전화음성에 강인한 문장종속 화자인식에 관한 연구*

On a robust text-dependent speaker identification over telephone channels

정 의 상 · 최 홍 섭**

(Eu-Sang Jung · Hong-Sub Choi)

ABSTRACT

This paper studies the effects of the method, CMS(Cepstral Mean Subtraction), (which compensates for some of the speech distortion caused by telephone channels), on the performance of the text-dependent speaker identification system. This system is based on the VQ(Vector Quantization) and HMM(Hidden Markov Model) method and chooses the LPC-Cepstrum and Mel-Cepstrum as the feature vectors extracted from the speech data transmitted through telephone channels. Accordingly, we can compare the correct recognition rates of the speaker identification system between the use of LPC-Cepstrum and Mel-Cepstrum. Finally, from the experiment results table, it is found that the Mel-Cepstrum parameter is proven to be superior to the LPC-Cepstrum and that recognition performance improves by about 10% when compensating for telephone channel using the CMS.

Keywords: speech distortion, speaker identification, Mel-Cepstrum, Hidden Markov

1. 서 론

현재까지 개발된 화자인식기술은 실험실에서 잡음이 적은 환경에서 얻은 양질의 음성을 대상으로 한 제한적인 실험에 불과하며, 상용화할 수 있는 인식률을 갖는 시스템도 사실은 오염되지 않은 음성을 대상으로 제한적인 응용범위에서만 사용되고 있다. 이와 같이 화자인식시스템은 주변환경의 변화와 잡음이 심한 경우에 적용하였을 때는 상당한 인식률의 저하를 가져오기 때문에 상용화에 큰 어려움이 있다. 특히 전화망을 통한 음성에 대하여 화자인식을 하는 시스템에서는 전화망에 의해 통과음성에 미치는 영향으로 인식기의 성능이 많이 저하됨을 알 수 있다. 최근 들어 주변 환경과 잡음을 고려한 강인한 음성인식에 관한 연구가 활발히 진행되고 있으며 이는 음성 및 화자인식시스템을 상용화하는데 있어 매우 중요한 기반기술이다. 본 논문에서는 전화음성에 실린 전화망의 영향을 제거하기 위해 기존의 제안된 알고리즘들 중에서 가

* 이 논문은 1996년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.

** 대진대학교 공과대학 전자공학과

장 뛰어난 성능을 보이는 것으로 보고된 켑스트럼 평균차감법을 이용하여 숫자음을 대상으로 하는 문장종속 화자인식시스템을 구현하여 실험하였다. 화자인식방법은 VQ와 HMM을 이용하였으며, 사용한 음성신호의 특징벡터는 LPC 켑스트럼과 Mel 켑스트럼 두 개의 특징벡터로서 각각의 성능을 비교하였다. 본 논문의 구성은 1장 서론에 이어 2장에서 전화음성이 화자인식기의 성능을 저하시키는 원인에 대하여 살펴보고, 이에 대한 처리방법을 지금까지 제안된 방법들을 중심으로 고찰한다. 그리고 3장에서는 화자인식에 사용하는 특징벡터와 전화채널에 의해 생긴 왜곡을 보상하는 방법으로 켑스트럼평균 차감법을 설명하였으며, 4장에서 실험에 사용한 화자인식시스템의 구성을 그리고 5장에서는 실험 및 결과를 보여준다. 화자인식의 결과표에서 보듯이 전화채널의 특성을 보상한 음성신호의 경우에는 보상 안한 경우에 비해 성능이 약 10 % 정도 향상되었음을 확인할 수 있었다.

