論文 92-29B-10-11

$\nabla^2 G$ 연산자의 신호 분석 특성을 이용한 음성 인식 신경 회로망에 관한 연구

(Neural Network for Speech Recognition Using Signal Analysis Characteristics by $\nabla^2 G$ Operator)

李鍾赫,*** 鄭容近,* 南基坤,** 尹台焄,** 金在昌,** 朴義烈,** 李梁成**

(Jong Hyeok Lee, Young Guen Jung, Ki Gon Nam, Tae Hoon Yoon, Jae Chang Kim, Ei Yul Park, and Yang Sung Lee)

要 約

본 논문에서는 음성 인식을 위한 신경회로망 모델을 제안한다. 제안한 모델은 특징 추출부와 인식부로 구성된다. $\nabla^2 G$ 연산자에 기초한 Interconnection 모델이 주파수 분석에 사용되었다. 두가지 특징들, 즉, 개괄 특징과 국부 특징이 이 모델에 의해 추출되었다. 인식부는 개괄 분류단과 국부 분류단으로 구성되었다. 입력패턴을 경사치 방법(Slope method)에 의해 부호화했을 때 화자 A와 B의 인식율이 각각 100%이었다. 9명의 화자를 대상으로 인식 실험을 수행했을 때의 인식율은 91.4%이었다.

Abstract

In this paper, we propose a neural network model for speech recognition. The model consists of feature extraction parts and recognition parts. The interconnection model based on $\nabla^2 G$ operator was used for frequency analysis. Two features, global feature and local feature, were extracted from this model. Recognition parts consist of global grouping stage and local grouping stage. When the input pattern was coded by slope method, the recognition rate of speakers, A and B, was 100%. When the test was performed with the data of 9 speakers, the recognition rate of 91.4% was obtained.

*準會員, **正會員, 釜山大學校 電子工學科
(Dept. of Elec. Eng., Pusan Nat'l Univ.)
***正會員, 慶星大學校 컴퓨터工學科
(Dept. of Com. Eng., Kyungsung Univ.)

接受日字 : 1992年 1月 24日
(※본 연구는 1991년도 한국학술진흥재단 학술연구 조성비(자유 공모 과제)에 의하여 연구 되었음.)

I. 서 론

의사 표현의 수단 중에 가장 보편적인 것은 말과 글이다. 글은 기록이 남지만 속도가 느리며, 이와 반대로 말은 속도가 빠르지만 기록을 남기지 못한다. 기록을 남기지 못하는 말의 단점을 해결하여 말을 기계와 인간 사이의 정보전달 매개 수단으로 사용하고자 하는 연구는 음성에서 추출된 특징 파라미터를 이용하여 수행할 수 있다.

특징 파라미터를 추출하는 방법으로는 음성의 발성 과정을 모델링하여 추출하는 방법,[1] 음성 신호 자체를 주파수 영역이나 시간 영역에서 해석하여 추출하는 방법,[2],[3] 인지 과정을 모델링하여 추출하는 방법으로 귀는 스펙트럼 분석기라는 이론을 도입하여 필터 뱅크(filter bank)의 출력을 이용하는 방법을 들 수 있다.[4]

사람과 비슷한 수준의 음성 인식 시스템을 구현하는데는 세 가지 문제점이 있다. 첫째, 음성 자체의 다양성과 둘째, 실용화를 위해서는 인식 대상의 단어가 많아야 한다는 것과 셋째, 조음결합(coarticulation)으로 인하여 같은 글자라 할지라도 발음이 현저히 달라져 버리는 점이다. 이와 같은 문제를 해결하려면 인식 단위를 음소 혹은 음절 단위로 해야 하며 인식 단위가 작아지므로 발생하는 음성 인식의 규칙을 스스로 추출할 수 있는 방법이 마련되어야 할 것이다.

최근에 연구가 활발한 신경 회로망(neural network)은 자율 조직(self organization) 특성, 학습(Learning) 특성, 그리고 적응(adaptation) 특성 등이 밝혀지고 있고, 이들 특성을 연속 음성 및 화자 독립의 경우에 적용하려는 연구가 활발히 진행되고 있다.[5]-[9] 국내에서의 신경 회로망을 이용한 음성 인식 방법으로는 흡필드 신경회로망, 다층 신경 회로망, TDNN, 자율 조직 신경 회로망을 이용한 경우가 있지만[10]-[12] 기존의 신경 회로망을 한국어의 특징에 맞도록 적용시키고 있다. 일반적으로 신경 회로망을 음성 인식에 이용하는 경우 인식부만 신경 회로망으로 구성하고 특징 파라미터 추출부는 FFT 방법이나 필터 뱅크를 이용하여 처리하고 있다. 이 경우 특징 파라미터 추출부를 신경 회로망으로 처리하지 않기 때문에 음성 인식 기구는 사람의 음성 인식 기구와 달라지게 된다.

오래전부터 신경 생리학(Neurophysiology)과 정신물리학(Psychophysics)적 관점에서 특징 추출부인 감각 기관을 취급한 연구들이 행해졌다. 이 중 신경 세포의 배열 상태와 비슷한 구조로서 중심부의 자극 영역(excitatory region)과 주변부의 억제 영역(inhibitory region)으로 나누어져 있는 $\nabla^2 G$ 연산자는 최근까지도 영상에서의 경계점 검출에 널리 사용되고 있다.[13] 반면 인간의 음성 인식 기구는 청각 기관의 하위 뉴런에서 개괄 특징(global feature)을 추출하여 신경 섬유(nerve fiber)를 통해 상위 뉴런으로 전달하고, 추출된 개괄 특징이 신경 섬유를 통해 청각 피질(auditory cortex)로 전달되면서 정보 처리가 행해져 국부 특징(local feature)이 추출된다고 알려져 있으나 어떠한 정보 처리가 행해지는지는 분명하지 않다. 그러나 장소설(place theory)[14]에 의하면 입력 음성의 주파수에 따라 달팽이관(cochlea)에서 음위상도(tonotopic map)가 형성된다고 하며 이를 기반으로 한 음성 인식 연구들이 최근 많이 발표되고 있다.[15]-[17] 그러나 음성 신호에서 특징 파라미터의 추출과정을 신경 회로망으로 모델화하지는 않았다.

