지능형 서비스 로봇을 위한 문맥독립 화자인식 시스템

Context-Independent Speaker Recognition in URC Environment

지 미 경[1] · 김 성 탁[2] · 김 회 린[3]

Mikyong Ji[1] · Sungtak Kim[2] · Hoirin Kim[3]

Abstract This paper presents a speaker recognition system intended for use in human-robot interaction. The proposed speaker recognition system can achieve significantly high performance in the Ubiquitous Robot Companion (URC) environment. The URC concept is a scenario in which a robot is connected to a server through a broadband connection allowing functions to be performed on the server side, thereby minimizing the stand-alone function significantly and reducing the robot client cost. Instead of giving a robot (client) on-board cognitive capabilities, the sensing and processing work are outsourced to a central computer (server) connected to the high-speed Internet, with only the moving capability provided by the robot. Our aim is to enhance human-robot interaction by increasing the performance of speaker recognition with multiple microphones on the robot side in adverse distant-talking environments. Our speaker recognizer provides the URC project with a basic interface for human-robot interaction.

Keywords : Speaker Recognition, Speaker Identification, Speaker Verification

1. 서 론

로봇의 시각과 청각은 서로 보완적인 역할을 한다. 시각정보는 물체의 기하학적 정보를 정확히 전해주지만 계산량이 많고 조명 등의 환경 변화에 열악하다. 이와 반대로 청각정보는 사물의 움직임이나 인간의 행동에 따라 특이한 패턴을 내며, 어두운 곳이나 시야가 미치지 않는 곳 또는 원거리에서도 감지할 수 있다는 장점이 있다. 또한 인간이나 로봇에게 위험한 사건들은 굉음을 동반하는 경우가 많아 청각은 인간과 로봇의 자신 보존에도 중요한 역할을 한다. 화자인식은 로봇과 사용자 사이의 상호작용을 시작하는 가장 첫 번째 단계로써 사용자의 호출에 응답하도록 하는 것이 로봇의 가장 기본적인 기능을 구현하는 필수 기술이다.

화자인식은 입력음성으로부터 추출된 정보로부터 화자, 즉 말을 한 사람의 음성을 인식하는 것으로 크게 두 가지로 나뉘는데 등록화자 중 한 사람으로 인식하는 화자식별 (Speaker Identification) 기술과 제시된 화자가 입력음성을 발성했는지를 승인하거나 거절하는 화자검증 (Speaker Verification) 기술로 나눌 수 있다. 또한, 인식대상이 되는 음성의 발성방법에 따라 문맥 종속형 (text-dependent) 과 문맥 독립형 (text-independent), 그리고 문맥 제시형 (text-prompted) 으로 나뉘어진다 [1]. 화자가 발성할 문장이 미리 정해져 있을 때 문맥 종속형이라 하고, 자유롭게 발성하는 경우를 문맥 독립형이라 하며 화자인식 시스템이 사용자에게 발성할 문장내용을 제시하는 방식이 문맥 제시형으로 녹취를 효과적으로 막을 수 있는 대안으로 부각되고 있다.

본 논문에서는 지능형 서비스 로봇의 가장 기본적인 사용자 인터페이스의 핵심기술인 화자인식 기술을 로봇에 적용하였다. 특히 잡음에 강인한 화자인식 시스템을 제안하고 4인 가족을 기준으로 하여 가정 내 잡음 환경과 로봇 자체 소음을 고려하여 3m 이내의 거리에서의 문맥 독립형 화자인식 기술을 적용하여 로봇 사용자 개개인에게 적합한 서비스를 제공하고자 한다. 2장에서는 지능형 로봇을 위한 기반 화자시스템에 대한 설명과 화자식별 및 제안된 화자검증 방법에 관해 얘기하고 3장에서는 로봇 환경에서의 실험결과를 제시할 것이다.

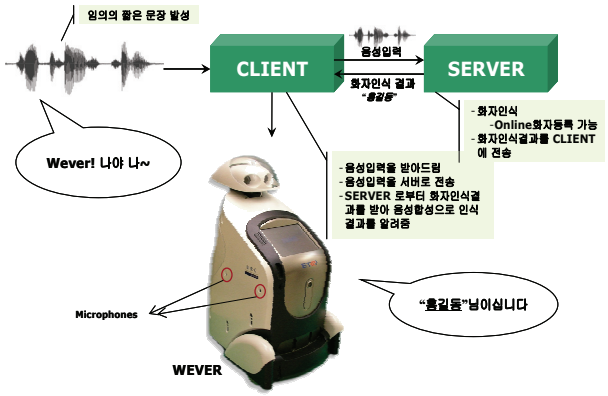
※ 본 연구는 정보통신부의 URC 프로젝트 지원사업의 연구결과로 수행되었음.

