시청각 화자식별에서 신뢰성 기반 정보 통합 방법의 성능 향상

Md. Tariquzzaman(전남대), 김진영(전남대), 홍준희(경원대)

<차 례>

<Abstract>

Improvement of Reliability based Information Integration in Audio-visual Person Identification

Md. Tariquzzaman, Jinyoung Kim, Joonhee Hong

In this paper we proposed a modified reliability function for improving bimodal speaker identification(BSI) performance. The convectional reliability function, used by N. Fox[1], is extended by introducing an optimization factor. We evaluated the proposed method in BSI domain. A BSI system was implemented based on GMM and it was tested using VidTIMIT database. Through speaker identification experiments we verified the usefulness of our proposed method. The experiments showed the improved performance, i.e., the reduction of error rate by 39%.

* Keywords: Multi-modal speaker recognition, Late integration, Reliability, PSO.

1. 서 론

화자인식(speaker recognition)은 발화자의 음성정보를 이용하여 발화자를 식별(identification) 또는 확인(verification)하는 기술이다. 화자인식 기술은 유선통신망, 인터넷망 그리고 모바일 서비스 환경에서 보안을 위한 중요한 기술로 받아들여지고 있어, 활발하게 연구가 이루어지고 있다[2]. 뿐만 아니라 휴머노이드(humanoid) 로봇과 같은 지능형 인터페이스를 요구하는 시스템의 증가는 기계에 의한 자동 화자인식 기술을 요구하고 있다.

그러나 음성정보는 통신망을 통한 채널왜곡 및 주위환경의 잡음에 쉽게 노출되어 화자인식 성능이 저하되므로, 실제 서비스 환경에 성공적으로 적용되지 못하고 있다. 물론, 성능개선을 위해 왜곡에 강한 특징 파라미터 및 인식모델 적응(adaptation) 기술이 꾸준히 개발되고 있지만[2]-[8], 여전히 활용에 어려움이 있다. 이러한 문제를 해결하기 위한 한 방법이 멀티모달(multi-modal) 화자인식 기술이다[1][9]-[11]. 즉 음성 외에 부가 모달리티로서 얼굴, 지문, 홍채 등을 이용하는 것이다. 특히 얼굴과 음성 정보를 이용하는 바이모달(bimodal) 인식을 시청각 개인 인식(audio-visual speaker recognition)이라고 한다.

시청각 개인 인식을 포함한 멀티모달 화자인식에서 가장 중요한 문제는 다양한 모달리티 정보를 어떻게 통합하는 가이다. 기본적으로 초기통합(early integration)과 후통합(late integration)이 사용되고 있으며, 후통합 방법이 초기통합에 비해 쉽고, 다양한 모달리티에 적용 가능하기 때문에 널리 채택된다. 후통합 방법은 각 모달리티에 대한 분류기(classifier)의 출력(확률)들의 가중합(weighted sum)을 구하여, 최종적으로 판단을 하는 방법이다. 일반적으로 가중합 계산에 필수적인 가중인자(weighting factor)는 각 모달리티 정보의 신뢰성(reliability)에 의하여 결정되며, 신뢰성은 입력 정보의 왜곡량을 측정하거나, 개인별 관측 확률의 통계적 성질에 의하여 결정된다[1].

