An improvement of the GMM Speaker Identification Method by using Two-state HMM and Discriminative Training

Yih-Ru WANG, and Shih-Ming FAN

Department of Communications Engineering,
National Chiao Tung University, Hsinchu
yrwang@cc.nctu.edu.tw

ABSTRACT

In this paper, the GMM-based text-independent speaker identification system for Mandarin speech is modified by adding an upper layer to form a two-state HMM system. The two-state HMM aims at modeling the initial-final phonetic structure of Mandarin syllables for assisting in speaker identification. The GPD/MCE training algorithm is also applied to further improve the system. The performance of the proposed system was examined by using a 300-speaker speech database. Error rate reductions of 25-50% were achieved for the proposed two-state HMM system over the conventional GMM system.

1. INTRODUCTION

GMM (Gaussian mixture model) classifiers have been shown to be effective and robust on text-independent (TI) speaker identification/verification [1,2]. In a GMM classifier, a GMM with several mixture components is used to model the acoustic characteristics of each speaker. The GMM can be regarded as using a single-state HMM to model the property of all phones in the speech signals of the speaker. But it is generally difficult to use HMM with multiple states in this case because the input text is unconstrained and the training data is usually limited and not large enough to be used to model the detail of the phonetic structure of speech signals.

In Mandarin, each character is pronounced as a syllable which can be decomposed into two parts: an optional initial and a final. By taking advantage of the simple initial-final syllable structure, we propose in this paper a method to improve the GMM TI speaker identification system for Mandarin speech by adding an upper layer to form a two-state HMM system. The two-state HMM aims at modeling the initial-final phonetic structure of Mandarin syllables. With the incorporation of upper-level acoustic information, it is expected that the performance of the GMM speaker identification system can be improved.

The other improvement discussed in this paper is via the use of discriminative training. A segmental Generalized Probabilistic Descent (GPD) discriminative training algorithm was used in GMM speaker identification to minimize the classification error [3]. About 10% reduction of speaker identification error rate was reported. The MCE/GPD discriminative training algorithm is also used in the study to re-train the two-state HMM speaker identification system for further improving its performance.

The remainder of the paper is organized as follows. In Section 2, the principle of the conventional GMM-based speaker identification system and the GPD/MCE training algorithm are described first. Then, the proposed two-state HMM system is presented. In Section 3, the performance of the proposed method is examined by using a 300-speaker microphone speech database (TCC-300). Some conclusions are given in the last section.

2. THE GMM SPEAKER IDENTIFICATION SYSTEM AND THE GPD/MCE ALGORITHM

2.1 The Basic GMM System

In a conventional GMM speaker identification system, a GMM with several mixture components is used to model the speech signals of a speaker, i.e.,

\[ p(x | \lambda_j) = \sum_{i=1}^{N} c_i^{(j)} p(x_i | \mu_i, \Sigma_i) = \sum_{i=1}^{N} c_i^{(j)} N(\mu_i^{(j)}, \Sigma_i^{(j)}) , \]

where \( x \) is the input feature vector with dimension \( D \), \( N(\mu_i^{(j)}, \Sigma_i^{(j)}) \) is a Gaussian distribution with mean vector \( \mu_i^{(j)} \) and covariance matrix \( \Sigma_i^{(j)} \), and \( \lambda_j \) is the model parameters of speaker \( j \). Here, diagonal covariance matrices are used.

The GMM model can be trained by the EM (expectation-maximization) algorithm with parameters being updated by the following formulas,

\[
\hat{c}_i^{(j)} = \frac{1}{T} \sum_{n=1}^{T} p(\{x_n, \lambda_j\}) = \frac{1}{T} \sum_{n=1}^{T} c_i^{(j)} p(x_n | \mu_i, \lambda_j) / \sum_{i'} c_{i'}^{(j)} p(x_n | \mu_{i'}, \lambda_j),
\]

\[
\hat{\mu}_i^{(j)} = \frac{1}{T} \sum_{n=1}^{T} c_i^{(j)} p(x_n | \mu_i, \lambda_j) x_n / \sum_{i'} c_{i'}^{(j)} p(x_n | \mu_{i'}, \lambda_j).\]

The GPD/MCE discriminative training algorithm was used in the study to re-train the two-state HMM speaker identification system for further improving its performance.
\[ \hat{c}_i^{(j)} = \frac{1}{T} \sum_{n=1}^{T} c_i^{(j)} p(x_n | \hat{\lambda}_j) \]  \\
\[ \sum_{i} c_i^{(j)} p(x_n | \hat{\lambda}_j) \]  \\
\[ \frac{[A]_{mm}}{[v]_m} \]  \\
\[ \text{where } [A]_{mm} \text{ is the } m\text{th diagonal element of matrix } A \text{ and } [v]_m \text{ is the } m\text{th element of vector } v. \]

For an input testing speech signal, \( X = (x_1, x_2, \cdots, x_T) \), the maximum likelihood (ML) classification criterion is used to determine the speaker, i.e.,

\[ \hat{S} = \arg \max_j \log p(X | \hat{\lambda}_j) \]

\[ = \arg \max_j \sum_{n=1}^{T} \log p(x_n | \hat{\lambda}_j) \]

where \( S \) is the total number of speakers.

