□특집□

Voice Interface 및 인식

김 회 린† 이 영 직††

◈ 목 차 ◈

1. 서 론

20년 전만 해도 음성을 이용하여 컴퓨터를 조작하거나, 사람이 말을 하면 이를 그대로 컴퓨터가 받아 적는 장면은 공상 과학 영화에서나 볼 수 있었다. 그동안 많은 연구자들의 노력에 힘입어 현재 음성 명령어의 경우 그 인식률이 95%를 넘는다. 또한 미국의 경우 음성을 받아들여 이를 문장으로 인식하는 소프트웨어들이 속속 출시되고 있다. 이러한 현상으로 미루어 짐작해 보면, 공상 과학 영화와 동일한 장면은 아직 멀었지만, 음성을 이용한 제한된 성능의 기술들이 우리에게 생각보다 가까이 왔음을 실감할 수 있다.

음성 처리 기술 하면 바로 음성 인식이나 음성 합성이 생각난다. 그만큼 이 단어들이 우리 귀 주변을 맴돌았다는 얘기다. 그러나 이 한 마디의 단어를 차분히 생각해 보면 의외로 많은 서로 다른 기술 분야가 그 안에 포함되어 있음을 알 수 있다. 이러한 기술 분야를 표 1에 정리하였다.

<표 1> 음성 처리 기술의 분류

대분류	기술 분야
음성 인식	고립 단어 인식
	연속 음성 인식
	대화체 음성 인식
음성 합성	편집 합성
	무제한 TTS*
	음색 변환
화자 인식	화자 검증
	화자 식별
음성 인터페이스	사용자 모델링
	편의성 평가
기타	대화 관리

*TTS: Text-to-Speech, 문장-음성 변환기

† 정회원 : 한국전자통신연구원 음성신호처리연구실 선임연구원
†† 정회원 : 한국전자통신연구원 음성신호처리연구실장, 책임연구원

이 표는 기술의 대상에 근거하여 분류한 것으로, 한 기술 분야라 하더라도 그 안에 서로 다른 여러 기술이 존재한다. 무제한 TTS의 경우를 예를 들면, 합성 대상 어휘가 무제한이라 할지라도 일기예보를 음성으로 합성하는 기술과 동화를 읽어주는 기술은 서로 달라서, 하나의 기술로 다른 것을 대치할 수 없다. 음성 처리 분야에 있어 같은 기술이라는 말을 하기 어려운 이유는 다음과 같다.

하나의 음성 처리 기술의 개발에는 눈에 보이지 않는 여러 세부 기술들이 필요하다. 먼저 음성 데이타베이스를 개발 목적에 맞도록 설계하고 수집해야 한다. 이어서 이 데이타베이스를 기반으로 개발 목적에 맞는 알고리즘을 이용하여 인식, 혹은 합성기를 개발한다. 음성 데이타베이스가 다르면 인식 대상 어휘도 달라지며, 같은 알고리즘을 활용한다 하더라도 데이타베이스 확보에 필요한 시간 및 학습에 필요한 시간이 소요되므로 같은 기술이라 보기 어렵다. 또 처리 대상 어휘가 같다 하더라도 주변 잡음 상황이 달라지면 데이타베이스 및 처리 알고리즘이 달라진다. 무잡음 상황에서 인식이 잘 되던 기술이 약간의 소음만 더해지더라도 성능이 현격히 떨어지는 경우가 이에 해당한다. 즉 위 표의 같은 항목에 분류됨에도 불구하고 서로 다른 기술이 다수 존재한다.

음성 처리 기술에 대한 다양한 응용의 욕구는 일반 사람들의 상상 속에 많이 퍼져 있다. 그러나 본 논문에서는, 아직 음성 기술이 완전히 해결된 분야가 아니므로, 음성 처리 기술이 일반적으로 사용될 응용 분야보다는 현재의 기술 수준을 조금만 더 발전시킴으로써 응용이 가능한 분야를 주로 고려하였다. 이를 크게 나누어 보면 음성 명령, 받아쓰기, 통신망 응용 등으로 구분 지을 수 있다. 각 응용 분야 및 그 분야에서 기술적으로 핵심적으로 처리해야 할 대상을 표 2에 나열하였다.

