□ 기술해설 □

# 음성인식

한국과학기술원 유하진* · 오영환**

목 차

1. 서 론
2. 음성인식 과정 및 기본 개념
3. 음성인식 알고리즘
3.1 Hidden Markov Model
3.2 신경회로망
3.3 지식처리와 퍼지이론
4. 음성을 통한 대화
5. 결 론

## 1. 서 론

최근 컴퓨터가 소형화 되면서 자판 (keyboard)을 이용한 컴퓨터에의 자료 입력 방법은 컴퓨터와 사용자 간의 거리를 좁히는데 큰 장애가 되고 있어, 이 문제를 해결하기 위한 문자 인식 및 음성 인식 등의 연구가 활발하게 진행되고 있다. 특히 최근에 음성 정보 처리에 대한 관심이 증가하고 있으며, 많은 실용화 시스템이 발표되고 있다.

음성인식기술의 응용분야로는 받아쓰기, 개인용 컴퓨터의 명령어 입력, 자동 전화 서비스와 그 밖에 특수한 공업적 이용 등이 있다. 받아쓰기 기능으로는 편지나 신문기사 등과 같이 무제한의 단어를 사용하고 일정한 형식을 지정할 수 없는 것과, 신청서나 방사선 보고서 등과 같이 특정의 분야에만 적용될 수 있는 어휘만을 사용하고, 형식에 제한을 가할 수 있는 것으로 나눌 수 있다. 받아쓰기의 응용분야에서는 한 사람의 사용자가 오랜시간에 걸쳐 같은 시스템을 사용하게 되는 경우가 많으므로, 사용자를 한 사람 또는 제한된 수로 한정하거나 사용자에 따라 적응하게 하는 것이 가능하며, 구분발성 단어인식 기술만으로도 실용화가 가능하다. 현재 상품화된 시스템으로는 Dragon-Dictate, IBM Speech Server, Kurzweil Voice 등이 있다[1]. 이들 시스템에서는 40,000 단어 정도를 사용할 수 있으며, 사무실 환경에서 고성능 마이크를 통해 음성을 입력한다. 받아쓰기 시스템의 실용성은 새로운 사용자의 목소리에 적응하는 성능 (화자적응), 새로운 단어를 입력하는 기능, 새로운 응용 분야에 적응 할 수 있는 기능 등에 따라 향상될 수 있다. 이와같은 기능으로 현재 실용 가능한 분야 들을 살펴보면 다음과 같다.

- 컴퓨터에서의 함축된 명령어로 사용할 수 있다. 예를 들어, 파일을 열때 깊은 디렉토리 구조를 따라가서 열 필요없이, "예산 열어" 등과 같이 처리할 수 있다.
- 서류를 타자하는 도중 글꼴을 바꾸거나, 도형을 작도하면서 그림 도구를 바꾸는 등의 작업을 간단하게 할 수 있다.
- 정보 검색시 자연언어와 유사한 형태로 요구를 할 수 있다.

현재 수준의 음성인식 기술로는 이미 가장 일반적으로 쓰이고 있는 키보드와 마우스의 기능 모두를 대신하는 것이 불가능 할 수도 있다. 따라서, 앞으로는 이들 기능 모두를 이용하여 사용자에게 편리함을 주는 방향으로 사용자 입출력 장치가 발전할 것이다. 그런데, 음성의 처리에 있어서는 자연어처리나 지식 처리 등의 과정이

*준회원
**종신회원

반드시 필요하게 되므로, 간단히 기존의 그래픽 인터페이스의 대화 방식을 대치하는 것으로는 음성의 특성을 충분히 이용할 수 없을 수도 있다.

한편, 전화를 통한 음성의 입력은, 음성이 가지는 많은 정보의 양을 입력할 수 있는 다른 입력 방법을 찾기 어렵기 때문에 큰 매력과 잠재성을 지니고 있다. 그러나, 전화음성의 인식은 여러가지 음성인식의 적용 분야 중에서 가장 어려운 과제이다. 전화음성이 가지는 어려움중에는 전송선 잡음, 좁은 대역폭, 다양한 전화 송화기의 특성, 예측할 수 없이 많은 화자수 등 음성인식의 주요 과제의 대부분을 포함한다. 현재 가장 성공적으로 사용되고 있는 전화음성 시스템은 10개 내지 20개 정도의 제한적인 단어만을 사용하고 있다. 중요한 기능만을 고려한다면 "예", "아니오"와 같은 2개의 단어만으로도 실용화 시스템을 구성할 수 있다. 또한 전화선의 이용은 신호처리의 어려움외에 사용자가 시스템이 마치 사람처럼 대응해 주기를 기대하는 데에 있다. 예를 들어, 시스템의 응답이 미처 끝나기 전에 말을 하거나, 어순에 맞지 않게 말하거나, 등록되지 않은 단어를 발성할 수 있다. 이러한 경우에는 단어추출(word-spotting) 방식을 이용하면 높은 성능을 기대할 수 있다.

음성인식의 또다른 응용 분야로는 서로 다른 언어간의 통역을 들 수 있다. 여기에는 인식과 번역, 합성의 기술을 필요로 한다. 현재 다국어간의 통신을 목표로 하는 프로젝트로 Waibel이 중심이 된 JANUS(CSTAR) 프로젝트가 있다[2]. 이미 미국, 독일, 일본 세 나라의 위성을 통한 통역 실험에 성공하였고, 여기에 한국과 이태리가 새로 참여하게 된다. 다양한 언어를 처리하기 위해서 언어를 식별하기 위한 연구의 필요성도 증대되고 있다.

현재의 음성인식 연구는 음성인식에 자연어처리를 도입하는데 관심이 모아지고 있으며, 음성을 통한 컴퓨터와의 대화도 활발히 연구되고 있다. 실용되고 있는 대화 시스템으로는 PEGASUS와 VOYAGER 등이 있다. PEGASUS는 American Air Line의 예약에 실제로 사용되고 있으며, VOYAGER는 전화를 통하여 Boston 시의 호텔, 도로, 식당 안내 등의 서비스를 제공해 주고 있다.

본 고에서는 최근에 주로 연구되고 있는 음성인식 방법 및 응용사례를 살펴보고자 한다. II장에서는 음성을 인식하는 기본 과정을 살펴보고, III장에서는 주요 인식 알고리즘의 개요, IV장에서는 음성을 이용한 대화 시스템을 알아본다.

## 2. 음성인식 과정 및 기본 개념

본 장에서는 음성인식에 필요한 기본적인 단계들을 살펴본다. 일반적인 음성인식 과정은 그림 1과 같다. 마이크를 통하여 입력된 아나로그 신호인 음성은 컴퓨터로 처리하기 위하여 디지탈 신호로 변환되어야 한다. 근래에는 대부분의 개인용 컴퓨터나 워크스테이션(workstation)에 사운드카드가 기본적으로 장착되어 판매되므로 하드웨어의 구성에 큰 어려움이 없지만, 보다 정

