한국음향학회지 제20권 제6호 pp. 26~33 (2001)

지속시간항을 갖는 AR HMM을 이용한 잡음환경에서의 강인 화자인식 시스템 구현

Implementation of a Robust Speaker Recognition System in Noisy Environment Using AR HMM with Duration-term

이 기 용*, 임 재 열**
(Ki Yong Lee*, Jae Yeol Rheem**)

*숭실대학교 정보통신 전자공학부, **한국기술교육대학교 정보기술공학부
(접수일자: 2000년 3월 28일; 수정일자: 2001년 7월 19일; 채택일자: 2001년 8월 2일)

기존의 AR HMM (auroreg ressive hidden morkov model)에 의한 화자인식 방법은 그 성능이 우수하나, 잡음에 대한 것이 고려되지 않아 실제 환경에 적용시 성능저하가 문제가 된다. 본 논문에서는 실제 환경에 맞추기 위하여 관측 신호 모델에서 잡음을 고려하고, 화자인식 성능을 개선하고자 지속시간항 (duration-term)을 포함하는 AR HMM을 이용하여 잡음환경에서의 강인한 화자인식 시스템을 제안한다. 100명의 화자 (남자 77명, 여자 23명)가 2주에 걸쳐 6번 발성한 숫자음 데이터베이스을 가지고, 백색잡음 및 자동차 잡음하에서 실험한 결과, 제안된 방법으로 성능이 향상됨을 확인하였다.
핵심용어: 화자검증, 지속시간항, AR HMM, 잡음, Kalman 필터링, IMM (Interacting Multiple Model)
투고분야: 음성처리 분야 (2.5)

Though speaker recognition based on conventional AR HMM shows good performance, its lack of modeling the environmental noise makes its performance degraded in case of practical noisy environment. In this paper, a robust speaker recognition system based on AR HMM is proposed, where noise is considered in the observation signal model for practical noisy environment and duration-term is considered to increase performance. Experimental results, using the digits database from 100 speakers (77 males and 23 females) under white noise and car noise, show improved performance.
Keywords: *Speaker verification, Duration-term, AR HMM, Noise, Kalman filtering, IMM*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

최근 인터넷의 보급으로 인하여 인터넷 구매, 인터넷 뱅킹, 사이버트레이딩, 등 전자상거래의 이용이 급속도로 늘어가고 있어, 해킹 방지 등을 목적으로 암호화 기법에 의한 전자인증 방식이 널리 보급되고 있다. 그러나 이러한 전자인증 방식의 경우에도 ID (identification)및 암호 (패스워드)가 도용되는 것과 같은 근본적인 문제는 막을 수가 없어 전자상거래에서의 보안문제는 앞으로 더욱 심각해질 것이다. 이에 따라 음성, 지문, 홍체, 안면 (face), 손모양, 등등 신체의 일부 정보를 복합적으로 이용하여 사용자 본인임을 확인하는 생체인식 개념이 최근 각광을 받고 있다. 이러한 기법은 ID와 패스워드가 도용되어도 본인임을 확인하는 이중 절차에 의하여 불법사용을 근본적으로 막을 수 있기 때문이다. 특히 음성의 경우에는 지

책임저자: 이기용 (kylee@saint.soongsil.ac.kr)
156-743 서울시 동작구 상도동
숭실대학교 정보통신 전자공학부
(전화: 02-820-0908; 팩스: 02-821-7653)

문, 홍채, 안면, 손모양 등의 신체적 특성과는 다르게, 개인의 건강상태나 감정상태에 따른 변화가 반영되기 때문에, 이를 활용할 경우에는, 위협 등에 의하여 타의로 인증을 시도하는 경우에도 배제할 수 있는 장점이 있다.

음성을 이용하여 사용자를 확인하기 위한 방법이 화자인식 (Speaker Recognition)기법이다. 화자인식은 음성신호에 들어있는 사람 개개인의 독특한 정보를 기본으로, 말하는 사람이 누구인지를 자동적으로 인식하는 기술로, 등록된 화자 중에서 가장 유사한 화자를 골라내는 화자식별 (Speaker Identification)과 제시된 화자를 승인 (acceptance)하거나 거절 (rejection)하는 화자검증 (Speaker Verification)으로 나누어진다[1,5]. 화자검증 방식을 이용하면 전자인증방식에서 한층 진보된 화자인증이 가능해진다. 이러한 기술은 전화망 또는 인터넷을 이용하는 은행송금/이체기능, 전화 쇼핑, 정보 검색, 정보 서비스, 음성 우편, 정보시스템의 보안유지와 같은 여러 서비스에서 사용자의 신분을 확인 가능하도록 하여, 네트워크 시대의 개인 식별 및 검증에 잘 맞는 방법이라 할 수 있다.

