神經網을 利用한 韓國語 數字音 認識에 관한 硏究

A Study on the Spoken Korean-Digit Recognition Using the Neural Network

朴 賢 和* 姜 海 東** 裵 建 星**

(Hyun Hwa Park,* Hae Dong Gahng,** Keun Sung Bae**)

요 약

한국어 숫자음이 단음절인 특성을 이용하여 각 숫자음에 대해 시간정합을 필요로 하지 않으면서 일정한 수를 갖는 특징벡터를 추출하여 다층구조 신경망으로 인식실험을 하였다. 음성신호의 시작점/끝점과 더불어 모음의 최대 피크점을 기준으로 해석구간을 초성, 중성, 종성의 세 부분으로 나누었으며, 음성신호의 특징벡터로는 반사계수, 켑스트럼, △켑스트럼, △에너지 등을 이용하여, 각 특징벡터 및 입력층과 은닉층의 노드 수에 따른 인식율 및 학습속도 등을 비교하였다. 신경망의 입력층의 특징벡터로서 반사계수를 사용한 경우보다 켑스트럼을 사용했을 때가 더 좋은 인식율을 보였다. △켑스트럼의 특성이 전체 인식율에 미치는 영향이 그다지 크지않았는데, 이는 한국어 숫자음이 단음절로 구성되어 있는 특징을 이용해 분석 구간을 stationary한 특성을 갖는 세 부분으로 구분하였기 때문이라 생각된다. 각 숫자음에 대해 150개의 켑스트럼을 사용한 경우에 97.8%의 인식율을 얻었다.

ABSTRACT

Taking advantage of the property that Korean digit is a mono-syllable word, we proposed a spoken Korean-digit recognition scheme using the multi-layer perceptron. The spoken Korean-digit is divided into three segments (initial sound, medial vowel, and final consonant) based on the voice starting/ending points and a peak point in the middle of vowel sound. The feature vectors such as cepstrum, reflection coefficients, △cepstrum and △energy are extracted from each segment. It has been shown that cepstrum, as an input vector to the neural network, gives higher recognition rate than reflection coefficients. Regression coefficients of cepstrum did not affect as much as we expected on the recognition rate. That is because, it is believed, we extracted features from the selected stationary segments of the input speech signal. With 150 ceptral coefficients obtained from each spoken digit, we achieved correct recognition rate of 97.8%.

※이 論文은 1990年度 敎育部 學術硏究助成費에 의하여 硏究되었음.

I. 서 론

음성은 인간의 가장 자연스러운 통신 방법으로서, 인간과 기계사이의 통신을 위해 컴퓨터를 이용한 음성신호의 분석 및 합성, 음성인식 등에 대한 연구가 꾸준히 진행되고 있다. 그 중에서도 음성인식의 문제는 인간이 기계를 사용하여 일을 수행하기 시작한 때부터 제기된 과제로서 컴퓨터 및 통신 기술의 발전에 힘입어 많은 연구가 있었고, 그 결과로 제한적인 분야에서는 음성인식 기술이 실용화 되어가고 있다고 그러나 여타분야의 비약적인 기술의 발전에 비해 음

* 한국전자통신연구소 TDX개발단
** 경북대학교 전자공학과
접수일자 : 1992. 2. 17.

성인식 기술의 발전은 만족스럽지 못하였으며, 이는 음성신호가 매우 다루기 어려운 stochastic process로서 음성의 요소가 유일하게 정해지지 않으므로, 이를 적절하게 모델링하는데 어려움이 많기 때문이다.

특히, 1980년대 중반부터는 벡터 양자화(vector quantization)와 HMM(hidden Markov model)을 이용한 음성인식에 대한 연구가 주종을 이루어 왔다. 그러나 기존의 방법은 인식대상어휘가 커질수록 또 말하는 사람이 일반적일수록 그 복잡도는 매우 커지고, 실시간의 음성인식이 어려워지는 문제점이 있다. 이러한 문제를 해결하기 위해서는 인식단위를 음소 또는 음절 단위로 해야 하는데 인식단위가 작아질수록 추출할 수 있는 정보량이 작아지므로 인식결과에 대한 신뢰도가 떨어진다. 그러므로 인식 기술은 음성 파형의 변화에 대하여 안정되고, 학습을 통하여 음성 인식의 규칙을 스스로 추출할 수 있어야 한다.

최근에는 인간의 두뇌는 대량의 복잡한 데이타를 병렬 처리할 수 있을 뿐만 아니라 학습능력이 있다는 사실에 근거하여 새로운 패턴인식 방법으로 제시된 인공신경망(artificial neural network)을 이용한 음성인식에 대한 연구가 활발히 진행되고 있다. 신경망을 패턴인식의 방법으로 이용하여 음성인식을 할 때 음성신호가 갖는 동적특성을 적절히 표현해야 하는데, 신경망은 정적패턴(static pattern)의 인식에는 우수한 성능을 보이지만 시간에 따라 변하는 동적패턴(dynamic pattern)의 인식에는 취약한 점이 있다. 이를 해결하기 위해 기존의 신경망을 변형시킨 TDNN(time delay neural network), INN(integrated neural network) 등에 대한 연구가 활발히 진행되고 있다[3,4].