2. 전화음성의 분석

1) 전화선로에 의한 영향

R. Stern은 깨끗한 음성신호와 전화음성을 인식기에 각각 사용하여 전화망에 의해서 인식률이 나빠지는 요인을 비교 분석하였다[1]. 실험에 의하면 전화선로에 의해서 생기는 통과대역폭의 제한이나 선형필터링에 의한 영향 뿐 만 아니라, 다른 복합적인 요소에 의해서도 상당 부분의 인식오류를 발생시킴을 알 수 있다. 즉, 전화음성의 낮은 주파수(약 180 Hz)에 존재하는 톤신호, 음성이 전화망을 통과할 때 부가적으로 유입되는 정상상태의 잡음과 간혹 발생하는 스파이크성 임펄스 잡음, 또한 다른 통화자와의 혼신음에 의한 왜곡, 그리고 진폭과 위상 지터 등의 원인들을 들 수 있는데 특히 이중에서도 톤신호, 정상잡음과 임펄스 잡음 등은 인식률에 많은 영향을 주는 요소들이다. 따라서 화자인식 시스템의 주파수 전처리과정에서 이를 반영하여 음성신호의 켑스트럼의 주파수대역을 200 Hz - 3700 Hz까지로 제한하여 사용할 경우에 상당한 성능향상을 보인다. 또한 전화채널의 전달함수의 진폭특성에 의하여 생기는 영향은 주로 진폭특성이 평탄치 못한 주파수구간에 의해 크게 영향을 받지만, 위상특성에 의한 성능저하는 거의 없음을 알 수 있다.

2) 전화음성의 처리방법

(1) RASTA(RelAtive SpecTrAl processing): 음성신호 내에서 시간에 따라 천천히 변화하는 성분은 배제하고 대신 빠르게 변화하는 성분의 특징을 반영해 주는 파라메타 추출방법으로 RASTA 방법이 제안되었다[2][3][4]. 이 방법은 일반적인 짧은 구간 스펙트럼 대신 스펙트럼성분중 시간에 따라 천천히 변화하는 성분을 배제하는 대역통과 스펙트럼을 사용한다. 필터링 블록은 각 주파수대역을 IIR필터를 사용하여 대역통과 필터링하는 것과 같다.

(2) 스펙트럼 차감법(Spectral Subtraction Method): 스펙트럼 차감법은 주변 잡음에 의해 손상된 음성스펙트럼에서 잡음의 성분만을 제거하는 방법이다[5][6]. 이러한 스펙트럼 차감법은

배경잡음의 스펙트럼 형태를 미리 알고 있거나, 잡음의 스펙트럼을 추정하기에 충분한 묵음구간(약 300 msec)이 있어야 한다. 또한 배경잡음이 최소한 부분적으로 안정된 특성을 가져야 하며, 통계적 특성이 서서히 변화하는 환경에서는 음성이 존재하는 구간과 잡음만이 존재하는 구간을 검출할 수 있는 방법이 필요하다.

(3) 켑스트럼평균 차감법(CMS: Cepstral Mean Subtraction): 켑스트럼 평균 차감법은 전체 구간에 대하여 켑스트럼의 평균을 구하고, 이를 음성의 켑스트럼에서 차감하므로써 채널의 영향을 제거하는 방법이다[7]. 이에 대한 내용은 3장에서 설명한다.

(4) 신호편의제거(SBR: Signal Bias Removal): 신호편의제거 방법은 여러 가지 환경에 의하여 왜곡된 입력신호에 생기는 바이어스를 음성신호로부터 분리한 후, 이를 제거함으로써 채널 왜곡이나 잡음에 의한 영향 등을 효과적으로 억제할 수 있다[8]. 입력신호의 특징벡터열 $X=\{x_1,x_2,\cdots,x_t,\cdots,x_T\}$와 특징벡터에 대한 대표 모델 $\Lambda=\{\lambda_i, i=1,2,\cdots,M\}$ (λ_i는 Markov 모델, t는 프레임 수)에 대하여 λ_i가 동일한 확률을 갖는다면, likelihood는 벡터양자화(VQ)와 동일하게 된다. 여기서 Λ는 바이어스가 없는 음성의 모델이라고 가정하면, λ_i는 VQ 코드북에서 코드워드가 된다. 이 코드워드들이 바이어스가 없는 신호를 대표하므로, 입력신호의 특징벡터와 코드워드들 사이의 차이를 바이어스라고 할 수 있다. 입력신호의 전체구간에 대하여 각 바이어스의 평균을 구하고 이를 차감하는 과정을 반복적으로 수행함으로써 바이어스를 제거한다. SBR 방법은 바이어스를 제거한 후에도 기존 모델에 대한 의존성이 유지되기 때문에, 모델을 새로 훈련하지 않고, 기존의 모델을 사용하면서 인식단계에서만 처리해 주면 되는 장점이 있다.

