SNR을 이용한 프레임별 유사도 가중방법을 적용한 문맥종속 화자인식에 관한 연구*

최홍섭(대진대)

<차 례>

<Abstract>

A Study on the Context-dependent Speaker Recognition Adopting the Method of Weighting the Frame-based Likelihood Using SNR

Hong Sub Choi

The environmental differences between training and testing mode are generally considered to be the critical factor for the performance degradation in speaker recognition systems. Especially, general speaker recognition systems try to get as clean speech as possible to train the speaker model, but it's not true in real testing phase due to environmental and channel noise. So in this paper, the new method of weighting the frame-based likelihood according to frame SNR is proposed in order to cope with that problem. That is to make use of the deep correlation between speech SNR and speaker discrimination rate. To verify the usefulness of this proposed method, it is applied to the context dependent speaker identification system. And the experimental results with the cellular phone speech DB which is designed by ETRI for Koran speaker recognition show that the proposed method is effective and increase the identification accuracy by 11% at maximum.

* Keywords : Context-dependent speaker identification, Speaker discrimination rate, Frame-based likelihood, Frame SNR, Cellular phone speech DB

* 이 논문은 2005학년도 대진대학교 학술연구비지원에 의한 것임.

1. 서 론

화자인식기술은 음성을 인터페이스로 사용하여 보안시스템, 폰뱅킹, 개인의 정보검색 등 온라인과 오프라인 상에서 사용자의 신분을 확인하는 유용한 기술로 그 중요성은 정보사회로 진행될수록 커질 수밖에 없겠다. 이러한 화자인식기술의 성능을 저해하는 주요 원인으로는 화자모델을 훈련시킬 때의 음성데이터와 실제 현장에서 인식기를 작동시킬 때 들어오는 입력데이터 간의 잡음특성과 같은 주변 환경의 불일치를 꼽을 수 있다. 특히 화자모델 훈련용으로 준비하는 데이터는 가능한 깨끗한 음성데이터를 사용할 수 있지만, 인식실험을 적용하는 단계에서는 주변 소음, 전화채널 잡음 그리고 시스템 잡음 등이 입력음성에 혼입되어 시스템의 인식성능이 급감하게 됨을 알 수 있다. 이러한 잡음이 많이 혼합되는 음성데이터에 대한 처리는 보통 두 가지 방향으로 처리를 하고 있는데, 그 하나는 인식의 전처리 단계에서 가능한 잡음을 제거하여 음질을 향상시키거나, 또는 추출한 특징벡터에서 잡음의 영향을 제거하는 방법이고[1][2], 두 번째는 시스템의 모델을 변화된 환경에 맞게 적응시키는 방법이다[3][4]. 현재는 음질 및 특징벡터를 잡음에 강인하게 향상시키는 첫째 방법이 선호되고 있으며, 대표적인 방법으로는 CMS(Cepstral Mean Subtraction)와 RASTA필터링 등이 있다[1][2]. 그러나 이러한 여러 방법이 제안되고 사용되고 있지만, 실제로 잡음에 대한 대처 방법에는 한계가 있어 아직도 이에 대한 많은 연구가 진행되고 있는 것이 사실이다.

본 논문에서는 화자인식 단계에서 인식기의 입력으로 들어오는 음성데이터의 신뢰도를 추정하여 이를 인식결정 단계에서 추출되는 확률 값에 가중치를 주어 잡음의 영향을 가능한 줄이는 방법을 제안하였다. 일반적으로 데이터의 신뢰도는 잡음에 의해 손상되므로 본 논문에서는 음성데이터의 프레임별 SNR (Signal-to-Noise Ratio)을 기준으로 데이터의 신뢰도를 추정한 후, 이를 토대로 가중치를 구하여 사용하였으며, 제안한 방법을 검증하기 위하여, ETRI에서 수집한 한국어 화자인식용 휴대폰 음성DB를 사용하여, 문맥종속 화자식별 시스템에 적용하여 실험을 하였다.

