특 집

음성정보처리기술의 현황과 전망

김 형 순

부산대학교 전자전기정보컴퓨터공학부

I. 서 론

음성은 사람의 가장 자연스러운 의사전달 수단이며, 이 음성을 사람과 기계 사이의 인터페이스 수단으로 사용하고자 하는 것은 인류의 오랜 꿈 중의 하나다. 음성정보처리기술 또는 음성정보기술이란 음성신호에 포함된 정보를 자동 분석하는 기술과 이러한 정보로부터 음성을 생성하는 기술을 통칭한다. 여기에는 사람의 말을 알아듣는 음성인식, 문장을 말로 바꾸어주는 음성합성, 그리고 말하는 사람이 누구인지 식별하는 화자인식 기술 등이 포함된다.

스탠리 큐브릭 감독의 1968년작 '2001년 스페이스 오디세이'에 등장하는 HAL 컴퓨터를 비롯하여, 많은 공상과학영화에서 사람과 자유롭게 대화하는 컴퓨터 또는 로봇 등이 등장한다. 그러나 지난 수십 년에 걸친 연구에도 불구하고 사람과 유창하게 대화를 나눌 수 있는 시스템은 개발되지 못하고 있으며, 앞으로도 가까운 장래에 이루어질 전망은 보이지 않는다.

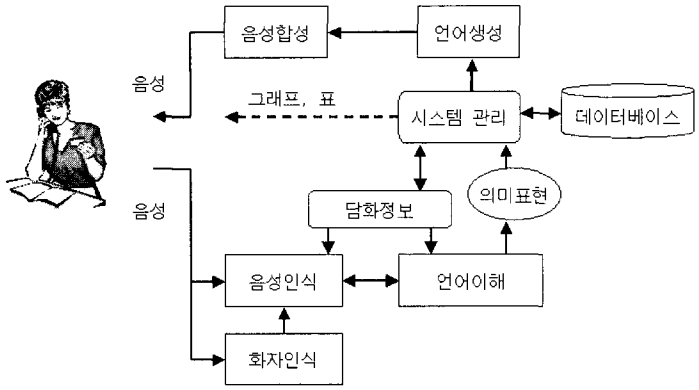
그러나, 기대수준을 조금만 낮추어 본다면, 현재의 기술 수준으로도 제한된 영역에서 사람의 말을 알아듣고 말로 정보를 제공하는 실용적인 제품들이 등장하여 이미 사용되고 있는 단계에 이르렀으며, 수년 안에 이러한 제품의 보급률이 크게 높아질 전망이다. 미국에서는 전화사용과 관련한 모든 문제에 대해 자유롭게 말한 내용을 인식하여 처리하는 "How may I help you?"와 같은 서비스가 진행되고 있고[1], 국내에서도 음성인식에 의한 증권거래 및 음성 다이얼링 서비스, 그리고 음성에 의한 문서작성(dictation) 프로그램 등이 상품화되어 사용되고 있다. 그리고 전화를 통한 기상예보 등에서는 이미 음성합성이 그 자리를 차지한지 오래다.

통신기술의 발전과 더불어 휴대단말기를 통해 언제 어디서나 정보를 주고받는 시대가 도래하면서, 키보드 사용이 용이하지 않은 상황에서의 인터페이스 수단으로서 음성의 필요성이 증대되고 있다. 운전 중 자동차 내부에서의 각종 정보처리를 가능케 하는 텔레매틱스 서비스에서도 음성 인터페이스는 필수 요소이다. 폭증하는 멀티미디어 데이터의 자동검색을 위해서 음성에 의한 자동색인을 통한 오디오 마이닝(audio mining)이라는 용어가 등장하고 있으며[2][3], 언어교육의 수단으로서의 음성기술의 유용성도 부각되고 있다.

본 고에서는 음성정보처리기술의 전반적인 현황을 살펴보고 향후의 발전방향을 전망해 본다. 서론에 이어 II장에서 음성정보처리기술의 개요를 살펴보고, III장 및 IV장에서 음성정보처리기술의 핵심요소인 음성인식 및 합성기술 및 그 현황을 기술한다. 그 다음으로 V장에서 음성정보처리기술의 향후 전망을 언급한 후 VI장에서 결론을 맺는다.

II. 음성정보처리기술 개요

음성신호와 관련된 정보처리를 다루는 음성정보처리기술은 이미 언급한 바와 같이 음성인식, 음성합성 및 화자인식 등과 같은 세부 요소기술



〈그림 1〉 Voice User Interface(VUI)의 구성도

들로 이루어진다. 음성신호를 압축/재생하는 음성부호화 기술도 음성신호처리 분야의 대표적인 기술이지만, 음성신호에 포함된 정보의 의미를 직접 다루지는 않으므로 좁음 의미에서의 음성정보처리기술에서는 제외하기로 한다.

