신경망에 의한 초성자음(ㄱ, ㄷ, ㅂ)의 인식방법

* * 김석동 ** 이행세

* 호서대학교 전자계산학과 ** 아주대학교 전자과

The methods of recognition of consonants(voiced stops) by Neural Network

* Suk-Dong Kim ** Haing Sei Lee

* Dept. of Computer Science , Hoseo University
** Dept. of Electronics Eng., Ajou University

ABSTRACT

As the basic analysis to solve the stop consonants in phoneme based speech recognition using Back Propagation learning algorithm, changes in hidden units , training set and iteration. Also we propose an efficient processing method of separation between consonants and vowels.

1. 서론

인간과 기계의 의사전달방법으로서 여러가지 매체가 개발되고 사용되어 왔다. 그중에 키보드가 가장 대표적인 매체이고 근래에 들어와 시각과 청각을 이용한 매체가 개발되고 있다. 음성 인식의 경우, 각 언어마다 음운의 특징이 다르고, 사용 빈도나 음운 사이의 결합 규칙등이 복잡하여 극복해야 할 많은 내용이 있다. 실제 음성의 경우 개인차가 심하고, 지방마다 사투리와 억양이 다르며, 특정한 개인의 기분에 따라서도 달라지며, 건강상태의 변화에도 영향을 받으므로 음운적 특징에 상당한 변화가 있게되는 것을 생각 할 수 있다. 또한 시각이나 촉각에 비해서 잡음요소의 개인을 막기가 용이하지 않다. 현재 도시생활의 경우 상당한 큰 소음공해를 겪고 있는 것을 감안하면, 이와같은 여러가지 주변 조건이 음성인식 문제에 장애 요인으로 등장하게 된다.

또한 언어마다 음운 체계가 다르고 문자 음성의 특징이 달라서 다른 언어에서 개발한 내용을 쉽게 수용 할 수 없는 장벽이 있다. 따라서 한글의 경우 우리가 보다 유리한 조건하에서 음성 인식체계나 음성이해 기구를 개발하고 구현하는 것이 필요하다고 생각된다. 복잡한 음성신호처리는 디지탈 컴퓨터가 대체로 수행하여 왔고[1], 컴퓨터가 급격히 발전되므로서 이 분야의 연구가 더욱 활기차게 되었다. 특히 반도체 회로의 발전과 아울러 DSP 칩이 VLSI 소자로 실용화 됨으로써 실장화(implementation)의 가능성이 높아졌다. 인공지능(AI) 기법[2]과 애매 이론(Fuzzy theory)[3]및 신경회로망(Neural Network)[4]의 보급에 힘입어 구현 알고리즘도 다양해지게 되었다. 앞으로는 컴퓨터가 중앙집중식 단일 명령어 처리방식인 Von Neumman 방식에서부터 분산처리 방식과 병렬처리 방식으로서의 전환이 기대된다. 이를 해결하는 가장 촉망받는방식이 신경회로망이다. 흔히들 우리는 근접 4세대 컴퓨터를 사용하고 있고, 앞으로 인공지능을 이용한 제 5세대 컴퓨터가 출현하고, 신경회로망을 이용한 6세대가 뒤를 이을 것이라고 생각하고 있다. 따라서 이 분야의 연구는 차세대 컴퓨터의 개발에 이바지하게 된다고 말할 수 있다.

음성을 컴퓨터에 의해 자동적으로 인식하기 위해서는 음성신호를 분석하여 얻은 파라미터를 인식의 기본 요소로 사용한다. 음성인식을 위한 특정 파라미터 추출 방법에는 크게 시간영역분할방법[5]과 주파수영역분석방법[6]으로 나눌수 있다. 시간영역 분석방법에는 영교차율, 음성에너지, 선형예측계수등이 있고 주파수영역방법에는 포르만트(formant)주파수, 셉스트럼(cepstrum)분석, 필터뱅크분석등이 있다. 본연구는 파라미터 추출방법으로 시간영역분석방법을 이용하였고, 자음을 음성에서 분리하는 방법은 일종의 주파수를 이용한 영교차율을 이용하였다. 또한 인식은 신경망으로, 각각의 구조를 변화시키면서 인식률을 조사하였다.

