수정된 EM알고리즘을 이용한 GMM 화자식별 시스템의 성능향상

Performance Enhancement of Speaker Identification System Based on GMM Using the Modified EM Algorithm

김 성 종* · 정 익 주**
Seong-Jong Kim · Ik-Joo Chung

ABSTRACT

Recently, Gaussian Mixture Model(GMM), a special form of CHMM, has been applied to speaker identification and it has proved that performance of GMM is better than CHMM. Therefore, in this paper the speaker models based on GMM and a new GMM using the modified EM algorithm are introduced and evaluated for text-independent speaker identification. Various experiments were performed to evaluate identification performance of two algorithms. As a result of the experiments, the GMM speaker model attained 94.6% identification accuracy using 40 seconds of training data and 32 mixtures and 97.8% accuracy using 80 seconds of training data and 64 mixtures. On the other hand, the new GMM speaker model achieved 95.0% identification accuracy using 40 seconds of training data and 32 mixtures and 98.2% accuracy using 80 seconds of training data and 64 mixtures. It shows that the new GMM speaker identification performance is better than the GMM speaker identification performance.

Keywords: GMM, EM, CHMM, Speaker Identification

1. 서 론

사람이 발성한 음성은 여러 가지 형태의 정보를 지니고 있다. 가장 주된 정보는 화자의 배경(출신지역, 교육수준 등), 화자의 감정 및 건강 상태 그리고 성도(vocal tract)의 물리적 특성이 음성 신호에 내재되어 있다. 이들 중 많은 부분이 화자에 종속적인 것이고 화자를 구분하는 데 사용할 수 있는 정보이다. 화자인식은 이러한 음성정보를 이용하여 화자가 누구인가를 판별하는 기술이다. 화자인식(speaker recognition)기술은 화자식별(speaker identification)기술과 화자검증(speaker verification)기술로 나눌 수 있다. 화자식별 기술은 입력된 음성에 대해서 등록된 화자들과 비교하여 그중에서 가장 유사한 화자를 골라내는 것이며, 화자검증 기술은 핵심어 인식에서와 같이 기준 패턴과 입력 패턴을 서로 비교하여 미리 정해 놓은 발생 확률 값을 넘어서면 승인하고 그렇지 않으면 거절하는

* 강원대학교 전자공학과
** 강원대학교 전기전자정보통신공학부

것이다.

본 논문에서는 화자인식 기술 중에서 화자식별의 성능향상을 위한 방법을 제안한다.

최근 화자식별에 가장 많이 적용되는 대표적인 알고리즘은 GMM이며, 다른 알고리즘에 비해 높은 인식성능을 발휘한다. 따라서 본 논문에서는 기존의 GMM을 이용한 화자식별 시스템을 구현하고, 새롭게 제안한 수정된 EM 알고리즘이 적용된 GMM을 이용한 화자식별 시스템을 구현하여 두 화자식별 시스템의 성능 비교실험을 하였고, 이를 통하여 새롭게 제안한 알고리즘이 기존의 GMM에 비해 높은 인식성능을 발휘한다는 것을 보였다.

본 논문의 구성은 다음과 같다. 2 장에서는 GMM 화자식별 시스템을 설명하고, 3 장에서는 새롭게 제안한 수정된 EM 알고리즘을 적용한 GMM 화자식별 시스템의 구현에 대해 설명하고, 4 장에서는 두 화자식별 시스템의 성능비교실험 및 결과를 제시하며, 마지막으로 5 장에서는 실험결과를 종합적으로 검토하고 결론을 맺는다.

2. GMM 화자식별 시스템

화자인식은 여러 해 동안 연구주제였고, 다양한 화자모델이 연구되었다. HMM은 음성인식 및 화자인식을 위한 도구로서 가장 많이 사용되었으며, 그 중에서도 가장 좋은 성능을 나타낸 화자인식 알고리즘은 CHMM(continuous HMM)이다[5]. 이후 하나의 상태(one state)만을 갖는 CHMM 즉, GMM(gaussian mixture model)이 화자를 모델링 하는데 널리 사용되었는데[4][7][8], GMM이 다중상태(multi-states)를 갖는 CHMM에 비하여 훨씬 더 좋은 성능을 보인다는 것이 증명되었다[6]. 따라서 본 장에서는 GMM 화자식별 시스템의 구현에 대해 알아본다.

