論文 98-23-9-43

상태 종속 관측 확률 밀도 함수를 세분화한 CHMM에서의 화자적응

정회원 김 광 태*, 한 유 수**, 홍 재 근**

Speaker Adaptation in CHMM Having the Subdivided State-Dependent Observation Density

Kwang-Tae Kim*, Yoo-Soo Han**, Jae-Keun Hong** *Regular Members*

요 약

본 논문에서는 CDHMM과 ARHMM을 이용하여 화자적응화 하는 방법을 각각 제안하였다. 기존 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 modified k-means 알고리듬을 이용하여 각 상태마다 하나의 가지를 갖는 모델로 적응시키는 방법을 제안하였다. 15개의 한국어 지역명으로 구성된 음성데이터와 40개의 단어로 구성된 ETRI의 셈돌이 데이터에 대하여 인식실험한 결과 기존의 방법들에 비해 높은 인식율을 얻을 수 있었다.

ABSTRACT

In this paper, we proposed the method of speaker adaptation in CDHMM and ARHMM respectively. In conventional CDHMM, speaker adaptation had been performed using one mixture in each state by the method of MAPE (maximum a posteriori estimation). In this paper, we proposed the method using variable mixtures to represent properly various speech information of the speaker in each state. We determined the number of mixtures 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 mixture in one state. The input

* 상주대학교 전자전기공학과
** 경북대학교 전자전기공학부
論文番號 : 98137-0324
接受日字 : 1998年 3月 24日

utterance was divided into each states by Viterbi algorithm using speaker independent model and then transformed into a typical vector by the modified k-means algorithm. The 15 Korean domestic name database and the ETRI 'Samdol' database consisted of 40 words were evaluated, the recognition rate of proposed methods were improved higher than the conventional methods.

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. 화자독립인식기는 불특정화자에 의해 훈련되는 인식기이다. 인식기를 사용할 때 별도의 훈련과정이 필요치 않으나 화자종속인식기에 비해 낮은 인식성능을 나타내게 된다.

화자적응(speaker adaptation)방법은 소량의 훈련데이터를 사용하여 충분한 훈련데이터로 훈련된 화자독립모델을 인식시스템을 사용하려는 특정화자에 적응시키는 방법이다[1～10]. 따라서 인식시스템을 사용하려는 화자가 화자종속인식기를 훈련시키기 위한 훈련데이터보다 적은 훈련데이터로도 화자종속인식기의 인식성능을 얻을 수 있으며 인식시스템이 사용되어지는 환경이 바뀌거나 잡음이 있는 곳에서도 우수한 인식성능을 나타내게 된다.

화자적응에는 1) 화자독립모델을 새로운 화자독립데이터를 사용하여 최신화하는 적응 클러스터링(adaptive clustering), 2) 특정화자에 맞춰 훈련된 모델을 조금의 훈련데이터를 사용하여 새로운 화자의 모델로 변환하는 화자변환(speaker conversion), 3) 화자독립모델이나 여러 화자의 모델로부터, 특정화자의 훈련데이터를 사용하여 그 화자로 적응시키는 화자적응(speaker adaptation), 4) 특정화자의 훈련데이터가 긴 시간에 걸쳐서 들어올 때, 매 시간 새로운 훈련데이터로 특정화자 모델을 순차적으로 적응시키는 순차적 적응(sequential adaptation) 등이 있으며 이들을 그림 1에 나타내었다[3]. 본 논문에서는 화자독립모델로부터 특정화자로 적응시켜 화자종속모델을 만드는 화자적응에 대해서 다룬다.

음성인식기가 어떤 음성 패턴들로 모델링되어 있는가에 따라 사용될 화자적응기술이 정해진다. 즉, 템플릿으로 구성된 인식기에서의 적응은 새로운 템플릿의 추가 또는 수정으로 이루어지고, 벡터 양자화에 기준을 둔 인식기에는 코드북 적응기술이 사용된다[5～7]. DHMM으로 구성된 인식기에서는 히스토그램 적응과 같은 이산관측심벌 분포의 수정을 사용한다[8].

