휴대폰음성을 이용한 화자인증시스템에서 배경화자에 따른 성능변화에 관한 연구

A Study on the Performance Variations of the Mobile Phone Speaker Verification System According to the Various Background Speaker Properties

최 홍 섭*
Hong Sub Choi

ABSTRACT

It was verified that a speaker verification system improved its performances of EER by regularizing log likelihood ratio, using background speaker models. Recently the wireless mobile phones are becoming more dominant communication terminals than wired phones. So the need for building a speaker verification system on mobile phone is increasing abruptly. Therefore in this paper, we had some experiments to examine the performance of speaker verification based on mobile phone's voices. Especially we are focused on the performance variations in EER(Equal Error Rate) according to several background speaker's characteristics, such as selecting methods(MSC, MIX), number of background speakers, aging factor of speech database. For this, we constructed a speaker verification system that uses GMM(Gaussin Mixture Model) and found that the MIX method is generally superior to another method by about 1.0% EER. In aspect of number of background speakers, EER is decreasing in proportion to the background speakers populations. As the number is increasing as 6, 10 and 16, the EERs are recorded as 13.0%, 12.2%, and 11.6%. An unexpected results are happened in aging effects of the speech database on the performance. EERs are measured as 4%, 12% and 19% for each seasonally recorded databases from session 1 to session 3, respectively, where duration gap between sessions is set by 3 months. Although seasons speech database has 10 speakers and 10 sentences per each, which gives less statistical confidence to results, we confirmed that enrolled speaker models in speaker verification system should be regularly updated using the ongoing claimant's utterances.

Keywords: GMM(Gaussian Mixture Model), speaker verification, background speaker, MSC, MSF, mobile phone speech DB

1. 서 론

음성은 정보의 전달뿐만 아니라 보안이 필수적인 유비쿼터스 환경에서 중요한 생체인식 도구 중의 하나이다. 현재 사람에게 가장 쉽고 편리한 인터페이스는 음성으로, 이를 이용한 개인 인증 방

* 대진대학교 공과대학 전자공학과

법은 카드나 열쇠 보다 매우 편리하고, 분실위험이 전혀 없어 매우 안전하며, 손이나 다른 도구를 필요로 하지 않으므로 유비쿼터스 정보화시대의 중요한 인터페이스 기술로 자리매김하고 있다. 사람이 발성한 음성은 여러 가지 형태의 정보를 지니고 있다. 이것은 음성 신호가 전달하고자 하는 언어적 내용 외에도 화자간의 차이에 의한 음향적 특징을 포함하고 있기 때문이다. 이들 중 많은 부분이 화자에 종속적인 것이고 화자를 구분하는데 사용할 수 있는 정보이다. 아직까지 이들 정보가 화자마다 어느 정도까지 독특한 특징을 나타내는지 알려지지 않았지만 화자간의 변이가 화자내의 변이에 비해 상대적으로 크다는 성질을 이용하여 어느 정도의 신뢰성을 가지고 화자를 구분할 수 있다. 이러한 화자간의 변이를 이용하여 발성한 사람을 알아내는 것을 화자인증시스템이라 한다[1][2]. 일반적으로 화자인증 시스템에서는 검증하고자 하는 화자모델과 그의 배경화자모델을 이용하여 log likelihood ratio의 값을 정규화하고, 요구하는 화자의 임계값과 비교해서 화자의 인증을 수락할 것인지, 거절할 것인지를 결정하는 것이다. 이때 배경화자모델의 사용은 배경화자를 사용하지 않는 경우에 비해 전체 시스템의 성능을 안정화시키며, 인식률을 향상 시켜주고 있다[3].

