Proceedings of KFIS 2001 Spring Conference, 2001. 5.

자기 조직 신경망을 이용한 모음 인식

Recognize voiced vowel using self organizing map

장성환 · 강훈
중앙대학교 전자전기공학부

Sung-Hwan Jang and Hoon Kang
School of Electrical and Electronic Engineering, Chung-Ang University
E-mail : bigloud@sirius.cau.ac.kr

Abstract

본 논문은 Self Organizing Map을 이용한 한국어의 모음 10개를 인식하는 것을 다루고 있다. 분류기로서 우수한 성능을 보이고 있는 Self Organizing Map의 출력 층을 2차원으로 구성하여 짧은 시간 간격으로 주파수 도메인에서 벡터화 되어진 음성을 입력 층에 인가하여 유사한 출력 층의 분포를 이용하여 모음 10개를 인식하는 분류기로서의 가능성을 보여줄 것이다.

Key Words : 음성인식, Self Organizing Map, Neural network.

1. 서 론

Self Organizing Map 을 이용하여 한국어의 모음 중 '아', '야', '어', '여', '오', '요', '우', '유', '으', '이'를 인식할 수 있게 하는 것이 이 논문의 주제이다. 사람의 음성을 인식하는 시스템들이 많이 나오고 있으며 간단한 장난감에서부터 인터넷이 이르기까지 점점 그 사용빈도가 높아지고 있으며 앞으로도 이러한 사용은 더욱 증가할 것으로 보인다. 사람의 말을 패턴으로 본다면 아주 다양한 형태를 가지고 있는 것으로 볼 수 있다. 이 다양한 패턴 때문에 음성에 대한 연구는 계속되고 있고 그 음성을 인식하여 응용되는 분야 또한 다양하다. 인간의 가장 큰 특징 중에 다양한 소리를 인식하여 다양한 의사 소통을 하는 것은 인간이 사용하는 도구에 바로 이러한 기능을 이식하려는 기술을 만들게 되는 것이다.. 이에 먼저 병렬 처리로 속도를 빠르게 할 수 있는 신경망 중에 kohonen 이 제시한 self organizing map을 분류기로 사용하였다. 단 2개의 입력 층과 출력 층을 가지고 있는 self organizing map 은 간단한 학습 알고리즘과 간단한 구조를 가지고 있으면서도 분류기로서 우수한 성능을 보여주고 있다. 출력 층 뉴런들을 2차원 형태로 구성하고 학습을 통한 이웃 뉴런들과의 상호 작용으로 유사한 출력 층 사이에서는 모이는 현상과 유사하지 않은 출력 층 사이에는 멀어지는 형태를 이루게 된다. 이러한 형태로 구성되어지면 유사한 입력 패턴에 대해서는 출력 층의 뉴런의 분포를 2차원으로 보았을 때 어떤 특정한 지역에서의 출력 층의 분포가 나타나게 되는 것이다.

또한 이 논문에서 사용한 음성 분석 방법으로 일반적으로 음성 신호 처리에 많이 사용되는 방법인 short time fourier transform을 사용하여 시간 축에서 다양한 길이를 가지고 있는 음성 신호를 적절한 길이로 샘플링 하여 음성 신호의 주파수적인 특징을 추출하였으며 통상적으로 받아들여지는 모음 주파수의 구별에 지배적인 신호의 성분 주파수 영역만을 벡터로 만들었다. 사람의 음성은 시간 축에서 보면 우리가 사용하고 있는 라디오 주파수 대역에 비하면 아주 주파수가 낮은 신호에 해당한다. 이것은 상대적으로 느린 변화를 보이고 있는 신호로 받아들여지며 그러한 이유로 시간 축에서 일정한 신호의 길이 만큼을 short time fourier transform을 하고 일정 음성 신호의 길이를 겹쳐서 벡터로 만들어서 초기 음성 신호의 시작

감사의 글 : 본 연구는 한국과학재단 목적기초연구 사업(2000-2-30300-003-3)에 의해 일부 지원 받았습니다.