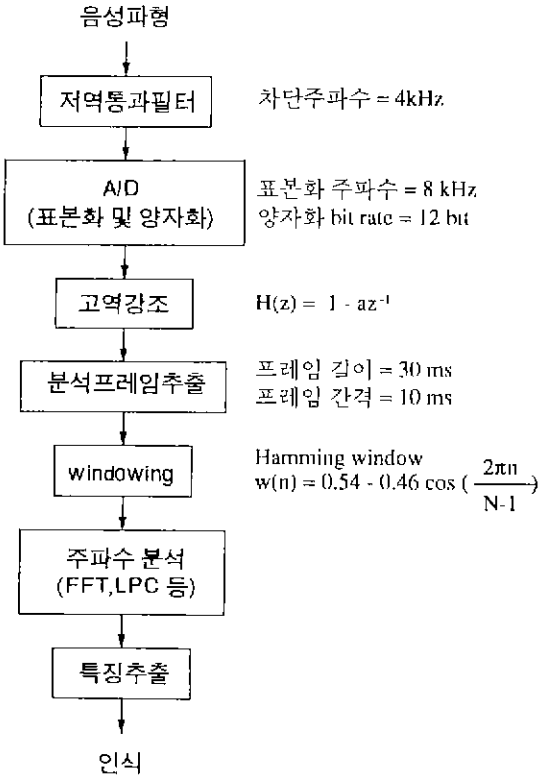


그림 1 일반적인 음성분석 과정의 예

밀한 분석을 위해서는 전용 기기를 사용하기도 한다. 수집되는 음성자료의 상태는 주변 잡음이나 마이크의 종류에 따라 크게 좌우되지만, 얼마나 정밀한 분석을 할 수 있는가는 표본화주파수(sampling frequency)와 양자화 비트수로 결정된다. 일반적으로 10 KHz의 주파수를 많이 사용하나, 최근에는 인식률 향상을 위해 16 KHz를 사용하기도 하며, 전화선을 대상으로 하는 경우에는 8 KHz 이하를 사용해야 하는 경우도 있다. 양자화 비트수로는 일반적으로 12 bit를 사용한다. 표본화를 하기 전에는 aliasing 효과를 없애기 위해 사용 주파수 대역에 맞추어 저역통과필터(low pass filter)를 사용한다. 입력된 음성은 고역의 에너지가 저역에너지에 비하여 매우 작으므로, 평탄한 스펙트럼 특성을 가지도록 하기 위해 고역강조(preemphasis)를 행한다. 다음에 음성을 분석하기 위하여 일정한 길이로 나누는데, 여기서 하나의 구간을 프레임(frame)이라고 한다. 연속된 음성자료에서 N개의 표본을 취하여 스펙트럼을 구하는 등의 처리를 하기 위해서는, 양 끝점에서의 급격한 변화를 제거하기 위하여 창(window)함수를 적용하여야 한다. 창함수는 절단되는 구간의 양 끝점이 서서히 감소하도록 하며, 해밍창(Hamming Window) 이나 해닝창(Hanning Window)을 많이 사용한다.

음성의 특성을 나타내는 파라미터로는 영교차율(zero crossing rate), 에너지, 주파수 스펙트럼(spectrum), 포만트(formant) 주파수, 자기상관계수 (autocorrelation), 선형예측계수 (linear prediction coefficient), 부분자기상관계수(partial autocorrelation coefficient), 켑스트럼계수 (cepstrum), PLP(perceptual linear predictive) 계수 등 필요한 특성에 따라 여러가지가 사용된다[3-5]. 이중 한가지만으로 시스템을 구성하기도 하지만, 필요한 부분에 따라 몇가지의 특징이 조합적으로 사용되기도 한다. 에너지는 시스템의 앞부분에서 음성의 시점, 종점의 탐색과 유성, 무성 구분에 사용되며, 영교차율은 유무성의 구분 등에 사용되는 간단한 특징들이다. 포만트 주파수는 성도의 공명특성을 나타내며, 모음의 인식에 주로 사용된다. 자기상관계수는 음성의 높이를 나타내는 피치(pitch)를 추출하는데 주로 사용된다. 선형예측계수(LPC)는, 음성이 정상적(stationary)이라고 가정하고, 하나의 표본을 앞의 몇개의 표본의 상수곱의 합으로 예측할 때, 오차를 최소화하는 상수들의 집합이다. 현재, 많은 파라미터 추출 과정은 이 선형예측을 기반으로 하고 있다. 부분자기 상관계수는 반사계수(reflection coefficent)라고도 하며, 표본들간의 잔차파형들의 상관으로 정의되는데, 잔차파형은 두 표본 사이의 데이타를 이용하여 예측된 값과 실제파형과의 차인 오차파형을 말한다. 부분자기상관계수는 LPC에 비하여 안정된 하드웨어의 구성이 가능한 장점이 있다. 켑스트럼계수는 푸리에 변환된 신호를 로그변환하고 다시 역푸리에 변환한 것으로, 동적정보를 잘 표현할 수 있는 델타켑스트럼(delta cepstrum)과 함께 현재 인식에 비교적 많이 사용된다.

최근에 발표된 특징파라미터로는 PLP(perceptual linear predictive) 계수와 rasta-PLP계수가 있다. PLP계수는 인간의 청각의 특징을 이용한 계수로 화자 개개의 정보를 줄여주어 화자독립 음성인식에 유효한 특성을 가진다[6]. Rasta-PLP계수는 PLP계수에 변형을 가한 것으로 전화선상에 존재하는 convolution 잡음을 제거하는 기능을 첨가시킨 특징 파라미터이다. rasta-PLP 계수를 구하는 중요한 과정은 다음과 같다.

1. 입력을 FFT하여 주파수 영역에서 Bark scale에 따라 스펙트럼을 재샘플링한다.
2. 대수를 취하고 적절한 필터링을 한다.
3. Equal-loudness곡선에 따라 pre-emphasis를 한다.
4. 사람의 청력의 주파수에 따른 소리크기에 대한 비선형적인 민감도를 근사한다.
5. 주파수 영역에서 세제곱근을 취한다.
6. 역대수를 취한다.
7. 역 FFT를 수행한다.
8. rasta-PLP의 cepstral계수를 구한다.

이 계수는 전화선 등의 전송선에 의한 잡음에 강한 특징을 가지고 있다. 이와같이 많이 쓰이는 특징 파라미터들을 구하기 위한 프로그램은 여러 문헌에서 쉽게 구할 수 있다.

음성인식과정에서 고려해야 할 또 다른 사항은

인식단위의 선정이다. 인식단위는 응용분야에 따라 달라지며, 시스템의 설계에 큰 영향을 미친다. 주로 사용되어 온 단위로는 단어, 음소(phone), 음절, diphone, triphone 등이 있다. 단위를 선택하는 데는 다음 세가지를 고려하여야 한다[7]. 첫째는 음소가 환경에 따라 변화하는 조음결합(co-articulation) 현상을 흡수할 수 있는가(sensitivity), 둘째는 적은 자료로 충분히 학습시킬 수 있는가(trainability), 세째는 드물게 나타나는 단어에 포함된 단위라도 기존에 학습된 단위를 이용하여 보간될 수 있는가이다(sharing).

단어 단위는 음성에서 추출하고자 하는 가장 궁극적이고 자연스러운 단위이다. 단어는 조음결합 현상을 포함하여 음향학적 변화를 흡수하므로, 단어를 단위로 하는 인식 시스템은 가장 좋은 성능을 낸다. 그렇지만, 많은 어휘를 필요로 하는 응용분야에서는 학습에 필요한 자료의 양이 많아져 상당한 분량의 기억용량과 시간이 소요되는 문제점이 있다. 예를 들어, HMM 모델로 단어를 충분히 학습하기 위해서는 하나의 단어에 대하여 20개 이상의 자료를 필요로 한다. 단어 단위의 인식 시스템을 상용화하였을 경우에는 사용자가 필요에 따라 새로운 단어를 등록하도록 하는 작업이 극히 어렵게 된다. 또한, 응용분야가 바뀌게 되면 많은 양의 어휘를 추가해야 하므로 시스템의 유연성이 없어진다.