음성정보처리기술은 음성을 통한 인간과 기계 사이의 의사소통을 가능케 하며, 이를 그래픽 사용자 인터페이스(graphic user interface, GUI)에 대응하는 표현으로서 VUI(voice user interface)라고 부르기도 한다. 〈그림 1〉에 VUI의 일반적인 구성도가 나타나 있다[4].

그림에서도 확인할 수 있듯이 음성에 의한 사람과 기계 사이의 대화를 위해서는 직접 음성신호를 다루는 음성인식, 음성합성 및 화자인식 기술 이외에도 음성언어를 다루기 위한 다양한 언어처리 기술이 포함된다. 일례로 적절한 언어처리 기술은 시스템 주도형의 단순한 대화를 시스템과 사용자 사이의 상호주도형 대화로 바꾸어 줌으로써 편의성을 증진시킬 수 있다. 시스템 주도형 대화와 상호주도형 대화의 예를 들면 다음과 같다.

〈시스템 주도형 대화의 예〉
시스템 : ○○항공입니다. 목적지 도시명을 말씀해 주십시오.
사용자 : 부산
시스템 : 출발지 도시명을 말씀해 주십시오.
사용자 : 서울
시스템 : 서울에서 부산까지 여행하실 날짜를 말씀해 주십시오.
사용자 : 7월 17일

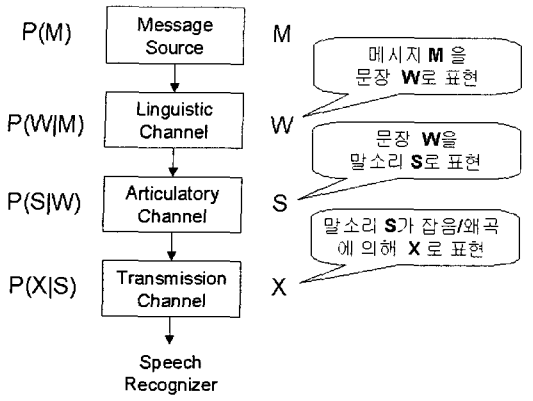
〈상호주도형 대화의 예〉
시스템 : ○○항공입니다. 어디로 여행하시겠습니까?
사용자 : 서울에서 부산으로 가려는데요.
시스템 : 서울에서 부산까지 언제 가시겠습니까?
사용자 : 7월 17일 11시에 일반석 좌석이 있나요?

실제로 음성정보처리기술은 전자공학과 컴퓨터공학 분야에 속하는 디지털 신호처리, 패턴인식, 자연언어처리 기술 이외에도 음성학, 언어학, 심리학, 생리학, 통계학 등 다양한 학문분야의 전문지식이 요구되는 대표적인 학제간 연구주제이다. 음성정보처리기술의 전반적인 내용을 짧은 지면에 모두 다루는 것은 무리이고, 본 고에서는 음성정보처리의 핵심 요소기술인 음성인식 및 합성기술에 대해서만 기술하기로 한다.

Ⅲ. 음성인식 기술

1. 개 요

음성발생 과정을 통신이론의 개념으로 단순화시켜 표현하면 〈그림 2〉와 같다[7]. 먼저 사용자



〈그림 2〉 음성발생 과정

는 P(M)라는 사전확률분포에 의해 자신이 의도하는 메시지 M을 결정한다. 동일한 메시지라도 대화문맥 등에 따라 다양한 문장으로 표현될 수 있으므로 문장W의 생성 과정은 조건확률분포 P(W|M)로 모델링 된다. 문장 W는 사람의 조음기관(성대, 혀, 입술 등)의 동작에 의해 말소리 S로 표현된다. 같은 사람이라도 말할 때마다 음성에 차이가 있고, 사람에 따른 차이는 더욱 큰데 이러한 발성과정에서의 변화요인을 P(S|W)로 나타낸다. 음성신호가 배경잡음 및 통신채널 왜곡 등에 의해 최종적으로 음성인식 시스템의 입력형태인 X로 표현되며, P(X|A)는 이러한 외부적 변화요인을 표현한다.

