한국음향학회지 제24권 제7호 pp. 410~415 (2005)

화자간 변별력 최대화를 위한 혼합 모델 방식과 심볼 확률 가중함수에 관한 연구

A Study on the Mixed Model Approach and Symbol Probability Weighting Function for Maximization of Inter-Speaker Variation

진 세 훈*, 강 철 호*
(Se-Hoon Chin*, Chul-Ho Kang*)

*광운대학교 전자통신공학과

(접수일자: 2005년 6월 23일; 수정일자:2005년 8월 11일; 채택일자: 2005년 9월 12일)

최근 대부분의 화자확인 시스템은 패턴 인식 접근방식에 기인하고 있다. 패턴 분류기의 성능은 화자의 특징 파라미터를 어떻게 분류하는가 하는 데에 기인한다. 그 특징 파라미터를 잘 분류하기 위해서는, 화자간 변이를 최대화하고 특징 파라미터 간 거리를 효과적으로 측정하는 것이 매우 중요하다. 따라서, 본 논문에서는 개인 모델과 월드 모델을 동시에 배치함으로써 화자간 변이를 최대화 할 수 있는 개선된 혼합 모델 구조를 제안한다. 결정 과정 시 제안한 혼합 모델 방식을 사용함으로써 화자간 변별력을 최대화 할 수 있었다. 또한, 입력데이터에 대한 개인 모델과 월드 모델의 거리비율에 따라 심볼 확률 값을 가중하여 벡터 양자화 에러를 줄이는 가중치 함수를 제안 한다. 실험 결과, 이 두 가지 방법을 취함으로써 DCF (Detection Cost Function)를 2.37%에서 1.16%로 낮출 수 있었다.

핵심용어: 패턴 인식, 화자확인 시스템, 화자간 변별력, 혼합 구조의 모델, 심볼 확률 가중

투고분야: 음성처리 분야 (2,5)

Recently, most of the speaker verification systems are based on the pattern recognition approach method. And performance of the pattern-classifier depends on how to classify a variety of speakers' feature parameters. In order to classify feature parameters efficiently and effectively, it is of great importance to enlarge variations between speakers and effectively measure distances between feature parameters. Therefore, this paper would suggest the positively mixed model scheme that can enlarge inter-speaker variation by searching the individual model with world model at the same time. During decision procedure, we can maximize inter-speaker variation by using the proposed mixed model scheme. We also make use of a symbol probability weighting function in this system so as to reduce vector quantization errors by measuring symbol probability derived from the distance rate of between the world codebook and individual codebook. As the result of our experiment using this method, we could halve the Detection Cost Function (DCF) of the system from 2.37% to 1.16%.

Keywords: GPattern recognition, Speaker verification, Inter-speaker variation, Mixed model scheme, Symbol probability weighting

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

화자확인 시스템은 발성 내용에 따라 크게 문맥독립, 문맥종속, 문맥 제시형의 3가지 방식으로 나뉜다. 그 중, 문맥 독립 화자확인 시스템은 GMM (Gaussian Mixture Modeling) 기법을 기반으로 하여 활발히 연구되고 있으나, 계산량과 성능면에서 상용화 되는 단계에는 이르지 못한 실정이다[1]. 현재 활발히 연구되고 있는 문맥종속 화자확인 시스템의 많은 적용분야에서는, VQ/DHMM (Vector Quantization/Discrete Hidden Markov Model) 기법이 정확도 측면이나 계산량 측면에서의 우수성 때문에 채택되어 왔다. 그러나, 비록 VQ/DHMM 기반 화자확인 시스템이 CHMM (Continuous density Hidden

책임저자: 진 세 훈 (shchin@explore.kw.ac.kr)
139-701 서울시 노원구 월계동 447-1
광운대학교 전자통신공학
(전화: 02-940-5136; 팩스: 02-917-5136)