부분 설정의 상대적 error를 줄이고 음성 주파수의 한 음성에 대한 특징을 더 분명하게 나타내도록 하였다.

2. 본 론

2.1 preprocessing

모음 중 10개 즉 '아', '야', '어', '여', '오', '요', '우', '유', '으', '이'를 구별할 모음으로 선정하였다. 모음은 유성음에 들어가고 유성음은 주파수 축에서 강한 파워의 주파수 성분으로 구별할 수 있게 나타난다. 이 특징을 이용하여 시간 축에서 PCM 방식으로 8KHz의 샘플링 주파수와 모노방식을 사용하며 샘플링한 값은 8bit 값으로 저장을 하였다. 3초간의 녹음 시간을 이용하여 위의 모음을 한 개 씩 녹음하였으며 녹음되는 신호에서 있어서 음성의 초기 시작 부분은 시간 축에서 power의 레벨을 설정하여 그 설정 값 보다 큰 부분을 시작으로 생각하고 끝나는 부분 또한 단순한 파워의 비교를 통하여 얻어내었다. 3초간의 녹음 시간에서 음성이 차지하는 부분은 약 0.5초~0.7초 정도 되며 이 시간축의 신호를 short time fourier transform을 취한다.

2.2 short time fourier transform

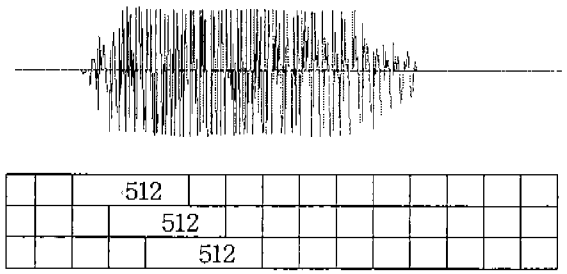

그림 1. 시간 축에서의 연산 방법

일반적인 fourier transform은 상대적으로 시간 축에서 느리게 변하는 음성 신호를 해석하는 방법으로는 적당하지 못하다. 그래서 사용하는 방법이 연속적으로 들어오는 신호를 시간 축에서 디지털화 시킨 후에 512 개의 시간 축 점을 FFT한다. 이 FFT한 신호의 한 점간의 주파수 차이는 8KHz/512이므로 15.625 Hz에 대응된다. 그리고 시간적으로 느리게 변한다는 것을 이용하기 위해 시간 축에서 384포인트 즉 128 포인트 쉬프트 한 신호의 형태로 FFT를 취하므로 주파수를 더 정확하게 해석할 수 있도록 한다. 그림 1과 수식 1이 그 예이다. 이렇게 얻어진 주파수 값을 파워를 구하여 이 파워 값들을 가지고 위에 제시한 모음 10개를 구별하도록 한다. 여기서 유성음중 모음의 주파수 특징을 이용하여 15.625 * 8 = 125 의 주파수부터 15.625 * 78 = 1218.75 부분만 데이터로 사용하여 저주파 노이즈와 상대적 고주파 부분은 데이터에서 제거한다. 이 부분은 모음에서 큰 파워를 가지고 있는 부분이 아니기 때문에 제거하기로 했다. 음의 길이에 따라 FFT를 한 값의 길이도 다르게 된다.

$$X_n\left(e^{jw}\right)=\sum_{m=-\infty}^{m=\infty} w(n-m)x(m)e^{-jwm} \qquad (1)$$

