상태당 가지 수를 가변시킨 ARHMM을 이용한 화자적응

김 광 태[†] · 서 정 일[††] · 홍 재 근[†††]

요 약

CDHMM(continuous density hidden Markov model)에서는 MAPE(maximum a posteriori estimation) 방법을 이용하여 평균과 분산을 조정함으로써 화자에 적응된 모델을 만든다. 그러나, ARHMM(autoregressive hidden Markov model)에서는 특징벡터로 선형예측계수(LPC; linear prediction coefficient)를 사용하기 때문에 MAPE 방법을 사용할 수 없게 된다. 따라서, 본 논문에서는 화자독립모델을 이용하여 적응화자에 대한 음성을 Viterbi 알고리듬으로 상태별로 분할한 후 각 상태마다 하나의 가지를 가지는 모델로 만들어 적응시키는 방법을 제안한다. 덧붙여, 상태마다 여러 개의 가지를 사용하는 방법에 대해서도 실험해 보았다. 이때 훈련 데이타 수가 적기 때문에 어떤 상태에서는 여러 개의 가지를 사용하는 것이 타당하지 않을 수도 있다. 따라서 각 상태마다 속하는 프레임 수에 따라 가지의 수를 달리하는 방법을 사용하였다. 그리고 상태 지속시간 분포도 적응시켜 화자고유의 발음속도와 길이 등의 특성을 흡수하도록 하였다. 15개 한국 지역명에 대해 제안한 방법으로 화자독립모델을 화자적응시켰을 때 에러율이 50% 이상 감소함을 확인할 수 있었다.

Speaker Adaptation Using ARHMM Varied Number of Branches in Each State

Kwang-Tae Kim[†] · Jeong-Il Seo[††] · Jae-Keun Hong[†††]

ABSTRACT

It is made the speaker adaptation model by adjusting both the mean and the variance of the Gaussian state observation densities of a CDHMM to use the MAPE method. However, we can't use the MAPE method in ARHMM because the components of LPC vector are used as the feature vector of ARHMM. Therefore, in this paper, we propose a speaker adaptation method of ARHMM to adapt the speaker adaptation model having one branch in each state after it is divided the input utterance, which is spoken to an adapted speaker, into states by Viterbi algorithm and then make a typical vector using modified k-means algorithm. In addition we have experimented another method in which each state is represented by several branches. If the training data is insufficient, this method is not proper to train. So we vary the number of branch in proportion to the number of frame stayed in each state, and make to absorb the characteristics of speaker's pronunciation speed and duration by using the distribution of the state duration adapted to the speaker. When testing 15-word Korean domestic name isolated word model, using the proposed method, the recognition performance was found to reduce the error rate of speaker-independent systems more than 50%.

† 정 회 원:상주산업대학교 전자전기공학과
†† 준 회 원:경북대학교 전자전기공학부
††† 정 회 원:경북대학교 전자전기공학부
논문접수:1997년 12월 1일, 심사완료:1998년 1월 9일

1. 서 론

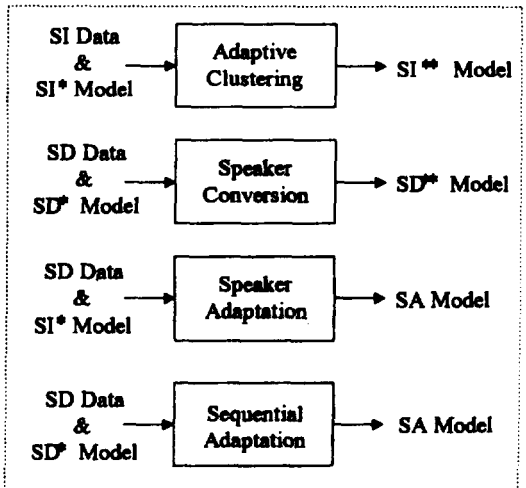
디지털 신호처리 기술과 통신기술 및 컴퓨터의 발달로 인간과 기계간의 의사소통을 보다 자연스럽고 정확하게 하는 man-machine interface 기술이 현실적 문제로 부각되고 있는 가운데, 최근에는 멀티미디어나 이동통신과 같은 다양한 정보매체를 통한 통신분야에 있어서도 음성신호처리의 중요성이 대두되고 있다. 이처럼 음성이 중요한 분야로 연구되고 있는 이유는 음성신호가 가지는 정보전달 측면에 있어서의 신속성 및 사용의 용이함 때문일 것이다. 음성인식은 대어휘, 화자독립, 연속음성인식을 최종목표로 하고 있지만 화자독립, 연속음성인식기의 인식률을 우수하게 하는 것은 매우 힘든 일이며, 또 평균적인 인식률이 좋다고 하더라도 사용하는 화자에 따라 인식률이 많은 차이를 보인다. 이런 문제를 해결하기 위해 미리 훈련되어 있는 화자독립 인식시스템을 특정화자에 적응시키는 화자적응기술을 사용할 수 있다.

