발성변화에 강인한 화자 인식에 관한 연구☆

Safety Robust Speaker Recognition Against Utterance Variationsed

이 기 용*
Ki-Yong Lee

요 약

화자인식 시스템에서 화자 모델은 여러 세션동안 수집된 많은 양의 데이터 집합으로 등록한다. 많은 양의 데이터 집합은 많은 양의 메모리와 계산을 필요로 할 뿐 아니라, 게다가 사용자가 음성 등록을 위하여 여러 번에 걸쳐서 발성해야 하는 문제점이 있다. 최근, 이러한 문제를 보완하기 위해서 많은 적응 방법들이 제안되었다. 그러나, 여러 세션동안 모아진 데이터 집합은 불규칙한 발성 변화와 잡음 같은 이상치에 취약하고, 그것은 부정확한 화자 모델을 만든다. 본 논문에서는, GMM에 기초를 둔 화자 모델에 이상치들의 영향을 최소화하기 위한 적응 방법을 제안하였다. 강인한 적응은 M-추정의 점진적인 방법으로부터 얻어진다. 화자 모델은 초기에 적은 양의 데이터로 등록되어지고, 각각의 세션에서 얻어진 데이터로 반복적으로 적응시킨다. 실험 결과는 7개월에 걸쳐서 수집된 데이터 집합으로부터 제안된 방법이 이상치에 강인하다는 것을 보여준다.

Abstract

A speaker model in speaker recognition system is to be trained from a large data set gathered in multiple sessions. Large data set requires large amount of memory and computation, and moreover it's practically hard to make users utter the data inseveral sessions. Recently the incremental adaptation methods are proposed to cover the problems. However, the data set gathered from multiple sessions is vulnerable to the outliers from the irregular utterance variations and the presence of noise, which result in inaccurate speaker model. In this paper, we propose an incremental robust adaptation method to minimize the influence of outliers on Gaussian Mixture Model based speaker model. The robust adaptation is obtained from an incremental version of M-estimation. Speaker model is initially trained from small amount of data and it is adapted recursively with the data available in each session. Experimental results from the data set gathered over seven months show that the proposed method is robust against outliers.

Keyword : Speaker Recognition, GMM, M-estimation, Adaptation

1. 서 론

화자 인식 시스템에서 화자 모델은 여러 세션에서 발성된 많은 데이터 집합을 사용해서 등록한다[1,2]. 많은 데이터 집합은 화자 모델을 등록하는데 상당히 많은 양의 메모리와 계산을 필요로 한다. 또한, 실제 시스템에서, 사용자를 여러 세션에 걸쳐 많은 양의 데이터를 발성하게 하는 것은 어렵다. 최근에, 화자 적응 방법[3,4]과 점진적인 적응방법[5]들이 이러한 문제를 해결하기 위해 제안되었다. 점진적인 적응 방법에서, 화자 모델은 처음에 한 세션에서 발성된 적은 양의 데이터로 등록되어지고, 다음 세션들에서 발성되는 새로운 데이터로 증가적으로 갱신된다. 그러나, 다음 세션에서 발성된 음성 데이터가 이상치를 포함하고 있을 때, 기존의 적응 방법으로 얻어진 화자 모델은 부정확하게 된다. 이상치는 불규칙한 발성 변화와 잡음에서 발생 할 수 있다. 본 논문에서는, GMM[2]을 사용하고, 화자 모델에서 이상치의 영향을 최소화 시키기 위해서 점진적인 강인한 적응 방법을 제안하였다. GMM의 강인한 적응 방법은 M-추정의 점진적인 방법으로 얻을 수 있다. 초기의 화자 모델은 적은 양의 데이터로 등

* 정 회 원 : 숭실대학교 정보통신 전자공학부 교수
kylee@ssu.ac.kr(제 1 저자)
☆ 본 논문은 2004학년도 숭실대학교 교내학술연구비 지원에 의하여 수행되었습니다.

록되어지고, 새로운 데이터를 이용가능할 때마다 GMM의 파라메터들은 반복적으로 적응시킨다. 실험 결과는 제안된 방법이 이상치에 강인하다는 것을 보여주고 있다.

