□특 집□

음성인식 기술의 현황과 전망

한국통신 소프트웨어연구소 구 명 완

목 차

Ⅰ. 서 론

인간과 동물사이의 가장 큰 차이중의 하나는 인간은 언어능력을 가지고 있는 반면 동물은 언어능력이 없다는 점이다. 인간 사회에서 언어는 인간사이의 통신을 가능하게 해 주었으며 이로 인해 기술의 교류 및 진보가 쉽게 이루어졌다. 사실 현대의 문명은 언어가 존재하였기 때문에 이루어졌다고 하여도 지나친 말이 아닐 것이다. 이러한 언어는 실제로 문자 혹은 음성으로 표현될 수 있다. 문자는 후대를 위한 기록으로서 음성은 인간사회의 통신으로서 언어를 주로 표현하여 주었다.

음성인식이란 음성속에 내재되어 있는 언어정보를 자동으로 추출하는 과정이다. 물론 이러한 과정은 컴퓨터를 이용하여 수행된다. 단어의 넓은 의미만을 고려한다면 음성인식의 범주에는 누가 말을 하고 있는지를 알 수 있는 개인 정보를 추출하는 과정인 화자 인식(speaker recognition)도 포함될 수 있다.

음성인식에 대한 연구는 음성을 인식할 수 있는 로보트와 타이프라이타(typewriter)를 개발할 목적으로 수십년 전부터 수행되었다. 음성인식에 관한 최초의 논문은 1952년에 미국 벨 연구소와 숫자음 인식기인 Andrey에 관한 것이었지만 최근에서야 비로서 인간과 기계사이의 음성통신이 부분적으로 가능하게 되었다. 현재의 음성인식 기술수준은 인간의 능력에 비하면 보잘것이 없다. 이와같은 이유중의 하나는 아직 인간이 음성을 인식하는 정확한 방법을 모른다는 것이다. 음성이 인간의 귀에 도달하면 외이, 중이, 내이를 거쳐 주파수 성질을 나타내는 신호로 바뀐다는 사실은 알고 있지만 이 신호가 뇌에 도달하여 어떻게 언어정보로 변하는 지는 아직 알려지지 않고 있다. 그렇지만 음성인식 학자들은 인간의 음성인식방식이 알려질 때까지 기다리지 않고 연구를 계속하고 있다. 그들은 다음과 같은 예를 들어가며 스스로의 연구방식을 고수하고 있다.

"인간이 하늘을 날고자 하는 꿈은 새를 보고 키워졌지만 현재의 비행기는 새가 날개를 젓는 방식으로 비행하지 않으면서도 새보다 훨씬 높이, 보다 빠르게 날 수 있다. 비슷하게 음성의 인식도 이와같은 현상이 가능할 수 있다."

그러면 이와같이 어려운 음성인식 능력을 기계에 부여하는 이유는 무엇인가. 그 이유는 음성인식이 다음과 같은 장점을 가지고 있다는 사실이다.

(1) 자연스러움: 음성은 인간의 자연스러운 통신수단이므로 음성으로 기계에 다 명령을 입력하는 것은 매우 쉽다. 즉 기존의 타이핑 혹은 푸쉬버튼을 이용할 때 필요한 전문기술이 필요없다는 것이다.

(2) 신속성: 음성은 글로 쓰는 것보다 8~10배 정도 빠르며 타이프라이터보다 3~4배 정도 빠르게 명령을 입력할 수 있다.

(3) 동시성: 사람이 눈, 귀, 손, 다리를 사용하여 다른 행동을 하고 있을 경우에도 음성으로 명령을 입력할 수 있다.

(4) 경제성: 먼 지역에서도 마이크나 전화를 통하여 음성으로 명령을 내릴 수 있으므로 명령을 입력하기 위하여 특별한 비용이 들지 않는다.

본 고에서는 최근의 음성인식 기술의 현황 및 응용사례를 살펴보고 앞으로의 발전방향을 전망하고자 한다. 먼저 II장에서는 국내외의 음성인식연구 및 음성인식 시스템 개발 현황을 알아보고 음성인식 시스템을 이루는데 필수적인 특징추출, 인식알고리즘 및 언어처리 알고리즘에 대한 연구동향을 파악한다. III장에서는 음성인식 기술의 향후전망을 기술적 측면과 응용시스템 측면으로 나누어서 예상하고자 한다. 그리고 마지막으로 III장에서 결론을 짓는다.

II. 음성인식 기술의현황

2.1 음성인식연구 현황

2.1.1 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 progam)과 문자언어 프로그램(written language program)으로 나누어진다. 음성언어 프로그램은 대용량 음성인식 시스템과 음성언어 이해에 관한 연구를 추축으로하여 특정 task 영역에서 자연스러운 음성을 실시간으로 인식하는 화자독립 혹은 화자적응 음성인식 시스템을 개발하는 것을 목표로 한다. 시스템의 성능평가를 위하여 낭독체(reading speech) 연속음성으로 구성된 RM(resource management) 데이타베이스와 항공기 여행정보에 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, Dragon, Lincoln, MIT, SRI, Texas Instruments, Unisys, AT&T 등 이다. 각 기관에서 개발한 음성인식 시스템는 동일한 음성 데이타베이스를 사용하여 성능을 비교하며, 최근의 성능비교 결과가 〈표 1〉에 나타나 있다[1,2].