화자적응에는 1) 화자독립 모델을 새로운 화자독립 데이타를 사용하여 최신화하는 적응 클러스터링(adaptive clustering), 2) 특정화자에 맞춰 잘 훈련된 모델을 소량의 훈련 데이터를 사용하여 새로운 화자의 모델로 변환하는 화자변환(speaker conver-sion), 3) 화자독립 모델이나 여러 화자의 모델로부터, 특정화자의 훈련 데이터를 사용하여 그 화자로 적응시키는 화자적응(speaker adaptation), 4) 특정화자의 훈련 데이터가 긴 시간에 걸쳐서 들어올 때, 매 시간 새로운 훈련 데이터로 특정화자 모델을 적응시키는 순차적 적응(sequential adaptation) 등이 있으며 이들을 (그림 1)에 나타내었다. 또한 이런 적응화 기술은 채널 특성, 주변 잡음 등에 인식기를 적응시키는 환경 적응화에도 이용될 수 있다[1, 2, 3, 4, 5].

가우스 분포를 사용하는 CDHMM(continuous density hidden Markov model)[6]의 경우에는 MAPE (maximum a posteriori estimation) 방법[7]을 이용하여 화자적응을 하지만 가우스 분포를 선형예측계수값(linear predictive coefficient, LPC)의 자기상관계수를 이용해 표현하는 ARHMM(autoregressive hidden Markov model)[8]에서는 각각의 선형예측계수들이 상호 의존적이므로 MAPE 방법을 사용할 수 없게 된다. 따라서 본 논문에서는 음성 데이터의 각 상태마다의 선형예측계수값의 자기상관계수값의 평균값만을 그 화자에 적응시키는 방법을 사용하였다.

또한, 본 논문에서는 화자적응시에 각 상태를 하나의 가지로 화자적응 모델을 만드는 기존 방법과는 달리 상태마다 여러 개의 가지를 사용하여 음성의 다양한 특성을 반영할 수 있는 방법을 제안한다. 이 경우 문제점은 훈련 데이터의 양이 충분치 않으므로 인해 특정 상태에 머무는 프레임의 수가 상당히 적을 경우에는 제한된 데이터 양으로 여러 개의 가지를 사용하는 것이 타당하지 않을 수도 있다. 따라서 각 상태마다 머무는 프레임 수에 따라 가지 수를 달리하는 방법을 사용하였다. 음성의 상태당 길이 정보의 분포는 가우스 분포, 감마 분포, 히스토그램 등이 이용되고 있는데 본 논문에서는 가우스 분포를 사용하였고 이를 화자적응시켜 화자 고유의 발음속도와 길이 등의 특성을 흡수하도록 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 ARHMM의 원리와 특징을 설명하고, 3장에서는 연속 HMM에서의 화자적응 방법을 설명한 후 CDHMM에서와 ARHMM에서의 화자적응 방법을 설명한다. 4장에서는 제안된 알고리듬으로 고립단어인식실험을 한 결과를 나타내고, 5장에서는 결론을 맺는다.



(그림 1) 4가지 화자적응 방법들
(Fig. 1) Block diagrams of four different speaker adaptation setups.

