한국어 연결 숫자음 인식을 위한 시공간 신경회로망의 개발

[O]이 종식 정 재호
인하대학교 전자공학과

Development of Spatio-Temporal Neural Network for Connected Korean Digits Recognition

Jong Sik Lee Jae Ho Chung
Dept. of Electronic Eng. Inha University

ABSTRACT

In this paper, a new approach for Korean connected digits recognition using the Spatio-Temporal Neural Network (STNN) is reported. The data of seven digits phone numbers are used in the recognition of connected words, and in the initial experiment, digit recognition rate of 28% was achieved. In this paper, to increase recognition rate, two different approaches are analyzed. In the first system, to compensate the STNN's own defect and to emphasize the Korean word's phonic characters, the starting point of phone is pointed by comparing the average magnitude and zero-crossing rate and the ending point is pointed by comparing only zero-crossing rate. The digit recognition rate increased to 61%. Also, in the second system, to consider fact that same word's phone is varied severally, the number of STNNs of each word is increased from one to five, and then the varied same word's phones can be included to the increased STNNs. The digit recognition rate of connected words increased to 89%.

1. 서 론

정보화 사회로 변해가는 현대 사회에, 인간과 기계와 의사 소통을 보다 자연스럽고, 정확하게 하고자 하는 욕구가 높아 있다. 인간과 기계의 대화를 인간과 인간의 대화에 가깝게 하기 위하여, 인간의 가장 기본적인 통신 수단인 음성을 이용하는 것이 그 하나의 방법이다. 그리고, 음성을 이용한 사람과 기계의 인터페이스를 이루기 위한 핵심 기술이 음성 인식이다. 음성 인식이란, 인간의 음성을 인식알고리즘을 통해 단어나 문장으로 전환하는 것인데, 본 논문에서는 인식알고리즘으로 신경회로망을 도입했다.

신경회로망은 인간의 생물학적 뉴런시스템에 기초를 두고 있으며, 병렬 처리 기능과, 적응 학습 능력, 그리고, 결함 극복 능력 등의 장점을 가지고 있어, 최근 들어 신경망을 이용한 음성인식에 대한 연구가 활발히 진행되고 있다 [1,2]. 본 논문에서는, 시간에 따라 순차적으로 변화하는 음성 신호의 동적 특성을 고려하여 제안된 시공간 신경회로망 즉 STNN (Spatio-Temporal Neural Network)을 이용하여, 한국어 숫자음 인식을 시도하였다 [3,4,5]. STNN은 인식 과정에서 화자의 발성 간의 20% 정도의 증감에는 크게 영향받지 않으며, 음성 신호의 부분적인 시간변화와 주파수변화의 영향에도 민감하지 않은 장점을 가진 것으로 알려져 있다. 현재까지의 연구에서는, STNN을 이용한 고립 숫자음 인식에 대한 연구가 진행되어 왔으나, 본 논문에서는 연결 숫자음 인식을 위하여 STNN을 도입하였다.

본 논문은 한국어 연결 숫자음 인식에 대한 연구로서, 100개의 서로 다른 7자리 전화 번호를 연속적으로 발음한 데이타를 사용하여 인식 실험을 하였다. 7개의 시작점과 7개의 끝점 검출을 위하여 에너지값과 zero-crossing rate을 비교하였고, 화자가 의식적으로 숫자 사이를 끊어서 발음함으로 시작점과 끝점 검출을 가능케 하였다. 그러나, 7자리 연결음의 초기 인식률은 28%로 매우 저조하였다. 본 논문에서는, 연결음의 인식률 향상을 위하여 두 가지 제안을 하였다. 첫번째는, 연결음에 대하여 새로운 시작점, 끝점 검출법을 시도하였다. 즉, 마지막 몇 개의 weight 구간들과 입력 구간들의 유사성은 처음과 중간의 weight 구간들과 입력 구간들의 유사성보다 상대적으로 최종 출력값에 큰 영향을 미치는 STNN 자체의 미비점을 보완하고, 한국어 단어의 발음 특성에 착안하여, 시작점은 에너지값과 zero-crossing rate의 비교로써 검출하고, 끝점은 에너지값만으로 검출하였다. 새로운 끝점의 검출로 7자리 연결음의 인식률은 61%로 향상시켰다. 그리고, 두번째는, 각 단어에 대한 STNN의 갯수를 증가시켰다. 즉, 같은 단어 임지라도, 발음이 여러 가지로 변하는 경우를 생각하여 각 단어의 STNN의 갯수를 1개에서 5개로 늘려서, 같은 단어가 여러 가지로 달리 발음되어도, 충분히 그 다른 발음을 별도로 구별된 STNN에서 포함할 수 있게 하였다. STNN의 확장으로, 7자리 연결음의 인식률은 89%로 향상되었다.

