모듈구조 신경망을 이용한 한국어 단어 인식에 관한 연구

(Korean Isolated Word Recognition Using Modular Structureed Neural Network)

최 관 진*, 오 영 환

한국과학기술원전산학과

요 약

음소단위로 구성된 음소군들 각각에 대해 구성된 신경 회로망을 하나로 통합하는 모듈구조 신경망을 이용하여 일반적인 예약 시스템에서 사용할 수 있는 어휘인 시간명, 월명, 지역명등 총 34 단어에 대한 인식 실험내용을 기술한다. 구문회로망 (context net)를 이용하는 경우에 약 91.2 %의 인식율을, 단순히 음소단위를 기반으로하여 인식할 경우에 약 72%의 인식율을 얻으므로써, 음소 단위 인식 시스템의 경우에 보다 높은 인식율을 얻기 위해서는 상위 level의 처리가 수반되어야 함을 확인 할 수 있었다.

1. 서론

인간의 음성을 기계를 이용하여 인식하려는 노력은 1940년대 구분 발성된 숫자음을 자동적으로 인식하는 연구를 시초로 지금까지 계속되고 있다. 음성인식은 화자가 발생한 음성신호로부터 언어 정보를 추출하여 인간이 해독가능한 방법으로 변환하는 작업을 말하다. 이를 위해 음향학적 분석이나 언어적 분석 방법들이 사용되고 있으나, 음성이 가지는 정보의 중첩성이나 변이, 고도의 계산능력의 필요성, 음운/음소/구문 분석등의 계층적 분석과 음성 자체에 대한 기초 이론의 부족등으로 음성인식에는 여전히 많은 어려움이 남아있다.

널리 사용되는 음성 인식 방법은 크게 4 가지 정도로 나누어 볼 수있다

(1) 동적정합법(DTW : Dynamic Time Warping)에 의한 방법이 있다. 이는 화자간 지속시간의 차를 정규화시키는 방법으로, 많은 계산이 필요한 단점이 있다.

(2) 지식에 기반을 둔 방법으로, 이는 인간의 음성인식방법을 규칙의 형태로 정형화시킨 것을 말한다. 그러나, 인간의 음성인식 방법을 규칙의 형태로 표현 불가능한 경우도 있으며, 규칙 저장을 위해 많은 기억장소가 필요하므로 일반적으로 사용하기는 어렵다.

(3) 확률에 기초를 둔 hidden Markov model을 이용하는 방법이 있다. 이는 Markov model중 left to right 모델에 기반을 둔 방법으로, 단어나 음소의 이전 시점의 상태에만 의존하는 모델의 단순성으로 인하여 복잡한 현상에는 적용시키기 어려운 단점이 있다.

(4) 신경회로망을 이용하는 방법이 있다. 이는 인간이 정보를 부호화하고 해독하는 방법을 모델링한 것으로, 인공적으로 만든 뉴우런들을 상호 밀도 높게 연결시킨 뒤, 이들 간의 연결방식에 따라 학습하도록 되어 있다. 고도의 병렬 계산능력, 적응적 특성, 학습능력등의 다양한 장점을 지니고 있으며, 음성과 같이 시간적인 변화특성이 크거나, 인식을 위해 단위 시간당 많은 비교를 수행해야하는 경우 신경망을 사용하는 것이 유리하다 판단되는 경우도 많다.(3)

신경망을 사용하여 음성을 인식하는 경우, 음성에 나타나는 특성을 잘 표현할 수 있는 구조를 갖춘 신경망의 사용이 바람직하다.일반적으로 Elimann 이나 Watrous 등이 제안한 순환구조 신경망이나 Gori 가 제안한 BPS(Back Propagation for Sequence), Waibel 이 제안한 TDNN(Time Delayed Neural Network)등, 연전파 구조를 변경한 신경망을 많이 사용하고 있다. 음성 인식 시스템에서 인식의 기본단위로 음소를 사용하는 경우, 각 음소들이 가지는 고유의 특성을 최대로 반영하면서, 신경망을 설계하는 것이 바람직하다. 음소의 특성구분 없이 한번에 모든 음소를 학습시키기 보다는 특성이 비슷한 것끼리 묶어서 그 들간의 정밀한 분할이 이루어지도록 학습을 시킨뒤, 음소군간의 식별을 위한 제어망을 사용하는 modulc 구조 방식의 신경망의 사용이 음성인식을 위해서 타당하리라 판단된다

