상태당 가지수를 가변시킨 HMM을 이용한 화자적응화에 관한 연구

A Study on the Speaker Adaptation in HMM Using Variable Number of Branches in Each State

김 광 태*, 서 정 일**, 한 유 수**, 홍 재 근**
(Kwang Tae Kim*, Jeong Il Seo**, Yoo Soo Han**, Jae Keun Hong**)

요 약

본 논문에서는 CHMM인 CDHMM과 ARHMM을 이용하여 화자적응화 하는 방법을 각각 연구하였다. CDHMM에서는 최대사후확률 추정법에 의하여 각 상태마다 하나의 가지를 이용하여 화자에 적응시킨다. 본 논문에서는 음성의 다양한 음향학적 특징을 표현하기 위하여 상태마다 여러 개의 가지를 갖는 방법을 제안하였다. 상태마다의 적절한 가지 수를 결정하기 위하여 각 상태에 속하는 프레임 수와 특징벡터들의 분산행렬의 행렬식값을 이용하였다. ARHMM에서는 특징벡터로 선형예측계수를 사용하기 때문에 최대사후확률 추정법을 사용할 수 없게 된다. 따라서 화자독립모델을 이용하여 적응화자에 대한 음성을 Viterbi 알고리듬으로 상태별로 분할한 후 k-means 알고리듬을 이용하여 각 상태마다 하나의 가지를 갖는 모델로 적응시키는 방법을 제안하였다.

ABSTRACT

In this paper, we have studied the method of speaker adaptation using CDHMM and ARHMM in CHMM respectively. In CDHMM, speaker adaptation had been performed using one branch in each state by the method of MAPE (maximum a posteriori estimation). In this paper, we proposed the method using variable branches to represent properly various speech information of the speaker in each state. We determined the number of branches in each state depending on the number of frames and the determinant of the variance matrix in the state. In ARHMM, because the feature vector is used as the components of LPC vector, the MAPE method could not be used. So, we proposed the method of ARHMM to adapt the speaker adaptation model with one branch in one state. The input utterance was divided into each states by Viterbi algorithm using speaker independent model and then transformed into a typical vector by the k-means algorithm.

Ⅰ. 서 론

HMM(hidden Markov model)을 이용한 음성인식기는 훈련에 참가한 화자와 인식기를 사용하는 화자에 따라 화자종속(speaker dependent)인식기와 화자독립(speaker independent)인식기로 나눌 수 있다. 화자종속인식기는 특정화자에 의해 훈련된 모델을 이용하여 훈련에 참가한 화자가 사용하는 인식기로서 훈련음성이 충분하다면 화자독립인식기에 비해 인식성능이 항상 우수하다. 그러나 훈련에 참가하지 않은 화자가 발음하였을 경우 인식성능이 급격히 저하되며 인식시스템의 사용자가 바뀌었을 경우 모델을 다시 훈련시켜야 하는 번거로움이 있다.

화자적응(speaker adaptation)방법은 소량의 훈련데이터를 사용하여 충분한 훈련데이터로 훈련된 화자독립모델을 인식시스템으로 사용하려는 것으로 특정화자에 적응시키는 방법이다[1, 2].

CDHMM(continuous density hidden Markov model)[3]을 사용하는 인식기에서는 최대사후확률 추정(MAPE; maximum a posteriori estimation) 방법 즉, 베이적응(Bayesian adaptation)[4, 5] 방법을 이용하여 화자적응을 시킨다. CDHMM에서 베이적응 방법을 이용하여 화자에 적응된 모델을 만들 때, 각 상태마다 하나의 가지를 갖는 모델로 만들어 적응시킨다. 이것은 이미 만들어져 있는 인식기의 특정상태에 속하는 각 가지들의 평균값과 분산값들의 분포를 적응시킬 데이터의 사전분포로 이용하는데, 이 분포를 각 상태마다 한 개밖에 구할 수가 없기 때문이다. 그러나 각 상태마다 하나의 가지로 나타내면 화자의 다양한 음성정보를 적절히 나타내지 못하여 적응에 한계를 나타내게 된다. 본 논문에서는 CDHMM을 이용한 화자적응시에 화자의 다양한 음성정보를 잘 나타내기 위하여 상태마다 여러 개의 가지를 사용하는 방법을 제안한다.