2.1 GMM 화자식별 시스템

화자식별 시스템은 크게 두 부분으로 나눌 수 있다. 하나는 훈련부로서 각 화자로부터 음성특징벡터를 추출하여 가우시안 혼합모델 λ를 구하게 되며, 다른 하나는 인식부로서 입력음성에 대해 최대의 확률을 갖는 가우시안 혼합모델 λ에 해당하는 화자를 찾아낸다.

GMM 파라미터 λ는 다음처럼 혼합가중치(mixture weight) p_i, 평균벡터(mean vector) $\vec{\mu_i}$, 공분산행렬(covariance matrix) Σ_i으로 구성된다[1][9].

$$\lambda = \{p_i, \vec{\mu_i}, \Sigma_i\} \quad i = 1, \ldots, M \tag{1}$$

여기서, 혼합가중치(mixture weight) p_i는 다음 식을 만족해야 한다.

$$\sum_{i=1}^{M} p_i = 1 \tag{2}$$

2.1.1 훈련부

각 화자에 대한 훈련 음성 데이터가 주어져 있을 때, GMM 훈련부의 목표는 훈련에 사용되는 특징벡터의 분포를 가장 잘 표현해 줄 수 있는 가우시안 혼합모델의 파라미터를 추정하는 것이다. GMM의 파라미터를 추정하는 데 가장 많이 사용되는 방법은 Expectation-Maximization(EM) 알고리즘[3]이며, 반복적인 방법으로 최적의 파라미터를 구해낸다.

GMM 훈련부에 대한 블록도가 <그림 1>에 나타나 있다.

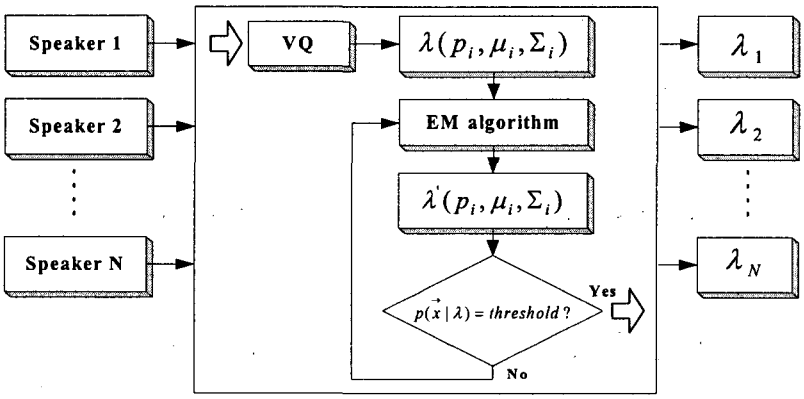


그림 1. GMM 훈련부의 블록도

2.1.2 GMM 인식부

N 명의 화자 그룹 $S=\{s_1, s_2, \ldots, s_N\}$ 은 GMM의 $\lambda_1, \lambda_2, \ldots, \lambda_N$ 으로 표현된다. 인식부의 목표는 주어진 관측열(즉, 입력음성)에 대해서 다음의 식(3)에 주어진 것처럼 최대사후확률(maximum a posteriori probability)을 갖는 화자모델을 찾아내는 것이다. 다시 말하면 훈련부에서 구한 각 화자모델에 대한 입력음성의 확률이 가장 큰 모델을 찾는 것으로서 입력된 음성이 어떤 화자로부터 나올 확률이 가장 큰 것인지를 판단하는 것이다.

$$\hat{S}=\arg\max_{1\le k\le N}\Pr(\lambda_k|X)=\arg\max_{1\le k\le N}\frac{p(X|\lambda_k)\Pr(\lambda_k)}{p(X)} \tag{3}$$

인식부에 대한 블록도는 다음 <그림 2>와 같다.

