1999年度 大韓電子工學會 秋季綜合學術大會 論文集 第22卷 第2號 99/11

# 한국어 음절 인식을 위한 MLP 신경망 구조 및 특징 추출에 관한 연구

금 지 수(琴 智 秀), 이 현 수(李 顯 洙)
경희대학교 전자계산공학과
전화 : (0331) 201-2947 / 팩스 : (0331) 202-1723

# A Study on MLP Neural Network Architecture and Feature Extraction for Korean Syllable Recognition

Ji Soo Kum, Hyon Soo Lee
Dept. of Computer Engineering, KyungHee University
E-mail : {tbno, leehs}@cann.kyunghee.ac.kr

## Abstract

In this paper, we propose a MLP neural network architecture and feature extraction for Korean syllable recognition.

In the proposed syllable recognition system, firstly onset is classified by onset classification neural network. And the results information of onset classification neural network are used for feature selection of input patterns vector. The feature extraction of Korean syllables is based on sonority. Using the threshold rate separate the syllable. The results of separation are used for feature of onset, nucleus and coda. ETRI's SAMDORI has been used by speech DB. The recognition rate is 96% in the speaker dependent and 93.13% in the speaker independent.

## I. 서론

인간의 정보처리 기술을 모방한 신경망(Neural Network)은 스스로 학습하는 능력과 초고속 병렬처리 능력을 바탕으로 음성인식을 비롯한 패턴분류, 적응제어 등의 분야에서 널리 사용되고 있다[1-2].

MLP(Multilayer Perceptron)는 오류 역전파(Error Back propagation) 학습 방법을 이용하는 대표적인 정적구조 신경망으로 음성인식에 적용하려면 시간에 따라 변화하는 음성의 신호로부터 정적인 특징을 추출해야 한다[3-4].

본 연구에서는 한국어 음성의 발화 및 인지 단위인 음절을 인식단위로 하는 음성인식 시스템의 MLP 신경망 구조와 특징 추출 방법을 제안한다. 제안하는 MLP 신경망의 구조는 음성인식에서 부정확한 끝점 검출이나 연결 또는 연속음 발음시 발생하는 음운현상에 따른 오인식을 최소화할 수 있도록 구성하였다. 그리고 MLP 신경망의 학습 및 실험에 사용되는 특징 패턴은 음성의 모든 정보가 음성인식에 필요한 것은 아니므로 음절에서 일부 구간의 정보만 추출하여 사용하였다[5].

MLP 신경망의 학습 성능에 미치는 여러 가지 파라미터 중 학습 속도 향상을 위해 각 학습 단계마다 학습률(learning rate)을 적응적으로 변화시키면서 학습하였다[1]. 음성인식기의 성능 평가를 위해 ETRI의 샘플이 숫자음을 인식 실험에 사용하였다.

본 논문의 구성은 2절에서 제안하는 음성인식 시스템을 소개하고, 3절에서는 음절 인식을 위한 특징 추출 방법에 대하여 기술하였다. 그리고 4절에서는 인식 실험과 결과 분석을 하였으며, 마지막으로 5절에서는 결론과 향후 연구 방향에 대하여 기술하였다.

## II. 제안한 음성인식 시스템

음성인식 시스템은 그림 1과 같이 음성 신호에서 특징 추출을 위한 전처리 단계와 음절 단위 인식을 위한 인식 단계로 구성된다.

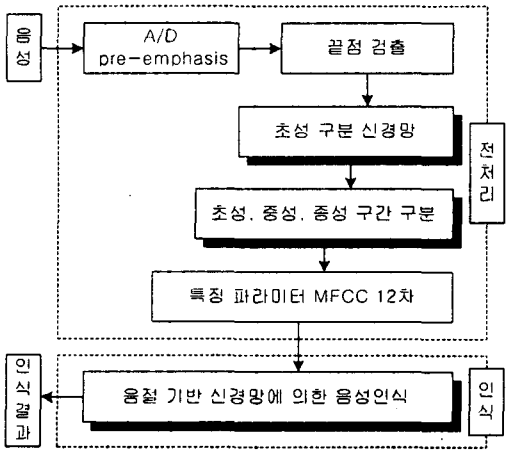


그림 1. 제안한 음성인식 시스템 구성

### 1. 전처리 단계에서의 초성 구분 신경망

초성 구분 신경망은 오류 역전파 학습을 통하여 초성의 음가가 존재하는 숫자음 "삼, 사, 칠, 팔, 구"와 음가가 존재하지 않는 숫자음 "영, 일, 이, 오, 육"의 두 개의 그룹으로 구분한다. 초성을 구분하는 이유는 초성이 존재하지 않는 숫자음의 초성 입력으로 중성을 사용하여 인식률을 높이고, 연결이나 연속음 발음시 발생하는 연음 현상을 고려하기 위해서이다.

