Abstract - In this paper, we focus on the research of robust speaker verification system in adverse noise conditions. To improve the performance of conventional single PLDA model in the practical environment, two SNR-multicondition strategies are adopted to solve the SNR mismatch problem. The first approach uses different PLDA models to match the relative SNR conditions of training and testing speech utterances. The other uses one PLDA model trained by speech utterances with different SNR and a multicondition score normalization method to adjust the score distribution. The performance of the two systems are demonstrated on a 120-speakers test set for each gender. According to the result of the experiment, the proposed approaches are proof to reduce the EER to 7.39%, 7.17% for male and 10.22%, 10.06% for female respectively in the random SNR noise condition, which means the SNR-multicondition systems are more robust in the real noisy environment.

Keywords: speaker verification, PLDA, SNR-multicondition, score normalization

1 Introduction

The state-of-art speaker verification system is able to obtain high recognition accuracy in the lab environment after many years of research on both speech features and speaker models. However, in real environments the variant noise will dramatically deteriorate the system performance. Many research focused on the front-end process, such as enhancing noisy speech signals [1], extracting more robust features [2], or warping the feature vectors [3]. These methods are proof to be valid but the improvement is very limited, far from the demands of real applications.

Since NIST Speaker Recognition Evaluations [4] put more emphasis on channel mismatch problem, many back-end compensation methods like JFA [5] and i-vector/PLDA [6, 7] has been proposed. These methods use subspaces to describe speaker and channel models separately, suppress channel-variability in the speaker factor, making channel-independent recognition possible. The same idea can be implemented in noise mismatch compensation, so in this paper we mainly use i-vector/PLDA model to build baseline system.

The i-vector is a low-dimension vector representing speaker characteristics. It is extracted from mean supervector of a classical GMM-UBM [8] speaker model. PLDA conducts factor analysis on the i-vector space by grouping i-vectors derived from the same speaker in different channels or noise conditions. On the foundation of i-vector/PLDA framework, advanced research has been done in recent years. For example, the mixture of PLDA models was firstly proposed in [9] to be a solution to make systems independent of the speaker gender, and the mixture of models was later used in [10] to describe speaker utterances of different SNR. However, a simpler method is employing multicondition training PLDA as is described in [11, 12], using the classification system to adapt to different noise conditions. Based on the idea of SNR-multicondition, in this paper, two classification systems are designed and their performance is analyzed as well.

This paper is organized as follows. Section 2 introduces the baseline i-vector/PLDA architecture. Section 3 describes the two systems and their respective design principles. Section 4 provides the experiment procedures and results. The last section 5 concludes the findings.

2 i-vector/PLDA Baseline System

2.1 i-vector extraction

I-vector is a new approach of front-end factor analysis proposed by Dehak in [7]. Instead of separate speakers and channel subspaces in JFA, the i-vector extractor employs a total-variability subspace to describe both speaker and channel characteristics simultaneously. Therefore a target speaker’s GMM supervector can be represented by

\[ M = m + T \omega \]  

where \( m \) is the UBM supervector and \( T \) is the low rank total-variability matrix, \( \omega \) is the speaker factor, as called i-vector. The total-variability matrix is trained by EM algorithm with a large amount of speech utterances.

2.2 Gaussian PLDA modeling

Probabilistic Linear Discriminant Analysis is an approach firstly proposed in [13] to compensate the within-individual variation of face images. This approach was later used in speaker verification to directly model session and speaker variability within the i-vector space by Kenny. In modified Gaussian PLDA model, the i-vector can be further considered as a variable following the generative model
\[ \omega = m + Vz + \epsilon \quad (2) \]

where \( m \) is a global offset, usually using the mean of all the i-vectors. \( V \) is the eigenvoice matrix and \( z \) is the speaker factor. \( \epsilon \) is the noise residuals. In Gaussian PLDA, \( z \) follows the standard normal distribution and \( \epsilon \) is assumed to be Gaussian with full covariance \( \Sigma \). The model parameters \( \{m, V, \Sigma\} \) are obtained from the collection of speaker i-vectors in a specific background condition by EM algorithm.