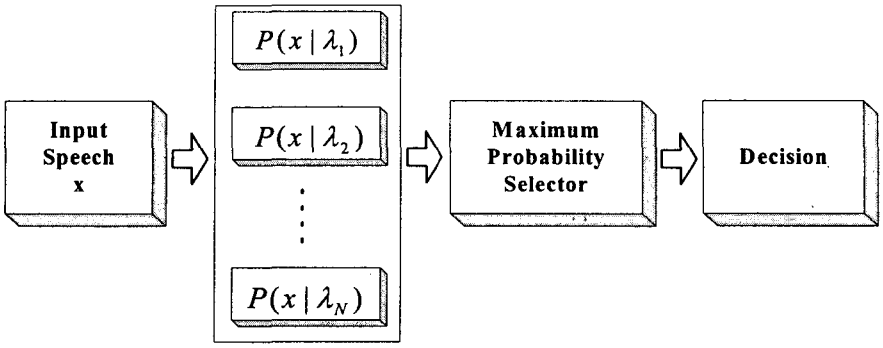


그림 2. GMM 인식부의 블록도

3. 수정된 EM 알고리즘을 적용한 GMM 화자식별 시스템

GMM 화자모델의 최적의 파라미터는 EM iteration을 통하여 반복적인 방법으로 구한다. 그런데, GMM 화자모델 $\lambda(p_i, \vec{\mu_i}, \Sigma_i)$를 재추정할 때, 기존의 EM 알고리즘은 어떤 mixture이건 간에 모든 관측 프레임(즉, $v_1, v_2, \ldots, v_T$)을 적용하여 해당 mixture를 갱신하게 된다. 따라서 해당 mixture와는 상관관계가 매우 희박한 프레임들도 포함될 수 있다. 따라서 본 논문에서 새롭게 제안한 수정된 EM 알고리즘은 상관관계가 낮은 프레임들을 제외시켜서 각각의 mixture를 갱신하는 것이다. 다시 말하면, 각각의 mixture를 갱신할 때 해당 mixture와 가장 멀리 떨어져 있는 mixture를 찾아낸 다음 그 mixture에 속한 프레임들은 제외시킴으로서 mixture간의 변별력을 극대화시키고자 하는 것이다.

수정된 EM 알고리즘의 기본적인 수행 절차는 다음과 같다.

1) 모델 초기화(initialization): EM 알고리즘을 적용하기 전에 초기 화자모델이 필요하게 되는데, 본 논문에서는 적용할 mixture 수(mixture size)만큼의 셀 블록을 벡터양자화(VQ)[2]를 이용하여 구하고 각 셀당 평균과 분산을 구하여 이것을 초기 모델로 정하였다.

2) 순서조정(ordering): 각 셀 블록의 평균벡터간의 거리를 Euclidean distance로 구한 다음, 거리가 먼 순서 즉, 상관관계가 낮은 순서대로 정렬한다.

3) 셀 블록 제외(except the cell block): 해당 셀 블록의 파라미터를 갱신할 때 상관관계가 낮은 셀 블록에 해당하는 프레임들은 제외한다. 따라서 상관관계가 높은 프레임들만을 이용하여 파라미터를 갱신함으로서 다른 화자와 보다 변별력 있는 화자모델을 구하는 것이다.

4) 클러스터링(clustering): EM 알고리즘을 수행할 때마다 p_i, $\vec{\mu_i}$, Σ_i이 갱신된다. 따라서 이 갱신된 파라미터 중에서 평균벡터 $\vec{\mu_i}$을 이용하여 M개의 셀 블록을 다시 형성시킨다.

5) 정규화(normalization): 수정된 EM알고리즘 적용시 상관관계가 낮은 프레임을 제외하게 되는데, 이에 대한 영향으로 갱신된 혼합가중치 $\hat{p}_i$가 식 (2)를 정확히 만족하지 못하는 일종의 혼합가중치 에러(mixture weight error)가 발생한다. 따라서 혼합가중치 $\hat{p}_i$가 식 (2)의 조건을 만족할 수

있도록 하기 위해 다음식과 같이 정규화를 취하였다.

$$\hat{p}_i = \frac{\hat{p}_i}{\sum_{k=1}^{M} \hat{p}_k}, \quad i = 1, 2, \ldots, M \tag{3}$$

6) 반복(recursion) 및 종료(termination): 주어진 반복회수에 도달할 때까지 2)~5)를 반복하여 최종 GMM 파라미터를 구한다.

다음 <그림 3>은 본 논문에서 제안한 수정된 EM 알고리즘을 적용한 GMM 화자식별 시스템의 훈련부를 나타낸 것이다.