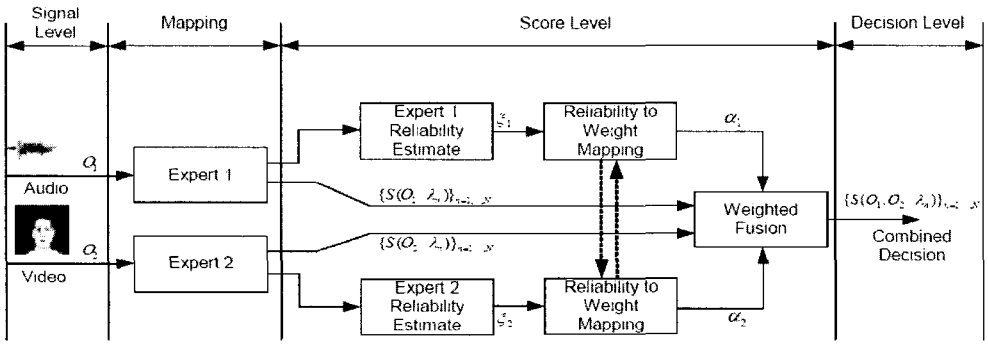
본 논문에서는 음성과 얼굴 정보를 이용한 바이모달(bimodal) 화자식별 문제에서 신뢰성 기반 통합 방법의 향상에 대하여 기술한다. 신뢰성 측정은, 신호의 왜곡을 측정하는 것이 어렵기 때문에, 관측 확률의 통계적 성질에 기반 한 방법을 사용한다. 그러나 지금까지 알려진 신뢰성 측정 방법들의 문제점은 인식기의 성능 향상을 위하여 최적화 할 여지가 부족하다는 점이다. 즉, 관측 확률들의 엔트로피, 또는 관측 확률들의 순위 정보에 기반하고 있으며, 신뢰도 변환함수가 고정적이어서 성능 향상이 제약적이다[1]. 본 논문에서는 기존의 신뢰성 변환 함수에 최적화 변수를 도입하고, 최적화 이론에 따라 인식률이 최대화 되도록 변수 최적화를 시행하여, 화자인식기의 성능 향상을 이루고자 한다.

본 논문의 구성은 다음과 같다. 2장에서 시청각 화자인식 기준 시스템(baseline system)에 대하여 설명하고, 3장에서 기존 신뢰도의 문제점을 도출하고, 최적화를

위한 변수를 도입, PSO(particle swarm optimization)에 의한 최적화 방법에 대하여 설명한다. 4장에서는 실험 환경과 결과를 제시하고, 5장에서 결론을 맺는다.

2. 기준 시청각 화자인식 시스템

본 논문에서는 기준 시스템으로 최근에 발표된 참고논문[1]에서 제시된 신뢰성을 이용한 후통합 기반 시청각 화자인식 시스템을 사용한다. 다음 <그림 1>은 기준 시스템을 보여준다.



<그림 1> 신뢰성을 이용한 후통합 기반 시청각 화자인식 흐름도[1]

<그림 1>에서 보듯이 화자인식은 다음과 같은 순서에 의하여 결정된다.

- 첫째, 입력된 음성 및 얼굴 영상에 대해 각 모달리티를 위한 전문가(expert: classifier)에 의하여 모달리티별 그리고 개인 별 관측 확률을 계산하다.
- 둘째, 각 전문가의 관측 확률로부터 모달리티별 신뢰성 값을 추정한다.
- 셋째, 각 모달리티의 신뢰성 함수로부터 후통합을 위한 가중인자 값을 결정한다.
- 넷째, 결정된 가중인자를 이용하여 각 모달리티의 관측 확률 값으로부터 가중합을 구하고, 최종적으로 판단한다.

물론, 최종 판단은 가중합의 결과 가장 큰 확률을 보이는 화자로 결정되게 된다. 각 모듈에 대한 설명은 다음과 같다.

2.1 가우시안 혼합 모델 기반 분류기

본 논문에서 사용한 전문가는 가우시안 혼합 모델(Gaussian mixture model,

GMM) 기반 분류기이다. GMM은 알려진 바와 같이, 특징벡터 공간을 다수의 가우시안 분포의 가중합으로 표현하는 것으로써[1], 다음의 식으로 표현된다.

$$f_X(x) = \sum_{i=1}^{K} w_i N(x;\mu_i,\Sigma_i) \tag{1}$$

여기서 w_i는 가중 값, $N(\cdot)$는 정규분포, μ_i는 i번째 정규분포의 평균 그리고 Σ_i는 i번째 정규분포의 공분산행렬이다. GMM 모델의 파라미터들은 EM 알고리듬에 의하여 학습되어 진다.

한편 식 (1)의 GMM 모델 학습을 위한 특징 파라미터로는 음성신호의 경우 멜켑스트럼(mel-cepstrum) 그리고 얼굴영상에 대해서는 블록단위 이산코사인변환(discrete cosine transform, DCT)를 이용하였다. 멜켑스트럼 분석은 프레임별로 수행되어지며, DCT 분석은 블록 윈도우를 정하고 윈도우를 주어진 얼굴 영상에 대하여 스캔(scan)함으로써 이루어진다. 단, 주어진 영상에서 얼굴탐지는 자동으로 이루어졌는데, adaboost 기반 방법[12]을 사용하였다.

