論文2002-39SP-2-13

# CDHMM의 화자적응에 관한 연구

# (A Study on the Speaker Adaptation in CDHMM)

金 光 泰 *

(Kwang-Tae Kim)

요 약

본 논문에서는 CDHMM 음성인식기의 인식성능을 향상시키기 위해 상태 당 관측밀도함수 수 변화에 의한 화자적응 알고리듬을 제안하였다. 제안한 방법은 CDHMM의 각 상태마다 관측 확률밀도함수의 가지 수가 두 개 이상이 될 수도 있게 하여 발음특성의 다양성을 반영할 수 있게 하였다. 가지 수는 각 상태에 속하는 적응음성의 프레임 수에 따라 정하는 방법과 특징벡터 행렬식에 따라 정하는 방법으로 하였다. 이 두 방법중의 어느 하나로 관측 확률밀도함수의 가지가 결정되면, 세분화된 각 가지로부터 MAP 파라미터를 추출함으로써 정밀한 화자적응모델의 파라미터를 구할 수 있었다. 아울러 적응음성을 상태분할 할 때 기존의 화자독립모델을 사전정보로 이용함으로써 ML 추정시의 초기 상태분할 오류의 영향을 줄여 기존 상태분할 방법의 단점을 개선하였다. 그리고 상태지속분포를 화자에 적응시킴으로써 화자 고유의 발음속도와 발음 패턴 등의 음성특성을 흡수하도록 하였다. 제안한 방법들의 타당성을 확인하기 위한 실험에서 제안한 방법이 기존 방법에 비해 높은 인식률을 얻음을 확인하였다.

**Abstract**

A new approach to improve the speaker adaptation algorithm by means of the variable number of observation density functions for CDHMM speech recognizer has been proposed. The proposed method uses the observation density function with more than one mixture in each state to represent speech characteristics in detail. The number of mixtures in each state is determined by the number of frames and the determinant of the variance, respectively. The each MAP parameter is extracted in every mixture determined by these two methods. In addition, the state segmentation method requiring speaker adaptation can segment the adapting speech more precisely by using speaker-independent model trained from sufficient database as a priori knowledge. And the state duration distribution is used for adapting the speech duration information owing to speaker's utterance habit and speed. The recognition rate of the proposed methods are significantly higher than that of the conventional method using one mixture in each state.

**Key Word** : Speaker adaptation, speech recognition, speaker dependent, speaker independent, Viterbi algorithm.

## Ⅰ. 서 론

최근에 디지털 신호처리기술과 컴퓨터 응용기술, 대규모 집적회로 등 제반 기술의 발전으로 인해 언어인식(speech recognition)이 실용화 단계에 와 있다.[1] 음성인식 방법에는 크게 패턴정합(pattern matching)에 의한 방법[1-2]과 신경회로망(neural network)을 이용하는 방법,[1, 3] 그리고 HMM (hidden Markov model) 방법[1] 등이 있다.

음성인식은 화자종속(speaker dependent) 인식과 화자독립(speaker independent) 인식으로 나눌 수 있다.

* 正會員, 尙州大學校 電子·電氣工學部
(SangJu National University)

接受日字:2001年2月6日, 수정완료일:2002年3月5日

화자종속인식은 특정화자에 의해 훈련된 모델로 그 화자의 음성을 인식하는 것이고, 화자독립인식은 다수화자에 의해 훈련된 모델로 불특정 화자의 음성을 인식하는 것이다. 화자마다 발성방법과 환경적인 요인이 다양하기 때문에 일반적으로 화자독립 인식기는 화자종속 인식기에 비해 인식 성능이 떨어지는 단점이 있다. 그러나 제한된 수의 음성으로 훈련된 화자종속 인식기는 충분한 수의 음성으로 훈련된 화자독립 인식기보다 인식성능이 떨어진다. 따라서 훈련음성의 수가 적은 경우에는 충분히 훈련된 화자독립모델을 특정화자의 음성특성에 적응시키는 방법이 필요하다.

화자적응(speaker adaptation)[5-10]은 충분한 음성데이터로 훈련된 화자독립모델 인식기를 사전정보로 하여 특정화자의 소량의 음성데이터를 사용하여 그 화자에 적응시키는 것이다. 그러므로 인식기를 사용하려는 화자가 화자종속 인식기를 훈련시키기 위한 훈련데이터보다 적은 수의 음성데이터로도 화자종속 인식기에 가까운 인식성능을 얻을 수 있다. 또한 화자적응은 인식기가 사용되는 환경이 바뀌거나 잡음이 있는 곳에서도 우수한 인식성능을 나타낸다.

화자적응기법[5]에는 다음과 같은 것들이 있다. 화자독립모델을 새로운 화자독립 데이터를 사용하여 최신화하는 적응 클러스터링(adaptive clustering), 특정화자에 맞춰 훈련된 모델을 약간의 훈련데이터를 사용하여 새로운 화자의 모델로 변환하는 화자변환(speaker conversion), 화자독립모델이나 여러 화자의 모델로부터 특정화자의 훈련데이터를 사용하여 그 화자로 적응시키는 화자적응, 특정화자의 훈련데이터가 긴 시간에 걸쳐서 입력될 때, 매 시간 새로운 훈련데이터로 특정화자 모델을 순차적으로 적응시키는 순차적 적응(sequential adaptation) 등이 있다.

화자적응방법은 음성인식기가 어떤 음성 패턴들로 모델링되어 있는가에 따라 달라진다. 이산관측 HMM에서는 히스토그램(histogram) 적응과 같은 이산관측심벌 분포의 수정방법[7-9, 11]이 사용된다. 최근에는 컴퓨터의 계산속도의 향상으로 인식률이 좋은 연속관측 HMM의 화자적응을 많이 사용한다. 연속관측 HMM의 화자적응에서는 최대사후확률 추정(maximum a posteriori estimation, MAPE) 방법 즉, 베이적응(Bayesian adaptation) 방법[5, 12]이 이용된다.

연속관측 HMM을 화자적응시키는 기존의 방법에서는 적응음성의 수가 적기 때문에 각 상태마다 하나의 가지를 갖는 모델로 만들어 적응시킨다.[13] 그러나 각 상태마다 하나의 가지로는 화자의 다양한 음성정보를 적절히 나타내지 못하여 적응에 한계를 나타내게 된다. 그러므로 각 상태의 음성정보를 자세히 나타내기 위하여 관측 확률밀도함수의 가지를 여러 개 사용하는 방법이 필요하다. 그리고 적응음성을 상태분할할 때 ML (maximum likelihood) 추정법[4]을 사용하는데, ML 추정법에는 훈련데이터의 양이 무한하다는 가정이 전제되어 있으나 실제로는 적응음성 데이터의 양이 많지 않으므로 상태분할이 정확하지 못하다. 또한 화자들 사이의 성도 길이, 구강 크기, 발성습관 등에 차이가 있으므로 화자고유의 발음 습성을 포함할 수 있도록 상태지속분포의 적응방법이 요구된다.