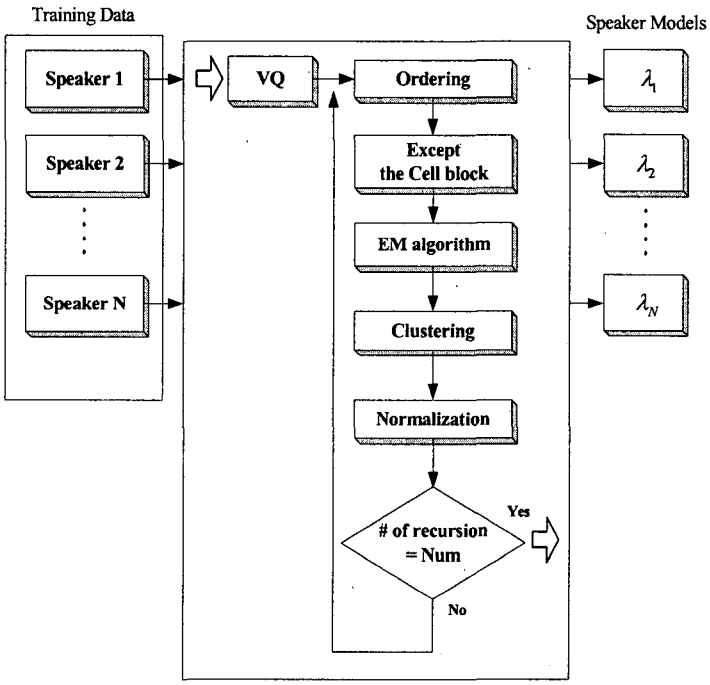
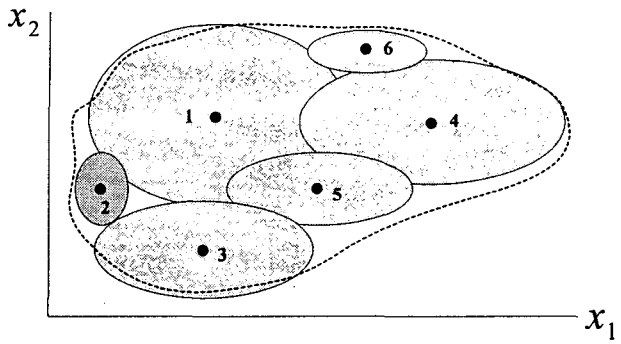
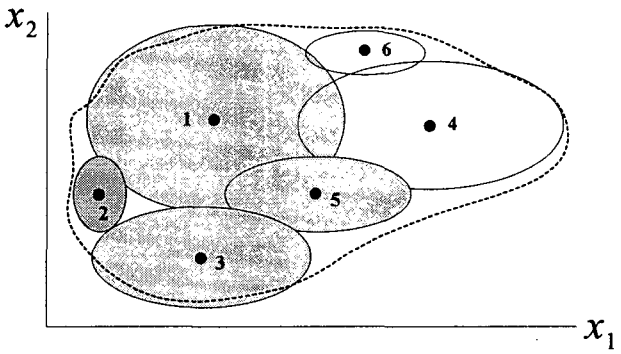


그림 3. 수정된 EM 알고리즘을 적용한 GMM 화자식별 시스템의 훈련부

다음 <그림 4>는 기존의 EM 알고리즘과 새롭게 제안한 수정된 EM 알고리즘을 비교한 것으로서, mixture의 수를 6 개로 가정하였다. (a), (b) 모두 진하게 표시된 2 번 mixture를 갱신하는 예를 보인 것으로서, (a)는 기존의 EM 알고리즘을 이용하여 2 번 mixture를 갱신할 때 자신을 포함한 다른 모든 프레임을 이용한다는 것을 보여준다. 그러나 (b)는 새롭게 제안한 수정된 EM 알고리즘을 적용하기 때문에 2 번 mixture와 상관관계가 매우 낮다고 판단된 4 번과 6 번 mixture에 속한 프레임은 제외 시켜서 해당 mixture를 갱신한다.



(a) EM algorithm



(b) Modified EM algorithm

그림 4. EM 알고리즘과 수정된 EM 알고리즘의 비교

4. 실험 및 결과

기존의 GMM 화자식별 시스템과 새롭게 제안한 수정된 EM 알고리즘이 적용된 GMM 화자식별 시스템의 성능비교 실험을 위하여 <표 1>과 같은 실험 파라미터를 구축하였다.

표 1. 실험 파라미터

훈련데이터의 길이	15 초, 40 초, 80 초
특징벡터	LPCC 14 차
알고리즘	GMM, 수정된 EM을 적용한 GMM
음성데이터베이스	SITEC PBW 16 kHz, 16 bits - 남자화자 38 명
입력 음성 데이터	각 화자 당 3 음절, 60 단어

기존의 실험에서는 남자 38 명, 여자 32 명으로 총 70 명을 적용하였는데, 성별이 다른 경우에는 화자간 특징차이가 뚜렷하기 때문에 성별간 오인식은 나타나지 않았다. 따라서 본 실험에서는 남자 화자 38 명 만을 이용하였다. 실험에 사용된 인식률의 단위는 백분율(%)로 나타내며 다음 식과 같다.

