휴대용 화자확인시스템을 위한 배경화자모델 설계에 관한 연구*

A Study on Background Speaker Model Design for Portable Speaker Verification Systems

최 홍 섭**
Hong Sub Choi

ABSTRACT

General speaker verification systems improve their recognition performances by normalizing log likelihood ratio, using a speaker model and its background speaker model that are required to be verified. So these systems rely heavily on the availability of much speaker independent databases for background speaker model design. This constraint, however, may be a burden in practical and portable devices such as palm-top computers or wireless handsets which place a premium on computations and memory. In this paper, new approach for the GMM-based background model design used in portable speaker verification system is presented when the enrollment data is available. This approach is to modify three parameters of GMM speaker model such as mixture weights, means and covariances along with reduced mixture order. According to the experiment on a 20 speaker population from YOHO database, we found that this method had a promise of effective use in a portable speaker verification system.

Keywords: Portable Speaker Verification System, Background Speaker Model, GMM Speaker Model

1. 서 론

사람이 발성한 음성에는 언어적 의미 외에도 화자간의 음향적 차이를 포함하고 있다. 가장 주된 차이는 화자의 출신지역, 교육수준, 모국어 사용여부에 따라 다르며, 화자의 감정 및 건강상태 그리고 성도의 물리적 특성 또한 음향적 특성으로 나타난다. 이들 중 많은 부분이 화자에 종속적인 것이므로 이를 이용하여 화자를 구분하는데 사용할 수 있다. 아직까지 이들 정보가 화자마다 어느 정도까지 독특한 특징을 나타내는지 알려지지 않았지만 화자간의 변이가 화자 내의 변이에 비해 상대적으로 크다는 성질을 이용하여 어느 정도의 신뢰성을 가지고 화자를 구분할 수 있는데 이러한 화자간의 변이를 이용하여 발성한 사람의 진위를 알아내는 것을 화자확인(speaker verification)이라 한다.

* 이 논문은 2002학년도 대진대학교 학술연구비지원에 의한 것임.
** 대진대학교 이공대학 전자공학과

일반적인 화자확인 시스템에서는 확인 받고자 하는 화자인 의뢰인이 발성한 음성신호와 함께 의뢰인의 신원을 밝히는 ID를 입력으로 받아들인다. 확인시스템은 입력으로 들어온 ID에 해당하는 화자모델에 대한 입력 음성데이터의 조건부 발생 확률인 likelihood 값을 계산하고 이를 임계치와 비교하여 likelihood 값이 임계치보다 크면 화자의 신분이 맞는 것으로, 그렇지 않으면 틀린 것으로 판정하게 된다. 이 방법은 입력으로 들어오는 발성음성이 화자의 성대 상태, 언어적 내용 그리고 발성된 음질에 따라 likelihood 값이 많은 변화를 하게 되어 임계치를 일정하게 유지하기에 어려움이 있다.

이러한 문제를 개선하는 방법으로 likelihood 값의 정규화가 있는데, 이는 화자확인시스템이 확인을 받으려는 화자의 모델에 대한 likelihood 값만이 아니라 그 화자의 배경화자모델에 대한 likelihood 값을 구한 후, 그들의 비율인 likelihood ratio의 값으로 화자에 대한 진위 판단의 근거로 사용하는 방법이다. 이러한 정규화방법은 배경화자에 대한 의뢰인의 모델 특성이 보다 변별력을 갖도록 하는 동시에 앞에서 말한 문제점인 발성 음성에 의한 영향에 덜 민감하게 되어 시스템이 안정화되고 인식률도 향상시키는 장점이 있다[1]. 따라서 화자확인 시스템에서는 어떠한 배경화자모델을 사용하느냐가 시스템의 인식 성능에 많은 영향을 주고 있음을 알 수 있다. 지금까지 이러한 배경화자 모델선정에 많이 이용되는 방법들로는 검증하고자 하는 화자와 가장 유사도가 높은 화자모델들을 배경화자 모델로 선정하는 MSC(Maximally Spread Close) 방법과 이와 반대로 가장 유사도가 작은 화자모델들을 배경화자모델로 택하는 MSF(Maximally Spread Far)방법이 있다[2]. 이들은 모두 데이터베이스에 있는 화자들로부터 배경화자를 구하는 화자기반 cohort 방법으로 많이 사용되고 있지만, 데이터베이스에 등록된 화자의 수와 음성데이터의 크기에 따라 시스템의 성능이 좌우되는 경향이 있다. 또한 최근에는 화자기반 cohort와는 달리 확인을 원하는 의뢰인 주위의 여러 화자들의 확률분포모델에서 의뢰인의 확률분포와 가까운 확률분포들만을 따로 골라서 새로운 가상의 화자모델을 합성하는 방법인 가상조합(virtually synthesized) 배경화자 모델방법이 제안되어 우수한 결과를 보여주고 있다[3].