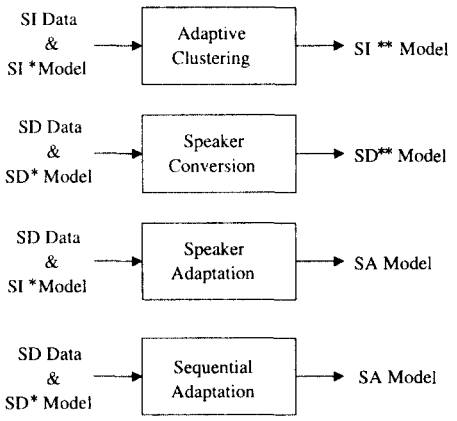


그림 1. 4가지 화자적응 방법들
Fig. 1 Block diagrams of four different speaker adaptation setups.

CDHMM(continuous density hidden Markov model)[13]을 사용하는 인식기에서는 최대사후확률 추정(MAPE; maximum a posteriori estimation) 방법 즉, 베이적응(Bayesian adaptation) 방법[10～12]을 이용한다. CDHMM에서 베이적응 방법을 이용하여 화자에 적응된 모델을 만들 때, 각 상태마다 하나의 가지를 갖는 모델로 만들어 적응시킨다. 이것은 이미 만들어

져 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. 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 셈돌이 데이터에 대해 인식실험한 결과 제안한 방법이 가지를 한 개 사용했을 때에 비해 높은 인식률을 얻을 수 있었다.

가우스 분포를 사용하는 CDHMM의 경우에는 MAPE 방법을 이용하여 화자적응을 하지만 가우스 분포를 선형예측계수값(linear predictive coefficient)의 자기상관계수를 이용해 표현하는 ARHMM(auto-regressive hidden Markov model)[14]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 수행하면 오인식율이 50% 이상 감소함을 확인하였다.

II. CDHMM 파라미터들의 화자적응

MLE(maximum likelihood estimation) 방법과 베이 학습(Bayesian learning) 방법 사이의 차이점은 추정될 파라미터의 적절한 사전분포의 가정에 있다.

$Y=\{y_1, y_2, \cdots, y_T\}$가 확률밀도함수 $P(Y)$를 가지는 주어진 관측열이고 λ는 그 분포를 정의하는 파라미터일 때, 만약 λ가 미지의 상수라면 λ의 ML 추정치는 다음의 조건으로 구할 수 있다.

$$\frac{\partial}{\partial\lambda} P(y_1, y_2, \cdots, y_T|\lambda) = 0 \tag{1}$$

λ가 사전분포함수 $P_o(\lambda)$를 가지는 불특정값이라면 λ의 MAP(maximum a posteriori) 추정치는 다음의 조건으로 구한다.

$$\frac{\partial}{\partial\lambda} P(\lambda|y_1, y_2, \cdots, y_T) = 0 \tag{2}$$

베이정리를 이용하면, $P(\lambda|Y)$는 다음 식과 같다.

$$P(\lambda|Y) = \frac{P(Y|\lambda)P_o(\lambda)}{P(Y)} \tag{3}$$

식 (3)에서의 λ의 분포함수 $P_o(\lambda)$를 사전에 알고 있어야 한다.

본 논문에서는 위의 MAPE(maximum a posteriori estimation) 방법을 segmental k-means 알고리듬에 적용시켜 사용한다.

N개의 상태를 가지는 1차 Markov chain에서, 상태열이 $s=(s_0, s_1, \cdots, s_T)$로 주어지고, 초기상태확률이 $\pi^t=[\pi_1, \pi_2, \cdots, \pi_N]$, 상태천이확률행렬이 $A=[a_{ij}]$, $i, j = 1, 2, \cdots, N$이라고 하면 상태열 s가 관측될 확률은 다음과 같다.

$$P(s|\pi, A) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \tag{4}$$

여기서 π, A, $B=\{b_i\}_{i=1}^{N}$의 세 파라미터값은 HMM의 정의값들이고 $\lambda=(\pi, A, B)$로 표시한다. 상태열 s를 가지는 Y가 관측될 확률은 다음과 같다.

