한국산업정보학회논문지 제6권 제4호 (2001. 12)

다층 퍼셉트론에 기반한 한국어 숫자음 인식시스템 구현을 위한 특징 연구

(A Study on the Features for Building Korean Digit Recognition System Based on Multilayer Perceptron)

김 인철*, 김 대영**
(In-Cheol Kim Dae-Young Kim)

요 약 본 논문에서는 한국어 숫자음 인식을 위해 다층 퍼셉트론을 이용한 인식시스템을 구현하였으며 음성인식 분야에서 일반적으로 널리 사용되는 여러 종류의 특징을 인식시스템의 입력으로 적용하여 각각의 인식 성능 및 특성을 알아보았다. 이를 위해 Mel-scale-Filterbank 계수, MFCC, LPCC, 그리고 PLP 계수를 입력 특징으로 사용하였다. 본 논문에서는 제한된 환경이 아닌 여러 종류의 잡음이 존재하는 일반적인 환경에서도 견실한 성능을 보일 수 있는 인식시스템을 구현하기 위해 잡음이 거의 포함되지 않은 음성 데이터뿐만 아니라 잡음이 첨가된 음성 데이터에 대해 인식 실험을 각각 수행하였다. 실험에서는 20개의 한국어 숫자음에 대한 인식 실험을 수행하였으며 그 결과로부터 Mel-scale Filterbank 계수가 잡음의 첨가 유무에 관계없이 화자 종속 및 화자 독립적인 음성 데이터에 대해 가장 견실한 인식 성능을 보임을 확인할 수 있었다.

Abstract In this paper, a Korean digit recognition system based on a multilayer Perceptron is implemented. We also investigate the performance of widely used speech features, such as the Mel-scale filterbank, MFCC, LPCC, and PLP coefficients, by applying them as input of the proposed recognition system. In order to build a robust speech system, the experiments for demonstrating its recognition performance for the clean data as well as corrupt data are carried out. In experiments of recognizing 20 Korean digit, we found that the Mel-scale filterbank coefficients performs best in terms of recognition accuracy for the speech dependent and speech independent database even though noise is considerably added.

1. 서 론

현재의 인간과 컴퓨터간의 상호 작용(human computer interaction: HCI)은 대부분 GUI(graphical user interface) 환경 하에서 키보드, 마우스, 그리고 조이스틱 등을 이용하여 이루어지고 있다. 그러나 컴퓨터의 활용 분야가 광범위해지고 다양하고 많은 양의 정보들을 효과적으로 처리하고자 하는 요구가 급등하고 있는 현대의 지식 정보화 사회에서는 기존의 인터페이스 기술이 정보의 원활한 흐름을 저해하는 주 요인이 되고 있다. 특히, 최근의 컴퓨팅 환경은 휴대성과 편의성 위주로 급변하고 있는데 이러한 변화에 적절히 대처하기 위해서는 좀더 자연스럽고 지능적인 인터페이스 기법이 매우 요구된다. 인간의 자연스러운 통신 수단인 음성을 이용하여 자연스러운 HCI 환경을 구현하고자 하는 연구가 지난 수십 년 동안 활발히 이루어져 왔으며 현재도 꾸준히 그 연구가 계속되고 있다. 음성은 인간의 가장 자연스러운 통신 수단으로서, 정보통신 시스템의 발전에 따라 기계와 인간사이의 정보 교환의 필요성이 점점 증가하게 되면서 원활하고 자연스러운 통신을 위해 음성을 사용하는 통신 수단 즉, 기계가 음성을 인식할 수 있도록 하는 연구가 활발히 진행되어 왔다.

*경북대학교 전자전기컴퓨터학부
**계명문화대학 멀티미디어계열

음성 인식시스템의 구현을 위해 DTW(dynamic time warping)[1], 벡터 양자화[2] 및 GMM(gaussian mixture model)[3] 등의 여러 가지 방법이 사용되고 있으나 최근 들어 음성 신호를 통계적인 방식으로 모델링 하는 은닉 마르코프 모델(hidden Markov model: HMM)[4][5]을 이용한 방법과 오차 역전파 학습 알고리듬에 기반한 다층 퍼셉트론(multilayer perceptron: MLP)[6] 및 음성의 동적 특성을 반영하기 위한 TDNN(time-delay neural network)[7] 등과 같은 신경회로망 기법이 널리 사용되고 있다. HMM은 현재 음성 인식 연구에서 가장 널리 이용되는 인식 모델로서 음성 신호의 변동을 통계적으로 처리하고 이 통계량을 확률 형태의 모델에 반영하여 음성을 인식하는 방법이다. 신경회로망은 문자인식과 같은 다른 형태의 인식 분야에서는 널리 사용되고 있으나 주어진 입력의 동적 특성 즉, 시간에 따른 입력의 변화를 제대로 반영하지 못하는 문제가 있어 음성인식 분야에서는 HMM에 비해 그 적용이 활발히 이루어지지 않았다. 최근 들어서는 신경회로망의 분별(discriminant) 특성과 HMM의 시공간적 신호에 대한 통계적 모델링 특성을 결합한 하이브리드 인식 기법이 많이 연구되고 있다[8][9].