본 연구에서는 기존의 다층구조 신경망(multilayer perceptron)으로 한국어 숫자음을 인식할 수 있는 시스템에 대해 연구하였다. 먼저 한국어 숫자음이 단음절로 구성된 성질을 이용하여 음성신호의 시작점/끝점과 더불어 모음의 최대 피크점을 기준으로 해석구간을 초성, 중성, 종성의 세부분으로 나누었다. 나누어진 각 구간에서 특징벡터로 반사계수(reflection coefficient), 켑스트럼(cepstrum), 스켑스트럼(Δcepstrum)등을 구하였으며, 각 특징벡터에 대한 신경망의 학습속도 및 인식율에 대해 실험하였다. 또한 은닉층의 노드 수가 신경망의 학습속도, 인식율 등에 미치는 영향에 대해서도 비교 검토하였다.

Ⅱ. 음성데이타 수집 및 특징벡터 추출

1. 음성데이타 수집

실험에 사용한 데이타는 한국어 숫자음 /영/에서 /구/까지와 /공/을 포함해서 모두 11개의 숫자음을 균등하게 포함하도록 임의로 작성한 전화번호를 20대 남성화자 5명이 20회 발음한 1100개의 음성으로서, 550개는 신경망을 학습시키기 위한 데이타로, 나머지 550개는 인식실험을 위한 데이타로 사용하였다.

그림 1은 음성데이타의 수집 과정을 보인 것이다. 먼저, 소음이 없는 상태의 일반 연구실에서 화자가 발성한 음성신호를 Sony ECM-220T 콘덴서 마이크를 통해 Inkel DD-2130C 녹음기에 dolby B type으로 녹음한다. 녹음된 신호는 다시 anti-aliasing을 위한 3.4 kHz 저역통과 여파기를 거친 후 8 kHz로 샘플링하고 12 bits/sample로 양자화하여 각 숫자음에 대해 화일명을 붙여 차후의 분석을 위해 컴퓨터의 디스크에 저장한다.

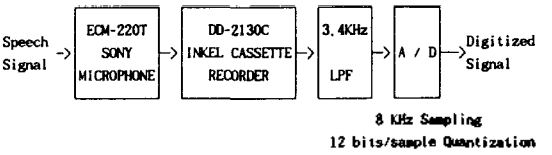


그림 1. 음성데이타 수집 과정
Fig. 1. The procedure of speech data collection.

2. 음성신호의 특징벡터 추출

음성인식 시스템에서는 샘플된 디지틀 신호에서 음성부분과 묵음부분을 분리하고, 시간정합을 필요로 하지 않으면서 일정한 수를 갖는 특징벡터를 추출하기 위해 단음절 음의 초성, 중성, 종성의 특성을 포함하는 세 부분으로 분리한 후, 그림 2와 같은 여러 단계의 음성신호 처리과정을 거친다. 선형예측계수(LPC : Linear Predictive Coefficient)로부터 얻어지는 여러가지 특징벡터 중에서 켑스트럼이 높은 인식율을 나타낸다는 사실에 근거하여 LPC 켑스트럼

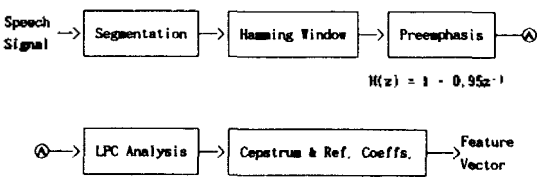


그림 2. 음성신호의 특징벡터 추출과정
Fig. 2. The procedure of feature vector extraction.

을 대표적인 특징벡터로 사용하였다[8]. 또한, 스펙트럼의 천이특성과 같은 파라미터의 동적특성을 잘 반영한다고 알려져 있는 △켑스트럼[9]과 파라미터 값이 −1에서 1까지로 제한된 특성을 갖는 반사계수도 특징벡터로 하여 실험하였다.

한국어 숫자음이 단음절로 구성된 점을 이용하여 해석구간을 음성신호의 시작점/끝점과 더불어 단음절 음의 가장 안정된 부분인 중성에 해당하는 모음의 최대 피크점을 기준으로 그림 3과 같이 초성, 중성, 종성의 특성을 포함하는 세 부분으로 나누었다.

