전화선 채널이 화자확인 시스템의 성능에 미치는 영향

The Effect of the Telephone Channel to the Performance of the Speaker Verification System

조 태 현*, 김 유 진*, 이 재 영**, 정 재 호*
(Tae Hyun Cho*, Eu Gene Kim*, Jae Young Lee**, Jae Ho Chung*)

※본 논문은 한국과학재단 '98 핵심전문연구비 지원으로 수행되고 있습니다.

요 약

본 논문에서는 깨끗한 환경에서 녹음된 음성데이터와 채널환경에서 수집된 음성데이터의 화자확인 성능을 비교하였다. 채널데이터의 화자확인 성능을 향상시키기 위하여 채널환경에 강인한 특징 파라메타 및 전처리에 대해 연구하였다. 실험을 위한 음성 DB는 어구지시(text-prompted) 시스템을 고려하여 두자리의 한국어 숫자음으로 구성하였다.

적용한 음성 특징은 LPCC(Linear Predictive Cepstral Coefficient), MFCC(Mel Frequency Cepstral Coefficient), PLP-(Perceptually Linear Prediction), LSP(Line Spectrum Pair)이며, 채널 잡음을 제거하기 위한 전처리 과정으로는 음성신호에 대한 필터링을 적용하였다. 추출된 특징으로부터 채널의 영향을 제거 또는 보상하기 위해 cepstral weighting, CMS(Cepstral Mean Subtraction), RASTA(RelAtive SpecTrAl)를 적용하였다. 또한 각각의 특징 및 처리 방법에 대한 음성인식 성능을 제시함으로써 화자확인에서의 성능과 음성인식에서의 성능을 비교하였다.

적용한 음성 특징 및 처리 방법들에 대한 성능 평가를 위해 HTK(HMM Tool Kit) 2.0을 이용하였다. 남자, 여자 화자별로 임계값을 다르게 주는 방법으로 깨끗한 음성데이터와 채널 데이터에 대한 **EER(Equal Error Rate)**을 구하여 비교하였다.

실험결과 전처리 과정에서 대역통과 필터(150~3800Hz)를 적용하여 저대역 및 고대역의 채널 잡음을 제거하고, 이 신호로부터 MFCC를 추출하였을 때 EER 측면에서의 화자확인 성능이 가장 좋게 나타났다.

ABSTRACT

In this paper, we compared speaker verification performance of the speech data collected in clean environment and in channel environment. For the improvement of the performance of speaker verification gathered in channel, we have studied on the efficient feature parameters in channel environment and on the preprocessing. Speech DB for experiment is consisted of Korean doublet of numbers, considering the text-prompted system. Speech features including LPCC(Linear Predictive Cepstral Coefficient), MFCC(Mel Frequency Cepstral Coefficient), PLP(Perceptually Linear Prediction), LSP(Line Spectrum Pair) are analyzed. Also, the preprocessing of filtering to remove channel noise is studied. To remove or compensate for the channel effect from the extracted features, cepstral weighting, CMS(Cepstral Mean Subtraction), RASTA(RelAtive SpecTrAl) are applied. Also by presenting the speech recognition performance on each features and the processing, we compared speech recognition performance and speaker verification performance.

For the evaluation of the applied speech features and processing methods, HTK(HMM Tool Kit) 2.0 is used. Giving different threshold according to male or female speaker, we compare EER(Equal Error Rate) on the clean speech data and channel data.

Our simulation results show that, removing low band and high band channel noise by applying band pass filter(150~3800Hz) in preprossing procedure, and extracting MFCC from the filtered speech, the best speaker verification performance was achieved from the view point of EER measurement.

* 인하대학교, 전자공학과, 디지털 신호처리 연구실
** 나래이동통신, 기술연구소

접수일자 : 1998년 12월 21일

I. 서 론

일반적으로 음성신호가 전화선 채널을 통과할 때 발생되는 전화선 채널의 영향은 다음과 같다[1]. 첫째, 전화선 채널은 약 300~3400Hz의 통과대역을 갖는다. 이것은 저주파 및 고주파 영역의 음성 정보가 손실됨을 의미한다. 또한, 이 통과대역은 통화가 연결될 때마다 달라질 수 있으므로 훈련에 사용된 데이터와 테스트를 위한 데이터 사이에 불일치(mismatch)를 초래한다. 이러한 점들은 전화선 채널을 이용하는 화자인식 시스템의 성능을 저하시키는 가장 중요한 요인이다. 둘째, 채널의 통과대역에서의 불균일(non-flat)한 주파수 응답으로 인한 음성신호의 스펙트럼 왜곡을 들 수 있다. 셋째, 채널 자체에서 부가되는 채널 잡음을 들 수 있다. 반면에 전화선 채널은 통화할 때마다 조금씩 그 특성이 달라지지만, 주어지 통화시간 동안에는 그 특성이 느리게 변하거나 안정적이다. 이는 채널의 특성이 켑스트럼 영역에서 저주파 영역에 분포하는 것을 의미한다.

본 논문에서는 전화선 채널이 화자확인 시스템의 성능에 미치는 영향을 수치적으로 제시하고, 전화선 채널의 왜곡을 줄일 수 있는 적절한 채널 보상 방법에 대해서 연구하였다. 이를 위해서 배경 및 채널 잡음이 포함되지 않은 깨끗한 음성데이터와 이를 전화선을 통하여 수집한 채널 데이터를 사용하였다. 이 두 가지 데이터를 비교 실험함으로써 채널이 화자확인 성능에 어느 정도의 영향을 미치는지 알 수 있고, 기존에 제안되었던 여러 가지 특징 추출 방법 및 잡음제거 기법이 채널의 영향을 어느 정도 보상할 수 있는지 이해할 수 있다. 이러한 과정을 통하여 현재 국내의 전화망 환경이 화자인식 시스템에 미치는 영향을 분석하고자 한다.

본 연구에서는 이러한 전화선 채널 환경에서 화자확인 시스템의 성능개선을 위하여 LPCC, MFCC, PLP, LSP등의 특징 파라메타와 cepstral weighting, CMS, RASTA등의 잡음처리 방법을 적용하였을 때의 성능을 비교하였다.

