# 신경망을 이용한 음성인식의 안내
# Introduction to Speech Recognition using Neural Networks

정 홍
Hong Jeong

포항공과대학 전자전기공학과
Pohang Institute of Science and Technology
산업과학기술연구소
Research Institute of Science and Technology
인공지능연구센타
Center for Artificial Intelligence

요약

한국의 HAN 人工知能컴퓨터과제나 일본의 NIPT나 성사를 가름할 수 있는 기술 중의 하나가 컴퓨터에 의한 音聲認識의 성공여부이다. 그러나 자동음성인식은 話者獨立, 連續音聲, 無制限 語彙처리라는 세가지 난관을 아직 극복하고 있다. 현재 DTW나 HMM 시스템은 계속 개선되고 있으나 근본적으로 한계가 있다고 보인다. 이와 같은 이유로 신경망을 이용한 음성인식연구가 급속히 확산되고 있다. 이와 같은 추세에 따라 본 심포지움에서는 신경망을 이용한 음성인식에 대해 소개한다.

## I. 서 론

1980년대에 들어서 신경망 연구가 다시 활기를 띄게 되었다. 아울러 난제에 부딪혀 좀처럼 실마리를 찾을 수 없었던 많은 기존의 연구분야들이 신경망 연구 분야에 눈을 돌리고 큰 기대를 걸게 되었다. 일반적으로 신경망은 생체기능 중 특히 知覺과 制御기능을 인공적으로 실현하는데 탁월한 성능을 발휘하고 있는 것으로 알려지고 있다. 따라서 지각기능 중 음성인식 기능을 기계적으로 구현하는 데 신경망이 큰 기대를 모으고 있는 것이다. 즉, 1950년대부터 활발히 연구되어 오던 기존의 음성인식 시스템이 한계를 보이고 있는 화자독립, 연속음성 인식문제를 해결해 줄 수 있는 가능성을 보여주고 있는 것이다.

한편 차세대 전자제품으로 서서히 부상할 음성컴퓨터, 음성로보트 멀티미디어에도 음성인식 시스템이 중요한 역할을 차지할 것이 분명하다. 이 분야에 대해 엄청난 연구활동을 기울이고 있는 미국과 일본 등 외국에 비해 이미 나와 있는 것에만 따라 가고 새로운 기술개발을 주저한다면 영원히 선진국을 따라 가기에 급급할 것이 분명하다. 이런 맥락에서 현재 국책적으로 추진하고 있는 HAN프로젝트 중 인공지능 컴퓨터과제에도 음성인식 시스템 구현이 중요한 역할을 차지해야 할 것이다.

## II. 음성인식의 기초

### II.1 음성인식

먼저 음성인식이란 여러 학문분야가 복합적으로 협동해서 이루어 내야할 분야이다. 즉 음성을 체계적으로 연구하기 위해서는 音聲學, 音韻學, 音韻배열론, 詩形論, 句文論, 意味論 語形論의 모든 지식을 총 동원 해야 하는 것이다. 이 것은 음성이 文章, 句, 單語, 音節, 音素라는 원소로 유기적으로 구성되어 있기 때문이다.

일반적으로 음성처리(Speech Processing)분야는 다시 음성처리(Speech Processing), 음성인식(Speech Recognition), 음성합성(Speech Synthesis)으로 구분된다. 음성처리분야에서는 음성신호의 개선(Enhancement) 및 코딩(Coding)을 연구하고, 음성인식분야에서는 음성인식 및 화자인식을 연구하며, 음성합성분야에서는 음성합성 및 음색변환을 연구한다. 이 세 분야는 시스템의 입출력 신호의 관점에서 보면 더욱 명확히 이해할 수 있다. 즉, 음성처리

1. 본 연구는 92-93년도 산업과학기술연구소, 인공지능연구센타, 전자통신연구소 지원에 의한것임.
2. 포항시 포항공과대학 전자전기공학과 (우) 790-600,
전화:(0562)79-2223, FAX:(0562)79-2903

| 문제 \ 시기 | 초 기 | 중 간 | 완 성 |
|---|---|---|---|
| 화자독립 | O | O | O |
| 무한단어 | X | O | O |
| 연속음성인식 | X | X | O |

<표 1> 음성인식의 발전단계

시스템의 입출력은 모두 음성신호이고 음성인식 시스템의 입력은 음성신호이지만 출력은 심볼이며 음성합성시스템은 이와 반대로 입력이 심볼이고 출력이 음성신호이다.

