최적화된 관측 신뢰도와 변형된 HMM 디코더를 이용한 잡음에 강인한 화자식별 시스템*

Md. Tariquzzaman(전남대), 김진영(전남대), 나승유(전남대)

<차 례>

<Abstract>

A Robust Speaker Identification Using Optimized Confidence and Modified HMM Decoder

Md. Tariquzzaman, Jinyoung Kim, Seungyu Na

Speech signal is distorted by channel characteristics or additive noise and then the performances of speaker or speech recognition are severely degraded. To cope with the noise problem, we propose a modified HMM decoder algorithm using SNR-based observation confidence, which was successfully applied for GMM in speaker identification task. The modification is done by weighting observation probabilities with reliability values obtained from SNR. Also, we apply PSO (particle swarm optimization) method to the confidence function for maximizing the speaker identification performance. To evaluate our proposed method, we used the ETRI database for speaker recognition. The experimental results showed that the performance was definitely enhanced with the modified HMM decoder algorithm.

* Keywords: Speaker identification, HMM decoder, Confidence measure, Membership function, PSO.

* 이 논문은 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임.

1. 서 론

음성은 듣는 사람으로 하여금 발성 화자가 누구인지를 쉽게 알 수 있도록 개개인의 특성을 잘 전달해주는 가장 중요한 통신매체중 하나이다. 그러므로 화자 정보를 추출하여 화자를 인식하는 시스템 관련 기술 개발에 관심이 모아져 왔다 [1]-[4]. 화자인식(speaker recognition) 기술은 일반적으로 화자식별(speaker identification)과 화자확인(speaker verification)으로 분류된다. 화자인식기술은 국가안전, 통신 시스템 보호, 컴퓨터 네트워크 보호, 사이버 거래 등과 같은 다양한 분야에 활용되고 있다. 또한 휴머노이드와 같은 인공지능 로봇의 등장으로 자동화자인식 시스템에 대한 요구가 더욱 절실하여졌으며, 화자인식에 대한 지속적인 연구개발이 이루어지고 있다.

그러나 음성정보를 사용한 화자인식 성능은 전송매체의 채널 왜곡이나 주변 환경의 잡음, 코덱의 왜곡 등에 의해 쉽게 손상되며, 실제 응용영역에서 인식률이 상당히 저하되고 있다. 이러한 문제점을 극복하기 위한 많은 알고리듬이 연구되었으며, 크게 두 가지 접근 방법으로 분류할 수 있다. 첫째, 잡음 또는 채널 외곡에 강인한 파라미터를 추출하는 방법으로써 CMS(cepstral mean subtraction) 방법과 RASTA(relative spectra) 방법이 대표적이다[6]-[8]. 둘째, 화자의 모델을 잡음에 맞도록 적응시키는 모델적응 방법이다[9]. 모델적응 방법은 특히 음성인식에서 매우 일반적으로 사용되고 있다. 한편 최근에 새로운 접근 방법이 소개 되었는데, 관측 신뢰도의 개념을 도입하여 부정확한 관측을 지닌 문제를 해결하고자 하는 방법이다[1][2]. 논문 [1][2]에서는 관측 신뢰도라는 개념을 가우시안 혼합모델 (Gaussian mixture model, GMM)에 적용하여 GMM 학습과 인식방법을 제안하였으며, 화자인식 영역에서 검증한 결과 기존의 GMM에 비하여 우수한 성능을 얻었다. 특히 논문 [1][2]에서는 관측 신호의 측정에 대한 신뢰성이 잡음에 의하여 저하된다는 개념 하에 신호대잡음비(signal-to-noise ratio, SNR)의 함수로 관측 신뢰도를 표현하였다.

본 논문에서는 논문 [1][2]에 적용된 관측 신뢰도 기반 변형된 GMM 방법을 은닉 마코프 모델(hidden Markov model, HMM)에 확장 적용하고자 한다. 관측 신뢰도 개념은 HMM의 학습과 인식 단계에 모두 적용할 수 있는데, 본 논문은 오직 인식 단계에서 사용되는 HMM 디코더(decoder)의 변형에 대하여 다룬다. 이는 화자 식별 시스템의 경우, 잡음이 없는 깨끗한 음성을 가지고 화자 모델이 학습되어진다는 가정에 근거한 것이다. 실제, 화자 모델을 구축할 때 깨끗한 음성을 사용하는 경우가 많으므로 타당한 가정이라고 생각된다. 물론, 학습 DB에도 잡음이 있는 경우에 대한 방법론을 개발할 필요가 있지만 이는 차후의 연구로 남겨두고자 한다. 본 논문에서는 제시된 변형된 HMM 디코더 방법론을 문맥종속 화자식별 영역에서 검증하고자 한다. 한편, 관측 신뢰도에 대한 적절한 평가 및 최적화가 필

요한데, SNR로 표현되는 관측 신뢰도의 최적화를 위해 PSO(particle swarm optimzation)를 통한 최적화 실험을 수행하고 그 결과를 기술한다.