2. Autoregressive HMM

최근 가장 많이 사용되고 있는 음성인식 알고리듬인 HMM(hidden Markov model)[9]은 Baum등의 연구에 기초하여 Baker와 IBM의 연구진에 의해 제안되었다. HMM은 관측할 수 없는 확률과정을 관측이 가능한 다른 확률과정을 통해 추정하는 이중 확률처리 과정이다. 음성에서는 음성의 발성구조를 상태로 보고 성도가 이러한 상태 중의 하나에 있다고 가정하면, 성도의 전달특성의 변화는 관측이 불가능한 확률과정에 해당되며 관측이 가능한 음성신호로부터 그 변화를 추정하게 된다.

HMM은 관측벡터의 관측밀도함수 형태에 따라서 이산 HMM과 연속 HMM이 있으며 연속 HMM은 가우스 분포함수의 표현방법에 따라 CDHMM과 ARHMM으로 나눌 수 있다. CDHMM은 평균과 분산을 이용하여 가우스 분포를 나타내는 반면에 ARHMM에서는 선형예측계수의 자기상관계수를 이용하여 간접적으로 가우스 분포를 표현한다. 따라서 ARHMM이 CDHMM에 비하여 인식성능은 다소 떨어지지만 관측밀도함수의 계산이 간단하기 때문에 인식속도가 빠르며 메모리도 적게 사용하는 장점이 있다.

자기회귀과정(autoregression process)을 통해 얻어지는 ARHMM(autoregressive HMM)의 관측벡터 $o=(x_0, x_1, x_2,\cdots,x_{k-1})$는 여러 개의 음성 표본 x_i로 구성되며 다음과 같은 선형예측식을 만족한다고 가정한다.

$$x_k=-\sum_{i=1}^{P} a_i x_{k-i}+e_k \tag{1}$$

여기서 a_i는 선형예측계수이고, e_k는 선형예측오차이며, P는 AR모델의 차수이다. 이 ARHMM으로부터 입력신호의 특징벡터를 관측할 확률밀도함수는 다음과 같다.

$$b_j(o)=\sum_{m=1}^{M} c_{jm}\, b_{jm}(o), \qquad 1\le j\le N \tag{2}$$

여기서 M은 확률함수를 구성하는 기본확률밀도함수의 수, 즉 상태당 가지(branch)수이고, $b_{jm}(o)$는 가우스 기본 밀도 함수이며, c_{jm}은 기본밀도함수를 조합하는 하중값이다. ARHMM에서 $b_{jm}(o)$는 근사적으로 다음과 같이 나타낼 수 있다[9, 10].

$$b_{jm}(o)\simeq(2\pi)^{-K/2}\exp\left\{-\frac{1}{2}\,\delta(o:a_{jm})\right\} \tag{3}$$

여기서

$$a_{jm}=[1, a_1, a_2,\cdots, a_P]^T, \ (a_0=1) \tag{4}$$

$$\delta(o;a_{jm})=r_a(0)\,r(0)+2\sum_{i=1}^{P} r_a(i)\,r(i) \tag{5}$$

$$r_a(i)=\sum_{n=0}^{P-i} a_n\, a_{n+i}, \ 1\le i\le P \tag{6}$$

$$r(i)=\sum_{n=0}^{K-i-1} x_n\, x_{n+i}, \ 0\le i\le P \tag{7}$$

이며 a_{jm}은 모델의 j번째 상태의 m번째 가지를 나타내는 선형예측계수 벡터이고, $\delta(o;a_{jm})$은 관측벡터 o와 모델사이의 Itakura-Saito 거리이며, 모델이 관측벡터를 얼마나 잘 표현하고 있는가를 나타내는 척도이다. $r_a(i)$는 모델의 선형예측계수들의 자기상관함수이고, $r(i)$는 음성표본들의 자기상관함수이다.

3. 연속 HMM에서의 화자적응

3.1 화자적응(speaker adaptation) 방법

불특정화자에 의해 훈련된 화자독립 음성인식시스템을 사용하려는 특정화자가 발음한 소량의 음성 데이터로써 화자적응을 수행하여 인식성능을 높이는 화자적응기술은 음성인식시스템의 실용화 단계에서 꼭 필요한 기술중의 하나이다.

