特輯

음성인식기술의 현황과 전망

丘 明 完
韓國通信 소프트웨어研究所 自動通譯研究

Ⅰ. 서론

음성은 인간사이의 중요한 통신 수단이지만 인간과 기계 사이의 주요 통신 수단은 되지 못했다. 그 이유는 아직까지 음성인식 기술이 자유롭게 기계와 사람 사이의 통신을 할수 있을 정도로 발전되지 못하였기 때문이다. 음성인식이란 시스템에 입력된 음성을 정확히 인식하여 문자로 바꾸어 주거나 혹은 입력된 음성을 이해하여 적절히 요구에 대응하여 주는 것을 말한다. 음성을 정확히 인식하여 문자로 변환시켜 주는 일은 주로 제한되고 고립된 단어 인식시스템에서 처리되며 음성을 이해하는 일은 주로 대용량 연속단어 인식시스템에서 수행된다.

음성인식기술은 1970년 초부터 연구되어 왔지만 현재까지 사람의 음성을 정확히 인식할 수 있는 시스템은 개발되어 있지 않다. 현재까지 개발된 시스템은 다음과 같은 네가지의 제약성을 가지고 부분적으로 음성을 어느정도 인식하고 있는 실정이다. 첫번째의 제약으로서 화자 독립성(speaker independence)을 들 수 있다. 대부분의 음성인식 시스템은 특정화자 전용으로 사용될 경우 높은 인식률을 나타내지만 불특정 화자일 경우는 인식률이 특정화자 인식시스템에 비하여 떨어진다. 이런 현상의 중요한 이유중의 하나는 음성인식을 위하여 음성으로부터 추출하는 음성특징이 화자에 너무 종속된다는 것이다. 즉 동일한 단어를 여러사람이 발음하였더라도 현재의 기술로 추출된 음성특징은 사람에 따라서 많은 차이점이 있다는 것이다. 두번째 제약은 연속음성 인식이다. 연속음성 인식은 고립단어 인식에 비하여 훨씬 어렵다. 왜냐하면 연속음성은 고립단어에 비하여 단어구간을 알기 어려우며 조음현상을 규명하기도 어렵기 때문이다. 그러나 사람과의 통신은 주로 연속음성형태로 이루어지므로 연속음성 인식시스템을 구현하는 것이 필요하다.

세번째의 제약으로는 대용량 단어 인식시스템을 구성하는 것이다. 대용량이란 대략 1000단어 이상을 말하는데 단어가 증가할수록 유사한 단어가 많아지며 인식시스템 훈련을 위한 데이타 베이스도 상대적으로 증가된다. 이러한 현상을 극복하기 위하여 기본 유니트로서 단어 대신 음소와 같은 서브워드 유니트(subword unit)를 사용한다. 그러나 서브워드 유니트는 조음현상을 나타내기가 단어보다는 좋지 않기 때문에 통상 인식률이 저하된다. 마지막의 제약으로서 무제한 문법(unconstrained grammar)을 들 수 있다. 보통 음성은 특정한 문법에 국한되어 있지않다.

그러나 현재의 음성인식 시스템은 제한된 문법을 사용함으로써 대용량 단어 인식 시스템의 인식률을 높히고 있다. 제한된 문법이란 인식시스템의 대상을 특정업무로 국한시켜 입력 가능한 문장의 틀을 고정시키는 것이다.

본 고에서는 최근의 음성인식 기술의 현황을 파악하고 앞으로의 발전 방향을 전망하고자 한다. 먼저 Ⅱ장에서는 음성인식연구 현황, 음성인식시스템 기술 및 개발 사례를 살펴보고 음성인식시스템 개발에 필수적인 특징추출, 인식 알고리즘 및 언어처리 알고리즘에 대한 기술 현황을 파악한다. Ⅲ장에서는 음성인식 기술의 기술적 측면과 음성인식 기술을 이용한 응용시스템 측면으로 나누어서 음성인식 기술의 향후 전망을 예상하고자 한다. 그리고 마지막으로 Ⅳ장에서 결론을 맺는다.

Ⅱ. 음성인식 기술의 현황

1. 음성인식연구 현황

1) 미국

미국의 음성인식연구는 국방성의 주도로 연구되고 있다. 1971년에서 1976년까지 SUR(speech understanding research)이라는 음성이해연구 프로젝트가 수행되었으며 최근에는 1984년부터는 5년에서 10년 기간으로 음성 및 자연언어처리에 관한 새로운 프로젝트가 수행되고 있다.

이 프로젝트는 크게 음성언어 프로그램(spoken language program)과 문자언어 프로그램(written language program)으로 나누어진다. 음성언어 프로그램은 대용량 음성인식시스템과 음성언어 이해에 관한 연구를 추축으로하여 특정 task 영역에서 자연스러운 음성을 실시간으로 인식하는 화자독립 혹은 화자적응 음성인식시스템을 개발하는 것을 목표로 한다. 시스템의 성능평가를 위하여 낭독체(reading speech) 연속음성으로 구성된 RM(resource management) 데이타베이스와 항공기 예약에 관련된 회화체(spontaneous speech) 연속음성으로 구성된 ATIS(air travel information system) 데이타베이스를 사용한다. 이 프로그램에 참가하고 있는 기관은 BBN, Brown 대학, Boston 대학, CMU 대학, Drag-on, Lincoln, MIT, SRI, Texas Instruments, Unisys, AT&T등 이다. 각 기관에서 개발한 음성인식 시스템는 동일한 음성 데이타베이스를 사용하여 성능을 비교하며, 최근의 성능비교 결과가 표 1.에 나타나있다. [1]

표 1의 결과에 따르면 낭독체 음성인식 시스템의 성능은 CMU의 음성인식 시스템이 96.4%의 인식률로 가장 우수하였으며 회화체 음성인식 시스템은 BBN의 인식시스템이 84.3%로 가장 우수하였다. 그리고 자연언어처리 기술과 통합된 음성이해 인식시스템의 성능은 SRI의 시스템이 58.6%로 가장 우수하였다.

