한국음향학회지 제21권 제2호 pp. 190~198 (2002)

MLLR 화자적응 기법을 이용한 새로운 화자확인 디코딩 알고리듬

A Noble Decoding Algorithm Using MLLR Adaptation for Speaker Verification

김 강 열*, 김 지 운*, 정 재 호*
(Gang Youl Kim*, Ji Un Kim*, Jae Ho Chung*)

*인하대학교 전자공학과 디지털 신호처리 연구실
(접수일자: 2001년 9월 14일; 채택일자: 2001년 12월 10일)

화자확인에서 사용되는 디코딩 방법에는 음성인식에서 주로 사용되는 비터비 알고리듬을 사용하여 왔다. 그러나 화자확인에서는 화자의 특성을 최대한 발휘하여 같은 음소라도 화자마다 다르게 인식해야 하는 어려움이 있다. 본 논문에서는 기존 화자확인 디코딩에서 사용하는 비터비 알고리듬을 대신하는 새로운 알고리듬을 제안하였다. 제안된 알고리듬은 음성인식에서 사용되고 있는 화자 적응 알고리듬을 화자의 특성에 따라 모델 파라미터로 변환하는 것을 응용한 방법이다. 본 논문에서는 여러 적응 알고리듬 중 MLLR (Maximum Likelihood Linear Regression)과 MAP (Maximum A-Posterior) 적응 알고리듬을 사용하였고 제안된 알고리듬이 기존의 비터비 알고리듬을 사용하였을 때보다 평균 30%의 EER (Equal Error Rate) 향상을 이루었다.
핵심용어: 비터비 알고리듬, 화자적응, MLLR 알고리듬, MAP 알고리듬, 화자확인
투고분야: 음성처리 분야 (2.5)

In general, we have used the Viterbi algorithm of Speech recognition for decoding. But a decoder in speaker verification has to recognize same word of every speaker differently. In this paper, we propose a noble decoding algorithm that could replace the typical Viterbi algorithm for the speaker verification system. We utilize for the proposed algorithm the speaker adaptation algorithms that transform feature vectors into the region of the client' characteristics in the speech recognition. There are many adaptation algorithms, but we take MLLR (Maximum Likelihood Linear Regression) and MAP (Maximum A-Posterior) adaptation algorithms for proposed algorithm. We could achieve improvement of performance about 30% of EER (Equal Error Rate) using proposed algorithm instead of the typical Viterbi algorithm.
Keywords: *Viterbi algorithm, Speaker adaptation, MLLR algorithm, MAP algorithm, Speaker verification*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

지금까지 화자확인은 전처리 과정, 훈련방법 그리고 정규화 등에서 새로운 알고리듬들이 개발되어 인식률의 향상에 기여하고 있다. 그리고 음성인식에서 사용되는 여러 알고리듬들이 화자확인에 또한 활용되고 있다. 음성인식을 위해서 개발된 화자 적응 알고리듬인 MLLR 적응 알고리듬과 MAP 적응 알고리듬 역시 화자확인에도 활용되고 있다[1].

화자 독립 모델로 만들어진 인식기는 화자 종속 모델로 만들어진 인식기보다 일반적으로 성능이 떨어진다. 따라서 잘 훈련된 화자 독립 모델로 인식기를 구성하고 화자의 소량의 데이터를 이용하여 화자 종속 모델로 변

책임저자: 김강열 (kgyms@dreamwiz.com)
402-751 인천광역시 남구 용현동 253
인하대학교 전자공학과 디지털 신호처리 연구실
(전화: 032-860-7420; 팩스: 032-860-3564)

환하는 것이 화자적응이다. 적은 양의 데이터가 제공될 때 화자 종속 모델을 구성하면 훈련을 할 때 너무 적은 관측 데이터로 인하여 부정확한 재추정이 실행될 수 있다. 화자확인의 인식 특성상 실용성을 위해 많은 훈련 데이터를 구하는 것은 쉽지 않다. 그래서 적은 양의 데이터로도 큰 효과를 보기 위한 화자 적응을 이용한 방법들이 연구되고 있다[1].

화자확인에서 사용되는 디코딩 방법에는 음성인식에서 주로 사용되는 비터비 알고리듬을 사용하여 왔다. 화자확인에서 디코딩을 수행할 때는 화자의 특성을 최대한 발휘하여 유사한 음소라도 화자마다 다르게 인식해야 한다. 그러나 비터비 알고리듬은 음성인식에서 사용될 때 화자의 특성보다는 다른 화자의 발성이라도 유사한 음소라면 같은 결과를 출력한다. 따라서 화자확인의 특성을 비터비 알고리듬보다 잘 표현할 수 있는 새로운 디코딩 알고리듬이 요구된다. 본 논문은 최근 화자의 특성을 추정해서 HMM 모델을 빠르게 변환하는 화자 적응을 이용하여 비터비 알고리듬을 대신할 수 있는 새로운 디코딩 알고리듬을 제안하였다.

본 논문은 II절에서 MLLR과 MAP 적응 알고리듬의 기본적인 설명을 하였다. III절에서는 비터비 알고리듬과 제안된 알고리듬을 비교하여 성능 향상 요인에 대해 언급을 하였다. IV절에서는 실험을 위해서 제작된 음성데이터베이스에 대해 설명을 하였고 V절에서는 음성데이터베이스와 제안된 방법을 이용한 구체적인 실험 방법을 설명하였다. VI절에서는 여러 가지 실험의 결과 제시와 결과 분석을 하였고 VII절에서 결론을 맺었다.