*상주산업대학교 전자전기공학과
**경북대학교 전자전기공학부
접수일자 : 1998년 2월 28일

이때 훈련데이터 수가 적기 때문에 어떤 상태에서는 여러 개의 가지를 사용하는 것이 타당하지 않을 수도 있으므로 각 상태에서 적절한 가지 수를 결정하여야 한다. 본 논문에서는 각 상태에 속하는 프레임 수에 따라 가지 수를 달리하는 방법과 상태내의 특징벡터들의 분산행렬의 행렬식값을 이용하는 방법을 사용하였다. 제안한 방법을 이용하여 15개의 한국 지역명으로 구성된 음성데이터와 40개의 단어로 구성된 ETRI의 샘돌이 데이터에 대해 인식실험한 결과 제안한 방법이 가지를 한 개 사용했을 때에 비해 높은 인식률을 얻을 수 있었다.

CDHMM의 경우에는 MAPE 방법을 이용하여 화자적응을 하지만 가우스 분포를 선형예측계수값(linear predictive coefficient)의 자기상관계수를 이용해 표현하는 ARHMM (autoregressive hidden Markov model)[6]에서는 각각의 선형예측계수 값들이 상호 의존적이므로 베이적응 방법을 사용할 수 없게 된다. 따라서 본 논문에서는 음성 데이터의 각 상태마다의 선형예측계수값의 자기상관계수값의 평균값만을 그 화자에 적응시키는 방법을 사용하였다. 한국어 지역명 음성데이터를 이용하여 제안된 방법으로 화자적응을 수행하여 오인식률이 감소함을 확인하였다.

Ⅱ. CDHMM 파라미터들의 베이적응

정규분포의 평균과 분산에 대하여 베이적응방법을 사용하여 단순 좌우구조 CDHMM에 대하여 전개한다. 대각성분값만을 갖는 분산 행렬을 가지는 정규분포에 대해서만 고려한다. 적응은 평균과 분산에 대해 각 상태마다 독립적으로 이루어지므로 적응 수식은 하나의 상태에 대해서만 기술한다.

2.1 CDHMM의 화자적응방법

2.1.1 평균의 Bayes 적응

평균 μ가 사전분포 $P_0(\mu)$를 가지는 불특정값이고, 분산 σ^2이 상수일 때, $P_0(\mu)$가 평균 ν와 분산 τ^2을 가지는 정규분포라고 가정하면 μ의 MAP 추정치는 다음과 같다.[4]

$$\hat{\mu}_{MAP}=\frac{n\tau^2}{\sigma^2+n\tau^2}\ \bar{y}+\frac{\sigma^2}{\sigma^2+n\tau^2}\ \nu \tag{1}$$

여기서 n은 훈련데이터의 개수이고, $\bar{y}$는 샘플 데이터의 평균이다.

사전분포의 평균 ν와 분산 τ^2은 다음과 같이 추정된다.

$$\nu=\sum_{m=1}^{M} w_m\nu_m \tag{2}$$

$$\tau^2=\sum_{m=1}^{M} w_m(\nu_m-\nu)^2 \tag{3}$$

여기서 ν_m, w_m은 각각 화자독립모델의 m번째 가지의 평균과 가중치이다.

σ^2은 다음 식과 같이 각 가지의 가중된 분산을 사용하여 구할 수 있다.

$$\sigma^2=\sum_{m=1}^{M} w_m\sigma^2_m \tag{4}$$

여기서 σ^2_m은 m번째 가지의 분산이다.

2.1.2 분산의 Bayes 적응

평균 μ가 미지의 값이고, 분산의 사전분포가 σ^2_{min} 이상에서 일정한 값이라면, 분산 σ^2의 MAP 추정치는 다음과 같다.[7]

$$\hat{\sigma}^2_{MAP}=\begin{cases} S_y^2 & S_y^2\geq\sigma^2_{min} \\ \sigma^2_{min} & o.w. \end{cases} \tag{5}$$

여기서, σ^2_{min}값은 화자독립모델로부터 추정되며, 또 평균에 대한 사전정보는 없으므로 샘플 평균 $\bar{y}$로 평균 μ를 추정한다. 샘플의 분산을 나타내는 S_y^2는 다음 식과 같다.

$$S_y^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n} \tag{6}$$

여기서 n은 샘플 수이다.