〈표 1〉의 결과에 따르면 낭독체 음성인식 시스템의 성능은 CMU의 음성인식 시스템이 96.4%의 인식률로 가장 우수하였으며 회화체 음성인식 시스템은 BBN의 인식시스템이 94.2%로 가장 우수하였다. 그리고 자연언어처리 기술과 통합된 음성이해 인식 시스템의 성능은 SRI의 시스템이 67.9%로 가장 우수하였다.

문자언어 프로그램은 대용량 텍스트(text)처리에 필요한 기술을 개발하는 것을 목표로 하며 메시지 이해(message understanding), 자연언어 학습(natural language learning) 및 데이타베이스구축 등에 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.

프로젝트의 성공을 위하여 매년 음성 및 자

〈표 1〉 DARPA 음성인식 시스템의 성능비교

음성 데이타베이스 종류	인식률(%)	기관
RM (낭독체 1000단어 연속음성인식) * 1991년 2월 실험	96.4	CMU
	96.2	BBN
	95.6	MIT
	95.5	AT&T
ATIS (회화체 연속음성인식) 음성인식 부문 * 1992년 2월 실험	94.2	BBN
	92.7	SRI
	89.6	CMU
	87.5	MIT
ATIS (회화체 연속음성인식) 음성인식 + 자연언어처리부문 * 1992년 2월 실험	67.9	SRI
	64.2	BBN
	60.0	MIT
	48.3	CMU

연언어 워크샵을 개최하여 연구에 참여하고 있는 연구원들이 최신의 정보를 교류하고 앞으로의 연구방향을 모색하도록 하고 있다. 1992년에 개최된 제 5차 워크샵에서 토의된 기술적인 문제는 항공기 예약정보에 관련된 인식 시스템의 성능향상, 대용량 데이타 수집을 위한 여러기관의 공동노력, 대용량 음성인식 시스템의 개발 및 통계적인 자연언어처리 연구에 관한 것이었다[2]. 항공기 예약정보에 관련된 인식 시스템의 인식율은 '91년도 결과보다 약 8% 이상 향상되었으며 특히 1991년에서는 문맥에 관계없는 문장의 인식율에 중요성을 두었는데 반해 1992년도에는 문맥을 고려하여야만 답을 낼 수 있는 문장에 대한 인식률도 점차 고려하고 있었다. 대용량 데이타수집을 위한 대표적인 노력은 MADCOW(Multi-Site ATIS Data COllection Working group)을 중심으로 이루어지고 있다. 이 그룹은 음성언어처리 시스템의 평가를 위해서 1991년 5월 AT&T, BBN, CMU, MIT 및 SRI 등이 중심이 되어 결정되었으며 현재 이들 기관을 통하여 12,000 회화체 문장을 수집하여 배포하고 있다.

대용량 음성인식 시스템의 개발을 위해선 Wall Street Journal을 통해 얻어진 낭독체 문장을 이용하여 인식실험을 수행하고 있는데 초기단계로서 5,000단어로 한정하고 있다. 1992년 2월에 실험하였을 때 화자독립 인식율은 82.9%, 화자종속 인식율은 89.3%이었는데 1992년 12월에는 화자독립 인식율은 94.7%, 화자종속 인식율은 95.5%로 나타나 〈표 1〉의 1000단어 인식 시스템과 유사한 결과를 나타내었다.

2.1.2 일본

일본에서의 음성인식기술은 1982년부터 추진한 제 5세대 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 실험 시스템을 데모하는 것으로 1단계 연구를 끝내고 음성번역통신 연구소를 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, 합성, 지식처리, 잡음에서의 음성처리 및 평가기술 등 8가지의 핵심기술로 나누어서 약 185명의 연구자가 연구를 수행하고 있다.

2.1.3 유럽

유럽에서의 음성인식 기술연구는 유럽국가들이 모여서 공동으로 수행하는 연구와 각 나라에서 자체적으로 수행하는 연구로 나누어진다. 범유럽국가들이 수행하는 연구는 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 세단계로 나누어서 진행되는데 음성인식에 관한 연구는 매 단계마다 주요 추진 과제

〈표 2〉 음성인식에 관련된 ESPRIT 프로젝트 내용

연구단계	Project 이름	내 용
I (1984-1989)	SIP (advanced algorithms 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

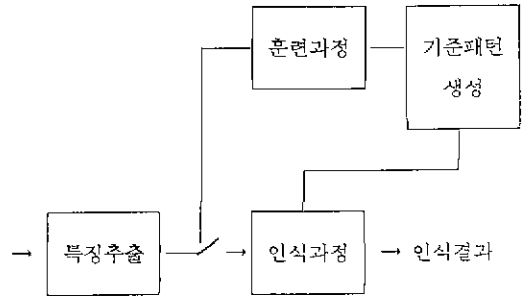
였다. 각 단계에서 음성인식에 관련된 주요 프로젝트의 내용이 표 2에 나타나 있다.

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(Siemens-Philips-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.

