화자식별을 위한 파라미터의 잡음환경에서의 성능비교*

Parameters Comparison in the speaker Identification under the Noisy Environments

최 홍 섭**
(Hong-Sub Choi)

ABSTRACT

This paper seeks to compare the feature parameters used in speaker identification systems under noisy environments. The feature parameters compared are LP cepstrum (LPCC), Cepstral mean subtraction(CMS), Pole-filtered CMS(PFCMS), Adaptive component weighted cepstrum(ACW) and Postfilter cepstrum(PF). The GMM-based text independent speaker identification system is designed for this target. Some series of experiments show that the LPCC parameter is adequate for modelling the speaker in the matched environments between train and test stages. But in the mismatched training and testing conditions, modified parameters are preferable the LPCC. Especially CMS and PFCMS parameters are more effective for the microphone mismatching conditions while the ACW and PF parameters are good for more noisy mismatches.

Keywords : speaker identification, GMM, LPCC, CMS, PFCMS, ACW, PF

1. 서 론

컴퓨터의 발달로 현재의 화자인식 기술은 실험실 환경에서와 같이 잡음이 적은 곳에서는 매우 좋은 성능을 보이고 있다. 그러나 이러한 화자인식 시스템을 실제의 현장에서 사용할 경우에는 인식성능이 매우 현격하게 떨어지는 것을 알 수 있다. 이는 주로 화자를 등록하기 위해서 화자인식 시스템을 훈련시킬 때의 환경과 실제로 인식을 위해 인식기를 사용하는 환경 사이의 차이에 의해 발생하는 것이다. 예로 전화를 이용한 화자인식의 경우에, 어떤 이용자가 일반 전화기를 이용하여 자기의 음성을 화자인식 시스템에 등록한 후에, 나중에 본인임을 확인시키려고 휴대폰을 이용하여 화자인식 시스템에 접근하게 되면 인식률이 현저히 저하된다. 이는 실제 화자인식을 수행할 때와 화자 등록할 때의 환경의 차이, 즉 전화채널, 수화기의 종류, 배경 잡음 그리고 주변 환경에 의한 화자 음성의 변화 등에서 많은 차이가

* 이 논문은 1998학년도 대진대학교 학술연구비지원에 의한 것입니다.
** 대진대학교 이공대학 전자공학과

있기 때문이다.

이와 같이 화자인식을 수행하는 실제의 환경은 매우 다양하고 예측할 수 없으므로 이에 대한 방안으로 강인한 화자인식 방법들이 제안되고 있는데 이들을 세 종류로 분류하면 다음과 같다. 첫째는 음향학적인 단계에서 음질 향상 방법을 이용하여 입력 음성신호의 SNR을 높이는 방법이며[1], 둘째는 잡음이 섞인 음성을 인식하기 위하여 깨끗한 음성과 잡음을 결합한 혼합 모델링을 이용하는 방법이 있고[2], 마지막 방법으로는 입력 음성에서 화자인식에 사용할 특징 파라미터를 추출할 때, 잡음에 강인한 파라미터를 구하고자 하는 노력이 그것이다[3][4].

본 논문에서는 세 번째 접근방법에 주안점을 두고, 잡음 환경에서 여러 음성 파라미터들이 화자인식기의 성능에 어떠한 영향을 주는 가를 비교, 검토하고자 한다. 특히 음성에 부가된 잡음과 녹음에 사용한 마이크의 차이에 의해 발생하는 문제에 대처하는 파라미터들의 성능을 비교하였다. 이를 위해 GMM(Gaussian Mixture Model)을 이용한 문장독립 화자식별 시스템을 사용하였고, 비교할 음성 파라미터들은 기본적인 LPC 켑스트럼(LPCC)을 포함하여 Cepstral mean subtraction(CMS), Pole-filtered cepstral mean subtraction(PFCMS), Adaptive component weighted cepstrum(ACW), 그리고 Postfilter cepstrum(PF) 등이다.

논문의 구성은 다음 2장에서 실험에 사용한 화자인식 알고리즘인 GMM에 대한 간단한 설명을 하였으며, 3장에서는 잡음에 강인한 파라미터에 대한 기본적인 개념을 설명하였으며, 이어 4장에서 인식실험과 그 결과를 그리고 마지막 5장에서는 결론을 기술하였다.

2. 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[5]. 다음은 GMM을 어떻게 정형화하고 그 파라미터는 어떻게 추정하며 그 결과를 화자인식에 어떠한 방법으로 사용하는 지를 간략히 설명한다.

