관측신뢰도 적용에 의한 투표기법 기반의 화자인식시스템의 성능향상*

Performance Improvement of Voting-based Speaker Identification System by using the Observation Confidence

최 홍 섭**
Hongsub Choi

ABSTRACT

Recently demands for the speech technology-based products targeted for the mobile terminals such as cellular phones and PDA are rapidly increasing. And voting-based speaker identification algorithm is known to have a good performance in the mobile environment, since it works well with small amount of speaker training data. In this paper, we proposed a method to improve the performance of this voting based speaker identification system by using the observation confidence value which is derived from the function of SNR each frame. The proposed method is evaluated with ETRI cellular phone DB which is made for the speaker recognition task. The experimental results show that the proposed method has better performance of 2-3% identification rate than the conventional GMM method.

Keywords: speaker identification, GMM, voting, observation confidence, mobile, SNR

1. 서 론

오늘날 유비쿼터스 기술이 급진전됨에 따라 음성신호를 사용한 자동 화자인식기술은 모바일 또는 인터넷 응용역역에서 그 수요가 증대되고 있다. 그러나 모바일 환경이 보편화되면서 음성인식 및 화자인식 시스템은 보다 잡음에 많이 노출이 되어, 성능저하를 초래하였고, 또한 모바일 단말기의 특성상 제한된 하드웨어 자원으로 복잡한 알고리즘의 구현이 어렵다는 제약으로 실제 상용화된 서비스 제공에 걸림돌이 되고 있다.

이러한 모바일 환경의 잡음 관련 문제점에 대하여 현재 많은 연구가 진행되고 있는데, 주로 다음과 같은 접근 방법 등이 연구되고 있다. 먼저 CMS(Cepstral Mean Subtraction)와 같이 잡음과 채널왜곡에 강인한 파라미터를 추출하는 방법이다[1]. 이는 인식에 사용될 특징벡터의 전체 평균값을 특징벡터 값에서 차감하는 방법으로 채널왜곡과 잡음에 매우 효과적이다. 둘째는 화자 모델을 잡음에 맞도록 적응하는 모델적응 방법[2]으로 잡음의 특성에 따라 화자모델을 적응적으로 학습하는 방법이다. 마지막으로 집중성(burst) 잡음에 효과적인 투표(voting) 기법의 화자인식 방법이 있다. 투표기법의 화자인식기는 기존의 GMM

* 이 논문은 대진대학교 2007년도 대진대학교 학술연구비지원에 의한 것임.
** 대진대학교 공과대학 전자공학과

화자모델을 사용하되, 매 프레임 마다 계산된 조건부 확률값을 기반으로 프레임별 분류기를 구성하는 방법으로, 이는 인터넷에서 VoIP 통신이나 Voice mail과 같이 데이터가 패킷 형태로 전송될 때, 발생하는 집중성(burst) 잡음에 대해 효과적인 방법으로 사용될 수 있을 것이다. 또한 투표기법은 화자모델 학습을 위한 데이터의 양이 작은 경우, 기존 GMM방법에 비해 효과적이어서 휴대폰과 같은 모바일 환경에 사용하는 단말기에 내장하는 화자인식 알고리즘으로 적당하다고 판단된다[3].

본 논문에서는 이러한 휴대폰 응용을 위한 투표기법의 화자인식 시스템에 관측신뢰도 개념을 적용하여 성능을 향상하는 알고리즘을 제안하였다. 관측신뢰도(observation confidence)는 관측된 데이터가 잡음 등에 의하여 오염되었을 경우, 부정확한 데이터 값으로 인식성능이 저하되는 것을 막기 위하여 관측 데이터의 정확도를 평가하여 이를 인식과정에 가중치를 주어 반영하는 방법이다[5]. 그리고 제안한 화자인식 알고리즘의 타당성을 검증하기 위하여 ETRI에서 제작한 화자인식용 휴대폰 음성DB에 대하여 문맥종속 화자실험을 하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 GMM 모델을 사용하는 화자인식기의 일반적인 구조에 대해서 설명하고, 이어 3 장에서 투표기법 화자인식방법과 관측신뢰도 추정에 대하여 설명한다. 본 논문에서 제시한 알고리즘에 대한 타당성을 4 장에서 문맥종속 화자식별 실험을 통하여 검증하며, 마지막 5 장에서 결론을 맺는다.

