Spatio-Temporal Pattern Recognition Neural Network를 이용한 전동 휠체어의 음성 제어에 관한 연구

°백 승우, 김 승법, 권 장우, 이 응혁, 홍 승홍
인하 대학교 전자 공학과

A Study on the Voice-Controlled Wheelchair using Spatio-Temporal Pattern Recognition Neural Network

S.W.BAEK, S.B.KIM, J.W.KWON, E.H.LEE, S.H.HONG
Dept. of Electonic Eng., INHA UNIV.

ABSTRCT

In this study, Korean speech was recognized by using spatio-temporal recognition neural network. The subjects of speech are numeric speech from zero to nine and basic command which might be used for motorized wheelchair developed in own Lab.

Rabiner and Sambur's method of speech detection was used in determining end-point of speech, speech parameter was extracted by using LPC 16 order. The recognition rate was over 90%.

1. 서론

음성에 대한 연구는 40년대 Bell 연구소를 중심으로 시작하였고 60년대 디지탈 신호 처리(Digital Signal Processing, DSP)의 발전과 함께 음성의 신호처리 연구가 급속히 발전하였으며, 1968년 Atal과 Schroeder에 의해서 선형 예측법(Linear Prediction Method)이 제안되었다. 72년에는 Itakura와 Saito가 편자기 상관법을 제안하였고 이 두가지 방법은 음성의 인식및 합성 분야에서 많이 이용되어 지고 있다.

신경회로망(Neural Netwrok)은 기존의 폰 노이만(Von Neumann)방식의 컴퓨터가 문자나 음성등의 애매모호한 정보의 처리에 어려움을 느끼는 것에 대한 보완책으로 나온 것이다.

신경회로망의 특징은 기존의 디지탈 컴퓨터에서는 0,1의 2진 기호를 이용하여 Yes/No의 판단을 하는 것이나 신경회로망은 아날로그 신호를 이용하여 불일치하거나 모순된 데이터에 대해서도 가중 판단을 내림으로서 문제를 해결한다는 것이다. 또 데이터의 분류에서도 완전 정합 보다는 근접 방향의 결합을 행함으로써 일부의 파손된 데이터나 애매모호한 데이터에 대하여서도 근접한 결과를 얻을 수 있기에 Fault-Tolerance가 강하다.

본 연구에서는 많이 알려져 있는 신경회로망은 아니지만 음성 신호의 인식에 용이한 시공간 차원 패턴 인식기(Spatio-Temporal Pattern Recognition, SPR) 신경 회로망을 사용하여 한국어 숫자음과 본 연구실에서 개발하고 있는 전동화 수동 휠체어에 사용할 수 있는 간단한 명령어 대해서 인식을 시도하였다.

여기서 특징 파라메터를 추출하는 방법으로는 해밍 윈도우를 사용하여 16차 LPC계수를 추출하여 사용하였으며, PC를 통해서 신경회로망을 시뮬레이션 하여 인식을 시도하였다.

여기서 인식한 숫자음마다 각각의 기능을 부여한다면, 전동휠체어의 확장된 명령어로 사용할 수 있다. 즉, 예를 들어서 1과 2에 '빨리', '천천히'를 3에는 '180도 회전'등의 기능 명령어로 준다면 음성 제어 휠체어의 명령어를 확장 할 수 있다.

2. Spatio-Temporal Pattern Recognition Neural Network

이 네트워크는 시공간 차원 패턴 인식기라고 부르며, 1980년대 중반 Bart Kosko와 Harry Klopf에 의해서 학습률이 제안되었으며, *Kosko/Klopf learnimg law* 이라고 부르기도 한다. 이 네트워크의 주 사용분야는 음성의 인식과 networks의 제어 분야이다. 특히 음성인식에서 화자의 발성 길이의 20% 정도의 증감은 변화하지 않은 것으로 인식하여 처리 가능하다.

이 네트워크를 사용해서 음성인식을 할때 기본적인 어려움은 speech의 classification을 잡는 것이다. 고립단어의 인식에서는 한번의 입력으로 이 네트워크를 학습시키게 된다. 그러므로 좋은 classification을 선택해서 학습시키는 것이 대단히 중요하다.

Spatio-Temporal warping은 다음식 (1)과 같은 전달 함수를 가지게 된다.

$$T : \mathcal{F} \subset \mathcal{A} \rightarrow \mathcal{A} \qquad (1)$$

매핑한 각 Spatio-Temporal Pattern의 $\mathcal{A}$ ($\mathcal{F}$는 부분집합) 는 모든 가능한 Spatio-Temporal Warping이다. Time warping은 패턴 $x(t)$를 $x(\theta(t))$로 전달한다. 여기서 θ는 단조증가 함수이다. Time warping은 속도의 증감에 움직여서 R^N의 궤도에 앞뒤로 움직이게 된다.

