kim*)

*부산대학교 전자공학과, **(주)보이스웨어

(접수일자: 2002년 12월 16일; 채택일자: 2003년 1월 3일)

Eigenvoice 방법은 고속화자적응에 적합하다고 알려져 있지만, 이 방법은 발화수가 증가하더라도 추가적인 인식성능 향상이 이루어지지 않는 단점이 있다. 본 논문에서는 이 문제를 해결하기 위해 음성 특징벡터의 차원별로 eigenvoice의 가중치를 구하여 적응시키는 방법과 또한 적응 데이터 수에 따라 높은 인식률을 얻는 적응 방식을 선택하는 방식을 제안한다. 화자독립모델 및 eigenvoice들을 구성하기 위해 POW (Phonetically Optimized Words)데이터베이스를 사용하였으며, PBW (Phonetically Balanced Words) 452 단어 중 50개까지 발화 수를 변화시키면서 교사방식 (Supervised mode)로 적응에 사용하고 나머지 중 400개를 인식실험에 사용하였다. 차원별 eigenvoice 방법이 발화수가 증가함에 따라 기존의 eigenvoice 나 MLLR 방법보다 높은 성능을 보였으며, eigenvoice와 차원별 eigenvoice방법 사이의 적응 모드 선택을 통해 기존의 eigenvoice 방식에 비해 최고 26%의 단어 오인식률 감소를 얻었다.

핵심용어: 화자적응, 고속화자적응, Eigenvoice, MLLR, 음성인식
투고분야: 음성처리 분야 (2,5)

Eigenvoice method is known to be adequate for fast speaker adaptation, but it hardly shows additional improvement with increased amount of adaptation data. In this paper, to deal with this problem, we propose a modified method estimating the weights of eigenvoices in each feature vector dimension. We also propose an adaptation mode selection scheme that one method with higher performance among several adaptation methods is selected according to the amount of adaptation data. We used POW DB to construct the speaker independent model and eigenvoices, and utterances (ranging from 1 to 50) from PBW 452 DB and the remaining 400 utterances were used for adaptation and evaluation, respectively. With the increased amount of adaptation data, proposed dimensional eigenvoice method showed higher performance than both conventional eigenvoice method and MLLR. Up to 26% of word error rate was reduced by the adaptation mode selection between eigenvoice and dimensional eigenvoice methods in comparison with conventional eigenvoice method.

Keywords: *Speaker adaptation, Fast speaker adaptation, Eigenvoice, MLLR, Speech recognition*
ASK subject classification: *Speech signal processing (2,5)*

I. 서 론

사용자가 불편을 느끼지 않고 발성할 수 있는 최소한의 적응 데이터만으로 인식성능을 개선할 수 있는 가변어휘 인식시스템을 개발하기 위해 고속화자적응 기술은 필수적이다. 현재 화자적응 기법으로 사용되는 방법들은 새로운 화자에 대한 모델 파라메터를 어떻게 추정하느냐에 따라 크게 Maximum A Posteriori (MAP) 계열[1], Maximum Likelihood Linear Regression (MLLR) 계열[2], 화자 군집화 계열 등으로 구분할 수 있다. 이 중 고속화자 적응에 유리한 화자 군집화 계열의 한가지 방법인 eigenvoice 적응 방법[3]이 최근 개발되었으며, 이를 기반으로 성능 향상을 위해 여러 가지 방법들이 제안되고 있다[4-6].

책임저자: 송화전 (hwajeon@pusan.ac.kr)
609-735 부산시 금정구 장전동 산 30
부산대학교 전자공학과 음성통신연구실
(전화: 051-510-1704; 팩스: 051-515-5190)

Eigenvoice 방법의 단점은 발화수가 증가하더라도 다른 적응 방법들에 비해 추가적인 인식 성능 향상이 이루어지지 않는다는점이다. 이를 해결하는 방법으로 eigenvoice 적용 후에 MAP 적응 방식을 적용하는 경우가 있다[3].

본 논문에서도 eigenvoice 방법을 기본으로 하여 그 성능을 향상시키고자 음성 특징벡터 차원별로 eigenvoice의 가중치를 추정하는 방법을 제안하였으며, 또한 적응 모드 선택을 통하여 발화 수에 따라 가장 좋은 성능을 가지는 적응방법이 선택되도록 하였다.