본 논문의 구성은 서론에 이어, 2장에서는 화자인식에 있어서 음성데이터의 신뢰도를 입력신호의 SNR을 근거로 추정하고, 이를 프레임별로 계산되는 유사도 값에 적용하는 가중치로 사용하는 타당성을 실험적으로 보인다. 3장에서는 제안한 방법을 사용하여 구성한 문맥종속 화자식별 시스템의 구성을 자세히 설명하고, 이어 4장에서는 인식실험에 사용한 음성DB와 실험의 구체적인 과정 및 결과를 그리고 마지막 5장에서는 논문의 결론과 앞으로의 연구진행 방향 등에 대해 기술한다.

2. SNR에 따른 프레임별 유사도 가중 방법

화자식별은 등록된 화자를 적당한 방법으로 모델링한 다음, 인식시스템에 입력으로 들어오는 음성데이터를 이용하여, 미리 구한 화자모델에 대한 조건부 확률을 계산하여 이를 근거로 화자를 판단하게 된다. 화자식별에 이용하는 기본 식은 아래와 같이 나타낼 수 있다.

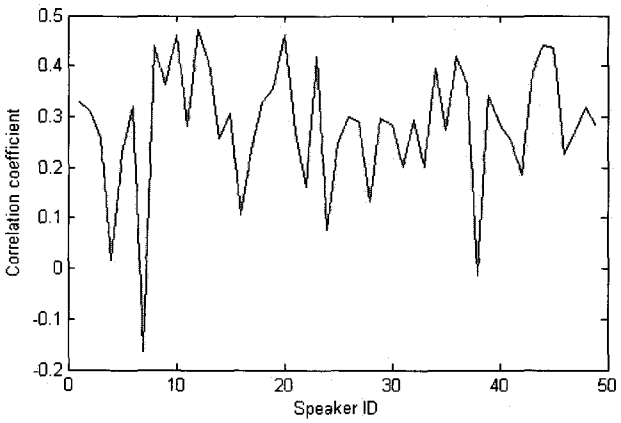
$$\hat{S} = \arg\max_k p(X|\lambda_k) = \arg\max_k \left[\prod_{t=1}^{T} (p(\vec{x_t}|\lambda_k)\right] \tag{1}$$

여기서, λ_k는 화자모델을 나타내며, $X = \{\vec{x_1}, \vec{x_2}, \cdots, \vec{x_T}\}$는 음성데이터의 프레임들에서 추출한 특징벡터 $\vec{x_t}$들의 집합이고, T는 음성데이터의 총 프레임 개수이다. 위의 식에서 보면 화자모델에 대한 조건부 확률인 $p(X|\lambda_k)$을 등록된 모든 화자모델에 대하여 계산한 다음 이중에서 최대 확률을 갖는 인덱스 k의 화자를 인식의 결과로 결정하게 되는 것이다. 일반적인 화자인식시스템의 성능저하의 주요 원인으로 꼽히는 것이 바로 앞에서 얘기한 화자모델을 구하는 훈련과정에서 사용하는 음성데이터와 실제의 인식기를 사용하는 실험과정에서 사용하는 음성데이터의 통계적 특성의 불일치이다. 즉, 화자모델의 훈련 시에는 가능한 잡음이 적은 깨끗한 음성데이터를 사용할 수 있지만, 실제 인식기의 사용단계에서는 주변의 잡음이나, 채널로부터 발생되는 잡음 등으로 입력데이터의 특성이 변화되어 훈련과정에서 구한 화자모델과의 통계적 특성의 차이를 나타내게 되어 인식성능에 영향을 주게 된다. 따라서 논문에서는 먼저 음성데이터의 SNR이 화자인식기의 성능과 어느 정도의 상관관계를 갖는 지를 화자분해능(Speaker Discrimination Rate)이라는 파라미터를 만들어서 확인하여 보았다. 화자분해능은 다음과 같은 식으로 표현할 수 있다.

$$SDR(\vec{x}) = \frac{p(\vec{x}|\lambda_{true})}{\sum_{i=1}^{N} p(\vec{x}|\lambda_i)} \tag{2}$$

여기서, N은 시스템에 등록된 전체 화자의 수이고, λ_{true}는 입력된 음성데이터의 실제 화자모델을 나타낸다. 즉, 위의 식은 임의의 프레임의 특징벡터 $\vec{x}$에 대한 화자모델의 조건부 확률을 모든 등록된 화자에 대해서 계산한 다음, 이의 전체 합에 대한 참인 화자모델의 조건부 확률의 비를 의미하게 되어, 이 파라미터의 값이 큰 경우는 결국 참이 아닌 다른 화자모델에 대해서 참인 화자의 변별력이 높다는 것을 의미하게 되어, 화자식별기의 성능과 상관이 있다고 할 수 있겠다. 실제로 화자모델을 GMM(Gaussian Mixture Model)로 모델링하여 음성데이터의 프레임별

SNR과 화자분해능인 SDR 파라미터 사이의 상관관계를 구하였더니, 다음 <그림 1>과 같은 결과를 볼 수 있었다.