2 . GMM에 기초를 둔 강인한 화자모델

길이가 T_n인 N개의 등록 음성의 집합을 $Y_n = y_n(t), t=1,\cdots,T_n$이라고 두자. 여기에서, $y_n(t) \in R^L$이며, L차원 벡터이다. Y^N에 이상치가 존재할 때, 기존의 GMM의 파라메터 추정 과정은 이상치에 민감하다는 문제점을 공통적으로 가지고 있다. 그래서, GMM의 믿을 수 있는 추정을 얻기 위해서, M-추정 방법에 기초를 둔 강인한 추정을 아래와 같이 표현한다.

$$J = \sum_{n=1}^{N}\sum_{t=1}^{T_n} \rho[\log p(y_n(t) \mid \theta)] \tag{1}$$

위 식에서, $\rho[\cdot]$는 손실 함수이고, 이상치의 영향을 감소시키기 위해서 사용된다. 식(1)에서 가우시안 혼합성분 밀도 $p(y_n(t) \mid \theta)$는 M개의 다중 가우시안 함수들의 가중치 합으로 표현한다.

$$p(y_n(t) \mid \theta) = \sum_{i=1}^{M} p_i\, b_i(y_n(t)) \tag{2}$$

위식에서, $b_i(y_n(t))$는 평균이 μ_i이고 분산이 Σ_i인 가우시안 밀도 함수이다.

$$b_i(y_n(t)) = \frac{1}{(2\pi)^{\frac{L}{2}}|\Sigma_i|^{\frac{1}{2}}} \times \exp\left\{-\frac{1}{2}(y_n(t)-\mu_i)^T\Sigma_i^{-1}(y_n(t)-\mu_i)\right\}$$

화자 모델을 위한 강인한 GMM은 모든 성분 밀도로부터 구한 혼합 성분 가중치, 평균 벡터, 공분산 행렬로 파라메터화 된다. 이러한 파라메터들은 θ로 정의된다.

$$\theta = \{p_i, \mu_i, \Sigma_i\},\ i=1,\cdots,M$$

각각의 p_i, μ_i, Σ_i, $i=1,\cdots,M$에 대해서 (1)식의 J를 최소화 시킴으로써, θ의 재추정식을 찾을 수 찾을 수 있다. 각각 $\frac{\partial J}{\partial p_i}=0$, $\frac{\partial J}{\partial \mu_i}=0$, $\frac{\partial J}{\partial \Sigma_i}=0$ 일 때, 강인한 GMM을 위한 재추정식은 아래와 같이 얻어진다.

- 혼합성분의 가중치

$$p_i^N = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t)p(i \mid y_n(t), \theta)}{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t)} \tag{3.a}$$

- 평균

$$\mu_i^N = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t)p(i \mid y_n(t), \theta)y_n(t)}{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t)p(i \mid y_n(t), \theta)} \tag{3.b}$$

- 분산

$$\Sigma_i^N = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t)p(i \mid y_n(t), \theta)(y_n(t)-\mu_i)(y_n(t)-\mu_i)^T}{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t)p(i \mid y_n(t), \theta)} \tag{3.c}$$

위 식에서, $p(i \mid y_n(t), \theta)$는 클래스 i에서의 사후 확률이며, $p(i \mid y_n(t), \theta) = \frac{p_i b_i(y_n(t))}{\sum_{j=1}^{M} p_j b_j(y_n(t))}$이다. $w_{n(t)}$은 가중치 함수로 $z_n(t) = \log p(y_n(t) \mid \theta)$일 때, $w_n(t) = \frac{\partial \rho[z_{n(t)}]}{\partial z_n(t)}$로 정의된다. 본 논문에서는 $w_n(t) = 1/(1+z_n^2(t)/\beta)$로 주어지는 코시 가중치 함수(Cauchy's weight function)을 사용하였고, β는 스케일 파라메터이다[6]. 여기에서, 큰 $z_n(t)$는 작은 $w_n(t)$값을 가지기 때문에, 식(3)에서 이상치의 영향이 감소될 수 있다.