2. GMM 화자인식시스템의 구조

GMM(Gaussian Mixture Model)은 여러 개의 가우시안 확률밀도(Gaussian probability density) 함수들에 각각의 가중치를 준 다음, 이를 선형 결합함으로써 임의의 모양을 갖는 확률밀도 함수를 표현할 수 있는 확률분포 모델이다. 그리고 음성의 특징 파라미터 벡터의 확률분포는 화자마다 그 모양이 다르다는 사실을 화자인식에 이용할 수 있다. 따라서 GMM 화자인식시스템은 화자모델로 GMM을 사용하여 화자인식을 수행하는 시스템이다. 다음은 GMM을 어떻게 정형화하고 그 파라미터는 어떻게 구성되어 있는지를 간략히 설명한다. <그림 1>은 GMM의 구성을 보여주고 있다[7].

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

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

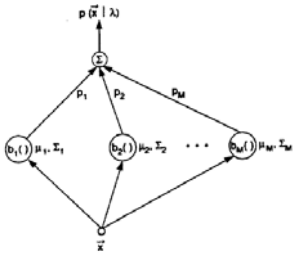

그림 1. GMM의 구성

식(1)에서 $\vec{x}$는 D차원의 랜덤 벡터이며

$b_i(\vec{x}), \quad i=1,...,M$은 성분 가우시안 분포이고 $p_i, \quad i=1,...,M$ 은 혼합 가중치(mixture weight)라고 불리며 각각의 가우시안 분포에 대한 가중치이다. 각 분포는 D차원의 가우시안 분포이며 식(2)와 같이 표현되며

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu_i})'\Sigma_i^{-1}(\vec{x}-\vec{\mu_i})\right] \tag{2}$$

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

$$\sum_{i=1}^{M} p_i = 1 \tag{3}$$

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

$$\lambda = \left\{p_i, \vec{\mu_i}, \Sigma_i\right\}, \quad i=1,...,M \tag{4}$$

GMM을 이용한 일반적인 화자식별 실험과정은 다음과 같다.

(1) 입력음성으로부터 특징파라미터 멜켑스트럼(MFCC)을 추출한다. 사용하는 멜켑스트럼은 지금까지 화자인식 실험에서 우수한 성능을 갖는다고 검증된 파라미터이다.

(2) 추출된 멜켑스트럼에 대해 채널보상방법인 CMS(Cepstral Mean Subtraction)를 수행한다. CMS는 채널에 의해 발생하는 채널왜곡을 제거하고, 동일화자가 시차를 두고 발성한 음성데이터의 평균스펙트럼의 편차를 제거하는 성질을 가지고 있다.

(3) CMS 처리된 특징파라미터에 대하여 학습을 통해 GMM모델을 구한다. 화자식별을 위하여 등록된 모든 화자에 대하여 GMM모델을 구하여 저장한다.

(4) 화자인식실험 과정에서는 입력음성에서 특징파라미터를 추출한 다음, 위에서 구한 모든 등록화자의 GMM모델에 대한 조건부 확률을 계산하고, 이들 중 가장 높은 확률을 갖는 모델의 화자를 인식결과로 출력한다.

위의 화자식별 시스템의 알고리즘은 <그림 2>와 같다. 실선으로 구성된 부분이 기존의 화자식별 알고리즘이고, 점선으로 표시된 부분은 본 논문에서 제안한 부분이 첨부된 것이다. 즉, 신호대잡음비(SNR)을 추출하여, 이를 근거로 관측데이터인 특징벡터의 관측신뢰도를 추정한 후, 이를 투표기법에 가중치로 적용하는 알고리즘이다.