본 논문에서는, 공간적으로는 신경 세포의 배열 상태와 비슷한 구조를 하고 주파수 영역에서는 대역 통과 특성을 가지는 $\nabla^2 G$ 연산자를 이용하여 구성한 신경 회로망 형태의 Interconnection 모델[18]에서 음성 신호의 개략적인 주파수 특성을 나타내는 개괄 특징과 보다 자세한 주파수 특성을 나타내는 국부 특징을 추출한다. 또한 Interconnection 모델에서 추출되는 개괄 특징과 국부 특징을 특징 파라미터로 이용하여 사람의 음성을 인식하고자 하며, 특징 추출부에서 인식부까지를 신경 회로망 형태로 구성한 INNA (Integrated Neural Network with Attention) 모델을 제안하고, 제안된 INNA 모델이 모음을 인식할 수 있음을 보이고자 한다.

Ⅱ. 음성 인식

음성은 모음과 자음으로 분류되며 필스음이 조음기관을 지날 때 큰 장애를 받지 않고 입 밖으로 나오는 소리가 모음이며 표 1과 같이 단모음과 복모음으로 구분된다.

표 1. 한국어 모음
Table 1. Korean vowels.

monophthong	아 어 오 우 으 이 애 에 외 위
Diphthong	야 여 여 유 애 예 야 와 왜 워 웨 의

단모음의 분류법중 혀의 최고점으로 단모음을 분류하는 방법은 복잡한 조음 관계를 간략하면서도 효과적으로 형식화한 것이다. 한국어 단모음 /아·어·오·우·으·이·애·에/를 발음할 때, 혀의 최고점은 발성 기관의 측면에서 촬영한 X선 실험에 따르면 그림 1과 같다.[19] 각 단모음에서 혀의 최고점은 각각 다르지만 Y축을 기준으로 대별하여 보면 /아/는 가장 밑에 있으며 /이·으·우/는 윗쪽에 있고 /어·오·애/는 중간에 있음을 알 수 있다.

콜티(corti) 기관은 청각 기관으로서 기저막 위에 놓여 있으며 기계적인 자극에 민감하게 반응하는 유모 세포(hair cell)와 지지 세포로 구성되어 있다. 기저막은 특별한 주파수에 잘 진동하며 이런 특성은 콜티 기관의 유모세포에 연결된 신경 섬유를 통해 생성 전류를 보내 청각 신경(auditory nerve)에 구심성 신경 흥분을 전달한다. 청각 신경 흥분이 청각 피질로 전달되는 경로에 대해서는 잘 알려져 있다. 그러나 청각 신경 흥분이 청각 피

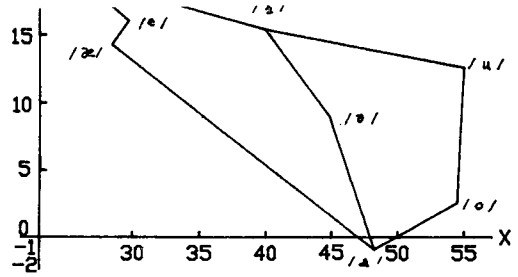


그림 1. 혀의 최고점과 기본 모음의 관계
Fig. 1. Relation of fundamental vowels and the mixiumum-point positions of tongue.

질로 전달되는 과정에서 일어나는 생리학적 과정에 관해서는 아직 정설이 없으며 장소설[14]과 빈도설(frequency theory)[20]의 두 가지 가설이 있다.

청각 신경계에서 선택도는 하위 뉴런에서 낮고 상위 뉴런으로 올라갈수록 높아진다.[21] 이것은 대역 통과 특성을 가지며 주파수 1~2KHz범위에서 선택도가 2정도로 알려져 있는 신경 섬유[22]에서 완전히 분석된 주파수 정보를 청각 피질로 전달하여 주는 것이 아니라 하위 뉴런은 개괄 특징을 추출하여 신경 섬유로 전달하고 개괄 특징이 신경 섬유에서 청각 피질로 전달되는 과정에서 정보 처리가 수행되어 국부 특징으로 변환됨을 뜻한다.

발음 지속 시간이 음성의 인식율에 미치는 영향을 연구한 결과[23] 사람의 경우 모음은 56%(평균 지속시간 117ms), 자음과 모음이 결합된 CV형태의 음성은 63%(평균 지속시간 117ms), 단음절은 81%(평균 지속시간 498ms), 문장은 98%(평균 지속시간 2.4sec)를 인식한다고 발표하였으며, 화자 인식을 위하여 특징 파라미터를 2구간 이상 평균한 값으로 했을 때 인식율이 0.1초 동안은 68%에서 80%로 향상되며 약 0.3초 이상 되었을 경우는 별 차이를 발견하지 못하였다고 발표하였다.[24] 이는 사람이 음성을 인식할 때 짧은 순간 동안 입력되는 소리로서 각각을 구별한 후 그것을 조합하여 전체적인 말을 이해하는 것이 아니라 전후 구간에서 들려오는 소리를 참고하여 말을 인식한다는 것을 나타낸다.