2.1.4 국내

국내에서의 음성인식 연구는 1980년 초부터 일부 대학을 중심으로 연구가 수행되었으며 최근에는 많은 대학과 연구소를 중심으로 활발히 진행되고 있다. 최근의 연구결과에 따르면 한국과학 기술원에서는 100개의 어휘로 구성된 연속음성인식 시스템을 개발하였으며[3], 한국통신에서는 전화망을 통한 음성을 인식하여 연구센타내의 전화번호를 자동으로 알려주는 시스템을 개발하였다[4]. 1991년부터는 한국통신과 전자통신 연구소가 공동으로 자동통역전화 요소기술연구를 수행하고 있으며 이 연구결과는 향후 한·일간 자동통역전화 시스템 개발에 이용될 것이다. 최근에는 기업체에서도 음성인식 기술을 이용한 여러가지 제품개발을 시도하고 있다.

2.2 음성인식 시스템



(그림 1) 음성인식 시스템의 개념도

〈표 3〉 음성인식 시스템의 분류

	화자 종속 여부	
	화자종속	화자독립
고립단어인식시스템	상용화	상용화
연속음성인식시스템 연결단어인식 대화체음성인식	상용화 연구중	상용화 연구중

2.2.1 기본 개념도

현재의 음성인식 기본 개념도는 (그림 1)과 같이 기본적으로 음성으로부터 음성 패턴(단어, 음소 등)의 특징을 추출하여 기준 패턴을 만드는 훈련과정과 미지의 음성이 입력되면 저장된 기준 패턴과 비교하여 가장 유사한 기준 패턴을 찾아내는 인식과정으로 나눌 수 있다. 이러한 알고리즘을 일반적으로 패턴 매칭(pattern matching) 알고리즘이라고 부른다.

2.2.2 분류

음성인식 시스템은 〈표 3〉과 같이 단어를 인식하는 고립단어 인식 시스템과 연속적으로 발음된 문장을 인식하는 연속음성 인식 시스템으로 크게 나누어질 수 있다. 연속음성 인식 시스템은 연결단어(connected word) 인식 시스템과 대화체음성(conversational speech) 인식 시스템으로 세분화 될 수 있다. 연결단어 인식 시스템은 상대적으로 작은 단어를 매 단어마다 또박 또박 발음하는 단어를 인식하는 시스템인 반면 대화체음성 인식 시스템은 상대적으로 대용량 단어를 매 단어마다 인식하는 것이 아니라 문장의 의미를 파악하는 것이다. 이와같은 대화체 음성인식

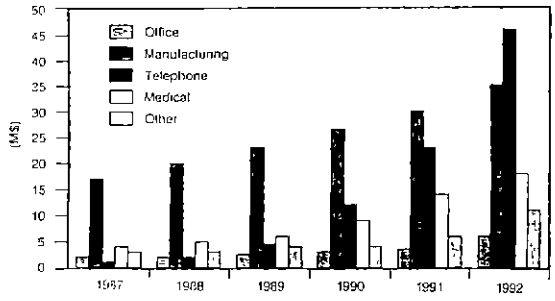
시스템은 음성이해 시스템이라고도 불리우며 언어지식을 사용하는 것이 중요하다. 대화체 음성인식 시스템은 상당히 어렵기 때문에 문장내에서 필요한 단어만 선별하여 인식할 수 있는 단어선별(word spotting) 인식 시스템의 연구가 최근 활발히 진행되고 있다.

또한 음성인식 시스템은 화자종속 인식 시스템과 화자독립 인식 시스템으로도 나눌 수 있다. 화자종속 인식 시스템은 훈련 된 특정화자 만을 인식할 수 있는 시스템이고 화자독립 인식 시스템은 어떠한 사람의 음소도 인식할 수 있는 시스템이다. 일반적으로 화자독립 인식시스템이 화자종속 인식 시스템보다 훨씬 어려우나 이용범위가 넓기 때문에 많이 연구되어지고 있다.

한편 기준 패턴의 단위를 무엇으로 사용하는냐에 따라서 음성인식 시스템의 특징이 구분될 수 있다. 단어를 기준 패턴으로 사용하면 단어내의 연음(coarticulation)현상을 고려할 필요가 없기 때문에 인식율이 높은 반면 인식대상 단어수가 많아질수록 메모리와 계산량이 증가한다는 단점도 있다. 또한 연속음성에 내재되어 있는 단어사이의 연음현상을 표현할 수 없다는 단점도 있다. 반면 기준 패턴으로서 음소를 사용하게 되면 대상 단어수가 늘어난다고 하여도 계산량 및 메모리 사용량이 많이 증가되지 않으며 훈련과정도 간단하다. 또한 단어사이의 연음현상도 쉽게 표현될 수 있다. 그러나 음소의 발음규칙이 명확히 알려져 있지 않기 때문에 인식율이 떨어지는 단점이 있다. 최근에는 이러한 현상을 극복할 수 있는 문맥종속음소(context-dependent phoneme)를 기준 패턴으로 사용하기도 한다. 이와같이 음소를 기준 패턴으로 사용하게 되면 훈련과정에서 전혀 훈련이 되지않는 어휘도 인식할 수 있는 장점이 있다. 이러한 음성인식 시스템을 단어독립(vocabulary-independent) 인식 시스템이라고 한다.