The scoring method of PLDA is based on likelihood ratio. Given the target speaker and test i-vectors, the decision score can be calculated as follows

\[ \text{score} = \ln \frac{P(\omega_{\text{target}}, \omega_{\text{test}} \mid H_1)}{P(\omega_{\text{target}} \mid H_0)P(\omega_{\text{test}} \mid H_0)} \quad (3) \]

where \( H_1 \) is the hypothesis that the speakers are the same, \( H_0 \) is the hypothesis that the speakers are different.

### 3 SNR-multicondition Architecture

In [10], a mixture of PLDA speaker verification system was proposed to solve the noise mismatch problem. However, the single mixture PLDA model is hard to train sufficiently and isn’t able to match the wide range of SNR in the real environment. So in this paper, we shared some key ideas with [10] in the baseline system and designed two SNR-multicondition architectures with multi PLDA models to improve the noise robustness.

#### 3.1 PLDA model classification

Single Gaussian PLDA model assumes that all the speaker factor follows a Gaussian distribution. But in a wide range of signal-to-noise ratio (SNR) noisy conditions, the single distribution is rather limited. Usually a PLDA model trained by i-vectors of a specific SNR condition is more valid to compensate the SNR mismatch in the same condition.

Table 1. Speaker verification performance across different SNR conditions test set for different PLDA models

<table>
<thead>
<tr>
<th>PLDA model</th>
<th>Test Set</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean-15dB</td>
<td>5.39</td>
<td>8.31</td>
<td>19.50</td>
<td></td>
</tr>
<tr>
<td>clean-10dB</td>
<td>5.50</td>
<td>7.33</td>
<td>13.00</td>
<td></td>
</tr>
<tr>
<td>clean-5dB</td>
<td>6.22</td>
<td>7.44</td>
<td>9.44</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 shows the performance of different single PLDA models in different SNR test set. The PLDA model is trained by a group of i-vectors containing both clean and specific SNR conditions. In the test set, the training data is clean and the testing data is noisy. As we can see in the table, in every test set, the PLDA model which matches the SNR of the test set obtained the best performance. Based on this result, we designed a PLDA model classification structure to match different SNR conditions dynamically.

The classification system employs a group of SNR-dependent PLDA models and each model is responsible for the test data in a specific range of SNR. To cover the wide range of SNR in practical environment, we choose four conditions: clean, 15dB, 10dB and 5dB. And the SNR estimator module decides which model to use. After each PLDA scoring module there is also an individual Z-norm process, because the scores of different models are in different ranges. The score normalization can be obtained as follows

\[ z = \frac{s_i - \mu_i}{\sigma_i} \quad (4) \]

where \( \mu_i \) and \( \sigma_i \) is the mean and standard of the impostor scores in relative SNR condition. The final scores are normed to follow the same distribution and can use one threshold to decide whether to reject or accept.

#### 3.2 Score normalization classification

The other strategy is using a pooling PLDA model trained by i-vectors in different noise levels. In this paper, we pool the training data of clean, 15dB, 10dB and 5dB conditions together to obtain the model parameters. Although we aim to employ a single pooling PLDA to match all the conditions, we still need to classify the system in the score field.
3.3 SNR estimator

In both of the systems above, the classification is based on a SNR estimator module. SNR is usually calculated from both the clean and noisy speech data. But in real environment, the clean speech isn’t available to get, so we should use another method to estimate the SNR. In this paper, we proposed an SNR estimator based on noise estimation algorithm to solve this problem.

\[
SNR_{est} = 10 \log_{10} \frac{\sum |y(t) - \bar{n}(t)|^2}{\sum \bar{n}^2(t)} \tag{5}
\]

In equation 5, \( y(t) \) is noisy speech signal while \( \bar{n}(t) \) is the estimated noise signal. In this paper, we use the MMSE-SPP algorithm proposed in [13] to estimate the noise power spectral density. Because the estimated noise level is often lower than real, which means the result of SNR estimator is greater, the decision of which SNR condition to use is based on the table as follows.

