PRUNING ABNORMAL DATA FOR BETTER MAKING
A DECISION IN SPEAKER VERIFICATION

Jun-Hui LIU and Ke CHEN
National Laboratory of Machine Perception and The Center for Information Science
Peking University, Beijing 100871, P. R. China
E-mail: chen@cis.pku.edu.cn, URL: http://speech.cis.pku.edu.cn

ABSTRACT
In development of a speaker verification system, a priori threshold estimation is often needed based on a training set for decision making. Such a threshold critically determines performance of a speaker verification system. From a statistical viewpoint, a speaker's voice could be modeled by a certain distribution. Thus, data for training are only some samples of this distribution in a subspace, and the statistical information acquired from the training set is usually biased to that of the whole space. In this paper, we propose a method for better estimation of underlying statistics by abnormal data elimination. Without use of more data, our method provides an alternative way to improve performance of those statistics-based a priori threshold estimation methods in terms of generalization capability. On the basis of a benchmark database, KING, and a baseline system with a priori threshold estimation, we demonstrate the effectiveness of our method.

1. INTRODUCTION
As one of the most important fields in biometrics, speaker recognition is to automatically recognize a speaker based on his/her voice. Recently, it has been increasingly demanded for security in information access of an intelligent speech information system [1]. Speaker recognition includes verification and identification. Speaker verification is to accept or reject the identity claim of a specific speaker. Furthermore, a speaker verification system could be either text-dependent or text-independent. Text-dependent means that the same text is used in training and test. In contrast, any text is allowed in process of either training or test in a text-independent system.

A central issue in speaker verification is how to make a decision. Given a threshold, a speaker verification system might make two types of mistakes: one is false reject (FR) where a true speaker's identity claim is denied and the other is false accept (FA) where an imposter is accepted. For evaluation of a speaker verification system, a criterion, equal error rate (EER), is usually used, where a threshold makes the FR error equal to the FA error. Although this threshold can ensure the optimal performance in a given data set, it fails to produce good performance for other data sets beyond training due to overfitting. To tackle this problem, a priori threshold estimation methods have been proposed based on the statistical information in decision space acquired from a training set [2][3]. The use of such a threshold leads to better performance in other sessions in contrast to the use of the EER threshold through the performance based on the EER threshold is better in the training set [2][3].

From a statistical viewpoint, a speaker's voice could be modeled as a certain distribution. Thus, data for training are only samples of this distribution in a subspace, and the statistical information acquired from the training set is usually biased to that of the whole space. However, a priori threshold has to be estimated based on only an arbitrary data set; i.e. a training set. Thus, the use of a priori threshold causes the performance of a speaker verification system to be degraded in field application due to inaccurate statistics. In order to improve performance of a speaker verification system, parameter update is usually based on the data available beyond training. Obviously, an update method needs more data and expensive maintenance since a supervised way is often needed in update. Unlike previous methods, we propose a method based on only the training set to improve generalization capabilities of statistics-based a priori threshold estimation. The basic idea underlying our method is to tailor training data so that statistics estimated from only the training set can be closer to the true one. For this purpose, we develop an abnormal data elimination procedure in the decision space. To demonstrate the effectiveness of our method, we apply the proposed method to text-independent speaker verification by means of a Gaussian mixture model (GMM) baseline system and a benchmark database, KING. Simulation results show that
the use of our method yields considerable improvement in all sessions in comparison with those without pruning data.

The rest of this paper is as follows. Section 2 presents our method. Section 3 describes our baseline system. Section 4 reports simulation results. Conclusions are drawn in the last section.

2. ABNORMAL DATA ELIMINATION

In this section, we first pose a problem in statistics-based threshold estimation. Then, we present our idea to tackle the problem. As a result, we give a procedure to better estimate underlying statistics based on only a training set.

In order to build a speaker verification system, a set of data have to be collected first for training. Obviously, such a training data set collected within a limited period carries only some speaker’s characteristics on a certain condition. It is well know that a person’s voice always changes over time and, moreover, is affected by many factors; e.g. environment and mood. To make this point intuitive, we take a sketch score map to give an explanation in terms of a speaker verification system. Assume that we could gather all the voice of speakers, we could have a global scenario of scores produced by the speaker verification system, as illustrated in Figure 1(a). Obviously, the global map is unavailable or unobservable. In reality, observable scores forms a local map since they are achieved from only the training set collected in a certain period, as shown in Figure 1(b). Although this local map contains some general features of speakers, it also conveys other special features resulting from specific utterances, mood, and environment variation in the recording period. Our observation indicates that scores in a local map are often scattered in a certain small region and deviated to a certain direction in contrast to the global map. Thus, statistics estimated from the local map must be biased to that of the global one, which poses a problem – how to achieve relative accurate estimation to underlying statistics carried by the global map only based on a local map. Here we emphasize that this problem is unavoidable in statistics-based a priori threshold estimation.