Markov Model)보다 적은 양의 훈련데이터에 대해 우수한 성능을 보여왔지만, VQ/DHMM 기법은 여전히 벡터 양자화 에러와 같은 많은 성능 열화 요인들을 내포하고 있다[2], [3], [4]. 지금까지 문맥종속 화자확인 시스템의 성능향상을 위하여 화자간 변별력 향상을 위한 다양한 접근방식으로는 먼저, 개인성 정보의 가중화을 통한 화자확인 성능 향상, 양자화 에러를 줄이기 위한 개선된 클러스터링 기법들과 같이 전처리 단계 및 분류 단계 등의 다양한 접근방식으로 연구되어 왔다[5-6]. 본 논문에서는 패턴 인식기의 분류 단계에서 소량의 학습데이터를 가지는 문맥 종속 화자확인 시스템의 성능열화 요인들을 찾아내어 해결점을 제시하고자 한다. 첫째로, 화자간 변별력을 최대화 함으로써 화자확인 시스템의 성능향상을 꾀하는 새로운 모델 구조를 제안하고, 두 번째로, 입력되는 테스트 벡터의 중요도를 자동으로 체크하여 프레임별 중요도를 가중하는 심볼 확률 가중 함수를 제안한다. 본 논문은 다음과 같이 구성되어 있다. 2장에서는 기존의 화자확인 시스템과 개선점 및 문제점들을 기술하고 3장에서는 제안하는 혼합 모델 구조와 심볼확률 가중 함수를 자세히 기술할 것이다. 4장과 5장에서는 모의 실험 결과 및 결론을 기술하겠다.

II. 기존의 화자 확인 시스템에서의 배경 모델

화자확인 시스템은 화자 모델 뿐아니라 다른 화자들을 표현하는 배경모델 또한 필요로 한다. 그 이유는 관측열 $O=\{O_1,O_2,\cdots,O_T\}$가 주어졌을 때 화자 S_i의 확률을 결정하는 우도비 $p(S_i|O)$가 직접 구해질 수 없기 때문이다. 이 우도비를 구하기 위하여 다음 식 1과 같은Bayes의 이론을 필요로 한다.

$$p(S_i|O)=\frac{p(O|S_i)p(S_i)}{P(O)} \tag{1}$$

그러나, 이 이론은 정확히 구해질 수 없는 두 가지 사전 확률을 가지고 있다. 즉, 화자 S_i가 존재하는 사전확률과$p(S_i)$와 관측열 O가 발생하는 사전확률, $P(O)$이다. 만일 $p(S_i)$가 모든 가능한 화자들의 집합에 대해 균일하게 분포되었다고 가정하고, 즉, $p(S_i)=p_S,\ \forall i\in I$ 그리고 $P(O)$를 조건부 확률의 합으로 쓴다면, 다음과 같이 나타낼 수 있다.

$$p(S_i|O)=\frac{p(O|S_i)p(S_i)}{\sum_{j\in I}P(O|S_j)P(S_j)}=\frac{p(O|S_i)}{\sum_{j\in I}P(O|S_j)} \tag{2}$$

그렇다면 식 2의 모든 부분은 계산되거나 근사화 될 수 있다. 물론, 모든 화자들의 완전집합 I를 구한다는 것은 실현 가능하지는 않다. 따라서, 식 2의 분모를 근사화하는 2가지 기술이 존재한다[7]. 한가지 방법은 화자 S_i와 비슷한 화자들의 그룹의 대표 값인 코호트 모델이라 불리는 유한집합 I를 추정하는 기술이다. 다른 하나는 많은 다른 화자들로부터 취득한 데이터로부터 학습된 모든 화자를 표현할 수 있는 가상의 화자모델 S를 가지는 월드모델링 기법이다. 문맥종속 화자확인에서 두 가지 기술가운데 어떤 방법이 더 좋은가는 확인되지 않았다. 하지만, 월드모델링은 코호트 모델링에 비해 계산상과 이론상 몇 가지 장점을 가지고 있다[6].

2.1. 월드모델을 배경모델로 사용하는 기존 화자 확인 시스템에서의 개선점

월드모델이나 군중모델과 같은 배경모델은 기존의 화자 확인 시스템에서 스코어 정규화 기법의 수단으로 사용된다. 하지만, 만약 배경모델이 스코어 정규화의 수단뿐만 아니라 화자간 변이를 최대화 시키는 방법으로도 사용된다면 시스템의 성능은 크게 향상될 것이다. 따라서, 본 연구에서는 화자 모델의 구조를 변형함으로써 화자간 변별력을 향상시키는 효과적인 방법을 제안한다.

