한국음향학회지 제22권 제5호 pp. 403~407 (2003)

가중치 분포 특성을 이용한 Eigenvoice 기반 고속화자적응

Rapid Speaker Adaptation Based on Eigenvoice Using Weight Distribution Characteristics

박 종 세*, 송 화 전*, 김 형 순*
(Jong-Se Park*, Hwa-Jeon Song*, Hyung-Soon Kim*)

*부산대학교 전자공학과

(접수일자: 2003년 3월 5일; 수정일자: 2003년 5월 12일; 채택일자: 2003년 6월 30일)

최근 고속화자적응 기법으로 eigenvoice 방식이 많이 사용되고 있다. Eigenvoice 적응 방식에서도 적응화자의 적응 데이터가 매우 적은 경우에는 적절한 가중치의 추정이 어렵기 때문에 적응 데이터가 어느 정도 많은 경우에 비해 인식성능 향상이 크지 않다. 본 논문에서는 적응 데이터가 적을 때의 성능향상을 위하여 eigenvoice의 가중치 분포 특성을 이용한 eigenvoice 기반 고속화자적응을 제안한다. PBW 452 데이터베이스를 사용한 어휘독립 단어인식 실험 결과에서 가중치 문턱치 (threshold) 적용 방식을 사용하여 적응 데이터가 매우 적은 경우의 상대적인 성능 저조 문제를 완화시켰다. 적응단어를 단 1개만 사용한 경우 가중치 문턱치 적용 방식을 사용하여 단어 오인식률을 9-18% 정도 감소시켰다.

핵심용어: 음성인식, 화자적응, 고속화자적응, Eigenvoice
투고분야: 음성처리 분야 (2.5)

Recently, eigenvoice approach has been widely used for rapid speaker adaptation. However, even in the eigenvoice approach, performance improvement using very small amount of adaptation data is relatively small in comparison with that using somewhat large adaptation data because the reliable estimation of weights of eigenvoice is difficult. In this paper, we propose a rapid speaker adaptation method based on eigenvoice using the weight distribution characteristics to improve the performance on a small adaptation data. In the Experimental results on vocabulary-independent word recognition task (using PBW 452 database), the weight threshold method alleviates the problem of relatively low performance for a tiny small adaptation data. When single adaptation word is used, word error rate is reduced about 9-18% by the weight threshold method.

Keywords: *Speech recognition, Speaker adaptation, Rapid speaker adaptation, Eigenvoice*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

음성인식에 있어서 화자적응을 통하여 사용자의 특성을 더욱 잘 반영한 모델을 구성함으로써 인식성능 향상을 얻을 수 있다. 화자적응 방식은 크게 최대사후확률 (MAP: Maximum A Posteriori)[1], 최대 우도 선형회귀 (MLLR: Maximum Likelihood Linear Regression)[2], 그리고 화자 군집화 (speaker clus-tering) 방식 등이 있다. 그 중에서 화자 군집화 방식의 하나인 eigenvoice 기법 [3]이 고속화자적응에 유리한 것으로 알려져 있다.

Eigenvoice 기반 화자적응 방식에서는 새로운 화자의 적응데이터를 이용하여 각 차수별 eigenvoice의 가중치를 추정하고, 각 eigenvoice의 가중 합으로 적응모델을 구성한다. 이 방식에서도 발화수가 매우 적은 경우에는 가중치의 추정이 어려워 상대적으로 인식성능이 저하된다. 본 논문에서는 eigenvoice 가중치의 분포 특성을 이용하여 발화수가 적을 때의 인식성능 저하를 보완하였다.

책임저자: 송화전 (hwajeon@pusan.ac.kr)
609-735 부산시 금정구 장전동 산30
부산대학교 전자공학과 음성통신연구실
전화: 051-510-1704; 팩스: 051-515-5190)

본 논문의 구성은 다음과 같다. 2장에서는 가중치 분포 특성을 이용한 eigenvoice 기반 고속화자적응 방식에 대하여 소개하고, 3장에서는 실험 및 성능평가를 정리하고, 4장에서 결론을 맺는다.