화자검증 방법은 세가지 유형 – 문장종속 (text dependent), 문장독립 (text independent), 문장지정 (text prompted approach) – 으로 분류된다. 문장종속방법은 미리 문장을 정해서 화자가 같은 문장을 발성했을 때 화자를 인식하는 방법으로 그 구현이 간단하고 성능이 우수하나 개인의 비밀이 노출되는 단점을 가지고 있다. 문장독립 방법은 임의의 문장을 사용하는 방식으로 비밀이 보장되나 구현이 상대적으로 어렵고, 성능이 떨어지는 단점을 가지고 있다. 문장지정 방식은 한정된 단어들을 정하고 사용할 때마다 순서를 다르게 조합하여 디스플레이 장치를 통하여 화자에게 제시하고 그 문장을 발성하게 함으로 개인의 비밀을 보장하고 인식률을 높일 수 있는 장점을 가지고 있는 방법이다. 문장종속이나 문장독립 방식의 경우에는, ID와 패스워드 도용과 같이 화자의 목소리를 녹음하여 사칭할 경우에는 근본적으로 막을 방법이 없다.

화자검증 방법에 많이 사용되는 기법은 Dynamic Time Warping (DTW)[1], Gaussian Mixture Models (GMM) [15]–[17], Vector Quantization[4], Hidden Markov Model (HMM)[2–5], Neural Network (NN) 등이며, 최근 GMM과 HMM방법을 많이 사용하고있는 추세이다.

HMM에서의 관측열을 auroregressive (AR) Gaussian source로 가정하는 AR HMM에 의한 방법은 벡터 양자화기나 DTW에 의한 방법보다 우수한 성능을 가지고 있으나 실용화하는 경우, 사용하는 음성신호가 잡음이 없는 환경을 가정하고 있어, 전화망 등 주변 잡음이 있는 실제환경에서 사용될 때 상대적으로 성능이 저하되는 문제점을 가지고 있다. 본 논문에서는 이러한 현실 적용의 문제점을 해결하기 위하여 잡음을 모델링하여 잡음환경에 대한 강인성을 확보하고, 아울러 화자인식 성능을 개선하고자 관측신호모델에 지속시간항 (duration-term)을 포함하는 AR HMM을 이용하여 잡음환경에 강한 문장지정 방식의 강인 화자인식 시스템을 제안하고 구현했다. 지속시간이 모델링될 경우, 화자개인의 발성습관에 따른 단어 지속시간이 추가되어 화자간의 변별력이 높아짐으로써 화자인식률이 향상된다[12].

기존의 AR HMM에서는 기하분포 (geometric distribution)로 지속시간 확률을 암시적으로 모델하고 있으나, 단어의 지속시간에 관한 모델은 포함하고 있지 않는다. 주변 잡음을 가정하지 않은 경우, 즉 깨끗한 음성신호에 대한 AR HMM에 지속시간항을 추가할 경우에는, 수정된 전향 알고리즘 (forward algorithm) 및 후향 (backward algorithm) 알고리즘을 이용하여, 파라미터 재추정 (parameter reestimation)이 가능하고, 수정된 Viterbi 알고리즘을 이용하여 인식 가능하다[18,19]. 그런데 잡음이 고려된 경우에는 지속시간 추정과정 중에 잡음 신호에서 깨끗한 신호에 대한 추정이 동반되야 하므로, 앞서 언급된 수정된 알고리즘을 사용할 수 없다. 따라서, 본 논문에서는 관측 잡음 신호의 추정과정 중 지속시간 추정이 가능한 Kalman 필터링과 IMM (Interacting Multiple Model)에 의거한 방법을 제안하고, 100명의 화자(남자 77명, 여자 23명)가 2주에 걸쳐 6번 발성한 숫자음 데이터베이스를 가지고, 백색잡음 및 자동차 잡음하에서 실험한 결과, 제안된 방법이 우수함을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서 음성신호의 AR HMM 모델링을 다루며, 3장에서는 이에 근거한 잡음환경하에서의 추정과 지속시간에 의한 추정을 다룬다. 그리고, 4장에서 실험 및 결과를 살펴보고, 5장에서 결론을 맺는다.

II. 음성신호의 AR HMM 모델링

잡음에 오염되기 이전의 깨끗한 음성신호가 시각 t에서 상태 $M(t)$와 각 상태에 대한 가우시안 AR 프로세스에 의해 필터 (HFM: Hidden Filter Model)로 모델링된다고 가정하자. 이때 음성신호의 상태가 1차 Markov 체인으로

모델링된다고 가정하고 상태 전이 확률을 $p_{M(t-1)M(t)}$, 상태수를 N이라 하면, 잡음에 오염되기 전의 음성신호는 식 (1)과 같이 모델링되어 나타낼 수 있다.

$$x(t) = \boldsymbol{B}_{M(t)}^{T}\, \boldsymbol{y}(t-1) + v_{M(t)}(t) \tag{1}$$

여기서

$$\boldsymbol{B}_{M(t)} = [b_{M(t)}(1), b_{M(t)}(2), \cdots, b_{M(t)}(p)]^{T} \tag{2.b}$$

는 AR 계수 벡터,

$$\boldsymbol{y}(t-1) = [y(t-1), \cdots, y(t-p)]^{T} \tag{2.a}$$

는 p개의 관측열 벡터, 그리고 $v_{M(t)}(t)$는 드라이빙열(driving sequence)로 평균 0, 분산이 $\Sigma_{M(t)}$인 가우시안 프로세스이다.

