관찰 확률 최대화에 의한 화자 적응 알고리즘

Speaker Adaptation Algorithm Based on a Maximization of the Observation Probability

양 태 영*, 신 원 호*, 전 원 석*, 김 지 성*, 김 원 구**, 이 충 용*, 윤 대 희*, 차 일 환*
(Tae Young Yang*, Won Ho Shin*, Won Suk Jun*, Ji Sung Kim*,
Weon Goo Kim**, Chung Yong Lee*, Dae Hee Youn*, and Il Whan Cha*)

※본 논문은 1997년 한국과학재단 과제번호 95-0100-22-01-3의 지원으로 연구되었습니다.

요 약

본 논문에서는 SCHMM에 적용된 관찰 확률 최대화에 의한 화자 적응 알고리즘을 제안한다. 제안된 알고리즘은 SCHMM의 관찰 확률 밀도들이 새로운 화자의 음성 특징을 잘 표현하지 못하는 경우 인식 성능이 저하되는 것을 막기 위하여, 적응 데이터의 각 특징 벡터들이 최대의 관찰 확률을 가질 수 있도록 관찰 확률 밀도를 결정하는 평균 벡터 μ와 분산 행렬 Σ를 기울기 탐색(gradient search) 알고리즘에 의해 반복적으로 적응시켜 주는 방법이다. SCHMM의 상태 천이 확률 A와 혼합 밀도 계수 C는 관찰 확률 밀도 적응 과정을 거친 후, 적응 데이터로부터 구한 확률과 기존 확률의 가중 평균을 취하는 과정을 반복하여 적응시켜 주었다. 제안된 화자 적응 알고리즘을 사용하여 단독음 인식 실험을 수행한 결과, 화자 적응을 수행하지 않았을 때와 비교하여 화자 독립 시스템에서는 평균 9.8%, 남성 화자 종속 시스템에서는 평균 46.0%, 여성 화자 종속 시스템에서는 평균 52.7%의 인식률 향상을 보였다.

ABSTRACT

This paper proposes a new speaker adaptation technique which maximizes the mixture probability of an input speech. It is applied to semi-continuous hidden Markov model (SCHMM) speech recognizers. When there is a large acoustic mismatch between the input speech and the mixtures in SCHMM, the recognition performance degrades. To avoid such a problem, the proposed algoridtm changes the mean vector μ and the variance vector Σ iteratively by the gradient search algorithm so that the features of the adaptation speech data could achieve maximum observation probabilities. After the adaptation process of the means and the variances, the state transition probability matrix A and the mixture weight matrix C are adapted by an iterative process which interpolates the original parameters and the new parameters calculated from the adaptation speech data. The proposed adaptation algorithm was evaluated by a speaker-independent, a male speaker-dependent, and a female speaker-dependent recognizers. The experiment results on the isolated word recognition showed that the proposed adaptation algorithm achieved 9.8% average enhancement in the speaker-independent recognizer, 46.0% in the male speaker-dependent recognizer, and 52.7% in the female speaker-dependent recognizer.

I. 서 론

음성 인식 시스템은 인식할 수 있는 화자의 범위에 따라 화자 종속 시스템(speaker-dependent system)과 화자 독립 시스템(speaker-independent system)으로 나눌 수 있다[1]. 화자 종속 시스템은 한 화자의 음성으로 학습되어 학습된 화자에 대해서는 높은 인식률을 보이지만 다른 화자의 경우는 인식률이 매우 낮다. 화자 독립 시스템은 여러 화자의 음성을 인식할 수 있는 인식 시스템으로, 인식 시스템의 학습을 위해서는 다수의 화자로부터 얻은 많은 양의 음성 데이터를 필요로 하며, 각 개인마다 서로 다른 음성 특징을 갖고 있기 때문에 인식 시스템을 이용하고자 하는 화자의 음성 특징이 인식 시스템의 학습된

* 연세대학교 전자공학과
** 군산대학교 전기공학과
접수일자 : 1998년 5월 20일

음성 특징과 다를 경우 인식 성능은 저하된다. 이러한 단점을 보완하기 위하여 화자 적응 시스템(speaker-adaptive system)이 사용된다. 화자 적응 시스템은 새로운 화자로부터 약간의 음성 신호를 제공받아 이를 이용하여 새로운 화자의 음성 특징에 알맞도록 인식 시스템의 모델을 수정함으로써, 음성 특징의 차이에 의해 발생하는 인식 성능 저하를 방지해 주는 인식 시스템으로, 화자 종속 시스템과 화자 독립 시스템 모두에 화자 적응 기법을 적용하는 것이 가능하다.

