

화자 인증에서의 효과적인 화자 적응과 *a priori* Threshold Updating에 관한 연구

조영훈*, 이수호*, 홍대희*, 고한석**

* 고려대학교 기계공학과, ** 고려대학교 전자공학과

A Study for Effective Speaker Adaptation and *a priori* Threshold Updating in Speaker Verification

Young-Hoon Cho*, Su-Ho Lee*, Dae-Hie Hong*, Han-Seok Ko**

*Department of Mechanical Engineering, Korea University

**Department of Electronics Engineering, Korea University

yhcho@ispl.korea.ac.kr

Abstract

실제 화자 인증기를 설계함에 있어서 발생하는 가장 큰 문제는, 적은 Enrollment data로 화자 모델이 만들어 지므로 화자 인증기의 성능이 시간이 지남에 따라 굉장히 줄어들게 되는 것과, 미리 훈련된 데이터 만으로 Threshold를 설정함에 따라 차후 실제 사용 시에 발생하는 변이를 고려하지 못하여 역시 성능 저하의 문제를 발생시킨다는 것이다.

위의 문제를 해결하기 위해 이 논문은 화자 모델을 구성하는데 있어 MAP 방법을 적용하고, threshold를 Resetting하는 방법을 적용했다. 본 논문에서 제안한 방법으로 HTER값이 23%정도 줄어들을 보여준다.

1. 서 론

근래에 지문이나 흉채, 얼굴, 음성, DNA 정보 등 개인의 고유한 특성을 이용하여 사람의 신원에 대한 정보를 알아내고자 하는 연구가 활발히 진행되고 있다. 그 중에서 인간의 고유한 음성을 이용하여 발성한 화자에 대한 정보를 알아내는 연구가 바로 화자인식 기술이다. 화자인식 기술은 task의 성격에 따라 화자식별과 화자확인으로 나뉘어지는데, 본 실험은 이중 화자확인에 대한 내용이고, 화자 확인 시스템의 기본 구조는 그림 1과 같다.

테스트 할 화자의 음성은 먼저 끝점 검출이 되고 특징 파라미터의 추출을 통해 분석된다. 추출된 특징 파

라미터는 기존에 저장되어 있는 확인을 바라는 의뢰인의 화자 모델과 비교를 한다.

이렇게 해서 나온 Likelihood를 계산하여 미리 정해진 threshold와 비교하여 그 의뢰인을 수락(Accept)할 것인지 거부(Reject)할 것인지를 판단한다[3].

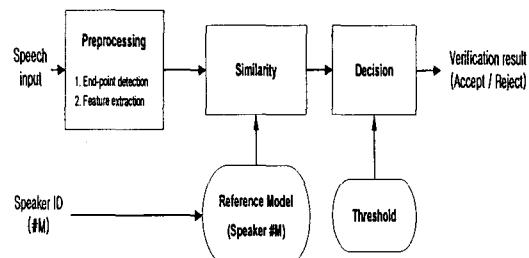


그림 1. 화자확인 시스템의 기본 구조

위의 화자 확인 과정에서 발생하는 문제는 적은 훈련데이터로 화자 모델이 만들어지므로 화자의 발음 변이를 잘 반영하지 못하고, 또 미리 정해진 threshold를 이용해 수락, 거부를 결정함으로 인해 차후 의뢰인의 사칭여부를 판단할 때 시간이 지남에 따라 성능 저하가 현격히 증가하게 된다[2].

이런 문제를 해결하기 위해, 먼저 효과적인 화자 모델을 만들기 위해 MAP method를 이용 화자 모델을 upgrade하고 새롭게 들어오는 data를 이용 threshold value를 updating하는 방법을 제안한다.

2. 화자 적용

음성인식이나 화자확인 시스템이 학습된 환경과 실제 적용되는 환경이 다를 경우 성능이 크게 감소한다. 이와 같은 문제를 해결하기 위해 사용되는 방법이 적용(adaptation) 방법이다. 그 중 화자 적용이란 먼저 만들어 놓은 화자 독립 모델을 원하는 특정 화자의 적용은 학습 데이터를 이용하여 특정화자에 대한 모델로 만드는 방법이다.

