지능형 서비스 로봇 환경에서의 화자 인식 연구

*반규대, 곽근창, 정연구
과학기술연합대학원대학교, ETRI 지능형 로봇 연구단
e-mail : *kdban@etri.re.kr, kwak@etri.re.kr, ykchung@etri.re.kr*

Speaker Recognition in the Intelligent Service Robot

*Kyu-Dae Ban, Keun-Chang Kwak, Yun-Koo Chung
University of Science & Technology,
Intelligent Robot Research Division, ETRI

Abstract

Speaker Recognition for the Intelligent Service Robot is implemented in this paper. For this purpose, we perform speaker recognition based on Gaussian Mixture Model(GMM) and use robot platform called WEVER, which is a Ubiquitous Robotic Companion(URC) intelligent service robot developed at Intelligent Robot Research Division in ETRI. The experimental results reveals that the approach presented in this paper yields a good identification (89.00%) performance within 2 meter distance.

I. 서론

지능형 서비스 로봇은 조명이나 잡음, 거리변화 등의 매우 그 변화가 큰 사용 환경에 위치하고 있으므로 연구하기가 쉽지 않다. 또한 일반 사용자가 기본적으로 생각하는 로봇의 능력과 사회적인 관심에 비해 실제 서비스를 제공해 줄 수 있는 영역은 그리 넓지 않은 상태이다. 화자인식은 호출한 사람의 음성에서 특징을 추출한 후, 각 화자의 특징을 모델링하여, 저장된 데이터베이스와 비교한 후, 호출한 사람이 누구인지 인식하는 음성 신호처리 기술 중의 하나이다. 지능형 로봇이 호출한 사람이 누구인지 알 수 있다는 것은 그 사용자에게 다양한 서비스를 제공해 주기 위해 필수적인 요소이다. 그러나 지능형 로봇이 위치하는 환경이 가정이나 사무실 혹은 공공장소라고 가정한다면, 그 환경은 로봇의 자체 잡음이나, 주위 환경의 소음, 로봇과 호출한 사람과의 거리변화 등으로 인해 매우 불안정한 상태라고 할 수 있다. 본 논문에서는 전통적인 MFCC-GMM기반의 화자인식 방법을 지능형 로봇상황에 적용하여 본다.

II. 본론

2.1 Mel-Frequency Cepstrum Coefficient

음성인식에서 사용되는 특징은 Linear predictive coding(LPC) cepstrum, Mel - frequency Cepstrum coefficient (MFCC)등 이다. 화자인식 역시 이러한 특징들을 사용하여 발성한 화자를 인식한다. 본 논문에서 사용한 MFCC 기법에서는 먼저 발성한 음성을 hamming window를 씌운 프레임으로 나눈다. 프레임 단위의 음성신호는 FFT를 이용하여 주파수 영역으로 변환 시킨 뒤 주파수 대역에 따라 필터뱅크로 나누어 각 뱅크에서의 에너지를 취한다. 밴드 에너지에 로그

를 취한 후 Discrete Cosine transform(DCT)를 하여 최종적으로 MFCC 값을 얻게 된다.

2.2 Gaussian Mixture Model

앞서 설명한 MFCC에 의해 추출된 특징들은 GMM을 통하여 분류되게 된다. 본 논문에서 사용하는 화자인식 시스템은 가정의 맞춤형 서비스를 목표로 하고 있기 때문에 문맥 독립형 화자 인식기를 구축하였다. GMM은 여러 개의 가우시안 분포를 가중치의 곱을 통해 혼합하여 임의의 비선형 확률분포를 표현할 수 있도록 한 방법이다. GMM의 혼합계수를 늘려가면서 최적의 화자 확률분포 모델을 선택할 수 있다. 음성 특징을 x라고 할 때 각 가우시안 분포를 $g_i(x)$, 가중치를 w_i로 하면 M개의 Gaussian Mixture를 통해 음성의 분포 $P(x)$를 추정할 수 있고 그 식은 아래와 같다.

$$p(\boldsymbol{x}) = \sum_{i=1}^{M} w_i \, g_i(\boldsymbol{x})$$

GMM의 각 파라미터들은 Expectation-Maximization (EM) 알고리즘에 의해 얻을 수 있다.

Ⅲ. 실험

본 논문에서는 전자통신연구원에서 개발한 MIM-Board를 탑재한 Wever에서 데이터를 취득하였다. 실험에서 사용한 데이터 베이스는 10명의 화자가 발성한 음성을 기반으로 제작되어졌다. 실험은 가정에서의 서비스 로봇을 고려하여, 로봇과 발화자의 거리(0.5m, 1m, 2m)에 따라 3세트로 구성하였고, 한 세트에서 10명의 사람에 대해 각각 20문장씩 데이터를 획득하였다. 그 중 0.5m에서 얻은 데이터를 등록용으로, 2m,3m에서 얻은 음성 데이터를 테스트용으로 사용하였다. 각 문장의 길이는 약 2~3초이다.

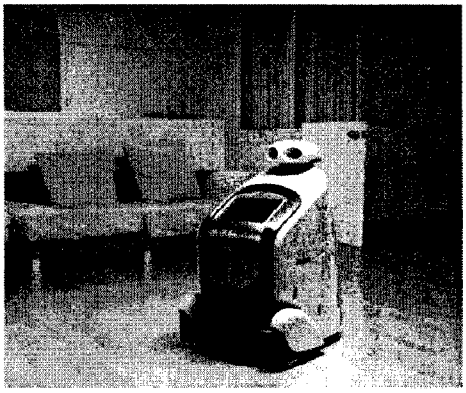

그림 1. DB구축에 사용된 로봇(WEVER)

		로봇과 발화자 간의 거리	
		1m	2m
Gaussian Mixture 개수	16	88.5	72.0
	32	89.0	70.5
	64	89.0	68.0
	128	89.0	67.0

표1. 거리와 Gaussian Mixture 개수에 따른 인식 결과(%)

표 1에 실험결과를 나타내었다. 실험결과를 참조하면 로봇과 발화자간의 거리가 멀어질수록 인식률이 낮아짐을 알 수 있다. 또한 Gaussian Mixture의 개수를 증가시킨다고 해서 인식률의 증가를 기대하는 것은 무리라는 결과를 얻을 수 있다. 2m에서 얻어진 테스트 DB에서는 Gaussian Mixture의 개수가 증가할수록 인식률이 떨어짐을 알 수 있다.

Ⅳ. 결론 및 향후 연구 방향

지능형 서비스 로봇이 서비스를 제공하기 위한 방법 중 화자인식은 필수적인 요소이다. 논문에서는 가정환경과 같은 실험실에 위치한 로봇에서 학습, 테스트 DB를 구축하여 로봇에서의 화자인식에 관하여 살펴보았다. 특징 벡터로서 MFCC를 이용하였으며, 얻어진 특징벡터들을 GMM을 통하여 모델링하고, 인식률을 얻었다. 전처리를 통한 음성파형의 변화는 오히려 인식률을 떨어트리는 결과를 나타내어 논문에서 제외하였으나, 이는 MFCC-GMM방법이 로봇상황에서 적합한 화자인식 기법 중 하나임을 알 수 있게 한다.

참고문헌

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. on Speech and Audio Processing, vol. 3, no. 1, pp. 72 - 83. 1995.

[2] J. P. Campbell, "Speaker recognition: a tutorial," Proceedings of IEEE, vol. 85, no. 9, pp. 1437-1462, 1997.

[3] 정현열, "음성을 이용한 화자인식 시스템 기술의 현황과 전망", 정보과학회지 제 19 권, 제 7 호 pp32-44 2001.