화자 확인을 위한 다중대역에 기반한 주성분 분석 공분산 모델

PCA Covariance Model Based on Multiband for Speaker Verification

최 민 정* · 이 윤 정** · 서 창 우***

Minjung Choi · Younjeong Lee · Changwoo Seo

ABSTRACT

Feature vectors of speech are generally extracted from whole frequency domain. The inherent character of a speaker is located in the low band or high band frequency. However, if the speech is corrupted by narrowband noise with concentrated energy, speaker verification performance is reduced as the individual characteristic is removed.

In this paper, we propose a PCA Covariance Model based on the multiband to extract the robust feature vectors against the narrowband noise. First, we divide the overall frequency band into several subbands. Second, the correlation of feature vectors extracted independently from each subband is removed by PCA. The distance obtained from each subband has different distribution. To normalize against the different distribution, we moved the value into the normalized distribution through the mapping function. Finally, the represented value applying the weighting function is used for speaker verification. In the experiments, the proposed method shows better performance of the speaker verification and reduces the computation.

Keywords: PCA, covariance model, multiband, speaker verification, Bhattacharyya Distance

1. 서 론

화자 확인 시스템은 발성된 음성 모델에서 계산된 대표값이 등록된 화자의 문턱값보다 큰 경우 화자로 수락하고, 그렇지 않으면 거절하는 시스템이다. 이 시스템은 크게 전처리 과정과

* (주)인스모바일

** 국방과학연구소

*** ㈜에스씨디 정보통신연구소

모델링 과정으로 나눌 수 있다. 먼저 전처리 과정은 화자의 특징벡터를 추출하는 과정으로 발성된 음성을 이용하여 화자의 발성기관을 모델링한 LPCC(Linear Prediction Cepstral Coefficients)와 멜 척도로 변환하여 화자의 청각 특성을 고려한 MFCC(Mel Frequency Cepstral Coefficients)가 사용된다. 그리고 추출된 특징벡터를 이용해서 모델링을 하게 되는데, 모델링 방법으로는 일반적으로 HMM(Hidden Markov Model)(Juang, 1985), GMM(Gaussian Mixture Model)(Reynold & Rose, 1995; Lee et al., 2003), 공분산 모델(Covariance Model) (Zilca, 2002) 등이 사용된다.

기존의 화자 확인 시스템은 주파수 영역, 즉 전대역에 골고루 분포되어 있는 백색 잡음의 영향을 감소시키기 위해서 노력해왔다. 하지만 실생활 잡음(자동차 잡음, 공장 잡음 등)은 주파수 영역에서 일정한 영역에 에너지가 집중된 협대역 잡음의 형태로 주로 나타나게 되는데, 이러한 경우 특징벡터 전체를 손상시켜 성능을 저하시키는 문제점이 있다(Yoshida, Takagi & Ozeki, 2004).

본 논문에서는 주파수 영역에서 음성 스펙트럼이 존재하는 전대역을 여러 개의 부대역으로 분할하고, 분할된 각각의 부대역에서 독립적으로 추출된 특징벡터에서 벡터간 상관관계를 제거하여 사용하는 PCA(Principal Component Analysis) (Reynolds, 2003; 이윤정, 서창우, 이기용, 2002) 공분산 모델을 제안한다. 또한, 부대역의 프레임별 값의 분포는 부대역마다 평균이 다른 분포를 가지고 있어, 가중치 함수를 적용하여 화자의 대표값을 계산하면 큰 값을 갖는 부대역과 작은 값을 갖는 부대역에 객관적인 가중치를 적용할 수 없다. 이 문제를 해결하기 위해 사상 함수를 적용하여 부대역의 프레임별 값을 기준이 되는 분포로 각 부대역별로 이동하여 동등한 위치에서 가중치를 적용하여 화자의 대표값을 얻어 화자 확인을 수행하였다.