문자언어 프로그램은 대용량 텍스트(text)처리에 필요한 기술을 개발하는 것을 목표로 하며 메시지 이해(message understanding), 자연언어학습(natural language learning) 및 데이타베이스구축등에 관한 연구를 한다. 또한 기계번역에 관한 연구도 포함한다. 현재 이 프로그램에 참가하고 있는 기관은 BBN, Columbia 대학, New Mexico State 대학, Pennsylvania 대학, Rochester 대학, SRI, University of California at Berkeley 등이 있다.

표 1. DARPA 음성인식시스템의 성능비교

음성 데이타베이스 종류	인식률(%)	기관
RM (낭독체 연속음성인식)	96.4	CMU
	96.2	BBN
	95.6	MIT
	95.5	AT&T
ATIS (회화체 연속음성인식) 음성인식 부문	84.3	BBN
	82.0	SRI
	73.9	MIT
	71.0	CMU
ATIS (회화체 연속음성인식) 음성인식 + 자연언어처리부문	58.6	SRI
	42.8	BBN
	34.5	CMU
	18.6	MIT

2) 일본

일본에서의 음성인식기술은 1982년부터 추진한 제5세대 컴퓨터 프로젝트의 일부인 "음성과 자연언어를 통한 컴퓨터 입출력"이라는 제목으로 연구가 진행되었으나 연구결과의 대외발표는 거의 없었다. 최근에서의 음성인식 관련 프로젝트는 ATR(advanced telecommunitions research institute) 산하 자동통역연구소에서 1986년부터 수행하고 있는 자동통역전화(automatic telephone interpretation) 프로젝트와 1987년부터 교육, 과학, 문화성의 자금지원을 받고 있는 "Advaced man-machine interface through spoken language"이라는 국가 프로젝트가 있다.

자동통역전화 프로젝트는 1993년 1월, 7년 동안 수행하여온 연구결과인 자동통역전화 실험시스템을 데모하는 것으로 1단계 연구를 끝내고 음성번역통신연구소를 새로 만들어서 2단계 연구를 시작하였다. 자동통역전화 실험은 일본, 미국, 독일 사이의 국제회의에 관한 문의내용에 대한 것인데 세계 최초로 국제전화를 사용한 음성번역 실험이었다는 면에서 의의가 있었다.

국가 프로젝트는 음성에 관한 기술을 분석, 특징추출, 인식, 이해, 합성, 지식처리, 잡음에서의 음성처

리 및 평가기술등 8가지의 핵심기술로 나누어서 약 185명의 연구자가 연구를 수행하고 있다.

3) 유럽

유럽에서의 음성인식 기술연구는 유럽국가들이 모여서 공동으로 수행하는 연구와 각 나라에서 자체적으로 수행하는 연구로 나누어진다. 범 유럽국가들이 수행하는 연구로는 ESPRIT(European strategic program for research and development in information technology)라는 정보통신에 관련된 유럽국가들의 공동 프로그램이 있다. 이 프로그램은 ESPRIT Ⅰ(1984~1989), ESPRIT Ⅱ(1988~1993), ESPRIT Ⅲ(1992~1997)의 세 단계로 나누어서 진행되는데 음성인식에 관한 연구는 매 단계마다 주요 추진 과제였다. 각 단계에서 음성인식에 관

표 2. 음성인식에 관련된 ESPRIT 프로젝트 내용

연구단계	Project 이름	내 용
I (1984-1989)	SIP (advanced algorthms and architetures for speech and image processing)	o 음성 및 화상신호을 인식하고 이해하기 위한 알고리즘 및 구조 개발, 적당한 응용사례 제시 o 목표 : 1000단어 연속음성인식 시스템 개발
	IKAROS (intellegence and knowledge-aided recognition of speech)	o 음성이해를 위한 인공지능 기술 개발 o 목표 : 대화관리 기능을 갖으며 다국어(불어, 영어, 독어), 다화자 인식이 가능한 1000단어 연속음성 인식 시스템 개발
	SAM 1 (multilingual speech input-output assessment methodology and standard	o 음성기술의 평가를 위한 범 유럽의 기반 조성 o 목표 : 다국어 EUROM database를 CD Rom으로 제작 배포
II (1988-1993)	SUNDIAL (speech understanding and dialogue)	o 정보통신 서비스 구현에 필요한 컴퓨터와의 정합을 위하여 음성인식 기술을 이용한 대화에 관한 연구 o 목표 : 4개국어(영어, 불어, 독어, 이탈리아어)를 이해하고 전화를 통하여 자연스럽게 발음한 1000~2000단어급의 연속음성을 인식하는 시스템 개발 o 정보서비스 응용 : 호텔업무(이탈리아어), 항공기 예약(영어, 불어), 기차 시간표 안내(독일어)
	SUNSTAR (integration and design of speech understanding interface)	o 음성 입·출력을 사용한 human computer interface 장점을 연구 o 목표 : 전화망 및 전문 OA 환경하에서 음성 입·출력을 사용하는 데모시스템 개발
	SAM 2 (speech assement methodology)	o 음성 입·출력 평가 및 데이타베이스 구축 tool 개발
	POLYGLOT (multilanguage speech-to-text and text-to-speech system)	o 다국어 음성 입·출력의 타당성 검토 o 목표 : 원거리 전자 우편함 검색 및 전화번호부의 음성검색을 할수있는 유럽 6개 국어에 대한 대용량 화자적응 고립단어 인식시스템과 합성시스템 개발
	ARS (adverse recognition of speech)	o 잡음이 있는 음성의 인식 알고리즘 개선과 실시간 데모 시스템 개발 o 목표 : 자동차 및 공장에서의 음성을 인식하는 시스템 개발 및 음성 데이타베이스 구축

