A NOVEL TEXT-INDEPENDENT SPEAKER VERIFICATION METHOD USING THE GLOBAL SPEAKER MODEL

Yiying Zhang, Xiaoyan Zhu, IEEE Member
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University, Beijing 100084
Email: zxy-dcs@mail.tsinghua.edu.cn

ABSTRACT

In this paper a new text-independent speaker verification method is proposed based on likelihood score normalization and the global speaker model, which is established to represent the universal features of speech and environment, and to normalize the likelihood score. As a result the equal error rates are decreased significantly, verification procedure is accelerated and system adaptability is improved. Two possible ways of establishing the global speaker model, one of which can meet the real-time requirement, are also suggested and discussed. Experiments demonstrate the effectiveness of this novel verification method and its improvement over the conventional method and other normalization methods.

1. INTRODUCTION

In the conventional speaker verification method (noted as CSV in the following), the decision rule of acceptance or rejection is based on the score of a test utterance for the claimed speaker and a predefined threshold [1]. The CSV method has three limitations. The first is the loose distribution of likelihood score, which leads to vague boundaries between speakers. The second is the burden to set a proper threshold, which is a direct consequence of scattered likelihood score. The third is low system adaptability to protean input utterances with different duration and different content.

The proposed text-independent speaker verification method using the global speaker model (called as GSMSV method in this paper) is based on our previous work [2] and aims at the limitations of the conventional method. The global speaker model is established to represent all of the common information between speakers, such as the pronunciation characteristics, background noises, common features of texts with different speakers, and the pronunciation, speaking environment and text contents. If these information is removed out from speech, the differences between speakers will be emphasized. Thus GSMSV utilizes the common information to normalize likelihood score and to exclude all common information contained in speech.

Let $S_{GSM}$ be the normalized likelihood score of GSMSV method. To an input utterance $X$, it is computed as the following.

$$ S_{GSM} = P(X | \lambda) - P(X | \lambda_{GSM}) $$

By subtracting the score produced by $\lambda_{GSM}$, the common information is obliterated from speech. As a result, the interference of unimportant factors is avoided and the distinction between different speakers is clearer. Therefore the decision rule for GSMSV is:

$$ P(X | \lambda) - P(X | \lambda_{GSM}) \geq \eta, \text{ Accept the claim} $$

$$ P(X | \lambda) - P(X | \lambda_{GSM}) \leq \eta, \text{ Reject the claim} $$

in which $\eta$ is a threshold.

One point should be noted. Although $\lambda_{GSM}$ is obtained from the training data of reference speakers, it is a universal speaker model.
model, not only representing common characteristics of reference speakers, but also embodying the common information of outside impostors. Therefore theoretically GSMSV has a powerful ability to distinguish both reference speakers and outside impostors.

In order to avoid overflow in computation, log likelihood score is utilized and thus the decision rule becomes

$$\log P(X | \lambda_1) \cdot \log P(X | \lambda_{GSM}) \begin{cases} > \eta', & \text{Accept the claim} \\ \leq \eta', & \text{Reject the claim} \end{cases}$$

in which $\eta'$ is a threshold.

To further improve the adaptability of text-independent verification system and alleviate the influence of utterance duration, likelihood score is further normalized by the duration of test utterance. Thus decision rule (2) is improved by formula (3), in which $T_X$ is the number of frames and $\eta''$ is a threshold.

$$\frac{\log P(X | \lambda_1) \cdot \log P(X | \lambda_{GSM})}{T_X} \begin{cases} > \eta'', & \text{Accept the claim} \\ \leq \eta'', & \text{Reject the claim} \end{cases}$$

Compared with CSV, GSMSV improves the distribution of likelihood score, making it more centered and is more adaptive since it removes the influence of speaking speed and thus further improve system adaptability. In addition, setting threshold is much easier owing to more compact score distribution.

Compared with ASMSV method, establishing $\lambda_{GSM}$ does not need the procedure to find the typical impostors for a reference speaker. It is obtained as to create the usual speaker model, and the difference only lies in the training data used. Furthermore two calculations of likelihood scores during verification phase are also very fast and can satisfy real-time verification requirements.

2.2. Comparison with Other Methods

One of the popular methods for normalizing likelihood score is to establish Anti-Speaker Model (noted as ASM) [3]. The ASM of a reference speaker is the subset of reference speaker models. ASM represents the speech characteristics of impostors. Thus speaker space is divided into two sub-spaces, claimed reference speaker and his impostors.

Let $\lambda_i$ and $\hat{\lambda}_i$ represent the speaker model of certain reference speaker $i$ and his impostors respectively. Thus

$$\hat{\lambda}_i = \{\lambda_{u(1)}, \lambda_{u(2)}, \ldots, \lambda_{u(L)}\}$$

in which $s(k) \in \{1..N\}, s(k) \neq i, k = 1, 2, \ldots, L$

Maximum Likelihood criterion, in which $Y$ and $Y'$ are training data of reference speaker and impostors. Given test data $X_i$ the decision rule for deciding whether $X$ is uttered by the $i$-th reference speaker is:

$$\frac{P(X | \lambda_i)}{P(X | \lambda_{GSM})} > \eta, \text{ Accept the claim}$$

$$\frac{P(X | \lambda_i)}{P(X | \lambda_{GSM})} \leq \eta, \text{ Reject the claim}$$

From the above description, we can see that there exists great differences between CSV, ASMSV and GSMSV methods. As a summary, their differences are listed in Table 1.

