잡음 환경에서 화자 확인을 위한 다중대역에 기반한 공분산 방법

최민정, 이기용
숭실대학교 정보통신공학과

Covariance Model Based on Multi-Band for Speaker Verification in Noise

Min Jung Choi and Ki Yong Lee
School of Electronics Engineering, Soongsil University / cmj1109@ctsp.ssu.ac.kr, kylee@ssu.ac.kr

요약

기존의 전대역(Full-Band)에서 특징 파라미터를 추출하는 화자 확인(Speaker Verification) 시스템은 저대역이나 고대역에서 화자 정보의 특징이 제거되기 쉽다. 또한, 주파수 스펙트럼에 부분적으로 오염이 되는 경우, 특징 파라미터를 왜곡시켜 화자 확인 시스템의 성능을 저하시킨다. 본 논문에서는 이러한 문제점을 해결하기 위해 다중대역 공분산 모델(Covariance Model)을 제안한다. 제안한 방법은 주파수 역영에서 전대역을 여러 개의 부대역(Sub-Band)으로 분할하고, 부대역별로 독립적으로 특징 파라미터를 추출하여 공분산 모델을 구한다. 제안된 방법의 성능 확인을 위하여 공분산 모델 간의 거리를 측정하는 화자 확인 실험을 하였다. 잡음 환경에서 기존의 방법인 전대역에 기반한 공분산 모델과 제안한 방법을 비교 분석한 결과, 제안한 방법이 기존 방법보다 2%정도 성능이 향상되었다. 또한, 제안된 방법은 전대역에 기반한 파라미터 차원 수를 다중대역의 개수로 분할하여 사용하므로 계산량의 감소와 저장 공간면에서 효율적이다.

1. 서론

음성신호에는 음운정보 뿐만 아니라, 개개인의 독특한 생체 정보를 가지고 있다. 개개인의 생체적 정보로서의 음성 신호를 이용하여, 이전에 등록된 본인 모델과 비교하여 기준값(Threshold)보다 크면 본인임을 확인(Accept)하고, 아니면 거절(Reject)하는 방법을 화자 식별이라 한다.

기존에는 전대역(Full-Band)에 골고루 분포되어있는 백색 잡음을 제거하기 위해 노력해 왔다. 하지만, 실생활에서는 주파수 스펙트럼 영역에서 잡음이 부분적(협대역)으로 분포되어 나타난다. 기존의 전대역에 기반한 화자 확인 시스템은 스펙트럼의 일부분만이 손상된 경우에도 특징 파라미터 전체가 왜곡되어 화자 확인률이 저하된다. 이러한 문제를 "각 주파수들은 독립적으로, 부분적으로 발생한 오류는 다른 주파수 영역에서의 인식에 영향을 미치지 않는다"라는 HSR(Human Speech Recognition)의 결과를 적용하여 협대역의 잡음에 오염되지 않은 부대역들의 재결합으로 화자 확인률의 저하를 줄일 수 있다.

다중대역(Multi-Band) 분석법에서의 중요한 논점은 부대역의 개수, 부대역의 범위, 부대역의 특징 파라미터의 종류, 부대역 인식 결과에 대한 가중 및 통합 방식들이 있다. 이 중에서 인식 결과를 재결합하는 방법은

화자 확인 시스템의 성능에 영향을 미치는 가장 핵심적인 문제이다. 재결합하는 방법으로는 부대역 인식 결과에 동일한 가중치(EW: Equal Weighting)를 주거나 부대역의 인식 정확도 가중치(AW: Accuracy Weighting)를 주는 방법 또는 특징 파라미터들을 재결합하는 방법(FR: Feature Recombination), 부대역 상호 정보(ML: Mutual Information)또는 최대 유사도(ML: Maximum Likelihood)에 기반한 부대역 신뢰도 측정 방법이 있다.

화자 인식에는 입력 음성신호가 기준 패턴과 일치하는 정도를 이용하는 방법인 시간축 정합법(DTW: Dynamic Time Warping)알고리즘, 출력 밀도 함수가 여러 개의 혼합성분으로만 이루어지는 GMM(Gaussian Mixture Model) 방법, 그리고, 평균을 고려하지 않고 전체 공분산(Covariance Model)을 사용하는 방법 등이 있다.