본 연구에서는 한국어에 나타나는 음소들을 조음방식에 따라 분류하고, 이들 음소군에 대한 개별적인 신경망을 구성한 뒤, 분할된 음소로부터 얻어진 유/무성 정보를 이용하여 해당 음소군을 선택하는 모듈구조 방식의 신경망에 대해서 기술하고자 한다.

2. 시•종점 검출과 음소분할

2.1 음성의 시점•종점 분할을 위한 방법

음성인식 시스템내에서 가장 먼저 할 일은 발성된 음성으로부터 실제로 음성이 존재하는 구간을 검출하는 일이다. 이는 실제 음성인식기의 성능에 큰 영향을 미치는 부분으로, 초기에 적절한 부분을 검출하지 못 할 경우, 좋은 인식율을 기대하기는 어렵다. 본 실험에서는 음성의 시점과 종점을 구하기 위해서 에너지와 영교차율을 사용하였다. 먼저, 전 음성구간에 대해서 에너지값에 대한 정규화를 수행한뒤, 0.3 % 이상되는 부분을 검출한다. 다음 검출된 구간에 대해서 2 차

정규화를 수행한다. 구해진 2개의 정규화 값들간의 차이를 구한뒤, 1차 정규화시킨 구간에서의 시작부와 끝나는 부분에서 2차 정규화시킨 값과의 차이가 큰 경우, 해당 frame 들을 제거한다. 이러한 방법은 모음이나 비음/유음등이 단어의 맨 끝부분에 오는 경우, 불필요한 부분을 제거해 준다. 이렇게 검출된 음성의 시•종점 검출 결과를 음성 분석용 소프트웨어인 CSL(Computerized Speech Lab)를 이용하여 실제 음성 파형을 보면서 분할한 구간과 비교해 본 결과 검출된 결과와 거의 동일하였다.(표 1 참조)

[표1] 시/종점 비교결과

오차범위	1% 미만	1-3% 미만	합
비율	98.7%	1.3%	100%

음성의 시점에 대한 검출시 무성자음의 경우, 몇 frame 정도 앞부분이 시점에 포함되지 못한 경우도 있었으나, 인식에 크게 영향을 끼치지는 않았다. 유성음인 경우, 시점과 종점을 잘 검출하였다.

2.2 인식단위의 선정

음성인식을 위해서는 적절한 음성단위의 사용이 필요하다. 한국어의 경우, 자음과 모음등의 음소가 서로 결합되어 하나의 음절을 형성하며, 2 내지 3 음절이 모여서 하나의 단어를 형성하는 형태로 되어 있다. 한국어 음성 인식기 설계시 인식의 기본단위로 단어, 음절, 음소등이 사용될 수 있는데, 단어인 경우, 모든 음절의 조합을 고려하여야 하므로, 인식단위로는 부적절하며, 음절을 기본단위로 하는 경우, 모든 음소의 결합을 고려하여야 한다. 반면, 음소를 인식단위로 사용하는 경우, 약 50 개의 음소만 있으면, 한국어에 나타나는 모든 음절을 표현할 수 있으므로, 인식기를 적은 크기로 구성할수 있는 장점이 있으나, 인식율이 음절이나 단어에 비해 크게 떨어지는 단점이 있다. 음소단위의 인식기를 구현하는 경우, 한국어에 나타나는 (모음+자음)형태의 음절에서 유성음의 특성을 가지면서 그 지속시간이 극히 짧은 자음 음소를 분리하는 일과 연속된 모음에서의 모음경계분리, (모음+비음•유음)에서의 비•유음 분할 등은 어려운 문제로 남아있어 음소단위의 인식기를 구현하기는 어렵다. 따라서, 본 연구에서는 기본단위로 무성음, 유성음과 위에서 언급한 특성을 가지는 음절들(단음절군)로 세분화하고, 이들 각각을 하나의 음소군으로 설정하여 인식을 수행하고자 한다. 무성음군은 무성자음으로, 파열음(ㄱ,ㄷ,ㅂ,ㅋ,ㅌ,ㅍ) 과 마찰음(ㅅ,ㅈ,ㅊ,ㅎ)으로 구성되어 있으며, 유성음군은 단모음, 유성자음(ㄴ,ㄹ,ㅁ,ㅇ), 복모음 및 혼합모음으로 구성되어 있다. 단음절군은 모음+모음, 모음+비음•유음, 모음+파열음으로 구성되어 있으며, 각 음소군에 대해서 별도의 신경망을 구현하였다.

