特輯

음성처리기술의 응용 현황 및 전망

韓敏洙, 鄭裕鉉, 李恒燮
韓國電子通信研究所 音聲應用研究室

Ⅰ. 머릿말

인간이 기계, 대부분의 경우 콤퓨터와 가장 자연스럽고 편리하게 의사소통을 할 수 있는 방법은 사람의 말, 즉 음성을 사용하는 것이다. 이를 위하여는 기계가 사람의 말을 알아 들을 수 있어야 하며 또 사람처럼 유창하게 말할 수 있어야 한다. 이런 기술들을 개발하여 인간과 기계가 음성어를 이용하여 대화할 수 있도록 하자는 것이 음성공학자들의 꿈이며 이 꿈을 실현하기 위하여 선진 각국에서는 지난 수십년간 막대한 인적, 물적 투자를 아끼지 않고 연구를 계속하여 왔다.[1-3] 그러나 이런 많은 노력에도 불구하고 우리 주변에서 음성인식 기술이나 음성합성 기술, 또는 자연어 처리 기술이 상용 시스템으로 개발되어 실생활에 직접 이용되는 사례를 찾기는, 말하는 전자 사전이나 장난감, 또는 몇몇 단어로 작동되는 제품 등 아주 간단하고 초보적인 것들을 제외하고는, 쉬운 일이 아니다. 즉, 그동안 음성처리 기술을 이용한 상용 시스템의 개발이 성공적으로 수행되었다는 보고들이 국내외에서 많았음에도 불구하고 실제로 큰 규모로 서비스되고 있는 것들은 찾아 보기가 힘들 정도인 것이다. 단적인 예로 AT&T Bell Lab에서 1976년도에 전화를 이용한 항공권 예약 데모 시스템이 실험실에서 상당한 가능성을 보여주는 수준까지 개발되었다고 발표하였으나 15년이 지난 지금까지도 이 시스템의 상용화가 성공적으로 이루어졌다는 보고는 없었던 것이다.[4]

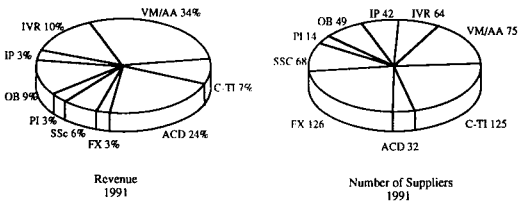
그러면 이렇게 음성처리 기술에 대하여 막대한 인적, 물적, 시간적 투자가 지속되어 왔음에도 불구하고 몇몇 제한된 분야에서 초보적인 단계의 상품들 만이 개발된 이유는 무엇일까? 음성공학자들이라면 이 질문에 첫째는 사람의 음성이 가지는 다양성을 제대로 표현할 수 있는 알고리즘이 아직 개발되지 않았기 때문이며, 둘째는 현재의 콤퓨터 기술로는 인간의 음성을 인식하기 위하여 필요한 막대한 정보를 처리할 수 없기 때문이다라고 대답할 것이다. 이와 다른 시각은, 첫째는 대부분의 사람들이 너무 이론적인, 즉 기술적인 연구에만 치중하고 실제 상용화가 가능한 기술들은 이미 진부한 것, 또는 더 연구해 보아야 논문거리도 안되는 것이라고 치부해 버리고 그에 대한 개선 연구를 등한히 했기 때문이며, 둘째는 시스템의 성능 향상에는 많은 노력을 기울여 왔으나 실제 시스템을 사용할 때의 사용자의 편리성이나 만족도가 간과된 채 시스템이 설계되어 왔으므로 이런 시스템을 상용화하였을 때 고객들로부터 외면당할 수 밖에 없었다는 주장이다.

이러한 모든 것을 고려해 볼 때, 본 고에서는 현재의 음성처리 기술의 응용 현황이 어떠한 지를 미국을 중심으로 먼저 살펴본 다음, 음성처리 기술을 이용한 상용 시스템을 개발하기 위하여, 무엇이 예상되는 문제이고, 어떤 것들을 개발 목표로 하여야 하며, 개발을 성공리에 완수하기 위해서는 어떤 점들을 고려해야 하는지를 나름대로 도출해 보고자 한다.