2.2. 기존 거리계산에서의 개선점

그림 1은 기존 벡터 양자화 기법의 명백한 단점을 보여준다. 실제 화자와 사칭자간에는 그림과 같이 코드북 센터와 입력 특징 파라미터간의 거리차이가 존재한다. 그럼에도 불구하고, 양자화 결과는 똑 같은 인덱스를 얻게 되어 결국 똑같은 인식 결과를 가져오게 된다. 이러한 양자화 오차를 줄이기 위하여 가우시안 혼합 확률을 심볼확률로써 사용하는 CHMM (Continuous Hidden Markov Model)과 SCHMM (Semi-Continuous Hidden Markov Model)기법이 개발되었다. 하지만, 많은 계산량을 요구하는 특성상, 실시간 학습 및 인식을 필요로 하는 화자 확인 시스템에서 적용되고 있지 못하다. 따라서, 적은 계산량으로 VQ/DHMM기반 화자 확인 시스템

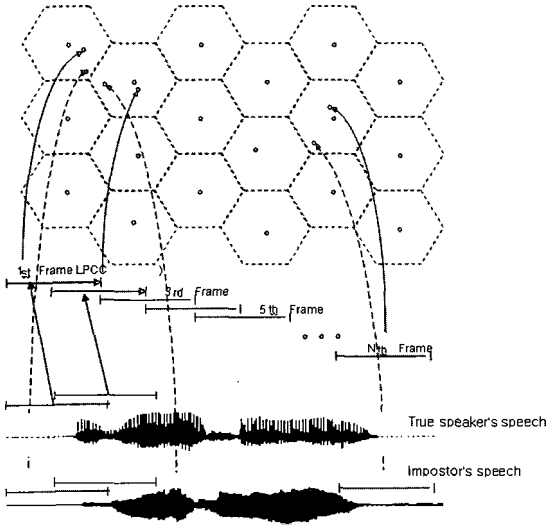

그림 1. 기존 VQ 과정의 단점
Fig. 1. Disadvantages of the conventional VQ procedure.

의 이러한 단점을 개선할 수 있다면 시스템의 성능은 크게 향상될 것이다.

III. 제안하는 화자 확인 시스템

3.1. 제안한 혼합 모델 기법

본 연구는 개인 모델과 월드 모델 모두를 포함하는 혼합된 모델에 대하여 비터비 (Viterbi) 연산을 수행하는 새로운 혼합 구조를 제안한다. 제안된 시스템에서의 로그 우도비 (log likelihood)는 그림 2와 3에서 보여주듯이 $\bar{p}(O(t)|\lambda_c) - \bar{p}(O(t)|\lambda_w)$ 의 차로써 계산된다. 여기서,

$$\bar{p}(O(t)|\lambda_c) = \sum_t \log(p(o_t|\lambda_c))$$

$$, b_j(o_t) = 0, \quad if\ o_t\ is\ quantized\ to\ C_w \tag{3}$$

$$\bar{p}(O(t)|\lambda_w) = \sum_t \log(p(o_t|\lambda_w))$$

$$, b_j(o_t) = 0, \quad if\ o_t\ is\ quantized\ to\ C_c \tag{4}$$

이렇게 함으로써, 화자의 특징을 잘 나타내는 중요한 심볼이 많을수록 $\bar{p}(O(t)|\lambda_c)$ 의 확률 값은 높아지게 되고, $\bar{p}(O(t)|\lambda_w)$ 의 확률 값은 반대로 낮아지게 된다. 따라서, 실제 화자 입력음성의 정규화된 스코어는 중요하지 않은 프레임의 심볼확률을 제거함으로써 상대적으로 증가하게 되고, 반대로 사칭자 입력음성의 스코어는 줄어들게 된다. 결국, 제안한 혼합 구조에서의 디코딩은 실제 화자의 입력 음성일 경우 LLR을 높여주고 사칭자의 음성일 경우 LLR을 더 낮춰줌으로써 화자간 변이를 극대화 시키는 효과를 나타내게 된다.

