ETRI

전자통신동향분석 제20권 제5호 2005년 10월

음성인터페이스 기술

Voice User Interface Technology

정보통신 미래기술 특집

이윤근 (Y.K. Lee) 음성처리연구팀 선임연구원
박 준 (J. Park) 음성처리연구팀 팀장
김상훈 (S.H. Kim) 음성인터페이스연구팀 팀장

목 차

음성인터페이스 기술이란 인간의 가장 자연스러운 의사소통 수단 중의 하나인 '말'을 이용하여 기계와 인간과의 대화를 가능하게 하는 기술이다. 음성인터페이스 기술에 대한 연구는 1960년대부터 이루어져 왔으며 1990년대 후반부터 제한적으로 상용화되기 시작하였다. 아직까지는 기술적 한계에 의해서 간단한 명령어를 알아들을 수 있는 수준이며 응용 분야도 극히 제한되어 있으나 향후 텔레매틱스, 지능형로봇, 홈오토메이션 등의 신성장동력 산업이 활성화됨에 따라 기존의 키보드, 마우스 등의 인터페이스 수단들이 충분히 만족스럽지 않은 환경으로 변화하면서, 음성인터페이스 기술은 매우 중요한 대안으로 떠오르고 있다. 본 고에서는 음성인터페이스 기술의 기본 원리 및 요소 기술을 설명하고 관련 산업동향 및 응용분야, 그리고 향후 신성장동력 산업을 중심으로 한 발전 전망을 예측해본다.

Ⅰ. 서론

컴퓨터 기술이 급속히 발전하면서, 이제는 컴퓨터가 단순한 반복 작업이나 복잡한 계산처리를 대신하는 차원을 넘어 사람처럼 보고, 듣고, 느끼고, 학습하고, 생각하여 종합적으로 판단을 내리는 지능의 일부를 대체할 수 있는 수준에 이르고 있다.

음성인터페이스(voice user interface) 기술이란 일상생활에서 사람들 사이의 음성언어를 사용하여 정보기기를 제어하거나 정보서비스를 받을 수 있도록 말과 글을 처리하기 위한 기술의 한 분야로서, 음성신호에 내재되어 있는 정보를 분석하여 글자나 문장으로 나타내기 위한 음성인식 기술, 주어진 단어나 문장을 음성으로 들려주기 위한 음성합성 기술 그리고 누구의 음성인지를 구분해내기 위한 화자인식 기술 등을 말한다. 즉, 사람의 음성을 마이크 등의 입력 장치를 통해 컴퓨터가 입력 받아 이를 분석하여 문자로 나타내주고 또한 문자를 스피커 등의 출력 장치를 통해 사람의 음성으로 변환하여 들려주는 일련의 절차라고 할 수 있다. 이와 같은 음성인식과 음성합성에 더하여 언어 기술과 결합함으로써, 인식된 결과를 문법적으로 분석하고 그 의미를 이해하여 질문에 대답하거나 서로 대화하면서 적절한 기능을 수행하게 할 수도 있다.

음성은 사람에게 주어진 가장 자연스럽고 편리한 정보교류의 수단으로서, 사람과 사람간의 의사소통뿐만 아니라 사람과 컴퓨터의 의사소통, 나아가서는 컴퓨터를 매개로 한 사람과 사람간의 자동통역에 이르기까지 그 활용영역을 무한히 넓혀가고 있다. 음성의 이와 같은 편리성 때문에 이를 응용서비스에 도입하려는 시도가 계속되어 왔고, 실제로 일부 제한된 영역에서는 실용서비스에 성공적으로 적용되고 있다. 국내에서는 1980년대부터 대학 및 대기업연구소를 중심으로 음성 기술에 대한 연구가 진행되기 시작했으며 1990년대 중반부터 정부주도로 본격적인 음성 기술 연구에 착수, 2000년대 초부터 CTI를 중심으로 음성 기술이 상용화되기 시작하였다.

음성인식의 완성도와 음성합성 출력음질의 자연성을 높여감에 따라 연관 기술과의 결합 하에서 음성 응용 제품 및 서비스는 계속 확산될 것이다. 특히, 정보 기술의 급속한 보급과 더불어 이를 기반으로 하는 각종 정보기기가 소형화되고 이동성이 강화됨으로 인하여, 음성 기술은 이들 정보기기를 효과적으로 제어하는 데 필수 불가결한 인터페이스로서 주목을 받고 있다. 또한 유비쿼터스 시대를 맞이하여 우리의 일상적 환경에 컴퓨터가 숨어 있는 상황을 가정하면 음성인터페이스의 요구는 향후 폭발적으로 증가될 전망이다.

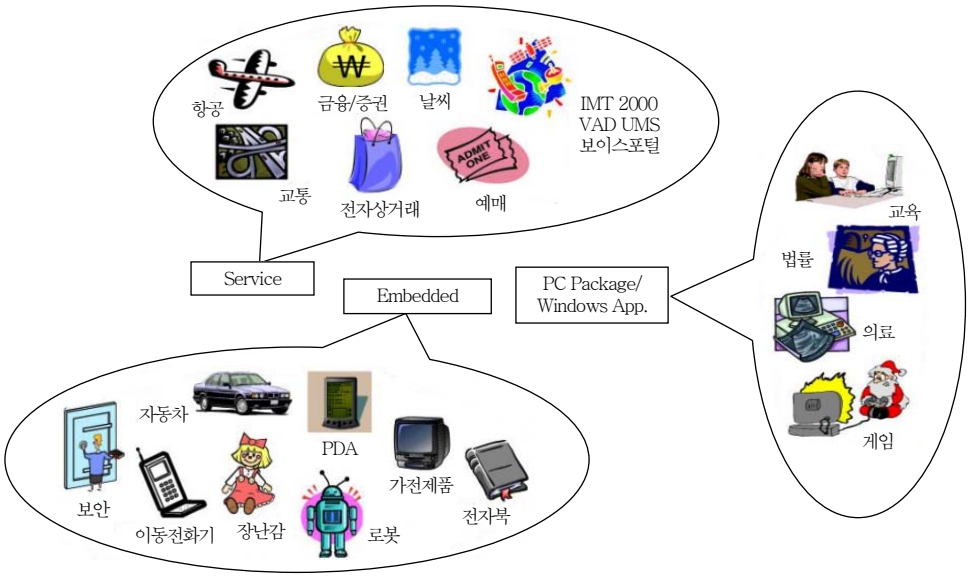
음성언어 정보처리 기술은 차세대 사용자 인터페이스 기술의 핵심요소로서 다보스 포럼 및 MIT의 미래예측에서 21세기 정보화 사회를 선도하는 10대 유망기술로 선정된 바 있으며, 선진국은 음성기술을 21세기 정보화 사회의 핵심 기술로 규정하고 관련 기술개발 및 음성 DB 자원 확보에 대규모 투자를 하고 있다. 우리나라에서도 신성장동력산업을 추진함에 있어서 음성 기술을 지능형로봇, 텔레매틱스, 홈네트워크, 차세대 PC 등 여러 산업분야에서 요구되는 공통 핵심 기술로서 지정하여 음성인터페이스 기술수요에 적극 대비하고 있다.

Ⅱ. 음성인터페이스 활용분야

현재까지의 음성인터페이스의 활용분야는 (그림 1)과 같이 크게 유/무선 통신망 환경 기반 서비스, 단말기 기반 응용 서비스, PC 기반 응용서비스로 나눌 수 있다.

1. 유/무선 통신망 환경 기반서비스

CTI 기반 무인 콜센터, 텔레뱅킹, 보이스웹 포털 등 응용서비스가 가장 활발한 분야이다. 현재까지 적용된 사례를 보면 증권정보 조회 시 종목 명을 발성하면 이를 인식하는 서비스, 항공 예약 시 출발지, 목적지 등을 발성하면 이를 인식하여 항공편을 조회하거나 예약할 수 있는 서비스, 폰뱅킹 시 계좌번호, 주민번호 등을 음성으로 입력할 수 있는 서비스 등



(그림 1) 음성인터페이스 기술의 활용분야

이 있다. 현재까지는 유/무선 음성 채널을 이용하여 음성인식을 시도하고 있으나 향후 유/무선 데이터 통신망의 전송속도가 비약적으로 증가함에 따라 데이터통신 채널을 이용한 분산형 음성인식(distributed speech recognition) 형태로 진화할 전망이다.