화자확인 시스템은 이전에는 주로 서버급 컴퓨터를 기반으로 구축되었는데, 오늘날은 컴퓨터 성능의 향상과 무선통신의 발달에 힘입어 개인휴대용 단말기(PDA), 팜탑컴퓨터 그리고 휴대폰과 같이 개인이 이동 중에 사용할 수 있는 휴대용 단말기에서도 화자확인 시스템을 적용하여 활용하고자 하는 필요성이 대두되고 있다. 이러한 휴대용 화자확인 시스템에는 기본적으로 메모리 용량과 같은 하드웨어적인 제약조건으로 인해 서버급의 시스템에서 사용하던 배경화자 모델을 이용할 수가 없다. 즉, 화자기반의 배경화자 모델에서는 가능한 많은 수의 화자와 음성 데이터로 데이터베이스를 구성하는 것이 시스템의 성능을 향상시키는 주요 요건인데 휴대용 시스템에서는 충분한 메모리를 갖지 못하기에 이러한 방법은 적당치 않음을 알 수 있다. 따라서 휴대용 화자확인 시스템에서는 사용자가 본인의 화자모델을 등록하기 위해서 발성한 음성데이터만을 이용해서 자신의 화자모델 뿐만 아니라 배경화자모델 또한 구성하는 방법을 고안하게 되었다. 이러한 노력의 일환으로 HMM기반의 화자확인시스템에서 HMM 화자모델의 상태 수를 줄이거나, 상태의 순서를 바꾸는 방법으로 그 화자의 배경화자 모델을 구성하는 방법과 등록 음성에 적당한 잡음을 부가하거나, 구성된 화자모델의 매개변수에 어느 정도의 왜란(perturbation)을 부가하는 방법 등이 제시되었다[4].

본 논문에서는 위의 방법에 착안하여 GMM기반의 휴대용 화자확인 시스템에서 등록화자의 화자모델로부터 배경화자 모델을 구축하는 여러 방법들을 제시하고 성능을 실험에 의해 비교 검증하였다. 즉, 등록시 입력된 음성 데이터를 이용하여 먼저 의뢰인인 기준 화자의 GMM모델을 구한 후, GMM모델의 mixture 차수, 평균과 분산 그리고 가중치에 각각 변화를 주어 배경화자 모델로 사용하는 방법이다. 화자확인 시스템을 구축하고 YOHO 데이터베이스를 이용하여 실험한 결과 제시된 방법이 휴대용 시스템에서 충분한 가능성이 있음을 확인하였다.

2. 배경화자모델의 설계

배경화자모델은 일반적인 응용에서 명확히 기대되는 사칭자들의 모임으로 구성되는 것이 타당하며, 이때 기대되는 사칭자들이란 비슷한 음성의 특징을 가지거나 적어도 동성의 화자들로 구성될 것이다. 이런 생각을 기반으로 하여 기준화자와 가까운 특징을 갖는 화자들로 배경화자를 구성하는 것이 MSC방법이다. 반면에 전화 기반의 응용 예에서는 보다 넓게 분포된 사칭자들의 접근을 가정할 수 있는데, 예를 들면 남성이 여성 사용자를 사칭하는 경우이다. 기존의 시스템에서는 등록된 화자와 가장 유사한 화자들을 배경화자로 선택하기 때문에 보통의 경우에 시스템의 신뢰도를 확보할 수 있었지만 위와 같이 매우 다른 음성특징을 갖는 사칭자에게는 취약한 약점이 있다. 이런 현상이 발생하는 이유는 기준화자와 매우 다른 통계적 특징을 갖는 사칭자는 유사도 수식의 분모, 분자인 배경화자 및 기준화자 모델 양쪽 모두 잘 모델링이 안 되기 때문이다. 이런 경우를 고려해서 제안한 방법이 기준화자와 되도록 거리가 먼 화자들을 묶어서 배경화자를 구성하는 MSF방법이다[2].