$$\text{인식률}(\%) = \frac{\text{정(正)인식 횟수}}{\text{입력 음성 데이터의 수}} \times 100 \tag{4}$$

각각의 화자에 대한 인식률은 위의 식을 이용하여 구하고, 전체인식률은 38 명의 화자에 대한 인식률의 평균을 취하였다. 즉, 다음 식과 같다.

$$\text{전체 인식률}(\%) = \frac{\text{각 화자의 인식률의 총합}}{\text{화자의 수(38명)}} \times 100 \tag{5}$$

본 실험에서는 훈련데이터의 길이를 15 초, 40 초, 80 초로 다양하게 설정하였고, mixture 수도 8, 16, 32, 64를 이용하였으며, 혼합가중치를 정규화해서 식 (2)를 만족시킨 경우와 정규화를 취하지 않고 그대로 적용한 경우에 대해 실험하였다.

4.1 실험결과

훈련데이터의 길이에 따라 mixture의 수를 다음과 같이 설정하여 인식실험을 수행하였으며, 결과 그래프에서 "Except 수"는 제외되는 셀 블록의 개수를 의미하며 Except수가 "0"이라는 것은 수정된 EM 알고리즘이 적용되지 않은 기존의 GMM 화자식별 시스템을 의미한다.

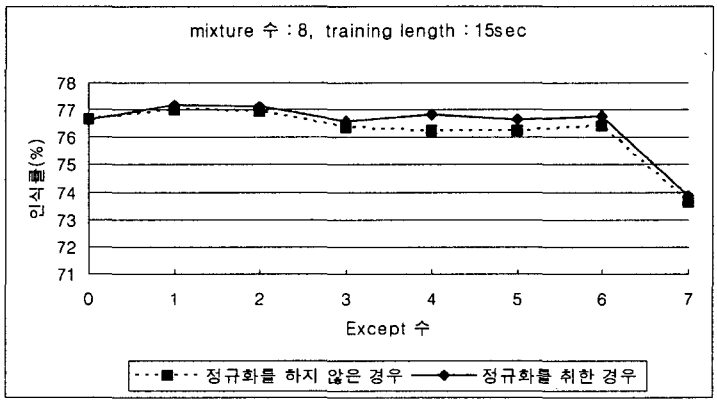


그림 5. 훈련데이터의 길이: 15 초, mixture수: 8

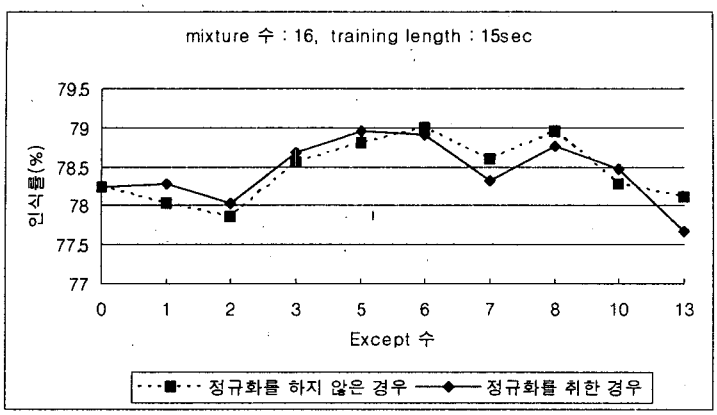


그림 6. 훈련데이터의 길이: 15 초, mixture수: 16

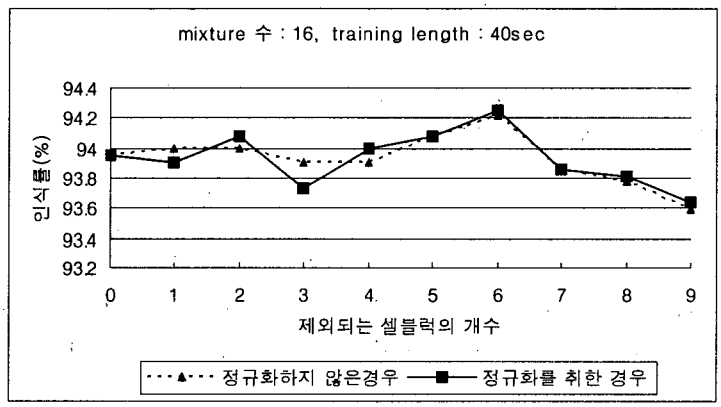


그림 7. 훈련데이터의 길이: 40 초, mixture수: 16

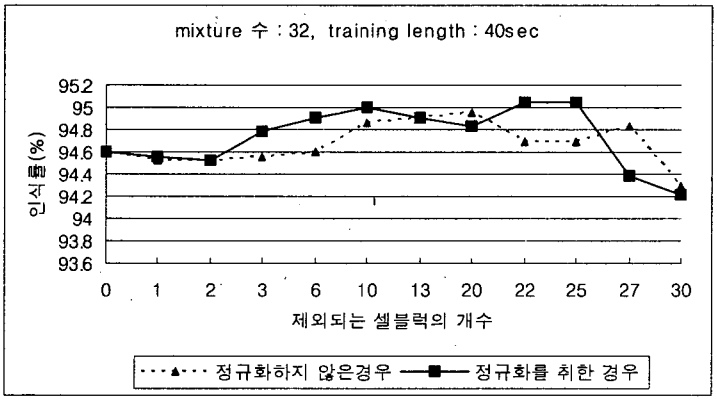


그림 8. 훈련데이터의 길이: 40 초, mixture수: 32

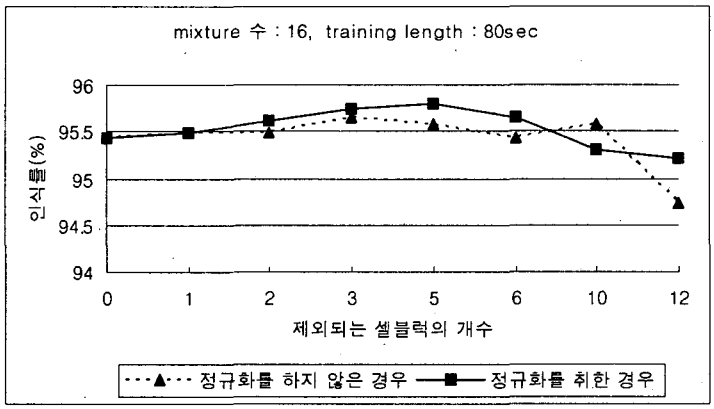


그림 9. 훈련데이터의 길이: 80 초, mixture수: 16