단어 단위의 단점을 해결하기 위하여 가장 일반적으로 고려되는 단위는 음소단위이다. 음소는 그 수가 적으므로, 몇백 문장 정도의 적은 자료만 가지고도 충분히 학습시킬 수 있다. 또한, 하나의 과제를 위해 만들어진 시스템은 다시 학습시키지 않고도 다른 일을 위해 사용할 수 있다. 그러나, 음소단위는 조음결합 현상에 의한 성능의 저하를 막기 힘들다. 특히, 조사 등 중요한 의미를 지니지 않는 단어가 빠르게 발성될 때에는 각각의 음소가 정확히 나타나지 않는 경우가 많다.

단어를 사용하지 않고 조음결합 현상을 모델링하기 위해서는 음절이나 반음절 등 몇 개의 음소가 조합된 형태를 사용한다. 그러나 국어에 필요한 음절의 수는 약 3000개 정도로, 단어의 수에 비하면 작은 수이지만, 여전히 큰 수이다. 음소간의 천이부분을 모델링하기 위해서는 두개의 음소를 포함하는 diphone이 많이 사용된다. triphone은 문맥종속적인 음소로서, 같은 음소라도 앞뒤 문맥에 따라 서로 다른 단위로 인식된다. triphone은 그 수가 매우 많아 학습이 어렵지만, 자주 사용되지 않는 triphone은 문맥독립적인 음소를 이용하여 보간될 수 있다. 또한, 조음결합현상을 잘 모델링할 수 있고, 과제가 바뀌어도 다시 학습할 필요가 없는 장점이 있다. 그러나, triphone은 많은 기억공간을 필요로 하며, 여러 개의 음소가 같은 효과를 다른 음소에 줄 수 있기 때문에 서로 다른 triphone이 같은 특성을 가지는 경우가 많게 된다. 따라서, 서로 유사한 triphone들을 묶어 하나로 취급하는 방법도 사용되고 있다. 조음결합을 포함하고 분할을 용이하게 하기 위하여, 하나의 단위에 포함되는 음소의 수를 지정하지 않고, 단지 스펙트럼 천이만을 고려하여 분할하는 불균일 단위(non-uniform unit)를 사용할 수도 있다[8].

구분발성 단위를 인식하는 경우에는 입력음성 신호로 부터 단어단위를 분할하기 위해 에너지, 영교차율 등을 사용할 수 있다. 음소 등의 단위로 인식을 할 경우에는 프레임 단위로 인식하기도 하고, 주파수 스펙트럼의 변화정도를 나타내는 스펙트럼 천이척도(spectral transition measure)를 사용하여 음소단위로 분할하기도 한다. 그렇지만, 음성을 음소단위로 정확히 분할하는 것은 어려우므로 HMM 모델을 이용한 인식에서는 segmental k-means 알고리즘을 이용하여 인위적인 분할이 필요없이 학습과 인식을 하기도 한다.

이상과 같이 특징 파라미터와 인식단위가 정해지면 다음장에서 설명되는 여러가지 방법에 의하여 인식이 행해진다. 기본적으로는 문자, 화상인식 등에 사용되는 패턴인식 기법이 그대로 적용될 수도 있지만, 음성의 특성을 고려한 방법이 만족한 결과를 줄 수 있다.

음성인식에서 최종적인 결과는 입력된 음성이 지니는 의미의 파악이라고 할 수 있다. 구분발성단어 인식의 경우에는 단어 인식결과가 그대로 명령어 입력 등의 응용분야에 적용되지만, 연속음성의 경우에는 인식된 단어열 자체만으로는

부족하며, 구문분석, 의미분석을 통해 화자가 의도하는 바를 인식하고, 적당한 결과를 내주어야 한다. 구문분석에는 문맥자유(context free) 문법이나 유한상태(finite state) 문법 등 문서입력에서의 자연어 처리에서 사용되는 문법이 많이 사용되지만, 문서 입력에 비하여 문법적 생략 및 오류가 많고, 오인식이 많은 음성 입력 응용분야의 실용화에는 문제가 많다. 주어진 문장의 집합으로부터 쉽게 구할 수 있는 문법 제약으로는 하나의 단어 뒤에 올 수 있는 단어들을 지정하는 word pair 문법이 있다. word pair 문법에서 각 단어마다 확률값을 지정한 것이 bigram 문법이다[7].

## 3. 음성인식 알고리즘

음성 인식에 사용되는 방법은 동적정합법(dynamic matching)[9], HMM(hidden Markov model), 벡터 양자화(vector quantization), 신경회로망, 퍼지 이론 등 다양하다. 본 장에서는 음성인식에서 현재 가장 많이 사용되고 있는 알고리즘에 관하여 살펴본다.

### 3.1 Hidden Markov Model

Hidden Markov Model(HMM)[10]은 현재 음성인식 모델 중 가장 성능이 좋은 것으로 알려져 있고 많이 연구되고 있다. HMM은 확률모델로서 많은 자료를 이용하여 학습될 수 있다. 음성생성 모델은 그림 2와 같이 몇개의 상태(state)로 이루어져 있으며, 하나의 상태에서 다른 상태로 천이하는 확률 및 어떤 상태에서 어떤 기호가 출력되는 확률이 학습에 의해 정해진다. 모델내에서는 확률값에 의해 상태를 천이해 가며 기호의 열을 출력하는데, 상태가 천이되는 과정은 밖에서 알 수 없으므로 은닉(hidden)이라고 한다. 외부에서 볼 수 있는 것은 모델에서 출력되는 기호의 열이다. 이 기호의 열로 부터 모델을 찾아내는 과정, 즉 어떤 기호의 열이 출력될 확률이 가장 큰 모델을 찾아내는 것이 인식과정이다. HMM이 사람의 조음기관을 모델링 한다고 할 때, 음성은 어느 하나의 음소를 표현하는 모

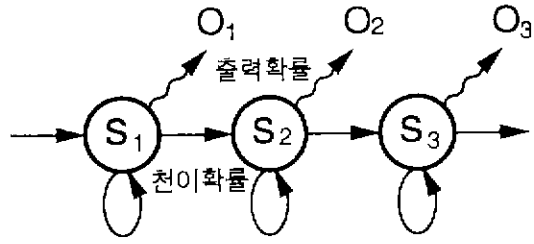


그림 2 HMM의 기본 구조

델의 출력이라고 할 수 있다. 음성은 근본적으로 정상적(stationary)이지 않고 계속 변화하지만, 음성신호를 단구간으로 나누면 각 구간내에서는 정상적이라고 할 수 있다. 그리고, 하나의 구간을 미리 정의된 구간으로 정의하면 이산(discrete) HMM의 출력과 같이 생각할 수 있다. 인식을 위해 입력된 음성신호는 벡터양자화(vector quantization)[11]에 의해 미리 정의된 기호로 변환된다. 벡터양자화는 연속이거나 이산적인 벡터를 하나의 숫자로 사상하는 것이다.

HMM에는 응용 분야에 따라, 또는 모델의 난이도에 따라 여러가지 형태의 모델이 사용된다. 또한, HMM은 출력되는 심벌의 type에 따라 이산형 HMM(discrete HMM)과 연속형 HMM(continuous HMM) 등으로 크게 나누어지는데, 이산형 HMM의 경우 벡터 양자화 과정을 통해서 얻어진 codebook을 사용하여 입력 패턴을 해당 색인(index)으로 부호화하는 전처리 과정을 필요로 하나, 연속형 HMM의 경우는 입력 자료의 값 그 자체를 그대로 입력하여 사용한다. 전자의 경우, 구현이 간단하고 학습시간 및 학습할 모델의 파라미터들이 적다는 장점이 있으나, 벡터 양자화시 정보가 유실되는 단점이 있다.