For the aforementioned problem, those special features may make statistics estimated biased to a specific circumstance, which leads to poor performance in a speaker verification system. Our further observation indicates that in a local map most of scores are concentrated in a small region, and a few of scores are scattered far from the small region as illustrated in Figure 1(b). The global map in Figure 1(a) could be viewed as the union of all the local maps. Therefore, the biased statistics results from the portion of scores farthest from the small region. In this paper, we name those scores caused by special conditions abnormal data. For achieving better estimation, the basic idea underlying our method is to prune a proper amount abnormal data. In order to prune abnormal data, we specify the following criteria based on our observation: a) Most of the scores on a training set should belong to normal data. (b) The remaining scores after pruning should be densely distributed. (c) Pruning should be done first from the farthest score points. According to the above criteria, Figure 2(b) depicts an expected result by pruning the local map illustrated in Figure 2(b) for instance. Note that here we take only mean of samples into consideration and, thus, the mean estimated from Figure 2(b) is closer to the true one estimated from Figure 1(a) in contrast to that estimated from Figure 2(a).

Following the aforementioned criteria, we propose an algorithm to prune abnormal data pruning as follows:

1. Calculate the mean of the current data set to be pruned.
2. Find a data point of the largest deviation by searching all the data points in the current data set. Hereinafter, we call the data point the most abnormal point.

Figure 1: Score map. (a) Global map. (b) Local map.

Figure 2: Local score map. (a) The map before pruning. (b) The resulting map after pruning.
3. Eliminate this data point from the current data set.
4. Re-calculate the mean of the new data set by excluding those data eliminated.
5. Repeat steps 1-4 until certain conditions are satisfied.

Note that termination in our algorithm critically determines the accuracy of statistics estimated after pruning. In our simulations, we simply use a number of data eliminated as a termination condition.

In our algorithm, direct calculation of statistics could be expensive when the number of data are large. To speed up the calculation, we introduce an incremental algorithm for calculating the mean of the remaining data fast. The incremental formula is as follows:

$$\mu_{n-1} = \frac{\mu_n \cdot n - D_n}{n-1},$$

where $D_n = \arg_{x_n} \max_{\mu} \| x_n - \mu \|$. $D_n$ is the most abnormal point in the current data set containing $n$ data points. $\mu$ is the mean of the data set, and $\mu_{n-1}$ is the mean of the data set excluding $D_n$. Apparently, data point $D_n$ is eliminated from the data set, and the new data set contains only $n-1$ data points. After an abnormal datum is pruned, our algorithm causes the re-estimated mean of the data set to be closer to the true one, as illustrated in Figure 2.

### 3. SYSTEM OVERVIEW

In this section, we briefly describe our baseline system – a GMM based speaker verification system.

In our baseline system, preprocessing and feature extraction are summarized as follows. An utterance is first segmented in to 25 ms frames. Then, each frame is examined through the use of an energy-based detector in order to rule out those unvoiced parts. The remaining frames corresponding to voice parts are normalized, and Mel-scaled cepstral vectors are derived from the speech frames. As a result, 16-order cepstrum is used as a feature vector in our baseline system.

For speaker verification, we employ GMM to specify speaker’s statistical characteristics. A GMM is described as follows [4]:

$$P(\tilde{X} | \lambda) = \sum_{i=1}^{M} p_i b_i(\tilde{X}),$$

where $p_i$ are mixture coefficients subject to $\sum_{i=1}^{M} p_i = 1$. $b_i(\tilde{X})$ are component Gaussian distributions in the following form:

$$b_i(\tilde{X}) = \frac{\exp\{ -\frac{1}{2} (\tilde{X} - \mu_i)^T \Sigma_i^{-1} (\tilde{X} - \mu_i) \}}{(2\pi)^{D/2} | \Sigma_i |^{1/2}}.$$  

Here, $\lambda = (p, \tilde{\mu}, \Sigma)$ is the set of all the parameters. $D$ is the dimension of $\tilde{X}$ and $M$ is the number of components. $\tilde{\mu}$ and $\Sigma$ are the mean vector and the covariance matrix of the $i$-th component, respectively. As a result, its log-likelihood function is defined in terms of all the samples:

$$L(\lambda) = \sum_{t=1}^{T} \log p(\tilde{X}_t | X, \lambda).$$

Here, $\tilde{X}_t$ is the $t$-th input vector and $T$ is the number of input vectors. Given a training set, parameter estimation in GMM is often performed by an Expectation Maximization algorithm [4].