2. 단말기 기반 응용서비스

휴대폰, PDA, 차량단말기 등 정보단말기에 내장형 음성인식, 합성기능을 제공하고 있다. 음성으로 휴대폰의 주소록을 검색하여 전화를 걸어주는 VAD 기능, 차량단말기를 음성명령으로 제어하는 기능 등이 대표적인 사례이다.

3. PC 기반 응용서비스

PC의 응용 소프트웨어에 음성인식 기술을 적용하여 새로운 개념의 제품을 만들거나 기존 제품에 새로운 기능을 추가할 수 있다. 기존의 문서편집기에 음성인식 기능을 추가한 음성인식 받아쓰기 소프트웨어, 음성인식 기술을 이용하여 영어 발음의 정확도를 측정해주는 영어발음교정 서비스, 멀티미디어 콘텐트에서 특정 오디오 부분을 검색하는 오디오 인덱싱, 음성인터페이스가 적용되는 게임 등이 대표적인 사례이다.

4. 음성인터페이스 응용 분야의 향후 전망

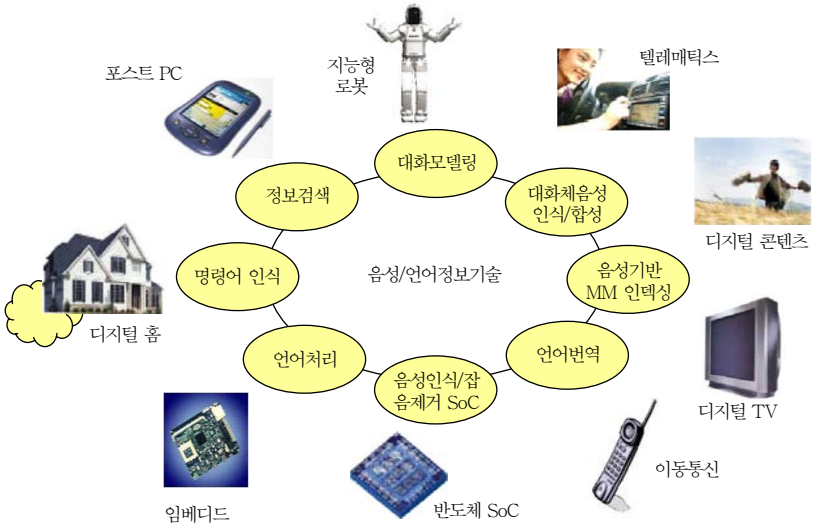
최근 신성장동력산업과 연계하여 지능형로봇의 대화형 음성인식 인터페이스, 텔레매틱스, 홈네트워크, 차세대 PC 등의 음성인터페이스, 디지털콘텐츠 검색 등의 응용분야가 주목받고 있다. 음성인터페이스 기술과 언어처리 기술이 융합되면 (그림 2)와 같은 여러 신성장동력산업의 요소 기술을 제공할 수 있다.

향후 음성인터페이스 기술은 기술적 완성도가 높아짐에 의해 (그림 3)과 같은 다양한 산업분야에 적용될 것으로 예측된다.

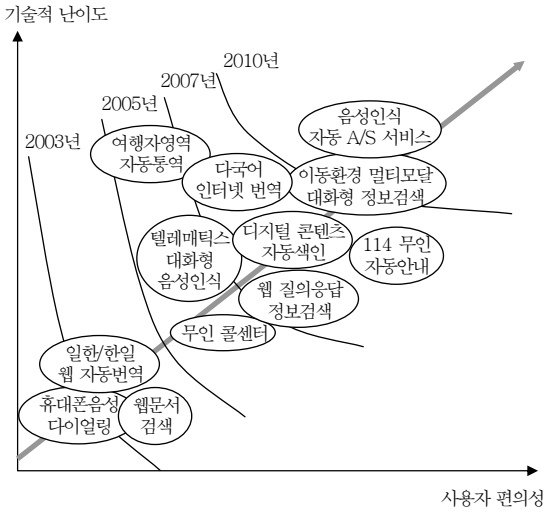
Ⅲ. 음성인터페이스 기술

1. 음성인식 기술의 개요

음성인식 기술은 컴퓨터가 인간의 음성을 알아들을 수 있는 기능을 구현하는 기술이다. 이는 인간의 귀와 초보적인 두뇌의 기능을 포함한다. 소리는 공



(그림 2) 신성장동력산업의 음성/언어 정보기술 응용분야



(그림 3) 음성인터페이스 기술 응용분야의 발전 전망

기의 진동에 의해 인간의 귀로 전달된다. 음성도 마찬가지이다. 귀는 음성의 주파수 특성을 분석하여 음성의 공진주파수, 주기성, 크기 등 여러 가지 정보를 추출해 낸다. 음성의 공진주파수는 음소의 특성과 밀접한 관계가 있으므로 이를 분석하여 발성된 음성을 인식한다. 사람이 음성을 인식함에 있어서 이와 같은 음향신호처리(acoustic signal processing) 측면만이 있는 것이 아니고 뇌에 저장되어 있는 여러 가지 지식 베이스(knowledge base)를 이용하는 측면이 중요하다. 즉 사람은 시끄러운 환경에서 발성되어 잘 들리지 않는 음성이나 너무 작게 발성되어 정확히 알아듣지 못할 정도의 음성인 경우에도 그 때의 상황이나 전후 문맥을 통해 유추 해석하여 음성을 인식할 수 있는 능력이 있다. 컴퓨터에 의한 음성인식의 경우에는 이와 같은 지식 베이스가 충분치 않아 아주 초보적인 수준의 기능만을 처리하므로 인식할 수 있는 능력이 인간에 비해 매우 떨어진다.

2. 음성인식 기술의 분류

음성인식 분야에 대한 연구는 약 1950년대부터 시작되었으므로 결코 짧지 않은 기간 동안 이루어져 왔다. 그러나 현재까지 선진국을 포함하여 그 기술 수준은 많은 한계를 가지고 있다. 음성인식의 궁극적인 목표는 인간과 마찬가지로 자연스럽게 발성되는 모든 음성을 인식할 수 있는 수준이어야 하지만 기술적인 한계가 있으므로 기술 수준에 따라 음성인식을 여러 단계로 나누어 연구를 수행하고 있다. 우선 인식할 수 있는 발성의 형태에 따라 다음과 같이 분류된다.

• 고립 단어 인식(isolated word recognition)

고립된 형태로 발성된 음성만을 인식할 수 있다. 즉 인식 가능한 대상 단어에 한하여 인식 가능하다. 음성인식의 가장 초보적인 단계이며 현재 가장 많이 상용화되어 있는 형태이다.

• 연결 단어 인식(connected word recognition)

여러 개의 단어를 연결시켜 발성하여도 인식 가능하다. 즉 인식 대상 단어의 연결 형태의 인식이 가능하다. 제한된 대상 단어의 조합으로 여러 형태의 음성인식이 가능하다. 고립 단어 인식에 비해 난이도가 높으며 인식률이 낮다. 대표적인 예가 연속 숫자 인식(connected digit recognition)이다. 현재 부분적으로 상용화되고 있다.

• 연속어 인식(continuous speech recognition)

자연스럽게 발성한 연속된 음성을 인식할 수 있다. 가장 난이도가 높은 단계이며 현재 선진국에서는 voice typewriter 등에 적용되어 출시되고 있다. 현재까지 인식률이 그다지 높지 못하며 특히 자연스러운 대화 형태의 발성의 경우 인식률이 매우 낮다.

• 핵심어 인식(keyword spotting)

자연스럽게 발성한 연속된 음성 중에서 인식 대상 단어만을 추출하여 이를 인식한다. 연속어 인식이 성능면에서 한계가 있으므로 이를 극복하여 특정한 분야의 상용화를 위해 이용된다. 즉 열차, 비행기 자동 예약 시스템 등에서 사용자가 발성한 여러 가지 정보 중에 지명에 해당하는 것만을 알고 싶을 경우 이러한 방식을 이용한다. 또는 상품 예약, 자동 콜센터 등에도 이용이 가능하다. 현재 부분적으로 상용화가 이루어지고 있다.