2. 파라미터 추출

2-1. 끝점검출

음성인식에서 음성부분과 묵음부분을 구별해내는 정확한 끝점검출은 인식률을 높이고 계산량을 줄이는 데 중요한 역할을 한다. 끝점검출에 사용되는 방법은 에너지, 피치(pitch), 영교차율등 여러가지 방법을 이용하여 구할 수 있으나 본 연구에서는 사용한 끝점검출 방법은 프레임내의 평균에너지와 피크 신호 크기를 이용한 것으로 이를 소개하면 다음과 같다.

음성신호는 시간에 따라 비교적 천천히 변화하는 사실(quasi-stationary)을 이용하여 묵음이 포함된 전 음성구간을 일정한 시간간격으로 분할(segmentation)하여 프레임(frame)으로 나누고, 각 프레임에서 평균에너지와 피크 신호크기를 구별하여 이것을 이용하였다. 일반적으로는 평균에너지를 이용하면 정확한 끝점검출이 용이하지 않으나 피크 신호 크기와 평균에너

지의 비를 이용하면 끝점검출을 용이하게 구할 수 있다 즉 두에너지의 비(rate)가 최대가 되는 프레임을 찾아 양쪽으로 가면서 어떤 일정한 값(threshold)이하가 되는 프레임이 끝점부분이 됨을 알 수 있다. 더우기 이 방법을 이용하면 천이구간과 안정구간의 분리가지도 문턱값(threshold)의 변화만으로 손쉽게 얻을 수 있다.

2-2. 자음과 모음의 분리

음성에서 자음과 모음을 정확하게 분리하는 것은 음성처리를 위해서는 해결해야 할 매우 중요한 일이다. 유성음과 무성음을 분리하는 방법을 크게 두가지로 나누어 볼수 있다. 하나는 음성신호(혹은 음성예측오차신호)에대한 자기상관(Autocorrelation)함수를 이용하여 분리하는 방법[8],과 또하나는 일종의 패턴인식의 한 방법으로 유성음과 무성음사이의 음성학적 특징을 이용하여 통계적 방법으로 분리하는 방법으로 나눌수 있다. 본연구에서는 음성학적 특징중에서 영교차율의 변화량을 이용하여 자음과 모음을 구분하였다.

자음은 성도의 일부분에서 수축작용을 일으켜 발생하는 것으로 폐에서 나온 공기가 매우 빠른 속도로 수축부위를 통과하면서 발생하는 고주파성분의 음성이며 모음은 성도진동에 의해 발생하여 크기가 비교적 큰 주기적인 성분을 갖는음성이다. 일반적으로 주파수에 비례하는 영교차율을 이용하여 정확하게 자음과 모음의 분리하는 것은 용이하지 않다. 왜냐하면 자음부분에는 Pop noise와 같은 잡음으로 진폭이 매우 큰 저주파성분이 들어가 영교차율을 모음보다 작게 할 수도 있고 모음부분에는 영레벨 근처의 신호에 외부잡음(백색잡음)이 들어가 영교차율을 크게 할 수가 있으므로 영교차율의 크기로 분리하는 것은 신뢰성이 없다. 그러나 그림 1과같이 음성자음에서 모음으로 천이되는 구간에서는 모음부분의 주파수보다 작은 저주파성분이 있음을 알수가 있다. 즉 영교차율의 변화를 이용하여 자음과 모음의 분리를 할 수가 있었다. 본논문에서 사용한 자음과 모음의 분리방법에 대하여 그림2에 나타내었다.

파열음과 같은 음성에는 자음부분에 pop noise가 실려있으므로 전처리과정에서 high pass filtering을 하여 이를 제거하였다. 실험에 사용한 음성중의 순수한 자음부분의 길이는 30 - 60 msec으로 매우 짧아서 구간(frame)의 길이를 10 msec로 하였고 구간사이의 간격은 6 msec가 중첩이 되도록 하였다.