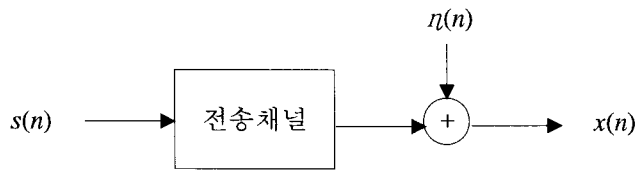
본 논문의 구성은 다음과 같다. 2장에서 관측신뢰도 및 이를 적용한 변형된 HMM 디코더를 제안하고, 3장에서는 관측 신뢰도 최적화를 위한 PSO 적용 방법에 대하여 설명한다. 4장에서는 실험 환경 및 실험 결과를 제시하고 5장에서 결론을 맺는다.

2. HMM 디코더 변형

본 절에서는 HMM 디코더의 변형을 위한 배경으로 관측 신뢰도에 대하여 설명하고, 이를 적용한 HMM 디코더에 대하여 제안하고자 한다.

2.1. SNR 기반 관측 신뢰도

음성신호뿐 아니라 모든 신호는 잡음 또는 전송 채널의 특성에 의하여 왜곡되는데, 잡음은 관측 신호에 더해지는 모습으로 모델링하는 것이 일반적이다. 이를 부가 잡음(additive noise)이라고 하는데, 신호의 왜곡 과정은 <그림 1>과 같이 설명된다. <그림 1>에 보인 바와 같이 임의의 신호 $s(n)$은 전송채널을 통해 왜곡되며, 부가잡음 $\eta(n)$이 더해져 한 번 더 교란된다.

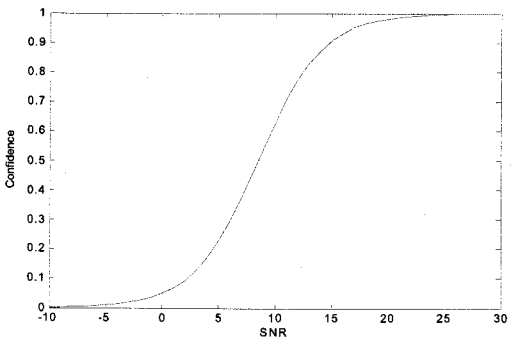


<그림 1 > 일반적인 신호 왜곡 과정

본 논문에서는, 발생신호가 전송채널을 통과하지 않고 측정되는 경우, 즉 관측된 신호가 잡음에 의해서만 왜곡되는 경우에 대하여 고려하였다. 이러한 가정 하에 관측된 신호의 부정확도 또는 신뢰도는 SNR에 의하여 측정이 가능하다고 할 수 있다. 그런데 관측 신뢰도는 애매모호한(fuzzy) 값이며, 멤버쉽(membership) 함수라는 퍼지이론을 동원하여 표현할 수 있다. 즉, 최저값 0과 최고값 1로 표현되는 적절한 함수에 의하여 신뢰도는 기술된다. 본 논문에서는 논문 [1][2]에서 성공적으로 적용된 시그모이드(sigmoid) 함수 기반의 신뢰도 함수를 사용하였다.

$$\rho(SNR) = \frac{1}{1 + e^{-a(SNR-b)}} \tag{1}$$

위 식에서 a는 스케일(scale) 파라미터이고, b는 이동(shift) 파라미터이다. 다음 <그림 2>는 $a = 0.35$, $b = 8.5$인 경우, 관측 신뢰도의 예를 보여주고 있다.



<그림 2> 관측 멤버쉽 함수의 예 (a=0.35, b=8.5)

위 그림에서 보듯이 시그모이드 함수를 사용한 신뢰도 함수는 퍼지 이론에서 정의되는 멤버쉽 함수의 성질을 만족하고 있다.

2.2. HMM 디코더 변형

HMM 모델은 연속(continuous) HMM과 이산(discrete) HMM으로 나눠지는데, 일반적으로 연속 HMM이 음성인식, 화자검증 또는 식별 문제에서 우수한 성능을 보인다. 따라서 본 논문에서는 연속 HMM 모델을 채택하였다. 연속 HMM은 연속 관측확률밀도함수를 사용하는데, 주어진 상태에서 관측확률은 GMM을 이용하여 모델링된다. 연속 HMM은 $\lambda = (A, B, \pi)$로 정의되는데, A는 상태천이행렬, π는 초기 확률 벡터, 그리고 B는 방사모델인데 방사모델은 다음 식들로 표현된다.

$$b_i(x_t) = \sum_{m=1}^{N_M} c_{im}\, \Phi_{im}(x_t; \mu_{im}, \Sigma_{im}) \text{ for } 1 \le i \le N_S \tag{2}$$

$$\begin{aligned} &c_{im} \ge 0,\ 1 \le i \le N_S,\ 1 \le m \le N_M \\ &\sum_{m=1}^{N_M} c_{im} = 1,\ 1 \le i \le N_S \end{aligned} \tag{3}$$

위 식에서 x_t는 관측벡터이며 Φ_{im}는 커널함수이다. 커널함수에서 i번째 상태의 가중치가 c_{im}, 평균벡터는 μ_{im}이고 공분산은 Σ_{im}이다. 그리고 N_M은 가우시안 믹스쳐 수이고, 커널함수 Φ_{im}은 가우시안 분포이다. 또한, N_S는 HMM 모델의 전체 상태의 수이다. 주어진 관측 패턴 $\{X_t\}(t=1,2,....,T)$와 HMM 모델 $\lambda=(A,B,\pi)$에 대하여 $P(\boldsymbol{X}|\lambda)$를 구하기 위해 비터비(Viterbi) 디코더 또는 순방향(forward) 알고리듬이 사용된다. 본 연구에서는 순방향 알고리듬을 확률계산을 위하여 사용하였는데 다음과 같다.