2.1 MSC (Maximal Spread Close) 방법

음성 DB에서 등록용 데이터를 이용하여 모든 화자의 GMM모델을 만들고, 화자모델 사이의 거리를 계산한다. 모델 (λ_i, λ_j)과 훈련음성 (X_i, X_j)을 가진 화자 i, j 간의 거리는 다음 식(1)과 같이 정의한다.

$$d(\lambda_i, \lambda_j) = \log \frac{p(X_i \mid \lambda_i)}{p(X_i \mid \lambda_j)} + \log \frac{p(X_j \mid \lambda_j)}{p(X_j \mid \lambda_i)} \quad (1)$$

거리가 가장 가까운 화자 N 명(N>B, B는 최종적인 배경화자군의 크기)을 근접 배경화자군(close cohort)으로 선택하며, 최종 배경화자군 $b(i)$는 가장 근접한 화자 중에서 최대한 고루 분포된 화자들로 구성된다.

$C(i)$: 화자 i와 가장 근접한 화자 N 명으로 구성

$b(i)$: 최종적인 배경화자 군

(단계 1) $C(i)$에서 가장 근접한 화자를 $b(i)$로 이동, N = N-1, B′ = 1
(B′는 $b(i)$에 현재 있는 화자의 수).
(단계 2) 다음 조건을 만족하는 화자 c를 $C(i)$에서 $b(i)$로 이동,

$$c = \underset{c=C(i)}{\arg\max} \left\{ \frac{1}{B'} \sum_{j \in b(i)} \frac{d(\lambda_j, \lambda_c)}{d(\lambda_i, \lambda_c)} \right\} \tag{2}$$

$$N = N - 1,\ B' = B' + 1$$

(단계 3) 단계 (2)를 B′=B가 될 때까지 반복한다.

이렇게 하면 비슷한 화자를 중복해서 배경화자로 선택하는 것을 방지함으로써, 크기가 한정된 배경화자모델에 최적의 배경화자들을 선택할 수 있게 된다. 이와 같은 방법을 MSC (Maximally-Spread Close)라고 한다[2].

2.2 MSF (Maximal Spread Far) 방법

예상되는 사칭자군이 기준화자와 비슷하지 않은 화자들을 포함하고 있을 때, 배경화자 선택은 멀리 떨어져 있는 화자들을 일부 포함하고 있어야 할 것이다. 이 경우 배경화자군의 구성은 거리가 근접한 화자군 일부와 멀리 떨어진 화자군 일부를 혼합하게 되는데, 이때 가장 멀리 떨어져 있는 화자 N 명을 추출한 다음, 이중 고루 분포된 B/2 명을 배경화자의 일부로 선정한다. 선정방법은 다음의 알고리즘을 따른다[2].

$F(i)$: 화자 i로부터 가장 멀리 떨어진 화자들의 집합
$b(i)$: 최종적인 배경화자 군

(단계 1) $F(i)$ 중에서 가장 멀리 떨어져 있는 화자를 $b(i)$로 이동, N = N -1, B′ = 1
(단계 2) $F(i)$에서 다음 조건을 만족하는 화자 f를 $b(i)$로 이동, 이때 f는 다음 방법에 의해서 선정,

$$f = \underset{f \in F(i)}{\arg\max} \left\{ \frac{1}{B'} \sum_{j \in b(i)} d(\lambda_j, \lambda_F) * d(\lambda_i, \lambda_F) \right\} \tag{3}$$

(단계 3) 단계 (2)를 B′ = B/2일 때까지 반복한다.

2.3 가상 배경화자 방법

앞의 방법들은 모두 데이터베이스 내에 들어 있는 화자들 중에서 기준화자와의 유사도를 거리개념으로 계산한 다음, 거리의 원근에 따라 화자들을 추출하여 배경화자모델에 포함시키는 화자기반의 배경화자 모델 구성방법들이다. 이에 비해 좀더 기준화자와의 유사도를 높이기 위하여 화자모델을 그대로 사용하는 것이 아니고 화자모델을 구성하는 확률분포 성분들

중 기준화자와 거리가 가장 가까운, 또는 거리가 가장 먼 확률분포 성분들만을 추출하여 가상의 배경화자모델을 구성하는 방법이 제안되었는데, 이것이 가상 배경화자 모델이다[3].