2. 다중대역에 기반한 PCA 공분산 모델

2.1 특징벡터 추출

특징벡터를 추출하기 위해서 입력된 음성은 프레임 단위로 처리되며, 각각의 프레임들은 FFT(Fast Fourier Transform)를 통해 주파수 영역으로 변환된다. L-point FFT의 데이터들에 N 개의 채널을 가진 필터 뱅크(Filter Bank)를 적용시켜 전력 스펙트럼 $e(i)$ 의 로그 에너지를 얻는다. M 개의 부대역을 갖는 다중대역은 식(1)을 이용하여 <그림 1>처럼 얻을 수 있다.

$$e_i^{(k)} = B\left(i \in (k\frac{N}{M} + l) \right) \cdot e(i) \qquad 0 \le k < M, 1 \le i \le N, 1 \le l \le \frac{N}{M} \tag{1}$$

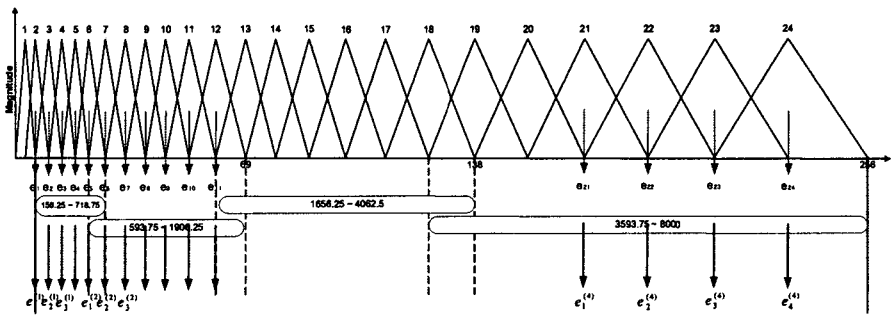

그림 1. 부대역 M(=4)을 갖는 다중대역으로 분리된 필터 뱅크

여기서, $B(\cdot)$는 i번째 필터 뱅크 에너지가 k번째 부대역에 속한 삼각 대역 통과 필터 뱅크이면 '1'이고, 그렇지 않으면 '0'의 값을 갖는다. 그리고, k번째 부대역의 j번째 멜 켑스트럼 계수 $c_j^{(k)}$는 필터 뱅크 $e_i^{(k)}$의 로그 에너지에 DCT(Discrete Cosine Transform)를 적용하여 p_{sc}차원의 멜 켑스트럼을 계산한다(Mak, 2002).

$$c_j^{(k)}(t) = \sqrt{\frac{2}{Q_k}} \sum_{i=1}^{Q_k} \log[e_i^{(k)}] \cdot \cos[\frac{\pi j}{Q_k}(i-0.5)] \qquad 1 \le k \le M, 1 \le j \le p_{sc} \tag{2}$$

식(2)로 얻어진 특징벡터는 CMS(Cepastrum Mean Subtraction)(Garcia & Mammone, 1999)로 채널 왜곡을 제거하고, 화자 확인률을 향상시키기 위해 스펙트럼의 천이 정보를 사용하여 시간이 t일 때, k번째 부대역의 특징벡터는 $P(2*p_{sc}+1)$차원으로 식(3)과 같이 멜 켑스트럼과 델타 에너지, 델타 켑스트럼으로 구성된다.

$$c^{(k)}(t) = \left\{c_1^{(k)}(t), c_2^{(k)}(t), ..., c_{p_{sc}}^{(k)}(t), \Delta_e^{(k)}, \Delta c_1^{(k)}(t), \Delta c_2^{(k)}(t), ..., \Delta c_{p_{sc}}^{(k)}(t)\right\} \tag{3}$$

2.2 PCA(Principal Component Analysis)

식(3)에서 얻어진 특징벡터들 간의 상관관계를 줄이고, 차원 수를 감소시키기 위해서 PCA을 적용하였다. 즉, PCA는 식(3)에서 얻어진 k 번째 부대역의 특징벡터 $X^{(k)} = \left\{c^{(k)}(t) \in \Re^P; 1 \le t \le T, 1 \le k \le M\right\}$가 $c^{(k)}(t) = [c_1, ..., c_P]$로 구성되어있을 때, 정보의 손실없이 상관관계가 줄어든 $Q(<P)$차원의 변환된 $y^{(k)}(t) = [y_1, ..., y_q]$를 다음과 같이 구할 수 있다.

