개선된 경쟁학습을 이용한 음성인식

송 준 규°, 이 동 욱, 김 영 태
동국대학교 전기공학과

A Study on the Speech Recognition using Advanced Competitive Learning

Joon-Gyu Song°, Dong-Wook Lee, Young T. Kim
Department of Electrical Engineering, Dongguk University

Abstract

This paper presents the speaker-dependent Korean isolated digit recognition system using advanced competitive learning. Since competitive learning algorithms are easy and simple to implement, they are used in various fields.

The proposed recognition algorithm consists of three procedures: comparing winning number of codebook vectors, selecting the representative vector out of codebook vectors, and generating a new codebook with the representative vectors.

In this paper, we use a sound blaster 16 for obtaining speech data. Speech data are sampled by 16 bits and 11 kHz sampling rate.

제 1 장 서 론

경쟁 학습 알고리즘은 입력 벡터들을 신경 회로망에 계속적으로 입력시키면서 자율적으로 연결가중치들을 변경시키는 방법으로, 알고리즘이 단순하고 하드웨어 구현시 그 구조가 간단하여 널리 이용되고 있다. 경쟁 학습은 본래 통계학에서 데이터 분류를 목적으로 출발한 K-means 군집화 알고리즘을 근본으로 한다. K-means 알고리즘은 주어진 데이터들을 k개의 클래스로 원하는 오차 수준이하로 구분될때까지 반복 수행함으로써 패턴을 분류하는 방법이다.[5]

단순 경쟁학습에서는 효율적인 군집화가 필수적인데, 초기 가중치 벡터들의 분포에 따라, 전혀 학습이 이루어지지 않는 출력 뉴런들이 생기는 문제(dead neuron problem)와 인식시 잘못 학습된 뉴런과 학습되지않은 뉴런에 의한 오인식 등의 문제가 발생할 수 있다. 이러한 문제들을 해결하기 위한 여러 방법들이 논의 되었는데, 그 중 많이 사용되는 방법으로는 코호넨 학습과 Frequency Sensitivity Competitive Learning (FSCL)등이 있다.

본 논문에서는 각 클래스에 속하는 벡터들의 학습 횟수를 이용하여 각 클래스를 대표하는 대표 벡터를 선정, 위에서 언급한 문제들을 해결하는 방법을 제시하였다. 또한 한정된 고립어로부터 얻어진 LPC 계수를 이용하여 음성 특징을 추출하고 LVQ와 개선된 경쟁학습을 이용하여 결과를 출력하는 방법을 사용하였다.

음성의 입력을 위해 사운드 카드를 이용하였으며, 음성 데이터들은 16 bit, 11 kHz로 샘플링하였다.

제 2 장 음성 분석

본 장에서는 음성 인식에 앞서 알아야 할 음성의 기본적인 지식과 음성 신호의 전처리 과정, 음성 파라미터 추출 방법을 논하게 될 것이다. 음성 파라미터 추출 과정은 사운드 카드를 이용한 음성신호의 입력, 음성 신호의 끝점 추출, 윈도우, LPC분석으로 이어진다.

2.1 음성 신호의 특징

음성은 발성자의 대뇌에서 만들어진 언어 메시지에 따라 폐에서 시작된 공기의 흐름이 성대 또는 구강내의 이, 입술에 의해 형성된 좁은 공간의 부위에 각각 주기적 또는 잡음성의 공기 진동을 일으켜 최종적으로 음성 파형으로 생성된다. 그림 2.1에서 보듯이 유성음은 주기적인 성문 펄스열이 나타나지만 무성음의 경우에는 이러한 주기열 대신에 랜덤 잡음이 나타난다. 이렇게 여기된 유성음 또는 무성음 신호는 성도관을 통해 성도관 파라미터에 의한 공진을 일으키는 과정을 겪게 되고 다시 입술 부근에서 방사하게 된다.[1] 따라서 음성 신호의 주기적 성분의 유무에 따라 유성음과 무성음을 구별하게 되며 성도관의 공진 주파수의 분포에 따라 음소를 구별할 수 있다.

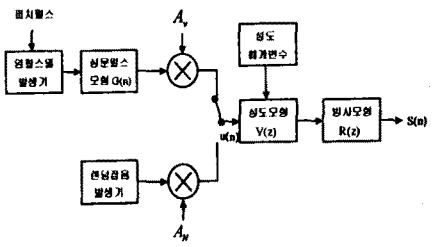

그림 2.1 음성 생성 모델의 블록도

음성 신호는 이웃한 샘플들간에 강한 상관관계를 가지며, 이를 음성 신호의 압축이나 인식에서 사용한다. 음성 신호는 대부분의 에너지가 4 kHz 이내에 존재하고, 20 ~ 30 msec 내에서 통계적 분포가 일정한 특성을 가지고 있다. 따라서 이 구간으로 음성 신호를 나눈 뒤 각 구간 음성 신호에 대해서 선형 시불변 신호로 간주해 해석하게 된다.

