가정용 로봇의 호출음 등록 및 인식 시스템

A Name Recognition Based Call-and-Come Service for Home Robots

오유리, Yoo Rhee Oh*, 윤재삼, Jae Sam Yoon*, 박지훈, Ji Hun Park*, 김민아, Mina Kim*,
김홍국, Hong Kook Kim*, 공동건, Donggeon Kong**, 명현, Hyun Myung**, 방석원, Seokwon Bang**

요약 본 논문에서는 Call-and-Come 서비스를 제공하는 가정용 로봇의 호출음 등록 및 인식 시스템을 구축하고, 음성 기반의 효율적인 로봇 호출음 등록 및 인식 알고리즘을 제안한다. 본 논문에서는 음성을 이용하여 로봇 호출음을 효율적으로 등록하기 위해 monophone 음향모델을 이용하여 탐색 범위를 줄이고, 줄어든 탐색 범위 내에서 triphone 음향모델을 이용하여 호출음을 등록을 한다. 또한, 잘못된 호출이 인식되는 것을 줄이기 위한 발화 검증에 필요한 파라미터를 구한다. 원거리 음성인식률을 향상시키기 위해서 근거리 음성에 최적화된 음향모델을 원거리 음성 데이터베이스로 적응시켰으며, 마이크로폰 배열을 이용하여 사용자의 위치를 추정한다. 제안한 시스템의 성능 측정을 위해 수행된 로봇 호출음에 대한 등록 및 인식 실험에서 98.3%의 음성 인식률을 얻었다.

Abstract We propose an efficient robot name registration and recognition method in order to enable a Call-and-Come service for home robots. In the proposed method for the name registration, the search space is first restricted by using monophone-based acoustic models. Second, the registration of robot names is completed by using triphone-based acoustic models in the restricted search space. Next, the parameter for the utterance verification is calculated to reduce the acceptance rate of false calls. In addition, acoustic models are adapted by using a distance speech database to improve the performance of distance speech recognition, Moreover, the location of a user is estimated by using a microphone array. The experimental result on the registration and recognition of robot names shows that the word accuracy of speech recognition is 98.3%.

핵심어: Home robot, Call-and-Come service, Robot name registration, Distance speech recognition, Voice interface, Speaker location detection

* 주저자 : 광주과학기술원 정보통신공학과 e-mail: yroh@gist.ac.kr
* 공동저자 : 광주과학기술원 정보통신공학과 e-mail: jsyoon@gist.ac.kr
* 공동저자 : 광주과학기술원 정보통신공학과 e-mail: jh_park@gist.ac.kr
* 공동저자 : 광주과학기술원 정보통신공학과 e-mail: kma58@gist.ac.kr
* 교신저자 : 광주과학기술원 정보통신공학과 교수 e-mail: hongkook@gist.ac.kr
** 공동저자 : 삼성종합기술원 e-mail: dgkong@samsung.com
** 공동저자 : 삼성종합기술원 e-mail: h_myung@samsung.com
** 공동저자 : 삼성종합기술원 e-mail: banggar_bang@samsung.com

1. 서론

로봇 등 고부가가치의 가전기기에 대한 수요가 날로 증가함에 따라, 로봇 및 가전기기에 음성인터페이스를 접목시키는 연구가 진행되어 오고 있다 [1]. 특히, 주부의 가사노동을 보조하는 로봇, 청소 로봇, 잔디 깎기 로봇 등과 같은 가정용 로봇에 Call-and-Come 서비스를 제공하기 위해서는, 사용자의 편의성과 로봇과의 친밀성을 얻기 위해 음성으로 명령하여 작동하는 것이 효율적이다. 이를 위해서는 먼저, 가정용 로봇에 대한 호출음을 실시간으로 등록하고, 인식하는 기능이 필요하다. 이때 실시간은 사람과 가정용 로봇 사이의 대화가 사람과 사람 사이에 이루어지는 대화인 것처럼 느껴질 수 있도록 로봇이 음성을 처리하는 시간을 사용자가 느끼지 못하여야 한다. 또한, 가정용 로봇의 Call-and-Come 서비스의 경우, 원거리에서 사용자가 자연스럽게 발화한 음성을 인식하는 경우가 빈번히 발생하므로, 이에 대한 인식 성능을 개선해야 한다.