<table>
<thead>
<tr>
<th>Estimated SNR</th>
<th>PLDA model</th>
<th>Z-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 20dB</td>
<td>Clean</td>
<td>clean</td>
</tr>
<tr>
<td>&gt; 15dB and &lt; 20dB</td>
<td>clean-15dB</td>
<td>15dB</td>
</tr>
<tr>
<td>&gt; 10dB and &lt; 15dB</td>
<td>clean-10dB</td>
<td>10dB</td>
</tr>
<tr>
<td>&lt; 10 dB</td>
<td>clean-5dB</td>
<td>5dB</td>
</tr>
</tbody>
</table>

4 Experiments and Results

4.1 Experimental setup

All the experiments were conducted on a 276 male and 276 female speaker data set. The speech is in Chinese language and each speaker has 16 segments of 30s length. We chose 120 speakers data for each gender as test set and the remaining as development set to train PLDA model parameters. Among the 16 segments, one is used to train speaker model while the others are used to test, which obtains a total of 1800 target trials and 1800 non-target trials for each gender.

The speaker verification baseline system is based on UBM with 512 mixtures and i-vector with 400 dimensions. The UBM and total variability matrix is trained from a large scale of gender-dependent development data. The speaker factors of PLDA models are 150 dimensions, and the mode parameters \( \{m,V,\Sigma\} \) are trained from 156 speakers data distinguished from the data in test set.

For noisy data, we obtained speech data in 3 SNR conditions by adding noise in NOIZEUS corpus to clean speech. We randomly picked 1 from 8 different kinds of noise: car, train, street, station, airport, exhibition, restaurant and babble for each speech utterance, in order to simulate the real environment. These noisy speech were used to train 3 SNR-dependent PLDA models, make up test sets and calculate the statistics.

In the front-end, all speech were used to extract 60 MFCCs as acoustics features, which were obtained every 10ms from a 25ms hamming window. CMS and feature warping were then utilized to the feature vectors.

4.2 Results and analysis

Table 3 shows the performance of two SNR-multicondition systems and 5 single PLDA systems as comparison in EER evaluation. The speech data in the test set contained randomly picked noise and the SNR range from 3 to 20dB. The test set is aimed to simulate the real environment because the practical noise varies a lot both in kind and SNR.

From table 3, we can see that the performance of single PLDA systems were dramatically deteriorated by additive noise, especially when only clean speech were used in training PLDA model. The best performance of single PLDA model is the pooling PLDA, because the model parameters describe the characteristic of i-vectors in all SNR conditions to some extent. However, the SNR-multicondition proposed in this paper were proof to perform better than all the single PLDA systems in this test situation, the EER were reduced to 7.39% and 7.17% for male and 10.22% and 10.06% for female respectively. The advantage of the SNR-multicondition systems is the dynamic selection of the models or algorithm, making the system always in the SNR matching state.

Figure 4 shows the DET curves for two SNR-multicondition systems for each gender, in order to evaluate the detailed performance of two proposed approach. The result show that the PLDA model classification system is slightly inferior to the score normalization classification system. The reason is that the PLDA models in the first proposed system don’t strictly match in SNR because we only trained four SNR conditions to cover a wide range, while the
pooling model in the second approach has a global characteristic to avoid this problem.

Table 3. Performance of single PLDA and SNR-multicondition systems in random SNR test set

<table>
<thead>
<tr>
<th>System</th>
<th>EER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>male</td>
</tr>
<tr>
<td>single PLDA : clean</td>
<td>23.67</td>
</tr>
<tr>
<td>single PLDA : clean-15dB</td>
<td>12.06</td>
</tr>
<tr>
<td>single PLDA : clean-10dB</td>
<td>12.94</td>
</tr>
<tr>
<td>single PLDA : clean-5dB</td>
<td>13.22</td>
</tr>
<tr>
<td>single PLDA : pooling</td>
<td>8.33</td>
</tr>
<tr>
<td>multicondition : PLDA model classification</td>
<td>7.39</td>
</tr>
<tr>
<td>multicondition : Score normalization classification</td>
<td>7.17</td>
</tr>
</tbody>
</table>

5 Conclusions

In this paper, two SNR-multicondition strategies are used on the base of i-vector/PLDA speaker verification system. The performance of these systems were evaluated on the randomly created test set of 120 speakers for each gender, and both of them show improvement to the baseline system, successfully reduce the SNR mismatch problem which may happen in the real environment. The result of the experiment also shows that in the test set of this paper, the score normalization classification approach is more valid than the PLDA model classification approach for the global characteristic of the pooling model.

6 Acknowledge

This work was supported by National Natural Science Foundation of China (grant Nos. 61371147, and 11433002), and Shanghai Academy of Spaceflight Technology (grant No. SAST2015039).

7 References