대부분의 음성 처리 기술 응용은 앞에서 나열한 음성 처리 기술 중 두 개 이상의 기술이 공동으로 활용되어 만들어 진다. 예를 들어 자동 응답 시스템을 보면 전화선 잡음 환경의 소규모 어휘 고립 단어 인식 기술, 정보 전달용 음성 편집 합성 기술 및 대화 관리 기술이 통합되어 만들어 진다. 여기에서 매우 중요한 점은 하나의 기술이 그 가치를 인정받기 위해서는 그 기술이 실제로 응용되는 환경에서 제 성능을 발휘해야 한다는 점이다.

다음 장에서는 표 1에 나열된 기술의 분류에 따라 ETRI의 음성 처리 기술 연구 동향을 설명한다.

<표 2> 음성 처리 기술의 응용 분야 및 핵심 처리 대상

응용 분야	핵심 처리 대상
음성 명령	잡음처리 - 사무용 컴퓨터 - 자동차 - 제품 조립 라인 - 게임
	인식 대상 어휘 - 음성 다이얼 - 웹 브라우저 - 컴퓨터 제어
받아 쓰기	대어휘 인식
	언어 처리
통신망	전화 음성 인식
	대화 관리
	화자 검증
음성언어 번역	대화체 음성 인식
	인식 결과의 번역

2. ETRI의 연구 동향

본 장에서는 ETRI의 음성 처리 연구 동향을 대화체 음성 언어 번역, 음성 명령어 인식, 음성 합성, 음성 인터페이스 분야로 나누어 기술한다.

2.1 대화체 음성언어 번역[1,5,6]

대화체 음성 인식 연구는 다국간(한국어/일어/영어) 자국어를 이용한 멀티미디어 원격 회의가 가능한 대화체 음성언어 번역 시스템 개발을 위해 수행하고 있다. 이 연구의 결과를 바탕으로 1999년 9월 미국, 일본과 한영, 한일 대화체 음성언어 번역 시스템을 통합하여 여행 계획 분야의 대화를 대상으로 국제간 공동 시연을 할 예정이다.

대화체 음성은 낭독체 음성에 비해 많은 차이를 가지고 있다. 우선 발성 중 비문법적인 표현이 많이 있다. 또 대화 상황이므로 무의미어의 발성이 많으며, 다양한 억양이 추가된다. 현재 대화체 음성 인식 수준은 그 인식 대상 어휘가 5,500 단어이며, 단어 인식률이 79%이다. 현재 인식률의 향상을 위해 화자 적응, 메타 정보 이용, 발음 사전 보완 등을 시도하고 있다.

인식된 문장을 일본어 및 영어로 번역하는 데에는 두 가지의 문제가 있다. 우선 대화체 문장이 비문법적인 면이 많으므로 이의 이해가 어렵다는 점이고, 번역 대상이 인식 결과이므로 평균적으로 10 단어 중 2개가 오류라는 점이다. 첫번째 문제를 대처하기 위해 문장 전체를 대상으로 번역을 시도하는 대신, 부분 부분으로 나누어 번역을 시도하는 부분 파서를 채용하였다. 현재 인식 결과 번역률은 75%이다. 현재 인식과 언어 처리의 인터페이스 분야에 단어 격자를 사용할 경우 언어 지식을 이용하여 정인식 단어를 골라 내는 연구를 수행하는 중이며, 다국어 번역에 대비하여 중간언어 방식의 번역을 연구하고 있다.

2.2 음성 명령어 인식[2]

본 연구는 음성 처리 기술의 실용화를 목적으로 수행하는 연구이다. 이 연구에서는 개인용 컴퓨터 윈도우즈 95/NT 상에서 음성 명령을 인식하여 컴퓨터를 구동하는 기술을 개발하고 있다.