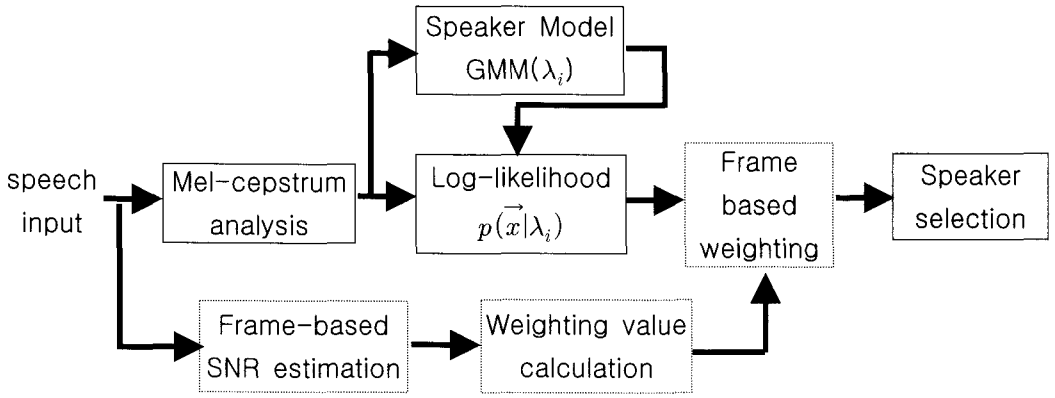


<그림 1> 화자별 SNR과 화자분해능의 상관계수

위의 그림에서 가로축은 화자의 인덱스를 나타내고, 세로축은 SNR과 화자분해능 SDR값과의 상관계수(correlation coefficient)를 표시한다. 이때 상관계수는 모든 프레임에서 구한 프레임별 SNR과 SDR 값을 랜덤변수의 표본 값으로 하여 계산하였으며, 화자별로 구한 SNR과 SDR 사이의 상관계수의 평균은 위의 그림과 같이 약 0.3 근처의 값을 갖고 있음으로 이는 SNR이 높은 음성데이터일수록 화자분해능이 높아짐을 보여주고 있다. 본 논문에서는 이러한 성질을 이용하여 주어진 화자모델에 대한 조건부 확률인 프레임별 유사도(likelihood) 값에 프레임별 SNR 값에 따른 가중치를 적용함으로써 SNR이 높은 프레임의 유사도 값은 가중치를 높이고, SNR이 낮은 프레임은 가중치를 줄이는 방법으로 인식기의 성능을 향상시키는 방법을 제안하였다.

3. 제안한 화자식별 시스템의 구성

프레임별 SNR에 의한 유사도 가중치를 적용하는 방법의 화자식별 시스템의 전체 구성은 <그림 2>와 같다. <그림 2>에서 실선으로 표시된 부분은 기존의 화자식별 시스템의 구조이고 점선으로 표시된 부분이 논문에서 제안한 부분이다. 화자식별을 위한 화자모델은 근래 많이 사용되는 GMM 모델을 사용하였다. GMM 모델은 M개의 Gaussian 분포의 가중치 합으로 구성되며 다음과 같은 식으로 표시된다[5][6].



