로봇 시스템의 음성 인식

최 동 진[1], 안 호 석[2], 최 진 영[3]
삼성전자[1], DR Robots[1,2], 서울대학교[2,3], PIRC[2,3], ASRI[2,3]

Speech Recognition in the Robot System

DongJin Choi[1], Ho Seok Ahn[2], Jin Young Choi[3]
Samsung Electronics Co[1], DR Robots[1,2], Seoul National University[2,3], PIRC[2,3], ASRI[2,3]

Abstract - 최근 음성인식률이 영어를 기준으로 97% 이상 높아짐에 따라 음성 인식을 사용한 여러 서비스들이 등장하고 있다. 지능 로봇도 예외가 아니어서 많은 서비스 로봇, 지능형 로봇에 음성 인식 기술이 적용되고 있다. 본 논문에서는 수많은 음성 인식 기술 중에서 로봇 시스템에 맞는 시스템을 선택하고, 이를 효율적으로 운용할 수 있는 방법을 제시한다. 또한 직접 지능형 서비스 로봇을 만들면서 음성 인식이 로봇에서 얼마나 중요한 위치를 차지하고 있는지를 확인하였으며 단순한 단어, 문장을 인식하는 것에 그치는 것이 아니라 향후 대화를 통해 사람과 로봇의 커뮤니케이션이 가능한 시스템을 제시한다.

1. 서 론

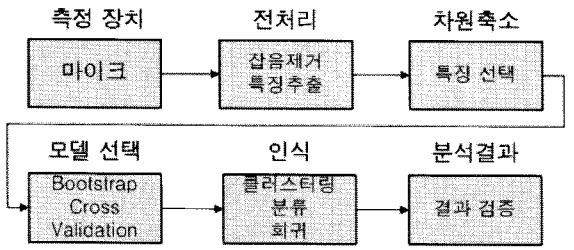
음성 인식 시스템은 1950년에 AT&T에서 음성학을 기초로 화자 종속의 숫자음 인식이 그 시초이다. 이후 50년간 음성 인식은 비약적으로 발전하였고 그 정확도가 상당히 높아졌다[1]. 음성 인식이 비약적으로 발전 할 수 있었던 이유는 무엇보다도 새로운 알고리즘이 개발과 더불어 S/W의 발전과 큰 관계가 있다. 특별한 하드웨어가 필요 없이 소프트웨어만으로도 음성 인식이 가능하게 되었고 비용도 많이 줄어들게 되었다. 그래서 최근 음성 인식이 지능형 로봇의 필수적인 기능 중 하나로 인식되고 있다. 그 예로 최근 국내외에서 발표된 거의 모든 지능형 로봇들은 음성 인식 기능을 가지고 있다.

하지만 국내 대부분의 로봇들에 적용되어 있는 음성 인식 시스템은 로봇을 위해 따로 개발된 것이 아니라 한 두 곳의 회사 또는 연구소에서 개발한 음성 인식 라이브러리를 이용하고 있는 실정이다. KAIST에서 개발한 휴보도 음성인식은 물론 TTS(Text to Speech) 기능도 예외는 아니다. 이러한 방식은 로봇 개발과 비용을 시간을 줄일 수 있다는 장점이 있지만, 사용된 라이브러리가 범용적으로 사용되기 때문에 로봇을 위해 개발된 음성 인식 소프트웨어가 필요하다. 본 논문에서는 로봇 시스템을 위한 음성 인식 방법을 소개한다. 2장에서는 음성 인식 시스템을 설명하고, 3장에서는 로봇의 음성 인식 시스템을 소개한다. 4장에서 본 논문의 결론을 맺는다.

2. 음성 인식 시스템

2.1 음성 인식

대부분의 음성 인식 원리는 미리 데이터화되어 저장한 음성 정보와 입력되고 있는 음성이 동일한지 비교하는 것이다. 즉, 입력 받은 패턴을 저장된 패턴과 비교하여 가장 유사한 패턴을 찾아내는 것이다. 그림 1은 이를 위한 패턴 인식 시스템의 처리 과정을 간단히 나타낸 것이다.