- 순방향 과정

확률 $P(\boldsymbol{X}|\lambda)$를 구하기 위해 일반적인 상태 열을 $Q=[q_1,q_2,.....q_t,.....q_T]$라 하자. 순방향 변수 $\alpha_t(i)$는 식 (4)로 표현된다.

$$\alpha_t(i)=P(x_1,x_2,.....,x_t,\ q_t=S_i|\lambda) \tag{4}$$

위 식은 주어진 모델 λ에 대하여 시간 t에서 상태변수가 S_i일 때 시간 1에서 t까지 주어진 특징 벡터들의 관측 확률이다. 다음 식은 $\alpha_t(i)$를 구하기 위한 반복적인 과정이다.

$$\text{i)}\ \ \alpha_1(i)=\pi_i b_i(x_1),\ 1\le i\le N_S \tag{5}$$

$$\text{ii)}\ \ \alpha_{t+1}(j)=\left(\sum_{i=1}^{N_s}\alpha_t(i)a_{ij}\right)b_j(x_{t+1}),\quad 1\le t\le T-1,\ 1\le j\le N_s \tag{6}$$

$$\text{iii)}\ \ \alpha_T(i)=P(x_1,x_2,...x_t,....,x_T,\ q_T=S_i|\lambda) \tag{7}$$

또한 $P(\boldsymbol{X}|\lambda)$은 $\alpha_T(i)$의 전체 합이므로 $P(X|\lambda)=\sum_{i=1}^{N_s}\alpha_T(i)$과 같다.

본 논문에서는 2.1 절에서 기술된 관측 신뢰도를 이용하여 식 (5)에서 식 (7)로 표현되는 디코더 알도리듬을 수정하여, 화자식별 성능을 향상시키고자 하는 것이다. 그런데 위 반복과정에서 관측된 벡터들과 직접 관계되는 변수는 관측 확률 식 (5)의 $b_i(x_1)$과 식 (6)의 $b_j(x_{t+1})$이다. 따라서 관측 신뢰도를 고려한 순방향 알도리듬의 변형은 식 (5)와 (6)의 관측 확률을 변형하여 이루어질 수 있다. 관측 벡터

x_t의 신뢰도를 ρ_t라고 하자. 그러면, 관측 확률을 ρ_t로 가중하여 다음 식 (8) 및 (9)를 얻을 수 있다.

$$\alpha_1(i) = \pi_i(b_i(x_1))^{\rho_1},\ 1 \le i \le N_S \tag{8}$$

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N_S}\alpha_t(i)a_{ij}\right)(b_j(x_{t+1}))^{\rho_{t+1}} \tag{9}$$

식 (8)과 (9)를 살펴보면, 관측 신뢰도 $\rho_t = 0$인 경우 가중 확률이 1이 되어, log 확률이 0이 되므로, 확률 $P(\boldsymbol{X}|\lambda)$를 계산함에 있어 의미를 상실함을 알 수 있다. 즉 파라미터 x_t는 모든 화자에 대하여 확률이 1로 적용되기 때문에 화자식별에서 화자를 판별함에 아무런 기여를 하지 않는다.

위와 같은 변형은 물론 비터비 디코더에 대해서도 똑같은 방식으로 이루어질 수 있다. 화자식별의 경우, 음성인식과 달리 역추정(back tracking)을 통한 상태 열 결정이 필요 없으므로, 순방향 디코더에 적용하였다.

3. 신뢰도 함수 최적화

식 (1)로 표현된 관측 신뢰도 함수는 변수로서 스케일 파라미터 a와 이동 파라미터 b를 포함하고 있다. 따라서 각 파리미터는 화자식별의 성능을 최대화하기 위하여 최적화 되어야 한다. 최적화를 위해서 두 가지의 문제가 정의 되어야 하는데, 하나는 적절한 목표함수를 정의하는 것이고, 다른 하나는 최적화 방법을 적절히 선택하는 것이다.