본 논문의 구성은 우선 음성 특징 및 다양한 채널 잡음 처리 방법에 대해서 설명하고, 음성 데이터 베이스 구축에서부터 HTK를 이용한 실험방법을 구체적으로 기술한다. 그리고, 깨끗한 음성 데이터와 채널 데이터에 대한 실험결과를 설명하고, 마지막으로 결론과 앞으로의 연구 방향을 제시한다.

II. 음성 특징 및 잡음처리 방법

본 절에서는 전화선 채널의 모델링에 대해서 간단히 설명하고, 잡음에 강인한 것으로 알려진 특징추출 방법 및 잡음을 효과적으로 제거(subtraction) 또는 보상하는 기법을 설명한다.

2.1 전화선을 통한 음성의 전달 모델 [2]

일반적으로 전화선을 통해 전달되는 음성 신호 $y(t)$는 발성 음성 $s(t)$에 발성 환경으로부터 주변잡음 $n(t)$가 가산되고, 이것이 전화선 채널의 전달함수 $h(t)$와 컨벌루션 된다. 이것을 수식으로 나타내면 다음과 같다.

$$y(t) = x(t) * h(t) = [s(t) + n(t)] * h(t) \quad (1)$$

전화망의 경우에는 주변잡음, 채널왜곡, 마이크로폰 왜곡 등으로 인식 성능이 크게 저하된다. 그러나, 본 논문에서는 방송국 스튜디오에서 깨끗한 음성데이터를 수집하였으므로 주변잡음 $n(t)$를 무시할 수 있다. 그러므로, 식 (1)은 다음과 같게 된다.

$$y(t) = s(t) * h(t) \quad (2)$$

식 (2)를 FFT 변환한 후 로그를 취하여 얻어지는 로그 스펙트럼 영역이나 켑스트럼 영역에서는 이러한 채널 성분이 음성신호에 가산되는 특성을 갖는다. 즉, 순수한 음성의 켑스트럼 벡터를 전환(translation) 시키게 된다. 이러한 채널의 특징은 로그 스펙트럼 영역이나 켑스트럼 영역에서 필터링이나 가중함수를 사용하여 제거되거나 보상될 수 있다.

2.2 음성 특징

2.2.1 LPCC(Linear Predictive Cepstral Coefficient) [3]

LPC 계수가 원래의 음성신호를 충분히 잘 반영하고 있다고 가정하면, LPC 스펙트럼으로부터 켑스트럼 계수를 추출해낼 수 있으며, LPC 계수로부터 얻어진 켑스트럼 계수를 특별히 LPCC라 한다. LPCC $c(n)$은 순환식(recursive formula)에 의해서 LPC 계수 a_n으로부터 간단히 구할 수 있다.

$$c(n) = a_n + \sum_{k=1}^{n-1}\left(\frac{k}{n}\right)c(k)a_{n-k} \qquad 1 \le n \quad (3)$$

2.2.2 MFCC(Mel Frequency Cepstral Coefficient)

인간의 청각은 주파수 영역에서 음성을 비선형적으로 분석하는 특징을 갖는다. 그러나, 1000Hz 이상에서는 로그 주파수(logarithmic frequency) 영역에서 거의 선형적으로 분석한다. 이러한 청각 특성을 이용하여, 음성의 스펙트럼 영역에서는 비선형적으로 분포하지만 멜 도메인에서는 균등하게 분포하는 필터뱅크를 통하여 음성 특징을 추출하는 방법이 MFCC이다. 이러한 필터뱅크들은 멜 도메인에서는 동일한 분해능을 나타내며 인식 성능을 높여준다. 실제 구현에 있어서는, 음성의 분석구간 만큼을 Fourier 변환한 후 각각의 스펙트럼 값에 대해서 멜 스케일의 필터뱅크를 취하여 에너지를 얻는다. 여기서 얻어진 각각의 에너지를 다음과 같이 DCT변환하여 MFCC c_i를 얻는다.

$$c_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right) \quad (4)$$

식 (4)에서 N은 필터뱅크의 개수이고, m은 로그 필터뱅크 에너지이다.

2.2.3 PLP(Perceptually Linear Prediction) [4]

음성 신호의 단구간 스펙트럼(short time spectrum) $P(w)$를 선형 예측(Linear Prediction) 분석하여 얻어지는 전극 모델 $A(w)$는 모든 주파수대역에 대해 동일한 정밀도로 나타낸다. 그러나, 인간의 청각 특성은 800Hz 이상의 주파수를 갖는 음성에 대해서는 주파수가 높아질수록 점점 둔감해지며, 가청 주파수의 중간 대역의 주파수를 갖는 소리를 민감하게 감지한다. 따라서, 음성의 단구간 스펙트럼과 실제로 지각되는 소리는 다소 차이가 있다. 이러한 차이를 극복하려는 것이 PLP 분석 방법이다.

2.2.4 LSP(Line Spectrum Pair) [5][6][7]

디지털 음성 부호화에서 LPC 스펙트럼 정보를 효율적으로 표현하기 위해서 사용되는 것이 LSP이다. 성도 전달함수에서 역필터 $A(z)$의 p개의 영점(zero)이 다음의 두 다항식 $P(z)$와 $Q(z)$를 통해서 단위원상으로 매핑된다.

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{5}$$

$$P(z) = A(z) - z^{-(p+1)}A(z^{-1}) \tag{6}$$

LSP는 $P(z)$, $Q(z)$의 영점에 해당하는 주파수이다. $P(z)$와 $Q(z)$의 영점들은 단위원상에서 서로 엇갈려 위치함으로써 LSP파라메타 사이에 다음과 같은 관계가 존재한다.

$$0 = w_0 < w_1 \cdots < w_p < w_{p+1} = \pi \tag{7}$$

$P(z)$, $Q(z)$의 주파수 값들이 비슷하면 역필터 $A(z)$의 영점은 단위원에 가깝게 위치하고 포먼트는 이 두 주파수 사이에 있게 된다. 그러나, $P(z)$, $Q(z)$의 주파수가 차이가 크면 $A(z)$는 넓은 대역폭의 영점을 갖게 된다. LSP계수는 LP필터의 영점에 해당하는 주파수 값을 나타내며, 효율적인 음성코딩뿐만 아니라 LPC계수를 평활화하는데 효과적이다.