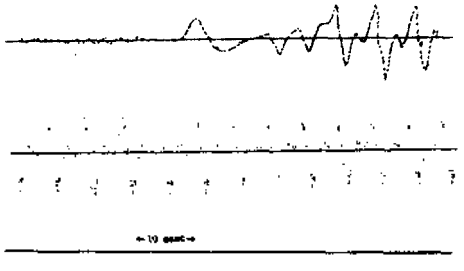

그림 1 음성 /디/의 전반부 파형

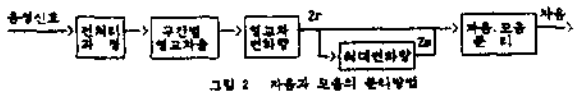

그림 2 자음과 모음의 분리방법

그림 3은 음성 /기/ 에대한 구간별 영교차율을 나타낸 그림으로 최대변화량이 나타나는곳이 자음과 모음의 과도부분임을 알수 가 있다.

2-3. 특징 검출

끝점검출을 행한후 음성부분만을 대상으로 음성의 특징을 잘 묘사하는 계수를 추출한다. 원래의 음성을 샘플링(sampling)한 많은 양의 데이타가 음성의 특징을 묘사해 주는 작은 수의 특징계수로 전환되는 의미에서 특징추출을 데이타의 압축으로 볼수 있다.

음성신호의 특성을 나타내는 특징계수의 추출방법은 에너지와 영교차율과 같은 간단한 것에서부터 LPC(linear predicitive coding)와 HP(homomorphic processing)방법등이 있다. 인식을 위한 특징 계수의 추출방법을 선택할 때 계산시간, 필요한 메모리의 양, 구현의 용이성등을 고려해야한다. 궁극적인 선택기준은 물론 인식률이지만, 이 인식율은 시스템을 시스템을 구성하는 모든 변수의 복합된 함수이다. 본 논문에서 사용한 LPC 분석에 대해서 간단히 설명하면 다음과 같다. 현재의 음성 샘플(sample)은 과거의 음성샘플(sample)들의 선형 조합으로 예측할 수 있다는 것이 LPC의 기본이 된다. 어떤 정해진 구간에서 실제값과 선형예측한 값 사이의 차의 제곱을 최소화 시킴으로써 그 구간에서의 예측계수(predicion coefficient)가 정해진다. 또한 음성신호는 유성음의 경우 준주기 펄스(quasi-periodic pulse)에 의해, 무성음의 경우에는 랜덤 잡음(random noise)에 의해 여기되는 선형 시변 시스템(linear time-variant system)의 출력으로 모델화하는 음성 발생 모델과 LPC(linear predicitive coding)은 밀접한 관계가 이다. LPC 계수의 추출과정 중 첫번째 과정은 프리엠파시스(preemphasis)를 수행함으로써 음성신호의 에너지를 저주파 대역에서 감소시키고 고주파 대역에서는 증가시키므로 음성신호 스펙트럼(spectrum)의 동적영역(dynamic range)이 감소하게 되어 성대의 스펙트럴 형태(spectral shape)을 정확하게 추정하는데 도움이 된다. 그 다음 과정으로 음성 신호는 N 개의 샘플(sample)의 블럭(block)에 대

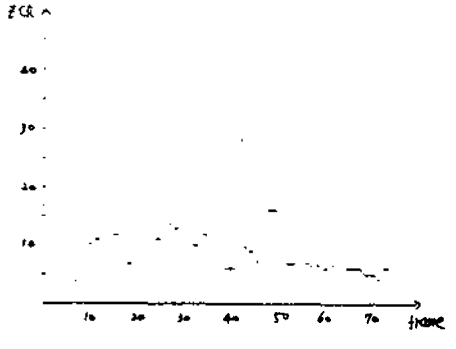

그림 3. 음성 /디/의 영교차율

하여 tapering window를 사용함으로써 스펙트럴 왜곡(spectral distortion)을 제거하게 된다. 일반적으로 해밍 창 (Hamming window)을 많이 사용하여 창 함수(window function)는

$$W(n) = \begin{cases} 0.56 - 0.46\cos(2\pi n / N), & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}$$

이다. 또한 overlapped windowing(N>M)을 사용함으로써 창(window)사이의 널(null)에 있는 데이타 샘플(data sample)을 잃어 버리는 것을 방지한다. 다음과 과정으로 창 씌운(windowing) 프레임(frame)에 대해 PARCOR 계수 (partial autocorrelation coefficient)를 구하고 나서 LPC 분석에 의해 예측계수를 구한다.