본 논문에서는 한국어 고립 숫자음 인식을 위해 전술한 인식 방법 중에서 오차 역전파 학습 알고리듬[10]에 기반한 다층 퍼셉트론을 이용해 인식시스템을 구현하였으며 LPCC (Linear predictive coding cepstrum)[11], Mel-scale Filterbank 계수, MFCC(Mel scale frequency cepstral coefficient)[12], 그리고 PLP(Perceptual linear prediction) 계수[13]를 입력 특징으로 적용하여 각각의 인식률을 비교 분석하였다. 다층 퍼셉트론은 HMM에 비해 신호의 동적 특성을 잘 반영하지는 못하나 분별 특성이 상대적으로 뛰어나 각 입력 특징별 인식 성능을 분석하기 위한 인식기로 적절히 사용될 수 있다. 또한 본 논문에서는 여러 종류의 잡음이 존재하는 일반적인 환경에서도 높은 인식 성능을 얻을 수 있는 인식시스템을 구축하기 위해 잡음이 섞인 음성 데이터 대한 인식 실험도 수행하였다.

부가적으로 본 논문에서는 일반적인 인식시스템이 학습 과정에 참가한 화자로부터 획득한 음성 데이터에 대해서만 인식을 수행하는 화자 종속적인 환경에서는 높은 인식률을 나타내나 임의의 화자에 대해 인식을 수행하는 화자 독립적 인식 환경에서는 음성의 화자 의존성으로 인해 그 인식률이 상대적으로 떨어지는 것을 고려해 화자 독립적인 음성 데이터에 대한 인식 실험도 수행하였다. 인식 실험에서는 화자 20명으로부터 하나에서 열 및 영에서 구까지의 20개의 한국어 고립 숫자음에 대한 음성 데이터를 획득하여 전술한 각 입력 특징 별로 인식 실험을 수행하고 그 결과를 분석하였다.

2. 데이터베이스

본 논문에서는 동일 지역에 거주하는 성인 남녀 20명이 20번씩 발음한 하나에서 열 그리고 영에서 구까지의 20개의 한국어 고립 숫자음을 데이터베이스로 사용하였다. 음성 데이터의 획득은 비교적 주변 잡음이 적은 연구실에서 이루어졌으며 각 음성은 16kHz 샘플링 과정과 16 bit 양자화 과정을 거쳐 입력 데이터로 사용된다. 추출된 각 음성 데이터를 이용하여 다층 퍼셉트론 기반의 음성 시스템에 대해 학습 및 인식 실험을 수행하기 위해서는 먼저 신호의 길이에 관계없이 동일한 크기의 입력 특징 벡터를 추출하는 것이 필요하다. 실험에서는 각 음성 데이터를 음성 신호의 길이에 따라 가변적인 폭을 가지는 19 개의 프레임으로 나눈 후에 Hamming 창을 반 프레임씩 겹치도록 적용하여 생성된 총 37개의 프레임으로부터 특징을 추출한 뒤에 전체 특징값을 프레임의 순서별로 정렬하여 인식기에 대한 하나의 입력 특징 벡터로 사용하였다. 본 논문에서는 인식시스템의 학습을 위해 15명의 화자가 10번 발음한 총 3000개의 숫자음을 학습용 데이터로 사용하였으며 동일 화자의 나머지 3000개의 숫자음을 화자종속 인식 실험을 위한 검사용 데이터로 사용하였다. 또한 학습에 참가하지 않은 또 다른 5명의 화자로부터 획득한 1000개의 음성 데이터는 인식시스템의 화자 독립에 대한 인식 성능 검사를 위해 사용하였다.