공분산 행렬 $\Sigma = \left[\sigma_{ij} \mid 1 \le i, j \le P\right]$는 $X^{(k)}$의 평균벡터로 계산된다. 여기서 ij번째 공분산 요소는 $\sigma_{ij} = \frac{1}{T}\sum_{t=1}^{T}\left(c^{(k)}(t) - \mu_i^{(k)}\right)^T \cdot \left(c^{(k)}(t) - \mu_j^{(k)}\right)$ 이다. 또한, 공분산 행렬은 고유벡터(Eigen Vector)와 고유값(Eigen Value)으로 식(4)와 같다.

$$\Sigma = \sum_{i=1}^{P} \lambda_i \upsilon_i \upsilon_i^T \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_P \tag{4}$$

여기서, λ_i는 공분산 행렬 Σ의 i번째 고유값이고, υ_i는 고유값 λ_i에 대응되는 정규화된 고유벡터이다. 이들은 $P \times P$인 직교 행렬($\Omega\Omega^T = I$)을 이룬다. 이로부터 t번째 프레임의 P차원 특징벡터 $c_t^{(k)}$와 PCA 변환을 수행한 t번째 프레임의 Q차원 특징벡터 $y_t^{(k)}$의 관계는 $y_t^{(k)} = \upsilon_i^T c_t^{(k)}$이며, 성분 전체의 벡터 $X^{(k)}$와 $Y^{(k)}$와의 관계는 $Y^{(k)} = \Omega^T X^{(k)}$이다.

이 때, 주성분 Q차원이 가지는 정보 비율(I_R)은 다음과 같다.

$$I_R = \frac{\sum_{i=1}^{Q} \lambda_i}{\sum_{i=1}^{P} \lambda_i} \tag{5}$$

여기에서, $P = Q$이면 정보의 손실없이 $X^{(k)}$에서 $Y^{(k)}$로 변환된다. 이와 같이 정보 비율에 의해 고유값이 큰것부터 Q차원만 선택한 Ω_Q^T를 사용하여 식(6)과 같이 특징벡터의 주성분을 얻을 수 있다.

$$Y^{(k)} = \Omega_Q^T X^{(k)} \tag{6}$$

2.3 공분산 모델 및 거리 왜곡 측정 방법

k번째 부대역의 특징벡터 $Y^{(k)} = \left\{ y_t^{(k)} \in \Re^Q ; 1 \leq t \leq T, 1 \leq k \leq M \right\}$와 k번째 부대역의 화자 평균 $\mu^{(k)} = \sum_{t=1}^{T} \left(y_t^{(k)} / T \right)$으로 공분산 모델 $\Sigma^{(k)}$를 계산된다.

$$\Sigma^{(k)} = \frac{1}{T} \sum_{t=1}^{T} \left(y_t^{(k)} - \mu^{(k)} \right)^T \cdot \left(y_t^{(k)} - \mu^{(k)} \right) \tag{7}$$

공분산 모델들의 거리 측정을 위한 BD(Bhattacharyya Distance) 방법은 Q차원의 특징벡터들이 가우시안 확률 분포를 가질 때, 두 공분산 사이의 거리 왜곡 측정을 위해 Cambell에 의해 제안되었다(Petry, Zanus & Barone, 2000). 여기서 Σ_1^k는 등록 단계에서 계산된 k번째 부대역의 공분산이고, Σ_2^k는 테스트 단계에서 계산된 k번째 부대역의 공분산이다.

$$BD_{1,2}^k = BD(\Sigma_1^k, \Sigma_2^k) = \frac{1}{2} \ln \frac{\left| \frac{\Sigma_1^k + \Sigma_2^k}{2} \right|}{\left| \Sigma_1^k \right|^{1/2} \left| \Sigma_2^k \right|^{1/2}} \tag{8}$$

3. 다중대역에 기반한 대표값 산출

각 부대역의 프레임별 값의 분포는 부대역마다 평균이 다른 분포를 가지고 있어, 가중치 함수를 적용하여 화자의 대표값을 산출하면, 각 부대역의 인식 결과를 충분히 대응시키지 못하기 때문에 사상 함수를 적용한 뒤 가중치 함수를 적용하여 화자의 대표값을 얻는다.