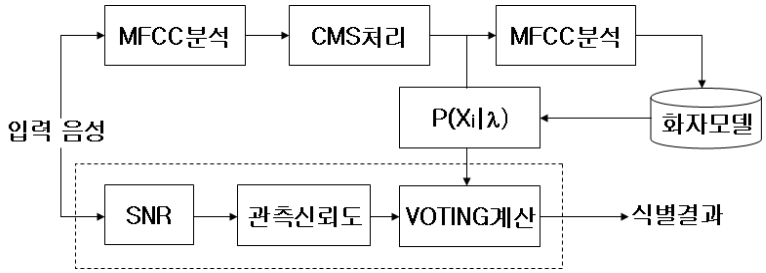

그림 2. 화자식별 시스템의 구조

3. 투표기법(Voting method)과 관측신뢰도 추정방법

3.1 투표기법에 의한 화자식별

GMM 기반의 화자인식 실험에서 발생하는 오류를 분석하여 보면, 인식에 실패한 음성데이터의 대부분의 프레임에서는 화자모델에 대한 조건부 확률값 $p(\overrightarrow{x_i}|\lambda_k)$ (i는 프레임 번호, k는 화자식별 번호)이 다른 화자모델에 비해 높아도, 몇몇 프레임에서 매우 작은 값을 갖게 되면, 오류가 발생하는 것을 확인할 수 있다. 이는 GMM 화자인식에서 판정에 사용하는 확률값을 다음 식(5)와 같이 모든 프레임의 확률값들을 곱해서 산출하기 때문이다.

$$p(\overrightarrow{X}|\lambda_k) = \prod_i p(\overrightarrow{x_i}|\lambda_k) \tag{5}$$

이와 같이 확률값들을 곱할 경우에 매우 낮은 확률값을 갖는 프레임이 5-6개 정도만 되어도 전체 데이터의 확률값 $p(\overrightarrow{X}|\lambda_k)$이 작게 나올 수 있어 인식오류가 생기게 된다. 이러한 프레임은 주로 잡음에 심하게 오염되어 있는 경우, 또는 등록된 화자모델과의 편차가 큰 경우에 발생할 수 있으며, 이러한 프레임을 인위적으로 찾아서 제거하면 인식 오류가 없어짐을 통해서 이런 사실을 확인할 수 있다.

따라서 위와 같이 일부 오염된 프레임이 전체 확률에 높은 가중치를 주는 문제를 해결하기 위하여 투표기법이 제안되었다[3,4]. 즉, 모든 프레임이 화자인식 판정에 미치는 영향을 균등하게 부여하는 방법으로, 매 프레임 마다 모든 화자모델에 대하여 계산된 확률값 $p(\overrightarrow{x_i}|\lambda_k)$ 중 가장 높은 값을 갖는 화자모델에 투표를 하는 방법이다. 식(6)은 프레임의 조건부 확률값에서 화자모델을 찾는 관계식이다.

$$\hat{S}_i = \arg\max p(\overrightarrow{x_i}|\lambda_k),\ 1 \le k \le M \tag{6}$$

즉, 식(6)에 의해서 확률값이 가장 높은 화자모델에 대한 계수기의 값을 1 증가하는 것으로 투표를 대신한다. 이를 매 프레임 마다 반복하여 최종적으로 가장 많은 투표를 받은, 즉 계수기의 값이 가장 큰 화자모델을 최종 화자로 식별하는 방법이다[3,4]. 따라서 투표기법의 화자인식은 각 프레임마다 인식판정을 내리는 분류기들의 집합체로 볼 수 있겠다. 아래 <그림 3>은 투표기법을 수행하는 알고리즘의 순서도이다.

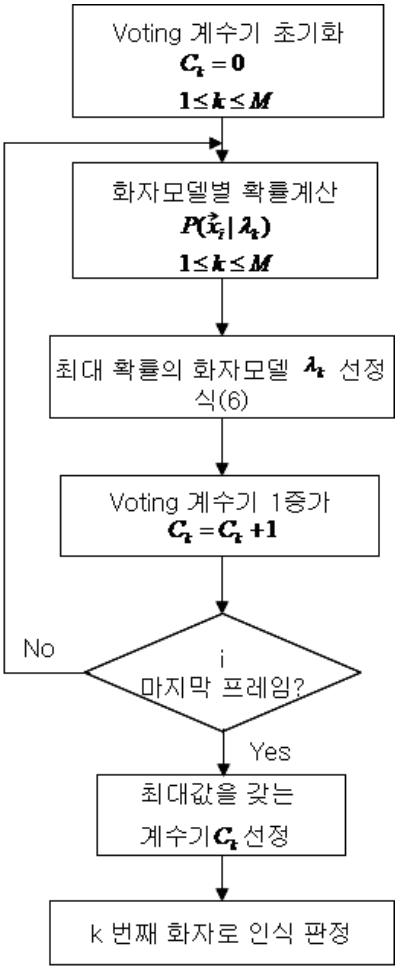

그림 3. 투표기법 알고리즘의 순서도

3.2 관측신뢰도 추정

관측신뢰도란 주어진 관측 파라미터에 대한 신뢰를 정하는 척도로, 얼마나 정확하게 측정된 파라미터인가를 정량적으로 표현하는 값이다. 일반적으로 관측신뢰도는 잡음과 가장 밀접한 관계가 있으므로 잡음의 함수로 표현하는 것이 타당할 것이다. 본 논문에서는 시그모이드(sigmoid)함수를 잡음과 신뢰도의 매핑함수로 도입하였다. 즉, 관측신뢰도 함수는 식(7)과 같이 정의한다[5].

