GMM 기반의 문맥독립 화자 검증 시스템의 성능 향상

함 성 준, 신 광 호, 김 민 정, 김 주 곤, 정 호 열, 정 현 열
영남대학교 정보통신공학과

Performance Improvement in GMM-based Text-Independent Speaker Verification System

Seong-Jun Hahm, Guang-Hu Shen, Min-Jung Kim, Joo-Gon Kim,
Ho-Youl Jung, Hyun-Yeol Chung
Dept. of Information and Communication Eng., Yeungnam University
E-mail: branden65@yumail.ac.kr

요 약

본 논문에서는 GMM(Gaussian Mixture Model)을 이용한 문맥독립 화자 검증 시스템을 구현한 후, arctan 함수를 이용한 정규화 방법을 사용하여 화자검증실험을 수행하였다. 특징파라미터로서는 선형예측방법을 이용한 켑스트럼 계수와 회귀계수를 사용하고 화자의 발성 변이를 고려하여 CMN(Cepstral Mean Normalization)을 적용하였다. 화자모델 생성을 위한 학습단에서는 화자발성의 음향학적 특징을 잘 표현할 수 있는 GMM(Gaussian Mixture Model)을 이용하였고 화자 검증단에서는 ML(Maximum Likelihood)을 이용하여 유사도를 계산하고 기존의 정규화 방법과 arctan 함수를 이용한 방법에 의해 정규화된 점수(score)와 미리 정해진 문턱값과 비교하여 검증하였다. 화자 검증 실험결과, arctan 함수를 부가한 방법이 기존의 방법보다 항상 향상된 EER을 나타냄을 확인할 수 있었다.

1. 서 론

화자 검증 시스템에서 검증 요구는 임의의 화자에 의해 이루어지며 임의의 화자의 발성은 사전에 등록된 화자의 모델과 비교된다. 화자 모델과의 점수(score)가 충분히 크면(미리 정해진 문턱값보다 크면) 검증 요구는 수락된다. 화자 검증 기술은 어떤 보호된 시스템에 접근하려는 사람들에 의한 검증 요구를 처리하기 위해 사용될 수 있다.[1] 이러한 화자 검증 시스템에는 수락오류(False Acceptances; FA)와 거부오류(False Rejections; FR) 두 가지 에러가 존재한다. 거부오류는 시스템이 등록된 화자를 거부할 때 발생하고, 수락오류는 사칭자(Imposter)를 수락할 때 발생한다. EER(Equal Error Rate)은 이 두 가지 에러가 같을 때의 에러율을 가리키며 화자 검증 시스템의 성능을 나타내는데 자주 이용된다[2].

화자 검증 시스템의 성능은 신호특성의 변이에 매우 크게 영향을 받으므로 화자 간의 변이를 보상하기위해 정규화 방법이 많이 이용되고 있다[3]. 또한, 문맥독립 화자인식방법으로서는 화자특성변이의 표현과 화자검증 율면에서 좋은 결과를 나타내고 있는 GMM에 의한 접근 방법이 가장 유리한 것으로 알려져 있다[4]. 대부분의 화자 검증 시스템의 정규화 방법은 검증을 요구한 화자와 백그라운드 화자(검증할 화자를 제외한 나머지 화자)의 유사도비에 기반을 둔다. 이러한 유사도비에 기반한 여러 가지 정규화 방법이 제안되었고, Dat Tran 등[5]은 정규화 방법으로 arctan[log(x)]에 기반한 방법을 적용하여 기존의 방법보다 향상된 결과를 얻고 있다.

따라서 본 연구에서는 이러한 유사도를 이용한 일반적인 정규화 방법과 arctan[log(x)]에 기반한 방법을 이용하여 GMM을 이용한 문맥독립 화자 검증 시스템을 구성한 후 검증실험을 수행하고 그 결과를 보고한다. 2장에서는 GMM 기반의 화자검증시스템에 대해 설명하고, 3장에서는 정규화방법에 대해 설명한다. 4장에서는 시스템의 구성 및 실험 결과에 대해 기술하고, 마지막으로 5장에서 결론을 맺는다.

2. 화자 검증 시스템

2.1 Gaussian Mixture Model(GMM)

화자 검증 시스템을 위한 GMM은 출력확률밀도함수가 가우시안밀도(Gaussian density)의 혼합인 1개의 상태만으로 이루어진 CHMM(Continuous HMM)의 한 형태이다. GMM의 장점 중 하나는 임의의 형태를 갖는 밀도를 부드럽게 근사화된 형태로 모델을 구성하는 것이다. 단일모드(Unimodal) 가우시안 화자모델은 평균벡터(mean vector)와 공분산(covariance)으로 화자의 특징분포를 표현하고, VQ-distortion 모델은 특징벡터의 이산집합으로 화자분포를 표현한다. 이들 두 모델의 특징을 고려하여 구성된 GMM은 각각의 평균과 공분산을 가진 가우시안 함수의 이산집합을 사용함으로써 화자의 특징을 더 잘 표현하는 모델의 생성을 가능하게 한다[6].