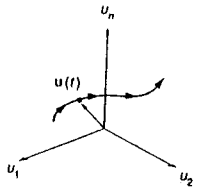


그림 1 Spatio-Temporal Pattern의 시간에 따른 움직임

그림 1은 n-dimension 공간의 Spatio-Temporal Pattern을 보인다. 이 패턴의 Time Warp Transformation에서 이 패턴은 같은 궤도로 움직인다. 그러나 원 패턴과는 다른 속도로 이 궤도상을 움직이게 된다. 이 스피드의 변화률은 $d\theta/dt$로 Warping되어 진다. Time Warps $\theta(t)$는 다음의 식 (2)와 같은 일반적인 감도를 가지게 된다.

$$0.5 \le d\theta/dt \le 2.0 \qquad (2)$$

Spatio-Temporal Pattern Neural Network는 다음 그림 2와 같은 일반적인 모델로 설계되어 진다. 여기서 L줄은 processing elemants 상수이다. 각 i개의 행이 있다면 V_i개의 패턴 인식훈련에 사용되어진다.

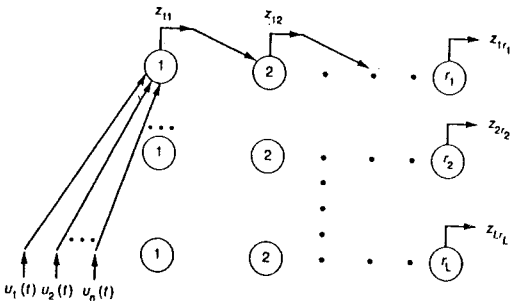


그림 2 SPR 네트워크의 회로 모델링

그림 3은 Spatio-Temporal Pattern 네트워크에서 전형적인 Processing Element의 동작을 표현한 것이다. 그림 3의 Processing Element를 l행에 i번째 열이라고 할때 이 PE 신호의 출력을 $Z_{li}(t)$라고 하자. 이 PE는 모든 하위열의 PE로 부터 입력을 받는데 여기서,

$$Z_{li} = U\ (\ x_{li}(t) - \sigma_{li}\ ) \qquad (3)$$

여기서,

$$x_{li}(t) = \alpha_{li}\ (\ -c_{li}\ x_{li}(t-1)$$
$$+ d_{li}\ U(\ [\varphi_{li} - |v_{li} - u(t)|]\ z_{l(i-1)}(t-1))) \quad (4)$$

여기에서,

$$0 \le x_{li}(t) \le 1,$$

$$z_{l0}(t) = 1$$

$$U(\zeta) = \begin{cases} 1 & \text{if } \zeta > 0 \\ 0 & \text{if } \zeta \le 0 \end{cases} \qquad (5)$$

그리고,

$$\alpha_{li}(\xi) = \begin{cases} \xi & \text{if } \xi \ge 0 \\ \phi_{li}\,\xi & \text{if } \xi < 0 \end{cases} \qquad (6)$$

여기서 α는 Attack 함수이다. ϕ의 범위는 0부터 1사이의 값으로 그 Attack함수의 특성 곡선은 그림 4와 같다. 식 (4)는 이 SPR 신경회로망의 전달 함수이다.

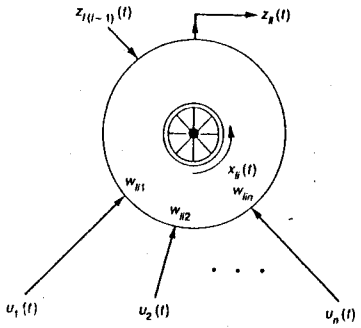


그림 3 Spatio-Temporal Pattern의 Processing Element

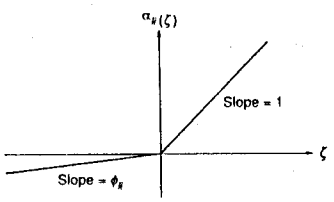


그림 4 Attack 함수의 특성 곡선

3. 실험 구성

본 연구에서 구성한 음성 인식 시스템의 구성도는 그림 5와 같다.

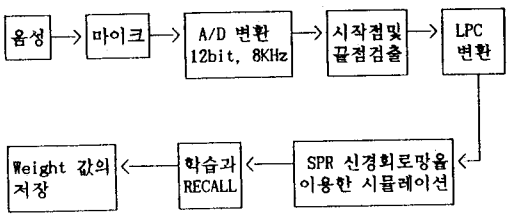


그림 5 음성 인식 시스템의 구성도

음성 인식 데이터를 얻기 위해서 12bit 양자화 레벨을 갖는 A/D 인터페이스 카드를 IBM 486 호환 기종에 연결하여 사용하였으며, 표본화 주파수는 8KHz를 사용하였고, 한 프레임을 12.5 msec로 하였다. 또한 마이크로 입력된 음성 파형은 A/D 변환을 위하여 증폭단과 저역 통과 필터를 사용하였다. 입력 받은 데이터의 시작점과 끝점 검출을 하여 정확한 음성 데이터를 추출해내었다.