2.1.3 평균과 분산의 Bayes 적응

평균과 분산 모두 어떤 사전분포를 가지는 불특정값이라고 가정하고 적응된 모델을 구할 수 있다. 이때 분산의 역수값인 precision($\theta=1/\sigma^2$)을 감마분포로 가정하여 사용할 수 있다.

평균과 precision 파라미터가 불특정값이고 사전분포 $P_0(\mu,\theta)$를 joint 정규감마분포로 가정하면 평균과 분산의 MAP 추정치는 다음과 같다.[5]

$$\hat{\mu}_{MAP}=\frac{\frac{\sigma^2}{\tau^2}\nu+n\bar{y}}{\frac{\sigma^2}{\tau^2}+n} \tag{7}$$

$$\hat{\sigma}^2_{MAP}=\frac{1+\frac{n}{2}S_y^2+\frac{n\frac{\sigma^2}{\tau^2}(\bar{y}-\nu)^2}{2(\frac{\sigma^2}{\tau^2}+n)}}{\frac{1}{\sigma^2}+\frac{n}{2}} \tag{8}$$

2.2 상태마다 여러 개의 가지를 사용하는 방법

CDHMM에서 화자에 적응된 모델을 만들 때, 각 상태마다 하나의 가지를 가지는 모델로 만들어 적응시킨다. 그러나 상태마다 하나의 가지로는 적응시키려는 화자의 다양한 음성정보를 적절히 나타내지 못하기 때문에 모델을 그 화자에 적응시키는데 한계를 가지게 된다. 이점을 해결하기 위하여 상태마다 여러 개의 가지를 사용하는

방법을 생각해 볼 수 있다.

분산만을 베이적응시키는 방법에서는 사전분포로서 분산의 하한값만을 이용하기 때문에 상태마다 여러 개의 가지를 갖게 하는 방법을 적용하기가 용이하다.

2.2.1 프레임 수에 따른 가지 수 결정방법

여러 개의 가지를 갖게 하기 위하여 입력 벡터열을 k-means 알고리듬을 사용하여 몇 개의 클러스터(cluster)로 분리한 후 이 각각에 대한 데이터 분산인 S^2_{ym}을 구한다. 이 값과 분산의 하한값 σ^2_{min}을 이용하여 각 가지의 적응된 분산을 다음과 같이 구한다.

$$\hat{\sigma}^2_{MAP,m} = \begin{cases} S^2_{ym}, & S^2_{ym} \geq \sigma^2_{min} \\ \sigma^2_{min} & o.w. \end{cases} \tag{9}$$

그러나 프레임 수가 많은 상태에서는 여러 개의 가지를 사용하는 방법이 타당하지만 프레임 수가 적은 상태에서는 여러 개의 가지를 사용할 수가 없다. 특히 화자적응에서는 훈련데이터 수가 적기 때문에 상태에 속하는 프레임 수가 적은 경우가 많이 생긴다. 그래서 상태에 속하는 프레임 수에 따라 가지 수를 달리하는 방법을 사용하였다. 가지 수를 결정하는 수식은 다음과 같다.

$$m_j = \left\lceil \frac{N \times \sum_k n_{jk}}{\sum_k \sum_j n_{jk}} \right\rceil \tag{10}$$

여기서, m_j는 상태 j에서의 가지 수를, n_{jk}는 k번째 훈련음성의 상태 j에서의 프레임 수를 나타낸다. 그리고 N, M은 각각 모델의 상태 수, 평균 가지 수를 나타낸다.

이러한 방법으로 상태에 속하는 프레임 수가 많은 경우에는 가지 수를 많게 하고 상태에 속하는 프레임 수가 적은 경우에는 가지 수를 적게 하면 가지 수를 일률적으로 한 개로 할 때보다 프레임 수가 많이 불리는 상태에서의 분포를 좀 더 세밀히 나타낼 수 있기 때문에 적응시키고자 하는 화자의 특성을 잘 나타낼 수 있게 된다.

상태에 속하는 프레임 수가 많다고 하여 무조건 많은 가지를 사용하는 것은 좋지가 않으며 프레임 수에 따라 사용할 적절한 가지 수를 찾는 것이 중요하다. 즉, 식 (10)에서 사용할 평균 가지 수 M의 적절한 값을 찾아야 한다.