다음은 인식 대상 화자에 따라 다음과 같이 분류된다.

• 화자 종속 인식(speaker dependent recognition)

특정 화자 또는 사용자가 자신의 음성으로 미리 인식기를 훈련시키는 과정이 필요하다. 이 경우 인식기는 훈련된 음성만을 인식할 수 있다. (물론 목소리가 비슷한 경우 다른 사용자의 음성도 인식 가능하나 성능이 저하될 가능성이 있다.) 비교적 구현이 간단하여 단말기 등에 탑재되어 응용되고 있으나 사용자가 훈련하는 과정을 거쳐야 하는 불편함이 있으므로 제한된 분야에만 사용된다.

• 화자 독립 인식(speaker independent recognition)

임의의 화자의 발성을 인식할 수 있다. 미리 수백 또는 수천 명의 음성에 관한 정보를 추출하여 데이터베이스화 함으로써 별도의 훈련 과정 없이 어떤 사용자라도 사용 가능하다. 화자 종속 인식기에 비해 구현이 어려우나 현재 대부분의 상용화 시스템은 이 방식을 이용한다.

• 화자 적응(speaker adaptation)

화자 적응 방식은 화자 종속 및 화자 독립의 절충으로써 화자 독립 인식기의 경우 사용자가 자신의 목소리에 대한 인식률을 높이기 위해 화자 독립 인식기를 자신의 목소리에 적응시키는 방식이다. 이를 위하여 사용자는 인식기를 사용하기 전에 인식기가 요구하는 약간의 훈련 과정을 거쳐야 한다.

마지막으로 인식 대상 단어에 따라 다음과 같이 분류된다.

• 고정 단어 인식(fixed vocabulary recognition)

인식할 수 있는 대상 단어가 고정되어 있다. 가장 간단히 구현 가능한 방식이나 대상 단어를 바꾸려면 새로 갱신되는 대상 단어에 대해 여러 사람의 음성 데이터를 녹취, 분석하여 음성 모델을 구축하는 과정을 거쳐야 하므로 시간과 비용이 많이 소모되는 단점이 있다.

• 가변 단어 인식(flexible vocabulary recognition)

인식 대상 단어를 수시로 갱신할 수 있다. 즉, 음성인식기는 모든 음소에 대한 정보를 갖추고 있으면서 대상 단어가 갱신될 경우 음소의 조합으로 인식 대상 단어의 모델을 생성한다.

3. 음성인식 기술의 원리

음성인식 기술은 큰 범위에서 패턴 매칭 기법의 응용이라 볼 수 있다. 즉 인식 대상 단어 또는 음소의 특징 파라미터를 미리 저장하여 놓고 음성이 입력되면 이를 분석하여 특징을 추출한 후 미리 저장되어 있는 단어 또는 음소의 특징들과 유사도(likelihood)를 측정하여 가장 유사한 것을 인식 결과로 출력한다. 음성은 시간의 진행에 따라 변화하므로 음성의 특성은 짧은 구간(frame) 동안에만 안정적(stationary)인 특성을 갖는다. 따라서 음성의 특징은 각 프레임별로 분석되어 특징벡터가 생성되며 이 특징벡터들의 열로써 표현된다.

음성인식 시 특징 파라미터로 사용되는 것은 여러 가지 형태가 있으나 대부분 성도 특성을 나타낸다. 인간의 발음은 입 모양과 혀의 위치, 기타 유/무성 분류, 에너지 변화패턴 등에 의해 결정되며 이중 가장 중요한 요소인 입 모양과 혀의 위치에 의해 성도 특성이 결정된다. 반면 음의 높낮이 또는 음의 크기 등은 사람마다 또는 발성 환경에 따라 고유한 특성을 가지므로 발음을 인식하는 측면에는 좋지 않은 영향을 미친다. 발음 특성을 잘 나타내는 파라미터가 음성인식용으로 좋은 파라미터이며 대표적인 것으로 선형예측 분석에 의해 추출하는 LPCC, 귀의 인지특성을 고려한 MFCC 등이 있으며 이 외에 여러 가지 변형된 형태들의 파라미터들도 사용된다[1],[2].

음성인식의 방법은 크게 두 가지로 분류된다. 첫번째는 음성을 일종의 패턴으로 간주하여 등록되어 있는 패턴과 입력되는 패턴과의 유사도를 측정하여 인식하는 방법이 있다[3]. 두번째는 음성이 발성되는 과정을 모델링하여 각 대상 단어 또는 음소마다 고유의 모델을 할당하여 입력되는 음성이 어떤 음성 모델로부터 발생되었을 확률이 가장 높은지를 측정하여 인식하는 방법이 있다[4]. 이 밖에 신경회로망을 이용한 방법, 여러 가지 방법의 혼합형태 등이 있다. 첫번째 방법은 과거에 화자 종속 고립어 인식기에 많이 적용되었으나 화자 독립 인식기나 연속 음성인식기의 경우 성능 면에서나 계산량 면에서 문제가 있어서 최근에는 대부분 두번째 방법을 이용하고 있다. 음성을 모델링하는 방법으로는 HMM 기법을 주로 사용한다[5]. 이는 확률 모델의 일종으로써 Markov Model의 형태를 가지며 상태열(state sequence)이 은닉되어 있는 특징을 갖는다. 예를 들어 "ㄱ"이라는 음소가 N개의 상태를 갖는 HMM으로 모델링 될 경우 각각의 상태는 현재 상태에서 특정한 스펙트럼 특징을 나타낼 확률, 즉 출력 확률과 현재 상태에서 임의의 상태로 천이할 확률, 즉 천이 확률을 갖는다. 모든 음소의 모델은 위와 같은 고유의 출력 확률과 천이 확률을 갖는다. 음성인식 과정은 입력된 음성이 임의의 음소 모델로부터 생성되었을 확률을 계산하는 과정이다. 즉 입력된 음성이 "ㄱ", "ㅏ", "ㅇ"이라는 각각의 음소 모델로부터 순차적으로 생성되었을 확률이 가장 크다면 입력된 음성은 "강"이라는 말로 인식된다.

음성인식 과정에는 위와 같은 신호처리 측면만이 있는 것은 아니다. 인간이 말을 인식하는 과정을 생각해보자. 인간은 말을 들으면서 다음에 나올 말을 어느 정도 예측한다. 또한 앞에 나왔던 말이 정확하게 인식되지 않았을 경우 뒤의 말을 듣고 정확하게 인식하는 경우도 많다. 연속 음성인식기를 생각해보자. 출현 가능한 단어 수를 10만 개라고 했을 경우 인식기는 매 단어가 입력될 때마다 10만 개의 후보 중에서 한 개를 골라내야 한다. 이는 많은 오류율을 발생시키며 연산양도 매우 크다. 사람의 경우는 어떤 단어가 인식되었을 경우 다음에 따라 나올 단어의 종류는 문법상, 의미상 제한 조건에 의해 그렇게 많지 않다. 그 이유는 사람의 경우 뇌에 저장되어 있는 지식 정보를 이용하기 때문이다. 음성인식 과정에도 이와 같은 지식 정보를 초보적인 수준이나마 적용시키고 있는데 이를 언어 모델(language model)이라 한다. 대용량 연속 음성인식기의 경우 언어 모델로써 단어 천이 확률을 적용시킨다. 즉 인식 대상 단어를 N개라 했을 경우 임의의 n개의 단어가 연속하여 발생할 확률을 미리 데이터베이스화 하여 갖고 있다. 가장 간단한 경우, 즉 n이 2인 경우의 확률 모

델을 바이그램이라 하는데 N×N개의 가지수가 생기며, n이 3인 경우 트라이그램이라 하며 N×N×N개의 가지수가 생긴다. 이러한 확률 모델을 이용하여 연속하여 발생할 후보 단어의 범위를 축소시킴으로써 인식률도 향상시키고 연산양도 감축시킨다.