화자 인식을 위해서, 각각의 화자 S는 강인한 GMM들인 $\theta_1, \ldots, \theta_s$로 표현한다. 변형된 특징벡터열이 주어져 있을 때 사후 확률이 최대가 되는 화자 모델을 찾는다.

$$\hat{s} = \max \sum_{t=1}^{T} \log p(y_t \mid \theta_t) \tag{4}$$

3. 화자인식을 위한 점진적인 적응

만약 모델 파라메터 θ^N이 음성 데이터 Y^N의 초기 집합으로 등록되어 지고, 새로운 데이터 $Y_{N+1} = y_{N+1}(1), \ldots, y_{N+1}(T_{N+1})$이 주어진다면, $(N+1)$번째 반복적인 재추정식은 (3)식으로부터 얻을 수 있다.

- 혼합 성분 가중치

$$p_i^{N+1} = \frac{p_i^{N+1} W(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p(i \mid y_{N+1}(t), \theta^N)}{W(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t)} \tag{5.a}$$

- 평균

$$\mu_i^{N+1} = \frac{\mu_i^N W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p(i \mid y_{N+1}(t), \theta^N) y_{N+1}(t)}{W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p(i \mid y_{N+1}(t), \theta^N)} \tag{5.b}$$

- 분산

$$\Sigma_i^{N+1} = \frac{\Sigma_i^N W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p(i \mid y_{N+1}(t), \theta^N)(y_{N+1}(t) - \mu_i^N)(y_{N+1}(t) - \mu_i^N)^T}{W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p(i \mid y_{N+1}(t), \theta^N)} \tag{5.c}$$

위 식에서,

$$W(N) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} w_n(t)$$

$$W_p(N) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} w_n(t) p(i \mid y_{N+1(t)}, \theta^{N+1})$$

이다. $Y_{N+1} = y_{N+1}(1), \ldots, y_{N+1}(T_{N+1})$가 이상치를 포함하고 있을 때, $w_{N+1}(t)$이 작은 값을 갖게 되어 이상치의 영향이 감소하게 된다. Y_{N+2}가 주어져 있을 때, (5)식에서 $W(N+1)$과 $W_p(N+1)$은 반복적으로 아래식과 같이 얻을 수 있다.

$$W(N+1) = W(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) \tag{6.a}$$

$$W_p(N+1) = W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p(i \mid y_{N+1}(t), \theta^{N+1}) \tag{6.b}$$

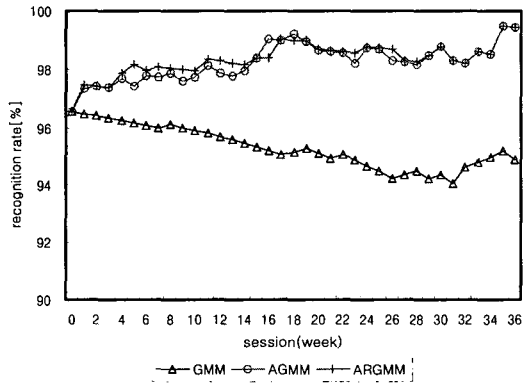
4. 실험 및 결과

문장 종속 화자 인식에서, 제안된 방법의 성능(ARGMM: Adaptive Robust GMM)을 검증하기 위해서 제안된 방법과 기존의 GMM에 기초한 방법들(GMM:GMM without Adaptation, AGMM: Adaptive GMM)을 사용하여 실험하였다. 음성 데이터는 12명의 화자(남자:7, 여자:5)로부터 얻어진 것이다. 각각의 화자에 대한 파일은 77개이며, 1세션을 1주일로하여, 총 37세션(37주)동안 처음에는 등록을 위해서 5문장을, 그 다음 세션는 한 세션에 2문장씩 발성했다. 음성의 샘플링 주파수는 11,025 kHz이고, 12차 LPC 켑스트럼 계수와 13차 델타켑스트럼을사용하였다. 음성 분석 창의 크기는 50% 중첩을 가진 20ms를 사용하였다. 초기 화자 모델은 첫번째 세션에서 발성된 5문장을 가지고 등록시켰다. GMM에서 혼합성분의 개수는 8이고, 손실 함수로 코시 가중치 함수를 사용하였다. 등록후, 초기의 테스트 과정에서는, 화자 인식 후 성공한 경우에만 각각의 세션에서의 발성데이터를 모델에 적응시켰다. 화자 적응은 발성 화자가 화자 모델로 맞게 확인 된 경우에만 이루어지는 슈퍼바이즈드 모드에서 이루어졌다.