$$\mu(snr) = \left[1 + e^{a(snr-b)}\right]^{-1} \tag{7}$$

위 식에서 a는 스케일 파라미터이고, b는 이동 파라미터로서 $a = -0.22$, $b = 10.0$인 경우에 관측신뢰도 그래프는 <그림 4>와 같이 나타난다.

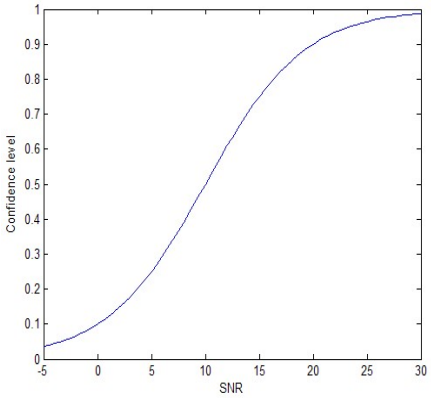

그림 4. 관측신뢰도 함수 그래프
($a = -0.22$, $b = 10.0$)

본 논문에서 사용한 관측신뢰도 함수의 파라미터 값은 실험을 통해서 최적화된 값을 구하여 사용하였다. 앞의 <그림 2>는 논문에서 제안하는 GMM 기반의 화자인식기의 구조로서, 여기서 점선으로 표시된 부분이 SNR을 이용하여 관측신뢰도를 추정하고, 이를 가중치로 하여 투표기법에 적용하는 부분이다. 즉, SNR이 낮은 프레임의 경우에는 가장 높은 확률값을 얻은 화자모델에 1을 주는 것이 아니라 관측신뢰도 함수에서 구한 가중치를 곱해서 1과 0사이의 값을 가산하는 방법이다.

4. 실험 및 결과 고찰

4.1 화자인식 DB 및 문맥종속 화자인식 실험개요

본 논문에서는 제안된 방법의 성능을 확인하기 위하여 ETRI에서 만든 한국어 화자인식용 휴대폰 음성DB을 사용하여 문맥종속 화자식별 실험을 하였다. 사용된 음성데이터의 샘플링 주파수는 8 kHz이며, 8 비트 μ-law PCM방식으로 음성코딩 되었다. 그리고 DB의 전체 화자의 수는 남녀 모두 50 명으로 구성하였으며, 음성데이터는 일주일 단위로 20 명, 월간 단위로 20 명, 그리고 3 개월 계절 단위로 10 명씩 구분하여 화자당 4 회의 세션을 거쳐 시간 간격을 두고 녹음하였다. 세션별로 5 회씩 녹음하였으므로 문장당 모두 20 개의 데이터가 수집되었다. 실험에서는 5 개의 단문장을 선정하여 학습과 테스트에 사용하였다. <그림 5>는 인식실험에 사용한 음성데이터의 파형으로 가로축은 샘플수로 약 2.5 초 정도 길이의 데이터임을 알 수 있다.

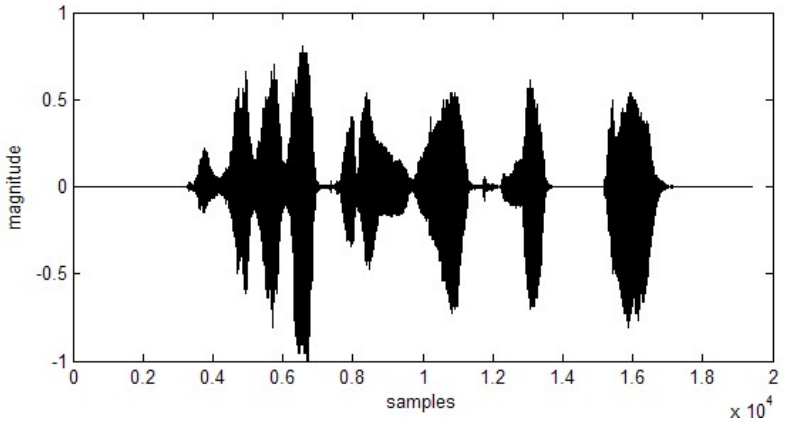

그림 5. 음성신호의 파형(발성: “무지개가 구름 뒤에 숨었다”)