음성신호의 시작점과 끝점을 찾기 위해, Rabiner와 Sambur가 제안한 방법[10]을 이용하였으며, 단음절 음에서 가장 안정된 부분인 중성에 해당하는 모음의 최대 피크점은, 초성이 파열음인 경우에 중성의 모음보다 진폭이 커질 수 있으므로, 초성부분의 최대 피크점이 모음의 최대 피크점으로 잘못 선택되지 않도록 시작점에서 30msec 지난 위치에서부터 신호의 최대 피크점을 찾았다. 그림 3과 같이 나누어진 각 구간에서 특징벡터가 추출되는 프레임 수를 5개, 3개, 1개로 하여 각 경우에 대해 실험을 하였다.

음성신호의 각 프레임의 크기는 200 샘플로서 Hamming창을 사용하였고, 한 프레임씩 중첩시켜 가면서 분석하였다. 0.95로 preemphasis한 후 Durbin 알고리즘을 이용한 선형예측계수로 부터 10차 켑스트럼[9]을 구하였으며, △켑스트럼은 식(1)과 같이 인접한 프레임 사이의 켑스트럼의 linear regression 값으로 표시된다. △에너지도 같은 방법으로 구해진다.

$$\triangle_i X(m) = \frac{\sum_{n=-M}^{M} X_m(n)n}{\sum_{n=-M}^{M} n^2} \qquad (1)$$

여기서 M은 (2m+1)로서 regression 계수를 계산하기 위해 사용된 프레임수를 나타내고 n은 인접한 프레임의 순서, m은 파라미터 X의 차수를 나타낸다.

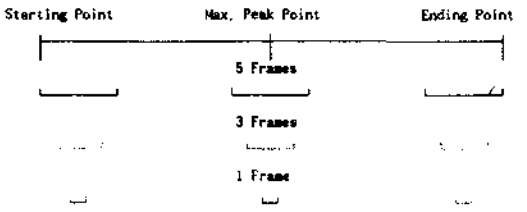


그림 3. 특징벡터 추출을 위한 숫자음의 해석구간
Fig. 3. The analysis frames for feature vector extraction.

Ⅲ. 실험 및 고찰

1. 신경망의 학습 및 인식실험

본 연구에서는 그림 4에 주어진 것과 같이 입력층과 출력층 사이에 1개의 은닉층을 갖는 다층구조 신경망을 사용하였다. 입력층의 노드 수는 각 숫자음에서 추출되는 특징벡터의 수에 의해 결정되며 출력층의 노드 수는 인식하고자 하는 숫자음의 갯수로서 /영/에서 /구/까지와 /공/을 포함한 11개가 된다. 은닉층의 노드 수는 16에서 32까지 4씩 증가시켜 가면서 실험하였다. 원하는 출력값은 각 숫자음에 해당하는 출력층 노드에는 0.9, 그 외의 노드에는 0.1로 하였다. 신경망의 학습을 위해서는 Rumelhart 등이 제안한 역전파 학습알고리즘[11]을 이용하였다.

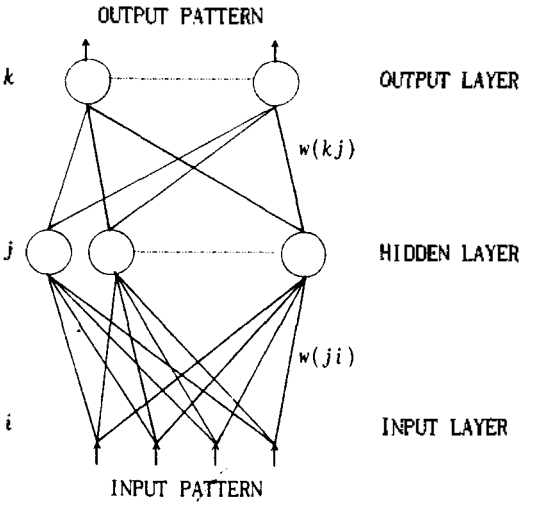


그림 4. 다층구조 신경망
Fig. 4. Multi Layer Perceptron.

신경망을 학습시킬 때의 여러 변수값들은 실험에 의해 학습율(learning rate) η는 0.25, 관성항(momentum) α는 0.7, 그리고 각 패턴에 대한 오차의 한계값은 0.0001, 전체 학습오차의 한계값은 0.025로 하였다. 입력층의 노드 수가 156, 은닉층의 노드 수가 28이며, 특징벡터는 5 프레임에 대한 켑스트럼을 사용하였을 때 신경망의 학습에 따른 전체 학습오차의 변화를 보인 것이다. 학습횟수가 증가함에 따라 전체 학습오차가 처음에는 급격히 감소하

다가 점차 전체 오차의 설정된 값에 수렴해 가는 것을 볼 수 있는데 이는 실험에 의해 정해진 여러 변수의 값이 적절히 선정되었음을 나타낸다.

