1999년도 한국음향학회 학술발표대회 논문집 제18권 제1(s)호

음절 기반 신경망을 이용한 한국어 숫자음 인식에 관한 연구

금지수, 이현수
경희대학교 전자정보학부 컴퓨터구조·신경망 연구실

A Study on Korean Digit Recognition Using Syllable Based Neural Network

Ji Soo Kum, Hyon Soo Lee
Computer Architecture and Neural Network Lab,
School of Electronics and Information, Kyung Hee University.
E-mail : {tbno, leehs}@cann.kyunghee.ac.kr

요 약

본 논문에서는 인간의 정보처리 기술을 모방한 신경망과 한국어 음절 구성의 특성을 이용하여 음절을 기반으로 하는 신경망 음성인식 방법을 제안한다.

제안한 방법에서는 임계비율을 정의하여 한국어 음절을 구성하는 초성·중성·종성을 구분하였고, 구분된 음절의 일부 구간 특징을 학습 및 인식의 특징 패턴으로 사용하여 음성인식 시스템의 전체적인 처리 단계를 줄였다. 한국어 숫자음 인식에 대한 성능 평가에서 20대 남성과 여성을 대상으로 화자 종속에서 96.5%의 인식률을 화자 독립에서 93%의 인식률을 얻었다.

1. 서론

인간의 뇌가 갖는 정보처리 능력을 인공적으로 실현하는 신경망(Neural Network)은 스스로 학습하는 능력과 초고속 병렬처리, 고장 허용(fault tolerance)과 같은 특성을 바탕으로 패턴분류, 적응제어, 음성인식 등과 같은 분야에서 널리 사용되고 있다[1]. 특히, MLP(Multilayer Perceptron)는 대표적인 정적구조 신경망으로서 오류 역전파(Error Back Propagation)라는 강력한 학습 방법을 이용하여 높은 성능을 발휘하고 있다[2][3].

그러나 정적 구조의 MLP 신경망은 화자가 발음하는 음성의 길이나 조음현상에 의해 나타나는 동적인 특성을 적절하게 흡수할 수 없으므로 음성인식 적용에 부적합하다. 따라서, 음성인식에 MLP 신경망을 적용하려면 음성의 동적인 특성에서 정적인 특징을 추출해야 한다[4].

본 연구에서는 음성의 특징을 추출하기 위해 단구간 에너지(short term energy)[5]와 영교차율(zero crossing rate)[5]에 임계값을 적용하여 한국어 음절을 구성하는 초성·중성·종성 구간을 구분하였다. 그리고 구분된 구간으로부터 음의 길이를 고려하여 신경망의 학습 및 인식 패턴을 추출하였다.

신경망의 학습 속도를 향상시키기 위하여 각 학습 단계마다 학습률(learning rate)을 적응적으로 변화시키면서 학습하였으며 제안한 음절 기반 신경망의 성능을 평가하기 위하여 20대 남녀 각 10인의 화자가 실험실 환경에서 2회 발음한 한국어 숫자음에 대하여 인식 실험을 하였다.

본 논문의 구성은 2절에서는 제안한 음절 기반의 신경망 구조와 적응적 학습 방법에 대하여 3절에서는 음성의 특징 패턴을 추출하는 방법에 대하여 기술하였다. 그리고 4절에서는 한국어 숫자음에 대한 인식 실험과 결과에 대한 분석을 하였으며, 마지막으로 5절에서는 제안한 신경망에 대한 결론과 향후 연구 방향에 대하여 기술하였다.

2. 음절 인식을 위해 제안한 MLP 신경망

오류 역전파 학습 방법에 의한 MLP 신경망은 정적인 패턴인식 분야에서 뛰어난 성능을 보이고 있다. 본 연구에서는 이러한 신경망의 학습 능력과 초성·중성·종성으로 구성되는 한국어 음절의 특성을 이용하여 음성의 동적 특성에서 정적 특징을 추출하여 인식할 수 있도록 신경망을 구성하였다.

2.1 음절 인식을 위한 MLP 신경망 구조

한국어 음절은 초성·중성·종성으로 구성되므로 이러한 구간에서 특징을 추출하여 동시에 입력하면 음절의 부정확한 끝점 검출로 인해 발생하는 오인식도 줄일 수 있고 음소 단위로 인식하는 경우처럼 후처리 과정을 통하여 음절을 구성하지 않아도 되므로 효율적이다. 신경망을 구성하는 입력 뉴런의 개수는 초성·중성·종성 각각에 대하여 전처리 단계인 신호처리 과정을 거쳐 얻은 MFCC(Mel Fourier Cepstral Coefficient)[6]의 차수와 같도록 하였다. 중간 뉴런의 개수는 인식률과 학습 속도를 고려하여 정하였고 출력 뉴런의 개수는 한국어 숫자음 음절을 인식 대상으로 하였으므로 '0'-'9'까지 10개로 하였다. 음절 인식을 위한 MLP 신경망의 구조는 그림1과 같다.