4. 음성인식 기술의 상용화 문제

음성인식 기술이 상용화되기 위해서는 해결해야 할 여러 가지 문제점이 있지만 그 중 가장 중요한 것이 음성의 왜곡 문제이다. 조용한 환경에서 성능 좋은 마이크로폰을 이용하여 입력된 음성은 비교적 인식이 잘 되는 편이다. 그러나 주변에 소음이 많은 환경, 즉 길거리, 버스나 지하철 안, 공장 등에서 음성인식을 시도하거나 선로 특성이 좋지 않은 유/무선 전화를 통하여 음성인식을 시도할 경우 인식 성능은 급속히 저하된다. 이러한 문제점을 극복하기 위해 다양한 기술이 응용된다. 우선 환경 잡음 또는 선로 왜곡 등에 강인한 성질을 갖는 음성 특징 파라미터를 이용한다. LPCC에 비해 MFCC의 특성이 잡음에 덜 민감하며 채널 왜곡을 없앨 수 있는 CMS, RASTA 필터링 기법 등도 이용된다. 또한 통계적 특성이 일정한 차량 잡음 등은 스펙트럼 차감법 등을 이용해 제거할 수 있다. 현재까지도 잡음 환경에서의 음성인식 분야는 중요한 이슈로서 연구가 지속적으로 이루어지고 있다.

5. 한국어 음성인식

음성인식 기술은 언어 특성에 관련된 부분이 많이 존재한다. 음성인식 연구가 선진국에서 출발된 분야이므로 현재는 영어에 대한 연구가 가장 많이 이루어져 있으며 유럽 각국 및 일본, 중국 등에서도 자국어에 대한 연구가 활발히 이루어지고 있다. 한국에서도 약 20여 년 전부터 학교, 국영 연구 기관, 대기업 연구소를 중심으로 한국어 음성인식에 관한 연구가 이루어지고 있으며 현재는 상당한 수준에 도달해 있다. 각 나라의 언어마다 고유한 음소 체계 및 문법 체계가 있으므로 각 언어의 음성인식기 개발은 그 나라에서 추진하는 것이 여러 가지 면에서 경쟁력이 있다. 한국어의 경우도 마찬가지이다. 음성인식 핵심 알고리듬은 이미 대부분이 개방되어 있는 상태이며 선진국의 수준과 국내 수준이 큰 차이가 없다. 한국어 음성인식기의 성능 향상을 위해서는 한국어 음성 및 텍스트 데이터의 축적, 한국어 언어 특성에 관한 지식 및 노하우 등이 필요하며 한국어를 모국어로 사용하고 있는 개발자가 유리하다. 또한 한국어는 영어와 음소 구성이 다를 뿐만 아니라 어미가 여러 가지 형태로 활용된다든지 매우 짧은 발음을 가진 조사에 의해 문장의 뜻이 완전히 달라지는 등 영어에 비해 어려운 점이 많이 있다. 또한 중요한 정보를 전달하는 숫자의 경우 모두 단음절로 이루어져 있으며 비슷한 발음이 많이 존재해 음성인식을 적용하기가 매우 어렵다.

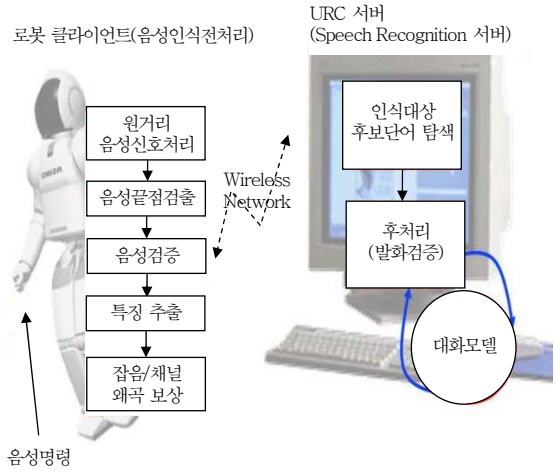
6. 음성인터페이스 요소 기술

음성인터페이스 기술은 신호처리 기술부터 언어처리 기술까지 다양한 영역의 기술의 조합으로 구성된다. 각 기술영역별 요소 기술은 <표 1>과 같다.

이러한 요소 기술들이 어떠한 형태로 구성되고 어떠한 절차를 거쳐 음성인식이 수행되는지 로봇의 음성인터페이스를 예로 설명하면 (그림 4)와 같다. 각 요소 기술별 자세한 설명은 다음과 같다.

〈표 1〉 음성인터페이스 기술영역별 요소 기술

기술영역	음성인터페이스 요소 기술
음성신호분석 및 전처리	특징추출, 잡음처리, 채널보상, 음성향상, 원거리 음성인식, 어레이 신호처리, 음성검출, 음원분류, 음원위치식별
음향모델링 및 탐색	음향모델훈련, 화자적응, 핵심어검출, 가변어휘 인식, 탐색구조, 미등록어 처리
음성인식 후처리	발화검증, 음성인식거부
대화모델링	대화모델링, 언어모델링, 대화주제탐색, 의사형태소 해석, 무의미어 처리, 문법구조, 시소러스, 언어생성
음성합성	DB 설계, 음소분할, 피치추출, 합성단위 선정 및 접합, 합성단위 압축, 음성신호처리, 음색변환, 텍스트 전처리, 발음변환, 운율모델제어, 양태표현, 감정표현



(그림 4) 로봇의 음성인터페이스 기술 흐름도

가. 음성신호분석 및 전처리 기술

입력된 음성신호를 분석하여 음성인식에 사용될 특징파라미터를 추출한다. 이 경우, 음성발화 시 주변 환경소음, 음성입력 시 녹음장비의 특성 등이 일반적인 서비스 환경에서의 음성인식 성공률을 저하시키는 주요 원인이 된다. 따라서 소음이나 녹음장비에 따른 음성왜곡을 제거하기 위해 잡음/채널왜곡 제거 기술이 요구된다. 특히 로봇의 경우, 가정환경 소음을 제거해야 하며, 텔레매틱스의 경우, 자동차 운전중 발생하는 주행 소음을 제거해야 한다. 그 외에 발화된 음성구간만 검출하여 음성인식 엔진으로 전송하는 음성구간검출 기술, 음성/비음성 검증기술도 실제 응용환경에서 해결해야 할 매우 중요한 기술이다.

나. 음향모델링 및 탐색 기술

불특정 다수의 화자로부터 다양한 발음 특성을 분석하여 이를 모델링한다. 이렇게 생성된 음향모델은 음성인식 과정에서의 기준 패턴(또는 모델)의 역할을 수행한다. 음성이 입력되면 입력음성과 확률적으로 가장 유사한 인식후보를 선정하여 인식 결과로 출력한다. 인식 대상 어휘 수가 많을 경우, 인식후보 고속탐색 기술이 요구된다. 인식후보 탐색 시 중요 어휘만 추출하는 핵심어검출 기능에 의해 사용자의 편의성을 증대시킬 수 있다.(예: “어 오늘 날씨가 어때”라고 발화했을 때 “오늘 날씨”만 핵심어로 검출하여 인식결과 출력)

다. 음성인식 후처리 기술

인식결과가 어느 정도 신뢰를 가지는지 최종 인식결과를 출력하기 전에 한번 더 검증하는 과정이 필요하다. 오인식 검증을 통해 신뢰도가 낮은 결과는 사용자에게 다시 되물어보거나 인식 실패로 간주하기도 한다.

라. 대화모델링 기술

사용자가 음성인터페이스 시스템에 대한 사전 지식이 없더라도 사용자가 원하는 작업을 빠른 시간 내에 완료할 수 있도록 시스템이 협력해주는 대화모델 기술이 요구된다. 음성인터페이스 시스템이 좀더 사용자에게 친숙해질 수 있기 위해서는 사용자의 시스템 사용행태를 분석하여 대화모델에 반영하여야 한다.