이 연구에서 중요한 점은 실제 상황에서 적용이 가능한 기술을 개발하고자 하는 것이다. 고립 단어의 경우 실험실에서 미리 잘 정리된 음성 데이타베이스로 고립 단어를 인식하면 99%가 넘는 성능을 보인다. 그런데 실제 컴퓨터가 사용되는 상황은 사무실 환경으로, 각종 소음이 상존한다. 이러한 소음만 섞여도 인식 성능이 현저하게 떨어져 쓸모가 없게 된다. 이 연구에서는 15 내지 25 dB의 신호대잡음비 상황에서의 소규모 어휘 인식을 목표로 한다. 본 연구에서는 잡음 상황에서 동작되는 비음성 검출기를 개발하였으며, 현재는 미등록어 검출기를 개발하고 있다.

한국 사람이 컴퓨터를 사용할 때, 한글도 사용하지만 영어를 사용하는 경우도 매우 많다. 이러한 경우의 영어 발성은 미국인의 영어 발성과 현저히 다르다. 이를 처리하기 위해 본 연구에서는 한국형 영어 발음 사전을 개발하였다.

음성 명령 기술의 또 다른 중요한 관점은 인식 속도이다. 명령 후 1초만 지나가도 매우 지루한 느낌을 받는 것이 컴퓨터 사용자들의 특징이다. 본 연구에서는 인식 시간 단축을 위해 인식의 전 과정을 파이프라인으로 처리하였다.

이러한 결과들을 활용하여 음성 웹 브라우저를 개발하였다[3]. 이것은 웹 브라우저의 기본 명령어를 음성으로 입력할 뿐만 아니라, 웹 페이지의 한글 링크를 음성으로 검색하는 기능을 가진다. 영문 혹은 그레픽 링크는 한글 숫자로 대치하여 화면에 표시하고, 사용자가 해당되는 한글 숫자를 발성하면 해당 화면을 표시하게 된다.

여기에서 특기할 만한 사항은 매 웹 페이지마다 인식 대상 어휘가 바뀐다는 점이다. 대부분의 인식기는 그 인식 대상 어휘가 고정되어 있는 것이 보통이다. 당 연구실에서는 인식 대상 어휘가 바뀌어도 실시간으로 인식 어휘 사전을 바꾸어

인식하는 가변어휘 인식기를 개발하여 여기에 사용하였다. 현재 가변어휘 인식기는 사무실 잡음 환경에서 94%의 인식률을 보인다.

2.3 음성 합성

당 연구실에서는 1990년대 초부터 무제한 문장-음성 변환기를 개발해 왔다. 초기의 TTS에는 그 합성 단위로 반음절 단위를 사용하였으나, 최근에 이를 음절 단위 및 triphone 단위로 바꾸었다.

합성음의 자연성 향상에는 운율의 조절이 필수적이다. 이를 위해 음성의 피치 궤적 및 각 음소/음절별 지속 시간을 조절하여, 자연성을 향상시켰다. 이러한 운율 조절을 위해 한국어 ToBI(tone and break indices)를 사용하였다[4].

발성의 상황에 따라 운율이 매우 달라진다. 예를 들어 대화 상황에서의 운율은 일기예보의 운율과 매우 다르다. 대화체 음성언어 번역은 대화체 문장을 번역하는 것이므로 대화체의 운율을 가지는 것이 바람직하다. 당 연구실에서는 이를 위해 대화체 운율을 가진 한국어 음성 합성기를 개발하였다.

하나의 음색을 가진 합성기의 개발 시간을 단축하고자, 현재는 trainable TTS방식에 주력하고 있다. 이 방식은 합성기 개발을 위해 한 사람의 음성을 녹음한 뒤, 음소 분할, 피치 마킹, 운율 추출 등의 과정을 자동으로 처리하여 짧은 시간 안에 합성기를 완성하는 방법이다.