본 논문에서는 이러한 화자인증 시스템이 휴대폰을 사용하는 환경에서의 성능을 측정하기 위하여 배경화자의 선정방법과 배경화자 수, 그리고 시간의 변화에 의한 인증성능의 변화 등을 실험에 의해 측정하였다. 이를 위해서 GMM(Gaussin Mixture Model)을 사용하는 화자인증 시스템을 구성하였고, ETRI에서 제작한 휴대폰 음성시료를 사용하여 인증실험을 하였다. 논문의 구성은 2 장에서는 화자인증 시스템의 구성에 대한 간단한 소개와 3 장에서는 음성DB의 소개, 4 장에서는 실험방법과 그 결과, 그리고 마지막 5 장의 결론 및 향후 과제 등으로 되어있다.

2. 화자인증 시스템의 구성

2.1 GMM(Gaussian Mixture Model)

GMM은 여러 개의 가우시안 확률밀도(Gaussian probability density) 함수들에 각각의 가중치를 준 다음, 이를 선형 결합함으로써 임의의 모양을 갖는 확률밀도 함수를 표현할 수 있다. 그리고 음성의 특징 파라미터 벡터의 확률분포는 화자마다 그 모양이 다르며, 이러한 확률분포를 GMM을 이용하여 모델링하여 인식하고자 하는 화자의 모델로 사용함으로써 화자인식에 이용할 수 있다. 그리고 지금까지의 실험 결과는 이것이 단지 가정이 아니라 사실임을 보여주고 있다. 다음은 GMM을 어떻게 정형화하고 그 파라미터는 어떻게 구성되어 있는지를 간략히 설명한다[5][6].

GMM의 혼합 확률분포는 M개의 가우시안 분포의 가중치 합으로 구성되며 다음과 같은 식으로 표시된다.

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{1}$$

식(1)에서 $\vec{x}$는 D차원의 랜덤 벡터이며 $b_i(\vec{x}), i=1,\ldots,M$은 성분 가우시안 분포이고 $p_i, i=1,\ldots,M$

은 결합 가중치(mixture weight)라고 불리며 각각의 가우시안 분포에 대한 가중치이다. 각 분포는 D차원의 가우시안 분포이며 식(2)와 같이 표현되며

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} \exp[-\frac{1}{2}(\vec{x}-\vec{\mu})' \sum_i^{-1}(\vec{x}-\vec{\mu})] \quad (2)$$

이때 $\vec{\mu}$는 평균 벡터이며, $\sum_i$는 공분산 행렬이다. 그리고 결합 가중치는 식(3)을 만족한다.

$$\sum_{i=1}^{M} p_i = 1 \quad (3)$$

가우시안 혼합분포는 위에서 기술한 파라미터들 즉 평균 벡터와 공분산 행렬 그리고 결합가중치에 의해 완전히 표현되며 식(4)와 같이 표현된다.

$$\lambda = \{p_i, \vec{\mu}_i, \sum_i\}, i=1,\ldots,M \quad (4)$$

2.2 화자인증 시스템의 결정논리와 배경화자 선정

화자인증은 발성된 음성이 원하는 화자, 즉 의뢰인인지 또는 사칭자인지를 구분해 내는 것으로 의뢰인에 대한 초기 등록이 요구된다. 시스템을 사용할 의뢰인들에 대한 화자모델이 사전에 모두 만들어지고, 또한 각각의 의뢰인에 대한 배경화자들을 코호트(cohort)방법으로 선정을 한다[4]. 다음으로 화자인증 시스템에 입력으로 들어온 음성데이터(X)를 사용하여 다음의 식과 같이 유사도(likelihood ratio) $\Lambda(X)$를 계산한다.

$$\Lambda(X) = \log p(X|\lambda_c) - \log p(X|\lambda_{\bar{c}}) \quad (5)$$

여기서 λ_c 는 의뢰인의 화자모델, $\lambda_{\bar{c}}$ 는 의뢰인의 배경화자모델이다.
다음으로 계산된 유사도 값을 정해진 문턱값 Θ와 비교하여 다음과 같이 인증을 수락 또는 거부하게 된다[5][10].

$$\begin{aligned} &\text{화자 인증 : } \Lambda(X) > \Theta \\ &\text{화자 거부 : } \Lambda(X) < \Theta \end{aligned} \quad (6)$$

유사도에 사용하는 배경화자를 이용할 때 고려해야 할 문제는 배경화자를 선택하는 방법과 배경화자의 수에 관한 것이다. 일반적으로 많이 사용하는 배경화자 선정방법으로는 등록화자 집단에서 화자간의 거리를 이용하여 선정하는 코호트(cohort) 방법을 사용하는데 이에는 MSC(Maximal Sperad Close)와 MSF(Maximal Spread Far) 방법이 있다[4].