가우시안 혼합밀도는 성분(component) 밀도 M개의 가중합이고, 그 형태는 다음과 같다[7].

$$p(x|\lambda) = \sum_{i=1}^{M} c_i N(x; \mu_i, \Sigma_i) \qquad (1)$$

여기서 x는 d차원 랜덤벡터이고, $N(x;\mu_i,\Sigma_i)$, $i=1,2,\cdots,M$은 성분 밀도, $c_i, i=1,2,\cdots,M$은 혼합가중치(mixture weight)이다. 각 성분 밀도는 평균 μ_i와 공분산 Σ_i를 갖는 d차원 가우시안 함수이다.

$$N(x;\mu_i,\Sigma_i) = \qquad (2)$$
$$\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)\right\}$$

식 (1)의 혼합 가중치는 다음의 제한조건을 만족한다.

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

가우시안 혼합밀도는 각 성분 밀도의 혼합 가중치, 공분산행렬, 평균벡터로 구성되고, 다음과 같이 표현된다.

$$\lambda = \{c_i, \mu_i, \Sigma_i\} \quad i=1,2,\cdots,M \qquad (4)$$

화자모델학습의 목적은 주어진 학습음성으로부터 추출된 학습특징벡터의 분포를 가장 잘 표현하는 GMM, λ의 파라미터를 추정하는 것이다. GMM의 파라미터를 추정하기 위한 방법에는 여러 가지가 있지만, MLE(Maximum Likelihood Estimation) 방법이 가장 널리 쓰인다. MLE의 목적은 주어진 학습데이터와 GMM의 유사도를 최대화하는 모델 파라미터를 찾는 것이다. T개의 학습벡터열을 $X=\{x_1, x_2, \cdots, x_T\}$라고하면, GMM의 유사도는 다음 식과 같다.

$$P(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \qquad (5)$$

2.2 Speaker Verification

화자 검증은 화자의 발성이 검증이 요구된 화자에 속하는 지 아닌 지를 결정하는 것이다. 주어진 발성의 검증 요구에 대한 수락/거부는 유사도비 $L(X)$와 미리 정해진 문턱값 θ에 의해서 결정된다.

$$L(X)\begin{cases} \geq \theta \; accept \\ < \theta \; reject \end{cases} \qquad (6)$$

최소의 위험률을 갖는 Bayes decision rule에 따른 유사도비는 다음의 식이 사용된다.

$$L_1(X) = \frac{P(X|\lambda_0)}{P(X|\lambda)} \qquad (7)$$

여기서 λ_0는 검증이 요구된 화자모델이고, λ는 λ_0를 제외한 백그라운드 화자의 모델이다. 이 비율(ratio)은 화자발성변이에 덜 민감한 상대적인 점수(score)가 된다. 로그영역에서 식 (7)은 Higgins 등[3]에 의해 제안된 유사도 방법과 동일하다.

$$L_2(X) = \log P(X|\lambda_0) - \log P(X|\lambda) \qquad (8)$$

위 식에서 $\log P(X|\lambda)$는 정규화항을 나타낸다.

3. 정규화 방법

화자 검증 시스템에서 EER을 감소시키기 위한 여러 가지 정규화방법들이 제안되어 왔다. 화자 검증 시스템의 정규화 방법은 검증을 요구한 화자와 백그라운드 화자의 유사도비에 기반을 둔다. 이러한 방법 중의 하나로 기하평균을 이용한 정규화 방법이 Liu 등[8]에 의해 제안되었다.

$$L_3(X) = \log P(X|\lambda_0) \qquad (9)$$
$$-\frac{1}{B}\sum_{i=1}^{B} \log P(X|\lambda_i)$$

여기서 B는 백그라운드 화자의 수이다. 본 연구에서

사용한 방법은 식 (9)의 기존 log(x) 기반의 정규화 방법과 arctan[log(x)] 함수에 기반한 Dat Tran 등[5]에 의해 제안된 방법으로 그 식은 다음과 같다.

$$L_4(X) = \frac{\arctan\left[\log P(X|\lambda_0)\right]}{\frac{1}{B}\sum_{i=1}^{B}\arctan\left[\log P(X|\lambda_i)\right]} \qquad (10)$$

기존의 정규화 방법과 arctan[log(x)] 기반 방법의 평균과 분산에 따른 등록화자와 백그라운드 화자의 점수(score) 분포를 그림.1에 각각 나타내었다. 그림.1에서 나타난 바와 같이 arctan 함수 기반 정규화 방법의 분포가 기존의 분포보다 EER이 낮음을 알 수 있다. 이러한 결과는 arctan 함수를 부가함으로써 두 분포의 분산은 증가하지만, 각 분포의 평균이 더 멀어지는 데서 기인하며 이 결과로부터 화자검증시 검증률의 향상이 기대된다.