주어진 음성인식 시스템을 기능면에서 판정하는데는 여러가지 판정기준이 있겠으나 크게 나누어 화자종속(Speaker Dependant)과 화자독립(Speaker Independant), 고립단어(Isolated Word)와 연속단어(Continuous Word)라 볼 수 있다. 기타 처리 어휘수, 무제한 문법(Unconstrained Grammar)여부 및 잡음특성 등의 판정기준이 있으나 앞의 세 기준이 가장 주요한 기준이다. 현재 이 세 가지 조건을 모두 만족하는 시스템은 아직 없으며, 이 조건 들을 모두 만족하는 시스템을 구현하려는 것이 이 분야 연구의 궁극적인 목표라 할 수 있을 것이다.

고전적인 알고리즘 하에서는 이들 세 가지의 문제는 서로 상충되는 성격을 갖고 있다. 연속음성을 인식하기 위해서는 단어 수를 몇 개로 제한해야 하고 인식할 수 있는 화자의 수도 몇 명으로 한정해야만 가능하다. 즉 한 문제를 해결하기 위해서는 다른 두 문제의 희생이 불가피하였다. 이 중 화자독립의 문제가 비교적 쉬우며 연속음성을 인식하는 문제가 제일 어렵다. 지금까지의 음성인식 분야에서 이들의 문제를 해결해 온 과정을 <표 1>에 나타냈다.

음성인식 시스템을 구현하기 위해서는 전처리(Preprocessing), 패턴인식(Pattern Recognition), 후처리(Postprocessing)의 3단계가 필요하다. 전처리 시스템은 청각 시스템(Auditory System)에서와 같이 음성신호로부터 시간 및 주파수 영역의 피처(Feature)를 추출해내는 작업이라고 볼 수 있다. 이 피처로 간략히 표현된 음성신호를 패턴인식부에서 음소, 음절, 단어라는 원소를 인식해 낸다. 다음에 후처리부에서 이 원소를 재구성해 문장을 복원해 내는 것이다.

이 중 전처리부는 청각시스템의 와우각(Cochlea)의 기능에 해당하는 것으로서 음성신호의 주기성(Periodicity)과 동기(Synchrony)성의 정보를 꺼집어 내는 것이다. ([6] Lippmann)

특히 패턴인식 알고리즘이 현재 가장 활발히 연구되고 있는데 크게 템플릿기반(Template-Based Approach)로서 DTW(Dynamic Time Warping), HMM(Hidden Markov Model), 지식기반 시스템(Knowledge-Based Approach), 신경망(Connectionist Approach)으로 구분할 수 있다. DTW는 Dynamic Programming을 HMM은 확률추정(Stochastic Estimation)을 지식기반 시스템은 인공지능을 이용한 추론(Inference)을 신경망은 패턴분류의 기능을 이용해 동일한 문제를 각기 다른 방법으로 풀고 있다고 볼 수 있다. 현재 가장 앞선 시스템이 HMM이라 할 수 있고, 신경망 시스템은 아직 연구단계에 있다.

후처리 시스템으로서 언어처리 알고리즘은 구문규칙 모델과 통계적 모델이있다.[7] 구문규칙 방식은 구문론 규칙에 따라 매 단어 다음에 올 수 있는 단어를 제한해 문장을 구성하는 방식이다. 한편 N-gram 으로 표현되는 통계적 모델은 매 단어에 대해 이전의 N개의 단어가 발생할 확률을 고려해 문장을 인식한다.