### Table 1: Differences between three methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>CSV</th>
<th>ASMSV</th>
<th>GSMSV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Procedure</td>
<td>Train N models</td>
<td>Train N models and establish L ASM models</td>
<td>Train N models and 1 global speaker model</td>
</tr>
<tr>
<td>Whether to Normalize Score</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Score Calculations for Verification</td>
<td>1 time</td>
<td>L+1 times</td>
<td>2 times</td>
</tr>
</tbody>
</table>

3. ESTABLISHMENT OF THE GLOBAL SPEAKER MODEL

Since the global speaker model is critical to the performance of verification system, the parameter estimation of $\lambda_{GSM}$ is provided in this section.

3.1. General Estimation

The speaker model employed is Gaussian mixture model [4]. Let the parameters for $\lambda_{GSM}$ be:

$$\lambda_{GSM} = (\mu_{GSM}^{c(1)} \Sigma_{GSM}^{c(1)}), \ldots, (\mu_{GSM}^{c(M)} \Sigma_{GSM}^{c(M)})$$

in which $\mu_{GSM}^{c(k)}$ and $\Sigma_{GSM}^{c(k)}$ are mean vector and covariance matrix of the $k$-th Gaussian density function respectively, $c_k$ is the corresponding weight, $M$ is the number of mixture components.

Assume there are $N$ users currently, whose training data is represented as

$$Y_i = \{y_{i(1)}, y_{i(2)}, \ldots, y_{i(T(i))}\} \quad i = 1, 2, \ldots, N$$

after being transformed to feature vectors, in which $i$ denotes the $i$-th speaker and $T(i)$ denotes the total number of feature vectors. The training data for the $(N+1)$-th speaker is

$$Y_{N+1} = \{y_{(N+1)}, y_{(N+1)}, \ldots, y_{(N+1)}\}$$

where

$$y_{(i)} = \{y_{i(1)}, y_{i(2)}, \ldots, y_{i(T(i))}\} \quad i = 1, 2, \ldots, N$$

$$y_{(N+1)} = \{y_{(N+1)}, y_{(N+1)}, \ldots, y_{(N+1)}\}$$

Let $\lambda_{i}$ and $\hat{\lambda}_{i}$ represent the speaker model of certain reference speaker $i$ and his impostors respectively. Thus

$$\hat{\lambda}_{i} = \{\lambda_{u(1)}, \lambda_{u(2)}, \ldots, \lambda_{u(L)}\}$$

in which $s(k) \in \{1..N\}, s(k) \neq i, k = 1, 2, \ldots, L$
Estimation of $\lambda_{GSM}$ is an iterative procedure starting from the initial values obtained by Segmental K-Means Procedure [5][6]. The re-estimation formulas for $\lambda_{GSM}$ are as follows.

$$\hat{c}^GSM_j = \frac{\sum_{n=1}^{N(t)} \alpha_j^n(t)}{\sum_{n=1}^{N(t)} \alpha_j^n(t) + \beta_j^n(t)} \quad j = 1, 2, \ldots, M$$

(4)

$$\theta_j^{n}(t) = \begin{cases} \sum_{jGSM} p_j jGSM(n) \beta_j^{n} & t = 1 \\ \sum_{jGSM} p_j jGSM(n) \hat{p}_j^{n} \beta_j^{n} & t = 2 \sim T(n) \end{cases}$$

(5)

$$\alpha_j^{n}(t) = \begin{cases} p(y_t^{n}, j \beta_j^{n}) & t = 2 \sim T(n) \\ p(y_t^{n}, j \beta_j^{n}) & t = 1 \end{cases}$$

(6)

$$\beta_j^{n}(t) = \begin{cases} p(y_t^{n}, j \beta_j^{n+1}) & t = 1 \sim (T(n) - 1) \\ 1 & t = T(n) \end{cases}$$

(7)

$$\hat{\mu}^{GSM}_j = \frac{\sum_{n=1}^{N(t)} \theta_j^{n}(t) y_t^{n}}{\sum_{n=1}^{N(t)} \theta_j^{n}(t)} \quad j = 1, 2, \ldots, M$$

(8)

$$\hat{\Sigma}^{GSM}_j = \frac{\sum_{n=1}^{N(t)} \theta_j^{n}(t) (y_t^{n} - \hat{\mu}^{GSM}_j)(y_t^{n} - \hat{\mu}^{GSM}_j)^T}{\sum_{n=1}^{N(t)} \theta_j^{n}(t)} \quad j = 1, 2, \ldots, M$$

(9)

Since $\lambda_{GSM}$ is obtained by using all of the training data of current users, the training time is long and sometimes cannot meet real-time needs, especially when system has a large number of users. Thus the real-time estimation is provided in the following sub-section.