본 논문의 구성은 다음과 같다. 2절에서는 기존 eigenvoice 적응 방법에 대해 간략히 살펴본 후, 3절에서 발화수 증가에 따라 성능 향상을 위해 음성 특징벡터 차원별로 가중치 추정에 대해, 4절에서 기존 시스템과 여러 적응 방법들에 대한 실험 및 결과를 기술하였다. 그리고 5절에서 적응 모드 선택 방법을 제안하였고 이에 대한 실험결과를 기술하였으며, 마지막으로 6절에서 결론을 맺는다.

II. Eigenvoice 적응 방법[3]

우선, T개의 잘 훈련된 화자 종속 (speaker-dependent, SD) 모델 파라메터들을 각각 차원 D의 벡터로 구성한다. 이러한 벡터를 "수퍼 벡터 (supervector)"라고 하며, HMM 파라메터가 수퍼 벡터에 저장되는 순서는 상관없지만, T개의 수퍼 벡터가 저장되는 순서는 동일해야 한다. 그리고 수퍼 벡터들의 평균을 구하고 나서 각 수퍼 벡터에서 평균을 차감한 후 주성분 분석 (PCA: Principal Component Analysis)를 적용해서 차원 D를가지는 T개의 "eigenvoice"를 얻을 수 있다. 최초 몇 개의 eigenvoice들이 주어진 데이터가 가진 변동의 대부분을 설명하기 때문에 T개의 eigenvoice 중 최초의 K개, 즉, $e(1),\ldots, e(K)$만으로 전체 변동을 대표할 수 있다 ($K < T \ll D$). 이와 같이 선택된 K개의 eigenvoice는 K-space를 생성한다. 적응 데이터가 주어지면 새로운 화자는 다음 식과 같이 K개의 eigenvoice로 나타낼 수 있다.

$$s = e(0) \times \sum_{k=1}^{K} w(k)\, e(k) \qquad (1)$$

여기서 $e(0)$는 수퍼 벡터 (supervector)들의 평균을 나타낸다. 그리고, 가중치 $w(k)$는 최대 우도 고유치 분석 (MLED: Maximum Likelihood Eigen Decomposition) 방법[3]을 통해 구한다.

III. 차원별 Eigenvoice

Eigenvoice 적응 방법은 아주 적은 적응 데이터 수에서 기존 시스템이나 MLLR 적응방법보다 좋은 성능을 보이지만 반면에 적응 데이터 수가 증가하더라도 성능이 추가적으로 향상되지 않는 단점이 있다. 이는 $e(0)$를 새로운 화자의 모델로 적응시키기 위해 단지 K개 eigenvoice의 global 가중치만 구하여 모든 음소와 모든 차원에 동일하게 식 (1)과 같이 적용하므로 적응 데이터 수가 많아지더라도 각 음소나 차원별로 세밀한 변화를 반영하지 못하기 때문이다. 본 논문에서는 기존의 eigenvoice 방법의 단점을 보완하고자 음성 특징벡터 각 차원별로 MLED 방법을 통해 eigenvoice의 가중치를 추정하는 방법을 제안하였다. 이는 MLLR 방법에서 새로운 화자의 평균 파라메터를 구하기 위해 일반적으로 SI 모델의 평균 파라메터의 차원별 선형 결합 (linear combination)으로 표현하는 가중치 행렬을 구하는 것과 유사하다. 이와 같은 각 차원별 가중치를 적용함으로써 발화수가 증가하는 경우에 고유 벡터 공간 (eigen space) 상에서 적응 화자의 위치를 좀더 세밀하게 표현할 수 있다.

IV. 실험 및 결과

4.1. 실험환경

본 논문에서는 고속 화자적응 방식을 가변어휘 고립단어 인식 실험에 적용하였다. 본 논문에서 사용한 음성 특징 파라메터로는 20 ms 해밍 창을 10 ms씩 이동시키면서 12차 MFCC 및 1, 2차 마분치를 구하여 총 36차의 파라메터를 사용하였다. 그리고, 46 유사음소 (PLU) set[7]을 기본으로 결정트리 군집화 (TBC: tree-based clustering)를 사용한 트라이폰 (triphone)을 기본 모델로 사용하였으며 모델 당 상태수는 3개로 정하였다.