마. 음성합성 기술

최근 음성합성 기술은 자연스럽게 발성된 문장단위 음성을 대용량(일반적으로 30시간 이상 분량)으로 녹음, 이로부터 음소 단위를 반자동으로 추출하여 합성단위로 사용하는 코퍼스 기반 음성합성 기술이 널리 사용되고 있다. 문장단위 발화음성에는 이미 생동감 있는 운율이 내재되어 있고, 합성단위간 연결 시 부조화를 최소화하도록 합성단위를 선정하여 연결한다면 기존 운율처리로 인한 음질열화 없이 사람이 발성하는 듯한 자연스러운 합성음을 생성할 수 있다. 따라서 코퍼스기반 합성 기술에서는 대용량 음성데이터베이스를 구축하는 기술과 복수 개의 합성단위로부터 최적 연결 합성단위를 선정하는 기술, 그리고 구단위로 끊어 읽고, 제대로 된 발음으로 변환하는 기술이 음질을 좌우하게 된다.

〈표 2〉 음성인터페이스 기술 로드맵

		2004년 이전	2005년	2006년	2007년	2008년	2009년~2012년
환경변화	통신인프라	유무선 전화망 기반 CTI		고속 무선데이터망(WIBRO) 기반 DSR			
	User Interface	개별 모달리티 입력(예: 음성, 키보드, 펜)		멀티 모달리티로 융합 입력(예: 음성+ 키보드+ 펜)			멀티모달 확장(예: 제스처, 비디오)
기술변화	음성인식	중규모(수천~수만 단어급) 명령형 단어/핵심어인식		대규모(수십만~수백만) 대화형 핵심어/연속음성인식			감정인식
응용변화	지능형 로봇/홈네트워크	제한영역 근거리 명령어 인식		제한영역 멀티모달 기반 원거리 대화형 인식		자유영역 멀티모달 기반 원거리 대화형 감정인식	
	텔레매틱스	소규모 제어명령		멀티모달 기반 대규모(수십만 단어급) 목적지 인식		멀티모달 대화형 정보검색	
사용자변화		명령(수동적)		대화(협동적)			공감(능동적)

7. 음성인터페이스 기술 로드맵

지능형로봇 및 텔레매틱스를 중심으로 음성인터페이스 기술 및 환경변화 등을 전망해보면 <표 2>와 같다.

환경 변화의 측면에서 보면 기존의 음성인식 서비스가 CIT 기반에서 주로 이루어짐으로써 음성통신망을 사용하던 것이 향후 무선 데이터통신 속도의 향상에 의해 DSR 형태의 음성인식 구조로 진화할 전망이다. 데이터 통신망 기반에서 음성인식이 이루어질 경우 타 모달리티와의 융합이 용이해지므로 다양한 멀티모달 인터페이스가 발전할 것으로 전망된다. 기술 변화의 측면에서는 인식대상 단어의 규모가 증가하며 사용자의 발화형태의 제한이 완화되는 핵심어, 연속어 기반의 음성인식으로 발전할 것으로 예측된다. 이러한 환경 및 기술의 변화에 의해 지능형로봇, 홈네트워크, 텔레매틱스 등 산업에서의 음성기술 응용분야는 점차 확대되고 사용자 편의성도 증대할 전망이다.

Ⅳ. 음성인터페이스 산업동향

1. 산업동향

음성인식 콜센터, 보이스포털, 단말기 내장형 음성인식, 텔레매틱스, 딕테이션, 교육, 게임, 완구, 가전, 로봇 등 다양한 산업에서 음성인터페이스를 적용하고 있다. 음성인식 딕테이션 소프트웨어는 1997년부터 시판되기 시작했으며 Dragon Systems의 NaturallySpeaking과 IBM의 ViaVoice 등이 대표적인 제품이다. Scansoft는 다국어 음성인식 엔진을 개발하여 CTI 분야 및 embedded 분야에서 세계시장 점유율 1위를 차지하고 있다. HMIHY("How May I Help You?") 서비스는 미국 AT&T사가 자사 고객센터를 위한 콜센터에 도입한 음성인식 콜센터 서비스로 음성인터페이스의 대표적인 성공사례이다.

삼성과 스프린트사는 최근에 개발한 VI660 핸드폰에 보이스시그날사의 내장형 음성인식을 탑재, 숫자를 연속으로 말해서 전화를 걸도록 하는 기능을 내장하였으며 향후 음성으로 SMS를 입력할 수 있는 기능을 개발할 예정이다. 현대모비스의 텔레매틱스 단말기인 eXride에 보이스웨어의 50단어 수준의 음성인식 엔진이 채택되어 상용화 되었으며 SK는 2002년 3월부터 "네이트 드라이브"를 미국 스피치웍스의 음성인식 솔루션과 국내 코아보이스의 음성합성 솔루션을 채용하여 서비스중이다. 국내외 연구기관도 음성인터페이스의 실상용화를 위해 국가적인 프로젝트를 수행하고 있다. 산학연관 공동으로 지능형로봇, 텔레매틱스, 차세대 PC, 홈네크워크 등 신성장동력산업에서 소요되는 음성인터페이스 개발(사업명: 언어정보처리기술 개발, 2004~2006년)에 착수하였으며 텔레매틱스의 경우, 산업자원부 지원 자동차용 음성인터페이스 기술 개발(사업명: 자동차용 음성 HMI 시스템 기술 개발, 2001~2004년) 사업을 수행하고 있다.

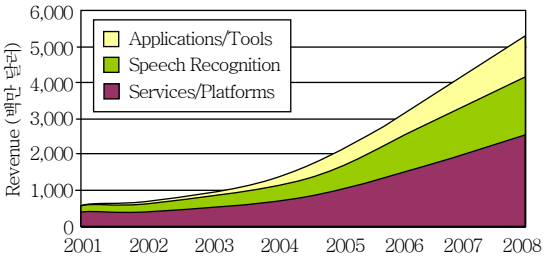
미국은 DARPA 프로젝트(1992~1999년)를 수

행하여 항공편 문의시스템, 전화망 대화 연속음성인식, 방송뉴스인식, 지능형 대화인터페이스 기술을 개발하였으며 2004년 유럽에서는 6th Framework 프로젝트에 착수, 미국 CMU 대학과 독일 Karlsruhe 대학이 공동으로 휴먼인터페이스 기술(CHIL)을 개발하였다.

음성인식 관련 표준화 작업도 진행중이다. W3C에서는 IBM 주도로 전화망 음성인터페이스 표준화안인 VXML 버전 2.0을 발표하였다. XML 기반의 음성인터페이스는 현재 단어와 단문을 인식할 수 있는 단계에 있으며 향후 자연스러운 대화체 인식 및 합성이 가능해 질 것으로 보인다. XML 2.0은 전화망환경 음성인터페이스 애플리케이션을 위해 설계되었고 IBM, Motorola, AT&T, Lucent Technologies 등이 참여하였다. 텔미네트웍은 XML 기반 플랫폼을 개발, 음성인식 정보서비스를 ASP 사업화하였고 KT는 외부 전화망을 통해 가정 내의 가전제어가 가능한 XML 기반 음성인터페이스 서비스를 개발하였다. 콜센터 기반 CRM 전문업체 MPC는 VXML 기반의 음성인터페이스 서비스 플랫폼을 개발, KTF와 삼성카드에 VXML 기반 포털서비스 시스템을 공급하였다. 마이크로소프트가 주도하는 SALT 포럼을 통해 음성을 포함한 멀티모달 인터페이스 표준화가 추진되고 있으며 Cisco Systems, Intel, Philips Electronics 등 20여 회사가 참여하고 있다. SALT는 PC용 비주얼 웹브라우저를 위한 HTML 또는 XHTML이나, 휴대전화 또는 PDA용 웹 브라우저를 위한 WML에 내포될 수 있도록 설계되었다. 웹 애플리케이션에 음성인터페이스용 태그를 첨가할 수 있어 마우스나 키보드 사용 외에 음성명령으로 소프트웨어를 제어할 수 있다. 마이크로소프트사는 Visual Studio.Net과 인터넷 익스플로러 .Net initiative에 SALT 호환 음성 인터페이스 엔진을 개발, 제공할 계획이다. 유럽 ETSI Aurora 그룹에서는 지난 2000년부터 음성인터페이스 전처리부인 ETSI standard DSR front-end를 표준화하였으며 휴대폰 통신단말에서는 음성전처리만 수행하고 음성인식은 서버를 사용하는 DSR 방식의 음성인터페이스를 개발하였다[6].