2.2.3 음성인식 시스템의 시장성

현재의 음성인식 기술은 3~4살 어린이가 음성을 이해하는 것 보다도 이해능력이 떨어지지만 부분적인 응용분야에서는 이미 상용화 제품이 출현되고 있다. 최초의 음성인식 기술을 이용한



(그림 2) 미국의 음성인식 시장

상용화 제품은 1980년 초에 출현하였는데 키이보드없이 컴퓨터에 명령을 입력시키는 장치 및 시각 장애자들을 위한 것이었다. 그러나 1980년 중반의 기술적인 진보에 힘입어 1990년도부터 전화망을 통한 음성인식 시스템 및 음성 받아쓰기(dictation) 시스템들이 나타내기 시작하였다.

음성인식 기술이 성공하기 위해서는 기술적인 문제 뿐만 아니라 응용분야를 잘 선정하여야 한다. 예를 들면 판매 혹은 관심을 끄는 새로운 분야가 아니라 음성인식 기술을 이용하여 실제적으로 사용자에게 이익이 되는 응용분야를 선정하여야 한다. 또한 사용자에게 편리하도록 사용자 편리성에 신경을 써야하며 실 시간으로 동작하여야 한다. 그리고 인식율은 95% 이상을 얻어야 한다. (그림 2)에는 1987년부터 1992년까지 미국의 음성인식 시장에 관한 내용이 나타나 있다. 1991년까지는 공장자동화 응용분야의 시장성이 가장 높았지만 1992년에는 전화망을 통한 시장이 가장 크다. 이것은 전화망을 통한 음성인식 기술이 확보되면서 응용분야가 넓어지고 있다는 것을 나타낸다.

2.3 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 이와같은 특징이 무엇인지는 정확히 알려져 있지 않다. 음성으로부터 특징을

추출하는 방법은 크게 네가지로 나눌 수가 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 변화시킬 때에는 안정된 특성을 어느정도 만족할 수 있는 구간(예를들면 10~30 msec)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 것이데 이 방식은 모든 음성이 구강의 모양에 따라 구분될 수 있으며 구강의 형태는 혀의 위치 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 비음 등은 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 AR 모델을 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 특징은 추출하는 방법에는 FFT에 의하여 주파수 영역으로 변환시킬 때 weighting을 가하는 방식과 LPC에 의해 추출된 파라미터를 weighting시켜 파라미터로 추출하는 방식이 있다[5].

네번째 방법은 동적 특징(dynamic feature)을 주파수 영역의 특징(spectral feature)들과 동시에 사용하는 것이다. 앞에서 설명한 음성특징은 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) 등을 표현되는데 최근의 연구결과에 의하면 mel scale된 LPC cepstrum의 차(differential LPC cepstrum), energy 및 energy차를 음성특징으로 사용하였을 때 높은 인식율을 얻었다고 한다[5].

2.4 인식 알고리즘

음성인식 알고리즘 (그림 1)의 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.

2.4.1 DTW 알고리즘

DTW 알고리즘은 인식과정에서 사용되는 알고리즘으로서 입력 음성 패턴과 기준음성 패턴간에 거리를 측정할 때 dynamic programming의 기법을 이용한다. 음성은 동일한 사람이 같은 단어를 여러번 발음하더라도 음성 특징이 각기 달라지며 특히 감정, 분위기에 따라 발음 지속시간(프레임 길이)이 달라진다. 기준음성 패턴과 입력음성 패턴의 발음시간의 차이가 있을 경우 두 패턴사이의 거리(distance)를 측정하기 위해서 우선 기준음성 패턴의 각 프레임과 그에 대응하

는 입력음성 패턴의 프레임 번호 사이의 쌍(pair)을 찾아야 한다. 이 대응쌍은 warping함수에 의하여 구해지며 이때 dynamic programing기법이 이용된다. Dynamic programin기법에 따르면 warping함수에 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 기준 패턴을 쉽게 만들 수 있기 때문에 사용자의 요구에 따라 음성인식 시스템의 업무 내용을 용이하게 변경할 수 있다.

2.4.2 HMM 알고리즘

HMM 알고리즘은 음성인식 시스템 개념도에서 훈련과정 및 인식과정을 수행하는 알고리즘으로서 1970년말부터 음성인식 알고리즘으로 많이 사용되었다. 최근에는 높은 인식률과 빠른 인식시간 때문에 대용량 음성인식 시스템이 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 사용한데 그 이유는 음성패턴의 다양한 변화를 수용하기 위해서이다. Hideen Markov 모델이란 이중 stochastic process로서 state 선정에 관한 stochastic process와 매 state마다 음성 패턴이 발생될 출력 확률(output probability)에 관한 stochastic process로 구성된다. 즉 음성 패턴의 각 특징은 state의 선정 확률과 출력 확률 등으로 표현하여 준다. 여기서 hidden이란 의미는 state가 음성 패턴에 관계없이 모델속에 숨어 있다는 것을 말한다.

HMM 알고리즘은 기준 패턴을 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 음성인식 시스템에서 기준 패턴의 기본단위로 단어를 선정하였다면 1000개의 단위가 필요하지만 음소를 선정하였다면 40~50개의 음소의 파라미터만 저장하면 된다.