최대사후확률추정법(MAP)과 ML(maximum likelihood) 방법과의 차이는 추정되어질 파라미터에 대한 사전확률분포의 적절한 가정에 있다. 입력벡터 Y가 확률분포 $P(Y)$를 갖고 λ는 주어진 확률분포에 의해 정의되는 파라미터라고 할 때 λ의 ML 추정치는 아래와 같은 식을 풀어서 구할 수 있다.

$$\frac{\partial}{\partial\lambda}\,P(y_1, y_2,\cdots, y_T|\lambda)=0 \tag{8}$$

만일 λ가 사전분포확률 $P_0(\lambda)$를 가진다면 λ의 MAP 추정치는 식 (9)를 풀어서 구할 수 있다.

$$\frac{\partial}{\partial \lambda} P(y_1, y_2, \cdots, y_T|\lambda)=0 \tag{9}$$

베이정리(Bayes theorem)를 이용하여 $P(\lambda|Y)$를 식 (10)과 같이 쓸 수 있다.

$$P(\lambda|y_1, y_2, \cdots, y_T|\lambda)=\frac{P(y_1, y_2, \cdots, y_T|\lambda)P_0(\lambda)}{P(y_1, y_2, \cdots, y_T|\lambda)} \tag{10}$$

3.2 CDHMM에서의 화자 적응

CDHMM에서는 사전분포확률을 이용하는 MAPE 방법으로 평균과 분산을 화자에 적응시킨다.

평균 μ가 사전 분포 $P_o(\mu)$를 가지는 랜덤값이고 분산 σ^2이 고정된 값으로 알려져 있을 때, $P_o(\mu)$가 평균 v와 분산 τ^2을 가지는 정규 분포라고 가정하면 μ의 MAP 추정치는 다음과 같다[6]. 여기서 n은 훈련데이타의 개수이고, $\bar{y}$는 데이타의 평균이다.

$$\hat{\mu}_{MAP}=\frac{n\tau^2}{\sigma^2+n\tau^2}\bar{y}+\frac{\sigma^2}{\sigma^2+n\tau^2}\mathrm{v} \tag{11}$$

평균 μ가 알려지지 않고, 분산의 사전 분포가 식 (12)와 같이 알려져 있다면, 분산 σ^2의 MAP 추정치는 식 (13)의 조건으로부터 식 (14)와 같이 구할 수 있다[4].

$$P_o(\sigma^2)=\begin{cases}\text{constant} & \sigma^2 \geq \sigma^2_{\min} \text{ 인 경우}\\ 0 & \text{그 외의 경우}\end{cases} \tag{12}$$

$$\max_{\sigma^2}\left\{-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right\} \tag{13}$$

$$\hat{\sigma}_{MAP}=\begin{cases}S_y^2, & S_y^2 \geq \sigma^2_{\min} \text{ 인 경우}\\ \sigma^2_{\min}, & \text{그 외의 경우}\end{cases} \tag{14}$$

여기서, 샘플 데이터의 분산을 나타내는 S_y^2는 식 (15)와 같다.

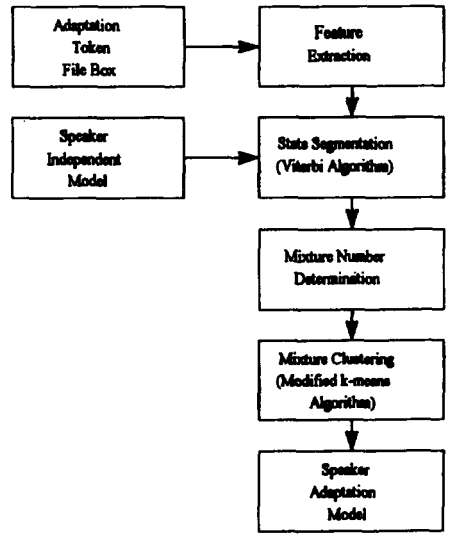
$$S_y^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n} \tag{15}$$

단, $\sigma^2_{\min}$ 값은 화자독립 모델로부터 추정되며, 또 평균에 대한 사전 정보는 없으므로 샘플 평균 $\bar{y}$로 평균 μ를 추정한다.