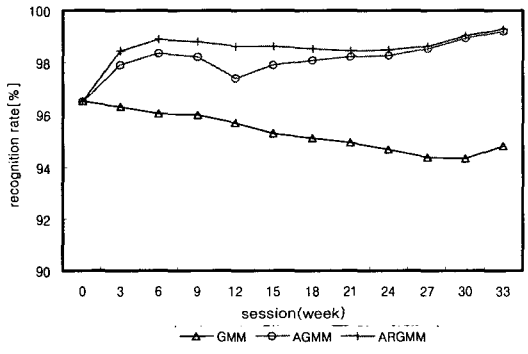
그림 1은 모든 세션에 대해서 GMM, ARGMM, AGMM의 화자 인식 성능을 보여주고 있다. 그림

에서 보면, GMM은 세션이 진행될수록 성능이 감소하고 있으나, ARGMM과 AGMM은 세션이 진행될수록 성능이 증가하는 것을 볼 수 있다. 초기의 세션에서, 제안된 방법이 AGMM 보다 더 좋은 성능을 보였다. 이것은 초기의 불안정한 훈련 모델이 이상치에 민감하기 때문이다. 그러나, 세션이 지날수록 ARGMM과 AGMM의 성능이 비슷함을 알 수 있었다.

그림 2는 3개의 세션마다 각각의 방법에 대한 화자 인식 성능을 보여주고 있다. 3주의 기간을 두고 발성한 데이터는 큰 발성변화를 포함하게 되는데, 이 경우에 ARGMM 방법이 AGMM보다 더 좋은 성능을 보였다.



〈그림 1〉 화자 인식 성능(매주마다) [%]



〈그림 2〉 화자 인식 성능 (3주마다) [%]

5. 결 론

본 논문에서는, GMM의 정확성에 발성변화의 영향을 최소화 시키기 위해서 점진적인 강인한 적응 방법을 제안하였다. 점진적인 강인한 적응 방법은 M-추정의 점진적인 방법으로부터 얻을 수 있다. 화자 모델은 초기에 적은 양의 음성 데이터로 훈련되어지고, 다음 세션에서 발성되는 음성 데이터를 이용가능한 경우만 반복적으로 적응시킨다.

참 고 문 헌

[1] Furui, S., "Cepstral analysis technique for automatic speaker verification", IEEE Trans. ASSP-29, vol 2, pp.254-272, 1981.

[2] Reynolds, D.A. and Rose, R.C., "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans Speech Audio Process., vol.3, no.1, pp.72-83, 1995.

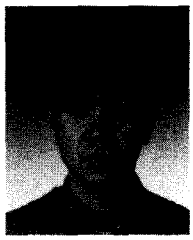
[3] Gauvain, J.L. and Lee, C.H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291-298, 1994.

[4] Ahn, S. and Ko, H., "Speaker adaptation in sparse training data for improved speaker verification", Electronics Letters, vol. 36, n0.4, pp.371-373, 2000.

[5] Fredouille, C. and Mariethoz, J., "Behavior of a Bayesian adaptation method for incremental enrollment in speaker verification", IEEE ICASSP 2000, vol.2, pp.1197-1200, 2000.

[6] Huber, P., Robust Statistics, New York: Wiley, 1981.

◐ 저 자 소 개 ◑



이 기 용

1983년 숭실대학교 전자공학과 졸업(학사)
1985년 서울대학교 대학원 전자공학과 졸업(석사)
1991년 서울대학교 대학원 전자공학과 졸업(박사)
1991년~1997년 8월 창원대학교 전자공학과 교수
1997년 9월~현재 숭실대학교 정보통신 전자공학부 교수
관심분야 : 음성신호처리, 음성향상, 화자인식, 신경망, 적응 신호처리 etc.
E-mail : kylee@ssu.ac.kr