지능형 서비스 로봇을 위한 잡음에 강인한 문맥독립 화자식별 시스템

Noise Robust Text-Independent Speaker Identification for Ubiquitous Robot Companion

김성탁, Sungtak Kim*, 지미경, Mikyoung Ji*, 김회린, Hoirin Kim*,
김혜진, Hye-Jin Kim**, 윤호섭, Ho-Sub Yoon**
*한국정보통신대학교 공학부
**한국전자통신연구원 지능형로봇연구단

요약 본 논문은 지능형 서비스 로봇의 여러 기술들 중에서 기본적인 기술인 화자식별 기술에 관한 내용이다. 화자식별 기술은 화자의 음성신호를 이용하여 등록된 화자들 중에서 가장 유사한 화자를 찾아 내는 것이다. 기존의 mel-frequency cepstral coefficient 를 이용한 화자식별 시스템은 무잡음 환경에서는 높은 성능을 보장하지만 잡음환경에서는 성능이 급격하게 떨어진다. 이렇게 잡음환경에서 성능이 떨어지는 요인은 등록환경과 식별환경이 다른 불일치문제 때문이다. 본 논문에서는 불일치문제를 해결하기 위해 relative autocorrelation sequence mel-frequencycepstral coefficient 를 사용하였다. 또한, 기존의 relative autocorrelation sequence mel-frequency cepstral coefficient 의 제한된 정보문제와 잔여잡음문제를 해결하기 위해 멀티스트리밍 방법과 멀티스트리밍 방법에 특징벡터 재결합 방법을 결합한 하이브리드 방법을 제한 하였다. 실험결과 제한된 방법들이 기존의 특징벡터보다 잡음환경에서 높은 화자식별 성능을 보여주었다.

Abstract This paper presents a speaker identification technique which is one of the basic techniques of the ubiquitous robot companion. Though the conventional mel-frequency cepstral coefficients guarantee high performance of speaker identification in clean condition, the performance is degraded dramatically in noise condition. To overcome this problem, we employed the relative autocorrelation sequence mel-frequency cepstral coefficient which is one of the noise robust features. However, there are two problems in relative autocorrelation sequence mel-frequency cepstral coefficient: 1) the limited information problem. 2) the residual noise problem. In this paper, to deal with these drawbacks, we propose a multi-streaming method for the limited information problem and a hybrid method for the residual noise problem. To evaluate proposed methods, noisy speech is used in which air conditioner noise, classic music, and vacuum noise are artificially added. Through experiments, proposed methods provide better performance of speaker identification than the conventional methods.

핵심어: *Ubiquitous robot companion, Speaker Identification, Mel-frequency cepstral coefficient, Relative autocorrelation sequence mel-frequency cepstral coefficient, Multi-streaming method, Hybrid method*

1. 서론

본 논문에서는 지능형 서비스 로봇의 여러 기술 중에서 기본적인 기술인 화자식별기술에 관한 내용이다. 화자식별기술은 화자의 음성정보를 이용하여 등록된 화자들 중에서 가장 유사한 화자를 찾아내는 기술이다. 최근에는 가우시안 혼합모델 (Gaussian mixture model)을 이용한 문맥독립 화자식별 기술 [1]이 주된 추세이다. 화자모델링을 위한 특징벡터로는 MFCC (Mel-Frequency Cepstral Coefficient)를 많이 사용한다. 기존의 MFCC 를 특징벡터로 이용한 화자식별 시스템은 무잡음 환경에서는 높은 성능을 보장하지만, 잡음환경에서는 성능이 급격히 떨어진다. 이렇게 잡음환경에서 성능이 떨어지는 요인은 등록환경과 식별환경이 다름에서 오는 불일치 문제 때문이다. 이와 같은 불일치 문제를 해결하기 위해 많은 기술들이 연구되고 있다. 본 논문에서는 불일치 문제를 특징벡터차원에서 해결하기 위해 제안된 RAS-MFCC (Relative Autocorrelation Sequence-MFCC)[2]를 이용하였다. RAS-MFCC 는 자기상관 영역에서 잡음이 정형적 (stationary)이라는 가정과 시간필터링 (temporal filtering)을 이용하여 얻는다. 비록 RAS-MFCC 가 기존의 MFCC 보다 잡음에 강인한 특징벡터이지만, 두 가지 문제점이 있다: 1) 제한된 정보문제, RAS-MFCC 를 구할 때 시간필터링된 신호만 이용한다. 2) 잔여잡음 문제, 실제 잡음은 정형적이지 않으므로 시간필터링 후에도 잡음이 존재한다. 본 논문에서는 제한된 정보문제를 해결하기 위해 멀티스트리밍 방법을 이용하여 시간 필터링하지 않은 자기상관신호로부터 구한 AS-MFCC (Autocorrelation Sequence-MFCC)를 RAS-MFCC 와 같이 사용하였다. 또한 RAS-MFCC 의 잔여잡음 문제를 해결하기 위해 다중밴드 (multi-band)방법중의 하나인 특징벡터 재결합 (Feature recombination)방법을 멀티스트리밍 방법과 결합한 하이브리드 방법을 사용하였다.