3.3 ARHMM에서의 화자적응

CDHMM과는 달리 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, modified k-means 알고리듬을 이용하여 대표되는 선형예측계수의 자기상관계수로 결정하는 방법을 통해 발음 화자에 적응시키는 방법을 제안한다. 상태를 분할하는 방법은 Viterbi 알고리듬을 사용하였다.

평균을 상태마다 하나로 나타내면 화자의 다양한 음성정보를 제대로 나타내지 못하므로 입력 벡터열을 modified k-means 알고리듬을 사용하여 몇 개의 클러스터로 분리한 후 이 각각에 대한 데이터 평균을 구하여 사용한다. 이때 훈련 데이타가 적기 때문에 어떤 상태에서는 여러 개의 가지를 사용하는 것이 적



(그림 2) 여러 개의 가지를 갖는 화자적응 알고리듬의 블록도

(Fig. 2) The block diagram of speaker adaptation algorithm with variable branches.

합치 않으므로 식 (16)과 같이 프레임수에 비례하여 가지 수를 달리하는 방법을 사용한다.

$$m_j = \left\lceil \frac{N \times \sum_k n_{jk}}{\sum_k \sum_j n_{jk} \times M} \right\rceil \quad (16)$$

여기서, m_j는 상태 j에서의 가지 수를, n_{jk}는 k번째 훈련 음성의 상태 j에서의 프레임 수를 나타낸다. 그리고 N, M은 각각 모델의 상태수, 평균 가지 수를 나타낸다. 여러 개의 가지를 가지는 방법의 전체적인 블럭도는 (그림 2)와 같다.

4. 인식 실험 및 결과 고찰

본 실험에 사용한 음성데이타는 10명(남자 5명, 여자 5명)의 화자가 15개 한국 지역명을 10번씩 발음한 것이다. 이 음성을 5.6 kHz의 저역필터를 통과시켜 12 kHz로 샘플링한 뒤 12 bit로 양자화 하였다. 100샘플씩 이동하면서 300샘플씩 해밍 윈도우를 취하였고, 12차 LPC를 특징인자로 사용하였다. 단순 left-to-right 구조의 ARHMM을 사용하였으며, 상태수는 6으로 하였다. 상태천이확률의 변화는 인식률에 별다른 영향을 못미쳐 화자독립모델의 확률값을 그대로 사용하였다.

화자독립모델의 상태지속정보를 정규분포로 나타내어 사용하였고, 모든 화자적응실험에서 이 분포의 평균값을 다음 수식과 같이 그 화자에 적응시키는 방법을 사용하였다.

$$\mu_d = \frac{n}{\text{token 수}} \quad (17)$$

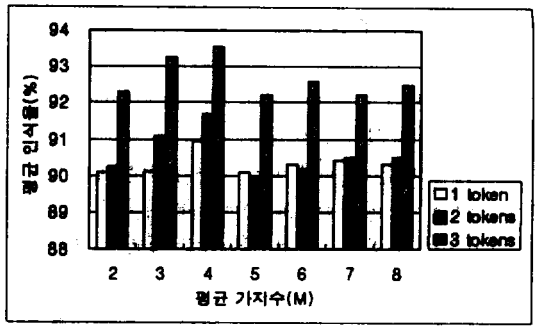
여기서, n은 상태에 속하는 벡터의 수이고 μ_d는 상태지속길이의 평균이다. 상태지속길이의 분산은 화자독립모델에서 구한 값을 그대로 사용하였다.

한국어 지역명 고립단어에 대한 화자독립 인식시스템의 인식률은 〈표 1〉과 같다. 이때 인식실험은 테스트하고자 하는 1명을 제외한 나머지 사람들로 훈련시킨 후 인식테스트를 하는 round-robin 방식을 사용하였다. 화자종속 인식시스템의 인식률도 〈표 1〉에 함께 나타내었다. 화자독립 및 화자종속인식 두 경우 모두 5개의 가지를 사용하였다.

〈표 1〉 화자독립/종속 인식기의 인식률(%)
〈Table 1〉 The recognition rate (%) of speaker independent & dependent system.