하나의 모델을 학습시키기 위해서는 각 단위별로 양자화된 기호의 열을 제시해 주어야 한다. 인식단위가 단어인 경우에는 단위별 분할이 용이하지만, 음소나 음절 등 단어 보다 세분화된 단위일 경우에는 단위 분할이 용이하지 않으며, 수작업이 필요하게 된다. 실제로 HMM을 충분히 학습시키기 위해서는 많은 양의 자료가 필요하므로 모든 자료를 수작업으로 분할하는 일은 비용과 시간이 많이 들게 된다. 최근에는 수작업을 전혀 하지 않고 모델을 학습할 수 있는 segmen-

tal K-means 알고리즘[12]이 사용된다. 이 알고리즘은 다음과 같다.

1. 초기화－모든 학습용 음성 자료를 단위와 HMM상태별로 균일하게 분할한다.

2. 군집화(clustering)－주어진 단위의 상태 $S_j$에 해당하는 관측벡터들을 $M_j$ 군집으로 분할한다.

3. 추정(estimation)－상태 $S_j$ 내의 각각의 군집 $m(1 \le m \le M)$에 대하여 평균벡터, 분산, mixture weights를 구한다. 모든 단위와 상태들에 대하여 단계 2와 3을 반복하면, HMM집합이 만들어진다.

4. 분할(segmentation)－각각의 HMM에서 Viterbi 알고리즘을 이용하여 학습 자료들을 단위와 상태들로 분할한다.

5. 반복(iteration)－결과가 수렴할 때까지 단계 2에서 4까지를 반복한다. 여기서 수렴은 정합의 평균 확률값이 더이상 증가하지 않는 것을 말한다.

HMM을 이용한 시스템 중 잘 알려진 것으로는 Tangora system, SPHINX system 등이 있다. Tangora 시스템[13]은 80년대 중반에 HMM의 가능성을 잘 보여준 시스템으로 IBM은 화자종속 구분발성 단어인식 성능을 5,000 단어에서 20,000 단어까지 확장하는데 성공하였다. 5,000 단어의 실시간 인식 시스템은 IBM PC에서의 보드로 제작되었다. 본 시스템은 코드북 크기 200의 VQ와 이산형 HMM으로 구성되었다. 언어 모델로는 n-gram 확률 모델을 사용하였다. 즉, 문장은 하나(unigrams), 둘(bigrams), 또는 세개의 단어의 조합(trigrams)으로 표현되며, 많은 문서 자료로 학습된다. 미리 정의되지 않은 단어의 입력을 위해서는 철자로 입력하기 위한 모드도 포함되어 있다.

SPHINX 시스템[7]은 CMU에서 개발된 화자독립 연속음성인식 시스템으로 triphone을 기본으로 하고 있다. 여기서는 DARPA의 자원 관리 과제로 1000단어 어휘와 7,000개의 노드와 65,000개의 연결선을 가지는 유한상태 오토마타로 구성된 시스템에 관하여 설명한다. 실시간을 위한 시스템은 Weitek 다목적 프로세서를 이용하여 구현되어 있다. SPHINX 시스템은 VQ와 이산형 HMM을 기본으로 하여 구성되었다. 켑스트럼, 차분켑스트럼(differential cepstral), 에너지를 특징으로 사용하며, 각각이 서로 다른 코드북으로 양자화된다.

### 3.2 신경회로망

신경회로망은 학습에 의해 음성패턴을 분류할 수 있고 병렬처리가 가능하여 음성인식의 실시간 인식을 가능하게 하는 잠재력을 가지고 있어, 꾸준히 연구되고 있다.

음성인식을 위한 신경회로망은 지연요소와 회귀연결의 유무에 따라 정적구조 신경망과 동적구조 신경망으로 구분할 수 있다. 정적구조 신경망은 음성을 정적인 패턴으로 간주하고 기존의 신경회로망의 구조를 이용하는 것으로, 다층 퍼셉트론(MLP), SOFM(self organizing feature map), 뉴럴예측모델(neural prediction model) 등이 있다. 동적구조 신경망은 음성의 동적 특징을 추출하기 위하여 지연요소, 회귀요소를 추가한 것으로 TDNN(time delayed neural network). Recurrent 신경회로망 등이 있다.

Kohonen의 SOFM[14]은 입력패턴을 스스로 분류하여 2차원 망위에 feature map을 형성하는 구조로 되어있다. 학습은 비교사 학습(unsupervised learning)으로 이루어진다. 출력측 뉴런과 뉴런간의 연결강도는 억제방향(inhibition)으로 작용하며, 입력과 출력층 뉴런간 연결강도는 증대방향(exitation)으로 작용한다. 학습과정은 먼저, 연결강도를 임의의 값으로 초기화 한 후, 입력을 입력노드에 인가한다. 다음은 연결 weight벡터에 의하여 가장 강하게 반응하는 하나의 출력 노드가 정해지고, 정해진 노드와 그 노드에 이웃하는 노드와 입력 노드와의 연결강도가 입력에 따라 변화한다.

Waibel의 TDNN[15]은 음성의 동적인 특성을 위치에 무관하게 감지할 수 있도록 MLP를 변형시켜 여러 개의 작은 수용영역(receptive field)을 중첩시킨 구조이다. 시간상에서 N개의 서로 다른 위치에 있는 입력패턴을 검출할 수 있으며, 음성신호의 현재 입력에 대하여 앞뒤사이의 관

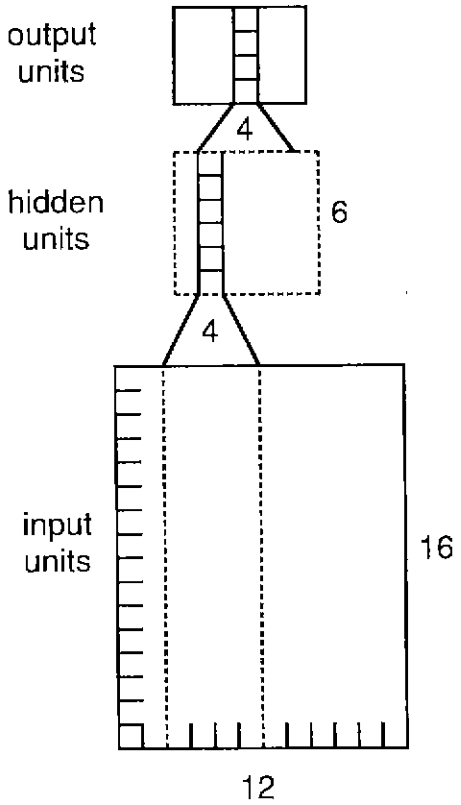


그림 3 TDNN의 기본 구조

계를 감지할 수 있어 패턴의 이동에 불변한 특성(shift invariant)을 가진다. 그리고, 정렬되지 않은 패턴도 명확히 구분할 수 있고, 학습패턴 중 단어를 구분하는 데 필요하지 않은 부분을 기각할 수 있다. 이 구조는 음성의 분할(segmentation)과 시간정합(time warping) 문제를 해결할 수 있다. 그림 3은 TDNN의 예를 보여준다. 이러한 구조로, 낮은 층에서는 입력의 국부적 특성들을 감지할 수 있고, 높은층에서는 시간적으로 더 길고 복잡한 특성들을 감지할 수 있다. 전체적으로는 입력에 포함된 각 음소들의 특징들을 시간위치에 무관하게 감지할 수 있다. TDNN의 학습은 약간 수정된 오류역전파(error backpropagation) 알고리즘을 이용한다. TDNN은 유성파열음(B, D, G) 인식 실험에서 HMM보다 향상된 인식률을 보여준다.