(1) 모델 파라미터

GMM의 혼합확률분포는 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.

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

화자인식을 위해서는 각각의 화자들을 위의 λ를 이용해 GMM으로 모델링할 수 있다. GMM은 공분산 행렬을 어떻게 구성하는가에 따라서 몇 가지 형태로 나눌 수 있다. 즉, 화자 모델의 각 가우시안 부분분포 마다 서로 다른 공분산 행렬을 가지는 경우(nodal covariance)와 각 화자 모델 당 하나의 공분산 행렬을 가지는 경우(grand covariance), 마지막으로 모든 화자 모델에 대해서 동일한 공분산 행렬을 가지는 경우(global covariance)인데 일반적으로 nodal covariance를 가지고 모델을 만들 경우에 가장 좋은 결과를 내는 것으로 알려져 있으며[5], 본 논문의 실험에서도 nodal covariance를 사용했다.

(2) 파라미터 추정

주어진 학습 데이터로부터 λ로 표현되는 GMM 모델의 파라미터들을 추정하는 것이다. 파라미터를 추정하는 데에는 여러 가지 방법이 있으나 가장 많이 사용되며 널리 알려져 있는 방법은 *maximum likelihood*(ML) 추정방법이다. ML 추정의 목표는 주어진 학습 데이터로부터 GMM의 우도함수(likelihood function)를 최대로 하는 파라미터 λ를 찾는 것이다. 만약 T개의 학습 데이터 $X=\{\vec{x_1}, \ldots, \vec{x_T}\}$가 주어져 있다면 GMM의 우도함수는 식(5)과 같이 표현되는데

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x_t}|\lambda) \quad (5)$$

이 식은 λ에 대한 비선형 관계식이므로 직접 최대값을 갖는 추정식을 얻기는 불가능하다. 하지만 EM(*expectation maximization*) 알고리즘을 사용하면 ML 파라미터를 반복적으로 추정할 수 있다. EM 알고리즘의 기본적인 생각은 주어진 초기 모델 λ로부터 $p(X|\overline{\lambda}) \geq p(X|\lambda)$를

만족하는 새로운 모델 $\bar{\lambda}$를 추정하는 것이다. 이 새로운 모델은 다음 반복과정에서 다시 초기 모델값으로 사용되며 이러한 과정은 특정 수렴조건을 만족할 때까지 반복적으로 한다. EM 알고리즘을 반복적으로 사용하여 모델의 likelihood 값을 단조증가시키는 재추정 관계식은 다음 식(6)과 같으며 결합가중치, 평균 그리고 분산 순으로 정리했다.

$$\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda) \tag{6a}$$

$$\vec{\bar{\mu}}_i = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)\,\vec{x}_t}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} \tag{6b}$$

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)\,x_t^2}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} - \bar{\mu}_i^2 \tag{6c}$$

그리고 i번째 성분에 대한 사후확률은 식 (7)과 같이 주어진다.

$$p(i|\vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \tag{7}$$

GMM 화자 모델 파라미터 추정에서 아주 중요한 요소는 mixture 모델의 차수 M과 EM 알고리즘의 초기 모델 파라미터의 선택이다. 그것들을 선택하는 데에는 기준이 될만한 이론이 없기 때문에 주어진 문제에 대해서 가장 적합한 값들을 실험적으로 찾을 수밖에 없다. 이 논문에서는 모델의 차수를 M=32로 하고 모델 파라미터의 초기값은 K-means 알고리즘을 이용하여 특징벡터값들을 대표적인 32개의 묶음으로 분류한 다음, 이런 묶음으로부터 각 부분 확률분포함수의 평균, 분산과 가중치들을 산출하였다.

3. 화자인식 파라미터

(1) Cepstrum (LPCC)

LP 캡스트럼 계수(LPCC) $c(n)$은 선형예측계수 a_i로부터 다음의 관계식을 이용하여 계산한다.

$$c(n) = a_n + \sum_{i=1}^{n-1}\left(\frac{i}{n}\right)c(i)a_{n-i} \tag{8}$$

여기서 $c(n)$은 무한대의 구간에서 존재하며, n의 값이 커질수록 계수는 $1/|n|$에 비례해서

작아지고, 대신 n이 값이 작을수록 계수의 중요성은 커진다. 따라서 켑스트럼의 차수를 p로 정하면, $c(1)$에서부터 $c(p)$까지의 계수만을 사용하게 된다. 이렇게 전체 켑스트럼의 일부만을 사용하여도 물리적으로는 파라미터가 추출된 음성구간의 스펙트럼 포락선에 관한 정보를 갖고 있음을 보여준다. 즉, 두 켑스트럼 계수의 차의 제곱의 평균은 해당하는 음성구간들의 스펙트럼 포락선의 차이를 나타내는 좋은 척도가 된다[6].