Ⅲ. ∇^2G연산자에 의한 주파수 분석 신경 회로망

감각 세포는 입력 신호에서 개괄 특징을 추출하여 신경 섬유로 전달하고 추출된 개괄 특징이 신경 섬유에서 청각 피질로 전달되는 과정에서 정보 처리가 수행되어 개괄 특징보다 특징 정보가 좀 더 세분화된 국부 특징으로 변환된다. 이와 같은 과정을 모델화하기 위해서는 신경 세포의 배열 상태와 비슷한 구조를 하고 있으며 개괄 특징과 국부 특징을 동시에 추출할 수 있는 모델이 필요하게 된다. 이 장에서는∇^2G 연산자의 구조 및 특징을 알아보고 ∇^2G 연산자를 이용하여 입력 신호에 포함된 각 주파수의 진폭 특성을 나타내는 개괄 특징과 국부 특징을 추출할 수 있는 신경 회로망으로서 Interconnection 모델에 대해서 기술한다.[18]

1. ∇^2G 연산자

1차원 가우스 함수 G(t)의 2차 미분 연산자

$$-\nabla^2 G(t) = G''(t)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma^2}\left(1-\frac{t^2}{\sigma^2}\right) exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{1}$$

과 같이 표시된다. 이의 퓨리어 변환한 것을 LoG(ω)라 하면

$$LoG(\omega) = \omega^2\, exp\left(-\frac{\sigma^2\omega^2}{2}\right) \tag{2}$$

로 표시된다. 이 때의 중심 주파수 ω_0 및 진폭 A(ω_0)는

$$\omega_0 = \frac{\sqrt{2}}{\sigma} \tag{3}$$

$$A(\omega_0)=LoG(\omega_0)=\frac{2}{\sigma^2}\, exp(-1) \tag{4}$$

이 된다.

한편, 중심 주파수가 f_0인 신호 성분을 찾기 위한 ∇^2G 연산자의 표준 편차 σ는 식 (3)에서 다음과 같이 유도된다.

$$\sigma = \frac{\sqrt{2}}{2\pi f_0} \tag{5}$$

식 (1)은 t=0인 축에 대하여 좌우 대칭이며, $|t| < \sigma$인 경우에는 양의 값을 가지는 자극 영역으로 나머지 부분은 음의 값을 가지는 억제영역으로 나누어짐을 알 수 있으며, 식 (2), (3)과 (4)에서 ∇^2G 연산자는 대역 통과 특성을 가지며, 중심 주파수 ω_0는 가우스 함수의 표준 편차 σ에 반비례하며 또한 중심 주파수가 커질수록 중심 주파수에서의 진폭도 커짐을 알 수 있다.

2. Interconnection 모델

음성 신호 s(t)의 퓨리어 변환을 S(ω)라고 하면 이

신호에서 특정 신호만을 추출하기 위해서는 대역 통과 특성을 가지는 $\nabla^2 G$ 연산자를 작용시키면 되며, 다음과 같이 대응된다.

$$s(t) * \nabla^2 G(t) \Leftrightarrow s(\omega) \cdot LoG(\omega)$$

여기서 "*"은 콘볼류션 연산기호이다. 즉, 주파수 영역에서 특정 주파수 성분의 검출은 시간 영역에서 입력 신호 s(t)와 $\nabla^2 G$ 연산자의 콘볼류션 결과로 대응된다. 그리고 $LoG(\omega)$를 정규화하면 입력 신호의 각 주파수에 대한 진폭 특성은 콘볼류션의 최대값으로 대치할 수 있다.

n개의 $\nabla^2 G$ 연산자에서의 콘볼류션 최대값 Cnmax를 수식으로 나타내면 다음과 같다.

$$\mathbf{C} = \mathbf{L}\mathbf{A} \tag{6}$$

여기서

$$\mathbf{C} = [\ C_{1max}, \cdots, C_{kmax}, \cdots, C_{nmax}\ ]^T \tag{7}$$

$$\mathbf{L} = \begin{bmatrix} L_{11}, L_{12}, \cdots L_{1n} \\ L_{21}, L_{22}, \cdots L_{2n} \\ \cdots\ \cdots \\ L_1, \ L_2, \ \cdots L_{nn} \end{bmatrix} \tag{8}$$

$$\mathbf{A} = [\ A_1, A_2, \cdots, A_n\ ]^T \tag{9}$$

이다. 벡터 A는 입력 신호에 포함된 각 주파수의 진폭을 나타내며 L행렬의 요소 L_{jk}는 j번째 $\nabla^2 G$ 연산자의 콘볼류션 결과값에 k번째 $\nabla^2 G$ 연산자의 중심 주파수 성분이 미치는 영향을 나타낸 것이다. $\nabla^2 G$ 연산자는 정규화되어 있으므로 대각 요소는 항상 1이 되고 나머지 요소는 1보다 작은 실수가 된다. 벡터 C는 각 필터의 중심 주파수에 해당하는 $\nabla^2 G$ 연산자와 입력 신호를 기본 주기의 1/2보다 큰 구간동안 콘볼류션한 값 중 최대값을 나타낸다. 따라서 입력 신호의 각 주파수에 대한 진폭 특성은 식 (6)에서 다음과 같이 구할 수 있다.

$$\mathbf{A} = \mathbf{L}^{-1}\mathbf{C} \tag{10}$$

지금까지 해석한 결과를 이용하여 입력 신호에 포함된 각 주파수의 진폭 특성을 나타내는 개괄 특징과 국부 특징을 추출할 수 있는 신경 회로망 형태의 Interconnection 모델을 구성하면 그림 2와 같다. 여기서 입력층과 중간층(intermediate layer) 사이의 연결세기 분포는 $\nabla^2 G$ 연산자의 개형과 같도록 하며 중간층과 출력층(output layer) 사이의 연결 세기 분포는 역행렬 L^{-1}로 결정한다. 중간층의 출력은 선택도가 2정도로 낮은 $\nabla^2 G$ 연산자와 입력 신호의 콘볼류션 결과값 중 최대값(Cnmax)을 나타낸 것이므로 입력 신호의 개략적인 주파수 특성을 나타내는 개괄 특징이 되며, 출력층의 출력은 중간층의 출력인 개괄 특징에 역행렬의 연산이 수행되어진 것이므로 주파수의 진폭 특성을 나타내는 중간층의 출력과 비교했을 때 좀 더 세분화된 국부 특징이 된다.