2.3 잡음처리 방법

2.3.1 Cepstral Weighting [2][8][9]

Cepstral weighting은 리프터링(liftering) 이라고도 하며 켑스트럼 계수 $c(n)$에 가중 함수 $w(n)$을 곱함으로써 이루어진다. 이것은 켑스트럼 계수가 저차의 경우 전체적인 스펙트럼의 기울기(slope)에 예민하고, 고차의 경우 잡음에 민감하므로 이 성분들을 제한하거나 감쇄시킴으로써 인식률을 높일 수 있다.

가장 간단한 형태는 quefrency 리프터링인데 다음과 같은 선형 가중함수를 사용한다. 이것은 고차로 갈수록 선형적으로 증가하는 가중치를 곱하여 강조하는 방법이다.

$$w(n) = \begin{cases} n, & n = 1, 2, \ldots, L \\ 0, & otherwise \end{cases} \tag{8}$$

위 식에서 L은 음성의 분석 길이이다. 또한, 대역통과 리프터링에 적용하는 가중 함수는 다음 식 (9)와 같다. 이것은 켑스트럼의 낮은 차수와 높은 차수의 바람직하지 못한 변화를 제거하는 목적으로 사용될 수 있다.

$$w(n) = \begin{cases} 1 + \frac{L}{2}\sin\left(\frac{n\pi}{L}\right), & n = 1, 2, \ldots, L \\ 0, & otherwise \end{cases} \tag{9}$$

2.3.2 CMS(Cepstral Mean Subtraction) [10][11]

CMN(Cepstral Mean Normalization)의 일종인CMS는 모델 훈련과정과 테스트 과정에서 조금씩 변하는 채널의 왜곡성분을 켑스트럼 영역에서 제거하는 방법이다. 즉, 전체 구간에 대하여 켑스트럼의 평균을 구하고, 이를 차감하여 전화선 채널에서 발생하는 선형왜곡을 제거한다. CMS에서 필터링된 음성의 켑스트럼을 평균하여 채널 켑스트럼을 추정하는 것은 순수한 음성의 켑스트럼에 대한 장구간 평균이 0이라는 가정에 근거한다.

이때, 순수한 음성에 대한 켑스트럼 평균이 0이 되기 위해서는 음성구간의 유성음, 무성음, 파열음(plosive sound) 등이 음향학적 균형을 이룰 때 가능하다. 이것은 실제로 불가능하므로 단구간의 켑스트럼 평균을 이용하여 채널의 켑스트럼 추정치를 구하여 이를 차감했을 때 음성 정보까지 왜곡시킬 수 있다. 다만, 훈련 데이터와 테스트 데이터 모두에 대해 CMS 처리를 하게 된다면 이러한 점은 크게 문제되지 않는다.

2.3.3 RASTA(RelAtive SpecTrAl) [11][12][13]

1990년 Hermansky에 의해 제안된 기법인 RASTA는 PLP분석과 함께 음성신호를 분석하는 방법 중 시간에 따라 천천히 변하는 성분(steady state factor)에는 영향을 덜 받는 음성 신호의 특징을 추출하는 방법이다. RASTA-PLP는 PLP과정에서 임계대역 스펙트럼(critical band spectrum)을 로그 영역으로 변환한 후 여기에 시간궤적에 대한 대역통과 필터링이나 고역통과 필터링을 하게 된다. 특히, 대역통과 필터는 채널 성분뿐만 아니라 빠르게 변하는 성분도 함께 제거하므로 잡음에 강인한 특성을 갖는다. 필터링 블럭은 각 주파수 대역을 IIR 필터를 사용하여 대역통과 필터링을 수행하며, 이 대역통과 필터의 전달함수는 다음과 같다.

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - az^{-1})} \tag{10}$$

식 (10)에서 a 는 저역차단 주파수를 결정하는 상수이며 일반적으로 0.9에서 0.98사이의 값을 사용한다. 이 필터의 저역 차단주파수는 RASTA분석에서 고려할 가장 느린 로그 스펙트럼의 변화를 결정하고, 고역 차단주파수는 로그 스펙트럼의 가장 빠른 스펙트럼 변화를 결정하게 된다. 위의 전달함수는 0.26Hz의 저역 차단 주파수를 갖고, 12.8Hz에서부터 전달함수의 기울기가 감소하기 시작하여 28.9Hz와 50Hz에서 0이 되는 모양을 나타낸다.

이와는 달리 저주파 부분만을 억제하기 위한 고역 통과 IIR 필터는 다음과 같다.

$$H(z) = \frac{1 - z^{-3}}{1 - \alpha z^{-1}} \quad (11)$$

식 (11)에서 α 는 보통 0.95에서 0.98사이의 값을 사용한다.

2.3.4 Filtering

저주파 영역에 존재하는 채널잡음을 제거하기 위해서는 고역통과 필터를 적용할 수 있고, 고주파 영역에서 포함될 수 있는 양자화 잡음 등을 제거하기 위해서는 저역통과 필터를 적용할 수 있다. 이때 필터는 음성신호의 제1차 포먼트(formant) 정보를 잃지 않도록 통과대역의 범위를 정해야 한다. 또한, 전화선 채널의 대역이 일반적으로 300~3400Hz으로 알려져 있지만, 300Hz이하나 3400Hz 이상의 대역에서도 음성의 스펙트럼 성분이 존재할 수 있다는 실험적인 결과가 제시되었다[10]. 그러므로, 채널 데이터를 필터링 하여 채널의 대역을 제한할 때는 필터의 대역폭은 300~3400Hz보다 넓어져야 한다.