II. 적응 알고리듬

적응 알고리듬은 여러 가지가 있지만 본 논문은 MLLR과 MAP를 이용한 모델 적응을 사용하였다. MLLR은 원래의 모델과 적응 데이터 사이에 불일치를 줄일 수 있는 변환행렬을 이용하는 것이다. 즉 Gaussian mixture HMM 시스템의 평균과 분산 파라미터들에 대한 선형 변환행렬 세트를 추정하는 것이다. 새롭게 적응된 평균을 구하는 데 사용되는 변환행렬은 다음과 같은 방정식으로 표현된다[2,3].

$$\mu_{new} = W\hat{\mu} \tag{1}$$

변환행렬 W는 $n\times(n+1)$ 행렬이며 $\hat{\mu} = [\omega\ \mu_1\ \mu_2\ \cdots\ \mu_n]^T$이고 여기서 ω는 바이어스 옵셋값이고, 그림 1에서 이 바이어스 옵셋에 대한 영향을 잘 나타내고 있다.

변환행렬 W는 EM (Expectation-Maximization) 기법을 이용한 최대화에 대한 문제를 푸는 것에 의해 구할 수 있다. 또한 MLLR 알고리듬에서 중요한 것은 회귀 클래스 트리를 사용함으로써 적은 양의 데이터를 갖고도 적응 속도를 빠르게 하였다는 것이다. 즉, 비슷한 mixture 벡터들을 하나로 묶어서 트리구조의 회귀 클래스로 만든다. 각 클래스들에 대한 공통의 변환행렬을 구하여 적응 데이터에 의한 영향이 적은 음소나 단위에 대해서도 충분히 적응을 시킨다. 그림 2와 그림 3은 이 회귀 클래스 트리 개념을 도식적으로 보여주고 있다. 그림 3에서는 이해를 쉽게 하기 위해서 특징을 2차원으로 하였고 각 mixture들을 한 클래스로 묶어서 변환한다.

위의 변환행렬 W를 구하는 방법은 다음과 같다[4,5].

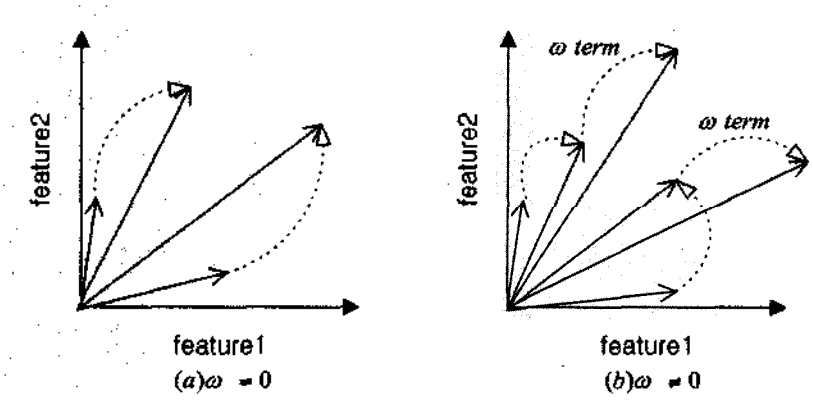

그림 1. ω의 영향
Fig. 1. Effects of ω.

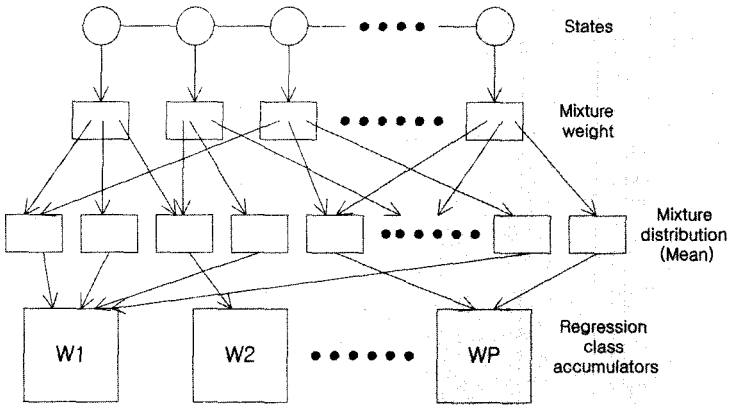

그림 2. 회귀 클래스 군집화와 변환 행렬
Fig. 2. Regression classes clustering and transformation matrix.

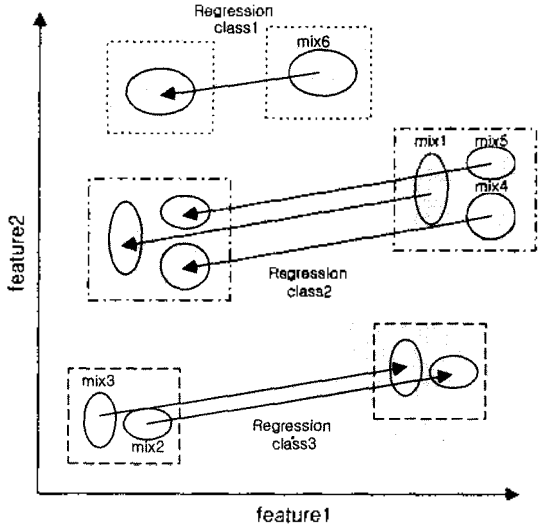

그림 3. 회귀 클래스
Fig. 3. Regression classes.