$$P(Y|\lambda) = \sum_{\{s\}} P(Y, s|\lambda) \tag{5}$$

여기서 $\{s\}$는 모든 가능한 상태열을 의미한다. 모델 λ의 파라미터들을 추정하는 여러 가지 방법 중에서, 본 논문에서는 식 (6)과 같은 반복적 방법으로 최대의 확률값을 가지는 모델을 추정한다.

$$\hat{\lambda} = \arg\max_{\lambda} [\max_{s} P(Y, s|\lambda)] \quad (6)$$

여기서 HMM 훈련 방법으로 잘 알려진 segmental k-means 알고리듬을 사용한다. segmental k-means 알고리듬을 확장하여 MAPE 방법에 적용하면 상태열 s를 포함하는 MAP 추정치를 다음과 같이 구할 수 있다.

$$\frac{\partial}{\partial \lambda} P(\lambda, s|Y) = 0 \quad (7)$$

베이정리를 사용하여 위 식의 joint 확률을 다음 식으로 구할 수 있다.

$$P(\lambda, s|Y) = \frac{P(Y, s|\lambda) P_o(\lambda)}{P(Y)} \quad (8)$$

여기서 $P_o(\lambda)$는 파라미터 λ의 사전분포이다. 베이적응과 결합된 적응 segmental k-means 알고리듬은 다음의 두 단계로 구성된다.

1) 주어진 모델 $\hat{\lambda}$에 근거하여 최적 상태열을 다음 식으로 구한다.

$$\hat{s} = \arg\max_{s} P(Y, s|\hat{\lambda}) P_o(\hat{\lambda}) \quad (9)$$

2) 위에서 찾은 최적 상태열 $\hat{s}$에 근거하여 MAP 추정치를 다음 식으로 구한다.

$$\hat{\lambda} = \arg\max_{\lambda} P(Y, \hat{s}|\lambda) P_o(\lambda) \quad (10)$$

λ가 어떤 값에 수렴할 때까지 이 두 단계를 반복 수행한다. 반복 수행할 때 같은 데이터 Y에 대해서는 식 (9)과 (10)의 사전분포함수 $P_o(\lambda)$가 같다. 훈련데이터를 한꺼번에 처리하지 않고 몇 개의 작은 그룹으로 만들어서 한 번에 한 개씩 적응시킬 수도 있다. 이때 한 개의 훈련데이터 그룹이 적용될 때마다 사전분포함수 $P_o(\lambda)$가 변하게 된다. 이것은 순차적 적응인데, 본 논문에서는 훈련데이터를 한꺼번에 이용하는 방법을 사용한다.

2.1 정규분포 파라미터들의 베이적응

이번 절에서는 관측벡터 분포로 정규분포를 가지는 CDHMM의 파라미터에 대한 베이적응 방법에 대해서 기술한다[3]. 수식은 그림 2와 같은 단순 좌우구조 HMM에 대하여 전개하며, 적응시킬 파라미터 수를 감소시키기 위해 각 HMM의 상태천이확률값은 화자독립모델로부터 구한 값을 사용한다. 그러므로 적응은 정규분포의 평균과 분산에 대해서만 이루어진다. 또 대각성분값만을 갖는 분산 행렬을 가지는 정규분포에 대해서만 고려할 것이다. 적응은 평균과 분산에 대해 각 상태마다 독립적으로 이루어지므로 수식은 하나의 상태에 대해서만 전개한다.

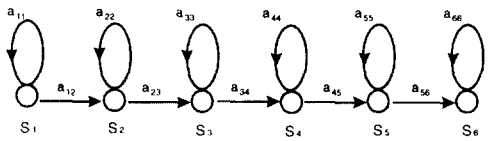


그림 2. 단순 좌우구조 HMM
Fig. 2 Simple left-to-right HMM.