가변어휘 음성 인식기의 훈련을 위해서는 음운 현상이 다양하게 포함된 데이터 베이스를 사용하여야 우수한 성능을 얻을 수 있다. 본 실험에서는 훈련을 위하여 POW 음성 데이터베이스[8] 중에서 남성 40명분의 음성 데이터베이스를 이용해서 모델을 훈련시켰다.

그리고 화자적응 및 인식 실험을 위해서는 훈련용 POW 음성 데이터베이스와는 어휘 내용이 다른 PBW 452 데이터베이스[9]의 일부를 사용하였다. 남성화자 10명의 1회 발성분 (test set A)에 대해서 처음 50개까지 단어 수를

늘려가면서 적응에 사용하였고, 나머지 중 400개 단어를 성능 평가에 사용하였다. 그리고 본 논문에서는 SI 모델에 대해 가장 인식성능이 낮은 10명의 화자 (test set B)들에 대해서도 성능 평가를 실시하였다.

Eigenvoice를 생성시키기 위해 먼저 POW 데이터베이스를 사용하여 SI 모델을 구성한 후 40명의 각각의 화자에 대해 MAP 적응 방식을 사용하여 40개의 SD 모델을 구성하였다. 각각을 수퍼 벡터로 만들고 화자 평균으로 차감한 후 PCA를 통하여 40개의 eigenvoice를 구성하였다. 본 논문에서 사용한 tied state 수는 4050개이며, 수퍼 벡터의 총 차원은 믹스쳐 (mixture)가 4개인 경우 D = 583200 (=4050*4*36)이다.

4.2. 실험결과 및 검토

먼저 상태당 믹스쳐가 1개인 SI 모델에 의한 기준 시스템의 결과와 기존의 적응 방법인 MLLR 및 MAP와 고속 적응 방법의 한가지인 eigenvoice 사이의 인식결과가 그림 1에 나타나 있다. MAP 방법은 적응 데이터 크기가 적절하지 않으면 오히려 인식성능의 하락이 발생됨을 알 수 있고, MLLR의 경우도 아주 적은 데이터 (10개 미만의 발화수)의 경우 그 성능을 보장하지 못한다. Eigenvoice의 경우는 적응 데이터수가 매우 적더라도 기존의 방식보다 더 좋은 성능을 보여준다. 그러나 데이터 수가 증가하더라도 추가적인 인식성능 향상이 이루어지지 않는 단점을 확인할 수 있다. 본 논문에서는 40명의 훈련화자로부터 5개 또는 30개의 eigenvoice를 구하여 사용하였다.

그리고 본 논문에서 제안한 차원별 eigenvoice의 가중치를 추정하는 방법 (그림 1에서 EV_DIM)을 적용함에 의해 적응 데이터가 증가함에 따라 MLLR이나 기존의 eigenvoice 방법에 비해 성능 향상이 이루어지는 것을 알 수 있다. 그러나 MLLR과 마찬가지로 적응 데이터 수가

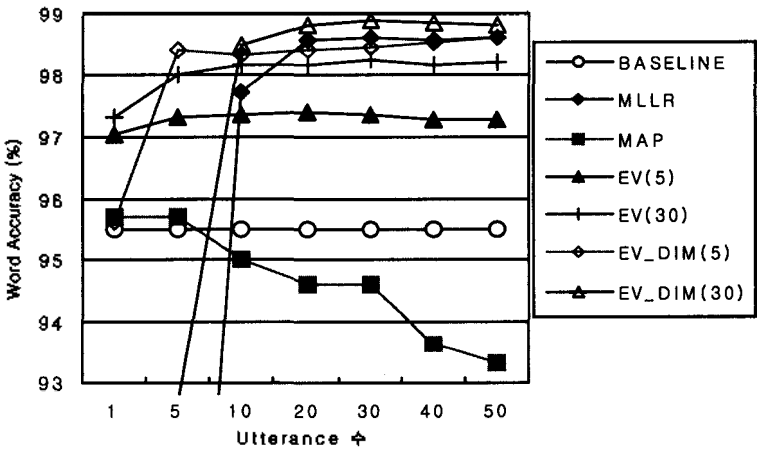

그림 1. 여러 가지 적응 방법들의 성능 비교
Fig. 1. Performance comparison of various speaker adaptation methods.