3. 특징파라메타 추출

1) LPC 켑스트럼

LPC 켑스트럼은 LPC 분석에 의해 얻은 LPC 계수로부터 변환식을 이용해 구한 켑스트럼으로, 음성인식 및 화자인식에 널리 사용되고 있는 특징벡터이다[9][10][11]. LPC 방법은 음성신호처리에서 가장 널리 쓰이고 있는 알고리즘의 하나로 음성을 전극(all pole)모델로 가정하고 그에 따른 필터의 계수를 이용하여 여러 가지 음성신호처리를 하는 방법이다. 또한 LPC 계수에 의하여 구성되는 필터는 음성이 어떻게 생성되는가 하는 것을 전극특성을 가정하여 분석한 것으로, 성도의 특성을 모델링하게 된다. 여기서 성도모델에 관한 필터가 LPC 계수에 의해 구성되며, 이 필터구성의 차를 음성인식 및 화자인식에서 유용하게 이용할 수 있다. 이러한 성도모델에 적당한 필터의 차수는 음성신호를 디지털로 나타낼 때의 표본화 주파수에 따라서 달라지는데 표본화 주파수가 크면 클수록 필터차수도 커야한다. 그리고 LPC 분석의 구간은 10-45 msec 정도이며, 10-20 msec 정도씩 이동하면서 분석하는 것이 보통이다.

관측된 음성샘플은 시간영역에서 음원과 성도의 임펄스응답의 컨벌루션 결합으로 나타내어

질 수 있다. 또한 주파수영역에서 그 음성의 스펙트럼은 음원스펙트럼과 성도필터의 곱으로 나타내어진다. 이러한 스펙트럼의 곱을 합으로 (즉 로그를 취한다) 바꾼 다음, 이 주파수 영역을 다시 시간영역으로 변환하면 켑스트럼이 구해진다. 이렇게 하여 얻은 켑스트럼 계수의 낮은 차수에는 성도모델에 관한 정보가 들어있고, 높은 차수에는 음원모델에 관한 것이 들어 있음을 알 수 있다. 켑스트럼을 구하는 방법으로 2가지 형태가 있는데 그 하나는 FFT를 이용하여 구하는 것이고 다른 하나는 LPC분석을 통해서 구하는 것이다. 후자의 것을 LPC 켑스트럼이라고 한다. LPC 계수로부터 켑스트럼 계수를 구하는 식은 다음과 같다[12].

$$c_1 = a_1$$

$$c_n = a_n + \sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c_{n-k}, \quad 1<n<P \tag{1}$$

여기서 a_k 는 LPC 계수, P는 LPC 차수, 그리고 c_k 는 켑스트럼계수이며, 만일 켑스트럼의 차수가 P를 넘을 때에는 위의 식에서 $a_k=0$, $k>P$을 사용한다. LPC 계수에 비해 켑스트럼 계수를 화자인식에 많이 쓰는 가장 중요한 이유는 켑스트럼 계수들 사이에는 상관관계가 매우 작아 단순화된 거리오차를 사용할 수 있다는 잇점이 있기 때문이다. 그런데 켑스트럼의 계수 중 낮은 차수의 계수는 전체적인 스펙트럼 경사도에 지나치게 민감하게 반응을 하는 한편, 높은 차수의 켑스트럼 계수는 잡음에 민감하게 반응하는 단점이 있다. 이러한 켑스트럼의 지나친 민감성을 둔화시키기 위하여 창함수를 이용하여 켑스트럼의 각 계수에 서로 다른 가중치를 주는 방법이 일반적으로 많이 사용되고 있다[9][13]. 이와 같이 가중함수를 켑스트럼에 곱한 것을 가중 켑스트럼이라 부른다. 이러한 가중함수를 곱함에 따라 두 특징벡터의 거리척도를 단순한 유클리드 거리로서 나타내는 것이 가능해진다.