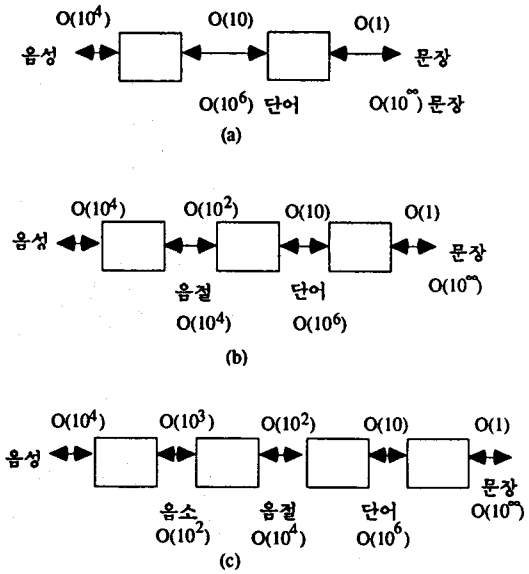
II.2 음성인식 시스템의 3가지 기본구조

이제 <그림 1>를 보면서 음성인식 시스템의 3가지 기본구조를 살펴 보자. 각 서브 시스템은 입력신호를 받아 미리 준비된 테이블 또는 사전에 기록된 아이템과 비교하여 가장 근사한 아이템을 선정해 출력으로 내놓게 된다. 이 입력신호와 사전내의 아이템과 비교하는 것이 음성인식 알고리즘의 가장 기본 기능으로서 패턴매칭 또는 분류(Pattern Matching, Classification)라 한다.

단어인식 시스템은 음성을 단어로 단어를 문장으로 바꾸는 두 서브 시스템으로 구성되어 있다. 첫번째 블록은 초속 10000개의 샘플로 들어오는 음성신호를 1백만개의 어휘를 저장한 테이블을 이용해 인식해내어 초속 10개의 단어로 변환해 내는 데이타 압축이 1000인 시스템이다. 두번째 블록은 초속 10개의 단어 시퀀스로부터 초속 1개의 문장으로 바꾸므로 데이타 압축이 10인 시스템이다.

음절인식 시스템은 음성에서 단어를 뽑아내는 블록을 다시 두 부분으로 나눈 것이다. 처음 블록은 음성신호로 부터 100개의 음절목록을 수록하고 있는 테이블을 이용해 초속 100개의 음절 시퀀스로 바꾸어 100배 데이타 압축을 한다. 뒤이어 이 음절 시퀀스로부터 1백만개의 단어 테이블을 이용해 초속 10개의 단어 시퀀스를 발생 10배 데이타 압축을 한다.

음소인식 시스템은 음절인식 시스템의 음절인식 블록을 다시 두 부분으로 나눈 것이다. 여기서 처음 블록은 음성신호로부터 100개의 음소목록을 수록하고 있는 테이블을 이용해 초속 1000개의 음소시퀀스를 발생해 10배 압축한다. 이어 초속 100개의 음절을 발생 10배 압축한다.




(그림 1) 음성인식 시스템의 3가지 기본 구조:
(a) 단어 (b) 음절 (c) 음소 인식 시스템.

이 세가지 시스템은 각기 장단점이 있다. 단어인식 시스템은 인식 어휘수가 적은 경우에 대단히 유리하다. 아울러 시스템 구조가 대단히 간단해 실용적이다. 그러나 일반적인 음성신호를 처리하기에는 단어 사전의 크기가 너무 커 구현하기가 힘들다. 그리고 음소인식 시스템은 대용량 어휘 화자독립 연속단어 음성인식을 위한 궁극적인 해결책이 될 것이다. 그러나 이 방법은 가장 복잡한 시스템으로서 음소와 음절을 정확하게 뽑아 내기 위해서는 앞으로도 많은 연구가 뒤 따라야 하는 형편이다. 그리고 음절인식 시스템은 이 두 방식의 중간적인 특성을 띤 것으로 볼 수 있다. 현재 상품화되어 있는 시스템들은 거의 전부 단어인식 시스템을 채택하고 있다고 볼 수 있다.

한국어는 특히 음절이 발달하여 음절을 인식한 후 단어를 인식하려는 연구가 진행되고 있다. 음소를 인식하여 단어를 인식하는 시스템은 몇 개 안되는 음소를 정확히 인식하여 무한단어, 화자독립의 문제를 해결할 수 있고, 음소들의 변화과정을 잘 처리하여 연속음성인식의 해결 방법으로 이용할 수 있다.