HMM(Hidden Markov Model)의 화자 적응 방법은 DHMM(Discrete HMM) 코드북의 평균 벡터(mean vector) μ나 SCHMM(Semi-Continuous HMM), CHMM(Continuous HMM)의 평균 벡터 μ와 분산 행렬(covariance matrix) Σ 같이 입력 특징 벡터를 정확히 모델링하기 위한 변수들을 적응시켜주는 과정과 DHMM의 관찰 확률(observation probability) B와 SCHMM, CHMM의 혼합 밀도 계수(mixture coefficient) C와 같이 모델링된 특징 벡터들을 확률적으로 나타내기 위한 변수들을 적응시켜주는 과정으로 구분될 수 있다.

입력 특징 벡터를 모델링하기 위한 HMM 변수들을 적응시켜주는 과정으로는, 학습된 HMM의 평균 벡터와 새로운 화자의 특징 벡터간의 대응 관계를 찾아 히스토그램을 구하고 이것을 가중치로 하여 입력 특징 벡터들과의 가중 평균으로 새로운 평균 벡터를 구하는 방법[2]이 있고, 히스토그램 대신 멤버쉽(membership)값을 이용하는 퍼지 벡터 양자화(fuzzy vector quantization)[3]가 있으며, 현재의 평균 벡터와 새로운 화자의 특징 벡터와의 대응 관계를 행렬로 나타내어 새로운 평균 벡터를 기존의 평균 벡터들의 선형합으로 구하는 방법[4][5]과 신경회로망(neural network)을 통한 비선형 대응을 통해 새로운 평균 벡터를 구하는 방법[6] 등이 있다. DHMM의 관찰 확률과 SCHMM, CHMM의 혼합 밀도 계수를 적응시켜주는 과정으로는, 학습되어 있는 모델과 새로운 화자의 음성 데이터를 비터비 상태 역추적(Viterbi state back tracking)을 통해 최적의 대응 관계를 찾은 후 각 상태(state)에서 관찰되는 평균 벡터들의 횟수를 누적시킨 히스토그램을 구하고 이를 이용하여 새로운 관찰 확률, 또는 혼합 밀도 계수를 구하는 방법[2]이 있으며, 대응 관계를 변환 행렬(transition matrix)로 나타내어 학습되어 있는 모델의 관찰 확률, 또는 혼합 밀도 계수를 변환시키 줌으로써 새로운 화자에 적응시켜 주는 방법[3]이 있다. 이 밖에, 사후 확률(a posteriori probability) 최대화에 의하여 SCHMM과 CHMM의 파라메터들을 적응시키는 베이시안 화자 적응(Bayesian speaker adaptation)[8][9][10]이 있으며, 화자군 모델(speaker Markov model)을 설정해서 새로운 화자와 비슷한 특성을 갖는 화자군의 HMM을 이용하여 인식하는 방식[11][12]등이 있다.

본 논문에서는 SCHMM을 기반으로 하는 음성 인식 시스템의 화자 적응을 위하여 관찰 확률 최대화에 의한 화자 적응 알고리즘을 제안한다. 제안된 알고리즘은 SCHMM의 평균 벡터와 분산 행렬을 적응시키는 과정과, 상태 천이 확률(state transition probability)과 혼합 밀도 계수를 적응시키는 과정의 두 단계로 나뉜다. SCHMM의 평균 벡터와 분산 행렬을 적응시키는 과정에서는 평균 벡터와 분산 행렬에 의해 결정되는 관찰 확률 밀도인 가우시안 확률 밀도를 이용하여 새로운 화자의 적응 데이터의 관찰 확률을 구할 때, 최대의 관찰 확률이 얻어지도록 평균 벡터와 분산 행렬을 기울기 탐색(gradient search) 알고리즘에 의해 반복적으로 바꿔주게 된다. 상태 천이 확률과 혼합 밀도 계수를 적응시키는 과정에서는 적응된 평균 벡터, 분산 행렬, 기존의 상태 천이 확률, 혼합 밀도 계수와 적응 데이터간의 최적의 대응 관계를 비터비 상태 역추적에 의해 찾은 후 새로운 상태 천이 확률과 혼합 밀도 계수를 계산하고, 적응 데이터의 양이 적기 때문에 발생될 수 있는 문제들을 막기 위하여 새로운 상태 천이 확률, 혼합 밀도 계수와 기존의 상태 천이 확률, 혼합 밀도 계수의 가중 평균을 취하여 적응된 상태 천이 확률, 혼합 밀도 계수를 구한다.