2절에서는 STNN의 구조와 동작원리에 대하여 언급하였고, 3절에서는 실험 구성에 대하여 서술하였다. 4절에서는 본 논문에서의 제안점과 그 결과에 대해 논하였고, 마지막으로 5절에서는 결론을 내었다.

2. STNN (시공간 신경 회로망)

2.1 구조와 동작원리

STNN은 여러 개의 층으로 구성되어 있으며, 각 층은 시간 지연 단위에 대한 weight (가중치 또는 특징값)의 connection line으로 연결된 neuron들로 구성되어 있다. 각 층의 구조는 동일하며, 각 층의 neuron의 수는 입력 신호 전체의 길이를 일정시간으로 나눈 구간의 갯수와 같다. 한 단어가 입력으로 들어오면, 시간지연 순서에 따라 첫번째 구간의 LPC-cepstrum 계수들이 층 전체의 neuron들을 활성화시키고, 출력값을 낸다. 시간의 지남에 따라 두번째 구간의 입력신호가 들어오면, 두번째 구간의 LPC cepstrum 계수들을 계산하고, 다시 층 전체의 neuron들을 활성화하여 출력값을 계산한다. 이와 같은 과정을 마지막 입력신호의 구간까지 반복 수행하여, 최종 출력값을 구한다. 각 층의 구조를 보면 그림 1에 설명되어져 있으며, 각 층을 구성하고 있는 i번째 neuron의 입력값을 수식으로 나타내면 다음과 같다.

$$I_i = \overline{Q_j} \cdot \overline{W_i} + d\sum_{k=1}^{i-1} x_k \qquad (2\text{-}1)$$

식 (2-1)에서, I_i는 i번째 neuron의 입력값을, $\overline{Q_j}$는 j번째 입력구간의 LPC-cepstrum 벡터를, $\overline{W_i}$는 i번째 neuron의 weighting 벡터를, x_k는 k번째 neuron의 출력값을 나타낸다. 또한, 식 (2-1)에서 d (< 1)는 i번째 neuron에 전달되는 i번째 이전 neuron들의 출력량의 크기를 조절하는 상수이다.

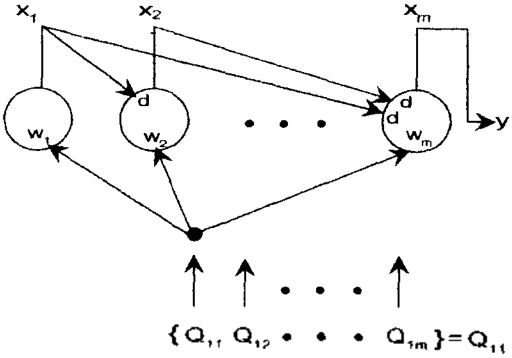

그림 1 STNN을 이용한 한 층의 구조

한편, i번째 neuron의 출력값 x_i는 다음과 같은 미분방정식을 통하여 나타내어 진다.