련된 주요 프로젝트의 내용이 표 2에 나타나 있다.

한편 영국에서는 국가주도의 Alvey program 내에서 국가 연구소와 산업체 연구소가 협력하여 음성인식 관련연구를 수행하였으며 현재는 ITI(information technology initiative) 프로젝트가 시작되어 음성인식 및 데이타베이스 구축에 대한 연구가 진행되고 있다. 프랑스에서는 CNRS(national research agency)와 상공부에서 후원하는 "Human-machine communication" 이라는 프로젝트에서 음성통신, 자연어 처리에 관한 연구를 수행하고 있다. 독일에서는 SPICOS(SiemensPhilips-IPO continuous speech recognition)라는 대형 프로젝트에서 연속음성인식 기술에 관한 연구를 수행하였으며 최근 1991년 1월부터 ASL(architecture for speech and language research)라고 불리우는 새로운 프로젝트가 4년 계획으로 시작되어 음성 및 텍스트 데이타베이스 구성 및 대용량 음성인식 알고리즘 개발에 역점을 두고 있다. 이 프로젝트의 연구결과는 실시간으로 음성의 자동통역을 실현하는 VERBMOBIL이라는 야심찬 프로젝트에 사용될 것이다. VERBMOBIL은 1991년부터 시작되어 20년간 지속될 대형 프로젝트이다.

4) 국내

국내에서의 음성인식 연구는 1980년 초부터 일부 대학을 중심으로 연구가 수행되었으며 최근에는 많은 대학과 연구소를 중심으로 활발히 진행되고 있다. 그러나 연구내용은 아직 수십단어 혹은 수백단어를 인식하는 고립단어 음성인식 시스템 개발의 수준에 머물러 있다. 1991년 부터는 한국통신과 전자통신연구소가 공동으로 자동통역전화 요소기술연구를 수행하고 있으며 이 연구결과는 향후 한 일간 자동통역전화 시스템 개발에 이용될 것이다. 최근에는 기업체에서도 음성인식 기술을 이용한 여러가지 제품개발을 시도하고 있다.

2. 음성인식 시스템

현재까지의 음성인식시스템은 그림 1에 나타나 있듯이 기본적으로 음성으로부터 음성패턴(단어, 음소 등)의 특징을 추출하여 기준 음성패턴을 만드는 훈련과정과 미지의 음성이 입력되면 저장된 기준음성 패턴의 특징과 비교하여 가장 유사한 기준 음성패턴을 찾아내는 인식과정으로 나눌 수 있다. 이러한 알고리즘은 일반적으로 pattern matching 알고리즘이라고 부른다.

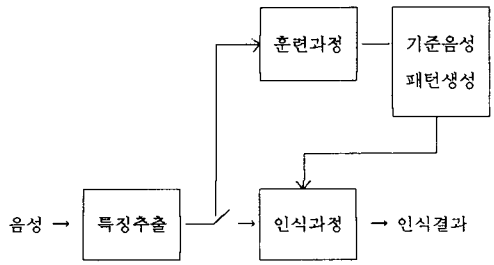

그림 1. 음성인식 시스템 개념도

음성인식 시스템은 입력 음성의 특성(고립단어, 연결단어, 연속음성), 화자(화자종속, 화자독립) 및 어휘량에 따라 구분할 수 있다. 현재 상용화되고 있는 음성인식 시스템을 표 3에, 연구기관이나 대학에서 최근 개발된 음성인식 시스템을 표 4에 나타내었다.[2] 상용화된 제품과 연구실에서 개발된 음성인식 시스템의 가장 큰 차이점은 상용화된 제품은 주로 DTW 알고리즘에 근거하였으며 고립단어 및 화자종속 인식시스템이 대부분이나 연구실내의 음성인식 시스템들은 대용량 연속 음성인식 시스템들이 주축을 이루고 있다.