2. 시장동향

Allied Business Intelligence(ABI)사는 (그림 5)에서 보는 바와 같이 음성인식 세계시장은 2004년 12억 달러에서 2008년에는 53억 달러까지 신장할 것으로 예측하였다.



<자료>: Allied Business Intelligence Inc.

(그림 5) 음성인식 세계시장 규모

국내 음성인식 시장은 <표 3>에서 보는 바와 같이 2002년도에 27억3,000만 원, 2004년도에 49억 4,0000만 원, 2009년도에 780억 원 시장을 형성하며 연평균 64%의 성장률을 보일 전망이다. 특히, 2003년도에서 2004년도 사이에 63.1% 성장률을

〈표 3〉 국내 음성인식 시장 규모 – Moderate 전망

단위(백만 원)	2002	2003	2004	2005	2006	2007	2008	2009
텔레매틱스	93	248	560	1,387	2,253	4,562	11,126	25,590
휴대전화 단말기	132	275	662	2,080	3,781	7,392	10,559	14,117
서버시장(CTI)	2,509	2,509	3,721	4,666	7,711	16,632	27,202	39,188
인식 Total	2,734	3,032	4,944	8,133	13,745	28,586	48,887	78,895

<자료>: CNET Research & Consulting

기록하고, 2006년도에서 2007년 사이 108% 성장을 보임으로써 2004년과 2007년은 주요 터닝 포인트가 될 것으로 보인다. 이는 휴대폰 기반 시장과 텔레매틱스 기반 시장이 가파르게 성장하고 CTI 시장이 완만한 성장세를 보이며, 음성인식 솔루션의 장착 비율이 증가하기 때문으로 분석된다.

3. 음성인터페이스 기술의 주요 응용분야

가. 지능형 로봇

지금까지 대부분의 로봇은 산업용으로 사용된 예가 많았다. 그러나 21세기에 접어든 지금, 다양한 지능형 로봇이 선을 보이고 있다. 정보통신을 기반으로 한 디지털 사회의 구현을 위해 사람과 함께 동일한 공간에서 생활하면서 사람에게 즐거움과 유익한 서비스를 제공할 수 있는 로봇의 필요성이 대두되기 때문이다[7].

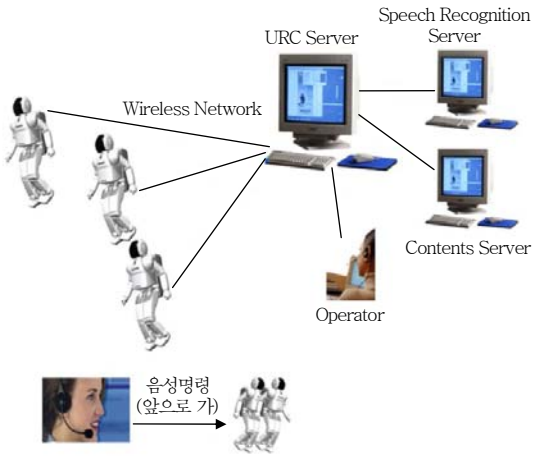
로봇이 사람과 같이 생활하면서 우리 생활에 편리를 제공하기 위해서는 사람들과의 상호작용 능력이 중요하며 그러기 위해서는 이동이 자유로워야 하고, 주변 환경뿐만 아니라 사람도 인식할 줄 알아야 하며, 통신 및 대화가 가능한 인간친화적인 지능형 로봇이어야 한다.

로봇과 사람이 말로서 의사소통을 하기 위해서는 음성인터페이스 기술이 필수적이다. 나아가서 현재까지 대부분의 음성기술 적용분야가 명령어 중심의 단어인식 형태였음에 반해 로봇의 음성인터페이스는 사람과 대화하듯이 자연스러운 대화 형태의 발성을 인식하고 그로부터 핵심적인 단어를 인식하여 이해할 수 있는 기능이 필요하다. 또한 대화를 이끌어 나갈 수 있는 능력도 필요하다. 즉, 핵심어검출 기능, 대화모델링 등의 기술이 접목되어야만이 로봇의 음성인터페이스가 가능해진다. 현재의 기술 수준으로는 수백 단어 정도의 핵심어 검출이 가능하며 제한된 영역에서의 대화모델링이 가능하고, 2007년도에는 수천 단어급의 핵심어 검출이 가능한 정도로 발전될 전망이다.

로봇은 일반 가정환경에서 생활하므로 음성인식도 일반 가정 소음 환경에서 이루어진다. 또한 로봇에 마이크로폰이 장착되어 있는 상태에서 원거리에서 사람이 말을 하는 것을 알아들어야 하므로 잡음에 많이 노출된 상태에서 음성이 입력되게 된다. 이 경우 음성인식 성능이 떨어지게 되므로 입력장치에 마이크로폰 어레이를 사용하거나 능동 잡음 제거 기술 등을 접목시켜서 인식 성능을 향상시키는 노력을 진행하고 있다.

지능형 로봇의 대표적인 예로써 URC가 있다. URC는 (그림 6)처럼 환경인식이나 음성인식 등과 같이 로봇이 수행하는 핵심 기능을 통신 네트워크를 통해 외부 서버에 분담, 로봇의 하드웨어 구성을 단순화하고 네트워크를 통해 교육, 무인방범, 맞춤정보 등 일상생활에 필요한 다양한 정보와 서비스를 제공한다[8]. 이는 로봇을 네트워크에 연결된 움직이는 정보통신 단말기로 보는 개념이다. 이 경우에는 다양한 콘텐트가 디지털네트워크상의 여러 콘텐트 서버로부터 제공되므로 다양한 정보서비스가 가능하고, 음성인식 기능을 담당하는 별도의 서버가 네트워크상에 존재하며 여기에서 음성인식 기능을 담당할 수 있으므로 컴퓨팅파워의 제약 없이 고성능의 음성인식 기능을 제공할 수 있다.

현재 URC는 정통부의 IT839 전략의 하나인 지능형 로봇 연구개발 과제로서 진행되고 있으며 정보콘텐츠 로봇 개발, 공공도우미 로봇 개발, URC 인프



(그림 6) 지능형로봇 URC의 개념도

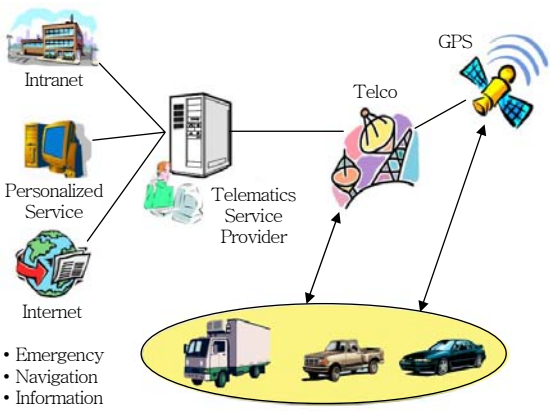
라 시스템 개발, 미들웨어 등 다양한 핵심 요소기술, 지능형 로봇 센서 개발 등으로 구성되어 있고 음성기술은 대부분 로봇의 핵심 인터페이스 기능으로 적용될 예정이다.

나. 텔레매틱스

텔레매틱스 산업은 우리가 강점을 가지고 있는 자동차 산업과 정보통신 산업의 기술 융합을 통해 자동차 산업의 부가가치를 높일 수 있으며, 이동통신 등 정보통신 산업에 투자확대 계기를 마련함으로써 우리 경제에 새로운 성장동력을 제공해 줄 것으로 기대되고 있다. 특히, 텔레매틱스 산업은 수출 주력 산업으로 중요성이 강조되고 있는 이동통신 산업과 자동차 산업의 진화를 선도하고 있다는 점, 그리고 소프트웨어 및 첨단 단말 산업에도 기여할 수 있다는 점 등 여러 측면에서 차세대 성장동력 중에서도 전략적 중요성이 매우 높은 분야로 인식되고 있다[9].