$$\dot{x}_i = A(-ax_i + b[I_i - \Gamma]^+) \qquad (2\text{-}2)$$

식 (2-2)에서, a와 b는 양의 상수값이다. 식(2-2)에 포함되어 있는 함수 $[I_i - \Gamma]^+$는 다음과 같이 정의된다. 즉,

$$[I_i - \Gamma]^+ = \begin{cases} I_i - \Gamma, & \text{if } I_i - \Gamma > 0 \\ 0, & \text{if } I_i - \Gamma \le 0 \end{cases} \qquad (2\text{-}3)$$

따라서 식 (2-3)에서 Γ는 임계값(threshold)의 역할을 한다. 식 (2-2)에서 함수 $A(\cdot)$는 attack function이라 부르며, 다음과 같이 정의 된다.

$$A(u) = \begin{cases} u, & \text{if } u > 0 \\ cu, & \text{if } u \le 0 \end{cases} \qquad (2\text{-}4)$$

그림 2에 attack function의 시간에 따른 변화, 즉 neuron의 시간에 따른 출력값이 설명되어져 있다.

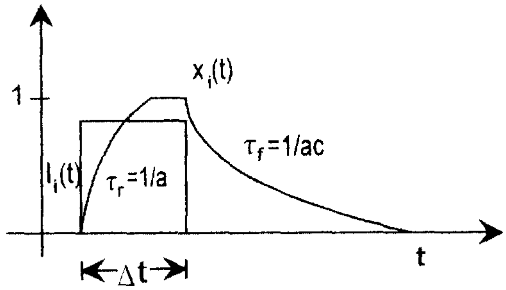

그림 2 Neuron의 시간에 따른 출력값의 변화

시간이 완료되었을 때에, 마지막 neuron의 최종 출력값이 입력신호와 입력신호를 적용한 STNN 사이의 닮은 정도를 나타낸다. 따라서, 입력된 숫자음을 인식하기 위하여서는, 입력신호의 LPC-cepstrum을 10개의 서로 다른 STNN 즉, 숫자 0부터 9까지 각각의 숫자음에 해당하는 STNN에 각각 입력신호로 사용하고, 10개의 최종 출력값을 얻은 후, 이들을 비교하여 최종 winner를 결정한다.

3. 실험 구성

본 연구에서 구성한 음성 인식 시스템은 그림 3과 같다. 마이크를 통해 받아 들여진 음성은 12bit 양자화 레벨을 갖는 A/D converter를 통하면서, 표본화 주파수 10KHz로 sampling 된다. 디지털로 바뀐 음성신호에 Rabiner와 Sambur가 제안한 average magnitude와 zero-crossing measurement algorithm[6,7]을 적용하여 시작점과 끝점 검출을 한다. 시작점과 끝점이 검출된 신호는 시간 영역에서 전체의 길이를 10개의 선형적 프레임으로 나누어 분석한다. 각 프레임으로 부터 autocorrelation 방법을 이용한 Durbin algorithm을 통해서 16차 LPC 계수를 추출한다 [8]. 계산된 LPC 계수들로부터 식(3-1)과 식(3-2)를 사용하여 16차 LPC-cepstrum 계수들을 계산한다.

$$c[1] = -a[1], \qquad (3\text{-}1)$$

$$c[n] = -a[n] - \sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right) a[k]\, c[n-k], \quad 1 < n \le p \qquad (3\text{-}2)$$

$a[n]$: LPC 계수

$c[n]$: LPC-cepstrum 계수

p : LPC 계수의 차수

그리고, 추출된 LPC-cepstrum 계수들을 STNN의 입력으로 사용하기 위해서 0 과 1 사이의 값들로 정규화한다.