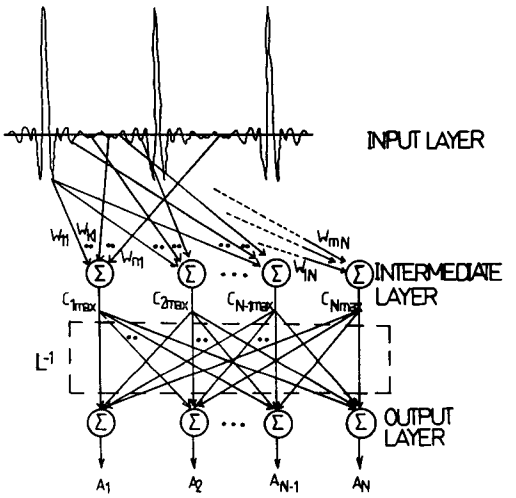


그림 2. Interconnection 모델
Fig. 2. Interconnection model.

Ⅳ. INNA 모델의 구성

신경 회로망은 자율적으로 판정 경계(decision boundary)를 정하며, 스스로 학습할 수 있는 기능을 가지고 있으며, 새로운 입력 패턴에 대해서도 가장 비슷한 범주를 찾을 수 있고, 학습받은 패턴과 약간 다른 형태이더라도 인식할 수 있으며, 병렬 구조를 하고 있기 때문에 처리 속도가 빠르다는 특징 등으로 인하여 최근에 연구가 활발히 진행되고 있다.

전형적인 다층 신경 회로망은 그림 3과 같다. 출력층과 입력층, 그리고 하나의 은닉층을 갖는 3층 구조를 이루고 있다. 그러나 전형적인 다층 신경 회로망은 인식하여야 할 대상의 수가 증가할 때 은닉층의 뉴런 수가 증가하게 되고 이로 말미암아 학습시간이 더욱 증가하는 단점이 있다. 이 점을 개선하기 위한 신경 회로망이 INN(Integrated Neural Networks)[9]이다.

INN은 제어 회로망과 각 그룹내의 대상을 인식하는 다수의 부회로망으로 이루어져 있다. 입력 패턴이 제어

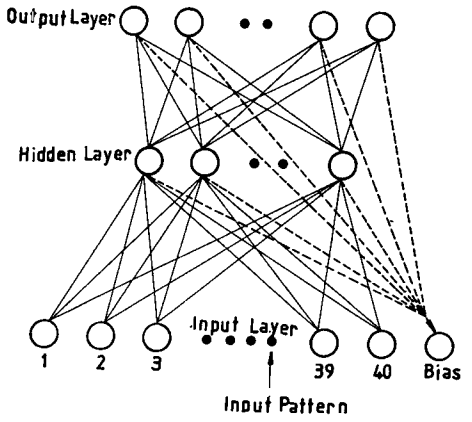


그림 3. 다층 신경회로망
Fig. 3. Multilayer neural network.

회로망과 부회로망에 인가되면 제어 회로망은 그룹내의 자음만을 학습하고 또한 입력 패턴이 속하는 그룹을 식별하여 다수의 부회로망이 최종적인 인식 결과를 출력할 수 있도록 게이트를 조절한다. 은닉층과 출력층 사이의 각 뉴런들이 부분적으로 서로 연결되어 있으므로 인식 대상이 많은 경우에도 학습 시간이 많지 않으면서 좋은 인식율을 얻을 수 있었다. 그러나 사람의 청각계는 대역 통과 특성을 가지며 주파수 1~2KHz범위에서 선택도가 2정도로 알려져 있는 신경 섬유[22]에서 완전히 분석된 주파수 정보를 청각 피질로 전달하여 주는 것이 아니라 하위 뉴런은 개괄 특징을 추출하여 신경 섬유로 전달하고 개괄 특징이 신경 섬유에서 청각 피질로 전달되는 과정에서 정보 처리가 수행되어 국부 특징으로 변환된다. 따라서 같은 특징 정보를 제어 회로망과 부회로망에 각각 인가하여 음성이 인식되도록 하는 INN 모델로써는 사람의 청각과정을 모델화할 수는 없을 것으로 생각된다.

신경 회로망 형태의 Interconnection 모델에서 중간층의 출력은 입력 신호의 개략적인 주파수 특성을 나타내는 개괄 특징이 되며 출력층의 출력은 개괄 특징에 역행렬의 연산이 수행되어 나온 것이므로 중간층의 출력과 비교했을 때 국부 특징이 된다. 이 두 가지 특징 파라미터를 이용하여 사람의 청각과정을 모델화할 수 있는 INNA 모델을 구성하면 그림 4와 같다.

