

# CDHMM의 상태당 가지 수를 가변시키는 화자적응에 관한 연구

## (A Study on the Speaker Adaptation in CDHMM Using Variable Number of Mixtures in Each State)

金光泰\*, 徐廷一\*\*, 洪再根\*\*

(Kwang-Tae Kim, Jeong-Il Seo, and Jae-Keun Hong)

### 요 약

CDHMM(continuous density hidden Markov model)에서 최대사후확률 추정법에 의하여 각 상태마다 하나의 가지를 이용하여 화자에 적응시킨다. 본 논문에서는 음성의 다양한 음향학적 특징을 표현하기 위하여 상태마다 여러 개의 가지를 갖는 방법을 제안한다. 적응음성의 훈련데이터 수가 적기 때문에 어떤 상태에서는 여러 개의 가지를 갖는 것이 타당하지 않을 수도 있으므로 각 상태에서 적절한 가지 수를 결정하여야 한다. 본 논문에서는 각 상태내에 속하는 프레임 수에 따라 가지 수를 달리하는 방법과 상태내의 특징벡터들의 분산행렬의 행렬식값을 이용하는 방법을 사용하였다. 15개의 한국 지역명으로 구성된 음성데이터와 40개의 단어로 구성된 ETRI의 샘플이 데이터에 대하여 인식실험한 결과 제안한 방법이 가지를 한 개 사용한 기존 방법에 비해 높은 인식률을 얻을 수 있었다.

### Abstract

When we make a speaker adapted model using MAPE (maximum a posteriori estimation), the adapted model has one mixture in each state. This is because we cannot estimate a number of a priori distribution from a speaker-independent model in each state. If the model is represented by one mixture in each state, it is not well adapted to specific speaker because it is difficult to represent various speech information of the speaker with one mixture. In this paper, we suggest the method using several mixtures to well represent various speech information of the speaker in each state. But, because speaker-specific training data is not sufficient, this method can't be used in every state. So, we make the number of mixtures in each state variable in proportion to the number of frames and to the determinant of the variance matrix in the state. Using the proposed method, we reduced the error rate than methods using one branch in each state.

### I. 서 론

HMM(hidden Markov model)을 이용한 음성인식은 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

\* 正會員, 尙州産業大學校 電子電氣工學科  
(Sangju National Polytechnic University)

\*\* 正會員, 慶北大學校 電子電氣工學部  
(Kyungpook National University)

接受日字: 1998年1月3日, 수정완료일: 1998年3月5日

별도의 훈련과정이 필요치 않으나 화자중속인식기에 비해 낮은 인식성능을 나타내게 된다.

화자적응(speaker adaptation)방법은 소량의 훈련 데이터를 사용하여 충분한 훈련데이터로 훈련된 화자 독립모델을 인식시스템을 사용하려는 특정화자에 적응시키는 방법이다<sup>[1~10]</sup>. 따라서 인식시스템을 사용하려는 화자가 화자중속인식기를 훈련시키기 위한 훈련데이터보다 적은 훈련데이터로도 화자중속인식기의 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 있다<sup>[3]</sup>. 또한 이런 적응화 기술은 채널 특성, 주변 잡음 등에 인식기를 적응시키는 환경적응화에도 이용될 수 있다. 본 논문에서는 화자독립모델로부터 특정화자로 적응시켜 화자중속모델을 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<sup>[5]</sup>, 벡터 양자화에 기준을 둔 인식기에는 코드북 적응기술이 사용된다<sup>[6]</sup>. DHMM으로 구성된 인식기에서는 히스토그램 적응과 같은 이산관측심벌 분포의 수정을 사용한다<sup>[7]</sup>.