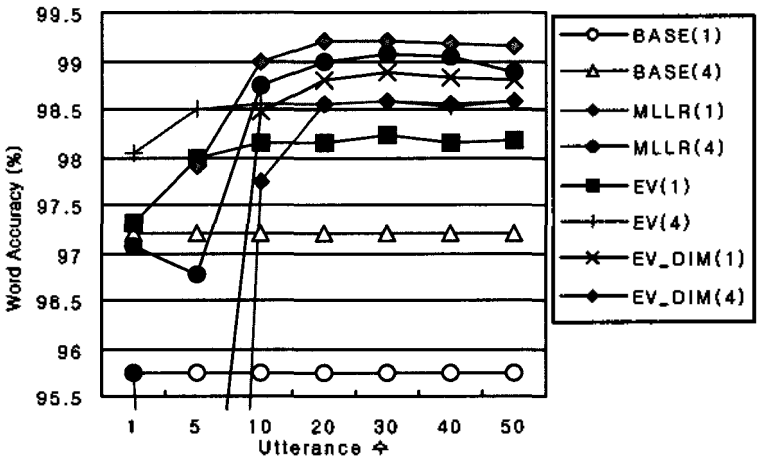

그림 2. 믹스쳐 수 변화에 따른 성능 비교
Fig. 2. Performance comparison according to the number of mixtures.

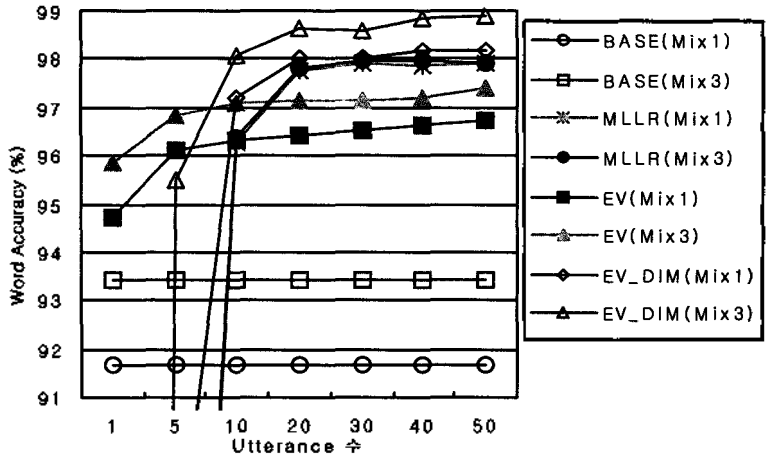

그림 3. 화자 독립 인식률이 가장 낮은 10명의 화자에 대한 성능 비교
Fig. 3. Performance comparison for 10 speakers that have the lowest SI performance.

개우 적을 경우 추정할 가중치 수가 많아질수록 제대로 로델이 적응되지 않는 경우가 발생하였다.

그림 2는 믹스쳐 증가에 따른 인식 결과이며, 1개와 4개일 때의 결과를 나타내고 있다. 그림 2에서도 eigenvoice가 발화수가 적은 경우 MLLR보다 성능이 높음을 알 수 있으며, 발화수가 증가함에 따라 차원별 eigenvoice를 사용한 경우가 가장 높은 성능을 나타내었다. 여기서는 30개의 eigenvoice를 사용하였다.

그림 3에는 PBW 데이터베이스 중 가장 인식성능이 낮은 10명에 대해 적응 및 인식실험을 실시하였다. 그림 3에서도 앞선 실험과 동일한 결과를 보여주고 있으며, 특히 믹스쳐가 3개인 경우 MLLR에 비해 높은 성능 향상률을 보인다. 이 실험에서도 eigenvoice를 사용하였다.