III. 실험방법 및 과정

3.1 어구지시 숫자음 [14]

본 연구에서는 어구지시 화자확인 시스템을 고려한 한국어 숫자음 DB를 구축하였다. DB는 21(이십일)에서부터 98(구십팔)까지의 3음절의 2자리 숫자 중에서 십단위의 숫자(30, 40, 50, 60, 70, 80, 90) 및 첫 음절과 마지막 음절이 중복되는 것(22, 33, 44, 55, 66, 77, 88)을 제외한 64개의 숫자음을 사용하였다. 십단위의 숫자 및 첫음절과 마지막 음절이 중복되는 숫자들은 그 외 숫자들에 비해 다양한 음운 현상을 포함하지 않으므로 훈련 및 인식 대상에서 제외하였다. 훈련 및 인식에서 사용된 어구는 위의 숫자들 중에서 3개를 조합한 것이다. 예를 들면, 27-83-91 (이십칠-팔십삼-구십일)과 같은 방법의 combination lock 어구이다.

훈련과정에서는 18개의 인식단위(십단위 8개, 일단위 9개, silence 모델 1개)를 훈련하여 각각의 화자를 모델링 하도록 하였다. 각 인식 단위는 표 1에 나타나 있다.

표 1. 인식 단위
Table 1. Recognition unit.

십 단위	이십, 삼십, 사십, 오십, 육십, 칠십, 팔십, 구십
일 단위	일, 이, 삼, 사, 오, 육, 칠, 팔, 구
기타	Silence

본 연구에서는 HTK를 이용하므로 훈련을 위해서는 음성 데이터에 대해 각 인식 단위별로 레이블링 하는 작업이 필요하다. 실제의 레이블링에서는 많은 경우에 십단위와 일단위를 경계로 하여 구분하기가 어려웠다. 그래서, 모든 숫자음에 공통적으로 존재하며 변이가 적은 마찰음 'ㅅ'을 경계로 하여 나누는 방법을 적용하였다. 예를 들면, "27(이십칠)"을 (이십+칠)로 나누는 것이 아니라 (이ㅅ+ 입칠)로 나누어 레이블링 하였다. 표 2에는 각 인식단위별 평균 길이(duration)가 나타나 있다.

표 2. 각 인식단위별 평균길이(sec)
Table 2. Average duration of each recognition unit (sec).

Twen (이십)	Thir (삼십)	For (사십)	Fif (오십)	Six (육십)	Seven (칠십)	Eigh (팔십)	Nine (구십)	
0.144	0.206	0.203	0.155	0.178	0.201	0.183	0.135	
one (일)	two (이)	three (삼)	four (사)	five (오)	six (육)	seven (칠)	eight (팔)	nine (구)
0.365	0.355	0.422	0.410	0.345	0.320	0.430	0.414	0.4.7

HTK에서 레이블링할 때 한국어 표기를 사용할 수 없으므로 한국어 발성에 해당하는 영문자를 사용하였다. 한 번의 훈련 세션(session)에서는 21개의 어구를 발성하였으며 이 21개의 어구에는 각 인식단위 당 6~10회의 빈도수를 갖는다. 각 인식단위의 개수가 차이를 보이는 것은 오인식 가능성이 높은 인식단위를 1~2회 더 많게 구성하였기 때문이다. 각 화자는 두 번의 훈련 세션을 가지며, 한 번의 훈련 세션에서 모든 화자는 동일한 어구를 발성하였다. 21개의 어구로 구성된 훈련 set은 다음과 같다.

27-23-23	31-38-37	35-37-36	43-45-49	46-42-48	51-59-69
53-56-54	56-54-59	59-53-64	68-62-71	68-65-76	75-74-72
78-87-82	79-71-75	79-89-81	85-81-87	87-85-81	91-98-94
96-92-92	97-97-95				

각 화자에게는 세 번의 테스트 세션이 주어지며 각 세션에서는 20개의 어구를 발성하였다. 테스트 어구는 화자, 세션에 따라 같지 않도록 랜덤하게 구성하였다.

3.2 음성 데이터 베이스 [15][16]

앞에서 설명된 훈련 set 및 랜덤하게 구성된 테스트 set을 사용하여 대학교 방송국 스튜디오 환경에서 깨끗한 데이터를 수집하였다. 본 연구에서는 음성신호가 채널에 의해 받는 왜곡 정도를 분석하는 것이므로 배경잡음은 포함되지 않도록 하였다. 발성된 음성은 48KHz 샘플링 주파수로 A/D 변환하여 DAT에 녹음하고, 이것은 실험을 위해 다시 8KHz, 16bit의 해상도를 갖는 신호로 변환하였다. 깨끗한 음성 데이터는 각 화자 당 102어구(훈련: 21어구*2세션, 테스트: 20어구*3세션)를 수집하였고, 총 1,224어구(102어구*12명)를 수집하였다.

깨끗한 음성데이터와 동일한 발성의 채널 데이터를 수집하는 과정은 다음과 같다. 본 연구에서는 전화선 채널과 PC 인터페이스가 가능한 Dialogic 보드를 사용하였다. 또한 PC의 OS를 Unix 환경으로 전환하였다. 전화선 채널에서는 음성의 코딩이 8KHz의 샘플링 주파수에서 8bit μ-law로 이루어지므로, 8KHz, 16bit의 데이터를 μ-law 테이블을 이용하여 8bit의 해상도를 갖는 μ-law 데이터로 변환하였다. 이때, bit 수를 줄이는 압축과정에서 양자화 에러가 발생한다. 이것은 본 연구에서 의도하는 채널잡음 외에 새로운 형태의 잡음이 될 수 있으나 크게 문제되지 않으리라 판단된다. 변환된 μ-law 데이터를 Unix 환경의 PC(sender)에 저장을 한다. 결과적으로 그림 1과 같은 전체적인 시스템을 구성하였다.

그림 1. 채널데이터 수집 구성도
Fig. 1. The framework for collecting channel data.

음성 데이터를 본 대학교에 있는 시스템(sender)에서 전화선을 통하여 보내면 서울의 나래이동통신 본사에 있는 시스템(receiver)에서 수집하였다. Receiver에서 수집된 데이터 형태는 8bit μ-law이다. 수집된 데이터에는 전화선채널에 의한 왜곡 및 채널 잡음이 포함되었다. 데이터가 다양한 채널의 변화를 포함할 수 있도록 2주동안 시간대별로 다르게 수집하였다. 이러한 과정을 통하여 깨끗한 음성 데이터와 동일한 발성의 채널 데이터를 1,224개 수집하였다.