INNA 모델은 특징 추출부와 인식부로 구성되어 있으며 특징 추출부는 Interconnection 모델로 구성하고 인식부는 개괄 분류단(global grouping stage)과 국부 분류단(local grouping stage)으로 구성한다. 개괄 분류단과 국부 분류단은 전형적인 다층 신경 회로망 구조를 하고 있으므로 BP(Back Propagation) 알고리듬으로 학습시킨다. 따라서 INNA모델은 사람의 청각 과정을 좀

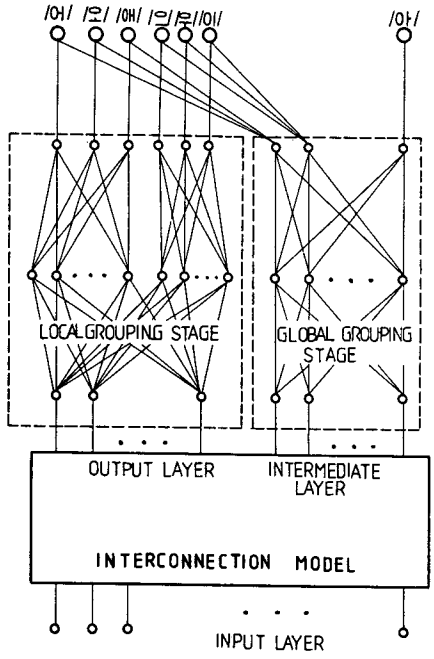


그림 4. INNA 모델
Fig. 4. INNA model.

더 정확히 모델화한 것이 되며 은닉층과 출력층 사이의 각 뉴런들이 부분적으로 서로 연결되어져 있으므로 학습 시간이 많지 않으면서도 좋은 인식율을 얻을 수 있는 신경망이 된다.

INNA 모델에서의 음성 인식 과정은 다음과 같다. 음성 신호가 Interconnection 모델의 입력층에 인가되면 음성 신호에 대한 개괄 특징이 추출되어 개괄 분류단 신경 회로망에 인가되고 개괄 분류단 출력은 입력 음성의 인식 범주를 결정한다. 한편 개괄 특징에 역행렬 연산이 수행되어 국부 특징이 추출되고 입력 음성에 따라 서로 다른 국부 특징이 국부 분류단의 각 신경 회로망에 인가되면 신경 회로망의 각각의 출력값은 결정된다. 입력 음성의 인식 범주를 결정하여 주는 개괄 분류단의 출력과 국부 분류단의 출력을 이용하여 최종적으로 입력 음성에 대한 인식을 수행하게 되므로 개괄 분류단의 출력이 최종 출력에 주위 집중 능력을 갖도록 하는 결과가 된다.

V. 시뮬레이션

남성 화자 9명(부산 대학교 방송국 어나운서 2명, 학생 7명)이 발음한 한국어 모음(아·어·오·우·으·이·애)을 방송국 스튜디오에서 녹음하여 저장하였다. 저장된 모음중 시간적으로 정상적인 부분을 분석 구간으로

하였으며 각 모음에서 5 프레임(frame) 구간만 특징 파라미터를 추출하였다. 이와 같이 하여 만들어진 전체 음성 패턴은 315개이다. 시뮬레이션 과정에서 어나운서와 학생을 구별하기 위하여 어나운서 2명은 M_A, M_b 혹은 화자 A, 화자 B로 나타내었다.

신경 회로망에 모음을 학습시키기 위하여 Interconnection 모델에서 얻어지는 개괄 특징과 국부 특징을 코딩하여 입력 패턴을 만들었으며 이 때 사용한 코딩 방법으로는 Interconnection 모델의 출력 신호에서 평균을 구한 후 평균값보다 큰 경우만 "1"로 하는 평균치(average) 방법, 중간값을 구한 후 중간값보다 큰 경우만 "1"로 하는 중간치(median) 방법, 포락선을 구한 후 증가하는 부분만 "1"로 하는 경사치(slope) 방법을 사용하였다.

모음을 인식하기 위하여 그림 3과 같은 다층 신경 회로망의 입력층 뉴런 수는 41개로 하였고 이 중 40개를 입력 패턴의 각 비트에 대응시켰으며 나머지 1개의 뉴런은 바이어스 항을 처리하기 위해 도입하였다. 출력층의 뉴런수는 모음의 갯수가 7개이므로 7개로 하였다. 그리고 은닉층의 뉴런 갯수를 5, 10, 20개로 각각 변화시키면서 시뮬레이션하였다.

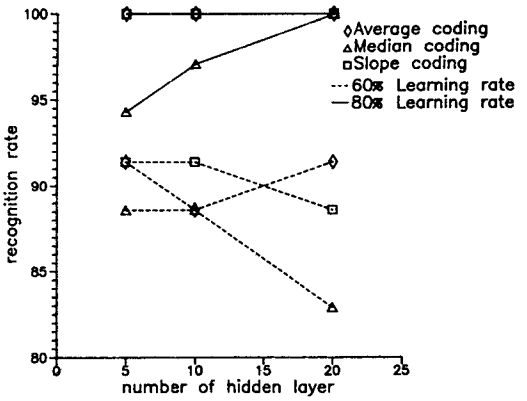
혀의 최고점으로 모음을 분류하면 기본 모음을 3그룹(/아/, /어·오·애/, /우·으·이/)으로 분류할 수 있는 것을 이용하여 INNA 모델에서 개괄 분류단 신경 회로망의 은닉층 뉴런 수는 10개로 하고 출력층의 뉴런 수는 모음 3그룹을 위하여 3개로 하였다. 모음 3그룹 중 /아/는 개괄 분류단에서 이미 분류되므로 국부 분류단 신경 회로망의 수는 2개로 하고 은닉층 뉴런 수는 5개로 하며 출력층의 뉴런 수는 각 그룹내에서 분류하여야 하는 모음의 갯수만큼인 3개로 두었다.[19]

출력에서 0.2이하는 "0.0"으로, 0.8이상은 "1.0"로 하여 출력 패턴을 구성하고 원하는 출력 패턴과 비교하여 인식 여부를 결정하였으며 출력값이 0.8이상되지는 않지만 다른 출력 노드와 비교해서 크면 인식할 수 있다고 생각하고 후처리를 하여 인식율을 계산하였다.