### 3.2. Real-time Estimation

When a new user comes, the training time is consumed mainly on re-training $\lambda_{GSM}$. Therefore accelerating the training procedure for $\lambda_{GSM}$ is important. The real-time estimation updates $\lambda_{GSM}$ parameters in one step. The initial values are same as those corresponding values of the last new user. The new re-estimation formulas for $\lambda_{GSM}$ are as follows.

$$\hat{c}^GSM_j = \frac{(1 - \rho) \sum_{n=1}^{N(t)} \theta_j^{n}(t) + \rho \sum_{n=1}^{N(t)+1} \theta_j^{n+1}(t)}{(1 - \rho) \sum_{n=1}^{N(t)} \alpha_j^{n}(t) + \rho \sum_{n=1}^{N(t)+1} \alpha_j^{n+1}(t)} \quad j = 1, 2, \ldots, M$$

(10)

Two main points lead to the decrease on computation overhead. In this method, updating is not an iterative procedure. The initialization is completed in one phase using formulas (10)-(12). On the other hand updating starts from rather better initial values, avoiding the burden of setting proper initial values, a time-consuming procedure.

### 4. EXPERIMENTS

#### 4.1. Database and Settings

Data used comes from a Mandarin speech database 863Bag provided by the State Education Commission. Speech data of 50 persons (25 females and 25 males) is used. Each person uttered 50 sentences, ranging from 5.6 seconds to 1.2 seconds.

There are 30 reference speakers (15 females and 15 males). Training data includes 15 sentences. Test data is one sentence. Tests on data of reference speakers constitute closed set test. Tests on data of 20 other speakers who are regarded as outside impostors constitute open set test. The average duration of training data is 60 seconds, and the average duration of every test utterance is 3.5 seconds.

The feature used is 16 cepstrum, 16 delta cepstrum, and delta energy. The general estimation (formulas (4)-(9)) of the global speaker model is adopted in the following experiments.

Equal error rate is used to measure the performance of speaker verification for different scoring methods. The equal error rate is
a posterior error rate, and at this equal error rate, the decision boundary is set to make the error rate of false rejection be equal to that of false acceptance. The posterior equal error rate is a convenient measure of the degree of separation between true and false speaker scores and, therefore, a useful predictor of speaker verification performance.

4.2. Statistical Analysis

In this experiment, the likelihood scores of closed set test for GSMSV and CSV are recorded and analyzed. The corresponding statistical results are listed in Table 2.

From Table 2, the following conclusions can be obtained:

1) The variance of GSMSV likelihood scores is much smaller than that of CSV method, making the distribution of likelihood scores more compact.

2) The difference between the likelihood scores of GSMSV valid users and impostors is greater than that of CSV method, enlarging the distance between valid users and impostors. Therefore it is more convenient for GSMSV to set a proper threshold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Valid speakers</th>
<th>Impostors</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean $(\hat{M}_v)$</td>
<td>Variance $(\hat{V}_v)$</td>
<td>Mean $(\hat{M}_i)$</td>
</tr>
<tr>
<td>CSV</td>
<td>3354.79</td>
<td>584.36</td>
<td>1762.85</td>
</tr>
<tr>
<td>GSMSV</td>
<td>285.32</td>
<td>188.13</td>
<td>-1312.71</td>
</tr>
</tbody>
</table>

Table 2: Statistical analysis of likelihood scores.

4.3. Performance Comparison of Different Methods

In this experiment, the performance of CSV, ASMSV and GSMSV are compared. ASMSV method has the lowest equal error rates when the anti-speaker model consists of all of other reference speakers[3], thus in this experiment set $L = 29$. Table 3 lists the equal error rates.

<table>
<thead>
<tr>
<th>Method</th>
<th>Closed set (%)</th>
<th>Open set (%)</th>
<th>Speed (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSV</td>
<td>6.19</td>
<td>1.69</td>
<td>0.57</td>
</tr>
<tr>
<td>ASMSV ($L=29$)</td>
<td>0.19</td>
<td>1.06</td>
<td>17.26</td>
</tr>
<tr>
<td>GSMSV</td>
<td>0.59</td>
<td>0.51</td>
<td>1.15</td>
</tr>
<tr>
<td>ASMSV ($L=1$)</td>
<td>7.80</td>
<td>2.16</td>
<td>1.15</td>
</tr>
</tbody>
</table>

Table 3: Performance comparison.

5. CONCLUSION

A novel speaker verification method GSMSV is proposed in this paper. Compared to the CSV and ASMSV methods, GSMSV has the following characteristics. 1) The differences between speakers are enlarged. 2) The system ability to distinguish reference speakers is improved. 3) Impostors can be more easily detected. 4) Verification speed is fast. 5) System is adaptable to speaking speed.

As a summary, the equal error rates of GSMSV method are significantly low, especially for open set test, which is vital for practical applications. GSMSV provides a promising way for realizing verification systems.

6. REFERENCES