2.3 음소단위의 분할

인식을 위한 단위로, 유성음, 무성음 그리고 유성음적인 특성을 가지는 단음절을 사용하였으므로, 음소분할시 세가지 형태로 분할하면된다. 분할을 위해서 에너지와 영교차에 대한 정규화값을 이용하였다. 에너지에 대해서 4 단계(level) 의 값으로 표시하고, 영교차는 3 단계로 표시한뒤, 이들간의 연관 관계를 이용하여 음소단위로 분할하였다. 에너지는 정규화 값이 10 인경우 1 단계, 9인 경우 2 단계, 8이나 7 인경우 3 단계, 6 이하인 경우 4단계로 표시하였고, 영교차율의 경우, 정규화값이 6 이상되면 3 단계, 3 이상이면 2 단계, 3 이하면 1 단계라 표시하여 사용하였다.

이는 유성음 사이에서의 무성음 검출과, 유음과 비음등의 유성자음 검출을 위해 보다 세분화된 형태를 가지도록 하였다.

3. 인식 신경망의 구성

사용하는 신경망은 EBP 를 기초로하여 만들어진 신경망으로, 그림 1,2 에 보인 바와 같이, 은닉층의 노드는 소수의 인접 frame 들과 연관되어 있으며, 일부노드는 특정 벡터(vector)를 구성하는 각 요소의 시간적인 변화 패턴을 학습하도록 설계하였다. 이러한 은닉층의 노드 설정은 음성에 나타나는 국부적인 연관성과 시간에 따른 특정 벡타의 각 요소의 변화 패턴을 비선형적으로 연관시키는데 있다. 순환구조 신경망의 경우, 인접 frame들간의 시간적인 연관 구조를 학습하여 실제 자료 자체의 연관성을 덜 중시하는 반면, 일반적인 BP는 시간적인 개념을 내포하지 않고 있다. 따라서, 음성과 같이 시간적인 변화 특성을 가지는 패턴의 처리시, 신경망의 구조상 순환구조 신경망이 적합하다고 말할 수 있다.

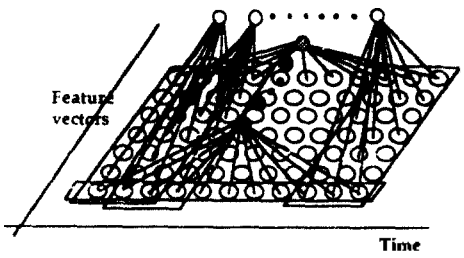

Fig1. Input Layer and Hidden Layer Configuration

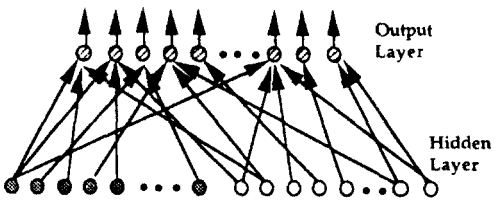

Fig2. Hidden and Output Layer Configuration

그러나, 음성자료의 수가 많아지고, 그 길이가 길어질수록, 실제 학습을 위해서 필요한 계산량과 시간이 많이 필요하므로, 순환신경망의 구현은 그리 효율적이지 못하다. 특히, 음소단위로 분할된 패턴의 경우, 그 길이가 일반적으로 짧아, 충분한 시간적인 연관성을 학습할 만큼 그 길이가 길지 못하므로, 순환구조 신경망은 효율적이지 못하다. 반면, 시간적으로 길면서, 시간적인 패턴의 변화구조가 큰 단어단위 인식에는 순환구조 신경망이 적합할 것으로 판단된다. 결과적으로,