3.2. 제안한 심볼 확률 가중기법

기존의 화자확인 시스템에서는 모든 음성 프레임 $O = \{O_1, O_2, \cdots, O_T\}$들은 스코어과정에서 똑같은 중요도를 가진다. 즉, 양자화된 모든 시퀀스의 사전 정보도 VQ와 EM알고리즘이 수행되는 동안 고려되지 않으며, 인증과정에서, 화자에 대한 어떠한 사전 정보도 포함하지 못하고, 단순히 식 5와 같이 월드모델과 화자모델 사이의 로그 우도비의 평균값으로 정규화된다.

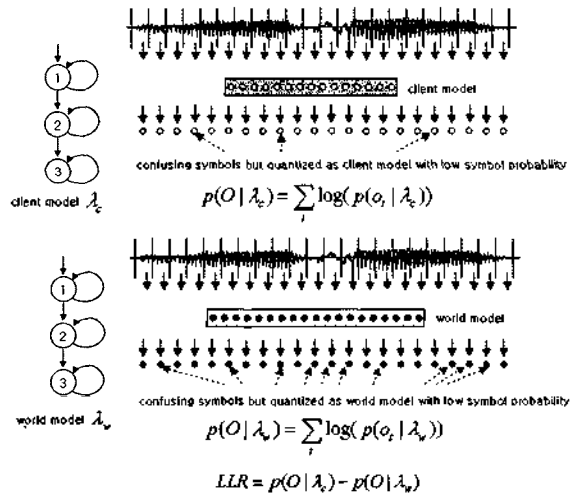

그림 2. 기존의 월드 모델을 사용한 스코어 정규화 과정
Fig. 2. Conventional score normalization procedure using world model.

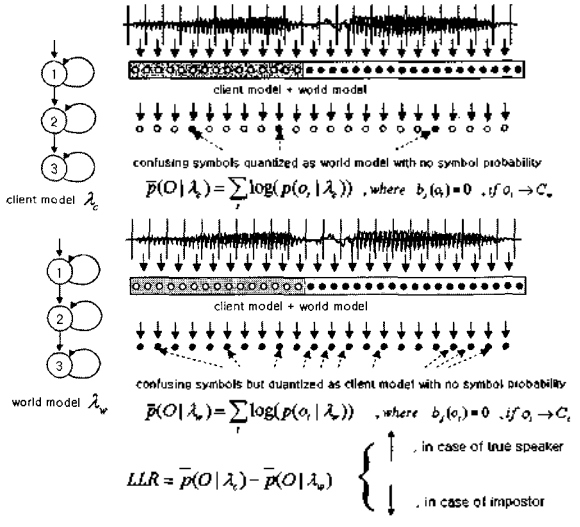

그림 3. 화자간 변별력 최대화를 위한 제안된 스코어 정규화 과정
Fig. 3. The proposed score normalization procedure for maximizing inter-speaker variation.

$$P(O)=\frac{1}{T}\sum_{t=1}^{T}[\log(o_t\mid\lambda_c)-\log(o_t\mid\lambda_w)] \tag{5}$$

여기서, T는 입력 시퀀스의 총 길이이며, o_t는 시간 t에서의 특징벡터이고 λ_c와λ_w는 각각 개인모델과 월드 모델을 의미한다. 한편, 각 음성 프레임들은 같은 중요도를 갖지는 않는다. 예를 들어 그림 4에서 보듯이, 1번 특징 벡터는 2번 보다 더 큰 화자 특징을 포함한다. 그러나 1번과 2번 특징 벡터는 모두 월드 모델로 양자화 된다. 1번과 같은 화자 특징을 많이 포함하는 중요 특징벡터는 월드 모델로 양자화 되어서는 안 된다.

그러므로, 이와 같은 문제를 해결하기 위해 각 테스트 벡터의 중요도를 측정하여 프레임별 가중치, ω_t를 이용한 다음과 같은 새로운 계산법을 제안한다 .