본 논문의 구성은 II장에서 제안된 관찰 확률 최대화에 의한 화자 적응 알고리즘을 설명하고, III장에서는 인식 실험을 위한 설정을 다룬다. 이에 대한 실험 결과는 IV장에서 논하며, V장에 결론을 기술하였다.

II. 관찰 확률 최대화

2.1 평균 벡터와 분산 행렬의 적응

SCHMM에서는 모든 모델(model)과 상태에서 공유되는 L개의 가우시안 확률 밀도들과, 각 상태에서 가우시안 확률 밀도들의 가중치를 결정하는 혼합 밀도 계수(mixture coefficient)에 의해 입력 음성의 특징을 확률적으로 모델링하는 혼합 확률(mixture probability)를 구한다. k번째 평균 벡터 μ_k와 분산 행렬 Σ_k에 의해 결정되는 가우시안 확률 밀도 $N(\mu_k, \Sigma_k)$, 그리고 이에 대한 i번째 상태에서의 혼합 밀도 계수 c_{ik}를 $\theta_i = \{c_{ik}, \mu_k, \Sigma_k\}$, $1 \le k \le L$로 나타낼 때, 시간 t에서의 입력 특징 벡터 x_t에 대한 상태 i에서의 혼합 확률 $P_i(x_t|\theta_i)$는 다음과 같다.

$$P_i(x_t|\theta_i) = \sum_{k=1}^{K} c_{ik} N(x_t|\mu_k, \Sigma_k). \quad (1)$$

일반적으로 혼합 확률을 구할 때, 전체 L개의 혼합 밀도 계수가 곱해진 가우시안 확률 밀도를 모두 사용하지 않고, 이 중 가장 큰 확률을 갖는 K개만을 사용한다. (1)을 보면, 입력 특징 벡터 x_t는 일차적으로 평균 벡터 μ와 분산 행렬 Σ에 의해 확률적으로 모델링된다. 만일, 새로운 화자의 입력 특징 벡터 x_t와 평균 벡터 μ, 분산 행렬 Σ의 차이가 크다면 정확한 모델링이 될 수 없으며, 옳지 못한 가우시안 확률 밀도가 혼합 확률을 구하는데 사용될 수 있다. 따라서, 입력 특징 벡터 x_t를 잘 표현할

수 있도록 평균 벡터 μ와 분산 행렬 Σ를 수정해 주어야 하며, 이를 위해 혼합 밀도 계수가 곱해지지 않은 가우시안 확률 밀도들만의 합인 관찰 확률 $P\{x_t|N(\mu, \Sigma)\}$를 다음과 같이 정의한다.

$$P\{x_t|N(\mu, \Sigma)\} = \sum_{k=1}^{K} N(x_t|\mu_k, \Sigma_k) \quad (2\text{-a})$$

$$= \sum_{k=1}^{K} \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x_t-\mu_k)^T \Sigma_k^{-1}(x_t-\mu_k)\right\}. \quad (2\text{-b})$$

여기서, k번째 평균 벡터 μ_k는 D차원의 벡터이고, k번째 분산 행렬 Σ_k는 $D\times D$차원인 대각 행렬(diagonal matrix)이며, 전체의 평균 벡터와 분산 행렬을 나타내는 μ와 Σ는 $\mu=\{\mu_1, \mu_2, \cdots, \mu_L\}$, $\Sigma=\{\Sigma_1, \Sigma_2, \cdots, \Sigma_L\}$이다. (2-a)와 (2-b)에서 정의된 관찰 확률 $P\{x_t|N(\mu, \Sigma)\}$는 가우시안 확률 밀도들이 새로운 화자의 음성 특징을 잘 나타내줄 수 있는 경우 큰 확률값을 갖지만, 음성 특징의 차이가 커서 정확한 모델링이 이루어지지 못하는 경우에는 작은 값을 갖게 된다. 그러므로, (2-a)와 (2-b)에서 사용된 K개의 평균 벡터 μ_k와 분산 행렬 Σ_k를 관찰 확률 $P\{x_t|N(\mu, \Sigma)\}$이 최대화될 수 있도록 입력 특징 벡터 x_t의 방향으로 이동시켜 줌으로써 화자 적응을 수행한다.