우선 상태당 단일 가우시안 분포를 갖는다고 가정을 하면 각 상태의 확률밀도 함수는 다음과 같다.

$$b_s(o) = \frac{1}{(2\pi)^{\frac{n}{2}} |C_s|^{\frac{1}{2}}} e^{-\frac{1}{2}(O-W_s\hat{\mu}_s)^T C_s^{-1}(O-W_s\hat{\mu}_s)} \quad (2)$$

여기서 W_s는 적응 데이터로 발생하는 적응된 모델들의 가능성을 최대로 하도록 선택된다. 적응 데이터가 T 관측의 열이라 가정하면 $O = o_1 \cdots o_T$라 하고 현재의 모델 세트를 λ라 하고 재추정된 모델 세트를 $\bar{\lambda}$로 놓는다. 그리고 O를 만드는데 사용되는 상태들의 열은 $\theta = \theta_1 \cdots \theta_T$라 한다. 그러면 보조 함수를 다음과 같이 정의할 수 있다.

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} F(O, \theta \mid \lambda) \log(F(O, \theta \mid \bar{\lambda})) \quad (3)$$

$$F(O, \theta \mid \lambda) = a_{\theta_T N} \prod_{t=1}^{T} a_{\theta_{t-1}\theta_t} b_{\theta_t}(o_t) \quad (4)$$

$$F(O \mid \lambda) = \sum_{\theta \in \Theta} F(O, \theta \mid \lambda) \quad (5)$$

여기서 보조 함수를 최대로 하는 각각의 구성요소를 구하면 된다. 재추정된 파라미터를 문자 상단에 윗줄로 표시를 하면 다음과 같이 출력밀도 함수로 변형할 수 있다.

$$Q(\lambda, \bar{\lambda}) = \sum_{i=1}^{N} Q_{a_i}[\lambda, \{\bar{a}_{ij}\}_{j=1}^{N}] + \sum_{i=1}^{N} Q_{b_i}(\lambda, \bar{b}_j) \quad (6)$$

$$Q_a[\lambda, \{\bar{a}_{ij}\}_{j=1}^{N}] = \sum_{\theta \in \Theta} F(O, \theta_T = i \mid \lambda) \log \bar{a}_{iN} + \sum_{\theta \in \Theta} \sum_{t=1}^{T} \sum_{j=1}^{N} F(O, \theta_{t-1} = i, \theta_t = j \mid \lambda) \log \bar{a}_{ij} \quad (7)$$

$$Q_b(\lambda, \bar{b}_j) = \sum_{\theta \in \Theta} \sum_{t=1}^{T} F(O, \theta_t = j \mid \lambda) \log \bar{b}_j(o_t) \quad (8)$$

변환 행렬 W_s만이 재추정되기 때문에 단지 $Q_b(\lambda, \bar{b}_s)$만 최대화되면 된다. 여기서 S를 시스템에서 모든 상태 분포 세트로 정의하고, $r_s(t)$를 시간 t에서 상태 s의 총 아커페이션 (total occupation) 확률로 정의하면 다음과 같이 표현된다.

$$r_s(t) = \frac{1}{F(O \mid \lambda)} \sum_{\theta \in \Theta} F(O, \theta_t = s \mid \lambda) \quad (9)$$

$$Q_b(\lambda, \bar{b}_s) = F(O \mid \lambda) \sum_{t=1}^{T} r_s(t) \log \bar{b}_j(o_t) \quad (10)$$

$\bar{W}_s$에 관해서 $Q(\lambda, \bar{\lambda})$을 최대화하기 위해서 $\bar{W}_s$로 미분을 하면

$$\frac{dQ(\lambda, \bar{\lambda})}{d\bar{W}_s} = \frac{d}{d\bar{W}_s}\left[\sum_{k \in S} Q_b(\lambda, \bar{b}_k)\right] = \frac{d}{d\bar{W}_s} Q_b(\lambda, \bar{b}_s) = 0 \quad (11)$$

적당한 표기법을 이용하여 정리하면 다음 식과 같다.

$$\sum_{t=1}^{T} r_s(t) C_s^{-1} o_t \mu_s^T = \sum_{t=1}^{T} r_s(t) C_s^{-1} \bar{W}_s \hat{\mu}_s \mu_s^T \quad (12)$$

위 식이 $\bar{W}_s$의 최적화된 일반식이 된다. 각 분포가 분리된 회귀 변환을 갖고 있다면 적응된 평균에 대한 값은 쉽게 유도된다. 즉,

$$\bar{\mu}_s = \bar{W}_m \hat{\mu}_s = \frac{\sum_{t=1}^{T} r_s(t) o_t}{\sum_{t=1}^{T} r_s(t)} \quad (13)$$

이고 이것은 평균 벡터에 대한 표준 최대 가능성 재추정 식이라고 한다.

MAP 적응은 Bayesian 적응이라고도 하며 모델 파라미터 분포에 대해서 사전 지식을 사용하는 것이 MLE와 차이점이다. λ에 대한 MLE는 다음과 같은 식으로 구해진다[6,7].

$$\frac{\partial}{\partial \lambda} P(o_1 \cdots o_T \mid \lambda) = 0 \quad (14)$$

여기서 λ가 priori 분포 함수, $P_0(\lambda)$를 갖는 랜덤함수라고 가정을 하면 MAPE는 다음 식을 푸는 것이 된다.

$$\frac{\partial}{\partial \lambda} P(\lambda \mid o_1 \cdots o_T) = 0 \quad (15)$$

Bayes theorem을 사용하여 다시 쓰면

$$P(\lambda \mid o_1 \cdots o_T) = \frac{P(o_1 \cdots o_T \mid \lambda) P_0(\lambda)}{P(o_1 \cdots o_T)} \quad (16)$$