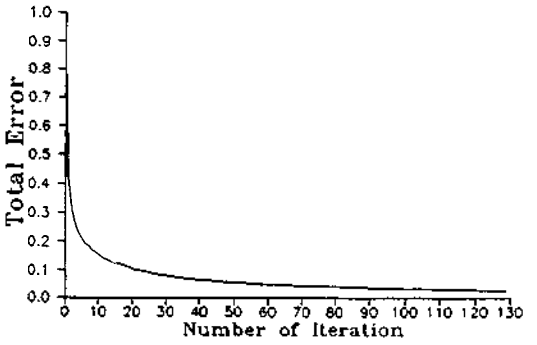


그림 5. 학습횟수에 대한 전체 학습오차

Fig. 5. The total error versus the number of iteration
No. of nodes in the input layer : 150(Cepstrum in 5 frames)
No. of nodes in the hidden layer : 16

음성신호가 그림 3과 같이 세 부분으로 나누어진 각 구간에 대해 다음의 두 경우와 같이 특징벡터를 구하고, 각 경우에 대해 입력층의 노드 수를 바꾸어 가면서 인식율 및 학습속도에 대한 실험을 하였다.

(1) 특징벡터로 켑스트럼 또는 반사계수를 이용한 경우

Case Ⅰ. 각 구간에서 5 프레임에 대한 특징벡터를 사용한다.(N=150)

Case Ⅱ. 각 구간에서 3 프레임에 대한 특징벡터를 사용한다.(N=90)

Case Ⅲ. 각 구간에서 1 프레임에 대한 특징벡터를 사용한다.(N=30)

(2) 특징벡터로 켑스트럼과 △켑스트럼 등을 이용한 경우

Case Ⅰ. 각 구간에서 5 프레임에 대한 켑스트럼을 사용한다.(N=150)

Case Ⅱ. 각 구간에서 1 프레임의 켑스트럼, $\triangle_5$켑스트럼, $\triangle_5$에너지, $\triangle_3$켑스트럼, $\triangle_3$에너지를 사용한다.(N=96)

Case Ⅲ. 각 구간에서 1 프레임의 켑스트럼, $\triangle_5$켑스트럼, $\triangle_5$에너지를 사용한다.(N=63)

Case Ⅳ. 각 구간에서 1 프레임의 켑스트럼, $\triangle_5$켑스트럼을 사용한다.(N=60)

Case Ⅴ. 각 구간에서 1 프레임의 켑스트럼, $\triangle_5$에너지를 사용한다.(N=33)

Case Ⅵ. 각 구간에서 1 프레임의 켑스트럼을 사용한다.(N=30)

여기서 N은 입력층의 노드 수를 나타낸다. 반사계수와 regression 계수 값은 −1에서 1로 제한되어 있는데 비해 켑스트럼은 그 범위가 상대적으로 크기 때문에 모든 켑스트럼의 값을 1/2로 줄여서 사용하였다.

은닉층의 노드 수를 16에서 32까지 4씩 증가시켜 가면서 위의 각 경우에 대해 신경망을 학습시키고 인식실험을 통해 그 결과를 비교 검토하였으며, 위의 각 경우에 대한 인식실험 결과와 더불어 화자 및 각 숫자음에 따른 오인식을 분석하였다.

2. 실험결과 및 고찰

(1) 특징벡터로 켑스트럼 또는 반사계수을 사용한 경우

표 1은 특징벡터로 켑스트럼과 반사계수를 사용한 각각의 경우에 대해 은닉층의 노드 수에 따른 전체 학습오차가 설정된 한계값까지 수렴하는데 소요된 학습횟수 및 인식율의 변화를 나타낸 것이다. 표 1(a)에 의하면, 입력층의 노드 수가 증가할수록 전반적으로 신경망의 학습속도가 빨라짐을 볼 수 있는 반면, 은닉층의 노드 수 증가는 반드시 학습속도의 개선으로 이어지지 않음을 볼 수 있다. 입력층의 특징벡터로 각 분석 구간에 대해 1프레임의 반사계수만을 사용했을 경우에는 신경망을 학습시킬 때에 전체 학습오차가 설정된 값까지로 수렴되는 결과를 얻을 수 없었다. 표 1(b)에서 입력층의 특징벡터로서 반사계수를 사용한 경우보다 켑스트럼을 사용했을 때가 더 좋은 인식율을 나타냄을 볼 수 있다. 입력층의 노드 수를 150에서 90으로 줄였을 경우에도 켑스트럼 및 반사계수 모두 인식율에서 별 변화가 없었다. 하지만 입력층의 노드 수를 30으로 하여 켑스트럼을 사용하였을 경우에는 평균 인식율이 94.55%로 입력층 노드 수가 150인 경우에 비하여 3%정도 감소되었다. 은닉층의 노드 수 변화가 학습속도 및 인식율에 미치는 영향은 그리 크지 않은 것으로 나타났다. 그림 6은 표 1을 그래프로 나타낸 것이다.

