PERGAMON

Pattern Recognition 36 (2003) 329–346

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Towards better making a decision in speaker verification

Ke Chen*

School of Computer Science, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

Abstract

Speaker verification is a process that accepts or rejects the identity claim of a speaker. How to make a decision is a critical problem; a threshold for decision-making critically determines performance of a speaker verification system. Traditional threshold estimation methods take only information conveyed by training data into consideration and, to a great extent, do not relate it to production data. It turns out that a speaker verification system with such threshold estimation suffers from poor performance in reality due to mismatches. In this paper, we propose several methods towards better decision-making in a practical speaker verification system. Our methods include the use of additional reliable statistical information for threshold estimation, elimination of abnormal data for better estimation of underlying statistics, and on-line incremental threshold update. To evaluate the performance of our methods, we have done simulations based on a baseline system, Gaussian Mixture Model, in both text-dependent and text-independent modes. Comparative results show that in contrast to the recent threshold estimation methods our methods yield considerably better performance, especially on miscellaneous mismatch conditions, in terms of generalization. Thus our methods provide a promising way for real speaker verification applications. © 2002 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Speaker verification; Decision-making; Threshold estimation; Abnormal data elimination; Incremental update; Mismatch; Generalization; KING corpus

1. Introduction

As one of the most important fields in biometrics, speaker recognition is a process that automatically authenticates a personal identity based on his/her voice. Although speaker recognition includes diverse tasks that discriminate people in terms of their voices, most of studies focus on speaker identification and verification. While speaker identification is to classify an unknown voice token as one of reference speakers, speaker verification is to accept or reject an identity claim. Moreover, a speaker recognition system often works in either of two operating modes: text-dependent and text-independent. By text-dependent, the same or known text is used for training and test. In contrast, any text is allowed to be uttered in the process of either training or test in the text-independent mode. As argued by Doddington et al. [1], it seems that the speaker verification task is of the greatest application potential though the speaker identification task appears to attract substantial scientific interests. Recently, speaker verification has been increasingly demanded for security in miscellaneous information systems [1–3]. Therefore, the development of effective speaker verification technologies for use in reality is of utmost importance, which recently has been recognized by both academic and industrial communities [1,2,4].

A central issue in speaker verification is how to make a decision. Essentially, a speaker verification system could make two types of mistakes during decision-making; one is the *false acceptance* (FA) which causes an impostor to be accepted, and the other is the *false rejection* (FR) which causes a genuine speaker's identity claim to be denied. A substantial task in decision-making is somehow to minimize both FA and FR errors during decision-making. Unfortunately, decision-making was regarded as a simple task and diminished in speaker recognition researches [1]. More recently, however, this problem has been found to be very

* Corresponding author. Tel.: +44-121-414-4769; fax: +44-121-414-4281.

E-mail address: k.chen@cs.bham.ac.uk (K. Chen).

PII: S0031-3203(02)00034-1

challenging for those who actually have been working for real operational systems [1,2]. It has been realized that the actual decision process must be considered to be a part of any comprehensive speaker verification study and the capability of a system to make proper decisions should be an integral part of evaluation [1].

Although the decision-making studies were not highlighted in previous speaker verification researches, there have been a few approaches developed for decision-making. These approaches could be roughly classified into two categories: a priori threshold setting [2,5–12,14] and a posteriori threshold setting [2,13,15]. The basic idea of a priori threshold setting is to find a proper threshold somehow based on a training set, then the resultant threshold will be applied to make a decision during test for any claimed voice token. The a priori threshold setting gives a feasible way to create real speaker verification systems. On the other hand, the a posteriori threshold is often set by finding the threshold of equal error rate (EER) that makes the FA rate equal to the FR rate for a given speaker system. In contrast to the a priori threshold setting, the a posteriori threshold setting provides a way to evaluate the discrimination capabilities of a particular speaker model in terms of a certain data set. Although such a method allows us to compare objectively the modeling performance, it is ultimately unrealistic for a real application.

On the basis of statistical hypothesis-test theory, some normalization approaches have been proposed to help the aforementioned threshold setting methods against mismatches [2,16–27]. The idea underlying such a sort of approaches is to create a reference model either associated with a speaker model or independent of speaker models. Thus scores produced by a speaker model are somehow normalized by the reference model. For building a reference model, there are usually two approaches; i.e. cohort and world models. The cohort model approach is to find a set of speakers whose characteristics in speech are similar to a specific speaker such that a cohort model can be built for modeling cohort speakers' ensemble characteristics [16], while the world model approach is to build a universal speaker model on a pool of speech utterances produced by various speakers [17]. It has been widely reported that by incorporating a reference model the performance of a priori or a posteriori threshold setting methods may yield the improved performance since the score variability due to mismatches is minimized by the reference model. As pointed out by Doddington et al. [1], however, speaker normalization sometimes may lead to worse performance. Therefore, it is a non-trivial task to build an effective reference model for the normalization in decision-making [2,4].

From a statistical viewpoint, acoustic characteristics of a speaker could be modeled with a certain distribution and, thus, training data for building a system are viewed as samples drawn from a speaker distribution in some subspace. All the a priori threshold setting methods work based on a given training set. Therefore, statistical information acquired from training data plays a critical role for setting a decision threshold, which poses a central problem how to take advantage of available statistical information effectively. As a matter of fact, there are only some limited data available for training in the development of a real operational system. Thus, it is unavoidable that the statistics estimated from the training data may be unreliable and inconsistent with those underlying a specific distribution. In the previous a priori threshold setting studies, most efforts were made in finding the reliable statistics and utilizing them to estimate a decision threshold [2,5–9,11,14]. To our knowledge, however, there is little consideration on the consistency of statistics between the training and the production data in those methods. Apparently, the inconsistent statistics could produce an improper decision threshold, which results in poor generalization for those utterances used beyond training. On the other hand, a fixed decision threshold is always subject to limitation in generalization since it is achieved based on only limited statistical information conveyed by a training set. Fortunately, new data should be always available as long as a system is used in reality. The availability of new data provides possibilities to update a decision threshold and, therefore, threshold update is also a basic issue in a real speaker verification system [9,14].