CDHMM(continuous density hidden Markov model)<sup>[11]</sup>을 사용하는 인식기에서는 베이적응(Bayesian adaptation) 방법<sup>[12~13]</sup>을 이용한다. CDHMM에서 베이적응 방법을 이용하여 화자에 적응된 모델을 만들 때, 각 상태마다 하나의 가지를 가지는 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 가지게 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 상태내에 속하는 프레임 수에 따라 가지 수를 달리하는 방법과 상태내의 특징벡터들의 분산행렬의 행렬식값을 이용하는 방법을 사용하였다. 제안한 방법의 타당성을 입증하기 위하여 15개의 한국 지역명으로 구성된 음성데이터와 40개의 단어로 구성된 ETRI의 셸들이 데이터에 대하여 인식실험한 결과 제안한 방법이 가지를 한 개 사용했을 때에 비해 높은 인식률을 얻을 수 있었다.

## II. CDHMM 파라미터들의 화자적응

이 장에서는 CDHMM 파라미터들의 화자적응에 대하여 기술한다<sup>[3]</sup>.

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

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

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

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

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

베이정리를 이용하면,  $P(\lambda | Y)$ 는 다음 식과 같다.

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

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

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

$N$ 개의 상태를 가지는 1차 Markov chain에서, 상태열이  $s=(s_0, s_1, \dots, s_T)$ 로 주어지고, 초기상태확률이  $\pi'=[\pi_1, \pi_2, \dots, \pi_N]$ , 상태전이확률행렬이  $A=[a_{ij}]$ ,  $i, j=1, 2, \dots, N$  이라고 하면 상태열  $s$ 가 관측될 확률은 다음과 같다.

$$P(s|\pi, A) = \pi_{s_0} \prod_{i=1}^T a_{s_{i-1}s_i} \quad (4)$$

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

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

여기서  $\{s\}$ 는 모든 가능한 상태열을 의미한다. 모델  $\lambda$ 의 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 알고리즘을 사용한다. segmental k-means 알고리즘을 확장하여 MAPE 방법에 적용하면 상태열  $s$ 를 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  $\lambda$ 의 사전분포이다. 베이적응(Bayesian adaptation)과 결합된 적응 segmental k-means 알고리즘은 다음의 두 단계로 구성된다.

1) 주어진 모델  $\hat{\lambda}$ 에 근거하여 최적 상태

열을 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 다음식으로 구한다.

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

$\lambda$ 가 어떤 값에 수렴할 때까지 이 두 단계를 반복 수행한다. 반복 수행할 때 같은 데이터  $Y$ 에 대해서는 식 (9)와 (10)의 사전분포함수  $P_o(\lambda)$ 가 같다. 훈련데이터를 한꺼번에 처리하지 않고 몇 개의 작은 그룹으로 만들어서 한 번에 한 개씩 적용시킬 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.

### III. 정규분포 파라미터들의 베이적응

이번 장에서는 관측벡터 분포로 정규분포를 가지는 CDHMM의 파라미터에 대한 베이적응 방법에 대해서 기술한다<sup>[3]</sup>. 수식은 그림 1과 같은 단순 left-to-right HMM에 대하여 전개하며, 적용시킬 파라미터 수를 감소시키기 위해 각 HMM의 상태전이확률값은 화자독립모델로부터 구한 값을 사용한다. 그러므로 적용은 정규분포의 평균과 분산에 대해서만 이루어진다. 또 대각성분값만을 갖는 분산 행렬을 가지는 정규분포에 대해서만 고려할 것이다. 적용은 평균과 분산에 대해 각 상태마다 독립적으로 이루어지므로 수식은 하나의 상태에 대해서만 전개한다.

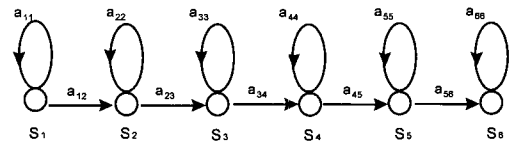


그림 1. 단순 left-to-right HMM  
Fig. 1. Simple left-to-right HMM.

#### 1. 평균의 Bayes 적응

평균  $\mu$ 가 사전분포  $P_o(\mu)$ 를 가지는 불특정값이고, 분산  $\sigma^2$ 이 상수일 때,  $P_o(\mu)$ 가 평균  $\nu$ 와 분산  $\tau^2$ 을 가지는 정규분포라고 가정하면  $\mu$ 의 MAP 추정치는

다음과 같다.