음성의 시작점과 끝점의 검출은 실 시간 처리를 고려하여 영교차율과 에너지만으로 음성 검출을 하는 Rabiner와 Sambur의 음성 검출 방법을 채택하였다. 여기에서 사용하는 에너지는 식 (7)과 같다.

$$E(n) = \sum_{i=-50}^{50} | s(n+i) | \qquad (7)$$

여기서 s(n)은 음성 샘플들이고, 본 연구에서는 12.5 msec 구간에 대한 음성 진폭의 합이다.

해밍 윈도우를 사용한 LPC 방법을 사용하였고 LPC 차수는 16차를 사용하였다. 예측 차수를 10차에서 20차까지 분석한 결과 13-16차까지는 비슷한 특성이 나타남을 알 수 있었다.

SPR 신경회로망의 구성은 SPR 전달 함수의 파라메터인 Firing Rate는 0.4를, Input Clampt는 0.7을, Mod Factor는 0.8을, Gain은 0.8을 사용하였다. 이 파라메터는 여러번의 반복 실험을 통해서 얻었으며 한국어 단음의 인식을 LPC 변환과 SPR 신경회로망을 사용해서 할때 가장 최적화된 파라메터라고 생각된다.

이 SPR 신경회로망에서 Hidden Layers의 갯수를 최적화 시키는 것이 인식률 향상에 많은 도움이 됨을 알 수 있었다. 이 신경망의 특성상 Hidden Layers의 개수를 가급적 적게 하는 것이 인식률을 향상 시킬 수 있으나 너무 적은 개수를 취하게 되면 음성 데이터의 파손및 인식률의 저하가 생기므로 최적화한 Hidden Layers를 구하기 위해서 많은 반복 실험하여 산출하였다.

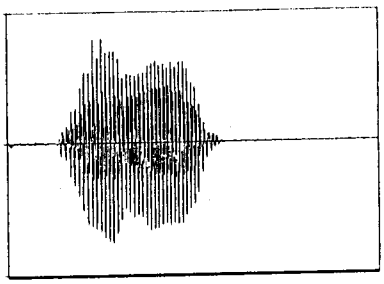
이 파라메터로 네트워크를 학습및 Recall을 하여 각 음성의 인식을 시뮬레이션 하는 것이다. 각각의 출력에 나온 가중치 결과를 비교하여 인식 여부를 결정한다.

음성 인식을 위해서는 한국어 숫자음 10개를 사용하였다. 어느 정도 인식률을 확인한 다음에 본 연구실에서 개발중인 전동화 수동 휠체어에서 응용 할 수 있는 음성 명령어들을 수행하여 보았다.

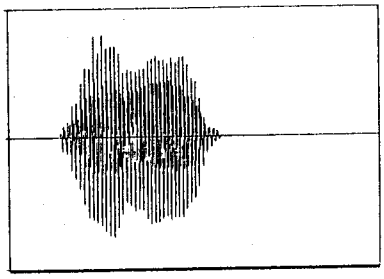
실험에는 20대 후반의 남성 화자를 대상으로 하여 실험 하였으며, 숫자음 인식 실험과 휠체어용 명령어 인식 실험에는 각기 다른 화자를 사용하여 인식 실험을 하였다.

4. 실험 결과

이 실험을 위해서 입력 받은 음성 데이터와 끝점 검출된 데이터는 그림 6과 같다. 그림 6에서 보면 쉽게 볼 수 있는데, 끝점 검출이 상당히 잘되고 있음을 알 수 있다.



(a) 끝점 검출전의 파형



(b) 끝점 검출후의 파형

그림 6 끝점 검출전의 파형과 끝점 검출후의 파형

입력 받는 음성 데이터를 LPC 변환을 한 다음에 학습용 표준 패턴을 생성하였다. 그림 7은 LPC 변환한 것의 한 예이다. 아래 그림은 한개의 프레임의 LPC 변환한 그래프이다.

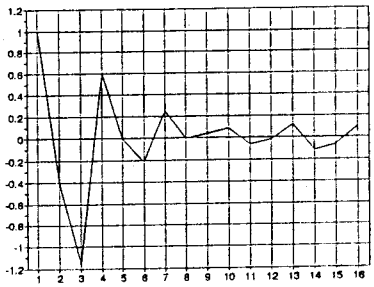


그림 7 1개 프레임을 LPC 변환한 그래프

한국어 숫자음의 인식 결과는 표 1과 같다. 총 실험 횟수는 250회 이다.