NPM(neural prediction model)[16]은 동적정합법(DP)과 다층 퍼셉트론을 이용하여 패턴을

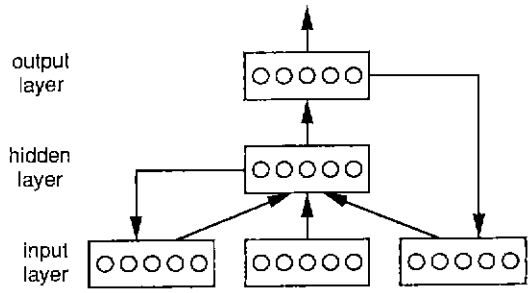


그림 4 Recirrent 신경망의 구조

비선형 예측하는 구조로 되어 있다. 적은 학습 데이타로 학습이 가능하며, 새 클래스의 추가가 용이하고, 연속음성에 적용할 수 있다. NPM의 입력은 t-1 시각 이전의 연속된 특징벡터이며, 출력은 시각 t에서의 예측벡터이다. 예측과정에서 입력음성은 N개로 분할되고, n번째 MLP가 n번째 부분열을 예측한다. 이때, 누적예측오차를 최소화하기 위하여 동적정합법을 사용하며, 누적예측오차가 입력음성과 단어간의 거리척도가 되어 인식에 사용된다.

Recurrent 신경회로망[17]은 그림 4와 같이 회로망의 입력과 출력단 사이에 궤환(feed back) 연결이 존재하여 시간에 따라 그 특성이 변하는 구조를 가진다. 현재 입력에 대해 과거 신경 회로망의 반응 혹은 상태를 반영하여 현재 신경회로망의 상태를 결정한다. 궤환 연결선은 일종의 history기능으로, 현재 입력이 과거의 상태를 반영하여 분류될 수 있는 효과를 지닌다. 이러한 동적구조 신경 회로망을 사용한 음성인식은 신경회로망의 자체의 동적특성에 따라 정적 구조 신경 회로망에서와 같이 공간상에서의 연관관계보다는 시간에 따른 입력패턴의 변화형태 즉 시간적인 궤적(trajectory)을 학습하게 된다. 따라서, 새로운 입력이 제시되는 경우, 학습된 자료들과의 공간적인 인접성으로 패턴을 분류하는 것이 아니라, 입력패턴의 시간축상에서의 궤적들간의 근접성으로 패턴을 분류하게 된다. 동적 신경회로망을 사용하므로써, 입력 패턴의 길이를 일정하게 고정시킬 필요가 없으며, 인식률면에서도 정적구조의 신경망보다는 좋은 것으로 알려져 있다. 그러나, 수렴이 완전히 보장되지는 못하며,

학습에 사용되는 입력패턴의 변이가 크거나, 자료의 길이가 긴경우 학습을 위해 많은 시간이 필요하고, 기억장소나 계산량이 큰 단점이 있다. 이러한 동적구조를 갖는 신경 회로망으로는, 은닉층의 활성화 패턴을 다음 시점에서 입력과 함께 반영하여 처리하는 Elman의 SRN(Simple Recurrent Neural Network), 은닉층과 출력층의 노드들 각각이 재귀적 연결선을 갖는 Watrous가 제안한 Temporal Flow Model, Gori가 제안한 BPS(Back-Propagation for Sequence)등이 있다.

### 3.3 지식처리와 퍼지이론

일반적으로 음성인식 시스템에서는 기본적인 수법 이외에 개발자가 독자적으로 얻은 경험적 지식(knowhow)이 중요한 역할을 하여, 이것이 시스템의 성능을 크게 좌우한다. 최근에는 지식공학적 수법을 음소의 인식 자체에 적용하는 연구가 여러 곳에서 진행되고 있다. 음성인식의 핵심이 되는 부분에의 지식처리기술의 적용은 음성인식 연구에서의 새로운 방법론으로 그 성과가 주목된다. 음성인식 전문가 시스템은 다음과 같은 특징을 가진다. 첫째로, 경험적 지식이 규칙화되어 시스템의 단계적인 구현이 가능해진다. 즉, 음성연구자가 얻은 경험적 지식을 추론규칙의 형태로 표현할 수 있으므로, 이러한 지식의 지식베이스화를 통해 단계적인 시스템의 구성이 가능해진다. 또, 시스템의 동작과 연구자의 인식과정간의 상호 대응이 가능하므로, 시스템의 개량 및 오류수정이 쉬워진다. 종래의 수법으로 구성된 시스템은 모두 블랙박스 (black box)와 같아서, 오인식이 어떠한 과정을 거쳐 일어나는가 또는 어떠한 방법으로 이것을 수정할 수 있는가를 알기가 어려웠다. 그러나, 전문가 시스템에서는 지식의 실행 하나 하나가 음성 연구자의 인식과정과 대응되어, 시스템의 동작 상태를 알기 쉬우므로 오류의 수정이 쉬워진다. 이러한 특징들은 기존의 시스템에서 실현하기 어려운 것으로, 전문가 시스템의 장점이라고 할 수 있다.

전문가 시스템에서는 전문가의 지식을 수집하는 일이 필요한데, 보통 이러한 일은 목적하는 분야의 전문가와 상담함으로써 이루어진다. 연속음성의 인식에 관해서는 모든 사람이 전문가이지만, 자신이 어떻게 음성을 인식하고 있는지를 구체적으로 표명할 수 있는 사람은 없다. 그러나, 전문 수사관이나 음성 연구자에 의해 많이 사용되고 있는 스펙트로그램 (spectrogram)에 의한 음성의 인식에 관한 지식은, 사람의 노력에 의해 의식적으로 체득되는 것이므로 표현이 비교적 용이하다. 또한, 훈련에 의해서 매우 높은 인식률을 얻을 수 있다는 사실이 알려져 있다. 이러한 사실에 기초하여, 현재 개발 중에 있는 많은 연속음성인식 전문가 시스템이 그림의 형태로 표현된 특징 파라미터의 변화를 읽어 추론규칙으로 표현하는 방법을 실현하고 있다. 주파수 스펙트럼 등의 파라미터의 시간변화로 시각적인 인식을 하는 소위 spectrogram reading은 오랜 역사를 가지고 있으며, 많은 음성연구자가 어느 정도는 이와 같은 방법으로 연구하고 있다고 할 수 있다. 최근, 컴퓨터의 성능이 비약적으로 향상되고, 지식공학 분야의 연구가 진척됨에 따라, 시각에 의한 지식을 이용하는 음성인식 전문가 시스템의 연구가 활발해지고 있다. 음성인식 전문가 시스템의 처리의 흐름은 특징추출부와 인식부로 나누어진다. 특징추출부에서는 주파수 분석과 같은 특징추출 후에 이러한 특징을 명암의 변화와 같은 화상 표시로 변환하는, 보다 고차원적인 특징 기술이 이루어진다. 인식부에서는 특징추출부에서 얻어진 특징기술에, 추론규칙으로 표현된 인식지식을 적용하여 음소를 인식한다. 여기서 알 수 있는 바와 같이 음성인식 전문가 시스템을 구현하기 위해서는, 전문가가 실제로 주목하는 특징을 기술하고, 이것을 기초로 구성된 인식 지식을 지식 베이스 (knowledgebase)에 저장하는 작업이 필요하다.