(2) Cepstral mean subtraction (CMS)

전화망을 통한 음성신호는 $T(z)=S(z)H(z)$로 표현할 수 있다. 여기서 $S(z)$는 깨끗한 음성, $H(z)$는 전화망의 전달함수 그리고 $T(z)$는 전화음성을 의미한다. 위의 식에 로그함수를 취하면

$$\log T(z) = \log S(z) + \log H(z) \tag{9}$$

이 되고, 이를 역변환하면 전화음성의 켑스트럼은 $c_T(n)=c_S(n)+c_H(z)$로서 깨끗한 음성의 켑스트럼 $c_S(n)$과 전화망의 영향인 $c_H(z)$의 합으로 표현된다. 그리고 깨끗한 음성의 켑스트럼의 평균이 영이라고 가정하면, 전화망의 켑스트럼의 값은 전화음성의 켑스트럼의 평균값으로 추정할 수 있음을 알 수 있다. 따라서 전화음성에서 전화망의 영향을 제거한 켑스트럼 $c_{cms}(n)$은 다음과 같이 나타낼 수 있다[3].

$$c_{cms}(n) = c(n) - E[c(n)] \tag{10}$$

여기서 $c(n)$은 전화음성의 켑스트럼 계수이며, 평균은 전체 음성구간에 대하여 계산한다. CMS는 화자를 등록시킬 때와 인식실험을 할 때의 전화음성이 통과한 전화망이 서로 다른 경우에는 확실히 인식성능이 향상되나, 그렇지 않고 동일한 전화망을 사용한 음성의 경우에는 인식성능이 떨어짐을 알 수 있다. 녹음에 사용하는 마이크의 경우도 음성에 선형필터링의 효과를 준다고 간주하면 이러한 CMS 방법은 서로 다른 마이크에 의한 인식률 저하를 보상할 수 있겠다.

(3) Pole-filtered cepstral mean subtraction (PFCMS)

LP 분석에 의해서 구한 극점중 단위원에 가까이 있는 것일수록 대역폭이 좁은 포먼트 성분이 되며 이는 채널과 잡음의 영향에 덜 민감하다. 즉 단위원에 가까이 있는 극점이 보다 많은 음성의 특성을 갖고 있다는 것인데, 이는 CMS 방법에서 전체 프레임의 켑스트럼의 평균으로 채널추정을 하는 경우에는 오히려 추정치의 정확도를 떨어뜨리는 원인임을 알 수 있다. 따라서 PFCMS는 이러한 음성의 포먼트에 해당하는 극점을 인위적으로 단위원에서 멀리 위치시킴으로서, 대역폭이 넓은 극점으로 변형시켜서 음성의 특성은 약화시키고 채널의

특성을 부각하여 CMS보다 상대적으로 정확한 채널특성을 구하는 방법이다. 여기서 극점을 필터링(수정)하는 방법은 극점의 크기가 미리 정해 놓은 임계값(r_{th})보다 큰 경우에, 즉 단위원에 가까이 존재하는 극점의 크기를 임계값으로 대치하는 것이다[7]. PFCMS 캡스트럼은 다음 식으로 계산할 수 있다.

$$c_{pfcms} = c(n) - E[c_m(n)] \tag{11}$$

여기서 $c_m(n)$은 필터링된 극점으로부터 구한 켑스트럼이며 이를 전체 프레임에 대해 평균을 구하면 보다 정확한 채널특성을 추정할 수 있다. 이를 원래의 켑스트럼으로부터 감산한 것이 PFCMS 켑스트럼이다[7].