2) Mel-켑스트럼

Mel 켑스트럼은 인간의 청각계를 모델링한 것이다. Mel은 음색의 피치나 주파수의 측정단위로서 사람의 청각 인지도 실험을 통하여 그 스케일을 정한다. Mel 켑스트럼은 파워 스펙트럼을 물리적인 주파수, 즉 선형 주파수 축으로 표현하는 것이 아니라 Mel 척도를 사용하여 파워 스펙트럼을 표시한 다음, 이로부터 켑스트럼을 구해낸다. Mel 척도란 인간의 청각특성을 고려한 주파수 특성인데 1000 Hz는 1000 Mel을 대응하고 이로부터 실험적으로 값을 결정한 것이다. 일반적으로 Mel 척도는 1000 Hz이하에서는 물리적인 주파수와 선형적으로 비례하며 그 이상에서는 대수적으로 비례한다. Fant는 Mel 주파수와 일반적인 주파수 사이의 근사식을 식(2)에 제시하고 있다[14].

$$F_{mel} = \frac{1000}{\log 2} \log\left(1 + \frac{F_{Hz}}{1000}\right) \quad (2)$$

Mel 켑스트럼을 구하고자 할 때에는 먼저 주어진 음성신호를 pre-emphasis하고 창함수를 취해서 한 프레임 길이의 음성신호를 선택한 후, 이 신호를 FFT해서 스펙트럼을 얻는다. 이렇게 해서 얻은 스펙트럼의 주파수 축을 위의 Mel 척도로 바꾼 후 역 FFT하면 얻을 수 있는데 이때 스펙트럼의 절대값을 취한 다음 역 FFT를 하므로 데이터값은 모두 실수 값을 가지게 되므로 Cosine Transform을 이용하여 구하는 것이 편리하다[15]. Mel 켑스트럼은 물리적인 주파수를 Mel 척도의 주파수로 바꿀 때, 일반적으로 필터뱅크를 사용한다. 일반적으로 이때 사용하는 필터는 삼각필터이나 본 실험에서는 8 kHz 표본화인 경우에 한 프레임의 데이터갯수가 적으므로 사각필터를 이용하였다. 이 출력을 $S(\omega)$라고 하면 Mel 켑스트럼은 식(3)과 같은 관계식으로부터 계산된다[16].

$$c_n = \sum_{k=1}^{K} (\log S_k) \cos[n(k-\frac{1}{2})\frac{\pi}{K}], \quad n = 1, 2, , \ldots L \quad (3)$$

위에서 K는 필터뱅크의 개수이고 L은 사용하려는 켑스트럼의 차수이다. 필터 뱅크를 사용하지 않고 이를 계산하려면 Mel 척도를 나타내는 다음 관계식 (4)-(5)를 이용할 수 있다.

$$\hat{\Omega} = \theta(\Omega) = \Omega + 2\arctan[\frac{0.35\sin\Omega}{1-0.35\cos\Omega}] \quad (4)$$

$$c_m = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{S_k}(\hat{\Omega})\cos(j\hat{\Omega}m)d\hat{\Omega} \quad \text{(III-17)} \quad (5)$$