식 (1)으로 나타내지는 음성신호에 대한 AR HMM의 파라미터는 $\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \Sigma\}$로 나타낼 수 있으며, 여기서 $\boldsymbol{A} = \{p_{ij}\}, \boldsymbol{B} = \{\boldsymbol{B}_i\}, \Sigma = \{\Sigma_i\}, i, j = 1, 2, \cdots, N$이다.

잡음이 고려된 경우에는 지속시간 추정과정 중에 잡음신호에서 깨끗한 신호에 대한 추정이 동반되야 하므로, 앞서 언급된 수정된 알고리즘을 직접 사용할 수 없어 새로운 알고리즘이 필요하다.

III. 지속시간이 고려된 잡음환경에서의 강인 화자인식 방법

3.1. 잡음환경에서의 강인한 화자인식

잡음음성으로부터 얻은 음성 파라미터 열은 다음과 같이 관측된 신호로 표현할 수 있다.

$$y(t) = x(t) + w(t) \tag{3}$$

여기서 $y(t)$는 관측된 잡음에 오염된 신호이고 $w(t)$는 잡음 신호로 평균이 0이고 분산이 Σ_w인 백색잡음으로 가정한다.

시각 t까지 관측된 잡음 신호를 $Y^t = \{y(1), y(2), \cdots, y(t)\}$로 나타내고, 지정된 화자 음성에 대한 모델 파라미터 λ가 주어진 경우, 잡음환경하의 화자인식 방법을 얻기 위해서는 아래와 같이 관측된 잡음 신호에 대한 유사도 함수 (likelihood function)를 정의 할수 있다.

$$\log p(Y^t|\lambda) = \log \sum_{M^t} \int p(M^t, X^t, Y^t|\lambda) dX^t \tag{4}$$

여기서 $M^t = \{M(1), M(2), \cdots, M(t)\}$은 Markov 상태 열이다.

위의 유사도 함수를 풀기 위하여 EM (Expectation and Maximization) 알고리즘을 적용하면 위 수식은 다음과 같이 표현될 수 있다.

$$\begin{aligned} Q(Y^t|\lambda) = & \sum_{\tau=1}^{t} \sum_{M(\tau)} p(M(\tau)|Y^t, \lambda) \int p(x(\tau)|M(\tau), Y^t, \lambda) \\ & \log p(x(\tau), M(\tau), |Y^t, \lambda) dx(\tau) \\ = & \sum_{\tau=1}^{t} \sum_{M(\tau)} p(M(\tau)|Y^t, \lambda)[\log p_{M(\tau-1)M(\tau)} \\ & + \log E\{p(x(\tau)|M(\tau), \lambda)|Y^t\} \\ & + \log E\{p(y(\tau)|x(\tau))|Y^t\}] \end{aligned} \tag{5}$$

최종적으로 화자인식을 위한 규칙은 다음과 같이 얻어진다. 입력된 잡음음성을 화자 i의 모델 파라미터 λ_i와 비교하여 일정 문턱값을 넘으면 화자 i로 인식하고 아니면 거절한다.

$$\begin{cases} \text{if } \dfrac{1}{T} Q(Y^T|\lambda_i) > \Theta, & \textit{accept speaker} \\ \textit{otherwise}, & \textit{reject speaker} \end{cases} \tag{6}$$

여기서 T는 입력된 단어의 전체 시간이고 Θ는 화자를 결정하기 위한 문턱값이다.

식 (5)에서 $p(M(\tau)|Y^t, \lambda)$, $E\{p(x(\tau)|M(\tau), \lambda)|Y^t\}$, $E\{p(y(\tau)|x(\tau))|Y^t\}$은 Kalman 필터링에 의해서 구할 수 있다[20,21]. 여기서 $E\{p(y(\tau)|x(\tau))|Y^t\}$은 주어진 관측신호에서 잡음신호에 대한 분산에 해당되어 직접 추정이 가능하다.

식 (1)과 (3)을 이용하여, 시각 t에서 AR HMM의 상태 i에 있다면 다음과 같이 공간상태 (state-space) 모델식을 만들 수 있다.

$$x(t) = \Phi[M(t) = i]\, x(t-1) + G[M(t) = i] v(t) \tag{7}$$

$$y(t) = Hx(t-1) + w(t) \tag{8}$$

여기서,

$$\boldsymbol{x}(t) = [x(t), x(t-1), \cdots, x(t-p+1)]^{T}, \tag{9}$$

$$H = [1\ 0\ \cdots\ 0], \tag{10}$$

$$\Phi[M(t)=i]=\begin{bmatrix} \sum_{m=0}^{M} b^1_{m,i}t^m & \sum_{m=0}^{M} b^2_{m,i}t^m & \cdots & \sum_{m=0}^{M} b^p_{m,i}t^m \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{11}$$

$$G[M(t)=i]=[1\ 0\ \cdots\ 0]^T \tag{12}$$

이다. 그리고 $M(t)$는 시각 t에서 모델의 상태를 나타내며, 1과 모델의 상태수 N 사이의 값을 가지며, $M(t)\in\{1,2,\cdots,N\}$, 여기 신호(excitation process) $v(t)$와 관측(measurement) 잡음 신호 $w(t)$는 서로 독립이라고 가정한다.