본 논문에서는 가정용 로봇의 Call-and-Come 서비스에서 사용자의 편의성 및 로봇과의 친밀성을 제공하기 위하여, 음성을 이용한 로봇 호출음의 등록 및 인식 알고리즘을 제안한다. 먼저 로봇 호출음을 등록하기 위하여, monophone 기반 음향모델로 탐색 범위를 줄이고, 줄어든 탐색 범위 내에서 triphone 기반 음향모델로 로봇 호출음을 최종적으로 등록하는 2-pass decoding 방법을 제안한다 [2,3]. 또한, 로봇 호출음 등록 및 인식 과정에서 잘못 인식된 경우를 처리하기 위하여, 발화 검증 기법을 적용한다. 다음으로, 가정용 로봇의 사용 환경에 적합한 인식 시스템을 설계하기 위하여, 가정용 로봇이 사용되는 환경 및 마이크 특성, 사용자가 음성을 발화하는 다양한 거리 등에 맞게 MAP/MLLR 기법을 이용하여 음향모델을 적응시킨다 [4]. 마지막으로, 가정용 로봇에 장착된 마이크로폰 배열을 이용하여 사용자의 위치를 추정하여 Call-and-Come 서비스를 실현한다.

본 논문의 구성은 다음과 같다. 2장에서는 구현한 Call-and-Come 서비스에 대해 간단히 기술하고 3장에서는 로봇 호출음의 인식 및 등록 알고리즘을 제안한다. 4장에서는 원거리 음성인식에 대한 성능 향상 방법을 기술한다. 5장에서는 구현한 시뮬레이터와 호출음 등록 및 인식 성능을 보이고, 6장에서 결론을 맺는다.

2. 가정용 로봇의 Call-and-Come 서비스

Call-and-Come 서비스는 사용자가 가정용 로봇을 호출하면, 가정용 로봇은 사용자의 위치를 추적하여 호출한 사용자에게 접근하는 것을 말한다. 본 논문에서는 사용자 친화적인 Call-and-Come 서비스를 위하여, 음성인터페이스를 이용한 Call-and-Come 서비스를 효율적으로 제공하도록 한다. 이를 위하여, 임의의 호출음 음성을 실시간으로 등록 및 인식하고 사용자의 방향을 추정하는 알고리즘을 제안한다.

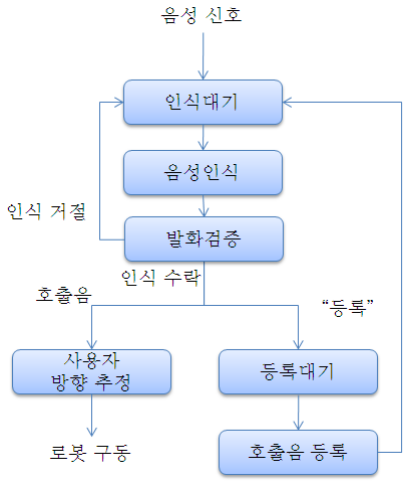

그림 1. 음성 인터페이스 기반 Call-and-Come 서비스 시스템.