II.3 음성인식의 문제점

일찍부터 연구되기 시작한 음성인식 시스템은 다음과 같은 어려운 점 때문에 실용화가 늦어지고 있다. 즉, 음성은 화자의 신체적인 상태와 시간에 따라 음성신호의 크기가 심하게 변한다. 또한, 음의 앞뒤의 관계에 따른 다양한 음운변화(Coarticulation) 현상이 관측된다. 더구나 같은 단어를 한 사람이 발음하더라도 시간적인 길이의 변화가 심하며, 시간지연(Time-delay) 현상이 빈번함에 따라 단어의 파형이 크게 변화한다. 따라서 음성인식 시스템은 이러한 점에 영향을 받지 않기 위해서 Scaling invariance, Learning capability, Time-warping 그리고 Tim-shift invariance 기능을 갖고 있어야 한다. 이러한 기능은 IV장에서 설명하겠지만 신경망이 특히 장점을 발휘하는 부분인 것이다.

III. 음성인식의 응용분야

이제 음성인식의 응용분야를 설명하고 음성인식 제품시장 조사결과를 소개한다.

III.1 음성인식의 응용분야

응용분야를 난이도에 따라 크게 음성제어(Voice Control), 음성타자기(Voicewriter), 음성컴퓨터(Voice Computer) (또는 대화형 컴퓨터) 및 음성로보트(Voice Robot), 자동통역(Automatic Translation)의 네 가지로 분류할 수 있다. (<표 2> 참조) 음성인식이 실현되면 여러가지 전자제품을 제어할 수 있을 것이기 때문에 이런 제품군을 음성제어라 하자. 이 음성인식기능에 문법처리 기능까지 부가할 수 있다면 음성타자기를 구현할 수 있을 것이다. 여기에 음성합성 및 자연어처리 기능을 더하면 음성컴퓨터나 로보트를 구현할 수 있는 것이다. 더우기 자연어 이해 기능이 부가되면 자동통역장치도 구현할 수 있을 것이다. 여기에서 음성인식 단독으로 이러한 시스템이 완성되는 것은 아니지만 음성인식 시스템이 이들 모든 응용 분야의 핵심 부분을 차지하고 있다.

III.2 음성인식 제품

이제 시장에 나온 음성인식 제품을 살펴보자. 1989년과 1990년에 나온 상품을 조사해 보면 아직 화자종속 제품이 화자독립 제품보다 많이 나와 있다는 것을 알 수 있다. 현재 시중에 나와 있는 첨단 제품은 화자독립 연속단어 제품인데 겨우 10개의 어휘를 90%정도 인식 할 수 있는 시스템인 것을 알 수 있다. 그리고 최근에 개발된 음성인식 시스템을 <표 3>에 나열했다.

IV. 음성인식용 신경망
IV.1 신경망의 특성

1980년대에 접어 들면서 사람의 정보처리 능력이 음성, 화상, 제어분야에서 컴퓨터보다 탁월하다는 사실에 근거하여 연구되기 시작한 신경회로망은 사람의 정보처리 과정을 모델링하여, 간단하고 많은 처리요소들을 병렬로 상호 연결하여 학습을 통해 입력패턴에 내재하는 정보를 스스로 찾아내어 처리할 수 있도록 고안되었다. ([6] Lippmann)

| 응용 분야 | 사용기술 |
|---|---|
| 음성제어 | 음성인식 |
| 음성타자기 | 음성인식 및 문법처리 |
| 음성컴퓨터/로보트 | 음성인식, 합성 및 자연어처리 |
| 자동통역 | 음성인식, 합성, 자연어 이해 |

<표 2> 음성인식의 응용 분야

| 연구실 | 국명 | 시스템 | 특징 | 단어수 | 인식률 |
|---|---|---|---|---|---|
| CMU | 미국 | SPHINX | 화자독립연속음성 | 1000 | 95.8 |
| BBN | 미국 | BYBLOS | 화자종속연속음성<br>화자적응 | 1000 | 88.7<br>94.8 |
| Lincoln Lab | 미국 | | 화자독립연속음성 | 1000 | 87.4 |
| ATR | 일본 | | 화자종속연속음성<br>화자적응 | 1035 | 88.4<br>81.6 |
| IBM | 미국 | Tangora | 화자종속고립단어 | 20,000 | 95 |
| NEC | 일본 | | 화자종속고립단어 | 1,800 | 97.5 |

<표 3> 최근에 개발된 음성인식시스템

최근들어 신경망을 이용한 음성인식이 다양하게 시도되고 있으며 음성인식에 있어서 몇 가지 새로운 가능성을 보여주고 있는데 그 이유는 다음과 같다. 첫째는, 사람의 자연스러운 음성을 인식하기 위해서는 무엇보다도 매우 높은 계산속도가 요구되는데 신경망은 간단하고 많은 처리 요소들을 병렬로 연결하여 높은 계산속도를 제공해 줄 수 있다.