길이가 짧고 충분한 시간적인 연관성을 가지지 못하는 음소 단위의 인식인 경우, 순환구조 신경망보다는 BP 가 바람직하나, BP구조에는 시간적인 특성을 부여할수 없다. 여기서, BP구조에 시간적인 특성을 부여하기 위해, 특징 벡터의 시간적인 변화 형태를 하나의 패턴으로 간주하고, 이를 입력 가운데에서 일부의 연관성만을 학습하는 다른 은닉층의 노드 값과 비선형적으로 통합하도록 신경망을 설계하였다.

실제로 각 음소군에 대해 이러한 신경망을 구성할 경우, 은닉층의 노드수는 가변적이다. 특징 vector의 수가 12 차이므로, 특징 벡터의 각 요소를 담당하는 은닉층의 노드수는 12 로 고정되나, 음소의 지속시간이 가변적이므로, 구현의 편의상 실제적인 지속시간을 고려하여 모든 음소군에 대해서 최대 15 frame 로 한정하여 사용하였다.

4. 실험 및 검토

4.1 실험 자료

녹음은 비교적 조용한 실험실에서 녹음하였으며, 2 회씩 3 번에 걸쳐 녹음 하였으며, 학습을 위해서 3 회 발성분을, 나머지 3 회 발성분에 대해서 인식 실험을 하였다. 전처리과정을 [그림 3] 에 보인다. 녹음된 음성을 4.9 KHz 저역 filter 를 먼저 통과시킨 뒤, pre-emphasis 하였다. 256 ms Hamming window 를 사용하여 128 ms 씩 이동시키면서, 자기상관방법에 의한 12 차 선형 예측 계수를 구했다. 이를 cepstrum 계수로 바꾸고, 인간의 귀의 특성을 반영하기 위해서 Oppenheim 이 제안한 Bilinear transformation 을 사용하여 mel-cepstrum 계수로 변환하였다. 다음으로, 신경망의 입력으로 사용하기 위하여 시점 t-1 과 시점 t+1, 시점 t 에서의 mel-cepstrum 계수의 평균값으로 대치한후, +1 과 -1 사이의 실수값을 가지도록 변형하였다.

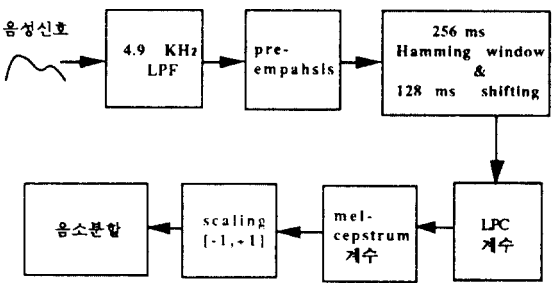

그림3. 전처리 과정

실험을 위하서 남자 화자 1 인이 6 회 반복하여 발성한 시간(12 개), 월명(12 개), 지역명(10 개)등 총 34 개의 단어를 대상으로 인 식 실험을 하였다.[표 2참조]

4.2 신경망의 학습

인식실험에 앞서서 신경망을 학습시키기 위해서, 학습 자료로 사용된 3 회분의 자료에 대해서 음소 분할을 수행한 뒤, 이들 각각을 음소 신경망의 입력으로 사용하여 신경망을 학습시켰다. 신경망은 2.2절 설명한 바와 같이, 4 개의 음소군을 위한 신경망과 무성 자음중 파열음과 마찰음의 구별을 위해 사용하는 제어 신경망 1 개와 유성음과 단음절의 구분을 위한 제어망 1 개를 포함하여 총 6 개의 신경망으로 구성되어 있다.

학습은 개별적으로 하였고, 학습이 완료된 뒤, 이들을 통합하였다.