그림 1은 본 논문에서 사용한 음성인터페이스 기반 Call-and-Come 서비스 시스템의 구성을 보여준다. 먼저, 마이크로폰으로부터 입력받은 신호에서 음성 구간이 검출될 때까지 '인식대기' 상태에 놓인다. 이때, 음성 구간이 검출되는 경우, 입력된 신호에 대하여 음성 인식을 수행한다. 또한 인식 결과에 대한 신뢰도를 바탕으로 발화 검증을 수행함으로써, 가정용 로봇의 오작동을 줄이도록 한다. 인식된 결과에 대하여, 발화 검증을 실패하면 음성 구간을 다시 검출하기 위하여 '인식대기' 상태로 돌아간다. 반대로 발화 검증에 성공한 경우, 호출음에 대한 처리와 '등록' 에 대한 처리로 나뉜다. 즉, 호출음을 인식한 경우, 호출음을 발화한 사용자의 방향을 추정한 후 가정용 로봇을 구동하도록 한다. 반대로 '등록' 을 인식한 경우, 시스템은 '등록대기' 상태에 놓이고 호출음을 등록하는 과정을 수행한다.

3. 로봇 호출음 등록 및 인식 알고리즘

본 논문에서 음성인터페이스를 이용한 Call-and-Come 서비스를 제공하기 위하여, 총 4개의 마이크로폰을 90도 간격으로 하여 원형으로 배치하였다. 여기서 마이크로폰 배열의 4 채널 음성 신호는 특징 벡터를 추출하거나, 사용자의 위치를 추정하는 데 이용된다. 그림 2는 본 논문에서 구축한 로봇 호출음 등록 및 인식과정을 보며, 크게 대기모드, 인식모드, 등록모드 등으로 구성된다. 먼저, 대기모드는 마이크로폰으로부터 입력받은 신호에서 Teager energy를 기반으로 음성 구간을 검출한다. 대기모드에서 음성구간을 검출하면, 음성구간에 대해 10ms 마다 특징벡터인 39차 MFCC를 추출한 후, 설정된 모드에 따라 인식모드 또는 등록모드로 전환한다.

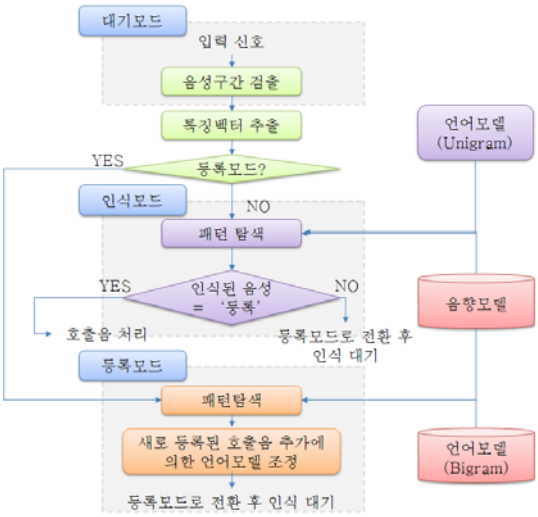

그림 2. 로봇 호출음의 등록 및 인식 과정.

먼저 설정된 모드가 인식모드인 경우, triphone 기반 음향모델과 등록된 호출음으로 구성된 언어모델 (unigram)을 이용하여 단어인식을 수행한다. 단어인식 결과가 등록된 호출음인 경우 2장에서 언급한 바와 같이 호출음을 발화한 사용자의 방향을 추정한 후 가정용 로봇을 구동하도록 한다. 단어인식 결과가 "등록" 인 경우 등록모드로 전환한 후, 등록할 호출음에 대한 음성을 입력받아 음성인식을 수행한다. 그 후 기존의 호출음에 대한 언어모델에 새로운 호출음을 추가하는 과정을 수행한다.

3.1 음성구간검출 알고리즘

음성의 진폭과 주파수를 이용한 Teager energy를 기반으로 음성구간을 검출함으로써 마찰음과 파열음 등 진폭이 작은 발음으로 시작하는 단어에 대하여 우수한 성능을 보이며, 근거리 발성음성에 비해 상대적으로 작은 진폭을 갖는 원거리 발성음성 검출에도 효과적이다 [5]. 뿐만 아니라 에너지와 영교차율을 이용하는 기존의 음성구간검출 방법[6]과 달리 하나의 파라메터를 사용함으로써 계산량을 줄일 수 있어 가정용 로봇에의 적용이 용이하다 [7].