식 (15)를 만족시키는 MAPE를 구하기 위해서 최소 Bayes risk를 갖도록 한다. 상태를 갖는 시스템에서 다시 생각을 해보면

$$P(O \mid \lambda) = \sum_{\theta} P(O, \theta \mid \lambda) \quad (17)$$

MLE의 상태 최적 가능성은 다음과 같다.

$$\bar{\lambda} = \arg \max_{\lambda} [\max_{\theta} P(O, \theta \mid \lambda)] \quad (18)$$

이고 이 식을 풀기 위해서 분절 k-means 알고리듬에 의한 반복식으로 구할 수 있다. MAPE에 대해서는 다음과 같이 된다.

$$\frac{\partial}{\partial \lambda} P(\lambda, \theta \mid O) = 0 \quad (19)$$

$$P(\lambda, \theta \mid O) = \frac{P(O, \theta \mid \lambda) P_0(\lambda)}{P(O)} \quad (20)$$

$$\theta = \arg \max P(O, \theta \mid \lambda) P_0(\lambda) \quad (21)$$

$$\bar{\lambda} = \arg \max_{\lambda} P(O, \bar{\theta} \mid \lambda) P_0(\lambda) \quad (22)$$

식 (21)과 (22)를 이용하여 MAPE를 풀게 된다. MAPE에서 중요한 것은 사전 지식을 선택하는 것이다. 우리가 추정되는 파라미터들에 관한 지식을 갖고 있다면 그러한 사전 지식을 priori 분포에 포함시킬 수 있게 되고 이러한 것을 유익한 사전 지식 (informative priori)이라고 한다. 일반적으로 prior 분포의 선택은 주어진 데이터를 특징지울 수 있게 사용되는 음향학적 모델들에 의존되며 이전의 경험, 데이터의 물리학적 중요성 그리고 수학적 접근을 토대로 구해진다. 가우시안 평균의 MAP 적응은 다음과 같은 식으로 표현된다[8].

$$\bar{\mu}_{MAP} = \frac{n\tau^2}{\sigma^2 + n\tau^2} \bar{y} + \frac{\sigma^2}{\sigma^2 + n\tau^2} v \quad (23)$$

앞의 식은 접합 사전 지식을 사용하고 평균 v와 분산 τ^2을 가질 때 유도된다. n은 HMM 상태에서 관찰되는 훈련 샘플들의 총 수이고 $\bar{y}$는 표본 평균이다. 더이상의 훈련 대

이터가 없으면 n이 0이 되고 추정되는 단순한 priori 평균이 된다. 또 훈련 샘플들이 상당히 많아지면 ($n\to\infty$) MAP 추정되는 MLE에 수렴하는 식이 된다. 식 (23)에서 priori 파라미터들, v, τ^2, σ^2은 상태당 많은 수의 mixtures를 갖는 화자 종속 모델들이나 단일 화자 독립 모델을 통해서 추정된다.

$$v=\sum_{m=1}^{M} w_m v_m \tag{24}$$

$$\tau^2+\sum_{m=1}^{M} w_m (v_m-v)^2 \tag{25}$$

여기서 w_m은 m번째 믹스쳐 구성요소에 할당되는 weight이고 v_m은 m번째 믹스쳐 구성요소의 평균이다.

III. 제안된 알고리듬

비터비 알고리듬은 HMM에서 주어진 관측 열에 대한 가장 최적화된 상태 열을 찾기 위한 알고리듬이다. 비터비 알고리듬이 최적화된 상태 열을 찾는 간단한 예를 그림 3에서 나타내고 있다. 우선 각 관측 시간에서 최고 확률을 계산한다. 이것을 마지막까지 수행을 한 후 다시 역추적을 통하여 최적의 상태 열을 찾는 것이다[9,10].

화자 적응 알고리듬은 화자 독립 모델에 어떤 특정한 화자의 데이터를 적용하여 그 화자 종속 모델로 변환한다. 또 MLLR과 MAP 적응 알고리듬은 각각 수식에서도 알 수 있듯이 주어진 적응 데이터의 양에 따라서 그 성능에 차이가 있다. MLLR 적응은 회귀 클래스 트리를 사용하여 비교적 적은 양의 데이터로도 비교적 좋은 성능을 나타내지만 MAP 적응은 다량의 데이터가 요구된다. 그러나 충분한 양의 데이터가 제공될 경우 MAP 적응의 성능은 MLE에 수렴하지만 MLLR 적응의 성능은 그렇지 않다. 화자확인 시스템의 특성상 훈련 데이터와 테스트 데이터는 소량이다. 그러므로 MAP 적응 알고리듬보다는 MLLR 적응 알고리듬을 이용하는 것이 화자확인 시스템의 성능 향상에 유리하다. 제안된 알고리듬의 중요한 것은 적응 알고리듬을 훈련하는 부분에 대해 사용하는 것이 아니고 디코딩 부분에 대해 사용한다는 것이다. 제안된 알고리듬과 비터비 알고리듬을 적용과정을 그림 4에서 블록 다이어그램을 이용하여 나타내고 있다. 객관성을 유지하기 위해 디코딩을 하고 스코어를 계산하는 부분을 제외하고 나머지 부분에 대해서는 동일하게 수행을 하였다.

