# 원거리 음성명령어 인식시스템 설계

*오유리[1], 윤재삼[1], 박지훈[1], 김민아[1], 김홍국[1]
공동건[2], 명현[2], 방석원[2]
[1]광주과학기술원 정보통신공학과
e-mail: {yroh, jsyoon, jhpark, kma58, hongkook}@gist.ac.kr
[2]삼성종합기술원 Micro Systems Lab.
{dgkong, h.myung, banggar.bang}@samsung.com

# Performance Evaluation of an Automatic Distance Speech Recognition System

*Yoo Rhee Oh[1], Jae Sam Yoon[1], Mina Kim[1], Hong Kook Kim[1],
Donggeon Kong[2], Hyun Myung[2] and Seokwon Bang[2]
[1]Department of Information and Communications, Gwangju Institute of Science and Technology
[2]Micro Systems Lab., Samsung Advanced Institute of Technology

## Abstract

In this paper, we implement an automatic distance speech recognition system for voiced-enabled services. We first construct a baseline automatic speech recognition (ASR) system, where acoustic models are trained from speech utterances spoken by using a cross-talking microphone. In order to improve the performance of the baseline ASR using distance speech, the acoustic models are adapted to adjust the spectral characteristics of speech according to different microphones and the environmental mismatches between cross-talking and distance speech. Next, we develop a voice activity detection algorithm for distance speech. We compare the performance of the baseline system and the developed ASR system on a task of PBW (Phonetically Balanced Word) 452. As a result, it is shown that the developed ASR system provides the average word error rate (WER) reduction of 30.6 % compared to the baseline ASR system.

## I. 서론

가정용 로봇과 같은 사용자의 편리성을 높인 고부가가치의 가전기기에 대한 수요가 날로 증가함에 따라 사용자의 편리성 향상을 위하여 전화 교환 시스템, 휴대폰의 전화번호 검색 등에서 제공되던 음성 명령 기능을 가정용 로봇 등의 가전기기에 적용하는 연구가 진행되어 오고 있다 [1]. 특히, 가까운 거리에서 발화하는 음성을 인식하는 휴대폰, 전화기 등과 달리 가전기기에서는, 사용자가 먼거리에서 발화하는 음성을 등록 및 인식하는 경우가 빈번히 발생하게 된다. 이 경우 음성인식률이 크게 저하되므로 이를 개선해야 한다. 또한, 인식할 명령어를 키보드 등으로 직접 등록하는 기존의 인식 시스템은 사용자 편의성 저하 등을 초래하므로 음성을 통하여 명령어를 등록할 필요가 있다.

본 논문에서는 사용자의 편의성을 위하여 음성 입력에 대한 거리 제한을 두지 않는 원거리 음성명령어 인식 시스템을 설계한다. 우선, 원거리 음성인식을 위한 음향모델의 설계를 위해, 가전기기가 사용되는 환경 및 마이크 특성, 사용자가 음성을 발화하는 다양한 거리 등을 고려한 적응 데이터로부터, 기존의 무제한 연속음성인식시스템에서 설계된 음향모델을 MAP/MLLR 기법을 이용하여 적응시킨다 [2]. 또한, 원거리 음성 명령어 인식 시스템의 성능을 향상시키기 위하여 Teager energy 기반의 음성구간검출 알고리즘을 설계 구현한다.

본 논문의 구성은 다음과 같다. 제 I 장의 서론에 이어, 제 II 장에서는 음향모델 적응 및 원거리 음성명령어 인식시스템에 대하여 기술한다. 이어 제 III 장에서는 본 논문에서 구현한 원거리 음성인식시스템의 인식성능을 보이고, 제 IV 장에서 결론을 맺는다.

## II. Baseline 음성 인식시스템

Baseline 음성 인식시스템은 음성정보기술산업지원센터에서 제공하는 낭독문장 음성 DB (CleanSent01) [3]로 학습되었다. CleanSent01은 형태소 빈도를 고려한 20806 문장을 남녀 200명이 발화한 음성으로 구성되었다. 또한 발화 음성은 방음실 환경에서 제작되었으며, AKG C414-ULS와 Sennheiser 마이크로 동시에 녹음되어 16 kHz의 샘플링레이트, 16 bit로 저장되었다. 음성특징벡터로는 39차 특징벡터가 사용되었으며, 이를 위하여 12차 멜-켑스트럼 계수(MFCC), 로그 에너지를 추출하였고, 1차 및 2차 미분계수를 사용하였으며 에너지 정규화 기법이 적용되었다. 음향모델로는 3개 상태의 천이를 left-to-right로 하는 HMM과 4개의 혼합밀도를 갖는 Gaussian 분포의 문맥독립 cross-word triphone 모델을 사용하였다. 결론적으로, baseline 음성인식 시스템은 14,901개 triphone과 15,182개 상태의 음향모델로 구성되었다.