### 2.2 The GPD/MCE Training Algorithm

In the above GMM identifier, all speaker models are trained with the ML criterion. By using the minimum classification error (MCE) criterion, the GPD discriminative training algorithm [4] can be used to retrain these GMM models. The GPD/MCE algorithm is discussed as follows.

First, a misclassification measure is defined by

\[ d(X; k, \Lambda) = -g(X; \hat{\lambda}_k) + \ln \left( \frac{1}{r_m} \sum_{i=1}^{r_m} \exp[g(X; \hat{\lambda}_{n(i)}) \cdot \eta] \right)^{1/\eta} \]

where \( \Lambda = (\hat{\lambda}_i; i = 1, \cdots, S) \), \( g(X; \hat{\lambda}_k) = \log p(X | \hat{\lambda}_k) \) is the log-likelihood score of the input \( X \) evaluated using the GMM model of the correct speaker \( k \), \( g(X; \hat{\lambda}_{n(i)}) \) is the log-likelihood scores of the top-\( r \)-th competing speaker \( n(r) \), and \( r_m \) is the total number of competing speakers used in the discriminative training. In this study, \( r_m \) is empirically set to be 5.

Then, a loss function \( l(d) \) is used to transfer the misclassification measure into error counts. The loss function is sign-function like, but differentiable, and is defined by

\[ l(d) = \frac{1}{1 + \exp(-\gamma (d + \theta))} \]

where the constants \( \gamma \) and \( \theta \) can be properly chosen to make the differential of the loss function \( l'(d) = l(d) \cdot (1 - l(d)) \) be large enough in the range that we want to adjust the misclassification measure of the training data.

By using the steepest descent method to minimize the loss function, the following parameter re-estimation formulas were used to find the new model \( \lambda' = (c', \mu', \Sigma') \) iteratively.

\[ c_m^{(i)} = c_m^{(i)} - \varepsilon \frac{\partial l(d(X;k,\Lambda))}{\partial c_m} \]

\[ \mu_m^{(i)} = \mu_m^{(i)} - \varepsilon \frac{\partial l(d(X;k,\Lambda))}{\partial \mu_m^{(i)}} \]

\[ \Sigma_m^{(i)} = \Sigma_m^{(i)} - \varepsilon \frac{\partial l(d(X;k,\Lambda))}{\partial \Sigma_m^{(i)}} \]

where \( i = k, n(1), \cdots, n(r_m) \), then

\[ c_m^{(i)} = c_m^{(i)} - \sum_{n=1}^{r_m} \frac{c_m^{(i)} p(x_n | m, \hat{\lambda}_i)}{c_m^{(i)}} \]

\[ \mu_m^{(i)} = \mu_m^{(i)} - \sum_{n=1}^{r_m} \frac{c_m^{(i)} (x_n - \mu_m^{(i)})}{c_m^{(i)}} \]

\[ \Sigma_m^{(i)} = \Sigma_m^{(i)} - \exp(-\Delta) \sum_{i=1}^{r_m} \left( \frac{x_n - \mu_m^{(i)}}{\Sigma_m^{(i)} \cdot \Delta \cdot \eta} \right)^{1/2} \]

and

\[ \Delta = \varepsilon \frac{\sum_{n=1}^{r_m} \frac{c_m^{(i)} p(x_n | m, \hat{\lambda}_i)}{c_m^{(i)}} \left( \frac{x_n - \mu_m^{(i)}}{\Sigma_m^{(i)}} \right)^{1/2}}{\sum_{i=1}^{r_m} \exp[\eta \cdot g(X; \hat{\lambda}_i)]} \rightarrow \text{iteration} \]

2.3 The Two-state HMM System

In Mandarin TI speaker identification, the testing speech signal should comprise both the initial and final parts of some syllables even its length is very short. So we can model the initial and final parts of each speaker’s speech separately by a GMM model. A 2-state HMM system, as shown in Fig. 1, is therefore constructed for Mandarin TI speaker identification. Its performance should be better because it models the characteristics of Mandarin speech more precisely.

The training of the two-state HMM now becomes to find the models with maximum likelihood score, i.e.,
\[ \hat{s} = \arg \max_{j} \sum_{1 \leq j \leq S} \log \left( p(x, q | \lambda_{ini}^{j}, \lambda_{fin}^{j}) \right) \]

where \( \lambda_{ini} \) and \( \lambda_{fin} \) are the initial and final GMM models, and, \( q_n \in \{\text{initial, final}\} \). The optimization can be found by using the Viterbi algorithm.