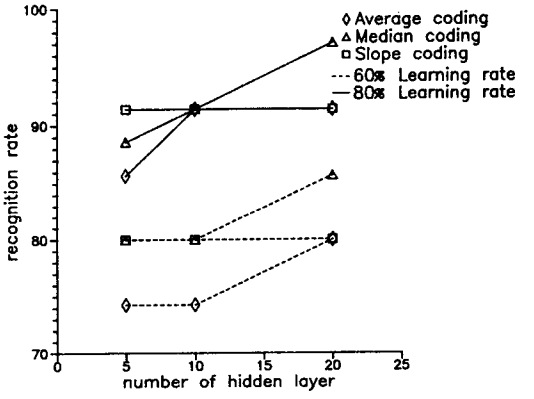
학습 대상 모음의 설정은 전체 음성 패턴중 임의로 60%, 80%씩 선택하여 학습시켰으며 이를 학습 샘플수(%)로 나타내었다. 수렴의 기준인 tss(total sum of error)는 0.05로 하고 학습율 η는 0.25, 모멘텀 항 α는 9로 하였으며 최대 반복 횟수는 학습에 필요한 전체 시간을 고려하여 500회로 하였다. 알고리듬은 IBM 386-DX 호환 기종에서 MSC를 사용하여 구현하였다.

각 화자(화자 A 또는 B)를 인식 대상으로 하고 다층 신경 회로망에서 실험한 결과를 그림 5에 나타내었다.

그림 5에서 화자 A의 경우, 60% 학습 샘플수에서 평균치 방법과 경사치 방법의 평균 인식율이 모두 90.5%이었으며 80% 학습 샘플수에서는 100%의 인식율을 나타내었다. 화자 B의 경우 중간치 방법이 좋은 결과를 나타내었으며 80% 학습 샘플수에서 평균 인식율이 92.4%로써 화자 A와 비교해 보았을 때 다소 떨어졌다. 이는 화자 B의 목소리가 화자 A와 비교해서 상당히 저음이었으며 이로 인하여 주파수 정보가 잘 추출되지 못하므로 발생된 것으로 생각된다.



(a)



(b)

그림 5. 코딩방법에 따른 각 화자의 인식결과
(a) 화자 A
(b) 화자 B

Fig. 5. Results of recognition for each speaker by coding methods.
(a) speaker A,
(b) speaker B.

두 화자(M_A, M_B)를 동시에 인식 대상으로 하고 다층 신경 회로망에서 실험한 결과를 그림 6에 나타내었다.

그림 6에서 중간치 방법, 80% 학습 샘플수에서 평균 인식율이 89.5%로 가장 좋았지만 그림 5에서 화자 B만

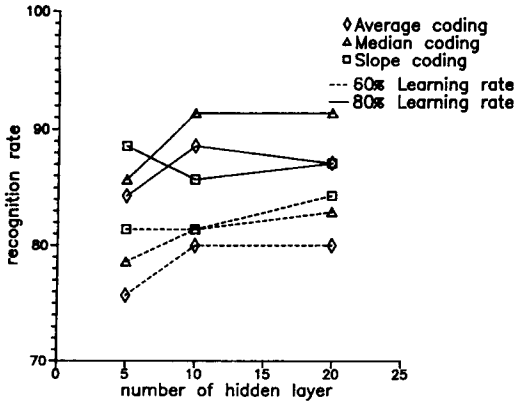


그림 6. 코딩방법에 따른 화자 2명의 인식결과
Fig. 6. Results of recognition for two speakers by coding methods.

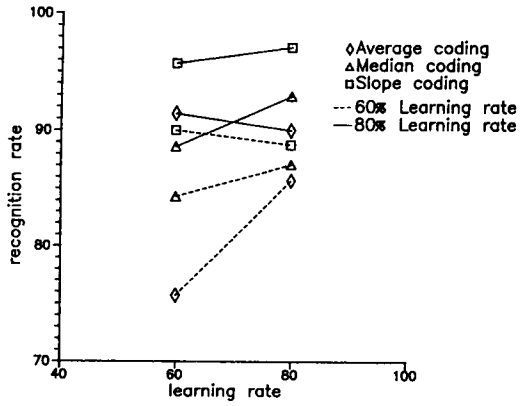


그림 7. 평균에 의한 화자의 인식결과
Fig. 7. Results of recognition for speakers by average.

의 결과보다 인식율이 떨어졌다. 이는 두 화자를 동시에 학습시켰을 때 인식하여야 할 범주가 넓어지므로 인식율이 떨어지는 것으로 생각된다.

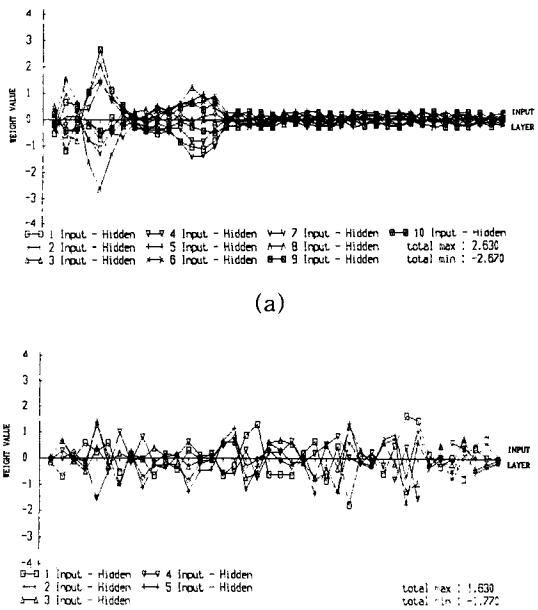
두 화자를 동시에 인식대상으로 하고, 시간에 따른 입력 패턴의 변화를 적게하기 위하여, 3 프레임에서 각각 구한 개괄 특징을 평균하여 현재 프레임에서의 개괄 특징으로 한 후 다층 신경 회로망에서 실험한 결과를 그림 7에 나타내었다.