2.4 음성 인터페이스 분야

음성 인식 기술이 존재한다 하더라도 사용자가 이 기술을 즐거이 사용할 것인가 하는 문제는 또 다른 문제이다. 특히 음성 인식 기술은 그 완성도가 낮기 때문에 이 문제가 더욱 심각하다. 따라서 사용자가 어떠한 때에 음성 입력 방법을 필요로 하게 될 것인가, 그리고 어떠한 방법으로 음성 입력 수단을 제공하는 것이 가장 효율적으로 사용될 수 있는가 결정을 해야 한다. 지금까지 몇 개의 음성처리 기술이 선보였지만, 대부분 이 사용자 인터페이스 측면이 충분히 고려되지 않은 경우가 많이 있었다. 당 연구실에서는 음성 웹 브라우저를 대상으로 사용자가 선호할 음성 명령어를 선택하였으며, 사용자 측면에서의 편의성을 측정하였다.

이러한 인터페이스의 사용자 편이성을 높이기 위해 음성 합성 시에 소리 뿐만 아니라 소리에 따라 움직이는 입술 모양을 보여 줄 수 있다. 당 연구실에서는 음성 합성기의 출력에 입술 모양을 지정하는 출력을 추가하여, 입술 모양 동기를 가능케 하였다.

3. 국외 음성 기술의 실용화 동향

본 장에서는 국외의 음성 처리 기술 실용화 동향을 음성 명령어, 받아쓰기, 화자 검증의 분야로 나누어 살펴 본다.

3.1 음성 명령어[8]

미국 Kurzweil사(L&H사에 통합됨)에서는 Voice-Commands라는 음성 명령 소프트웨어를 출시하였다. 이 제품은 MicroSoft Word 문서를 음성 명령으로 편집하는 기능을 가진 소프트웨어로, 연속 음성 인식으로 여러 단계(6~7 단계)의 마우스 조작을 한 음성 명령으로 제어할 수도 있다. 이 시스템의 성능은 화자 독립 90%, 화자 적응 97%이다. 이 소프트웨어는 20 MB의 하드디스크 및 16 MB의 메모리를 가진 Pentium 90 MHz 시스템에서 Win95/NT 4.0 상에서 동작한다. 이 시스템은 잡음에 강한 특정 마이크를 사용한다.

3.2 받아쓰기[7, 8]

음성 받아쓰기 기능은 사용자 자신이 타이프를 잘 하지 못한다고 생각하고 있기 때문에, 예로부

터 요구가 많던 기능이었다. 1992년 말에는 변호사 혹은 의료 분야에 약간씩 쓰이던 이 기능은 1996년 만 카피 정도가 팔렸으며, 1997년에는 수십만 카피가 팔릴 것으로 예상된다. 소비자들은 좀더 인식이 잘되고, 자신의 타이프 속도보다 빠른 음성 받아쓰기 소프트웨어를 바라고 있다.

현재 가장 우수한 성능의 받아쓰기 제품은 Dragon사의 NaturallySpeaking 이다. 보통의 미국 사람은 오타 수정 시간을 포함하여 분당 50 개의 영어 단어를 타이프 하는데 반하여 이 소프트웨어는 114 단어를 처리한다. 이 시스템은 화자 적응 기능이 있어, 사용 전에 훈련 과정을 거치면 인식 성능이 향상되어, 95% 이상의 단어를 인식한다. 인식 대상 어휘는 3만 단어이고, 20만 단어 크기의 back-up 사전을 가지고 있다. 또 텍스트 혹은 알파벳을 음성으로 지정하거나 텍스트를 입력하는 방식으로 새 어휘를 추가할 수 있다. 이 소프트웨어는 60 MB의 하드디스크 및 32 혹은 48 MB의 메모리를 가진 Pentium 133 MHz 시스템에서 Win95/NT 4.0 상에서 동작한다. 음성 받아쓰기를 위해서는 특정한 마이크를 사용해야 한다.