평균 벡터 μ_k와 분산 행렬 Σ_k를 이동시키는 방법으로는 기울기 탐색 알고리즘을 사용한다. 적응된 후의 평균 벡터 μ_k와 분산 행렬 Σ_k의 d번째 차수의 값을 $\hat{\mu}_{kd}$와 $\hat{\sigma}^2_{kd}$로 나타내면 적응 식은 다음과 같다.

$$\hat{\mu}_{kd} = \mu_{kd} + \alpha_\mu \frac{\partial P\{x_t|N(\mu, \Sigma)\}}{\partial \mu_{kd}}, \quad (3)$$

$$\hat{\sigma}^2_{kd} = \sigma^2_{kd} + \alpha_\Sigma \frac{\partial P\{x_t|N(\mu, \Sigma)\}}{\partial \sigma^2_{kd}}. \quad (4)$$

α_μ와 α_Σ는 $0<\alpha_\mu, \alpha_\Sigma<1$인 수렴상수이며, 그 값이 1보다 크면 적응 식이 수렴하지 않을 수도 있다. (3)과 (4) 우변의 편미분 식은 다음과 같이 계산된다.

$$\frac{\partial P\{x_t|N(\mu, \Sigma)\}}{\partial \mu_{kd}} = \frac{\partial}{\partial \mu_{kd}} \sum_{k=1}^{K} N(x_t|\mu_k, \Sigma_k) \quad (5\text{-a})$$

$$= P\{x_t|N(\mu, \Sigma)\}(x_{td}-\mu_{kd})/\sigma^2_{kd}, \quad (5\text{-b})$$

$$\frac{\partial P\{x_t|N(\mu, \Sigma)\}}{\partial \sigma^2_{kd}} = \frac{\partial}{\partial \sigma^2_{kd}} \sum_{k=1}^{K} N(x_t|\mu_k, \Sigma) \quad (6\text{-a})$$

$$= \frac{1}{2} P\{x_t|N(\mu, \Sigma)\}\{(x_{td}-\mu_{kd})^2 - \sigma^2_{kd}\}/(\sigma^2_{kd})^2. \quad (6\text{-b})$$

여기서, x_{td}는 입력 특징 벡터 x_t의 d번째 차수의 값이다. 평균 벡터 μ와 분산 행렬 Σ를 적응시켜주는 식을 정리하면 다음과 같다.

$$\hat{\mu}_{kd} = \mu_{kd} + \alpha_\mu P\{x_t|N(\mu, \Sigma)\}(x_{td}-\mu_{kd})/\sigma^2_{kd}, \quad (7)$$

$$\hat{\sigma}^2_{kd} = \sigma^2_{kd} + \frac{1}{2}\alpha_\Sigma P\{x_t|N(\mu, \Sigma)\}\{(x_{td}-\mu_{kd})^2 - \sigma^2_{kd}\}/(\sigma^2_{kd})^2. \quad (8)$$

(7)과 (8)에 의해 평균 벡터와 분산 행렬을 적응시켜주는 과정을 새로운 화자로부터 얻은 전체 적응 데이터에 대해 반복적으로 수행하여 적응된 평균 벡터 $\hat{\mu}$와 분산 행렬 $\hat{\Sigma}$를 구한다. 분산 행렬의 적응에서는 적응 데이터가 충분하지 못하기 때문에 $\hat{\sigma}^2_{kd}$가 지나치게 작은 값을 갖게 될 수 있다. 이러한 문제를 방지하기 위해 $\hat{\sigma}^2_{kd}$의 최소값을 $\sigma^2_{\min}$으로 제한하여 준다.

2.2 상태 천이 확률과 혼합 밀도 계수의 적응

SCHMM의 화자 적응을 위해서는 가우시안 확률 밀도를 나타내는 평균 벡터 μ와 분산 행렬 Σ뿐만 아니라 상태 천이 확률 A와 혼합 밀도 계수 C도 적응시켜주어야 한다. 본 논문에서는 상태 천이 확률 A와 혼합 밀도 계수 C의 적응을 위해 가중 평균 방법을 적용하였다.

상태 i에서 j로의 천이 확률을 a_{ij}라고 하고, 상태 j에서의 l번째 가우시안 확률 밀도에 대한 혼합 밀도 계수를 $c_j(l)$이라 할 때, 기존의 a_{ij}, $c_j(l)$과 적응된 평균 벡터 μ와 분산 행렬 Σ을 이용하여 새로운 화자로부터 얻은 적응 데이터와 인식 모델과의 최적의 대응관계를 비터비 역추적으로 구한 후, 새로운 화자의 적응 데이터에 대한 상태 천이 확률 $\bar{a}_{ij}$와 혼합 밀도 계수 $\bar{c}_j(l)$을 구한다. 여기서 얻은 $\bar{a}_{ij}$와 $\bar{c}_j(l)$은 적은 양의 적응 데이터로부터 구해진 값이므로, 이를 그대로 인식 과정에 적용한다면 좋은 인식 성능을 기대할 수 없다. 따라서 $\bar{a}_{ij}$, $\bar{c}_j(l)$와 기존의 a_{ij}, $c_j(l)$의 가중 평균을 취하여 적응된 상태 천이 확률 $\hat{a}_{ij}$와 혼합 밀도 계수 $\hat{c}_j(l)$을 구한다.