본 논문에서는 최적화 목표 함수를 화자식별 문제의 식별률로 정의하였다. 즉, 최적화 목표 함수는 다음과 같은 식으로 정의 된다.

$$x(f(Z)) = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L_k}\delta(\arg Max_m(P_m(X_{kl})),k)}{K\sum_{k=1}^{K}L_k} \tag{10}$$

위 식에서 $x(f)$는 주어진 관측신뢰도 함수 $f(Z)$에 대한 화자 식별률이며, Z는 최적화 대상 파라미터 벡터인 $[a,b]^T$이고, $\delta(i,j)$는 $i=j$일 때 1이고 그렇지 않으면 0인 함수이다. X_{kl}는 k번째 화자의 l번째 음성 데이터이다. K는 전체화자의 수이며, L_k는 화자 당 주어진 발화음성의 개수이다. P_m은 주어진 시료에 대한 m번째

화자 모델에 대한 관측확률이다. 그리고 $\arg_m \max P_m$는 가장 큰 확률을 갖는 화자의 인덱스 (index)를 의미한다. 그리고 m번째 화자의 관측 열 X에 대한 확률 P_m은 $P(X|\lambda_m) = \sum_{i=1}^{N_M} \alpha_T(i)$으로 정의되며, λ_m은 m번째 화자의 모델이다. 식 (10)으로 정의된 최적화 목표함수는 비선형 함수로서 최적화 변수 f에 대하여 닫힌 해(closed solution)를 구할 수 없다. 따라서 비선형 목표함수를 최적화하기 위한 방법을 채택하여야 한다. 본 논문에서는 여러 가지 최적화 방법 중 구현이 매우 간단한 PSO 방법을 채택하였다[9][10]. PSO 방법은 Eberhart와 Kennedy에 의하여 1995년 제안된 방법으로서, 새의 무리 또는 물고기 떼들의 먹이를 찾는 움직임을 모방하여 개발된 방법이다[10]. PSO는 초기 불규칙한(random) 해들의 모임으로 시작한다는 면에서 유전자 알고리듬과 유사하지만, 각 잠재적인 해들이 다시 불규칙한 속도와 이전 잠재적인 해들의 결합으로 구성된다는 측면에서 다르다. 이 잠재적인 해들의 모임을 입자무리(particle swam)이라고 한다.

일반적인 문제로서 파라미터 Z에 의하여 최적화 되어야 할 함수 $f()$가 있다고 하자. 물론 Z는 위에서 기술한 바와 같이 파라미터 a와 b를 갖는 벡터이다. 그러면 PSO 방법은 다음과 같다.

1) Random하게 잠재적인 해들 $\{Z_{i0}\}$를 결정한다. 단, $i = 1,2,...,I$이고 I는 전체 particle의 수이다.
2) 각 iteration j에 대하여 다음을 반복한다.
   2-1) 각 Z_{ij}에 대하여 $f(Z_{ij})$를 구한다. 그리고 j 단계에서 입자들에 대하여 $f(Z_{ij})$에 따른 식별률 x_i를 구한다. $x_i = x(f(Z_{ij}))$이다.
   2-2) 최적 $x(f)$값 즉 $Max_i x_i$의 변화를 계산하고, 수렴한 경우 루프를 빠져나간다.
   2-3) 각 i에 대하여 $\{0,...,j-1\}$에 대하여 가장 최적인 해를 저장한다. 이를 $Zbest_{ij}$라고 하자.
   2-4) 모든 $Zbest_{ij}$를 대상으로 particle 인덱스 i를 대상으로 가장 최적인 해를 저장한다. 이를 $Zgbest_j$라고 하자.
   2-5) 각 particle의 속도를 다음과 같이 계산한다.

$$v_{ij} = v_{ij-1} + c_1 r_1 (Zbest_{ij} - Z_{ij-1}) + c_2 r_2 (Zgbest_j - Z_{ij})$$

   위 식에서 c_1과 c_2는 상수이며 r_1과 r_2는 random한 수이다.
   2-6) 각 particle의 값을 갱신한다.

$$Z_{ij} = Z_{ij-1} + v_{ij}$$

3) $Zgbest_j$를 최적의 해로 결정한다.

위 PSO 알고리듬은 전체 식별률을 최대화시키도록 파라미터 Z를 구하게 되므로 알고리듬 2-3)의 '가장 최적인 해'는 가장 큰 식별률을 갖는 해를 의미한다.

4. 실험 및 고찰

4.1. 실험 DB 및 화자인식 시스템

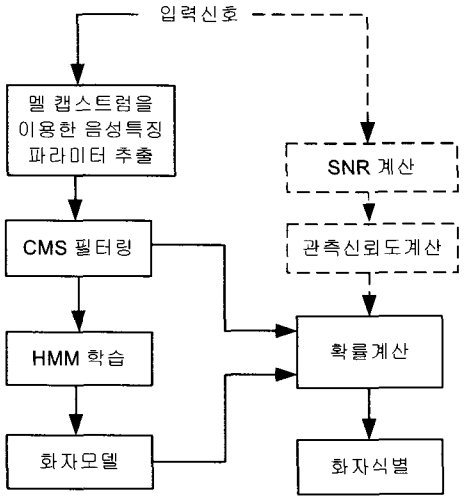
본 논문에서는 제안된 방법의 성능을 확인하기 위하여 ETRI에서 만든 한국어 화자인식용 휴대폰 음성 DB를 사용하여 문맥종속 화자식별 실험을 하였다. 음성 데이터의 샘플링 주파수는 8 kHz이며, 8 비트 μ-law PCM 방식으로 코딩되어 제공되었고 DB의 전체 화자의 수는 남녀 모두 49명이고, 화자 당 음성파일은 모두 20개로 이 중 10개씩을 학습용과 실험용으로 나누어 사용하였다. 문맥종속 인식실험에 사용한 음성데이터의 파형으로 발성시간이 약 3초 정도로 화자모델 학습에 사용된 음성데이터는 파일 10개를 합친 평균 약 30초 정도의 분량이다. 실험에서 입력 음성데이터의 한 프레임은 40 ms로 하였고, 20 ms씩 중첩되어 처리되도록 하고, 음성의 특징벡터는 12차 멜-켑스트럼(mel-cepstrum) 계수와 로그 에너지를 포함하였으며, 채널 왜곡을 보상하기 위하 3절에서 설명한 바와 같이 CMS 방법을 적용하였다.