2. 멀티스트리밍 기반의 화자식별 시스템

RAS 는 자기상관영역에서 음성신호를 대신해서 사용할 수 있는 잡음에 강인한 신호이다. RAS 를 구하는 방법은 아래와 같다. 무잡음 음성, $x(m,n)$ 이 부가잡음, $w(m,n)$ 으로 왜곡이 되면 잡음음성, $y(m,n)$ 은 식 (1)과 같이 표현된다.

$$y(m,n) = x(m,n) + w(m,n) \quad 0 \le m \le M-1, 0 \le n \le N-1 \tag{1}$$

여기서 m 은 프레임 인덱스를 나타내고, n 은 한 프레임내의 시간 인덱스를 나타낸다. M 과 N 은 발성 (utterance)내의 프레임 수와 프레임내의 샘플 수를 나타낸다. 만약 자기상관영역에서 음성과 잡음이 상관성이 없다고 가정하면 잡음음성을 자기상관영역에서도 식 (2)와 같이 음성과 잡음의 합으로 나타낼 수 있다.

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{ww}(m,k) \quad 0 \le m \le M-1, 0 \le k \le N-1 \tag{2}$$

여기서 $r_{yy}(m,k)$, $r_{xx}(m,k)$ 그리고 $r_{ww}(m,k)$ 는 각각 잡음 음성, 무잡음 음성 그리고 잡음의 자기상관계수를 나타내고, k 는 자기상관 인덱스이다. 잡음이 정형이라고 가정하면 식 (2)에서 잡음의 자기상관계수 $r_{ww}(m,k)$ 에서 인덱스 m 을 제거할 수 있다.

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{ww}(k) \quad 0 \le m \le M-1, 0 \le k \le N-1 \tag{3}$$

식 (3)의 양변에 시간 필터링을 취하면 아래 식 (4)와 같다.

$$\begin{aligned} &\Delta r_{yy}(m,k) \\ &= r_{yy}(m+1,k) - r_{yy}(m-1,k) \\ &= r_{xx}(m+1,k) + r_{ww}(k) \\ &\quad - r_{xx}(m-1,k) - r_{ww}(k) \\ &= \Delta r_{xx}(m,k) \end{aligned} \tag{4}$$

여기서 $\Delta r_{xx}(m,k)$ 와 $\Delta r_{yy}(m,k)$ 는 무잡음 음성과 잡음 음성의 RAS 이다. 식 (4)에서와 같이 무잡음 음성과 잡음 음성의 RAS 가 같으므로 RAS 를 음성신호대신 사용하여 MFCC 를 구하면 기존의 MFCC 보다 잡음에 강인한 MFCC 를 구할 수 있다.