근래에 시스템의 성능 향상을 위해 화자 적용에 대한 연구가 활발히 진행되고 있다[1][5]. 그 중에서도 Maximum A Posteriori(MAP)와 Maximum Likelihood Linear Regression(MLLR), 그리고 Stochastic Matching 방법 등이 많이 연구되고 있는데 여기에서는 이번 논문에서 사용된 MAP 방법에 대해서 알아보겠다.

2.1 MAP 적용 방법

MAP Adaptation의 기본 아이디어는 학습과정에서 prior 정보를 이용하여 불충분한 학습 데이터의 문제를 해결하는데 있다. 즉, 새롭게 얻어지는 특정화자의 데이터를 기준의 존재하는 모델과 융합해 optimal한 값을 얻어내는 방법이다[6].

이 논문에서는 효과적인 화자 모델을 만들기 위해 segmental MAP adaptation을 적용하였다. 그리고 초기 상태 존재확률과 상태 천이 확률, Gaussian mixture의 covariance는 고정시켜 놓고 단지 Gaussian mixture의 mean과 weight만 adaptation하였다. 그 이유는 적은 데이터로 variance를 adaptation하는 것은 어렵고, 오히려 성능 저하를 초래할 수도 있기 때문이다[4].

Adaptation data의 숫자가 V 일 때, 주어진 학습 데이터를 이용하여 batch 형태의 adaptation을 적용했고 이때 사용한 식은 다음과 같다.

$$\hat{w}_{ik} = \frac{(v_{ik} - 1) + \sum_{v=1}^V \sum_{t=1}^T c_{ikt}^{(v)}}{\sum_{k=1}^K (v_{ik} - 1) + \sum_{k=1}^K \sum_{v=1}^V \sum_{t=1}^T c_{ikt}^{(v)}} \quad (1)$$

$$\hat{m}_{ik} = \frac{\tau_{ik}\mu_{ik} + \sum_{v=1}^V \sum_{t=1}^T c_{ikt}^{(v)}x_t}{\tau_{ik} + \sum_{v=1}^V \sum_{t=1}^T c_{ikt}^{(v)}} \quad (2)$$

식 (1)에서 w_{ik} 는 state i 의 k 번째 mixture의 weight이고, c_{ikt} 는 시간 t 에서 state i 의 mixture k 에 있을 확률을 나타낸다.

식 (2)에서 볼 수 있듯이 adaptation된 mean \hat{m}_{ik} 은 prior distribution의 mean μ_{ik} 와 입력데이터 x_t 의 mean과의 τ_{ik} 에 의한 weighted sum이다. 여기서 τ_{ik}

는 adaptation의 빠르기를 나타내는 변수이다. 이 값이 크면 adaptation이 천천히 이루어지고, 반대로 작으면 adaptation이 빠르게 이루어 진다. 그리고 v_{ik} 는 식 (3)과 같이 estimation된다.

$$v_{ik} = \tau_{ik} + 1 \quad (3)$$

위의 식 (1)과 식 (2)를 이용한 MAP adaptation 을 화자확인의 화자 모델을 만드는데 적용했다.

3. Resetting the Threshold Value

화자 확인 시스템 구현 시 미리 결정된 threshold value를 이용해 의뢰인의 사칭 여부를 판단하므로 발생되는 시간 흐름에 따른 지속적인 성능 저하를 막기위한 방법으로 threshold value를 resetting하는 방법을 제안 한다.

그림 2에서의 False Reject Error Rate(FRR) curve의 shift width는 데이터의 양이 많아짐에 따라 줄어든다 [2]. 즉, 발음 변이에 따른 robustness가 증가함을 알 수 있다. 그러므로 updating을 위한 데이터로부터 계산되는 threshold value는 점진적으로 high False Accept Error Rate(FAR)을 통과했던 값으로부터 EER을 통과하는 값으로 근접하게 된다(식 (4)).