2.1.1 평균의 Bayes 적응

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

$$\hat{\mu}_{\text{MAP}} = \frac{n\tau^2}{\sigma^2 + n\tau^2} \bar{y} + \frac{\sigma^2}{\sigma^2 + n\tau^2} \nu \quad (11)$$

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

μ의 MAP 추정치는 사전분포의 평균 ν와 샘플 평균 $\bar{y}$의 하중값을 달리한 평균이다. 이 하중값은 n이 0이면 부가적인 훈련데이터의 사용없이 ν의 하중값만 1로 만들게 되어 기존의 평균값을 그대로 사용하는 것이 된다. 훈련데이터의 수가 많아지면, μ의

MAP 추정치는 ML 추정치와 거의 같게 된다. 사전분포의 분산 τ^2이 σ^2/n보다 훨씬 클 경우에도 μ의 MAP 추정치는 ML 추정치와 거의 같게 된다.

사전분포의 평균 ν와 분산 τ^2은 다음과 같이 추정된다.

$$\nu = \sum_{m=1}^{M} w_m \nu_m \tag{12}$$

$$\tau^2 = \sum_{m=1}^{M} w_m (\nu_m - \nu)^2 \tag{13}$$

여기서 ν_m, w_m은 각각 화자독립모델의 m번째 가지의 평균과 가중치이다.

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

$$\sigma^2 = \sum_{m=1}^{M} w_m \sigma^2_m \tag{14}$$

여기서 σ^2_m은 m번째 가지의 분산이다.

적응된 후의 사후분포 $P(\mu|Y)$도 또한 정규분포로 나타난다. 이 정규분포에서 평균은 $\hat{\nu} = \hat{\mu}_{MAP}$이고 분산은 다음 식과 같다.

$$\hat{\tau}^2 = \frac{\sigma^2}{\sigma^2 + n\tau^2}\ \tau^2 \tag{15}$$

사후분포의 분산 $\hat{\tau}^2$은 사전분산 τ^2보다 항상 크지 않다. 샘플수 n이 증가하면 사후분포는 샘플 평균 주위로 몰리게 된다. 만약 순차적 적응이라면 이 사후분포가 사전분포로 대치되어 추정을 반복하게 된다.

2.1.2 분산의 Bayes 적응

평균 μ가 미지의 값이고, 분산의 사전분포가 다음 식과 같다면, 분산 σ^2의 MAP 추정치는 식 (17)의 조건으로부터 식 (18)과 같이 구할 수 있다.

$$P_o(\sigma^2) = \begin{cases} \text{constant}, & \sigma^2 \geq \sigma^2_{\min} \\ 0 & o.w. \end{cases} \tag{16}$$

$$\max\left\{ -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 \right\} \tag{17}$$

$$\hat{\sigma}^2_{MAP} = \begin{cases} S_y^2 & S_y^2 \geq \sigma^2_{\min} \\ \sigma^2_{\min} & o.w. \end{cases} \tag{18}$$

여기서, $\sigma^2_{\min}$ 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.

$$S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} \tag{19}$$

이 방법은 훈련데이터가 충분하지 않은 경우에도 분산 파라미터를 추정하는데 매우 효과적이다.

2.1.3 평균과 분산의 Bayes 적응

평균과 분산 모두 어떤 사전분포를 가지는 불특정 값이라고 가정하고 적응된 모델을 구할 수 있다. 이 때 분산의 사전분포를 식 (16)과 같이 가정하지 않고 분산의 역수값인 precision($\theta = 1/\sigma^2$)을 감마분포로 가정하여 사용할 수 있다.

평균과 precision 파라미터가 불특정값이고 사전분포 $P_o(\mu, \theta)$를 다음 수식과 같은 joint normal-gamma 분포로 가정한다.

$$P_o(\mu, \theta) = \frac{\sqrt{\omega\theta}}{\sqrt{2\pi}}\ e^{-\frac{\omega\theta}{2}(\mu-\nu)^2}\ \frac{\beta^\alpha}{\Gamma(\alpha)}\ \theta^{\alpha-1}e^{-\beta\theta} \tag{20}$$

여기서 θ가 주어졌을 때 μ의 조건부 확률분포는 평균 ν와 분산 $\tau^2 = 1/\omega\theta$를 가지는 정규분포이고, θ의 marginal 분포는 파라미터 α, $\beta(>0)$를 가지는 감마분포이다.