통합된 신경망의 입력으로는 음소식별을 위한 UVchain 와 2.3절 설명한 음소분할 방법에 의해서 분할된 음소들이 입력으로 들어가게 된다. 신경망의 선택은 UV chain 의 각 요소들이 담당하며, 선택된 신경망의 입력으로는 분할된 음소가 들어가게 된다.

UV chain code 의 생성은 음소분할과 동시에 이루어지며, 유성음보다는 무성음구간의 검출에 비중을 두어, 분할된 음소에 대한 code 를 부여하게 된다.

인식실험을 위해 사용한 단어의 총수는 102 개이고, 음절수는 2 - 3 음절이였으며, 단어는 35 - 55 frame의 지속시간을 가지고 있었다. 여기서, frame length 는 128 ms이다. 학습을 위해서 사용한 분할된 음소의 평균 길이를 보면, 파열음의 평균 길이는 4.3 frame, 마찰음의 경우는, 7.8 frame 로써, 마찰음 다음에 이어지는 모음의 시작부분 이전에 나타나는 과도음을 포함하기 때문에, 파열음보다는 다소 긴 지속시간을 가지게 된다. 모음의 경우, 유성 자음은 4.2 frame 이 였고, 모음은 11.3 frame 의 긴 길이를 가지고 있었다. 마지막으로, 모음과 자음, 모음과 비음의 연결 형태를 가지는 음소군의 경우, 14.3 frame 으로 가장 길었다.

4.3 실험결과 및 분석

음성인식 실험에 앞서서 각 음소군에 대해서 인식 실험을 하였다. 음소군이 갖는 특성에 따라 마찰음이나 파열음의 경우 최대 지속시간을 8 frame 으로 하였고, 유성음이나 단음절에 대해서는 15 frame 을 사용하였으며, 모든 음소군에 대해서 시간창(time window) 크기는 3 으로 고정하여 사용하였다.

각 음소군에 대한 신경망은 초기 학습율을 0.8로 사용하고, 100 번 반복마다 0.01씩 감소시켜 나갔으며, 학습의 가속화를 위한 모멘텀 계수는 0.55 를 사용하였다.

[표 3]는 각 음소신경망이 허용오차 0.01 까지 도달하는데 걸린 반복횟수이다.

[표 3] 음소군별 반복횟수

파열음	마찰음	유성음	단음절
2470	1784	11100	12560

[표2] 음성 데이타

일월	칠월	한시	일곱시	서울	수원
이월	팔월	두시	여덟시	안양	천안
삼월	구월	세시	아홉시	성환	옥천
사월	시월	네시	열시	평택	부산
오월	십이월	다섯시	열한시	대구	
유월	십일월	여섯시	열두시	대전	

pattern 에 대한 인식실험을 하였다. 무성음이나 유성음에 해당하는 신경망의 선택을 위해서 UVchain code가 사용될 때마다 제어 신경망이 작동한다. 제어망에 의해서 해당 음소의 신경망이 선택된 경우, 최대 출력값이 0.8을 초과하는 출력 node만을 저장한다.

그렇지 못한 경우, 제어망이 제어하는 모든 신경망의 최대 출력값을 모두 저장한다. 각 음소에 대해 위의 과정을 반복 적용한 뒤, 모든 음소에 대한 처리 과정이 끝나게 되면, 구문 회로망을 사용하여 최종적인 단어인식을 하게 된다. 사용하는 구문회로망은 해당어휘 set에 나타나는 음소들의 전이 관계를 graph 형태로 표시한 것으로, edge는 전이 확률을 나타내게 된다.

(표 4)에 음소군별 인식실험를, (표 5)에는 구문회로망을 이용한 최종 인식결과를 나타낸다.