음성인식(speech recognition)은 잡음과 채널특성에 의해 왜곡된 음성신호 X를 문장 W로 복원하는 과정이며, 발화자가 의도했던 메시지 M으로 최종복원하는 과정을 음성인식과 구분하여 음성이해(speech understanding)라고 부른다. 문장 W가 음성신호 X로 표현되는 과정에서 〈그림 2〉에서 보는 것처럼 다양한 변화요인들이 존재하기 때문에, 그 역과정인 음성인식은 어려운 문제가 된다.

이러한 문제 때문에 음성인식의 궁극적인 목표, 즉, 잡음이 있는 실제적인 환경에서 임의의 화자가 어휘에 제한없이 자연스럽게 발음한 연속음성을 실시간에 인식 및 이해하는 수준을 만족시키는 시스템은 아직 개발되지 못하고 있으며, 현재의 음성인식 시스템들은 몇 가지 제약조건 하에서 운용됨을 전제로 하고 있다. 음성인식의 난이도를 결정하는 대표적인 요인들은 다음과 같다.

(1) 발성 형태 : 각 단어와 단어를 또박또박 띄어 발음하는 고립단어의 인식에 비해, 자연스럽게 연결시켜 발음한 연속음성의 인식이 어려우며, 말하는 속도가 빨라질수록 더 어려워진다.

(2) 화자 독립성 : 자신의 목소리를 미리 등록시킨 특정인의 음성을 인식하는 화자종속 인식에 비해, 훈련에 참여하지 않은 임의의 다수 화자의 음성을 인식하는 화자독립 인식이 어렵다.

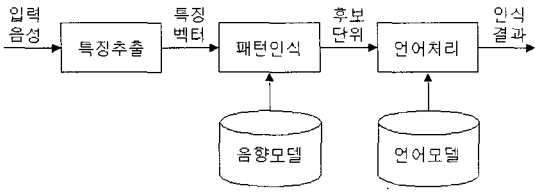
(3) 인식대상 어휘규모 : 일반적으로 인식하고자 하는 어휘 수가 많아질수록 난이도가 높아진다. 물론 동일한 어휘 수라고 할지라도 단어들의 음성학적 유사성에 따라 난이도는 다르다.('불 켜'와 '불 꺼'의 구별이 '예'와 '아니오'의 구별보다 어렵다.)

(4) 문법구조 및 주제 : 자유발화 대화음성은 입력음성에 문법적 제약을 많이 두는 경우에 비해 인식하기 매우 어렵다. 또한 다수의 주제를 대상으로 하는 경우가 호텔예약 등 특정한 단일 주제로 내용을 한정할 때보다 인식하기 어렵다.

(5) 음성통신 환경 : 전화망과 같이 미지의 채널 왜곡이 있거나 자동차 소음 등 배경잡음이 부가되어 신호대잡음비(SNR)가 낮을 경우 인식 성능이 현저하게 떨어진다.

2. 음성인식 시스템의 기본 구성

일반적인 음성인식 시스템의 구성도가 〈그림 3〉에 나타나 있으며, 그 동작을 개략적으로 설명하면 다음과 같다. 마이크를 통해 입력된 음성이 디지털 신호로 변환되어 음성인식 시스템으로 들어오면, 먼저 특징추출 단계에서 일정 시간(보통 매 10 ms)마다 음성학적 특징을 잘 표현해 줄 수 있는 특징벡터를 추출한다. 추출된 음성 특징벡



〈그림 3〉 음성인식 시스템의 기본 구성

터들은 패턴인식 과정으로 넘겨져서 미리 저장된 단어 또는 음소들의 음향모델과 비교하게 되며, 그 결과는 일련의 후보단어 또는 후보음소들의 형태로 언어처리 과정에 전달된다. 언어처리 과정에서는 후보단어 또는 후보음소들의 정보를 토대로 하여, 인식대상어휘 및 문법구조, 그리고 특정 주제에의 부합 여부를 판단하여 최종 인식된 문장을 출력시키게 된다. 현재 주류를 이루고 있는 통계모델 기반의 음성인식 방식에서는 실제로 패턴인식 과정과 언어처리 과정이 하나의 최적화 과정으로 통합된다. 이하에 각각의 요소기술에 대해 간단히 언급한다.

1) 특징추출

특징추출 과정의 목적은 음성학적 정보는 잘 표현하면서 그 이외의 음향적 변화(배경잡음, 채널왜곡, 화자 차이, 발음 태도 등)에는 둔감한 특징벡터들을 추정하는 것이다. 그러나 이러한 이상적인 특징추출 방법은 아직까지 알려져 있지 않으며, 일반적으로 매 10 ms 정도의 시간 간격으로 20-30 ms의 단구간 음성을 분석하여, 로그 스펙트럼의 포락선 정보를 표현해 주는 켑스트럼(cepstrum) 영역의 특징 벡터들이 주로 사용된다.