$$w(n)=1 \quad 0\le n\le N-1$$
$$=0 \quad otherwise$$

2.3 Self Organizing Map

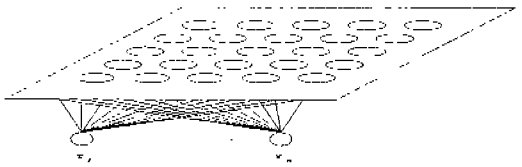

그림 2. Self Organizing Map

self organizing map은 kohonen에 의해서 제안 됐으며 2층 구조 즉 입력 층과 출력 층을 사용하고 있다. 입력 층은 입력하려는 패턴과 같은 차원으로 만들고 출력 층은 우리가 원하는 차원으로 만들 수 있다. 또한 입력 층의 모든 노드는 출력 층의 모든 노드에 연결된다. 이웃하는 뉴런들과 경쟁을 통하여 학습이 이루어지고 자기의 웨이트를 점진적으로 고쳐나감으로써 다른 패턴들을 구별한다. 또한 비감독자 학습의 형태를 가지고 있다. 비감독자 학습 형태의 신경망 구조는 우리가 분류하고 싶은 형태를 학습을 통해서 출력 값을 모르는 상태에서 얻어 낼 수 있는 방식이다. 또한 Self Organizing Map은 2차원이상의 벡터를 2차원으로 대응시킬 수 있는 특징을 가지고 있어서 다차원 vector의 분류에 강점을 가지고 있으며 병렬 처리가 가능하므로 분류하고자 하는 수만큼 뉴런의 수를 증가시키는 방식으로 구현하면 비교하는 방식의 데이터 양이 증가하더라도 데이터를 처리하는 부분에서의 시간적 증가는 피

할 수 있다.

$$m_c(t+1) = m_c(t) + \alpha(t)[x(t) - m_c(t)]$$
$$m_i(t+1) = m_i(t) \qquad for\ i \neq c \qquad (2)$$

알고리즘: Self organizing map

출력 층의 뉴런 수 : N

i 번째 출력 층의 웨이트 벡터 : Wi

웨이트 벡터 Wi 와 입력 벡터 Vj 사이의 거리 : d(Vi , Wj) 거리는 Euclidean distance를 말한다.

이웃하는 뉴런과의 반지름 NE

Step 1: 초기화. 모든 웨이트 벡터를 초기화 해준다. Wj , j= 1,2,·.N 를 0에서 1사이 값으로 랜덤하게 초기화시킨다. 총 학습 시간 단계 T를 설정한다. 학습 계수 $\alpha(0)$, $0<\alpha(0)<1$를 초기화한다. α는 시간이 증가하면서 총 학습 시간 T까지 선형적으로 줄어든다. 이웃하는 노드의 반지름 NE(0)을 초기화한다. NE는 시간이 증가하면서 총 학습 시간 T까지 선형적으로 줄어든다. 시간 단계를 t=0으로 초기화한다.

Step 2: 패턴을 입력한다. 처음 입력 패턴 Vi, i = 1 를 입력을 시작으로 i를 1 증가시킨다.

Step 3: 거리를 계산한다. 거리 d(Vi,Wj)를 모든 j=1,2,·,N에 대하여 계산한다.

Step 4: 출력 뉴런을 선택한다. 출력 뉴런 j'를 선택한다. 거리 d(Vi,Wj')는 step 3 에서 계산된 거리 중 가장 작은 값이다.

Step 5: 웨이트 벡터를 갱신한다. 뉴런 j'의 웨이트 벡터와 이웃하는 모든 뉴런 (NEj'(t) 에 속하는 모든 뉴런)의 웨이트 벡터에 $\alpha(t)(Vi-Wj)$ 를 더한다.

Step 6: 모든 Vi에 대해서 반복한다. 모든 입력 Vi가 입력 될 때까지 Step 2. 에서 Step 5 를 반복한다.

Step 7: 시간 단계를 증가하고 반복한다. 시간 t를 1 증가시킨다. 와 NE를 시간 t에 따라 감소시킨 값을 설정한다. t = T 가 되기 전까지 step 2. 에서 stcp 6. 까지 반복하고 멈춘다.