1) MSC(Maximal Spread Close)방법

모델 (λ_i, λ_j)과 훈련음성 (X_i, X_j)을 가진 화자 i,j간의 거리는 다음 식(7)과 같이 정의 한다.

$$d(\lambda_i, \lambda_j) = \log \frac{p(X_i \mid \lambda_i)}{p(X_i \mid \lambda_j)} + \log \frac{p(X_j \mid \lambda_j)}{p(X_j \mid \lambda_i)} \tag{7}$$

거리가 가장 가까운 화자 N명(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 N명으로 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 $b(i)$에서 현재 화자의 수).

(단계 2) 다음 조건을 만족하는 화자 c를 $C(i)$에서 $b(i)$로 이동,

$$c = \arg\max_{c \in C(i)} \left\{ \frac{1}{B'} \sum_{j \in b(i)} \frac{d(\lambda_j, \lambda_c)}{d(\lambda_i, \lambda_c)} \right\} \tag{8}$$

N = N - 1, B′ = B′ + 1

(단계 3) 단계(2)를 B′ = B일 때까지 반복한다.

2) MSF(Maximal Spread Far)방법

이 방법에서 배경화자군의 구성은 거리가 근접한 화자군 일부와 멀리 떨어진 화자군 일부를 혼합하게 되는데, 이때 가장 멀리 떨어져 있는 화자 N명을 추출한 다음, 이중 고루 분포된 B/2 명을 배경화자의 일부로 선택한다. 선택방법은 다음의 알고리즘을 따른다[4].

$F(i)$: 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,

$$f = \arg\max_{f \in F(i)} \left\{ \frac{1}{B'} \sum_{j \in b(i)} d(\lambda_j, \lambda_F) * d(\lambda_i, \lambda_F) \right\} \tag{9}$$

(단계 3) 단계(2)를 B′ = B/2일 때까지 반복한다.

직관적으로 볼 때 배경화자는 일반적인 응용에서 명확히 기대되는 사칭자들의 모임으로 구성될 것이다. 이러한 사칭자들은 비슷한 음성의 특징을 가지거나 적어도 동성의 화자들로 구성될 것이라는 기준에서 배경화자를 선정하는 방법이 MSC 방법이다. 반면에 전화 기반의 응용 예에서는 보다 넓게 분포된 사칭자들의 접근을 가정할 수 있는데 예를 들면 남성 사칭자가 여성 사용자를 사칭하는 경우

가 이에 해당된다. 기존의 시스템에서는 등록된 화자에 최대한 접근한 화자들을 배경화자로 선택하기 때문에 보통의 경우에 시스템의 신뢰도를 확보할 수 있었지만 위와 같이 매우 다른 음성특징을 갖는 사칭자에게는 취약한 약점이 있다. 이런 경우에 적용하게 되는 방법이 MSF로 의뢰인과 거리가 먼 등록화자를 배경화자로 선택하는 방법이다. 이외의 배경화자 선정방법으로 의뢰인의 화자모델을 기준으로 인위적으로 가상의 화자모델을 생성하는 방법이 제안되어 등록화자 기반의 MSC, MSF 방법에 비해, 간단히 의뢰인의 음성데이터만을 이용하여 배경화자를 선정할 수 있다는 장점이 있다[7].