Zue가 개발한 시스템은 스펙트럼을 처리의 대상으로 한다[18]. 음소단위로 인식함을 목표로 하며, 각 음소의 고유한 음성학상의 특성을 기술하고, 이것을 기초로 하여 인식 규칙(rule)을 작성한다. 추론 엔진은 MYCIN 류와 같이 후향추론을 사용하였다. 이 시스템은 종래의 음성연구의 흐름에 충실하여, 음성연구에 의한 지식의 축적에 크게 의존하는 전통적인 시스템이라 할

수 있다.

일본의 Mizoguchi에 의해 개발된 음성인식 전문가 시스템 SPREX(a SPeech Recognition EXpert)은 여러가지 음성파라미터를 기호화 하여 표현한 후, 규칙으로 음소를 인식하고 있다[19]. 일반적인 전문가 시스템이 주파수 스펙트럼을 기본적인 특징 파라미터로 하여 음소 단위의 인식을 목표로 하는 데 반하여, 본 시스템에서는 포만트 주파수, 대수에너지, 영교차율, 고역/저역 에너지 비, 피치 등의 비교적 간단한 파라미터를 사용하여 그룹화된 음소군을 인식함을 목표로 하고 있다. 음소군 내에서의 음소 구분은 의미정보를 이용하여 처리할 수 있다. 이렇게 문제를 간소화함으로써 시스템의 현실화, 화자 독립성 등을 이룰 수 있다. 전문가는 음성분석도를 보고 음운 경계를 알아낼 수 있으며, 음운을 인식할 수 있다. 이때, 각 특징 파라미터의 절대값으로 판단하지 않고, 파라미터의 시간적 변화의 경향과 극대값 등의 특이한 부분에 주목한다. 음성분석도의 각 파라미터는 상태기술부에 의해 C, I, D, N, A, E, G (정상, 증가, 감소, 잡음, 출현, 소실, 불연속구간) 등의 기호로 표시된다. 이렇게 하여 얻어진 음성 데이타는 기호처리부에 보내져, If-Then 규칙으로 표현된 지식에 의한 인식이 이루어진다. 모음의 인식은 제 1, 제 2 포만트 주파수를 사용한다. 이때, 화자 의존성이 문제가 되므로, 인식시에 표준패턴을 선택, 학습하는 방법을 병용하여 불특정화자에 대응하고 있다.

한편, 사람이 영상, 글자, 음성 등의 패턴을 인식할 때, 다른 패턴 과의 차이를 명백하게 결정지을 수 없는 경우가 많다. 특히 음성을 인식하는 데 있어서, 발음이 불명확하거나, 잡음이 포함된 경우, 그 판단이 애매모호 해지는 경우가 많다. 이러한 경우에는 전후 문맥이나 상황에 따라서 다른 판단을 하게 된다. 지식처리를 이용한 많은 음성인식 시스템은 인간의 지식을 표현하기 위하여 "크다", "작다" 등의 언어를 사용한다. 물리적인 수치로 표현되는 파라미터의 정량적인 값을 이와 같이 규칙의 작성을 위한 언어로 표현하기 위해서는, 보통 임계값을 사용하여 어느 물리적인 값 이상은 "높은 값"으로 정의하게 된다. 그런데, 이 임계값에 가까운 물리량은 어느쪽에 포함시켜야 하는지 애매해지는 경우가 많으며, 이러한 애매한 판단에 의해 최종 결과가 크게 영향을 받게 된다. 그러므로, 발음과 인식 자체에 있어서 애매모호한 음성을 표현하는 데 있어서 퍼지논리를 사용하는 것은 자연스러운 일이라고 할 수 있다. 퍼지 (fuzzy) 집합 A는 전체집합 X의 원소 x와 적합도 (membership) 함수의 순서쌍의 집합으로, 적합도는 원소 x가 집합 X에 속하는 정도를 표시하는 함수이며, 일반적으로 그 최대값이 1로 정규화 되어있다[20].

S. K. Pal은 모음과 화자 인식에서의 의사결정에 퍼지 집합론을 사용하였고[21], 퍼지 척도를 이용하여 급내(intraclass)와 급간(interclass)의 애매성을 이용하여 특징을 선정하는 방법을 제안하였다[22]. Renato de Mori는 음성 이해 시스템에 의하여 가설이 만들어지고, 퍼지 알고리즘을 이용하여 이것을 검증하는 시스템을 구현하였다[23]. G. Hirsch는 퍼지관계를 이용하여 음소의 구별을 하였고[24], Michel Lamotte는 불확실성 척도(uncertainty measure)를 이용하여 음소의 인식을 하였다[25]. Francisco Casabuberta는 지식기반을 여러단계로 구조화 하고, 퍼지 오토마타를 이용하여 음소의 분류를 하였다[26]. Fujimoto는 퍼지 패턴 정합을 이용한 단어 인식 시스템의 구현을 기술하였다[27].

## IV. 음성을 통한 대화

사람과 사람이 음성을 통해 의사를 전달할 때는 한편이 일방적으로 말하고 다른 한 쪽이 들을 때보다 서로간의 주고 받는 대화를 할 때 이해가 빨라진다. 상대방이 말한 내용중 몇 개의 단어가 확실하게 인식이 되지 않거나, 말의 의미가 이해가 되지 않거나, 동의할 수 없는 내용이 있을 때, 듣는 쪽은 이에 대응하게 된다. 컴퓨터에 의한 음성의 인식은 사람보다 오인식 되는 경우가 더 많으므로 대화에 의한 확인이 사람보다 더 필요하다고 할 수 있다. 사용자와 컴퓨터가 대화하기 위해서는 사용자의 발성내용을 구문분석하고 의미분석하여 사용자의 의도를 파악하여야 한다. 그런데, 대화에 사용되는 음성 언어(spoken

language)는 문서 언어(written language)에 비하여 의미를 분석하는 데 많은 어려움이 따른다. 음성 언어는 보다 많은 생략(ellipsis)과 단편적인 표현(fragmental expression)들이 사용된다. 일반적으로 음성 언어는 표준 학교 문법에 비교하여 생략(ellipsis), 어순의 전도(inversion), 머뭇거림(hesitation), 자신이 했던 말에 대한 수정(self-correction)이 두드러지게 많이 나타난다. 음성언어 상에서는 이러한 언어적 현상 뿐만 아니라 발음상의 오류(enunciation error), 형태소상의 오류(morpheme error), 구문상의 오류(syntax error) 및 그 외의 지협적인 오류들이 빈번하게 발생한다. 이렇게 대화 도중에 발생하는 오류들은 대화 도중 화자가 같은 내용을 반복해서 이야기 해 줌으로써 제거되어질 수 있다. 이러한 현상들에 대하여 좀 더 자세히 살펴보면 다음과 같다[28].

(1) 생략(ellipsis)

발화의 내용이 전달하고자 하는 정보만을 포함하고 있는 경우로서 구문구조의 대부분이 생략되어 있지만 대화 상황에서 의사소통을 위해 필요로 하는 충분한 정보를 가지고 있으며 다른 구문구조와의 구별이 가능하다.

(2) 머뭇거림(hesitation in speaking)

머뭇거림은 화자가 적절한 단어나 어구를 생각해내기 위해서 이루어지는 현상이다.

(3) 어순의 전도(inversion)

관점 및 주제를 공유하고 있다면, 구문상에 큰 구애를 받지 않고 어순이 전도될 수 있다.

(4) 했던 말에 대한 수정(self-correction)

머뭇거리는 것과 유사하게 화자의 전달하고자 하는 내용을 적절하게 유지하기 위하여 자신이 했던 말에 대하여 수정을 가하게 된다.