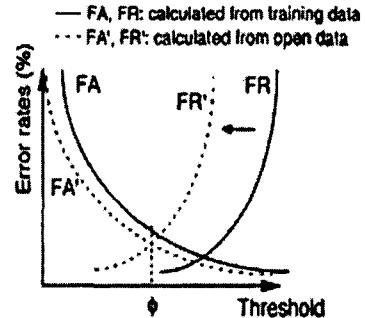


그림 2. training & open data로부터 계산된 threshold 와 FA & FR rates 의 관계

$$\tilde{\phi} = w\phi_1 + (1-w)\phi_0 \quad (4)$$

여기서, ϕ_0 는 updating을 위한 데이터가 EER값을 지난 Threshold value를 나타내고, ϕ_1 은 high FA rate를 지나는 threshold value(set experimentally)이다(그림 3).

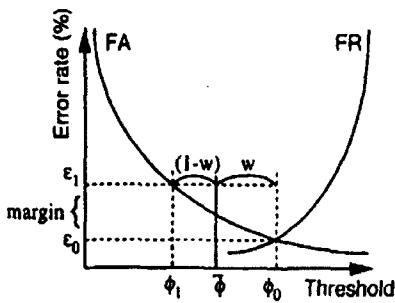


그림 3. *a priori* threshold를 updating하는 방법

$$w = \frac{2}{1 + \exp(a \cdot k)} \quad (5)$$

여기서, w 는 threshold가 EER을 지나는 value에 접근하는 속도를 조절하는 파라미터이고, a 는 free 파라미터이며, k 는 화자모델에 대해 발생한 updating의 횟수를 나타낸다(그림 4).

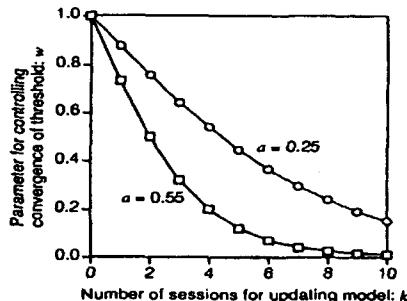


그림 4. Threshold의 수렴을 조절하는 parameter : w

4. 실험 환경

본 논문에서는 성능 향상을 평가하기 위해 문맥 종속형 화자확인 실험을 수행하였다.

사용된 음성 데이터베이스는 남자 25명, 여자 25명, 총 50명이 발음한 한국어 단어로 구성되었다. 각 화자는 자신의 비밀단어에 대해 15번 발음하였고, 다른 사람의 비밀 단어는 3번씩 발음하였다. 음성은 마이크를 통해 컴퓨터로 녹음되었고, 8 kHz로 샘플링 되었다.

음성 특징으로는 12차원의 MFCC와 1차 미분, 2차 미분 그리고 zero mean으로 총 36차의 특징벡터를 사용하였다. 훈련에는 자기 비밀 단어 5회 발음 분을 사용했고, adaptation을 위해 자기 비밀 단어 5회 발음 분을 그리고 테스트에는 FAR을 얻기 위해 다른 사람

단어 3회 발음분, FRR을 얻기 위해 자기 비밀 단어 5회 발음 분을 사용하였다.

Normalization factor로는 world 모델을 사용했고, 화자모델과 world 모델은 3개의 state를 가지는 Left-Right CDHMM으로 각 state는 여러 개의 Gaussian mixture로 이루어졌다.

이번 실험에서는 S/W로 HTK 3.0을 사용하였다.

5. 실험 결과

5.1 Gaussian Mixture의 수 변화에 따른 성능 평가

이 실험은 제안된 방법을 사용하지 않고 화자모델을 만들어서 수행한 화자 확인 실험으로 다른 실험에 대한 기준실험으로 수행되었다.

3개의 state를 가지는 Left-Right CDHMM을 이용해 각 state가 가지는 mixture 수 변화에 따른 성능을 비교해 보았다.

# of mix.	2	4	8	16
EER(%)	1.99	1.60	2.45	3.29

표 1. Mixture수 변화에 따른 EER(%)

Mixture의 수가 4일 때 가장 좋은 성능을 보여줄 수 있다. 차후 모든 실험은 mixture를 4로 놓고 실험을 진행하였다. 위의 결과로 mixture의 개수가 많아지면 적은 학습 데이터로부터 각 모델 파라미터들을 추정하는데 있어 정확성을 오히려 손상시킬 수 있음을 알 수 있다.