그리고 배경화자의 수는 사칭자 집합을 충분히 모델링 할 수 있을 만큼 커야 하지만 실제적으로는 계산량과 컴퓨터의 저장용량을 고려하여 제한된 작은 수의 배경화자를 이용하게 된다. 요즈음은 계산량을 고려하여 배경화자의 수를 동적으로 변화시켜 적용하는 방법과 심지어 배경화자를 사용하지 않는 방법 등도 제안되고 있다[8][9].

본 논문에서는 휴대폰 음성을 대상으로 화자인증 시스템을 구축하였을 때, 위와 같은 화자인증 시스템을 구성하는 여러 매개변수들이 성능에 미치는 영향을 파악하기 위하여 배경화자선정은 MSC 방법과 MIX 방법 두 가지를 비교하였으며, MIX 방법은 MSC에서 절반, MSF에서 절반의 배경화자를 선정하는 방법으로 배경화자의 성격이 혼합되었음을 나타내기 위해 MIX 방법이라고 본 논문에서 정하였다. 배경화자 수에 있어서는 6 명, 10 명, 16 명으로 변화시켜 보았으며, 또한 측정된 결과의 통계특성을 고려하여 적은 음성데이터이지만, 가능한 표본크기를 늘리기 위해 각각의 화자는 사용자 또는 의뢰자가 되는 한편 사칭자로도 이용되어 모든 화자에 대해 교차 실험을 하였다.

3. 음성 DB

실험에 사용한 음성 DB는 ETRI의 음성정보연구센터에서 만든 한국어 휴대폰 화자인식용 DB를 사용하였다. 음성데이타는 휴대폰 전화망 환경에서 총 257 명의 화자가 발성한 2 연 숫자, 4 연 숫자, 문장으로 구성되어있다. 문장음성의 발성목록은 개인정보와 관련된 10 개의 질문과 3 어절 이내로 구성된 단문 10 개로 구성되었으며, 한 화자당 동일한 목록을 5 회 발성하고, 주차/월차/3 개월차로 구분하여 각각 4 회 반복하여 녹음 수집하였다. 그러나 본 연구실에서 구입한 비영리용으로 만들어진 음성DB의 경우에는 원래의 영리용 DB에 비해 등록된 화자의 수가 많이 제한되어 있는데, 주차, 월차에 각각 20 명의 화자와 계절별 화자로 10 명으로 총 50 명의 화자가 등록되어 있다. 물론 주차, 월차 및 계절별 등록화자는 모두 다른 화자로 구성되어 있다. 음성시료의 녹음환경은 다양하여, 사무실, 집, 거리 지하철, 백화점 그리고 자동차안에서 휴대폰을 이용한 전화음성을 수집하였으며. 음성데이타는 8KHz/8 비트의 μ-law 압축 PCM방식으로 코딩되어 있다. 인증실험에 사용한 50 명은 남자 25 명, 여자 25 명으로 구성되어 있으며, 사용한 음성 데이타는 3 어절 이내로 구성된 단문 10 개이다. 단문 한 개의 평균 지속시간은 3 초 정도이며, 화자등록을 위해서 5 개의 단문을 합친 다음 끝점검출을 하여 평균 12 초 정도 분량의 데이터를 사용하였으며, 그리고 나머지 5 개의 단문으로 인증실험의 테스트용 화자음성으로 사용하였다. 다음은 인증에 사용한 화자군의 분류와 음성시료의 개수를 <표 1>에서 보여주고 있다.

표 1. 실험에 사용한 화자 및 음성시료의 구성

	남자	여자	전체
화자수	25	25	50
의뢰인 음성시료	125	125	250
사칭인 음성시료	3,000	3,000	12,250

<표 1>에서 의뢰인 음성시료의 개수는 등록 화자당 5 개의 단문을 사용하므로 총 50 명의 화자의 경우 250 개가 나오며, 사칭인 음성시료의 개수는 우선 의뢰인 본인을 제외한 나머지 화자는 사칭인으로 사용하므로 남자화자인 경우, 24 명의 사칭자가 각각 5 개의 음성시료가 있고, 이를 25 명의 의뢰인에 대한 전체 개수로 계산하면 실험이 표본크기는 $24 \times 5 \times 25 = 3,000$ 개가 된다.