한편 각 모달리티 분류기의 결과는 로그 확률 값이며, $S(O_i|\lambda_n)$와 같이 정의한다. 이 때 O_i는 음성 또는 얼굴정보의 특징들을 나타내는데, $i=1$이면 음성신호 정보의 관측 특징 행렬이며, $i=2$이면 얼굴 영상의 특징 행렬이다. 또한 λ_n는 n번째 화자의 GMM 모델을 의미한다.

2.2 신뢰성 측정 및 시청각 정보 통합

신뢰성 측정은 크게 개별 신호 분석을 이용한 방법과, 관측확률들의 통계적 성질을 사용하는 방법이 있다. 개별 신호 분석의 대표적인 방법은 입력 음성신호의 신호대잡음비(SNR)를 측정하는 방법이 있다. 그러나 SNR을 정확하게 예측하는 것이 쉽지 않을 뿐 아니라, 영상 신호의 경우 적절한 방법이 없기 때문에 그 사용이 지양된다. 한편, 관측확률의 통계적 성질에 기반을 둔 방법은 관측확률들의 엔트로피 기반 방법, Min-Max 정규화에 기반한 rank 정보 기반 방법 등이 사용된다. 본 논문에서는 논문 [1]에서 성공적으로 이용한 rank 정보 기반 방법을 사용하였다. 이 경우 신뢰성 계산은 다음의 절차에 의하여 이루진다.

- 첫째, 모달리티별로 확률분포에 대해 MinMax 정규화를 수행한다. MinMax 정규화는 주어진 관측 심볼 열에 대하여, 최대 확률은 1로 그리고 최소 확률은 0으로 정규화 시키는 것이다. 즉,

$$\tilde{S}(O_i|\lambda_j) = \frac{S(O_i|\lambda_j) - MinP_i}{MaxP_i - MinP_i} \tag{2}$$

이고, 여기서 $MinP_i = \min_j S(O_i|\lambda_j)$이고 $MaxP_i = \max_j S(O_i|\lambda_j)$이다.

- 둘째, 모달리티별 정규화 된 확률분포를 이용하여 신뢰성 값을 rank 정보에 기반 하여 다음과 같이 계산한다.

$$\rho_i = \widetilde{MaxP_i} - \widetilde{Max2P_i} \tag{3}$$

여기서 $\widetilde{MaxP_i}$는 i번째 모달리티에서 가장 큰 정규화 된 확률을 의미하며, $\widetilde{Max2P_i}$는 두 번째로 큰 확률을 의미한다. 그런데 $\widetilde{MaxP_i}$는 항상 1이므로 위식은

$$\rho_i = 1 - \widetilde{Max2P_i} \tag{4}$$

과 같이 된다.

- 셋째, 신뢰성 값 ρ_i로부터 변환함수를 이용하여 다음과 같이 가중값을 결정한다.

$$\alpha_1 = \frac{\rho_1}{\rho_1 + \rho_2}, \quad \alpha_2 = 1 - \alpha_1 \tag{5}$$

- 넷째, 각 관측확률은 다음과 같이 위의 가중값을 이용하여 합쳐진다.

$$P(O|\lambda_j) = \alpha_1 \tilde{S}(O_1|\lambda_j) + \alpha_2 \tilde{S}(O_2|\lambda_j) \tag{6}$$

그러면 최종적으로 화자식별은 $j^* = \arg Max_j P(O|\lambda_j)$로 결정된다.

3. 최적화 인자 도입과 신뢰성 계산

3.1 최적화 인자 도입 배경 및 신뢰도 함수

위절에서 기술한 바와 같이, 최종 통합 관측확률은 식 (2)-(6)에 의하여 결정되게 된다. 특히 식 (4)로 주어지는 신뢰성 지수는 청각 및 시각 모달리티 정보를 통합함에 있어 중요한 역할을 한다. 그런데 식 (4)의 신뢰성 값은 정규화 된 관측 확률들로부터 기계적으로 구해지게 된다. 즉, 화자인식의 최종 목표함수를 인식률이라고 할 때, 인식률을 극대화하기 위한 어떤 최적화 인자가 존재하지 않는다. 즉, $\rho_i = f(\tilde{S}(O_i|\lambda_1), \tilde{S}(O_i|\lambda_2), ..., \tilde{S}(O_i|\lambda_n))$이다. 그러므로 신뢰성 함수에 최적화 인자를

도입하고 화자인식률이 극대화 되도록 최적화를 시킨다면, 시청각 화자인식의 성능을 향상시킬 수 있으리라고 판단된다. 신뢰성 함수, 식 (4)로 표현된 신뢰성 값에 대하여 두 가지의 경우를 생각해 볼 수 있다.