2) 음향모델과 패턴인식

동일한 문장 W를 발성하더라도 앞서 〈그림 2〉에서 언급한 여러 요인들로 인해 최종적인 음성신호 X(여기서는 X를 음성의 특징벡터 열이라고 하자)는 매우 다양한 형태로 나타나며, 이를 확률모델 P(X|W)로 표현한 것을 음향모델이라 한다. 문장 W는 보통 여러 개의 단어들로 구성되고, 각 단어는 다시 음소(/ㄱ/, /ㄴ/, /ㅏ/, /ㅓ/ 등 말소리의 기본 단위)들로 이루어지므로, 각 단어 또는 음소별 음향모델이 주어진다면 이들로부터 P(X|W)를 구할 수 있다. 단어 또는 음소의 음향모델을 구성하는 효과적인 방법으로 hidden Markov model(HMM)이라는 통계 모델이 주로 사용된다[8].

3) 언어모델과 언어처리

음성인식에서 언어모델은 일차적으로 인식대상 어휘 및 문법 구조에의 제약을 통해 인식성능을 향상시키는 것을 목표로 한다. 문법 구조에 많은 제약을 두어도 무방한 응용 분야에는 유한상태 문법(finite state grammar)이 사용되며, 보다 자연언어에 가까운 문장을 인식하기 위해서는 N-gram 형태의 통계적 언어모델을 통해 문장(또는 단어열) W의 확률 P(W)를 구한다[9]. N-gram 모델이란 일련의 단어들로 구성된 문장이 생성될 확률을 N-1개의 단어로 구성된 단어열 다음에 특정 단어가 나타날 확률들로써 표현하는 것으로서, 보통 두 개 또는 세 개의 연속된 단어 사이의 연관관계를 확률적으로 나타내는 bigram 및 trigram 언어모델이 사용된다.

3. 음성인식 기술 현황

음성인식 성능은 인식대상 어휘 수, 언어모델의 복잡도, 배경잡음 및 채널왜곡 정도, 발화 자유도(낭독음성과 대화음성의 차이), 음성신호의 표본주파수, 그리고 훈련에 사용된 음성 데이터베이스 분량 등 여러 요인에 의해 영향을 받기 때문에, 현재의 음성인식 기술 수준이 어느 정도인지를 간단한 수치로 표현하기는 어렵다. 〈표 1〉에 영어를 대상으로 한 여러 가지 다른 인식 영역에서의 사람과 컴퓨터의 인식성능이 나타나 있다. 표의 결과로부터 컴퓨터가 사람에 비해 적어도 5배 이상 많은 오류를 범하며, 잡음이 섞이는 등 환경이 열악해지거나 자유발화(spontaneous speech)와 같이 문법적 제약이 적은 음성일 경우 성능 차이가 더 커짐을 알 수 있다.

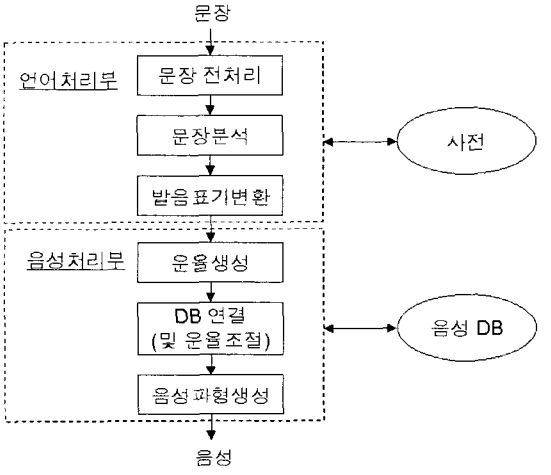
〈표 1〉의 마지막 항목(조용한 환경에서의 trigram 문장발화)은 흥미로운 예외로서, 잡음

〈표 1〉 사람과 컴퓨터의 영어 음성인식 성능 비교 (단어오인식률)[10]
(WSJ는 5000 단어 규모의 Wall Street Journal 기사를 낭독한 연속음성의 인식임)

인식대상 영역	어휘 수	사람	컴퓨터
연결숫자	10	0.009%	0.72%
알파벳 문자	26	1%	5%
자유발화 전화음성	2,000	3.8%	36.7%
조용한 환경 낭독음성 (WSJ)	5,000	0.9%	4.5%
10 dB SNR 환경 낭독음성 (WSJ)	5,000	1.1%	8.6%
조용한 환경에서의 trigram 문장발화	20,000	7.6%	4.4%

의 영향이 없는 상황에서 미리 정의된 문법에 생성된 문장을 인식하는 경우에는 컴퓨터가 사람보다 우수한 성능을 나타냈다. 이는 사람이 컴퓨터에 비해 인식성능이 뛰어난 것이 문법이나 의미 등 언어정보를 보다 잘 사용하기 때문임을 보여주며, 향후 음성인식의 성능향상을 위해 언어정보의 보다 효율적인 활용이 중요함을 시사한다.