4. 실험 및 결과

실험에서 각각의 음성시료는 끝점검출과 전처리과정을 거친 후, 20 ms(160 샘플)의 프레임 크기와 프레임 간격은 10 ms(80 샘플)로 하고, Hamming 창을 사용하여 프레임 단위로 특징벡터를 추출하였다. 음성의 특징벡터로는 12 차 MFCC(Mel-Frequency Cepstral Coefficient)를 사용하였다[11].

표 2. 화자인증 실험환경

실험환경	매개변수 값
GMM 차수	16 차
특징벡터	MFCC 12 차
프레임 길이	20 ms(160 샘플)
프레임 주기	10 ms(80 샘플)
Window	Hamming

화자모델링은 GMM방법을 사용하였으며, 가우시안 믹스처(mixture)의 개수는 16 개로 고정하여 사용하였다. 믹스처의 개수는 곧, GMM의 차수를 의미하는데, 이 차수의 선정은 시스템의 성능과 밀접한 관계가 있는데, 차수가 매우 작은 경우에는 화자모델링이 부정확해지고, 반대로 큰 경우에는 GMM의 파라미터를 추출하기 위해 보다 많은 훈련용 사용자의 음성이 요구될 뿐만 아니라 많은 계산량이 필요하게 되는 문제가 발생한다. 일반적인 GMM 화자모델의 차수 결정은 실험적으로 결정하게 되는데, 차수와 시스템의 성능 비교는 이미 여러 논문에서 실험을 통해 결과를 제시하였으며[5][6], 논문에서 30 초 분량의 훈련용 음성데이터에서 상대적으로 좋은 성능을 보여준 16 차수를 이번 실험에서 GMM 차수로 사용하였다[5].

화자인증의 성능을 나타내는 파라미터로 EER(Equal Error Rate)을 사용하였는데, 계산방법은 의뢰인 음성시료에 의해서 계산된 유사도 값들을 크기순서로 정렬하고, 마찬가지로 사칭자의 음성시료를 입력으로 한 경우의 유사도 값들 전체를 크기순서로 정렬한 다음, 인증수락과 거절을 결정

하는 유사도 값인 문턱값을 조절하여, FR(False Reject)율과 FA(False Accept)율이 동등하게 될 때의 FR율로 EER을 정하였다. 이에는 의뢰인 화자 개개인에 대한 EER이 있을 수 있지만, 본 논문의 실험에서 사용한 파라미터는 global EER 값으로 전체 등록화자의 데이터를 의뢰인, 사칭자 테스트에 사용한 결과를 모두 같이 정렬하여 통계의 표본크기를 증가시키므로 시스템의 성능을 나타내는 지표가 보다 신뢰성을 갖도록 하였다[10].

실험은 먼저 배경화자를 사용하여 유사도를 추정하는 경우와 그렇지 않은 경우를 비교하는 실험을 각각의 화자군에 대해서 실험을 하였다. 이때 배경화자가 없는 경우(NON)와 근접 화자로 구성된 MSC방법, 그리고 배경화자의 절반은 MSC방법으로 나머지 절반은 멀리 떨어진 화자 MSF방법을 사용하여 혼합하여 구성한 경우(MIX)로 구분하였다.

화자인증실험을 또 다시 두 종류로 나누었는데, 첫 번째가 문장종속 화자인증실험인 경우, 즉, 의뢰인으로 등록시에 사용한 발성목록과 같은 내용의 발성목록을 갖는 테스트 음성을 사용한 경우이고, 둘째로 등록시와 별개의 발성목록을 갖는 음성을 사용하는 문장독립 화자인증실험이다. 아래의 <표 3>은 문장종속인 경우의 성능을 요약하여 놓았다. 이때 사용한 배경화자의 수는 10 명으로 고정하였다.