표 3. 상용 음성인식 시스템

제조업체	국 명	시스템명	특 징	단어수	인식률(%)
Dragon System	미국	Voice Scribe 1000 Dragon Dictate	화자 종속 고립 단어 화자 적응 고립 단어	1000 30,000	
IBM	미국	Voice command	화자 종속 고립 단어	64	95~98
Kurzweil Applied Intelligence	미국		화자 종속 고립 단어	1000	
NEC Ameria	미국	SAR-10 SR-10 DP-200	화자 종속 고립 단어 화자 종속 고립 단어 화자 종속 연결 단어	250 128 150	98 98
Texas Instrument	미국	Speech Command	화자 종속 연속 음성	1000	
VOTAN	미국	Voice-Card Voice-Card Voice Key	화자 독립 연속 음성 화자 종속 연속 음성 화자 종속 고립 단어	13 640 64	93 98
MARCONI	영국	ASR-1000	화자 종속 연속 음성	200	
VECSYS	프랑스	DATAVOX	고립 단어 연결 단어	5000 300	

표 4. 최근에 개발된 음성인식 시스템

연 구 실	국 명	시 스 템 명	특 징	단 어 수	인식률(%)
CMU	미국	SPHINX	화자 독립 연속 음성	1000	96.4
SRI	미국	DECIPHER	화자 독립 연속 음성	1000	95.2
BBN	미국	BYBLOS	화자 종속 연속 음성 화자 적응	1000	98.7 94.8
Lincoln Lab	미국		화자 독립 연속 음성	1000	87.4
ATR	일본		화자 종속 연속 음성 화자 적응	1035	95.3 89.7
IBM	미국	Tangora	화자 종속 고립 단어	20,000	95
NEC	일본		화자 종속 고립 단어	1,800	97.5

3. 특징 추출

음성인식을 위한 pattern matching 알고리즘은 음성 패턴의 특징이 발성자 및 발음시간에 따라 변하는 것이 아니라 음성의 의미에 따라서만 변한다는 가정을 전제로 한다. 그러므로 동일한 의미를 갖는 음성을 여러사람이 발음하더라도 각 음성으로부터 추출한 음성특징은 동일하여야 한다.그러나 현재 이와 같은 특징이 무엇인지는 정확히 알려져 있지 않다. 음성으로부터 특징을 추출하는 방법은 크게 네가지로 나눌수가 있다.

첫번째로 음성파형 자체를 하나의 특징으로 생각하는 것이다. 그러나 음성파형은 시간축에 기준하여 많은 변화량을 갖고 있으며 데이타 양도 많으므로 주파수 영역으로 변환시켜 특징을 추출하는 방식을 사용한다. 주파수 영역에서 특징을 추출하기 위해서 Fourier 변환을 이용한다. Fourier 변환은 시간축에서 안정된(stationary) 신호를 분석하는데 주효한데 음성신호는 실제로 이러한 성질을 만족하지 못한다. 그래서 음성신호를 주파수 영역으로 변환시킬 때에는 안정된 특성을 어느 정도 만족할 수 있는 구간(예를 들면 10~30msec) 단위로 분석한다. 주파수 영역의 특징을 추출하기 위해서 FFT(fast fourier transform)을 사용한다.

두번째 방법은 음성이 구강(vocal tract)으로부터 발성된다는 사실을 근거로 구강의 형태를 필터(filter)로 가정하고 그 필터 계수를 음성의 특징으로 삼는 것이다. 일반적으로 필터는 AR(Auto Regressive)모델 혹은 ARMA(auto regressive moving average)모델에 의해 구성된다. AR 모델의 대표적인 것이 LPC(linear predictive coding)방식을 사용하는 것인데 이 방식은 모든 음성이 구강의 모양에 따라 구분될 수 있으며 구강의 형태는 혀의 위치에 따라 변한다는 가정을 이용하여 구강의 형태를 all pole filter로 모델링하는 것이다. 실제로 all pole filter의 pole은 음성 스펙트럼상의 피크인 포만트(formant)를 나타낸다. 그러나 음성중 자음과 비음등은 all pole filter로 모델링이 잘 되지 않는다. 특히 비음은 주파수 영역에서 zero 특성을 가지고 있는데 이를 위해서 ARMA 모델방식이 사용 되기도 한다. 그러나 ARMA 모델은 계산량이 많아서 음성인식을 위해서는 주로 AR모델을 이용한다.

세번째 방법은 귀가 음성을 분석하는 방식을 이용하는 auditory 분석방식이다. 인간의 귀는 외이, 중이, 내이로 이루어져 있는데 음성이 들어오면 주파수 영역으로 변환시켜 뇌로 전달되는 것으로 알려져 있다. 이때 저주파 영역에서는 상세히 분석하고 고주파수 영역에서는 상대적으로 개략적인 분석을 한다. 이러한 사실을 실험적으로 측정하여 주파수 영역내의 weighting 함수를 구하여 Bark scale 혹은 mel scale이라 명명하였다. 이러한 scale에 따라서 음성의 특징을 추출하는 방법에는 FFT에 의하여 주파수 영역으로 변환시킬 때 weighting을 가하는 방식과 LPC에 의해 추출된 파라미터를 weighting시켜 파라미터로 추출하는 방식이 있다. [3]