(4) Adaptive component weighted cepstrum (ACW)

음성의 선형예측모델로부터 구한 음성의 전달함수는 다음과 같다.

$$\begin{aligned} H(z) &= \frac{1}{A(z)} = \frac{1}{1-\sum_{i=1}^{p} a_i z^{-1}} \\ &= \prod_{i=1}^{p} \frac{1}{1-z_i z^{-1}} = \sum_{i=1}^{p} \frac{r_i}{1-z_i z^{-1}} \end{aligned} \tag{12}$$

이때 a_i는 선형예측계수, z_i는 극점이며 r_i는 잉여값(residues)이다.
위의 전달함수의 마지막 표현식에 들어있는 파라미터는 잉여값 r_i와 극점 z_i인데, 극점에 비해 잉여값 r_i가 음성이 전화채널과 같은 필터를 통과했을 때, 많은 변화를 보여준다고 알려져 있다[8]. 즉, 잉여값 r_i는 전화채널의 변화에 대해 매우 민감하게 영향을 받고 있음을 보여주는 것이다. 따라서 ACW는 이러한 잉여값 r_i를 정규화시켜서 채널의 영향을 줄이는 방법으로 다음과 같이 정규화된 식을 갖는다.

$$H_{acw} = \sum_{k=1}^{p} \frac{1}{1-z_k z^{-1}} = \frac{N(z)}{1+\sum_{i=1}^{p} a_i z^{-1}} = \frac{N(z)}{A(z)}, \tag{13}$$

여기서 $N(z) = \sum_{k=1}^{p} \prod_{i=1 \neq k}^{p} (1 - z_i z^{-1}) = p\left(1 + \sum_{i=1}^{p-1} b_i z^{-i}\right)$

이때 b_i는 a_i로부터 다음 관계식을 이용하여 간단하게 구할 수 있다.

$$b_i = \left(\frac{p-i}{p}\right) a_i, \text{ 여기서 } 0 \le i \le p \tag{14}$$

최종의 ACW 켑스트럼을 구하면 다음과 같다.

$$c_{acw}(0) = \log p$$
$$c_{acw}(n) = c_{lp}(n) - c_{nn}(n), \quad n>0 \tag{15}$$

위의 식에서 $c_{nn}(n)$은 채널 또는 잡음의 영향으로 생긴 켑스트럼의 변화량을 추정한 값으로 다음의 관계식을 이용하여 계산한다.

$$c_{nn}(n) = -b_n - \sum_{k=1}^{n-1}\left(\frac{n-k}{n}\right)b_k c_{nn}(n-k) \tag{16}$$

위의 식에서 $n-k=i$로 치환하여 다시 쓰면,

$$c_{nn}(n) = -b_n - \sum_{i=1}^{n-1}\frac{i}{n}c_{nn}(i)\,b_{n-i} \tag{17}$$

따라서 $c_{nn}(n)$을 구하는 위 식(17)은 앞에서 사용한 선형예측계수 a_i로부터 켑스트럼을 구하는 식(8)에서 단지 a_i를 $-b_i$로 바꾼 것을 제외하고는 같은 식임을 알 수 있다[3].

(5) Postfilter cepstrum(PF)

이 방법은 사람의 청각특성에 의하면 스펙트럼의 포먼트 부분이 스펙트럼의 계곡 부분보다 잡음에 대해서 상대적으로 영향을 덜 받는다는 사실에 근거한다. 이 PF 켑스트럼의 유도식은 다음과 같다[3].

$$c_{pf}(0) = 0, \quad c_{pf}(n) = c(n)[\alpha^n - \beta^n], \quad n>0 \tag{18}$$

이때 $0<\beta<\alpha\leq 1$이다. 따라서 이것은 LP 켑스트럼에 차수에 따른 가중치를 부과한 것이며 식에서 보듯이 낮은 차수의 켑스트럼의 계수보다 높은 차수의 켑스트럼의 계수를 강조해 주고 있다. PF 켑스트럼은 채널과 잡음에 대해서 강인한 특성을 보여준다.

4. 실험 및 결과

화자인식에 사용하는 음성의 특징 파라미터들이 부가된 잡음과 마이크의 필터링 효과에 대해서 얼마나 강인한가를 비교하는 실험을 하였다. 이를 위하여 기준 파라미터로 12차의 LPC 켑스트럼을, 그리고 이를 잡음 환경에 강인하도록 변형한 4가지의 다른 파라미터를 사용하였다. 화자식별 시스템은 GMM을 이용하여 구성하였으며, 이때 GMM의 혼합 가우시안 분포의 개수는 M=32로 하였다.

본 실험에서는 20대의 20명 화자(남자 16명, 여자 4명)를 대상으로 화자등록용으로 60초

정도 그리고 인식실험용으로 15초 분량의 음성을 4번씩 녹음하였는데, 대상 문장은 심리학 전공서적의 내용으로 모든 음성데이터의 내용은 서로 다른 문장을 사용하였다. 그리고 녹음에는 AKG D190와 SHURE SM58 두 종류의 마이크를 이용하여 마이크의 특성 차이를 고려하였다. 따라서 화자 개인당 마이크 각각으로 녹음하여 총 2분 정도씩의 음성을 실험실 환경에서 녹음하였다. 음성데이터는 8 kHz 샘플링하여 16비트로 저장하였으며 음성의 한 프레임의 길이는 20 ms로 160샘플/프레임이며, 10 ms씩 중첩하였다. 그리고 끝점검출과 전처리과정을 거쳐서 특징벡터를 추출하였다.