정리하면, 현재의 음성인식 기술이 아직 사람의 인식 능력에 비해 떨어지는 것이 사실이며, 특히 잡음 환경에서의 급격한 성능저하와 자유발화 음성에 대한 적절한 언어모델의 부재가 앞으로 해결해야 될 최대의 과제라고 판단된다.



〈그림 4〉 음성합성 시스템의 기본 구성

Ⅳ. 음성합성 기술

1. 음성합성 시스템의 기본 구성

음성합성 기술, 또는 Text-To-Speech(TTS) 기술은 임의의 문장을 음성으로 바꾸어 출력시키는 기술이다. 기존의 ARS 서비스 등에서 금융 및 증권정보를 제공할 때 사용하던 제한된 합성방식은 단어 또는 구 등을 녹음했다가 이들을 조합하는 방식이라서 합성하고자 하는 어휘 및 문장에 제한이 있었다. 이에 반해 TTS 기술은 입력된 문장을 언어학적 규칙을 통해 분석하고, 이로부터 추출한 음소 정보 및 운율 정보를 이용하여 합성음을 생성해 내기 때문에, 사용 어휘와 문장 형태에 대한 제약이 전혀 없는 무제한 합성이 가능하다. 예전에는 TTS 방식이 제한된 합성방식에 비해 음질 면에서 크게 뒤떨어졌으나, 최근 기술발전을 통해 매우 우수한 합성음질을 얻을 수 있게 되었다.

현재 대부분의 음성합성 시스템이 연결합성(concatenative synthesis) 방식, 즉, 음소와 같이 단어보다 작은 음성 단위들의 연결에 의해 음성을 만들어내는 방식을 사용하고 있다. 연결합성 방식에 의한 TTS 시스템의 일반적인 구성도는 〈그림 4〉와 같다. 그림에서 보는 같이 TTS 시스템은 언어처리부와 음성처리부의 두 부분으로 크게 나눌 수 있으며, 이들 각각에 대해 간단하게 설명하면 다음과 같다.

1) 언어처리부

언어처리부는 입력 문장을 분석하여 음성생성에 필요한 각종 정보들을 추출한다. 먼저 전처리 과정을 통해 입력 문장에 포함된 숫자, 기호, 약어 등을 일반 문자로 대체시킨다. (예 : '12.12 km'를 '십이 점 일이 킬로미터'로 바꿈) 그 다음으로, 문장분석 과정에서 문장구조를 분석하며 이 결과가 끊어 읽기와 억양 등 운율처리를 위한 정보로 사용된다. 마지막으로 발음표기변환 과정은 정서표기 문장을 사람이 발음하는 형태의 표기로 바꾸어 준다.

2) 음성처리부

음성처리부에서는 언어처리부에서 분석된 정보를 토대로 음성신호를 만들어낸다. 먼저 운율생성 과정에서는 문장의 분석 결과로부터 사람이 글을 읽을 때와 같은 고저, 장단, 강약 및 끊어 읽기 정보를 생성한다. 그 다음 DB연결 및 운율조절 과정에서는 발음표기에 따라 미리 저장된 음성의 기본단위들을 가져와서 이들을 연결하고, 앞서 생성된 운율정보에 의해 음높이와 길이 등을 조절한다. 최종적으로 음성파형생성 과정에서 만들어진 음성신호를 출력시킨다.

종래의 음성합성 방식에서는 메모리 용량의 제한 등으로 각각의 음성단위별로 하나씩 저장한 후 신호처리 기술을 이용하여 음높이와 길이의 조절 및 연결과정에서의 스펙트럼 보간 등을 수행하는 방식을 택했는데, 이 과정에서 불가피하게 신호왜곡이 생겨서 상당한 음질저하가 발생하였다. 이 문제를 해결하기 위해 최근 도입된 코퍼스 기반의 합성방식에서는 각각의 음성단위별로 좌우 음소문맥, 음높이, 길이 등을 고려한 다수의 음성 샘플(또는 세그먼트)들을 모두 저장해 두었다가, 합성하고자 하는 음소열과 운율정보에 가장 적합한 샘플들을 탐색하여 연결합성한다. 이 경우 추가적인 운율조절의 필요성이 최소화되며, 많은 경우 신호처리 과정을 아예 생략함으로써 이에 따른 음의 왜곡을 방지하게 된다. 그 결과 합성단위를 저장하기 위한 메모리 용량이 수백 MB에서 수 GB 수준으로 증가하고 이에 따른 탐색 소요시간도 늘어나기는 하지만, 이전에 비해 훨씬 명료하고 자연스러운 합성음의 생성이 가능해졌다.