2.4 휴대용 화자확인 시스템을 위한 배경화자모델 설계

휴대용 화자확인 시스템은 메모리의 제약으로 많은 음성 데이터베이스를 확보할 수 없으므로 사용자의 등록용 음성데이터를 사용하여 사용자 모델뿐만 아니라 그의 배경화자 모델을 구성해야 한다. 논문에서 사용한 화자모델은 문장독립 화자확인을 위하여 일반적으로 많이 사용되는 GMM (Gaussian Mixture Model) 모델이다. GMM 화자모델은 $\lambda = \{p_m, \mu_m, \Sigma_m\}$로 표시하며 모델 내의 각각의 매개변수는 다음과 같다. 여기서 m은 GMM의 차수로서 모델에 포함된 가우시안 분포의 개수를 말한다.

p_m : 가우시안 분포들에 대한 가중치
μ_m : 가우시안 분포들의 평균
Σ_m : 가우시안 분포들의 표준편차

위의 화자모델을 이용하여 휴대용 시스템에 사용할 배경화자를 설계하는 방법으로는 등록용 음성신호에 일정한 규모의 잡음을 부가한 다음에 이를 이용하여 배경화자모델을 구성하는 방법과 이미 구한 사용자의 기준모델의 매개변수에 왜란을 주어 사용자와 유사한 배경화자 모델을 구하는 방법 등이 있다. 본 논문에서는 후자의 경우로서 기준화자 모델의 가중치, 평균 그리고 공분산을 변형시켜서 사용하는 방법을 적용하여 실험하였다.

3. 화자확인시스템의 구성

GMM을 이용한 화자확인시스템의 구성은 아래의 그림 1과 같다. 화자확인은 발성된 음성으로부터 확인을 원하는 화자, 즉 의뢰인이 맞는지 또는 사칭자인지를 구분해 내는 것으로 의뢰인에 대한 초기등록이 요구된다. 먼저 본인임을 확인하고자 하는 의뢰인은 그림 1과 같이 입력 음성신호와 본인의 화자ID를 밝히게 되며, 확인시스템은 그 화자ID에 해당되는 화자모델과 배경화자모델을 참조하여 유사도를 구한 후, 이를 사전에 정해진 임계치와 비교하여 임계치보다 크면 본인과 일치라고 하고, 그렇지 않을 경우에는 불일치라는 판정을 내리게 된다.

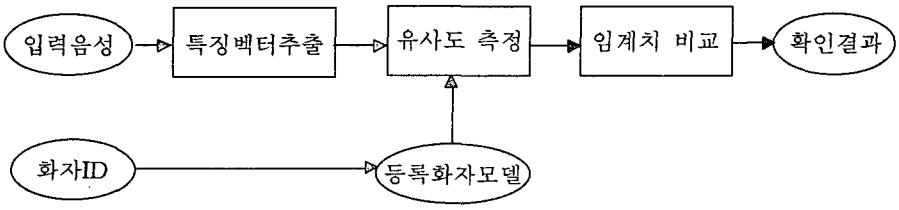

그림 1. 화자확인 시스템의 구성

4. 실험 및 결과

4.1 음성 DB

논문에서는 인식실험을 위하여 YOHO 표준 DB를 음성DB로 사용하였다. YOHO 음성DB는 화자인식 실험을 위해 만들어진 DB로 총 138 명의 화자로 구성되었다. 각각의 화자는 등록용으로 4 번의 세션을, 인식실험을 위하여 10 번의 테스트 세션을 갖고 녹음을 하였으며, 녹음은 고급 전화용 마이크를 사용하여 조용한 사무실 환경에서 하였다. 발성한 내용은 (35-72-41)과 같이 2 자리 숫자를 3 개씩 연이어 발음하는 형식으로, 등록 세션에는 24 개씩, 그리고 실험을 위한 테스트 세션에는 4 개씩의 녹음된 음성파일이 있다.

