신경망을 이용한 단어에서 모음추출에 관한 연구

이택준, 김윤중
한밭대학교

A study on the vowel extraction from the word using the neural network

Taek-Jun Lee*, Yoon-joong Kim**

Abstract

This study designed and implemented a system to extract of vowel from a word. The system is comprised of a voice feature extraction module and a neutral network module.

The voice feature extraction module use a LPC(Linear Prediction Coefficient) model to extract a voice feature from a word. The neutral network module is comprised of a learning module and voice recognition module. The learning module sets up a learning pattern and builds up a neutral network to learn. Using the information of a learned neutral network, a voice recognition module extracts a vowel from a word.

A neutral network was made to learn selected vowels(a, eo, o, e, i) to test the performance of a implemented vowel extraction recognition machine. Through this experiment, could confirm that speech recognition module extract of vowel from 4 words.

1.서론

오늘날의 사회가 멀티미디어 사회로 급격하게 전환이 되고 있다. 멀티미디어 사회는 양방향의 의사소통이 가능하기 때문에 인간과 기계사이의 의사소통이 중요시 되고 있다. 인간과 기계사이 보편적으로 사용되는 단말기는 물리적인 힘을 통해 제어한다. 이러한 물리적 힘이 없을 경우 제어를 할 수 없는 단점이 있다. 이와 같은 단점을 해결하기 위해 인간의 음성을 이용한 인터페이스 연구가 진행 중이다.

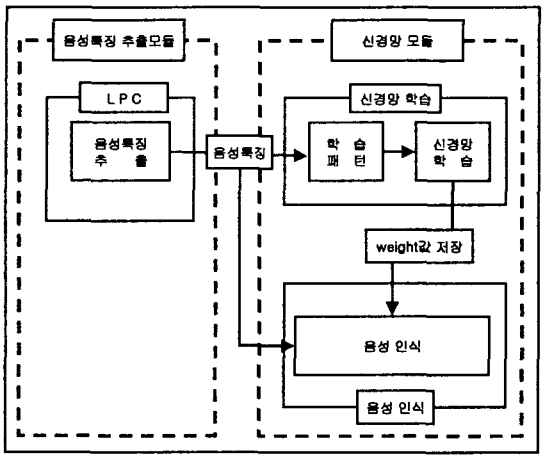
진행중인 연구들 중에 '역전파 학습 신경망을 이용한 고립단어 인식시스템에 관한 연구'[1]와 '신경망에 의한 초성자음(ㄱ, ㄷ , ㅂ)을 이용한 음소단위의 음성인식'[2] 등이 있다.

위의 첫 번째 연구는 신경망을 사용하여 독립단어를 인식하는 인식시스템이다. 이 시스템은 학습되지 않은 음성신호가 입력되면 음성을 인식하지 못한다. 두 번째 논문에서는 신경

망을 구성하여 음성데이터를 학습 시켰지만, 음소의 위치가 초성에 제한되어 있다.

본 연구에서는 상기의 문제점을 보완하기 위해 학습되지 않은 음성데이터를 음소단위로 인식하고, 중성이나 종성에 위치하는 모음을 추출하는 시스템을 구축하고자 한다. 또한 신경망 모듈을 라이브러리형태로 구성하여 다른 시스템에서 호출될 수 있는 장점을 제공한다.

본 연구에서 구현한 시스템은 (그림 1)과 같은 구성으로 되어 있다. 본 연구에서 구현한 시스템은 음성특징 추출모듈과 신경망 모듈부분으로 구분할 수 있다. 음성특징 추출모듈은 음성신호가 입력되면 LPC(Linear Prediction Analysis)를 통해 음성의 특징을 추출하는 부분이다. 신경망 모듈은 신경망 학습부분과 음성인식부분으로 구성되어 있다. 신경망 학습부분은 신경망을 구성한 후 음성특징 추출모듈에서 음성특징을 받아 신경망에 학습을 하는 과정이다. 음성인식부분은 신경망 학습을 통해 얻어진 가중치를 저장한 후, 음성특징 추출모듈의 음성특징을 받아 음성을 인식하는 부분이다.