In this paper, we present several methods to handle the aforementioned issues in terms of the statistics-based a priori threshold setting framework. Our efforts include the use of additional reliable statistical information for threshold estimation, elimination of abnormal data to achieve the consistent statistics, and on-line incremental threshold update. As a consequence, our methods provide several simple yet effective technologies towards better making a decision in an operational system. For performance evaluation, we have done simulations based on a GMM-based speaker verification system [28,29] in terms of different databases including a benchmark corpus KING [30]. In particular, different mismatch conditions have been considered in simulations to thoroughly evaluate the performance of our methods. Comparative results in different operating modes show the effectiveness and robustness of our methods in contrast to some sophisticated a priori methods. In particular, the results of our autonomous on-line threshold update indicate that our work provides promising decision-making technologies in the development of a real speaker verification system.

The remainder of this paper is organized as follows. Section 2 reviews several sophisticated threshold estimation methods. Section 3 presents our methodologies towards better making a decision in speaker verification. Section 4 reports simulation results. Further discussions are given in Section 5, and conclusions are drawn in Section 6.

2. Review of threshold estimation

To make this paper self-contained, we briefly review threshold estimation methods with an emphasis on the a

priori speaker-dependent threshold setting. We first describe the theoretical background of threshold setting and then present the statistics-based a priori threshold estimation methods where most of them will be applied in our simulations for comparison.

2.1. Background

On the basis of statistics, we model acoustic characteristics by probabilistic models. Let S and $\bar{S}$ denote a speaker and non-speaker.[1] Therefore, both speaker S and non-speaker $\bar{S}$ can be described by probabilistic models. Since speaker verification is a process of making a decision accepting or rejecting a claimed identity by a system, we denote the acceptance decision as $\hat{S}$ and rejection decision as $\hat{\bar{S}}$, respectively. According to Bayesian decision theory, an optimal decision can be made by minimizing the following cost function [31]:

$$C = P(\bar{S})P(\hat{S}|\bar{S})C_{\hat{S}|\bar{S}} + P(S)P(\hat{\bar{S}}|S)C_{\hat{\bar{S}}|S}. \quad (1)$$

Here, $P(S)$ and $P(\bar{S})$ are the a priori probabilities of the claimed speaker to be the speaker S and to be non-speaker $\bar{S}$. $P(\hat{S}|\bar{S})$ and $P(\hat{\bar{S}}|S)$ are, respectively, the probabilities of an FA and of an FR, while $C_{\hat{S}|\bar{S}}$ and $C_{\hat{\bar{S}}|S}$ denote the corresponding costs of a null hypothesis for a true acceptance and a true rejection.

We denote a speech utterance claimed as belonging to speaker, S, as $\mathscr{U}$. Thus, the minimization of the cost function in Eq. (1) leads to the Bayesian optimal decision rule [32]:

$$\text{If } \frac{p(\mathscr{U}|S)}{p(\mathscr{U}|\bar{S})} \geqslant T_B, \text{ then } \textit{accept} \text{ the claimed identity.} \quad (2a)$$

$$\text{If } \frac{p(\mathscr{U}|S)}{p(\mathscr{U}|\bar{S})} < T_B, \text{ then } \textit{reject} \text{ the claimed identity.} \quad (2b)$$

Here, $p(\mathscr{U}|S)$ stands for the value of the claimed speaker's probability density function (PDF) and $p(\mathscr{U}|\bar{S})$ denotes the PDF of the non-speaker distribution. T_B is the Bayesian threshold:

$$T_B = \frac{C_{\hat{S}|\bar{S}}}{C_{\hat{\bar{S}}|S}} \frac{P(\bar{S})}{P(S)}. \quad (3)$$

From Eq. (3), it is apparent that the optimal Bayesian threshold depends upon only the cost ratio of FA to FR as well as the a priori probability ratio of impostors to the speaker. When a real application domain is fixed, the costs, $C_{\hat{S}|\bar{S}}$ and $C_{\hat{\bar{S}}|S}$, could be estimated based on the prior knowledge and desires from a real application domain. In the circumstance of no prior knowledge, two costs are often assumed to be equal. Thus, the optimal threshold relies on only the a priori probability ratio. When speaker and impostors are further assumed to be a priori equal-probable, the value of the corresponding decision threshold is one, which results in a minimum of the half total error rate (HTER) as

$$\text{HTER} = \frac{P(\hat{S}|\bar{S}) + P(\hat{\bar{S}}|S)}{2}. \quad (4)$$

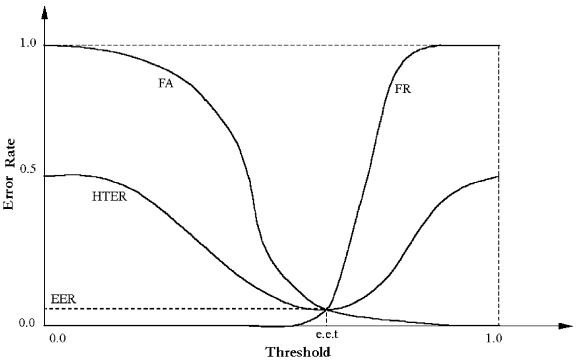


Fig. 1. The relationship among different error measures in speaker verification.

Intuitively, the HTER in Eq. (4) can be viewed as the average of the normalized FA and FR rates, which forms a useful measure to evaluate the overall performance of a speaker verification system. To understand different measures better, Fig. 1 illustrates the relationship among the FA, FR, HTER, and EER. Based on the definition in Eq. (4), the HTER will be equal to the EER of a system only if the threshold is set as an equal error threshold (e.e.t.). Since the change of FA and FR is not continuous as a threshold varies in reality, there may be no exact case that FA equals FR. Fortunately, the above fact provides a practical way to approximate the EER of a system. In this circumstance, the EER is approximated through use of the corresponding HTER on the condition that the absolute value of difference between FA and FR is minimal. As a result, we will take advantage of the HTER to approximate the EER and evaluate the performance of an operational system in our simulations.