<그림 1> 패턴 인식 시스템의 처리 과정

음성 인식은 그림 1의 처리 과정에서 차원 축소 단계가 필요 없다. 그 이유는 음성이 1차원의 데이터이기 때문에 더 이상 차원을 축소 할 수 없기 때문이다. 음성 인식에 최근 많이 사용되는 알고리즘은 DTW(Dynamic time warping), HMM(Hidden markov model), 신경망 구조 등이 있다[2]. HMM은 본 논문에서 구성한 시스템에서 사용한 알고리즘으로 1960년대 말부터 여러 가지 예측 문제를 해결에 적용하면서 시작되었다. 현재 HMM은 음성 인식 뿐만 아니라 비정상적인 복잡한 패턴(경제학, 패턴인식, DNA 분석)을 모델링하는데 많이 사용되고 있다. 음성 인식 소프트웨어가 널리 사용되면서 음성 인식에 관한 Open Source가 많이 생겨나게 되었다. 표 1은 대표적인 음성 인식 Open Source의 목록이다. CMU Sphinx는 BSD license를 채택하고 있기 때문에 많은 곳에서 사용되고 있다.

<표 1> 대표적인 음성 인식 Open Source

CMU Sphinx	open source under a BSD license
HTK	copyrighted by Microsoft, but altering the software for the Licensee's internal use is allowed.
Julius	BSD-style license
VoxForge	open source, GPL

2.2 음성인식 시스템 구축 과정

본 논문의 시스템에서는 HMM을 프로그래밍으로 구현한 HTK (HMM Took Kit)을 사용하고 있다. HTK를 사용해 음성인식 시스템을 구축하기 위한 과정을 정리 하면 다음과 같다. 1번부터 5번 과정은 준비 과정이며, 6번부터 8번 과정은 음성 인식 실행이 가능한 과정이다.

1. Feature Vecture Extract :
   레코딩된 음성에서 특징 벡터를 추출 한다.
2. Transcription Label Create
   발음사전을 바탕으로 각 특징 벡터에 label을 부여 한다.
3. Flat Starting with HCOMPV
   HTK의 HCOMPV Tool을 이용해 세그멘테이션을 한다.
4. Embedded Training using HEREST
   3회 이상 트레이닝을 한다. 7회 이상 트레이닝해도 인식율에는 차이가 없다. 2회 이하로 트레이닝을 하면 인식율이 급격하게 떨어졌다.
5. the Task Grammar
   * Dial three three two six five four
   * Call Steve Young

위와 같은 문장은 표 2와 같은 문법으로 정의 할 수 있다.

<표 2> HTK를 사용한 문법의 예

```
$digit = ONE | TWO | THREE | FOUR | FIVE |
         SIX | SEVEN | EIGHT | NINE | OH | ZERO;
   $name  = [ JOOP ] JANSEN |
            [ JULIAN ] ODELL |
            [ DAVE ] OLLASON |
            [ PHIL ] WOODLAND |
            [ STEVE ] YOUNG;
( SENT-START ( DIAL <$digit> | (PHONE|CALL) $name) SENT-END )
```

여기까지 진행했다면 음성 인식을 실행 해 볼 수 있다. HVite라는 HTK Tool로 앞에서 생성된 파일을 입력해서 음성 인식을 진행하고 결과를 분석할 수 있다. 하지만 이 단계에선 monophone으로 HMM 알고리즘을 실행한 것이기 때문에 인식율이 많이 떨어진다.

6. Making Triphones from Monophones
   생성된 Triphone으로 4번 단계인 Trainning을 실시한다.
7. Making Tied-State Triphones
8. Speech Recognition from Tied-State Triphones.

3. 로봇의 음성 인식 시스템

3.1 로봇에서 사용되는 인식 알고리즘의 성능 비교

2장에서 언급 했듯이 음성 인식 시스템은 로봇에 적용되어 있는 얼굴 인

식이나 물체 인식 등의 다양한 인식 기술에 비해서 인식률이 상대적으로 높다. 표 3은 본 시스템에서 구현한 로봇에서 구현되어 있는 인식 시스템들을 비교한 결과이다.