GMM 화자모델을 구하기 위하여 학습에 사용한 음성데이터와 화자식별에 사용된 테스트용 음성데이터의 크기는 모두 5 초, 10 초, 15 초 그리고 20 초로 구분하여 각각에 대하여 실험을 하였다. 그러나 한 개의 음성파일은 평균 2 초에서 3 초 분량의 데이터를 가지고 있어서, 주어진 시간 길이의 데이터를 얻기 위하여 5 개에서 10 개까지의 음성파일들을 합쳐서 학습용 및 테스트용 음성데이터를 구성하였다. 화자인식의 전처리 과정에서는 음성데이터의 한 프레임의 길이는 20 ms로 하고, 중첩은 10 ms로 하였고, 특징벡터는 12 차 멜켑스트럼 계수와 로그에너지를 포함하였다. 또한 채널 및 잡음 왜곡을 보상하고, 음성시료들의 세션간의 스펙트럼 편차를 제거하기 위하여 CMS 방법을 적용하였다. 실험에서 입력 음성데이터의 값은 최대값을 1로 정규화 하였으며, 음성데이터의 프레임 별 SNR 계산은 다음과 같이 하였다. 사실 SNR 계산을 위해서는 매 프레임에서 잡음신호를 음성신호와 분리해야 하지만 이는 실제 상황에서 불가능하다. 실험에서는 대신 입력음성의 시작부터 200ms 구간을 미리 묵음구간으로 설정하여 이 부분의 전력을 프레임 잡음전력으로 정하고, 신호전력은 프레임 내 음성데이터 값의 제곱평균으로 구한 후 이들의 비로서 신호대잡음비인 SNR을 계산하였다. GMM 화자모델에 포함된 가우시안 혼합 개수는 10 개이고, EM알고리즘을 사용하여 GMM 모델 λ_i의 파라미터를 반복적으로 훈련하여 구하였다. 이 과정에서 공분산은 Full covariance를 사용하였고, EM 알고리즘의 초기과정에서는 fuzzy C-means clustering 방법을 적용하였다. 논문에서 설정한 화자식별 실험의 조건은 다음 <표 1>과 같다.

표 1. 문맥종속 화자식별 시스템의 개요

항목	내용
음성DB	ETRI 화자인식용 휴대폰 DB
음성특징벡터	12 차 MFCC와 로그에너지
샘플링 및 음성코딩	8 KHz, 8 bits/sample, μ-law PCM
프레임길이/중첩	20 ms/10ms
화자수	50 명
채널보상	Cepstral Mean Subtraction
GMM모델	EM알고리즘, Full covariance 사용

Gaussian mixture 개수	10
화자당 학습데이터 길이	5 초, 10 초, 15 초, 20 초
화자당 테스트데이터 길이	5 초, 10 초, 15 초, 20 초

4.2 실험결과 및 고찰

일반적으로 화자인식기의 성능은 학습데이터의 양과 테스트 데이터의 양에 따라 차가 나타나므로, 본 논문에서는 각각 5 초, 10 초, 15 초, 20 초 분량으로 나누어서 실험을 하였다. 아래 <그림 6>부터 <그림 9>는 학습데이터가 각각 5 초, 10 초, 15 초, 20 초일 때의 인식실험 결과를 막대그래프로 나타냈다. 그래프의 가로축은 테스트에 사용한 음성데이터의 크기를 초단위로 나타내었고, 세로축은 인식 성능을 백분율로 표시하였다. 실험에 사용한 인식방법은 기존의 GMM 방식(GMM)을 기본으로 하여 맨 왼쪽의 막대로 표시하였고, 둘째는 투표기법을 적용한 화자인식기(VOTING) 그리고 마지막 세 번째 막대는 관측신뢰도를 적용한 투표기법의 화자인식기(VOTING+WT)를 표시하였다.