Although an optimal decision threshold is theoretically available by Eq. (3), it is hardly applicable in practice since the PDFs in Eq. (2) cannot be achieved directly and often need to be estimated based on the likelihood functions of statistical models. Due to limited data available during enrollment, the estimation from the likelihood functions does not exactly match true speaker and non-speaker distributions. Thus, we have to adjust the threshold for decision-making in order to compensate for mismatch between the model and the data. As a consequence, a proper speaker-dependent threshold T_S, which can be modeled as a function of the optimal threshold T_B, is demanded based on a registration population.

[1] Given a specific speaker, it often refers to the rest of registration speakers (pseudo-impostor) in an operational system.

2.2. Statistics-based a priori threshold estimation

If sufficient data are achieved during enrollment, a classical method for threshold setting is to find an e.e.t. Then the e.e.t. would be used for decision-making in a system. From the viewpoint of statistics, sufficient data could resist mismatch and lead to robust decision-making. Unfortunately, the enrollment materials in reality are limited so as to hinder the estimation of a speaker-dependent threshold reflecting the variability in modeling accuracy. In this circumstance, some statistics underlying those data convey reliable information indeed for threshold setting. As a consequence, a sort of the statistics-based a priori threshold setting methods have been proposed based on the statistics underlying speaker's and non-speaker's scores. In the sequel, we shall review some sophisticated statistics-based a priori threshold setting methods.

Given a set of data associated with speaker and non-speaker, intra-speaker's and inter-speaker's scores are achieved, as illustrated in Fig. 2. The simplest way is to assume that these score distributions are subject to Gaussian distributions. A Gaussian density is in the following form:

$$G(x|\tilde{\mu},\tilde{\sigma}) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp\left(-\frac{(x-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right), \tag{5}$$

where $\tilde{\mu}$ and $\tilde{\sigma}$ are mean and variance parameters. Thus intra-speaker and inter-speaker distributions can be determined by the first- and the second-order moments estimated from observed data. If we denote as x_i (respectively $\bar{x}_i$) the score of a speaker (respectively non-speaker) for the ith speech segment,[2] then the statistics can be estimated as follows:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2, \tag{6a}$$

$$\bar{\mu} = \frac{1}{\bar{N}}\sum_{i=1}^{\bar{N}} \bar{x}_i, \qquad \bar{\sigma}^2 = \frac{1}{\bar{N}}\sum_{i=1}^{\bar{N}}(\bar{x}_i-\bar{\mu})^2. \tag{6b}$$

Here, μ and σ (respectively $\bar{\mu}$ and $\bar{\sigma}$) denote the estimate of mean and the variance on intra-speaker's scores (respectively inter-speaker's), and N (respectively $\bar{N}$) is the total number of speech segments belonging to speaker (respectively non-speaker) available for threshold setting. Thus, the speaker-dependent decision threshold (c.f. Fig. 2) is achieved [11] by setting

$$T_S = \arg_x(G(x|\mu,\sigma) = G(x|\bar{\mu},\bar{\sigma})). \tag{7}$$

To simplify the presentation, hereinafter, this threshold setting method is called *Gauss* method.

In addition, there is a statistical principle especially for a Gaussian distribution $G(x|\tilde{\mu},\tilde{\sigma})$; that is, 99.7% out of all the samples drawn from this distribution should be located within only the interval $[\tilde{\mu}-3\tilde{\sigma},\tilde{\mu}+3\tilde{\sigma}]$ [33]. Thus, application of this principle leads to a so-called 3σ method for threshold setting in the following form:

$$T_S = \begin{cases} \mu - 3\sigma & \text{if } \mu - 3\sigma > \bar{\mu} + 3\bar{\sigma}, \\ (\mu\bar{\sigma} + \bar{\mu}\sigma)/(\sigma + \bar{\sigma}) & \text{otherwise.} \end{cases} \tag{8}$$

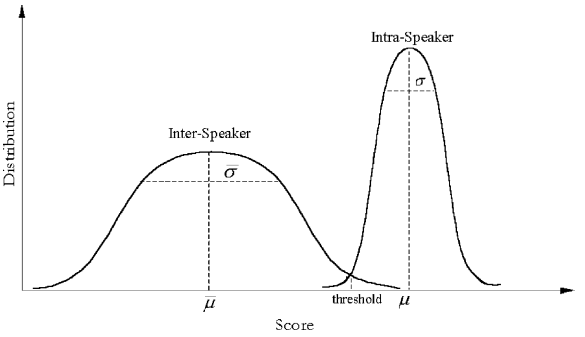


Fig. 2. Gaussian distributions of speaker's and non-speaker's scores.

The above threshold setting methods are often used for decision-making. However, the resultant decision threshold highly relies on the assumption of Gaussian distribution as well as training data. In general, the scores do not follow a Gaussian distribution and, furthermore, insufficient data may lead to a bias even though they follow a Gaussian distribution indeed. Thus, the Gaussian-distribution based methods often yield poor generalization. For better generalization, some statistics-based a priori threshold setting methods have been proposed without the assumption of Gaussian distribution.

To our knowledge, Furui first propose a statistics-based a priori threshold setting method regardless of the Gaussian distribution assumption [5]. The method uses the statistics of distances between utterances and templates in the framework of dynamic time warping (DTW), and the threshold setting becomes the determination of parameters in a linear combination of the estimate of mean and variance on the distances between different speakers in the non-speaker set. As a consequence, the threshold is set as

$$T_S = \alpha(\mu_D - \sigma_D) + \beta, \tag{9}$$

where μ_D and σ_D are the estimate of mean and variance on the distances between different speakers, and α and β are the speaker-independent parameters estimated based on inter-speaker's scores.