$$\hat{a}_{ij} = (1-\alpha_a)a_{ij} + \alpha_a \bar{a}_{ij}, \quad (9)$$

$$\hat{c}_j(l) = (1-\alpha_c)c_j(l) + \alpha_c \bar{c}_j(l). \quad (10)$$

여기서, α_a와 α_c는 $0<\alpha_a, \alpha_c<1$인 수렴 상수이다. 위의 과정을 전체 적응 데이터에 대해 반복적으로 수행하여 적응된 상태 천이 확률 $\hat{a}_{ij}$와 혼합 밀도 계수 $\hat{c}_j(l)$을 구한다.

III. 인식 실험 구성

3.1 데이터 베이스

본 논문에서는 한국어 단독음에 대한 인식 실험을 수행하였다. 인식 대상 단어로는 우리말을 구성하고 있는 음소들을 발음 형태에 따라 분류하여 선택한 61개의 단어들[13]을 사용하였다. (예 : "바람", "입술", "나비" 등)

인식 시스템의 학습을 위해서 데이터 베이스를 남성 화자 종속 데이터, 여성 화자 종속 데이터 및 화자 독립 데이터의 3종류로 구성하였다. 남성 화자 종속 데이터는 남성 화자 1명이 61 단어를 10번씩 발음한 데이터이고, 여성 화자 종속 데이터는 여성 화자 1명이 각 단어를 10번씩 발음한 데이터이다. 화자 독립 데이터는 남성 7명과 여성 5명이 각 단어를 2번씩 발음한 데이터이다. 인식 시스템의 성능 평가를 위한 테스트 데이터로는 남성 2명과 여성 2명이 각 단어를 3번씩 발음한 데이터를 사용하였으며, 화자 적응을 위한 데이터로 각 단어를 2번씩 발음한 데이터를 사용하였다. 음성 녹음은 소음이 없는 조용한 환경에서 이루어졌으며, 16 bit 10 kHz의 샘플링으로 A/D하였다.

3.2 전처리 및 특징벡터 추출

음성 데이터는 $1-0.95z^{-1}$의 전달 함수를 갖는 프리 엠퍼시스(pre-emphasis) 필터를 거친 후, 20 ms의 길이를 갖는 해밍 윈도우(Hamming window)를 사용하여 10 ms씩 이동하면서 각 음성 프레임(frame) 마다 14차의 LPC 켑스트럼 계수(cepstral coefficient)를 구하여 인식 시스템의 특징 벡터로 사용하였다.

3.3 인식 시스템 구성

음성 인식 알고리즘으로는 SCHMM을 사용하였다. 인식 단위는 단어이며, 각 단어마다 하나의 모델을 갖는다. 각 모델은 단어에 포함된 음소의 수에 비례하여 5에서 17개의 상태를 갖도록 구성하였다. 분산 행렬은 대각 행렬을 사용하였으며, 혼합 확률 밀도는 전체 128개의 가우시안 확률 밀도 중 7개의 가우시안 확률 밀도를 사용하여 구했다($L=128$, $K=7$).

인식 시스템의 학습 데이터에 따라서 남성 화자 종속 시스템, 여성 화자 종속 시스템, 화자 독립 시스템의 3가지 인식 시스템을 구현하였다. 남성 화자 종속 시스템은 남성 1명이 인식 대상 단어를 10번씩 발음한 데이터로 구성된 남성 화자 종속 데이터를 사용하여 학습시킨 시스템이고, 여성 화자 종속 시스템은 여성 화자 1명이 각 단어를 10번씩 발음한 데이터로 학습시킨 시스템이며, 화자 독립 시스템은 12명의 화자가 각 단어를 2번씩 발음한 화자 독립 데이터로 학습시킨 시스템이다.

IV. 인식 실험 결과 및 고찰

표 1은 화자 적응 과정을 수행하지 않은 경우 남성 2명과 여성 2명의 테스트 화자에 대한 화자 독립, 남성 화자 종속, 여성 화자 종속 시스템의 인식률이다.

표 1. 화자 적응을 수행하지 않은 경우의 인식률 [%]

인식 시스템 \ 테스트 화자	남1	남2	여1	여2	평균
화자 독립	87.4	89.1	84.1	88.0	87.4
남성 화자 종속	70.5	79.8	17.5	26.2	49.3
여성 화자 종속	13.7	20.2	56.3	72.7	40.7

저조한 인식 성능을 보이고 있으며, 특히 남성 화자 종속 시스템을 여성 화자가 사용한 경우나, 여성 화자 종속 시스템을 남성 화자가 사용한 경우에는 매우 낮은 인식률을 보이고 있다.