- 첫째, (Overestimation) 최적성능을 위해 신뢰성 값이 높게 추정된 경우: 이 경우 신뢰성 값을 낮추도록 신뢰도 함수를 조정한다.
- 둘째, (Underestimation) 최적 성능을 위해 신뢰성 값이 낮게 추정된 경우: 이 경우 신뢰성 값을 높여 주도록 조정한다.

위의 두 가지 경우를 처리할 수 있도록 최적화 인자를 도입해야 하는데, 결국 식 (4)를 고찰하여 볼 때, 신뢰도를 결정하는 인자는 $\widetilde{Max2P_i}$이므로, 최적화 인자가 도입된다면 $\widetilde{Max2P_i}$를 제어해야 한다. 단, 신뢰성 함수는 신뢰성 값의 범위가 $0 \le \rho_i \le 1$을 만족해야 하고, $\widetilde{Max2P_i}$이 0과 1사이에 존재해야 한다는 점을 고려하여 최적화 인자가 도입되어야 한다. 따라서 본 논문에서는 다음과 같이 신뢰성 함수를 변경하였다.

$$\rho'_i = 1 - (\widetilde{Max2P_i})^f \tag{7}$$

변형된 식 (7)에서 f는 최적화 변수이다. 위 식은 $\widetilde{Max2P_i}$의 범위를 고려하여 볼 때, 다음과 같은 의미를 갖는다.

- $0 \le f \le 1$: $\rho'_i < \rho_i$
- $f = 1$: $\rho'_i = \rho_i$
- $f > 1$: $\rho'_i > \rho_i$

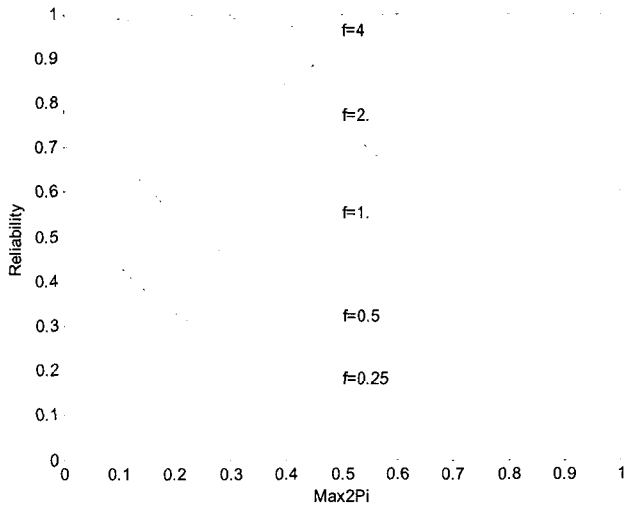
<그림 2>는 최적화 변수 f값에 따른 신뢰성 값을 나타낸 것으로, 위에서 설명한 물리적 의미를 시각적으로 보여 주고 있다.

일단, 식 (7)과 같이 최적화 함수가 도입되면, 목표함수를 결정하고 목표함수를 극대화하기 위한 최적화 방법을 적용하는 것이다. 본 논문에서는 이미 설명한 바와 같이 목표함수를 인식률(또는 식별률)로 정하였다.

3.2 최적화 목표함수

최적화 목표 함수를 화자식별 문제의 인식률로 정의하였다. 즉, 최적화 목표함수는 다음과 같은 식으로 정의 된다.

$$x(f) = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L_k}\delta(\arg Max_m(P_m(X_{kl})),k)}{K\sum_{k=1}^{K}L_k} \tag{8}$$



<그림 2> 최적화 변수 f따른 신뢰도 함수의 변화

위 식에서 $\delta(i,j)$는 $i=j$일 때 1이고 그렇지 않으면 0인 함수이다. X_{kl}는 k번째 화자의 l번째 음성시료이고, P_m은 주어진 시료에 대한 m번째 화자의 관측확률이다. 그리고 $\arg Max_m P_m$는 가장 큰 확률을 갖는 화자의 인덱스(index)를 의미한다. 그리고 P_m는 식 (6)으로 정의된다. 식 (8)로 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. 이에 대해서는 다음 절에서 설명한다.