본 논문에서는 연속관측 HMM 음성인식기의 인식성능을 향상시키기 위해 상태당 관측밀도함수 수 변화에 의한 화자적응 알고리듬을 제안하였다. 제안한 방법은 연속관측 HMM 언어인식기가 적응화자의 다양한 음성정보를 좀 더 세밀하게 표현할 수 있도록 구현되었으며 다음과 같다.

연속관측 HMM의 각 상태마다 관측 확률밀도함수의 가지 수가 두 개 이상이 될 수도 있게 하여 발음특성의 다양성을 반영할 수 있게 하였다. HMM의 상태에 따라 그 상태에 속하는 프레임의 수가 달라서 프레임 수가 적은 상태에서는 여러 개의 가지를 사용할 수가 없게 된다. 이는 상태마다 관측밀도함수의 가지 수를 달리하여 해결할 수 있다. 가지 수는 각 상태에 속하는 적응음성의 프레임(frame) 수에 따라 정하는 방법과 특징벡터 행렬식(determinant)에 따라 정하는 방법으로 하였다. 이 두 방법중의 어느 하나로 관측 확률밀도함수의 가지가 결정되면, 세분화된 각 가지로부터 MAP 파라미터를 추출함으로써 정밀한 화자적응모델의 파라미터를 구할 수 있었다.

아울러 적응음성을 상태분할할 때 기존의 화자독립 모델을 사전정보로 이용함으로써 ML 추정시의 초기 상태분할 오류의 영향을 줄여 기존 상태분할 방법의 단점을 개선하였다. 그리고 상태지속분포를 화자에 적응시킴으로써 화자 고유의 발음속도와 발음패턴 등의 음성특성을 흡수하도록 하였다.

본 논문에서 제안한 방법들의 타당성을 확인하기 위하여 한국 도시명 음성데이터와 ETRI(Electronics and Telecommunications Research Institute)의 샘돌 데이터에 대해 인식실험하여 제안한 방법이 기존 방법에

비해 높은 인식률을 얻을 수 있음을 확인하였다.

## Ⅱ. 제안한 화자적응방법

화자독립 HMM을 사전정보로 하여 소량의 적응음성 데이터를 가지고 음성인식기를 사용하려는 화자에 최적화된 새로운 화자종속 인식기를 만드는 것이 화자적응이다. 기존의 화자적응방법들은 제한된 훈련데이터로 인하여 상태당 관측 확률밀도함수를 한 개 사용하거나 일률적인 개수를 사용하여 적응모델을 만들었다.

화자의 다양한 음성정보를 반영하기 위하여 상태마다 여러 개의 가지를 사용하여야 하지만 화자적응에서는 훈련데이터 수가 적기 때문에 상태에 속하는 프레임 수가 적은 경우가 많이 생겨 모든 상태에서 여러 개의 가지를 사용할 수가 없다. 이때 상태마다 관측 확률밀도함수의 가지 수를 달리하면 이 문제를 해결할 수 있으며, 이를 위하여 관측 확률밀도함수의 가지 수를 변화시키는 알고리듬을 사용하였다. CDHMM에서는 세분화된 가지마다의 적응된 평균과 분산을 구하여 사용하였다. 그리고 적응음성의 상태분할방법으로 기존의 화자독립모델을 사전정보로 이용하여 적응음성을 상태분할함으로써 ML 추정시의 초기 상태분할 오류의 영향을 줄일 수 있고, 보다 정확한 상태분할을 할 수 있게 하였다.

### 1. 관측 확률밀도함수 수 변화 알고리듬

적응시킬 화자의 소량의 적응음성 데이터로부터 그 화자의 발성정보를 최대한 얻기 위해서는 각 상태마다의 관측 확률밀도함수를 좀 더 세밀히 표현할 필요가 있다.

화자종속모델을 만들 경우에는 화자독립모델에서와 같이 상태마다 관측 확률밀도함수의 가지 수를 많게 할 필요는 없으나 하나의 가지로는 한 상태내의 음성 특징벡터의 분포를 잘 표현할 수가 없다. 그래서 화자적응시에 적응시키려는 화자의 다양한 음성정보를 잘 나타내기 위하여 상태마다 여러 개의 가지를 사용할 필요가 있다. 그러나 적응음성의 프레임 수가 상태에 따라 많이 분포되어 있는 경우에는 상태당 여러 개의 가지를 사용하는 방법이 타당하지만, 프레임 수가 적게 분포되어 있는 경우, 즉 음성스펙트럼의 변화가 적은 경우에는 적은 수의 가지로도 음성정보를 잘 표현할 수 있기 때문에 여러 개의 가지를 사용할 필요가 없다.

관측 확률밀도함수를 정밀하게 표현하기 위하여 각 상태에서 가지 수를 가변하는 방법으로서는 상태에 속한 프레임 수에 따라 가지 수를 결정하거나 각 상태내에 속한 프레임들의 특징벡터들의 분산값을 사용하여 관측 확률밀도함수의 가지 수를 정하기로 한다.

첫 번째로 상태에 속하는 프레임 수에 따라 관측 확률밀도함수의 가지 수를 결정하는 알고리듬은

$$m_j = \left\lfloor \frac{N \times \sum_k n_{jk}}{\sum_k \sum_j n_{jk}} \times M \right\rfloor, \quad 1 \le j \le N \quad (1)$$

와 같다. 여기서, $m_j$는 상태 $j$에서의 가지 수를, $n_{jk}$는 $k$번째 적응음성의 상태 $j$에서의 프레임 수를 나타낸다. 그리고 $N$, $M$은 각각 모델의 상태 수, 평균 가지 수를 나타낸다. 위 식은 일단 적응음성 데이터들이 상태별로 나누어졌을 때, 그 상태에 속한 모든 데이터들의 프레임 수가 전체 적응음성 데이터의 프레임 수를 상태 수로 나눈 값보다 많으면 가지 수를 많이 배치하고, 그 보다 적으면 가지 수를 적게 배치하는 것이다. 각 상태에 속한 프레임 수가 많을수록 음성의 다양한 특징들이 나타날 가능성이 높은 것으로 추정하고, 그것이 나타날 확률을 나타내는 관측 확률밀도함수의 수를 많게 하고, 그 반대의 경우는 확률밀도함수의 수를 적게 하였다. 이러한 방법으로 가지 수를 정할 때 어떤 상태가 너무 많은 프레임을 포함하고 있다면 가지의 수가 너무 많이 배당되므로 이것을 제한하기 위해서 최대 가지 수를 7로 놓았다. 반대로 너무 적은 개수의 프레임을 포함하고 있으면 식 (1)에서 가지의 수가 1개 미만으로 되는 것을 방지하기 위하여 상태마다 적어도 1개의 가지를 포함하도록 올림을 취하였다.