## III. 원거리 음성명령어 인식시스템

### 3.1 음향모델 적응

원거리 음성명령어 인식시스템에 적합한 음향모델을 위하여, MAP/MLLR 적응기법을 이용한 음향모델 적응과정 수행한다. 첫째, 문장음성으로 학습된 baseline 음성인식시스템의 음향모델을 단어음성에 적합하게 단어음성 DB를 이용하여 적응한다. 다음으로, AKG와 Sennheiser 마이크로 녹음된 음성 DB로 학습된 음향모델을 본 논문에서 구축한 인식시

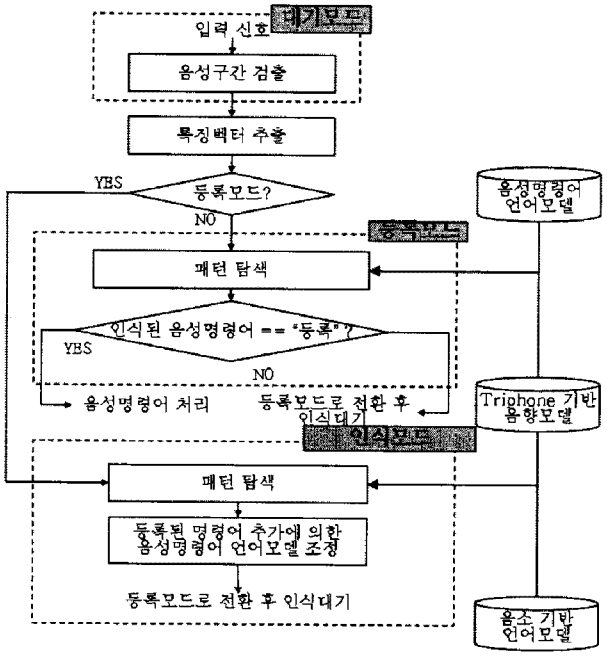


그림 1. 원거리 음성명령어 인식시스템 구성도.

스템의 마이크 특성에 맞게 적응한다. 또한, 가전기기의 사용자가 음성명령어를 발화하는 거리가 다양하기 때문에, 이들 다양한 거리의 음성으로 부터 음향모델을 적응시킨다. 또한, 문장 중심으로 학습된 baseline 음성 인식시스템의 음향모델을 단어 중심으로 적응시키기 위하여, ETRI의 한국어 헤드셋 음성인식용 단어 DB를 사용하였다.

### 3.2 음성구간검출 알고리즘

가전기기 등의 시스템에서 인식을 제공하기 위해서는 실시간으로 입력되는 신호 중에서 음성구간을 추출해야한다. 이를 위하여 본 논문에서는 Teager energy를 기반으로 한 음성구간검출 알고리즘을 사용한다. 음성의 진폭과 주파수를 이용한 Teager energy 기반 음성구간검출 알고리즘은 마찰음, 파찰음 등 진폭이 작은 발음으로 시작하는 단어에 대하여 우수한 성능을 보인다. 뿐만 아니라, 에너지와 영교차율을 이용하는 기존의 음성구간검출 방법과 달리 하나의 파라메타를 사용함으로써 계산량을 줄일 수 있어 가전기기에의 적용에 용이하다 [4][5].

### 3.3 원거리 음성명령어 인식시스템 구축

원거리 음성명령어 인식을 위하여 그림 1과 같이 대기모드, 인식모드, 등록모드로 구성되는 음성인식시스템을 구축한다. 대기모드에서는, 음성인식시스템의 입력신호로부터 음성구간을 검출한다. 검출된 음성구간에 대해 10ms 마다 특징벡터인 39차 MFCC를 추출한다. 마지막으로, 음성구간에서 추출된 특징벡터로 Viterbi 알고리즘을 기반으로 음성인식을 수행한다. 첫째, 인식모드인 경우, triphone 기반 음향모델과 등록된 음성명령어로 구성된 언어모델을 이용하여 단어인식을 수행한다. 단어인식 결과가 등록된 음성명령어인 경우 명령어를 처리하고, 단어인식 결과가 "등록"인 경우 등록모드로 전환한 후 대기모드로 다음 음성을 기다린다. 둘째, 등록모드인 경우, 음성인식시스템의 사용자가 어떠한 명령어를 등록할 지 알 수 없기 때문에 triphone 기반 음향모델과 음소기반 back-off bigram 언어모델을 이용하여 음소인식을 수행한다. 인식된 음소열을 명령어로 언어모델에 추가함으로써 등록모드를 종료하고 대기모드로 다음 음성을 기다린다.