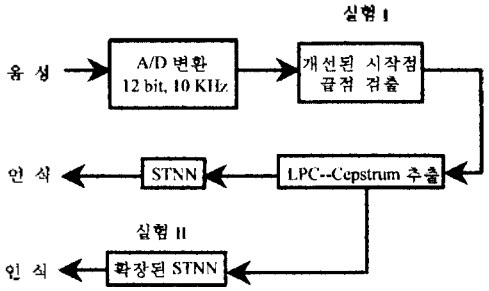

그림 3 음성 인식 시스템

4. 제안점과 실험 결과

4.1 연결 숫자음 인식

실험은 화자 종속 시스템이며, 한 명의 화자가 일주일 간격으로 3번, 100개의 서로 다른 7자리 전화 번호를 연속적으로 발음한 데이타를 사용하였다. 연결음(connected word)의 발음은, 화자가 의식적으로 숫자 사이를 끊어서 발음함으로써, 에너지값과 zero-crossing rate의 비교로써 7개의 시작점과 7개의 끝점을 숫자 사이의 침묵없이 검출할 수 있게 하였다. 첫 번째의 두 주, 200개의 전화 번호, 총 1400개의 숫자음을 가지고 network을 학습하였으며, 셋째 주의 100개의 전화 번호, 총 700개의 숫자음을 가지고 인식 실험을 하였다. 7자리 전화 번호의 초기 인식률은 28%로 매우 저조함을 보였다.

4.2 연결음의 끝점 검출 (실험 I)

실험 I 에서는, STNN의 인식과정에서의 비비점을 고려하고, 한국어 단어의 발음 특성에 맞추어, 시작점과 끝점을 검출하여, 인식 실험을 하였다. STNN에서는, 시간적으로 앞에 있는 뉴런의 출력값은 바로 다음 뉴런의 출력값에 영향을 미치도록 구성되어 있으나, 뉴런의 출력값은 앞에 있는 뉴런의 출력값보다 i 뉴런에 해당되는 weight구간과 입력 구간의 유사성에 더 많은 영향을 받는다. 즉 (4-1)식 에서 $\overline{Q_i} \cdot \overline{W_i}$ 가 $d\sum_{k=1}^{i} x_k$ 보다 시간의 흐름에 따라 상대적으로 I_i에 많은 영향을 미친다.

$$I_i = \overline{Q_i} \cdot \overline{W_i} + d\sum_{k=1}^{i} x_k \qquad (4\text{-}1)$$

I_i : i번째 neuron의 입력값

$\overline{Q_i}$: i번째 입력구간의 LPC-cepstrum 벡터

$\overline{W_i}$: i번째 neuron의 weighting 벡터

x_k : k번째 neuron의 출력값

그러므로, 시간이 경과할수록, 앞쪽에 있는 뉴런들의 출력값보다 뒤쪽에 있는 뉴런들에 해당되는 weight구간과 입력 구간의 유사성이 최종 출력값에 많은 영향을 미치게 된다. 결국, 마지막 몇 개의 weight구간들과 입력 구간들의 유사성은 처음과 중간의 weight구간들과 입력 구간들의 유사성 보다 상대적으로 최종 출력값에 큰 영향을 미치게 된다. 결국, 모든 입력 구간이 최종 출력값에 고르게 반영되지 않고, 음의 끝부분의 입력 구간일수록 최종 출력값에 많은 영향을 미친다. 이 점이 STNN의 비비점으로 지적받고 있다. 그러므로, STNN에서는 시작점보다는 정확한 끝점 검출이 인식률에 많은 영향을 미치게 된다. 또한, 한국어 단어의 발음 특성을 살펴 보면, 음의 앞부분에는 무성자음, 마찰음, 파열음, 그리고, 모음이 올 수 있고, 음의 끝부분에는 유지음(ㄱ,ㄷ,ㅂ)과 유성음(ㄴ,ㅁ,ㅇ,ㄹ), 그리고, 모음이 올 수 있다. 특히, 본 실험에서 사용한 숫자음(공, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구)은 표 1 과 같이 분석된다.