화자	화자독립	화자종속
M1	77.33	100.00
M2	77.33	100.00
M3	75.33	97.33
M4	87.33	96.00
M5	84.67	94.67
F1	82.00	100.00
F2	89.33	92.00
F3	82.00	98.65
F4	88.67	100.00
F5	56.00	89.23
평균 인식률	80.00	98.80

〈표 1〉에서와 같이 훈련음성이 충분하다면 화자종속 인식시스템의 성능이 화자독립 인식시스템보다 우수하다. 그러나, 불특정 화자들에 의해 훈련된 화자독립시스템의 범용성과 화자종속 인식시스템을 훈련시키기 위해서 여러 번씩 발음해야 하는 불편함 때문에 궁극적으로 화자독립 인식시스템이 바람직하다. 따라서 화자독립 인식시스템을 사용하려는 화자에 적응시키는 화자적응 인식시스템의 구현이 필요하다.



(그림 3) 평균 가지 수의 변화에 따른 화자적응기의 평균 인식률(%)
(Fig. 3) The recognition rate of speaker adaptation system in proportion to the number of the averave branches(%).

〈표 2〉 화자적응 방법에 따른 화자들의 평균 인식률(%)
〈Table 2〉 The recognition rate of speaker adaptation system (%)

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

본 논문에서 제안된 방법을 이용하여 ARHMM을 10명의 화자들에 대해 인식실험을 하였을때의 평균 가지 수 M에 따른 결과를 (그림 3)에 나타내었고, 인식실험에서의 평균 인식률을 〈표 2〉에 나타내었다.

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

(그림 2)에 나타난 바와 같이 4개의 평균 가지 수를 가질 때 가장 높은 인식률을 나타내었다. 〈표 2〉에서의 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. 이는 상태수가 충분하다면 상태당 1개의 가지만으로도 충분히 화자의 특징을 나타낼 수 있음을 보인다.

제안된 방법을 이용하여 상태수가 적을때의 화자적응 실험을 한 결과를 〈표 3〉에 나타내었다. 기본적인 화자독립 인식시스템의 상태수는 4로 하고 가지 수는 3으로 하였다.

〈표 3〉에서 나타난 바와 같이 여러 개의 가지를 갖는 EXP3 방법이 가장 우수함을 알 수 있다. 이는 훈련 데이터의 수가 적을 때나 상태수가 적을 때에는 여러 개의 가지를 사용하는 방법이 한 개의 가지를 사용하는 방법보다 화자의 특징을 보다 자세히 표현

〈표 3〉 4개의 상태수에서 화자적응 방법에 따른 평균 인식률 (%)
〈Table 3〉 The recognition rate of speaker adaptation system using 4 states

Tokens	EXP1	EXP2	EXP3
1	65.41	88.81	90.15
2	86.00	91.25	92.75
3	89.90	92.48	94.29

할 수 있으므로 인식성능이 우수하였다.

남자화자 중 가장 적응이 안 되는 화자 M5, 여성화자중 가장 적응이 잘 되는 화자 F1, 가장 적응이 잘 안 되는 화자 F5의 인식률을 각각 〈표 4, 5, 6〉에 나타내었다.

〈표 4〉 화자 M5의 인식률
〈Table 4〉 The recognition rate of M5 speaker.

Tokens	EXP1	EXP2	EXP3
1	61.48	91.11	92.59
2	85.83	95.83	95.00
3	91.43	93.33	97.14

〈표 5〉 화자 F1의 인식률
〈Table 5〉 The recognition rate of F1 speaker.

Tokens	EXP1	EXP2	EXP3
1	71.85	97.78	100.0
2	91.67	100.0	100.0
3	95.24	100.0	100.0

〈표 6〉 화자 F5의 인식률
〈Table 6〉 The recognition rate of F5 speaker.

Tokens	EXP1	EXP2	EXP3
1	45.93	74.07	74.81
2	78.33	84.17	76.67
3	84.76	88.57	83.81

화자에 따라서는 EXP3의 방법이 같은 수의 토큰으로 훈련한 EXP2의 방법보다 인식률이 우수한 경우가 있음은 여러 개의 가지 수를 갖는 방법이 화자에 따라 우수한 성능을 나타낼 수 있다는 것을 보인다. 화

자별로 적응의 차이가 있지만 다른 방법에 비해서 EXP2의 방법이 F1의 토큰 1개일 때를 제외하고는 성능이 가장 우수하였다.