대기 모드에서 Teager energy 기반으로 음성구간을 검출하기 위하여 다음 식 (1)과 같은 음성의 각 샘플에 대한 Teager energy 파라메터를 추출한다.

$$E_{Teager}(n) = x_n^2 - x_{n+1}x_{n-1} = A^2\sin(\Omega) \approx A^2\Omega^2 \quad (1)$$

여기서, A는 음성의 진폭을 Ω는 음성의 주파수를 나타내며, n은 음성의 샘플에 대한 인덱스이다.

Teager energy 기반 음성구간검출은 매 프레임별로 획득된 Teager energy 파라메터와 하위문턱값, 상위 문턱값과의 비교를 통하여 검출된다. 하위 문턱값 Th_{Lower} 과 상위 문턱값 Th_{Upper}은 10 프레임마다 식 (2)와 같이 계산한다. 이 때 이전 하위 문턱값과 새로 구한 하위 문턱값을 7:3의 비로 적용하여 새로운 하위 문턱값을 획득한다.

$$Th_{Lower} = \sigma_{Teager} + \frac{1}{10}\sum_{i=1}^{10} E_{Teager,i} \qquad Th_{Upper} = w \times Th_{Lower} \quad (2)$$

여기서 σ_{Teager}는 초기 10 프레임의 Teager energy에 대한 표준편차이며, w는 하위 문턱값에 따라 설정되는 가중치로 식 (3)과 같이 계산된다.

$$w = \begin{cases} 10 & , Th_{Lower} < L_1 \\ -\frac{1}{10000} \times Th_{Lower} + 25 & , L_1 \le Th_{Lower} < L_2 \\ 2 & , otherwise \end{cases} \quad (3)$$

가중치 값이 너무 커질 경우 상위 문턱값이 크게 설정되어 음성구간을 검출할 수 없기 때문에 이를 방지하기 위해 하위 경계값 L1을 설정하며, 잡음이 큰 환경에서 가중치 값이 1이 되지 않도록 상위 경계값 L2를 설정한다.

음성의 시작점 검출은 상위 문턱값보다 큰 Teager energy 파라메터를 갖는 프레임을 시작점으로 설정한다. 우선 하위 문턱값 보다 큰 에너지 파라메터를 가질 경우 임시 시작점으로 설정하고, 상위 문턱값보다 높은 에너지 파라메터를 가지면 임시 시작점을 실제 음성구간의 시작점으로 설정한다. Teager energy 파라메터가 상위 문턱값보다 큰 값을 갖기 전에 다시 하위 문턱값보다 낮은 값을 가질 경우는 설정된 임시 시작점을 삭제하고 새로운 임시 시작점을 검색한다.

음성의 끝점 검출은 시작점 검출과 반대로 하위 문턱값보다 낮은 Teager energy 파라메터를 갖는 프레임이 3 프레임 이상 지속될 경우 가장 끝 프레임을 끝점으로 설정한다. 또한 단어내의 휴지구간을 고려하여 이 후 10 프레임에 대해 새로운 시작점을 검색한다. 이 때 새로운 시작점이 검출되면 기존의 끝점에 해당하는 정보는 삭제되며 새로운 끝점을 검색하게 되고, 반대로 새로운 시작점이 검출되지 않으면 설정된 음성의 시작점과 끝점 사이의 구간을 최종 음성구간으로 설정한다.

마지막으로 음성구간의 신뢰성 향상을 위해 단어 유효성을 판단한다. 이 때 단어 평균길이 대비 최소, 최대 길이를 설정함으로써 검출된 음성구간이 기준보다 적거나 길면, 검출된 음성구간정보를 삭제하고 새로운 음성구간을 검색한다.