Ⅱ. 응용현황

국내에서 음성처리 기술이 상용화되어 이용되고 있는 것은 ARS (Audio Response System, 즉 700 서비스) 또는 몇몇 간단한 PC용 음성합성 보드 정도로서 시장규모나 응용분야가 매우 초보적인 단계이

다. 그러나 미국의 경우, 음성처리기술과 관련된 시장규모는 1990년도에 약 63억 달러였으며 매년 15 - 20% 정도의 신장세를 계속 유지할 것으로 전망되고 있다. [5] 이는 음성처리 기술에 관련된 국내 시장도 조만간 급속히 성장할 것이며 요구되는 상용화 기술도 보다 다양해질 것이라는 것을 간접적으로 시사해 준다고 볼 수 있는 것이다. 이에 본 장에서는 음성처리 기술의 응용현황을 음성합성 기술, 음성인식 기술 및 그 외의 기술의 세 분야로 나누어 미국을 중심으로 살펴보고자 한다. 참고로 그림 1에 미국 내의 음성처리 기술 응용 시장의 현황을 도표로 보였다. [6]



IVR: Interactive Voice Response
VMAA: Voice mail and auto attendant
SSC: Subsystem components
ACD: Automatic Call Distributor
IP: Information Providing
OB: Outbound
PI: Passive interception
C-TI: Computer telephony integration

그림 1. 미국 내의 음성처리 기술 응용 시장 현황

1. 음성합성 기술 응용현황

음성합성 기술이 상용화되어 이용되고 있는 분야는 크게 두가지로 나누어 진다. 하나는 analysis-to-synthesis technique, 즉 분석-합성 기술을 이용한 vocoder 분야로서 입력 음성신호를 압축하여 parameter로 저장하거나 전송하는 분야이다. 다른 하나는 text-to-speech conversion technique, 즉 문자로 입력된 정보를 여러가지 합성 규칙과 이미 저장되어 있는 부호화된 음성 분석 정보를 이용하여 대응되는 음성 출력으로 바꿔주는 문자-음성 변환 기술을 이용하는 분야이다.

전자는 주로 전화망이나 위성통신망에서의 음성신호의 압축 및 보안기술로 이용되고 있으며 대부분 LPC (Linear Predictive Coding)에 기초한 기술들이 이용되고 있다. 응용 사례로는 시장 점유율이 가장 높은 전자 사서함 서비스나 은행 잔고 조회 서비스 (때로는 문자-음성 변환 기술 또는 녹음-편집 방식 기술이 이용되기도 함), 또는 정부기관 내에서의 비화(秘話)통신 서비스 등이 있다.

후자는 인간과 기계 사이에 음성으로 의사 소통을 하기 위하여 꼭 필요한 기술로서, 즉 사람의 말을 기계가 이해한 다음 음성으로 사람에게 반응하기 위한 기술로서, 이용되고 있는 기법은 LPC 합성, 포만트 합성, 음편 합성 기술 등이다. 이 기술은 주로 전자전화 번호부 서비스나 음성 신문, 또는 주가 조회 등 audiotex 서비스나 다른 videotex 서비스와 결합된 형태로 많이 이용되고 있다.

2. 음성인식 기술 응용현황

음성합성 기술이 상대적으로 많은 분야에서 이용되고 있는 것과는 달리 선진국에서도 음성인식 기술이 상용화된 예는 그리 많지 않다. 이는 기술적으로는 사용자가 불편함을 느끼지 않을 정도로 음성인식기술을 상용화하기 위해서는 시스템의 최종 음성 인식률이 98% (rejection rate는 3% 이내)는 되어야 한다는 제약 때문이며 심리적으로는, 사람이 상대방의 말을 알아 듣는 것도 실제로는 96 - 97 % 밖에 안 됨에도 불구하고 대부분의 사용자가 콤퓨터로 실현된 음성인식기술을 이용하는 경우 잘못된 선입견 때문에 100%의 완벽한 인식률을 기대하는 경우가 많기 때문이다. [7] 현재의 음성인식 기술 수준으로 이 정도의 인식률을 화자독립으로 얻을 수 있는 분야는 선진국에서 조차 상당히 조용한 주변환경 하에서의 수십개의 고립단어 인식 기술 정도인 것이다. 따라서 음성인식 기술의 상용화는 아직 초보적인 단계에 머물고 있으며 그나마도 전화음성에 대해선 거의 전무한 형편이다.