2. 음성합성 기술 현황

음성합성 시스템의 성능은 합성음의 명료도와 자연성으로 표현될 수 있다. 지난 수십 년간의 음성합성 연구는 1차적으로 합성음의 명료성을 높이는데 주력해 왔으며, 그 결과 명료성 면에서는 사용자들이 어느 정도 큰 불만 없이 사용할 수 있는 수준에 이르렀다고 판단된다. 부분적으로는 아직도 명료성이 떨어지는 부분들이 있지만, 이는 DB tuning에 의해 해결 가능한 수준이다.

합성음의 자연성도 코퍼스 기반의 합성방식이 등장함에 따라 상당한 발전이 이루어졌다. 그 결과 낭독음성 형태에서는, 물론 앞으로도 개선의 여지가 많이 남아 있는 상태이지만, 상용 서비스가 가능한 수준에 도달했다. 이에 따라 일기예보, 교통정보 등 각종 뉴스와 이메일 등을 전화를 통해 음성으로 들려주는 서비스들이 이미 많이 보급되고 있다.

그러나 대화음성의 합성은 아직 자연성 면에서 많은 개선이 필요하며, 합성음에 감정정보를 포함시키거나 특정한 화자의 음색으로 변환하는 기술 등은 아직 기초연구 단계에 머물러 있다.

V. 음성정보처리기술의 향후 전망

음성정보처리기술의 미래를 예측하는 것은 쉬운 일이 아니다. 1970년대에 미국에서 음성인식 연구가 본격 추진되면서 2000년 이전에 이 기술이 완성되어 보편화 될 것이라는 낙관론과, 음성인식이 너무나 어려운 과제이므로 여기에 투자하는 것은 낭비에 불과하다는 비관론이 동시에 대두되었으나, 현재 시점에서 볼 때 이들 모두 잘못된 견해임이 판명되었다.

음성정보처리기술에 많은 진전이 이루어진 현 상황에서도 주제에 제한없이 사람과 자유롭게 대

화하는 컴퓨터가 언제 출현할 수 있을지는 미지수이다. 그러나, 제한된 영역의 음성인식 성능은 최근 들어 매년 10% 정도의 오류감소율(error reduction rate)을 보이는 수준으로 계속 향상되는 추세에 있으며, 이 추세가 계속된다면 구체적인 영역에 따라 차이가 있지만 향후 10-40년 사이에 사람에 버금가는 수준에 도달할 것이라는 전망이 나오고 있다[11].

이미 항공예약 등 특정 주제를 대상으로 한 전화망 대화 인터페이스가 선진국에서 이미 서비스되고 있으며, United Airline 항공사의 사례 중에서 음성 인터페이스 도입을 통한 실제적인 비용절감 효과들이 보고되고 있다. 국내에서도 이미 전화망을 통한 음성 서비스들이 실행되고 있으며, 최근의 경기침체의 여파를 고려하더라도 수년 이내에 보다 본격화되리라고 전망된다.

자동차 운전 중에 정보통신 서비스를 제공받기 위한 텔레매틱스 분야, 멀티미디어 정보검색, 외국어 교육 분야에도 음성기술의 도입이 곧 보편화 될 것으로 보이며, 머지않아 제한된 영역에서의 휴대용 통역기도 등장할 수 있을 것으로 전망된다.

또한 음성으로 웹 콘텐츠나 서비스 이용을 가능하게 하는 확장성 생성 언어인 VXML(Voice eXtensible Markup Language)의 등장은 음성응용 서비스의 개발을 용이하게 해 줌으로써, 음성 포털 서비스를 비롯한 음성정보처리기술의 상품화의 확대에 크게 기여할 것으로 보인다[12].

음성정보처리기술의 향후 연구 방향에 대해서도 여러 가지가 있겠지만, 본 고에서는 잡음환경처리, 대화음성 처리, 그리고 멀티모달 인터페이스의 3가지에 대해서만 언급하기로 한다.