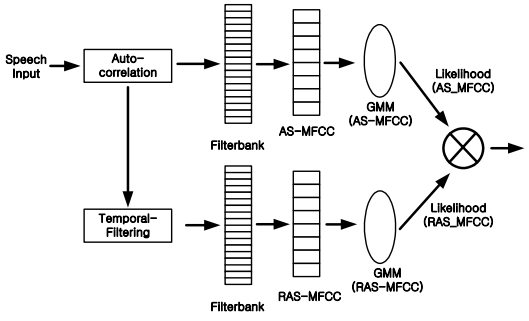

그림 1. 멀티스트리밍 기반 화자식별 시스템

하지만, 식 (4)에서와 같이 RAS-MFCC 는 시간 필터링된 신호만 이용하기 때문에 제한된 정보문제가 발생한다. 제한된 정보문제를 해결하기 위해 시간 본 논문에서는 멀티스트리밍 방법을 이용하여 필터링전의 신호인 자기상관신호를 이용해서 얻은 AS-MFCC 를 RAS-MFCC 와 같이 사용한다. 그림 1 은 제안된 멀티스트리밍 기반의 화자식별 시스템을 보여준다.

3. 하이브리드 기반의 화자식별 시스템

멀티스트리밍 방법을 이용해서 RAS-MFCC 가 가지고 있는 제한된 정보문제를 해결하였지만, RAS-MFCC 의 잔여 잡음과 AS-MFCC 의 잡음문제가 여전히 존재한다. 본 논문에서는 이런 잡음들이 각 서브밴드마다 주는 영향이 다른 점을 이용하기 위해 다중밴드 방법 [3][4]를 사용하였다. 그림 2 는 주파수 영역이 제한된 잡음환경에서 전체밴드를 이용하는 방법과 다중밴드를 이용하는 방법을 특징벡터관점에서 보여주고 있다. 그림에서와 같이 다중밴드방법을 이용하면 주파수영역이 제한된 잡음이 전체 특징벡터에 영향을 주는 것을 피할 수 있다.

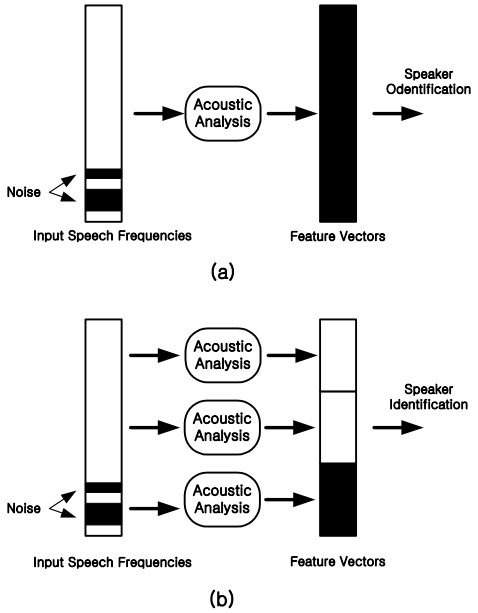

그림 2. 주파수 영역이 제한된 잡음환경에서의 화자식별
(a) 전체 주파수밴드를 이용한 경우
(b) 다중밴드 방법을 이용한 경우

다중밴드 방법에는 크게 유사도 재조합 방법(Likelihood recombination)과 특징벡터 재조합 방법(Feature recombination)방법으로 나누어진다. 특징벡터 재조합 방법은 유사도 재조합 방법과 달리 서브밴드간의 상관관계도 같이 모델링 할 수 있기 때문에 성능이 우수하다고 알려져 있다. 그래서 본 논문에서도 그림 4 와 같이 특징벡터 재결합 방법을 멀티스트리밍 시스템과 결합한 하이브리드 시스템을 제안하였다. 그림 3 은 다중밴드 방법을 보여준다.