### 2. 음절 단위 인식을 위한 MLP 신경망

음절 기반 신경망은 초성, 중성, 종성 각 하나의 프레임 정보만을 입력으로 사용하기 때문에 하나의 프레임 정보가 음의 특징을 잘 반영하지 못하는 경우 오인식이 발생하는 문제점이 있었다[5]. 이러한 문제점을 해결하기 위하여 입력으로 사용하는 구간의 프레임 개수를 확장하였다. 초성 구분 신경망의 인식 결과를 바탕으로 초성이 음가를 갖는 경우에는 초성 두 프레임을 입력으로 사용하고, 음가를 갖지 않는 경우에는 중성 두 프레임을 입력으로 사용한다.

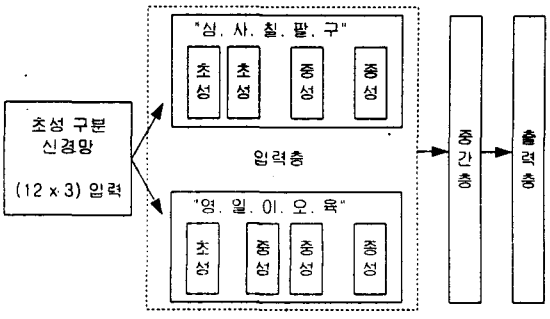


그림 2. 입력 프레임의 개수를 확장한 신경망 구조

### 3. 학습 속도 향상을 위한 적응적 학습 방법

본 연구에서는 신경망 학습에 영향을 미치는 여러 가지 학습 파라미터 중에서 학습률을 적응적으로 변화시켜는 학습 방법을 이용하여 학습시간을 단축시켰다.

적응적으로 학습률을 변화시키면서 학습하는 단계와 입력층과 중간층, 중간층과 출력층 사이의 출력값은 다음과 같다.

$$y = f(v^T x) = \frac{1-\exp(-v^T x)}{1+\exp(-v^T x)} \quad (1)$$

$$o = f(w^T y) = \frac{1}{1+\exp(-v^T x)} \quad (2)$$

$x$ : 입력 패턴 벡터 $y$ : 중간층 출력 벡터

$o$ : 최종 출력 벡터

$v$ : 입력층과 중간층의 연결강도 벡터

$w$ : 중간층과 출력층의 연결강도 벡터

목표치 벡터 $d$와 최종 출력 벡터 $o$ 사이의 제곱 오차 합 $E$는 식(3)과 같고 $K$는 출력 뉴런의 개수이다.

$$E = \frac{1}{2}(d_k - o_k)^2 + E \quad for\ k = 1, 2 \ldots K \quad (3)$$

출력층의 오차 신호 벡터 $\delta o$와 중간층에 전파되는 오차 신호 벡터 $\delta y$는 다음의 식(4), (5)와 같으며, $\delta o_k$는 $k$번째 출력 뉴런의 오차 신호이고 $w_{kj}$는 $j$번째 중간 뉴런과 $k$번째 출력 뉴런 사이의 연결 강도이다.

$$\delta o = (d-o)(1-o)o \quad (4)$$

$$\delta y = \frac{1}{2}(1-y^2)\sum_{k=1}^{K}\delta o_k w_{kj} \quad (5)$$

$p$ 학습 단계에서의 연결강도 변화량은 다음의 식(6), (7)과 같이 계산된다. 여기서 $\alpha$는 학습률이다.

$$\Delta w^p = \alpha\,\delta o y \quad (6)$$

$$\Delta v^p = \alpha\,\delta y x \quad (7)$$

$p+1$ 학습 단계에서의 중간층과 출력층간의 연결강도 $w^{p+1}$과, 입력층과 중간층의 연결강도 $v^{p+1}$은 식(8), (9)와 같다.

$$w^{p+1} = w^p + \Delta w^p \quad (8)$$

$$v^{p+1} = v^p + \Delta v^p \quad (9)$$

$p+1$ 학습 단계에서의 학습률은 $p$ 학습 단계에서의 제곱 오차 합에 대한 $p+1$ 학습 단계의 제곱 오차 합의 비율이 허용치 이하일 경우는 학습률을 증가시켰고,

허용치 이상일 경우에는 학습률을 감소시켰다.

$$\alpha^{n+1} = \alpha^{n} * increase\_rate \quad (10)$$

$$\alpha^{n+1} = \alpha^{n} * decrease\_rate \quad (11)$$

## III. 음절 단위 인식을 위한 특징 추출

음성인식을 하는데 있어서 음성의 모든 정보가 인식에 필요한 것은 아니다. 즉, 음성의 특성을 잘 표현할 수 있는 일부의 정보만으로도 인식이 가능하므로 한국어 음절을 구성하는 초성, 중성, 종성의 일부 특징만 사용하여 인식을 하였다.