[1] 한국정보통신대학교 공학부 박사과정
[2] 한국정보통신대학교 공학부 박사과정
[3] 한국정보통신대학교 공학부 부교수

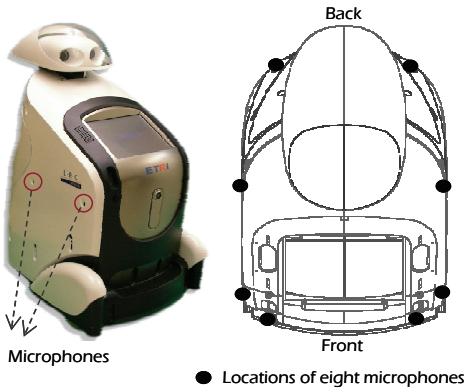
2. 지능형 로봇을 위한 화자인식 시스템

2.1 기반 화자인식 시스템

지능형 서비스 로봇을 위한 화자인식 과정은 [그림 1] 과 같다. 지능형 서비스 로봇 (웨버) 은 간단한 음성 인터페이스를 통해 사용자의 음성입력을 받아드리고 이를 서버에 전송한다. 서버에서는 상대적으로 계산량이 많은 화자인식 과정을 거쳐 인식결과를 웨버에게 전송하고 웨버는 인식결과를 받아 음성합성을 통해 화자인식 결과를 알려준다. 입력음성을 받아들이기 위한 웨버 내의 마이크 배열을 [그림 2]에 표시하였다.



[그림 1] 지능형 로봇을 위한 문맥독립 화자식별



[그림 2] 로봇의 마이크 배열

2.2 GMM 기반의 화자인식 방법

2.2.1 GMM 기반의 화자식별

화자를 특징벡터를 출력하는 랜덤소스로 가정하고 이 화자모델 안에는 화자의 성도의 특성을 나타내는 숨겨진 상태 (state) 들이 있다. 이 랜덤소스가 특정한 상태에 있을 때 특정한 성도 특성에 해당하는 특징벡터를 출력하게 된다. 각각의 상태는 특징벡터의 평균 μ_i와 공분산 Σ_i를 가지고 다차원 Gaussian 확률 분포에 의하여 특징벡터를 출력한다. 따라서 상태 i의 D차원 확률분포 함수 PDF 는 식 [1]과 같다.

$$b_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\} \quad (1)$$

따라서 화자모델 λ_s의 Gaussian Mixture Model (GMM) 은 식 [2]와 같이 표현된다. 결국 화자모델 λ_s는 식 [4]의 파라미터의 집합으로 표현된다 [2]-[4].

$$P(x|\lambda_s) = \sum_{i=1}^{M} c_i b_i(x) \quad (2)$$

$$\sum_{i=1}^{M} c_i = 1 \quad (3)$$

$$\lambda_s = \{c_i, \mu_i, \Sigma_i\} \quad (4)$$

2.2.2 GMM-UBM 기반의 화자검증

입력음성 X에 대해 제시된 화자의 모델이 λ_c이고, 제시된 화자가 아닌 모델 (Universal Background Model) 이 $\lambda_{c'}$ 라면, 유사도 비는 식 [5]와 같이 표현된다.

$$\frac{\Pr(X\text{가 제시된 화자의 음성})}{\Pr(X\text{가 제시된 화자이외의 음성})} = \frac{p(\lambda_c|X)}{p(\lambda_{c'}|X)} \quad (5)$$

여기에 Bayes' rule를 적용하고 사전 확률 ($P(\lambda_c)$ = $P(\lambda_{c'})$ = *constant*) 이 같다고 가정하면 로그영역의 유사도 비는 식 [6]과 같이 표현할 수 있다.

$$R(X) = \log P(\lambda_c|X) - \log P(\lambda_{c'}|X) \quad (6)$$

즉 이 유사도 비를 미리 정한 임계값, θ 와 비교해 제시된 화자를 승인할지 ($R(X)>\theta$) 또는 거부할지를 ($R(X)<\theta$) 결정한다.