3. 음성 특징

3.1 LPCC(Linear Predictive Coding Cepstrum)

음성 신호에 대한 LPC 해석은 아래 식 1에서와 같이 어느 시점에서의 음성 샘플값은 그 이전에 샘플링 된 음성 신호의 선형적인 결합으로 예측할 수 있다는 이론에 근거한 것이며 식 2와 3을 통해 음성 신호는 AR(auto regressive) all pole 모델로 모델링 된다[11].

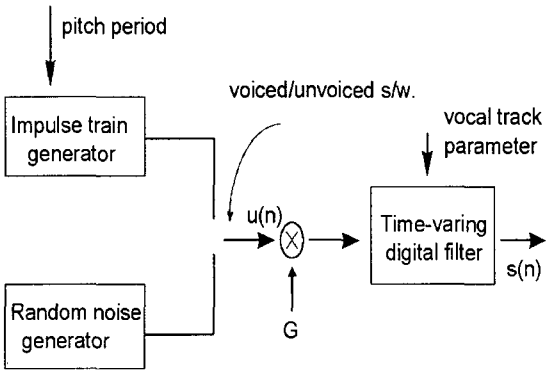
$$\hat{s}(n) = a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p) \quad (1)$$

$$s(n) = \hat{s}(n) + Gu(n) \quad (2)$$

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \quad (3)$$

LPC 방식에 의한 음성 생성 모델은 그림 1과 같이 나타낼 수 있으며 결과적으로 실제 음성 샘플값과 예측된 샘플값

사이의 차를 최소화하는 선형 예측 계수, a_p를 구하는 것은 음성 생성 모델에서 음성 구강(vocal track)의 형태를 필터로 가정하고 그 필터 계수를 구하는 것과 같다.



<그림 1> LPC 기반의 음성 생성 모델

선형 예측 계수를 구하는 방법은 Autocorrelation 방법과 Covariance 방법이 있으나 본 논문에서는 Autocorrelation 방법에 기반한 Levinson-Durbin 알고리듬을 이용하여 12차의 선형 예측 계수를 구하였다. 선형 예측 계수는 외부 잡음에 매우 민감한 특성을 나타내므로 본 논문에서는 선형 예측 계수로부터 잡음에 좀더 견실한 특성을 가지는 cepstral 계수, c_n를 식 4에서와 같이 구하여 입력 특징으로 사용하였다.

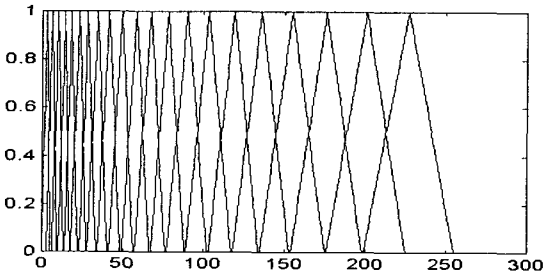
$$c_1 = -a_1$$
$$c_n = -a_n - \sum_{m=1}^{n-1}(1-\frac{m}{n})a_m c_{n-m} \qquad 1<n\le p \quad (4)$$

실험에서는 37개의 프레임으로 나누어진 음성 데이터에 대해 각 프레임별로 전술한 과정을 통해 12차의 cepstrum을 추출하였으며 결과적으로 하나의 음성 데이터에 대해 차원이 444인 특징 벡터를 추출하여 인식시스템 입력 특징으로 사용하였다.

3.2 Mel-scale filterbank 및 MFCC

인간의 귀는 음성이 들어오면 주파수 대역상에서 비선형적으로 음성을 분석하는 것으로 알려져 있다. 즉, 저주파 영역에서는 높은 해상도를 가지고 고주파 영역으로 갈수록 그 해상도가 떨어져 개략적으로 음성을 분석한다. 이러한 귀의 음성 분석 방식을 이용해 음성을 주파수 영역으로 변환하여 저주파 대역에서는 선형적이고 고주파 대역에서는 대수적인 간격을 가지는 Mel-scale 혹은 Bark-scale에 따라서 음성을 분석하고 특징을 추출하는 방법들이 제안되었으며 그 중에서 가장 널리 사용되는 방법이 Mel-scale filterbank이다[12]. 이 방법은 음성 데이터의 각 프레임을 퓨리에 변환 과정을 통해 주파수 영역으로 변환 한 후에 그 파워 스펙트럼을 구하고 아래 식 5에 정의된 Mel-scale을 이용하여 주파수 대역 상에서 동일한 Mel-scale 간격을 가지는 그림 2와 같은 심각(triangular) 대역 통과 필터를 적용함으로써 얻어진 각 필터별 출력값에 대한 log-energy 값을 주어진 음성 데이터의 특징으로 사용한다.

$$Mel(f) = 2595 \log_{10}(1+\frac{f}{700}) \qquad (5)$$



<그림 2> Mel-scale 삼각 대역 통과 필터

실험에서는 37개의 프레임으로 나누어진 음성 데이터에 대해 각 프레임 별로 퓨리에 변환을 수행한 후 그 파워 스펙트럼에 대해 258.18*Mel* 간격을 가지는 21개의 삼각 대역통과 필터를 적용해 각 필터별로 log-energy값을 추출함으로써 한 음성 데이터에 대해 총 777개의 특징값을 추출하였다.