잡음에서의 성능을 비교하기 위하여 마이크로 녹음한 원음에 백색가우시안 잡음을 첨가하여 여러 SNR값을 갖는 음성데이터를 만들었다. 모든 특징벡터의 차수는 12차로 일치시켰으며 PFCMS 파라미터 추출에서 극점을 필터링하는 크기 임계값(r_{th})은 0.85로 하였고, PF 파라미터에서는 $\alpha=1$, $\beta=0.9$를 사용하였다.

먼저 화자모델을 훈련할 때와 다른 잡음환경에서 인식실험을 할 때의 인식기의 성능이 얼마나 나빠지는 지를 확인하기 위하여 서로 다른 SNR을 갖는 음성데이터를 가지고 실험을 하였으며, 결과는 표 1에 요약했으며 이때 파라미터로는 LPCC를 사용하였다.

표 1. 부정합 잡음 환경에서 인식성능 비교 (LPCC 사용)

훈련 \ 실험	원음	20 dB	15 dB	10 dB	5 dB
원음	98.8	15.0	15.0	15.0	15.0
20dB	27.5	96.3	92.0	28.8	15.0
15dB	16.3	78.8	96.3	66.3	18.8
10dB	15.0	18.8	33.8	91.3	28.8
5dB	15.0	12.5	15.0	13.8	85.0

표 1에서 보듯이 훈련시와 인식실험시의 SNR 값이 정합된 상태의 인식 결과는 대각선 방향으로 98.8-85.0 % 사이의 값을 갖고 상대적으로 인식률이 높으나, 대각선 외의 항들은 서로 부정합된 환경에서의 인식률로 대각선에서 멀어질수록, 즉 부정합되는 정도가 클수록 인식률은 급격히 감소함을 보여주고 있다.

정합환경에서 SNR가 다른 음성들을 이용해서 인식성능에 미치는 파라미터의 영향을 비교하여 표 2에 보였다. 결과로부터 정합된 환경에서는 대체로 원래의 파라미터인 LPCC를 사용하는 것이 이를 변형한 다른 파라미터에 비해 성능이 좋았음을 알 수 있다. 그러나 잡음이 많이 부가될수록 ACW와 PF 파라미터의 경우는 인식률이 LPCC보다 개선되었다.

표 2. 정합 잡음 환경에서의 파라미터 성능 비교

파라미터	원음	20 dB	15 dB	10 dB	5 dB
LPCC	98.8	96.3	96.3	91.3	85.0
CMS	96.3	92.5	90.0	76.3	70.0
PFCMS	96.3	92.5	86.3	80.0	73.8
ACW	98.8	96.3	96.3	93.8	87.5
PF	98.8	96.3	96.3	92.5	85.0

부정합 환경에서의 파라미터의 효과를 살펴보려고 원음의 데이터로 훈련을 하고 잡음이 섞인 20 dB 음성데이터를 갖고 인식실험을 하였다. 표3에서 보면 부정합 상태에서는 전반적으로 인식률이 급격히 떨어지는 것을 알 수 있으나, 상대적으로 LPCC에 비해서 ACW와 PF 파라미터가 좋은 결과를 보여주는데, 특히 PF의 경우에는 약 12.5 % 정도의 인식률 향상이 있었다. CMS와 PFCMS는 예상한대로 잡음에 대해서는 원래의 LPCC보다 떨어진 인식률을 보였다.

표 3. 부정합 잡음 환경에서의 파라미터 성능비교

원음으로 훈련, 20 dB 음성으로 인식 실험				
LPCC	CMS	PFCMS	ACW	PF
27.5	20.0	21.3	28.8	40.0

다음으로 마이크의 차이에 의한 인식률의 차이는 표4에 나타난다.