2.2.3 SCR 기반의 화자검증

[그림 3]과 같은 입력음성에서 묵음구간을 제외한 구간에서만 특징 벡터열 $x_1, x_2, \ldots, x_N$ 을 추출한다고 가정하자. i번째 프레임의 특징벡터만을 고려했을 때 최대

유사도를 나타내는 화자모델 S_i^{max} 는 식 [7]과 같이 정의된다.

$$S_i^{\max} = \arg\max_S P(x_i \mid S) \tag{7}$$

$$S^c = \arg\max_S P(x_1, \cdots, x_N \mid S) \tag{8}$$

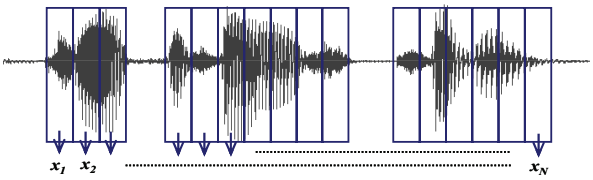
식 [8]과 같이 입력음성의 모든 프레임에 대하여 최대 유사도를 나타내는 화자 S^c에 대한 혼잡도, SCR (Speaker Confusion Rate) 는 식 [9]과 같다 [5].

$$R^c = \frac{1}{N}\sum_{i=1}^{N} d_f(S_i^{\max}, S^c) \tag{9}$$

식 [9]에서 거리함수 $d_f(S_i^{max}, S^c)$는 식 [10]과 같다.

$$d_f(S_i^{\max}, S^c) = \begin{cases} 1, \text{if } S_i^{\max} = S^c \\ 0, \text{ otherwise} \end{cases} \tag{10}$$

이 값을 미리 정의한 임계치, θ와 비교하여 화자를 승인할지 ($R^c>\theta$) 또는 거부할지를 ($R^c<\theta$) 결정한다.



[그림 3] 입력음성의 특징벡터

3. 실험 및 결과

기본적으로 화자식별/검증 성능을 평가하기 위해 일반적으로 사용되는 화자식별률과 Equal Error Rate (EER)을 사용하였다. 로봇환경에서 16명으로부터 30 문장을 두 번씩 발성하여 로봇으로부터 직접 7채널을 동시에 이용하여 DB를 수집하였으며, 이때 화자와 로봇의 거리는 약 1m 정이며, 비교적 조용한 상태에서 음성을 녹음하였다. 또한 보조적으로 Electronics and Telecommunications Research Institute (ETRI) 음성정보연구센터에서 배포하고 있는 화자인식용 DB를 이용하여 실험을 수행하였다. Mel Frequency Cepstral Coefficients (MFCC)를 화자인식용 특징으로 사용하였다.

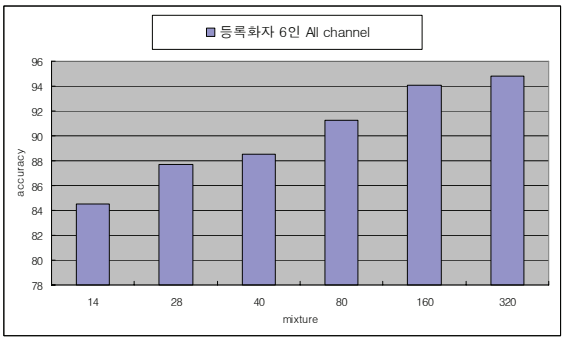
[표 1]은 문맥독립 화자인식을 수행한 결과 Gaussian mixture의 수에 따른 채널 별 화자식별 정확도를 보여주고 있다. 이때의 등록화자는 16명의 화자 중 남녀 각 2인씩 총 4인을 무작위로 선택하여 화자식별을 수행하였다. 훈련데이터는 7채널 모두 사용하였다. 실험결과에서 보는 바와 같이 마이크나 잡음의 위치, 거리, 마이크 특성 등에 따라 특정채널이 나머지 채널에 비해 우수한 인식률을 보였으며, 이는 효율적인 음성 데이터 선택을 통해 전체 인식률을 향상 시킬 수 있는 근거를 제시한다. 실험에서 사용된 데이터는 로봇을 정면으로 바라보고 발성했을 경우이며 실험결과와 같이 최소 mixture 수는 160개 이상이 적합하다고 판단된다.

추가적으로 등록화자를 4인에서 6인으로 확장하였을 때의 화자식별의 mixture 별 평균 정확도의 조사하였고 결과는 [그림 4] 와 같다. 그 결과는 [표 1]과 유사하나 전체적으로 약간 성능이 감소하였다.