3.3 PSO 기반 최적화

PSO 방법은 R. Eberhart와 J. Kennedy에 의하여 1995년 제안된 방법으로서, 새떼 또는 물고기 떼들의 먹이를 찾는 움직임을 모방하여 개발된 방법이다[13]. PSO는 초기 random한 해들의 모임으로 시작한다는 면에서 유전자 알고리듬과 유사하지만, 각 잠재적인 해들이 다시 random한 속도와 이전 잠재적인 해들의 결합으로 구성된다는 측면에서 다르다. 이 잠재적인 해들의 모임을 particle swam이라고 한다.

일반적인 문제로서 파라미터 P에 의하여 최적화 되어야 할 함수 $f()$가 있다고 하자. 그러면 PSO 방법은 다음과 같다.

1) Random하게 잠재적인 해들 $\{P_{i0}\}$를 결정한다.

2) 각 iteration j에 대하여 다음을 반복한다.

2-1) 각 P_{ij}에 대하여 $f(P_{ij})$를 구한다.

2-2) 최적 f값의 변화를 계산하고, 수렴한 경우 루프를 빠져나간다.

2-3) 각 i에 대하여 $\{0...j-1\}$에 대하여 가장 최적인 해를 저장한다. 이를 $pbest_{ij}$라고 하자.

2-4) 모든 $pbest_{ij}$를 대상으로 가장 최적인 해를 저장한다. 이를 $gbest_j$라고 하자.

2-5) 각 particle의 속도를 다음과 같이 계산한다.

$$v_{ij} = v_{ij-1} + c_1 r_1 (pbest_{ij} - P_{ij-1}) + c_2 r_2 (gbest_j - P_{ij}) \quad (9)$$

위 식에서 c_1과 c_2는 상수이며 r_1과 r_2는 random한 수이다.

2-6) 각 particle의 값을 갱신한다.

$$P_{ij} = P_{ij-1} + v_{ij} \quad (10)$$

3) $gbest_j$를 최적의 해로 결정한다.

PSO 방법은 수학적으로 해를 구하기 어려운 비선형 문제에 일반적으로 널리 응용되고 있다. 즉 최적화 함수 f가 비선형인 경우, PSO 방법은 해가 국부적인 최적값(local optimum)을 피하면서 전체 최적인 해(global optimum)를 구하는데 사용된다.

4. 실험 및 고찰

4.1 실험 데이터베이스 및 화자 식별 시스템

본 연구에서는 제안한 방법을 평가하기 위하여 화자식별 영역에 적용하였다. 실험은 시청각 기반 화자식별 실험이므로, 화자인식을 위한 데이터베이스가 필요하다. 본 논문에서는 시험용 데이터베이스로서 VidTIMIT 데이터베이스를 이용하였다. VidTIMIT 데이터베이스는 43명의 화자의 오디오 및 비디오 녹음으로 이루어있고, 녹음 문장은 NTIMIT 코퍼스(corpus)의 문장 셋에서 발췌한 것으로 서로 다른 10개의 문장이다[14]. 녹음은 일반 사무실 환경에서 이루어져 있어, 잡음이 자연스럽게 첨가 되어있다. 다음 <표 1>은 VidTIMIT 데이터베이스를 요약한 것이다.

한편 <표 2>는 시청각 화자식별 시스템의 주 모델인 GMM 모델 학습의 주 특징을 보인 것이다. 표에 보인 바와 같이 음성 모달리티의 모델링을 위해서는 17차의 멜캡스트럼을 사용하였고, 총 10 개의 가우시안 확률분포를 이용하였다. 또한 비디오 모달리티 모델링을 위해서는 8x8차의 DCT 변환을 이용하였으며, 화자별 가우시안 모델은 12개를 이용하였다.