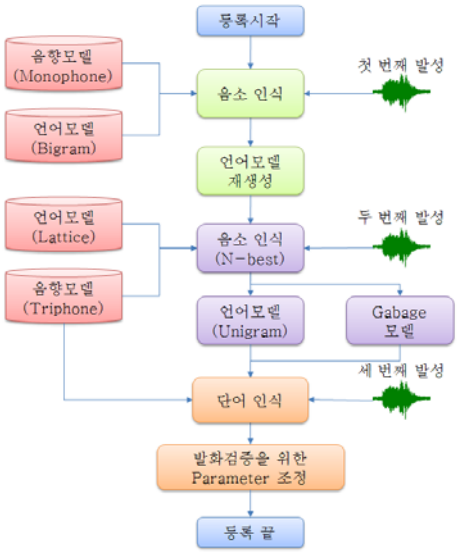

그림 3. 음성인터페이스를 이용한 Call-and-Come 서비스를 위한 호출음 등록 구성도.

3.2 로봇 호출음 등록 알고리즘

그림 3은 음성인터페이스를 이용한 Call-and-Come 서비스를 위한 호출음 등록 과정을 보여 준다. 첫 번째 발성한 호출음에 대한 등록 과정에서는, monophone 기반 음향모델과 두 음소간의 천이확률을 정의한 언어모델 (back-off bigram)을 이용하여 음소 인식을 수행하여 n-best 리스트 기반 언어모델(lattice)를 획득함으로써 탐색 범위를 줄인다. 동일한 호출음을 발성한 두 번째 음성에 대해서는, triphone 기반 음향모델과 첫 번째 호출음 발성 음성으로 생성된 음소기반의 언어모델 (lattice)을 이용하여 음소 인식을 수행함으로써, 호출음에 대한 언어모델을 갱신하고 발화검증을 위한 garbage 모델을 생성한다. 동일한 호출음을 발성한 세 번째 음성에 대해서는, 발화검증에 사용될 파라메터를 조정한다.

호출음 등록 과정에서 두 번의 발성을 통하여 호출음에 대한 언어모델을 생성하는 이유는, 임의의 호출음을 등록하기 위해서는 많은 계산량이 필요하므로 실시간 처리가 힘들기 때문이다. 일반적으로 음성인식의 성능 향상을 위하여 triphone 기반 음향모델을 사용한다. 그러나 약 70,000 개의 triphone 음소에 대한 음향모델과 두 음소 사이의 확률값을 고려하는 back-off bigram 언어모델을 사용할 경우, 70,000 개 음소의 제곱에 해당하는 수만큼의 탐색공간이 생기게 된다. 이로 인해 계산량이 증가하고 음소인식에 소요되는 시간이 크게 늘어나므로 실시간으로 호출음을 등록하는데 어려움이 있다. 이러한 문제를 해결하기 위해서 첫 번째 등록 호출음 발성을 통해서는 42 개의 음소로 구성된 monophone 기반 음향모델을 사용하여 back-off bigram 언어모델의 탐색 공간을 줄임으로써, 한정된 음소 및 음소간 천이 확률만으로 구성된 lattice를 생성한다. 그 후, 두 번째 등록 호출음 발성을 통해서는, 보다 정확한 음소 인식을 위하여 triphone 기반 음향모델을 사용한다. 특히, 첫 번째 등록 호출음 발성을 통해 재 생성된 언어모델인 lattice를 사용함으로써 실시간으로 수행이 가능하게 된다.

3.3 발화검증 알고리즘

발화검증은 garbage model을 바탕으로 하며, 인식된 호출음의 likelihood와 해당하는 garbage model의 likelihood에 대한 비를 계산하여 호출음의 유효성을 판단한다.