3) 전화망의 채널특성의 보상방법

켑스트럼 평균차감법을 이용하여 음성신호에 실린 전화망에 의한 채널특성을 다음의 방법으로 제거한다. 전화망을 통과한 음성신호는 channel의 필터링 작용으로 선형왜곡이 일어난다. 이는 $T(Z) = S(Z)G(Z)$로 나타낼 수 있는데 여기서 $S(Z)$는 순수한 음성신호, $G(Z)$는 전화선로의 채널 전달함수, 그리고 $T(Z)$는 필터링된 음성신호이다. 이에 양변 로그를 취하면 $\log T(Z) = \log S(Z) + \log G(Z)$이 된다. 즉, 채널의 영향은 음성신호의 켑스트럼에 부가적인 성분으로 나타난다. 이때 순수한 음성 켑스트럼의 전체구간평균이 0이라고 가정하면, 채널 켑스트럼의 추정치는 필터링된 음성의 켑스트럼들을 평균하여 구할 수 있다. 그리고 이러한 채널효과를 보상하기 위해서는 추정된 채널 켑스트럼을 전화음성의 켑스트럼에서 제거한다. 이를 켑스트럼 평균차감법이라 하며, 채널의 영향이 보상된 켑스트럼은 다음과 같이 표현된다.

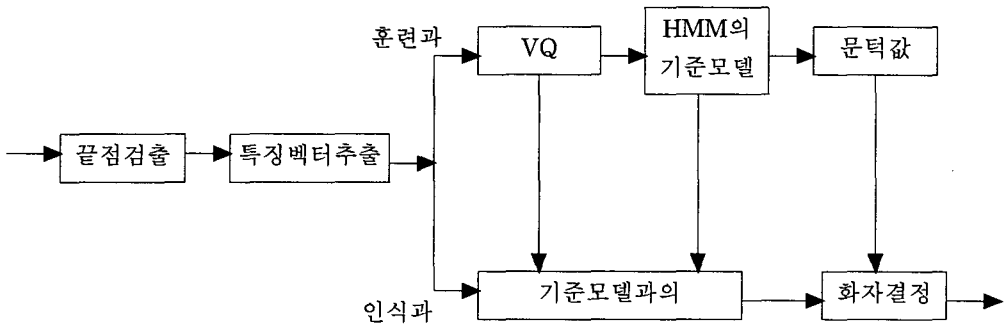
$$\hat{c}_t = c_t - E[c_t] \tag{6}$$

$$E[c_t] = \frac{1}{T}\sum_{t=1}^{T} c_t$$

여기서 $E(c_t)$는 채널 켑스트럼의 평균값, c_t는 t번째 프레임의 켑스트럼, T는 전체 프레임 수이다.

4. 화자인식 시스템 구성

문장종속 화자인식 시스템의 구성은 다음의 기본 구성도와 같다.



먼저 입력되는 모든 음성데이타는 끝점검출을 통해 데이터의 음성부분만 선택을 하고 이로부터 특징벡터의 추출을 위한 음성처리가 진행된다. 학습과정에서는 각각의 화자가 미리 정해진 숫자을 여러 번 반복 발음한 것을 갖고 시스템의 기준패턴을 작성하게 된다. 여기서는 화자인식 방법으로 VQ와 HMM을 사용하게 되므로 먼저 VQ를 통해 코드북을 작성한 후, 이를 이용하여 추출된 특징벡터의 인덱스를 구하고 이 인덱스를 이용하여 HMM을 훈련하여 각각의 화자에 필요한 HMM 모델을 구한다. 이때 학습과정은 Baum-Welch 알고리즘을 사용하였고, 인식과정에서는 학습과정에서 사용했던 것과 동일한 특징벡터를 추출하고, 이를 VQ 코드북의 인덱스열로 변환한 후, 이미 만들어진 각각의 화자에 대해 만들어진 HMM의 기준 모델에 대해서 사후확률값들을 각각 계산한다. 이때 Forward - backward procedure를 사용한다. 그리고 최고의 확률값을 갖는 모델을 인식된 화자로 선정하게 된다. 그러나 단순하게 최고의 확률을 얻은 것을 찾고자하는 화자로 인식을 하는 경우에 오인식할 가능성이 있으므로 상용화시에 문제가 발생할 수도 있다. 따라서 이런 경우에 대비하여 화자확인기법을 결합시켜야 하는데 화자확인에 요구되는 정도로 엄밀한 의미의 문턱값을 설정할 필요는 없지만 그에 준하는 정도의 문턱값을 설정하여 이 문턱값보다 낮을 경우에는 비록 특정 화자가 최고의 확률을 얻었더라도 그 화자로 인식하지 않고 인식화자가 없다는 결과를 내주어야 한다. 이와 유사하게 인식된 화자의 확률과

차점을 얻은 화자의 확률이 일정한 차이 이하이면 다시 한번 확인하는 방법이 요구된다. 이 차이에 대해서도 동일하게 문턱값이 요구되므로 많은 반복실험을 통해 이 값을 설정해야 할 것이다.