HMM 모델은 left-to-right 모델을 사용하였고, EM 알고리듬에 의해 HMM 모델 파라미터를 반복적으로 훈련하여 계산하였다. 이 과정에서 공분산 값은 full covariance를 사용하였으며, 알고리듬의 초기 과정에서는 fuzzy c-means clustering 방법을 사용하였다. 이를 정리하면 <표 1>과 같다.

한편, <그림 3>은 관측 신뢰도 기반 변형된 HMM 디코더를 사용하는 화자식별 과정을 보여주고 있다. 그림의 실선은 기존의 화자식별 과정이며, 점선은 변형된 화자식별 과정을 보여준다. 그림에 보인 바와 같이 학습단계는 기존의 HMM 기반 학습 방법과 동일하다. 반면 점선으로 보인 바와 같이, 식별 단계에서는 입력된 음성 신호에 대한 SNR을 추정하고, 추정된 SNR을 통해 관측 신뢰도를 계산한다.

<표 1> 문장종속 화자식별 실험의 개요

음성 DB	ETRI 휴대폰 화자인식용 음성 DB
샘플링/음성코딩	8kHz/8 bits μ-law PCM
화자 수	49
화자당 HMM 모델 학습 파일의 개수	10
화자당 관측 신뢰도 최적화 학습 파일의 개수	5
화자당 테스트 음성파일의 개수	5
프레임길이/중첩	40 ms/20 ms
음성특징벡터	12차 멜 켑스트럼과 로그에너지
채널보상	Cepstral Mean Subtraction
HMM 모델	left-to-right HMM
HMM 학습	EM 알고리듬, full covariance



<그림 3> 관측신뢰도 기반 화자식별 과정

4.2. 실험 결과 및 검토

제안한 방법을 검증하기 위해 먼저 위절에서 설명한 ETRI 음성 DB를 이용하여 다양한 상태수를 대상으로 실험을 하였다. 실험은 관측 확률을 모델링하기 위하여 한 개의 가우시안 분포를 사용하였는데, 실험결과, 문맥종속 HMM 화자식별 실험에서 여러 개의 가우시안 분포를 사용하는 것이 크게 식별률의 향상을 보이지 않았었기 때문이다. 실험에서는 백색 가우시안 잡음을 깨끗한 시험용 신호와 혼합하여 신호의 오염을 발생시켰다. 즉, 오염된 신호 $y[n]$은 식 (11)과 같은 방식으로 얻어지는데, $s[n]$은 깨끗한 음성신호이고, $\eta[n]$은 백색 가우시안 잡음이다.

실험을 위해서는 SNR에 대한 정보가 필요하다. SNR은 깨끗한 원신호 $s[n]$의 전력을 잡음의 전력으로 나누어 식 (12)와 같이 계산하였다.

$$y[n] = s[n] + \eta]n] \tag{11}$$

$$SNR(t) = 10\log\frac{\sigma_s^2(t)}{\sigma_\eta^2} \tag{12}$$

위 식에서 σ_η^2은 부가잡음의 전력이며, $\sigma_s^2(t)$는 t 번째 프레임의 전력이다.

제안된 알고리듬 검증을 위해 총 4 가지의 실험을 하였는데, 실험의 목적은 다음과 같다.

- (실험 1) : HMM의 상태수 결정을 위한 상태수에 따른 식별률 평가
- (실험 2) : 기존 디코더와 제안된 디코더 성능 비교, 단 경험적 파라미터 이용
- (실험 3) : 기존 방법과 변형된 방법의 최적화 이전 및 이후 성능 비교
- (실험 4) : (실험 3)의 최적화 실험들의 평균값 파라미터 이용시 성능 검증
- (실험 5) : (실험 4)의 결과를 추정된 SNR을 사용한 경우와 성능 검증

다음 <표 2>는 기존의 방법에 의한 화자인식 식별을 보여준다. <표 2>에서 신호대잡음비 SNR은 한 발화의 전체 음성구간의 평균 SNR을 말한다. 물론, 발화 전 구간의 평균 SNR이므로 프레임별 신호대잡음비 $SNR(t)$는 시간에 따라 다양한 값을 갖는다.