4명의 테스트 화자들이 각 인식 대상 단어를 1번씩 발음한 경우와 2번씩 발음한 경우의 두 가지 적응 데이터를 사용하여, SCHMM의 변수들 중 혼합 밀도 계수 C를 적응시켰을 때의 인식률을 표 2에 보였다. 수렴 상수 α_c는 0.5로 하였다. 혼합 밀도 계수 C의 적응은 화자 적응에서 큰 역할을 한다. 인식 성능이 크게 향상된 것을 볼 수 있다.

표 2. 혼합 밀도 계수 C를 적응시켰을 경우의 인식률 [%]

인식 시스템	적응 데이터 \ 테스트 화자	남1	남2	여1	여2	평균
화자 독립	1	94.0	97.8	88.5	95.1	93.9
	2	96.7	97.8	91.8	95.6	95.5
남성 화자 종속	1	96.2	92.9	80.9	90.7	90.2
	2	97.3	98.4	84.7	94.0	93.6
여성 화자 종속	1	80.9	91.8	91.8	95.1	89.9
	2	86.3	93.4	94.0	96.2	92.5

표 3은 혼합 밀도 계수 C와 상태 천이 확률 A를 적응시킨 경우의 인식률이다. 상태 천이 확률은 음성의 시간적인 변화를 모델링하기 때문에 화자가 달라져도 크게 변하지 않는다. 따라서 다수의 화자에 의해 학습된 화자 독립 시스템의 경우는 인식률의 향상이 거의 없으며, 한 화자에 의해 학습된 화자 종속 시스템의 경우에만 약간의 인식률 향상을 보이고 있다. 상태 천이 확률에 대한 수렴 상수 α_a는 0.5로 하였다.

표 4는 혼합 밀도 계수 C, 상태 천이 확률 A와 함께 가우시안 확률 밀도의 평균 벡터 μ를 적응시킨 경우의 인식률이다. 평균 벡터를 적응시키기 위한 수렴 상수 α_μ는 0.1을 사용하였다. 입력 음성 특징과 가우시안 확률 밀도와의 차이가 줄었기 때문에 정확한 혼합 확률을 구할 수 있어 인식 성능이 향상됨을 볼 수 있다.

표 3. 혼합 밀도 계수 C와 상태 천이 확률 A을 적응시켰을 경우의 인식률 [%]

인식 시스템 \ 테스트 화자 / 적응 데이터		남1	남2	여1	여2	평균
화자 독립	1	92.9	96.7	89.6	95.6	93.7
	2	97.8	97.8	91.8	96.2	95.9
남성 화자 종속	1	95.6	93.4	85.8	93.4	92.1
	2	98.4	98.4	87.4	95.1	94.8
여성 화자 종속	1	79.2	89.1	91.3	95.6	88.8
	2	87.4	93.4	94.5	96.7	93.0

표 4. 혼합 밀도 계수 C, 상태 천이 확률 A와 평균 벡터 μ를 적응시켰을 경우의 인식률 [%]

인식 시스템 \ 테스트 화자 / 적응 데이터		남1	남2	여1	여2	평균
화자 독립	1	94.0	99.5	93.4	96.2	95.8
	2	98.9	99.5	92.9	95.1	96.6
남성 화자 종속	1	95.6	96.2	85.3	92.4	92.4
	2	99.5	98.4	88.0	94.5	95.1
여성 화자 종속	1	82.0	92.4	92.4	97.8	91.1
	2	89.6	92.4	93.4	98.4	93.4

SCHMM의 모든 변수를 적응시킨 경우의 인식률을 표 5에서 보였다. 일반적으로 적응 데이터는 학습에 사용되는 데이터에 비해 상당히 적은 양이다. 적은 양의 데이터로 분산 행렬을 변화시키는 경우 지나치게 작은 값을 갖게 되어 인식 성능을 저하시킬 위험이 있으므로, 세심한 주의가 필요하며, 기존 분산 행렬의 값을 크게 변화시키는 것은 피해야 한다. 본 논문에서는 이러한 문제가 발생되지 않도록 분산 행렬의 수렴 상수 α_Σ를 다른 변수의 수렴 상수보다 크게 작은 값인 5.0×10^{-5}으로 하였다. 인식 결과를 보면 분산 행렬을 적응시키지 않은 표 4와 비교하여 일관된 인식률의 향상을 보이고 있지 못 하다. 화자 독립 시스템에서는 인식률이 약간 향상되었으나, 화자 종속 시스템에서는 인식 성능이 유사하거나 오히려 저하된 결과를 보였다.