이러한 방법으로 상태에 속하는 적응음성의 프레임 수가 많은 경우에는 가지 수를 많게 하고 상태에 속하는 프레임 수가 적은 경우에는 가지 수를 적게 하면, 관측 확률밀도함수의 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 (1)에서 사용할 평균 가지 수 $M$의 적절한 값은 실험적 방법 (heuristic method)으로 찾아야 한다.

상태에 속하는 프레임 수에 따라 가지 수를 결정하

는 알고리듬을 요약하면 다음과 같다.

① 적응음성 데이터를 연속관측 HMM 음성인식기에 입력시킨다.
② 제안한 상태분할 알고리듬으로 적응음성 데이터를 각 상태별로 분할한다.
③ 상태 분할된 적응음성 데이터로부터 각 상태에 속하는 적응음성의 프레임 수를 조사한다.
④ 실험적 방법으로 관측 확률밀도함수의 최적의 평균 가지 수 M을 구한다.
⑤ 각 상태마다의 관측 확률밀도함수의 가지 수를 식 (1)에 의해 결정한다.

위에서 설명한 프레임 수에 따른 가지 수 결정방법은 복잡한 연산과정이 필요 없으므로 수행과정이 간단하다. 그러나, 상태내의 프레임 수만으로 관측 확률밀도함수의 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 음성 스펙트럼 변화가 크다고 보고 보다 많은 가지를 갖게 하는 것이 타당하다.

두 번째로 대각성분값만을 갖는 분산행렬을 가지는 정규분포에 대해서만 고려할 때 관측벡터의 각 상태에서의 분산행렬은

$$V_j(k)=\frac{1}{T}\sum_{i=1}^{T}(c_i(k)-\mu_j(k))^2 \qquad (2)$$
$$j=1,2,\cdots,N,\quad k=1,2,\cdots,P$$

와 같이 구해진다. 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 상태 수와 캡스트럼 차수를 나타낸다. 상태 $j$에서의 분산의 특징벡터 행렬식(determinant)은

$$D_j=\prod_{k=1}^{P}V_j(k) \qquad (3)$$

로 표현할 수 있다.

행렬식이 작은 상태보다 행렬식이 큰 상태의 가우스 확률분포가 넓게 퍼져있다고 볼 수 있으므로 행렬식이 큰 상태가 보다 많은 가지를 갖도록 하였다. 행렬식이 최대인 상태가 가지를 가장 많이 갖게 하고, 다른 상태에서는 최대 행렬식과의 상대적인 비율로 가지 수를 결정하였다. 분산의 특징벡터 행렬식에 따라 가지 수를 결정하는 식은

$$m_j=\left\lceil \frac{\sum_{j}^{N}\log D_j}{N\times\log D_j}\times M \right\rceil,\quad 1\le j\le N \qquad (4)$$

이다. 여기서 $D_j$와 $M$은 상태 j에서의 분산행렬의 행렬식과 관측 확률밀도함수의 평균 가지 수를 나타내며, 최소 $m_j$가 1이 되도록 즉 최소한 한 개의 가지를 갖도록 각 상태에서의 가지 수를 결정하였다.

분산의 특징벡터 행렬식에 따라 가지 수를 결정하는 알고리듬을 요약하면 다음과 같다.

① 적응음성 데이터를 연속관측 HMM 음성인식기에 입력시킨다.
② 상태분할 알고리듬으로 적응음성 데이터가 각 상태별로 분할한다.
③ 상태분할된 적응음성 데이터로부터 각 상태에 속하는 적응음성에 대한 분산의 특징벡터 행렬식을 식 (2)와 식 (3)으로 구한다.
④ 실험적 방법으로 관측 확률밀도함수의 최적의 평균 가지 수 M을 구한다.
⑤ 각 상태마다의 관측 확률밀도함수의 가지 수를 식 (4)에 의해 결정한다.

상태마다 관측 확률밀도함수의 가지 수를 달리하는 경우에는 각 가지마다의 평균과 분산을 적응시켜야 한다. 이때 기존의 화자독립모델로부터 구한 사전정보는 각 상태마다 하나의 평균과 분산을 가진다. 이 값을 이용하여 적응할 음성 데이터의 각 가지에서의 평균과 분산을 적응시킨다. 앞에서 제안한 두 가지 방법에 의하여 상태당 가지 수가 결정되면, 적응하는 각 가지의 확률분포의 가중치(weighting value)를 상태당 가지에 속하는 프레임의 수에 비례하도록 하였다. 또한, 가지마다의 MAP 파라미터 적응을 다음과 같이 하였다.

관측 확률밀도함수를 여러 개의 가지로 나눈 후 각 가지의 MAP 적응된 평균 $(\hat{\mu}_m)_{\mathrm{MAP}}$는

$$(\hat{\mu}_m)_{\mathrm{MAP}}=\frac{n_m\tau^2}{\sigma^2+n_m\tau^2}\bar{y}_m+\frac{\sigma^2}{\sigma^2+n_m\tau^2}\nu \qquad (5)$$

와 같이 나타낼 수 있다. 여기서 $n_m$은 $m$번째 가지에

속하는 적응할 표본음성 데이터의 개수이고, $\bar{y}_m$은 $m$번째 가지에 속하는 표본음성 데이터의 평균이다.

각 상태에 속하는 $m$번째 가지의 표본음성 데이터 $y_m$에 대한 분산인 $S_{ym}^2$은

$$S_{y_m}^2 = \frac{\sum_{i=1}^{n_m} (y_{im} - \bar{y}_m)^2}{n_m} \tag{6}$$

와 같이 구할 수 있다.

이 값과 기존의 화자독립모델로부터 구한 분산의 최소값 $\sigma^2_{\min}$을 사전정보로 하여 각 가지의 MAP 적응된 분산 $(\hat{\sigma}^2_m)_{MAP}$는

$$(\hat{\sigma}^2_m)_{MAP} = \begin{cases} S^2_{y_m}, & S^2_{y_m} \geq \sigma^2_{\min} \\ \sigma^2_{\min}, & S^2_{y_m} < \sigma^2_{\min} \end{cases} \tag{7}$$

와 같이 나타낸다.

