DETERMINATION OF THRESHOLD FOR SPEAKER VERIFICATION USING SPEAKER ADAPTATION GAIN IN LIKELIHOOD DURING TRAINING

Toshiaki Uchibe, Shingo Kuroiwa, Norio Higuchi
KDD R&D Laboratories Inc.
2-1-15 Ohara Kamifukuoka Saitama 356-8502, Japan
to-uchibe@kdd.co.jp

ABSTRACT
This paper describes methods to determine thresholds for speaker verification. Setting an appropriate threshold a priori is difficult because likelihood verification covers a wide range and the appropriate threshold for each speaker is different. We propose new methods to determine the speaker verification threshold depending on the "adaptation degree" for each speaker. We use the gain in likelihood during the adaptive training process from speaker-independent models as the "adaptation degree" and determine the threshold by its linear function. We evaluate the proposed methods in text-prompted speaker verification experiments using connected digit speech to show that the estimated coefficients of the linear function are relatively constant regardless of the amount of training data and that thresholds set by our proposed methods are stable and reliable. Consequently, use of our new methods improves verification performance and reduces the error rate by 30 percent.

1. INTRODUCTION
In speaker verification systems, a threshold for distinguishing a true speaker from an impostor is required and this threshold must be set a priori. However, determining the appropriate threshold is difficult because the likelihood of verification is not stable enough to allow setting a common threshold for every speaker[1].

A method to set threshold for each speaker by DP matching or Vector Quantization (VQ) algorithm by calculating the mean and standard deviation of intra-speaker and inter-speaker distortions from training speech data has been proposed[2][3]. This method assumes that these two types of distortions can be approximated by Gaussian distributions, which means it requires a huge speech database of various speakers and consequently also requires a large amount of computation to find a more appropriate threshold. However, these distributions are not usually considered to demonstrate Gaussian distributions and setting an appropriate threshold is difficult unless there is a sufficient amount of training data.

To solve these problems, codebook-based threshold design algorithm without speech data was proposed[4]. This method is based on use of VQ codebooks and utilizes the correlation between inter-codebook distortion and appropriate threshold instead of calculation of intra-speaker or inter-speaker distortion using speech data. This method saves memory space and computation work, however still has some problems. Inter-speaker distortion for instance, is calculated by using registered speaker codebooks. An actual impostor might not be registered and additional speaker codebooks are required in advance to calculate threshold. Moreover, the amount of computation that is required, grows larger in proportion to the number of registered speakers.

In Hidden Markov Model (HMM)-based speaker verification, there is a proposal for updating thresholds by taking into account the speech intervals between registration and verification[5] and yet another proposal involves setting the threshold by calculating likelihood using impostor data[6]. However, effective methods have not yet been found for setting an initial threshold a priori that take into account the memory size and computation.

Likelihood normalization techniques have already been proposed as a method for setting stable thresholds regardless of variations in likelihood[7][8][9]. These techniques attempt to normalize the verification likelihood for claimed speaker models using antispeakers cohort models and consequently are effective versus variations in speech environments or styles for each utterance even by the same speaker. However, these techniques are inadequate for establishing stable thresholds common to every speaker.

In speaker verification using HMM especially in text-prompted style verification, creating speaker models by adaptive training from speaker-independent (SI) models has proven an effective method[7]. In this training process, the degree of adaptation to a speaker or "adaptation degree" is subject to change according to the speaker or training data. Therefore, the verification score varies over a wide range among speakers, making it difficult to set a stable threshold.

We therefore propose methods to determine a stable threshold a priori using the degree of adaptation to the speaker or so-called speaker adaptation gain in likelihood during training models for each speaker that function regardless of the amount of training data. We further design these methods to achieve high-performance verification compared to conventional methods.
2. LIKELIHOOD VARIATION

The difference in the appropriate threshold for each speaker is caused by a variation in likelihood during verification. To improve this variation, likelihood normalization techniques have been proposed. This method is expressed by:

\[ \hat{P} = P - \overline{P} \]

where \( P \) and \( \hat{P} \) are logarithmic likelihoods versus claimed speaker models and cohort models, and \( \overline{P} \) is the normalized likelihood.

We now investigate likelihood variations. Likelihood normalization assumes that the likelihoods for both true speaker models and antispeaker models vary to a similar extent. Eq.(1) shows that normalized likelihood is effective because it allows eliminating the relative change in likelihood for true speaker and antispeaker models. This likelihood variation is caused by a large difference due to range of different speech environments or styles even for utterances by the same person. In other words, normalized likelihood is effective for intra-speaker variations in likelihood.