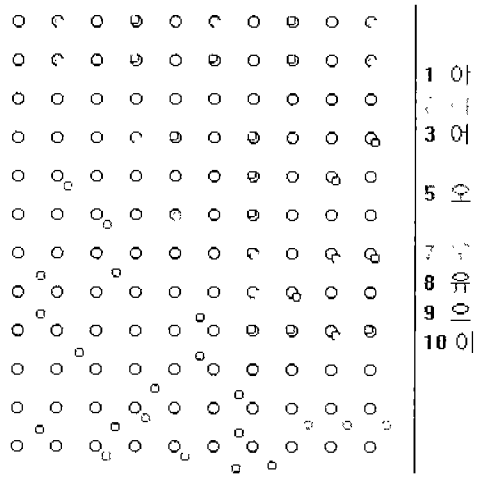

그림 3. 학습된 Self Organizing Map 1

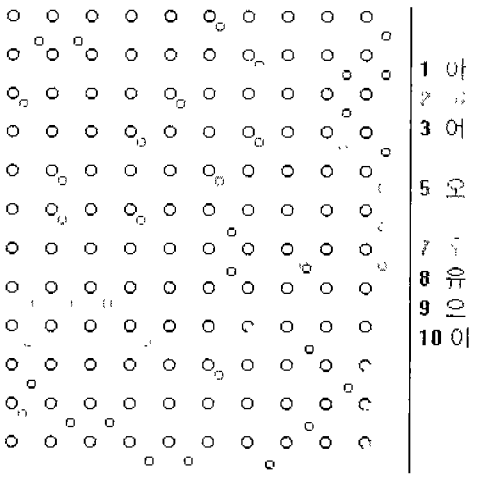

그림 4. 학습된 Self Organizing Map 2

2.4 final neural network

학습을 통하여 얻어진 self organizing map을 이용하여 입력된 모음을 구별하는 신경망으로 각각 음성에 대하여 승리하는 뉴런들을 마지막 10개의 뉴런에 연결한다. 웨이트는 전부 1로 설정한다. 이 마지막 신경망에서의 출력은 누적되는 형태를 가지고 있어서 출력된 값이 가장 큰 뉴런을 승리한 뉴런으로 한다.

3. 시뮬레이션 1

self organizing map 은 초기의 weight 들의 random 하게 설정하기 때문에 학습되는 형태가 다름을 위의 그림에서 알 수 있다. 위의 그림은 20*20개의 출력 노드 중에서 10*12개만 보여주고 있다. 아래의 표는 각 음성마다 승리하는 노드를 순서적으로 나열한 것이다. 밑줄은 같은 노드가 연속적으로 승리하는 것을 표

시하는 것으로 연속적으로 승리하는 노드는 각 음성마다 다르게 나오고 있다. 이것은 short time fourier transform을 사용하므로 얻어지는 효과로 더욱 겹치는 시간을 길게 하면 승리하는 노드의 연속되는 숫자가 더욱 많아지지만 연산의 회수가 증가하므로 서로 tradeoff 가 존재한다. 학습된 음성을 분류하는 데에는 아주 좋은 결과를 보여주고 있다. 한 사람의 음성 중에서 모음 중에서 10가지를 한 번 씩 발성한 것을 학습 시켜서 인지 이 음성으로 학습된 신경망에 그 학습 음성을 입력하면 월등하게 분류하는 결과를 보이고 있다. 하지만 학습시킨 음성이 아니면 같은 사람이 발성한 음성에 대해서는 좋은 결과를 보이고 있지 않다. 특히 '아' 와 '야', '어' 와 '여', '오' 와 '요', '우' 와 '유' 사이에서 서로 다른 것으로 인식하는 결과를 보이고 있다. 이것은 한 사람이 발성하더라도 같은 음을 발성하더라도 다른 음으로 즉 '야'의 경우 실제 발성 시 '야~아' 로 발음되기 때문에 위와 같은 경우 현재의 학습 음성의 개수로는 구별하기는 아주 힘들다.

4. 시뮬레이션 2

학습시키는 음성의 개수를 4배로 증가시키고 즉 '아' 부터 '으' 까지 각각 4개의 음성을 학습시켰다. 음성의 개수의 증가는 학습된 음성에 대하여 구별하는 능력을 올려주고 비 학습 음성에 대해서도 구별하는 능력을 향상을 보여주고 있다. 학습된 음성에 대하여 82.5%의 인식 결과를 보인다 인식에 실패하는 학습 음성을 보면 상대적으로 음성의 시간 축 길이가 짧은 음성에 대하여 인식에 실패하는 결과를 보여주고 있다.