1. 잡음에 강인한 음성인식 방식

현재 대부분의 음성인식 시스템이 조용한 환경에서는 우수한 성능을 나타내다가도 잡음과 채널왜곡이 있는 상황에서는 성능이 급격히 저하되는 문제점을 보인다. 성능저하를 야기하는 음향학적 변화요인들로는 배경잡음과 입력 마이크의 특성, 전화회선 등에서 비롯되는 채널왜곡 등을 들 수 있으며, 배경잡음이 화자의 발성에 영향을 미치는 Lombard 효과도 하나의 중요한 요인이다. 이 문제는 사용자와 마이크 사이의 거리가 떨어져 있는 distance-talking 환경에서 더 심각해진다.

지금까지 잡음환경에 강인한 음성인식 방식에 대한 많은 연구가 이루어졌고, 그 결과 많은 성능향상이 이루어졌다[13]. 그러나 음성인식의 유용성이 부각되는 많은 실제적인 환경(예를 들면, 자동차 운전 중)에서 무리없이 사용될 수준과는 아직도 많은 격차가 있으며, 이 격차를 줄이기 위한 시도가 앞으로 핵심 연구주제가 될 전망이다.

2. 대화음성 인터페이스

주어진 문장을 읽는 형태인 낭독음성과 실제 대화음성은 많은 차이가 있으며, 대화음성에는 "에, 저"와 같은 간투사, 망설임, 반복, 생략 등 다양한 표현변화가 등장하기 때문에 인식하기 훨씬 어렵다. 대화음성 인터페이스를 위해서는 구문분석, 의미분석 및 담화관리를 비롯한 다양한 언어처리 기술들이 사용되어야 하며, 대화음성처리에 적합한 언어모델의 개발에 많은 노력이 필요하다. 항공예약이나 증권거래와 같은 특정 영역에 대해서 집중적인 전문가적 분석을 토대로 사용 가능한 수준의 대화모델을 구성한다 하더라도, 이로부터 어떻게 새로운 대상 영역으로 쉽게 전환할 것인가 하는 점도 현실적인 관건이다. 또한 사람의 경우 대화가 정상적으로 이루어지고 있는지 여부를 쉽게 인지할 수 있으나, 컴퓨터가 이를 판단하여 문제가 있다면 대화전략을 바꾸는 단계까지 가려면 무엇이 필요한지도 향후 연구의 대상이다.

음성합성의 경우에도 낭독음성의 합성은 충분히 상품화 수준에 이르렀지만, 감정정보까지 표현할 수 있는 대화음성의 합성기술은 초보적인 단계에 머물러 있으므로, 이 부분에 대한 연구가 향후 중심 주제가 될 것이다.

3. 멀티모달 인터페이스

정보접근 수단이 기존의 데스크탑 PC나 전화기에서 PDA, tablet PC, 그 외의 차세대 이동통신 단말기로의 전환이 진행됨에 따라, 사람과 컴퓨터 사이의 상호작용(human computer interaction, HCI)에도 새로운 접근방식이 요청된다. 이에 따라 키보드도 없는 화면도 협소한 단말기 환경에서 음성, 펜, 제스쳐 등의 통합 활용에 의한 멀티모달 인터페이스에 대한 연구가 이슈가 되고 있으며, 목표는 사용자의 개인적인 선호도까지 고려하여 주어진 상황에 가장 적합한 인터페이스 방식을 제공하는 것이다. 음성정보처리기술 영역에서도 멀티모달 환경을 고려한 음성대화 인터페이스의 설계가 필요하다.

음성과 영상의 통합에 의해 음성인식의 한계를 극복하려는 시도도 본격화될 전망이다. 배경잡음이 심한 환경에서 음성신호는 왜곡되지만 영상정보(입술 움직임 정보)는 소리 형태의 잡음에는 영향을 받지 않기 때문에, 음성/영상 통합처리를 통해 성능향상을 기대할 수 있다[14]. 음성과 영상에 의한 바이모달 인식방식은 음성인식 분야 이외에도 화자인식과 얼굴인식 등을 통합한 생체인식의 수단으로도 각광을 받을 것이다.

또한 음성합성의 경우에도 단순히 음성만으로 정보를 제공하기보다 화면에 사람의 얼굴 모습을 가진 에이전트가 등장하여 입술모양과 음성이 동기화되어 자연스러운 음성/영상 정보를 제공하는 visual TTS가 시도되고 있다.