본 실험에서는 반복적인 실험을 위하여 계산량과 데이터량을 고려하여 대상 화자의 수를 20 명으로 제한하였다. 기준화자의 등록을 위하여 등록 세션에 있는 24 개 파일을 모두 합친 약 1 분 40 초 길이의 데이터를 사용하였고, 화자확인실험에서는 YOHO DB의 테스트세션 중 4 개의 세션을 뽑아 세션 내의 4 개의 파일을 합친 20 초 데이터를 만들어 의뢰인의 음성데이터로 사용하였다. 그리고 의뢰인이 아닌 나머지 음성데이터는 모두 사칭자 음성데이터로 보고 실험을 하였다. 따라서 20 명의 화자 각각에 대해 본인 데이터 4 개와 사칭자 데이터 76 개를 합한 80 번의 반복 실험을 하였다.

4.2 실험과정 및 성능평가 기준

선정된 음성데이터는 필터 $H(z)=1-0.95z^{-1}$를 이용하여 Pre-emphasis한 후, 30 msec 길이를 한 프레임으로 하는 Hamming창을 적용하여 20 msec를 중첩하여, 매 10 msec마다 특징벡터인 12 차의 LPC 켑스트럼을 추출하였다[5]. 화자모델에 사용한 GMM의 차수는 32차를 기본으로 하였으며, 각각의 가우시안분포의 공분산(covariance)값은 전체(full) 행렬을 사용하지 않고 대각선(diagonal)행렬로 제한하여 모델링하였다.

화자확인시스템의 성능을 평가하기 위하여 일반적으로 사용하는 EER을 쓰지 않고 시스템에서 의뢰인을 판단하는 기준이 되는 유사도 값인 log-likelihood ratio값을 큰 값부터 정렬하여 순위를 매겨서 평가하는 ranking point 값을 사용하였다[2]. 이는 선정된 화자마다 실험데이터가 4 개밖에 되지 않아 EER값을 구하는 것이 통계적인 신뢰도가 떨어지기 때문이다. ranking point 값은 한 명의 화자에 대해 본인 및 사칭자의 음성데이터를 합쳐 80 번의 실험결과로 나오는 log-likelihood ratio 값을 1 번부터 80 번까지 순위를 매긴 다음 의뢰인의 음성데이터 4 개의 순위들을 평균하여 구한다. 따라서 본 실험에서는 화자확인 시스템의 성능을 전체 화자 20 명에 대해 각각 4 개씩인 본인 데이터를 합한 80 개의 의뢰인에 해당하는 음성데이터에 대한 순위를 모두 합하여 평균을 구한 값으로 나타내었다. 예로써, ranking point가 1.0이라면 의뢰인 20 명의 데이터 80 개 모두 화자확인실험에서 1 순위를 차지한 것으로 인식률 100%에 해당하는 것과 같다. 이러한 평가방법은 똑같은 인식률을 갖는 경우에도 실패한 정도를 수치로 반영할 수 있는 장점이 있다. 즉, 본인 음성에 대한 확인이 거부된 횟수가 1 회일 경우에, 2 순위로 거부된 경우가 10 순위로 거부된 경우에 비해 성능이 좋다고 판단하므로, 똑같은 인식률을 보이는 경우에도 ranking point 값은 이를 수치로 구별할 수 있다.

4.3 실험결과

화자확인실험에서 유사도 값을 배경화자를 이용하여 정규화하는 경우의 장점을 알아보기 위하여, 비정규화한 유사도 값과 정규화한 유사도 값을 사용한 시스템의 인식성능을 ranking point 값으로 비교하여 보았다. 의뢰인의 화자모델만을 이용한 비정규화 유사도 값을 사용한 경우의 ranking point 값은 1.29인 반면, 정규화한 유사도 값을 사용한 경우에는 1.0으로 100%의 인식률로서, 성능이 확실히 좋아진 것을 확인할 수 있었다. 이때 사용한 배경화자모델은 일반적으로 많이 사용되는 MSC방법으로 5 명의 배경화자를 cohort로 선정한 경우이다.