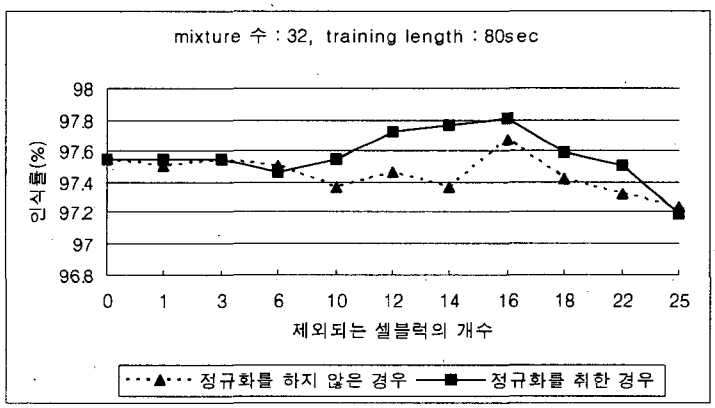


그림 10. 훈련데이터의 길이: 80 초, mixture수: 32

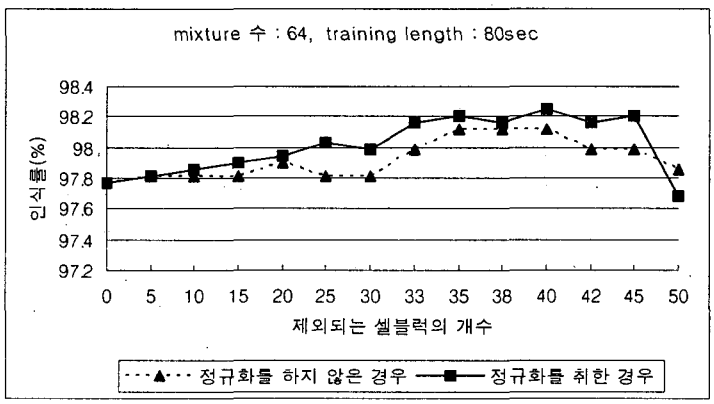


그림 11. 훈련데이터의 길이: 80 초, mixture수: 64

4.2 실험결과 분석

우선, 훈련데이터의 길이가 15 초일 경우를 보면, mixture 수가 8, 16에 대한 결과 그래프에서 보는 바와 같이 기존의 EM알고리즘(Except수가 0인 경우)에 비해 인식률 향상이 있음을 알 수 있다. 그러나 mixture의 수가 8일 경우와 16일 경우에 제외되는 셀블럭에 따른 인식률 추이에 큰 차이가 있음을 알 수 있다. Mixture의 수가 8일 경우는 비교적 선형적으로 진행되는 반면에 16일 경우는 변화정도가 급격하다는 것을 알 수 있다. 이러한 결과는 훈련데이터의 길이에 따른 적절한 mixture의 개수와 밀접한 관련이 있다고 할 수 있는데, 훈련데이터의 길이 15 초에 대한 적절한 mixture의 수가 8 정도이기 때문에 제외되는 셀블럭의 개수에 따라 비교적 선형적인 인식률 변화를 보이는 것이다. 하지만, mixture의 수가 16일 경우에는 훈련데이터의 길이에 비해 보다 많은 mixture를 적용하였기 때문에 결과 그래프에서처럼 급격한 인식률 변화를 보이게 되는데, 제외되는 셀블럭의 개수가 6일 경우를 보면, mixture의 수가 8일 경우보다 높은 인식률을 보인다는 것을 알 수 있다. 이것은 훈련데이터의 길이에 비해 보다 많은 mixture를 적용할 경우에는 인식률 변화가 급격한 반면, 보다 높은 인식성능을 발휘할 수 있다는 가능성을 보여준다.

다음으로, 훈련데이터의 길이가 40 초일 경우를 살펴보면, mixture의 수를 16, 32로 선택하여 실험하였는데, 결과 그래프에서 보듯이 32일 경우가 16에 비해 보다 선형적인 인식률 향상을 보인다는 것을 알 수 있다.