V. 적응 모드 선택

Eigenvoice는 발화수가 적은 경우에 인식성능을 향상시키는데 유리하고 MLLR이나 차원별 eigenvoice는 발화수가 증가함에 따라 유리하다는 것을 4절에서 기술하였다. 이런 특성을 이용하여 적응 데이터 수에 따라 높은 인식률을 얻는 적응 방식을 선택하도록 하여 발화수가 적은 경우나 증가하는 경우에 인식성능이 저하되지 않도록 하였다.

5.1. Eigenvoice와 MLLR

본 논문에서는 eigenvoice와 MLLR을 사용한 경우의 장점을 살리고자 eigenvoice나 MLLR 적응식 유도시 사용하는 Q-함수를 이용하여 적응 모드 선택 방식을 제안하였다. Eigenvoice와 MLLR을 사용하여 적응된 평균에 대해 Q-함수 값을 구하여 발화수로 정규화시킨 후의 값

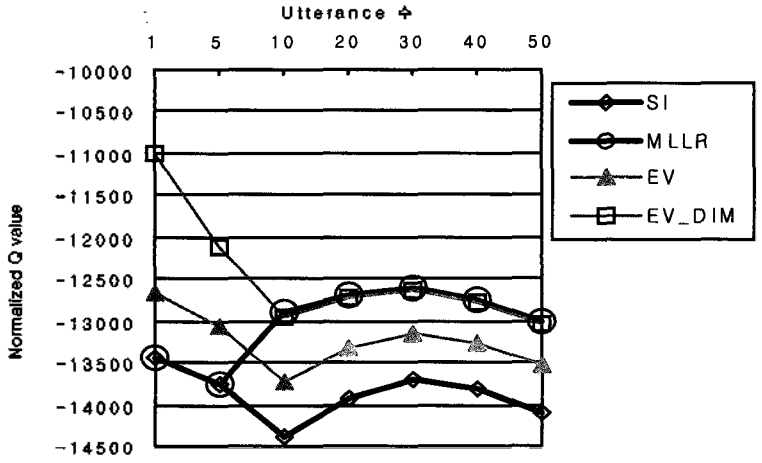

그림 4. 발화 수에 따른 각각의 화자적응 방식의 정규화된 Q-함수값
Fig. 4. Normalized Q-function value of speaker adaptation method according to the number of utterances.

을 그림 4에 나타내었다. 그림 4에 나타나 듯이 10개 미만의 경우는 eigenvoice의 Q-함수 값이 MLLR의 경우보다 높음을 알 수 있다.

이를 이용하여 다음과 같이 파라메터 평균을 적응시킬 수 있다.

$$\mu = \alpha \cdot \mu_{EV} + (1-\alpha) \cdot \mu_{MLLR} \tag{2}$$

여기서

$$\alpha = f(Q_{EV} - Q_{MLLR}) \tag{3}$$

이고, 이 때 $f(\cdot)$는

$$f(x) = sigmoid(x) = 1/(1+\exp(-\gamma x+\theta)) \tag{4}$$

또는

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & otherwise \end{cases} \tag{5}$$

의 형태를 사용할 수 있다. 식 (4)에서 γ는 *sigmoid* 함수의 기울기이고, θ는 축 방향으로의 *bias*이다.

5.2. Eigenvoice와 차원별 Eigenvoice

그림 4에서 나타나듯이 차원별 eigenvoice를 사용한 경우는 발화수가 적을 때 추정한 파라메터의 Q-함수 값이 기존의 eigenvoice 방법보다 훨씬 높으므로 5.1절과 동일한 방법을 사용할 수 없다. 이 문제를 해결하기 위해 먼저 발화수에 따라 기존의 eigenvoice 가중치 추정방법과 차원별 eigenvoice 가중치 추정방법으로 구한 가중치들의 크기를 조사하여 그림 5에 나타내었다. 발화수가 적은 경우는 식 (6)과 같은 관계가 나타난다.

$$\sum_{k=1}^{K} \sum_{d=1}^{D_f} \left| w_{\dim}^{(d)}(k) \right|^2 \gg \sum_{k=1}^{K} \left| w_{global}(k) \right|^2 \tag{6}$$