현재 음성인식 기술을 이용한 시스템 중 상용화되어 있는 대표적인 것들은 IBM의 Tangora system과 Dragon Dictate System이나 이들도 고객을 만족시키기 위한 98% 이상의 인식률이라는 위의 조건은 충족시키지 못하고 있는 실정이다. 이 중 Tangora System은 약 20,000 개의 영어 단어에 대하여 95% 정도의 인식률이 주장되고 있으며 [8], Dragon Dictate System은 약 25,000 개의 영어 단어에 대하여 다후보단어(multi-candidates)를 이용하면 약 93 - 95%의 인식률을 갖는다고 주장되고 있다. (불행히도 주장은, 특히 음성 인식 시스템의 최종 인식

률에 대한 주장은 주장만으로 끝나는 경우가 많다.) 앞의 두 시스템이 모두 화자독립 시스템을 지향한 반면, 금년초 NYNEX에서는 화자종속으로 voice dialing 서비스를 전화망에서 제공하기 시작하였으며 그 정확도는 미리 등록된 약 40개의 고립단어에 대하여 90% 내외인 것으로 알려져 있다.

또한 AT&T에서는 1991년 1월 CONVERSANT Voice Information System (VIS)을 발표하였다.[9] 이 시스템은 전화망 상에서 발성단어를 인식할 수 있으며 word spotting 기능도 가지고 있다. 이 시스템은 우편-판매 시스템 (mail order application), 신용카드 이용 시스템, 제 3자 과금 통화, 은행 잔고 조회 및 계좌 이체 등의 다양한 용도로 사용될 수 있을 것으로 기대되며 최종 성능은 key word 만을 정확히 발성한 경우 10 - 20 단어를 대상 어휘로 불특정화자에 대하여 99%라고 주장하고 있으며 과도한 음성 (extraneous speech)이 포함된 경우 같은 task에 대하여 약 95%의 인식률이 주장되고 있다.

음성인식이나 합성 기술 외에도 화자인식 (speaker-identification), 또는 화자인증 (speaker-verification) 기술도 제 3자 과금 통화나 신용카드의 조회, 또는 보안장치에 이용하기 위하여 많이 연구되고 있으나 아직은 상용화 단계까지는 이르지 못한 실정이다.[10] 또한 가장 널리 보급되어 있고 누구나 사용법을 알고 있는 전화를 이용한 여러가지 응용분야, 즉 전화 판매 시스템의 자동화, 전자 전화 번호부, 전화를 이용한 banking transaction 등을 실현하기 위하여 전화음성인식에 대한 연구도 활발히 이루어지고 있으나 전화음성이 갖는 회선 잡음이나 제한된 주파수 대역 등의 여러가지 제약 때문에 만족할 만한 연구결과는 보고되지 않고 있으며, 따라서 완전 자동화된 상용 시스템의 출현은 아직 시기상조이다.[11] (많은 음성공학자들이 전화망에서 "예", "아니오" 만을 100% 인식할 수 있어도 그 응용분야가 무척 다양하다고 주장한다는 사실이 전화음성인식이 얼마나 어려운가를 단적으로 말해 준다고 할 수 있다.)

한가지 재미있는 사실은 Bellcore에서 전화음성에 대하여 word-spotting 기술을 도입하여 전화번호 안내 시스템의 일부를 자동화하는데 상당히 성공을 거두고 있으며 계획대로라면 2 ~ 3 년 안에 교환수를 도와주는 시스템으로 실용화되어 1년에 약 2 ~ 3 억 달러의 인건비를 절약할 수 있을 것이라고 전망하고 있다는 점이다. 물론 이 시스템도 완벽한 것은 아니며 시스템이 자동으로 처리 못하는 입력은 alarm 신호를 보냄으로서 교환수가 대신 처리하게 하는 것이지만 현재의 음성처리 기술 수준으로도 응용분야만 잘 선택하면 성공할 수 있다는 가능성을 시사해 준다는 데에 그 의미가 크다고 할 것이다.