### 1. 공명도(Sonority)

공명도는 조음점의 간극과 성대 울림의 유무에 의해 결정되는 소리의 크기이다[8]. 표 1에는 울림도 등급에 따른 국어 음소들이 나타나있다. 발음하는 음성으로부터 울림도 등급을 구분할 수 있으면 하나의 음절에서 자음과 모음의 구간을 구분할 수 있다.

표 1. 국어 음소들의 공명도

| 울림도 등급 | 1도 | 2도 | 3도 | 4도 |
|---|---|---|---|---|
| 발소리의 부류 | 장애음 | 비음 유음 | 반모음 | 모음 |
| | 파열음 파찰음 마찰음 | | | 고모음 중고모음 중저모음 저모음 |

### 2. 음절에서의 특징 추출

초성·중성과 중성·종성 구간을 구분하는데 이용한 임계비율은 한국어 음절의 공명도에 기인하여 초성과 종성이 중성과 구분될 수 있는 비율로 하였다. 구분은 끝점 검출에 이용되는 단구간 에너지(Short Term Energy)와, 단구간 영교차율(Short Term Zero-crossing)을 사용하였다[6-7]. 먼저 음절에서 최대 에너지 프레임을 구하고 임계 비율을 적용하여 초성-중성 구분임계값 그림 3.(1)과, 중성-종성 구분 임계값 그림 3.(2)를 구한다.

- 초성_중성 구분 임계값 = 최대에너지 * 초성_중성 구분 임계비율
- 중성_종성 구분 임계값 = 최대에너지 * 중성_종성 구분 임계비율

구해진 임계값을 기준으로 얻어진 프레임들은 다시 단구간 영교차율 조건에 의해 만족하는 첫 번째 프레임을 신경망의 학습 및 실험에 사용되는 특징으로 결정한다.

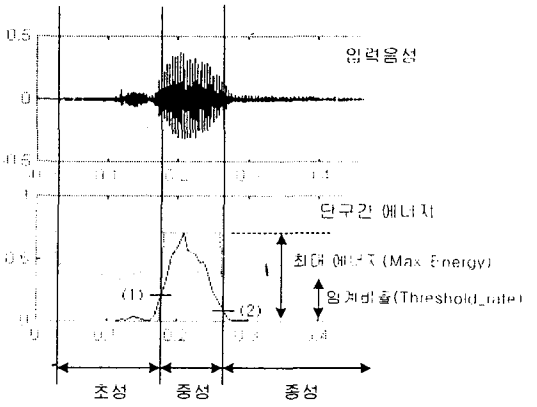


그림 3. 임계 비율에 의한 초성, 중성, 종성 구간 구분

## IV. 인식 실험 및 결과 분석

### 1. 음성 DB 및 분석조건

제안한 음성인식 시스템의 성능을 평가하기 위하여 ETRI의 샘플이 숫자음 "영, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구"를 사용하였다. 숫자음의 분석 조건은 표 3과 같다. 샘플이 숫자음에서 남녀 각 10인이 2회 발음한 400개의 데이터를 초성 구분 신경망과 음절 기반 인식 신경망의 학습 데이터로 사용하였다. 화자 종속으로 남녀 각 10인이 2회 발음한 400개의 데이터를 실험하였고, 화자 독립으로 학습에 사용하지 않은 남녀 각 10인이 4회 발음한 800개의 데이터를 인식 실험에 사용하였다.

표 3. 음성데이터 분석 조건

| 구분 | 분석조건 |
|---|---|
| 샘플링 주파수, 양자화 | 16kHz, 16bit |
| Filtering | LPF, 7kHz |
| Pre-emphasis | $1-0.97z^{-1}$ |
| 입력 패턴 프레임 길이 | 256 샘플 |
| 창함수 | Hamming |
| 특징 추출 | MFCC 12차 |

### 2. 신경망의 네트워크 구성

초성 구분 신경망은 끝점 검출된 음절에서 음의 시작점을 기준으로 3프레임을 입력으로 사용한다. 각 프레임은 256 샘플로 구성되고 128 샘플씩 중첩하였으며 인식 실험에서 중간 뉴런은 2개로 하였다.