[표 4] 음소군별 인식율

ㄱ	83%	ㄴ	100%	ㅓ	87%	열	100%	영	100%
ㄷ	84%	ㄹ	70%	ㅛ		영	100%	완	100%
ㅂ	93%	ㅏ	100%	ㅟ	90%	옵	100%	육	100%
ㅅ	77%	ㅕ	100%	얼	100%	액	100%		
ㅌ	96%	ㅗ	93%	안	100%	인	100%		
ㅅ	70%	ㅜ	94%	알	97%	일	84%		
ㅈ	83%	ㅡ		암	90%	월	93%		
ㅊ	77%	ㅣ	91%	양	100%	울	100%		
ㅎ	81%	ㅐ/ㅒ	100%	언	100%	원	100%		

[표 5] 인식율 비교

context net를사용하지 않은 경우	context net를사용한 경우
72.5 %	91.2 %

학습 데이타에 대한 인식율은 100 % 였으며 실험데이타에 대한 인식율은 91.2 % 였다.

음소별 인식 실험에서 마찰음과 파열음에 대한 혼동이 컸는데, 특 히, ㅎ과 (ㄱ,ㄷ), 'ㅅ' 과 'ㅈ'간의 혼동이 있었으며, 모음 '이' 와 '알' 간의 혼동이 있어 이들이 일률로 인식되는 경우도 있었다. 이러한 경우를 제외하고는 모든 음소에 대해서 높은 인식율을 보였다.

5. 결론 및 검토

지금까지 음소의 인식단위로 유성음과 무성음이외에 모음을 초성으로 갖는 단음절까지 포함되도록하여 각 음소군에 대한 신경망을 학습시킨 후, 이를 통합하는 모듈구조방식의 신경망에 대해서 기술하였다. 인식을 위해서 일반적인 예약 시스템에서 사용가능한 시간명, 월명, 지역명등 34 개를 대상으로하여 인식 실험한 결과, 구문 회로망을 사용하지 않는 경우 약 72 %의 인식율을 얻었으며, 구문 회로망을 사용하는 경우 92% 인식 결과를 얻을 수 있었다.

인식을 위한 기본단위로 음소를 사용하는 경우, 길이가 짧고 조음환경에 따라 음소의 성질이 변하므로, 이를 일반화 시키기 어려우므로, 음소 자체만의 인식을 기반으로 음절이나 단어 인식을 하는 경우 인식율이 낮은 반면, 어휘에 대한 문맥적인 의존관계를 나타내는 구문 회로망을 보완적으로 사용하여 비교적 높은 인식율을 얻을 수 있었다. 따라서, 인식대상 어휘수가 증가하고, 독립화자 인식으로 갈수록 구문정보를 이용하는 것이 매우 바람직하다고 판단된다.

모듈구조 신경망을 사용하는 경우, 학습 데이타 첨가시 신경망 전체에 대한 재학습이 필요없이 해당 음소에 해당되는 신경망만을 학습시킬수 있어 학습 시간을 줄일수있었고, 음소군별로 별도의 신경망을 구성하므로써, 해당 음소군내에서의 음소 분할 능력을 향상시킬 수 있었다. 그러나, 음소군간의 선택을 위해 제어망이 사용되어야 하며, 동시에 여러 신경망을 사용해야하므로, 많은 resource 가 필요하였다.

앞으로 보다 효율적인 음성인식을 위해서 적절한 크기의 음소군의 선정, 음소의 특성에 맞는 신경망의 설계등에 대한 연구가 필요하다고 생각한다.

6. 참고문헌

[1] A.V. Oppenheim et al.,"Discrete Representation of Signal," *Proc. of IEEE*, Vol. 60, pp. 681 - 691, 1972

[2] D. E. Rumelhart, G. E. Hinton, " Learning Internal Representations by Error Propagation", In *Parappel Distributed Processing*, MIT Press, 1986

[3] R. P. Lippmann, "Review of Neural Networks for Speech Recognition", *Neural computation 1*, pp. 1 - 38, 1989

[4] A. Waibel, "Modular Construction of Time Delayed Neural Networks for Speech Recognition," *Neural Computation 1*, Vol. 1, pp. 39 - 46, 1989

[5] Y. H. Pao, Adaptive Pattern Recognition and Neural Networks, Addison Wesley, 1989

[6] D. P. Morgan, C. L. Scofield, Neural Networks and Speech Recognition, KLUWER ACADEMIC PUBLISHERS, 1991