본 논문에서는 협대역 잡음이 부가된 음성을 멜영역에서 여러 개의 부대역(Sub-Band)으로 분할하여 독립적으로 추출된 특징 파라미터들을 공분산 모델을 이용하여 가중 및 사상 함수를 이용하여 재결합 방식을 적용하였다.

본 논문은 다음과 같이 구성되어 있다. 2 장에서는 제안된 방법인 다중대역에 기반한 특징 파라미터를 추출하는 방법과 공분산 모델에 대해서 설명하고, 3 장에서는 가중 및 사상함수에 대해 설명하고, 4,5 장에서는 실험 결론 및 고찰을 서술하였다.

2. 특징 파라미터 추출 및 공분산 모델

본 논문에서 제안한 다중대역에 기반한 화자 확인 시스템은 특징 파라미터를 추출하는 전처리단과 화자 모델을 생성하는 인식단으로 구분된다.

2.1 부대역별 특징 파라미터 추출

입력된 음성은 프레임 단위로 처리된다. 각각의 프레임들은 FFT(Fast Fourier Transform)를 통해 주파수 영역으로 변환된다. L-point FFT 의 데이터들에 N 개의 채널을 가진 필터뱅크를 적용시켜 m_i 의 로그 에너지를 얻는다. 여기서 우리는 부대역 간의 특징 파라미터가 독립적인 특성을 갖도록 주파수 전대역을 M 개의 다중대역으로 분할한다. 하나의 부대역 안에 존재하는 필터뱅크 수는 $S(=N/M)$ 개이다. k 번째 부대역 안에 존재하는 필터뱅크 로그 에너지를 $m_i^{(k)}$ 라 한다면, 식 (1)의 DCT 함수를 사용하여 k 번째 부대역의 P_{sc} 차원 멜 켑스트럼을 추출한다.[2]

$$c_j^{(k)} = \sqrt{\frac{2}{S}} \sum_{i=1}^{S} m_i^{(k)} \cos\left[(i-0.5) * \frac{j\pi}{S} \right] \qquad (1)$$

여기서 $j = \{1,...,P_{sc}\}$ 이다. 채널 왜곡을 제거하기 위해 켑스트럼 평균 차감법(CMS)을 적용하고, 화자 확인률을 향상시키기 위해 스펙트럼의 천이 정보인 델타 켑스트럼을 사용하여 P (켑스트럼+ 델타 에너지+ 델타 켑스트럼)차원의 특징 파라미터를 추출할 수 있다.

2.2 공분산 모델 및 BDR

k 번째 부대역에서 추출된 P 차원의 관측열 벡터를 $X^{(k)} = \{X_1^{(k)}, X_2^{(k)}, ..., X_T^{(k)}\}$ 라 하자. 공분산 모델은 식 (2)와 같이 관측열 벡터와 k 번째 부대역의 화자 평균 $\mu_{spk}^{(k)}$로써 계산될 수 있다.

$$C_{spk}^{(k)} = \frac{1}{T-1} \sum_{t=1}^{T} (X_t^{(k)} - \mu_{spk}^{(k)})^T (X_t^{(k)} - \mu_{spk}^{(k)}) \qquad (2)$$

여기서 공분산 모델은 각각의 부대역 별로 계산되므로 M 개의 공분산 모델로 구성된다.

공분산 모델들의 거리 측정을 위한 BD(Bhattachary-ya Distance)방법은 P 차원의 관측 벡터들이 가우시안 확률분포를 가질 때, 두 공분산 사이의 거리 측정 및 부동성을 측정하기 위해 Cambell 에 의해 제안되었다.[2][3]

k 번째 부대역에서 계산된 2 개의 공분산 행렬 $C_1^{(k)}$ 과 $C_2^{(k)}$ 을 가지고 있을 때, $BD_{1,2}^{(k)}$ 값은 식 (3)과 같이 계산된다.

$$BD_{1,2}^{(k)} = \frac{1}{2} \ln \frac{\left| \frac{C_1^{(k)} + C_2^{(k)}}{2} \right|}{\left| C_1^{(k)} \right|^{1/2} \left| C_2^{(k)} \right|^{1/2}} \qquad (3)$$

