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

Speaker adaptation is a technique that adjusts the model of a speaker to improve recognition accuracy. This paper proposes a new method of adapting a speaker's model using a priori threshold updating. The method is based on the likelihood ratio test, which compares the likelihood of a given speech sample under two different models: the enrollment model and the test model. The threshold is determined based on the prior knowledge of the expected performance of the system.

1. 서 론

이론적 지식으로는 음성프로세싱, 음성인식, DNA 정보 등과 개인의 고유한 특성을 이용하여 사람의 신원에 대한 정보를 알아내고자 하는 연구가 활발히 진행되고 있다. 그 중에서 인간의 고유한 음성을 이용하여 발생한 화자의 신원에 대한 정보를 알아내는 연구가 밀접히 연결되며, 이에 따라 화자인식 기술이 발전하고 있다. 본 논문은 이에 이어 화자인식에 대한 내용이고, 화자인식 시스템의 기본 구조는 그림 1과 같다.

그림 1. 화자인식 시스템의 기본 구조

위의 화자 확인 과정에서 발생하는 문제는 적출된 문맥 데이터로 화자 모델이 만들어지므로 화자의 발음 범위가 잘 반영하지 못하고, 또 다른 문제로 threshold를 이용해 수락, 거부를 결정할 경우, 착각이 발생하여 화자인식의 성능 저하가 현저하게 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 화자확 인식 시스템은 성능 차이의 다양한 결과가 발생하는 환경을 구현하기 위해 사용하는 방법 중 하나 'adaptation' 방법입니다. 이 방법은 화자의 특성에 따라 모델을 생성하거나 특정 화자의 특성을 이용하여 특정 화자에 대한 모델로 만드는 방법입니다.

그러면 이를 통해 화자에 따라 모델을 생성하거나 특정 화자의 특성을 이용하여 특성화자에 대한 모델로 만드는 방법입니다. 

2.1 MAP 적응 방법

MAP Adaptation의 기반 아이템은 학습과정에서 모델을 생성하거나 특성화자에 대한 모델이 생성되는 환경을 구현하기 위해 사용하는 방법입니다. 이를 통해 학습모델과 특정화자에 대한 모델이 생성되는 환경을 구현하기 위해 사용하는 방법입니다.

이 논문에서는 효과적인 모델 생성을 위해 segmental MAP adaptation을 적용하였습니다. 그리고 초기 환경화와 이에 따라 모델을 생성하거나 특정화자에 대한 모델이 생성되는 환경을 구현하기 위해 사용하는 방법입니다. 

\[
\hat{w}_k = \frac{(v_k - 1) + \sum_{t=1}^{T} \sum_{i=1}^{N} c_{ik}^{(t)}}{(v_k - 1) + \sum_{t=1}^{T} \sum_{i=1}^{N} c_{ik}^{(t)}} \quad (1)
\]

\[
\hat{m}_k = \frac{\tau_k \mu_k + \sum_{t=1}^{T} \sum_{i=1}^{N} c_{ik}^{(t)} x_i}{\tau_k + \sum_{t=1}^{T} \sum_{i=1}^{N} c_{ik}^{(t)}} \quad (2)
\]

식 (1)에서 \(w_k\)는 state \(i\)의 \(k\) 번째 mixture의 weight이고, \(c_{ik}^{(t)}\)는 시간 \(t\)에서 state \(i\)의 mixture \(k\)에 있는 확률을 나타냅니다.

식 (2)에서 \(m_k\)은 adaptation된 mean \(\hat{m}_k\)은 prior distribution의 mean \(\mu_k\)와 입력데이터 \(x_i\)의 mean과의 \(\tau_k\)에 의한 weighted sum입니다. 여기서 \(\tau_k\)

는 adaptation의 빠르기를 나타내는 변수이다. 이 값이 크면 adaptation이 천천히 이루어지고, 반대로 적으면 adaptation이 빠르게 이루어진다. 그리고 \(v_k\)는 식 (3)과 같이 estimation된다.

\[
v_k = v_k + 1 \quad (3)
\]

위의 식 (1)과 식 (2)를 이용한 MAP adaptation을 화자정의의 화자 모델을 만드는데 적용했다.

3. Resetting the Threshold Value

화자 확인 시 사용한 threshold value를 이용하여 동일한 사용자의 사양을 판단하기로 발생되는 시간 초과에 따른 지속적인 성능 저하를 막기위한 방법으로 threshold value를 resetting하는 방법을 제안한다.

