자동차 주행 환경에서의 화자인식 시스템 개발에 관한 연구

양준영* 장준혁**, 이창원***, 박기희***
*한양대학교 전자컴퓨터통신공학과
**한양대학교 융합전자공학부
***현대모비스 선행연구팀
e-mail : jchang@hanyang.ac.kr (corresponding author)

A Study on Developing Speaker Recognition System In Driving Car Environment

Joon-Young Yang*, Joon-Hyuk Chang**, Chang Won Lee***, Ki-Hee Park***
*Dept. of Electronic and Computer Engineering, Hanyang University
**Dept. of Electronic Engineering, Hanyang University
***Advanced Engineering Team, Hyundai Mobis

요 약

화자인식 기술은 등록된 화자 목록 내 화자 또는 사칭 화자의 발화로부터 발화자를 식별하는 기술로써, 음성 소스를 기반으로 동작하는 디바이스의 개인화를 위해 필요한 기술이다. 본 논문에서는 차량 잡음이 존재하는 자동차 주행 환경을 타겟으로 하는 화자인식 시스템 개발 방법을 제안한다. 차량 잡음에 의해 오염된 음성신호로부터 잡음 성분을 제거하기 위해 parametric multi-channel Wiener filter (PWMF)를 이용하여 실험한 결과, 남성화자 조건에서는 PMWF 의 내부 파라미터 조절을 통해 필터를 minimum variance distortionless response (MVDR) 빔포머로 동작하도록 설정하였을 때, 여성화자 조건에서는 잡음을 제거하지 않았을 때 가장 낮은 동일오류율을 보임을 확인할 수 있었다.

1. 서론

자동차의 전장화가 점차 진행됨에 따라 차량 내 장비들의 사용을 편리하게 하기 위한 음성인식 기술의 개발 및 적용 사례와 응용 기술의 수요가 증가하는 추세이며, 이러한 전장 장비들의 개인화된 사용을 위해 차량 주행 환경에서의 화자인식 기술 또한 필요한 상황이다.

화자인식 기술은 발화(utterance)의 내용에 대한 제약 유무에 따라 문장종속(text-dependent) 방식과 문장독립(text-independent) 방식으로 구분된다. 문장종속 방식 화자인식 기술은 정해진 내용의 패스프레이즈를 발화하는 방식으로, 문장독립 방식의 화자인식 기술은 임의의 발화를 통해 발화자를 등록하고 테스트하는 방식으로 동작한다. 따라서, 차량 주행 중 운전자로부터 임의의 발화를 입력받아 개인화된 스케쥴링 서비스를 제공하는 것을 목적으로 할 경우, 문장독립 방식의 화자인식 기술 개발이 필요하다고 할 수 있을 것이다.

본 논문에서는 차량 잡음에 의해 오염된 다채널 음성신호로부터 잡음 성분을 제거하기 위한 paramertric multi-channel Wiener filter (PMWF) [1]를 이용한 전처리(pre-processing) 기술 적용과, 기존에 문장독립 방식 화자인식 분야에서 가장 널리 사용되고 있는 i-vector [2] 및 probabilistic linear discriminant analysis (PLDA) [3] 기반의 화자인식 시스템 개발 방법을 제안한다. 실험을 통해 잡음을 제거하지 않은 음성신호, Wiener filter를 이용하여 잡음을 제거한 음성신호, 그리고 minimum variance distortionless response (MVDR) 빔포머를 이용하여 잡음을 제거한 음성신호에 대하여 화자검증 동일오류율(equal error rate; EER)을 비교한 결과, 남성화자 조건에서는 MVDR 빔포머를 사용하여 잡음을 제거한 경우에, 여성화자 조건에서는 잡음을 제거하지 않은 경우에 가장 낮은 동일오류율을 보임을 확인할 수 있었다.