A procedure of parameter estimation is as follows. An exhausted search for the parameters, α and β, is performed within an interval given initially. For each value of α and β, the threshold is estimated by Eq. (9) and then used to achieve an HTER. The optimal parameters, α and β, are obtained only if they yield the minimal HTER. As argued by Furui [5], the motivation underlying this method is of two-fold. On the

[2] Here a speech segment refers to as an acoustic unit, e.g. frame, that produces a score by a speaker model.

one hand, intra-speaker's scores for different speakers are located within a narrow interval in contrast to inter-speaker's. On the other hand, the number of intra-speaker's scores is much fewer in comparison with that of inter-speaker's where they can be achieved by cross-comparison between different speakers. In terms of the text-dependent mode, therefore, only inter-speaker's scores are reliable, and thus, able to be used in threshold setting. Note that Eq. (9) was designed especially for the distance measure like that used in the DTW template matching. When Eq. (9) is incorporated with a probabilistic speaker model, e.g. GMM, it should be adapted to the following form:

$$T_S = \alpha(\bar{\mu} + \bar{\sigma}) + \beta. \tag{10}$$

Hereinafter, threshold setting in Eq. (10) is named *Furui* method.

CAVE (caller verification in banking and telecommunication) was a famous European project to develop speaker verification technologies towards real applications [2]. Researchers in this project have highly realized the importance of a priori threshold setting for an operational system, and thus, systematically investigated the statistics-based a priori threshold setting [2,10–12]. In the sequel, we review a number of a priori threshold setting methods developed in CAVE. To simplify the presentation, the three methods developed in CAVE will be called *CAVE-1*, *CAVE-2*, and *CAVE-3*.

Motivated by Furui's idea, Lindberg et al. propose a threshold setting method [11], and the threshold is set by a linear combination of the estimates of means on intra-speaker's and inter-speaker's scores. As a result, the threshold, in CAVE-1, is in the following form:

$$T_S = \gamma\mu + (1-\gamma)\bar{\mu}. \tag{11}$$

Like the parameters in Furui's method, γ is a speaker-independent parameter and optimized on a registration population.

Bimbot et al. develop a threshold setting method especially for a Gaussian model of the utterance logarithmic likelihood-ratio distribution [10]. Based on the optimal decision threshold in Eq. (3), a speaker-independent correction, δ, is introduced to adjust the estimate of the mean on intra-speaker's scores only. Thus, the statistics of intra-speaker scores are estimated as $\hat{\mu} = \mu - \delta$ and $\hat{\sigma} = \sigma$. Here, δ is optimized on a registration population as done in Furui's method. Thus, the threshold is obtained, in CAVE-2, by

$$T_S = \arg_x \left(\frac{G(x|\hat{\mu}, \hat{\sigma})}{G(x|\bar{\mu}, \bar{\sigma})} = T_B \right). \tag{12}$$

The speaker-independent threshold setting [2] is a classical way to achieve a decision threshold by optimizing the cost function in Eq. (1). Unfortunately, the speaker-independent threshold setting method demands sufficient data. Otherwise, the use of such a threshold tends to suffer from the poor generalization. In order to tackle this problem, Lindberg et al. propose a speaker-dependent adjustment [11]. In this method, a speaker-independent threshold, T_{SI}, is adjusted by considering the difference between the estimates of means on intra-speaker's and inter-speaker's scores. Thus, the improved threshold is obtained as a linear combination of the previous speaker-independent threshold and the adjustment, in CAVE-3, as follows:

$$T_S = T_{SI} + \eta(\bar{\mu} - \mu). \tag{13}$$

Note that both the speaker-independent threshold, T_{SI}, and parameter, η, in Eq. (13) are optimized on a registration population.

In summary, the basic idea underlying such a sort of methods is to estimate a speaker-dependent threshold by taking advantage of reliable statistics (either the first- or the second-order moments and even both) of intra-speaker's and inter-speaker's scores. Therefore, it should be emphasized that those statistics used in threshold setting critically determine whether a resultant threshold can generate the good performance or not.

3. Our methodologies

In this section, we present our methodologies towards better making a decision in the statistics-based a priori threshold setting framework. Our methods include an alternative statistics-based a priori threshold setting method by utilizing more reliable statistics for threshold setting, a pruning method for data selection in order to achieve better generalization, and an on-line incremental threshold update method.

3.1. Alternative a priori threshold setting

The idea underlying the a priori speaker-dependent threshold setting reviewed in Section 2.2 is to use the reliable statistics associated with speaker and non-speaker (pseudo-impostor) to customize a proper threshold for a specific speaker. Thus, the statistics of speaker's and non-speaker's scores would play a critical role in such a decision-making process. For use of statistics, it becomes a critical issue to investigate which one is reliable; that is, only reliable statistics are able to convey useful information. According to statistical properties, the first- and the second-order moments of intra-speaker's and inter-speaker's scores, μ, σ, $\bar{\mu}$, and $\bar{\sigma}$, are common candidates for threshold setting. Empirical studies including ours [2,5,29] indicate that for most of the speakers in an operational system, the estimate of variances on intra-speaker's scores are quite similar such that they are hardly distinguishable in contrast to the estimate of their means. On the other hand, the variances of intra-speaker's scores corresponding to a few speakers may be quite large indeed, which leads to large EERs. In particular, the estimate of variance is biased, in

particular, as only limited data are available. Therefore, the estimate of variance on intra-speaker's scores is unreliable, and thus, unable to be used in the a priori threshold setting, which has been realized by a few of researchers [2,5]. In contrast, the empirical studies including ours indicate that the rest three statistics, μ, $\bar{\mu}$, and $\bar{\sigma}$, perform reliably, which provides useful information for threshold setting.