3.1 사상 함수(Mapping Funtion)

식(8)에서 계산된 k 번째 부대역의 분포는 $N(\mu_k,\sigma_k)$ 로 평균이 $\mu_1,...,\mu_M$ 으로 서로 다르기 때문에 사상 함수를 적용하여 각 부대역의 값들을 평균 μ_R 를 가진 참조 분포 $N(\mu_R,\sigma_R^k)$ 로 이동한다. 사상 함수는 거리 왜곡값 $d^{(k)}$ 가 나올 수 있는 빈도수 $(freq_k)$ 와 최고점 $(\max_k)$ 의 비율이 참조 분포에 사상되는 값 $d_{map}^{(k)}$ 가 나올 수 있는 빈도수 $(freq_R)$ 와 최고점 $(\max_R)$ 의 비율이 같다는 가정으로 유도된 식(9)이다. 식(9)로 변화된 값의 분포는 본인과 사칭자와의 분포를 더욱 멀리 위치하게 하여 화자 확인의 문턱값에 민감하지 않아 화자 확인 성능을 향상시킬 수 있다.

$$d_{map}^{(k)} = \mu_R + \frac{\sigma_R^k}{\sigma_k}(d^{(k)} - \mu_k) \tag{9}$$

3.2 가중치 함수(Weighting Function)

협대역 잡음이 부가된 음성은 잡음이 부가된 대역의 에너지가 다른 부대역에 비해 상대적으로 높기 때문에 가중치 함수는 무음 구간에서 계산된 부대역 에너지의 반비례하도록 계산하였다. 이는 잡음의 영향에 오염된 부대역을 제거하는 효과로 나타나기 때문에 화자 확인률을 향상시킬 수 있었다.

화자 발성 구간이 시작되기 전인 무음 구간에서 계산된 k 번째 부대역에서 계산된 로그 에너지의 합이 SBE_k 이고, k 번째 부대역의 로그 에너지의 역수가 φ_k 이면 k 번째 부대역에 곱해지는 가중치는 다음과 같다.

$$w_k = \frac{\varphi_k}{\sum_{l=1}^{M}\varphi_l} \qquad k = 1,...,M \tag{10}$$

따라서 화자의 대표값(D)은 식(9)로 사상된 k 번째 부대역의 값 $d_{map}^{(k)}$ 에 식 (10)번의 가중치 w_k 를 적용한 값으로 얻을 수 있다.

$$D = \sum_{k=1}^{M} w_k \cdot d_{map}^{(k)} \tag{11}$$

4. 실험 및 결과

4.1 실험 환경

실험에 사용된 데이터는 대학원 실험실 환경에서 수집된 한국어 문장 종속 연속음 "열려라 참깨"이다. 데이터는 1 주 간격의 시간차로 3 주에 걸쳐 수집하였다. 매주 1 회 발성에서는 각 5 번 발성을 하였으며, 개인별 전체 발성된 데이터 수는 15 개이다. 화자 인원수는 20 대 남여 각각 100 명이다. 샘플링 주파수는 16 kHz이고, 분해능은 16 bit이다. 등록 데이터는 처음 2 주간 발성된 10 번의 음성 데이터를 사용하였다. 음성 분석을 위하여 Hamming Window를 적용하였으며, 50% 중첩을 갖는 25.6 ms를 한 프레임으로 사용하였다. 특징벡터를 추출하기 위해 512-point FFT와 멜 스케일에서 균등한 24개의 채널을 가진 필터 뱅크를 사용하였다.

4.2 실험 방법 및 결과

실험은 기존의 방법인 전대역에서 추출된 특징벡터와 본 논문에서 제안한 방법인 다중대역에서 추출된 특징벡터의 성능을 비교하였다. 성능 비교는 화자를 사칭자로 판단하는 FRR (False Reject Rate)과 사칭자를 화자로 오인 판단하는 FAR(False Accept Rate)의 교차점 중 최소가 되는 EER(Equal Error Rate)이다.