3.3 음성 분석

수집된 채널데이터는 8KHz, 8bit μ-law이므로 실험을 위하여 8KHz, 16bit linear PCM 형태로 변환하였다. 이때 μ-law 테이블을 이용하였다. HTK에서 음성신호로부터 특징을 추출하기 위해서 20ms의 길이를 갖는 Hamming 윈도우를 사용하여 10ms씩 이동하면서 분석 하였으며, 0.97의 factor를 사용하여 프리엠퍼시스(pre-emphasis) 하였다.

3.4 HTK (HMM Tool Kit) [8]

본 연구의 인식 단위 훈련 및 인식은 HMM을 기반으로 한 음성 인식 시스템 구현 및 실험을 위한 상용 도구인 HTK를 사용하였다.

HTK에서는 음성 데이터의 레이블링 정보를 이용하여 인식 단위 HMM 모델의 초기화를 위한 HInit, 재추정을 위한 HRest 도구를 제공한다. HInit를 통한 초기화 과정은 각 인식 단위 HMM 모델을 음성을 표현한 특징 벡터와 대응된다는 개념으로 수행된다. 따라서 각 모델에 해당되는 훈련 데이터가 주어질 때 모델의 각 상태와 대응된 모든 훈련 데이터의 특징 벡터의 통계를 계산하여 출력분포의 평균과 분산을 추정하게 된다. 또한 전이 확률도 각 상태에 대응되는 특징 벡터의 개수를 통해 추정된다. 이때 훈련 데이터의 특징 벡터와 모델을 구성하는 각 상태 사이의 대응은 균등 상태 분할(uniform state segmentation)로부터 시작하여 Viterbi 알고리즘을 통한 최대 유사 상태열에(maximum likelihood state sequence) 근거한 확률값 $P(O|\lambda)$를 계산하여 다시 특징 벡터에 대응하는 상태를 분할하게 된다. 이 과정은 $P(O|\lambda)$값이 문턱값(threshold) 이하로 수렴되었을 때 완료된다.

한편, 초기화 과정을 거친 하위 단어 단위 HMM 모델의 재추정은 전향 및 후향 확률을 이용한 Baum-Welch 알고리즘을 통해 수행된다. 이 과정은 최대 유사 상태열이 아닌 모든 상태열을 고려한 $P(O|\lambda)$값의 수렴에 의해 완료된다.

3.5 인식

본 연구에서는 음성인식 툴인 HTK를 이용하여 깨끗한 음성데이터와 채널데이터에 대한 성능을 비교, 평가하는 화자인식 실험을 수행하였다. 즉, HTK가 화자인식 실험을 위해서 설계된 것은 아니지만, 음성인식 실험을 위한 과정을 그대로 사용하여 화자인식 성능을 평가할 수 있다.

이것은 음성인식과 화자인식의 기본 골격이 동일하다는 것을 의미한다. 음성인식과 화자인식 모두 인식 알고리즘으로 HMM을 사용한다면, HMM에서 출력되는 확률값으로써 인식을 수행하는 기본적인 과정은 유사하다. 다만, 인식 알고리즘의 전반부와 후반부에서 음성인식(또는 단어인식)과 화자인식 각각의 경우에 있어서 다른 처리가 포함될 수 있다. 전반부에서는 음성인식과는 달리 화자인식에 적합한 특징 파라메타를 선정해야 하고(물론, 음성인식 및 화자인식에 동일하게 높은 성능을 보일 수도 있다), 후반부에서는 화자인식의 경우 각각의 화자에 적응적인 임계값를 구하는 과정이 필요하다.

본 연구에서는 하나의 테스트 어구가 입력되었을 때 각각의 인식단위를 인식하는 단어인식 과정을 수행한다. 이 과정에서 발생되는 각각의 인식단위별 인식 스코어(log probablity)를 길이(duration)로 나누어 프레임당 평균 인식 스코어를 구한다. 한 테스트 어구에는 6개의 인식단위가 있으므로, 이 6개의 인식단위에 대한 스코어를 더해서 입력된 데이터의 주체인 화자에 대한 인식 스코어를 계산한다. 이 인식스코어가 바로 화자확인을 위한 확률값이 된다.

speaker1의 모델에 대한 speaker2의 테스트 어구 "21-46-51"에 대한 인식결과가 포함되어 있는 MLF(Master Label File)를 예로 들면 다음과 같다.

"speaker1/test1/214651.rec"			
0	3100000	sil	-108.261902
3100000	5000000	Tw	20.877529
5000000	8300000	one	23.614771
8300000	10600000	For	21.869333
10600000	13900000	six	27.544411
13900000	16100000	Fif	25.604769
16100000	20800000	on	25.905426
20800000	22900000	sil	0.444044

각 인식단위의 시작과 끝이 100nsec단위로 나타나 있고, 인식단위, 그리고 프레임당 인식 스코어가 포함되어 있다. 인식단위가 영문으로 나타난 것은 한국어 숫자음을 영문자로 레이블링 하였기 때문이다. 처음과 마지막 행에는 silence(sil)에 대한 스코어가 있지만 여기에는 화자정보가 담겨 있다고 보기 어려우므로 최종 스코어에서는 제외하였다. 결과적으로, 어구 '21-46-51'에 대한 인식 스코어는 6개의 인식단위 (Twen, one, For, six, Fif, one)들에 대한 인식 스코어를 합하여 145.415가 된다. 이 값은 일차적으로 각 인식단위에 대한 단어인식 결과를 수행하고, 이 과정에서 발생한 로그 확률값을 더한 것으로서 화자확인을 위한 된다. 그런데, 만일 speaker1에 대한 임계값이 인식스코어보다 높으면 이 어구는 화자speaker1이 아닌 사칭자의 발성으로 판단하고, 인식스코어보다 낮으면 이 어구는 speaker1의 발성으로 판단한다.

HTK를 사용하는 인식실험에서 HMM 모델들에 대한 음성데이터의 복호화(decoding), 인식 과정은 HTK의 HVite 도구에 의해 수행된다.