식 (7)-(12)의 계산은 Kalman 필터에 의하여 최소평균자승에러 (minimum mean square error: MMSE) 또는 최대사후확률 (maximum a posteriori: MAP)로 계산될 수 있다.

3.2. 상태지속확률 크기함수와 조건 천이확률

식 (5)에서 상태지속 확률함수와 조건 천이확률을 구하기 위하여, 모델의 스위칭 프로세스로는 semi-Markov class에 속하는 sojourn-time-dependent-Markov (STDM) chain을 고려한다. 여기서, 모델의 상태변동을 나타내는 스위칭 프로세스의 상태변화는 천이확률 $p_{ij}(\tau_i)$로 나타낼 수 있으며, 이것은 상태 모델 i에서 시스템의 체류시간 (sojourn-time)을 나타내는 상태 지속시간 (state duration) τ_i의 함수이다[10,12]. 따라서 모델의 상태에 대한 천이확률을 추론하기 위해서 관측상의 조건인 지속시간의 확률밀도함수 (pdf)가 필요하다.

시스템 모델을 나타내는 프로세스 $M(t)$, $t=1,2,\cdots$은 N개의 가능한 상태들 중의 하나에 속하며, STDM chain에 대한 천이의 현재 확률은 지속시간 τ의 함수로 다음과 같이 정의 가능하다.

$$p_{ij}(\tau)=P\{M(t)=j|M(t-1)=i,\tau_i(t-1)=\tau\} \tag{13}$$

여기서, $\tau_i(t-1)$은 $t-1$시점의 상태 i에서 지속시간으로, t시점에서 j상태로 천이할 때의 확률이다. 지속시간 τ의 값은 1에서 최대 k까지 가능하다고 가정하고, $t=0$에서는 초기값으로 시스템 모델의 상태에 관계없이, $\tau=1$이라고 가정한다.

식 (5)의 $y(t)$를, 모델이 위에서 설명된 STDM chain을 따르는 천이를 경험하는 동적시스템 상태의 잡음 관측 (noisy observation)이라 하자. 시점 t에서, 이용 가능한 정보 Y^t가 주어지고, 그 때 상태 i에 속할 모델 프로세스의 확률을 $\mu_i(t)$라 하면 다음과 같이 정의할 수 있다.

$$\mu_i(t)\equiv P\{M(t)=i|Y^t\},\quad i=1,\cdots,N \tag{14}$$

그리고, 시점 t에서 이용할 수 있는 정보 Y^t가 기본인 상태 $M(t)=i$에서 지속시간 τ에 대한 조건확률은 다음과 같다.

$$\begin{aligned} g_i^t(\tau)&=P\{\tau_i(t)=\tau|M(t)=i,Y^t\}\\ &=P\{\tau_i(t)=\tau|M(t)=i,Y^{t-1}\}\\ &=P\{M(t-1)=i,\ \cdots,M(t-\tau+1)=\tau,\\ &\quad M(t-\tau)\neq i|M(t)=i,Y^{t-1}\} \end{aligned} \tag{15}$$

여기서, 상태 $M(t)$의 완벽한 지식은 조건 Y^{t-1}에서 한 지수를 따라가는 것을 허락함을 알 수 있다.

식 (15)에서 관측 Y^{t-1}이 주어진 $t-1$시점에서 i에서 j까지 추정된 천이의 조건 확률은 다음과 같다.

$$\begin{aligned} \hat{p}_{ij}(t-1)&\equiv P\{M(t)=j|M(t-1)=i,Y^{t-1}\}\\ &=\sum_{\tau=1}^{k}P\{M(t)=j|M(t-1)=i,\tau_i(t-1)=\tau,\\ &\quad Y^{t-1}\}\cdot P\{\tau_i(t-1)=\tau|M(t-1)=i,Y^{t-1}\}\\ &=\sum_{\tau=1}^{k}p_{ij}(\tau)g_i^{t-1}(\tau) \end{aligned} \tag{16}$$

식 (16)에서 정의된 $\hat{p}_{ij}$가 현재 시점의 함수인 반면, 식 (13)에서 정의된 p_{ij}는 지속시간 τ의 함수이다.

시점 t의 상태 i에서 지속시간 τ의 조건확률크기함수는 다음과 같이 나타낼 수 있다:

$$g_i^t(1)=1-\frac{\mu_i(t-1)}{a_i(t,1)}b_i(t,1) \tag{17.a}$$

$$g_i^t(s)=\left[1-\frac{\mu_i(t-s)}{a_i(t,s)}b_i(t,s)\right]\prod_{m=1}^{t}\frac{\mu_i(t-m)}{a_i(t,m)}b_i(t,m),\quad s=2,\cdots,t \tag{17.b}$$

$$g_i^t(t+1)=\prod_{m=1}^{t}\frac{\mu_i(t-m)}{a_i(t,m)}b_i(t,m) \tag{17.c}$$

여기서, $b_i(t,s)$는 프로세스가 $t-s$시점에서 i상태에 있을 때 계속해서 s시간 단계 동안 i상태에 머무를 조건확률이며,

$$\begin{aligned} b_i(t,s)&\equiv P\{M(t)=i,\cdots,M(t-s+1)=i\\ &\quad |M(t-s)=i,Y^{t-s}\}\\ &=\sum_{n=1}^{t-s+1}\prod_{j=n}^{t+s-1}p_{ij}(j)g_i^{t-s}(n),\quad s=1,\cdots,t \end{aligned} \tag{18}$$

a_i는 $t-s$ 시점에서 이용할 수 있는 정보 Y^{t-s}가 주어질 때, 다음 s시간 단계에 대해서 같은 상태 i에 머무를 조건확률이다.