단계 1〉 k 개의 가장 가까운 월드 코드북을 찾고 거리의 평균값을 구한다 :

$$d_{x_t w_i}=d(x_t,\ C_{w_i}) \tag{6}$$

여기서, $d(*)$는 유클리디안 거리이며, x_t는 t번째프레임의 특징파라미터 이고, C_{w_i}는 i번째 월드 코드북의 중심값이다.

k 개의 가장 가까운 $d_{x_t w_i}$를 구한후,

$$d_{x_t w}=\frac{\sum_{i=1}^{k} d_{x_t w_i}}{k} \tag{7}$$

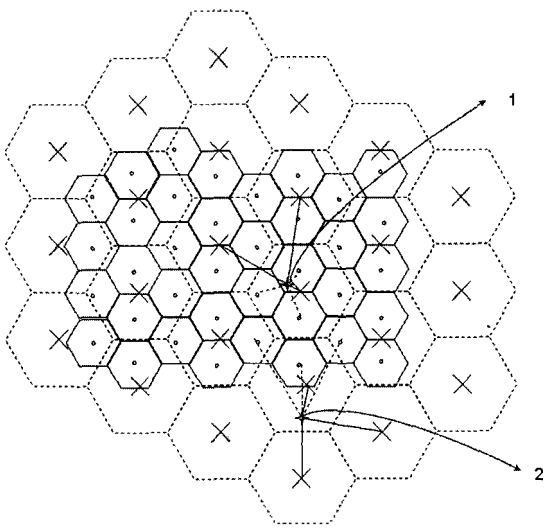

그림 4. 양자화 과정에서의 개인과 월드 코드북간의 관계
Fig. 4. Relationship between world and individual codebook in quantization procedure.

단계 2〉 k 개의 가장 가까운 개인 코드북을 찾고 거리의 평균값을 구한다 .

$$d_{x_t s_i}=d(x_t,\ C_{s_i}) \tag{8}$$

여기서, C_{s_i}는 i번째 개인 코드북의 중심값이다.

k 개의 가장 가까운 $d_{x_t s_i}$를 구한후,

$$d_{x_t S}=\frac{\sum_{i=1}^{k} d_{x_t s_i}}{k} \tag{9}$$

본 실험에서는 k=3에서 최적의 성능을 나타내었다.

단계 3〉 심볼 확률 가중치 ω_t는 0에서 1의 값을 갖으며 개인 코드북과의 거리 평균 $d_{x_t S}$가 작을수록 월드 코드북과의 거리 평균 $d_{x_t w}$가 클수록 가중치 ω_t가 커지게 해야 하므로 다음과 같이 구한다 :

$$\omega_t=\exp\left(-\frac{d_{x_t S}}{d_{x_t w}}\right) \tag{10}$$

단계 4〉 ω_t를 이용하여 가중된 비터비 연산을 수행한다 :

$$\begin{aligned}\tilde{\mathrm{b}}_j(\mathrm{O}_t)&=\omega_t\times \mathrm{b}_j(\mathrm{O}_t)\\ \bar{\delta}_t(j)&=\max_{1\le i\le N}\left[\bar{\delta}_{t-1}(i)+a_{ij}\right]+\tilde{\mathrm{b}}_j(\mathrm{O}_t)\end{aligned} \tag{11}$$

여기서, $b_j(O_t)$는 j 상태에서의 심볼관측 확률이며, a_{ij}는 i상태에서 j상태로 천이할 상태 천이확률이다.

단계 5〉 변형된 모델 $\bar{\lambda}_c$, $\bar{\lambda}_w$를 사용하여 프레임 별로 가중된 로그 우도비를 정규화 한다:

$$P(O)=\frac{1}{T}\sum_{t=1}^{T}\left[\log(o_t\mid\bar{\lambda}_c)-\log(o_t\mid\bar{\lambda}_w)\right] \tag{12}$$

IV. 실험 및 결과고찰

본 논문에서 사용된 월드모델을 구성하기 위한 데이터

베이스는 연령별 (10대에서 40대까지) 남녀 각 100명씩 4가지 단어로 미리 구성 되었다. LBG (Linde, Buzo, Gray) 알고리즘에 의해 256 사이즈의 월드 코드북을 생성하였으며, EM (Expectation-Maximization) 알고리즘에 의해 5상태의 HMM모델을 생성하였다[8]. 그림 5는 코드북 사이즈에 따른 EER (Equal Error Rate)의 변화를 나타낸다. 본 논문에서는 계산량 대비 성능면에서 가장 우수한 256 사이즈를 사용하였다. 제안한 알고리즘을 증명하기 위한 테스트 데이터베이스의 구성은 다음과 같다. 사용된 단어는 "안녕하세요", "다녀왔습니다", "문열어 나야", "열려라 참깨"의 총 4가지 단어로 각 단어는 남녀 각 50명의 화자로부터 6개월에 걸쳐 매달 20번씩 수집하였다. 모든 음성은 16kHz의 주파수로 샘플링 되었고 프레임 분석 구간은 20msec이며 1/3 중첩하였다. 프리-엠퍼시스된 음성은 20차의 LPC 켑스트럼으로 변환되었다. 각 단어별로 120회의 발성음이 인증실험되었으며, 자신을 제외한 총 11,880번의 발성음이 사칭실험에 참여하게 되었다. 실험은 기존의 VQ/HMM 시스템과 제안한 방식의 VQ/HMM 시스템에 대해 동일한 데이터베이스로 수행되었다.