![Block diagram of the two-state HMM system](image)

3. EXPERIMENTS AND RESULTS

3.1 Speech Database

The database used in the following experiments is a large 16-kHz sampled, Mandarin microphone-speech database, called TCC-300 (recorded by Taiwan, Cheng Kung, and Chiao Tung Universities, Taiwan). The database contains paragraphic and sentential utterances of 300 speakers including 150 females and 150 males. Texts of the database were collected from newspaper and textbooks of elementary school. The length of the whole database is 26.55 hours, and there are in total 232,708 syllables.

All speech signal were segmented into frames for feature extraction. A 25-dimensional vector containing 12 MFCC, 12 delta-MFCC and one delta-energy was extracted for each frame. The frame length is 30 ms and the frame-shift is 10 ms.

3.2 Experiments

First, the GMM system was tested. According to [1], a speech segment of 24 seconds was used to train a GMM of 32 mixtures for each speaker. The testing data contained 1500 utterances, five for each speaker. Notice that silences in all training and testing data were pre-removed. It is also noted that the testing utterances may contain incomplete syllables because they are of fixed length.

The recognition results of the GMM system for different numbers of mixtures and different length of testing data are shown in Table 1. It can be seen from Table 1 that the speaker recognition rate is above 90% when the length of the testing data was equal to or larger than 2s and the number of mixtures was greater than 16.

Table 1. Speaker identification results of the GMM system. (unit of recognition rate : %)

<table>
<thead>
<tr>
<th>Number of mixtures</th>
<th>Length of testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1sec</td>
</tr>
<tr>
<td>8</td>
<td>67.4</td>
</tr>
<tr>
<td>16</td>
<td>74.8</td>
</tr>
<tr>
<td>32</td>
<td>79.0</td>
</tr>
</tbody>
</table>

Then, the proposed two-state HMM system was tested. First, 22 initial and 40 final HMM models were constructed from all the speech data of 300 speakers in TCC-300. Then, a GMM was trained for each speaker using a 24-sec training data pre-segmented into initial and final parts by the HMM syllable recognizer. In order to be fair for performance comparison, the total number of mixtures used in the two-state HMM system was set to the same value as the GMM system. The numbers of mixtures used for initial and final GMMs were empirically determined with ratio of 1 to 3. In testing the two-state HMM system, minimum duration constraints of 3 and 5 frames were respectively used for initial and final states in the Viterbi search. The recognition results of the proposed system are shown in Table 2. Here 8(I:2,F:6) means 2 and 6 mixtures were respectively used for initial and final GMMs, and so on. Figure 2 shows the error rate reductions of the two-state HMM system over the GMM

Table 2. Speaker identification results of the two-state HMM system. (unit of recognition rate : %)

<table>
<thead>
<tr>
<th>Number of mixtures</th>
<th>Length of testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1sec</td>
</tr>
<tr>
<td>8(I:2,F:6)</td>
<td>67.4</td>
</tr>
<tr>
<td>16(I:4,F:12)</td>
<td>74.8</td>
</tr>
<tr>
<td>32(I:12,F:24)</td>
<td>79.0</td>
</tr>
</tbody>
</table>
system. Can be seen from Fig. 2 that 10-20% error rate reductions were achieved for the proposed two-state HMM system.

Table 2. Speaker identification results of the two-state HMM system. (unit of recognition rate : %)

<table>
<thead>
<tr>
<th>Number of Mixtures</th>
<th>Length of testing data</th>
<th>1sec</th>
<th>2sec</th>
<th>4sec</th>
<th>6sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>8(I:2,F:6)</td>
<td></td>
<td>70.4</td>
<td>87.5</td>
<td>95.4</td>
<td>96.5</td>
</tr>
<tr>
<td>16(I:4,F:12)</td>
<td></td>
<td>76.5</td>
<td>91.3</td>
<td>95.6</td>
<td>96.8</td>
</tr>
<tr>
<td>32(I:10,F:22)</td>
<td></td>
<td>81.8</td>
<td>92.5</td>
<td>96.7</td>
<td>97.2</td>
</tr>
</tbody>
</table>

Lastly, the GPD/MCE discriminative training algorithm was applied to both the GMM system and the two-state HMM system. The recognition results are shown in Table 3 and the corresponding error rate reductions are displayed in Figure 3. It can be found from Fig. 3 that 25-50% error rate reductions were achieved for the proposed two-state HMM system using the MCE/GPD training algorithm.

4. CONCLUSIONS

In this paper, a new GMM-based Mandarin TI speaker identification system with an upper-level two-state Markov chain was proposed. The proposed system models the phonetic structure of Mandarin speech more accurately without losing the text independent property. Besides, the GPD/MCE algorithm was used to improve the training of the system. About 25-50% error rate reductions were achieved for the proposed two-state HMM system when different numbers of mixtures in GMMs and different lengths of testing data were used.

REFERENCES