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

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

$\mu$ 의 MAP 추정치는 사전분포의 평균  $\nu$ 와 샘플 평균  $\bar{y}$ 의 하중값을 달리한 평균이다. 이 하중값은  $n$ 이 0이면 부가적인 훈련데이터의 사용없이  $\nu$ 의 하중값만 1로 만들게 되어 기존의 평균값을 그대로 사용하는 것이 된다. 훈련데이터의 수가 많아지면,  $\mu$ 의 MAP 추정치는 ML 추정치와 거의 같게 된다. 사전분포의 분산  $\tau^2$ 이  $\sigma^2/n$ 보다 훨씬 클 경우에도  $\mu$ 의 MAP 추정치는 ML 추정치와 거의 같게 된다.

사전분포의 평균  $\nu$ 와 분산  $\tau^2$ 은 다음과 같이 추정된다.

$$\nu = \sum_{m=1}^M w_m \nu_m \quad (12)$$

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

여기서  $\nu_m$ ,  $w_m$ 은 각각 화자독립모델의  $m$ 번째 가지의 평균과 가중치이다.

$\sigma^2$ 은 다음 식과 같이 각 가지의 가중된 분산을 사용하여 구할 수 있다.

$$\sigma^2 = \sum_{m=1}^M w_m \sigma_m^2 \quad (14)$$

여기서  $\sigma_m^2$ 은  $m$ 번째 가지의 분산이다.

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

$$\hat{\tau}^2 = \frac{\sigma^2}{\sigma^2 + n\tau^2} \tau^2 \quad (15)$$

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

## 2. 분산의 Bayes 적용

평균  $\mu$ 가 미지의 값이고, 분산의 사전분포가 다음 식과 같다면, 분산  $\sigma^2$ 의 MAP 추정치는 식 (17)의

조건으로부터 식 (18)과 같이 구할 수 있다.

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

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

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

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

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

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

## 3. 평균과 분산의 Bayes 적용

평균과 분산 모두 어떤 사전분포를 가지는 불특정값이라고 가정하고 적용된 모델을 구할 수 있다. 이때 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.

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

여기서  $\theta$ 가 주어졌을 때  $\mu$ 의 조건부 확률분포는 평균  $\nu$ 와 분산  $\tau^2 = 1/\omega\theta$ 를 가지는 정규분포이고,  $\theta$ 의 marginal 분포는 파라미터  $\alpha, \beta(>0)$ 를 가지는 감마분포이다.

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

$$\hat{\nu} = \frac{\omega\nu + n\bar{y}}{\omega + n} \quad (21)$$

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

리고 샘플 데이터가 주어졌을 때  $\theta$ 의 marginal 분포는 감마분포이고, 이것의 파라미터  $\hat{\alpha}$ 와  $\hat{\beta}$ 는 다음과 같다.

$$\hat{\alpha} = \alpha + \frac{n}{2} \quad (23)$$

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

위 분포로부터 구한 평균과 분산의 MAP 추정치는 다음과 같다.

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

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

위 식에서 사용된 파라미터들은 다음과 같다.

$$\nu = \sum_{m=1}^M w_m \nu_m \quad (27)$$

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

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

$$\beta = 1 \quad (30)$$

여기서  $\nu$ 는 각 가지에 하중값을 부여하여 구한 평균 값이고,  $\alpha$ 는 precision 파라미터이다.  $\omega$ 는 각 가지의 분산을 평균하여 구한 값과 각 가지의 평균으로부터 구한 분산과의 비율을 나타낸다.  $\beta$ 는 1로 두었다.