Motivated by the previous work of Furui [5] and Lindberg et al. [11], we propose an alternative statistics-based a priori method for threshold setting. Unlike their methods, ours attempts to make a proper use of all the reliable statistics available. We believe that more reliable statistics provide more useful information and thus, the proper use of more reliable statistics may lead to a better threshold for decision-making. As a consequence, an alternative speaker-dependent threshold is estimated by a linear combination of all there reliable statistics mentioned above:

$$T_S = b(\bar{\mu} + a\bar{\sigma}) + (1 - b)\mu, \tag{14}$$

where a and b are two speaker-independent parameters and optimized on a registration population. Thus, Eq. (14) encodes the useful information conveyed by the reliable statistics, and the decision threshold becomes a monotonically increasing function of μ, $\bar{\mu}$, and $\bar{\sigma}$.

3.2. Pruning abnormal data for better generalization

From a statistical viewpoint, different speakers' voice could be modeled as different distributions, which makes them discriminable. In reality, the data for building a speaker verification system are only some samples drawn from this distribution corresponding to a speaker in a certain subspace. For training, only the data recorded in a couple of sessions are usually available during enrollment. It implies that such a training data set collected within a short period carries only some speakers' characteristics on a certain condition. It is well known that speakers' voice always changes over time, and moreover, may be affected by many factors, e.g. verbal change, change of vocal tract, mood, healthy status, and channels as well as other mismatch conditions. It is highly agreed that a proper threshold should not only perform well on the data used for estimating it but also be of good generalization for the other data drawn from the same distribution on mismatch conditions. Therefore, it poses a problem how to capture the intrinsic characteristics underlying a specified speaker from only the limited data for better generalization. Although it is, more or less, mentioned in literature [1,2,4], this problem has not been emphasized yet. Our study has indicated that it is critical to set a proper threshold for good generalization in decision-making.

In order to understand the above problem better and facilitate the presentation of our idea, we figure out some sketch score maps to give an intuitive explanation in term of a speaker verification system. Assume that we could collect all the voice of a speaker, we could have a global scenario of scores, hereinafter called *global map*, as depicted in Fig. 3(a). Obviously, it is unrealistic to have such a global map since there is no way to observe all the data in a short period and predict all possible mismatch conditions. In reality,

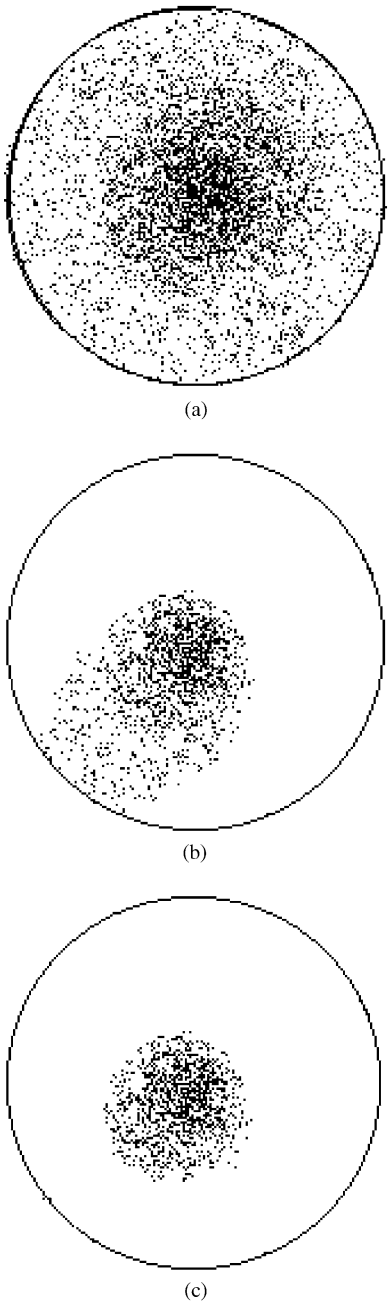


Fig. 3. Sketch score maps to explain the process of pruning abnormal data. (a) Global score map; (b) local score map before pruning; (c) local score map after pruning.

the observable scores by a speaker model form only a local landscape of scores, hereinafter called *local map*, as depicted in Fig. 3(b). Although this local map contains some intrinsic speaker characteristics indeed, it conveys some aforementioned session-dependent features as well. Moreover, these session-dependent features may be quite distinct from those recorded in other sessions beyond those observed currently (cf. Figs. 3(a) and (b)). Thus, statistics estimated from the local map are biased to those of the global map, which again poses the problem in question in another concrete form as follows. From the local map, how can we obtain the estimate of statistics relatively consistent with those of the global map? Here, we emphasize that the problem is unavoidable in any statistics-based a priori speaker-dependent threshold setting.

For the problem, the session-dependent features often make an estimate of statistics biased in order to fit a specific circumstance, which results in the poor generalization. Our further observation indicates that in a local map most of the scores are often concentrated in a certain small region and the rest of them are scattered from the small region, as depicted in Fig. 3(b). Thus, the global map in Fig. 3(a) could be regarded as the union of all the local maps. From this observation, we find that the bias of statistics, e.g. an estimate of mean, from a local map results from those scores farthest from the dense portion (a small region around the center of the circle in Fig. 3(a)). In this paper, we name those scores, caused by session-dependent conditions, *abnormal data*. Thus, a set of session-dependent data can be divided into two categories; one is the portion corresponding to the dense region, hereinafter called *normal data*, and the other is the abnormal data. The former likely carries the intrinsic characteristics of a speaker, and the latter probably indicates the special session-dependent features. For generalization, our idea is to prune a proper amount of abnormal data, to a great extent, for reducing the influence of session-dependent features.

In order to prune the abnormal data, we specify the following criteria based on our observation: (i) most of the scores on a training set should belong to normal data, (ii) the remaining portion after pruning should be densely distributed, thus the previous centroid should move towards that of all the data, as illustrated in Fig. 3(a), and (iii) pruning should be performed in a sequential way starting from the furthest away from the centroid of the global map (cf. Fig. 3(a)). According to the above criteria, Fig. 3(c) depicts an example of a desired outcome by pruning the local map in Fig. 3(b). Note that here we take only the estimate of the mean on intra-speaker's scores into consideration during pruning, thus the estimate of mean in the local map, as illustrated in Fig. 3(c), is more consistent with that of the global map in Fig. 3(a) in contrast to that of the local map in Fig. 3(b). It is also worth mentioning that the estimate of variance on intra-speaker's scores after pruning changes accordingly but does not affect threshold setting since it is never used in the a priori methods.