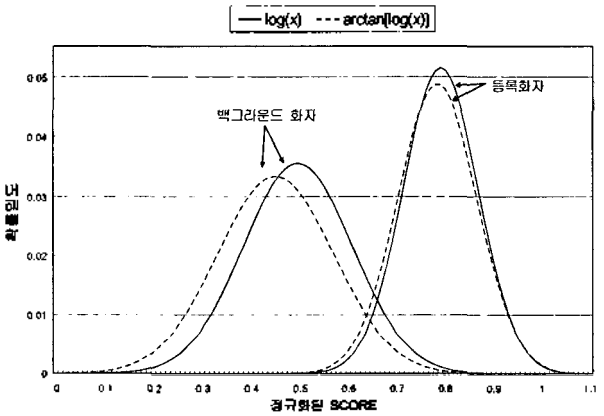

그림 1. 등록화자와 백그라운드 화자의 score 분포 (KLE452 학습프레임 4000, 테스트프레임 350의 예)

4. 시스템 구성 및 실험결과

4.1 데이터베이스(DB) 및 전처리

arctan 함수 기반 정규화법의 유효성을 확인하기 위하여 화자검증실험을 실시하였다. 실험에는 KLE452 DB와 ETRI445 DB, 이들 두 데이터와 KAIST 무역상담 DB를 혼합한 세 가지 형태의 데이터를 이용하였다. 화자검증 실험은 ETRI445의 남성 20명, KLE452 70명(남성 38명, 여성 32명), ETRI445와 KLE452, KAIST 무역상담 DB의 화자를 이용한 242명(KLE452 DB, 남녀 각각 1명 추가)에 대해서 수행되었다.

GMM 모델 학습을 위해서는 ETRI445와 KLE452의 경우는 250단어, KAIST 무역상담 데이터베이스의 경우는 60문장을 사용하여 최대10000프레임을 추출하여 파라미터로 이용하였다. 테스트를 위해서는 학습에 사용되지 않은 단어(또는 문장) 중에서 최대 1000프레임을 추출하여 이용하였다. 전처리단의 분석조건은 표 1과 같다.

표 1. 전처리단계의 조건

Sampling Rate	16kHz
Pre-emphasis coefficient	0.98
Hamming Window	yes
Frame length	320 samples
Frame Shift	160 samples
Cepstrum vector dimension	10

4.2 시스템의 구성

시스템은 화자모델 생성 및 화자 검증의 두 단계로 구성하였다.

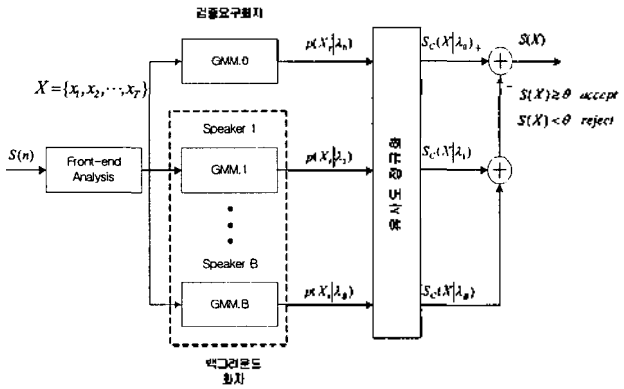

그림 2. 화자 검증시스템

그림 2는 화자 검증 시스템의 구성을 나타낸다. 화자 검증 실험에서 GMM의 혼합(mixture) 수는 32로 고정하였으며, GMM을 위한 학습프레임은 4000프레임부터 10000프레임까지, ML(Maximum Likelihood) 테스트를 위한 테스트프레임은 350프레임부터 1000프레임까지 증가시키면서 실험하였다. 화자 수에 따른 실험 결과를 표2, 3, 4에 각각 나타내었다.