IBM사에서는 고립단어 인식 기능을 가진 SimplySpeaking을 출시하였다. 이 제품의 성능은 22,000 어휘에서 단어 인식률이 97%이며, 분당 100 단어를 입력할 수 있다. 이 시스템을 사용하려면 약 30분에 걸쳐 50 문장을 훈련시켜, 자신의 목소리에 적응시켜야 한다. 이 소프트웨어는 30 MB의 하드디스크 및 16 MB의 메모리를 가진 Pentium 100 MHz 시스템에서 Win95 상에서 동작한다. 이 경우 역시 특정한 마이크를 사용해야 한다.

IBM사에서는 또 연속음성 인식 기능을 가진 ViaVoice를 출시하였다. 이 제품의 성능은 22,000 어휘(최대 4.2만 단어)에서 단어 인식률이 97%이며, 분당 100 단어를 입력할 수 있다. 이 시스템 역시 화자 적응을 위해 30분간 훈련을 시켜야 한다. 이 소프트웨어는 100 MB의 하드디스크 및 32 혹은 48 MB의 메모리를 가진 Pentium 166 MHz 시스템에서 Win95/NT 4.0 상에서 동작한다. 이 경우 역시 특정한 마이크를 사용해야 한다.

3.3 화자 검증

화자 인식 시스템에 있어서 미국의 경우에는 1000명의 다른 사람이 시험하여서 1명 이하의 사람을 잘못 승인하고, 본인이 100번 발성하여서 1번 이하의 잘못된 거절을 화자 인식 시스템의 최소 규격으로 삼고 있다. 최근의 국제 학술지 등의 보고에 의하면 여러 기관들이 위에서 언급한 성능을 이미 넘어섰다는 발표를 하고 있다. 화자 검증기는 그 용도에 따라서 승인 에러를 대폭 줄여야 할 경우와 거절 에러를 대폭 줄여야 할 경우가 있으며 시스템의 성능 평가 기준으로서는 일반적으로 승인 및 거절 에러가 같은 지점에서의 에러율(EER: Equal Error Rate)을 사용한다. 현재 발표된 성능은 AT&T의 경우 EER이 0.4%이며, Kurzweil의 경우 0.22%이다. 그럼에도 불구하고, 잡음이 섞이는 경우 오류가 급격히 심해지므로, 아직 실용화는 안되고 있다.

4. 음성 처리 기술의 전망

4.1 받아쓰기

앞 장에서 살펴본 바와 같이, 음성 처리 기술은 이미 우리 옆에 가까이 와 있다. 우선 음성 받아쓰기 기능은 사용자의 요구가 크기 때문에 활용이 점차 커질 것으로 보인다. 물론 사용자의 요구에 따라 인터페이스가 더욱 간편하고 활용도가 높도록 바뀌게 되고, 그 인식률이 점차 증가하게 될 것이다.

4.2 통신망 응용

본 논문에서 실용화의 예가 언급되지는 않았지만, 통신망을 이용한 자동 응답 서비스 분야는 그 대상 시장이 매우 크므로, 앞으로 시장이 크게 성장할 것으로 보인다. 예를 들어 증권정보 안내, 열차나 항공권, 호텔 등의 예약 시스템 및 원격 예금 처리에 음성 처리 기술이 활용될 것이며, 이 경우 화자 검증 기술 역시 많이 활용될 것이다.

4.3 음성 게임

음성을 이용한 게임 구동은 사용자의 편의성 향상을 목표로 하기 보다는 음성이라는 입출력 수단이 제공되어 게임에 더욱 몰두하게 되는 방향으로 발전될 것이다. 주의할 점은 게임을 사용하는 환경이 일반적인 사무실 환경보다 훨씬 시끄러우며, 특히 주변에 구경하는 사람의 발성도 많은 상황이라는 점이다. 또 게임은 실시간 동작이 요구된다. 이러한 상황에서의 인식 기술도 매우 효용성이 높은 일이라 본다.