등록 모드에서 호출음에 대한 garbage model을 생성하고 이에 대한 파라메터를 조정하는 과정은 다음과 같다. 첫째, 호출음 등록 과정 중 두 번째 음성으로 호출음을 등록한 후 garbage model이 생성된다. 여기서 garbage model은, 등록된 호출음에서 사용되지 않는 음소들에 해당하는 HMM를 조합하여 mean, transition matrix, mixture weight에 대해서는 평균값을, variance에 대해서는 큰 값을 계산하여 garbage model 을 생성한다. 둘째, 새로 등록된 호출음을 추가하여 로봇이름 인식용 언어모델을 재생성하고 등록된 로봇이름에 대한 garbage model에 대한 언어모델을 생성한다. 그 후, 두 번째 음성에 대하여 로봇이름 인식용 언어모델로 인식을 수행하여 등록된 로봇이름에 대한 likelihood를 저장하고, 두 번째 음성에 대하여 garbage model에 대한 언어모델로 인식을 수행하여 garbage model에 대한 likelihood를 저장한다. 셋째, 세 번째 음성에 대하여 두 번째 과정을 반복 수행함으로써 등록된 호출음에 대한 likelihood와 garbage model에 대한 likelihood를 저장한다. 마지막으로, 두 번째 음성과 세 번째 음성을 인식 후 호출음에 대한 likelihood와 garbage model에 대한 likelihood를 이용하여 등록된 호출음에 대한 파라메터를 조정한다.

인식 모드에서 발화검증을 수행하는 과정은 다음과 같다. 첫째, 인식된 호출음에 대한 likelihood를 획득한다. 둘째, 인식된 호출음에 해당하는 garbage model에 대한 언어모델을 재생성 한 후, garbage model 용 음향모델과 네트워크로 다시 인식을 수행하여 garbage model에 대한 likelihood를 획득한다. 마지막으로 인식된 호출음에 대한 likelihood와 garbage model에 대한 likelihood를 호출음의 등록 과정에서 조정된 파라메터와 비교함으로써 인식된 호출음에 대한 유효성을 검사한다. 만약 인식된 호출음이 유효한 것으로 판단된 경우 사용자의 방향을 추정하고, 그렇지 않은 경우 대기 모드로 전환한다.

4. 인식률 향상 방안

가정용 로봇의 경우, 사용자가 원거리에서 자연스럽게 발화하는 음성에 대한 Call-and-Come 서비스를 효율적으로

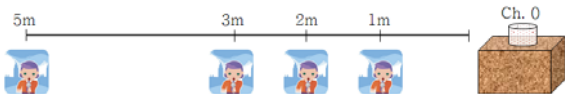

그림 4. 원거리 음성 데이터베이스 수집 환경.

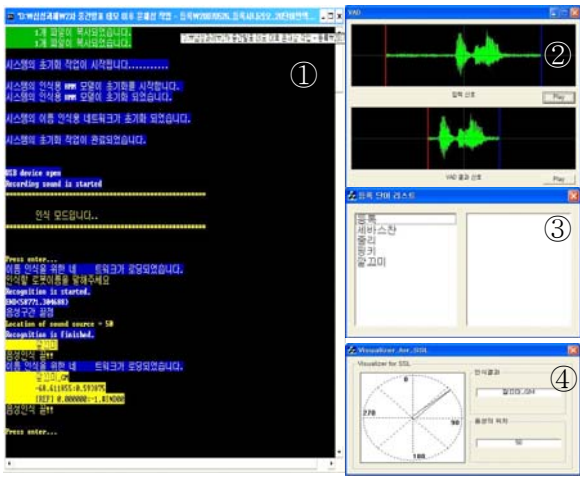

그림 5. 가정용 로봇 호출음 등록/인식 시뮬레이터.

제공해야 한다. 그러나, 음성인식시스템의 학습을 위하여 구축된 음성데이터베이스는 방음실 환경에서 연속 문장을 발화한 음성으로 구성된 경우가 많다. 그러므로 본 논문에서는 Call-and-Come 서비스에 적합한 호출음 인식 시스템의 인식성능 개선을 위하여, MAP/MLLR 적응기법을 이용한 음향모델의 적응과정을 수행한다.