표 1. 한국어 숫자음의 분석

음의 앞부분	음의 끝부분
무성자음 : ㄱ,ㅅ	유성자음 : ㅇ,ㄹ,ㅁ
마찰음 : ㅊ	유지음 : ㄱ
파열음 : ㅍ	모음 : ㅣ, ㅓ,ㅗ,ㅜ
모음 : ㅓ,ㅗ,ㅠ	

단어의 발음 특성을 살펴볼 때, 음의 앞부분은 에너지값의 비교와 zero-crossing rate의 비교로써 검출이 용이함을 알 수 있다. 반면에, 음의 끝부분은 파열음과 마찰음이 없으므로, zero-crossing rate의 비교함으로써의 검출이 불필요하며, 에너지값의 비교로써 검출이 용이함을 알 수 있다. 연결음은 음이 서로이 붙어 있어, 다음에 파열음과 마찰음이 오면, 음의 끝부분에서의 zero-crossing rate의 비교는 무성화한 끝점 검출의 원인이 된다. 또한, 연결음을 각 단어로 구분할 때는, 음의 앞부분과 음의 끝부분을 검출하는 방법을 서로 달리함으로 보다 정확한 단어 사이의 경계를 설정할 수 있다. 따라서, STNN의 비비점과 한국어 단어의 발음 특성에 착안하여, 시작점은 에너지값과 zero-crossing rate의 비교로써 검출하고, 끝점은 에너지값만으로 검출하여, 인식 실험을 하였다. 이와 같이, 새롭게 시작점과 끝점을 검출하므로, 7자리 연결음의 인식률은 61%로 향상시켰다.

4.3 STNN의 확장(실험 II)

실험 II 에서는, 실험 I 에서 제시한 방법으로 시작점과 끝점을 검출하고, 각 단어의 STNN의 수를 늘려서, 새로운 network을 구성한 후 인식 실험을 하였다. STNN의 전체 구성을 살펴 보면, 각 단어에 따라 개개의 STNN으로 구성되어 있다 (예 : " 0 " 의 STNN, " 1 " 의 STNN, ...). 그런데, 같은 의미의 단어 인식라도, 발음이 여러 가지로 변하는 경우를 생각할 수 있다. 예를 들면, "6"의 발음은 "육", "융", "뉵", "륨" 등

으로 적절 반응할 수 있다. 따라서, 한 개의 STNN이 같은 단어의 연속된 발음을 모두 포함하고 있어야 한다. 같은 이유로 나, 따라서, 실험 II 에서는, 각 단어의 STNN의 개수를 1개에서 5개로 늘려서, 같은 단어가 여러 개로 되어 발음되어도, 충분히 그 다른 발음을 반영한 STNN에서 포함할 수 있게 하였다.

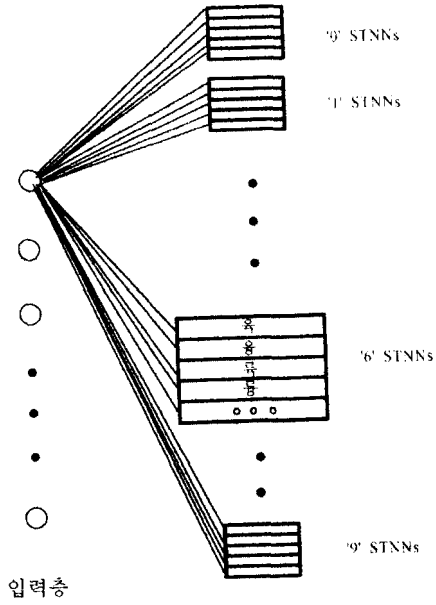

그림 5 확장된 STNNs 의 구조

그래서, 총 50개의 STNNs으로 network을 구성하였다. Weight를 초기화할 때는, 같은 단어의 5개의 STNNs을 모두 동일하게 초기화하였다. training할 때는, 전체 50개의 STNNs 중에서 최종출력값이 가장 큰 STNN만을 training하게 된다. 또한, 인식할 때에도, 전체 50개의 STNNs 중에서 최종출력값이 가장 큰 STNN에 해당되는 단어를 인식하게 하였다. 이로써, 7자리 연결음의 인식률을 80%로 향상시켰다.