II. 가중치 분포 특성을 이용한 Eigenvoice 기반 고속화자적응

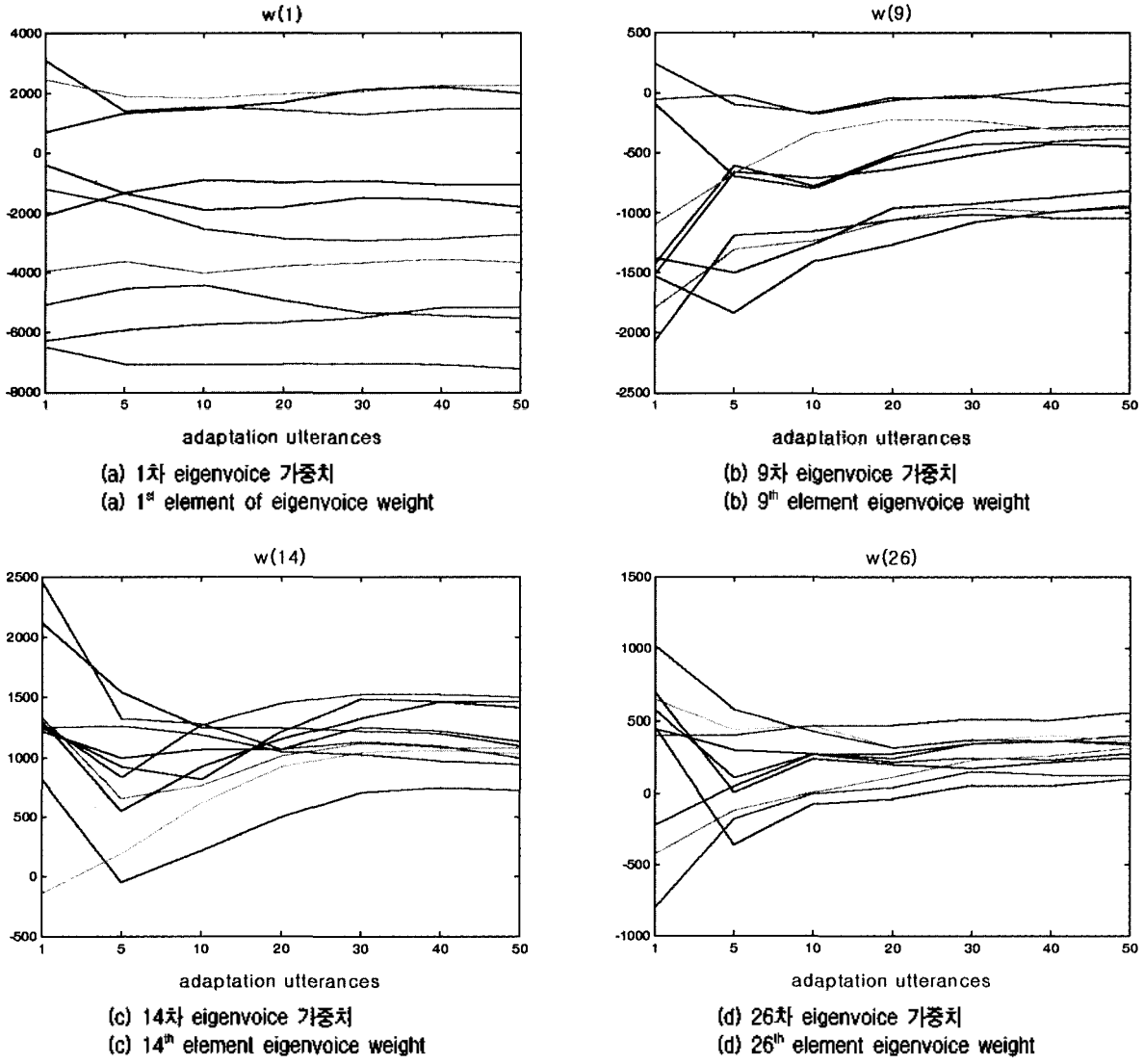
2.1. Eigenvoice 가중치의 분포 특성

Eigenvoice 기반 화자적응 방식에서 새로운 화자의 모델은 식 (1)과 같이 K개의 eigenvoice의 가중합으로 나타낼 수 있다.