$$
\begin{aligned}
&a_i(t,s) \equiv P\{M(t)=i,\cdots,M(t-s+1)=i|Y^{t-s}\} \\
&=\sum_{j=1}^{N} P\{M(t)=i,\cdots,M(t-s+1)=i \\
&|M(t-s)=j,Y^{t-s}\}P\{M(t-s)=j|Y^{t-s}\} \\
&=b_i(t,s)\mu_i(t-s)+\sum_{j\neq i}P\{M(t)=i,\cdots, \\
&M(t-s+1)=i|M(t-s)=j,Y^{t-s}\}\mu_j(t-s) \\
&=b_i(t,s)\mu_i(t-s)+\sum_{j\neq i}\Big[\sum_{n=1}^{t-s+1}P\{M(t)=i,\cdots, \\
&M(t-s+1)=i|M(t-s)=j, \\
&\tau_j(t-s)=n,Y^{t-s}\}g_j^{t-s}(n)\Big]\mu_j(t-s) \\
&=b_i(t,s)\mu_i(t-s)+\sum_{j\neq i}\Big[\sum_{n=1}^{t-s+1}p_{ji}(n)p_{ii}(1)p_{ii}(2) \\
&\cdots p_{ii}(s-1)g_j^{t-s}(n)\Big]\mu_j(t-s) \\
&=b_i(t,s)\mu_i(t-s)+\sum_{j\neq i}\Big[\sum_{n=1}^{t-s+1}p_{ji}(n) \\
&\prod_{l=1}^{s-1}p_{ii}(l)g_j^{t-s}(n)\Big]\mu_j(t-s),\quad s=1,\cdots,k
\end{aligned}
\tag{19}
$$

3.3. 상태 추정 알고리즘

Markov 모델 점프를 가지는 선형시스템에 대한 최적 추정기는 상태수에 따라서 지수적으로 증가하는 메모리를 요구하여 현실적으로 구현하기 어려우며, 준최적 접근방법 (suboptimal approach)에서는 IMM이 구현상 상대적으로 적은 계산량 때문에 효과적이다[8]. IMM 방법에 의한 지속시간 독립 천이확률을 가지는 선형시스템에 대한 상태 추정은 다음과 같다.

IMM 접근방법에서는, 시점 t에서 상태 추정을 이전모델-조건 추정의 다른 조합을 사용하는 각 필터를 가지는 N개의 필터(모델 상태수가 N일 때)를 사용하는 각 가능한 모델 가정하에 계산한다. 그리고 각 모델 천이확률은 식 (13)에서 주어진 지속시간 τ의 알고 있는 함수이나, 각 모델은 알지 못하는 상태 i에서 지속시간 $\tau_i(t)$를 가지므로, 식 (17)의 지속시간에 대한 조건확률함수로부터 식 (16)의 조건 천이확률을 계산할 수 있다.

동적 시스템 상태의 조건확률을 구하기 위하여, 시점 t에서 이용 가능한 잡음 관측열 Y^t로부터 잡음이 배제된 신호 $x(t)$에 대한 조건확률은, 병렬로 동작하는 N개의 필터에 대한 전체 확률로 다음과 같이 나타낼 수 있다.

$$
\begin{aligned}
p[x(t)|Y^t] &= \sum_{j=1}^{N} p[x(t)|M(t)=j, \\
&\quad y(t),Y^{t-1}]P\{M(t)=j|Y^t\} \\
&= \sum_{j=1}^{N} p[x(t)|M(t)=j,y(t),Y^{t-1}]\mu_j(t)
\end{aligned}
\tag{20}
$$

식 (20)에서 상태 모델-조건 사후확률은 다음과 같이 다시 쓸 수 있다.

$$
\begin{aligned}
&p[x(t)|M(t)=j,y(t),Y^{t-1}] \\
&=\frac{p[y(t)|M(t)=j,x(t)]}{p[y(t)|M(t)=j,Y^{t-1}]}p[x(t)|M(t)=j,Y^{t-1}]
\end{aligned}
\tag{21}
$$

여기서, 식 (21)의 마지막 항은 사전조건확률이다.

$$
\begin{aligned}
&p[x(t)|M(t)=j,y(t),Y^{t-1}] \\
&=\sum_{i=1}^{N} p[x(t)|M(t)=j,M(t-1)=i, \\
&Y^{t-1}]p\{M(t-1)=i|M(t)=j,Y^{t-1}\} \\
&=\sum_{i=1}^{N} p[x(t)|M(t)=j,M(t-1)=i, \\
&Y^{t-1}]\mu_{ij}(t-1|t-1)
\end{aligned}
\tag{22}
$$

여기서,

$$
\mu_{ij}(t-1|t-1)\equiv P\{M(t-1)=i|M(t-1)=j,Y^{t-1}\}
\tag{23}
$$

이다.