Ⅲ. 연구방향

지금까지 음성처리기술이 어떻게, 또 어느 정도 응용되고 있는지 미국을 중심으로 간략히 살펴보았다. 이를 바탕으로 본 장에서는 음성처리기술들을 실용화하기 위하여 고려해야 할 사항들을 먼저 생각해 보고자 한다.

첫째, 어떤 응용 분야를 target으로 시스템을 개발할 것인지를 신중히 결정하여야 한다. 만일 고객이 별로 흥미를 느끼지 못하는 응용분야를 선택한다면 시스템에 실현된 음성인식이나 합성 기술이 거의 완벽하게 작동하는 경우라 할 지라도 그 사업은 실패할 수 밖에 없을 것이다.

둘째, 사용자들이 시스템을 손쉽게 사용할 수 있도록 설계되어야 한다. 단적으로 사용하기가 불편한 시스템은 차라리 동작하지 않는 것이 사업이라는 측면에서는 더 바람직하다고 할 수 있다.

셋째, 음성인식시스템의 성패는 사용하는 어휘에 달려있다 해도 과언이 아니다. 즉 사용자에게 친근감을 주는 동시에 높은 인식률을 보장하기 위하여 음향적 특성이 가능한 한 서로 다른 어휘들을 신중하게 선택하여 시스템을 설계하여야 한다.

넷째, 현재 상용화되어 있는 거의 모든 음성합성 시스템은 완벽한 명료성이나 자연성을 갖춘 음성을 합성하지 못한다. 따라서 합성음성 출력시 사용자가 알아듣지 못한 부분을 쉽게 반복하여 들을 수 있는 방법을 제공할 수 있도록 설계되어야 한다.

이러한 고려사항들을 반영하여 향후 어떤 방향으로 음성처리기술들을 연구 개선하는 것이 바람직한 가를 생각해 보고자 한다.

우선 음성합성 기술을 살펴보면, 선진국의 경우, 상용화된 문자-음성 변환 기술들 중 성능이 좋은 것들은 명료도, 즉 어떤 내용인지 알아 듣는데는 거의 문제가 없으나, 자연성이라는 측면에서는 아직 해결해야 할 문제가 많다. 국내의 경우는 아직 명료도에도 약간의

문제가 있으나 조만간 해결될 것으로 믿어지며, 따라서 연구의 중점은 역시 합성음의 자연성 개선에 두어져야 될 것이다. 이를 위하여 합성 알고리즘의 개선에 대한 연구와 병행하여 한국어 자체의 고유성질, 즉 음의 장단이나 고저, 조음결합 효과, 문장의 발성패턴 등에 대한 보다 기초적인 연구가 선행되어 많은 통계자료가 축적되어야 할 것이다.

또한 점점 더 다양화되고 고급화되어가는 사용자들의 욕구를 충족시키고 보다 양질의 음성 서비스를 제공하기 위하여 꼭 개발해야 할 기술이 음색변환 기술일 것이다. 이 기술의 필요성은 음성합성 기술을 등장인물이 여러 명인 콤퓨터 만화나 콤퓨터 동화에 이용한다고 가정해 보면 쉽게 알 수 있을 것이다. 또 할머니를 대상으로 하는 음성 서비스에 젊은 여자의 카랑카랑한 목소리를 합성하여 서비스를 하게 되면 그 사업은 반드시 실패할 수 밖에 없다는 음성 심리학적 측면을 고려해 보면 자명하다 할 것이다.