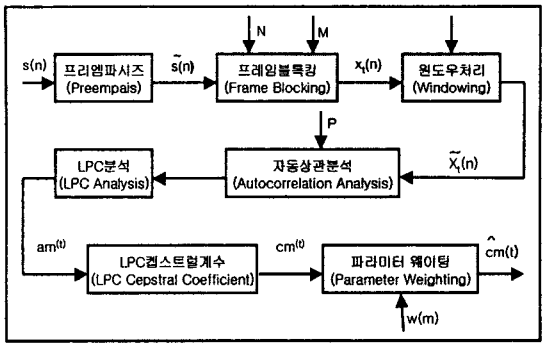
(그림 1) 시스템 구성도

본 연구의 2장에서는 음성특징 추출모듈로 음성의 특징을 추출하는 과정에 대해 설명하고 3장에서는 본 연구에서 사용된 신경망모듈내의 신경망의 구조에 대해 설명한다. 4장에서는 실험결과를 기술한다. 마지막장으로 5장에는 본 논문의 결론과 향후 추가로 연구될 과제를 제시한다.

2.음성특징추출

본 연구에서는 음성인식에 필요한 음성특징을 얻기 위해 LPC(Linear Prediction Analysis)를 사용하였다.[3-5] (그림 2)는 LPC를 통해 Cepstral 계수를 추출하는 과정이다.[5] 본 연구에서는 인식률을 고려하여 신경망 학습에 사용되는 음성 신호에서 처음과 끝부분에 포함된 묵음과 잡음을 음성 편집기를 이용하여 제거하였다.

(그림 2)에서 사용되는 음성데이터는 11 Khz로 샘플링된 16bit의 mono를 사용한다.



(그림 2) LPC Cepstral계수 추출 과정

프리엠파시스 과정은 전송선로나 마이크의 성능, 배경 잡음 등의 영향으로 인한 신호 왜곡을 보정한다. 프레임 블록킹 과정은 프레임을 블록화한는 단계이다. 이때 프레임의 크기는 300, 프레임 시프트의 크기는 100으로 설정한다.

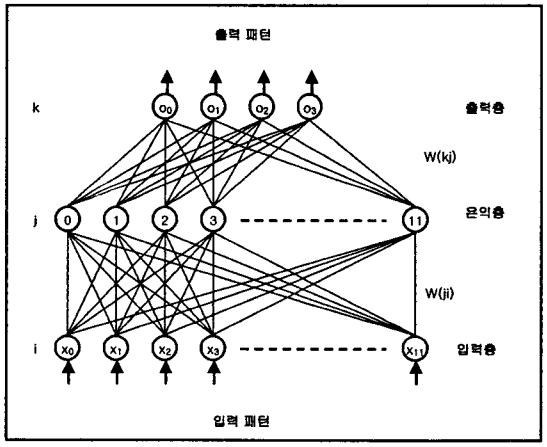
원도윙 과정은 각 프레임들의 시작과 끝에서 에러 신호를 최소화하기 위해 해밍원도우(Hamming Window)함수를 사용한다.

자동상관분석은 원도윙 과정을 거친 신호를 각 프레임별로 자동상관분석을 수행한다. 이때 LPC분석차수 P는 보통 8에서 16사이의 값을 갖는다. 본 연구에서는 10으로 설정하였다. LPC 분석은 자동상관분석 과정에서 적용된 수식을 Durbin방법[6]을 이용하여 분석을 수행한다. LPC 켑스트럴 계수는 LPC 계수에서 12차 로 산출한다.

파라미터 웨이팅과정은 차수 중 낮은 차수는 전체 스펙트럼기울기에 민감하고 높은 차수는 소음에 민감하기 때문에 이러한 민감성을 감소시키기 켑스트럴 계수들을 웨이팅 해주어야 한다.

3.신경망

본 연구에서 사용된 역전파 신경망은 (그림 3)의 구조를 지닌다.



(그림 3) 신경망의 구조

신경망에 사용되는 음성신호는 2장의 과정을 거치면 12차의 음성데이터의 형태로 변환이 된다. 따라서 입력층의 입력 패턴의 수는 12로 설정을 한다. 은익층의 층수는 1개로 하고 유닛의 수는 12로 설정을 한다. 출력층의 출력 패턴은 4로 설정한다.

신경망의 학습과정[7]에서 연결강도의 변화 Δw_{ji}에 영향을 주는 파라미터로 모멘텀(momentum term)과 학습률(learning rate)이 있다.

식 1의 η은 학습률이고, α은 모멘텀이다. 학습률을 통해 연결 가중치를 조절하고, 모멘텀을 통해 수렴속도를 조절한다.

$\Delta w_{ji}(n+1) = \eta(\delta_j o_i) + \alpha \Delta w_{ji}(n)$ 식 1

4. 실험 및 결과

4.1 실험 환경

본 연구에서는 사용된 운영체제는 윈도우 XP이다. 음성데이터는 11Khz로 샘플링된 16bit의 mono를 사용한다.