여기서 사용한 예측계수는 9차의 예측계수를 사용하였다. 또한 자모분리에서 얻어진 구간별 영교차율의 평균값과 최대변화량과 자음부분의 길이를 첨가하여 특징의 갯수를 총 12개로 하였다.

3. 신경망 알고리즘

패턴과정에서 신경회로망 모델중에서 다층망으로 구성된 퍼셉트론(perceptron) 학습방법을 사용 하였다. 그림 에 신경 회로망의 구조가 나타나 있다. 이 모델은 입력층과 출력층 사이에 중간(hidden)층이 있고, 하나 이상의 계층으로 구성된다. 여기서는 하나의 중간층을 갖는 모델을 나타내었다.

노드 끼리의 연결은 상위를 향하는 방향(feed forward)이며 훈련을 통해서 모델에 입력된 값이 원하는 출력값이 나오도록하여 파라미터를 적응 수정하는 알고리즘으로서 D.E Rumelhart,G.E Hinton,R.J. Williams의 BP(back propagation)방법을 이용하였다 [7]. 이를 소개하면 다음과 같다. 각각의 입력 자료에 대하여 원하는 출력값과 실제 출력값의 차이를 계산하고 그 값은 입력에 궤환시켜 출력오차를 최소한으로 줄이기 위해 모든 노드들 사이의 가중치를 변경한다. 이 방법의 주요과정은 다음과 같다.

1단계: 모든 Wij,j를 초기화 시킨다. Wij는 노드 j와 노드 i사이의 가중치값이다.

2단계: 훈련시킬 12차의 PARCOR계수이고 입력 데이타에 대응하는 출력 데이타는 8개의 노드중 하나를 1의 값으로 그리고 나머지는 0이 되도록 선택한다.

3단계: 현재의 Wij값에 따라 모든 노드들의 실제 출력값을 계산한다. j번째 노드의 출력값을 yj로 표시할때 그 값은 그 노드의 총입력에 대한 비선형 함수로써 다음과 같이 나타난다.

$$yj = \frac{1}{1 + \exp(-\sum yj\ wij)} \qquad (1)$$

4단계: 모든 노드에 대하여 오차항인 δj를 구한다. 즉 dj를 원하는값,yj를 실제 출력값이라 할때 출력 계층에서 오차항은 다음(2) 식과 같다.

$$\delta j = (dj - yj)\ yj\ (1 - yj) \qquad (2)$$

중간층에서의 오차는 다음과 같이 구해진다.

$$\delta j = yj\ (1 - yj) \sum_k \delta k\ wjk \qquad (3)$$

여기서 k는 노드 j의 상위 계층의 모든 노드를 가리킨다.

5단계: 가중치 값을 다음식에 의해서 수정한다.

$$wij(n+1) = wij(n) + \alpha\,\delta i yj + \beta(wij(n) - wij(n-1)) \qquad (4)$$

여기서 (n+1) 은 다음 상태,n은 현재 상태,(n-1)은 과거를 나타내고 α는 학습률이다. β는 0과1사이의 값으로 과거의 변화량이 현재의 변화량에 영향을 끼치는 값이다.

6단계: 다른 입출력 데이타를 받아 들여 2단계부터 되풀이 한다.

이 모델은 신경세포의 연결에 의한 신호 종합 기능을 나타내며, 노드는 뉴런을 나타내고 연결선은 결합강도를 나타내는 역손과 시냅스의 성질을 나타낸다. 이를 이용하여 전단 신호가 다음단에서 합해지도록 하였으며 또한 출력이 포화되는 특징을 시그모이드(sigmoid) 함수로 대치한 것이다. 모든 신호는 전향적으로 전파되는 것을 가정 하였고 각 연결선의 결합도를 학습에 의하여 적응 하도록 하였다. 이와같은 모델을 음성인식에 적용하였다.

