HCI를 위한 음성 입출력 처리 기술 개발

Speech Input/Output Processing Technology for Human-Computer Interface

이영직, 김회린, 이정철, 양재우

Youngjik Lee, Hoi-Rin Kim, Jung-Chul Lee, and Jae-Woo Yang

ylee@zenith.etri.re.kr, hrkim@zenith.etri.re.kr, jwyang@media.etri.re.kr

요약

본 논문은 정보통신부 출연의 "HCI를 위한 음성 입출력 처리 기술 개발" 과제에 대하여 기술한다. 이 과제의 주 목적은 PC 윈도우 환경에서 사람과 기계 간의 음성 입출력 기술을 개발하는 것이다. 이를 위해 음성 인식 분야에서는 화자 적응, 잡음 적응, 및 인식 대상 어휘 적응 기술을 개발하며, 합성 분야에서는 시스템 메시지 합성 기술을 개발한다. 또, 음성이 기존의 입출력 수단인 키보드나 마우스를 모두 대치할 수 없으므로, 본 과제에서는 음성이 추가됨으로써 입출력이 편리해지는 다중 모드 입출력 기술의 개발에 초점을 맞추어 기술을 개발하고 있다. 인식 분야의 주요 연구 내용은 음성검출 및 비음성 제거, 인식 속도 향상, 인식 성능 향상이며, 합성 분야 주요 연구 항목은 학습형 합성기 알고리즘 및 이의 문제점 해결이다. 본 논문은 이러한 점을 정리하여 발표한다.

1. 서론

20년 전만 해도 음성을 이용하여 컴퓨터를 조작하거나, 사람이 말을 하면 이를 그대로 컴퓨터가 받아 적는 장면은 공상 과학 영화에서나 볼 수 있었다. 그 동안 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.

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 더해 지더라도 성능이 현격히 떨어지는 경우가 이에 해당한다. 현재는 음성처리 기술의 활용이 시작되는 시기이므로 음성처리 기술의 활용에 많은 요구가 있으나, 음성처리 기술의 적용 상황에 따라 요구되는 세부 기술이 많이 차이가 나므로 주의를 요한다.

본 논문에서는, 음성처리 기술의 실용화를 위해 정보통신부에서 1997년부터 1998년까지 출연한 "HCI를 위한 음성입출력 처리기술 개발" 과제의 기술적 측면을 정리하였다. 제2장에서는 음성명령어 인식 기술에 대하여 기술하였으며, 제3장에서는 음성합성 분야의 연구 내용 및 결과를 설명하였다. 제 4장에는 사용자 인터페이스에 관련된 내용을 요약하였으며, 제 5장에 결론을 맺었다.

2. 음성 명령어 인식

본 연구는 음성 처리 기술의 실용화를 목적으로 수행하는 연구이다. 이 연구에서는 그림 1과 같이 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[1], 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. 이 연구에서는 입력을 영어 알파벳으로 하고 출력을 해당 발음기호로 하는 다층신경망의 학습을 이용하였다. 그러나 하나의 단어가 여러 개의 알파벳으로 이루어져 있으므로, 알파벳-발음기호 정변환율이 90%가 되더라도 단어-발음기호 정변환율은 50%가 되지 못한다. 이러한 다층신경망 기반 영어 발음사전 생성기의 성능을 개선시키기 위하여 기존의 영어 알파벳과 발음기호를 대응시키는 전처리 과정을 DTW 알고리즘을 이용하여 개선시키고, 하나의 다층신경망을 알파벳 별로 나누어 훈련시킴으로써 단어-발음기호 정변환율은 약 6% 개선됨을 볼 수 있었다.

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[2]. 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 인식율을 보인다.

3. 음성 합성

당 연구실에서는 1990 년대 초부터 무제한 문장-음성 변환기를 개발해 왔다. 초기의 TTS 에는 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[3].

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 개발한 바 있다.

하나의 음색을 가진 합성기의 개발 시간을 단축하고자, 현재는 trainable TTS 방식에 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.

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 음성 처리 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[4].

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.

5. 결론

이 논문은 음성의 실용화를 촉진시키기 위해 1997 년부터 1998 년까지 정보통신부에서 출연한 "HCI 를 위한 음성입출력 처리기술 개발" 과제의 연구 내용 및 결과를 기술한 것이다. 음성 명령어 인식 분야에서는 전처리 과정에서 음성 검출률을 높이기 위해 비음성 거절 기술을 개발하였고, 한국식 영어 인식을 위해 영어 발음 생성 기술을 개발하였으며, 인식 속도 개선을 위한 파이프라이닝 기술을 개발하였다. 음성합성 분야에서는 학습형 합성기술을 개발하였으며, 음성 인터페이스를 사용할 때 사용자 편의성 평가 기준을 마련하였다. 전자통신연구원은 개발된 기술을 바탕으로 음성처리 기술의 폭 넓은 실용화를 위해 계속 노력하고 있다.

6. 감사의 글

이 연구는 정보통신부 출연 "HCI 를 위한 음성입출력 처리기술 개발" 과제로 수행되었습니다.

7. 참고문헌

[1] 안영목, "비음성 검출 기술," 음향학회 논문지, 1998. 게재 예정.

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

[3] 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.

[4] 어홍준, 김범수, 한성호, 이영직, 이항섭, "음성인식을 이용한 사용자 인터페이스의 평가지침," HCI'98 학술대회, pp. 336-341, 1998. 2.

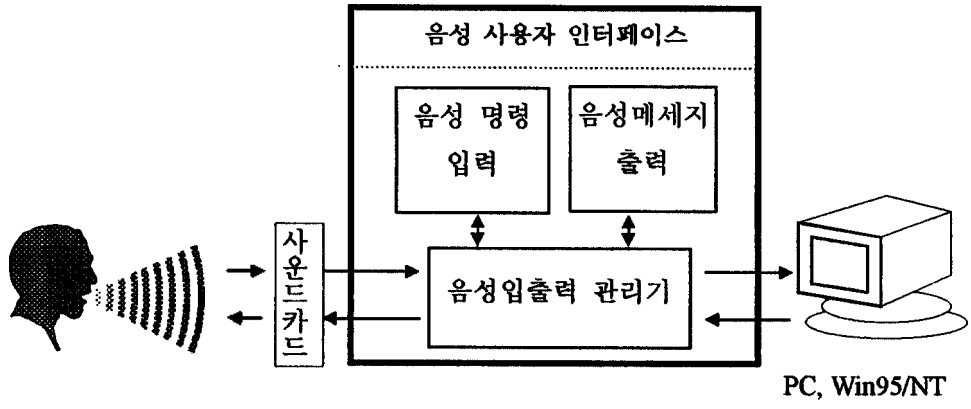

그림. 1 음성 사용자 인터페이스 소프트웨어의 구성도.