식 (22)는 (4)와 (5)의 잡음 항에서 전형적인 가우시안 가정 (Gaussian assumptions)하의 가우시안 혼합 (Gaussian mixture)을 나타낸다. 그러므로 모델 상태 $j, j=1,\cdots,N$에 일치된 필터에서 입력은 이들 N개 필터의 반복으로부터 얻어지며, 이들 반복은 가중치(확률) $\mu_{ij}(t-1|t-1)$를 따르는 추정 $\hat{x}^i(t-1|t-1)$의 혼합을 구성한다. STDM 상황에서 확률 (14)와 (23)은 이런 전형적인 모델 스위칭에 대한 순환 상태 추정 알고리즘을 얻기 위해서 필요한 중요한 결과이다. 한 사이클의 알고리즘은 다음과 같다.

공분산 $P^i(t-1|t-1)$과 관련있는 모델-조건 추정 $\hat{x}^i(t-1|t-1)$로 시작하는 것은 식 (22)을 따르는 $M(t)=j$에 일치된 필터에 대한 혼합된 초기 조건을 계산한다:

$$
\hat{x}^{0j}(t-1|t-1)=\sum_{i=1}^{N}\hat{x}^i(t-1|t-1)\mu_{ij}(t-1|t-1)
\tag{24}
$$

식 (24)로부터

$$\mu_{ij}(t-1|t-1) = \frac{1}{c_j} P\{M(t)=j|M(t-1)=i, Y^{t-1}\}P\{M(t-1)=i, Y^{t-1}\} = \frac{1}{c_j}\hat{p}_{ij}(t-1)\mu_i(t-1) \quad (25)$$

여기서 식 (14)와 (23)이 사용되었으며, c_j는 정규화 상수를 나타낸다.

$$\hat{x}^i(t-1|t-1) \equiv E[x(t-1)|M(t-1)=i, Y^{t-1}] \quad (26)$$

위 식은 $t-1$시점에서 모델-조건 상태 추정이다. 체류시간 확률을 포함하는 항을 사용하는 STDM 경우에 대한 $\hat{p}_{ij}$의 표현은 (16)에서 얻어진 것이다. 식 (25)에 응답하는 공분산은

$$P^{0j}(t-1|t-1) = \sum_{i=1}^{N}\mu_{ij}(t-1|t-1)\{P^i(t-1|t-1) + [\hat{x}^i(t-1|t-1) - \hat{x}^{0j}(t-1|t-1)][\hat{x}^i(t-1|t-1) - \hat{x}^{0j}(t-1|t-1)]\} \quad (27)$$

이다.

추정 (25)와 공분산 (27)은 모델 조건 추정 $\hat{x}^j(t|t)$와 공분산 $P^j(t|t)$을 산출하기 위한 $M(t)=j$에 일치된 표준 Kalman 필터에 입력으로서 사용된다.

r개 필터에 응답하는 유사도함수는 다음과 같이 계산된다.

$$\Lambda_j(t) = p[y(t)|M(t)=j, Y^{t-1}] \approx p[y(t)|M(t)=j, \hat{x}^{0j}(t-1|t-1), P^{0j}(t-1|t-1)] \quad (28)$$

여기서 과거 데이타는 IMM의 중요한 단계를 따르는 식 (25)과 (27)에 의해서 대체된다. 모델 확률은 다음과 같이 갱신된다:

$$\mu_j(t) = P\{M(t)=j, Y^{t-1}\} = \frac{1}{c}\Lambda_j(t)\sum_{i=1}^{N}\hat{p}_{ij}(t-1)\mu_i(t-1) \quad (29)$$

여기서 조건 천이확률 $\hat{p}_{ij}$은 (16)에서 주어지며, 그리고 c는 정규화 상수이다.

$\hat{p}_{ij}$을 가지는 조건에서 방정식 (26)과 (28)은 지속시간-독립 모델 천이를 가지는 시스템에 대한 상태 추정을 가능하게 하는 중요한 결과이다.

마지막으로, 출력에 대한 마지막 상태 추정 공분산은 다음과 같이 (20)과 (14)에 따라서 얻어진다.

$$\hat{x}(t|t) = \sum_{j=1}^{N}\hat{x}^j(t|t)\mu_j(t) \quad (30)$$

$$P(t|t) = \sum_{j=1}^{N}\mu_j(t)\{P^j(t|t) + [\hat{x}^j(t|t) - \hat{x}(t|t)][\hat{x}^j(t|t) - \hat{x}(t|t)]\} \quad (31)$$

잡음이 배제된 신호에 대하여는 기존의 AR HMM에 지속시간이 포함된 수정된 파라미터 재추정식에 의한 학습 알고리즘으로 모델을 구하고, 실제로 잡음이 포함된 음성이 입력되어 들어올 때는 식 (13), (14), (26)을 이용하여 유사도 함수를 구하여 화자의 인정 거절을 정하게 된다.