$$\hat{\mu} = e(0) + \sum_{k=1}^{K} w(k)e(k) \tag{1}$$

여기서 $e(0)$는 T명의 화자종속 (SD) 모델의 평균이고, $w(k)$는 k차 eigenvoice인 $e(k)$에 대한 가중치이다. 가중치, $w(k)$는 새로운 화자의 적응 데이터로부터 최대 우도 고유치분석 (MLED: Maximum Likelihood Eigen Decomposition)[3]방법을 통해 추정된다.

그림 1은 10명의 화자 각각에 대하여 적응 데이터가 증가함에 따라 추정된 eigenvoice 가중치 중에서 $w(1)$, $w(9)$, $w(14)$, 그리고 $w(26)$을 나타낸 것이다. 그림 1에서 w(1)과 같이 eigenvoice 차수가 낮을 때에는 가중치의 범위가 크고 화자간의 차이가 많이 나타나며, 적응 데이터가 증가하더라도 가중치의 변화가 크게 나타나지 않는다.



(a) 1차 eigenvoice 가중치
(a) 1st element of eigenvoice weight

(b) 9차 eigenvoice 가중치
(b) 9th element eigenvoice weight

(c) 14차 eigenvoice 가중치
(c) 14th element eigenvoice weight

(d) 26차 eigenvoice 가중치
(d) 26th element eigenvoice weight

그림 1. 적응단어 수에 따른 eigenvoice 가중치 분포
Fig. 1. Eigenvoice weights distribution as the number of adaptation word.

이는 주성분 분석법 (PCA: Principal Component Analysis)을 이용하여 eigenvoice를 구성하므로 저차 eigenvoice가 화자들 사이의 변이 특성을 잘 표현하기 때문이다. 반면에 그림 1에서 $w(14)$ 또는 $w(26)$과 같이 eigenvoice의 차수가 클 때에는 적응 데이터가 적은 경우의 화자별 가중치가 많이 퍼져 있으나, 적응 데이터가 증가하면 화자간 가중치 분포가 밀집되어 있음을 볼 수 있다. 이와 같이 eigenvoice 차수가 증가할수록 적응데이터의 양에 따른 가중치 분포의 수렴 속도가 빨라지는 경향을 가진다.

그림 1에서 적응 데이터가 적은 경우에 가중치의 분포가 많이 퍼져 있는 것은 적은 양의 데이터로부터 가중치를 추정함으로 인한 추정오차 때문이며, 이는 인식성능의 저하를 초래한다. 본 논문에서는 eigenvoice 가중치의 분포특성 정보를 이용하여 적응 데이터가 적은 경우의 인식성능 저하를 보상하는 방법으로 eigenvoice 가중치에 대한 문턱치 적용 방식을 제안하였다.

2.2. Eigenvoice 가중치에 대한 문턱치 적용

앞에서 살펴본 바와 같이 적응 데이터의 수가 적은 경우에 추정된 eigenvoice 가중치의 신뢰도가 낮다. 따라서 그림 2와 같이 추정된 가중치가 정해진 범위 내에 들도록 식 (2)와 같이 eigenvoice 가중치에 대한 문턱치를 적용할 수 있다.

$$w'(k) = \begin{cases} w(k)_{TH\_U}, & if\ w(k) > w(k)_{TH\_U} \\ w(k)_{TH\_L}, & if\ w(k) < w(k)_{TH\_L} \\ w(k), & otherwise \end{cases} \quad (2)$$