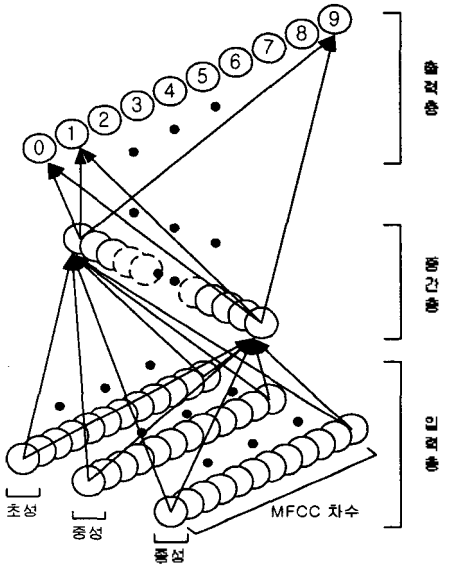


그림 1. 음절 인식을 위한 MLP 신경망

2.2 학습 속도 향상을 위한 적응적 학습 방법

학습 데이터에 대한 출력 오차를 최소화하는 방향으로 가중치를 조절하여 허용하는 오차 범위 내로 될 때까지 학습을 반복하는 오류 역전파 방법은 여러 응용분야에서 뛰어난 성능을 나타내고 있지만 실제 학습에 사용되는 초기 연결 강도와 학습률 등이 신경망 학습 성능에 커다란 영향을 준다[7]. 본 연구에서는 학습에 영향을 미치는 학습 파라미터 중에서 학습률을 적응적으로 변화시켜 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 d와 최종 출력 벡터 o 사이의 제곱 오차 합 E는 식(3)과 같고 K는 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 δo와 중간층에 전파되는 오차 신호 벡터 δy는 다음의 식(4), (5)와 같으며, δo_k는 k번째 출력 뉴런의 오차 신호이고 w_{kj}는 j번째 중간 뉴런과 k번째 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 α는 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.

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

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

3. 음절 인식을 위한 특징 추출

신경망의 학습 및 인식에 사용할 특징 패턴은 끝점 검출된 음절의 단구간 에너지와 영교차율에 임계 비율(threshold rate)을 적용하여 추출하였다. 적용되는 임계 비율은 조음점의 간극과 성대 울림의 유무에 의해 결정되는 소리의 상대적인 크기인 공명도(sonority)[8]의 등급 경계를 구분하는 역할로 음절에서 초성·중성·종성 구분이 가능하다.

음절의 끝점 검출을 위해 128 샘플을 한 프레임으로 64 샘플씩 중첩하면서 단구간 에너지를 구하였고 현재 프레임을 기준으로 끝점 검출 임계값을 넘는 프레임이 5프레임 이상일 경우에만 음성 구간으로 하였다.

초성은 끝점 검출된 음절로부터 시작점을 기준으로 첫 번째 프레임을 음절의 초성 특징 패턴으로 사용하였다.

중성은 음절에서 최대 크기를 갖는 단구간 에너지와 영교차율에 임계비율을 적용하여 임계값을 만족하는 구간의 첫 번째 프레임을 특징 패턴으로 사용하였다.

그리고, 음절의 종성은 중성과 종성이 연결되는 부분에서 단구간 에너지가 감소하는 성질을 이용하였고, 초성·중성을 구분하는 것과 다른 임계비율을 적용하여 임계값을 만족하는 구간의 첫 번째 프레임을 종성의 특징 패턴으로 사용하였다.

초성·중성과 중성·종성 구간을 구분하는데 이용한 임계비율은 한국어 음절의 공명도에 기인하여 초성과 종성이 중성과 구분될 수 있는 비율로 하였다. 식(12)는 n개의 프레임으로 구성된 음절의 최대 단구간 에너지 값을 나타내고 식(13)은 임계비율을 적용하여 임계값을 구하는 수식이다.