표 1. 음향모델 적응에 의한 단어인식률 (%) 비교.

| 음향모델 \ 발성 거리 | 1m | 3m | 5m | 평균 인식률 |
|---|---|---|---|---|
| 연속음성인식용 음향모델 (baseline) | 78.7 | 69.3 | 49.3 | 65.8 % |
| 단어인식을 위한 음향모델 적응 | 85.3 | 93.3 | 73.3 | 83.0 % |
| 인식시스템 환경 및 다양한 거리 음성에 대한 적응 | 98.7 | 94.7 | 96.0 | 96.4 % |

표 2. 음성 기반 명령어 등록에 대한 단어인식률 (%).

| 음향모델 \ 발성 거리 | 1m | 2m | 3m | 5m | 평균 인식률 |
|---|---|---|---|---|---|
| 음성 기반 등록된 20개 명령어에 대한 인식성능 | 93.8 | 100 | 100 | 92.5 | 96.6 % |

## Ⅳ. 음성인식 실험 및 결과

본 절에서는 제 Ⅲ 절에서 설계한 원거리 음성명령어 인식시스템에 대한 성능을 평가한다. 먼저, 4.1절에서는 원거리 음성명령어에 적응시킨 음향모델의 성능을 평가한다. 그 후, 4.2절에서는 음성으로 등록된 명령어에 대한 단어인식 성능을 보인다.

### 4.1 음향모델 성능 평가

구축된 원거리 음성명령어 인식시스템의 환경에 대한 적응 및 다양한 거리에서의 음성인식 성능 향상을 위하여, 남자 5명과 여자 3명이 1m, 2m, 3m, 5m 거리에서 각각 PBW 452 단어를 발성한 음성을 음향모델 적응에 이용하였다. 표 1은 음향모델 적응에 의한 인식성능 향상을 나타낸다. 음향모델 적응용 음성 DB 구축에 참가하지 않은 남자 1명이 1m, 3m, 5m 거리에서 PBW 452 단어를 발성한 음성으로 인식실험을 수행하였다. 먼저, 단어 DB로 baseline 음성 인식시스템의 연속음성인식용 음향모델을 적응시킴으로써 음성명령어에 대한 거리별 평균 인식률이 65.8%에서 83.0%로 향상되었다. 또한, 음성 인식시스템 환경 및 다양한 거리 음성에 대하여 추가적인 적응을 수행함으로써 96.4%로 향상되었다. 결과적으로, baseline 음성인식시스템의 연속음성용 음향모델을 적응시킴으로써 30.6%의 인식성능 향상을 보였다.

### 4.2 등록된 음성명령어 인식시스템 성능 평가

구축한 원거리 음성명령어 인식시스템을 통하여 음향모델 적응용 음성 DB 구축에 참가하지 않은 남자 1명이 1m 거리에서 발성한 20개의 명령어를 등록하였다. 남자 1명이 발성한 20개의 음성명령어 6 세트로 인식실험을 한 결과, 표 2에서 보는 바와 같이 평균 96.6%의 단어인식률을 보였다. 음성으로 등록한 명령어에 대한 인식률이 96.6%로서 음성 기반 명령어 등록의 신뢰성을 확인할 수 있었다.

## 참고문헌

[1] 장길수, "지능형로봇의 기술 및 산업동향," *전자부품연구원 전자정보센터(EIC)*, 2005년 12월.

[2] G. Zavagliakos, R. Schwartz, and J. McDonough, "Maximum a posteriori adaptation for large scale HMM recognizers," in *Proc. ICASSP*, pp. 725-728, May 1996.

[3] 김봉완, 최대림, 김영일, 이광현, 이용주, "SiTEC의 공동 이용을 위한 음성 코퍼스의 구축 현황 및 계획," *대한음성학회 말소리*, 제 46호, pp. 175-186, 2003년 6월.

[4] G.S. Ying, C.D. Mitchell, and L.H. Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement," in *Proc. ICASSP*, pp. 732-735, Apr. 1993.

[5] 이재한, 백성중, 성굉모, "변형된 Teager 에너지에 기초한 음성끝점 검출 알고리듬에 관한 연구," *한국 음향학회 학술발표대회 논문집*, 제17권, 제2(s)호, pp. 407-410, 1998년 11월.