각 가지의 평균과 분산을 동시에 적응하는 경우 사후확률분포를 구하면 그 역시 정규-감마분포가 된다. 그래서, $\theta$와 표본음성 데이터가 주어져 있을 때, $\mu$의 조건부 확률분포는 역시 정규분포이고 평균 $\hat{\nu}$와 분산 $\hat{\tau}^2$는

$$\hat{\nu} = \frac{\omega\nu + n_m \bar{y}_m}{\omega + n_m} \tag{8}$$

$$\hat{\tau}^2 = \frac{1}{(\omega + n_m)\theta} \tag{9}$$

와 같다.

그리고 표본음성 데이터가 주어져 있을 때 $\theta$의 여유분포 (marginal distribution)는 감마분포이고, 이것의 파라미터 $\hat{\alpha}$와 $\hat{\beta}$는

$$\hat{\alpha} = \alpha + \frac{n_m}{2} \tag{10}$$

$$\hat{\beta} = \beta + \frac{n_m}{2} S^2_{y_m} + \frac{n_m \omega (\bar{y}_m - \nu)^2}{2(\omega + n_m)} \tag{11}$$

와 같이 구할 수 있다. 위 분포로부터 구한 가지마다의 평균과 분산의 동시 MAP 추정치는

$$(\hat{\mu}_m)_{MAP} = \frac{\omega\nu + n_m \bar{y}}{\omega + n_m} \tag{12}$$

$$(\hat{\sigma}^2_m)_{MAP} = \frac{\hat{\beta}}{\hat{\alpha}} \tag{13}$$

와 같이 구할 수 있다.

2. 적응음성의 상태분할 알고리듬

ML 추정 알고리듬을 이용한 상태분할방법은 ML 추정할 때 충분한 데이터가 없으므로 초기분할을 어떻게 하느냐에 따라 인식성능에 많은 영향을 받지만 사전정보로 기존의 화자독립모델을 사용하는 방법은 그런 영향은 배제할 수 있다. 또한, 화자들 사이에는 성도 길이, 구강 크기 등의 해부학적 차이와 액센트, 발성 속도, 발성 크기 등의 화자 발성 습관에 따른 차이를 최대한 살릴 수 있다는 장점이 있다.

입력되는 적응음성 데이터를 기존의 화자독립모델을 사용하여 Viterbi 알고리듬에 의하여 상태분할을 할 수 있다. 이러한 방법으로 분할된 음성 데이터 열들을 사용하여 각 상태마다의 평균과 분산을 구하고, 이 값들을 기존의 화자독립모델로부터 구해놓은 사전정보를 이용하여 MAP 파라미터를 추정할 수 있다. 이 방법을 상태 분할 과정에서부터 기존의 화자독립모델을 이용함으로 인해 일차적으로 기존 정보값이 이용된 후에 다시 또 SI모델의 정보를 이용하는 방법이라 할 수 있다.

사전정보로 기존의 화자독립모델을 사용하는 제안한 상태분할 알고리듬을 이용하는 화자적응의 절차는 다음과 같으며, 이에 대한 블럭도는 그림 2와 같다.

① 적응할 음성데이터를 음성인식기에 입력시킨 후, 입력된 데이터를 기존의 화자독립모델을 사용하여 Viterbi 알고리듬으로 상태 분할을 수행한다.
② 상태 분할된 적응음성 데이터로부터 관측 확률밀도함수 수 변화 알고리듬으로 가지 수를 결정한다.
③ 각 상태의 각 가지마다 관측 확률밀도함수의 평균과 분산을 각각 구한다.
④ 기존의 화자독립모델로부터 평균과 분산을 각각 추출한다.
⑤ 단계 ③과 단계 ④에서 구한 값을 결합하여 MAP 파라미터를 추정한다.
⑥ 추정된 값으로부터 새로운 화자적응 모델을 만들고, 이것을 갱신된 화자독립모델로 사용한
⑦ 적응할 표본음성 데이터가 전부 입력될 때까지 단계 ①에서 단계 ⑥까지 반복 수행한다.

## Ⅲ. 상태지속분포의 적응

음성은 발성하는 화자에 따라 음성신호의 길이가 달라지므로 훈련음성에 대한 상태지속분포를 통계적으로 고려해야한다. 그러나, 화자의 실제 발음의 길이와 발성패턴이 다를 수 있으므로 상태지속분포도 차이를 가진다. 더구나, 소량의 음성데이터로 HMM모델을 훈련시키는 화자적응의 경우는 상태지속분포를 정확히 표현하는데 한계가 있다. 그러므로 적응화자의 음성적 특징을 자세하게 표현하기 위해서는 화자의 상태지속분포를 적응할 필요가 있다.

화자독립인식기에서 정규분포로 표현한 상태지속분포의 파라미터를 특정화자에 맞게 적응시켜 화자 고유의 발음속도와 발음패턴 등의 음성특성을 흡수하도록 하였다. 이 분포의 평균값은

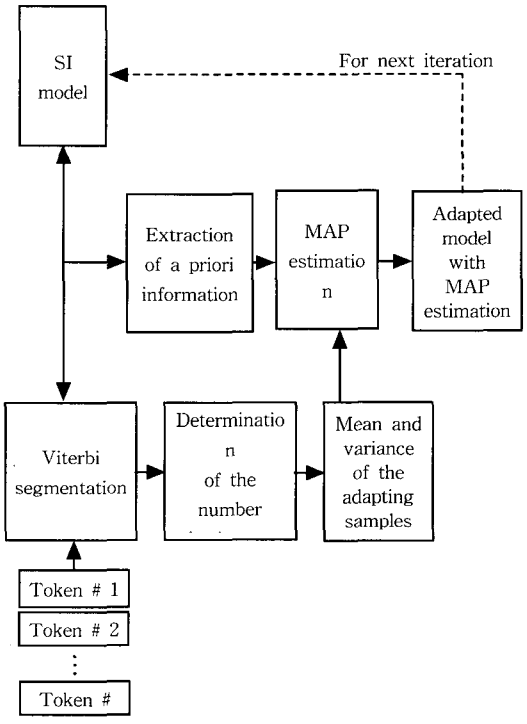


그림 2. 사전정보로 기존의 화자독립모델을 사용하는 상태분할 알고리듬의 블럭도

Fig. 2. The block diagram of state segmentation using conventional speaker independent model as a priori information.

$$\mu_d(j)=\frac{\sum_k n_{jk}}{K}, \quad 1\le j\le N \tag{14}$$

와 같이 구하고, 인식기를 사용하려는 특정화자에 적응시켰다. 여기서 $\mu_d(j)$는 상태 $j$에서의 상태지속길이의 평균이고, $K$는 훈련 음성 데이터의 수를 나타낸다. 그리고 $n_{jk}$는 상태 $j$에 속하는 프레임의 수이다. 화자적응에 사용할 적응음성 데이터 수가 충분치 않으므로 상태지속분포의 분산은 추정할 수가 없다. 그래서 상태지속길이의 분산은 화자 독립 모델에서 구한 값을 그대로 사용하였다.