표 1. 학습된 Self Organizing Map의 출력 뉴런의 순서와 패턴

음성	출력 뉴런 순서
아	(2 , 9) (0 , 9) (3 , 16) (2 , 16) (2 , 17) (0 , 19) (0, 19) (0, 19) (1 , 19) (2 , 19) (3 , 19) (6 , 19) (7 , 19) (8 , 19) (6 , 16) (5 , 16) (4 ,14) (4 , 14) (3 , 12) (2 , 11) (2 , 9) (2 , 0)
야	(0 , 2) (1 , 3) (4 , 4) (7 , 5) (16 ,10) (19 , 11) (19,12) (19,13) (0,16) (0,17) (0,17) (0,17) (3 ,19) (4 ,18) (5 ,18) (8 ,17) (8 , 15) (6 , 14) (5 , 12) (5 ,12) (2 ,11) (2 ,10) (2 , 0)
어	(0 ,10) (0 ,14) (13 ,19) (13 ,19) (19 ,19) (19 , 19) (18 ,19) (18 ,19) (17 ,19) (16 ,19) (15 ,19) (15 , 18) (15 ,17) (14 ,16) (12 ,17) (11 ,18) (10 ,19) (11 , 19) (2 ,14) (2 , 13) (1,12) (1 , 11)
여	(3 , 2) (5 , 4) (7,4) (15 , 9) (15 , 9) (18 ,12) (18 ,13) (19 ,14) (19 ,15) (19 ,16) (19 , 17) (17 , 17) (17 , 17) (16,15) (13 ,15) (12 ,16) (10 , 16) (10 , 16) (1 ,14) (0 ,13) (0,12) (0 ,9) (1 ,9) (2 , 0)

3. 결 론

음성 인식에 있어서 모음은 아주 중요한 인식의 기준이 되어 준다. 음성 모음 인식에 Self Organizing Map 분류기를 사용한 결과를 보면 우수한 결과를 보여주고 있다. 자음 보다 음성의 pitch 주파수가 뚜렷한 모음에 대하여 출력 뉴런들은 학습에 따라 유사 출력 뉴런들이 이웃하여 형성되고 다른 출력 뉴런들은 이웃해서 형성되지 않는 Self Organizing Map 의 특성 때문이다. 현재의 학습된 음성보다 수를 증가하고 음성 시간 축 길이를 더 자세히 분해해서 주파수 특성을 구하면 아주 좋은 분류기가 될 것이다. 시뮬레이션 결과에서 본 것처럼 현재의 음성 길이의 겹치는 시간 타임을 증가시키면 더욱 좋은 결과가 예상된다. 또한 Self Organizing Map 의 크기를 변경하여 모음 음성 분류에 가장 적당한 사이즈를 찾는 것도 차후 연구해야 하며, 또한 마지막 신경망에서 단순한 언결 형태가 아닌 감독자 학습 형태로 학습을 시킨 신경망 형태로 구성했을 때의 성능 또한 차후 연구되어야 할 부분이다.

참 고 문 헌

[1] L.R. Rabiner and R.W. Schafer "*Digital Processing of Speech Signals*" Prentice-Hall, 1978.

[2] Roman Kuc. "*Introduction to Digital Signal Processing*" Mcgraw-Hill, 1982.

[3] Deller and Hansen and Proakis , *"Discrete-Time Processing of Speech Signals"* IEEE Press, 2000

[4] L, Rabiner and B,H. Juang, *"Fundamentals of Speech Recognition"* Prentice-Hall, 1993

[5] S. Haykin *"Neural Networks A Comprehensive Foundation"* Prentice-Hall, 1999

[6] Ritter and Martinetz and Schulten *"Neural Computation and Self-Organizing Maps"* Addison-Wesley, 1992