표 4. 부정합 인식 환경에서의 파라미터 성능비교

마이크 M1으로 훈련, M2로 인식 실험(원음)				
LPCC	CMS	PFCMS	ACW	PF
91.3	97.5	97.5	92.5	97.5

마이크 M1(AKG D190)와 마이크 M2(SHURE SM58) 각각으로 녹음한 깨끗한 음성데이터를 대상으로 인식실험을 하였다. LPCC의 경우에는 마이크의 차이로 인해 정합인 경우(98.8 %)에 비해 7.5 % 정도 인식률이 떨어지는데 비해 CMS와 PFCMS의 결과는 97.5 %로 향상된 것을 알 수 있다. 이는 CMS와 PFCMS가 채널의 영향에 효과가 있음을 보여주는 것이다. 이 실험에서도 PF의 파라미터는 좋은 결과를 보이고 있어서 잡음과 채널 모두에 효과적임을 알 수 있었다.

5. 결 론

본 논문에서는 화자인식에 미치는 잡음의 영향을 살펴보고 그 동안 제안되었던 여러 LPCC 계열의 변형 파라미터를 사용하여 그들의 성능을 비교하였다. 인식환경의 변화의 요인으로는 음성에 삽입되는 일반적 부가잡음과 마이크나 전화채널에 의한 채널 필터링 특성을 들 수 있는데, 논문에서는 부가잡음과 마이크에 의한 채널의 영향만으로 국한하였다. 그리고 화자인식기는 화자모델 훈련시와 다른 환경에서 인식실험을 할 경우, 즉 부정합 환경에서는 인식률이 떨어지는데 이에 대한 실험을 위하여 인위적으로 녹음된 음성데이터에 가우시안 백색 잡음을 부가하여 서로 다른 SNR을 갖는 데이터를 만들었다. 실험에는 GMM을 이용한 문장독립 화자인식기를 사용하였으며, 화자의 수는 20명이었다. 화자의 수가 적어서 결과로 나온 인식률이 통계적인 신뢰도면에서 충분치는 않았지만 전체적인 경향은 살펴 볼 수 있다고 본다.

실험 결과를 보면, 일반적으로 잡음의 영향이 화자인식에 많은 걸림돌이 되는 것을 알 수 있는데, 특히 부정합 환경에서는 약간의 SNR 차이도 인식률을 급격히 감소시킴을 알 수 있어서 인식환경의 부정합에 대한 대처 방안이 필요함을 알 수 있었다. 실험에 의하면 정합환경에서는 LPCC가 변형된 다른 파라미터들에 비해 충분히 화자의 특성을 모델링하고 있음을 볼 수 있었다. 그러나 부정합 환경에서는 잡음과 마이크 채널의 영향에 의해 인식률이 떨어지는데, 이에 반해 변형된 파라미터는 보다 효과적임을 보여준다. 이를 잡음과 마이크에 대해 따로 생각하면, 잡음에 대해서는 ACW와 PF 파라미터가 더욱 효과적이며 마이크와 같은 채널의 변화에는 CMS와 PFCMS 파라미터가 상대적으로 좋은 결과를 보여주는데 이러한 결과는 다른 논문에서도 확인할 수 있었다[8].

참 고 문 헌

[1] J. Ortega-Garcia et al., 1996. "Overview of speech enhancement techniques for automatic speaker recognition," *Proc. ICSLP-96*, vol. 2, 929-932.

[2] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds. 1994. "Integrated models of signal and background with application to speaker identification in noise," *IEEE trans. on Speech and Audio processing*, 245-257, Apr.

[3] R. J. Mammone, X. Zhang and R. P. Ramachandran, 1996. "Robust speaker recognition, a feature-based approach," *IEEE signal processing magazine*, 58-71, Sep.

[4] V. Ramanujam, R. Balchandran and R. J. Mammone, 1999. "Robust speaker verification in noisy conditions by modification of spectral time trajectories," *Eurospeech-99*, vol. 2, 791-794.

[5] D. A. Reynolds and R. C. Rose. 1995. "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE trans. on Speech and Audio processing*, vol. 3, no. 1, 72-83, Jan.

[6] L. R. Rabiner and B. H. Juang, 1993. *Fundamental of speech recognition*, Prentice-Hall, Englewood Cliffs, NJ.

[7] D. Naik, 1995. "Pole-filtered cepstral mean subtraction," Proc. ICASSP-95, vol. 1, 157-160.

[8] K. T. Assaleh and R. J. Mammone, 1994. "New LP-derived features for speaker identification," *IEEE trans. on Speech and Audio processing*, 630-638, Oct.
[9] Carlos Alonso-Martinez and Marcos Faundez-Zanuy. 2000. "Speaker identification in mismatch training and testing conditions," *Proc. ICASSP-00*, vol. 2, 181-1184.

접수일자: 2000. 7. 26.
게재결정: 2000. 9. 4.

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