(5) 발음상의 잘못을 복구할 수 있는 듣는쪽의 능력

듣는쪽에게는 화자가 말하는 도중 종종 범하게 되는 발음상의 잘못을 복구할 수 있는 능력이 있다. 또, 듣는쪽이 복구할 수 없는 경우에는 화자에게 잘못을 수정하거나 확인해 줄 것을 요청할 수 있다. 듣는쪽의 이러한 능력은 대화의 상황을 정확히 이해하고 있는 지식을 이용한 것이다.

지금까지 개발된 대화 시스템 중 실제 입출력으로 음성을 사용하는 시스템 상에서 구현된 몇가지 예를 살펴보면 다음과 같다.

(1) Toshiba의 Y. Takebayashi, H. Tsuboi, H. Kanazawa에 의해 구현된 TOSBURG[29,30]는 가장 널리 알려져 있는 대화 시스템 중의 하나이다. 이 시스템은 자연 발성언어(spontaneous speech)를 이해하고 응답해주는 작업 중심적 음성 대화 시스템(task-oriented speech dialogue system)으로서 화자독립(speaker-independent)의 키워드를 기반(keywork-based)으로 한 자연 발성 이해방법을 사용하여 음식을 주문받는 작업을 수행할 수 있도록 개발되었다. 사용자의 의도를 이해하기 위해 잡음에 견고한 키워드추출기(noise-robust keyword-spotter), 의미분석용 키워드 격자 구문분석기(a semantic keyword lattice parser), 사용자주도 대화 관리기(user-initiated dialogue manager), 다형식 응답 제조기(multimodal response generator)로 구성되어 있다. 키워드추출기에 의하여 키워드가 추출되면 키워드 격자 구문분석기를 이용하여 입력된 음성의 의미를 분석하게 된다. 그리고, 이전에 이루어졌던 대화의 내용과 문맥을 참조하여 대화관리기는 입력된 음성의 의미적인 내용을 해석하게 된다. 해석이 모호하거나 불분명한 경우 대화관리기는 사용자에게 확인을 요구하게 된다. 시스템의 응답은 사용자에게 친근하면서도 시스템이 이해하고 있는 내용을 잘 전달하기 위해 합성된 음성만을 이용하지 않고 음성, 텍스트, 그림으로된 얼굴 표정, 주문된 음식의 그림 등이 사용되었다.

(2) 일본의 NTT의 K. Arai, M. Kitai 등에 의해서 구현된 음성 호출 응답 시스템(voice activated call answering system)[31] 상에서 구현된 대화처리부는 인식의 오류를 감소시키기 위한 목적으로 만들어졌다. 여기서는 시스템이 사용자에게 하는 질문의 형식을 최적화시킴으로서 사용자 시스템에게 가장 적합한 형식으로 응답을 하게 하였다. 시스템의 질문은 사용자에게 사용

하는 단어를 제한하도록 만든다. 또, 사용자가 시스템에 적합한 형태로 정보를 제공하도록 만들며, 강조하고자 하는 단어의 전후에 휴지기간을 두며 발음을 하도록 유도한다. 이와 유사하게 시스템의 인식을 돕기 위한 방법으로 같은 NTT의 H. Nishi, M. Kitai는 시스템이 사용자에게 확인을 요구하는 후보 어휘들에 대하여 제한을 두는 방법을 연구하였다[32]. 또, 미국에서도 Ronald A. Cole, David G. Novick에 의해서 사용자의 응답 형식을 유도하는 방법이 연구되었다[33].

(3) T. Yamamoto, Y. Ohta에 의해서 구현된 MASCOTS[34,35]이라 불리는 대화관리 시스템은 사용자와 문제해결 시스템(problem solving system) 사이의 대화를 관리하고 음성이해 시스템(speech understanding system)이 언어를 처리하는 것을 돕고 있다. 이 시스템에서는 특별히 사용자의 다음 발화 내용을 예측하고자 노력하고 있다. 이 시스템에서는 SR-계획(SR-plan)이라 불리는 계획 정보와 시스템 스택 및 사용자 스택의 두 스택을 가지고 있는 구조를 취하고 있다. SR-계획의 정보와 두 스택의 내용을 사용하여 시스템은 사용자가 다음에 하고자 하는 발화의 내용을 미리 예측함으로써 언어 처리를 효과적으로 하고 문제 해결 시스템을 돕고 있다. MASCOTS에는 발화의 종류를 확인하고 문제해결 시스템에서 처리하기에 유용한 표현 형식(representation form)으로 표준화하는 과정이 포함되어 있다.

(4) MIT의 Victor Zue, Stephanie Seneff 등이 개발한 VOYAGER 음성언어 시스템은 사용자들이 여러 언어로 사용할 수 있도록 지원하고자 하고 있다. 여기서는 시내의 교통 안내와 여행 계획에 관한 영역으로 제한된 범위 내에서 각 언어로된 입력으로부터 공통적인 의미 표현(semantic representation)을 추출하게 된다. 이 시스템은 언어 의존적인 정보(language dependent information)와 다른 부분을 가능한한 분리하고 있다. 즉, 내부 시스템의 관리기(internal system manager), 담화와 대화 부분(discourse and dialogue component), 데이타베이스들이 언어와 독립적으로 관리된다. 이 VOYAGER 시스템은 현재 일본어로 사용가능하고, 프랑스어, 이태리어, 독일어 등에 대해서도 연구 중이다.

## 5. 결 론

본 고에서는 음성인식의 기본 과정과 최근에 주로 사용되고 있는 인식 알고리즘 및 대화 시스템을 소개하였다. 현재 대용량의 음성인식 시스템은 주로 HMM을 기본으로 하여 개발되고 있지만, 많은 알고리즘이 각각의 장단점을 가지고 있고, 응용분야에 따라 서로 다른 성능을 보여주고 있으므로, 응용분야에 적합한 인식방법을 선택하는 일도 중요하다고 할 수 있다. 최근에는 여러가지 인식 방법을 조합하여 각각의 장점을 이용하려는 시도가 증가하고 있다. 신경회로망과 HMM[36], 신경회로망과 퍼지이론의 조합 등 두가지 방법 이상의 조합이 많이 발표되고 있다. 한편, 음성인식에 관한 연구는 인공지능의 분야뿐만 아니라, 신호처리, 음향학, 언어학, 자연어처리, 지식공학, 정보이론, 생물학 등 다양한 분야의 지식을 필요로 하며, 이중 부족한 부분이 있다면 만족할 만한 성능을 기대하기 힘들다. 그러나, 이들을 모두 동시에 연구하기는 어려우며, 현재 서로 독립적으로 연구되고 있어, 다양한 지식을 이용하기 힘든 실정이다. 따라서, 각 분야의 연구자 간의 공동연구의 필요성이 증대되고 있으나, 각 분야의 관점의 차이로 완벽한 협동이 어렵다. 보다 진보된 연구를 위해서는 서로 다른 분야에 관심을 가지는 것도 중요하다고 판단된다.

## 참고문헌

[1] Alexander I. Rudnicky, Alexander G. Hauptmann, and Kai-Fu Lee, "Survey of Current Speech Technology," Communications of the ACM, Vol. 37, No. 3, pp. 52~57, March 1994.

[2] A. Waibel, M. Woszczyna, "Recent Advances in JANUS: A Speech Translation System," ATR International Workshop on Speech Translation, IWST'93, 199.

[3] Sadaoki Furui, Digital Speech Processing, Synthesis, and Recognition, Marcel Dekker, INC.,

1992.

[4] John R. Deller, John G. Proakis, John H. L. Hansen, Discrete-time processing of speech signals, Macmillan Publishing Company, 1993.