#### IV. 상태마다 프레임 수에 따라 여러 개의 가지를 사용하는 방법

앞장에서 살펴본 방법을 이용하여 CDHMM에서 화자에 적응된 모델을 만들 때, 각 상태마다 하나의 가지를 가지는 모델로 만들어 적용시킨다. 그러나 상태마다 하나의 가지로는 적용시키려는 화자의 다양한 음성정보를 적절히 나타내지 못하기 때문에 모델을 그 화자에 적용시키는데 한계를 가지게 된다. 이점을 해결하기 위하여 상태마다 여러 개의 가지를 사용하는 방법을 생각해 볼 수 있다. 하지만 앞장에서 설명한 방법 중, 첫 번째 방법과 세 번째 방법에서는 이 방법

을 적용하기가 어렵다. 이것은 이미 만들어져 있는 인식기의 특정상태에 속하는 각 가지들의 평균값과 분산값의 분포를 적용시킬 데이터의 사전분포를 이용하는 데, 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. 또한 가지 수를 여러개 사용하여도 적응에 쓰이는 음성의 개수가 적고, 대각성분의 분산 행렬식값만을 이용하기 때문에 계산량은 크게 증가하지 않는다.

##### 1. 프레임 수에 따른 가지 수 결정방법

여러 개의 가지를 갖게 하기 위하여 입력 벡터열  $Y$ 를 k-means 알고리즘을 사용하여 몇 개의 클러스터(cluster)로 분리한 후 이 각각에 대한 데이터 분산인  $S_{ym}^2$ 을 구하여 사용한다.  $S_{ym}^2$ 를 구하는 식은 다음과 같다.

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

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.

$$m_j = \lceil \frac{N \times \sum_k n_{jk}}{\sum_k \sum_j n_{jk}} \times M \rceil \quad (32)$$

여기서,  $m_j$ 는 상태  $j$ 에서의 가지 수를,  $n_{jk}$ 는  $k$ 번째 훈련음성의 상태  $j$ 에서의 프레임 수를 나타낸다. 또한, 기호  $\lceil \cdot \rceil$ 는 올림을 의미한다. 그리고  $N$ ,  $M$ 은 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 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 좋지 않으며 프레임 수에 따라 사용할 적절한 가지 수를 찾는 것이 중요하다. 즉, 식 (32)에서 사용할 평균 가지 수  $M$ 의 적절한 값을 찾아야 한다.

## 2. 분산행렬의 행렬식값에 따른 가지 수 결정방법

앞절에서 설명한 프레임 수에 따른 가지 수 결정방법은 복잡한 연산과정이 필요 없으므로 수행과정이 간단하다. 그러나, 상태내의 프레임 수만으로 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.

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

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

여기서  $T$ 는 상태  $j$ 에 속하는 프레임 수이고  $c_i$ 는 상태  $j$ 에서의  $i$ 번째 프레임의 캡스트럼 값이며  $\mu_j$ 는 상태  $j$ 에서의 평균값이다.  $N$ 과  $P$ 는 각각 상태수와 캡스트럼 차수를 나타낸다. 상태  $j$ 에서의 분산행렬의 행렬식값(determinant)은 식 (34)와 같이 표현할 수 있다.

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

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.

$$m_j = \lceil \frac{D_{\max} M}{10 D_j} \rceil \quad (35)$$

여기서  $D_{\max}$ 와  $M$ 은 최대 행렬식값과 평균 가지 수를 나타내며 최소한  $m_j$ 가 1이 되도록 즉 최소한 1개의 가지를 갖도록 각 상태에서의 가지 수를 결정하였다.

## V. 실험 결과 및 고찰

본 논문에서는 두 개의 음성데이터를 사용하였다. 첫번째 음성데이터는 10명(남자 5명, 여자 5명)의 화자가 15개 한국 지역명을 10번씩 발음한 것이다. 두번째 음성데이터는 ETRI의 샘돌이 음성데이터로 40명(남자 20명, 여자 20명)의 화자가 40개의 단어를 4번씩 발음한 것이다. 샘돌이 데이터는 한국어 고립사와 고립단어들로 구성되어 있다. 첫 번째로 프레임 수에 따라 가지 수를 결정하는 방법에 대해 실험해 보았다.

### 1. 지역명 음성데이터

다음은 지역명 음성데이터에 대해 사용된 값이다.