<표 3> 다양한 인식 알고리즘의 성능 비교

	얼굴인식 (PCA)	물체인식 (SFIT)	음성인식 (HMM)
인식률	61%	75%	85% (한글)
인식시간	800ms	490ms	380ms

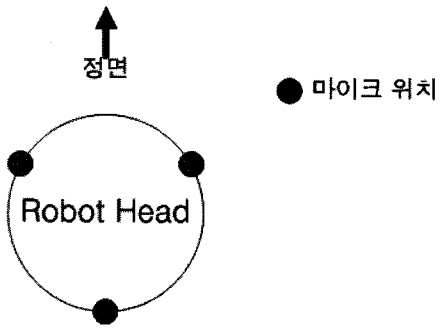
음성 인식은 표 3에서 알 수 있듯이 다른 인식 알고리즘에 비해서 비교적 빠르고 정확한 결과를 제공한다. 로봇이 음성 인식 시스템을 기본적으로 채택하고 있는 이유는 바로 로봇에게 입력 하는 수단으로 가장 알맞기 때문이다. 일반적으로 컴퓨터 시스템에 입력을 하기 위해선 키보드와 마우스를 사용하지만 로봇에게 키보드와 마우스를 사용하는 것은 많은 어려움이 따른다. 터치스크린을 사용해서 사용자의 입력을 받기도 하지만 이런 경우에는 로봇을 사용하기 위해서 로봇에 가까이 다가가야 한다는 단점이 있다. 음성 인식은 비교적 멀리 떨어져 있어도 사용자의 입력을 받을 수 있으며, 마이크 이외의 특별한 장치를 필요로 하지 않기 때문에 앞으로도 로봇에게 음성 인식은 중요한 기능이다.

3.2 로봇 음성인식 시스템 구축 시 주의 해야 할 점

로봇에서 음성 인식 시스템을 구축하기 위해선 다음의 특징을 알아 둘 필요가 있다.

1. 대부분의 지능형 서비스 로봇은 이동 로봇이다.
2. 현재 로봇은 제한된 H/W 성능을 가지고 있다.

음성 인식은 사용자의 변화는 물론이고 마이크의 종류나 로봇을 사용하는 장소에 따라서도 인식률이 다를 수 있다. 이동 로봇은 사람으로부터 멀리 떨어져 있을 수 있고, 소음이 심한 장소에서 사용 될 수도 있다. 또한 사람이 로봇의 뒤나 옆에 있을 수도 있다. 이러한 조건은 음성 인식의 성능에 치명적인 결과를 가져오기 때문에 사람이 가지고 있는 무선 마이크를 이용해서 사용자의 음성을 로봇에게 전달한다. 혼다의 아시모나 KAIST의 휴보가 이러한 방법을 사용한다. 본 논문에서 구현한 음성 인식 시스템에서는 다음과 같은 방법으로 이런 단점을 보안 했다.



<그림 2> 로봇의 마이크 배치도

그림 2와 같이 로봇의 머리 부분에 마이크 3개를 부착 했다. 지향성 마이크를 120도 각도로 배치해서 전방향에서 입력되는 음성 신호를 모두 받을 수 있다. 하지만 이와 같은 경우, 같은 음성이 여러 번 입력되는 경우가 발생한다는 단점이 있다. 입력된 신호를 분석해서 음성 인식 시스템에 전달해 주는 과정을 Negotiation 이라고 하고, negotiation 방법은 음압의 세기, 주파수 분석을 통해 결정하게 된다.

<표 4> 마이크의 위치에 따른 음성 인식률의 변화

	정면	후면	옆면
마이크가 정면에만 있음	83%	55%	80%
마이크를 120도 각도로 배치	82%	77%	77%

표 4는 마이크의 위치에 따른 음성 인식률의 변화이다. 마이크가 정면에만 있을 경우에는 인식률의 편차가 크지만, 마이크를 120도 각도로 배치했을 때에는 편차가 적어지는 것을 알 수 있다. 그리고 Negotiation 알고리즘을 개선하면 더 좋은 결과가 있을 것이다.

현재 Robot의 하드웨어는 계속 좋아지고 있지만 아직까지 로봇에 필요한 각종 소프트웨어를 제대로 구동시키기 위해선 모자란 부분이 있다. 로봇은 영상 처리, SLAM (Simultaneous Localization and Mapping), Motor control 등의 매우 다양한 소프트웨어 자원을 사용하기 때문에 음성 인식의 성능에 문제가 생길 수 있다. 또한 음성 인식 시스템은 상당히 많은 기억 공간을 필요로 한다. 따라서 메모리가 작은 임베디드 시스템의 경우에는 음성 인식이 필요로 하는 Data가 메모리에 Loading 될 수 없는 경우가 발생할 수 도 있다. 따라서 본 논문에서 구현한 시스템에는 로봇이 필요로 하는 인식 시스템 혹은 프로그램을 미리 예측해 필요한 수간에 음성 인식이 동작하는데 방해가 없도록 하는 스케줄링 기능을 탑재 했다[3-5]. 또한 로봇에 전화를 걸어 음성으로 명령을 내리는 기능이 있다[6]. 휴대폰의 음성을 소스로 음성 인식을 실행하면 표 5와 같은 결과가 나온다. VoIP일 경우, 휴대폰 보다 더 낮은 인식률이 나온다고 한다. 만약 로봇 음성 인식 중 휴대폰과 VoIP를 이용한 기능이 있다면 이를 고려해야 한다[7].