$$Max\_Energy = Max(Energy(l)) \quad l = 1, 2, \cdots n \quad (12)$$

$$threshold = Max\_Energy * threshold\_rate \quad (13)$$

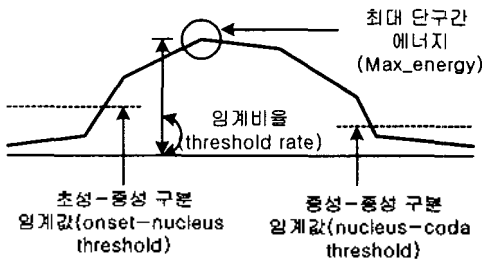


그림 2. 임계비율에 의한 음절 구분 임계값

이러한 방법으로 음절을 구분하면 초성과 종성의 길이가 짧아서 적절하게 프레임을 적용할 수 없는 경우와 음절이 처음부터 중성으로 구성되는 경우가 나타난다. 그리고 음절이 파찰음으로 시작하여 초성 구간에서 단구간 에너지가 높게 나타나는 경우도 나타난다. 음절의 특징 패턴을 추출하는 방법을 정리하면 다음과 같다.

[경우1] 초성·중성·종성으로 구성된 음절(그림3-a)
초성 : 음절의 시작점을 기준으로 첫 번째 프레임
중성 : 초성·중성 임계값을 넘는 첫 번째 프레임
종성 : 중성·종성 임계값 이하의 첫 번째 프레임

[경우2] 중성으로만 구성된 음절(그림3-b)
초성 : 음절의 시작점을 기준으로 첫 번째 프레임
중성 : 음절의 시작점을 기준으로 첫 번째 프레임
종성 : 음절의 마지막 프레임

[경우3] 초성·중성으로 구성된 음절(그림3-c)
초성 : 음절의 시작점을 기준으로 첫 번째 프레임
중성 : 초성·중성 임계값을 넘는 첫 번째 프레임
종성 : 음절의 마지막 프레임

[경우4] 중성·종성으로 구성된 음절(그림3-d)
초성 : 음절의 시작점을 기준으로 첫 번째 프레임
중성 : 음절의 시작점을 기준으로 첫 번째 프레임
종성 : 중성·종성 임계값 이하의 첫 번째 프레임

[경우5] 초성이 파찰음으로 시작하는 음절(그림3-e)
초성 : 음절의 시작점을 기준으로 첫 번째 프레임
중성 : 단구간 에너지와 단구간 영교차율 모두
각각의 임계값을 만족하는 첫 번째 프레임
종성 : 중성·종성 임계값 이하의 첫 번째 프레임

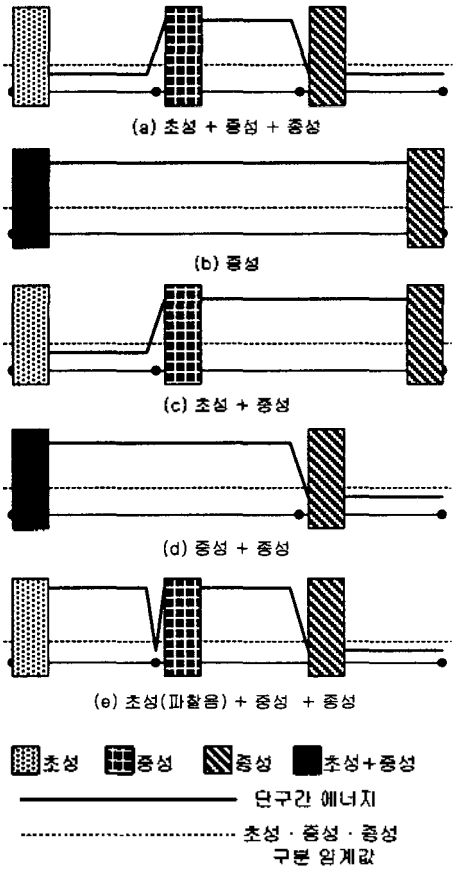


그림 3. 음절의 특징 패턴 추출 구간

4. 인식 실험 및 결과 분석

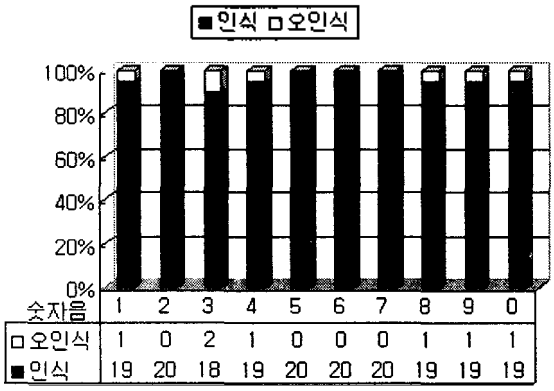
제안한 음절 기반 신경망의 성능을 평가하기 위하여 20대 남녀 각 10인이 한국어 숫자음 '0'-'9'를 실험실 환경에서 4회씩 발음하여 음성 데이터베이스를 구축하였고 구축된 데이터베이스로부터 남녀 각 5인이 2회 발음한 음성을 선택하여 신경망의 학습 패턴으로 사용하였다. 인식 실험으로는 화자 종속으로 남녀 각 5인이 2회 발음한 음성을 실험하였으며, 화자 독립으로는 남녀 각 5인이 2회 발음한 음성을 가지고 성능을 평가하였다. 음성데이터의 분석 조건은 표1과 같다