여기서 $w(k)_{TH_U}$와 $w(k)_{TH_L}$은 각각 k차 eigenvoice 가중치에 대한 상한 문턱치와 하한 문턱치이다. 이들 가중치 문턱치는 인식대상 화자를 제외한 여러 화자로부터 적응 데이터가 많을 때에 추정된 가중치의 분포를 구한 뒤, 각 eigenvoice 차원별로 아래의 식 (3)과 같이 결정한다.

$$w(k)_{TH\_U} = Mean(w(k)) + \beta\, STD(w(k)),$$
$$w(k)_{TH\_L} = Mean(w(k)) - \beta\, STD(w(k)),\ \beta > 0 \quad (3)$$

여기서 $Mean(x)$와 $STD(x)$는 x의 분포로부터 구한 평균 및 표준편차이다. 그리고 β는 임의의 상수이다. 위 식에서 β가 커지면 문턱치의 효과가 줄어들고, 반면에 β가 너무 작아지면 제대로 추정된 가중치에 대해서 잘못된 왜곡을 가져온다. 따라서 인식성능 향상을 위해서는 최적의 β의 값을 설정하는 것이 중요하다. 본 논문에서는 식 (3)에서의 β를 실험적으로 바꿔가면서 성능평가를 하였다.

III. 실험 및 결과

3.1. 실험 환경

본 실험에서 20 ms 해밍창 (Hamming window)을 10 ms씩 이동시키면서 얻은 12차 MFCC (Mel-Frequency Cepstral Coefficients)와 1, 2차 미분치를 구하여 총 36차의 음성 특징벡터를 사용하였다. 그리고 HMM 구성에서는 결정트리 기반 군집화 (tree-based clustering) 기법을

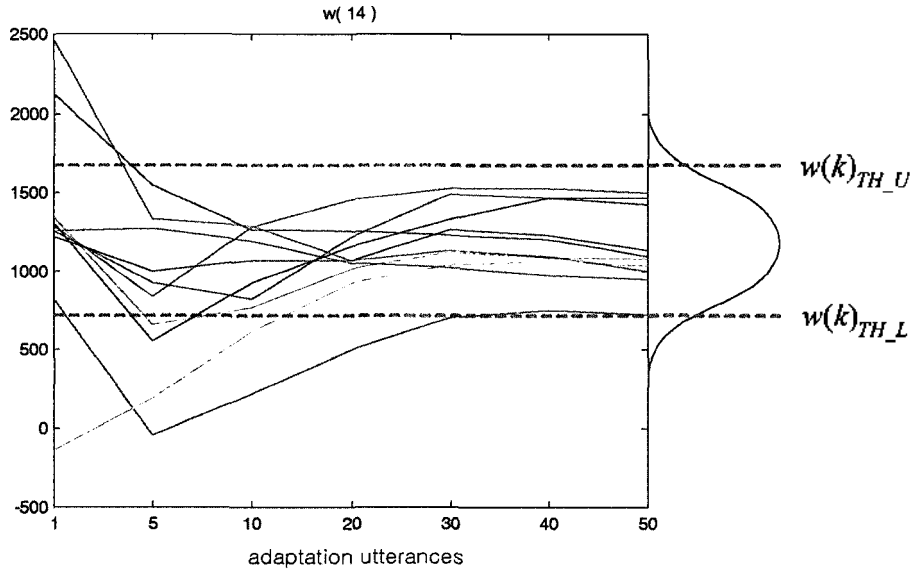


그림 2. Eigenvoice 가중치에 대한 문턱치 적용
Fig. 2. Applying threshold to eigenvoice weights.

적용한 트라이폰을 기본 모델로 사용하였으며 모델당 상태수는 3개이다.

훈련에 사용한 데이터베이스는 ETRI에서 구축한 3음소열 최적화 단어 (POW: Phonetically Optimized Words) [4] 음성 데이터베이스이며, 그 중에서 남성화자 40명의 음성 데이터베이스로부터 화자독립 (SI) 모델을 구성하고 각 화자에 대하여 MAP 적응 방식을 적용하여 40개의 화자종속 (SD) 모델을 구성하였다. 그리고 40개의 화자종속 모델에 대하여 주성분 분석법 (PCA)을 적용하여 30차의 eigenvoice를 구성하였다.