Following the proposed criteria, we develop an algorithm to prune abnormal data as follows:

1. For a specific speaker, estimate the mean and the variance from a given data set, $\{x_i\}_{i=1}^{N}$, say μ_N and σ_N^2.
2. For a sample x_i $(i = 1,\ldots,n+1)$, let $d_i = |x_i - \mu_N|$. Find the sample, x_{i^*}, which satisfies the condition $i^* = \arg\max_{1\leqslant i\leqslant n+1} d_i$, from the current data set of $n+1$ $(n < N)$ samples. Hereinafter, this sample is called *the most abnormal datum* in the current data set.
3. Eliminate the most abnormal data found in Step 2 from the current data set of $n+1$ samples.
4. Re-estimate the mean of the resulting data set of n samples.
5. Repeat Steps 2–4 until all the remaining samples satisfy the condition of $d_i < \kappa\sigma_N$ $(\kappa > 0)$.

In this algorithm, κ is a parameter to control termination. Note that a termination condition in Step 5 of our algorithm plays an important role for better generalization. In our simulations, we use the same parameter value of $\kappa = 2.0$ for all the speakers. How to tune the parameter κ for maximal generalization is a non-trivial issue that will be discussed later.

In our algorithm, the direct estimation of statistics suffers from expensive computational costs, in particular, as the number of training data is large. In order to speed up calculation, we propose an incremental algorithm to efficiently estimate the mean of the remaining data set after pruning. Suppose that μ_{n+1} is the estimate of the mean on the data set of $n+1$ samples. After a pruning operation, the new estimate of mean μ_n is as follows (for details, see Appendix A):

$$\mu_n = \frac{(n+1)\mu_{n+1} - x_{i^*}}{n}, \tag{15}$$

where $i^* = \arg\max_{1\leqslant x_i\leqslant n+1} |x_i - \mu_N|$ refers to the most abnormal datum in the data set containing $n+1$ samples before pruning. The incremental estimation in Eq. (15) is applied in Step 4 of our algorithm. After one of abnormal data is eliminated, apparently, our algorithm causes the estimate of the mean on the remaining data to be closer to the centroid in the global map (cf. Fig. 3).

3.3. On-line incremental threshold update

As pointed out previously, a session-dependent data set always contains only the limited useful information no matter how we manipulate these observable data. Thus, a fixed threshold based on the limited data does not always fit to a new landscape for most of speakers. On the other hand, an operational system should be accessed frequently. Therefore, new data for claimed speakers are available during running time, which provides a new information source to improve decision-making. Taking Fig. 3, we can give an intuitive explanation to the fact. When new data are available, they might change the local map in Fig. 3(b), and the estimate of statistics from the new local map tends to

vary. Thus, the re-estimation of statistics is required for a statistics-based a priori method, which leads to a problem how to update a threshold in terms of new data.

There have been some methods developed for threshold update [4,9,14]. The basic idea underlying those methods are to re-estimate statistics for threshold setting based on all the history data and the new data, which is equivalent to threshold setting based on a new training set consisting of both historical and appended data. Although they lead to the improved performance, such methods need to pool all the historical data, and hence, suffer from high spatial and computational costs as the number of speakers increases. For the statistics-based a priori threshold setting, a threshold update implies the need of the re-estimation of statistics used for threshold setting. Unlike the previous methods, we propose an on-line incremental threshold update method that needs only the estimate of statistics of historical data prior to threshold update and a coming datum. As a consequence, the statistics are re-estimated on-line as (for details, see Appendix A)

$$\tilde{\mu}_{n+1} = \frac{n\tilde{\mu}_n + \tilde{x}_{n+1}}{n+1} \tag{16}$$

and

$$\tilde{\sigma}_{n+1} = \frac{n(n+1)\tilde{\sigma}^2 + n(\tilde{x}_{n+1} - \tilde{\mu}_n)^2}{(n+1)^2}. \tag{17}$$

Here, $\tilde{\mu}_n$ and $\tilde{\sigma}_n$ are the estimates of the first- and the second-order moments used to set the old threshold. $\tilde{\mu}_{n+1}$ and $\tilde{\sigma}_{n+1}$ are the estimates of the statistics after a new datum $\tilde{x}_{n+1}$ is added.

In contrast to other threshold update methods [4,9,14], ours is of the following salient features: (i) instead of historical data themselves the estimate of their statistics is merely needed, (ii) the threshold is incrementally re-estimated on-line, and (iii) our on-line update illustrates an evolutionary process how the change of speakers' voice is captured for decision-making.

In summary, we have presented our efforts towards better making a decision in speaker verification. Note that the last two methods are also applicable to other statistics-based a priori speaker-dependent threshold setting methods to improve their performance. Here, we emphasize that all of our methodologies presented in this section merely work on the output space of a speaker model, and thus, are applicable to miscellaneous speaker models. In other words, parameters in a speaker model are unchanged and our pruning and update methods merely take effect on its output space.

4. Simulations

In this section, we present simulation results by applying our methods to both text-dependent and text-independent speaker verification tasks. In order to demonstrate the effectiveness of our methods, we also apply some sophisticated decision-making methods in literature to the same problems for comparison. For simulations, we use a GMM to model speaker's characteristics from the viewpoint of statistics. In the sequel, we first present a brief description on our simulations, including databases, preprocessing, and feature extraction in different operating modes. It follows by a brief review on the GMM-based speaker verification baseline system and a related standard performance evaluation method used in simulations. Finally, simulation results on different methods are reported.

4.1. Brief description

4.1.1. Databases

For different operating modes, we adopt two different databases in our simulations. We use the PKU-TD database in the text-dependent mode and the KING speech corpus [30] in the text-independent mode.