우선 사용자 (client)의 데이터로 훈련시킨다. 그 후 디코딩 부분에서 사용자의 테스트 데이터와 사칭자 (imposter)의 테스트 데이터로 화자 적응 알고리듬을 적용한다. 그러면 사용자의 테스트 데이터로 적응된 모델은 변화가 크지 않고 사칭자의 테스트 데이터로 적응된 모델은 훨씬 많이 변화한다. 즉, 원래 모델의 평균 파라미터들은 테스트에 사용된 데이터들이 사용자의 것이냐 아니면 사칭자의 것이냐에 따라서 변화된다. 따라서 이 변화된 양을 적당한 방법으로 측정을 하면 사칭자의 테스트 데이터들의 값과 같은 화자의 테스트 데이터들의 값은 많은 차이를 보이므로 이것을 새로운 스코어로 사용하였다. 측정 방법으로는 벡터 곱, 스칼라 곱, 기하학적 거리 등 여러 가지가 있지만 여기서는 기하학적 거리를 사용하였다. 그림 6은 제안된 알고리듬에서 사용자의 데이터와 사칭자의 데이터간에 적응 후에 차이를 보여주고 있다.

여기서 제안된 알고리듬과 비터비 알고리듬을 비교하여 성능이 향상될 수 있는 원인은 다음과 같이 두 가지로 말할 수 있다. 첫째, 적응 알고리듬이 주어진 데이터를

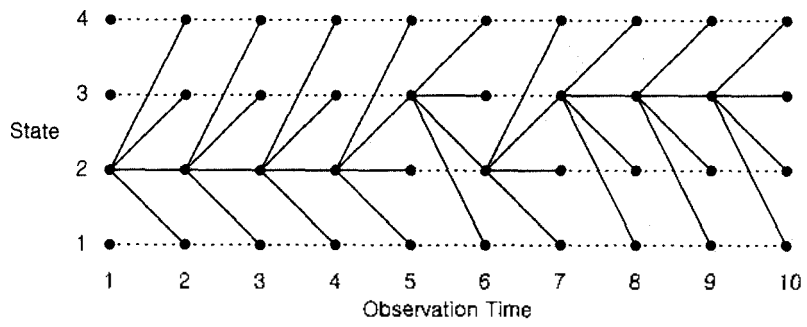

그림 4. Viterbi 알고리듬
Fig. 4. The example of viterbi algorithm.

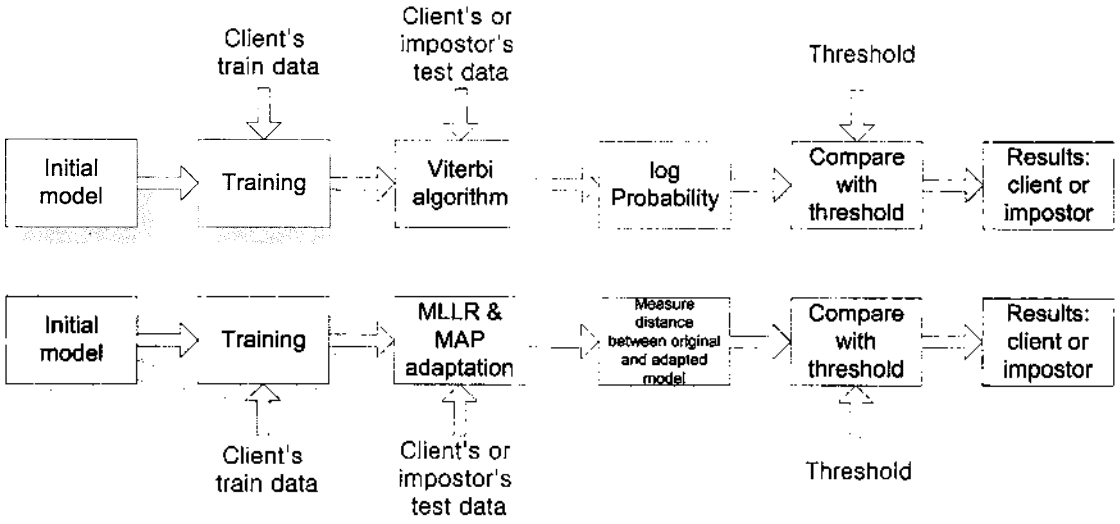

그림 5. 제안된 알고리듬의 블록 다이어그램
Fig. 5. The block diagram of the proposed algorithm.

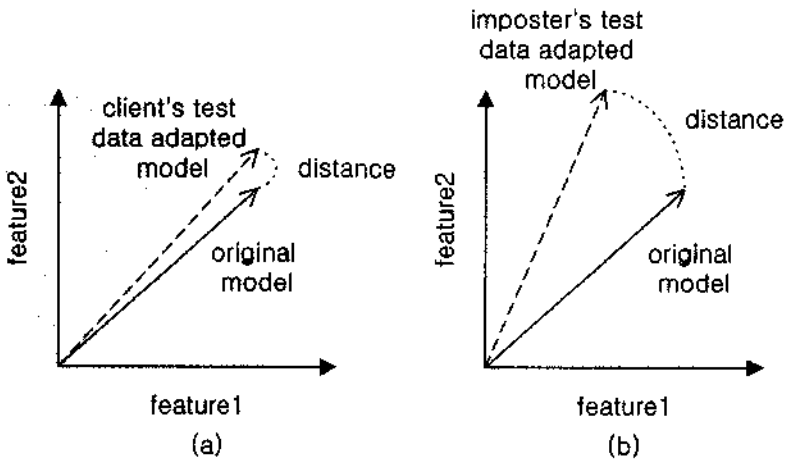

그림 6. 테스트 데이터를 이용한 적응 결과
Fig. 6. The results of adaptation with test data.