4.4 멀티미디어 정보의 저장 및 검색

앞으로 대부분의 정보는 멀티미디어 정보가 될 것이며, 그 정보의 양 또한 매우 방대해 질 것이다. 이러한 멀티미디어 정보를 검색하는 데에 텍스트가 주로 사용될 것이다. 그러나 뉴스 등 음성이 주 역할을 하고 있는 정보를 검색하기 위해서는 우선 이 정보를 텍스트 형태로 바꾸어 요약해 저장해야 한다. 이러한 의미에서 음성 처리 기술이 많이 활용될 것으로 보인다.

5. 결 론

본 논문은 음성 처리 기술의 동향을 기술하였다. 우선 음성 처리 기술을 분류하고, 그 응용 예를 살펴 보았으며, 이어서 ETRI의 음성 처리 연구 동향을 살펴 보았다. 또, 국외의 음성 처리 기술 실용화 동향을 기술하고, 마지막으로 앞으로의 음성 처리 기술 응용 전망을 기술하였다.

본 논문에서 강조하고 싶은 점은, 음성 처리 기술의 경우 그 사용 환경 및 목적에 따라 비슷해 보이는 기술이라 할지라도 실제로는 서로 많이 다른 기술이라는 점이다. 많은 사람들이 하나의 음성 처리 기술을 확보하고, 이를 이용하면 많은 분야의 문제를 해결할 수 있다고 생각하는데, 이는 사실이 아니다. 음성 처리 기술이 실험실 수준에서 좋은 성능을 보이고 있지만, 실제 기술의 개발은 지금부터라는 생각으로 실제 필요한 기술을 개발해 나가야 할 것이다.

6. 감사의 글

이 연구는 정보통신부 출연의 다중매체 환경하에서의 대화체 음성 번역 통신 기술 개발 및 HCI를 위한 음성 입출력 처리 기술 개발 과제로 수행되었습니다.

참고문헌

[1] Y. Lee, J. Park, and J.-W. Yang, Spontaneous speech translation under multimedia environment, Proc. of ICSP 97, Seoul, pp. 683-686, Aug. 1997.

[2] J.-W. Yang, Y. Lee, and H.-R. Kim, Speech input/output processing technology for human-computer interface, Proc. of ICSP 97, Seoul, pp. 417-420, Aug. 1997.

[3] H.-S. Lee and H.-R. Kim, Internet surfing with the Korean spoken language, Proc. of ICSP 97, Seoul, pp. 687-690, Aug. 1997.

[4] J.-C. Lee and K.-M. Sung, Improvement of the

synthesized speech intonation with stylization and neural network learning, Electronic Letters, vol. 33, no. 19, pp.1600-1601, Sept. 1997

[5] J.-W. Yang, et al., Multimedia spoken language translation, IEICE Transactions on Information & Systems, vol. E79-D, no. 6, pp.653-658, June 1996.

[6] J.-W. Yang and Y. Lee, K-E/K-J spoken language translation system prototype, Proc. Of the C-STAR II 96 ATR International Workshop on Speech Translation, Sept. 1996.

[7] N. Simon, "Voice recognition market evolution from discrete to continuous," Proc. of AVIOS 97, pp.1-6, Oct. 1997.

[8] K. Bouwers, "Kurzweil VoiceCommands and the market for continuous dictation," Proc. of AVIOS97, pp.7-12, Oct. 1997.

김 회 린

1984년 한양대학교 전자공학과 (학사)
1987년 KAIST 전기 및 전자공학과 (석사)
1992년 KAIST 전기 및 전자공학과 (박사)
1994년-1995년 일본 ATR 음성번역통신연구소 방문연구원
1987년-현재 한국전자통신연구원 음성신호처리연구실 선임연구원
관심분야 : 음성인식, 음성신호처리, 음성합성, 신경회로망, 음성언어번역

이 영 직

1979년 서울대학교 전자공학과 (학사)
1981년 한국과학기술원 산업전자공학과 (석사)
1989년 Polytechnic University 전기 및 전산과 (박사)
1981년-1984년 삼성전자주식회사 컴퓨터개발실
1989년-현재 한국전자통신연구원 음성언어연구실장, 책임연구원
관심분야 : 음성인식, 음성합성, 자동통역, 신경회로망, 패턴인식