The PKU-TD database is composed of 35 speakers (30 male and 5 female) uttering ten isolated digits from zero to nine in Chinese (Mandarin). These data were collected in five separate sessions, labeled as $S_1, \ldots, S_5$, by speech information processing laboratory at Peking University. The interval between sessions varies from 1 week to 3 months. In each session, a speaker was asked to utter 10 times for any of 10 digits. All the digit utterances were recorded with an ordinary microphone.

The KING is an English speech corpus collected at ITT around 1987 and re-sampled in 1992 [30], in particular, for text-independent speaker recognition. It contains utterances from 51 male speakers in two versions; i.e. wide-band and narrow-band sets. The speakers are further divided into two groups, consisting of 25 and 26 people, in terms of different locations. There are 10 sessions, labeled by $S_{01}, \ldots, S_{10}$, and the interval between sessions varies from 1 week to 1 month. Typically, an utterance of each speaker is within a duration from 30 to 60 s. It is worth pointing out that some speech segments were missing in the wide-band set and only 49 speakers' utterances in all the 10 sessions are available. To facilitate comparison, therefore, we use the wide-band set of only those 49 speakers' utterances in our simulations. In the narrow-band set, the limited bandwidth and distorted transmission channel cause speech quality to be degraded severely. In particular, there are differences in spectral characteristics between sessions S_{01}–S_{05} and sessions S_{06}–S_{10} since speech is passed through different local telephone channels [30], which leads to miscellaneous mismatches. Signal-to-noise ratio for sessions S_{06}–S_{10} is about 10 dB worse than that for sessions S_{01}–S_{05}.

4.1.2. Acoustic analysis

For different databases, we use similar methods for acoustic analysis; i.e. preprocessing and feature extraction.

For the PKU-TD database, we use the following acoustic analysis: (i) pre-emphasis with the filter $H(z) = 1 - 0.95z^{-1}$, (ii) Hamming windowing speech by a frame size 25.6 ms

with a frame shift 12.8 ms, and (iii) extracting 16-order Mel-scaled cepstrum vectors.

For the wide-band set of KING database, the acoustic analysis is as follows: (i) pre-emphasis with the filter $H(z) = 1 - 0.95z^{-1}$, (ii) Hamming windowing speech by a frame size 25.6 ms but without any frame shift, (iii) unvoiced frame removal based on an energy measure, and (iv) extracting 16-order Mel-scaled cepstrum vectors. In simulations on the narrow-band set, we adopt a preprocessing procedure similar to that for the wide-band set. Besides, the mean subtraction technique [28] is applied to preprocessing and the weighted Mel-scaled cepstrum is further used for feature extraction, which results in the robustness to noise and degraded speech [4].

4.2. GMM-based speaker verification

Gaussian Mixture Model (GMM) has been used to characterize speaker's voice in the form of probabilistic model. It has been reported that the GMM approach outperforms other classical methods for text-independent speaker recognition [28]. Our recent work has shown that the GMM also performs well for text-dependent speaker recognition [29].

For a feature vector denoted as $\mathbf{x}_t$ belonging to a specific speaker s, the GMM is a linear combination of K Gaussian components as follows:

$$P(\mathbf{x}_t|\lambda_s) = \sum_{k=1}^{K} \omega_{s,k} P(\mathbf{x}_t|\mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k}), \tag{18}$$

where $P(\mathbf{x}_t|\mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k})$ is a Gaussian component parameterized by a mean vector $\mathbf{m}_{s,k}$ and covariance matrix $\mathbf{\Sigma}_{s,k}$. $\omega_{s,k}$ is a linear combination coefficient for speaker s ($s=1,2,\ldots,S$). Usually, a diagonal covariance matrix is used in Eq. (18). Given a sequence of feature vectors, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots\}$, from a specific speaker's utterances, parameter estimation for $\lambda_s = (\omega_{s,k}, \mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k})$ ($k = 1,\ldots,K,\ s = 1,\ldots,S$) is performed by the Expectation–Maximization (EM) algorithm [34]. Thus, a specific speaker model is built through finding proper parameters in the GMM based on the speaker's own feature vectors.

For text-dependent speaker identification, the log-likelihood value for each feature vector is used as a score for decision-making. In contrast, a sequence of feature vectors in text-independent speaker identification is divided into overlapping segments of T feature vectors, as suggested by Reynolds [28]:

$$\overbrace{\mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+T-1}}^{\text{segment } l}, \mathbf{x}_{l+T}, \ldots,$$

$$\mathbf{x}_l, \overbrace{\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+T-1}, \mathbf{x}_{l+T}}^{\text{segment } l+1}, \mathbf{x}_{l+T+1}, \ldots .$$

For a testing segment $X^{(l)} = \{\mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+T-1}\}$ of T frames, the log-likelihood value of a GMM is calculated as follows:

$$\mathscr{L}(X^{(l)}, \lambda_s) = \sum_{t=l}^{l+T-1} \log P(\mathbf{x}_t|\lambda_s), \quad s = 1,\ldots,S. \tag{19}$$

Thus, the smoothed likelihood value corresponding to a segment $\mathscr{L}(X^{(l)}, \lambda_s)$ becomes a score for decision-making. Note that the length of a segment, T, in the above testing method provides a way to evaluate the performance of a GMM system on utterances of different lengths in simulations. In other words, this parameter may simulate utterances of arbitrary lengths as its value varies.

In our simulations, we use 16 and 32 Gaussian components in GMMs for text-dependent and text-independent cases, respectively.

4.3. Simulation results

We have done simulations to evaluate the performance of our methods on the basis of two aforementioned databases, PKU-TD and KING. For comparison, we also apply other decision-making methods including those reviewed in Section 2 and the e.e.t. method, where an equal error threshold achieved from a validation set is used for decision-making on other testing sets, to the same task. Moreover, the overall EERs on all testing sets are also reported as a reference. In simulations, the HTER measure is employed for performance evaluation.