두번째는 신경회로망의 연결강도의 일부분이 여러가지의 에러에 의해서 훼손되어도 신경회로망은 특별한 오동작을 하지 않으며, 이는 역으로 잡음이나 생각지 않은 요인에 의해 입력패턴이 변형될 경우에도 정보를 처리할 수 있는 것을 의미한다.

또한 신경회로망은 내부구조를 적절하게 변형시킴으로써 특수한 분야에서 기존의 알고리즘이 갖지 않는 특별한 기능을 갖게 된다. 연속음성인식에 가장 중요한 Time-shift invariance 특징이 TDNN(Time Delay Neural Network)[9]에 존재하는 이유는 TDNN의 특별한 내부구조에 기인하며, 새로운 문제나 현상이 관측되면 그 문제를 해결할 수 있도록 신경망을 변형 발전시켜 최종 목적에 도달할 수 있다.

IV.2 음성인식용 신경망

음성인식용 신경망은 동작 특성에 따라 靜的(Static)인 신경망과 動的(Dynamic)인 신경망으로 구분할 수 있다. 또한 학습(Learning) 능력에 따라 지도학습(Supervised Learning)과 자율학습(Unsupervised Learning) 신경망으로 구분하기도 한다.[6]

동적 특성이 중요한 이유는 패턴분류(Pattern Classification)의 관점에서 보면 음성신호로 부터 추출할 패턴의 성질이 정적인 것인지 동적인 것인지 먼저 대별되기 때문이다. 정적 신경망의 대표적인 것으로 MLP(Multilayer Perceptron), SOFM(Self Organization Feature Map) ART가 있으며, 동적신경망으로는 TDNN(Time Delay Neural Net)과 Recurrent Neural Net이 있다. 아울러 지도학습의 대표적인 예는 MLP, Hopfield Net, Hamming Net, RCE이고 자율학습의 대표적인 예는 SOFM, ART, Darwin II이다.

음성인식을 구현하기 위해서는 음성신호로부터 특징되는 패턴을 안정적으로 추출할 수 있는 능력이 있어야 한다. 우선 음성신호의 시간축 왜곡현상(Time Warpin)을 처리해야 하므로 Time Warp Invariant해야 한다. 다음에 단어의 시작과 끝을 미리 알 수가 없으므로 Window시간 구간내 어느 부분에 단어가 위치하더라도 찾아 낼 수가 있어야 하므로 Time Invariant해야 한다. 아울러 특징추출 능력이 음성신호의 절대적 크기에 관계해야 하므로 Scale Invariant해야 하며, 특징추출 능력이 화자에 따라 변해서는 않되므로 Speaker Invariant해야 한다. 그리고 패턴인식 능력에 있어서 음성신호로부터 포만트(Formant)분포와 같은 정적인 특징과 포만트 천이(Formant Transition)와 같은 동적인 특징을 추출할 수가 있어야 한다.

이런 기능을 모두 구비하도록 설계된 신경망 중의 하나가 현재 이 분야에서 많이 쓰고 있는 TDNN이다. 그러나 음성인식 시스템을 구현하기 위해서는 이것 이외에도 여러 다른 구조의 신경망이 복합적으로 연결되어 동시에 참여해야 한다.

V. 기존의 연구결과

음성인식을 위한 신경망은 주로 동적 특성을 학습할 수 있도록 정적구조 신경망에 음성의 동적인 특성을 추출할 수 있도록 동적구조(Delay, Integration)를 첨가하여 변형한 것이다. 대표되는 신경망은 지연요소와 회귀연결을 갖는 MLP(Multi-Layer Perceptron) 구조인 시간지연 신경망(Time Delay Neural Network)과 회귀구조 신경망(Recurrent Neural Network)이 있다. 특히 Waibel에 의해 제안된 시간지연 신경망은 영어와 일어의 고립단어 인식에 사용되어 높은 인식률을 나타내었다. <표 4>는 최근의 신경망을 이용한 기존의 연구 결과를 조사한 것이다.