갖고 디코딩할 때 화자의 특성을 잘 적용하여 비터비 알고리듬이 출력하는 스코어보다 신빙성 있는 결과를 출력한다. 즉 사칭자의 데이터에 대한 디코딩 결과에서는 비터비 알고리듬보다 더 높은 스코어를 출력하고 사용자의 데이터에 대한 것은 오히려 더 낮게 출력된다는 것이다. 둘째는 비터비 알고리듬은 가장 최적의 상태 경로를 찾고 거기에 맞는 로그 확률을 계산하고 이 로그 확률을 화자확인 시스템의 스코어로 사용한다. 그러나 비터비 알고리듬상에서 이 스코어를 계산하는 과정에서 사용자의 테스트 데이터를 이용하여 적용할 경우는 비터비 알고리듬과 제안된 알고리듬 사이에 차이가 없지만 사칭자의 테스트 데이터로 적용을 할 경우는 테스트 데이터를 이용하여 추정되는 상태의 위치가 상당히 다르게 변화가 되는 경우가 발생한다. 이러한 경우 비터비 알고리듬은 입력된 관측 벡터와 가장 가까운 곳에 위치한 상태로부터 스코어를 계산하게 된다. 그러나 제안된 알고리듬에서는 그런 경우에도 가까운 상태와의 스코어를 계산하는 것이 아니고 원래의 상태와 급격히 변화된 상태의 거리를 계산을 하기 때문에 보다 큰 스코어를 보이게 된다. 그림 7에 위에서 설명된 경우에 대하여 예를 보여주고 있다.

IV. 음성 데이터베이스

본 논문에서 사용한 음성 데이터베이스는 남자 53명, 여자 53명, 전체 106명에 대해 수집하였다. 이중 남자 22

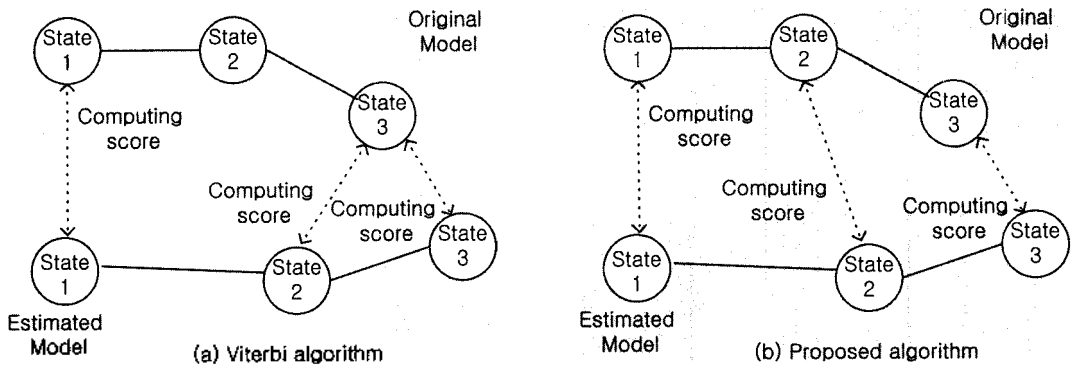

그림 7. 두 알고리듬에서의 스코어 계산
Fig. 7. Score computation of two algorithms.

명과 여자 20명은 동일한 음성, "범일 정보통신입니다." 라고 발성하였고 남자 26명, 여자 33명은 다른 음성, "○○○ (이름)은 ○○월 ○○일 (몇 월 몇 일)입니다."라고 발성하였다. 또한 시간에 따른 화자의 발성 변화를 포함하도록 날짜별로, 요일별로, 오전, 오후, 저녁으로 나누어 3개월간 수집하였다 (오전 : 9시 ~ 11시, 오후 : 2시 ~ 5시, 저녁 : 7시 ~ 10시). 각각의 어구에 대해 화자별로 훈련 데이터는 6 세션으로 구성되어 있고, 각 세션당 6개의 훈련 데이터가 있다. 테스트 데이터는 화자별로 최대 20 세션으로 구성되어 있고, 각 세션당 4회의 테스트 데이터가 있다[11].

음성 데이터를 각 화자가 있는 직장이나 가정에서 전화기로 전화선을 통하여 보내고 서울의 범일 정보통신 본사에 있는 시스템에서 수집하였다. 단, 이동전화나 공중전화는 사용하지 않도록 하였다. 그러나 실제생활을 하다가 전화를 하였기 때문에 주변 소음이나 전화기 혹은 전화선에서의 무수히 많은 잡음이 함께 수집되었기 때문에 실제 환경과 같은 데이터베이스를 구성하려고 노력했다. 수집된 데이터의 형태는 8 KHz 8 bit μ-law 데이터이다.

V. 실험 방법

실험에 사용된 데이터베이스가 실제 환경과 동일한 환경에서 제작되었기 때문에 묵음 (silence) 모델을 사용하는 것보다는 성능이 양호한 EPD (End-Point Detection) 처리한 것을 선택하였다. 채널 데이터에서 주로 사용되는 MFCC (Mel Frequency Cepstral Coefficient)를 사용하고 채널 성분의 제거를 위해서 CMS (Cepstral Mean Subtraction)처리를 하였다. 전화선 채널이기 때문에 300 ~ 3400 Hz의 BPF (Band Pass Filtering)를 하였고 25 ms의 허밍 창 (Hamming windowing)을 하였다. 타겟 속도는 5 ms로 하였다. 그리고 인식 시스템으로는 HTK3.0을 사용하였다. 상태의 수는 24개에서 48개까지 변화를 주었고 mixture의 수는 3~8개로 하였다. 적응은 특징 벡터 중 평균과 분산을 모두 적응을 하지만 화자의 특성은 일반적으로 평균 벡터가 갖고 있기 때문에 분산 벡터는 사용하지 않았다.