Joint 사후확률분포도 normal-gamma 분포가 된다. 그래서, θ와 샘플 데이터가 주어져 있을 때, μ의 조건부 확률분포는 역시 정규분포이고 평균 $\hat{\nu}$와 분산 $\hat{\tau}^2$는 다음 식과 같다.

$$\hat{\nu} = \frac{\omega\nu + n\bar{y}}{\omega + n} \tag{21}$$

$$\hat{\tau}^2 = \frac{1}{(\omega + n)\theta} \tag{22}$$

그리고 샘플 데이터가 주어져 있을 때 θ의 marginal 분포는 감마분포이고, 이것의 파라미터 $\hat{\alpha}$와 $\hat{\beta}$는 다음과 같다.

$$\hat{\alpha} = \alpha + \frac{n}{2} \tag{23}$$

$$\hat{\beta} = \beta + \frac{n}{2} S_y^2 + \frac{n\omega(\bar{y} - \nu)^2}{2(\omega + n)} \tag{24}$$

위 분포로부터 구한 평균과 분산의 MAP 추정치는 다음과 같다.

$$\hat{\mu}_{\text{MAP}} = \frac{\omega\nu + n\bar{y}}{\omega + n} \tag{25}$$

$$\hat{\sigma}^2_{\text{MAP}} = \frac{\hat{\beta}}{\hat{\alpha}} \tag{26}$$

위 식에서 사용된 파라미터들은 다음과 같다.

$$\nu = \sum_{m=1}^{M} w_m \nu_m \tag{27}$$

$$\alpha = \frac{1}{\sigma^2} = \frac{1}{\sum_{m=1}^{M} w_m \sigma_m^2} \tag{28}$$

$$\omega = \frac{\sigma^2}{\tau^2} = \frac{1}{\alpha \sum_{m=1}^{M} w_m (\nu_m - \nu)^2} \tag{29}$$

$$\beta = 1 \tag{30}$$

여기서 ν는 각 가지에 하중값을 부여하여 구한 평균값이고, α는 precision 파라미터이다. ω는 각 가지의 분산을 평균하여 구한 값과 각 가지의 평균으로부터 구한 분산과의 비율을 나타낸다. β는 1로 두었다.

2.2 세분화된 상태 종속 관측 확률 밀도 함수를 구하는 방법

앞절에서 살펴본 방법을 이용하여 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. 이러한 단점을 해결하기 위하여 상태마다 여러 개의 가지를 사용하는 방법을 제안하여 세분화된 상태 종속 관측 확률 밀도 함수를 구하였다. 하지만 앞절에서 설명한 방법 중, 첫 번째 방법과 세 번째 방법에서는 이 방법을 적용하기가 어렵다. 이것은 이미 만들어져 있는 인식기의 특정상태에 속하는 각 가지들의 평균값과 분산값의 분포를 적응시킬 데이터의 사전분포로 이용하는데, 이 분포를 각 상태마다 한 개밖에 구할 수가 없기 때문이다.

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

2.2.1 상태당 프레임 수에 따른 가지 수 결정방법

여러 개의 가지를 갖게 하기 위하여 입력 벡터열 Y를 k-means 알고리듬을 사용하여 몇 개의 클러스터(cluster)로 분리한 후 이 각각에 대한 데이터 분산인 S_{ym}^2을 구하여 사용한다. 이 값과 분산의 하한값 $\sigma^2_{\min}$을 이용하여 각 가지의 적응된 분산을 다음 식과 같이 구한다.

$$\hat{\sigma}^2_{\text{MAP},m} = \begin{cases} S_{ym}^2, & S_{ym}^2 \geq \sigma^2_{\min} \\ \sigma^2_{\min} & o.w. \end{cases} \tag{31}$$

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.

$$m_j = \left\lceil \frac{N \times \sum_k n_{jk}}{\sum_k \sum_j n_{jk}} \times M \right\rceil \tag{32}$$

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. 여기서 기호 $\lceil\cdot\rceil$는 올림을 의미한다.