2.2.2 분산행렬에 따른 가지 수 결정방법

CDHMM에서 음성의 통계학적 특성은 평균과 분산에 의해 결정되는데, 분산값이 작은 상태보다 분산값이 큰 상태에서 보다 많은 가지 수를 갖게 하는 것이 타당하다.

관측벡터의 각 상태에서의 분산행렬은 다음과 같이 구해진다.

$$V_j(k) = \frac{1}{T} \sum_{i=1}^{T} (c_i(k) - \mu_j(k))^2 \tag{11}$$

$$j = 1, 2, \cdots, N, \quad k = 1, 2, \cdots, P$$

여기서 T는 상태 j에 속하는 프레임 수이고 c_i는 상태 j에서의 i번째 프레임의 켑스트럼 값이며 μ_j는 상태 j에서의 평균값이다. N과 P는 각각 상태수와 켑스트럼 차수를 나타낸다. 상태 j에서의 분산행렬의 행렬식값(determinant)은 다음 식으로 표현할 수 있다.

$$D_j = \prod_{k=1}^{P} V_j(k) \tag{12}$$

행렬식값이 작은 상태보다 행렬식값이 큰 상태의 가우스 분포가 넓게 퍼져있다고 볼 수 있으므로 행렬식값이 큰 상태일 때 보다 많은 가지 수를 갖도록 하였다. 최대의 행렬식값을 갖는 상태에서 최대의 가지 수를 갖게 하고 다른 상태에서는 최대 행렬식값과의 상대적인 비율로 가지 수를 결정하였다. 분산행렬의 행렬식값에 따른 가지 수를 결정하는 수식은 아래와 같다.

$$m_j = \left\lceil \frac{D_{max} M}{10 D_j} \right\rceil \tag{13}$$

여기서 D_{max}와 M은 최대 행렬식값과 최대 가지 수를 나타내며 최소한 m_j가 1이 되도록 즉 최소한 1개의 가지를 갖도록 각 상태에서의 가지 수를 결정하였다.

Ⅲ. ARHMM에서의 화자적응화

ARHMM은 상태내에서 가우스 분포를 나타내는 특징벡터가 평균과 분산이 아닌 선형예측계수의 자기상관계수를 사용한다. 선형예측계수는 성도를 전극필터(all-pole filter)라 가정하였을 때 전극필터들의 계수가 되므로 각각의 계수들 간에 밀접한 상관관계가 존재한다. 따라서 이들 중 한 개의 파라미터가 변한다면 전극필터 전체의 특성이 변하게 된다. 따라서 MAP를 사용하여 특징벡터를 화자에 적응시킬 수 없게 된다. 따라서, 본 논문에서는 불특정화자들에 의해 훈련된 화자독립모델을 이용하여 입력음성을 상태별로 나눈 후, k-means 알고리듬을 이용하여 대표되는 선형예측계수의 자기상관계수로 결정하는 방법을 통해 발성화자에 적응시키는 방법을 제안한다.

평균을 상태마다 하나로 나타내면 화자의 다양한 음성정보를 제대로 나타내지 못하므로 입력 벡터열을 k-means 알고리듬을 사용하여 몇 개의 클러스터로 분리한 후 이 각각에 대한 데이터 평균을 구하여 사용한다. 이때 식 (10)를 사용하여 상태마다 가지 수를 달리 하였다.

Ⅳ. 실험 결과 및 고찰

본 실험에서는 두 그룹의 음성데이터를 사용하여 적응화 실험을 하였다. 첫번째 음성데이터는 10명(남자 5명, 여자 5명)의 화자가 15개 한국 지역명을 10번씩 발음한 것으로 구성된 지역명 음성데이터이고, 두번째 음성데이터는 ETRI의 셈돌이 음성데이터로 40명(남자 20명, 여자 20명)의 화자가 40개의 단어를 4번씩 발음한 것이다. 셈돌이 데이터는 한국어 고립숫자와 고립단어들로 구성되어 있으며, 특징벡터로는 12차의 LPC 켑스트럼을 사용하였으며 HMM의 상태수는 지역명 음성데이터는 6개, ETRI 셈돌이 음성데이터는 4개로 하였다.

4.1 CDHMM에서의 화자적응

4.1.1 지역명 음성데이터

모든 실험에서 단순 좌우구조 HMM을 사용하였으며, 상태천이확률의 변화는 인식률에 별다른 영향을 못미쳐 화자독립모델의 확률값을 그대로 사용하였다.