IV. 실험 및 결과

성능 테스트를 위하여 기존의 방법과 본 논문에서 제안한 방법을 비교하였다. 실험에 사용된 음성 DB는 전체 100명이며 이중에서 남자 77명과 여자 23명이 각각 숫자음 '영', '일', '이', '삼', '사', '오', '육', '칠', '팔', '구' 를 한번에 3번씩 발성을 하였으며, 2주후에 다시 한번 녹음하였다. 이중 학습과정에서 사용된 음성은 각 화자별 첫번째 3번 발성한 음성으로 화자학습을 시켰으며, 테스트 과정은 두 번째 발성한 3번의 음성 DB 중에서 6단어를 임의로 지정해서 사용함으로써 문장지정의 효과를 보도록 했다. 잡음에서 제안된 방법의 성능을 알아보기 위하여 잡음은 0dB, 5dB, 10dB, 15dB, 30dB의 입력 신호 대 잡음비 (SNR)를 가지는 백색잡음과 자동차잡음을 사용하였다. 자동차 잡음은 시속 60km 주행시 마이크로 녹음한 것을 사용하였다.

실험에 사용된 음성 데이터의 샘플링 주파수는 표준 윈도우즈의 사운드 시스템에 맞추어 11.025 kHz를 사용했고, 256 샘플을 한 프레임으로 하여 128샘플을 이동하여 중첩되도록 해밍 윈도우를 취하였다. 화자학습을 위해서 제안된 방법으로는 12차 AR 필터 ($p=12$), 5개의 상태 ($N=5$)를 갖는 AR HMM이 사용되었다. 비교를 위해 사용된 기존의 방법은 5개의 상태 ($N=5$)를 가진 AR HMM과 64개의 mixture를 가진 GMM을 사용하였다. GMM에 사용된 파라미터는 24차 LPC-켑스트럼을 사용하였다. 이때, 켑스트럴 평균 차감법 (CMS)과 스펙트럼 차감법 (SB)이 사전처리과정으로 기존의 방법과 같이 사용되었다.

표 1. 백색잡음에서의 화자 오인식률 (%)
Table 1. False speaker verification rate (%) under white noise.

	SNR 0	SNR 5	SNR 10	SNR 15	SNR 30
GMM	12.5	10.3	8.1	7.2	4.1
AR HMM	12.4	10.1	7.9	6.9	3.8
제안된방법	10.9	8.8	7.1	6.1	2.7

표 2. 자동차잡음에서의 화자 오인식률 (%)
Table 2. False speaker verification rate (%) under car noise.

	SNR 0	SNR 5	SNR 10	SNR 15	SNR 30
GMM	12.9	10.8	8.1	6.8	3.9
AR HMM	12.8	10.5	7.9	6.2	3.4
제안된방법	11.5	9.3	7.1	5.8	2.5

표 3. 화자오인식률 (%)과 duration의 관계
Table 3. False speaker verification rate (%) with duration-term.

		SNR 0	SNR 10	SNR 30
AR HMM	no duration	12.4	7.9	3.8
	duration	11.5	7.1	3.2
제안된방법	no duration	11.1	7.5	2.7
	duration	10.9	7.1	2.4

표 1은 백색잡음환경에서 기존방법과 제안된 방법의 화자오인식률을 나타낸다. 잡음에 대한 고려가 없는 기존의 HMM과 GMM에서도 인식률이 유지되는 것은 사전처리로 스펙트럼 차감법 방법으로 잡음이 제거되었기 때문이다. 제안된 방법이 약 1.0 ~ 1.5% 정도 오인식률이 줄어들어 성능이 향상되었음을 보여준다.

표 2는 자동차잡음환경에서 기존방법과 제안된 방법의 화자인식 성능을 나타낸다.

표3은 백색잡음에서 각각의 방법에 미치는 지속시간의 영향을 나타낸다.

위의 결과에서 알 수 있듯이 제안된 방법이 두 잡음의 환경에서 기존의 방법보다 화자오인식률이 줄어듦을 볼 수 있다. 그러나 제안된 방법에서 사용되는 파라미터의 수가 기존의 방법보다 늘어나는 문제점을 가지고 있다. 지속시간의 사용은 각 방법의 성능을 개선시킬 수 있음을 보였다. 백색잡음의 환경에서 얻은 화자인식의 성능이 상대적으로 자동차잡음의 환경에서 보다 개선됨을 보이는데, 이는 잡음 모델링시 백색잡음을 가정하였기 때문이다. 백색잡음을 유색잡음으로 확장시에는 자동차 잡음에서도 성능향상을 기대할 수 있을 것이다.

본 실험에서는 화자인식을 위한 문턱값은 일반적인 값을 사용하였다. 따라서 문턱값을 위하여 최근에 제안된 방법들을 사용한다면 제안된 방법의 화자오인식률은 더 작아질 수 있을 것으로 보인다[14].

V. 결 론

기존의 AR HMM에 의한 화자인식 방법은 비교적 그 성능이 우수하나, 잡음이 고려되지 않아 실제의 시스템 구현시 두드러진 성능 저하가 문제가 된다. 본 논문에서는 화자인식 성능을 개선하고자 지속시간항을 포함시키고, 잡음을 고려한 AR HMM을 이용하여 잡음환경에 강한 문장지정 방식의 강인 화자인식 시스템을 제안했다.