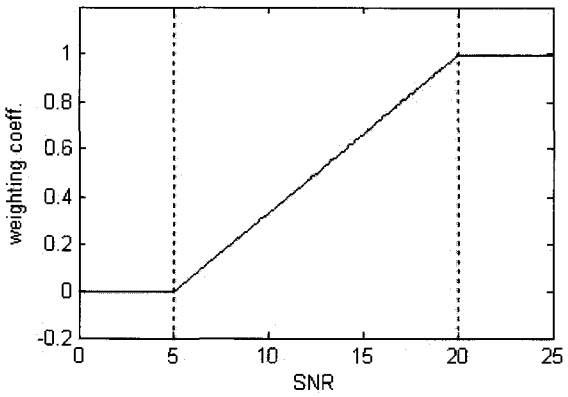
<그림 2> 프레임별 SNR에 의한 유사도 가중치를 적용하는 방법의 화자식별 시스템의 전체 구성도

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{3}$$

식 (3)에서 $\vec{x}$는 D차원의 랜덤 벡터이며 $b_i(\vec{x})$은 성분 Gaussian 분포이고 p_i은 결합 가중치(mixture weight)로 $\sum_{i=1}^{M} p_i = 1$의 관계를 가진다. 따라서 GMM 모델은 위에서 기술한 파라미터들 즉 평균 벡터 $\vec{\mu_i}$와 공분산 행렬 Σ_i 그리고 결합가중치 p_i에 의해 다음과 같이 완전하게 표현이 되며,

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

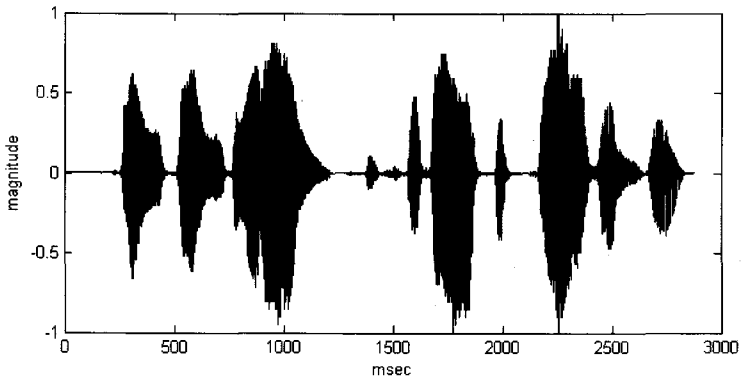
이들 GMM의 파라미터는 EM(Expectation and Maximization) 알고리즘을 이용하여 반복적으로 추정하게 된다[7][8]. 그리고 그림 2의 점선으로 표시된 가중치를 구하는 과정에서는 먼저 프레임의 SNR을 계산한 후 이를 근거로 프레임에서 출력되는 유사도에 적용할 가중치를 얻게 되는데, 이때 SNR과 가중치와의 변환관계는 여러 함수관계를 고려할 수가 있을 것이다. 논문의 실험에서는 우선 가장 간단한 방법인 상한과 하한이 있는 일차 선형변환 방법을 적용하였으며, 이의 변환관계는 다음 <그림 3>과 같다. 즉, SNR값이 5dB 이하인 음성프레임의 유사도 값은 가중치를 0으로 하여 배제하고, SNR값이 20dB 이상인 경우는 가중치를 1로 균일하게 적용하며, 5dB와 20dB 사이의 SNR을 갖는 프레임의 경우는 가중치를 SNR에 비례하여 0과 1사이의 값으로 정하는 것이다.



<그림 3> SNR과 가중치와의 변환 관계

4. 실험 및 결과

제안한 방법의 성능을 확인하기 위하여 ETRI에서 만든 한국어 화자인식용 휴대폰 음성DB를 사용하여 문맥종속 화자식별 실험을 하였다. 음성데이터의 샘플링 주파수는 8KHz이며, 8비트 μ-law PCM 방식으로 코딩되어 제공되었다. 그리고 DB의 전체 화자의 수는 남녀 모두 49명으로 구성하였으며, 화자 당 음성파일은 모두 20개로 이중 10개씩을 훈련용과 실험용으로 나누어 사용하였다. <그림 4>는 문맥종속 인식실험에 사용한 음성데이터의 파형으로 발성 시간이 약 3초 정도로 화자 모델 훈련에 사용된 음성데이터는 파일 10개를 합친 평균 약 30초 정도의 분량임을 알 수 있다. 실험에서 입력 음성데이터의 한 프레임은 40ms로 하였고, 20ms씩 중첩되어 처리되도록 하였으며, 음성의 특징벡터는 12차 Mel-cepstrum 계수와 로그