2. PMWF

PMWF 는 다채널 음성신호를 입력으로 하여 잡음성분을 제거할 시 음성왜곡정도와 잔여잡음량을 내부 파라미터를 통해 조절할 수 있는 알고리즘이다 [1]. 잡음성분 제거를 위해 Fourier transform (FT) 도메인에서 계산되는 이득 함수(gain function)의 식은 다음과 같다.

$$h(jw) = \frac{\Phi_{vv}^{-1}(jw) \cdot \Phi_{yy}(jw) - I_{N\times N}}{\beta + \xi(w)}$$

이 때, w와 N은 각각 radian 단위의 주파수와 마이

크로폰 채널 입력 수를 의미하며, $\Phi_{vv}^{-1}(jw)$, $\Phi_{yy}(jw)$, $\xi(w)$는 각각 w에 해당하는 주파수 빈(frequency bin)에서의 잡음신호 power spectral density (PSD) 행렬, 마이크 입력신호 PSD 행렬, 그리고 다채널 a priori SNR 을 의미한다 [1]. 이득 함수 식의 분모에 있는 파라미터 β값을 0 으로 설정할 경우 PMWF 는 음성왜곡을 허용하지 않고 잡음 성분을 제거하는 MVDR 빔포머로써 동작하고, β값을 0 보다 큰 값으로 설정할 경우, 잡음 제거 이후의 음성왜곡정도와 잔여잡음량 사이의 trade-off 를 조절할 수 있는 Wiener filter 로써 동작한다.

3. I-vector/PLDA 기반 화자인식 프레임워크

I-vector [2]는 임의의 발화에 대한 저차원의 벡터 표현식으로, 전체 발화에 대한 평균적인 특징을 기준으로 하였을 때 각 발화의 특징에 포함된 변이(variability) 성분을 모델링하는 방법이다. I-vector 추출 모델 학습 방법은 다음과 같다. 먼저, 전체 배경 화자의 발화로부터 프레임 단위의 특징을 추출하여 배경화자모델(universal background model; UBM)로써 사용할 Gaussian mixture model (GMM)을 구성한 뒤, GMM 의 각 Gaussian 성분과 각 발화로부터 추출한 특징들을 이용하여 각 발화에 대한 supervector 를 구성한다 [2].

$$F_c = \sum_{t=1}^{T} \gamma_c(t) \cdot (X_t - X_m)$$

위의 식은 특정 Gaussian 성분을 이용하여 supervector 의 구성 벡터를 계산하는 식으로, t와 c는 각각 시간별 프레임 인덱스 및 특정 Gaussian 성분 인덱스를 나타내며, $\gamma_c(t)$, X_t, X_m은 각각 c번째 Gaussian 성분에 대한 사후확률, t번째 프레임에서 추출한 특징벡터, 그리고 전체 특징벡터의 평균값을 의미한다. I-vector 추출 모델은 이와 같이 구성한 각 발화별 supervector 를 factor analysis [2] 모델을 이용하여 차원을 감소시키는 방향으로 선형 분해함으로써 각 발화에 포함된 변이 성분들 중 중요한 성분들만을 저차원의 벡터로 추출해내는 역할을 한다. 또한, PLDA [3] 모델은 각 발화별로 추출한 i-vector 특징으로부터 화자 성분을 분리해내어 비교함으로써 임의의 두 발화에 대한 발화자의 일치/불일치 정도를 나타내는 점수를 계산하는 역할을 한다. 본 논문에서는 simplified PLDA [4] 모델을 사용하여 실험을 진행하였다. .

4. 실험 내용 및 결과

실험에서 사용한 학습 데이터는 남성화자 250 명 및 여성화자 250 명으로 구성하였으며, 각 화자당 약 180 개 정도의 조용한 환경에서 근거리에 위치한 단일 마이크로폰을 향해 발성한 깨끗한 음성 녹음 발화를 사용하였다. 또한, 근거리 단일 마이크로폰을 통해 녹음한 깨끗한 음성 발화를 자동차 환경에서의 2 채널 발화로 시뮬레이션 하기 위한 방법으로 room impulse response (RIR) generator [5] 툴을 사용하여 시뮬레이션 데이터를 생성하였다. 실험에서 사용한 RIR generator 의 파라미터 설정값은 <표 1>과 같다.