4.2 학습데이터

학습데이터는 사무실에서 남성화자 2명이 모음(아, 어, 오, 에, 이)5개를 10회씩 발음한다. 채취된 학습데이터는 2장의 과정을 거쳐 12차의 벡터를 생성한다.

<표 1> 학습데이터의 화자당 발음횟수 및 벡터수

모음	화자 1		화자 2		모음별 벡터합
	발음 횟수	벡터수	발음 횟수	벡터수	
아	10	348	10	344	702
어	10	334	10	371	705
오	10	343	10	384	737
에	10	370	10	351	721
이	10	345	10	381	726
합계	50	1770	50	1,831	3,591

<표 1>에선 화자별로 발음한 모음의 횟수와 벡터수를 표시하고 있다. 그리고 모음과 벡터의 합을 나타내고 있다.

4.3 학습

학습을 위한 신경망 구조는 3장의 구조와 같은 입력층의 입력패턴은 12로 설정한다. 은익층의 층수는 1개로 하고 유닛의 수는 12로 설정한다. 출력층의 출력패턴은 4로 설정한다.

학습에 사용되는 자료 중 입력 패턴($x_0,x_1,x_2,x_3,,...,x_{11}$)은 4.2절의 데이터를 사용하고, 출력패턴(t_0,t_1,t_2,t_3)은 <표 2>과 같은 목표값을 설정하여 패턴에 사용한다. 이 두개의 패턴으로 하나의 학습패턴($x_0,x_1,x_2,x_3,...,x_{11},t_0,t_1,t_2,t_3$)을 생성한다.

<표 2> 신경망 학습시 모음의 목표값

모음	목표값
아	0 0 0 0
어	0 0 0 1
오	0 0 1 1
에	0 1 1 1
이	1 1 1 1

신경망의 학습에 사용되는 모멘텀과 학습률은 3장의 수식 1을 참조한다. 모멘텀은 0.9로 학습률은 0.5로 설정하고 학습 오차값 합계는 0.000001로 설정한다.

4.4 실험

4.4.1 학습데이터와 인식률

학습데이터와 인식률의 관계를 알아보기 위해 학습에 사용되는 모음의 수를 증가시키면서 인식률을 조사한다. 이때 인식률은 모음마다 학습에 이용된 음성 2개와 이용되지 않은 음성 2개를 사용하여 총 20개의 음성을 이용하여 측정한다.

<표 3> 모음 수에 따른 인식률의 변화

아	어	오	에	이	모음의 합계	벡터의 합계	인식률 (%)
2	2	2	2	2	10	341	10
4	4	4	4	4	20	692	25
8	8	8	8	8	40	1395	60
10	10	10	10	10	50	1740	65
12	12	12	12	12	60	2103	65
16	16	16	16	16	80	2875	65
20	20	20	20	20	100	3591	65

<표 3>은 인식률을 알아보기 위해 모음별 사용된 음성의 수와 그 음성에 대한 벡터수가 표시되어 있다. <표 3>을 보면 학습에 사용된 모음의 합계가 50일때 까지는 인식률이 증가한다. 그러나 50이상부터는 인식률이 더 이상 증가하지 않는다.

4.4.2 모음 판단

음성인식모듈을 통해 단어에서 모음을 추출한다. 이때 추출된 모음이 정확한지를 판단하기 위해 식. 2를 사용한다.

$$E = \sum_{i=0}^{3} min\{|o_i|, |1-o_i|\} \qquad \text{식 2}$$

식 2에서 E의 값은 0.0025를 기준으로 삼는다. 기준보다 값이 작으면 음성인식모듈에서 추출된 모음이 정확한 것으로 판단한다. 그러나 기준보다 값이 크면 잘못된 것으로 판단한다.