이제 음성인식 기술에 대해 생각해 보도록 하자. 현재의 음성인식 기술 수준과 상용화를 위하여 필요한 인식률이 98%라는 사실을 감안할 때, 2 - 3년 내에 실용화가 가능한 분야는 고립단어 인식기술을 이용하는 초보적인 단계로 생각된다. 즉, 숫자음과 수십개 정도의 고립단어를 포함하는 어휘로 가능한 서비스 분야를 생각할 수 있다. (물론 현재의 기술 수준으로 연결단어를 인식하는 상용 시스템의 개발이 불가능한 것은 아니지만 성공할 확률이 상대적으로 상당히 낮은 것 또한 사실이다.) 이런 간단한 기술을 활용할 수 있는 서비스로 먼저 현재 국내에서 음성처리 기술이 상용화되어 활발히 이용되고 있는 700 서비스를 생각할 수 있다. 현재 제공되고 있는 모든 700서비스는 #, *, 0-9의 push button에 의해서 정보검색이 가능하다. 이 경우 숫자음과 적은 어휘의 음성인식 기능의 첨가만으로도 push-button의 기능을 대체할 수 있으므로 시스템의 부가가치를 비약적으로 높일 수 있을 뿐만 아니라 보다 다양한 서비스를 사용자에게 제공할 수 있다. 이외에도 은행의 잔고조회, 특정기관의 내선 전화번호 안내 시스템, 장애자용 (car phone용) voice dialing 시스템, 음성작동 가전기기 등을 들 수 있을 것이다. 따라서 현재의 음성인식 기술 즉, HMM, 신경회로망 및 DTW 중 상용화 음성기술로는 고립단어, 소어휘 인식에 효과적인 DTW - based 기술이 바람직하다고 생각된다. 물론 현재 실용화 가능성이 높은 것들이 고립단어 인식 기술을 이용하는 것이기 때문에 연속음성 인식 연구가 필요없다는 의미는 결코 아니다. (향후 연속음성인식에는 HMM, 신경회로망, 혹은 Hybrid type이 필수적일 것이다.)

실용화를 목적으로 하는 경우 key word spotting 기술에 대한 연구도 병행되어야 할 것이다. 실제로 사용자들이 발성을 할 때 인식 대상 어휘 뿐만 아니라 그 단어 전후로 불필요한 단어나 소리를 내는 경우가 많으므로 이 중에서 인식 대상이 되는 어휘만을 찾아내는 key word spotting 기술의 개발은 매우 중요하다. 실제로 AT&T의 CONVERSANT Voice Information System (VIS)에서는 key word spotting 기술을 주변 잡음 (background noise)과 과도한 음성을 제거하는데 사용하였다.

전화망에서의 음성인식 기술의 상용화를 위하여 간과해선 안될 문제점 중 하나는 잡음제거 기술에 대한 것이다. 즉 회선잡음과 주변잡음, 나아가서 차량전화의 경우, 자동차로 인하여 발생되는 소음의 제거 기술이 확보되어야만 인식 시스템이 현장에서 제 성능을 발휘할 수 있을 것이다. 이를 위하여 적응신호처리 기술을 이용한 잡음제거 기술 , 또한 필요하다면 여러 개의 마이크를 이용한 잡음제거 기술 등이 꾸준히 연구되어야 할 것이다.

음성처리기술의 응용에 대하여 연구한다면, 이 외에도 신용화, 정보화되어가는 사회 추세에 대응하기 위하여 꼭 확보되어야 할 기술들이 화자인식 및 화자인증 기술이다. 이 기술의 효용성은 음성 key 뿐 만 아니라, 이 기술과 몇 십 단어 수준의 고립단어 인식 기술을 결합하면 비밀번호 없이도 전화만으로 banking transaction이 가능하다든지, 신용카드를 분실하더라도 고객 관리를 담당하는 주전산기에 이미 등록되어 있는 본인의 음성 입력이 아니면 사용이 불가능하기 때문에 다른 사람이 사용할 수 없으므로 그다지 초조해 할 필요가 없어진다는 사실을 생각해보면 쉽게 이해할 것이다.

Ⅳ. 맺음말

본 고에서는 선진국에서, 특히 미국에서는 음성인식, 음성합성 등의 음성처리 기술들이 어느 분야에서 상용화되었으며 어느 정도 성공을 거두고 있는지, 또