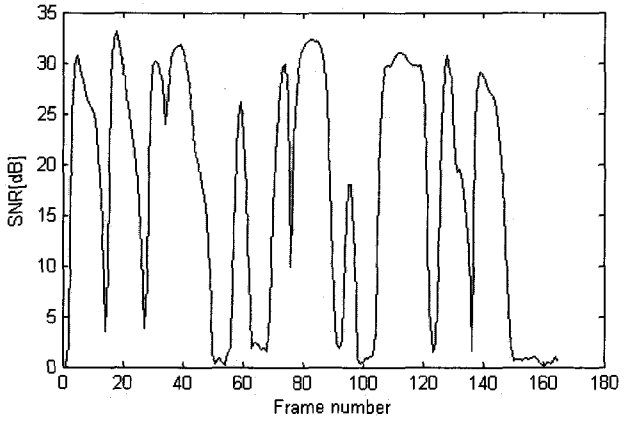
<그림 4> 실험에 사용한 음성신호의 파형

에너지를 포함하였고, 채널의 잡음을 보상하기 위하여 CMS 방법을 적용하였다. GMM 화자모델에 포함된 Gaussian 개수는 모두 10개이고, EM알고리즘에 의해 GMM 모델 λ_i의 파라미터를 반복적으로 훈련하여 구하였다. 이 과정에서 공분산 값은 full covariance를 사용하였고, 알고리즘의 초기과정에서는 fuzzy C-means clustering 방법을 사용하였다. 이에 대한 내용은 다음 <표 1>과 같다.

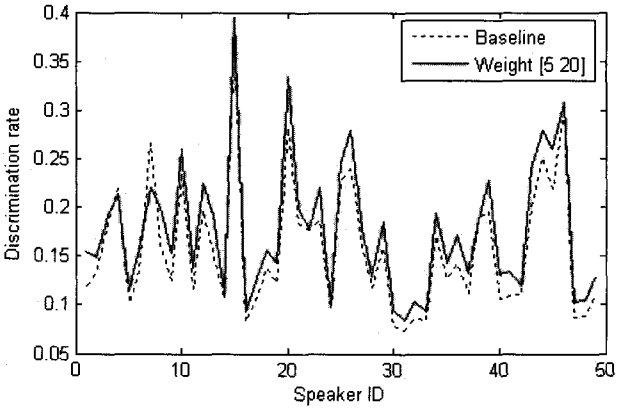
<표 1> 문장종속 화자식별 실험의 개요

음성 DB	ETRI 휴대폰 화자인식용 음성DB
Sampling rate/speech coding	8000 Hz / 8 bits μ-law PCM
화자수	49
화자당 Training 음성파일의 개수	10
화자당 Testing 음성파일의 개수	10
음성특징벡터	12차 mel-cepstrum과 log energy
Frame length/Frame shift	40ms/20ms
Channel compensation	Cepstral Mean Subtraction
GMM modeling	EM algorithm, full covariance
Gaussian mixture 개수	10

우선 화자모델을 구하기 위한 훈련과정에서는 휴대폰 음성 DB 원래의 상대적으로 깨끗한 음성데이터만을 사용하여 모델을 구하였고, 인식실험 과정에서는 화자식별 시스템의 잡음에 대한 영향을 평가하기 위하여, 입력 음성신호에 인위적으로 발생시킨 다양한 전력레벨의 Gaussian 잡음을 부가하여 실험을 진행하였다. 잡음을 부가하기 전에 원래의 음성신호는 1로 정규화를 하였으며, 아래의 <그림 5>는 평균 0이고 분산 값이 1인 Gaussian 잡음에 이득을 0.01로 하여 음성신호에 부가하여 만든 입력신호의 프레임 별 SNR 값의 변화를 보여주고 있다. 이때 프레임별 SNR을 구하는 과정에서 잡음전력은 인위적으로 부가한 Gaussian 잡음의 전력을 사용하고, 신호전력은 각 프레임의 음성신호의 전력을 사용하였다. 그러나 실제 인식기를 사용하는 환경에서는 잡음을 음성신호에서 분리하는 것이 불가능하므로, 이 경우에는 미리 시작 프레임에서 몇 프레임 정도의 구간을 묵음구간으로 설정하여 이 구간에서의 신호전력을 잡음전력으로 가정하는 방법을 사용할 수 있겠다. 다음으로 <그림 5>의 SNR 분포를 갖는 음성신호에 제안한 방법을 적용하였을 때, 화자분해능의 변화 여부를 살펴보았다. 프레임별 가중치는 <그림 3>의 변환 관계를 적용하여 프레임 SNR로부터 구하였으며, 이를 로그 유사도(log likelihood) 값에 곱한 다음에 이로부터 화자분해능을 계산하였다. 기존의 방법과 가중치를 적용한 제안한 방법에서의 화자분해능의 변화를 <그림 6>에 나타내었다.