3.6 성능평가 방법

본 연구에서는 HTK를 이용하여 각 특징 파라메타 및 처리 방법을 적용한 경우에 대해 깨끗한 음성데이터와 채널 데이터의 화자확인 성능을 평가하였다. 성능평가 척도는 의뢰인을 거부하는 오인 거부율(FR, false rejection)과 사칭자를 수락하는 오인 수락률(FA, false acceptance)이 같아지는 지점에서의 에러율인 EER(equal error rate)을 사용하였다. 확률 임계치가 높아지면(log probability가 낮아지면) 오인 거부율이 높아지는 반면 오인 수락률은 낮아진다. 즉, 오인 수락율과 오인 거부율은 동시에 향상될 수 없는 관계에 있다. 본 연구에서는 EER을 화자별로 구하지 않고, 남자 및 여자의 두 가지로만 하였다. 즉, 임계값은 남자, 여자에 대해 두 가지 값으로 고정된다.

IV. 실험 및 결과

HMM을 기반으로 한 인식 실험은 훈련 데이터의 양, 특징 파라메타의 종류, 음성을 표현하는 HMM 모델 등의 조건에 따라 많은 차이를 보인다. 본 연구에서는 훈련 및 인식 알고리즘, 음성을 표현하는 HMM 모델의 위상, 훈련 데이터의 양을 동일하게 하고, 특징 파라메타 및 잡음처리 방법을 변화시키면서 성능평가의 지표인 EER을 비교하였다.

4.1 실험의 구성

각 특징 파라메타의 화자 확인 성능비교를 위하여 차수를 12차로 동일하게 하였다. 다만, PLP는 12차가 아닌 6차를 적용했을 때 좋은 성능을 나타내었다. 본 실험에서 사용된 옵션은 다음과 같다.

BPL : band pass liftering
CMS : cepstral mean subtraction
BPF : 전처리에서의 band pass filtering
HPF : 전처리에서의 high pass filtering
E : log energy
D : cepstral derivatives (델타 성분)
A : cepstral acceleration (델타-델타 성분)
Q : quefrency liftering

화자확인 측면에서의 성능 척도인 EER 및 음성인식 측면에서 성능 평가 기준이 될 수 있는 인식단위의 인식 에러율(SWU_error, sub-word unit error rate)을 동시에 나타내었다. 이것은 본 화자확인 실험에서 적용했던 특징 파라메타 및 처리 방법들이 음성인식에서도 동일하게 사용될 수 있으므로, 음성인식에서 좋은 성능을 보이는 방법들이 채널환경의 화자확인에서도 좋은 성능을 나타내는지 확인해 볼 필요가 있기 때문이다.

4.2 결과 분석

표 3에는 각각의 특징 및 적용 방법에 대한 성능을 비교하여 나타내었다. 화자확인 성능의 척도인 EER 및 단어인식 성능의 척도인 SWU_error을 남, 여 화자별로 제시하고 그 평균치를 나타내었다.

표 3. 특징 및 적용방법의 성능 비교
Table 3. Performance of each feature and processing.

특징 파라메타	EER(%)			SWU_error(%)		
	남	여	평균	남	여	평균
Clean speech						
LPC_ED	2.08	7.34	4.71	2.56	4.67	3.62
LPCC_ED	***0.42***	***2.26***	***1.34***	***0.21***	***1.48***	***0.85***
MFCC_ED	1.11	4.55	2.83	0.35	2.85	1.60
Channel speech						
LPC_ED	11.95	19.70	15.83	5.29	11.21	8.25
LPCC_ED	3.93	7.21	5.57	1.25	5.86	3.56
LPCC_EDQ	4.01	6.85	5.43	1.28	5.76	3.52
MFCC_ED	5.09	8.31	6.70	2.08	8.74	5.41
MFCC_EDQ	4.30	8.46	6.38	2.02	8.81	5.42
MFCC_ED + BPL	5.04	8.47	6.76	2.08	8.57	5.33
MFCC_ED + CMS	4.28	8.69	6.49	2.23	8.42	5.33
BPF (150-3800Hz) + MFCC_ED	***2.50***	***8.02***	***5.26***	1.98	8.21	5.10
PLP_D	7.87	9.15	8.51	2.52	5.67	4.10
PLP-RASTA	8.36	12.55	10.46	2.58	6.70	4.64
LSP_D	4.72	6.80	5.76	***1.49***	***3.95***	***2.72***

깨끗한 음성데이터에 대한 화자확인 결과에서는 LPCC_ED를 적용하였을때 평균 EER이 1.34%(남자 0.42%, 여자 2.26%)로서 가장 좋은 성능을 보였다. 반면, 채널 데이터의 경우는 전처리 단계에서 대역통과 필터(150~3800Hz)를 적용하고, MFCC_ED를 추출하였을 때 평균 5.26%(남자 2.50%, 여자 8.02%)로 가장 우수하게 나타났다. 이것은 150Hz이하 및 3800Hz 이상의 대역에 존재하는 저주파 채널잡음 및 왜곡된 신호를 제한함으로써 화자들의 변별력을 높였기 때문으로 판단된다.

단어인식 결과에서는 깨끗한 음성데이터의 경우 LPCC_ED를 적용하였을 때 평균 EER이 0.85%(남자 0.21%, 여자 1.48%)로 가장 낮게 나타났고, 채널 데이터의 경우 LSP_D를 사용하였을 때 평균 2.72%(남자 1.49%, 여자 3.95%)로 가장 좋은 성능을 보였다. 전체적으로 보면, 단어인식 성능보다 화자확인 성능이 떨어짐을 알 수 있다. 각각의 특징 및 처리 방법에 대한 분석은 다음과 같다.

대역통과 리프터링의 적용은 화자인식 및 단어인식 측면에서 볼 때, 사용하지 않는 것에 비해 거의 향상되지 않는다. 이것은 대역통과 리프터링이 저주파 및 고주파 영역의 켑스트럼 성분을 동시에 감소시키기 때문이다. 반면, quefrency 리프터링은 LPCC_ED의 경우 평균 0.14%(남자 0.08%저하, 여자 0.36%향상), MFCC의 경우 0.32%(남자 0.79%향상, 여자 0.15%저하)의 향상을 가져왔다. 이것은 quefrency 리프터링이 저주파 영역보다는 고주파 영역을 강조하기 때문이다. 이것으로 보아 전화선 채널 데이터는 저주파 영역보다는 고주파 영역의 켑스트럼 성분에 화자의 정보를 많이 포함하고 있는 것으로 판단된다.