제안된 시스템은 관측신호를 깨끗한 음성신호와 백색 주변 잡음으로 모델링하고, 상태종속 지속시간을 명시적으로 AR HMM모델에서 구하였다. 화자인식을 위한 유사도 함수의 계산은 Kalman filtering과 IMM에 의한 다중 필터의 순차적인 계산에 의해서 구해지며, 이 과정에서 상태지속시간에 대한 값이 명시적으로 사용된다.

100명의 화자 (남자 77명, 여자 23명)가 2주에 걸쳐 6번 발성한 숫자음 데이터베이스를 가지고, 백색잡음 및 자동차 잡음 하에서 실험한 결과, 제안된 방법이 우수함을 확인하였다.

감사의 글

이 논문은 1997년 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음 (KRF-97-001-E00294).

참 고 문 헌

1. N. R. Dixon, and T. B. Martin, Automatic Speech & Speaker Recognition, *IEEE Press*, 1979.
2. B. S. Atal, "Effectiveness of linear predictive characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1034-1312, 1974.
3. A. E. Rosenberg, et.al., "Sub-word unit talker verification using HMM," *Proc. ICASSP-90*, pp. 269-272, 1990.
4. N. Tishby, "On the application of mixture AR HMM to text independent speaker recognition," *IEEE Trans. ASSP*, vol. 39, pp. 563-570, 1991.
5. T. Matsui, S. Furui, "Comparison of text-independent speaker recognition method using VQ-distortion and HMM," *IEEE Trans.*, SAP, vol. 2, pp. 456-459, 1984.
6. L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal,"

EURASIP Signal Prodssing, **27**, pp. 65-78, 1992.
7. K. Y. LEE, K. Shirai, "Efficient recursive estimation for speech enhancement in colored noise," *IEEE Signal Processing Letters*, vol. **3**, pp. 196-199, 1996.
8. K. Y. LEE, J. Y. RHEEM, and K. Shirai, "Recursive estimation based on the trend Hidden Markov Model in speech enhancement," *IEEE APCCAS-96*, Nov., 1996.
9. Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Trans., ASSP*, vol. **36**, no. 4, 1983.
10. D. Burshtein, "Robust parametric modeling of durations in Hidden Markov Models," *IEEE Trans., SP*, pp. 548-551, 1995.
11. T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM composition in noisy environments," *ECSA EUROSPEECH-95*, pp. 621-624, 1995.
12. L. Campo, P. Mookerjee, and Y. Bar-Shalom, "State estimation for systems with Sojourn-Time-Dependent Markov model switching," *IEEE Trans., AC*, pp. 238-243, 1991.
13. B. K. Sin and J. H. Kim, "Nonstationary hidden Markov model," *Signal processing*, pp. 31-46, 1995.
14. C. S. Liu, H. C. Wang, and C. H. Lee, "Speaker verification using normalized log-likelihood score," *IEEE Trans., Speech and Audio proc.*, vol. **4**, no. 1, pp. 56-60, Jan, 1996.
15. D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans., Speech and Audio proc.*, vol. **3**, no. 1, pp. 72-83, Jan. 1995.
16. D. A. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal processing letters*, vol. 2, no. 3, pp. 46-48, Mar. 1995.
17. D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech and Audio proc.*, vol. 2, no. 4, pp. 639-643, Oct. 1994.
18. L. R. Rabiner and B. H. Junag, Fundamental of Speech Recognition, Englewood Cliffs, New Jersey: Prentice-Hall, 1993.
19. C. Mitchell, M. Harper, and L. Jamieson, "On the Complexity of explicit Duration HMM's," *IEEE Trans. Speech and Audio Proc.*, vol. **3**, no. 3, pp. 213-217, May 1995.
20. B.-G. Lee, K. Y. Lee, and S. Ann, "An EM-based approach for parameter enhancement with an application to speech signals," *Signal Processing*, **46**, pp.1-14, 1995.
21. K. Y. Lee, B-G. Lee, and S. Ann, "Adaptive filtering for speech enhancement in colored noise," *IEEE Signal Processing Letters*, vol. **4**, 10, pp.277-279, Oct. 1997.

저자 약력

● 이 기 용 (Ki Yong Lee)
1983년 2월 숭실대 전자공학과 졸업
1985년 2월 서울대대학원 전자공학과 졸업 (공학석사)
1991년 2월 서울대대학원 전자공학과 졸업 (공학박사)
1994년 8월~1995년 8월: 일본 와세다대학/영국 에딘버러대학 Post-Doc.
1996년: 일본 와세다대학 방문연구원 (JSPS 초청)
1997년: 독일 뮌헨공대 방문연구원 (DAAD 초청)
1991년 9월~1997년 8월: 창원대학교 전자공학과 조교수
1997년 9월~현재: 숭실대 정보통신전자공학부 부교수

● 임 재 열 (Jae Yeol Rheem)

1986년 2월: 서울대학교 전자공학과 (공학사)
1988년 2월: 서울대학교 전자공학과 (공학석사)
1995년 2월: 서울대학교 전자공학과 (공학박사)
1995년 9월~현재: 한국기술교육대학교 정보기술공학부 (부교수)
※ 주관심분야: 음성신호처리, DSP