적응된 평균 $\mu_d(j)$와 분산을 이용하여 정규분포를 구한 후에 이를 식 (45)와 같이 관측열 $O$의 관측확률과 곱한 값을 인식과정의 최종 확률값으로 이용하였다.

$$P=f(O|\lambda)\ [\prod_{i=1}^{N} D_i(\tau)]^{\gamma_d} \tag{15}$$

## Ⅳ. 실험 결과 및 고찰

본 논문에서 제안한 연속관측 HMM음성인식기에서의 관측밀도함수의 수 변화에 의한 화자적응과 화자적응에 필요한 방법들에 대한 타당성을 검증하고, 또한 기존방법들과의 인식성능을 비교하기 위하여 인식실험을 수행하였다.

### 1. 음성 데이터 베이스

본 논문에서는 제안한 방법들의 성능을 평가하기 위하여 두 가지의 음성데이터를 사용하였다. 첫 번째 음성데이터는 10명(남자 5명, 여자 5명)의 화자가 15개 한국 도시명(서울, 부산, 대구, 인천, 광주, 대전, 수원, 춘천, 청주, 공주, 안동, 울산, 전주, 목포, 제주)을 10번씩 발음한 1500개의 음성데이터이다. 두 번째 음성데이터는 ETRI의 샘돌 음성데이터로 40명(남자 20명, 여자 20명)의 화자가 40개의 단어를 4번씩 발음한 1600개의 음성데이터이다. 이것은 한국어 고립숫자와 고립단어들로 구성되어 있다.

다음은 도시명 음성데이터에 대한 실험에 사용된 상수 값들이다.

1) 저역통과여파기의 차단주파수 : 3.4 kHz
2) 표본화 주파수 : 8 kHz
3) 양자화 비트수 : 12 bits/sample
4) 프레임 길이 : 300 samples
5) 프레임 이동거리 : 100 samples
6) preemphasis 전달함수 : $1-0.95\ z^{-1}$

7) LPC 캡스트럼 차수 : 12차

그리고, 다음은 ETRI의 샘돌 음성 데이터에 대한 실험에 사용된 상수 값이다.

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

2. CDHMM에서의 화자적응

본 논문에서 제안한 화자적응방법들을 CDHMM 음성인식기에 적용하고, 이들의 성능을 조사하기 위하여 한국 도시명 데이터베이스와 ETRI 음성 데이터베이스를 대상으로 하여 각각 인식실험을 수행하였다.

먼저, 도시명 음성데이터를 대상으로 하여 제안한 여러 가지 인식실험들을 수행하였다. 적응화 실험에 사전정보로 사용할 최적의 화자독립모델을 정하기 위하여 그 모델의 상태 수와 가지 수를 달리하며 인식실험을 수행하였다. 이 경우의 인식실험은 테스트하고자 하는 화자 한 명을 제외한 나머지 화자들로부터 훈련시킨 후 인식실험을 하는 라운드 로빈 방식(round robin method)을 사용하였고, 화자독립 모델의 구조는 단순 좌우구조 HMM을 사용하였다. 표 1에 사전정보로 사용되는 화자독립모델의 상태 수와 가지 수에 따른 인식기의 인식률을 나타내었다.

표 1에서 인식기의 인식률이 가장 우수한 경우인 6개의 상태와 3개의 가지를 갖는 최적화된 화자독립 HMM 인식기를 적응화의 사전정보로서 사용하였다. 이러한 조건을 갖춘 화자종속인식기의 평균인식률은, 98.4% 이고, 화자독립인식기의 경우 83.2%였다.

화자적응실험을 아래와 같은 5가지의 실험방법으로 나누어 각각 수행하였다.

EXP1 : MLE 방법
EXP2 : 적응음성 데이터의 평균과 화자독립모델로부터 구한 분산 $\sigma^2$ 사용
EXP3 : 화자적응된 평균(식 5)과 분산 $\sigma^2$ 사용
EXP4 : 가지마다의 적응음성 데이터 평균과 화자적응된 분산(식 5) 사용

표 1. CDHMM 인식기의 상태 수와 가지수에 따른 인식률 (%)

Table 1. Recognition rate (%) of CDHMM recognizer according to the number of states and mixturees.

| State(N) | mixture(M) | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 4 | 72.1 | 77.5 | 80.2 | 81.2 | 80.8 | 78.6 |
| 5 | 75.0 | 78.8 | 82.7 | 83.2 | 82.4 | 79.2 |
| 6 | 77.9 | 82.7 | 83.2 | 83.0 | 82.1 | 78.7 |
| 7 | 79.9 | 82.2 | 82.1 | 80.0 | 78.5 | 77.5 |

EXP5 : 화자적응된 평균(식 12)과 분산(식 13) 사용

상태천이확률의 변화는 인식기의 인식률에 별다른 영향을 미치지 못함으로 기존 화자독립모델의 확률값을 그대로 사용하였다.

각 상태마다의 적응화 파라미터들을 구하기 위하여 적응할 음성의 상태를 분할하는 방법을 아래와 같이 나누어 각각 실험하였다.

SEG1 : 적응할 음성의 상태분할시에 ML추정 알고리듬을 이용하는 방법(기존의 방법)
SEG2 : 적응할 음성의 상태분할시에 사전정보로 화자독립모델을 사용하는 방법(제안한 방법)

기존의 화자적응방법 즉 상태마다 하나의 가지를 사용하는 경우에 대하여 적응음성을 상태분할하는 방법에 따른 인식기의 인식률을 표 2 및 표 3에 각각 나타내었다.