<표 2>의 결과를 바탕으로 상태수를 9로 결정하였다. 상태 수가 9일 때 전반적으로 인식률이 수렴하였다고 판단하였기 때문이다. <표 3>은 상태수 9일 경우, 기존의 방법과 제안된 방법의 식별결과를 보이고 있다. 실험은 일단 경험적으로 사용된 스케일 파리미터 a=0.35, 이동 파리미터 b=8.5를 사용하였다.

<표 2> (실험 1) 기존의 방법에 의한 식별률
(학습 파일수 = 화자 당 10개, 테스트 파일수 = 화자 당 10개)

상태수 / 평균 SNR	3	5	7	9	11	13
8	29.59	24.49	27.55	26.53	27.55	31.63
12	45.92	45.92	41.83	43.88	45.92	38.79
16	51.02	51.02	47.96	52.04	52.04	48.98
20	63.26	63.26	60.20	75.51	65.30	63.26
30	78.57	78.57	83.67	84.69	88.78	89.80

<표 3> (실험 2) 기존의 방법과 제안된 방법 비교 (a=0.35, b=8.5)
(학습 파일수 = 화자 당 10개, 테스트 파일수 = 화자 당 10개)

평균 SNR	8	12	16	20	30
기존의 방법	26.53	43.88	52.04	75.51	84.69
제안된 방법	52.02	69.30	79.59	83.67	88.78

<표 3>은 변형된 HMM 디코더를 사용하는 것이 기존의 방법에 비하여 높은 화자 식별률을 얻을 수 있음을 보여 주고 있다. 특히, 신호대잡음비가 낮은 경우에 성능개선이 대폭 향상되고 있다. 다음 <표 4>는 각 신호대잡음비에 대하여 PSO 최적화를 수행한 경우, 각 파라미터 a와 b의 값 그리고 식별률을 보여주고 있다. 실험은 두 번의 학습을 필요로 한다. 즉, HMM 모델 학습과 최적 파라미터 학습이다. 따라서 <표 4>의 실험을 위해 HMM 모델 학습에 사용된 발화 파일들을 제외하고, 나머지 10개 중 5개는 최적 파라미터 학습에 사용하고, 다른 5 개는 테스트를 위하여 사용되었다. 즉, <표 1>의 실험에서 테스트로 사용된 파일들을 그룹 1과 그룹 2로 나누어 사용하였으며, 각 집단은 화자 당 5개의 파일을 포함하고 있다.

<표 4> (실험 3) 최적화 실험 결과 (학습 파일수 = 화자 당 10개, 최적화 학습 파일수 = 화자 당 5개, 테스트 파일수 = 화자 당 5개, 그리고 B, M, O는 베이스라인 시스템, 최적화 이전, 최적화 적용을 표시함)

평균 SNR			8	12	16	20	30
실험 3-1 최적화=그룹1 테스트=그룹2	B		24.49	44.90	51.02	77.55	83.67
	M		53.06	69.39	79.59	81.63	89.90
	O		57.14	75.51	81.83	85.71	89.80
	최적값	a	0.585	0.590	0.818	0.785	0.714
		b	3.32	3.33	7.36	12.84	21.16
실험 3-2 최적화=그룹2 테스트=그룹1	B		28.57	42.86	53.06	73.47	85.71
	M		51.02	69.39	79.59	85.71	87.76
	O		51.02	67.51	81.63	89.90	93.84
	최적값	a	0.364	0.521	0.951	0.838	0.814
		b	3.306	2.796	5.499	14.204	15.730

<표 4>의 결과는 PSO를 통해 구한 최적화 파라미터를 사용한 경우 경험적으로 결정한 파라미터를 사용한 경우에 비하여 화자식별의 성능이 향상되었음을 확인할 수 있다. 한편, <표 4>의 실험 3-1과 실험 3-2의 결과를 통합하기 위하여 각 경우에 구해진 파라미터의 값들을 평균하여, 통합 파라미터 값을 결정하고, 10개

의 테스트 파일에 대하여 화자식별 실험을 수행하였다. <표 5>는 이 경우에 대한 실험 결과를 보여준다. <표 5>의 실험 결과는 두 번의 실험을 통해 평균적으로 얻어진 스케일 파리미터와 이동 파라미터 값들을 사용하여도 성능 향상이 유지되고 있음을 알 수 있다.

<표 5> (실험 4) 공통 최적화 파라미터 사용시 성능 (학습 파일수 = 화자 당 10개, 테스트 파일수 = 화자 당 10 개)

평균 SNR		8	12	16	20	30
변형된 디코더		55.01	75.71	81.63	87.76	91.84
평균 최적값	*a*	0.474	0.553	0.885	0.812	0.764
	b	3.311	3.056	6.431	13.520	18.443

지금까지의 검토한 <표 2>에서 <표 5>는 식 (11)을 통하여 정확하게 계산된 SNR을 이용한 실험 결과이다. 그러나 실제 SNR을 정확하게 계산하는 것은 간단한 문제가 아니며, SNR을 추정할 경우 항상 오차를 수반하게 된다[11][12]. 본 논문에서는 다음 식 (13)을 이용하여 프레임별 SNR을 추정하였다.