네번째 방법은 동적 특징(dynamic feature)을 주파수 영역의 특징(spectral feature)들과 동시에 사용하는 것이다. 앞에서 설명한 음성특징은 모두 주파수 영역의 특징들인데 시간에 따른 주파수 영역의 특징 차(differential spectral feature)를 주파수 영역의 특징들과 동시에 같이 사용하기도 한다. 또한 음성의 세기(stress)와 억양(intonation) 등으로 특징지어지는 운율(prosody)를 음성특징으로 병행하여 사용한다. 음성의 운율은 시간에 따른 에너지, 에너지 차(differential energy) 및 피치(pitch) 등으로 표현되는데 최근의 연구결과에 의하면 mel scale된 LPC cepstrum, LPC cepstrum의 차(differential LPC cepstrum), energy 및 energy 차를 음성 특징으로 사용하였을 때 높은 인식율을 얻었다고 한다. [3]

4. 인식 알고리즘

음성인식 알고리즘은 그림 1의 훈련과정과 인식과정에 사용되는 알고리즘으로서 크게 DTW(dynamic time warping), HMM(hidden markov model) 및 neural network 등으로 나눌 수 있다. 각 알고리즘에 대한 상세한 설명은 다음과 같다.

1) DTW 알고리즘

DTW 알고리즘은 인식과정에서 사용되는 알고리즘으로서 입력 음성패턴과 기준음성패턴간에 거리를 측정할 때 dynamic programming의 기법을 이용한다. 음성은 동일한 사람이 같은 단어를 여러번 발음하더라도 음성 특징이 각기 달라지며 특히 감정, 분위기에 따라 발음 지속 시간(프레임 길이)이 달라진다. 기준 음성패턴과 입력 음성패턴의 발음 시간의 차이가 있을 경우 두 패턴사이의 거리(distance)를 측정하기 위해서 우선 기준 음성패턴의 각 프레임과

그에 대응하는 입력 음성패턴의 프레임 번호 사이의 쌍(pair)을 찾아야 한다. 이 대응쌍은 warping함수에 의하여 구해지며 이때 dynamic programing기법이 이용된다. Dynamic programing 기법에 따르면 warping 함수에 의해 구해진 경로는 모든 경로에 의한 거리중 최단의 경로라는 것을 전제로 한다.

DTW 알고리즘을 이용한 음성인식 시스템은 고립단어 인식에 주로 이용되며 대상단어가 소용량이며 인식시간이 많이 소요된다는 단점이 있지만 인식률이 높기 때문에 VLSI 기술에 의해 chip으로 제작되어 현재 많이 상용화되어 있다. 또한 기준패턴을 쉽게 만들 수 있기 때문에 사용자의 요구에 따라 음성인식 시스템의 업무내용을 용이하게 변경할 수 있다.

2) HMM 알고리즘

HMM 알고리즘은 음성인식 시스템 개념도에서 훈련과정 및 인식과정을 수행하는 알고리즘으로서 1970년말 부터 음성인식 알고리즘으로 많이 사용되었다. 최근에는 높은 인식률과 빠른 인식시간 때문에 대용량 음성인식 시스템에 많이 사용되고 있다. HMM 알고리즘의 기본적인 사상은 음성이 Markov 모델로 모델링될 수 있다는 가정하에 훈련과정에서 Markov 모델의 파라미터를 얻어 기준 Markov 모델을 만들고 인식과정에서는 입력음성과 가장 유사한 기준 Markov 모델을 찾아냄으로써 인식한다. Markov 모델로서 hidden Markov 모델을 사용하는데 그 이유는 음성패턴의 다양한 변화를 수용하기 위해서이다. Hidden Markov 모델이란 이중 stochastic process로서 state 선정에 관한 stochastic process와 매 state마다 음성패턴이 발생될 출력 확률(output probability)에 관한 stochastic process로 구성된다. 즉 음성패턴의 각 특징을 state의 선정확률과 출력확률등으로 표현하여 준다. 여기서 hidden이란 의미는 state가 음성패턴에 관계없이 모델속에 숨어있다는 것을 말한다.

HMM 알고리즘은 기준패턴을 음소, 음절 등과 같이 단어 이하의 발음 길이를 갖는 패턴으로 설정할 수 있으며 입력음성으로 단어, 문장들을 입력할 수 있기 때문에 대용량 음성인식 시스템에 주로 이용된다. 만약 1000 단어 음성 인식 시스템에서 기준패턴의 기본단위로 단어를 선정하였다면 1000개의 단위가 필요하지만 음소를 선정하였다면 40~50개의 음소의 파라미터만 저장하면 된다.

3) Neural Network

Neural network는 인간의 뇌세포를 간단히 모델링하고 모델된 뇌세포들을 연결시켜줌으로써 인간의 뇌가 하는 역할을 수행시켜 주는 알고리즘이다. 현재까지 개발된 neural network 알고리즘은 훈련시키는 방식에 따라 크게 supervised learning neural network와 unsupervised learning neural network로 나눌 수 있다.