표2. STNN에서의 초기 7자리 연결음, 실험 I, 실험 II 의 연결음에 대한 인식률

	인 식 률
초기 7자리 인속음	28 %
실험 I	61 %
실험 II	89 %

5. 결 론

이번 숫자음 인식에서는, 시작점과 끝점 검출이 연결 숫자음보다 용이하므로, 인식률 향상에 잇점을 갖고 있지만, 연결 숫자음 인식은 음이 서로에 붙어 있어, 시작점과 끝점 검출에 오차가 심하므로 인식률이 떨어진다. 인식률은 7자리 연결음이 모두 맞았을 경우만 인식된 것으로 간주하여 계산하였다.

본 논문에서, 제안한 STNN을 이용한 한국어 연결 숫자음 인식 실험에서는, 100개의 서로 다른 7자리 전화 번호를 연속적으로 발음한 데이타를 사용하여 인식 실험을 하였다. 에너지값과 zero-crossing rate의 비교로써 7개의 시작점과 7개의 끝점을 각 쌍 사이의 겹침없이 검출하였다. 그러나, 7자리의 연결음의 초기 인식률은 28%로 매우 저조함을 보였다. 본 논문에서는, STNN구조의 이점을 부각하고, 한 단어 단어의 발음 특성에 적합하여, 시작점은 에너지값과 zero-crossing rate의 비교로써 검출하고, 끝점은 에너지값만으로 검출하였다. 제안된 시작점, 끝점 검출로 7자리 연결음의 인식률은 61%로 향상시켰다. 또한, 같은 단어 인식라도, 발음이 여러 가지로 변화는 경우를 생각하여 각 단어의 STNN의 개수를 1개에서 5개로 늘려서, 같은 단어가 여러 가지로 달리 발음되어도, 충분히 그 다른 발음을 반영한 STNN에서 포함할 수 있게 하였다. 이로써, 7자리의 연결음의 인식률을 89%로 향상시켰다. 본 논문의 두 가지 실험은 전화 번호처럼 연결 단어의 인식에서는 각 단어가 겹침없이 정확한 끝점 검출이 선행되어야 하며, 음의 앞부분과 음의 끝부분을 검출하는 방법을 서로 달리함으로, 보다 정확한 단어 사이의 경계를 설정할 수 있음을 보였다. 또한, 같은 단어 인식라도, 발음이 여러 가지로 변하는 경우도 고려하여야 함을 지적하여, 각 단어에 대한 확장된 STNN들이 변화하는 음의 특성을 나타낼 수 있음을 보였다.

참고 문헌

[1] P. Demichelis, "On the Use of Neural Networkks for Speaker Independent Isolated Word Recognition," *Int. Conf. on Acoust., Speech, and Signal Proc.*, May 1989.

[2] 이중식, 이상욱, "신경망과 구분 분석을 이용한 한국어 연결 숫자음 인식," 대한 전자공학회 추계지, pp. 21-30, 1993.

[3] J. A. Freeman and D. M. Skapura, *Neural Networks*, Addison Wesley, 1990.

[4] 백승우, 홍승홍, "시공간 패턴인식 신경회로망을 이용한 한국어의 인식," 음향학회지 제 12권 3호, 1993.

[5] 이중식, 정재호, "시공간 신경회로망을 이용한 한국어 숫자음 인식," 한국통신학회지, pp. 771-779, March, 1995.

[6] L. R. Rabiner and M. R. Samber, "An Algorithm for Determining the Endpoints of Isolated Utterances," *Bell Tech. Journal*, vol. 54, no. 2, pp. 297-315, Feb. 1975.

[7] L. R. Rabiner and F. Soong, "Single Frame Vowel Recognition Using Vector Quantization with Several Distance Measures," *AT&T Technical Journal*, vol. 64, pp. 2319-2330, 1985.

[8] Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. ASSP-35, pp. 1414-1422, Oct. 1987.