<표 1>은 기존의 방법에서 혼합성분(Mixture)이 8과 17을 갖는 GMM과 BD 방법으로 거리 왜곡을 계산한 공분산 모델의 화자 확인의 성능과 다중대역에서 추출된 특징벡터에 사상 함수 및 가중치 함수를 적용한 EER로 화자 확인율을 나타내었다. 기존 방법에서 GMM은 혼합성분이 증가할수록 화자 확인률이 향상되었으며, 공분산 모델은 GMM보다 낮은 화자 확인률을 보임을 확인할 수 있다. 그리고, 다중대역에 기반한 방법에 사용된 가중치 함수는 각 부대역에 동일한 $1/M$ 을 사용한 EW(Equal Weight)와 [0:0.05:1]의 값들을 각 부대역에 적용하는데 그 가중치의 합이 1이 되는 1584개의 조합 중 가장 좋은 화자 확인률을 갖는 방법이며, 마지막은 본 논문에서 제안한 방법이다. GMM의 경우 기존의 방법을 사용한 성능보다는 저하되지만, 공분산 모델에서는 향상된 화자 확인률을 얻었다.

표 1. 기존의 특징벡터와 다중대역에 기반한 특징벡터의 화자 확인률(%)

	GMM		공분산 모델
	혼합성분 (8)	혼합성분(17)	BD
기존 방법	5.79	3.17	7.76
EW	7.03	6.55	3.81
1584	4.71	4.38	3.64
제안한 방법	5.04	4.78	3.76

<표 2>는 PCA 공분산 모델을 이용한 화자 확인율이다. 다중대역에서 각 부대역은 7차의 특징벡터로 6, 5, 4 차로 차원을 감소시키고, 기존의 방법은 다중대역 성능과 비교 분석에 편이성을 위해 24, 20, 16 차로 감소하였다.

표 2. PCA를 적용한 화자 확인률(%)

전대역의 PCA 차원 수			부대역의 PCA 차원수		
24	20	24	6	5	4
6.06	4.21	2.65	3.80	3.58	3.00

<표 3>은 기존 방법으로 추출한 특징벡터를 모델링한 경우와 논문에서 제안한 다중대역에 기반한 특징벡터로 모델링한 경우의 모델 파라미터 수를 비교하였다. 기존 방법인 경우, GMM은 $(2P_G+1)$차의 파라미터가 혼합성분 M_G개 만큼, 공분산 모델은 $(P_G * P_G)$차의 파라미터들을 요구한다. 본 논문에서 제안한 방법인 경우 PCA 공분산 모델은 $(P_p * P_p)$차의 파라미터들이 다중 대역의 수 M개 만큼 요구된다. 실험에서는 P_G=12, M_G=17, P_p=4, M=4이므로, 전대역의 GMM, 공분산 모델 그리고 다중 대역에서의 PCA 공분산 모델은 각각 425개, 144개, 64개의 특징 파라미터들이 요구되어 기존의 방법보다 약 85%의 계산량을 줄일 수 있는 이점을 갖는다.

표 3. 모델 파라미터 수

기존의 방법 (GMM)	기존의 방법 (공분산 모델)	제안한 방법 (PCA 공분산 모델)
$M_G(2P_G+1)$	$P_G * P_G$	$M(P_p * P_p)$

협대역 잡음이 부가된 음성 신호에서의 화자 확인 성능을 비교하기 위해서 랜덤 함수로 생성한 백색 잡음 신호를 FIR 필터(100차)를 통과시켜 400Hz의 대역폭의 협대역 잡음을 다중대역 중 첫번째 부대역 SB1(156.25~718.75 Hz)과 두번째 부대역 SB2(593.75~1906.25 Hz)에 SNR에 따라 부가하여 실험한 결과를 <표 4>에 나타내었다. SB1에 10[dB]의 협대역 잡음이 부가된 경우 기존 방법은 8.23%, 제안한 방법은 5.45%의 화자 확인률을 얻었으며, 동일한 방법으로 SB2인 경우 기존 방법은 11.01%, 제안한 방법은 5.62%의 화자 확인률을 갖음을 확인하였다. 동일한 대역폭의 협대역 잡음이 부가되더라도 기존의 방법은 부대역에 부가되는 협대역 잡음의 위치에 따라 성능 변화가 큰 반면 제안한 방법은 비슷한 성능을 얻을 수 있었다.