마지막으로 훈련데이터의 길이가 80 초일 경우를 보면, mixture의 수를 16, 32, 64로 선택하여 실험하였는데, 결과에서 보듯이 64일 경우에 선형적인 인식률 향상을 보인다는 것을 알 수 있다.

따라서, 훈련데이터의 길이에 따른 적절한 mixture의 수를 선택하면, 선형적인 인식률의 향상을 기대할 수 있으며, 그렇지 못할 경우에는 비선형적인 향상이 기대되지만, 인식성능이 더 높아질 가능성도 있는 것이다.

이상의 실험 결과로부터, 본 논문에서 제안한 수정된 EM 알고리즘을 적용한 화자식별 시스템의 성능이 기존의 GMM 화자식별 시스템보다 향상 되었다는 것을 알 수 있다. 결과를 종합하면 다음과 같다. 첫째, 기존의 화자식별 시스템과 마찬가지로 훈련데이터의 길이가 증가함에 따라 인식률이 향상되었으며, mixture 수가 커짐에 따라 인식성능이 향상되었다. 둘째, 제외되는 셀 블록의 개수가 증가함에 따라 인식률이 증가한다는 것을 알 수 있다. 그러나 이 인식률 증가가 어느 문턱치에 도달한 이후부터는 점차 떨어지는 것을 알 수 있는데, 이유는 제외되는 셀 블록의 개수가 많아질수록 EM 알고리즘 적용시 모델 파라미터(p_i , $\vec{\mu_i}$, Σ_i)를 갱신하는 데 이용되는 훈련데이터의 양이 현저하게 줄어들기 때문이다. 다시 말하면 화자모델을 구현하기위한 각 화자에 대한 정보량이 그만큼 줄어들게 되고, 결국에는 다른 화자와의 변별력이 떨어지게 되는 것이다. 특히, 훈련데이터의 길이가 80 초이고 mixture 수가 64일 경우에는 제외되는 셀 블록의 수에 따른 인식률의 향상이 선형적으로 증가하지만, 이 경우에도 제외되는 셀 블록의 개수가 지나치게 많아지면 인식률이 떨어지는 것을 알 수 있다. 따라서 제외되는 셀 블록의 개수를 훈련데이터의 길이에 따라 적절하게 조절하면 기존의 GMM 화자식별 시스템보다 우수한 인식성능을 발휘할 수 있음을 알 수 있다.