잡음에 강인한 문맥독립 화자검증의 정확한 성능측정을 위하여 ETRI 음성정보연구센터에서 배포하는 한국어 화자인식용 DB를 이용하여 화자검증 알고리즘의 성능을 측정하였다. 전체 등록화자는 40명이라 가정하고 각 화자모델은 320개의 Gaussian mixture 를 사용하여 훈련하였으며 UBM 도 마찬가지로 320개의 Gaussian

[표 1] 채널별 화자식별률(%) (등록화자 4인)

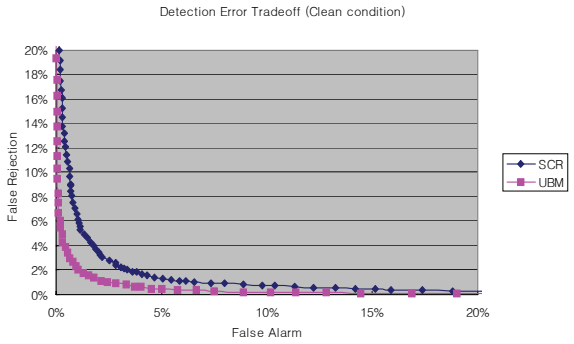
Mixture 수	CH1	CH2	CH3	CH4	CH5	CH6	CH7
14	83.3	91.7	94.4	83.3	100	100	91.7
28	86.1	91.7	100	91.7	100	100	100
40	86.1	94.4	100	94.4	97.2	100	100
80	86.1	97.2	100	97.2	100	100	97.2
160	91.7	100	100	100	97.2	100	100
320	94.4	100	100	100	97.2	100	100



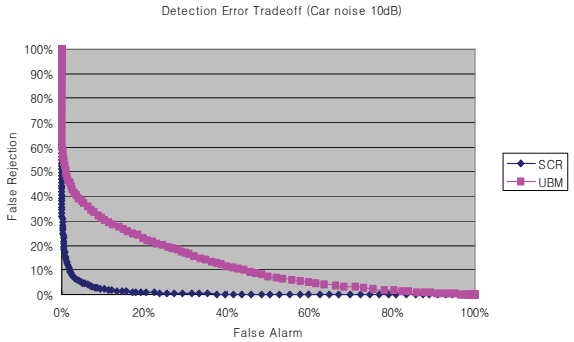
[그림 4] Mixture 별 평균 화자식별률(%) (등록화자 6인)

mixture를 사용하였다. 화자당 훈련에 사용된 문장은 1,280개 이고 테스트를 위하여 화자당 800개의 문장으로 테스트하였다. 또한 자동차 잡음과 로봇에서 직접 수집한 박수소리를 더하여 성능평가에 사용하였다.

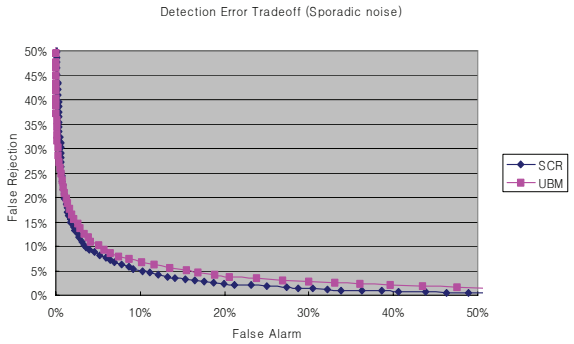
[그림 5]에서 보는 바와 같이, clean환경에서는 제안된 SCR기반 방법보다는 GMM-UBM방법이 성능이 우수함을 나타낸다. EER은 GMM-UBM방법이 약 1.5%였으며 제안된 SCR방법은 약 2.5%였다. 제안된 방법이 성능이 약간 저하되긴 하지만, 전체적인 성능은 두 가지 방법 모두 우수함을 볼 수 있다.



[그림 5] 화자검증 성능 비교 (Clean)

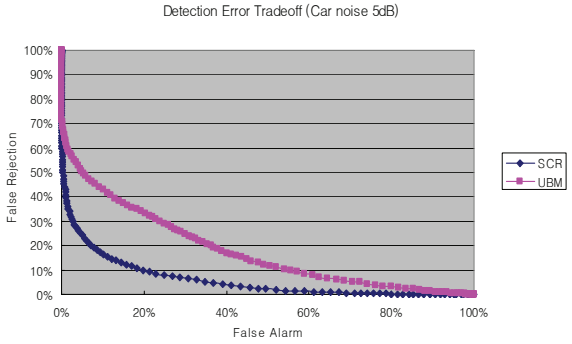


[그림 6] 화자검증 성능 비교 (자동차 잡음: 10dB)