음절 인식을 위한 신경망의 입력은 초성 구분 신경망의 인식 결과에 의해 초성 2프레임 중성 1프레임 종성 1프레임으로 사용하거나, 초성 1프레임 중성 2프레임 종성 1프레임으로 사용한다. 중간층의 뉴런 개수는

조절하면서 인식 실험한 결과 중간 뉴런 11개에서 가장 좋은 성능을 얻었다.

### 3. 인식 결과 및 분석

초성 구분 신경망 및 음절 인식 신경망의 숫자음 인식 결과(인식률)는 표 4와 같다. 그리고 표 5는 인식할 숫자음에 대한 오인식 된 숫자음의 개수를 나타낸다.

표 4. 인식 실험 결과(인식률)

| 구 분 | 화자 종속 | 화자 독립 |
|---|---|---|
| 초성 구분 신경망 | 98.75 % | 94.25 % |
| 음절 인식 신경망 | 96 % | 93.13 % |

표 5. 인식 실험 결과(화자 종속, 화자 독립)

| 구분 | | 인식할 숫자음 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 인식된 숫자음 | 0 | | 3 | | 1 | 2 | | 1 | | | |
| | 1 | 2 | | 2 | | | | 1 | 2 | 3 | |
| | 2 | | 9 | | | | | | | | |
| | 3 | 1 | | | | 12 | | | | | 1 |
| | 4 | | | | 6 | | | | | | |
| | 5 | | | | | | | | | | 1 |
| | 6 | 3 | 1 | | | | 1 | | | | |
| | 7 | | | | | | | | | 2 | 2 |
| | 8 | 4 | | | 2 | | | | 1 | | |
| | 9 | | | | | | 9 | | | | |

인식 결과를 분석해보면 초성 구분 신경망은 음의 유사성을 갖는 '5'와 '9'의 구분이 정확히 되지 않아 오인식 되는 경우가 많았고 부정확한 끝점 검출로 인한 오인식이 많았다.

숫자음의 음절을 인식하는 신경망에서는 중성과 종성을 구분하는 임계비율이 중성과 종성의 구분이 적절하지 못해 '1'과 '2', '3'과 '4' 같은 숫자음에서 오인식이 나타났다. 그리고 숫자음 '9'는 숫자음 '5'와 음의 유사성으로 인한 오인식이 발생하였고, 초성 구분 신경망의 오인식 결과에 영향을 받은 오인식이 있었다.

## V. 결론 및 향후 연구 방향

본 연구에서는 한국어 음성의 인지 단위인 음절 단위 음성인식을 위해 MLP 신경망을 구성하고 특징을 추출하여 인식 실험을 하였다. MLP는 정적인 패턴인식에서 높은 성능을 발휘하는 신경망이므로 시간에 따라 변하는 음성의 동적인 특성에서 정적인 특징을 추출해야 한다. 정적인 특징 추출을 위해 음절에서 초성, 중성, 종성을 구분할 수 있는 임계 비율을 설정하여 입력 패턴을 추출하였다. 기존의 음절 기반 신경망의 방법에서 발생했던 문제점을 줄이고 성능을 향상시키기 위하여 입력 프레임의 수를 확장하여 신경망을 구성하여 향상된 인식 결과를 얻었다.

향후 연구 방향으로는 초성 구분 신경망의 구분 그룹을 확장하여 보다 정확하고 많은 음절을 인식할 수 있도록 하고, 음절의 경계를 구분하는 보다 정확한 임계 비율의 결정과 입력 프레임을 개수를 확장하여 향상된 인식 결과를 얻는 것이다.

## 참고문헌


[1] Jacek M. Zurada, "Introduction to Artificial Neural Systems", PWS Publishing Company, 1995

[2] Simon Haykin, "Neural Networks A Comprehensive Foundation, Prentice Hall, 1999

[3] Carl G. Looney, "Pattern Recognition using Neural Networks", Oxford, 1997

[4] 박정선 외 3인, "KL 변환을 이용한 Multilayer perceptron에 의한 한국어 연속 숫자음 인식", 대한전자공학회 논문지, 제33권 B편 제8호, 1996 pp 105-113

[5] 곰지수, 이현수, "음절 기반 신경망을 이용한 한국어 숫자음 인식에 관한 연구", 한국음향학회 학술발표대회 논문집, 제18권 제1(s)호, 1999, pp 78-81

[6] Rabiner and Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993

[7] John R. Deller, Jr etc, "Discrete-Tim Processing of Speech Signals", Prentice Hall, 1987

[8] 이호영, "국어 음성학", 태학사, 1996