(2) 특징벡터로 켑스트럼, △켑스트럼 등을 사용한 경우

특징벡터로 켑스트럼, △켑스트럼, △에너지를 사

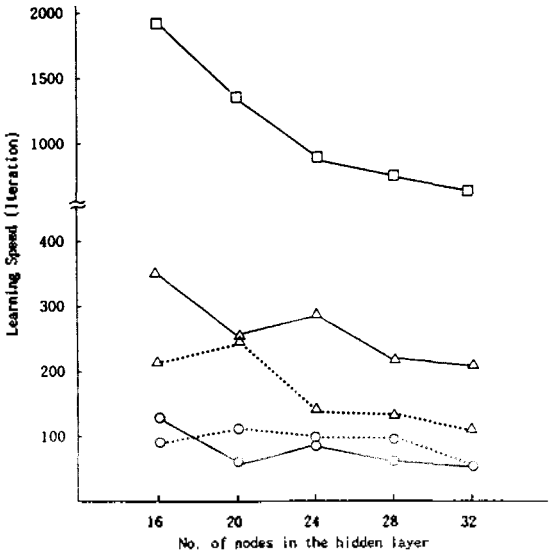
표 1. 켑스트럼 또는 반사계수를 사용한 경우의 실험 결과
Table 1. The results using cepstrum or reflection coefficients.

(a)Learning speed for the number of nodes in the hidden layer(No. of iteration)

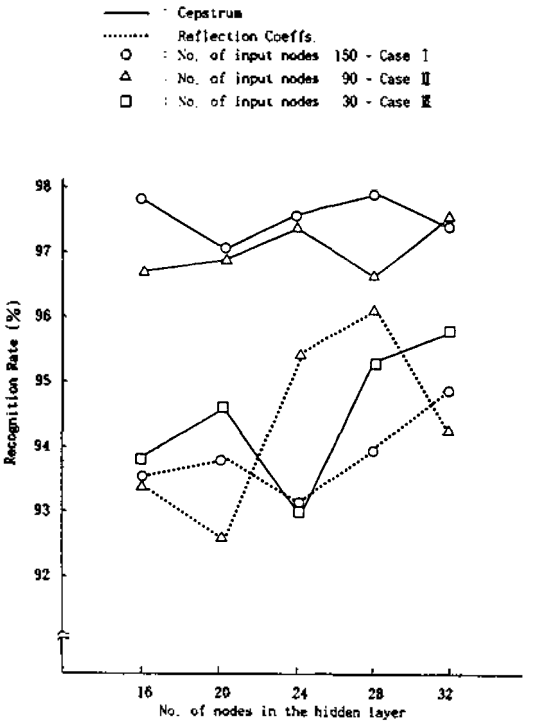
No. of nodes \ Feature	CEPSTRUM			REFLECTION COEFFICIENT		
	I	II	III	I	II	III
16	130	356	1879	92	214	·
20	66	244	1355	103	248	·
24	82	287	875	90	145	·
28	74	226	739	94	137	·
32	66	212	634	66	107	·

(b)Recognition rate for the number of nodes in the hidden layer(%)

No. of nodes \ Feature	CEPSTRUM			REFLECTION COEFFICIENT		
	I	II	III	I	II	III
16	97.82	96.73	93.82	93.64	93.45	·
20	97.09	96.91	94.73	93.82	92.73	·
24	97.64	97.45	93.10	93.27	95.45	·
28	97.82	96.73	95.29	94.00	96.18	·
32	97.45	97.64	95.82	94.91	94.36	·
AVG.	97.56	97.09	94.55	93.93	94.43	·



(a)Learning speed for the number of nodes in the hidden layer(No.of iteration)



(b)Recognition rate for the number of nodes in the hidden layer(%)

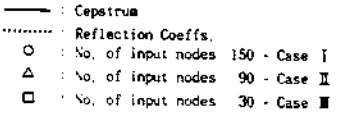


그림 6. 은닉층의 노드 수에 따른 학습속도 및 인식율
Fig. 6. Learning speed and recognition rate for the number of nodes in the hidden layer.