<표 5> 마이크의 위치에 따른 음성 인식률의 변화

	음성인식용 마이크	로봇의 마이크	휴대폰
인식률	88%	85%	71%

3.3 음성인식을 통한 대화 시스템

과거 음성 인식은 한 단어만 인식 할 수가 있었다. 하지만 현재는 알고리즘의 발전으로 문장을 인식 할 수 있고, 이로 인해 로봇과 대화를 할 수 있다. 만약 본 논문의 시스템처럼 HMM을 이용해 대화 시스템을 구축하려면 몇 가지 사항을 고려해야 한다.

첫째, 음성 인식 시스템 구축 과정의 5단계에서 정의한 문법은 음성 인식의 계산에 필요한 문법이다. 즉, 5단계의 문법에 맞지 않는 문장을 말해도 연속적인 단어의 입력(문장)을 인식 할 수 있다. 하지만 인식률은 낮아진다. 대화 시스템을 위해서는 인식할 문장을 미리 예상하고 문법을 정의해야 한다. 만약 기존 문법을 수정하려면 5단계부터 다시 진행하면 된다. 하지만 새로운 단어를 추가하게 되면 인식률에 영향을 미칠 수 있기 때문에 처음부터 다시 트레이닝 해주는 것이 좋다.

둘째, 본 시스템의 음성 인식 시스템은 입력되는 음성 중, 등록된 단어를 인식하게 되면 call-back 함수를 통해 로봇 시스템에 알려 주게 된다. 즉, 사용자가 문장을 말하게 되면 그 문장의 구성된 단어들이 하나씩 순차적으로 Robot system에 전송된다. 하지만 중간을 인식 못할 경우가 있거나 잘못된 단어로 인식 할 경우가 있다. 그렇기 때문에 문법과 비교를 해서 문장을 보정하는 알고리즘이 있어야 한다. 본 시스템에선 python의 dictionary 자료 구조와 라이브러리를 이용해 보정 모듈을 구현했다.

셋째, 정확도가 일정 수준 이하로 떨어지면 사용자에게 feedback을 받는다. 음성 인식의 결과 정확도가 떨어질 경우를 수치로 확인 할 수 있다. 이를 이용해 사용자에게 입력한 음성이 맞는지를 확인하는 메시지를 보여주거나 들려 줄 수 있어야 한다.

4. 결 론

현재 음성 인식 기능이 많이 활성화 되고 로봇에도 적극적으로 적용되고 있지만, 아직 음성 인식 기능은 많이 보완 되어야 한다. 특히 로봇에 적용되는 음성 인식에는 단지 음성 인식에 관련된 라이브러리를 아무런 분석 및 최적화 없이 사용하게 되면 효율적으로 사용 할 수 없게 된다. 본 논문에서 소개한 시스템과 같이 음성 인식 시스템을 구축하고 로봇에 적용시키면서 했던 일련의 작업들을 통해 로봇의 음성 인식률을 증가시키며 로봇이 수행하게 될 작업에 최적화된 음성 인식 시스템을 구축해야 한다.

[참 고 문 헌]

[1] Murray Hill, NJ, "An improved automatic lipreading system to enhance speech recognition," 1988.
[2] Young, S.J, "The general use of tying in phoneme-based HMM speech recognisers," ICASSP, 1992.
[3] Ho Seok Ahn, Young Min Beak, In-Kyu Sa, Woo Sung Kang, Jin Hee Na, and Jin Young Choi, "Design of Reconfigurable Heterogeneous Modular Architecture for Service Robots," In Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2008), pp.1313-1318, 2008.
[4] 안호석, 사인규, 백영민, 안윤석, 최진영, "유비쿼터스 환경에서 사용자의 일정에 따른 지능 정보 제공 시스템," 정보 및 제어 학술대회 2007 (CICS2007), pp.357-358, 2007.
[5] 안호석, 최진영, 김형준, 이형철, 백영민, 임종윤, 최한솔, 장혜진, "블록형 모듈화 서비스 로봇 Mom's Friend," 한국로봇공학회지, 제3권, 제4호, pp.67-74, 2006.
[6] 안호석, 사인규, 백영민, 안윤석, 최진영, "핸드폰을 통한 서비스 로봇 제어 시스템," 정보 및 제어 학술대회 2007 (CICS2007), pp.341-342, 2007.
[7] Evermann, G, "Development of the 2003 CU-HTK conversational telephone speech transcription system," ICASSP, 2004.