MFCC는 한 프레임에서 구해진 21개의 필터별 log-energy 값에 대해 아래 식 6에 정의된 이산 코사인 변환(discrete cosine transform)을 적용해 구할 수 있으며 본 논문에서는 각 프레임 별로 16차의 MFCC를 추출함으로써 한 음성 데이터에 대해 총 592개의 MFCC를 구해 인식시스템의 입력 특징으로 사용하였다.

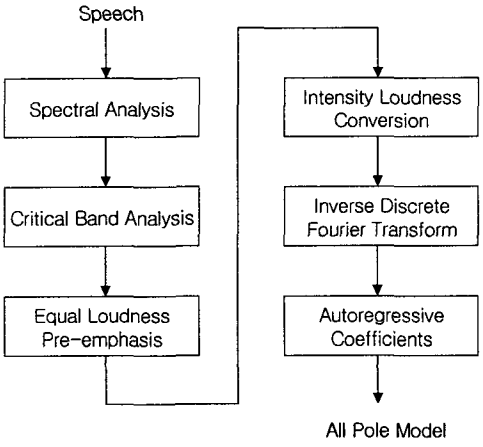
$$c_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{N} m_j \cos(\frac{\pi i}{N}(j-0.5)) \qquad (6)$$
where, i = 1, 2, . . . M

3.3 PLP(Perceptual Linear Prediction) 계수

LPC 모델은 Autoregressive all pole 모델로서 음성 신호 해석 시에 모든 주파수 영역에서의 신호 성분을 동일하

게 반영하는데 이것은 인간의 청각 특성과 맞지 않다. 인간의 청각 특성은 음성 신호의 저주파 성분을 자세히 반영하고 고주파 성분을 개략적으로 반영하는데 이러한 인간의 청각 특성에 맞게 LPC 모델을 수정한 것이 PLP 모델이다. PLP 모델은 음성 신호 해석 시 그림 3에 나타낸 과정을 통해 음성 신호의 파워 스펙트럼을 인간의 청각 특성과 유사하게 변형시킨 뒤 AR 모델로 근사화 시킨 것이다[13]. PLP 계수를 구하기 위해서는 먼저 음성 데이터의 각 프레임에 대해 퓨리에 변환을 적용하여 얻어진 파워 스펙트럼의 주파수 단위를 식 7을 이용하여 Bark-scale로 변환한다.

$$\Omega(\omega) = 6 \ln \omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{0.5} \quad (7)$$



<그림 3> PLP 모델을 이용한 음성 신호 해석

Bark-scale로 변환된 파워 스펙트럼은 식 8에 나타낸 일종의 청각 필터 형태인 임계 대역 곡선(critical band curve)과의 convolution을 통해 임계 대역 파워 스펙트럼인 $\Theta(\Omega)$으로 변환된다.

$$\psi(\Omega) = \begin{cases} 0 & \text{for } \Omega \leftarrow 1.3 \\ 10^{2.5(\Omega+0.5)} & \text{for } -1.3 \leq \Omega \leq -0.5 \\ 1 & \text{for } -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & \text{for } 0.5 \leq \Omega \leq 2.5 \\ 0 & \text{for } \Omega > 0.5 \end{cases} \quad (8)$$

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\psi(\Omega) \quad (9)$$

상대적으로 넓은 대역의 임계 대역 곡선 $\psi(\Omega)$과의 convolution은 원래의 파워 스펙트럼에 비해 스펙트럼 분해능의 상당한 저하를 가져오며 이로 인해 $\Theta(\Omega_i)$의 하향 샘플링(down sampling)이 가능해진다. 실험 과정에서는 Hermansky가 제안한 대로 대략 1 Bark scale 단위로 샘플링 하였다. 정확한 샘플링 간격은 식 9에 나타낸 바와 같이 적분 구간이 분석하려는 전 주파수 대역을 포함할 수 있어야 하며 일반적으로 분석 대역이 0에서 5kHz인 경우에는 0.994 Bark scale 단위로 18개의 $\Theta(\Omega_i)$ 샘플을 추출해 사용한다. 샘플링 된 $\Theta(\Omega_i)$는 서로 다른 주파수에서 인간의 청각 민감도가 다른 것을 근사화 한 함수 $E(\omega)$에 의해 pre-emphasis 과정을 거친다. 함수 $E(\omega)$와 pre-emphasis 과정을 식 10과 11에 나타내었다.