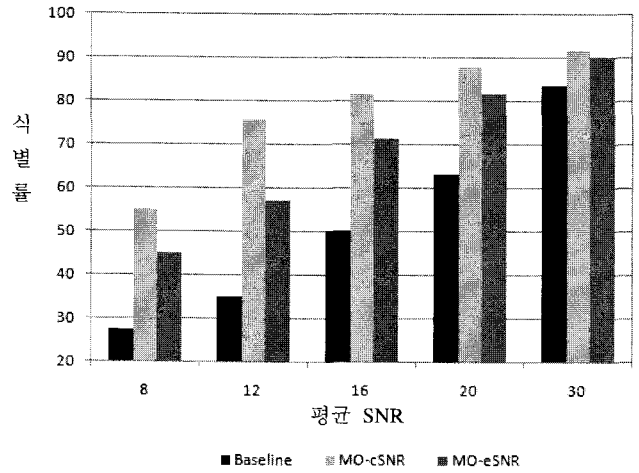
$$\widehat{SNR}(t) = 10\log\frac{\hat{\sigma}_y^2(t) - \hat{\sigma}_\eta^2}{\hat{\sigma}_\eta^2} \tag{13}$$

위 식에서 $\hat{\sigma}_\eta^2$는 추정된 잡음의 전력이고 $\hat{\sigma}_y^2(t)$는 프레임 t의 전력이다. $\hat{\sigma}_\eta^2$는 묵음구간에서 측정하였는데, 음성구간 앞에 존재하는 총 10 프레임의 잡음을 대상으로 결정하였다.

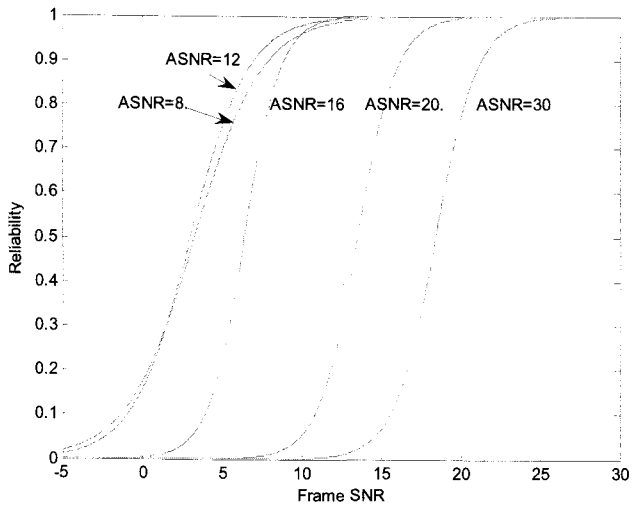
<그림 4>는 최종적으로 기존의 베이스라인 시스템, 제안한 변형된 디코더에 대해 정확한 SNR을 사용한 경우와 추정된 값을 사용한 경우에 대하여 성능을 비교한 것이다. <그림 4>에서 확인할 수 있는 바와 같이 추정된 SNR을 사용하는 경우, 정확한 SNR에 대한 정보를 이용하는 경우에 비하여 식별률이 저하되고 있음을 알 수 있다. 그러나 비록 추정된 SNR을 사용하여도, 기본 화자식별기를 사용하는 경우에 비하여 상당한 식별률 향상이 있음을 확인할 수 있다. 그러므로 본 논문에서 제안하는 최적화 관측신뢰도 기반 변형된 HMM 디코더 알고리듬은 잡음환경하 화자식별에서 성공적으로 적용될 수 있을 것으로 판단된다.

한편 <그림 5>는 각 평균 SNR에 따른 최적 신뢰도 함수들을 동시에 나타낸 것이다. <그림 5>에 따르면, 신뢰도 함수들은 SNR에 따라서 다음과 같은 특성을 갖는다.

첫째, 신뢰도 함수는 그 중심점(신뢰성 값이 0.5인 점)이 평균 SNR이 감소함에 따라서 점차 왼쪽으로 이동한다. 즉, 이동 파라미터의 값이 평균 SNR이 감소함에 따라서 줄어든다.



<그림 4> (실험 5) 베이스라인 시스템, 변형된 HMM 디코더 시스템의 SNR 추정에 따른 성능 비교 (Baseline: 기본식별기, MO-cSNR: 정확한 SNR 이용시, MO-eSNR: 추정된 SNR 이용시)



<그림 5> SNR에 따른 신뢰도 함수들 (ASNR은 발화별 음성구간의 평균 SNR)

둘째, 신뢰도 함수의 천이 구간은 평균 SNR 값이 감소함에 따라 증가하게 된다. 이는 <표 5>에서 스케일 파라미터가 평균 SNR이 감소함에 따라 줄어들고 있다는 사실과 일치한다.

한편, 본 논문에서는 <그림 5>에 정리한 바와 같이 평균 SNR이 30, 20, 16, 12

그리고 8dB의 경우에 대해서만 고려하였다. 그러나 실제 관측되는 평균 SNR은 다양한 값을 가지게 된다. 이 경우 내삽(interpolation)을 이용하여 간단하게 주어진 평균 SNR에 대한 신뢰도 함수를 추정할 수 있을 것이다.