같은 방법을 CDHMM에 적용 하였을 때의 결과를 〈표 7〉에 나타내었다. CDHMM에서는 토큰 수에 관계없이 여러 개의 가지를 사용하는 방법이 한 개의 가지를 사용하는 방법보다 인식성능이 우수하였다. CDHMM에서는 가우스 분포를 평균과 분산을 통하여 직접적으로 나타내므로 한 개의 가지를 사용할 때의 특성도 여러 개의 가지를 사용하는 방법이 흡수할 수 있으며 여러 개의 가지들로 화자별 음성의 특징을 보다 자세히 표현할 수 있기 때문이다.

〈표 7〉 CDHMM에서 화자적응 방법에 따른 평균 인식률 (%)

〈Table 7〉 The recognition rate of speaker adaptation system in CDHMM (%).

Tokens	EXP1	EXP2	EXP3
1	76.0	88.4	94.4
2	88.7	88.4	97.7
3	95.3	91.0	98.0

5. 결 론

ARHMM에서 Viterbi 알고리듬을 이용하여 상태를 분할한 후 각 상태에서 modified k-means 알고리듬을 이용하여 하나의 가지로 상태를 대표하게 하여 화자에 적응시키는 방법이 화자독립 인식방법에 비해 약 50% 오인식률의 감소를 나타내었다. 한 개의 가지를 갖는 방법 외에 여러 개의 가지를 갖는 방법은 화자의 다양한 음성정보를 포함할 수 있어서 화자에 따라서는 우수한 인식성능을 나타냄을 확인하였다. 또한 토큰수가 적을 때나 상태수가 적을 때는 여러 개의 가지를 사용하는 방법이 제한된 훈련데이타로 보다 자세히 화자의 특징을 표현할 수 있기 때문에 보다 우수한 인식성능을 나타냄을 알 수 있었다. 그리고, 상태지속시간분포의 화자적응으로 화자 고유의 발음속도와 길이 등의 특성을 잘 흡수함으로 인해 인식률을 2% 정도 향상시킬 수 있었다. 또한 같은 방법을 CDHMM에 적용하였을 경우 훈련 토큰의 수나 상태수에 관계없이 여러 개의 가지를 갖는 방법이 우수함을 확인하였다.

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," Vol. 10, No. 1, pp. 317-320, 1997.

[3] 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.

[4] Y. Hao and D. Fang, "Speech recognition using speaker adaptation by system parameter transformation," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, pp. 63-68, Jan. 1994.

[5] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 357-365, Sep. 1995.

[6] 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, Jul.-Aug. 1985.

[7] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[8] 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.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speeech rec-

ognition," *Proc. IEEE*, vol. 77, no. 2, pp 257-286, Feb. 1989.

[10] L. R. Rabiner and G. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, pp 4-16, Jan. 1986.

김 광 태

1985년 2월 경북대학교 공과대학 전자공학과 졸업(공학사)

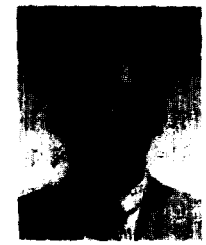
1987년 2월 경북대학교 대학원 전자공학과 졸업(공학석사)

1992년 3월~현재 경북대학교 대학원 전자전기공학부 박사과정

1989년~1993년 국방과학연구소 연구원

1994년~현재 상주산업대학교 전자전기공학과 조교수

관심분야: 음성인식, 음성신호 처리



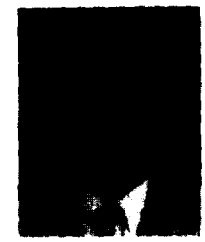
서 정 일

1994년 2월 경북대학교 공과대학 전자공학과 졸업(공학사)

1996년 2월 경북대학교 대학원 전자공학과 졸업(공학석사)

1996년 3월~현재 경북대학교 대학원 전자전기공학부 박사과정

관심분야: 음성인식, 음성 신호처리



홍 재 근

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

관심분야: 음성인식, 음성인식 합성, 음성신호 처리