표 2 및 표 3에서 MLE 훈련 방법을 사용한 EXP1의 결과를 분석하면 훈련 토큰 수가 적을 경우에는 인식률이 많이 떨어짐을 알 수 있다. 이것은 MLE 방법으로 훈련할 경우 적은 양의 훈련데이터를 가지고는 분산을 제대로 추정할 수 없기 때문으로 여겨진다. EXP1의 결과에서 적응음성 데이터 수가 증가함에 따라 다른 방법들에 비해 인식률이 많이 향상되는데, 특히 토큰 수가 3개일 때는 EXP1의 결과가 다른 방법보다 가장 우수함을 알 수 있다.

또한, SEG2로 상태분할한 경우의 인식률이 SEG1보다 전반적으로 0.2%~1.2% 정도 높게 나타남을 알 수 있었다.

표 2. SEG1을 이용한 화자적응방법의 평균 인식률 (%)
Table 2. Average recognition rate (%) of speaker adaptation methods using SEG1.

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 84.1 | 92.1 | 91.2 | 92.2 | 92.1 |
| 2 | 96.0 | 95.8 | 93.3 | 94.3 | 95.6 |
| 3 | 98.6 | 96.7 | 96.1 | 97.8 | 97.3 |

표 3. SEG2를 이용한 화자적응방법의 평균 인식률 (%)
Table 3. Average recognition rate (%) of speaker adaptation methods using SEG2.

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 84.1 | 93.3 | 92.1 | 93.4 | 92.3 |
| 2 | 96.0 | 96.1 | 95.5 | 96.3 | 95.8 |
| 3 | 98.6 | 97.0 | 96.6 | 98.5 | 97.6 |

이러한 결과는 ML알고리듬을 사용하여 상태분할하는 SEG1방법은 훈련데이터가 적을 경우에는 ML추정이 제대로 이루어지지 않으므로 상태가 제대로 분할되지 않는 반면, SEG2 방법은 상태분할시부터 충분한 훈련데이터들로부터 추정된 화자독립인식기의 파라미터들을 사전정보로 사용함으로써 적응음성의 상태를 보다 정확히 분할할 수 있기 때문이다.

제안한 관측밀도함수 수 변화에 의한 화자적응기의 인식률을 실험방법에 따라 각각 구분하여 표 4에서 18까지 나타내었다. 여기서 관측 확률밀도함수의 가지 수를 결정하는 방법에 의한 인식실험의 결과를 프레임 수에 따라 가지 수를 결정하는 방법, 즉 FRA(frame) 방법과 분산의 특징벡터 행렬식에 따라 가지 수를 결정하는 방법, 즉 VRA(variance) 방법으로 구분하여 각각 나타내었다.

표 4에서 표 7까지 나타난 결과를 분석하면 EXP1의 경우에는 인식률이 표 2 및 표 3의 결과보다 오히려 떨어지는 것을 볼 수 있는데, 이는 훈련데이터의 양이 충분치 않은 상태에서 가지 수를 늘리는 것은 분산을 구하는데 역효과를 내기 때문이다. EXP1을 제외한 다른 방법으로 실험했을 경우에는 가지를 여러 개 사용한 경우의 인식결과가 가지를 한 개 사용한 경우보다 인식률이 높음을 볼 수 있다. 또한 전반적으로 SEG2로 상태분할하고 프레임 수에 따라 가지 수를 결정한 경우의 인식률이 가장 높음을 볼 수 있다. 특히 평균과 분산을 동시에 적응시키는 EXP5의 토큰을 한 개 사용하였을 때의 결과가 98.9 %로 매우 높음을 볼 수 있다.

표 4. 화자적응방법에 따른 평균 인식률 (%) (SEG1+FRA)
Table 4. Average recognition rate (%) of speaker adaptation methods(SEG1+FRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 82.2 | 95.6 | 93.6 | 93.1 | 93.9 |
| 2 | 95.8 | 98.4 | 98.0 | 98.0 | 97.9 |
| 3 | 98.3 | 99.2 | 99.0 | 98.8 | 99.0 |

표 5. 화자적응방법에 따른 평균 인식률 (%) (SEG1+VRA)
Table 5. Average recognition rate (%) of speaker adaptation methods(SEG1+VRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 80.7 | 95.8 | 93.8 | 93.6 | 94.3 |
| 2 | 94.3 | 98.3 | 97.6 | 97.3 | 97.3 |
| 3 | 98.0 | 99.6 | 99.0 | 99.0 | 99.1 |

표 6. 화자적응방법에 따른 평균 인식률 (%) (SEG1+FRA)
Table 6. Average recognition rate (%) of speaker adaptation methods(SEG2+FRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 82.2 | 96.7 | 98.8 | 95.2 | 98.9 |
| 2 | 95.8 | 98.8 | 99.3 | 98.3 | 99.3 |
| 3 | 98.3 | 98.8 | 99.3 | 99.2 | 99.3 |

표 7. 화자적응방법에 따른 평균 인식률 (%) (SEG2+VRA)

Table 7. Average recognition rate (%) of speaker adaptation methods(SEG2+VRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 80.7 | 95.1 | 93.2 | 92.7 | 93.3 |
| 2 | 94.3 | 97.9 | 97.2 | 97.0 | 96.2 |
| 3 | 98.0 | 98.9 | 98.3 | 98.8 | 98.2 |

식 (14)를 사용하여 화자의 상태지속분포의 평균을 적응시킨 결과를 다음의 표들에 나타내었다.

표 8에서 표 11까지에서 적응음성이 한 개일 때 상태지속길이의 평균값을 적응시키면 인식률이 0.2%에서 최고 1% 정도 향상된다. 또 토큰 수가 많아지면 상태

표 8. 상태지속분포의 적응을 포함한 화자적응방법의 평균 인식률 (%) (SEG1+FRA)

Table 8. Average recognition rate (%) of speaker adaptation methods including the adaptation of state duration distribution(SEG1+FRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 82.2 | 95.8 | 93.9 | 93.5 | 94.1 |
| 2 | 95.8 | 98.4 | 98.2 | 98.0 | 98.2 |
| 3 | 98.3 | 99.2 | 99.0 | 98.8 | 99.0 |

표 9. 상태지속분포의 적응을 포함한 화자적응방법의 평균 인식률 (%) (SEG1+VRA)

Table 9. Average recognition rate (%) of speaker adaptation methods including the adaptation of state duration distribution(SEG1+VRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 80.7 | 95.8 | 94.0 | 93.6 | 94.5 |
| 2 | 94.3 | 98.3 | 97.7 | 97.3 | 97.6 |
| 3 | 98.0 | 99.6 | 99.0 | 99.0 | 99.1 |

표 10. 상태지속분포의 적응을 포함한 화자적응방법의 평균 인식률 (%) (SEG2+FRA)