BE 값은 발성된 음성들의 공분산들 값에 의존하는 점을 보완하기 위해 다양한 화자들로부터 수집된 음성들을 사용한 배경 모델을 적용하였다. 발성 모델 $C_{utt}^{(k)}$ 과 화자 모델 $C_{spk}^{(k)}$, 발성 모델과 배경 모델 $C_{world}^{(k)}$ 의 비

율로 정의되는 $BDR^{(k)}$ (Bhattacharyya Distance Ratio) 방법이다.

$$BDR^{(k)} = -\frac{BD^{(k)}(C_{utt}^{(k)}, C_{spk}^{(k)})}{BD^{(k)}(C_{utt}^{(k)}, C_{world}^{(k)})} \quad (4)$$

3. 가중 및 사상함수

식 (4)에서 계산된 각 부대역의 BDR 값은 화자마다 화자 평균이 다른 분포를 가지고 있기 때문에 동일한 화자 확인 시스템의 기준값을 적용시킬 수 없다. 이러한 문제점을 보완하기 위한 방법으로 화자의 BDR 값들의 분포가 $N(\mu,\sigma)$일 때, 우리가 보고 싶은 영역에 기준이 되는 가우시안 분포 $N(\mu_R,\sigma_R)$가 되도록 사상시키는 방법을 사용한다. BDR 값 x가 $N(\mu,\sigma)$의 분포에서 최고점과 x가 나올 수 있는 빈도 수의 비율이 기준이 되는 $N(\mu_R,\sigma_R)$의 분포에서의 최고점과 사상되는 지점의 빈도 수의 비율이 같다는 가정으로부터 다음 식을 얻을 수 있다. [5]

$$z_{map} = \mu_R + \frac{\sigma_R}{\sigma}(x-\mu) \quad (5)$$

각각의 부대역에서 계산되어진 z_{map} 값에 협대역 잡음에 영향을 미치지 않으면서 화자 확인률을 최적으로 하는 가중치 w_k를 적용하여 화자 확인 시스템의 값을 얻을 수 있다.

4. 실험 및 결론

실험을 위하여 사용된 데이터는 실험실 환경에서 수집된 한국어 문장 종속 연속음 "열려라 참깨"를 사용하였다. 데이터는 1 주 간격의 시간차를 가지고, 3 주에 걸쳐 수집하였다. 매주 1 회 발성에서는 각 5 번 발성을 하였으며, 개인별 전체 발성된 데이터 수는 15 개이다. 화자 인원 수는 200 명으로 남/여 각각 100 명이다. 샘플링 주파수는 16kHz 이고, 분해능은 16bit 이다. 등록 데이터는 처음 2 주간 10 번 발성한 음성 데이터를 사용하였다.

50% 중첩을 갖는 25.6ms 를 한 프레임으로 음성을 분석하였다. 추출된 특징 파라미터는 256-point FFT 를 사용하고, 멜 스케일에서 균등한 24 개의 채널을 가진 필터 뱅크를 다중대역 개수(M =4)로 분할한 뒤, 각각의 부대역에서 독립적으로 7 차 MFCC 를 사용하였다.

다중대역에 기반한 공분산 모델의 화자 확인률을 비교 분석하기 위해, 기존의 방법인 혼합성분 17개를 사용한 GMM을 사용하였다. 표 1은 깨끗한 음성에서의 화자 확인률이다. Full-Band는 전대역에서 하나의 특징 파라미터를 추출하는 방법이고, Sub-Band는 각각의 부대역에서 독립인 화자 모델을 만든 것이다. 주파수 축에서 SB1는 다중대역들 중 첫번째 대역으로 156.25 ~718.75Hz에서 추출된 특징 파라미터이고, 이와 같은 방식으로 SB2는 593.75~1906.25Hz, SB3는 1656.25 ~4062.5Hz, 마지막 대역인 SB4는 3593.75~8kHz 대역에서 추출한 파라미터를 사용하였다. Multi-Band는 Sub-Band의 부대역들의 결과를 가중 및 사상 함수를 적용시켜 재결합한 방법이다. 부대역들의 각 밴드별 확인률은 전체적으로 나쁘지만, 재결합한 다중대역 확인률은 기존의 화자 확인률의 성능을 유지하거나, 약간 향상되었다. 다중대역 인식에서 GMM은 적은 데이터에서 많은 모델 파라미터를 추출하므로 성능면에서 1.75% 저하되었다.