2.4.3 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 훈련 데이타의 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 기준 패턴이 된다. 그러나 single layer perceptron으로는 입력 데이타를 분류할 수 없는 경우(exclusive OR 상태)도 있기 때문에 실제로 MLP가 음성인식 시스템에 이용된다. 이때 MLP의 입력 데이타로는 2.3절에서 설명한 바 있는 특징추출된 데이타가 된다. 그런데 음성은 공간적인 특징(주파수에 의한 특징)과 시간적인 특징(음성발성 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[6]. 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 음성 데이타를 의미에 관계없이 훈련시키면 음소를 대표할 수 있는 특징이 저절로 나타난다는 것이다. 이 알고리즘은 실제 인간의 청각작용과 비슷하나 음성인식 시스템에 적용하였을 겨우 supervised learning 알고리즘 보다는 낮은 인식률을 보인다.

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

2.5 언어처리 알고리즘

언어처리 알고리즘은 문장을 인식할 때 주로 사용되는 알고리즘으로서 음성인식결과를 근거로 하여 문법규칙에 적합한지를 판별하는데 이용된다. 언어처리 알고리즘이 음성인식과정에 이용되는 근본적인 이유는 주로 신호처리에 의해 이루어지는 인식 알고리즘으로서는 완벽한 문장을 인식하기가 어렵기 때문에 언어정보(문법규칙 등)를 이용해서 인식율을 향상시키기 위해서이다. 언어의 특성상 미국에서는 단어를 기준한 언어규칙을 이용하고 있으며 일본에서는 구를 기본적 단위로 사용한 언어규칙을 사용하고 있다.

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 해준다. 대표적인 통계적 모델로서 bigram과 trigram이란 것이 있는데 이것은 매 단어에 대해서 이전 한 개(두 개)의 단어가 입력되었을때 이 단어의 발생가능성을 확률값으로 표현하여 문장인식에 사용한다. 이러한 알고리즘은 HMM 모델에 근거한 인식시스템과 쉽게 결합할 수 있으며 회화체 음성인식에 적합하다.

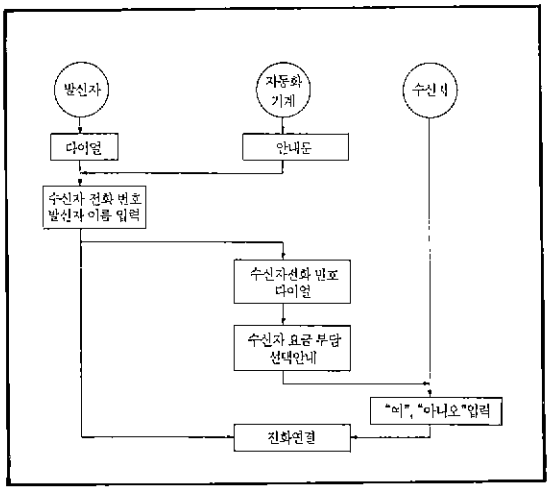
구문 규칙 방식은 언어학에서 연구된 구문론(syntax)에 따라 규칙을 만들어서 매 단어 다음에 올 수 있는 단어의 종류를 규제함으로써 문장을 인식하는 방식이다.

III. 음성인식 기술의 응용

3.1 요금부과 선택의 자동화

현재의 음성인식 기술을 실제의 업무에 적용하기 위해서는 적은 단어만 사용하여도 가능한 고객 서비스를 선택하여야 한다. 대표적인 업무로 요금 부담을 발신 전화번호로 부과시키지 않고 다른 방법으로 부과시키는 업무를 들 수 있다. 현재 미국에서는 발신 전화번호 이외로 요금을 부과시킬 수 있는 방식으로서 수신자 요금부담(collect call), 제 삼자 요금부담(third-number billed calls) 및 전화 카드(calling card) 등이 있다. 이러한 업무를 자동화시키기 위하여 Bellcore에서는 단지 예(Yes), 아니오(No)의 두 단어만을 인식할 수 있는 시스템을 구성하여 수신자 요금부담 및 제 삼자 요금부담의 수락 혹은 거절을 자동화하고 있다[7]. 전화망에 수신자 요금부담을 자동화하는 과정은 자동음성응답장치를 사용하여 대화형식으로 이루어진다. 수신자 요금부담을 위한 대화 과정이 (그림 3)에 그려져 있다. 먼저 시스템은 송화자의 이름을 묻고 그것을 녹음한다. 이 녹음된 송화자의 이름을 나중에 수신자에게 돌려 주어서 수신자가 요금을 부담할 것인지를 묻는다. 이때 수신자가 "예" 혹은 "아니오"라고 말하면 그 단어를 음성인식하여 처리한다[7].

한편 AT&T에서는 자동 호 형태 인식(automatic call type recognition)을 위하여 음성인식 기술을 사용하였다. 이 시스템은 안내양을 통하지 않고 사용자가 원하는 호 형태를 자동적으로 인식하여 처리해 준다. 선택할 수 있는 호 형태는 수신자 부담(collect call), 전화 카드(calling card), 지명통화(person to person) 및 제 삼자 요금부담(bill to third party)이다. 이러한 서비스의 제공은



(그림 3) 수신자 요금 부담의 동작 흐름도

안내양의 작업량을 줄여주며 현재 Northern Telecom에서도 개발되어 실험중에 있다.