그림 2에서의 False Reject Error Rate (FRR) curve의 shift width는 데이터의 양이 다르기 때문에 줄어들었다. 즉, 발생 범위에 따른 robustness가 증가할 수 있다. 그러므로 updating을 위한 데이터로부터 계산되는 threshold value는 점진적으로 high False Accept Error Rate (FAR)를 통과했던 값으로부터 EER을 통과하는 값으로 근접하게 된다 (식 (4)).

\[
\tilde{\phi} = w \phi_i + (1 - w) \phi_0 \quad (4)
\]

여기서, \(\phi_0\)는 updating을 위한 데이터가 EER값을 가지는 threshold value를 나타내고, \(\phi_i\)은 high FA rate를 가지는 threshold value(set experimentally)이다 (그림 3).
4. 실험 환경

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

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

음성 특징으로는 12차원의 MFCC와 1차, 2차 미분 그리고 zero mean으로 총 38차의 특징벡터를 사용하였다. 훈련에서는 자기 비밀 단어 5회 발음을 복을 사용했고, adaptation을 위해 자기 비밀 단어 5회 발음 복을 그리고 테스트에는 FAR를 얻기 위해 다른 사람

이 3회 발음본, FRR를 얻기 위해 자기 비밀 단어 5회 발음 복을 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 제한된 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.

<table>
<thead>
<tr>
<th># of mix.</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER(%)</td>
<td>1.99</td>
<td>1.60</td>
<td>2.45</td>
<td>3.29</td>
</tr>
</tbody>
</table>

표 1. Mixture수 변화에 따른 EER(%) の 수가 4일 때 가장 좋은 성능을 보여주므로 알 수 있다. 차후 모든 실험은 mixture를 4로 높고 실험을 진행하였다. 위의 결과로 mixture의 개수가 많아 지면 적은 학습 데이터로부터 각 모델 파라미터들은 추정하는데 있어 정확성을 오히려 저하시킬 수 있음을 알 수 있다.

5.2 MAP adaptation을 통한 성능 평가

3개의 state를 가지는 Left-Right CDHMM을 이용해 각 state가 가지는 mixture수를 4로 늘린 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 rate가 적은 것을 지나는 value를 설정하고, variance는 같은 값으로 고정하고, mean과 weight를 adaptation 하였다.
<table>
<thead>
<tr>
<th># of Adaptation</th>
<th>No MAP</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAR/FRR</td>
<td>3.52/1.00</td>
<td>3.63/0.80</td>
<td>3.58/1.00</td>
</tr>
<tr>
<td>HTER(%)</td>
<td>2.26</td>
<td>2.22</td>
<td>2.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># of Adaptation</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAR/FRR</td>
<td>3.36/0.80</td>
<td>3.40/0.80</td>
<td>3.07/0.80</td>
</tr>
<tr>
<td>HTER(%)</td>
<td>2.08</td>
<td>2.10</td>
<td>1.94</td>
</tr>
</tbody>
</table>

표 2. MAP을 이용한 화자 확인 HTER(%)

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

5.3 Resetting the threshold value의 효과

\( \phi_0 \)는 updating을 위한 데이터(5회 받음)에서 얻어진 EER값을 지나는 Threshold value로 설정해주고, \( \phi_1 \)은 high FA rate를 지나는 threshold value(set experimentially)로 설정해서 실험하였다[2].

Updation의 횟수는 data의 부족으로 5회 받음 data를 가지고 1회만 update 했다.

<table>
<thead>
<tr>
<th># of Update(K)</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAR/FRR</td>
<td>3.07/0.80</td>
<td>2.26/1.20</td>
</tr>
<tr>
<td>HTER(%)</td>
<td>1.94</td>
<td>1.73</td>
</tr>
</tbody>
</table>

표 3. Threshold resetting 방법을 사용했을 때의 HTER(%)

위의 표는 MAP만 수행하고 threshold value를 resetting하지 않은 결과와 threshold value를 resetting한 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. Conclusion

본 논문은 화자 확인 시스템을 상용화하는데 있어 높은 성능을 요구하는 화자 모델을 만들 때 야기되는 문제점을 해결하기 위해 MAP 화자 확인 방법을 적용했고, 또 기존의 데이터로 훈련 후 결정된 threshold value가 차후 새롭게 들어오는 응성 data를 잘 반영하지 못하는 점을 개선하기 위해 Threshold value를 Resetting하는 방법을 동시에 적용했다. 그에 따라 전체적으로 HTER값이 2.26%에서 1.73%로 향상되는 결과를 얻을 수 있었다.

차후 MAP와 MLLR를 조합하는 방법을 이용한 실험이 필요하며, 시간 끝에 따라 수집된 음성 db를 이용한 실험이 이루어져야 할 것이다.

참고 문헌