표 3. 문장종속 화자인증실험에서 배경화자에 의한 성능비교

(단위: EER(%))

배경화자선정	남자	여자	전체	평균
NON	16.0	18.4	13.6	16.0
MSC	7.2	1.6	5.7	4.8
MIX	5.6	1.6	4.0	3.7

결과에서 보면 문장종속인 경우에도 배경화자가 사용되지 않은 경우(NON)는 배경화자와의 상대적인 유사도 값이 아닌, 인증 의뢰자의 화자모델만 갖고 구한 유사도 값으로 판정을 하기 때문에 평균 16%의 저하된 성능을 보이고 있다. 이에 반해 배경화자를 사용한 경우 MSC는 4.8%, MIX는 3.7%의 성능을 나타내어 배경화자의 사용에 의하여 인증 시스템의 성능 향상이 뚜렷함을 확인할 수 있었다. 특히 여자로 구성된 화자군의 경우에는 그 차이가 심하게 나타나는 것을 볼 수 있다. 일반적으로 여자화자인 경우의 성능이 남자화자인 경우에 비해 떨어지는 경우를 볼 수 있는데, 이번 실험의 경우는 전반적으로 여자화자의 결과 값이 나은 것으로 나타난 점이 의외로 생각된다. 또한 배경화자 선정방법에서는 근접화자로만 구성된 MSC 경우보다도 절반씩 MSC와 MSF가 혼합된 MIX 방법이 일반적으로 더 나은 결과를 보여 주고 있다. 이는 남자와 여자가 모두 모여 있는 전체 화자군의 경우 1.7% 정도의 성능차를 보여 보다 뚜렷하게 개선되는 양상을 확인할 수 있겠다. 남자와 여자의 성별의 차이는 화자모델의 거리가 가장 먼 경우로 볼 수 있으며, 이와 같이 혼재된 경우에는 마땅히 MIX 방법이 보다 우수하다고 볼 수 있다.

아래는 문장독립인 경우의 인증실험의 결과인데, 문장종속인 경우의 4%대에 비해 10%대로 증가함을 보여주는데, GMM 화자모델 방법이 지금까지 제안된 화자인식 방법 중에서 가장 문장독립적인 응용에 효과적이라고 알려져 있지만, 위와 같이 문장종속과 독립의 차이가 성능에 영향을 미치고 있음은 분명하다고 하겠다.

표 4. 문장독립 화자인증실험에서 배경화자에 의한 성능비교

(단위: EER(%))

배경화자선정	남자	여자	전체	평균
NON	24.0	31.2	23.2	26.1
MSC	14.4	11.2	12.0	12.5
MIX	13.6	11.2	10.8	11.8

다음으로 배경화자의 수의 변화에 따른 화자인증 시스템의 성능을 보여주는 결과가 <표 5>에 있다. 전반적으로 배경화자의 수가 증가할수록 성능이 조금씩 향상됨을 확인할 수 있다. 배경화자가 6 명일 때, MSC는 평균 13.7%지만, 배경화자 수가 10 명과 16 명으로 증가함에 따라 12.5% 그리고 11.9%로 개선되었다. 마찬가지로 MIX인 경우에도 12.3%, 11.9%, 그리고 11.2%로 점차 향상됨을 보여준다. 이는 배경화자의 수가 많을수록 사칭자의 확률적 모델링이 향상됨을 의미하며, 이로 인해 EER이 개선된다는 것을 증명하고 있다.