5.2 MAP adaptation을 통한 성능 평가

3개의 state를 가지는 Left-Right CDHMM을 이용해 각 state가 가지는 mixture수를 4로 놓은 상태에서 MAP adaptation을 한 경우의 성능을 비교해 보았다.

본 실험에서는 총 15회의 자기 비밀 단어 발음 데이터 중 5회의 자기 비밀 단어 발음 데이터로 훈련하고, 5회의 발음 데이터로 테스트 하였으며, 나머지 5회의 발음 데이터를 1개씩 차례로 총 5회의 adaptation을 수행하였다.

실험에서 threshold value는 adaptation전에 결정된 EER보다 FA rate이 약간 높은 곳을 지나는 value를 설정했고, variance는 같은 값으로 고정하고, mean과 weight를 adaptation 하였다.

# of Adaptation	No MAP	1	2
FAR/FRR	3.52/1.00	3.63/0.80	3.58/1.00
HTER(%)	2.26	2.22	2.29

# of Adaptation	3	4	5
FAR/FRR	3.36/0.80	3.40/0.80	3.07/0.80
HTER(%)	2.08	2.10	1.94

표 2. MAP를 이용한 화자확인 HTER(%)

위의 결과로부터 MAP를 하지 않을 경우보다 MAP를 수행함으로써 HTER값이 2.26%에서 1.94%로 14%정도의 성능 향상이 있음을 알 수 있었다.

5.3 Resetting the threshold value의 효과

ϕ_0 는 updating을 위한 데이터(5회 발음분)에서 얻어진 EER값을 지나는 Threshold value로 설정해주고, ϕ_1 은 high FA rate를 지나는 threshold value(set experimentally)로 설정해서 실험하였다[2].

Updating의 횟수는 data의 부족으로 5회 발음 data를 가지고 1회만 update 했다.

# of Update(k)	0	1
FAR/FRR	3.07/0.80	2.26/1.20
HTER(%)	1.94	1.73

표3. Threshold resetting 방법을 사용했을 때의 HTER(%)

위의 표는 MAP만 수행하고 threshold value를 resetting하지 않은 결과와 threshold value를 resetting 한 결과를 비교한 것인데, 약 11%의 성능 향상을 얻을 수 있었다. 결과 5.2의 표 2를 참조해 MAP도 수행하지 않았을 때와 인식 결과를 비교해 보면 2.26%에서 1.73%로 23%의 성능 향상을 얻을 수 있음을 알 수 있다.

6. 결 론

본 논문은 화자 확인 시스템을 상용화하는데 있어 불충분한 학습 데이터를 가지고 화자모델을 만들 때 야기되는 문제점을 해결하기 위해 MAP 화자적응 방법을

적용했고, 또 기존의 데이터로 훈련 후 결정된 threshold value가 차후 새롭게 들어오는 음성 data를 잘 반영하지 못하는 점을 개선하기 위해 Threshold value를 Resetting하는 방법을 동시에 적용했다. 그래서 전체적으로 HTER값이 2.26%에서 1.73%로 23% 향상되는 결과를 얻을 수 있었다.

차후 MAP와 MLLR을 조합하는 방법을 이용한 실험을 할 예정이며, 시간 변이에 따라 수집된 음성 db를 이용한 실험이 이루어져야 할 것이다.

참 고 문 헌

- [1] H. Bourlard & N. Morgan, "Speaker Verification : A Quick Overview", *IDIAP Research report*, Aug. 1998.
- [2] T. Matsui, "A study of models and a priori threshold updating in speaker verification", *Systems and Computers in Japan*, Vol.30, No. 13, 1999.
- [3] Jayant M. Naik, "Speaker Verification: A Tutorial", *IEEE Communications Magazine*, Jan. 1990.
- [4] S. Ahn, "Effective Speaker Adaptations for Speaker Verification", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp.1081-1084, 2000.
- [5] S. Furui, "Recent Advances in Speaker Recognition", *Pattern Recognition Letters* 18, pp.859-872, 1997.
- [6] J.L. Gauvain and C. H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, April 1994.