자동 호 형태 인식을 위한 인식시스템의 상용화 실험은 1985년도에 처음으로 이루어졌다. 이 시스템은 사용자가 "수신자 부담", "전화 카드", "지명통화", "제 삼자 요금부담" 및 "안내양요구"와 같은 5개의 고립단어중 하나를 발음하면 인식하는 시스템인데 94%의 인식률을 나타내었다. 그러나 사용자의 20% 정도가 문장을 구성하는 등 5단어 이외의 단어(예를들어 "대화자부담 요금 방식으로 해 주세요")를 발음하였다. 이와같이 인식 시스템은 미리 정해진 단어 이외의 음성이 입력되면 인식률이 급격하게 떨어지게 된다. 최근에는 이러한 단점을 보완하기 위하여 필요한 단어만을 선별하여 인식시킬 수 있는 단어선별(word spotting)기술을 개발하여 시스템에 적용하고 있다[8]. 이 기술은 사용자가 문장을 발음하더라도 문장속에 포함되어 있는 대상단어만을 찾아내어서 인식하므로 매우 높은 인식률을 나타내게 된다. 실제 실험에서 사용자가 시스템의 요구에 따라 고립단어만을 발음했을 경우에 99.3%의 인식률을 나타내었으며, 사용자가 앞에서 제시한 5단어가 포함된 문장으로 발음하더라도 95.1%의 높은 인식률을 나타내었다.

현재 Ameritech 및 NYNEX 산하 전화국에서는 1989년 5월부터 36대의 자동 호 형태 인식 시스템이 설치되어 운영되고 있으며 한편 Bell Canada에서는 2개국 언어를 수용할 수 있는 연구를 진행하고 있다. Ameritech의 실험에 의하면 전화망을 통한 일반 가입자의 2,608호 시도 중 0.92%(24호)만이 잘못 연결이 되었으며, 1.8%(47호)가 연결이 거절되었다고 한다.

3.2 신용카드 조회 시스템 및 화자식별

신용카드 조회를 위해서 음성인식 기술을 사용할 수 있다. AT&T사의 CONVERSANT 시스템은 HMM과 단어별 기능이 첨가되어 연결단어 인식 시스템으로서 현재 서비스되고 있다. 이 시스템은 상점에서 구매자의 신용카드 조회하고자 할 때 상점 고유번호, 구매자의 카드번호 및 구매가격 등을 전화를 통하여 발음하면 카드를 조회하여 준다. 상점 고유번호 길이는 10자리인데 인식률은 81.6%이었고, 사용자가 다시 발음을 했을 경우에는 90.7%의 인식률을 나타내었고, 15자리의 카드번호일 경우에는 각각 82.4%, 89.4%의 인식률을 얻을 수 있었다. 그리고 구매 금액을 자연스럽게 발음했을 경우에는 70.9%의 인식률을 얻었으며 다시 한번 발음했을 경우에는 85.4%의 인식률을 얻었다. 최근에는 처음 발음했을 경우에 97% 정도의 인식률을 얻을 수 있는 알고리즘을 실험실에서 개발하고 있다.

화자식별이란 사람은 각자의 독특한 발음습관 및 음색을 가지고 있다는 전제하에 동일한 문장 및 단어를 발음시켜 실제로 등록된 화자를 찾아내는 기술이다. 벨 산하의 한 회사에서는 이 기술을 가택연금 확인에 이용하기 위하여 실험중에 있다. 법정에서 가택연금이 확정되면 연금기간 동안 연금자에게 자동적으로 전화를 걸게 되고 이때 연금자가 전화를 받으면 실제 연금자인지를 화자 식별기술을 이용하여 확인한다. 이러한 응용사례는 동일한 전화기를 사용하고, 하루에 여러번 확인전화가 있게 되므로 전송상의 변화가 적고 화자의 음성특징의 변화도 적게 되므로 현재의 기술수준으로도 가능한 분야이다. 또 다른 응용사례로서는 공중전화 카드에 본인의 음성을 등록시켜 본인만 전화 카드를 사용할 수 있게 한 것을 들 수 있다.

화자식별을 위하여 음성신호로부터 특징을 추

출하는 알고리즘은 음성인식을 위한 특징 추출 알고리즘과 비슷하며 인간이 인간을 구별할 수 있게 하는 특징을 현재의 화자식별 기술로는 정확히 파악하지 못하고 있다.

Bellcore에서는 전화망의 주파수 영역에서 화자식별기를 제작하여 단음절 단어로 실험한 결과 96%의 식별 능력을 나타냈으며 여러단어를 반복하여 말할 경우에도 98% 이상의 식별 능력을 얻을 수 있었다[9].

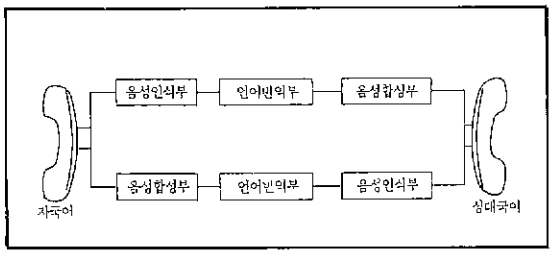
3.3 자동통역 시스템

자동통역 전화 시스템은 서로 다른 언어를 사용하는 통화자들이 상대방의 언어를 모르더라도 자유롭게 전화를 통화할 수 있게 해주는 전화시스템이다. 이러한 시스템이 국제간의 교류가 빈번해짐에 따라 상대방의 언어에 능숙하지 못한 많은 사람들의 관심을 끌고 있는 것은 당연하다.