PLP의 적용 결과를 보면 화자 확인 측면에서는 성능이 많이 저하되었으나, 단어 인식 측면에서는 MFCC에 대한 결과보다 좋게 나타났다(BPF+MFCC_ED 5.10%, PLP_D 4.10%). 이것으로 보아 PLP는 채널의 잡음환경에서는 효율적이지 못한 것을 알 수 있다. PLP과정에 RASTA를 적용한 PLP-RASTA는 화자확인 결과가 PLP를 적용한 것보다 좋지 않게 나타났다(PLP_D 8.51%, PLP-RASTA 10.46%).

본 실험에서는 채널의 효과를 켑스트럼 영역에서 제거하는 CMS가 좋은 성능을 나타내지 못했다. 이것은 수집된 숫자음 DB가 한 어구내에서 자음과 모음 등이 고른 분포를 갖지 않아 깨끗한 음성데이터의 켑스트럼 평균이 0이 되지 않음에도 불구하고 평균값을 차감함으로써 오히려 음성 정보를 왜곡시켰기 때문인 것으로 판단된다.

표 4에는 통과대역을 변화시키면서 음성 신호를 필터링하여 실험한 결과가 나타나 있다. 특징 파라메타는 모두 MFCC_ED를 적용하였다.

표 4. 필터의 통과대역에 따른 성능
Table 4. Performance of filters having different passband.

통과 대역	0 ~ 4000 Hz	0 ~ 3800 Hz	150 ~ 4000 Hz	150 ~ 3800 Hz	300 ~ 3400 Hz
평균EER (%)	6.70	6.25	5.265	5.26	6.19

음성신호의 전대역을 사용하여 특징을 추출하는 경우와 가장 성능이 우수하게 나타나는 150~3800Hz의 대역통과 필터링을 하였을 때의 성능은 평균 1.44% 차이를 보인다. 그리고, 일반적으로 전화선 채널의 통과대역이라고 알려진 300~3400Hz의 통과대역을 가진 필터를 적용하였을 때는 6.19%로 성능이 현저하게 저하된다. 이것으로 보아 300~3400Hz 이외의 저대역 및 고대역에도 화자확인에 필요한 정보가 포함되어 있음을 알 수 있다. 결과적으로, 통과대역이 150~3800Hz일 때 성능이 가장 좋고, 대역이 더 확장된다면 성능은 더 떨어진다.

위에서 LPC, LPCC, MFCC의 경우 모두 로그 에너지와 델타 계수들을 포함하는데, 델타-델타 계수는 사용하지 않았다. 이것은 실험결과 델타-델타 계수를 사용하는 것은 화자확인 및 단어인식 성능에서 좋지 않게 나타났기 때문이다. 표 5에 에너지, 델타, 델타-델타에 대한 성능비교가 나타나 있다.

표 5. 에너지, 델타, 델타-델타의 성능 비교
Table 5. Performance of enrgy, delta, delta-delta.

특징 파라메타	EER(%)			SWU_error(%)		
	남	여	평균	남	여	평균
MFCC	5.84	8.22	7.03	*1.67*	*9.00*	*5.34*
MFCC_E	5.54	8.22	6.88	3.55	10.55	7.05
MFCC_ED	*5.09*	*8.31*	*6.70*	2.08	8.74	5.40
MFCC_EDA	5.35	9.39	7.37	2.44	8.81	5.63

표 5로부터 에너지 및 델타성분을 특징 파라메타로 사용하는 것은 화자확인 성능을 향상시키지만, 델타-델타 성분을 추가하는 것은 오히려 EER을 증가시키는 것을 알 수 있다. 단어인식 결과에서도 이러한 사실을 확인할 수 있다.

일반적으로 단어 인식에서 시간 미분치인 델타 성분을 사용하는 것은 시간에 따라 변화하는 음성신호를 모델링하기 위함이다. 그러므로, 각각의 프레임에 해당하는 특징벡터의 시간에 따른 변화정도를 음성인식 및 화자인식에서 별도의 특징으로 사용할 수 있는 것이다. 그러나, 델타-델타의 경우는 델타 성분들로부터 구해진 델타 성분이므로, 단어인식 및 화자인식을 위한 적절한 특징을 포함하지 않는 것으로 판단된다. 결과적으로 델타-델타의 사용은 화자확인의 성능을 저하시켰다.

위의 실험에서 제시한 EER은 성별로 다른 임계값에 대해 결정되는 값이므로, 각 화자가 어떤 EER을 갖는지는 알 수 없다. 표 6에서는 MFCC_ED을 적용하였을 때 각 화자별로 EER을 비교해 보았다.

표 6. 각 화자의 EER 비교
Table 6. Performance of each speaker.

	화 자	EER(%)	평균(%)
남	Speaker 1	1.39	5.09
	Speaker 2	3.85	
	Speaker 3	*8.66*	
	Speaker 4	2.65	
	Speaker 5	2.50	
	Speaker 6	2.50	
여	Speaker 7	3.85	8.47
	Speaker 8	6.35	
	Speaker 9	*11.25*	
	Speaker 10	*18.11*	
	Speaker 11	5.13	
	Speaker 12	0.12	

표 6으로부터 남자 화자 중에서는 speaker3이, 여자 화자 중에서는 speaker9와 speaker10이 다른 화자들과의 변별력에서 상당히 떨어짐을 알 수 있다. 이 화자들은 다른 화자들에 비하여 발성이 분명하지 않아서 (특히, speaker9와 speaker10의 경우 '십'을 발성할 때 마찰음인 'ㅅ'음에 가깝게 발음하였다), 화자 확인에서 필요한 화자고유의 특성을 갖추지 못한 것으로 파악된다.