모의 실험 결과는 그림 6에 나타나 있다. 그림 6은 화자인증 실험의 DET (Detection Error Trade-off) 곡선이다. DET 곡선은 시스템에 의해 나타날 수 있는 모든 가능한 성능값을 표현해준다. 그림에서 볼수 있듯이 제안한 코드북 구조와 심볼확률 가중기법은 모두 우수한 성능을 나타낸다. 최소화되어질수록 좋은 성능을 나타내는 DCF (Detection Cost Function) 는 본인인증오류율과 타인사칭 오류율 (각각 P_{miss}와 P_{fa})의 가중합의 형태로 식 12에 잘 나타나 있다.

$$DCF = (C_{miss} \times P_{miss} \times P_{true}) + (C_{fa} \times P_{fa} \times P_{false}) \quad (13)$$

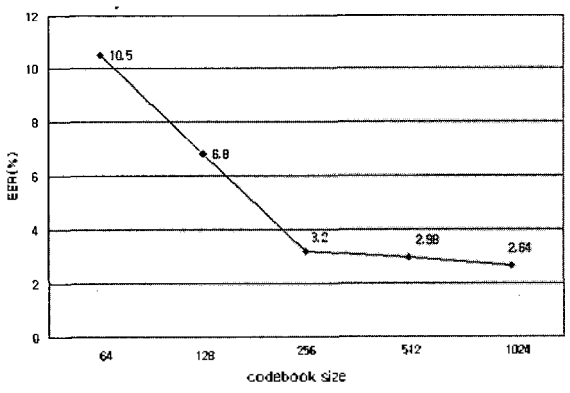

그림 5. 코드북 크기에 따른 평균 오차율
Fig. 5. EER according to codebook size.

표 1. 실험 결과에 대한 최소 DCF
Table 1. Minimum DCF for the experimental results.

시스템	Minimum DCF(%)
기존 시스템	2.37
제안한 혼합모델 방식	1.42
제안한 심볼가중 방식	1.58
제안한 시스템(혼합모델 + 심볼가중 방식)	1.16

여기서, C_{miss}는 본인인증오류 비용이며, C_{fa} 타인사칭오류 비용이다. P_{true}는 화자 모델의 사전확률이고, P_{false} ($P_{false} = 1 - P_{true}$)는 월드 모델의 사전확률이다. 최적의 DCF값은 각 곡선에 "○"으로 표시되어 있다. 본 실험에서는, Przybocki 와 Martin 의 실험에서와 같이 $P_{false} = 0.99$, $P_{true} = 0.01$, $C_{miss} = 10$, $C_{fa} = 1$로 정하였다[9]. 그림 6에서 우리는 제안한 화자간 변별력을 향상 시키는 2가지 방법들이 기존 VQ/HMM 시스템에 비하여 성능을 향상시킴을 확인할 수 있다. 즉, 표 I에서 보듯이 제안한 화자확인 시스템은 기존 화자확인 시스템에 비해 DCF를 2.37%에서 1.16%로 낮추는 것을 확인할 수 있다.