표 5. 배경화자수의 변화에 따른 인증성능의 비교

(단위: EER(%))

배경화자수	선정방법	남자	여자	전체	평균
B=6	MSC	14.4	12.0	14.8	13.7
	MIX	13.6	12.0	11.2	12.3
B=10	MSC	14.4	11.2	12.0	12.5
	MIX	13.6	11.2	10.8	11.9
B=16	MSC	13.6	10.4	11.6	11.9
	MIX	12.8	12.0	8.8	11.2

다음으로 등록화자의 GMM모델은 계속되는 시스템의 사용과 함께 입력으로 들어오는 의뢰인의 음성데이터를 이용하여 모델을 갱신해야 시간에 의한 사람의 음성의 변화에 대처할 수가 있을 것이다. 그래서 최초의 화자등록모델과 시간이 경과된 음성시료에 대한 인증실험의 결과를 알아보기 위해 ETRI 음성DB에 있는 3 개월차를 두고 녹음한 음성시료를 이용하여 실험을 하여 결과를 다음 <표 6>에 보였다.

표 6. 시간경과에 따른 인증성능의 변화

(단위: EER(%))

녹음수집 차수(3 개월차)	MSC	MIX	평균
세션 1 (최초 녹음)	4.0	4.0	4.0
세션 2 (3 개월 후)	10.0	14.0	12.0
세션 3 (6 개월 후)	20.0	18.0	19.0

표에 의하면 MSC나 MIX 경우 모두 세션 2와 세션 3의 음성데이터에 대해서 많은 성능의 저하를 보여주고 있다. 한 가지 특이한 사항은 일반적으로 MIX 경우가 MSC 보다 성능이 나은 경우를

보여 왔는데, 세션 2 실험에서는 오히려 MIX 경우가 4% 정도 열악한 것으로 결과가 나왔다. 이는 반복실험의 음성시료의 표본 개수가 작아서 일부 열악한 음성시료의 영향으로 생기는 것으로 파악된다. 그래도 실험에서 사용한 화자의 수가 불과 10 명이고 테스트 음성시료의 수도 각각 5 개로 의뢰인 실험결과가 50 개, 사칭자에 대한 실험결과가 450 개 정도이기에 통계 수치의 신뢰도가 많이 떨어진다는 것을 감안해야 하지만, 그래도 시간의 변화에 의한 인증 시스템의 성능변화를 살펴보려는 실험의 목적으로 보았을 때, MSC와 MIX의 평균 EER 값으로 4%에서 12%, 19%로 음성시료의 시간적 차에 의해서 분명한 성능의 감소가 있음을 확인할 수 있었다.

6. 결론 및 향후 과제

본 논문에서는 유선망 보다 가입자 수가 많은 휴대폰을 이용한 화자인증 시스템을 구축할 경우의 휴대폰 음성에 대한 인식시스템의 성능을 유사도 계산에 사용하는 배경화자의 선정방법과 배경화자의 수, 그리고 시간에 따른 화자인증시스템의 성능변화를 비교, 분석하였다. 실험에 사용한 음성DB는 ETRI에서 제작한 것으로 실험에 이용한 화자의 수는 남, 녀 각각 25 명씩 총 50 명이며, 발성목록은 3 어절내의 단문으로 구성된 10 개의 문장이며, 이중 문장 5 개는 화자모델을 등록할 때에 사용하였으며, 나머지 5 개는 화자인증 테스트를 할 때의 시료로 각각 사용하였다. 실험결과를 정리하면, 문장종속인 경우와 문장독립인 경우의 비교실험에서는 EER 값이 평균 4.2%와 12.1%로 문장종속인 경우가 성능이 좋았으며, 배경화자 선정방법인 MSC와 혼합형인 MIX인 경우에서는 문장종속에서 4.8%와 3.7%로, 문장독립일 경우에는 12.5%와 11.8%의 차이로 1% 범위내에서 MIX방법의 배경화자의 성능이 우수하였다. 또한 배경화자 수와 화자인증 시스템의 성능과의 관계를 살펴본 실험에서는 배경화자의 수가 6 명, 10 명, 16 명으로 증가하면서, EER 값은 MSC와 MIX의 평균으로 13.0%, 12.2%, 11.6%로 근소하게 향상됨을 알 수 있었다. 마지막으로 시간에 따른 음성의 변화를 살펴보기 위하여, 음성DB 중 3 개월차로 4 차례에 걸쳐 휴대폰 음성을 녹음한 데이터를 이용하였는데, 최초와 3 개월 그리고 6 개월 후에 수집된 음성들을 비교해서 실험한 결과, 4%, 12%, 19%의 EER 값이 계산되었다. 그러나 이 결과 값은 화자수가 10 명이고 실험횟수도 의뢰인의 경우 50 회, 사칭자의 경우 450 회로 EER 값을 정하기에 표본크기가 매우 적어서 FR값이 문턱값의 변화에 따라 최소 변동폭이 2%씩이 되므로 오차범위가 크다는 점을 감안하여도 예상보다 상당한 성능저하가 발생한다는 것을 확인하였다. 따라서 화자인증 시스템의 등록화자의 모델은 시스템의 사용과 더불어 꾸준히 사용자의 입력음성 데이터를 이용하여 모델을 갱신해야 성능의 저하를 막을 수 있으며, 이러한 화자모델의 효과적인 갱신과 관련된 연구가 진행되고 있다.