먼저 기존의 GMM방법과 투표기법의 화자인식 알고리즘의 성능을 비교하여 보았다. 결과의 그래프에서 보듯이 학습데이터의 크기가 커질수록 인식방법 모두 인식률은 지속적으로 증가함을 알 수 있다. 또한 테스트 데이터의 크기도 같은 크기로 증가시킴에 따라 인식률이 조금씩은 상승이 되지만, 학습데이터에 비해서는 변화율이 적음을 알 수 있겠다. 주목할 사항은 학습데이터가 20 초 정도일 경우에는 투표기법의 화자인식이 기존의 GMM방식에 비해 성능이 1-4% 정도 차이지만 학습데이터의 크기가 5 초일 경우에는 투표기법이 적게는 7%에서 많게는 18% 정도까지 성능차를 보임을 알 수 있다. 이를 통해 화자인식기의 성능을 위해서는 충분한 학습데이터를 확보하는 것이 좋은 방법임을 알 수 있으나, 모바일 단말기와 같이 제한된 메모리를 사용하는 응용분야에서는 투표기법이 효과적으로 적용될 수 있음을 보여 준다.

그러나 투표기법의 경우에는 테스트 데이터의 크기가 5 초일 경우에 보면, 오히려 기존의 방법보다 성능이 낮아지게 되는데, 이는 크기가 작은 데이터에서 작은 수의 프레임이 투표에 참가함으로써 판정의 기준이 되는 득표수에서 변별력이 떨어지기 때문에 발생하는 것이다. 특히 식별해야 할 등록된 화자모델이 많을 경우에 이러한 문제점이 더욱 두드러짐을 알 수 있고, 이 경우에는 기존의 GMM과 투표기법을 혼합하여 적용하는 방법으로 해결할 수 있다. 즉, 일차적으로 GMM모델을 이용하여 계산된 확률값을 토대로 N-best 화자후보들을 선발한 다음에 이들에 대해서만 투표기법을 적용하여 최종 인식을 수행하는 방법이다[3].

둘째로 투표기법 화자인식기와 이에 관측신뢰도를 적용한 제안한 투표기법 화자인식기의 성능에 대해 비교하였다. 전체적으로 투표기법에 관측신뢰도를 적용한 방법은 적용하지 않은 투표기법에 비해 모든 경우에서 1.5-3% 정도의 성능이 향상되었으며, 이는 SNR에 의한 관측신뢰도가 데이터의 정확도를 정량적으로 추정할 수 있음을 보여준다. 그러나 이 경우 관측신뢰도 함수에 따라 성능에 영향을 주게 되게 되므로 신뢰도 함수를 결정하는 것이 중요한 문제가 된다. 본 논문에서는 관측신뢰도 함수를 실험을 통해 경험적으로 결정하여 사용하였지만, 이를 찾는 최적화 방법에 대한 연구가 필요할 것이다.

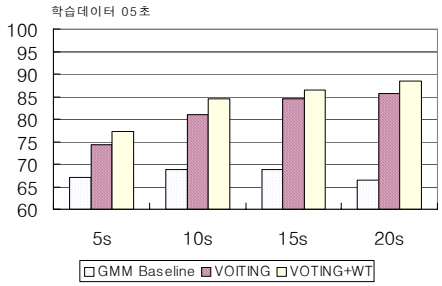

그림 6. 화자 식별 결과(학습데이터 5 초)

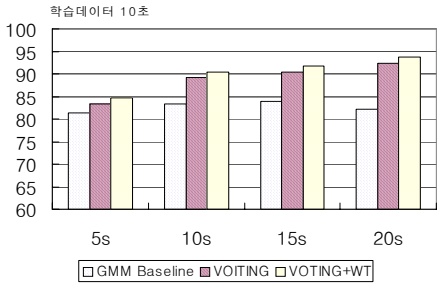

그림 7. 화자 식별 결과(학습데이터 10 초)

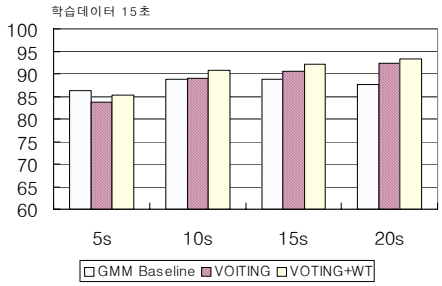

그림 8. 화자 식별 결과(학습데이터 15 초)