Supervised learning이란 neural network를 훈련시킬 때 훈련데이타(training data)와 훈련데이타의 의미를 모두 사용하는 훈련방식을 말하며 대표적인 알고리즘이 single layer perceptron과 MLP (multi-layer perceptron)이다. Single layer perceptron이란 한개의 neuron을 모델링한 것이며 입력 데이타와 출력 데이타(음성인식의 경우 음성의 의미)가 주어지면 weighting 값을 설정할 수 있다. 음성인식일 경우 이 weighting 값이 곧 기준패턴이 된다. 그러나 single layer perceptron으로는 입력데이타를 분류할 수 없는 경우(exclusive OR 상태)도 있기 때문에 실제로 MLP가 음성인식 시스템에 이용된다. 이때 MLP의 입력데이타로는 3절에서 설명한 바 있는 특징추출된 데이타가 된다. 그런데 음성은 공간적인 특징(주파수에 의한 특징)과 시간적인 특징(음성 발성 시간)이 모두 포함되어 있기 때문에 MLP를 발성시간이 긴 음성인식에 이용하기 위해서는 시간적인 특징도 포함할 수 있어야 한다. 이러한 문제점을 해결하기 위한 MLP의 변형이 TDNN (time delay neural network)이다. TDNN은 음성의 특징들을 잘 분류할 수 있게끔 MLP의 입력단에 시간적인 특징도 포함되도록 한 알고리즘으로서 음소, 단어 인식에 높은 인식률을 보인다.[4] 또 다른 supervised learing 알고리즘으로서 LVQ(learning vector quantization)가 있다. LVQ는 Kohonen이 제안하였는데 인간 뇌의 특성을 고전적인 VQ방식에 적용한 것으로서 음성인식 시스템에 이용할 경우 높은 인식률을 보인다.

Unsupervised learning 알고리즘의 대표적인 것은 Kohonen의 feature map이다. Feature map 알고리즘은 음성데이타를 의미에 관계없이 훈련시키면 음소를 대표할 수 있는 특징이 저절로 나타난다는 것이다. 이 알고리즘은 실제 인간의 청각작용과 비슷하나 음성인식 시스템에 적용하였을 경우 supervised learning 알고리즘 보다는 낮은 인식률을 보인다.

음성인식을 위한 neural network의 이용은 초창기에는 neural network만 사용하여 음성인식시스템 개발을 시도하였지만 기존의 알고리즘에 비해서 월등히 높은 성능을 나타내지 못하였기 때문에 최근에는 기존의 알고리즘과 neural network 알고리즘을 결합하는 방식이 연구되고 있다.

5. 언어처리 알고리즘

음성인식 시스템에서의 언어처리 알고리즘은 일반적으로 문장을 인식할 때 사용된다. 현재 대부분의 인식 시스템은 기본적으로 단어 혹은 구를 인식한 후 언어처리 알고리즘을 사용하여 문장을 인식한다. 기준 패턴이 음소단위일지라도 단어 혹은 구를 음소의 열로 나타낼 수 있으므로 실제로 검색하는 기본단위는 단어 혹은 구로 이루어진다. 언어의 특성상 미국에서는 단어를 기본으로 탐색하고 있으며 일본에서는 구를 기본적으로 사용한다. 일단 단어 혹은 구가 인식이 되면 모든 단어의 연결을 고려하여 최적의 문장을 찾아내야 한다. 최적의 문장이란 각 단어의 인식결과로 얻어진 단어들로 이루어진 문장을 말한다. 그런데 모든 언어에는 문법이라는 것이 있어서 각 단어의 전후에 올 수 있는 단어들을 한정시킨다. 언어처리 알고리즘이란 문법과 단어(구) 인식의 결과를 연결시켜 주어 문장을 인식할 수 있게 하여주는 역할을 한다. 그러므로 문법을 고려한다면 문장의 위치에 따라 모든 후보단어들에 대해서 pattern matching작업을 수행할 필요가 없게 된다. 실제로 언어처리 알고리즘은 문장을 인식할 때 문장내에서의 단어(혹은 구)의 위치에 따라 인식대상 단어(혹은 구)를 한정함으로써 인식시간을 단축시키도록 한다.

현재 연구되고 있는 음성인식을 위한 언어처리 알고리즘은 문법을 단어(혹은 구)인식기와 결합하는 방식에 따라 통계적 모델과 구문규칙 모델로 나눌 수 있다. 통계적 모델이란 단어와 단어 사이의 연관관계를 확률적 개념으로 표현하여 매 단어 다음에 어떤 단어가 나올 수 있는지를 확률로 표시하여 문장전체를 인식할 수 있게 해 준다. 대표적인 통계적 모델로서 bigram과 trigram이란 것이 있는데 이것은 매 단어에 대해서 이전 한개(두개)의 단어가 입력되었을 때 이 단어의 발생가능성을 확률값으로 표현하여 문장인식에 사용한다. 이러한 알고리즘들은 HMM 모델에 근거한 인식시스템과 쉽게 결합할 수 있으며 회화체 음성인식에 적합하다.

구문 규칙 방식은 언어학에서 연구된 구문론(syntax)에 따라 규칙을 만들어서 매 단어 다음에 올 수 있는 단어의 종류를 규제함으로써 문장을 인식하는 방식이다.

Ⅲ. 음성인식 기술의 전망

1. 기술적 측면

음성인식 기술은 향후 다음과 같은 분야에서 활발히 연구가 진행될 것이다.

첫째로 음성언어처리(spoken language processing)연구이다. 종래에는 신호처리학자, 음성학자들이 주축이 되어 음성처리(speech processing) 부문의 연구가 진행되었고, 이와는 별도로 전산학자 및 언어처리학자들이 중심이되어 언어처리(language processing)부문에 관한 연구가 진행되어 왔다. 그러나 최근에는 음성언어처리라고 하여 음성처리와 언어처리 부문을 통합한 연구가 시작되고 있으며 앞으로 더욱 활발해질 전망이다. 여기에는 음성신호 처리에서 얻은 지식과 언어처리에서 얻은 지식을 효율적으로 결합시키는 방법 및 상호 보완을 위한 새로운 특징 사용, 그리고 인식의 실시간 처리를 위한 알고리즘 개발에 대한 연구등이 포함될 것이다. 또한 언어처리는 회화체 음성에 관한 것이 주종을 이룰 것이다.