용하여 입력층과 은닉층의 노드 수를 변화시켜 가면서 인식실험을 한 결과가 표 2에 주어져 있다. 표 1에서와 마찬가지로 입력층의 노드 수가 많을수록 전반적으로 학습과정에서의 수렴속도는 빨라짐을 볼 수 있다. 표 2(b)에서 특징벡터를 추출할 때 각 분석 구간에서 5 프레임의 켑스트럼을 사용하는 경우에는 입력 층의 노드 수가 너무 많아지고 1 프레임의 켑스트럼을 사용하는 경우에는 인식율이 많이 저하되는 단점이 있다. 입력층의 노드 수를 줄이기 위하여 켑스트럼과 △켑스트럼을 사용한 III, IV의 경우 5 프레임의 켑스트럼을 사용한 경우보다 학습회수는 많이 증가하지만 입력층의 노드 수는 반이상 줄고 인식율도 1% 정도 감소하여 크게 변하지 않음을 알 수 있다. 또한 표 2(b)에서 II와 (III, IV)를 비교하면 △켑스트럼의 특성이 전체 인식율에 미치는 영향이 크다지

크지 않음을 볼 수 있는데, 이는 한국어 숫자음이 단음절로 구성되어 있는 특징을 이용해 분석 구간을 stationary한 특성을 갖는 세부분으로 구분하였기 때문이라 생각된다. 표 2(b)의 Ⅲ와 Ⅳ, Ⅴ와 Ⅵ을 비교하면 입력층의 특징벡터로 △에너지를 추가하였을 경우 학습속도및 인식율 모두가 △에너지를 사용하지 않은 경우보다 미미하나마 나빠짐을 볼 수 있다. 이것은 음성인식에서 일반적으로 에너지 정보를 적절한 방법으로 이용할 수 없는 경우에는 제외시키는 것이 나을 수 있다는 것을 의미하는데 이는 [11]에서 얻어진 비슷한 결과와도 일치한다.

표 2. 켑스트럼과 △켑스트럼 등을 사용한 경우의 실험결과
Table 2. The results using cepstrum and △cepstrum, etc.

(a) Learning speed for the number of nodes in the hidden layer (No. of iteration)

No. of nodes \ Feature	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ
16	130	933	1062	996	1662	1879
20	66	440	764	588	675	1355
24	82	341	682	544	721	875
28	74	391	638	532	1007	739
32	66	373	400	417	719	634

(b) Recognition rate for the number of nodes in the hidden layer(%)

No. of nodes \ Feature	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ
16	97.82	95.45	95.82	96.18	93.45	93.82
20	97.09	96.00	96.18	96.00	94.00	94.73
24	97.64	96.73	96.36	97.45	95.09	93.10
28	97.82	96.73	96.36	97.27	95.45	95.29
32	97.45	96.73	96.73	96.36	94.91	95.82
AVG.	97.56	96.33	96.30	96.65	94.58	94.55

3. 화자 및 인식 숫자음에 따른 오인식율 분석

표 3은 본 논문에서 행한 인식실험의 모든 경우에 대해 각 화자 및 숫자음 별로 오인식이 발생한 회수를 누적하여 구한 것이다. 입력층의 특징 벡터를 다르게 인가한 경우가 9가지, 은닉층의 노드 수 변화에 따른 인식실험이 5가지로 전체적으로 45회의 인식실험에서 발생한 오인식 회수를 합산한 것이다. 1회의 인식실험에서 550개의 숫자음이 사용되므로 전체실험에 사용된 숫자음은 550*9*5=24750개가 된다.

화자별로는 화자 B및 D가 상대적으로 더 많은 오인식을 하였음을 볼 수 있고, 각 숫자음에 대해서는 /구/에 대한 오인식이 제일 많고, 그 다음이 /오/와 /공/에 대한 오인식이 많음을 볼 수 있다. 특히 /구/에 대한 오인식은 대부분 /공/으로 인식되었으며, /오/에 대한 오인식은 많은 경우가 /구/로 인식되었다. 숫자음 /공/과 /구/의 오인식은 시작부분인 초성이 무성폐쇄음 /ㄱ/으로 같은데다 지속시간이 매우 짧으며 단음절에서 안정된 부분인 모음이 /오/와 /우/로서 첫째 포만트를 제외하고는 스펙트럼 특성이 비슷하며, 끝 부분에서의 신호 크기가 매우 작아 종성에서의 차이점을 제대로 검출할 수 없기 때문이라고 생각된다. /구/와 /오/의 오인식도 앞에서와 비슷한 이유에서 비롯된다고 생각된다. 표 3에 의하면 그 외의 대부분의 오인식은 공통 음소를 포함하는 숫자음의 쌍, 즉 /일/과 /칠/, /삼/과 /사/, /일/과 /이/ 등에서 발생함을 볼 수 있다.

표 3. 화자 및 숫자음에 대한 오인식 분포
Table 3. The error distribution for each speaker and digit

(a) The number of errors for each speaker.

Speaker \ Test Digit	0	1	2	3	4	5	6	7	8	9	공	Total Errors
A	1	3	4	8	30	16	1	1	5	82	3	154
B	1	2	34	23	5	27	21	0	11	200	26	350
C	17	51	2	0	3	47	18	2	11	10	32	193
D	1	4	0	6	1	7	9	16	1	135	20	200
E	6	8	0	14	0	5	15	0	6	94	19	164

(b) The number of errors for each digit.