표 2. ETRI445 20명의 화자 검증 실험 결과(EER(%))

테스트 프레임	학습 프레임							
	4000		6000		8000		10000	
	log	arctan	log	arctan	log	arctan	log	arctan
350	3.06	2.86	2.45	1.77	1.94	1.49	1.73	1.19
400	3.32	3.12	3.01	2.47	1.68	1.54	2.05	1.72
600	**1.71**	**1.44**	**1.51**	1.15	1.19	0.92	0.86	0.80
800	2.27	1.93	1.86	1.23	1.09	0.76	**0.53**	**0.47**
1000	2.33	1.80	1.71	**1.10**	**0.93**	**0.46**	0.68	0.55

표 3. KLE452 70명의 화자 검증 실험 결과(EER(%))

테스트 프레임	학습 프레임							
	4000		6000		8000		10000	
	log	arctan	log	arctan	log	arctan	log	arctan
350	4.33	3.29	3.22	2.53	2.71	1.95	2.35	1.62
400	4.32	3.32	2.97	2.19	2.59	1.88	2.25	1.56
600	2.33	1.99	1.99	1.38	1.91	1.23	1.43	1.04
800	2.11	1.46	1.29	0.89	1.11	0.61	1.14	0.54
1000	**1.23**	**1.00**	**0.96**	**0.74**	**0.81**	**0.54**	**0.68**	**0.47**

표 4. 총 242명(ETRI445 20명 + KLE452 72명+ KAIST 무역상당 DB 150명)의 화자 검증 실험 결과(EER(%))

테스트 프레임	학습 프레임							
	4000		6000		8000		10000	
	log	arctan	log	arctan	log	arctan	log	arctan
350	4.71	3.96	4.44	3.55	3.96	3.10	3.92	2.78
400	4.60	3.87	4.54	3.52	4.19	3.26	3.81	2.86
600	4.01	3.27	3.77	2.85	3.30	2.50	3.21	2.44
800	3.47	2.92	3.08	2.54	3.17	2.14	2.79	1.97
1000	**3.26**	**2.64**	**2.70**	**1.83**	**2.67**	**1.85**	**2.32**	**1.57**

표2-4로부터 화자수를 각각 20명, 70명, 242명으로 증가시킬 경우 평균 0.36%, 0.58%, 0.83%의 향상된 결과를 얻을 수 있음을 알 수 있다. 학습프레임과 테스트프레임이 증가함에 따라 EER율이 향상되지만, 프레임수를 증가시키지 않고도 arctan 함수 기반의 정규화 방법으로 최고 1.14%의 향상을 얻을 수 있었다. 또한 화자 수, 학습프레임, 테스트프레임에 관계없이 log(x)를 이용한 기존의 정규화 방법보다 항상 향상된 EER율을 나타냄을 알 수 있었다.

5. 결 론

본 논문에서는 GMM 기반의 문맥독립 화자 검증 시스템을 구성하여 기존의 log함수 정규화법에 arctan함수를 부가한 정규화 방법을 도입하고 이를 화자검증실험을 통하여 비교하였다. 화자 검증 시스템 구성을 위해 GMM을 이용하여 화자모델을 생성하였고, GMM의 파라미터 최적화를 위해서는 MLE(Maximum Likelihood Estimation)을 이용하였다. 이때 화자의 유사도 값은 ML(Maximum Likelihood) 방법을 이용하여 계산하였고, 이 유사도 값을 arctan 함수를 이용한 방법으로 정규화였다. 검증실험 결과를 비교한 검토한 결과, arctan 함수를 이용한 방법이 log 함수를 이용한 정규화 방법보다 항상 향상된 결과를 나타냄을 알 수 있어 화자 검증시스템의 정규화 방법으로 기존의 log함수에 arctan함수를 부가한 방법의 유효성을 확인할 수 있었다. 향후, cohort set을 이용한 화자 검증 방법에 대해 검토할 예정이다.

참고 문헌

[1] Sadaoki Furui, "Recent advances in speaker recognition ", Patt. Rec. Lett., Vol. 18, pp. 859-872, 1997.

[2] Konstantin Petrov Markov, 'Text-Independent Speaker Recognition based on Frame Level Likelihood Transformations', Toyohashi Univ. 1999.

[3] A. L. Higgins, L.Bahler and J. Porter, "Speaker Verification using Randomized Phrase Prompting", Digital Signal Processing, Vol. 1, pp. 89-106, 1991.

[4] S. Furui, "An overview of speaker recognition technology," in Acoustic speech and speaker recognition(C.-H Lee, F. K. Soong, and K. K. Paliwal, eds.), Ch. 2, pp. 31-56, Kluwer Acad. Pub., 1996.

[5] Dat Tran, Michael Wagner, "A Proposed Likelihood Transformation for Speaker Verification", in proc. IEEE International Conference on, vol. 2, pp. 1069-1072, 2000.

[6] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, Vol. 17, pp. 91-108, 1995.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. on SAP, Vol. 3, No. 1, pp. 72-83, 1995.

[8] C. S. Liu, H. C. Wang and C.-H. Lee, "Speaker Verification using Normalized Log-Likelihood Score", IEEE Trans. Speech and Audio Processing, Vol. 4, pp. 56-60, 1980.