표 1. 배경화자의 사용 유/무에 따른 휴대용 화자확인 시스템의 성능비교

배경화자 사용 유/무	배경화자 없는 경우	배경화자(MSC, 5명)
ranking point	1.29	1.0

아래의 표 2는 GMM모델의 차수를 조정하여 만든 휴대용 화자확인시스템의 배경화자모델의 예를 보인 것이다. 먼저 배경화자의 GMM모델의 차수, 즉, 모델에 포함된 가우시안 분포함수의 종류를 32 차에서부터 차례로 줄여서 실험한 결과 10 차 이상에서는 화자확인이 거의 되지 않는 결과를 보였으며, 여기서는 8차와 4차로 각각 줄인 배경화자모델을 사용한 경우의 결과만을 표로 비교하였다. 여기서 성능의 기준이 되는 것은 비정규화한 경우의 성능이므로 1.29보다 작은 값을 가져야만 정규화로서 역할을 하고 있다고 판단된다. 따라서 차수가 8인 경우는 비정규화한 경우인 1.29보다 큰 값을 보이므로 이 모델은 배경화자로서 의미가 없음을 알 수 있다. 결국 배경화자의 모델은 기준화자 모델과 어느 정도 이상의 거리가 있어야 배경화자로서의 역할을 하게 되는 것을 알 수 있다.

표 2. GMM차수를 감소시킨 휴대용 시스템의 배경화자모델들의 성능 비교표

배경화자모델	GMM(차수=8)	GMM(차수=4)	GMM(차수=8), 가중치 역순	GMM(차수=4), 가중치 역순
ranking point	1.46	1.19	2.32	1.14

그리고 GMM모델의 차수와 함께 가우시안 확률분포함수들에게 부가되는 가중치들의 변화를 크기 역순으로 재배열한 모델을 사용하는 경우에는 차수가 4인 경우는 1.14로 성능개선 효과가 있었으나, 차수가 8인 경우는 더욱 성능이 떨어지는 것을 볼 수 있다.

다음으로 GMM모델의 매개변수인 공분산과 평균값에 일정한 수준의 왜란을 인가하여 구성하는 휴대용 배경화자모델에 대한 성능을 알아보았다. 결과는 표 3에 보이듯이 차수가 8인 경우는 모두 비정규화한 경우에 비해 낮은 인식률을 보이므로 고려 대상이 아니고 차수가 4인 경우에는 공분산에 인가하는 왜란의 정도가 3%와 5% 모두 정규화 효과를 주는 것으로 나타났다. 그러나 이는 앞의 가중치 역순에 의한 경우에 비해 성능이 떨어지고, 왜란을 10% 이상을 인가하는 경우에는 전혀 배경화자로서의 역할을 하지 못함을 알 수 있었다. 여기서 5%의 왜란이란 난수발생기로 +0.05에서 -0.05까지를 발생시킨 후 이를 원래의 공분산 값에 합한 것을 말한다.

표 3. GMM의 공분산 값에 왜란을 주어 만든 휴대용 시스템의 배경화자모델의 성능비교

배경화자모델	GMM(차수=8), COV. 3%	GMM(차수=8), COV. 5%	GMM(차수=4), COV. 3%	GMM(차수=4), COV. 5%
ranking point	1.46	1.48	1.20	1.21

다음은 GMM모델의 평균값에 일정량의 왜란을 주어 만든 배경화자에 대한 성능을 표 4에 나타냈다. 표에서 보면 알 수 있듯이 GMM모델의 차수를 4로 하고 평균에 5%에서부터 30%까지 왜란을 줄 경우의 성능은 가중치나 공분산을 변형한 경우에 비해 ranking point가 상당히 낮고, 상대적으로 왜란의 정도의 차이에도 비교적 안정적인 성능을 보여주고 있다. 따라서 지금까지의 실험결과에 의하면 휴대용 화자확인시스템을 위한 배경화자 구성은 GMM의 차수를 4로 하면서 모델의 평균값에 일정한 왜란을 부가하여 만든 경우가 성능이 우수함을 알 수 있었다.

표 4. GMM의 평균값에 왜란을 주어 만든 휴대용 시스템의 배경화자모델의 성능 비교

배경화자모델	GMM(차수=4), MEAN 5%	GMM(차수=4), MEAN 10%	GMM(차수=4), MEAN 20%	GMM(차수=4), MEAN 30%
ranking point	1.15	1.16	1.14	1.11