2.2 음성 파라미터 추출

음성 신호의 분석은 음성 파형으로부터 인식에 유용한 특징 파라미터를 추출하기 위한 전처리 과정이다. 이 방법으로는 시간 영역 분석인 에너지, 영교차율, 단시간 자기상관 분석 등이 있고, 주파수 영역 분석 방법으로는 대역 필터 뱅크에 의한 분석, 포만트 주파수 분석, 퓨리에 변환, 캡스트럼 분석이 있으며, 선형 예측 분석 방법으로는 선형 예측계수, 반사 계수, 선형 예측 스펙트럼 등이 있다.

본 논문에서 사용하는 음성 파라미터 추출 방법은 다음과 같다.

2.2.1 단구간 에너지와 영교차율

본 논문에서는 가장 효율적이라 알려진 단구간 에너지(short-time energy)와 영교차율 (Zero-Crossing Rate, ZCR)을 사용하였다.

음성 신호의 에너지는 시간에 따라 변화한다. 특히 무성음의 경우는 유성음보다 낮은 크기의 에너지를 갖는다. 단구간 에너지는 이러한 변화량을 쉽게 표현하는 음성 파라미터이다. 보통 다음과 같이 정의한다.

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)\, w(n-m)]^2 = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m) \quad (2.1)$$

$$h(n) = w^2(n) \quad (2.2)$$

그러나 본 논문에서는 에너지의 민감성을 감안하여 식 (2.1)의 x^2 대신 절대값을 사용하였다.

h(n)은 window함수로써 보통 Hamming window를 사용한다. Hamming widow는 음성신호의 처리단위를 정해주며 passband 이상의 주파수 감쇄 특성이 좋아서 음성신호 처리에 많이 사용된다.

영교차란 두 개의 연속되는 샘플의 부호가 다를 때를 의미하며, 영교차율은 음성 신호의 가장 간단한 주파수적 요소이다. 단구간 영교차율은 다음 식과 같다.[1][2]

$$ZCR(i) = \sum_{n=0}^{N-1} |sgn[x(i,n)] - sgn[x(i,n-1)]| \quad (2.3)$$

$$sgn[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \quad (2.4)$$

2.2.2 배경잡음 추출

영교차율과 에너지를 이용한 끝점 추출을 위해 필요한 정규화 과정의 하나로서 주변환경의 잡음정도를 측정하여, 음성을 검출할 때 기준이 되는 임계값을 설정한다. 일정 시간동안 주변 잡음을 샘플링하여 일정수의 샘플을 한 분석구간으로 정하고 각 분석구간의 에너지와 영교차율, 그리고 평균[M]을 구한다.

$$M = \frac{1}{N} \sum_{n=0}^{N-1} x_n \quad (2.5)$$

여기서 M은 오디오 장치의 DC 성분을 제거하는데 사용된다. n은 분석구간의 index이고, N은 분석구간을 나타낸다. 배경잡음의 임계값은 일정 시간동안 배경잡음의 평균 에너지와 평균 영교차율을 사용하여 구한다.

2.2.3 선형 예측 계수

LPC 분석은 사람이 음성을 발생시킬 때 각각의 발음이 성대와 성도를 거쳐 입술에 이르는 과정을 전극필터로 모델링하여 특징 파라미터를 얻는 방법이다. 이는 음성 신호의 스펙트럼을 필터로 모델링할 때, 영점을 갖지 않는다는 가정하에서 음성 발생 모델을 전극 필터로 근사화 하여 분석하는 방법이다. 이러한 모델을 autoregressive 모델이라 하며, 전달 함수는 다음과 같이 표현된다.[1]

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z} \quad (2.6)$$

여기서 G는 이득계수이고, a_i는 예측계수로서 성도특성을 나타내는 계수가 된다. 예측 계수 a_i는 음성 신호의 현재값과 그 음성 신호의 과거값으로부터 추정하여 얻는다.

이와 같이 예측하는 선형 예측 방법은 계산 속도가 빠르면서도 전달 함수를 정확하게 추정하기 때문에 많은 분야에 사용되고 있다.

제 3 장 신경 회로망

신경 회로망이란 상호 연결된 많은 수의 인공 뉴런들을 이용하여 생물학적인 시스템의 계산 능력을 모방하는 소프트웨어나 하드웨어로 구현된 계산 모델을 말한다.[3][7]

3.1 경쟁 학습

경쟁 학습에서 각 뉴런은 연결강도 벡터와 입력 벡터가 얼마나 가까운가를 계산한다. 그리고 각 뉴런들은 학습할 수 있는 권리를 부여 받으려 서로 경쟁하는데 거리가 가장 가까운 뉴런이 승리하게 된다. 이 승자 뉴런만이 제시된 입력 벡터에 대하여 학습이 허용된다. 이 승자 뉴런이 출력 신호를 보낼수 있는 유일한 뉴런이다.