화자독립 및 화자종속인식기의 평균 인식률은 각각 83.2%와 98.4%이다.

10명의 화자에 대해 6가지의 방법으로 적응한 경우의 평균 인식률을 표 1에 나타내었다.

표 1. 화자적응방법에 따른 평균 인식률(%)
Table 1. Recognition rate of speaker adaptation.(%)

Tokens	EXP1	EXP2	EXP3	EXP4	EXP5	EXP6
1	84.1	93.3	92.1	93.4	92.3	95.2
2	96.0	96.1	95.5	96.3	95.8	98.3
3	98.6	97.0	96.6	98.5	97.6	99.2

EXP1 : MLE 방법
EXP2 : 훈련데이터의 평균과 화자독립모델로부터 구한 분산 σ^2 (식 4) 사용
EXP3 : 화자적응된 평균(식 1)과 분산 σ^2 (식 4) 사용
EXP4 : 훈련데이터의 평균과 화자적응된 분산(식 5) 사용
EXP5 : 화자적응된 평균(식 7)과 분산(식 8) 사용
EXP6 : 가지마다의 훈련데이터 평균과 화자적응된 분산(식 9) 사용, 여기서 평균 가지 수 M을 2로 하였다.

표 1의 결과를 보면 다른 방법에 비해서 제안된 EXP6 방법이 토큰 수에 상관없이 가장 우수하다. MLE 훈련 방법을 사용한 EXP1에서는 훈련 토큰 수가 적을 때는 인식률이 많이 떨어짐을 알 수 있다. 이것은 MLE 방법으로 훈련할 경우 적은 훈련데이터를 가지고는 분산을 잘 추정할 수 없기 때문이다. 같은 수의 토큰을 사용했을 때 EXP6 방법이 MLE 방법을 사용한 EXP1보다 항상 높은 인식률을 나타내는 것을 볼 수 있다.

EXP6 방법으로 할 때, 평균 가지 수 M에 따른 인식률을 표 2에 나타내었다.

표 2. EXP6 방법에서 평균 가지 수 M에 따른 인식률(%)
Table 2. Recognition rate according to M in EXP6 method.(%)

Tokens	M			
	1	1.5	2	2.5
1	94.7	94.4	95.2	94.8
2	97.8	98.1	98.3	98.1
3	99.1	99.2	99.2	99.1

표 2에서 토큰 수에 상관없이 M이 2일 경우에 가장 적응이 잘되지만 M값에 따라 큰 차이는 없음을 볼 수 있다.

4.1.2 ETRI 음성데이터

제안한 방법의 범용성을 입증하기 위하여 ETRI의 셈돌이 데이터에 대해서 같은 방법으로 실험하여 그 결과를 표 3에 나타내었다. 셈돌이 데이터를 사용하였을 때의 화자독립인식률은 71.5%이다. 실험은 40명 중 남자 10명과 여자 10명을 한 그룹으로 하여 20명씩 두 그룹으로 나누어 첫 번째 그룹에 속하는 화자들의 데이터로 훈련하여 각 단어마다 모델을 만들고 두 번째 그룹의 화자 20명으로 인식실험을 하였다. 표 3의 적응 결과도 첫 번째 그룹에서 만든 모델을 두 번째 그룹에 속하는 화자들에 적응시킨 결과이다.

표 3. 화자적응방법에 따른 평균 인식률(%)
Table 3. Recognition rate of speaker adaptation.(%)

Tokens	EXP1	EXP2	EXP3	EXP4	EXP5	EXP6
1	67.1	89.5	89.5	89.1	89.9	90.3
2	86.7	92.6	92.8	93.3	93.5	94.3

표 3의 결과를 보면 지역명 데이터의 경우보다 적응률은 조금 떨어지지만 전반적인 양상은 비슷하다. 이 결과들로부터, 본 논문에서 제안한 방법이 일반성을 가짐을 확인할 수 있었다.

4.1.3 분산행렬의 따른 가지 수 결정방법

셈돌이 음성데이터에 대해 EXP6 방법으로 화자적응 인식실험을 수행하였다. 가지 수는 식 (13)를 이용하여 결정 하였으며 최대 가지 수 M은 3으로 하였다. 인식 결과를 기존의 하나의 가지를 사용하는 방법과 프레임 수에 따라 가지 수를 결정하는 방법의 결과와 함께 표 4에 나타내었다.