$$E(\omega) = \frac{[(\omega^2 + 56.8 \times 10^6)\omega^4]}{[(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9)]} \quad (10)$$

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \quad (11)$$

음성신호를 All-pole 모델링 하기 전 마지막 단계로서 식 11을 통해 pre-emphasis 된 $\Xi[\Omega(\omega)]$에 대해 아래 식 12에 나타낸 세제곱근 크기 압축(cubic-root amplitude compression)을 적용한다.

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \quad (12)$$

이 과정은 청각의 전력법칙을 근사화 하고 소리의 실제 크기와 귀에 감지된 크기 사이의 비선형 관계를 모방한다. 또한 all-pole 모델링 시 상대적으로 낮은 차수로 모델링 될 수 있도록 임계 대역 스펙트럼의 크기 변화를 줄여준다. 식 12에 나타낸 과정을 통해 구해진 $\Phi(\Omega)$는 최종적으로 이산 푸리에 역변환을 통해 all-pole 모델로 모델링 된다. 본 논문에서는 주어진 음성 신호를 전술한 과정을 통해 각 프레임 별로 5차의 all-pole 모델로 모델링 하여 그 계수를 특징으로 사용하였으며 결과적으로 37개의 프레임으로 이루어진 음성 데이터에 대해 총 185개의 PLP 계수를 추출해 인식시스템의 입력 특징으로 사용하였다.

4. 실 험

본 논문에서는 하나에서 열 그리고 영에서 구까지의 20개의 한국어 숫자음 인식을 위해 다층 퍼셉트론에 기반한 인식시스템을 구현하였다. 이때 다층 퍼셉트론의 학습은 역전파 알고리듬에 의한 학습 과정에서 흔히 발생하는 조기 포화(premature saturation)[14] 현상으로 인한 학습 지연을

방지하기 위해 Fahlman[15]이 제안한 수정된 오차 역전파 알고리듬을 사용하여 수행되었다. 또한 인식시스템의 입력 특징으로는 전술한 과정을 통해 추출된 LPCC, Mel-scale filterbank 계수, MFCC, 그리고 PLP 계수를 사용하였으며 인식 실험을 통해 각각에 대한 인식 성능을 비교 분석하였다. 또한 본 논문에서는 잡음이 섞인 음성 데이터에 대한 인식 실험과 화자 종속 및 화자 독립의 경우에 대한 인식 시스템의 성능 변화에 대한 검토를 수행하였다.

4.1 학 습

인식 실험에 사용된 다층 퍼셉트론의 구조는 입력 특징에 따라 다르게 설정된다. 즉, LPCC를 입력 특징으로 사용하는 경우에는 한 음성 데이터로부터 차원이 444인 특징 벡터가 추출되므로 다층 퍼셉트론의 입력층 노드의 수는 444개가 되며 이때 은닉층 노드의 수는 30개가 사용된다. 또한 출력층 노드의 수는 인식하려는 숫자음의 개수와 동일한 20개로 설정하였다. 또한 Mel-scale filterbank의 경우에는 777개의 입력 노드와 35개의 은닉층 노드를 가지는 퍼셉트론이 사용되며 MFCC와 PLP 계수는 각각의 입력 노드가 592와 185개 그리고 은닉층 노드가 30개와 25개로 이루어진 다층 퍼셉트론으로 학습하였다.

다층 퍼셉트론을 이용한 인식시스템 구현 과정에서 은닉층 노드의 수를 적절하게 설정하는 문제는 쉬운 일이 아니다. 은닉층 노드의 수가 학습 데이터의 규모 및 인식의 복잡성에 비해 적은 경우에는 학습이 제대로 이루어지지 않으며 그 반대의 경우에는 학습은 좀더 원활하게 수행되나 다층 퍼셉트론의 일반화(generalization) 특성이 떨어져 그 인식 성능이 저하되며 가중치의 증가로 인해 더 많은 메모리를 필요로 하는 단점이 발생한다. 관련 연구 분야에서 은닉층 노드 수의 설정에 관한 연구 결과들이[16] 발표된 적은 있으나 실제 인식에서 다층 퍼셉트론의 비선형적인 특성 및 학습 데이터의 고차 특성, 그리고 인식의 복잡성 등을 고려한 최적의 노드 수를 구할 수 있는 이론적인 정립이 현재까지 명확하게 이루어지지 않고 있는 실정이다. 따라서 본 논문에서는 try-and-error로 개별 입력 특징별로 다층 퍼셉트론의 은닉층 노드 수를 설정하였다.