표 1. 음성데이터 분석 조건

구분	분석 조건
샘플링 주파수/양자화	11025Hz, 16Bit
Pre-emphasis	$1-0.95z^{-1}$
입력 패턴 프레임 길이	256 샘플
창함수	Hamming
특징 추출	MFCC 12차

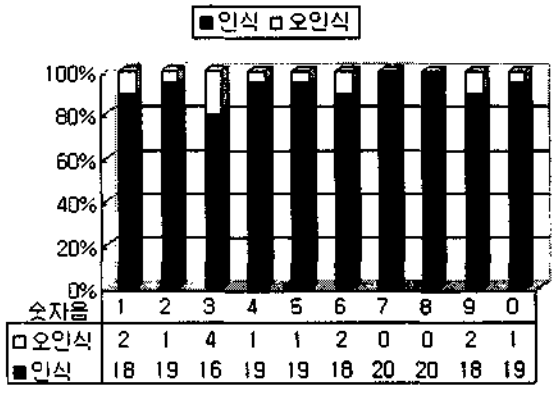
그림4와 그림5는 초성·중성·종성을 구분하는 임계비율을 조절하면서 실험한 결과 인식률이 높게 나타난 경우로 화자 종속에 대하여 96.5%의 인식률을 화자 독립에 대하여 93%의 인식률을 얻었다.

인식 결과를 분석하면 오인식이 주로 발생하는 숫자음은 '1'과 '3' 그리고 '9'였다. 숫자음 '1'은 '2'로 오인식 되었고 '3'은 '4'로 오인식 되었다. 이러한 오인식의 이유는 음절에서 중성·종성을 구분하는 임계비율이 높게 설정되어 음절에서 종성 부분의 특징 추출에 문제가 있다고 추측할 수 있다. 그리고 숫자음 '9'가 '5'로 인식되는 이유는 초성 보다 중성과 종성의 영향이 인식에 크게 반영된 결과로 볼 수 있다.



숫자음	1	2	3	4	5	6	7	8	9	0
□오인식	1	0	2	1	0	0	0	1	1	1
■인식	19	20	18	19	20	20	20	19	19	19

그림 4. 남녀 각 5인의 화자 종속 인식 결과
(임계비율 : 초성·중성: 28% 중성·종성 : 17%)



숫자음	1	2	3	4	5	6	7	8	9	0
□오인식	2	1	4	1	1	2	0	0	2	1
■인식	18	19	16	19	19	18	20	20	18	19

그림 5. 남녀 각 5인의 화자 독립 인식 결과
(임계비율 : 초성·중성: 28% 중성·종성 : 17%)

5. 결론

본 연구에서는 정적인 패턴인식에서 높은 성능을 발휘하는 MLP 신경망을 음절 기반의 음성인식이 가능하도록 구성하여 한국어 숫자음 인식을 하였다. 음절의 단구간 에너지와 영교차율을 이용하여 초성·중성·종성을 구분하였고, 구분된 구간에서 특징 패턴을 추출하여 인식 실험을 하였다. 한국어 숫자음을 인식 대상으로 실험한 결과 화자 종속 및 독립에서 높은 인식률을 얻을 수 있었다.

향후 연구 방향으로는 음절에서 특징 구간을 구분하는 임계비율의 적응적인 조절과 입력 프레임의 확장 및 화자의 나이와 성별 등의 차이에서 나타나는 인식률 차이를 줄이는 것이다. 그리고 음절을 기반으로 인식할 수 있는 실질적인 대상에 적용할 계획이다.

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, pp 105-113, 1996
5. Rabiner and Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993
6. John R. Deller, Jr etc, *Discrete-Time Processing of Speech Signals*, Prentice Hall, 1987
7. 김응명, 이현수, "에러 역전과 학습 성능 향상을 위한 초기 가중치 결정에 관한 연구", 한국정보과학회 가을학술발표논문집, Vol. 25, No.2, pp 333-335, 1998
8. 이호영, *국어 음성학*, 태학사, 1996
9. 이영호, 정홍, "음절을 기반으로한 한국어 음성인식", 대한전자공학회 논문지, 제31권 B편 제1호, pp 11-22, 1994