텔레매틱스(telematics)는 원격통신(telecommunication)과 정보과학(informatics)이 결합된 용어이다. 텔레매틱스는, 운전자가 원하는 목적지까지 운전경로를 찾아주는 단순한 기능의 차량 네비게이션 시스템과는 달리, (그림 7)과 같이 통신 및 방송망을 통하여 자동차를 사무실과 가정에 이은 제3의 인터넷 공간(connected car)으로 재구성한다. 이렇게 함으로써 위치추적, 인터넷 접속, 원격 차량진단, 사고감지, 교통정보 제공 등의 기능을 제공할 뿐만 아니라 홈 네트워크, 사무자동화 등과 연계함으로써 가정과 사무실에서 이용하는 서비스를 자동차에서도 단절없이 제공할 수 있는 첨단 시스템이다.

텔레매틱스 서비스는 차량 내에서 이루어지므로 대부분 사용자가 운전중인 상태에서 이루어지는 경우가 많다. 운전자는 시선은 전방을 주시하고 손은 핸들 위에 놓여 있는 상태이므로 정보서비스를 받기 위한 다양한 기기 조작을 하기 어려운 상태이다. 이 경우 가장 유용한 인터페이스 수단은 음성인터페이스이다. 현재 텔레매틱스 정보서비스 중에서 가장 중요한 기능은 길안내 및 교통정보 서비스이다. 길안내를 위해서는 목적지 설정을 해야 하는데 이를 음성인식 기술을 이용함으로써 편리한 기능을 제공할 수 있다.



(그림 7) 텔레매틱스의 개념도

〈표 4〉 국내 텔레매틱스 산업의 음성 기술 적용사례

단말기 적용 사례

업체 명	모비스	LG 전자(개발중)
모델 명	eXride	MTSII, I
음성솔루션	보이스웨어	보이스웨어
인식적용범위	명령어+ 주소록	명령어+ 주소록+ 등록목적지

서비스 적용 사례

업체 명	현대차(MOZEN)	SK	삼성화재
음성솔루션	보이스웨어	Scansoft (구 Speechworks)	보이스웨어
서비스 내용	• 음성인식 적용 - VAD - 생활 정보 - 유머, 영어 등 • 상담원 연결 - 길안내(음성인식 적용예정: 2005년)	• 음성인식 적용 - 길안내 • 상담원 연결	• 음성인식 적용 - VAD - 생활 정보 - 교통상황 안내 - POI 검색 • 상담원 연결 - 길안내 - SOS

국내에서는 SK가 최초로 음성인식 길안내 서비스를 상용화하여 현재 서비스중에 있으며 현대자동차도 뒤를 이어 준비중이다. 또한 텔레매틱스 전용 단말기에 음성인식 기능이 탑재된 예도 많이 있는데 주로 음성을 이용한 메뉴 선택, 전화걸기, 목적지 설정 등의 기능으로 사용된다(<표 4> 참조).

텔레매틱스 분야는 음성인터페이스가 가장 필요한 분야이면서도 기술적으로는 적용하기 어려운 분야이다. 차량 주행중에 발생하는 소음으로 인하여 음성인식 성능이 저하되는 현상이 발생하는데 이를 위하여 능동적 소음 제거에 대한 연구가 진행중에 있다. 또한 목적지(POI)의 개수가 수십만 개~백만 개 이상 되므로 대용량의 단어를 실시간으로 인식할 수 있는 기술에 대해서도 연구가 진행중이다.

다. 자동통역

자동통역 기술이란 언어장벽을 허물어 서로 다른 언어를 사용하는 사람 간에 대화가 가능하게 하는 기술이다. 자동통역 기술이 실현되면 현재 진행되고 있는 세계화가 가속되어 개인 생활이나 기업활동, 사회 전반에 지대한 영향을 미칠 것이다. 이러한 자동통역을 실현하기 위해서는 여러 요소 기술이 필요한데, 먼저 사람이 발성한 소리를 문자로 나타내는 음성인식 기술과, 이를 같은 의미에 해당하는 상대 언어의 문장으로 변환하는 언어번역 기술, 그리고 문자로 표기된 문장을 음성으로 읽어 주는 음성합성 기술 등의 요소 기술을 확보하여야 한다. 또한, 적어도 두 개 이상의 언어를 처리하여야 하며 각각의 언어에 대하여 요소 기술들을 개발하여야 한다. 최근 지난 수십 년에 걸친 기술 축적의 결과로 그간 상상 속에서만 존재하던 자동통역 기술의 실현이 현실로 다가오고 있다[10].

다국어간 자동통역에 필요한 기술을 효율적으로 개발하기 위하여 국제 자동통역 공동연구 컨소시엄(Consortium for Speech Translation Advanced Research, 이하 C-STAR라 칭함)이 결성되어 관련 연구를 진행하고 있다. 이 C-STAR에는 일본의 ATR 연구소, 미국의 카네기멜런대학교 등 자동통역 분야의 첨단 연구기관이 대거 참여하고 있으며, 한국전자통신연구원도 핵심그룹의 일원으로 참여하고 있다. 자동통역이 필요한 작업영역으로써 외국의 여행사 직원과 여행계획을 수립하는 상황을 설정하여, 기술 개발을 공동으로 추진하였으며, 1999년 7월 22일에 국제간 자동통역 실시간 시연을 통하여 그 연구결과를 공개하였다. 이어서 2000년부터는 자동통역 서비스의 상용화를 위한 기술 개발에 주력하고 있다[11].

자동통역에 필요한 음성인식 기술이나 음성합성기술은 기본적으로 위에서 다룬 내용과 같다[12],[13]. 그런데, 사람과 기계간에 사용되는 음성인터페이스와 달리 자동통역은 사람과 사람간의 대화가 처리 대상이다. 따라서, 대화체에 나타나는 여러 특징을 다루어야 한다. 대화체 발화에서는 반복, 도치, 수정, 생략 현상이 자주 발생하며, 간투사도 수시로 삽입된다. 또한, 하실래요, 할께요, 한가요 등과 같이 대화체에 고유한 어미가 다양하게 나타난다. 따라서, 기존의 음성인식 및 합성 기술에 이러한 현상을 구현하는 작업이 필요하다. 또한, 대화체 발화의 비정형성으로 인하여 특히 문장번역 기술에서는 기존의 문법기반의 방법의 적용이 매우 어렵다. C-STAR에서는 대화체 문장을 처리하기 위한 방법으로써 개념 기반의 중간언어를 통한 번역 방식에 대한 연구를 수행한 바 있으며, 최근 대규모 대역코퍼스를 기반으로 통계적 번역방식을 연구하고 있다.

중간언어 기반의 번역이란 각 언어 사이를 매개하는 중간언어를 정의하고, 이 중간언어를 통하여 언어번역을 수행하는 것이다[14],[15]. 특히, 이 방식은 각 언어에 대하여 음성인식 기술, 음성합성 기술, 문장해석 기술(인식결과를 해석하여 중간언어를 생성), 문장생성 기술(중간언어로부터 문장을 생성)을 개발하면, 중간언어를 지원하는 다른 어느 언어와도 번역이 가능하다는 장점을 갖는다. 이와 같은 장점에도 불구하고, 1) 문장 해석 및 생성 작업에 필요한 문법의 작성을 수작업으로 수행하여 개발 시간이 많이 걸리고, 2) 작업영역을 확대하면서 문법을 증강할 경우 기존 문법과 새로 작성한 문법 간

에 일관성 유지가 어렵게 되며, 3) 작업 영역이 바뀌는 경우, 문법을 대부분 재작성하여야 하기 때문에 작업영역에 대한 이식성이 떨어지는 문제점을 갖고 있다.

이러한 문제를 해소하기 위하여 최근 통계적인 접근 방식에 대한 연구를 진행하고 있다. 기본 아이디어는 대역 코퍼스로부터 번역단위인 '구'를 통계적으로 자동 추출하고 입력 문장을 '구' 단위로 자동 분할하여 이를 번역한 후, 번역된 '구'를 적절히 재배치하여 상대언어에 해당하는 문장을 생성함으로써 번역을 수행하고자 하는 것이다. 이때, 번역을 하고자 하는 언어간 '구'의 대응관계나, '구'의 재배치 규칙 등을 모두 대역코퍼스로부터 통계적인 방법으로 추출하여 사용한다. 대화체 문장에 대한 이러한 시도는 최근 시작되었으며, C-STAR 주도로 2004년부터 대화체 문장번역을 주제로 한 워크샵이 개최되고 있다[16].