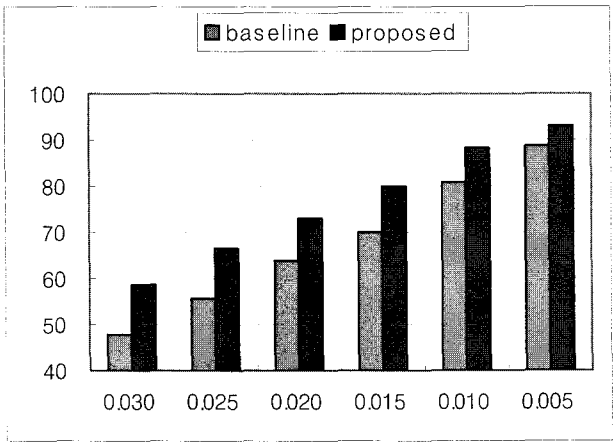
<그림 5> 잡음의 이득이 0.01인 음성의 프레임별 SNR 분포



<그림 6> SNR 가중에 의한 화자분해능의 비교

<그림 6>에서 점선은 기존의 방법을 사용한 경우이고, 실선은 논문에서 제안한 방법의 결과를 보이고 있으며, 전반적으로 기존의 경우보다 제안한 방법에서 화자분해능의 값이 좋아졌음을 확인할 수 있었다. 물론 화자분해능 값이 화자인식 시스템의 성능과 직접적인 비례관계에 있지는 않으나, 실험을 통해서 많은 상관도가 있음을 볼 수 있었다. <그림 7>은 여러 가지 이득 값을 갖는 잡음에 대하여 기존의 화자인식기와 제안한 방법을 적용한 화자인식기의 인식성능을 비교한 도표이다. 도표의 가로축은 잡음의 이득이며, 가로축은 화자식별기의 인식률을 백분율로 나타낸 것이다. 도표를 보면 전반적인 인식성능은 잡음이 적을수록 향상됨을 확인할 수 있고, 특히 제안된 방법에 의한 인식율의 향상 정도는 잡음의 이득이 클수록 개선이 많이 되어, 0.025의 경우에는 최대 11% 정도 인식률이 증가하였으

며, 최소는 잡음이 가장 적은 0.005에서 약 4.3% 정도 개선됨을 알 수 있었다. 이는 제안한 방법이 잡음이 많을 경우 더욱 효과적임을 알 수 있겠다.



<그림 7> 제안된 방법과 기존의 방법의 성능비교

마지막으로 SNR과 가중치의 변환에 따른 인식성능의 변화를 살펴보기 위하여 <그림 5>와 같이 SNR의 분포가 0~30dB로 분포하는 입력음성 데이터에 대해서 SNR의 상한과 하한 값을 바꾸어 가며 실험을 하여 다음 <표 2>와 같은 결과를 얻었다. 이때 실험에서의 잡음이득은 0.01로 고정하였으며, 상한과 하한 값의 변화에 따라 성능의 변화가 있음을 알 수 있으며, 전체적인 결과에서 인식 시스템의 최대의 인식률은 88.4%로 나왔으나, 이에 대한 상, 하한 값의 설정은 특별한 관계를 보이지 않고 대체로 하한은 5~10dB 내로 상한은 25~30dB 정도로 정하는 것이 적당할 것으로 보인다. 물론 이에 대한 것은 시스템의 입력데이터가 갖는 SNR 분포에 의해 영향을 받을 수밖에 없겠지만, SNR과 가중치의 변환 관계는 다양한 함수형태로 나타날 수 있을 것이며, 이에 대한 최적의 변환 관계를 추정하는 방법 및 그에 의한 시스템의 성능변화 등에 대한 연구는 앞으로 계속 진행될 것이다.