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 몰리는 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 (32)에서 사용할 평균 가지 수 M의 적절한 값을 찾아야 한다.

2.2.2 분산행렬에 따른 가지 수 결정방법

앞절에서 설명한 프레임 수에 따른 가지 수 결정 방법은 복잡한 연산과정이 필요 없으므로 수행과정이 간단하다. 그러나, 상태내의 프레임 수 만으로 가지 수를 결정하기 때문에 음성의 통계학적 특징을 무시한다는 단점이 있다. CDHMM에서 음성의 통계학적 특성은 평균과 분산에 의해 결정되며 특히 분산이 주된 역할을 하게 된다. 따라서 분산값이 작은 상태보다 분산값이 큰 상태에서 보다 많은 가지 수를 갖게 하는 것이 타당하다.

대각성분값만을 갖는 분산행렬을 가지는 정규분포에 대해서만 고려할 때 관측벡터의 각 상태에서의 분산행렬은 다음과 같이 구해진다.

$$V_j(k) = \frac{1}{T}\sum_{i=1}^{T}(c_i(k)-\mu_j(k))^2 \qquad (33)$$

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

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 식 (34)와 같이 표현할 수 있다.

$$D_j = \prod_{k=1}^{P} V_j(k) \qquad (34)$$

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.

$$m_j = \left\lceil \frac{D_{\max} M}{10\, D_j} \right\rceil \qquad (35)$$

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.

2.3 실험 결과 및 고찰

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

먼저 프레임 수에 따라 가지 수를 결정하는 방법에 대해 실험해 보았다.

2.3.1 지역명 음성데이터

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

각 화자별 화자독립인식률은 표 1과 같으며, 이때 인식실험은 테스트하고자 하는 1명을 제외한 나머지 사람들로 훈련시킨 후 인식테스트를 하는 round-robin 방식을 사용하였다. 화자종속인식률도 표 1과 같으며,

화자독립 및 화자종속인식 두 경우 모두 3개의 가지를 사용하였다.

표 1. 화자독립 및 화자종속인식기의 화자별 인식률 (%)
Table 1. The recognition rate of speaker independent and speaker dependent recognizer (%).

화자	화자독립	화자종속
M1	78.7	100.0
M2	70.7	93.3
M3	89.3	100.0
M4	87.3	100.0
M5	83.3	100.0
F1	87.3	100.0
F2	88.0	98.7
F3	92.7	92.0
F4	87.3	100.0
F5	67.3	100.0
평균 인식률	83.2	98.4

10명의 화자에 대해 6가지의 방법으로 적응한 경우의 평균 인식률을 표 2에 나타내었다.

표 2. 화자적응방법에 따른 평균 인식률(%)
Table 2. The average recognition rate of each speaker adaptation method (%).

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 14) 사용
EXP3 : 화자적응된 평균(식 11)과 분산 σ^2 (식 14) 사용
EXP4 : 훈련데이터의 평균과 화자적응된 분산(식 18) 사용
EXP5 : 화자적응된 평균(식 25)과 분산(식 26) 사용
EXP6 : 가지마다의 훈련데이터 평균과 화자적응된 분산(식 31) 사용, 여기서 평균 가지 수 M을 2로 하였다.

표 2의 결과를 보면 다른 방법에 비해서 제안된 EXP6 방법이 토큰 수에 상관없이 가장 우수하다. EXP6 방법으로 했을 때 다른 방법에 비해 토큰 수가 1개일 때는 오인식률이 25% 이상, 2개일 때는 50% 이상, 그리고 3개일 때는 40% 이상 감소되었다. 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 3에 나타내었다.

표 3. EXP6 방법에서 평균 가지 수 M에 따른 인식률(%)
Table 3. The 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 3에서 토큰 수에 상관없이 M이 2일 경우에 가장 적응이 잘되지만 M값에 따라 큰 차이는 없음을 볼 수 있다.