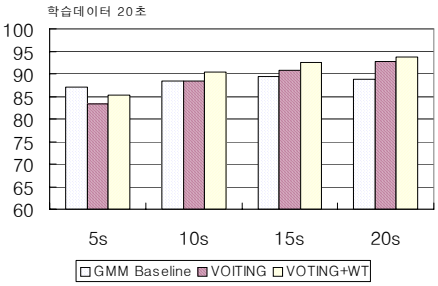

그림 9. 화자 식별 결과(학습데이터 20 초)

5. 결 론

본 논문에서는 잡음에 의한 화자인식기의 성능저하에 대응하기 위해 최근에 제안한 관측데이터의 신뢰도 추정방법을 투표(voting)기법을 기반한 GMM 화자식별기에 적용하여 성능을 검증하였다. 실험을 위하여 ETRI에서 만든 화자인식용 한국어 휴대폰 음성DB를 사용하여 문맥종속 화자식별 실험을 하였다. 총 화자수는 남녀 합쳐 50 명이었으며, 음성시료는 주간, 월간, 계절별로 4 회에 걸쳐서 그리고 1 회 녹음시 5 회 반복 발성된 음성을 녹음하였다. 화자모델 학습과정과 화자식별 테스트 과정에서 모두 5 초, 10 초, 15 초, 20 초 분량의 음성데이터를 사용하여 실험을 수행하였다.

먼저 기존의 GMM방식과 투표기법을 비교하는 실험의 경우에 화자모델 학습데이터의 크기가 20초인 경우에는 두 방식 사이에 성능차가 크지 않았지만, 학습데이터가 5 초인 경우에는 투표기법이 기존의 GMM방법에 비해 최대 18%까지 성능이 향상되었다. 이러한 특성은 모바일 단말기와 같은 제한적인 응용분야에서 화자인식기를 구성할 때 투표기법의 화자인식 방법이 효과적이라고 판단된다.

그리고 투표기법 화자인식과 관측신뢰도 개념을 적용한 투표기법의 화자인식에 대한 비교 실험에서는 관측신뢰도를 이용한 투표기법이 평균 약 1.5-3% 정도 인식률이 우수함을 확인하였다. 이

는 관측신뢰도 개념을 GMM 화자인식기에 적용하여 그에 대한 효과를 확인한 바와 같이[5,6], SNR을 변수로 하여 구한 관측신뢰도가 측정 데이터의 충실도를 정량적으로 나타내는 지표로 활용할 수 있음을 보여준다. 그러나 관측신뢰도 함수로 사용한 시그모이드 함수의 파라미터는 실험에 의하여 경험적으로 구한 값들을 사용하였으나, 추후 이를 찾기 위한 최적화 알고리즘에 대한 연구가 필요하겠다.

참 고 문 헌

[1] Rosenberg, A., et al. 1994. "Cepstral channel normalization techniques for HMM-based speaker verification," *Proc. ICSLP-94*, 1835-1838.

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

[3] Narayanaswamy, B. & Gangadharaiah, R. 2005. "Extracting additional information from Gaussian mixture model probabilities for improved text-independent speaker identification," *Proc. ICASSP 2005.*

[4] Narayanaswamy, B., Gangadharaiah, R. & Stern. R. 2006. "Voting for two speaker segmentation," *Proc. ICSLP 2006.*

[5] Kim, J. et al. 2007. "Modified GMM training for inexact observation and its application to speaker identification," *Speech Sciences* 14(1), 163-175.

[6] 최홍섭, 2007. "SNR을 이용한 프레임별 유사도 가중방법을 적용한 문맥종속 화자인식에 관한 연구," *말소리* 3, 113-123.

[7] Reynolds, D. A. & Rose, R. C. 1995. "Robust text-independent speaker identification using gaussian mixture speaker models", IEEE. Trans. *On Speech and Audio Processing*, 3(1), 72-83.

접수일자: 2008. 4. 3
게재결정: 2008. 5. 8

▲ 최홍섭
경기도 포천시 선단동 산 11-1번지 (우: 487-711)
대진대학교 전자공학과 교수
Tel: +82-31-539-1903 Fax: +82-31-539-1900
E-mail: hschoi@daejin.ac.kr