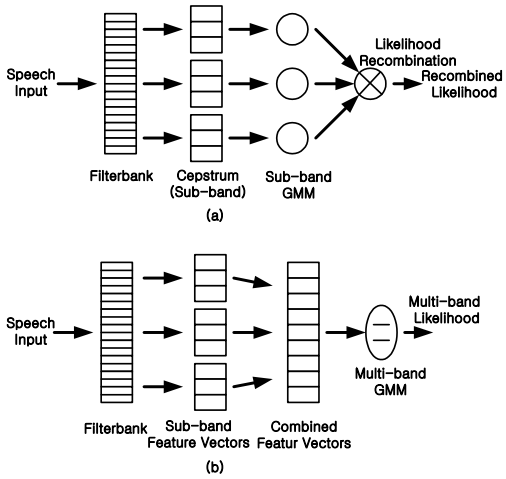

그림 3. 다중밴드 방법
(a) 유사도 재결합 방법
(b) 특징벡터 재결합 방법

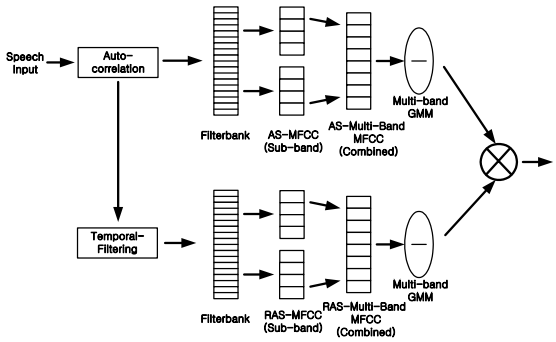

그림 4. 하이브리드 시스템

그림 5 는 특징벡터 재결합 방법에서 각 서브밴드 별로 MFCC 를 구하는 방법을 보여준다.

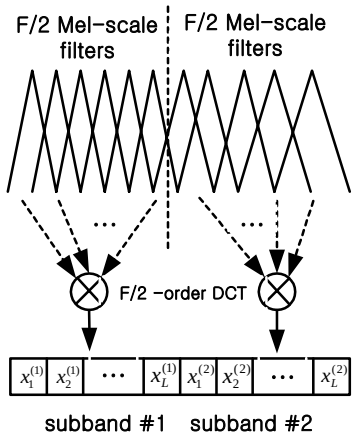

그림 5. 두 개의 서브밴드를 가지는 멀티밴드 시스템에서 MFCC 추출방법

I 개의 서브밴드와 F 개의 필터를 가지는 시스템에서 각 서브밴드 당 L 개의 MFCC 를 추출한다면, i 번째 서브밴드의 j 번째 MFCC 를 구하는 방법은 식 (5)와 같다.

$$x_j^{(i)} = \sqrt{\frac{2}{F/I}} \sum_{f=1}^{F/I} LFB_f^{(i)} \cos\left[(f-0.5) \frac{j\pi}{F/I} \right], \quad 1 \le j \le L \le \frac{F}{I} \tag{5}$$

여기서 $LFB_f^{(i)}$ 는 i 번째 서브밴드의 f 번째 필터에너지의 로그 값이다.

4. 지능형 서비스 로봇에서의 화자식별

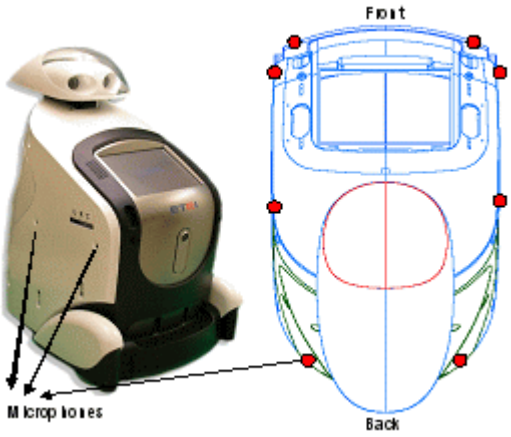

그림 6. 지능형 서비스 로봇의 마이크 배열

그림 6 은 지능형 서비스 로봇상의 마이크 배열을 보여준다. 그림과 같이 정면에 4 개, 옆면에 2 개, 후면에 2 개의 마이크가 배열되어 있다. 등록된 화자 그룹, $S = \{1,2,...,E\}$ 이 있고 c 번째 마이크의 테스트 특징벡터 열, $X_c = \{x_1^{(c)}, x_2^{(c)}, ..., x_{T_c}^{(c)}\}$ 이 주어 졌다면 8 개의 마이크를 가지는 지능형 서비스 로봇에서의 화자식별은 아래 식 (6)과 같다.

$$\hat{S} = \arg\max_{e \in S} \sum_{c=1}^{8} \left[\frac{1}{T_c} \sum_{t=1}^{T_c} \log(p(x_{t,AS}^{(c)} \mid \lambda_{e,AS})) + \frac{1}{T_c} \sum_{t=1}^{T_c} \log(p(x_{t,RAS}^{(c)} \mid \lambda_{e,RAS})) \right] \tag{6}$$

$x_{t,AS}^{(c)}$ 와 $x_{t,RAS}^{(c)}$ 는 c 번째 마이크로부터 나오는 음성신호를 이용해서 구한 AS-MFCC 와 RAS-MFCC 를 나타낸다. $\lambda_{e,AS}$ 와 $\lambda_{e,RAS}$ 는 AS-MFCC 와 RAS-MFCC 를 이용한 화자모델이다.