5. 결 론

실험 결과로부터, 본 논문에서 새롭게 제안한 수정된 EM 알고리즘을 적용하여 구성한 GMM 화자식별 시스템이 기존의 GMM 화자식별 시스템보다 우수한 인식성능을 발휘한다는 것을 알 수 있으며, 수정된 EM 알고리즘을 적용한 화자식별 시스템은 훈련 데이터의 길이가 고정되어 있을 경우 mixture 수(즉 mixture의 수)가 증가함에 따라 인식률이 높아지며, 훈련 데이터의 길이가 길어질수록 인식률이 향상된다. 그러나 훈련 데이터를 충분히 확보할 수 없는 상황에서는 훈련데이터의 길이에 따라 mixture 수를 적절히 선택해야 좋은 인식 성능을 발휘할 수 있다는 것을 실험을 통하여 알 수 있었다.

다음 그래프는 훈련데이터의 길이가 40 초, 80 초에 대한 두 화자식별 시스템의 최고 인식률을 비교한 것으로서, 본 논문에서 제안한 수정된 EM 알고리즘이 GMM 화자식별 시스템의 인식성능을 향상시킨다는 것을 알 수 있다.

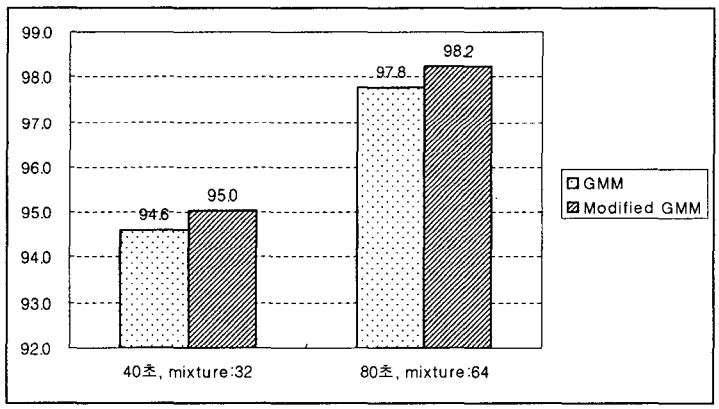


그림 12. 기존의 GMM과 수정된 GMM의 인식률 비교

결과적으로, 본 논문에서 제안한 수정된 EM 알고리즘 방식을 적용한 GMM 화자모델은 기존의 GMM 화자모델에 비해서 보다 변별력이 뛰어나다고 할 수 있다. 그리고 훈련데이터의 길이가 고정되어 있을 경우 기존의 GMM 방식보다 높은 인식률을 보인다. 따라서 훈련 데이터를 많이 확보할 수 없는 상황 또는 적은 훈련 데이터만으로 좋은 인식 성능을 발휘하고자 하는 시스템에 적용할 수 있는 방법이라 할 수 있다.

참 고 문 헌

[1] Reynolds, D. A. & Rose, R. C. 1995. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models." *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 72-83.

[2] Rabiner, L. & Juang, B. H. *Fundamentals of Speech Recognition*, Prentice Hall.
[3] Xuan, G., Zhang, W. & Chai, P. 2001. "EM Algorithms of Gaussian Mixture Model and Hidden Markov Model." Proc. *ICIP'01*, pp. 145-148.
[4] Reynolds, D. A. 1995. "Speaker Identification and Verification Using Gaussian Mixture Speaker Models." *Speech Communication*, vol. 17, pp. 91-108.
[5] Furui, S. 1991. "Speaker-dependent feature extraction, recognition and processing techniques." *Speech Comm.*, vol. 10, no. 5, pp. 505-520.
[6] Markov, K. & Nakagawa, S. 1995. "Text-Independent speaker identification on TIMIT database." *Proc., Acous. Soc. Jap.* pp. 83-84.
[7] Gish, H. & Schmidt, M. 1994. "Text-Independent speaker identification." *IEEE Signal Processing Magazine*, pp. 18-32.
[8] Bimbot, F., Magrin-Chagnolleau, I. & Mathan, L. 1995. "Second-order statistical measures for test-independent speaker identification." *Speech Communication*, vol. 17, pp. 177-192.
[9] Liu, L. & He, J. 1999. "On the use of orthogonal GMM in speaker recognition." *Proc. ICASSP'99*, pp. 845-848.

접수일자: 2005. 11. 08
게재결정: 2005. 11. 30

▲ 김성종
강원도 춘천시 효자2동
강원대학교 대학원 전자공학과 (우: 200-701)
Tel: +82-33-263-4345 (H)
E-mail: deniro1107@hotmail.com

▲ 정익주
강원도 춘천시 효자2동
강원대학교 공과대학 전기전자정보통신공학부 (우: 200-701)
Tel: +82-33-250-6322 (O)
E-mail: ijchung@kangwon.ac.kr