자동통역 기술이 실제 생활에서 보다 다양하게 응용되기 위해서는, 앞으로도 많은 연구개발 노력이 필요하다. 음성인식에서는 현재 초기상태에 있는 언어정보 활용 수준을 단순 문형정보나 n-gram을 활용하는 수준에서 의미분석을 통한 문맥정보를 활용할 수 있는 수준으로 끌어 올려야 하며, 응용분야에 대한 지식을 활용하는 기술도 접목되어야 할 것이다. 언어번역에서는 인식오류에 대하여 강인하게 대처, 수용하는 기술이나 응용 영역이 변경되어도 이를 유연하게 대처할 수 있도록 이식성이 보장되는 기술이 개발되어야 할 것이다. 이와 함께, 사용 리소스를 감축하여 소용량 플랫폼에서도 동작하게 하는 노력도 아울러 지속적으로 기울여야 할 것이다.

V. 음성인터페이스 산업전망 및 결론

1. 발전전망

음성인터페이스 기술수준이 높아짐에 따라 음성 응용서비스가 사회 전반적으로 확산될 것으로 예상되며 수요고객층이 세분화되어 다양한 음성인터페이스 응용제품 출시가 가속화 될 것으로 예상된다. 현재는 기술적 한계에 의해 수십~수백 단어 수준의 음성명령어 상용화에 집중되고 있으나, 향후 언제 어디서나 정보를 획득할 수 있는 유비쿼터스 환경에서는 단문 수준의 대화가 가능한 3,000~5,000단어급 음성인터페이스 기술이 개발될 것으로 예상된다. 2010년 이후에는 음성, 펜, 마우스, 제스처 등 다양한 입력장치를 통합한 멀티모달 인터페이스로 발전할 것이며, 음성인터페이스는 상황정보, 화자의 의도를 파악하여 고수준의 대화체 인식이 가능하게 될 것으로 전망된다.

2. ETRI 음성인터페이스 개발 성과 및 향후 기술개발 계획

ETRI는 1990년 중반부터 선도기술개발에 주력하여 대화체 음성인식, 대어휘 방송뉴스인식, 자동통역 기술을 개발, 1999년 다국간(미국, 일본, 독일, 프랑스, 이탈리아) 자동통역 국제시연을 성공적으로 수행하였다. 장애인용 음성인터페이스 기술, 국내업체 경쟁력 제고를 위한 음성인터페이스 애로 기술개발, 명령어 TTA 표준화, 표준형 공통음성 DB 구축 및 배포를 통해 국내 음성산업의 기반 강화에 노력하였으며 2004년부터 정보통신부 지원으로 신성장 동력산업에서 요구되는 공통핵심 기술인 음성인터페이스 기술을 산학연 공동으로 중점 개발하고 있다. 또한 음성인터페이스 실상용화 시 난제를 해결하기 위한 breakthrough 기술개발에도 힘을 기울이고 있다. 2007년에는 제한된 영역에서 상황판단, 화자의도분석 및 지식추론이 가능한 2,000단어급 단문수준의 대화형 음성인터페이스를 개발하고, 20만 단어 이상의 지명을 인식할 수 있는 대용량 인식기술을 개발하여 텔레매틱스 서비스에 적용할 계획이다. 또한 감정표현이 가능한 음성합성 기술도 개발할 계획이다. 2008년 국내 기술수준을 국내시장 점유율 90%, 2010년 세계시장 점유율 10%를 달성하며, 동 분야에서 세계 선도기업인 Scansoft, IBM과의 기술격차를 1년 이내로 축소하고자 한다.

3. 결론

음성인터페이스 산업은 모든 IT 산업에 새로운 인터페이스를 제공할 수 있는 기반산업이며, 현재 정부가 수립·추진하고 있는 IT839 전략 중 지능형 로봇, 텔레매틱스, 차세대 PC와 홈네트워크 등의 성공적인 사업수행을 위해 필요한 핵심기술이므로 음성솔루션 및 지적재산권을 국내기술로 확보하는 것이 해당 분야 산업의 육성에 절대적으로 필요하다. 또한 유럽의 Aurora, W3C의 VXML, MS가 주도하는 SALT 포럼 등 국제표준화 활동을 통해 음성인터페이스의 상용서비스를 준비하고 있으며 이로 인한 외국업체의 국내외 시장 선점이 우려되고 있어 국내업체/연구소/학계의 대처가 시급하다. 타산업에 대한 파급효과를 고려하고 핵심기술의 조기확보를 위해 정부의 적극적인 의지에 의한 육성책 마련과 기술개발에 대한 재정적 지원, 전문인력 양성, 기술표준화를 위해 관련법 제정이 필요하다.

약어 정리

ASP	Application Service Provider
CHIL	Computer-Human In the Loop
CMS	Cepstrum Mean Subtraction
CTI	Computer Telephony Integration
DARPA	Defence Advanced Research Projects Agency
DSR	Distributed Speech Recognition
ETSI	European Telecommunication Standards Institute
HMI	Human Machine Interface
HMM	Hidden Markov Model
LPCC	Linear Prediction Cepstral Coefficient
MFCC	Mel Frequency Cepstral Coefficient
POI	Point of Interest
RASTA	RelAtive SpecTrAl
SALT	Speech Application Language Tags
TTA	Telecommunication Technology Association
URC	Ubiquitous Robot Companion
VAD	Voice Activated Dialing
VXML	Voice eXtensible Markup Language
W3C	World Wide Web Consortium

참고문헌

[1] B.H. Juang, L.R. Rabiner, and J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Trans. ASSP*, Vol.35, July 1987, pp.947-953.

[2] H. Hermansky, "Perceptual Linear Predictive(PLP) Analysis of Speech," *Journal of the Acoustical Society of America*, Vol.87, No.4, 1990, pp.1738-1752.

[3] H. Ney, "The Use of a One-stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Tran. ASSP*, Vol.32, 1984, pp.263-271.

[4] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *Bell System Technical Journal*, Vol.62, No.4, Apr. 1983.

[5] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, Feb. 1989, pp.257-286.

[6] "신성장동력산업의 기반, 음성인터페이스 기술," ETRI CEO information 12호, 2004. 9.

[7] 김종환, "IT 기반의 유비봇(UbiBot)," 전자신문, 2003. 5. 13.

[8] 김민경, 김현, "URC 인프라 시스템 기술," ITFIND, Vol.1196, 2005.

[9] 최지훈, 장병태, "텔레매틱스 기술 및 서비스 동향," ITFIND, Vol.1157, 2004.

[10] 박준, "대화체 음성 자동통역 기술," 전자공학회지, 제30권 7호, 2003, pp.29-38.

[11] 박준, 이영직, 양재우, "대화체 음성언어번역 시스템 개발," 제 15회 음성통신 및 신호처리 워크샵, 한국음향학회, 1998, pp.281-286.

[12] Oh-Wook Kwon and Jun Park, "Korean Large Vocabulary Continuous Speech Recognition with Morpheme-based Recognition Units," *Speech Communication*, Vol.39, Issues 3-4, 2003, pp.287-300.

[13] 권오욱, 박준, 황규웅, "의사형태소 단위 대어휘 연속 음성인식기 개발," 제 15회 음성통신 및 신호처리 워크샵, 한국음향학회, 1998, pp.320-323.

[14] M. Mayfield et al., "Concept Based Speech Translation," *ICASSP*, Vol.1, 1995, pp.97-100.

[15] 최운천, "다국어 대화체 음성언어번역 시스템을 위한 IF (Interchange Format)와 IF 태깅," 제 15회 음성통신 및 신호처리 워크샵, 한국음향학회, 1998, pp.409-412.

[16] IWSLT 2004, CSTAR, Kyoto, 2004.