여기서 D_f는 음성특징 벡터의 차원을, K는 eigenvoice 수, $w_{\dim}^{(d)}$는 차원별 eigenvoice 가중치를, w_{global}은 기존의 eigenvoice 가중치를 나타낸다. 이는 아주 적은 데이터를 사용하여 신뢰할 수 있는 많은 수의 파라메터들을 추정하는 것이 어려운 작업임을 나타낸다. 추정된 가중치는 eigen space 상에서 적응화자가 화자들의 평균인 $e(0)$로부터 얼마만큼 떨어져 있는지에 대한 척도가 된다. 따라서 고유 벡터 공간상에서 적응화자의 위치를 표현하기 위해 w_{global}을 사용하는 경우 발화수에 따른 적응화자의 위치 변화가 적으므로 인식성능이 크게 차이를 나타내지 않지만, $w_{\dim}^{(d)}$을 사용한 경우는 발화수가 어느 정도 증가해야 적응 화자의 위치 이동변화가 적어지므로 추정된 가중치를 신뢰할 수 있게 된다. 본 논문에서는 실험적으로 식(7)과 같이 가중치를 선택하는 방법을 제안하였다.

$$w = \begin{cases} w_{global} & if \quad \frac{1}{D_f} \sum_{d=1}^{D_f} \sum_{k=1}^{K} \left| w_{\dim}^{(d)}(k) \right|^2 > \sum_{k=1}^{K} \left| w_{global}(k) \right|^2 + TH \\ w_{\dim} & otherwise \end{cases} \tag{7}$$

여기서 TH는 적절한 문턱치 값을 나타낸다. 이 방법은 발화수가 적은 경우에는 w_{global}을 사용하도록 하여 구해진 적응화자의 모델 파라메터의 신뢰도를 높이고, 발화수가 많아지는 경우는 $w_{\dim}^{(d)}$을 사용하여 MLLR을 사용한 경우보다 인식 성능 향상을 가져올 수 있다.

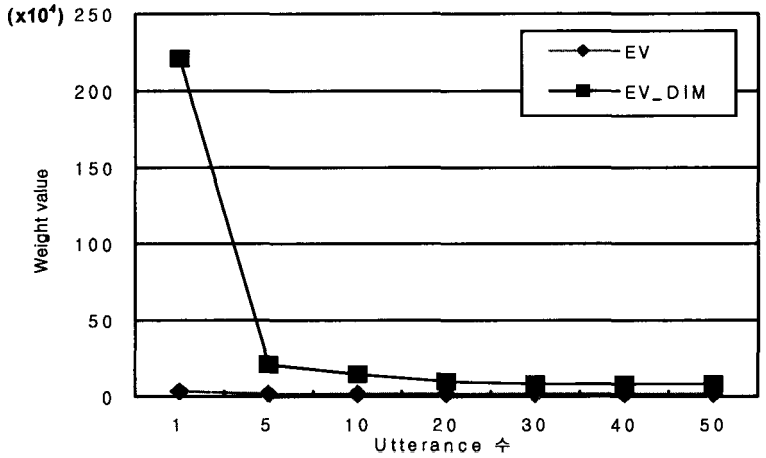

그림 5. 발화 수에 따른 eigenvoice와 차원별 eigenvoice방법의 가중치 크기 비교
Fig. 5. Comparison of magnitude of weights of eigenvoice and dimensional eigenvoice according to the number of utterances.

5.3. 적응 모드 선택에 대한 실험 및 결과

테스트 set A에 대해 적응 모드 선택에 의한 인식 결과를 서술하면, 그림 1에서 먼저 eigenvoice/MLLR 모드 선택의 경우는 식 (4), (5)를 사용한 경우에는 발화수 10개에서 좋은 성능을 보인 MLLR을 취하지 못하였으나, 식 (7)을 사용한 경우인 eigenvoice/차원별 eigenvoice 모드 선택에서는 모든 발화수에서 가장 좋은 성능만을 나타내었다. 이를 통해 테스트 set A에 대해서는 기존의 eigenvoice 방식에 비해 단어 오인식률이 21% 감소하였으며, 테스트 set B에 대해서는 26%의 단어 오인식률 감소를 얻었다.