다층 퍼셉트론의 학습 과정에서 학습율과 관성항은 전술한 4가지 입력 특징의 경우에 대해 동일하게 0.05와 0.8로 정하였으며 다층 퍼셉트론 내의 각 가중치는 $[-5\times10^{-3}, 5\times10^{-3}]$ 사이의 랜덤한 값으로 초기화하였다. 다층 퍼셉트론의 학습을 위해 15명의 화자가 20개의 숫자음을 10번 발음하여 획득한 총 3000개의 음성 데이터를 학습용 데이터베이스로 사용하였으며 각각의 음성 특징별로 출력층에서의 평균 자승 에러가 0.001이 될 때까지 반복적으로 학습을 수행하였다.

4.2 화자 종속 인식 실험

화자종속 한국어 숫자음 인식 실험은 다층 퍼셉트론의 학습에 참여한 15명의 화자가 10번씩 발음한 또 다른 음성 데이터를 검사용 데이터로 정하여 수행하였다. 또한 실험에서는 검사 데이터에 인위적으로 잡음을 첨가하여 잡음 환경에서의 입력 특징별 인식 성능을 비교 분석하고자 하였다. 이때 첨가된 잡음은 식 13에 정의된 신호 대 잡음비를 기준으로 한 백색 가우시안 잡음이다.

$$[S/N]_{db} = 10 \log_{10}(\overline{s^2(t)}/\overline{n^2(t)}) \qquad (13)$$

<표 1> 다층 퍼셉트론을 이용한 화자 종속 인식 실험 결과

	Clean Data	50db	40db	30db
Filterbank	99.73 (8)	99.67 (10)	99.37 (19)	94.83 (155)
MFCC	99.47 (16)	99.43 (17)	98.60 (42)	89.47 (316)
LPCC	97.13 (86)	96.50 (105)	89.63 (311)	59.47 (1216)
PLP	96.83 (95)	96.47 (106)	92.80 (216)	75.20 (744)

표 1에서는 검사용 음성 데이터에 대해 LPCC, MFCC, Mel-scale filterbank 계수, 그리고 PLP 계수를 각각 입력 특징으로 사용한 다층 퍼셉트론에서의 인식률 및 에러가 발생한 검사 데이터의 수를 나타내었다. 표 1의 결과로부터 각각의 입력 특징별 다층 퍼셉트론의 인식 성능은 잡음이 섞이지 않은 검사 데이터에 대해서는 모두 상당히 높은 인식률을 나타내나 잡음이 첨가 될수록 그 인식률이 떨어짐을 알 수 있다. 특히 LPCC의 경우에는 잡음에 대한 인식 성능의 저하가 가장 크게 나타남을 알 수 있다. 그러나 Mel-scale filterbank 계수를 입력 특징으로 사용한 경우에는 잡음이 섞이지 않은 검사 데이터뿐만 아니라 잡음이 어느 정도 섞인 음성 데이터에 대해서도 평균 99% 이상의 높은 인식률을 보여준다. 특히 다른 입력 특징들이 그 인식 성능이 급격히 저하되는 30db의 신호 대 잡음비 상태에서도 94% 이상의 인식률을 나타냄으로써 잡음에 가장 견실한 특성을 가지고 있음을 알 수 있다.

표 1에 나타낸 인식 결과는 인식시스템에 거절 기능을 포함하지 않고 다층 퍼셉트론의 출력층에서 최고값을 가지는 뉴런을 인식 결과로 선택하는 최고치 선택법을 이용해 얻은 결과이다. 이 경우에는 실제 출력층 뉴런에서의 출력값이 0에 가까운 낮은 값이라도 주어진 입력 특징에 해당하는 인식 결과로 선택 될 가능성이 있어 그 신뢰성에 문

재가 생길 수 있다. 표 2에서는 인식시스템의 최종 인식 단계에서 신뢰도 문턱값(threshold)을 이용한 거절 기능을 도입하여 인식 실험을 수행한 결과를 나타내었다. 인식률은 거절되지 않은 입력 벡터에 대한 인식 결과를 나타낸다. 즉, 출력층에서 얻어진 최고값이 지정된 신뢰도 문턱값보다 낮은 값인 경우에 그 값의 신뢰도가 낮음을 의미하므로 그 인식 결과에 관계없이 거절 기능을 적용하여 잘못된 인식으로 처리하고 문턱값 보다 높은 값을 나타내는 출력값에 대해서만 오인식 여부를 결정한다. 이때 신뢰도 문턱값은 0.5로 정하였다.