화자 적응 과정에서 각 변수들을 적응시키기 위한 반복 회수는 10회로 하였다. 즉, 모든 적응 데이터는 각 변수들을 적응시키기 위해 10번씩 사용되었다.

전체적인 인식 결과를 보면 제안된 화자 적응 알고리즘을 사용하여 화자 적응을 수행한 경우 화자 적응을 하지 않은 경우에 비해 화자 독립 시스템에서는 9.8%, 남성 화자 종속 시스템에서는 46.0%, 여성 화자 종속 시스템에서는 52.7%의 최대 인식률 향상을 보였다. 적응시켜준 변수의 종류에 따른 인식 성능을 표 6에 요약하였다.

표 5. 혼합 밀도 계수 C, 상태 천이 확률 A, 평균 벡터 μ와 분산 행렬 Σ를 적응시켰을 경우의 인식률 [%]

인식 시스템 \ 테스트 화자 / 적응 데이터		남1	남2	여1	여2	평균
화자 독립	1	94.0	99.5	94.0	96.2	95.9
	2	98.4	99.5	94.5	96.7	97.3
남성 화자 종속	1	96.7	95.1	85.8	91.8	92.4
	2	98.9	98.4	89.1	95.1	95.4
여성 화자 종속	1	82.0	89.6	90.2	96.7	89.6
	2	89.1	93.4	91.8	97.3	92.9

표 6. 적응시켜준 변수의 종류에 따른 인식 성능 [%]

인식 시스템 \ 적응시켜준 변수 / 적응 데이터		없음	C	C, A	C, A, μ	C, A, μ, Σ
화자 독립	1	87.4	93.9	93.7	95.8	95.9
	2		95.5	95.9	96.6	97.3
남성 화자 종속	1	49.3	90.2	92.1	92.4	92.4
	2		93.6	94.8	95.1	95.4
여성 화자 종속	1	40.7	89.9	88.8	91.1	89.6
	2		92.5	93.0	93.4	92.9

화자 독립 시스템은 SCHMM와 분산 행렬을 포함하여 모든 변수를 적응시켰을 경우 가장 높은 인식 성능 향상을 보였고, 남성 화자 종속 시스템은 분산 행렬을 적응시킨 경우와 적응시키지 않은 경우에 유사한 인식 성능을 보였으며, 여성 화자 종속 시스템은 분산 행렬을 적응시키지 않은 경우가 더 큰 인식률 향상을 보였다. 화자 독립 시스템의 학습에 사용된 여러 화자로부터 얻은 음성 특징은 한 화자의 음성 특징에 비해 분산이 크기 때문에, 새로운 한 화자에 대한 인식률을 높여주는 화자 적응 과정에서는 적응 데이터의 양이 적다는 문제점이 있더라도 분산 행렬을 적응시켜주는 것이 인식 성능을 향상시킨다고 판단된다. 그러나, 화자 종속 시스템과 같이 한 화자의 음성에 의해 학습된 분산 행렬은 다른 한 화자의 음성 특징의 분산 행렬과 큰 차이는 없다고 보여지며, 분산 행렬을 적응시키는 경우 적응 데이터 양의 부족 문제가 더 크게 작용하여 인식 성능의 향상이 거의 없거나 오히려 저하된다고 판단된다.

V. 결 론

본 논문에서는 관찰 확률 최대화에 의한 화자 적응 알고리즘을 제안하였다. 제안된 알고리즘은 SCHMM의 혼합 확률을 구하는 과정에서 새로운 화자의 음성 특징과 학습을 통해 결정된 가우시안 확률 밀도들과의 차이가 클 때 인식 성능이 저하되는 문제를 해결하기 위해 관찰

확률을 최대화하도록 SCHMM의 변수들을 적응시켜 주는 화자 적응 알고리즘이다. 제안된 알고리즘의 효용성을 검증하기 위해 단독음 인식 실험을 수행하였으며, 그 결과 화자 적응을 수행하지 않았을 때와 비교하여 화자 독립 시스템에서는 9.8%, 남성 화자 종속 시스템에서는 46.0%, 여성 화자 종속 시스템에서는 52.7%의 최대 인식률 향상을 보였다. 평균 벡터와 혼합 밀도 계수의 적응이 인식 성능 향상에 가장 큰 영향을 주는 것으로 판단되며, 분산 행렬은 화자 독립 시스템의 경우 적응하는 것이, 화자 종속 시스템의 경우 적응하지 않는 것이 더 높은 인식 성능을 기대할 수 있다고 판단된다.