먼저, 연속 문장 음성 데이터베이스[8]로 학습된 음성인식 시스템의 음향모델을 단어 위주인 호출음 음성에 적합하도록, ETRI의 한국어 헤드셋 음성인식용 단어 데이터베이스를 이용하여 적응한다. 다음으로, AKG와 Sennheiser 마이크로 녹음된 음성 DB로 학습된 음향모델을, 가정용 로봇의 Call-and-Come 서비스를 위하여 구축된 호출음 인식 시스템 마이크 특성에 맞게 적응한다. 마지막으로, 원거리 음성 데이터베이스를 이용하여 원거리 음성 특성에 맞게 음향모델을 재적응시켰다. 이때 사용된 원거리 음성 데이터베이스는 8명의 성인 (남자:5명, 여자:3명)이 그림 4와 같이 거리별 (1m, 2m, 3m, 5m)로 452개의 단어를 발성하여 구축되었다.

5. 시뮬레이터 구현 및 음성인식 실험 결과

그림 5는 본 논문에서 구축한 음성인터페이스 기반 Call-and-Come 서비스의 시뮬레이터를 보여준다. ①은 음성 인터페이스의 메인 화면으로서 인식 결과, 로봇 호출음의 등록 과정 및 발화 검증 결과 등을 출력해준다. ②는 자동으로 검출된 음성구간을 표시해주며, ③은 등록되어 있는 호출음을 보여준다. 마지막으로, ④는 마이크로폰 배열의 신호로부터 예측된 사용자의 위치를 표시해준다.

5.1 Baseline 음성인식시스템

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 멜켑스트럼 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.

음향모델 적응 및 음성구간검출의 성능을 평가하기 위해 452개의 phonetically balanced words (PBW) 단어리스트를 이용하였다. 마이크로부터 1, 3, 5m 떨어진 위치에서 각각 75단어에 대해 1회 발성하여 사무실 잡음환경에서 225 발화, 청소로봇 잡음환경에서 227 발화로 구성된 평가 DB를 구성하였다.

음성인터페이스 기반 Call-and-Come 서비스의 평가를 위하여, 가정용 로봇에 대한 호출음 20개에 대한 단어리스트를 선정하였다. 마이크로부터 각각 1, 2, 3, 5m 떨어진 위치에서 녹음하였으며, 각 20개 단어에 대해 다섯 번씩 발성하였다. 이 때 화자는 학습 및 적응에 참가하지 않은 남성 1명으로 선정하였다. 녹음환경은 사무실 잡음환경과 청소로봇 잡음환경으로 구성하였다. 또한, 잡음신호는 2m를 기준으로 7 dB의 신호 대 잡음 비 (SNR)를 갖도록 조정하였다.

5.2 Teager 기반 음성구간검출

본 시스템의 음성구간검출 알고리즘의 성능은 ETSI advanced front-end[9] 에서 제공되는 음성구간검출 알고리즘의 성능과 비교한다. 각각의 음성구간검출 알고리즘을 통해 추출된 음성구간에 대해 PBW 단어리스트 평가 DB를 이용한 offline 인식실험 통해 성능을 비교한다. 표 1은 음성구간검출 알고리즘을 통해 추출된 음성구간에 대한 인식률을 보여준다. ETSI advanced front-end에서의 음성구간검출 알고리즘을 사용하였을 때 평균 인식률이 79.7%인 반면,

표 1. Teager 기반 음성구간 검출에 대한 인식률(%).