5. 실험 및 결과

본 연구에서 사용한 음성 DB는 20대 남성화자 10명이 0(영)에서 9(구)까지 10개의 숫자음을 각각 20회 반복발음한 음성데이터로 구성된다. 전화선로의 영향을 고려하기 위하여 서울과 학교 실험실과의 전화통화를 통해서 음성데이터를 일주일에 걸쳐 녹음했다. 이때 입력으로 들어온 음성은 8 kHz, 8 bit μlaw로 표본화되었고, 또한 음성신호 파형의 DC-Baias를 제거하기 위하여 다음의 관계식을 이용하였다.

$$\hat{x}(n)=x(n)-\frac{1}{T}\sum_{t=1}^{T}x(t)\ ,\ 1\leq n\leq T \tag{7}$$

다음 전처리를 위하여 A/D변환된 음성을 필터 $H(z)=1-0.95z^{-1}$를 이용하여 Pre-emphasis하고 30 msec를 한 프레임으로 하는 Hamming창을 적용하여 매 10 msec마다 특징벡터를 얻었다. 특징벡터로는 앞서 설명한 LPC 켑스트럼과 Mel 켑스트럼를 사용하였다. 먼저 LPC 켑스트럼은 주어진 데이터에 대해서 Lebinson - Durbin 알고리즘을 이용하여 12차의 LPC 계수를 구한 다음, 이 LPC 계수로부터 LPC 켑스트럼을 얻는다. 두 번째 특징벡터인 Mel 켑스트럼을 구하기 위해 먼저 주어진 데이터를 DFT하고, 이 값에 절대값을 취한 후 로그를 취한다. 이렇게 얻은 신호는 다시 36차 Mel-Scale Filter Bank를 이용하여 Mel-Scale 스펙트럼으로 변환시킨다. 그리고 식(3)을 이용하여 Mel 켑스트럼의 계수를 계산한다. 그리고 위에서 구한 특징벡터들로부터 채널의 영향을 제거하기 위해 켑스트럼 평균차감법을 사용했으며 이의 관계식은 식(6)에 있다.

이렇게 처리된 특징벡터를 이산HMM에 사용하기 위해 VQ를 이용한다. 이때 VQ 코드북의 크기는 512로, 각 코드북은 LBG 알고리즘을 이용하여 구했으며, 이때 사용한 거리척도는 켑스트럼에 대한 거리관계식인 유클리드거리를 사용하였다.

각 화자에 의해 녹음된 DB중 10개는 HMM 모델의 훈련에 이용하였다. 즉, VQ를 사용하여 음성데이터로부터 코드북의 인덱스를 구하고 이 인덱스열을 이용하여 HMM 모델의 파라메타를 구하였다. 각 화자에 대한 HMM 모델을 훈련시키기 위해서 각 화자에 의해 발성된 숫자음을 임의로 결합시켜 하나의 문장을 만들고 이를 이용하여 각 모델을 구하였다. 이때 모델은 상태가 4개이고 출력심벌의 개수가 VQ 코드북의 크기로 하였다. 본 실험에서는 erogodic HMM모델에서 각 상태를 특성이 유사한 파라미터들의 집합으로 보고 VQ를 이용하여 이것에 대한 초기 확률을 구한다.

표 1. 화자인식 실험 결과표.