화자적응 및 성능평가에서는 452 균일 음소 분포 단어 (PBW: Phonetically Balanced Words)[5] 데이터베이스의 일부인 남성화자 10명에 대하여 각 화자별로 50개까지 단어 수를 늘려가며 적응에 사용하였고, 나머지 중 400개 단어를 성능평가에 사용하였다.

본 논문에서는 상태당 믹스쳐 (mixture)가 1개인 경우만 인식성능을 평가하였다. 그리고 10명의 화자에 대하여 적응단어 수를 증가시켜가면서 교사방식 (supervised mode)으로 30차의 eigenvoice를 사용하여 적응을 하였다.

3.2. 실험 결과

기존의 여러 가지 화자적응 방식에 대하여 인식성능을 비교하였다. 기준 시스템 (baseline)은 40명의 POW 데이터베이스로부터 구성한 화자독립 (SI) 모델을 사용하였다. 그림 3은 기준 시스템과 MAP, MLLR, 그리고 eigenvoice 방식으로 화자적응을 수행한 결과이다[6].

본 실험에서 화자독립 (SI) 모델을 사용하였을 경우에는 95.78%의 단어 인식률을 보였다. MAP 적응방식을 사용한 경우에는 화자독립 모델을 사용한 경우보다 성능이 떨어진다. 이는 적응데이터가 적어서 일부의 모델만 업데이트 되기 때문이라고 볼 수 있다. 그러나 실험에서 적

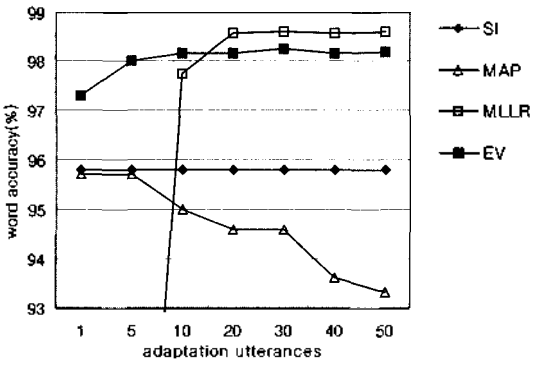


그림 3. 여러 가지 적응 방식들의 성능비교[6]
Fig. 3. Performance of several adaptation methods[6].

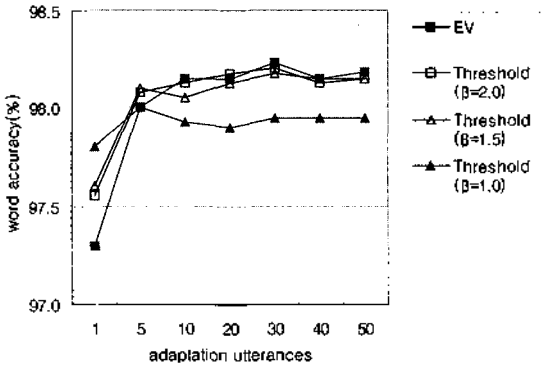


그림 4. Eigenvoice 가중치에 대한 문턱치 적용 실험결과
Fig. 4. Results of applying threshold to eigenvoice weight.

응단어 수를 충분히 많이 사용하면 MAP 적응방식에서도 성능이 향상되었다. MLLR 적응방식의 경우에는 적응단어의 수가 10개 이상인 경우에는 성능향상이 있지만, 10개 미만인 경우에는 성능을 보장할 수 없음을 알 수 있다. 그리고 Eigenvoice를 사용한 경우 (그림 3에서 EV)에는 적응데이터가 적은 경우에 성능향상이 있었다. Eigenvoice 적응방식은 적응화자에 대하여 추정할 파라메터 수가 적기 때문에 적은 고속화자적응에 유리함을 알 수 있다.