표 4. 화자적응방법에 따른 평균 인식률 (%)
Table 4. Recognition rate of speaker adaptation. (%)

Tokens	Method 1	Method 2	Method 3
1	67.1	90.3	88.4
2	86.7	94.3	92.5

표 4에서 Method 1은 기존의 한 개의 가지를 사용하는 방법이고, Method 2는 프레임 수에 따라 가지 수를 결정하는 방법이며 Method 3은 분산행렬의 행렬식값에 따라 가지 수를 결정하는 방법이다. 표 8에서 제안된 두가지 방법 모두 기존의 방법인 Method 1에 비하여 우수한 인식률을 나타내었다. 그러나, Method 3이 분산행렬의 행렬식값을 이용하여 행렬식값을 구하는 복잡한 과정에도 불구하고 Method 2에 비해 우수한 인식성능이 나타나지 않았다. 이는 분산행렬의 행렬식값만으로 분산행렬의 특징을 결정하기 어렵고 편차가 심한($10^{4\sim5}$) 행렬식값으로 가지 수를 결정하기란 매우 힘들기 때문이다. 따라서 프레임 수를 이용하여 가지 수를 결정하는 Method 2가 훈련과정시 오류가 생기는 것을 막을 수도 있으며 간단한 수식을 이용하여 효과적으로 가지 수를 결정할 수 있으므로 상태당의 가지 수를 결정하는 적절한 방법이라 할 수 있다.

4.2 ARHMM에서의 화자적응

한국어 지역명 고립단어에 대한 화자독립 및 화자종속 인식기의 평균 인식률은 각각 80.0%와 98.8%이다.

표 5. 화자적응 방법에 따른 화자들의 평균 인식률(%)
Table 5. Recognition rate of speaker adaptation.(%)

Tokens	EXP1	EXP2	EXP3
1	65.41	90.89	90.96
2	86.00	92.75	91.67
3	89.90	94.29	93.52

EXP1 : MLE(maximum likelihood estimation) 방법
EXP2 : 한 개의 가지를 갖는 방법
EXP3 : 여러 개의 가지 사용, 여기서 평균가지수 M을 4로 하였다.

표 5에서의 결과를 보면 다른 방법에 비해서 EXP2의 방법이 가장 우수함을 볼 수 있다. MLE를 사용한 훈련 방법을 사용한 EXP1에서는 훈련 토큰 수가 적을 때는 인식률이 떨어짐을 알 수 있다. MLE로 훈련할 경우 적은 훈련 데이터를 가지고는 각 상태를 대표하는 선형예측계수 값을 제대로 추정할 수 없기 때문이다. 여러 개의 가지를 갖는 EXP3 방법은 토큰 수가 1개일 때를 제외하고는 EXP2 방법에 비해 두드러진 인식성능의 향상을 나타내지는 못했다. ARHMM에서는 가우스 분포를 선형예측계수를 이용하여 간접적으로 나타내기 때문에 훈련 데이터가 부족할 때에는 가우스 분포를 적절히 나타내기가 힘들다. 따라서 부족한 훈련데이터의 영향으로 여러개의 가지를 갖는 방법이 두드러진 인식성능의 향상은 나타나지 않았다.

V. 결 론

본 논문에서는 CHMM인 CDHMM과 ARHMM을 이용하여 화자적응화 하는 방법을 연구하였다. CDHMM에서 최대사후확률 추정법에 의하여 화자적응화를 수행할 때 각 상태마다 하나의 가지를 사용함에 따라 적응성능의 한계를 나타내었다. 따라서 상태마다 여러 개의 가지를 사용하는 방법을 제안하였다. 이때 훈련데이터 수가 적을 경우 어떤 상태에서는 여러개의 가지를 사용하는 것이 타당하지 않을 수도 있으므로 각 상태에 속하는 프레임의 수와 상태내의 분산행렬의 행렬식값으로 가지 수를 결정하는 방법을 제안하였다. 상태내의 프레임 수로 가지수를 결정하는 방법이 분산행렬의 행렬식값으로 가지수를 결정하는 방법에 비해 수행과정이 간단할 뿐 아니라 인식성능도 우수하였다. 이와 같은 결과는 분산행렬의 행렬식 값의 편차가 너무 크기 때문에 적절한 가지수를 결정하기가 힘들고 프레임 수가 많은 상태에 가지수를 늘리는 것이 타당하기 때문이다.