[5] Shuzo Saito, Kazuo Nakata, Fundamentals of Speech Signal Processing, Academic Press, 1985.

[6] Hynek Hermansky, "Perceptual linear predictive(PLP) analysis of speech," Journal of Acoustical Society of America, pp.1738~1752, 1990.

[7] Kai-Fu Lee, "Automatic Speech Recognition: The Development of the SPHINX system," Kluwer Academic Publishers, Boston, 1989.

[8] Hajin Yu, Yung Hwan Oh, Yoichi Yamashita, Riichiro Mizoguchi, "Fuzzy Expert System for *Continuous Speech Recognition*," *Expert System with Application: An International Journal* Vol. 9, No. 2, 1995 (to be published).

[9] Harvey F. Silverman and David P. Morgan, "The Application of Dynamic Programming to connected Speech Recognition," IEEE ASSP Magazine, pp. 6~25, July 1990.

[10] Joseph Picone, "Continuous Speech Recognition Using Hidden Markov Models," IEEE ASSP Magazine, pp. 26~41, July 1990.

[11] Robert M. Gray, "Vector Quantization," IEEE ASSP MAGAZINE APRIL 1984.

[12] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic Modeling of Large Vocabulary Speech Recognition," Computer Speech and Language, Vol. 4, pp. 127~165, 1990.

[13] A. Averbuch, et. al. "An IBM-PC Based Large-Vocabulary Isolated Utterance Speech Recognizer," Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 53~56, Tokyo, Japan, April 1986.

[14] T. Kohonen, "Self-organizing feature maps," in Proc. IEEE Second International Conference on Neural Networks, San Diago, CA, Vol. I., tutorial 10, p. 13, July 1988.

[15] A. Waibel, T. Hanazawa, G.E. Hinon, K.Shikano *and K.J.Lang, "Phoneme recognition using* time-delay neural networks," IEEE Transacions on Acoustic, Speech, and Signal Processing, Vol. 37(3), pp. 328~339, March 1989.

[16] Ken-ichi Iso and Takao Watanabe, "Speaker-Independent Speech Recognition Using a Neural Prediction Model," 전자정보통신학회논문지 (일본), Vol. J73-D-II, No. 8, pp. 1316~1321, 1990년 8월.

[17] J.L Elman, "Finding structure in time," CRL Technical Report No. 8801, University of California, San Diego, 1988.

[18] Zue, V, and Lamel, L, "An expert spectrogram reader : A knowledge-based approach to speech recognition," Proceedings of International Conference on Acoustics, Speech, and Signal Processing '86, pp. 1197~1200, 1986.

[19] Mizoguchi, R., Tsujino, K., and Kakusho, O., "A Continuous Speech Recognition System *Based on Knowledge Engineering* Techniques," Proceedings of International Conference on Acoustics, Speech, and Signal Processing '86, pp. 1221~1224, 1986.

[20] H. J. Zimmermann, "Fuzzy Set Theory and Its Applications," Kluwer-Nijhoff Publishing, Boston, P363, 1986.

[21] Sankar K. Pal and Dwijesh Dutta Majumder, "Fuzzy Sets and Decisionmaking Approaches in Vowel and Speaker Recognition," IEEE Transactions on Systems, Man, and Cybernetics, pp. 625~629, August 1977.

[22] S. K. Pal and B. Chakraborty, "Intraclass and Interclass Ambiguities (Fuzziness) in feature evaluation," Pattern Recognition Letter, Vol. 2, pp. 275~279, 1984.

[23] Renato de Mori and Pietro Laface, "Use of Fuzzy Algorighms for Phonetic and Phonemic Labeling of Continuous Speech," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 2, pp. 136~148, March 1980.

[24] G. Hirsch, M. Lamotte, M. T. Mas and M. J. Vigneron, "Phoneme Classification Using a Fuzzy Dissimilitude Relation," Fuzzy Sets and Systems, Vol. 5, pp. 167~175, 1981.

[25] Michel Lamotte, Lucien Bour, Gerard Hirsch, "Fuzzy Phoneme Recognition," Fuzzy Sets and Systems, Vol. 28, pp. 363~374, 1988.

[26] Francisco Casacuberta and Enrique Vidal, "Interpretation of Fuzzy Data by Means of Fuzzy

Rules with Applications to Speech Recognition," Fuzzy Sets and Systems, Vol. 23, pp. 371~389, 1987.

[27] Jun-ichiroh Fujimoto, Tomofumi Nakatani and Masahide Yoneyama, "Speaker-independent Word Recognition Using Fuzzy Pattern Matching," Fuzzy Sets and Systems, Vol. 32, pp. 181~191, 1989.

[28] H. Iida, "Prospects for Advanced Spoken Dialogue Processing", IEICE Transactions on Information and Systems, Vol.E76-D, No.1, pp. 2~8, 1993.

[29] Yoichi Takebayashi, Hiroyuki Tsuboi, Hiroshi Kanazawa, 1"A Real-Time Speech Dialogue System Using Spontaneous Speech Understanding", IEICE Transactions on Information and Systems, Vol.E76-D, No.1, pp. 112~120, 1993.

[30] Sgugebibu Seto, Yoshifumi Nagata, Hiroshi Kanazawa, "Spontaneous Speech Dialogue System TOSBURG II and its Evaluation", Proceedings ISSD-93 International Symposium on Spoken Dialogue, pp. 41~44, 1993.

[31] K. Arai, M. Kitai, S. Nakajima, H. Nishi, "Effects of Question Style on Speech Dialog", Technical Report of IEICE. SP93-100 (1993-11).

[32] Hiroyuki Nishi, Mikio Kitai, "A New Confirmation Method using the Statistical Relationship between Likelihood and Accuracy", Technical Report of IEICE. SP93-101 (1993-11).

[33] Ronald. A. Cole, David, G. Novick, "Rapid Prototyping of Spoken Language Systems: The Year 2000 Census Project. Proceedings ISSD-93 International Symposium on Spoken Dialogue, pp. 19~23, 1993.

[34] T. Yamamoto, Y. Ohta, Y. Yamashita, O. Kakusho, R. Mizoguchi, "MASCOTS: Dialog Management System for Speech Understanding System", IEICE Transactions on Information and Systems, Vol.E74, No.7, pp. 36~43, 1991.

[35] Y. Yamashita, H. Yoshida, T. Hiramatsu, Y. Nomura, R. Mizoguchi, "MASCOTS II: A Dialog Manager neral Interface for Speech Input and Output", IEICE Transactions on Inf ormation and Systems, Vol. E76-D, No. 1, pp. 74~83, 1993.

[36] G. Zavaliagkos, Y. Zhao, R. Schwartz, and J. Kakhoul, "A hyubrid segmental neural net/hidden Markov model system for continous speech recogntion," IEEE transactions on Speech and Audio Processing, Vol. 2, No. 1, pp. 151-160, 1994

**오 영 환**

1972 서울대학교 공과대학(공학사)
1974 서울대학교 교육대학교(석사)
1980 Tokyo Institute of Technology 정보공학(박사)
1981 충북대학교 공과대학 조교수
1981 서울대학교 공과대학 강사
1983 Univ. of California, Davis연구 교수
1985 한국과학기술원 전산학과 교수
관심분야 : 음성인식, 음성합성, 패턴인식 등

**유 하 진**

1990 한국과학기술원 전산학과 학사
1992 한국과학기술원 전산학과 석사
1992 ~현재 한국과학기술원 전산학과 박사과정
관심 분야 : 음성인식, 퍼지이론, 신경회로망