<표 1> RIR generator 파라미터 설정값

Room size	(1.4, 2.3, 1.3) [m]
Source position	(0.4, 1.5, 1.0) [m]
Receiver position	mic1: (0.36, 1.8, 1.3) [m] mic2: (0.44, 1.8, 1.3) [m]
RT60	0.3 ~ 1.0 [sec]
Filter length	128 [samples]
Reflection order	-1 (maximum)
Polar pattern	Omnidirectional
Azimuth	0 [rad]
Elevation angle	π/6 [rad]
Apply HP filter	1 (True)
Gain	1.5 ~ 3.5

RIR generator 를 이용하여 생성한 2 채널 시뮬레이션 데이터에는 자동차 주행 시 발생하는 소음을 녹음한 잡음신호를 신호대잡음비를 -5 dB 에서 10 dB 사이에서 1 dB 간격으로 각 발화마다 랜덤하게 설정하여 합성하였으며, 이와 같은 방법으로 생성한 차량 주행 환경 시뮬레이션 2 채널 음성 데이터를 PMWF 에 통과시켜 획득한 음성신호를 모델 학습에 사용하였다. PMWF 는 $\beta = 0$으로 설정한 경우와 $\beta = 10$으로 설정한 경우에 대해 실험을 진행하였다.

평가에 사용한 데이터는 차량 주행 중 녹음한 남성화자 129 명 및 여성화자 129 명으로 구성하였으며, 각 성별별로 99 명의 화자를 등록화자로, 나머지 30 명의 화자를 사칭화자로 사용하여 화자검증을 위한 trial set 을 구성하는 데에 사용하였다. 등록발화로는 각 화자별로 1 개 또는 2 개의 발화를 사용하였으며, 테스트발화로는 각 화자별로 등록발화와 중복되지 않는 75 개의 발화를 사용하였다. 음성검출 알고리즘을 적용하여 평가에 사용한 발화들의 실제 음성구간을 조사해 본 결과, 약 1 초에서 10 초 사이의 길이를 가졌으며, 평균 약 3.8 초 정도 길이의 실제 음성구간으로 구성되어 있었다. 각 성별당 총 trial 의 개수는 957,825 개로, 등록발화와 테스트발화의 발화자가 일치하는 target trial 의 개수와 발화자가 일치하지 않는 nontarget trial 의 개수는 각각 7,425 개와 950,400 개이다.

실험은 Kaldi [6] 툴킷을 이용하여 진행하였으며, 사용한 모델 파라미터 설정은 다음과 같다. 먼저, 각 발화로부터 프레임 단위로 추출한 특징으로는 20 차원의 mel-frequency cepstral coefficient (MFCC)를 사용하였으며, 시간에 대한 변화량인 delta 및 delta-delta 특징을 덧붙여 총 60 차원의 특징을 배경화자모델 GMM 학습에 사용하였다. 음성검출 알고리즘으로는 Kaldi 에서 제공하는 에너지 기반 음성검출 알고리즘을 사용하였고, 잡음제거 방식의 종류에 따른 화자검증 성능을 공정하게 비교하기 위해 PMWF 의 파라미터 β를 10 으로 설정하여 잡음을 제거한 뒤에 얻은 음성신호에 대한 음성검출 결과를 잡음을 제거하지 않은 경우 및 $\beta = 0$으로 설정한 경우에 대해 동일하게 적용하였다. 배경화자 모델은 총 1,024 개의 Gaussian 성분을 갖는 GMM 을 사용하였고, i-vector 추

출 모델은 400 차원의 i-vector 를 추출하도록 학습하였다. PLDA 모델은 차원 감소 없이 400 차원의 eigenvoice space 를 구성하도록 학습하였다.