현재의 상용화 시스템들은 앞으로 어떻게 개선될 예정인지를 먼저 기술하였다. 다음에 이것을 바탕으로 우리는 음성처리 기술을 상용화하기 위하여 향후 어떤 것들을 중점적으로 연구해야하며 상용화 시스템을 설계할 때 어떤 점들을 간과해서는 안되는지 간단히 살펴 보았다. 현재의 기술 수준을 고려할 때 음성인식 기술의 상용화는 고립단어 인식기술의 상용화부터, 그것도 가능하다면 화자 종속으로 상용화가 가능한 분야부터 추진하는 것이 바람직할 것이다. 음성합성 기술의 경우, 빠른 시일 안에 합성음의 명료성 문제를 해결하면서, 한국어의 운율 및 강세 등에 대한 기초 연구를 수행하고 그 연구 결과들을 이용하여 자연성을 개선하는 방향으로 연구를 하는 것이 옳은 방향일 것이다. 즉, 일단 가까운 장래에 상용화의 가능성이 보이는 기본 기술들부터 연구하고 개선하여 상용화해 가는 한편, 보다 먼 장래를 위하여 한 차원 높은 음성처리 기술들을 동시에 연구해 나가는 것이 가장 바람직할 것이다. 마지막으로 실험실에서 성능이 우수한 음성인식 및 음성합성 기술을 이용하여 시스템의 개발을 성공적으로 마무리했다는 사실이 시스템을 상용화했을 때 사용자들의 호응을 보장해 주는 것이 아니며 따라서 사업의 성공을 의미하는 것은 절대 아니다라는 말을 다시 한번 인용하면서 본 고를 마치고자 한다.

參 考 文 獻

[1] J. L. Flanagan, "Voices of Men and Machines," *J. Acoust. Soc. Am.*, Vol. 51, pp. 1375-1387, May 1972.

[2] J. P. Haton, Automatic Speech Analysis and Recognition, D. Reidel Publishing Company, 1982.

[3] S. Furui, Digital Speech Processing, Synthesis, and Recognition, Marcel Dekker, 1992.

[4] J. L. Flanagan, "Computers that Talk and Listen: Man-Machine Communication by Voice," *Proc. IEEE*, vol.64, pp. 405-415, April 1976.

[5] T. Vitale, "Voice Output Systems and Technology," Pre-Conference Tutorial, AVIOS'92 Conference, Minneapolis, September 1992.

[6] W. Tetschner, "Voice Processing: State of Industry," Proc. of AVIOS'92 Conference, pp. 13-23, Minneapolis, September 1992.

[7] T. B. Schalk, "Voice Recognition Tutorials," Pre-Conference Tutorial, AVIOS'92 Conference, Minneapolis, September 1992.

[8] J. M. Lucassen, J. Gonzalez, E. Keppel, "Tangora: a large vocabulary speech recognition system for five languages," *Proc.* of AVIOS'92 Conference, pp. 281-287, AVIOS'92 Conference, Minneapolis, September 1992.

[9] S. A. Riederer, "CONVERSANT VIS Means Business," AT&T Technology, pp. 14-18, vol. 5, no.4, 1990.

[10] T. B. Schalk, "Speaker Verification Over the Telephone Network," *J. of Speech Technology*, pp. 32-35, Feb. 1991.

[11] D. B. Roe, D. P. Prezas, J. G. Wilpon, "AT&T's Speech Recognition in the Telephone Network," *J. of Speech Technology*, pp. Feb. 1991.

筆 者 紹 介

韓 敏 洙

1956年 11月 23日生
1979年 2月 서울대학교 전기공학과(학사)
1981年 2月 서울대학교 전기공학과(석사)
1989年 12月 Univ. of Florida 전기 및 전자공학과(박사)

1982年 4月 ~ 1985年 8月 한국표준연구원 연구원
1990年 2月 ~ 현재 한국전자통신연구소 음성응용연구실 실장

주관심분야: 음성분석, 음성 인식, 음성 합성

李 恒 燮

1967年 3月 12日生
1990年 2月 광운대학교 전자계산기공학과(학사)
1992年 2月 광운대학교 전자계산기공학과(석사)

1990年 4月 ~ 현재 한국전자통신연구소 음성응용연구실 연구원

주관심분야: 음성분석, 화자 인식, 음성 처리 시스템

鄭 柳 鉉

1956年 8月 10日生
1980年 2月 광운대학교 전자계산학과(학사)
1989年 2月 광운대학교 전자계산기공학과(석사)
1992年 9月 광운대학교 전자계산기공학과 박사과정

1980年 8月 ~ 현재 한국전자통신연구소 음성응용연구실 선임연구원

주관심분야: 음성 데이타베이스, 음성 분석, 음성 인식, 화자 인식