In view of the fact that likelihood normalization technique is inadequate for setting a stable threshold common to all speakers, we speculated that inter-speaker variations might also exist in verification likelihood. Inter-speaker variations are thought to depend on training speaker models. Training data is subject to various conditions of speech environments or styles and each speaker models have different "adaptation degree" during the process of adaptive training from SI models. Each speaker models therefore generate different likelihood in verification, which in turn causes inter-speaker variations. We then started treating this adaptation degree as "speaker adaptation gain" and propose methods here to determine the threshold based on speaker adaptation gain.

3. METHODS TO SET THRESHOLD BASED ON SPEAKER ADAPTATION GAIN

We now show how variations in likelihood are caused by the difference in speaker adaptation gain and show how to determine the threshold using speaker adaptation gain.

3.1 Definition of Speaker Adaptation Gain

Speaker adaptation gain \( \tilde{P} \) is defined as follows:

\[ \tilde{P} = \frac{1}{N} \sum_{n=1}^{N} (P_d(n) - P_a(n)) \]

where \( N \) is the number of training utterances, and \( P_d(n) \) and \( P_a(n) \) are the logarithmic likelihoods of \( n \)-th training utterance versus the speaker models after training and SI models respectively. A logarithmic likelihood is calculated by Viterbi algorithm and normalized by frame length.

3.2 Speaker Adaptation Gain and Verification Likelihood

If SI models characterize impostors, then speaker adaptation gain describes the likelihood distortion between a true speaker and impostors. From this point of view, speaker adaptation gain corresponds to inter-speaker distortion of VQ codebooks\[4\], which closely correlates with the appropriate threshold. Therefore, speaker adaptation gain will also correlates with the appropriate threshold. We now show the correlation between speaker adaptation gain and normalized likelihood in verification of true speakers in Fig.1 and impostors in Fig.2. These figures show that normalized likelihood still varies widely according to speaker adaptation gain.

In Fig.1 for true speakers, the range of likelihood distribution, which is small when speaker adaptation gain is small, becomes wider as speaker adaptation gain grows larger. In Fig.2, the impostor likelihood, which is rather high when speaker adaptation gain is small, becomes proportionately lower as speaker adaptation gain grows larger. From these two relations, we can see that if speaker adaptation gain is small, then the impostor likelihood becomes higher because the speaker models are still close to SI models. If speaker adaptation gain gets larger, the true speaker likelihood, which may become higher when the
speaker models are really adapted to the speaker, tends to become lower due to overtraining. On the other hand, the impostor likelihood becomes lower since speaker models get further from the SI models.

**Fig.1:** Correlation between speaker adaptation gain and normalized likelihood of true speakers.

**Fig.2:** Correlation between speaker adaptation gain and normalization likelihood of impostors.

We assume from these features that the threshold can be estimated by a linear function of speaker adaptation gain \( \tilde{P} \) as follows:

\[
\theta = \alpha \cdot \tilde{P} + \beta \tag{3}
\]

where \( \alpha \) and \( \beta \) are coefficients. If the threshold can be determined appropriately from Eq.(3), then we only need training data for the true speaker and no longer need the huge amount of training data and calculation required by conventional methods. We next estimate \( \alpha \) and \( \beta \) to set appropriate thresholds through the experiments in the following section.

4. EXPERIMENTS

4.1 Purpose

In speaker verification, we generally estimate the performance by means of two error rates, a false acceptance rate (FAR) and a false rejection rate (FRR) determined by distinguishing between true speakers and impostors using thresholds. Although the problem of which error rate is more important may differ according to the purpose of the verification, we aim to improve verification performance by setting thresholds that equalize the FAR and FRR whose values are equal error rate (EER).

Obtaining as much training data as possible is preferable in order to obtain high performance speaker models. However, it is initially difficult to obtain a huge amount of training data and in practice, improving verification performance by increasing the available data for training while also using the speaker verification system will prove useful.

Therefore we evaluated our new methods in terms of obtaining:

High performance verification compared with conventional methods using a common threshold for all speakers

An *a priori* threshold that is stable regardless of the amount of training data

4.2 Experimental Conditions

Our methods were evaluated in text-prompted speaker verification experiments using connected digit speech. The database comprised connected digit speech data uttered by 300 male speakers as true speakers and 1200 male speakers as impostors. The speech was recorded in three sessions at two month intervals over public telephone networks. The utterances were cut so as to include a one second silence before and after the speech. The sampling rate was 8kHz with 8bit \( \mu \)-law. The data parameters were calculated by LPC analysis with an order of 12, a frame period of 10ms, a frame length of 20ms and output 12-th order LPC mel cepstral coefficients and 12-th order delta LPC mel cepstral coefficients. Cepstrum Mean Normalization (CMN)[10] was also applied to eliminate effects from telephone networks or in other word, frequency characteristics of terminals and circuits.

Digit HMMs have 7 to 12 states, 15 Gaussian mixtures for each state, diagonal covariance and head-body-tail states, which means that the first and final states of each digit are composed of context-dependent utterances. SI models were learned with a Baum-Welch algorithm using 146750 utterances of connected 4-digit speech uttered by 1524 male speakers.