4.5 검토 및 결과

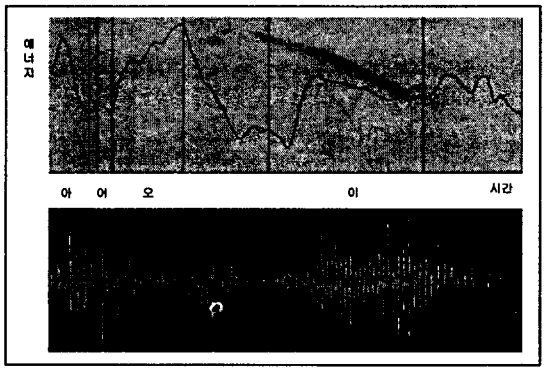
본 연구에서는 4.1.1절에서 인식률을 높이기 위해서 두 가지의 방법을 시도한다. 첫 번째는 인식률이 일정하게 유지하는 시점에서 학습데이터의 수를 다르게 증가시킨다. 인식률이 높은 모음은 학습데이터의 수를 줄이고 인식률이 낮은 모음은 학습데이터의 수를 늘린다. 두 번째는 사용되는 각 학습데이터의 처음과 끝부분을 1/3씩 잘라 음성에 포함된 불필요한 정보를 제거한다.

화자종속의 방법으로 단어(아버지, 세미나, 보고, 에너지)4개로 실험한 결과 "아버지"를 제외한 단어에서 정확하게 모음이 추출되었다.

<표 4> 모음추출 결과

단어	모음신호	추출된 모음신호
아버지	아, 어, 이	아, 어, 오, 이
세미나	에, 이, 아	에, 이, 아
보고	오, 오	오, 오
에너지	에, 어, 이	에, 어, 이

단어 "아버지"는 존재하지 않은 모음 "오"가 검출되었다.



(그림 4) 단어 "아버지"의 에너지량과 파형

이 같은 경우는 사용된 음성자료가 정확하게 발음되지 않았다고 해석 하였다.

5. 결론

본 연구에서는 단어에서 모음을 추출하기 위해 시스템을 설계하고 구현하였다. 이 시스템은 음성특징 추출모듈과 신경망 모듈로 나눌 수 있다.

음성특징 추출모듈에서는 음성의 특징을 LPC모델을 이용하여 12차 켑스트럴 계수를 출력한다. 이 켑스트럴 계수에는 음성을 인식할 때 필요한 음성의 정보가 담겨 있다.

신경망 모듈은 학습모듈과 음성인식모듈로 나눌 수 있다. 학습모듈은 음성의 특징이 담긴 켑스트럴 계수로 학습 패턴을 설정하고 신경망을 구성한 후 학습을 한다. 학습과정이 종료

되면 가중치를 저장한다. 음성인식모듈은 학습이 끝난 신경망을 통해 음성의 특징이 들어오면 학습과정에서 저장되어 있던 가중치로 음성의 특징을 분류하는 과정을 거친다. 앞의 과정을 수행한 후 신경망은 단어안에 들어있는 모음을 추출해 낸다.

구현된 모음추출 인식기의 성능을 알아보기 위해 모음(아, 어, 오, 에, 이)5개를 선택한다. 선택된 모음에서 음성특징을 추출한 후 신경망으로 학습시킨다. 학습이 종료된 후 단어(아버지, 세미나, 보고, 에너지)4개를 선정하여 음성인식모듈로 단어를 인식시킨다. 인식결과 단어 중 "아버지"를 제외한 3단어는 모음을 모두 찾아냈다. 단어"아버지"도 모음을 모두 찾았지만, 사용되지 않은 모음하나를 추출했다.

향후 연구에서는 현재 모음에 한정되어 있는 음성인식시스템을 확장하여 자음도 인식할 수 있는 음성인식시스템이 필요하다.

[참고문헌]

1. 김태정, "역전파 학습 신경망을 이용한 고립단어 인식시스템에 관한 연구", 한국 통신학회, 1990
2. 김석동, "신경망에 의한 초성자음(ㄱ, ㄷ, ㅂ)의 인식방법", 학국음악학회, 1991
3. Biing-Hwang juang, David. Y. Wong, "Distortion Performance of Vector Quantization for LPC Voice Coding", IEEE Transactions on Acoustic, Speech, and Signal Processing, vol Assp-30, no. 2,pp.294-303, Apr. 1982
4. Lawrence R. Rabiner, "LPC Prediction Error-Analysis of its Variation with the Position of the Analysis Frame," IEEE Transaction on Acoustic, Speech, and Signal Processing, vol.Assp-24, no.5, pp.434, Oct. 1987
5. L.R. Rabiner and B.H jung, 「Foundmentals of Speech Recognition」, PTR Prentic-Hall, 1993
6. R.E.Crochiere and L.R. Rabiner, Multirate, 「Digital Signal Processing」, Prentic Hall, Englewood Cliffs, NJ, 1983
7. Yoh-Han Pao, 「Adaptive Pattern Recognition and Neural Networks」, Addison-wisley, 1989