다음은 본 연구실에서 개발, 연구중인 전동화 수동 휠체어에 사용할 수 있는 기본적인 명령어에 대한 인식 결과이다. 이 실험에서 사용한 어휘는 가, 서, 좌, 우, 뒤의 5개 기본적인 명령어의 인식을 테스트 해 보았다.

	0	1	2	3	4	5	6	7	8	9	불인식	인식률
0	24										1	96 %
1		25										100 %
2			25									100 %
3				20		1					4	80 %
4			2		22						1	88 %
5						18				2	5	72 %
6			3				18				4	72 %
7		2						23				92 %
8									24		1	96 %
9										23	1	92 %

전체 인식률 222/250 * 100 = 88.8 %

표 1 한국어 숫자음 인식의 인식률 (화자 KJK)

	가	서	좌	우	뒤	인식불능	인식률(%)
가	15						100 %
서	2	13					87 %
좌	2		12			1	80 %
우				15			100 %
뒤					15		100 %

70/75 * 100 = 93.3%

표 2 휠체어 명령에 사용할 수 있는 음성의 인식률

5. 고찰및 결론

본 연구에서는 2명의 화자에게서 입력 받은 250개, 75개 즉, 모두 325개 데이터를 전부 사용하였고, 이 데이타를 Rabiner와 Sambur의 음성 검출 방법을 사용하여 끝점 검출을 하였으며, LPC 16차 계수를 사용해서 음성 파라메터를 추출하였다. 이 파라메터를 가지고 학습을 시킨 SPR 신경회로망에에 RECALL을 시켜서 가중치를 가지고 인식률을 구하였다. 각각 88.8%, 93.3%의 음성 인식률을 얻을수 있었다. 같은 음을 발음하여도 발음하는 화자의 감정에 따라서 그 길이가 상당히 변하게 된다. 그러나 이 SPR 신경회로망은 이러한 시간적인 변화에 많은 영향을 받지 않고 인식할 수 있다는 것을 보였다. 타 연구 논문과 비교하여서 인식률은 비슷하거나 약간 우수함을 알수 있다. 이번에 입력받은 화자는 훈련이 전혀 되어 있지 않은 화자로서, 음성 길이가 상당히 변화하였지만 비교적 좋은 인식률을 기록했다. 처음에 실험한 숫자음의 경우는 일주일의 간격을 두고 데이터를 두번에 나누어서 입력 받았는데, 발음 하는 사람의 감정등에도 좋은 적응성을 보임을 알수 있었다.

이번 연구는 Off-line 시스템이었지만 이 신경망을 이용한 음성 인식은 처리 속도가 무척 빠르기 때문에 신경회로망칩을 사용해서 구성한다면 실시간 처리도 충분히 가능하리라고 본다.

7. 참고문헌

1. Robert Hecht-Nielsen : "NeuroComputing", Addison Wesley, 164-192, (1990)

2. Casimer Klimasauskas, John Cuiver, and Garrett Pelton : "Neural Computing", NeuralWare, Inc., UG263-UG264, UG279-UG285, NC157-NC162, (1989)

3. James A. Freeman, David M. Skapura : "Neural Networks Algorithms, Applications, and Programming Techniques," Addison-Wesley, 341-371, (1991)

4. Veljko MiLutinovic, Paolo Antognetti : "Neural Networks Concepts, Applications, and Implementations", Prentice Hell, 54-72 Volume I, (1991)

5. Maureen Caudill, Charles Butler : "Understanding Neural Netwroks computer explorations", Massachusetts Institute of Technology, 44-46 Volume II, (1992)

6. Sadaoki Furui : "Digital Speech Processing, Synthesis, and Recognition", Dekker, 85-125, (1992)

7. L.R. Rabiner, R.W.Schafer : "Digital Processing of Speech Signals", Prentice-hall, 396-452, (1978)

8. L.R. Rabiner, M.R. Samber : "An algorithm for determining endpoints of isolates utterances," Bell syst. Tech. J. , VOL. 54 297-315, (1975)

9. A. Nejat Ince : "Digital Speech Processing Speech Coding, Synthesis and Recognition", Kluwer Ascdemic Publishers, 111-124, (1992)

10. Shozo Saito : "Speech Science and Technology", IOS Press, 5-9, (1992)

11. R Linggard, D J myers, and C Nightingale : "Neural Networks for Vision, Speech and Natural Language", Chapman & Hall, 129-134, (1992)

12. Russell C. eberhart, Roy W. Dobbins : "Neural Network PC Tolls", Academic Press, Inc., 98-102, (1990)

13. 한국 전자 통신 연구소, "데이워 연속 음성 인식을 위한 음소 인식 개발", 과학 기술처, 82-88, 104-115, (1990)