자동통역 전화기는 (그림 4)와 같이 음성인식부, 언어번역부 및 음성합성부로 구성된다. 음성인식부에서는 송화자의 음성을 인식하여 문자로 변환시키며 언어번역부에서는 이 문자를 상대국의 문자로 번역한다. 음성합성부에서는 변환된 문자를 가지고 음성을 합성하여 수화자에게 전달하여 준다. 각 부분별 상세기술은 다음과 같다[10].

3.3.1 음성인식부

자동통역 연구소에서 제안한 음성인식 알고리즘은 HMM에 근거한 것이다. 이 알고리즘 이외에도 음성인식률을 향상시키기 위하여 지속시간 제어(duration control), 개별 VQ(separate vector quantization) 및 퍼지 VQ(fuzzy vector quantization)기법을 채용하였다. 음성인식을 위한 훈련 데이타로 한명의 남자가 발음한 2,620 단어를 음소로 분할한 데이타를 사용하였을 때 화자종속 음소 인식률은 94.0%로 나타났다. 그러나 전화를 통한 일상 대화는 연속단어로 이루어진 문장이므로 이러한 문장을 인식할 수 있는 알고리즘이 구현되어야 한다. ATR에서는 문장을 이루는 기본 구조를 구문으로 가정하고 구문을 음소의 열로 구성하여 인식 작업을 수행한다. 문장이 입력되면 음소별 HMM 출력 확률에 의해서 가능한 음소를 찾아내고 LR parser에 의해 매 음소의 열을 찾아내어서 최고의 출력 확률을 갖는 음소의 열로 이루어진 구문을 선택하고 이 구문의 열이 문장이 되도록 한다.



(그림 4) 자동통역전화 시스템의 개념도

인식 실험을 위한 대상 업무는 1,035 단어로 이루어진 국제회의 등록업무로 한정하였으며 5,240 단어로 훈련시킨 후 화자종속 인식 실험을 한 결과 88.4%의 구문 인식률을 얻었다.

3.3.2 기계번역부

기계 번역부는 NADINE라고 명명되었으며 Symbolics 3620 및 SUN4 상에서 common LISP로 프로그램되었다. 이 시스템은 분석기, 변환기 및 생성기로 구성되어 있다. 분석기는 음성인식기에 의해 인식된 출력문장을 20여개의 구분 구조 법칙을 사용하여 분석하며 언어에 무관한 구조로 의미를 표현한다. 변환기에서는 이 의미구조를 수화자에 언어형태로 된 의미구조로 변환시킨다. 마지막으로 생성기에서는 수화자의 언어 문법 구조에 맞도록 언어를 구성하여 준다. 이 NADINE 시스템을 138 문장으로 구성된 대화 문장음으로 실험한 결과 126 문장이 영어로 적절하게 변환된 결과를 얻을 수 있었다.

3.3.3 음성 합성부

SL-TRANS 자동 통역기에 사용된 영어 음성 합성기로서 DECTALK를 구매하여 시스템을 구성하였다. 음성인식부, 기계번역부 및 음성합성부를 통합시켜 회의 등록에 관한 37 문장으로 구성된 대화로 실험한 결과 34 문장이 올바르게 통역되었다.

IV. 음성인식 기술의 전망

4.1 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 중심이 되어 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 연구 등이 포함될 것이다. 또한 언어처리는 회화체 음성에 관한 것이 주종을 이룰 것이다.

두번째로 HMM의 성능향상에 촛점이 맞추어질 것이다. HMM 알고리즘은 음성인식 시스템에 사용되어 매우 좋은 결과를 얻었지만 아직도 보완되어야 할 부문이 많다. 이를 위해서 neural network와 동시에 사용한다거나 HMM 알고리즘의 변형인 HMM-Net(hidden Markov network), stochastic segment model 등에 대한 연구가 진행될 것이다[11,12].

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 1000 단어를 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.

4.2 응용시스템 측면

음성인식 기술을 이용한 응용시스템 연구의 전망은 다음과 같다.