<표 2> 거절 기능을 가지는 다층 퍼셉트론을 이용한 화자 종속 인식 실험 결과

	Clean Data	50db	40db	30db
Filterbank	99.17 (25)	99.13 (26)	98.40 (48)	90.87 (274)
MFCC	98.87 (34)	98.73 (38)	96.77 (97)	84.47 (466)
LPCC	95.20 (144)	94.17 (175)	83.80 (486)	51.07 (1468)
PLP	94.50 (165)	94.93 (182)	88.90 (333)	68.17 (955)

Mel-scale filterbank 계수의 경우에 거절 기능을 적용한 인식 실험에서도 가장 높은 인식 성능을 보임을 알 수 있다. 인식 결과에 대한 신뢰도 지수는 아래의 식 14에 정의된 바와 같이 정인식률에 비례하여 나타나므로 Mel-scale filterbank 계수가 신뢰성이 요구되는 분야에 사용하기에 가장 적절한 입력 특징임을 알 수 있다. 그러나 LPCC와 PLP 계수는 그 인식률뿐만 아니라 신뢰성 측면에서도 다른 입력 특징에 비해 그 성능이 크게 떨어짐을 보여준다.

$$\text{신뢰도 지수} = \frac{\text{정인식률}}{\text{정인식률} + \text{오인식률}} \times 100 \qquad (14)$$

4.3 화자 독립 인식 실험

화자 독립 인식 실험은 음성의 화자 내 변질성(intra-speaker variability)으로 인한 인식 성능의 변화를 검사하기 위한 화자 종속 실험과는 달리 음성의 화자 간 변질성(inter-speaker variability)으로 인한 인식시스템의 성능 변화를 검사한다. 이를 위해 전술한 다층 퍼셉트론의 학습에 참여하지 않은 5명의 화자가 20개의 숫자음을 10번씩 발음한 총 1000개의 음성 데이터를 검사용 데이터로 사용하였다. 표 3에 나타난 결과로부터 LPCC 및 PLP 계수를 입력 특징으로 사용한 경우에는 화자 종속 인식 실험에 비해 화자간 변질성으로 인한 인식률 저하가 많이 나타날 뿐만 아니라 검사 데이터에 잡음이 많이 섞일수록 그 인식 성능 차이가 더 크게 나타남을 알 수 있다. 그러나 filterbank 계수 및 MFCC의 경우에 화자 종속인 경우와의 인식률 차이가 LPCC 및 PLP 계수에 비해 상대적으로 적게 나타남으로써 일반적인 환경에서의 인식시스템 구현 시 입력 특징으로 사용하기에 적당함을 알 수 있다.

<표 3> 다층 퍼셉트론을 이용한 화자 독립 인식 실험 결과

	Clean Data	50db	40db	30db
Filterbank	98.40 (16)	98.60 (14)	97.80 (22)	93.90 (61)
MFCC	98.30 (17)	97.90 (21)	96.9 (31)	88.50 (115)
LPCC	94.30 (57)	94.30(57)	85.20 (148)	52.40 (476)
PLP	92.80 (72)	92.00 (80)	88.40 (116)	70.90 (291)

5. 결 론

본 논문에서는 한국어 숫자음 인식을 위해 다층 퍼셉트론을 이용한 인식시스템을 구현하였으며 음성인식 분야에서 음성 특징으로 널리 사용되는 Mel-scale filterbank 계수, MFCC, LPCC, 그리고 PLP 계수를 인식시스템의 입력 특징으로 사용하여 그 인식 성능을 비교 분석하였다. 또한 본 논문에서는 여러 종류의 잡음이 존재하는 일반적인 환경에서 사용할 수 있는 인식시스템을 구현하기 위해 잡음이 섞인 음성 데이터에 대해 인식 실험을 각각 수행하였으며 부가적으로 학습에 참가하지 않은 화자의 음성데이터를 이용해 인식 시스템의 화자 독립성에 대한 실험을 수행하였다. 입력 특징별 인식 성능에서는 Mel-scale filterbank 계수를 입력 특징으로 사용한 경우에 잡음이 거의 없는 음성 데이터뿐만 아니라 잡음이 섞인 음성 데이터에서도 가장 높은 인식률을 나타내며 화자 독립적 음성 데이터에 대한 인식 실험에서도 가장 우수한 인식 성능을 얻어 일반적인 환경에서의 인식시스템 구현 시 가장 적합한 음성 특징으로 사용될 수 있음을 확인하였다.