VI. 결 론

음성정보처리기술의 현 상황을 한 마디로 요약한다면 "아직, 그러나 이미"라고 말할 수 있겠다. 음성기술이 앞으로도 많은 연구개발이 필요한, 아직 해결되지 않은 문제라는 사실은 분명하다. 이 기술의 궁극적인 목표인 사람과 아무런 제한없이 자유롭게 의사소통하는 음성 인터페이스는 아직 확보되지 못했을 뿐만 아니라 앞으로도 단기간에 달성될 가능성은 없다. 그러나, 음성정보처리기술은 현재의 기술수준으로도 이미 많은 실제적인 응용분야에 사용되기 시작하고 있으며, 실제로 그 유용성을 인정받고 있다.

앞으로의 음성정보처리기술 분야의 연구는 현재 기술의 한계를 극복하기 위한 기반기술 연구와 더불어, 현재의 기술 수준으로 해결가능한 특정 응용분야에 최적화된 솔루션을 개발하기 위한 시도가 병행될 것이다. 사람은 별다른 불편함 없이 말을 하고 말을 알아듣기 때문에, 음성정보처리기술에 대한 사용자의 기대수준이 높다는 점이 이 기술의 상품화에 부담으로 작용하고 있는 것도 사실이다. 현 단계에서의 상품화를 위해서는 기술상의 제약요인을 정확히 파악하고 응용 단계에서 이 문제를 최소화시킬 방법론의 확보가 필수적이다.

21세기에 들어서서 MIT나 다보스포럼 등에서 미래사회를 주도할 핵심기술로 음성정보처리기술을 선정한 것을 굳이 인용하지 않더라도, 이 기술이 앞으로 반드시 확보되어야 할 기반기술이라는 점에는 이론의 여지가 없다. 꿈의 인터페이스인 음성기술의 발전이 정보통신, 자동차, 가전 등 여러 산업분야에 미칠 커다란 파급효과를 기대해 본다.

참 고 문 헌

〔1〕 A. L. Gorin *et al.*, "Automated natural spoken dialog," IEEE Computer, pp. 51-56, Apr. 2002.

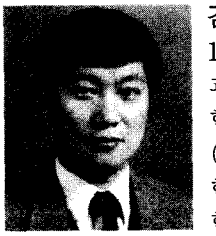
〔2〕 N. Leavitt, "Let's hear it for audio mining," IEEE Computer, pp. 23-25, Oct. 2002.

〔3〕 P. J. Moreno *et al.*, "From multimedia retrieval to knowledge management," IEEE Computer, pp. 58-66, Apr. 2002.

〔4〕 R. Cole *et al.*, Ed., *Survey of the State of the Art in Human Language Technology*, Cambridge University Press, 1997.

〔5〕 김회린, "음성인식 기술 개요 및 향후 과제," 대한전자공학회지, 제28권 제5호, pp. 580-586, 2001년 5월.
〔6〕 이정철, "음성합성 기술 개요 및 향후 과제," 대한전자공학회지, 제29권 제12호, pp. 1491-1497, 2002년 12월.
〔7〕 B. H. Juang and S. Furui, "Automatic recognition and understanding of spoken language-a first step toward natural human-machine communication," Proc. IEEE, vol. 88, no. 8, pp. 1142-1165, Aug. 2000.
〔8〕 L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.
〔9〕 R. Rosenfeld, "Two decades of statistical language modeling : where do we go from here?" Proc. IEEE, vol. 88, no. 8, pp. 1270-1278, Aug. 2000.
〔10〕 X. Huang, A. Acero and H. Hon, *Spoken language Processing*, Prentice Hall PTR, 2001.
〔11〕 X. Huang, "Making speech mainstream," *http://www.microsoft.com/speech/techinfo/articles*.
〔12〕 P. J. Danielsen, "The promise of a voice-enabled web," IEEE Computer, pp. 104-106, Aug. 2000.
〔13〕 J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition : Fundamentals and Applications*, Kluwer Academic Publishers, 1995.
〔14〕 T. Chen, "Audiovisual speech processing," IEEE Signal Processing Magazine, pp. 9-21, Jan. 2001.
〔15〕 R. V. Cox et al., "Speech and language processing for next-millennium communication services," Proc. IEEE, vol. 88, no. 8, pp. 1314-1337, Aug. 2000.

저 자 소 개



김 형 순

1983년 2월 서울대학교 전자공학과 (공학사), 1984년 2월 한국과학기술원 전기 및 전자공학과 (박사과정조기진학), 1989년 2월 한국과학기술원 전기 및 전자공학과 (공학박사), 1987년 1월~1992년 6월 : 디지콤정보통신연구소 (선임연구원), 1992년 7월~현재 : 부산대학교 전자전기정보컴퓨터공학부 (부교수), 〈주관심 분야 : 음성인식 및 합성, 음성신호처리〉