특징벡터	켑스트럼평균차감법	화자인식률
LPC-Cepstrum	불사용	82%
	사용	92%
Mel-Cepstrum	불사용	86%
	사용	95%

π_i = 상태가 j 인 벡터의 개수를 전체벡터의 개수로 나눈 값

a_{ij} = 상태가 i에서 j로 변한 벡터의 개수를 상태가 i인 벡터의 개수로 나눈 값, 천이확률

$b_{j(k)}$ = 상태가 j 이고 코드북 인덱스를 k로 가지는 벡터의 개수를 상태가 j 인 벡터의 개수로 나눈 값

이렇게 구한 초기 확률을 이용해서 Baum-Welch 재추정법을 사용하여 각 화자에 대한 HMM 모델을 만들었으며, 실제의 테스트과정에서는 음성데이타의 나머지 10회 분량을 가지고 실험을 했으며 이때는 연결된 4개의 숫자음을 만들어 이를 갖고 테스트에 사용하였다. 실제로 실험에 사용하는 숫자음의 스트링개수 즉 입력음성의 길이는 인식률에 상당한 영향을 주고 있음을 확인할 수 있었다. 인식실험의 결과는 다음의 표1과 같다.

실험결과를 통해 전화음성에 대한 화자인식기를 구현할 때, 켑스트럼 평균차감법을 이용함으로써 전화망을 통한 음성을 대상으로 하는 화자인식 시스템에서 전화망의 영향으로 생기는 인식률의 저하를 상당히 개선할 수 있음을 알 수 있었다. 사용한 두 개의 특징벡터 모두 켑스트럼 평균차감법을 이용하였을 때, 9-10 % 정도의 인식률이 개선되었다. 특히 LPC 켑스트럼을 사용한 경우와 Mel 켑스트럼을 사용한 결과를 보면 Mel 켑스트럼이 켑스트럼 평균차감법을 사용했을 때나 안했을 때나 모두 LPC 켑스트럼 보다 우수하다는 것을 확인할 수 있었다. 이는 이미 여러 논문을 통해 확인된 사항이지만, 실제의 화자인식시의 구현에서는 특징벡터를 구하는 데에 있어서 LPC 켑스트럼보다 Mel 켑스트럼은 몇 배의 계산이 소요되기 때문에, 그 인식성능의 차이가 크지 않은 경우에는 응용 예에 따라 두 가지 파라미터 중 어느 것을 사용해도 무방할 것으로 생각된다. 즉, 구성해야 할 시스템의 수준에 맞게 어떤 특징 벡터를 사용할 것인지를 결정해야 할 것이다. 그리고 실제 상용화할 경우에는 훈련과정의 주변 환경과 테스트 과정의 주변 환경이 다를 때 화자인식률이 급격히 떨어진다는 문제점이 있다. 실제로 실험결과표에는 나와 있지 않지만 두 개의 상황이 불일치하는 경우에는 인식의 결과가 급속하게 나빠지는 것을 확인했다. 따라서 이러한 경우 잡음에 강인한 특징 벡터를 사용하거나 잡음을 제거하기 위한 특별한 과정을 전처리 과정에 삽입하는 것이 필요하겠다. 또한 인식에 사용하는 입력 숫자열의 길이가 인식결과에 상당한 영향을 미친다는 것을 알았다. 본 논문의 실험에서는 4개

의 연결숫자음을 대상으로 테스트를 수행했지만 이를 늘려줄 경우에는 인식률이 향상됨을 보였다. 따라서 응용분야에 따른 적절한 입력음성 길이의 선택도 중요한 변수임을 알 수 있었다.

6. 결 론

본 논문에서는 전화망을 통한 음성신호에 대한 문장종속 화자인식시스템의 성능 개선에 대하여 논하였다. 즉, 전화망의 채널효과와 여러 잡음성분에 의해 통화자의 음성이 왜곡되므로 이를 대상으로 화자인식을 수행할 경우, 인식률의 상당한 저하를 가져온다. 이에 대해 본 논문에서는 음성인식에서 많이 이용되는 켑스트럼 평균차감법을 화자인식에 이용하여 전화망에 의한 영향을 제거하여 약 10 % 정도의 인식률 향상을 얻을 수 있었다. 이 실험에서 구현한 화자인식기는 VQ와 HMM 방법을 사용하였고, 또한 화자인식기에 사용한 음성의 특징벡터로는 LPC 켑스트럼과 Mel 켑스트럼 두 종류를 채택하여 화자인식용 파라메타로서 이들의 성능을 비교하였다. 실험결과로 LPC 켑스트럼 보다는 Mel 켑스트럼 파라메타를 사용했을 때 인식률이 더 좋음을 확인할 수 있었다.