4. 실험및 고찰

신경망의 인식능력은 은폐층의 노드의 갯수에 따라 영향을 받는다. 노드의 갯수가 충분치 못하면 신경망은 각 층사이에 적당한 상호관계를 유지하지 못하여 인식률이 떨어진다. 그림 4에서는 생략하였지만 은폐층을 1개로 하였을때 인식률은 30%이었다. 한편 노드의 갯수가 너무 많으면 훈련 데이타의 수가 매우 많아야 한다. 그이유는 훈련 데이타에서 추출해야할 파라미터가 많아야 하기때문이다. 또한 최적의 해를 구하기 위해서는 많은 파라미터가 오히려 정확한 판단을 흐리게 할 수가 있다. 즉 은폐층의 적절한 노드의 갯수는 인식률을 좋아지게 한다.

본논문에 사용한 음성데이타는 다음과 같다.

발성자	2사람의 남성 각사람의 5번 발음
sampling	10 KHz , 12 bits
분석방법	프레임 길이 : 10 msec 프레임 사이의 간격 : 6msec중첩 Hamming window
특징계수	9차의 PARCOR계수와 3개의 영교차에 관련된 특징

1) 은폐층의 노드 수에 따른 인식률

노드의 갯수를 1개에서 50 개까지 변화시키면서 인식률을 살펴보았다. 1개에서 11개 까지는 2개씩 증가시키고 15개부터 50개까지는 5개씩 증가시켰다. 그림 4에 자음별 인식률을 나타내었다./ㄱ /의 인식률은 노드의 갯수에 관계없이 일정하게 높은 인식률을 나타내고 있다. /ㄷ /의 인식률은 9개의 노드일때 가장 인식률이 좋고, /ㅂ /의 인식률은 3개, 5개, 7개일때 좋았다. 전체인식률은 5개와 7개일때 89%로 가장 높았다. 같은인식률을 나타내지만 신경망의 구조를 간단하게 하기위해 5개로 은폐층의 노드의 갯수를 결정하였다.

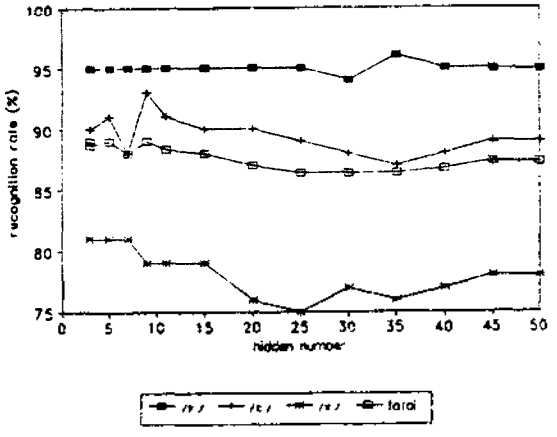

그림 4. 은폐층 노드의 수에 따른 인식률

2) 훈련 반복 횟수에 따른 인식률

실제 음성인식 시스템에 대하여 훈련시간은 가급적 적게 걸려야 한다. 훈련 할때 오차가 기준값 이하가 될때까지 반복시켰다. 본실험에 사용한 음성 조건에서는 1600방의 훈련을 반복하였을 때 오차가 기준값보다 적었다. 반복횟수가 100의 배수가 될때마다 인식률을 조사하였다. 그림 5에 나타난 바와 같이 은폐층의 노드갯수에 따른 인식률과 마찬가지로 /ㄱ /는 인식률이 처음부터 높으나 /ㄷ /와 /ㅂ /는 반복횟수에 비례하여 증가가 됨을 알수가 있다. 반복횟수가 1400번 이상일때 /ㄱ /과 /ㄷ /의 인식률은 더이상 좋아지지 않으나 /ㅂ /의 인식률이 전체 인식률에 영향을 끼침을 알수있다.

3) 훈련 데이타 수에 따른 인식률

음성인식 시스템에서 적은수의 훈련데이타로 높은 인식률을 얻을수 있다면 바람직하다. 가장 좋은 인식조건에서 실험한 결과가 그림 6에 있다. 인식률은 훈련 데이타 갯수가 증가할수록 높으나 인식률의 증가율은 약간의 변화가 있음을 알수가 있다. 훈련순서는 앞의 두실험순서와 달리 /ㄱ /, /ㄷ /, /ㅂ /를 번갈아 가면서 훈련을 시켰다. 하나의 쌍은 3개로 /ㄱ /, /ㄷ /, /ㅂ /로 구성되고, 15쌍 부터 150쌍까지 15쌍씩 증가시키면서 실험을 하였다.이때의 은폐층의 노드의 갯수는 5개, 훈련반복횟수는 1600번이었다.