For training, speaker models were re-estimated except for covariance by adapting the SI models as initial models with the Baum-Welch algorithm using training data selected from connected 4-digit speech recorded at the first session.

As verification, we used about 30 utterances for true speaker evaluation and 1200 utterances for impostor evaluation (1 utterance each impostor) from the speech data of connected 7-digit recorded at all sessions over a four month period. We applied likelihood normalization and cohort models as the SI models which were the same as the initial models in training. The likelihood normalization applied is expressed by:

\[
\tilde{P} = P - \bar{P} = \sum_{n=1}^{N} (p_n - \bar{p}_n) / \sum_{n=1}^{N} T_n \tag{4}
\]

where \( n \) means the \( n \)-th digit of \( N \)-digit speech (\( N=7 \) in this paper), \( T_n \) is the frame length of the \( n \)-th digit, \( p_n, \bar{p}_n \) are logarithmic likelihoods of the \( n \)-th digit for claimed speaker models and SI models are calculated with the Viterbi algorithm.

4.3 A Posteriori Threshold Decision

We performed experiments to evaluate our new methods versus conventional methods using a common threshold \( \Theta \) for every speaker. We used 15 training utterances and calculated the parameters \( \alpha \) and \( \beta \) in Eq.(3) and a common threshold \( \Theta \) *a posteriori* to get EER. Consequently, we obtained \( \alpha = -0.54, \beta = 1.02 \) and \( \Theta = -2.87 \), namely Eq.(3) is

\[
\theta = -0.54 \cdot \tilde{P} + 1.02 \tag{5}
\]
and the EER improved from 3.5% to 2.5%.

We also evaluated other conditions for a specified number of training utterances (20,30,50). The results of parameters are shown in Table 1 and the results of verification performance are shown in Fig.3. In terms of verification performance, the EER decreased in proportion to the amount of training data and our new methods improved verification performance by about 30% compared with the conventional methods with a common threshold in every condition (2.1% to 1.4% using 50 training utterances). This figure also shows that \( \alpha \) and \( \beta \) have less variety compared with a common threshold \( \Theta \). This means that \( \alpha \) and \( \beta \) are relatively constant coefficients for determining a priori thresholds.

### 4.4 A Priori Threshold Decision

To evaluate the methods to set a priori thresholds, we evaluated other conditions for specified number of training utterances (20,30,50) using the same values estimated by 15 training utterances for \( \alpha, \beta \) and \( \Theta \) (\( \alpha = -0.54, \beta = 1.02, \Theta = -2.87 \)). The results are shown in Fig.4. In the conventional methods, the two error rates FAR and FRR grew farther apart as the training data increased. This shows that the conventional methods are not robust for determining a priori thresholds versus various amounts of training data. In our new methods on the other hand, the FAR and FRR were nearly equal to the EER under every condition. This means that our new methods make it possible to determine stable a priori thresholds regardless of the amount of training data.

<table>
<thead>
<tr>
<th>Number of</th>
<th>15</th>
<th>20</th>
<th>30</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>-0.54</td>
<td>-0.54</td>
<td>-0.53</td>
<td>-0.53</td>
</tr>
<tr>
<td>( \beta )</td>
<td>1.02</td>
<td>1.01</td>
<td>0.96</td>
<td>1.01</td>
</tr>
<tr>
<td>( \Theta )</td>
<td>-2.87</td>
<td>-2.62</td>
<td>-2.33</td>
<td>-2.11</td>
</tr>
</tbody>
</table>

**Table 1:** Parameters for each number of training utterances.

**Fig.3:** Verification performance using a posteriori threshold.

**Fig.4:** Verification performance using a priori threshold.

### 5. CONCLUSION

We have proposed new methods to determine a priori thresholds in speaker verification using speaker adaptation gain in likelihood during training speaker models without an extremely large amount of speech data or computations. As one reason to explain why setting stable thresholds is difficult, we showed that verification score after likelihood normalization, which is also effective in speaker verification, still varies widely according to speaker adaptation gain. We also showed that there are some correlations between speaker adaptation gain and normalized likelihood. We therefore estimated the threshold by the linear function of speaker adaptation gain. In our experiments to set a posteriori thresholds for different amounts of training data, our new methods improved verification performance by about 30% compared with the conventional methods using a common threshold for all speakers. In our experiments to set a priori thresholds, we found that the estimated coefficients of the linear function \( \alpha \) and \( \beta \) are relatively constant regardless of the amount of training data and that the proposed methods make it possible to set stable a priori thresholds. We also demonstrated that our new methods to set a priori thresholds are robust regardless of the amount of training data and are effective in improving verification performance with only the training data for the speaker models.

### 6. REFERENCES