향후 연구과제로는 휴대폰과 같은 단말기에 화자인증 시스템을 구축하기 위해서는 단말기의 특성으로 인한 저전력과 가용 메모리가 적은 환경에서 유용하게 사용할 수 있는 배경화자 구축방법을 모색해야 하겠다. 물론 서버에 화자인증시스템을 구축하여 단말기의 계산량과 부담을 줄이는 방법이 제안되지만, 이는 이동망의 빈번한 사용과 더불어 비용의 증가를 가져온다는 것을 고려해야겠다. 화자인증이 앞으로 보다 활성화되는 유비쿼터스 통신환경에서 증대되는 개인 정보보호 측면에서 활용도가 높은 만큼 이에 대한 연구가 더욱 필요하겠다.

참 고 문 헌

[1] Gish, H. and Schmidt, M. 1994. "Text-independent speaker identification." *IEEE Signal Processing Magazine*, pp. 18-32, Oct.

[2] Reynolds, D. A. 2002. "An overview of automatic speaker recognition technology." *ICASSP*, pp. 4072-4075, Vol. IV.

[3] Liu, Chi-Shi., Wang, Hsiao-Chuan & Lee, Chin-Hui. 1996. "Speaker verification using normalized log-likelihood score." *IEEE Trans. On Speech and Audio Processing*, Vol. 4, No. 1, pp. 56-60, Jan.

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

[5] Reynolds, D. A. 1992. "A gaussian mixture modeling approach to text-independent speaker identification." Dissertation, Georgia Institute of Technology.

[6] Reynolds, D. A., Rose, R. C. 1995. "Robust text-independent speaker identification using gaussian mixture speaker models." *IEEE. Trans. On Speech and Audio Processing*, Vol. 3, No. 1, pp. 72-83, Jan.

[7] Isobe, T., and Takahashi, J. 1999. "Text-independent speaker verification using virtual speaker based cohort normalization." Eurospeech.

[8] 김성준, 계영철. 2003. "화자검증을 위한 새로운 코호트 선택방법." 한국음향학회지, 제22권, 제5호, pp. 383-387, 7월.

[9] Hsu, Chun-Nan., Yu, Hau-Chung., Yang, Bo-Hou. 2003. "Speaker verification without background speaker models." *ICASSP*, Vol. II, pp. 233-236.

[10] Therrien, C. W. 1989. *Decision, estimation and classification*, Wiley.

[11] Rabiner, L. R. & Schafer, R. W. 1978. *Digital processing of speech signals.*

접수일자: 2005. 07. 22
게재결정: 2005. 09. 01

▲ 최홍섭
경기도 포천시 선단동
대진대학교 공과대학 전자공학과 (우: 487-711)
Tel: +82-31-539-1903 Fax: +82-31-539-1900
E-mail: hschoi@daejin.ac.kr