우선 남자 화자 20명과 여자 화자 15명을 랜덤하게 선택을 하여 이들의 훈련 세션에서 EPD가 양호하게 된 것들을 골라서 한 세션의 6개 중에서 4개를 이용하여 훈련을 하였다. 테스트에 사용된 데이터는 106명 모든 화자에 대해서 각각 한 화자에서 랜덤하게 한 개만을 선택해서 사칭자 데이터로 하고 이와 같은 세트로 3개를 구성하여 인식결과의 평균을 구하였다. 사용자 데이터로는 훈련에 사용되지 않은 데이터와 테스트 세션에 있는 데이터들을 사용하였다.

성능 평가의 기준으로는 사용자를 거부하는 오인 거부율, FRR (False Rejection Rate)과 사칭자를 수락하는 오인 수락율, FAR (False Acceptance Rate)이 같아지는 지점인 EER (Equal Error Rate)을 사용하였다.

실험을 크게 비터비 디코딩과 제안된 알고리듬에 대해서 구분을 하였고 또 각각은 델타계수의 유무에 대해서 실험을 하였다. 그리고 제안된 알고리듬에서는 MLLR 적응만 사용하였을 경우와 MLLR과 MAP 적응을 모두 사용하였을 때도 실험을 실시하였다.

VI. 실험결과

위와 같은 실험방법으로 실시한 결과를 표 1에 정리하였다. 표 1에서 보듯이 MLLR만 사용하여 제안된 알고리

듬을 구현한 경우의 실험에서 최고의 EER 4.338%가 나왔고 MAP를 동시에 사용을 하여도 성능저하를 초래하는 것을 볼 수 있다. 이것은 적응 알고리듬의 수학적 고찰에서도 보았듯이 이번 실험과 같은 극소량의 데이터를 이용할 경우에는 MAP 적응이 비교적 안정적이고 믿을 수 있는 성능을 발휘하기가 어렵고 오히려 MAP 적응을 사용하지 않고 MLLR 적응만 단독으로 사용하는 것이 유리하다는 것을 보여주는 결과이다.

그리고 위의 결과에서 최고의 인식률을 나타내는 비터비 디코딩과 제안된 알고리듬을 비교하여 인식률의 향상 정도를 표 2에 정리하였다. 비터비 알고리듬과 비교를 해보면 제안된 알고리듬이 남자 화자에 경우에서는 48.618%의 향상율을 보였고 여자의 경우는 17.176% 정도의 향상율을 보였다. 이것의 원인은 남자의 음성 데이터들이 여자의 것들보다 더 급격한 변화하는 파형을 갖기 때문에 군집화 과정에서 mixture를 분류하는 과정에서 더욱 다양한 분류가 가능하고 그만큼 인식할 때 스코어 값들 간에 차이도 현저하게 나기 때문이라고 추정할 수 있다.

제안된 방법이 평균 30%정도의 향상을 나타냈고 MAP와 MLLR을 동시에 사용을 하는 것보다는 MLLR만 사용하는 것이 좀 더 우수한 성능을 보이고 델타계수가 첨가됨으로써 더 높은 인식률을 나타내었다.

표 1. 실험정리
Table 1. Summary of experiments.

ALGORITHM	Men's AVR	Women's AVR	AVR
VITERBI (MFCC_D, 26)	4.711	8.323	6.259
PROPORSED (MFCC_D, 26, MLLR ONLY)	2.420	6.894	4.338
PROPORSED (MFCC_D, 26, MLLR+MAP)	2.472	6.937	4.385
PROPORSED (Training 과정은 동일, Decoding 과정에서 DELTA계수 제외 계산, MLLR ONLY)	2.441	6.965	4.380
VITERBI (MFCC, 13)	5.031	8.913	6.972
PROPORED (MFCC, 13)	3.367	6.810	4.842

표 2. 최고 인식률 비교
Table 2. Comparison of the best EER.

ALGORITHM	VITERBI ALGORITHM	PROPOSED ALGORITHM	DIFF	DIFF(%)
MEN'S SPEAKER AVE	4.711	2.420	2.291	48.618
WOMEN'S SPEAKER AVE	8.323	6.894	1.429	17.176
TOTAL AVE	6.259	4.338	1.921	30.699

실험을 실시하는 과정에서 비터비 디코딩을 사용하여 인식을 할 경우 프레임이 길어지고 짧아짐에 따라서 스코어 값이 영향을 받기 때문에 최종 스코어는 프레임에 대한 정규화를 실시하였다. 그리고 인식하고자 하는 발성의 길이가 길어지면 제안된 알고리듬에서 사용되는 화자적응이 더 잘 된다. 인식하고자 하는 데이터가 사용자의 것이라고 가정을 하면 인식 결과에는 큰 변화가 없고 사칭자의 것이라면 발성의 길이에 영향을 받게 된다. 그러나 발성이 지나치게 짧은 것과 지나치게 긴 것을 실험에 사용하지 않았기 때문에 사칭자를 구별하게 되는 일정한 문턱 값에 의하여 사칭자라는 구별하는 것에는 어려움이 없다. 그러나 더욱 다양한 실험을 하게 되면 길이가 길면 길수록 더욱 분별력 있는 결과를 얻게 된다. 그러므로 실제 인증 시스템에 적용을 할 경우에는 지나친 길이의 차가 나는 발성이 들어오면 재요청을 하는 시스템으로 설계가 되면 이런 현상을 막을 수 있다.