<표 2> SNR 가중의 상한 값과 하한 값에 따른 인식성능 비교

하한 \ 상한	0	5	10	15	20	25	30	35
0	81.0	85.3	86.1	87.8	88.0	88.2	**88.4**	88.0
5	-	85.9	87.8	88.2	88.2	88.0	87.8	87.3
10	-	-	**88.4**	**88.4**	88.0	**88.4**	87.1	86.7
15	-	-	-	88.0	88.2	85.9	85.3	84.5
20	-	-	-	-	85.7	83.3	80.8	79.0

5. 결론 및 향후 과제

본 논문에서는 화자인식에서 훈련과정과 실험과정에 사용하는 음성데이터의 통계적 특성의 불일치에 의해 생기는 인식성능 저하에 대한 문제를 고찰하였다. 특히 실험과정에서 인식시스템의 입력으로 들어오는 음성신호의 경우에는 주변 환경, 전화채널 그리고 시스템 등의 영향으로 다양한 잡음이 유입되어 음질이 나빠지게 되어 인식시스템의 성능에 큰 영향을 주게 된다. 논문에서는 음성신호의 음질, 곧 SNR이 화자분해능과의 상관도가 있음을 이용하여, 프레임의 SNR에 따라 인식기의 출력으로 나오는 프레임의 유사도 값을 선별적으로 가중하여 인식 성능을 향상시키는 방법을 제안하였다. 이때 SNR과 가중치와 관계는 상한과 하한이 있는 일차함수의 형태로 만들어서 사용하였다. 제안한 방법의 성능을 평가하기 위하여 문맥종속 화자식별시스템을 ETRI의 한국어 휴대폰 화자인식용 음성DB에 대해 적용하였다. 실험에 사용한 총 화자는 남녀 합쳐 49명이며, 화자마다 10개씩의 훈련용과 실험용 음성파일을 사용하였다. 입력음성에 부가되는 잡음의 정도에 따른 성능을 살펴보기 위하여, 여러 잡음 레벨에서 실험을 한 결과, 대체로 잡음이 많은 경우일수록 제안한 방법에 의한 방법이 기존의 방법보다 높은 성능 향상을 보여, 잡음이득이 0.025일 때, 최대 11%차로 개선되었고, 잡음이득이 가장 작은 0.005에서는 약 4.3% 정도 성능개선이 있음을 확인하여, 제안한 방법이 화자인식에서 효과적으로 사용될 수 있음을 보여 주었다. 또한 SNR을 가중치로 변환하는 함수의 상한 값과 하한 값의 범위를 여러 경우에 대해서 적용한 실험에서는 입력 음성데이터의 SNR의 분포가 0~30dB인 경우에 대체로 하한은 5-10dB 정도이고 상한은 25-30dB 정도가 적당한 것으로 판단된다. 그러나 위의 실험에서 SNR의 상, 하한 값을 어떻게 정하는 가에 따라 시스템의 인식성능에 많은 변화가 생기는 것으로 보아, 앞으로 SNR에서 가중치를 구하는 변환함수의 가장 적절한 형태와 어떠한 방식으로 이를 유도하는 문제에 대한 연구가 계속되는 것이 필요하다고 본다.

참 고 문 헌

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

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

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

[4] C.-H. Sit, M.-W. Mak, S.-Y. Kung, "Maximum likelihood and maximum a posteriori adaptation for distributed speaker recognition systems", *Proc. 1st International Conference on Biometric Authentication*, 2004.

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

[6] D. A. Reynolds, *A gaussian mixture modeling approach to text-independent speaker identification*, Ph.D thesis, Georgia Institute of Technology, 1992.

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

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

접수일자 : 2007년 3월 13일
게재결정 : 2007년 3월 23일

▶ 최홍섭(Hong Sub Choi)
주소: 487-711 경기도 포천시 선단동 대진대학교 공과대학 전자공학과
소속: 대진대학교 공과대학 전자공학과
전화: 031) 539-1903
Fax : 031) 539-1900
E-mail: hschoi@daejin.ac.kr