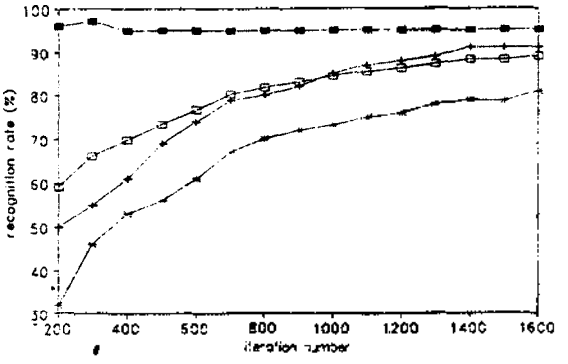

그림 5. 훈련반복 횟수에 따른 인식률

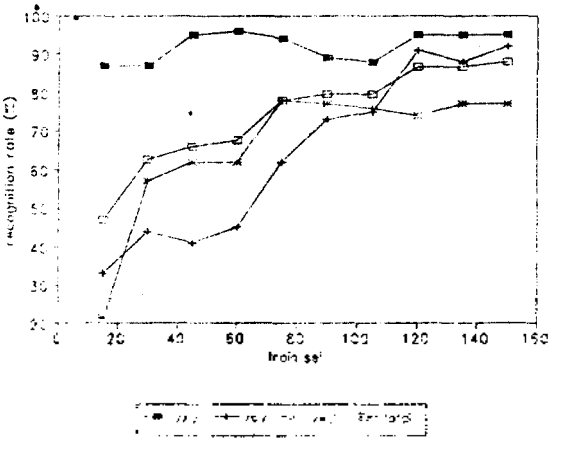

그림 6. 훈련 데이타 수에 따른 인식률

5. 결론

본논문은 초성자음중 /ㄱ /, /ㄷ / ,/ㅂ /에 대하여 BACK PROPAGATION에 의한 학습알고리즘을 이용하여 프레임별 인식률을 조사하였다. 실험에 의하면 /ㄱ /의 인식률이 95%로 가장높았고 /ㅂ /의 인식률이 떨어짐을 알수가 있었다. 본실험 조건에서 3개의 자음인식할때의 최적의 신경망의 구조가 은폐층의 노드의 갯수는 5개, 훈련데이타의 갯수가 150쌍, 훈련반복횟수가 1500번일때가 최상의 조건임을 보았다. 또한 영교차율을 이용해 자음과 모음의분리방법을 제안하였다.

앞으로 자음전체를 대상으로 인식실험을 확장하여 보다 일반적인 음성인식에 도대를 마련하여야 하겠다.

참고문헌

[1] J.L. Flanagan, "Speech Analysis, Synthesis, and Perception", Springer-verlag, New York, 1972.
[2] J.C.Jungua, " a Two pass Hybrid system for Isolated-words Automatic Speech Recognition", ICASSP-90, pp.41-44, 1990.

[3] Zadeh, L.A., " PRUF A Meaning Representation Language for Natural Languages", Int.J. Man-Machine Studies 10(1978), pp.395-460.

[4] Hell, D.O. "Organization of behavier", New York: science Edtion. 1961.

[5] G.M. White and R.B. Neely, "Speech Recognition Experiment with Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. Acoust., Speech and signal Processing, Vol. ASSP, April 1976, 183-188

[6] H.Hermansky, "an efficient Speaker-Indendent Automatic Speech Recognition by Simulation of some properities of Human auditory perception", ICASSP -87, pp.1159-1162, 1987.

[7] R.P. Lippman, "An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, Vol.4, No. 2, pp. 4-22, April 1987.

[8] H. Kobatake, "Optimization of Voiced/Unvoiced Decisions in Nonstationary Noise Environments", IEEE Trans. Vol. ASSP-35, No. 1, January 1987.