경쟁 학습의 목적은 입력 패턴들을 클래스별로 분류하는 것이다. 비슷한 입력들은 같은 클래스, 즉 같은 출력 뉴런에 속하게 된다. 이들 클래스는 신경회로망에 의해서 자율적으로 발견된다.

경쟁 학습 신경회로망은 한 개의 입력층과 출력층으로 구성된다. 그림 3.2는 가장 단순한 경쟁 학습 신경회로망의 구조를 보여준다.

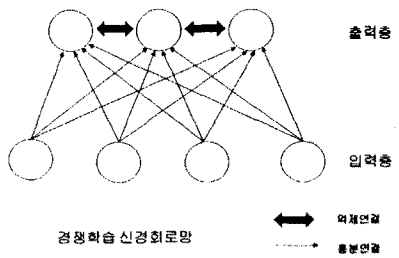

그림 3.2 경쟁학습 신경회로망

경쟁학습의 승자 뉴런 선정 기준과 대표적인 알고리즘은 다음과 같다. 일반적으로 경쟁 학습에서는 승자뉴런의 선정 기준으로 입력 벡터와 출력뉴런의 가중치 벡터간의 거리를 사용한다. 이를 수식으로 표현하면 식 (3.1)과 같다.

$$d[x, w_i(t)] = \|x - w_i(t)\| \quad (3.1)$$

이때 x는 입력벡터, w_i는 출력뉴런의 가중치 벡터, $\|\cdot\|$ 는 Euclidean norm을 나타낸다. 식 (3.1)의 출력뉴런들 중에서 최소거리를 가지는 뉴런을 승자뉴런으로 선정한다.

3.2 코호넨 학습

코호넨의 학습 알고리즘은 매우 간단하다. 오류 역전파 학습처럼 체층적인 시스템이 아니라 단지 2개의 층으로 되어있다. 첫 번째 층은 입력층이고 두 번째 층은 경쟁층이다. 모든 연결은 첫 번째 층에서 두 번째 층의 방향으로 되어있으면 두 번째 층은 완전 연결되어있다. 이 뉴런들은 경쟁층에서 고밀도로 연결되어있다.[7] 앞 절에서 설명한 바와 같이 승자 뉴런을 결정하고 난 후에는 코호넨의 학습 규칙에 따라 뉴런의 연결 강도를 조정해야 한다. 이 규칙은 다음과 같이 표현된다.

$$W_n = W_{old} + \alpha(X - W_{old}) \quad (3.2)$$

여기서 W_{old}는 조정되기 이전의 연결강도 벡터이며, W_n는 조정된 후의 새로운 연결강도 벡터이고, X는 입력패턴 벡터이며, α는 학습상수이다.

앞서 기술한 바와 같이 승자 연결강도 벡터는 기하학적으로 입력패턴에 가장 가깝다. 코호넨의 학습은 단순히 연결강도 벡터와 입력패턴 벡터의 차이를 구한 다음 그것의 일정한 비율을 원래의 연결강도에 더하는 것이다. 이때 승자 뉴런만이 그것과 관련된 연결강도 벡터를 조정하는 것이 아니라 그의 이웃 반경안에 드는 모든 뉴런들도 유사한 조정을 하게 된다. 승자 뉴런은 +1을 출력으로 내며, 승자 뉴런과 그것의 이웃 뉴런들은 각자의 연결강도 벡터를 입력벡터에 다소나마 가까이 접근하게 된다.[6]

제 4 장 신경회로망을 이용한 음성인식

4.1 Learning Vector Quantization(LVQ)

LVQ는 코호넨에 의해 제안된 신경망 모델이다. LVQ는 자기 구성 특징 지도와는 달리 학습을 통해 분류된 패턴들이 어떤 클래스에 속하도록 설계되어 있다. 그림 4.1에서 보듯이 LVQ 네트워크 구조는 출력층에 있는 각 뉴런이 몇 개의 클래스에 속해 있다는 것을 제외하고는 자기 구성 특징 지도와 유사하다.

LVQ는 학습과정에서 벡터 양자화를 수행하도록 되어있다. LVQ에서 가중치 벡터를 레퍼런스 또는 코드북 벡터라고 하는데 입력패턴을 올바른 클래스로 효과적으로 분류하기 위해서 잘 만들어진 코드북이 필수적이다.