Table 10. Average recognition rate (%) of speaker adaptation methods including the adaptation of state duration distribution(SEG2+FRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 82.2 | 97.0 | 98.9 | 96.2 | 99.1 |
| 2 | 95.8 | 98.9 | 99.3 | 98.8 | 99.3 |
| 3 | 98.3 | 98.8 | 99.3 | 99.2 | 99.3 |

표 11. 상태지속분포의 적응을 포함한 화자적응방법의 평균 인식률 (%) (SEG2+VRA)

Table 11. Average recognition rate (%) of speaker adaptation methods including the adaptation of state duration distribution(SEG2+VRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 80.7 | 95.7 | 93.8 | 93.5 | 94.3 |
| 2 | 94.3 | 98.0 | 97.2 | 97.3 | 96.5 |
| 3 | 98.0 | 98.9 | 98.3 | 98.8 | 98.2 |

지속길이의 효과가 별로 반영이 되지 않는데 이는 토큰 수의 증가로 인한 인식률의 향상에 이 효과가 포함되기 때문인 것으로 생각된다. 토큰이 한 개 일 때 가장 높은 인식률을 나타내는 EXP5의 결과가 98.9%에서 0.2% 더 올라 99.1%가 된 것을 볼 수 있다.

평균가지 수 $M$에 따른 인식률의 변화를 표 12에 나타내었다. 실험방법들에 따른 인식률은 조금씩 차이가 나지만 전반적인 양상은 비슷하므로 높은 인식률을 나타내는 SEG2 방법으로 상태분할하고 프레임 수에 따라 가지 수를 결정하는 경우의 EXP5방법에 대한 결과만 나타내었다.

표 12에서 토큰 수에 관계없이 $M$이 2.0일 때 가장 높은 인식률을 얻을 수 있었다. 적응음성이 한 개일 때 $M$값에 따른 영향을 많이 받는 것을 볼 수 있다.

제안한 방법의 범용성을 검증하기 위하여 ETRI의 샘돌이 데이터에 대해서 같은 방법으로 실험하였다.

표 12. 평균가지 수에 따른 화자적응기의 평균인식률 (%)

Table 12. Average recognition rate (%) of speaker adaptation system by the number of average mixturees.

| Number of tokens | Number of average mixtures ( $M$ ) | | | | |
|---|---|---|---|---|---|
| | 1 | 1.5 | 2.0 | 2.5 | 3.0 |
| 1 | 93.9 | 93.9 | 99.1 | 93.9 | 93.3 |
| 2 | 96.0 | 96.4 | 99.3 | 96.9 | 96.4 |
| 3 | 97.6 | 98.0 | 99.3 | 98.2 | 98.1 |

ETRI 샘돌이 데이터는 40개의 한국어 고립숫자 및 단어로 구성되어 있다.

샘돌이 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. 다음 표들의 적응 결과들도 첫 번째 그룹에서 만든 모델을 두 번째 그룹에 속하는 화자들에 적응시킨 결과이다.

표 13. SEG1을 이용한 화자적응방법의 평균인식률 (%)

Table 13. Average recognition rate (%) of speaker adaptation methods using SEG1.

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 67.1 | 88.4 | 88.1 | 87.3 | 88.9 |
| 2 | 86.7 | 92.4 | 92.5 | 93.0 | 93.7 |

표 14. SEG2를 이용한 화자적응방법의 평균인식률 (%)

Table 14. Average recognition rate (%) of speaker adaptation methods using SEG2.

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 67.1 | 89.5 | 89.5 | 89.1 | 89.9 |
| 2 | 86.7 | 92.6 | 92.8 | 93.3 | 93.5 |

표 15. 상태지속분포의 적응을 포함한 화자적응방법의 평균 인식률 (%) (SEG1+FRA)

Table 15. Average recognition rate (%) of speaker adaptation methods including the adaptation of state duration distribution(SEG1+FRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 65.3 | 91.0 | 90.7 | 89.9 | 90.5 |
| 2 | 85.9 | 94.3 | 93.3 | 93.8 | 93.8 |

표 16. 상태지속분포의 적응을 포함한 화자적응방법의 평균 인식률 (%) (SEG1+VRA)

Table 16. Average recognition rate (%) of speaker adaptation methods including the adaptation of state duration distribution(SEG1+VRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 66.8 | 90.7 | 90.3 | 89.9 | 90.3 |
| 2 | 87.9 | 94.3 | 93.2 | 92.8 | 93.0 |

표 17. 상태지속분포의 적응을 포함한 화자적응방 법의 평균 인식률 (%) (SEG2+FRA)

Table 17. Average recognition rate (%) of speaker adaptation methods including the adaptation of state duration distribution(SEG2+FRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 65.3 | 92.3 | 92.7 | 90.3 | 93.3 |
| 2 | 85.9 | 94.2 | 95.5 | 94.3 | 96.2 |

상태마다 관측밀도함수의 하나의 가지를 사용하는 경우의 인식률을 상태분할방법에 따라 다음의 표 14 및 표 15에서 각각 나타내었다.

표 13 및 표 14에서 MLE 훈련방법을 사용한 EXP1의 결과를 보면 도시명 데이터일 경우와 마찬가지로 다른 방법에 비해 인식률이 많이 떨어짐을 알 수 있다. 또한 전반적으로 SEG2 방법으로 상태분할한 표 15의

표 18. 상태지속분포의 적응을 포함한 화자적응방법의 평균 인식률 (%) (SEG2+VRA)

Table 18. Average recognition rate (%) of speaker adaptation methods including the adaptation of state duration distribution(SEG2+VRA).

| Number of tokens | EXP1 | EXP2 | EXP3 | EXP4 | EXP5 |
|---|---|---|---|---|---|
| 1 | 66.8 | 91.7 | 91.3 | 91.0 | 90.3 |
| 2 | 87.9 | 94.3 | 93.9 | 93.5 | 93.3 |

결과가 더 좋은 것을 볼 수 있다. 그리고 표 13 및 표 14에서 한국어 도시명 데이터를 대상으로 한 인식결과와 양상이 전반적으로 비슷함을 볼 수 있다.

표 15에서 표 18까지의 결과에서도 도시명 데이터의 경우와 마찬가지로 SEG2 방법으로 상태분할하고 프레임 수에 따라 가지 수를 결정했을 경우인 표 17의 결과가 가장 좋았고, 특히 평균과 분산을 동시에 적응하는 EXP5의 인식률이 가장 우수하였다. 또한, 적응음성의 상태당 길이 정보를 적응시키므로 인해 전반적으로 인식률이 1.5%정도 향상됨을 확인할 수 있었고, 특히 적응음성 데이터 수가 한 개일 때 인식률이 가장 높음을 알 수 있었다.