Recognized Digit \ Test Digit	0	1	2	3	4	5	6	7	8	9	공	Total Errors
0	0	27	17	14	0	0	41	0	2	20	3	121
1	5	0	15	0	0	0	5	14	4	0	0	43
2	6	34	0	0	0	0	0	0	0	1	0	41
3	0	0	0	0	35	0	6	1	0	5	0	47
4	0	0	0	10	0	0	0	1	1	0	0	12
5	0	0	0	0	0	0	0	0	0	52	27	79
6	13	0	0	0	0	0	0	0	1	0	0	14
7	0	7	0	0	0	0	0	0	8	0	0	15
8	0	0	0	2	4	0	0	3	0	1	2	12
9	0	0	0	8	0	63	3	0	1	0	65	140
공	2	0	8	17	0	39	9	0	17	442	0	534
Total Errors	26	68	40	51	39	102	64	19	34	521	97	1061

표 4는 인식실험에서 97.8%의 가장 높은 인식율을 얻은 경우에 대한 (입력 노드 수 150:5 프레임에 대한 켑스트럼, 은닉층 노드 수 16)오인식 분포를 보인 것이다. 화자별로는 표 3에서와 마찬가지로 화자 B 및 D가 전체 오인식의 90% 이상을 차지하고 있다. 숫자음별로는 거의 대부분의 오인식이 숫자음 /구/에서 생겼으며, 오인식된 /구/의 대부분이 /공/으로 인식되었다. 표 3과 표 4에서 보는 바와 같이 오인식의 대부분을 차지하는 /오/와 /구/가 /공/으로 인식되므로 수집된 음성데이타 중에서 /공/을 제외한 1000개의 데이타로서 학습과 인식실험을 하였다. 이 중 500개의 숫자음으로 신경망을 학습시키고 나머지 500개의 숫자음으로 인식실험을 한 결과 99.2%의 높은 인식율을 얻었으며, 표 5에 주어진 값은 그 때의 오인식 분포를 보인 것이다. 표 5에서 /공/으로 오인식된 /구/가 제대로 인식되는 것을 볼 수 있다.

표 4. 화자 및 숫자음에 대한 오인식 분포

Table 4. The error distribution for each speaker and digit.
No. of nodes in the input layer : 150 (Cepstrum in 5 frames)
No. of nodes in the hiddent layer : 16

(a) The number of errors for each speaker.

		Test Digit											Total
		0	1	2	3	4	5	6	7	8	9	공	Errors
Speaker	A	0	0	0	0	0	0	0	0	0	0	0	0
	B	0	0	0	0	0	0	1	0	0	4	1	6
	C	0	0	0	0	0	0	0	0	0	0	0	0
	D	0	0	0	0	0	0	0	0	0	5	0	5
	E	0	0	0	0	0	0	0	0	0	1	0	1

(b) The number of errors for each digit.

		Test Digit											Total
		0	1	2	3	4	5	6	7	8	9	공	Errors
Recognized Digit	0	0	0	0	0	0	0	1	0	0	0	0	1
	1	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	1	0	1
	6	0	0	0	0	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	0	0	0	1	1
	공	0	0	0	0	0	0	0	0	0	9	0	9
Total Errors		0	0	0	0	0	0	1	0	0	10	1	12

표 5. 화자 및 숫자음에 대한 오인식 분포(/공/을 제외한 데이타)

Table 5. The error distribution for each speaker and digit. (excluded 'gong')
No. of nodes in the input layer : 150 (Cepstrum in 5 frames)
No. of nodes in the hiddent layer : 16

(a) The number of errors for each speaker.

		Test Digit										Total
		0	1	2	3	4	5	6	7	8	9	Errors
Speaker	A	0	0	0	0	0	0	0	0	0	0	0
	B	0	0	1	0	0	0	1	0	0	1	3
	C	0	0	0	0	0	0	0	0	0	0	0
	D	0	0	0	0	0	1	0	0	0	0	1
	E	0	0	0	0	0	0	0	0	0	0	0

(b) The number of errors for each digit.