	Full -Band	Sub-Band				Multi -Band
		SB1	SB2	SB3	SB4	
GMM	6.01	14.50	16.96	14.08	21.69	7.76
BD	6.20	12.52	14.85	11.03	17.53	3.64
BDR	0.49	10.08	14.47	8.68	17.97	2.56

표 1). 실험실 환경에서의 전대역 및 다중대역 화자 확인 결과(EER, %)

		SNR			
		5 dB	10 dB	15 dB	20 dB
FB	BD	13.76	11.01	9.46	8.05
	BDR	2.44	1.93	1.73	1.36
MB	BD	6.79	5.62	4.86	4.38
	BDR	3.57	3.15	3.02	3.00

표 2). 협대역 잡음에서의 전대역 및 다중대역 화자 확인 결과(EER, %)

Conventional GMM	CM based on Fullband	Proposed CM based on Multiband
$M_G(2P_G+1)$	$P_G * P_G$	$M(P*P)$

표 3). 모델 파라미터 수

표 2 는 음성에 중심 주파수가 1.2kHz 이고, 대역폭이 400Hz 인 협대역 잡음이 부가된 것으로 다중대역의 두번째 대역을 오염시켜 실험한 결과이다. BD 방법에서 기존의 특징 추출 방법인 Fullband 는 잡음의 영향을 받아 성능이 많이 저하되었다. 그러나, 제안한 방법은 잡음에 영향을 받은 sb2 의 영향을 감소시켜 재결합하여 성능 저하를 줄일 수 있었고, 기존 방법보다 월등한 성능을 나타내었다. 표 1,2 에서 FB 은 배경 모델을 이용한 BDR 방법이 다중대역의 BDR 방법보다 우수한 성능을 갖는데, 이는 다중대역의 배경모델이 재결합하는 과정에서 각대역의 분별력을 저하시키기 때문이다.

전대역에 기반한 파라미터를 사용할 경우, 대각 공분산 행렬을 갖는 GMM 은 $(2P_G+1)$차원의 파라미터가 혼합성분 M_G개 만큼, 공분산 모델은 $(P_G * P_G)$차원의 파라미터들을 요구한다. 그리고, 다중대역에 기반한 공분산 모델은 $(P*P)$차원의 파라미터들이 다중 대역의 수 M 개 만큼 요구된다. 본 실험에서는 P_G=12, M_G=17, P=7, M=4 이므로, 전대역에서의 GMM, 공분산 모델 그리고 다중 대역에서의 공분산 모델은 각각 425 개, 625 개, 190 개의 특징 파라미터들이 요구된다. 따라서, 제안한 방법이 전대역에 기반한 모델보다 약 50% 정도의 계산량을 줄일 수 있고, 메모리 저장 공간을 줄일 수 있는 이점을 갖고 있다.

4. 고찰

전대역에서 추출된 특징 파라미터는 협대역의 오류가 전대역에 영향을 주므로 화자 확인 시스템의 성능을 저하되는 요인이 된다. 이를 해결하기 위하여 전대역을 여러 개의 부대역으로 분할한 후, 독립적으로 특징 파라미터를 추출하는 다중대역 공분산 모델을 사용하는 방법을 제안하였다. 제안된 방법이 기존의 GMM 확인률보다 성능이 우월하였으며, 혼합모델에 대한 계산량을 많이 줄일 수 있었다. 또한, 메모리 저장 공간을 줄일 수 있었다.

감사의 글

This work was supported by Biometrics Engineering Research Center,(KOSEF).

참고문헌

1. Brian Mak, "A Mathematical Relationship Between FullBand and Multiband Mel-Frequency Cepstral Coefficients", IEEE signal processing letter, vol 9, No 8, 2002
2. Campbell, J.P. Jr., "Speaker Recognition: a tutorial", Proceedings of the IEEE. 85(9), p.1437-p.1462, 1997
3. Petry, A., Zanus, A., Barone, D.A.C., "Bhattacharyya Distance Applied to Speaker Identification", Int., Conference on Signal Processing Applications and Technology, Dallas, Orlando, (1), 2000
4. Reynolds, D.A, "Channel Robust Speaker Verification via Feature Mapping", ICASSP'03, Vol(3), pp.II-53-6, 2003