5. 결론 및 향후 과제

본 논문에서는 휴대용 화자확인시스템을 위한 배경화자모델의 설계 방법의 여러 예를 실험을 통하여 실용 가능성을 확인하였다. 휴대폰과 PDA와 같은 휴대용 단말기는 메모리의 제약으로 인해 많은 데이터와 과도한 계산을 필요로 하는 응용프로그램을 내장하는 것이 어렵다. 따라서 화자확인기능을 이러한 휴대용 단말기에 이식하는 것은 기존의 서버급 컴퓨터에 적용하는 방법과 달라져야 한다. 특히 화자확인에서 배경화자의 구성은 가능한 많은 수의 화자와 많은 양의 음성 데이터베이스가 필요한데, 휴대용에서는 별도의 많은 화자들의 음성데이터를 확보하기 어려우므로 시스템을 사용하려는 화자의 등록용 음성데이터만을 갖고 등록화자의 배경화자를 인위적으로 생성하는 방법을 사용해야 한다. 논문에서는 GMM을 기반으로 하는 휴대용 화자확인 시스템의 배경화자를 GMM의 차수와 매개변수인 가중치, 공분산 그리고 평균을 인위적으로 변경시켜서 필요한 배경화자를 구성하는 방법을 각각 비교하였다.

실험에서 화자확인 성능은 ranking point를 사용하여 표시하였으며, 이는 동일한 %의 인식 성능이라도 세부적으로는 의뢰인의 유사도가 몇 번째 순위에 위치하고 있는 지를 수치로서 나타내주기 때문에 보다 자세한 알고리즘의 평가가 가능함을 알 수 있다. 휴대용 시스템에 사용하는 배경화자는 원래 기준화자의 GMM모델의 차수보다 낮은 값을 사용해야 하는데, 실험에서는 4 정도일 때, 정규화 효과가 나타남을 보였다. 그리고 매개변수의 변경에서는 가중치의 경우, 순서를 역순으로 재배열하여 만든 모델에서 ranking point 1.14로 변경하지 않은 경우에 비해 개선이 있었다. 공분산의 경우에는 변경시 부가하는 왜란의 정도에 따라 많은 변

화가 있었는데 일반적으로 10% 이상의 왜란에서는 전혀 기능을 하지 못하고 3-5%에서 정규화 효과가 나타났다. 그리고 평균의 경우는 위의 두 가지 매개변수에 비해 왜란의 정도에 덜 민감하고 성능도 비정규화일 때의 1.29에 비해 1.11−1.16 정도로 개선되었음을 알 수 있었다.

따라서 GMM을 기반으로 하는 휴대용 화자확인시스템에서는 등록용 음성데이터만을 이용하여 등록화자의 기준 화자모델을 추정한 후, 이를 바탕으로 GMM의 차수와 가중치, 공분산 그리고 평균 등을 적정 범위 내에서 변경하여 배경화자를 생성할 경우, 충분히 배경화자로써 유사도 값을 정규화하는 효과가 있음을 확인할 수 있었다. 이는 일반적인 배경화자 설계방식에서 많은 수의 화자들로부터 얻은 음성데이터를 사용해야하는 문제를 해결할 수 있는 방안의 하나로 충분한 가능성이 있다고 본다.

앞으로 과제로는 본 논문에서는 20 명의 화자만을 이용하였지만 전체 YOHO 데이터베이스의 138 명을 충분히 활용하는 실험과 휴대용 단말기에 대한 화자확인 시스템의 적용을 통해 실제 성능을 확인하는 과정이 필요하겠다.

참 고 문 헌

[1] Matsui, Tomoko & Sadaoki Furui. 1995. "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model." *Speech Communication*, Vol. 17, 109-116.

[2] Reynolds, Douglas A. 1995. "Speaker identification and verification using gaussian mixture speaker models." *Speech Communication*, Vol. 17, 91-108.

[3] Isobe, Toshihiro & Junichi Takahashi. 1999. "Text-independent speaker verification using virtual speaker based cohort normalization." *Eurospeech*.

[4] Siohan, Olivier, Chin-Hui Lee, Arun C. Surendran & Qi Li. 1999. "Background model design for flexible and portable speaker verification systems." *ICASSP*.

[5] O'Shaughnessy, Douglas. 1990. *Speech Communication-Human and Machine*. 345-346, Addison Wesley.

접수일자: 2003. 4. 29.
게재결정: 2003. 5. 29.

▲ 최 홍 섭
경기도 포천군 포천읍 선단리 (우: 487-711)
대진대학교 이공대학 전자공학과
Tel: +82-31-539-1903 Fax: +82-31-539-1900
E-mail: hschoi@daejin.ac.kr