- 1) 저역통과여파기의 차단주파수: 3.4kHz
- 2) 표본화 주파수: 8kHz
- 3) 양자화 비트수: 12 bits/sample
- 4) 프레임 길이: 300 samples
- 5) 프레임 이동거리: 100 samples
- 6) Preemphasis 전달함수:  $1 - 0.95z^{-1}$
- 7) LPC 캡스트럼 차수: 12차
- 8) HMM 상태수: 6개

모든 실험에서 단순 left-to-right HMM을 사용하였으며, 해밍 창함수를 사용하였다. 상태천이확률의 변화는 인식률에 별다른 영향을 못미쳐 화자독립모델의 확률값을 그대로 사용하였다.

각 화자별 화자독립인식률을 표 1에 나타내었다. 이때 인식실험은 테스트하고자 하는 1명을 제외한 나머지 사람들로 훈련시킨 후 인식테스트를 하는 round-robin 방식을 사용하였다. 화자중속인식률도 표 1에 나타나 있다. 화자독립 및 화자중속인식 두 경우 모두 3개의 가지를 사용하였다.

10명의 화자에 대해 6가지의 방법으로 적용한 경우

의 평균 인식률을 표 2에 나타내었다.

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

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  $\sigma^2$  (식 14) 사용

EXP3: 화자적응된 평균(식 11)과 분산  $\sigma^2$ (식 14) 사용

EXP4: 훈련데이터의 평균과 화자적응된 분산(식 18) 사용

EXP5: 화자적응된 평균(식 25)과 분산(식26) 사용

EXP6: 가지마다의 훈련데이터 평균과 화자적응된 분산(식 31) 사용, 여기서 평균 가지 수  $M$ 을 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  $M$ 에 따른 인식률을 표 3에 나타내었다.

표 3. EXP6 방법에서 평균 가지 수  $M$ 에 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  $M$ 이 2일 경우에 가장 적응이 잘되지만  $M$ 값에 따라 큰 차이는 없음을 볼 수 있다. 식 (32)에서  $m_i$  값이 1보다 작게 되는 경우를 방지하기 위하여 올림기호 ( $\lceil \cdot \rceil$ )를 사용하기 때문에  $M$ 값을 1로 하더라도  $\lceil 0.2 \times 1 \rceil = 1$ 이고,  $\lceil 1.6 \times 1 \rceil = 2$  이므로 실제로 쓰이는 가지 수의 평균은 1보다 크게 된다. 즉, 모든 상태는 하나이상의 가지를 갖게 된다 식 (32)에서  $m_i$  값이 1보다 작게 되는 경우를 방지하기 위하여 올림기호 ( $\lceil \cdot \rceil$ )를 사용하기 때문에  $M$ 값을 1로 하더라도  $\lceil 0.2 \times 1 \rceil = 1$ 이고,  $\lceil 1.6 \times 1 \rceil = 2$  이므로 실제로 쓰이는 가지 수의 평균은 1보다 크게 된다. 즉, 모든 상태는 하나이상의 가지를 갖게 된다

남자화자중 가장 적응이 안되는 화자 M2, 10명의 화자중 가장 적응이 잘 되는 여성화자 F1, 가장 적응이 잘 안되는 여성화자 F3의 인식률을 표 4, 5, 6에 각각 나타내었다.

화자별 인식률을 보아도 EXP6 방법이 같은 수의 토큰으로 훈련한 다른 방법들보다 항상 인식률이 좋은 것을 볼 수 있다. 또 토큰 수가 1개일 때 EXP1보다 EXP6의 결과가 특히 더 좋았다. 남자화자중 가장 적응이 잘 안되는 M2 화자일 경우에 EXP6 방법이 다

른 방법들에 비해 특히 더 높은 인식률을 보였다. 토큰 수가 적을 때 이 차이가 더 크게 나타났다. 적응이 가장 잘 되는 F1 화자의 경우에는 모든 방법에서 좋은 결과가 나왔지만 마찬가지로 다른 방법에 비해 EXP6 방법이 나쁘지 않은 것을 볼 수 있다. 화자 F3의 경우에도 M2의 경우와 비슷한 결과를 나타내었다.