첫째로 전화망을 통한 음성인식 기술을 이용한 음성정보 검색 시스템이 실용화될 것이다. 종래의 전화망을 통한 음성인식 시스템은 음성정보 검색 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 음성정보 검색 시스템의 전화망을 통한 입력 수단으로 음성인식이 중요하게 사용될 것이다. 현재 시험서비스중에 있는 대표적인 응용시스템으로 캐나다의 Northern Telecom에서 개발한 증권정보 검색 시스템(Stock-Talk)이 있다[13]. 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 10만단어에 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 응용 시스템으로 음성 다이얼링 시스템(voice dialing system)이 있다. 음성 다이얼링 시스템은 전화가입자가 상대방의 전화번호를 누르지 않고 상호나 이름을 음성으로 말하면 자동으로 전화가 걸리는 시스템이다[14]. 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(국제 전신전화회사)에서는 음성인식 기술을 이용한 구내 자동교환 시스템을 개발하여 연구소 내에서 시험서비스를 하고 있다. 이 시스템은 KDD 연구소로 걸려오는 외부 전화를 자동으로 받아 음성을 인식하여 연구소 내의 사람으로 자동교환시켜 주는 시스템이다[15]. 미국 BBN 회사에서도 전화를 통한 음성을 인식하여 회사원의 전화번호를 알려주는 시스템을 시험 운용하고 있다. 이러한 시스템은 시험 운용 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 타자기 등이다. 이러한 시스템은 적당한 응용 대상영역내에서 음성인식 기능을 수행하도록 하기 위한 것이다. 그러나 기존 방식을 이용하는 것보다 음성인식 기능에 의한 방식이 업무의 효율성을 증대시킬 수 있는 분야에서만 성공을 거둘 것으로 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. 결 론

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 있다. 미국은 회화체 연속음성인식에 주력하고 있으며 최근에는 초 대용량 음성인식 시스템 개발도 시작하고 있다. 중간 기술을 이용한 실용화에도 힘을 써 음성 다이얼링 서비스 및 요금부담 선택의 자동화 서비스 등을 개발하여 현재 사용중에 있다. 일본은 음성번역 통신연구소를 중심으로 자동통역 전화 시스템 개발에 주력하고 있으며 회사에서는 실용적인 시스템 개발을 수행하고 있다. 유럽은 범 국가 프로젝트를 중심으로 다국적 언어 인식이 가능한 실용적인 시스템 개발에 중점을 두고 있다.

앞으로의 음성인식 기술분야는 응용 위주의 시스템 개발이 보다 가속화되어 일상 생활에 침투될 것이다. 또한 음성인식 기술이 발전함에 따라 단순한 문장을 최종적으로 인식하는 것이 아니라 문맥에 따른 지식을 갖고 있어야만 대답이 가능한 음성이해 시스템 개발로 점차 변하고 있다. 이러한 시스템의 성공적인 개발을 위해선 전자공학자, 언어학자, 전산공학자 및 심리학자

등의 공동연구가 필수적이다.

국내의 음성인식 기술분야는 기술력 뿐만 아니라 인력과 연구비가 외국의 경우에 비하여 볼 때 부족한 실정이지만 관련 분야의 연구자들은 한국어의 음성인식 시스템 개발은 한국인이 해야 한다는 투철한 사명감을 갖고 연구에 매진하여야 할 것이다.

참 고 문 헌

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. D. S. Pallett, "DARPA february 1992 ATIS benchmark test result", *Proceedings of the DARPA speech and natural language Workshop*, pp. 15~27, Feb., 1992.
3. 김동영 외 4인, "한국어 연속음성인식 시스템 개발", 제 10회 음성통신 및 신호처리 워크샵 논문집, pp. 238~242, Aug. 1993.
4. 구명완 외 1인, "PABX를 통한 음성인식 시스템", 제 10회 음성통신 및 신호처리 워크샵 논문집, pp. 311~314, Aug. 1993.
5. K. F. Lee, *Automatic speech recognition: the development of the SPHINX*. Kluwer Academic Publisher, 1989.
6. A. Waibel *et al.*, "Phoneme recognition using time-delay neural networks", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 37, Mar. 1989, pp. 328~339.
7. R. W. Bossemeyer Jr., E. C. Schwab, "Automated alternative billing services at Ameritech: speech recognition performance and the human interface", *Speech Technology*, Feb./March 1991, pp. 24~30.
8. D. B. Roe *et al.*, "AT&T speech recognition in the telephone network", *Speech Technology*, Feb./March 1991, pp. 16~21.
9. G. elius *et al.*, "Bellcore effects in applying speech technology to telephone network services", *Int. Conf. on Speech Lang. Process.*, Kobe, 1991, pp. 20.2.1~20.2.4.
10. A. Kurematsu, "Language processing in connection with speech translation at ATR Interpreting Telephony Research Lab", *Speech Communication*, Vol. 10 Feb., 1991, pp. 1~9.
11. J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling", *Proceedings of Int. Conf. on Acoustics, Speech and Signal Processingal*, pp. 573~576, Mar. 1992.
12. M. Ostendort 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.
13. M. Lennig *et al.*, "Flexible vocabulary recognition for speech", *Proceedings of Int. Conf. on Spoken Lang Processing*, pp. 93~96, Oct. 1992.
14. G. G. Matison, "Emerging voice services in the NYNEX network", *Proceeding of Voice Systems Worldwide 1992* pp. 9~13, Feb. 1992.
15. S. Kuroiwa *et al.*, "Architecture and algorithms of a real-time word recognizer for telephone input", *Proceedings of Int. Conf. on Spoken Lang. Processing*, pp. 1523~1526, Oct. 1992.

구 명 완

1982 연세대학교 전자공학과 (학사)
1985 한국과학기술원 전기 및 전자공학과(석사)
1991 한국과학기술원 전기 및 전자공학과(박사)
1985 ~현재 한국통신 소프트웨어연구소 선임연구원
관심분야 : 음성인식 시스템 개발, 자동통역전화 시스템 개발
Neural Network, 음성합성