도시명데이터와 샘돌데이터에 대한 인식실험 결과 제안한 방법의 인식성능이 기존의 방법에 비해 우수하였다.

## V. 결 론

본 논문에서는 연속관측 HMM 음성인식기의 인식성능을 향상시키기 위해 상태당 관측밀도함수 수 변화에 의한 화자적응 알고리듬을 제안하고, 인식실험을 통하여 이의 성능을 확인하였다.

제안한 방법은 음성적 특징을 자세하게 나타내기 위하여 상태당 관측 확률밀도함수를 여러 개의 가지로 세분화하여 화자적응하였다. 각 상태의 가지를 정하는 방법으로 프레임 수에 따라 결정하는 방법과 특징벡터 행렬식에 따라 결정하는 방법을 각각 사용하였다. 이 두 가지 방법에 의해 결정된 가지 수만큼 가지마다 평균의 Bayes적응, 분산의 Bayes적응 및 평균과 분산의 동시 Bayes적응을 수행하여 MAP 파라미터를 추출함으로써 정밀한 화자적응모델의 파라미터를 구할 수 있었다. 제안한 방법의 인식률은 하나의 가지를 사용하는 기존의 화자적응방법보다 높은 수치를 보였으며, 특히 상태내의 프레임 수로 가지 수를 결정하는 방법이 훈련시 수행과정이 단순하면서 인식성능도 우수하였다. 이것은 각 상태의 음성정보를 좀 더 세밀히 나타낼 수 있기 때문이다. 그리고, 정규분포의 파라미터를 적용하는 실험방법에 따라 평균의 Bayes적응방법이 분산의 Bayes적응방법보다 인식성능이 우수하였으며, 특히 평균과 분산을 동시에 Bayes적응하는 경우가 인식성능이 가장 우수하였다. 이러한 화자적응방법은 적응음성 데이터를 한 개 사용할 때 기존의 ML 추정법을 사용하는 경우보다 인식률이 20%~30% 정도 향상되었다. 한편, 포만트 특성이 단순한 여성 화자들이 남성화자들보다 적응이 잘 안 되는 것으로 나타났다.

아울러, 화자적응에 필요한 상태분할 방법은 충분한 음성 데이터로 훈련된 화자독립모델을 사전정보로 이용함으로써 좀 더 정확한 상태 분할을 할 수 있었다. 특히, 적응음성 데이터가 한 개 사용되고 정규분포의 파라미터 중 평균의 적응시 상태분할이 가장 잘 되었다. 이것은 음성의 상태분할시 기존 화자독립모델을 사전정보로 이용함으로써 적응음성 데이터수가 적은 경우에도 정확한 상태분할을 할 수 있기 때문이다. 그리고 음성의 상태당 길이 정보를 화자에 적응시키는 방법으로 상태지속분포를 특정화자에 적응시키기 위해 상태지속시간의 평균은 상태에 속하는 적응음성의 프레임 수를 평균하여 구하였고, 분산은 화자독립모델에서 구한 값을 그대로 사용하였다. 이러한 방법을 음성인식기에 적용할 때 인식률이 전반적으로 향상되었고, 특히 적응음성 데이터가 한 개일 때 인식률이 가장 높았다. 이것은 제안한 방법이 화자고유의 발음속도와 발음패턴 등의 특성을 잘 흡수할 수 있기 때문으로 여겨진다.

위의 제안한 방법을 종합하여 화자독립 연속관측 HMM을 화자적응 시키는 경우, 한국 도시명 데이터에 대하여 인식률이 기존 방법의 92.3%에서 99.1%로 향상되었다. 또한 ETRI 데이터에 적응시켰을 때에도 한국 도시명 데이터에서와 비슷한 결과를 얻었다.

## 참 고 문 헌

[1] B. S. Atal and L. R. Rabiner, "Speech research directions," AT&T Tech. J., vol. 65, pp. 75~88, Sep.-Oct., 1986.

[2] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, pp. 43~49, Feb. 1978.

[3] "A neural network approach to speech recognition," Tech. Report, ETRI. Advanced Research Dept., Mar. 1990.

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

[5] K. F. Lee and R. Rebby, Automatic speech recognition, Kluwer Academic, 1989.

[6] 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.

[7] 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.

[8] Y. Shiraki and M. Honda, "Speaker adaptation algorithms for segment vocoder," IEICE, vol. SP87-67, pp. 49~56, Oct. 1987.

[9] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," Proc. ICASSP89, pp. 286~289, May 1989.

[10] 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.

[11] K. Shikano, K. F. Lee, and R. Reddy, "Speaker adaptation through vector quantization," Proc. ICASSP86, pp. 2643~2646, Apr. 1986.

[12] 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.

[13] P. F. Brown, C. H. Lee, and J. C. Spohrer, "Bayesian adaptation in speech recognition," Proc. ICASSP83, pp. 761~764, Apr. 1983.

[14] J. Chien, H. Wang, and C. Lee, "Improved Bayesian learning of hidden Markov models for speaker adaptation," Proc. ICASSP97, pp. 1027~1030, Apr. 1997.

[15] K. Ohkura, M. Sugiyama, and S. Sagayama, "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs," Proc. ICSLP, pp. 369~372, 1992.

[16] J. Takahashi and S. Sagayama, "Telephone line characteristic adaptation using vector field smoothing technique," Proc. ICSLP, pp. 991~994, 1994.

[17] S. Cox, "Predictive speaker adaptation in speech recognition", Computer Speech and Language, vol. 9, pp. 1~17, 1995.

[18] V. Nagesha and L. Gillick, "Studies in transformation-based adaptation," Proc. ICASSP97, pp. 1031~1034, Apr. 1997.

[19] 김광태, 서정일, 홍재근, "CDHMM의 상태당 가지수를 가변시키는 화자적응에 관한 연구," 대한전자공학회 논문집, 제35C권 제3호, pp. 166~175, 1998

## 저 자 소 개

金 光 泰(正會員)

1985년 2월 : 경북대학교 공과대학 전자공학과 졸업(공학사). 1987년 2월 : 경북대학교 대학원 전자공학과 졸업(공학석사). 1988년 8월 : 경북대학교 공과대학 전자공학과 졸업(공학박사). 1989년~1993년 : 국방과학연구소 연구원. 1994년~현재 : 상주대학교 전자전기공학부 부교수. <주관심분야 : 음성 신호처리, PDP 및 LCD 회로 설계>