표 4. 화자적응방법에 따른 화자 M2의 인식률 (%)

Table 4. The recognition rate of speaker M2 in each speaker adaptation method (%).

Tokens	EXP1	EXP2	EXP3	EXP4	EXP5	EXP6
1	80.0	83.3	80.0	85.6	80.0	97.8
2	96.7	94.4	92.2	95.6	92.2	97.8
3	97.8	94.4	95.6	98.9	94.4	98.9

표 5. 화자적응방법에 따른 화자 F1의 인식률 (%)

Table 5. The recognition rate of speaker F1 in each speaker adaptation method (%).

Tokens	EXP1	EXP2	EXP3	EXP4	EXP5	EXP6
1	94.4	100.0	97.7	100.0	98.8	100.0
2	99.6	100.0	100.0	100.0	100.0	100.0
3	100.0	100.0	100.0	100.0	100.0	100.0

표 6. 화자적응방법에 따른 화자 F3의 인식률 (%)

Table 6. The recognition rate of speaker F3 in each speaker adaptation method (%).

Tokens	EXP1	EXP2	EXP3	EXP4	EXP5	EXP6
1	60.0	84.4	82.2	85.6	81.1	92.2
2	86.7	85.5	85.6	85.6	83.3	95.6
3	96.7	94.4	94.4	97.8	96.7	98.9

2. 썸돌이 음성데이터

제안한 방법의 범용성을 입증하기 위하여 ETRI의 썸돌이 데이터에 대해서 같은 방법으로 실험하여 그 결과를 표 7에 나타내었다.

썸돌이 데이터는 40개의 단어로 한국어 고립숫자 및 단어로 구성되어 있다. 다음은 썸돌이 데이터에 사용된 값이다.

- 1) 지역통과여파기의 차단주파수: 7kHz
- 2) 표본화 주파수: 16kHz
- 3) 양자화 비트수: 16 bits/sample
- 4) 프레임 길이: 400 samples
- 5) 프레임 이동거리: 200 samples
- 6) Preemphasis 전달함수:  $1 - 0.95z^{-1}$
- 7) LPC 캡스트럼 차수: 12차
- 8) HMM 상태수: 4개

썸돌이 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 7의 적응 결과도 첫 번째 그룹에서 만든 모델을 두 번째 그룹에 속하는 화자들에 적용시킨 결과이다.

표 7. 화자적응방법에 따른 평균 인식률 (%)

Table 7. 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

표 7에서 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  $M$ 을 2로 하였다. 토큰 수가 1개일 때는 MLE 방법으로 실험한 EXP1의 결과가 지역명 데이터의 경우와 마찬가지로 매우 떨어지는 것을 볼 수 있다.

표 7의 결과를 보면 지역명 데이터의 경우보다 적응률은 조금 떨어지지만 전반적인 양상은 비슷하다. 이 결과들로부터, 본 논문에서 제안한 방법이 일반성을 가짐을 확인하였다.

3. 분산행렬의 행렬식값에 따른 가지 수 결정방법

ETRI 썸돌이 음성데이터를 이용하여 프레임 수에 따라 가지 수를 변화시킨 방법중 가장 우수한 인식성능을 나타내는 EXP6 방법에 대하여 화자적응 인식실



험을 수행 하였다. 가지 수는 식 (35)를 이용하여 결정 하였으며 평균 가지 수  $M$ 은 3으로 하였다. 인식 결과를 기존의 하나의 가지를 사용하는 방법과 프레임 수에 따라 가지 수를 결정하는 방법의 결과와 함께 표 8에 나타내었다.

표 8. 화자적응방법에 따른 평균 인식률 (%)

Table 8. 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

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

## VI. 결 론

본 논문에서는 이미 만들어져 있는 CDHMM으로 구성된 화자독립모델을 사용하려는 특정화자에 맞도록 적응시켜 화자종속모델을 만드는 화자적응방법에 대하여 살펴보았다. 이때 기존의 방법인 각 상태마다 하나의 가지로 적응시키는 방법 대신 각 상태에 속하는 적응음성의 프레임 수와 분산행렬의 행렬식값을 이용하여 상태마다 여러 개의 가지를 사용하는 방법을 제안

하였다.