<표 1> VidTIMIT 데이터베이스

모달리티	특징분석
<공통>	- 문장수 : 10문장 - 사람수 : 43명(남자: 24명, 여자: 19명)
음성	- 샘플링 : 16 bit 32 kHz sampling - 평균길이 : 4.25 sec
얼굴영상	- 샘플링 : 30프레임/sec - 영상크기 : 384x512

<표 2> GMM 기반 시청각 정보 학습

모달리티	특징분석	GMM 모델
음성	- 프레임 길이 : 20 msec - 파라미터 : 17차 멜캡스트럼	10 GMM
얼굴영상	- 입력영상 : 104x104 - 블록 윈도우 크기 : 8x8 - 윈도우 겹침 : 4	12 GMM

GMM 모델 학습을 위해서 음성의 경우 총 6개의 파일(첫 6개 파일)을 사용하였으며, 비디오 경우에는 문장별로 대표적인 얼굴 영상을 하나씩 추출하여 총 6개의 얼굴영상을 이용하였다.

4.2 신뢰도 함수 최적화 실험 및 결과

VidTIMIT 데이터베이스는 화자별 데이터의 개수가 10개로서 제안한 알고리듬 성능을 검증하기에는 데이터의 개수가 적다. 따라서 본 논문에서는 GMM모델 학습에 참여하지 않은 4개의 데이터를 6가지 조합으로 나누어 6번의 실험을 수행하여 결과를 보이고자 한다. 이렇듯 4개의 발화문장 셋을 두 개로 분류한 까닭은 최

적화 인자 f를 학습하기 위한 별도의 데이터 셋이 필요하기 때문이다. 물론, GMM 모델의 학습에 사용한 데이터를 사용하여, f값을 선정할 수도 있으나, 이 경우 식별률이 거의 100%에 이르러, 최적화 개념이 의미가 없어지기 때문이다.

<표 3> 실험별 데이터 셋 구분

	실험1	실험2	실험3	실험4	실험5	실험6
학습문장번호	7,8	9,10	7,9	8,10	7,10	8,9
검증문장번호	9,10	7,8	8,19	7,9	8,9	7,10

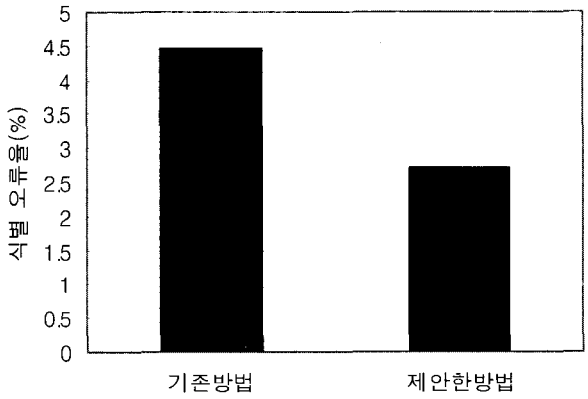
따라서 본 연구에서는 총 6회의 최적화 실험 및 검증을 하였으며, 결과적으로 검증을 위해 42*2*6=504개의 검증 데이터를 적용하게 되었다. 3.3에서 보였듯이 최적화는 PSO 방법을 이용하였는데, 200개의 particle을 이용하였고, 반복회수는 20회로 제한하여 실험하였다. 한편 학습 시에 200개의 particle 중 다수가 최적의 성능을 보일 수 있다. 본 논문에서는 이 경우 결과의 민감도(sensitivity)를 고려하여, 최적성능을 보이는 particle들의 평균값을 대표 값으로 결정하였다.

다음 <표 4>는 실험별 식별률과 최적화 인자의 값을 보인다. 표에서 알 수 있는 바와 같이 최적화 인자를 도입하여 신뢰성 함수를 최적화함으로써, 화작식별의 성능이 향상되고 있음을 확인할 수 있다. 최적화가 이루어진 경우 최적 f값은 평균치가 0.39로서 모두 1보다 작음을 알 수 있다. 이는 최적화 함수를 도입함으로써, 신뢰성 함수의 값들이 최적화 인자 도입이전에 비하여 작아졌음을 의미하고 있다(<그림 2>참조).