| 연구자 | 연구내용 | 적용언어 |
|---|---|---|
| Lang, Waibel, Hinton (1990) | TDNN이용<br>고립단어 "B","D","E","V" 인식<br>인식률 90.5-91.4 % | 영어 |
| Waibel, Hanazawa, Hinton, Shikano, Lang(1989) | TDNN 이용<br>유성파열음 /B, D, G/ 인식<br>인식률 98.5 % | 일어 |
| Waibel, Sawai, Shikano(1989) | TDNN 이용<br>18개의 자음인식<br>인식률 95.9 % | 일어 |
| Sawai, Waibel, Haffner, Miyatake, Shikano (1989) | TDNN 이용<br>음성인식<br>인식률 95 % | 일어 |
| Hataoka, Waibel (1990) | 확장된 TDNN 이용<br>화자독립모음인식<br>인식률 60.5 % | 영어 |
| Miyatake, Sawai, Minami, Shikano (1990) | TDNN과 predictive LR parsing<br>이용 5,240개의 고립단어,음소<br>고립단어 : 92.6-97.6 %<br>음소 : 98 % | 일어 |
| Hirai, Waibel (1990) | TDNN과 Dynamic Programming<br>중간용량 화자종속,<br>고립단어 인식,인식률 92 % | 일어 |

<표 4> 신경망을 이용한 음성인식 연구동향

이 조사 결과에서 볼 수 있 듯이 신경망을 이용한 음성인식이 영어와 일본어에 치우쳐 있음을 알 수 있다. 일본어의 구조가 단순해서 음성인식 시스템이 가장 먼저 성공을 거두리라는 기대 때문인지도 모른다. 그러나 일본어와 가장 유사한 한국어 음성인식의 진도가 늦다는 데는 문제가 있다. 이 와중에서도 국내 연구자들이 지금까지는 음성인식의 기본 기술을 축적하여 왔으며[1,2,3], 부분적으로 고립단어인식 분야에서는 실용화에 한걸음 다가섰다[2,3,4]. 현재 대학 연구소 등에서 이 분야의 연구를 시도하고 있으나 아직 대단히 연구활동이 부족하다. 많은 연구인력과 노력이 요구되고 있는 실정이라고 할 수 있다.

VI. 결론

이 논문에서는 음성인식의 제 문제를 고찰해 보고 신경망을 이용한 연구를 조사해 보았다. 신경망의 패턴분류 특성이 음성인식에서 요구되는 Invariance 특성 들과 일반화(Generalization) 조건 들을 만족 시킬 수 있으므로 화자독립 연속음성을 인식할 수 있는 시스템을 개발하는 데 보다 밝은 전망을 보여 주고 있는 것이다. 아직 이 분야에서 뒤떨어져 있는 국내의 연구활동도 신경망 분야에 보다 많은 관심을 기울여 하루 빨리 한국어 음성인식 연구를 진전시켜야 될 것이다.

참고문헌


[1] 김 동국, 정차균, 정홍, "신경회로를 이용한 한국어 음소 인식," *전기공학회 논문지*, 40권 4호, pp. 360-373, 1991년 4월.
[2] 음성명령환경에 관한 연구, 한국전자통신연구소 연구보고서, 1991.
[3] 정 차균, 이 영호, 최 동준, 정 홍, "음성명령을 위한 한국어 단어 인식 시스템," *신호처리합동학술대회 논문집*, 제 4권 제 1호, pp. 272-275, 1991년 9월.
[4] 정 홍, 음성명령응답 시스템개발, 산업과학기술연구소 연구보고서, 1992
[5] K. J. Lang and A. Waibel, "A Time - Delay Neural Network Architecture for Isolated Work Recognition," *Neural Networks*, Vol. 3, pp. 23 - 43, 1990.
[6] R. P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, 1987.
[7] Speech Recognition,Waibel and Lee, Morgan Kaufmann, 1990.


저자약력


1973-77 서울대 전기과 학사
1977-79 과학원 전기및전자과 석사
1979-1982 경북대학교 전자과 전임강사
1982-1984 MIT EECS 석사
1984-1986 MIT EECS E.E. 학위
1986-1987 MIT EECS 박사
1988-현재 포항공대 전자과 조교수

관심분야: 신호처리 음성처리 및 컴퓨터시각.