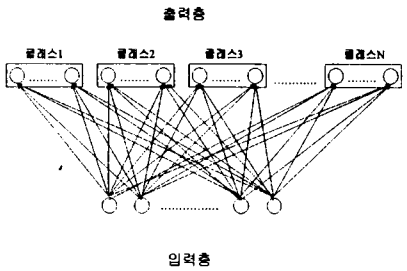

그림 4.1 LVQ의 구조

4.2 제안된 경쟁학습 알고리즘

본 논문에서 제안한 방법은 코드북의 각 벡터의 학습된 횟수를 이용하여 새로운 코드북을 만들어 인식에 사용하는 알고리즘이다.

먼저 식 (3.1)에 의해 승자 뉴런을 선정한다. 승자 뉴런과 가중치벡터의 클래스를 비교하여 식 (4.1)과 식 (4.2)의 학습규칙에 따라 뉴런의 연결강도를 조정하게 된다. 이때 입력패턴 벡터와 가중치 벡터의 클래스가 같을 경우 승자횟수를 증가시켜 승자횟수 비교에 의한 대표벡터 선정에 이용한다.

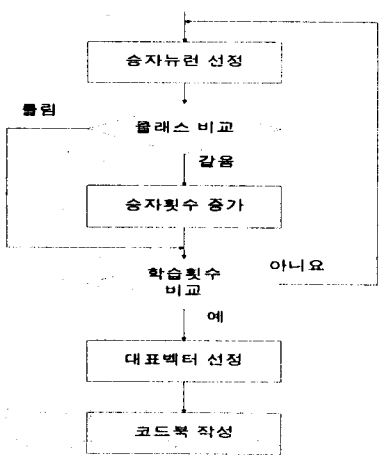

그림 4.2 제안된 알고리즘의 흐름도

$$W_c(t+1) = W_c(t) + \alpha(t)[x(t) - W_c(t)] \quad (4.1)$$

x(t)가 바르게 분류된 경우

$$W_c(t+1) = W_c(t) - \alpha(t)[x(t) - W_c(t)] \quad (4.2)$$

x(t)가 틀리게 분류된 경우

각 클래스별로 가장 많이 올바르게 학습된 가중치 벡터가 학습에 충실했다고 판단, 대표 벡터로 선정하여 인식에 사용하였다.

학습시 잘못 분류된, 또는 학습되지 않은 코드북 벡터에 의해 생길 수 있는 오인식을 방지할 수 있다.

그림 4.2는 제안된 알고리즘의 대략적인 흐름도이다.

제 5 장 모의 실험 및 결과

대표 벡터를 이용한 한국어 숫자음 인식을 위해 사운드 카드를 탑재한 개인용 컴퓨터에서 실험을 하였다. 음성의 입력은 사운드 카드를 통해 16 bit, 11 kHz로 샘플링 되었고, 끝점추출 알고리즘에 의해 음성 부분만 추출해 내었다. 코드북을 만들기 위한 정규화 과정으로 해밍 윈도우를 씌우는 동시에 음성 길이에 따라서 능동적으로 겹치는 부분을 조정하는 방법을 사용하였다.

코드북을 위한 음성특징 벡터로서 LPC 계수를 사용하였다. 코드북에 쓰인 음성 데이터는 0 부터 9 까지 10개씩, 총 100개이고, 테스트를 위해서 다시 100개의 데이터를 사용하였다.

그림 5.1은 제안한 알고리즘과 LVQ를 비교한 것으로 0부터 9까지의 인식률을 보여주고 있다. 전체적인 인식률은 각각 83%와 79%으로 제안된 알고리즘의 성능이 향상된 걸 알수 있다.

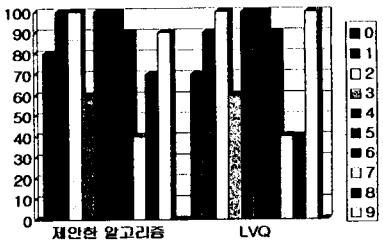

그림 5.1 인식률 비교

Reference

[1] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signal," *Prentice Hall*, pp. 116~135, 396~413, 1978.

[2] L. R. Rabiner and B. H. Jwang, "Fundamentals of Speech Recognition," *Prentice Hall*, pp. 97~132, 1993.

[3] Laurene Fausett, "Fundamentals of Neural Networks," *Prentice Hall*, pp. 169~195, 1994.

[4] Paolo Antognetti and Veljko Milutinovic "Nueral Networks," vol. 3 *Prentice Hall*, pp. 124~143, 1991.

[5] 김종완, "동적으로 출력뉴런이 생성되는 병렬 경쟁학습 신경회로망을 이용한 패턴인식," 서울대학교 박사학위 논문, 1994.

[6] T. Kohonen, "The Self-Organizing Map," Proceedings of the IEEE, vol. 78, no. 9, pp. 1464~1480, 1990.

[7] 김대수, "신경망 이론과 응용(I)," 하이테크정보, pp. 169~189, 1992.