		Test Digit										Total
		0	1	2	3	4	5	6	7	8	9	Errors
Recognized Digit	0	0	0	1	0	0	0	1	0	0	0	2
	1	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	1	1
	4	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	1	0	0	0	0	1
Total Errors		0	0	1	0	0	1	1	0	0	1	4

Ⅳ. 결 론

한국어 숫자음이 난음절인 특성을 이용하여 각 숫자음에 대해 시간정합을 필요로 하지 않으면서 일정한 수를 갖는 특징벡터를 추출하여 다층구조 신경망으로 인식실험을 하였다. 음성신호의 특징벡터로는 반사계수, 켑스트럼, 스켑스트럼, 스에너지 등을 이용하여, 각 특징벡터 및 입력층과 은닉층의 노드 수에 따른 인식율 및 학습속도 등을 비교하였다. 신경망의 입력층의 특징벡터로서 반사계수를 사용한 경우보다 켑스트럼을 사용했을 때가 더 좋은 인식율을

보였으며, 입력벡터로 켑스트럼만을 5 프레임 사용한 경우에 제일 높은 인식율을 얻었다. 켑스트럼과 △켑스트럼을 함께 입력벡터로 사용하는 경우에는 켑스트럼만을 5 프레임 사용하는 경우보다 학습횟수는 많이 증가하지만 인식율을 크게 감소시키지 않으면서 입력층의 노드수를 반이상 줄일 수 있었다. △켑스트럼의 특성이 전체 인식율에 미치는 영향이 그다지 크지 않았는데, 이는 한국어 숫자음이 단음절로 구성되어 있는 특징을 이용해 분석 구간을 stationary한 특성을 갖는 세 부분으로 구분하였기 때문이라 생각된다. /공/을 포함한 11개의 숫자음에 대해 각 구간의 5 프레임에서 추출된 입력벡터로 150개의 켑스트럼을 사용한 경우에 97.8%의 인식율을 얻었으며, /공/을 제외한 /영/에서 /구/까지의 열개의 숫자음에 대해서는 99.2%의 높은 인식율을 얻을 수 있었다. 따라서 세 부분으로 나누어진 각 분석 구간의 특성을 더 잘 나타낼 수 있는 방법에 대한 연구와 더불어 신경망에 입력되는 특징벡터의 수를 줄이면서, 인식율을 향상시키기 위한 연구가 필요하다고 본다.

참 고 문 헌

1. Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol.77, No.2, Feb. 1989.
2. L. R. Rabiner and S. E. Levinson, "Isolated and connected word recognition-theory and selected applications," *IEEE Trans. Comm*, vol.29, No.5, May 1981.
3. K. J. Lang, A. H. Waibel, and G. E. Hinton, "A time delay neural network architecture for isolated word recognition," *Neural Network*, vol.3, pp.23-43, 1990.
4. T. Matsuoka, H. Hamada, and R. Nakatsu, "Syllable recognition using integrated neural networks," *Proc. of the International Joint Conference on Neural Networks*, vol. SP87-101, Dec. 1989.
5. L. R. Rabiner, K. C. Pan., F. K. Soong, "On the performance of isolated word speech recognizers using vector quantization and temporal energy contours," *AT&T Bell Lab. Tech. J.* vol. 63, pp. 1245-1260, Sep. 1984.
6. Yoh'ichi Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Trans. ASSP.*, vol.35, No.10, Oct. 1987.
7. L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol.54, No.2, pp.297-315, Feb. 1975.
8. Sadaoki Furui, "On the use of hierarchical spectral dynamics in speech recognition," *Proceedings Int. Conf. ASSP.*, pp.789-792, 1990.
9. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, vol.1, D. E. Rumelhart and J. L. McClelland(Eds.), Cambridge, MA : MIT Press, pp.318-362, 1986.
10. 박현화, 강해동, 배건성, "多層構造 神經網을 利用한 韓國語 數字音 認識," 한일합동 음향학회 학술대회 논문집, pp.340-344, July 1991.
11. 이석우 "狀態遷移를 二項分布로 모델링한 HMM을 이용한 音聲認識," 경북대학교 석사학위논문 1991.

▲朴 賢 和(Hyun Hwa Park) 1967년 12월 5일생
1986년 3월～1990년 2월 : 경북대학교 전자공학과(공학사)
1990년 3월～1992년 2월 : 경북대학교 전자공학과(공학 석사)
1992년 2월～현재 : 한국전자통신연구소 TDX 개발단 연구원
주관심분야 : 음성신호처리, 디지틀 이동통신 등

▲姜 海 東(Hae Dong Gahng) 1961년 10월 9일생
1980년 3월～1987년 2월 : 경북대학교 전자공학과(공학사)
1987년 3월～1989년 2월 : 경북대학교 전자공학과(공학석사)
1989년 3월～현재 : 경북대학교 전자공학과(박사과정)
주관심분야 : 음성신호처리, 적응신호처리 등

▲裵 建 普(Keun Sung Bae) 1953년 11월 9일생
1973년 3월～1977년 2월 : 서울대학교 전자공학과(BS)
1977년 3월～1979년 2월 : 한국과학원 전기 및 전자공학과(MS)
1984년 8월～1989년 5월 : 미국 Univ. of Florida (Ph.D)
1979년 3월～현재 : 경북대학교 전자공학과 부교수
주관심분야 : 디지틀 신호처리, 음성신호처리, 디지틀통신 등