VII. 결 론

본 논문에서는 기존에 화자확인 디코딩에서 사용하는 비터비 알고리듬을 대신하는 새로운 알고리듬을 제안을 하였다. 화자확인 시스템의 특성상 많은 데이터를 구할 수 없다는 것과 화자의 특성을 최대한 잘 표현할 수 있는 알고리듬 개발에 초점을 맞추었다. 그래서 음성인식에서 사용이 되고 있는 화자 적응 알고리듬이 화자의 특성에 따라 모델 파라미터를 변환하는 것을 응용하여 새로운 디코딩 알고리듬을 제안하였다. 우리는 적응 알고리듬 중 MLLR과 MAP 적응 알고리듬을 사용하여 실험을 실시하였다. 수학적 고찰에서도 확인한 바와 같이 화자확인 시스템에서는 MLLR이 MAP보다 양호한 성능을 나타냈다. MLLR은 회귀 클래스 트리를 이용하여 주어진 테스트 데이터를 최대한 이용하여 화자의 특성을 잘 표현할 수 있게 동작하였다. 동일한 환경에서 동일한 데이터들을 이용하여 디코딩 방법을 달리한 결과 일반적인 비터비 단어 인식기를 사용하였을 때보다 제안된 알고리듬을 사용한 새로운 인식기의 결과가 남자 화자의 경우는 50% 여자 화자의 경우는 20%가 넘는 EER 향상을 보였고 평균적으로는 약 30%의 EER의 향상을 나타냈다.

본 논문은 어구종속 시스템에 대해서만 실험을 실시하였다. 또한 실험을 위해서 만들어진 음성 데이터베이스는 연구실에서 이번 연구를 위해서 제작한 것이다. 그러므로 제안된 알고리듬의 보다 공정한 결과를 추론하는

것에 문제가 있을 수 있다. 그러므로 앞으로의 과제로는 현재 화자확인 시스템을 구성하고 테스트에 널리 쓰이고 있고 NIST에서 제작된 YOHO 데이터베이스를 이용하여 어구지시 화자확인 시스템에 대해서도 지속적인 연구를 해야 할 것이다. 또 이번 실험에서는 채널 정규화를 제외한 스코어 정규화에 대해서는 실험을 실시하지 않았다. 따라서 어구지시 숫자 음을 이용한 화자확인 시스템 실험을 할 경우 스코어 정규화에 관해서도 심각한 문제점이 생길 것이다. 비터비 알고리듬은 스코어 정규화로 cohort 모델 정규화 방법과 world 모델 정규화 방법 등 여러 가지를 적용해 볼 수 있지만 제안된 알고리듬의 경우는 확률과는 차이가 있기 때문에 그와 같은 정규화 방법을 적용하는데 무리가 있다. 그러므로 스코어 정규화 방법의 접근도 하나의 과제가 될 것이다.

감사의 글

본 연구는 정보통신부의 2001년도 대학기초연구비 지원에 의하여 진행되고 있습니다.

참 고 문 헌

1. S. J. Ahn, S. M. Kang and H. S. Ko, "Effective speaker adaptations for speaker verification," *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings, 2000 IEEE International Conference*, vol. 2, pp. 1081-1084, 2000.
2. P. C. Woodland and M. J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression," *ICSLP 96., Proceedings, Fourth International Conference*, vol. 2, pp. 1133-1136, 1996.
3. M. J. F. Gales and P. C. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," *ICSLP 96., Proceedings, Fourth International Conference*, vol. 3, pp. 1832-1835, 1996.
4. C. J. Leggetter and P. C. Woodland, "Speaker adaptation of HMMs using linear regression," *CUED, F-INFENG, TR*. 181, June 1994.
5. T. Hain, P. C. Woodland, T. R. Niesler and E. W. D. Whittaker, "The 1998 HTK system for transcription of conversational telephone speech," *Acoustics, Speech, and Signal Processing, 1999. Proceedings, IEEE International Conference*, vol. 1, pp. 57-60, 1999.
6. J. L. Gauvian and C. H. Lee, "Maximum a Poster ori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains," *IEEE Transactions on Speech and Audio processing*, vol. 2, no. 2, pp. 291-298, 1994.
7. C. H. Lee and C. H. Lin, "A study on speaker adaptation of the parameters of continuous density Hidden Markov Models," *IEEE Transactions on signal processing*, vol. 39, no. 4, April 1991.
8. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification 2nd Edition*, WILEY, chap. 3, pp. 92-95, 2000.
9. A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260-269, 1967.
10. J. K. Omura, "On the Viterbi Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 15, no. 1, pp. 177-179., 1969.
11. 조태현, 김유진, 이재영, 정재호, "전화선 채널이 화자확인 시스템의 성능에 미치는 영향," 한국음향학회지, 제18권, 제5호, pp. 12-20, 1999.

저자 약력

● 김 강 열 (Gang Yeoul Kim)

2000년: 인하대학교 전자공학과 공학사
2000년~현재: 인하대학교 정보통신대학원 석사과정
※ 주관심분야: 음성인식, 화자확인, 화자적응

● 김 지 운 (Ji Un Kim)

1998년: 인하대학교 전자공학과 공학사
2000년: 인하대학교 전자공학과 공학석사
2000년~현재: 인하대학교 전자공학과 박사과정
※ 주관심분야: 음성인식, 화자확인, 화자적응

● 정 재 호 (Jae Ho Chung)

1982년: University of Maryland (BSEE)
1984년: University of Maryland (MSEE)
1990년: Georgia Institute of Technology (Ph. D.)
1984~1985년: 미국 국방성 산하 해군 연구소, 신호처리실, 연구원
1991~1992년: AT&T Bell Labs, 음성신호처리 연구실, 연구원 (MTS)
1992년~현재: 인하대학교 공과대학 전자공학과, (현)정교수
※ 주관심분야: 음성코딩, 음성인식, 화자확인, 화자적응