제안된 두가지 방법 모두 기존의 하나의 가지를 갖도록 적응시키는 방법보다 높은 인식률을 얻을 수 있었다. 이것은 제안된 방법이 프레임이 많이 물리는 상태에서의 분포를 좀 더 세밀히 나타낼 수 있기 때문이다. 적응음성의 프레임 수에 따라 가지 수를 결정하는 방법이 계산상으로 간단하면서도 분산행렬의 행렬식값을 이용하는 방법보다 우수한 인식성능을 나타내었다.

15개의 지역명으로 구성된 음성데이터를 사용하여 제안한 방법으로 적응음성을 각 단어마다 1개에서 3개까지 늘려가며 화자적응을 시켰을 때, 모든 경우에서 제안한 방법의 성능이 더 우수함을 확인하였다. 또 적응음성의 개수에 비례하여 인식률도 따라 증가함을 볼 수 있었다. 그리고 제안한 방법이 같은 개수의 음성데이터로 MLE 방법을 사용하여 혼련하고 테스트하였을 때보다 항상 더 높은 인식률을 얻는다는 것도 확인하였다. 특히 적응음성의 개수가 적을 때는 MLE 방법을 사용한 방법이 상당히 나쁜 결과를 보였는데, 이것은 데이터 수가 적을 때에는 MLE 방법으로 분산을 제대로 추정할 수가 없기 때문이다.

그리고 본 논문의 일반성을 입증하기 위해 ETRI의 샘플이 데이터에도 적용시켜 본 결과 지역명 음성 데이터에서와 비슷한 결과를 나타냄을 확인 하였다.

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

## 참 고 문 헌

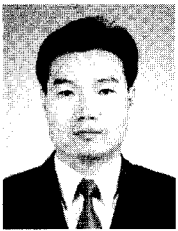
- [1] 김광태, 서정일, 홍재근, "ARHMM에서의 화자적응," *한국정보처리학회 추계학술발표 논문집*, Vol. 4, No. 2, pp 1184-1188, 1997.
- [2] 한우수, 서정일, 김광태, 홍재근, "연속 혼합 가우스 밀도를 가지는 HMM에서의 화자적응," *신호처리합동 학술대회*, Vol. 10, No. 1, pp. 317-320, 1997
- [3] C. H. Lee, C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. on Signal Processing*, vol. 39, no. 4, pp. 806-814, Apr. 1991.

- [4] Y. Shiraki and M. Honda, "Speaker adaptation algorithms for segment vocoder," *IEICE*, vol. SP87-67, pp. 49-56, Oct. 1987 (in Japanese).
- [5] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," in *Proc. ICASSP89 (Glasgow, Scotland)*, May 1989, pp. 286-289.
- [6] K. Shikano, K.-F. Lee, and R. Reddy, "Speaker adaptation through vector quantization," in *Proc. ICASSP86 (Tokyo, Japan)*, Apr. 1986, pp. 2643-2646.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
- [13] M. H. DeGroot, *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

---

 저 자 소 개
 

---



金 光 泰(正會員)

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 상주 산업대학교 전자전기공학과 조교수. 주관심분야는 음성 인식, 음성신호 처리



徐 廷 一(正會員)

1994년 2월 경북대학교 공과대학 전자공학과 졸업(공학사). 1996년 2월 경북대학교 대학원 전자공학과 졸업(공학석사). 1996년 3월 ~ 현재 경북대학교 대학원 전자전기공학부 박사과정

## 洪 再 根(正會員)

1975년 2월 경북대학교 공과대학 전자공학과 졸업(공학사). 1979년 2월 경북대학교 대학원 전자공학과 졸업(공학석사). 1985년 2월 경북대학교 대학원 전자전기공학과 졸업(공학박사). 1979년 ~ 1982년 경북산업대학교 조교수. 1983년 ~ 현재 경북대학교 전자전기공학부 교수