음성구간검출 알고리즘 / 녹음환경	ETSI 음성구간 검출	Teager energy 기반 음성구간검출
사무실 잡음환경	88.9	92.4
청소로봇 잡음환경	70.5	69.6
평균	79.7	81.0

표 2. 음향모델 적응에 의한 인식률 (%)

녹음환경/발성거리 / 음향모델	사무실 잡음환경				청소로봇 잡음환경			
	1m	3m	5m	평균	1m	3m	5m	평균
Baseline	93.8	100	92.5	96.6	64.0	45.3	24.7	39.3
적응된 음향모델	98.7	96.0	96.0	96.9	88.0	84.0	58.7	67.3

표 3. 음성인터페이스 기반 Call-and-Come 서비스에서의 로봇 호출음 등록에 대한 인식률 (%)

발성 거리 / 음향모델	1m	2m	3m	5m	평균
사무실 잡음환경	100.0	98.0.0	100.0	98.0	99.0
청소로봇 잡음환경	96.0	99.0	99.0	96.0	97.5
평균	98.0	98.5	99.5	97.0	98.3

Teager energy 기반 음성구간검출 알고리즘을 사용한 경우 81.0%로 인식률이 향상됨을 알 수 있다.

5.3 음향모델 성능 평가

음향모델 적응을 통한 인식성능 평가는 적응 전 baseline 음향모델과 적응 후 음향모델에 대하여, 사무실환경과 청소로봇 잡음환경에서 녹음된 PBW 단어리스트 평가 DB를 이용하여 offline으로 수행하였으며, 그 결과를 표 2에 정리하였다. Baseline 음향모델을 이용했을 때 평균 인식률이 사무실 잡음환경과 청소로봇 잡음환경에서 각각 96.6%와 39.3%인데 반해, 적응시킨 음향모델을 사용하였을 때 각각 96.9%와 67.3%로 인식률이 향상됨을 볼 수 있다.

5.4 등록된 로봇의 호출음에 대한 성능 평가

음성인터페이스 기반 Call-and-Come 서비스에서 로봇 호출음에 대한 인식 성능은 20단어 평가 DB를 이용한 online 인식실험을 통해 획득하였다. 즉, 20 단어 호출음을 등록하고, 등록된 호출음을 인식하였으며, 그 결과를 발성거리와 잡음환경에 따라서 표 3에 정리하였다. 표 3에서 보는 바와 같이 사무실 잡음환경과 청소로봇 잡음환경에서 각각 99.0%와 97.5%의 인식률을 보였으며, 98.3%의 평균 인식률을 획득함을 알 수 있다.

6. 결론

본 논문에서는 Call-and-Come 서비스를 제공하는 가정용 로봇의 호출음 등록 및 인식 시스템을 제안하였다. 제안된 시스템에서는 사용자가 임의의 로봇 호출음을 음성을 통하여 효율적으로 등록하기 위해 monophone 기반 음향모델을 이용하여 탐색 범위를 줄이고, 줄어든 탐색 범위 내에서 triphone 기반 음향모델을 이용하여 정확하게 호출음을 등록하였다. 또한, 잘못된 호출음의 인식을 줄이기 위해 등록 중인 호출음에 대한 발화 검증 파라미터를 구하였다. 더불어, 원거리 음성에 대한 인식률 향상을 위해 원거리 음성 데이터베이스를 구축하여 음향모델을 적응시켰으며, 사용자의 방향을 예측하여 로봇이 사용자에게 다가갈 수 있는 방법을 제공하였다. 로봇 호출음 등록 및 인식 실험 결과에서는 98.3%의 인식률을 보였다.

감사의 글

이 논문은 삼성종합기술원과 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2007-314-D00245).

참고문헌

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

[2] A. Lee, T. Kawahara, and S. Doshita, "An efficient two-pass search algorithm using word trellis index," in *Proc. ICSLP*, pp. 1831-1834, Nov. 1998.

[3] M. Novak, R. Hampl, P. Krbec, V. Bergl, and J. Sediw, "Two-pass search strategy for large list recognition on embedded speech recognition platforms," in *Proc. ICASSP*, pp. 6-10, April, 2003.

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

[5] 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.

[6] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Terminal Journal*, vol. 54, no. 2, pp. 297-315, Feb, 1975.

[7] 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.

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

[9] ETSI ES 202 050.1.1.3, *Seech Processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*, Nov. 2003.