VI. 결 론

본 논문에서는 전화선 채널이 화자확인의 성능에 미치는 영향을 분석하고, 채널의 영향을 보상 또는 최소화하기 위한 여러 가지의 전처리 및 특징 추출 방법을 적용해 보았다. 적용한 음성 특징 파라메타는 LPC, LPCC, MFCC, PLP, LSP이며, 전처리 과정으로는 필터링을 적용하였다. 추출된 특징으로부터 채널의 영향을 보상하기 위한 방법으로는 cepstral weighting, CMS, 그리고 RASTA를 적용하였다. 훈련 및 인식 실험은 HTK를 사용하였다.

전화선 채널의 대역폭은 보통 300~3400Hz이라고 알려져 있다. 그러나, 실험결과에서는 화자확인의 전처리 과정에서 300~3400Hz가 아닌 150~3800Hz로 대역제한하였을 때 성능이 더 좋게 나타났다. 이것을 통하여 일반적인 채널의 대역폭 150~3800Hz보다 확장된 영역의 음성신호에 화자확인을 위한 정보가 포함되어 있음을 알 수 있었다. 또한, 150~3800Hz의 대역을 갖는 대역통과 필터를 적용하고 특징 파라메타는 MFCC를 적용하였을 때 화자확인 성능이 가장 우수하였다.

대역통과 필터나 잡음처리 방법을 적용하지 않고 특징만을 추출하였을 경우, LPC기반의 LPCC가 화자확인 및 단어인식 성능에 있어서 FFT기반의 MFCC보다 우수하였다. LPCC는 깨끗한 음성데이터에서도 가장 성능이 좋았다. 이것은 LPC분석이 음성신호의 스펙트럼을 평활화하여 기울기(envelope) 정보만을 추출하는 반면, FFT분석은 스펙트럼의 기울기뿐만 아니라 하모닉(harmonic) 정보를 동시에 갖기 때문으로 판단된다. FFT기반의 멜 스케일(mel-scale) 필터 에너지로부터 얻어지는 하모닉 성분은 특히 저주파 영역에서 변이가 심하여 인식률을 저하시킨다.

화자확인 성능은 화자에 따라 차이를 보이며, 특히 발성이 부정확하고 명확하지 않은 화자의 EER이 높게 나타났다. 발성이 명확하지 않는 화자의 발성은 훈련과 인식에서 일정한 패턴을 유지하지 못하기 때문이다.

성별에 따른 화자확인 결과를 비교해 보면, 깨끗한 음성데이터 및 채널데이터 모두에서 남자 화자는 적용한 방법들에서 향상된 인식률을 보였지만, 여자 화자는 상대적으로 EER이 높게 나타났다.

참 고 문 헌

1. Sadaoki Furui, M.Mohan, *Advances in Speech Signal Processing*, Dekker, 1996
2. Richard J.Mammone, "Robust Speaker Recognition-A Feature-based Approach", IEEE Signal Processing Magazine, pp.58-64, 1996
3. 백상훈, 끝점검출이 내재된 실시간 고립단어 인식 알고리즘에 관한 연구, 인하대학교 전자공학과 대학원 석사학위 논문, 1996
4. Hynek Hermansky, "Perceptually Based Linear Predictive Analysis of Speech", *ICASSP*, pp.509-512, 1985
5. Joseph P.Campbell, "Speaker Recognition: A Tutorial", *Proceedings of the IEEE*, Vol 85, No. 9, pp.1437-1462, 1997
6. Noboru Sugamura, "Quantizer Design in LSP Speech Analysis-Synthesis", *IEEE, Journal on Selected Areas in Communications*, Vol 6, No 2, 1988
7. F.K.Soong, B.H.Juang, "Line Spectrum(LSP) and Speech Data Compression", *ICASSP*, Vol 1, pp.101-104, 1984
8. Steve Young, *The HTK Book*, Entropic, 1996
9. Sadaoki Furi, "Cepstral Analysis Technique for Automatic Speaker Verification", *ICASSP*, pp.254-271. 1981
10. D. OShaughnessy, "Speaker Recognition", *IEEE ASSP Magizine*, pp.4-17, Oct. 1986
11. 오영환, 음성언어 정보처리, 홍릉과학출판사, 1997
12. Hynek Hermansky, Nelson Morgan, "RASTA Processing of Speech", *IEEE Transactions on speech and audio processing*, Vol 2, No. 4, pp.578-589, 1994
13. Joachim Koehler, Nelson Morgan, "Integrating RASTA-PLP into Speech Recognition" *ICASSP*, Vol 1, pp.421-424, 1994
14. Joseph P.Campbell, "Testing with the YOHO CD-ROM Voice Verification Corpus", *ICASSP*, pp.1100-1106, 1995
15. Charles Jankowski, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database", *ICASSP*, pp.109-112, 1990
16. Pedro J.Moreno, Richard M.stern, "Sources of Degradation of speech recognition in the telephone network", *ICASSP*, Vol 1, pp.109-112, 1994

▲조 태 현(Tae-Hyun Cho)

1997년 2월 : 인하대학교 전자공학과 학사
1999년 2월 : 인하대학교 전자공학과 석사
현재 : LG 정보통신, 연구원
※주관심분야 : 화자인식, 잡음처리

▲김 유 진(Eu-Gene Kim)

1995년 2월 : 인하대학교 전자공학과 학사
1997년 2월 : 인하대학교 전자공학과 석사
1997년 2월~1998년 5월 : LG반도체 연구원
현재 : 인하대학교 전자공학과 박사과정
※주관심분야 : 화자인식, 음성인식

▲이 재 영(Jae-Young Lee)
1995년 2월 : 성균관대학교 전자공학과 학사
1997년 2월 : 성균관대학교 전자공학과 석사
현재 : 나래이동통신 연구원
※주관심분야 : 음성인식, 자연어처리, DSP

▲정 재 호(Jae-Ho Chung)

1982년 : University of Maryland (학사)
1984년 : University of Maryland (석사)
1990년 : Georgia Institute of Technology (박사)
1984년~1985년 : 미국 국방성 산하 해군 연구소, 신호처리 연구실, 연구원
1991년~1992년 : AT&T Bell Laboratories, 음성신호처리 연구실, 연구원 (MTS)
1992년~현재 : 인하대학교 공과대학 전자공학과, 현(부교수)