두번째로 HMM의 성능향상에 촛점이 맞추어 질것이다. HMM 알고리즘은 음성인식 시스템에 사용되어 매우 좋은 결과를 얻었지만 아직도 보완되어야 할 부문이 많다. 이를 위해서 neural network와 동시에 사용한다거나 HMM 알고리즘의 변형인 HM-Net (hidden Markov network), stochastic segment model 등에 대한 연구가 진행될 것이다. [5] [6]

세번째로 잡음에서의 음성인식 기술에 대한 연구가 더욱 활발히 진행될 것이다. 사람은 상당량의 잡음이 존재하는 곳에서도 음성을 잘 이해 하지만 기계는 아직도 제대로 음성을 인식하지 못하고 있다. 여기서 잡음이란 차량, 공장에서와 같이 소음이 많은 주변환경에 의한 것도 있으며 "에", "응" 등과 같이 발음습관 등에 의해 나타나는 의미없는 음성에 의한 것도 포함한다. 실제로 상용화를 위해서는 이 분야의 연구가 필수적이므로 앞으로 계속적인 연구가 이루어져야 할 것이다.

네번째로 초대용량 음성인식 기술에 대한 연구가

시작될 것이다. 현재 1000단어를 인식할 수 있는 시스템은 개발되어 있으나 수십만 단어를 인식할 수 있는 시스템은 아직 초보적인 연구단계에 있다. 이러한 연구를 위해선 고속 검색 알고리즘, 유사한 단어 사이의 변별력 향상을 위한 알고리즘에 대한 연구도 수행되어야 할 것이다 .

2. 응용시스템 측면

음성인식 기술을 이용한 응용시스템 연구의 전망은 다음과 같다.

첫째로 전화망을 통한 음성인식 기술을 이용한 음성정보검색 시스템이 실용화될 것이다. 종래의 전화망을 통한 음성인식 시스템은 음성정보검색 시스템을 이용할 수 없는 다이알식 전화기를 갖고 있는 가입자들을 위하여 전화망을 통한 숫자음을 인식하는 정도였다. 그러나 숫자음을 정확히 인식하기가 어렵고 버튼식 전화기의 확산에 따라 이러한 시스템의 활용도는 높지 않았다. 최근에는 음성인식 기술의 진보로 단순히 숫자음을 인식하는 것 이외에 다양한 종류의 단어, 문장을 인식할 수 있게 되었으므로 앞으로 음성정보검색 시스템의 전화망을 통한 입력 수단으로 음성인식이 중요하게 사용될 것이다. 현재 시험서비스 중에 있는 대표적인 응용시스템으로 캐나다의 Northern Telecom에서 개발한 증권정보검색 시스템(Stock-Talk)이 있다.[7] 이 시스템을 사용하기 위해서 전화 가입자는 전화번호(+1-154-765-7862)로 전화를 한 후 회사명을 음성으로 말하면 그 회사에 관련된 증권정보를 음성으로 들을 수 있다. 현재 이 시스템은 뉴욕 주식시장에 상장된 1561개의 회사이름을 인식할 수 있으며 새로운 회사가 상장이 되더라도 인식이 가능하기 때문에 이용 빈도수가 증가하고 있다고 한다. 다른 응용시스템으로 음성인식에 의한 전화번호안내 시스템을 들 수 있다. 일본 NTT에서는 전화번호안내 업무(한국의 114)의 효율화를 위해서 음성인식 기술을 이용한 시스템에 대한 연구를 수행하고 있다고 한다. 최근의 연구결과에 따르면 10만 단어에 대한 인식률로서 91%를 얻었다고 한다. 앞으로 인식률 향상을 위한 연구와 더불어 실용화 연구도 진행될 것이다.

두번째로 지능망에서 음성인식 기술을 이용한 IP (intelligent peripheral) 시스템 개발이 많아질 것이다. 현재 대표적인 응용시스템으로 음성 다이얼링 시스템(voice dialing system)이 있다. 음성 다이얼링 시스템은 전화가입자가 상대방의 전화번호를 누르지 않고 상호나 이름을 음성으로 말하면 자동으로 전화가 걸리는 시스템이다.[8] 미국 NYNEX 전화회사에서는 1993년 3월 중순부터 새로운 서비스로서 이 시스템을 사용하고 있으며 일본 NTT에서도 고도 정보화 시대를 향한 서비스 중의 하나로 음성 다이얼링 서비스를 선정하고 있다. 이 서비스의 장점은 상대방의 전화번호를 외울 필요가 없기 때문에 전화걸기가 쉽다는 것이다. 이 서비스는 세계에서 최초로 음성인식 기술을 이용하여 서비스 요금을 받는 전화서비스라는 면에서 의의가 있다. 현재 미국 Bell Atlantic 전화회사, Spirit 전화회사가 음성 다이얼링 서비스를 제공하기 위한 준비를 하고 있다.