2.3.2 ETRI 음성데이터

제안한 방법의 범용성을 입증하기 위하여 ETRI의 셈돌이 데이터에 대해서 같은 방법으로 실험하여 그 결과를 표 4에 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 4의 적응 결과도 첫 번째 그룹에서 만든 모델을 두 번째 그룹에 속하는 화자들에 적응시킨 결과이다.

표 4에서 EXP6 방법이 토큰 수에 관계없이 다른 방법에 비해 가장 우수함을 볼 수 있다. EXP6 방법으로 했을 때 다른 방법에 비해 토큰 수가 1개일 때는 4% 이상 오인식률이 감소되었고, 2개일 때는 12% 이상 감소되었다. 이때, EXP6 방법에서 평균 가지 수

M을 2로 하였다. 토큰 수가 1개일 때는 MLE 방법으로 실험한 EXP1의 결과가 지역명 데이터의 경우와 마찬가지로 매우 떨어지는 것을 볼 수 있다.

표 4의 결과를 보면 지역명 데이터의 경우보다 적응률은 조금 떨어지지만 전반적인 양상은 비슷하다. 이 결과들로부터, 본 논문에서 제안한 방법이 일반성을 가짐을 확인하였다.

표 4. 화자적응방법에 따른 평균 인식률(%)
Table 4. The average recognition rate of each speaker adaptation method (%).

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

2.3.3 분산행렬에 따른 가지 수 결정방법

ETRI 셈돌이 음성데이터를 이용하여 프레임 수에 따라 가지 수를 변화시킨 방법중 가장 우수한 인식성능을 나타내는 EXP6 방법에 대하여 화자적응 인식실험을 수행 하였다. 가지 수는 식 (35)를 이용하여 결정 하였으며 최대 가지 수 M은 3으로 하였다. 인식결과를 기존의 하나의 가지를 사용하는 방법과 프레임 수에 따라 가지 수를 결정하는 방법의 결과와 함께 표 5에 나타내었다.

표 5. 화자적응방법에 따른 평균 인식률 (%)
Table 5. The average recognition rate of each speaker adaptation method (%).

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 훈련과 정시 오류가 생기는 것을 막을 수도 있으며 간단한 수식을 이용하여 효과적으로 가지 수를 결정할 수 있으므로 상태당의 가지 수를 결정하는 적절한 방법이라 할 수 있다.

III. ARHMM에서의 화자적응화

3.1 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 불특정 화자들에 의해 훈련된 화자독립모델을 이용하여 입력 음성을 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 적합치 않으므로 식 36과 같이 프레임수에 비례하여 가지 수를 달리하는 방법을 사용한다.

$$m_j = \frac{N \times \sum_k n_{jk}}{\sum_k \sum_j n_{jk}} \times M \tag{36}$$

여기서, m_j는 상태 j에서의 가지수를, n_{jk}는 k번째 훈련 음성의 상태 j에서의 프레임 수를 나타낸다. 그리고 N, M은 각각 모델의 상태수, 평균 가지수를 나타낸다. 여러 개의 가지를 가지는 방법의 전체적인 블록도는 그림 3과 같다.

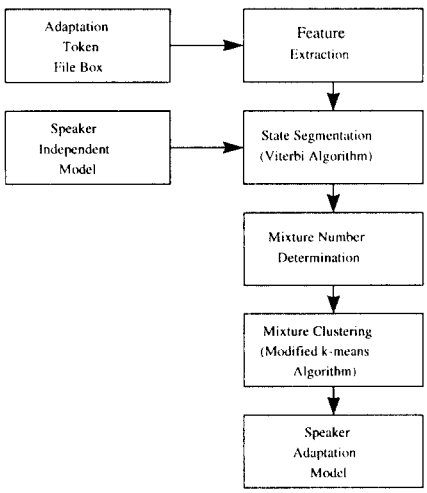


그림 3. 여러 개의 가지를 갖는 화자적응 알고리듬의 블록도
Fig. 3 The block diagram of speaker adaptation algorithm with variable mixtures.