표 4. 400Hz의 협대역 잡음 환경에서의 화자 확인 결과(%)

	SB1에 400Hz 대역폭을 가진 협대역 잡음 부가		SB2에 400Hz 대역폭을 가진 협대역 잡음 부가	
SNR[dB]	기존 방법	제안한 방법	기존 방법	제안한 방법
5	9.97	6.51	13.76	6.79
10	8.23	5.45	11.01	5.62
15	6.74	4.61	9.46	4.86
20	6.85	4.31	8.05	4.38

5. 요약 및 결론

본 논문에서는 다중대역 분석법을 화자 확인 시스템에 적용하는 방법과 특징 벡터들의 상관관계를 제거한 PCA 공분산 모델을 제안하였다. 또한, 사상 함수를 적용하여 각 부대역을 동일 기준으로 이동한 후 가중치 함수를 적용하여 대표값을 얻었다. 기존의 방법으로 추출된 특징벡터의 화자 확인률은 GMM을 사용한 경우보다 저하된 결과를 얻지만, 본 논문에서 제안한 방법을 사용할 경우 비슷한 성능을 나타냈으며, 협대역 잡음이 부가된 음성에서는 50% 이상 향상된 화자 확인률을 나타냄을 확인했다. 또한, 다중대역에 기반한 PCA 공분산 모델을 사용함으로써 계산량을 줄일 수 있었다. 향후 여러가지 실생활 잡음에 오염된 Data의 화자 확인률을 확인해 볼 계획이다.

참 고 문 헌

이윤정, 서창우, 이기용. 2002. "강인한 주성분 분석법을 갖는 화자인식." *한국 음향학회 하계학술발표대회* 21(1s), 225-228.

Garcia, A. & Mammone, R. 1999. "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping." *Proc. ICASSP*, 325-328.

Lee, Y. J., Lee, J. H. & Lee, K. Y. 2003. "GMM based on local fuzzy PCA for speaker identification" *LNCS 2690*, 1000-1007.

Mak, B. 2002. "A mathematical relationship between fullband and multiband Mel-frequency cepstral coefficients." *IEEE Signal Processing Letter* 9(8), 241-244.

Juang, B. H. 1985. "Maximum-likelihood estimation for mixture multivariate stochastic observation of markov chains." *AT&T Technical Journal* 64, 1235-1240.

Yoshida, K., Takagi, K. & Ozeki, K. 2004. "Improved model training and automatic weight

adjustment for multi-SNR multi-band speaker identification." *ICSAP'04*, 3.

Petry, A., Zanus, A. & Barone, D. A. C. 2000. "Bhattacharyya Distance applied to speaker identification." *Int., Conference on Signal Processing Applications and Technology, Dallas, Orlando*, (1).

Reynold, D. A. & Rose, R. C. 1995. "Robust text-independent speaker identification using Gaussian mixture speaker models." *IEEE Trans. SAP* 3(1), 72-83.

Reynolds, D. A. 2003. "Channel robust speaker verification via feature mapping. "*ICASSP'03* 3, II-53-6.

Zilca, R. D. 2002. "Text-independent speaker verification using level scoring and covariance modeling." *IEEE Trans. Speech, and Audio Processing* 10(6), 363-370.

접수일자: 2007. 4. 27
게재결정: 2007. 5. 29

▲ 최민정
경기도 성남시 분당구 야탑동 342-1 리더스 B/D 305호(우: 463-828)
㈜인스모바일 주임연구원
숭실대학교 정보통신전자공학부 졸업(공학석사)
Tel: +82-31-703-7301 H/P: 011-9891-9708
Fax: +82-31-703-7301
E-mail: cmj1109@korea.com

▲ 이윤정
서울시 송파구 거여동 산25번지 (우: 138-110)
국방과학연구소 선임연구원
숭실대학교 정보통신전자공학부 졸업(공학박사)
Tel: +82-2-3400-2684 H/P: 017-728-1085
Fax: +82-2-403-3512
E-mail: youn@add.re.kr

▲ 서창우
서울 금천구 가산동 60-19 SJ테크노빌 15층 1510호 (우: 153-801)
㈜에스씨디 정보통신연구소 책임연구원
숭실대학교 전자공학과 졸업(공학박사)
Tel: +82-2-2106-2428(O) H/P : 016-289-3735
Fax: +82-2-2106-2400
E-mail: cwseo@sscd.co.kr