그림 7에서 경사치 방법을 적용한 것을 입력 패턴으로 했을 때의 인식율이 97.1%로서 가장 좋았으며 이 결과는 평균하지 않았을 때의 가장 좋은 결과(중간치 방법, 91.4%)보다 약 6%의 증가를 보임을 알 수 있었다. 이는 사람의 경우 발음 지속 시간이 어느 정도 이상은 되어야 하며, 여러 구간의 입력 패턴에서 새로운 입력 패턴을 추출하여 인식하게 되므로 동일 음성일 경우 시간에 따른 입력 패턴의 변화가 작게 되어야 인식율이 높아진다는 것을 뜻한다.[23],[24]

두 화자(M_A, M_B)의 모음에서 추출된 평균 개괄 특징을 경사치 방법으로 코딩하고 국부 특징은 각각의 코딩 방법을 적용하여 얻은 70개의 패턴중 80%에 해당하는 56개의 패턴을 사람의 청각계와 비슷한 음성 인식 신경 회로망인 INNA 모델에 학습시켰다.

개괄 분류단과 국부 분류단에서 입력층과 은닉층간의 가중치 분포를 그림 8에 나타내었다. 개괄 분류단에서 가중치는 1.5KHz이내에서 크게 분포하고 있으며 국부 분류단에서 가중치는 주파수의 전영역에 분포하고 있음을 알 수 있었다.

표 2에 INNA 모델에서, Cnmax의 코딩 방법에 따른 각 그룹(/어·오·애/, /우·으·이/)별 인식 결과



(a)

(b)

그림 8. INNA 모델에서 가중치 분포
(a) global
(b) local grouping
Fig. 8. Weight distributions in INNA model.
(a) global,
(b) local grouping.

를 나타내었다. 평균 인식율이 96.2%로써 다층 신경 회로망의 결과인 그림 7의 학습 샘플수 80%때의 결과보다 인식율에서 약 3%의 증가를 나타내었고 학습 시간도 약

표 2. INNA 모델에서 두 화자의 인식 결과
Table 2. Results of recognition for two speakers by INNA model.

코딩 방법	구 룹	인식율	구 룹	인식율
Average	어·오·애	93.3	우·으·이	96.7
Median	어·오·애	96.7	우·으·이	86.7
Slope	어·오·애	100	우·으·이	100
	Global	100		

20% 감소함을 알 수 있었다. 그리고 경사치 방법에서는 100%의 인식율을 나타내었다.

음성 인식 시스템이 상용화 되기 위해서는 훈련받지 않은 다수의 사람들을 대상으로 했을 때에도 인식율이 좋아야 한다.

표 3에 9명(학생 7명 포함)의 화자를 대상으로 하고 Interconnection 모델의 출력인 개괄 특징과 국부 특징을 경사치 방법으로 코딩한 후 전체 패턴 수 315개(화자 9명 * 모음 7개 * 모음당 패턴 5개)에서 학습 샘플수 80%로 INNA 모델을 학습시킨 후의 인식 실험 결과를 나타내었다. /아·어·우·이/는 인식율이 95.6% 이상으로 잘 인식하고 있으나 /오·애·으/는 인식율이 좋지 못하였다. /오·애/는 학생 1명의 화자에 대해서 오인식을 했지만 /으/는 학생들 화자에서 주로 오인식이 일어났다. 이는 /으/발음을 보통 때는 잘 사용하지 않으며, 따라서 발음이 정확하지 못하였기 때문에 발생된 것으로 생각된다. 각 입력 음성의 출력 결과에서 다섯 프레임 중 한 프레임 혹은 두 프레임만 다른 음성이라고 인식하는 경우가 있다. 이것은 과거의 인식 결과를 현재의 인식 결정에 영향을 미치도록 INNA 모델을 수정한다면 인식율은 높아질 것으로 생각한다.

표 3. 경사치 방법에 의한 화자 9명의 인식 결과
Table 3. Results of recognition for nine speakers by slope method.

출력 / 입력	아	어 오 애	우 으 이	계	인식율(%)
아	45	0 0 0	0 0 0	45	100
어	0	44 0 1	0 0 0	45	97.8
오	0	0 40 0	0 5 0	45	88.9
애	0	0 1 39	0 5 0	45	86.7
우	0	0 0 0	45 0 0	45	100
으	0	0 8 3	0 33 1	45	73.3
이	0	0 0 0	1 1 43	45	95.6

Ⅵ. 결 론

$\nabla^2 G$ 연산자를 이용하여 구성한 Interconnection 모델에서 추출되는 개괄 특징과 국부 특징을 특징 파라미터로 이용하여 음성을 인식하였다. 특징 추출부에서 인식부까지를 신경 회로망으로 구성한 INNA 모델을 제안하였으며 모음의 인식율을 전형적인 다층 신경 회로망과 비교한 결과는 다음과 같다.

전형적인 다층 신경 회로망을 이용한 모음 /아·어·오·우·으·이·애/의 인식실험 결과 두 화자를 동시에 인식대상으로 할 경우 학습 샘플수 80%에서 인식율이 89.5%였으며, 시간에 따른 입력 패턴의 변화를 작게 하기 위한 방법에서 평균 인식율이 93.3%였다.

INNA 모델에 시간에 따른 입력 패턴의 변화를 작게 하기 위한 방법으로 두 화자를 동시에 인식 대상으로 할 경우 학습 샘플수 80%에서 평균 인식율이 96.2%로서 전형적인 다층 신경 회로망의 결과보다 인식율에서 약 3%의 증가를 나타냈으며 학습 시간도 약 20% 감소하였다. 그리고 경사치방법으로 입력패턴을 코딩했을 경우는 100%의 인식율을 나타내었다. INNA 모델에 9명이 발음한 7개의 모음을 인식 실험한 결과, 평균 인식율 91.4%를 얻을 수 있었다.