<표 4> 화자식별 성능(단위 %)

실험 차수	기저식별률 (학습데이터)	기저식별률 (검증데이터)	최적식별률 (학습데이터)	최적식별률 (검증테이터)	최적 f
1	95.3	95.3	98.8	96.5	0.285
2	94.2	95.4	96.5	96.5	0.374
3	95.3	97.6	97.7	100	0.284
4	95.3	94.2	97.7	96.5	0.476
5	96.5	96.5	97.7	98.8	0.316
6	96.5	94.2	96.5	95.4	0.599
평균	95.52	95.53	97.48	97.28	0.39

한편 <그림 3>은 검증데이터에 대하여 기존 신뢰성 함수에 의한 식별 오류율과 제안한 방법의 식별 오류율을 보여주고 있다. 오류율은 기존의 방법의 경우 4.47%이고, 제안한 최적화 방법의 경우 2.72%이다. 따라서 오류는 4.47%에서 2.72%로 약 39%정도의 성능향상이 이루어졌음을 확인할 수 있다.



<그림 3> 화자식별 오류율 비교

4. 결 론

본 논문에서는 멀티모달 화자식별 문제에서 신뢰성 기반 정보 통합의 향상에 관한 방법을 제시하였다. 기존의 신뢰성 변환 함수에 최적화 인자를 도입하고, PSO를 통한 최적화를 수행하여, 화자식별의 성능을 향상시킬 수 있었다. 제안된 방법은 VidTIMIT 데이터베이스 사용한 실험을 통해 검증하였다.

향후 좀 더 다양하고 많은 데이터베이스를 대상으로 검증하여, 실험의 통계적 신뢰도를 확보하고자 한다. 또한 본 논문에서는 화자식별 분야에 적용하여 성능을 검토하였는데, 화자확인 분야에도 적용하여 제안한 방법을 검증하고자 한다. 또한 신뢰성을 GMM 또는 HMM(hidden Markov model)에 적용하는 연구를 수행하여, 잡음환경에서 강인한 인식이 가능한 모델탐색에 대하여 탐구하고자 한다.

참 고 문 헌

[1] N. A. Fox, *Audio and video based person identification,* Ph.D. Thesis, University College Dublin, 2005.

[2J P. Campbell, "Speaker recognition: a tutorial", *Proceedings of the IEEE,* Vol. 85, No. 9, pp. 1437-1462, 1997.

[3] Z. Bin, W. Xihong, L. Zhimin, C. Huisheng, "An enhanced RASTA processing for speaker identification", *Proc. ICSLP*, pp. 251-254, 2000.

[4] 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.

[5] D. Stephane, R. Christophe, "Robust feature extraction and acoustic modeling at multitel: experiments on the Aurora databases", *Proc. EUROSPEECH,* pp. 1789-1792, 2003.

[6] A. Rosenberg, C.-H. Lee, F. K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification", *Proc. ICSLP*, pp. 1835-1838, 1994.

[7] J. Anguita, J. Hernando, A. Abad, "Improved Jacobian adaptation for robust speaker verification", *IEICE Transactions on Information and Systems,* Vol. E88-D, No. 7, pp. 1767-1770, 2005.

[8] 김유진, 정재호, "화자 인식을 위한 GMM기반의 이중 보상 구조", *말소리*, 제45호, pp. 93-105, 2003년.

[9] M. T. Chan, "HMM-based audio-visual speech recognition integrating geometric and appearance-based visual features", *Proc. IEEE Fourth Workshop on Multimedia Signal Processing*, pp. 9-14, 2001.

[10] A. Adjoudani, T. Guiard-Marigny, B. L. Goff, L. Reveret, C. Benoit, "A multimedia platform for audio-visual speech processing", *Proc. EUROSPEECH,* pp. 1671-1674, 1997.

[11] U. Dieckmann, P. Plankensteiner, T. Wagner, "SESAM: a biometric person identification system using sensor fusion", *Pattern Recognition Letters*, Vol. 18, No. 9, pp. 827-833, 1997.

[12] I Fasel, B. Fortenberry, J. Movellan, "A generative framework for real time object detection and classification", *Computer Vision and Image Understanding,* Vol. 98, No. 1, pp. 182-210, 2005.

[13] 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.

[14] C. Sanderson, *VidTIMIT dataset documentation,* http://users.rsise.anu.edu.au/ ~conrad/vidtimit/.

접수일자: 2007년 5월 22일
게재결정: 2007년 6월 10일

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

▶ 홍준희(Joonhee Hong) : 교신저자
주소: 461-70 경기도 성남시 수정구 복정도 산65번지 경원대학교
소속: 경원대학교 전자전기정보공학부
전화: 031) 750-8560
E-mail: hongpa@kyungwon.ac.kr