[그림 7] 화자검증 성능 비교 (박수소리: 10dB)

[그림 6]와 [그림 7]은 자동차 잡음과 박수소리를 섞어서 신호 대 잡음 비를 10dB로 만들었을 경우의 화자검증의 성능을 보여준다. 잡음이 부가됨에 따라 기존의 GMM-UBM 방법은 성능이 급격히 저하됨을 볼 수 있지만, 제안된 SCR기반의 방법은 성능저하가 적다. 즉, ordering characteristic 때문에 잡음에 의한 영향이 적다. 마찬가지로 [그림 8]은 신호 대 잡음 비가 5dB일 경우의 화자검증의 성능을 나타낸다.



[그림 8] 화자검증 성능 비교 (자동차 잡음: 5dB)

4. 결 론

본 논문에서는 가정용 지능형 서비스 로봇의 사용자 인터페이스의 핵심 기술인 화자인식 기술을 적용하였다. 최근 가장 높은 성능을 보여주며 가장 많이 사용하고 있는 화자 식별 및 화자검증 알고리즘인 GMM기반의 화자식별과 GMM-UBM기반의 화자검증[6]을 구현하여 로봇환경에서의 성능을 조사하였고, 채널별 성능을 조사하였다. 또한 잡음에 강인한 특성을 보여주는 SCR 기반의 화자검증 기법의 성능을 평가하였다. 실험결과, 계산량이 크게 증가하지 않으면서 잡음에 강인한 SCR기반의 화자검증 방법은 잡음환경에서 매우 우수한 결과를 보여주었다. 채널 별 화자인식 결과에서 보는 바와 같이 화자 또는 잡음의 위치, 거리, 마이크 특성 등에 따라 특정 채널이 다른 채널에 비해 더 나은 성능을 보였다. 따라서 앞으로 효율적인 채널선택 방법을 모색하여 전체적인 화자인식 성능을 높일 것이다. 또한 가장 시급하게 해결해야 할 과제인 잡음환경에 대한 대처방안에 대해서도 연구할 것이다. 한편, 화자 등록을 사용자 편의에 의해 빠르게 하면서 화자인식률의 정확도를 가능한 높일 수 있는 화자적응에 의한 등록방법도 연구중에 있다.

참고문헌

[1] P. Joseph, and Jr. Campbell, "Speaker Recognition: A Tutorial," Proc. of the IEEE, Vol. 85, No. 9, pp. 1437-1462, Sept. 1997.

[2] D. Reynolds and R.C. Rose, "Robust Text Independent Speaker Identification Using Gaussian Mixture Speaker Models", Proc. IEEE Trans. Speech and Audio Processing, vol. 3, pp. 72-83, Jan. 1995.

[3] B., Narayanaswamy, and Gangadharaiah, R., "Extracting additional information from Gaussian mixture model probabilities for improved text-independent speaker identification," ICASSP, Vol. 1, pp.621-624, Mar. 2005.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Model," Digital Signal Processing, Vol. 10, No. 1-3, pp. 19-41, 2000.

[5] Kyuhong Kim, Hoirin Kim, and Minsoo Hahn, "Utterance Verification Using Search Confusion Rate and Its N-Best Approach," ETRI Journal, Vol. 27, pp. 461-464, Aug. 2005.

[6] B. Tseng, F. Soong, and A. Rosenberg, "Continuous Probabilistic Acoustic Map for Speaker Recognition," Proc. ICASSP, Vol. 3, pp. 161-164, 1992.

지 미 경

2000 한성대학교 정보공학과(공학사)
2002 한국정보통신대학교 공학부 음성인식전공(공학석사)
2002~현재 한국정보통신대학교 공학부 박사과정
관심분야: 음성인식, 화자인식

김 성 탁

2000 울산대학교 전자공학과(공학사)
2003 한국정보통신대학교공학부 음성인식전공(공학석사)
2003~현재 한국정보통신대학교 공학부 박사과정
관심분야: 음성인식, 화자인식

김 회 린

1987 한국과학기술연구원 전자공학과(공학석사)
1992 한국과학기술연구원 전자공학과(공학박사)
1995 일본 ATR-ITL 방문연구원
~1999 ETRI 선임연구원
2001~현재 한국정보통신대학교 공학부 부교수
관심분야: 음성인식, 화자인식, 음향코딩