V. 결 론

본 연구에서는 화자간 변별력 향상을 위하여 혼합 구조를 갖는 모델과 벡터양자화 에러를 줄이기 위한 심볼 가중치 함수를 제안하였다. 화자모델과 월드모델을 합성

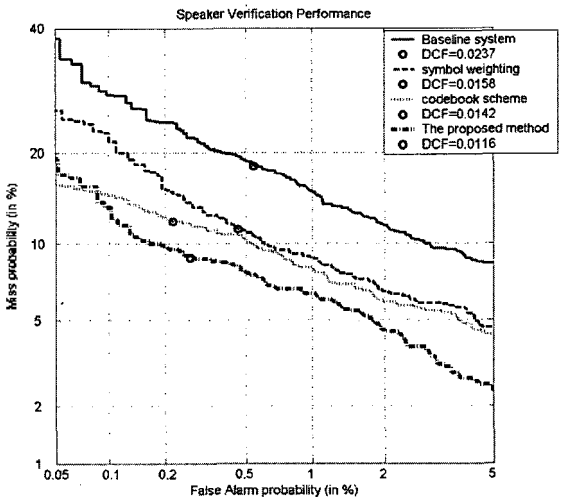

그림 6. 기존의 시스템과 제안한 시스템의 DET곡선 비교
Fig. 6. Comparison of DET curves for the conventional and proposed method.

한 혼합모델을 이용하여 비터비 연산을 수행함으로써, 화자본인의 입력음성에 대해서는 개인모델의 심볼에 대한 HMM 관측확률이 높아지게 되고, 상대적으로 월드모델의 관측확률은 낮아져 본인 인식율이 크게 증가한다. 반면, 타인에 의한 사칭의 경우 그 반대로 개인모델의 관측확률에 비해 월드모델의 관측확률이 증가하므로 오인식율은 줄어들게 된다. 결국, 혼합모델을 적용할 경우 기존의 방식에 비해 화자간 변별력을 향상시키는 효과를 보인다. 또한, 기존의 디코딩 과정이 전체 발음열에 대해 단순 평균값을 사용하는데 반해 제안한 심볼가중치 함수가 프레임별로 화자의 특징을 잘 나타내는 심볼에 큰 가중치를 줌으로써 VQ에러를 감소시키는 효과를 보여주었다. 문맥종속 화자확인 시스템에서 실험한 결과 제안한 방식들은 화자 모델과 월드모델을 효과적으로 구분함으로써 인식성능을 크게 향상시켰다.

참 고 문 헌

1. 이윤정, 서창우, 강상원, 이기웅, "화자식별을 위한 강인한 주성분 분석 가우시안 혼합 모델", 한국 음향 학회지, 22 (7), 519-527, 2003.
2. Christoph Neukirchen, Jörg Rottland, Daniel Willett, Gerhard Rigoll, "A continuous density interpretation of discrete HMM systems and MMI-neural networks", IEEE Trans. Speech Audio Processing, 9, 367-377, May 2001.
3. Matsui T., Furui, S., "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's", IEEE Trans. Speech Audio Processing, 2, 456-459,July 1994.
4. Zhiyuan He, Qixiu Hu, "A speaker identification system with verification method based on speaker relative threshold and HMM", 6th International Conference on Signal Processing, 1, 488-491, Aug. 2002.
5. Se-hyun Kim, Gil-Jin Jang, Yung-Hwan Oh, "Improvement of speaker recognition system by individual information weighting", ICSLP2000, 6, China, 1017-1020, Oct. 2000.
6. 정희석, 강철호, "문장종속 화자확인 시스템을 위한 개선된 군집화 알고리즘에 관한 연구", 한국음향학회지, 23 (7), 548-553, 2004
7. Reynolds, D.A., "Comparison of background normalization methods for text independent speaker verification", Proceedings of the European Conference on Speech Technology, 963-966, 1995.
8. Y. Linde et al., "An algorithm for vector quantizer design", IEEE Trans. Commun., COM-28, 84-95, Jan. 1980.
9. A. Martin, G. Doddington, T. kamm, M. Ordowski, and M. przybocki, "The det curve in assessment of detection performance", Proceedings of the European Conference on Speech Technology, 4, 1895-1898, 1997.

저자 약력

• **진 세 훈 (Se-Hoon Chin)**

1999년 2월: 광운대학교 전자통신공학과 (공학사)
2001년 2월: 광운대학교 전자통신공학과 (공학석사)
2001년 3월~현재: 광운대학교 전자통신공학과 박사과정

• **강 철 호 (Chul-Ho Kang)**

1975년 2월: 한양대학교 전자공학과 공학사
1979년 2월: 서울대학교 전자공학과 공학석사
1988년 2월: 서울대학교 전자공학과 공학박사
1982년~현재: 광운대학교 정교수