ARHMM에서는 특징벡터로 선형예측계수를 사용하기 때문에 최대사후확률 추정법에 의한 적응화 방법을 사용할 수 없다. 따라서 본 논문에서는 화자독립모델을 사용하여 적응음성을 각 상태로 분할한 후 k-means 알고리듬을 이용하여 하나의 가지로 대표하는 방법을 제안하였다. 제안된 알고리듬을 사용하였을 경우 화자독립인식에 비해 오인식률이 감소함을 확인하였다. CDHMM에서와 동일하게 프레임 수에 따라 여러 개의 가지를 갖는 방법에 대해서도 실험해 본 결과 CDHMM에서와 같은 두드러진 인식성능의 향상은 나타나지 않았다. 이것은 ARHMM에서 상태내의 가우스 분포를 선형예측계수를 이용하여 간접적으로 나타내기 때문에 훈련데이터가 부족할 경우 적절한 표현이 어렵기 때문이라 여겨진다.

제안한 방법은 음성인식 시스템의 구현시 채널의 특성이나 주변 잡음에 인식기를 적응시키는 환경적응화에도 효과를 보일 것으로 기대된다.

참 고 문 헌

1. 김광태, 서정일, 홍재근, "ARHMM에서의 화자적응," 한국정보처리학회 추계학술발표 논문집, Vol. 4, No. 2, pp. 1184-1188, 1997.
2. 한유수, 서정일, 김광태, 홍재근, "연속 혼합 가우스 밀도를 가지는 HMM에서의 화자적응," 신호처리합동 학술대회, Vol. 10, No. 1, pp. 317-320, 1997.
3. L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of isolated digits using hidden Markov models with continuous mixture densities," *AT&T Technical Journal*, vol. 64, no. 6, pp. 1211-1234, July-Aug. 1985.
4. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
5. M. H. DeGroot, *Optimal Statistical Decisions*. McGraw-Hill,

New York, 1970.

6. B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for Speech Signals," *IEEE Trans. on ASSP*, vol. 33, no. 6, Dec. 1985.
7. 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. on Signal Processing*, vol. 39, no. 4, pp. 806-814, Apr. 1991.

▲김 광 태(Kwang-Tae Kim)

1985년 2월 : 경북대학교 공과대학 전자공학과 졸업(공학사)
1987년 2월 : 경북대학교 대학원 전자공학과 졸업(공학석사)
1992년 3월~현재 : 경북대학교 대학원 전자전기공학부 박사과정
1989년~1993년 : 국방과학연구소 연구원
1994년 3월~현재 : 상주산업대학교 전자전기공학과 조교수
※주관심분야 : 음성인식, 음성신호처리, VLSI 설계

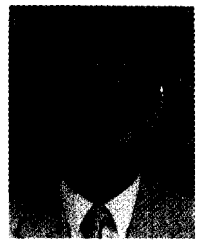
▲한 유 수(Yoo-Soo Han)

1996년 2월 : 경북대학교 공과대학 전자공학과 졸업(공학사)
1998년 2월 : 경북대학교 대학원 전자공학과 졸업(공학석사)
1998년 3월~현재 : 경북대학교 대학원 전자전기공학부 박사과정
※주관심분야 : 음성인식, 음성신호처리

▲서 정 일(Jeong-Il Seo)

1994년 2월 : 경북대학교 공과대학 전자공학과 졸업(공학사)
1996년 2월 : 경북대학교 대학원 전자공학과 졸업(공학석사)
1996년 3월~현재 : 경북대학교 대학원 전자전기공학부 박사과정
※주관심분야 : 음성인식, 음성합성

▲홍 재 근(Jae-Keun Hong)



1975년 2월 : 경북대학교 공과대학 전자공학과 졸업(공학사)
1979년 2월 : 경북대학교 대학원 전자공학과 졸업(공학석사)
1985년 2월 : 경북대학교 대학원 전자공학과 졸업(공학박사)
1979년~1982년 : 경북산업대학교 조교수
1983년~현재 : 경북대학교 전자전기공학부 교수
※주관심분야 : 음성인식, 음성합성, 음성신호처리