참 고 문 헌

1. X. D. Huang and K. F. Lee, "On Speaker-Independent, Speaker Dependent, and Speaker-Adaptive Speech Recognition," *Proc. ICASSP*, pp. 877-880, Toronto Canada, May 1991.
2. K. Shikano, K. F. Lee and R. Reddy, "Speaker Adaptation through Vector Quantization," *Proc. ICASSP*, pp. 2643-2646, Tokyo, Japan, Apr. 1986.
3. S. Nakamura and K. Shikano, "A Comparative Study of Spectral Mapping for Speaker Adaptation," *Proc. ICASSP*, pp. 157-160, Albuquerque, NM, Apr. 1990.
4. F. Class, A. Kaltenmeier, P. Regal and K. Trottler, "Fast Speaker Adaptation for Speech Recognition," *Proc. ICASSP*, pp. 133-136, Albuquerque, NM, Apr. 1990.
5. F. Class, A. Kaltenmeier and P. Regal, "Fast Speaker Adaptation Combined with Soft Vector Quantization in an HMM Speech Recognition System," *Proc. ICASSP*, pp. 461-464, San Francisco, CA, Mar. 1992.
6. Raymond L. Watrous, "Speaker Normalization and Adaptation Using Second-Order Connectionist Networks," *IEEE Trans. on Neural Networks*, Vol. 4, No. 1, pp. 21-30, 1993.
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, 1994.
8. C.-H. Lee, C.-H. Lin and B.-J. Huang, "A Study on Speaker Adaptation of Continuous Density HMM Parameters," *Proc. ICASSP*, pp. 145-148, Albuquerque, NM, Apr. 1990.
9. C.-H. Lee and J.-L. Gauvain, "Speaker Adaptation Based on MAP Estimation of HMM Parameters," *Proc. ICASSP*, pp. 558-561, Minneapolis, MN, Apr. 1993.
10. J.-T. Chien, C.-H. Lee, and H.-C. Wang, "Improved Bayesian Learning of Hidden Markov Models for Speaker Adaptation," *Proc. ICASSP*, Vol. II, pp. 1027-1030, Munich, Germany, Apr. 1997.
11. G. Rigoll, "Speaker Adaptation Using Improved Speaker Markov Models," *Proc. ICASSP*, Vol. II, pp. II-566-569, Mineapolis, MN, Apr. 1993.
12. H. Hattori, "Speaker Adaptation Based on Markov Modeling of Speakers in Speaker-Independent Speech Recognition," *Proc. ICASSP*, pp. 845-848, Toronto, Canada, May 1991.
13. 한국 방송공사, "표준 한국어 발음 대사전," 어문각, 1993

▲양 태 영(Tae-Young Yang)
한국음향학회지 1996년 15권 2호 참조
현재 : 연세대학교 대학원 전자공학과 박사과정

▲신 원 호(Won-Ho Shin)
한국음향학회지 1996년 15권 2호 참조
현재 : 연세대학교 대학원 전자공학과 박사과정

▲전 원 석(Won-Suk Jun)
한국음향학회지 1997년 16권 4호 참조
현재 : 연세대학교 대학원 전자공학과 석사과정

▲김 지 성(Ji-Sung Kim)

1992년 3월 ~ 1996년 2월 : 연세대학교 전자공학과(공학사)
1996년 9월 ~ 현재 : 연세대학교 대학원 전자공학과 석사과정
※ 주관심분야 : 잡음환경하에서의 음성인식

▲김 원 구(Weon-Goo Kim)
한국음향학회지 1992년 11월 2호 참조
현재 : 군산대학교 전기공학과 조교수

▲이 충 용(Chungyong Lee)

1983년 3월 ~ 1987년 2월 : 연세대학교 전자공학과(공학사)
1987년 3월 ~ 1989년 2월 : 연세대학교 전자공학과 대학원(공학석사)
1990년 3월 ~ 1991년 8월 : 연세대학교 산업기술 연구소 연구원
1991년 9월 ~ 1995년 12월 : Georgia Institute of Technology (Ph.D.)
1996년 2월 ~ 1997년 7월 : 삼성전자 선임연구원
1997년 8월 ~ 현재 : 연세대학교 기계전자공학부 조교수
※ 주관심분야 : 통신 신호처리, 음성인식, 레이다/소나 신호처리, 비선형 신호처리

▲윤 대 희(Dae Hee Youn)
한국음향학회지 1994년 13권 1호 참조
현재 : 연세대학교 기계전자공학부 교수

▲차 일 환(Il-Whan Cha)
한국음향학회지 1995년 14권 4호 참조
현재 : 연세대학교 기계전자공학부 교수