참 고 문 헌

[1] P. J. Moreno, R. M. Stern. 1994. "Sources of degradation of speech recognition in the telephone network." Proceeding of the International Conference on Acoustics, Speech and Signal Processing, 109-112.

[2] H. Hermansky, N. Morgan, H. G. Hirsch. 1993. "Recognition of speech in additive and convolutional noise based RASTA spectral processing." Proceeding of the International Conference on Acoustics, Speech and Signal Processing, 83-86.

[3] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, G. Tong. 1994. "Integrating RASTA-PLP into speech recognition." Proceeding of the International Conference on Acoustics, Speech and Signal Processing, 421-424.

[4] H. Hermansky, N. Morgan, A Bayya, P. Kohn. 1991. "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA - PLP)." Proc. Eurospeech, 3, 1367-1370.

[5] S. F. Boll. 1979. "Suppression of acoustic noise in speech using spectral subtraction." IEEE Trans. Acoust., Speech, Signal Processing, ASSP-27(2), 113-120.

[6] M. Berouti, R. Schwartz and J. Makhoul. 1983. "Enhancement of speech corrupted by acoustic noise." Signal Processing, 1: Speech Enhancement, Prentice-Hall, Englewood Cliffs, NJ, 69-73.

[7] Richard J. Mammone, Xiaoyu Zhang, Ravi P. Pamachandran. 1996. "Robust speaker recognition - A feauture-based approach." IEEE Signal Processing Magazine, 58-87.

[8] Mazin G. Rahim, Biing-Hwang Juang. 1996. "Signal bias removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition." IEEE Trans. Speech & Audio Processing, 4(1), 19-30.

[9] L. R. Rabiner and B. H. Juang. 1993. *Fundamental of Speech Recognition*, Englewood Cliffs, New Jersey: Prentice-Hall.

[10] F. K. Soong, A. E. Rosenberg, L. R. Rabiner and B. H. Juang. 1987. "A vector quantization approach to speaker recognition." AT&T Tech. J. 66, 14-26.

[11] H. Gish and M. Schmidt. 1994. "Text-independent speaker identification." IEEE Signal Processing Magazine, 11(4), 18-31.

[12] K. R. Farrel and R. J. Mammone. 1994. "Speaker recognition using neural networks and conventional classifiers." IEEE Trans. Speech and Audio Processing, 2(1), 194-205.

[13] L. R. Rabiner and B. H. Juang. 1993. *Fundamental of Speech Recognition*, Englewood Cliffs, New Jersey: Prentice-Hall.

[14] S. B. Davis and P. Mermelstein. 1980. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE Trans. Acoustics, Speech, Signal Proc., 28(4), 357-366.

[15] M. J. Hunt and C. Lefebvre. 1989. "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech." Proceeding of the International Conference on Acoustics, Speech and Signal Processing, 262-265.

[16] D. O'Shaughnessy. 1990. *Speech Communication - Human and Machine*, Addison Wesley, 420-422.

접수일자: '97. 10. 7.
게재결정: '97. 11. 12.

▲ 정 의 상
경기도 포천군 포천읍 선단리 산11-1
대진대학교 공과대학 전자공학과 487-711
전자공학과 대학원

▲ 최 홍 섭
경기도 포천군 포천읍 선단리 산11-1
대진대학교 공과대학 전자공학과 487-711
Tel : (0357) 539-1903 Fax : (0357) 539-1900
e-mail: hschoi@road.daejin.ac.kr