인간의 청각 기능은 주파수 분석 기능 뿐만 아니라 주파수 세분화 및 잡음이 있는 환경에서도 특별한 음성만을 인식할 수 있는 주위 집중 기능도 있다. 제안한 방법은 인간의 청각 기능을 모델화한 것으로서 필터 뱅크 특성뿐만 아니라 다른 기능도 있을 것으로 생각되므로 이에 대한 연구를 계속하여 인간의 청각 기능을 모방할 수 있는 음성 인식 시스템을 구현해 보려고 한다.

參 考 文 獻

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Engle-wood Cliffs, N. J. : Prentice Hall, 1978.

[2] Hyun-Yeol Chung, Shozo Makino, and Ken'iti Kido, "Analysis and recognition of isolated Korean vowels using Formant frequency," *J. Acoust. Soc. Jpn. (E)* vol. 9, no. 5, pp. 225-232, 1988.

[3] D. R. Reddy, "Computer recognition of connected speech," *J. Acoust. Soc. Amer.*, vol. 42, pp. 329-347, Aug. 1967.

[4] S. B. Davis and P. Mermelstein, "Comparision of parametric representations for

monosyllable word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, Aug. 1980.

[5] B. Gold, Hopfield Model Applied to Vowel and Consonant Discrimination, *MIT Lincon Laboratory Technical Report*, TR-747, AD-A169 942, June 1986.

[6] T. Kohonen, "The neural phonetic typewriter," *IEEE Computer*, pp. 11-22, Mar. 1988.

[7] B. R. Kammerer and W. A. Kupper, "Design of hierarchical perceptron structures and their application to the task of isolated-word recognition," *International Joint Conference on Neural Networks*, vol. 1, pp. I-243- 249, June 1989.

[8] A. Weibel, "Phoneme recognition using time delay neural network," 產學技報, SP 87-100, 1987.

[9] T. Matsuoka, H. Hamadah and R. Nakatsu, "Syllable recognition using integrated neural networks," *International Joint Conference on Neural Networks*, vol. 1, pp. I-251-258, June 1989.

[10] 이황수, "신경회로 컴퓨터 : 이론, 응용및 구현(음성인식)," 한국과학기술원 산학협동강좌, 1988.

[11] 이기영, 이인섭, 최승호, 김좌룡, 배철수, 최갑석 "Hopfield Network를 이용한 단모음 인식에 관한 연구," 대한전자공학회 하계종합학술대회논문집, 제12권 제1호, pp. 632-634, 1989.

[12] 이종혁, 심재형, 윤태훈, 김재창, 이양성, "신경망을 이용한 모음의 인식 및 방법," 대한전자공학회 논문지, 제 27권, 제 11호, pp. 144-151, 1990.

[13] D. Marr and E. C. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. Ser. B.* vol. 207, pp. 187-217, 1980.

[14] D. G. Sinex and L. P. McDonald, "Synchronized discharge rate represention of voice-onset time in the chinchilla auditory nerve," *J. Acoust. Soc. Am.* vol. 85, no. 5, 1989.

[15] M. Zhu and K. Fellbaum, "A connectionist for speaker independent isolated word recognition," *ICASSP*, pp. S10. 10 529-532, 1990.

[16] 김진영 외, "귀의 특성을 고려한 선형예측과 음성인식," 음성통신 및 신호처리 workshop, pp. 114-117, 1990.

[17] 이희규 외, "음성인식을 위한 청각신경 정보처리 모델링," 한국 음향 학회지, 9권, 3호, pp. 42-47, 1990.

[18] 이종혁, 정용근, 윤태훈, 김재창, 박의열, 이양성, "$\nabla^2 G$ 연산자를 이용한 음성 신호의 주파수 분석, 대한전자공학회 논문지," 제 28권 B편, 제 4호, pp. 24-32, 1991.

[19] Y. S. Kim, *Phonetics*, Pusan National University, 1986.

[20] M. C. Teich, "Fractal character of the auditory neural spike train," *IEEE Trans. on Biomedical Eng.* vol. 36, no. 1, pp. 150-160, 1989.

[21] P. Baldi and W. Heiligenberg, "How sensory maps could enhance resolution through orded arrangements of broadly tuned receivers," *Biol. Cybern.* vol. 59, pp. 313-318, 1988.

[22] N. R. S. KIANG, "Peripheral neural processing of auditory information," *Handbook of Physiology, The Nervious System*, vol. 3, Sensory Process Part 2.

[23] P. D. Bricker and S. Pruzansky, "Effect of stimulus content and duration on talker identification," *J. Acoust. Soc. Am.*, vol. 40, pp. 1441-1449, 1966.

[24] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304-1312, 1974.

著 者 紹 介

李 鍾 赫 (正會員) 第28卷 B編 第4號 參照
현재 경성대학교
컴퓨터공학과 조교수

●

鄭 容 近 (準會員) 第28卷 B編 第4號 參照
현재 럭키금성 통신개발단
CPE그룹 연구원

●

南 基 坤 (正會員) 第26卷 第1號 參照
현재 부산대학교
전자공학과 조교수

●

尹 台 熏 (正會員) 第28卷 B編 第4號 參照
현재 부산대학교
전자공학과 조교수

金 在 昌 (正會員) 第28卷 B編 第4號 參照
현재 부산대학교
전자공학과 교수

●

朴 義 烈 (正會員) 第20卷 第1號 參照
현재 부산대학교
전자공학과 교수

●

李 梁 成 (正會員) 第26卷 第1號 參照
현재 부산대학교
전자공학과 교수.