한편 일본 KDD(국제전신전화회사)에서는 음성인식 기술을 이용한 구내 자동교환 시스템을 개발하여 연구소 내에서 시험서비스를 하고 있다. 이 시스템은 KDD 연구소로 걸려오는 외부전화를 자동으로 받아 음성을 인식하여 연구소내의 사람으로 자동교환시켜주는 시스템이다.[9] 미국 BBN 회사에서도 전화를 통한 음성을 인식하여 회사원의 전화번호를 알려주는 시스템을 시험 운용하고 있다. 이러한 시스템은 시험운용 결과에 따라 상용화가 추진될 것이다.

세번째로 현재의 음성인식 기술의 수준으로 가능한 특정목적을 위한 음성인식 시스템의 개발이 증가될 것이다. 예를들면 음성인식 열차표 판매기, 음성인식 VCR remote control, 음성 타자 기등이다. 이러한 시스템은 적당한 응용 대상 영역내에서 음성인식 기능을 수행하도록 하기위한 것이다. 그러나 기존방식을 이용하는 것보다 음성인식 기능에 의한 방식이 업무의 효율성을 증대시킬 수 있는 분야에서만 성공을 거둘것으로 예상된다.

네번째로 PC의 부가기능으로서 사용될 수 있는 음성인식용 H/W 및 S/W 개발이 지속될 것이다. 이러한 종류의 음성인식 시스템은 주로 수백단어를 인식할 수 있는 고립단어 혹은 연속단어를 인식할 수 있으며 사용자의 요구에 맞도록 대상단어 및 문법을 쉽게 변형시킬 수 있는 장점을 가지고 있다. 최근에는 PC를 음성으로 명령하는 소프트웨어도 개발 되었으며 의료 보고서를 음성으로 작성을 할 수 있는 제품도 개발되고 있다.

마지막으로 자동통역 전화시스템 개발과 같은 장기적인 시스템 개발이 지속될 것이다. 일본 ATR의 연구 프로젝트와 독일의 VERBMOBIL 프로젝트 등은

시스템 개발 기간을 10년 이상으로 하여 연구를 수행하고 있다. 이러한 연구는 진행과정에서 창출되는 연구 부산물이 크기 때문에 나름대로 의미가 있다고 할 수 있다.

Ⅳ. 결론

본 고에서는 음성인식 기술의 현황과 전망에 대해서 기술하였다. 음성인식 기술은 능력면에서 아직 초보단계에 있지만 현재 미국, 일본 및 유럽국가들이 국가 주도로 매우 활발히 연구를 하고 있으며 특히 전화망을 통한 음성인식 시스템 개발, 대용량 음성인식 및 잡음에서의 음성을 인식하는 알고리즘에 대한 연구를 중점적으로 하고 있다. 이러한 연구의 결실로 최근에는 음성 다이얼링 서비스와 같은 새로운 서비스가 창출되었으며 자동통역 전화시스템과 같은 대형 프로젝트도 상당히 진척이 되고 있다. 앞으로 음성인식기술 분야는 응용 위주의 시스템 개발이 보다 가속화되어 일상생활에 침투될 것이다. 국내에서의 음성인식 기술분야는 기술력 뿐만아니라 인력과 연구비가 외국의 경우에 비하여 볼때 부족한 실정이지만 한국어의 음성인식 시스템 개발은 한국인이 해야 한다는 투철한 사명감을 갖고 끊임없는 투자와 지원이 있어야 겠다.

參 考 文 獻

[1] D. S. Pallett, "DARPA resource management and ATIS bench mark test poster session," Proceedings of the DARPA speech and Natural Language Workshop, pp.49-58, Feb., 1991.

[2] C. Delogu et al, "New directions in the evaluation of voice input/ouput systems," *IEEE Journal on Selected Areas in Comm.*, vol 9, pp.566-573, May 1991.

[3] K. F. Lee, Automatic speech recognition : the development of the SPHINX system. Kluwer Academic Publisher, 1989

[5] J. Takami and S. Sagayama, "A successive atate splitting algorithm for effficient allophone modeling," Proceedings fo Int.Conf. on Acoustics, speech and Signal Processing, pp.573-576, Mar. 1992

[6] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans.* on Acoust., Speech, Signal Processing, vol.37, pp.1857-1869, Dec. 1989.

[7] M. Lennig et al., "Flexible vocabulary recognition fo speech," Proceedings of and *Int. Conf.* on Spoken Lang. Processing, pp.93-96, Oct. 1992.

[8] G. G. Matison, "Emerging voice services in the NYNEX network," Proceeding of voice Systems Worldwide 1992 pp.9-13, Feb. 1992.

[9] S. Kuroiwa et al., "Architecture and algorithms of a real-time word recognizer for telephone input," Proceedings of and *Int. Conf.* on Spoken Lang. Processing, pp.1523-1526, Oct. 1992.

筆 者 紹 介

丘 明 完
1960年 4月 26日生
1982年 2月 연세대학교 전자공학과(학사)
1985年 2月 한국과학기술원 전기 및 전자공학과(석사)
1991年 2月 한국과학기술원 전기 및 전자공학과(박사)

1985年 4月 ~ 현재 한국통신 소프트웨어 연구소 자동통역 연구실

주관심분야: 음성 인식, 자동통역 시스템 개발, Neural Network