VI. 결 론

사용자가 불편을 느끼지 않고 발성할 수 있는 최소한의 적응 데이터만으로 인식성능을 개선할 수 있는 가변어휘 인식시스템을 개발하기 위해 여러 가지 화자적응 방법을 사용하여 성능을 비교하였다. 본 논문에서는 10단어 미만의 발성에 대해 가장 좋은 성능을 보인 것은 eigenvoice를 이용한 화자 적응 방법이며, 또한 이 방법의 단점을 개선하기 위해 발화수가 증가하는 경우에는 음성특징 벡터 차원별로 eigenvoice 가중치를 추정함으로써 기존의 방법들 보다 높은 인식성능을 얻었다. 그리고 적응 데이터 수에 따라 eigenvoice와 MLLR 또는 차원별 eigenvoice 방법 중 높은 인식률을 얻는 적응 방식을 선택하도록 하여 발화수가 적은 경우나 증가하는 경우 모두에 대해 인식성능이 저하되지 않도록 하였으며, 이를 통해 기존의 eigenvoice 방식에 비해 최고 26%의 단어 오인식률 감소를 얻었다.

참 고 문 헌

1. 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*, 39 (4), 806-814, April 1991.
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*, 9 (2), 171-185, September 1995.
3. R. Kuhn, P. Nguyen, J. C. Jungua, L. Goldwasser, N. Niedzielski, S. Finche, K. Field and M. Contolini, "Eigenvoices for speaker adaptation," *Proc. ICSLP*, 5, 1771-1774, 1998.
4. H. Botterweck, "Very fast adaptation for large vocabulary continuous speech recognition using eigenvoices," *Proc. ICSLP*, 4, 354-357, 2000.
5. H. Botterweck, "Anisotropic MAP defined by eigenvoices for large vocabulary continuous speech recognition," *Proc. ICASSP*, 1, 353-356, 2001.
6. R. Kuhn, F. Perronnin, P. Nguyen, J. C. Junqua and L. Rigazio, "Very fast adaptation with a compact context-dependent eigenvoice model," *Proc. ICASSP*, 1, 373-376, 2001.
7. 유재원, 연속음성인식을 위한 음성 단위 발음사전 구성방법 연구, 위탁과제 최종연구보고서, 한국전자통신연구소, 1995.
8. Yeonja Lim and Youngjik Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," *Proc. ICASSP'95*, 1, 89-91, 1995.
9. 김봉완, 김중진, 김선태, 김태환, 김영일, 이용주, "공동이용을 위한 단어음성 DB의 구축 및 PBS 설계에 관한 검토," 제 13회 음성통신 신호처리 워크샵 논문집, 256 -261, Aug, 1996.

저자 약력

● 송 화 전 (Hwa-Jeon Song)

1993년 2월: 부산대학교 공과대학 전자공학과 (공학사)
1995년 2월: 부산대학교 대학원 전자공학과 (공학석사)
1995년 2월~2001년 1월: 현대자동차(주) 근무
2001년 3월~현재: 부산대학교 대학원 전자공학과 (박사과정)
※ 주관심분야: 음성인식, 음성신호처리

● 이 윤 근 (Yun-Keun Lee)

1982년 3월~1986년 2월: 서울대학교 공과대학 제어계측공학과 졸업 (공학사)
1986년 3월~1988년 2월: 한국과학기술원 전기및전자공학과 졸업 (공학석사)
1994년 3월~1998년 8월: 한국과학기술원 정보및통신공학과 졸업 (공학박사)
1986년 1월~2000년 3월: LG 전자 기술원 책임연구원
2000년 4월~현재: (주)보이스웨어 연구소장

● 김 형 순 (Hyung-Soon Kim)

1983년 2월: 서울대학교 전자공학과 (공학사)
1984년 2월: 한국과학기술원 전기 및 전자공학과 (박사과정 조기진학)
1989년 2월: 한국과학기술원 전기 및 전자공학과 (공학박사)
1987년 1월~1992년 6월: 디지콤 정보통신연구소 연구부장
1992년 7월~현재: 부산대학교 전자공학과 부교수