<표 2>와 <표 3>은 각각 남성 및 여성화자별로 구성한 trial set 에 대한 화자검증 동일오류율을 나타낸 표이다. 남성화자에 대한 화자검증 결과에서는 잡음제거 시 음성왜곡을 허용하지 않는 MVDR 빔포머를 사용하여 전처리를 진행한 경우 가장 좋은 성능을 보였으며, 잡음을 제거하지 않은 경우와 Wiener filter 를 사용하여 잡음을 제거한 경우가 유사한 성능을 보였다. 여성화자에 대한 화자검증 결과에서는 Wiener filter 를 사용하여 잡음을 제거한 경우가 가장 좋지 않은 성능을 보였으며, 잡음을 제거하지 않은 경우가 MVDR 빔포머를 사용한 경우보다 조금 더 좋은 성능을 나타냈다. 이는 잔향과 잡음이 섞인 음성신호로부터 Wiener filter 를 이용하여 잡음을 제거한 결과 음성 구간에서 뮤지컬 노이즈(musical noise)와 같은 잔여잡음 및 음성의 왜곡이 발생하기 때문인 것으로 생각된다. 또한, 두 경우 모두 등록에 사용한 발화의 양을 증가시켰을 때에 많은 성능 향상을 보였다.

<표 2> 남성화자에 대한 화자검증 동일오류율 (%)

실험 조건	등록 발화 개수	
	1 개	2 개
PMWF 미적용	1.832	1.061
PMWF ($\beta = 10$)	1.806	1.060
MVDR ($\beta = 0$)	**1.697**	**0.972**

<표 3> 여성화자에 대한 화자검증 동일오류율 (%)

실험 조건	등록 발화 개수	
	1 개	2 개
PMWF 미적용	**4.183**	**2.663**
PMWF ($\beta = 10$)	4.579	2.964
MVDR ($\beta = 0$)	4.243	2.695

5. 결론

본 논문에서는 차량 주행 환경에서 동작하는 화자인식 시스템 개발을 위한 방법으로 PMWF 를 이용한 잡음제거 방법 및 이에 따른 i-vector/PLDA 프레임워크를 이용한 화자검증 결과를 비교하였다. 남성화자 집합에 대한 실험 결과로는 MVDR 빔포머를 사용한 결과가 가장 우수하였고, 여성화자 집합에 대한 실험 결과로는 잡음을 제거하지 않은 경우가 MVDR 빔포머를 사용한 경우보다 조금 더 우수한 성능을 보였으며, Wiener filter 를 사용한 경우가 가장 좋지 않은 성능을 나타냈다. 위와 같은 실험 결과를 통해 동일한 음성검출 결과를 이용하여 i-vector/PLDA 기반 화자검증을 수행하였을 때, 화자의 음성을 왜곡시킬 수 있는 Wiener filter 를 사용할 경우 화자검증 성능이 저하될 수 있다는 결론을 얻을 수 있었다. 또한, 여성화자의 발화에 대해 MVDR 빔포머를 사용할 경우, 남성화자의 발화와는 다른 내부 알고리즘 파라미터 설정을 고려해 볼 수 있을 것이다.

ACKNOWLEDGEMENT

이 논문은 과학기술정보통신부 재원으로 경찰청과 치안과학기술연구개발사업단의 지원을 받아 수행된 치안과학기술연구개발사업임. (PA-J000001-2017-101).

참고문헌

[1] M. Souden, J. Benesty, and S. Affes, “On optimal frequency-domain multichannel linear filtering for noise reduction,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260-276, Feb. 2010.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 19, no. 4, pp. 788-798, May, 2011.

[3] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *IEEE ICCV,* pp. 1-8, 2007.

[4] P. Kenny, “Bayesian speaker verification with heavy-tailed priors,” in *Odyssey 2010 – The Speaker and Language Recognition Workshop,* Brno, Czech Republic, 2010.

[5] J. Allen and D. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal Acoustic Society of America,* vol. 65, no. 4, pp. 943-950, Apr. 1979.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi Speech Recognition Toolkit,” in *Proc. ASRU,* 2011.