3.2 인식 실험 및 결과

한국어 지역명 고립단어에 대한 화자독립 인식시스템의 인식율은 표 6과 같다. 이때 인식실험은 테스트하고자 하는 1명을 제외한 나머지 사람들로 훈련시킨 후 인식테스트를 하는 round-robin 방식을 사용하였다. 화자종속 인식시스템의 인식율도 표 6에 함께 나타내었다. 화자독립 및 화자종속인식 두 경우 모두 5개의 가지를 사용하였다.

본 논문에서 제안된 방법을 이용하여 ARHMM을 10명의 화자들에 대해 인식실험에서의 평균 인식율을 표 7에 나타내었다.

표 7에서의 결과를 보면 다른 방법에 비해서 EXP2의 방법이 가장 우수함을 볼 수 있다. MLE를 사용한 훈련 방법을 사용한 EXP1에서는 훈련 토큰 수가 적을 때는 인식율이 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.

표 6. 화자독립 및 종속 인식기의 인식율(%)
Table 6. 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 인식율	80.00	98.80

표 7. 화자적응 방법에 따른 화자들의 평균 인식율(%)
Table 7. 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

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

IV. 결 론

본 연구는 CDHMM과 ARHMM을 이용하여 화자적응화 하는 방법을 제안하였다. 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 가지수를 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 연구에서는 화자독립모델을 사용하여 적응음성을 각 상태로 분할한 후 modified k-means 알고리듬을 이용하여 하나의 가지로 대표하는 방법을 제안하였다. 제안된 알고리듬을 사용하였을 경우 오인식율이 화자독립인식에 비해 50% 이상 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 때문이다.

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

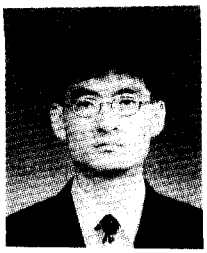
참 고 문 헌

1. 김광태, 서정일, 홍재근, "ARHMM 에서의 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. 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. 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.
5. Y. Shiraki and M. Honda, "Speaker adaptation algorithms for segment vocoder," *IEICE*, vol. SP87-67, pp. 49-56, Oct. 1987 (in Jananese).
6. S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," *in Proc. ICASSP89 (Glasgow, Scotland)*, May 1989, pp. 286-289.
7. 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.
8. K. Shikano, K.-F. Lee, and R. Reddy, "Speaker adaptation through vector quantization," *in Proc. ICASSP86 (Tokyo, Japan), Apr.* 1986, pp. 2643-2646.
9. R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 6, June 1987.
10. P. F. Brown, C. H. Lee, and J. C. Spohrer, "Bayesian adaptation in speech recognition," *in Proc. ICASSP83 (Boston, MA)*, Apr. 1983, pp. 761-764.
11. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
12. M. H. DeGroot, *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.
13. 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.

14. 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.

김 광 태(Kwang-Tae Kim) 정회원
1985년 2월 : 경북대학교 공과대학 전자공학과(공학사)
1987년 2월 : 경북대학교 대학원 전자공학과(공학석사)
1998년 8월 : 경북대학교 대학원 전자공학과(공학박사)
1989년～1993년 : 국방과학연구소 연구원
1994년 3월～현재 : 상주대학교 전자전기공학과 조교수
※주관심분야 : 음성인식, 음성신호처리, VLSI 설계



한 유 수(Yoo-Soo Han) 정회원
1996년 2월 : 경북대학교 전자공학과(공학사)
1998년 2월 : 경북대학교 대학원 전자공학과(공학석사)
1994년 3월～현재 : 경북대학교 전자공학과 박사과정
※주관심분야 : 음성인식, 음성신호처리

홍 재 근(Jae-Keun Hong) 정회원
1975년 2월 : 경북대학교 공과대학 전자공학과(공학사)
1979년 2월 : 경북대학교 대학원 전자공학과(공학석사)
1985년 2월 : 경북대학교 대학원 전자공학과(공학박사)
1979년～1982년 : 경북산업대학교 조교수
1983년～현재 : 경북대학교 전자전기공학과 교수
※주관심분야 : 음성인식, 음성신호처리, 음성합성