5. 화자식별 실험결과

지능형 서비스 로봇에서의 화자식별 성능을 알아보기 위해 30 명 (남자: 23 명, 여자: 7 명)이 30 문장을 2 회 발성하였다. 그리고 잡음환경에서의 화자식별을 위해 에어컨잡음, 클래식음악, 그리고 청소기잡음을 신호 대 잡음비가 약 10dB 정도로 인위적으로 무잡음 음성을 왜곡하였다. 화자모델을 훈련하기 위해 화자 당 5 문장으로 MAP (Maximum A Posteriori)방법을 이용하였다. 표 1 은 여러 가지 특징벡터들의 식별성능을 보여준다. 표 2 는 음성개선기술을 적용한 후, 특징벡터에 따른 화자식별 결과를 보여준다. 음성개선 기술은 MMSE-STSA (Minimum Mean Square Error Short Time Spectral Amplitude)방법을 적용하였다. 실험 결과 클래식음악 잡음환경에서는 RAS-MFCC 의 성능이 기존의 MFCC 에 비해 성능이 크게 떨어졌다. 클래식음악 잡음에서 성능이 떨어지는 이유는 RAS-MFCC 를 얻을 때 사용한 잡음이 정형적이라는 가정의 오류에 기인한 것으로 볼 수 있다. 하지만, 에어컨잡음이나 청소기잡음환경에서는 제안한 멀티스트리밍 방법과 하이브리드 방법 모두 기존의 MFCC 와 RAS-MFCC 에 비해 성능이 우수하였다.

표 1. 특징벡터에 따른 화자식별 성능 (%) [RAS-MFCC 대비 에러감소율]

Noise / Feature	AIRCON (10dB)	MUSIC (10dB)	VACUUM (10dB)
MFCC	75.9	87.9	32.8
RAS-MFCC	79.5	86.1	42.3
Multi-Streaming	79.6 [0.4]	83.4 [-19.4]	48.7 [11.1]
Hybrid	83.5 [19.5]	85.1 [-7.2]	56.2 [24.1]

표 2. 음성개선기술 적용 후, 특징벡터에 따른 화자식별 성능 (%) [RAS-MFCC 대비 에러감소율]

Noise / Feature	AIRCON (10dB)	MUSIC (10dB)	VACUUM (10dB)
SE+ MFCC	85.3	83.2	73.7
SE+ RAS-MFCC	91.1	76.8	80.7
SE+ Multi-Streaming	91.1 [0.0]	74.3 [-10.8]	83.2 [13.0]
SE+ Hybrid	91.3 [2.2]	76.7 [-0.4]	85.8 [26.4]

6. 결론

잡음환경에서 지능형 서비스 로봇의 기본적인 기능인 화자식별의 성능을 향상시키기 위해 잡음에 강인한 RAS-MFCC 를 사용하였다. 또한, 기존의 RAS-MFCC 의 성능을 개선하기 위해 멀티스트리밍 방법과 하이브리드 방법을 제안하였다. 실험결과, 제안한 멀티스트리밍 방법과 하이브리드 방법이 에어컨잡음과 청소기잡음환경에서 우수한 성능을 보여주었다.

참고문헌

[1] D. Reynold 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.

[2] K. Yuo, T. Hwang, and H. Wang, "Combination of Autocorrelation-Based Features and Projection Measure Technique for Speaker Identification," Proc. IEEE Trans. Speech and Audio Processing, Vol. 13, pp. 565-574, Jul., 2005.

[3] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-Band Speech Recognition in Noise Environments," In Proc. ICASSP, pp. 641-644, 1998.

[4] H. Hermansky, S. Tibrewala, and M. Pavel, "Toward ASR on Partially Corrupted Speech," In Proc. ICSLP, 1996.