Once GMM gets fit to training data, a decision rule is required for verification. In our baseline system, we use a statistics-based a priori threshold estimation method to achieve a proper threshold. The estimation formula is as follows [3]:

$$\theta = \beta \overline{\mu} + (1 - \beta) \bar{\mu}.$$  

Here, $\theta$ is the demanding threshold for a specific speaker. $\mu$ and $\overline{\mu}$ are the mean of speaker’s score and imposter’s score, respectively. $\beta$ is a weight achieved by optimization over all the speakers. Based on thresholds achieved, speaker verification is performed by comparing the likelihood value in (3) with a threshold.

### 4. SIMULATIONS

In this section, we report simulation results by incorporating our method into a text-independent speaker verification system. We first describe the database used and an experimental procedure. Then we present two sets of simulation results corresponding to those without and with update.

The database used in our simulations is a wide-band part of the KING database where there are 49 male speakers. The KING database is a collection of conversational speech and has
become a benchmark for speaker recognition [4]. For each speaker there are ten conversations of around 45 seconds. The speech was recorded in 10 sessions, denoted by S01-S10. For each speaker, we used utterances of 60 seconds belonging to two sessions (30 seconds for each), S01 and S02, to train GMM and utterances of 30 seconds in one session, S03, to achieve thresholds in our simulations. The remaining sessions, S04-S10, were used for test. In our pruning method, the termination condition is when 25% of abnormal data are eliminated. The effect of abnormal data pruning is evaluated on the KING database in terms of verification error rates.

Figure 3 illustrates error rates produced by our baseline system without and with use of abnormal data pruning. Evidently, the use of our pruning procedure makes our baseline system perform better consistently for different sessions. Overall, the averaging error rate on different sessions is reduced from 10.10% to 9.79%. It is worth pointing out that the equal error rate (EER) of all the testing data is 7.95%. Here we emphasize that the gain in error reduction is achieved without use of more data.

Figure 3: Performance of our baseline system without and with use of abnormal data pruning. Here black bar denotes results without pruning, while white bar indicates results with pruning.

To improve performance of a speaker verification system, a classical method is to update thresholds as long as more data are available. In our simulation, we also use our update method together with our pruning procedure to evaluate performance of our pruning method in that circumstance.

Now we briefly describe our update technique [5]. In our method, threshold update is performed via the update of statistics used in (4) based on new data available. The update of statistics is done in an incremental way. Once a new datum arrives, the statistics used in (4) is re-estimated in the following way:

\[
\mu_{n+1} = \frac{\mu_n \cdot n + x_{n+1}}{n+1}.
\]

Here, \( x_{n+1} \) is the score corresponding to an arrival utterance. \( \mu_n \) is the mean of original scores corresponding to \( n \) existing utterances. \( \mu_{n+1} \) is the updated mean. When a new utterance comes, both its owner’s score and all the imposter’s scores need to be updated simultaneously.

Due to the incremental property in our update method, error rates are estimated in the following way: Given a threshold, test is performed for all the coming utterances. The result is viewed as that without update. The threshold updated by an arrival utterance is used to make a decision for next coming utterance. The testing result is recorded as that with update. The final error rate for update is estimated by accumulation of all the testing results.

Figure 4 depicts error rates by our baseline system without and with update on the same condition for pruning, i.e. 25% abnormal data are eliminated. From Figure 4, it is evident that the use of our update method leads to significant improvement consistently for different sessions. Overall, the averaging error rate on different sessions is reduced from 8.14% to 7.84% by update. Once again, we point out that the EER of overall testing data belonging to different sessions is 7.95%.

5. CONCLUSION

We have presented an abnormal data pruning method for better making a decision in a speaker verification system. Simulation results demonstrate the effectiveness of our method. Incorporating by our update method, in particular, our method makes a GMM based speaker verification system yield significantly good performance. As a matter of fact, our method has also been evaluated on different databases and simulation results not reported show the effectiveness of our method [6]. As a result, our method provides an easy-to-use way to improve any statistics-based a priori threshold estimation without need of more data.

There are several issues for our method to be investigated. Our ongoing studies are towards solving the following problems: how to tailor abnormal data for estimating underlying higher-order statistics and how to automatically determine which data are abnormal and how many of them to be eliminated in terms of a given data set.

ACKNOWLEDGMENTS

We are grateful to X.K. Qing for his providing a GMM speaker verification system used as a baseline system for our simulations. This work was partially supported by a National Key Fundamental Research Project Grant (G1999032708) in China.
REFERENCES