또한 실험에 사용된 각 입력 특징들을 결합하여 통합 인식시스템을 구현할 경우에 좀더 개선된 인식 성능을 얻을 수 있을 것으로 판단된다. 따라서 각 입력 특징들을 효율적으로 결합할 수 있는 통합 기법에 대한 연구가 추후 연구 과제로서 계속 진행되어야 할 것이다.

참 고 문 헌

[1] L. Rabiner and S. Levinson, "Isolated and Connected Word Recognition-Theory and Selected Applications," IEEE Trans. Commun., vol. COM-29, May 1981.

[2] G.E. Kopec and M.A. Bush, "Network-based Isolated Digit Recognition Using Vector Quantization," IEEE Trans. ASSP, vol. ASSP-33, no. 4, pp. 850-867, Aug. 1985.

[3] L. Lamel and J.L. Gauvain, "Speaker Recognition with the Switchboard Corpus," Proc. IEEE Int'l Conf. ASSP, vol. 2, pp. 1067 - 1070, April 1997.

[4] L.R. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, pp. 4-16, Jan. 1986.

[5] J.F. Mari and J.P. Haton, "Automatic Word Recognition Based on Second-order Hidden Markov Models," IEEE Trans. SAP, vol. 5, no. 1, pp. 22-25, Jan. 1997.

[6] J.P. Hosom and R.A. Cole, "A Diphone-based Digit Recognition System Using Neural Networks," Proc. IEEE Int'l Conf. ASSP, vol. 4, pp. 3369-3372, April 1997.

[7] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," IEEE Trans. ASSP, vol. 37, no. 3, pp. 328-339, 1989.

[8] H. Bourlard and N. Morgan, "Continuous Speech Recognition by Connectionist Statistical Methods," IEEE Trans. NN, vol. 4, no. 6, pp. 893-909, Nov. 1993.

[9] C. Dugast, L. Devillers, and X. Aubert, "Combining TDNN and HMM in a Hybrid System for Improved Continuous-Speech Recognition," IEEE Trans. SAP, vol. 2, no.1, pp. 217-223, Jan. 1994.

[10] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning Internal Representations by Error Propagation," in Parallel Distributed Processing, Cambridge, MA: MIT Press, pp. 318-364, 1986.

[11] J.D. Markel and A.H. Gray, Linear Prediction of Speech, Springer Verlag, N.Y., 1976.

[12] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. ASSP, vol. ASSP-28, no. 4, pp. 357-366, Aug. 1980.

[13] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752, 1990.

[14] S.H. Oh, "Improving the Error Backpropagation Algorithm with a Modified Error Function," IEEE Trans. NN, vol. 8, pp. 799-803, 1997.

[15] S.E. Fahlman, "Faster-Learning Variations on Backpropagation: An Empirical Study," Proc. Connectionist Models Summer School, Carnegie Mellon University, pp. 38-51, 1988.

[16] G. Mirchandini and W. Cao, "On Hidden Nodes in Neural Nets," IEEE Trans. Circuits and Systems, vol. 36, no. 5, pp. 661-664,

김 인 철 (In-Cheol Kim)

1989년 2월 경북대학교 전자공학과 졸업(공학사)
1991년 2월 경북대학교 대학원 전자공학과 졸업(공학석사)
2001년 2월 경북대학교 대학원 전자공학과 졸업(공학박사)
1991년 - 1996년 (주) 카스 기술개발실 선임연구원
2001년 6월 - 현재 경북대학교 박사 후 연구원
관심분야 : 음성 인식, 컴퓨터비전, 신경회로망 등

김 대 영 (Dae-Young Kim)

1983년 2월 경북대학교 전자공학과 졸업(공학사)
1985년 2월 경북대학교 대학원 전자공학과 졸업(공학석사)
1992년 2월 경북대학교 대학원 전자공학과 졸업(공학박사)
1985년 1월~1986년 1월 금성전기(주) 연구원
1991년 9월~현재 계명문화대학 멀티미디어계열 교수
관심분야 : 컴퓨터비젼, 가상현실 등