5. 결 론

본 논문에서는 HMM 기반 화자식별 문제의 성능 개선을 위하여, GMM에 성공적으로 적용되었던 관측 신뢰도의 개념을 도입하였다. HMM 디코더는 관측 신뢰도를 이용하여 관측 확률을 가중하도록 변형되었다. 또한 관측 신뢰도 함수의 최적화를 위하여 PSO를 통한 최적화 실험을 수행하였다. 제안된 방법은 ETRI 문맥종속 화자인식용 DB를 대상으로 검증하였는데, 실험 결과 제안한 방법이 화자식별 성능을 크게 향상함을 확인할 수 있었다. 또한 SNR의 추정값이 정확하지 않을 때, 성능향상의 정도가 저하될 수 있음을 확인하였으나, 여전히 제안한 방법은 성공적인 결과를 보였다.

본 논문에서는 깨끗한 음성 DB를 이용하여 화자 모델을 구축하는 경우에 대해, 관측 신뢰도 개념을 적용하였다. 향후 학습용 DB가 잡음으로 오염된 경우에 대해서도 적용할 수 있는 학습 알고리듬을 개발하고자 한다. 또한 본 논문에서는 매우 간단한 SNR 추정 방법을 사용하였는데, 좀 더 우수한 SNR 추정 방법을 구현하여 성능을 검토하고 분석하고자 한다.

참 고 문 헌

[1] J. Y. Kim et. al., "Modified GMM training for inexact observation and its application to speaker identification", *Speech Science*, Vol. 14. No. 1, pp. 163-175, 2007.

[2] 민소희, 김진영, 송민규, 나승유, "Particle swarm 기반 최적화 멤버쉽 함수에 의한 잡음환경에서의 화자인식 성능향상", *음성과학회지*, Vol. 14. No. 2, pp. 105-114, 2007.

[3] J. P. Campbell, "Speaker recognition: a tutorial", *Proceedings of the IEEE*, Vol. 85, No. 9, pp. 1437-1462, 1997.

[4] D. A. Reynolds, "An overviews of automatic speaker recognition technology", *Proc. ICASSP*, Vol. 4, pp. 4072-4075, 2000.

[5] B. Zhen, X. Wu, Z. Liu, C. Huisheng, "An enhanced RASTA processing for speaker identification", *Proc. ICSLP*, pp. 251-254, 2000.

[6] R. J. Mammone, X. Zhang, R. P. Ramachandran, "Robust speaker recognition, a feature-based approach", *IEEE Signal Processing Magazine*, Vol. 13, No. 5, pp. 58-71, 1996.

[7] D. Stephane, R. Christophe, "Robust feature extraction and acoustic modeling at multitel:

experiments on the Aurora databases", *Proc. Eurospeech*, pp. 1789-1792, 2003.

[8] A. Rosenberg et al., "Cepstral channel normalization techniques for HMM-based speaker verification", *Proc. ICSLP*, pp. 1835-1838, 1994.

[9] E. Mengusoglu, "Confidence measure based model adaptation for speaker verification", *Proc. 2nd IASTED International Conference on Communications, Internet and Information Technology*, pp. 408-411, 2003.

[9] M. Tariquzzaman, 김진영, 홍준희, "시청각 화자식별에서의 신뢰성 기반 정보 통합 방법의 성능향상", *말소리*, 제62호, pp. 149-161, 2007.

[10] R. Eberhart, J. Kennedy, "A new optimizer using particle swarm theory", *Proc. Sixth International Symposium on Micro Machine and Human Science,* pp. 39-43, 1995.

[11] M. Vondrasek, P. Pollak, "Methods for speech SNR estimation: evaluation tool and analysis of VAD dependency", *Radio Engineering*, Vol. 14, No. 1, pp. 6-11, 2005.

[12] E. Nemer, R. Goubran, S. Mahmoud, "SNR estimation of speech signal using subbands and fourth-order statistics", *IEEE Signal Processing Letters*, Vol. 6, No. 7, pp. 171-174, 1999.

접수일자: 2007년 11월 12일
게재결정: 2007년 12월 18일

▶ Md. Tariquzzaman
주소: 500-757 광주광역시 북구 용봉동 300번지 전남대학교
소속: 전남대학교 전자컴퓨터공학부
전화: 062)530-0472
E-mail: tareq_ict_iu@yahoo.com

▶ 김진영(Jinyoung Kim): 교신저자
주소: 500-757 광주광역시 북구 용봉동 300번지 전남대학교
소속: 전남대학교 전자컴퓨터공학부
전화: 062)530-1757
E-mail: beyondi@chonnam.ac.kr

▶ 나승유(Seungyu Na)
주소: 500-757 광주광역시 북구 용봉동 300번지 전남대학교
소속: 전남대학교 전자컴퓨터공학부
전화: 062)530-1757
E-mail:syna@chonnam.ac.kr