PBW 데이터베이스 중에서 인식실험에 참가하지 않은 남성화자 28명에 대하여 각 화자별로 50개의 단어를 사용하여 eigenvoice 가중치를 추정하였다. 이 28명에 대하여 추정된 가중치로부터 식 (3)과 같이 각 eigenvoice 차수별로 eigenvoice 가중치에 대한 문턱치를 구한 뒤, 10명의 화자에 대하여 적응 및 인식성능 평가를 하였다.

그림 4는 기존의 eigenvoice를 사용했을 경우 (그림 4에서 EV로 표시)와 문턱치를 적용한 경우의 인식성능을 보여준다. 그림 4에서 살펴보면 적응단어 1개를 사용하였을 때 문턱치를 적용하지 않은 경우 (그림 4에서 EV)에는 단어 인식률이 97.3%였으나, 문턱치를 적용하였을 경우 (β =1.5)에는 97.6%로 인식성능이 향상되었다. 이와 같이 문턱치를 적용하여 적응단어 수가 적은 경우 (적응단어 5개 이하)에 성능향상이 있음을 알 수 있다. 이는 문턱치의 적용으로 적은 데이터로부터 추정된 가중치를 보상해주기 때문이다. 그러나 β 값이 너무 작은 경우 (그림 4에서 β =1)에는 가중치에 대한 왜곡으로 인하여 전체적으로 인식성능이 저하됨을 확인할 수 있다.

IV. 결 론

본 논문에서는 eigenvoice 가중치의 분포 특성을 이용

한 고속화자적응 방식을 제안하였다. Eigenvoice 가중치에 대한 문턱치 적용 방식을 사용하여 적응 데이터가 적은 경우에 가중치의 추정이 잘못되는 것을 보상할 수 있었다. 본 논문에서는 문턱치 선정을 위해 고정된 β 값을 사용하였으나, 적응단어의 수에 따라 β 값을 다르게 설정하는 것도 추가적인 성능향상을 위한 한 방법이라 생각된다. 또한 본 논문에서 eigenvoice 가중치의 분포를 구할 때 인식단계와 동일한 환경인 PBW 음성 데이터베이스를 사용하였다. 물론 테스트 화자와 동일 화자 음성을 사용하지는 않았다 하더라도 가중치 분포특성을 구하기 위해 동일한 환경의 데이터베이스를 확보해야 한다는 점은 제약요소로 작용한다. 앞으로 이러한 제약조건을 완화시킬 수 있는 eigenvoice 가중치 분포 특성 이용 방법에 대한 연구와 함께 eigenvoice 기반 고속화자적응의 성능을 향상시킬 수 있는 또 다른 방법에 대한 연구가 필요하다.

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 (1), 171-185, Sep, 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. Y. Lim and Y. Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," *Proc. ICASSP*, 1, 89-91, 1995.
5. 김용완, 김종진, 김선태, 김태환, 김영일, 이용주, "공동이용을 위한 단어음성 DB의 구축 및 PBS 설계에 관한 검토," 제13회 음성통신 및 신호처리 워크샵 논문집, 256-261, 1996.
6. 송화전, 이윤근, 김형순, "차원별 Eigenvoice와 화자적응 모드 선택에 기반한 고속화자적응 성능 향상," 한국음향학회지, 22 (1), 48-53, 2003.

저자 약력

● 박 종 세 (Jong-Se Park)

2000년 2월: 부산대학교 공과대학 전자공학과 (공학사)
2003년 2월: 부산대학교 대학원 전자공학과 (공학석사)
※ 주관심분야: 음성인식, 음성합성, 음성신호처리

● 송 화 전 (Hwa-Jeon Song)

현재: 부산대학교 전자공학과 박사과정
한국음향학회지 제22권 제1호 참조

● 김 형 순 (Hyung-Soon Kim)

현재: 부산대학교 전자공학과 부교수
한국음향학회지 제22권 제1호 참조