To create speaker models, we use the speech data recorded in session S_1 of the PKU-TD database and those in sessions S_{01} and S_{02} of the KING database, respectively, for parameter estimation of GMMs. Moreover, session S_2 in the PKU-TD and session S_{03} in the KING database are used as validation sets, respectively, for threshold setting in different operating modes. The remaining sessions in the PKU-TD and the KING databases are used for test. Thus, the total number of access in our text-dependent simulations is 1050 client access and 36,750 impostor access. For the text-independent case, the number of access depends upon the length of a testing segment and an utterance tested. For a segment of 5.12 s, typically, the overall number of client access is about 480,000 corresponding to utterances spanning from 30 to 60 s, while the total number of impostor access is around 24,500,000 corresponding to utterances of the same lengths.

4.3.1. Results of a priori threshold setting

First of all, we apply the e.e.t., several a priori threshold setting methods, and ours to two databases. During threshold setting, the data belonging to a specific speaker are used to estimate intra-speaker's statistics while those belonging to the other speakers in the same validation set are used as non-speaker or pseudo-imposter data. Thus, threshold setting is performed based on only the aforementioned validation sets for different operating modes in our simulations.

Table 1 shows the text-dependent performance of several methods on the PKU-TD database. Note that there are ten

Table 1
Comparative results on the PKU-TD database for several a priori threshold setting methods and the e.e.t. method

Method	S_3	S_4	S_5	Averaging
e.e.t.	9.34	12.86	10.93	11.04
Gauss	9.38	12.77	11.04	11.06
3σ	9.41	12.79	11.24	11.15
Furui	9.22	12.73	10.91	10.95
CAVE-1	9.12	12.54	10.82	10.83
CAVE-2	9.19	12.69	10.87	10.92
CAVE-3	9.16	12.63	10.84	10.87
Ours	8.47	11.70	10.33	10.17

GMMs, each corresponding to a digit, for each speaker. Here we report only the averaging HTERs in different sessions due to limited space. From Table 1, we observe that the e.e.t. and two Gaussian-distribution based methods (*Gauss* and 3σ) do not lead to good performance. In contrast, the statistics-based a priori methods yield better performance. In particular, ours results in the lowest HTERs in all three testing sessions. For reference, the overall EER of all three testing sets is 8.01%.

As mentioned in Section 4.2, we adopt a segment-based testing method for performance evaluation in the text-independent mode. Thus, testing segments of different lengths may result in different HTERs. For testing segments of 5.12 s on the wide-band set, we show their detailed HTERs in Table 2. Similarly, we also show the detailed performance on the narrow-band set in Table 3 by the use of testing segments of 7.68 s. It is evident from Tables 2 and 3 that the statistics-based a priori methods yield better performance though Furui's method, by using only the estimates of statistics of pseudo-impostor, leads to the relatively unsatisfactory performance. In contrast, the e.e.t. and two Gaussian-distribution methods produce poor performance. In particular, ours, by using the estimates of more reliable statistics, leads to the best performance on both the wide-band and the narrow-band sets.

Furthermore, Tables 4 and 5 shows the text-independent performance of different methods in terms of different testing segment lengths. In general, it is evident that the averaging HTERs for different methods decrease as the length of testing segments increases. By comparison, the statistics-based a priori threshold setting methods including ours yield much better performance while the other three methods produce poor performance. Note that the overall EERs on all seven wide-band testing sets are 17.94%, 7.95%, 5.13%, and 1.98%, respectively, corresponding to testing segments of 2.56, 5.12, 7.68, and 10.24 s, while the overall EERs on all seven narrow-band testing sets are 58.43%, 49.05%, 32.37%, and 27.92%, respectively, corresponding to testing segments of 2.56, 5.12, 7.68, and 10.24 s.

Table 2
Comparative results on the wide-band set of the KING database for several a priori threshold setting methods and the e.e.t. method

Method	S_{04}	S_{05}	S_{06}	S_{07}	S_{08}	S_{09}	S_{10}	Averaging
e.e.t.	13.95	13.64	13.67	14.89	14.68	15.50	14.01	14.33
Gauss	14.18	13.91	13.41	14.69	14.23	15.49	14.18	14.30
3σ	14.54	13.97	13.58	15.15	14.59	15.61	14.27	14.53
Furui	11.82	11.27	12.74	12.18	12.01	12.53	11.18	12.05
CAVE-1	11.34	9.17	10.68	11.87	11.24	12.04	10.25	10.94
CAVE-2	11.66	10.91	12.31	11.65	11.39	12.11	11.03	11.58
CAVE-3	11.43	9.63	10.97	11.93	11.33	12.17	11.07	11.22
Ours	11.26	8.86	10.43	11.57	10.91	11.78	10.68	10.78

The testing segment length is 5.12 s.

Table 3
Comparative results on the narrow-band set of the KING database for several a priori threshold setting methods and the e.e.t. method

Method	S_{04}	S_{05}	S_{06}	S_{07}	S_{08}	S_{09}	S_{10}	Averaging
e.e.t.	35.16	38.42	55.73	42.28	44.36	42.53	39.16	42.52
Gauss	36.44	37.98	54.28	43.96	44.13	44.29	38.31	42.77
3σ	37.25	39.10	57.13	44.54	45.56	45.17	40.48	44.18
Furui	32.31	33.64	50.59	40.79	41.18	42.34	38.89	39.96
CAVE-1	30.63	31.59	48.05	38.74	39.45	39.98	37.33	37.97
CAVE-2	31.45	33.04	49.93	39.54	40.57	40.53	39.09	39.16
CAVE-3	30.96	30.94	49.31	38.86	39.21	40.19	39.84	38.47
Ours	29.64	29.92	47.32	38.39	38.88	37.91	36.29	36.91

The testing segment length is 7.68 s.

Table 4
Comparative results on the wide-band set of the KING database for several a priori threshold setting methods at different testing segment lengths

Length (s)	e.e.t.	Gauss	3σ	Furui	CAVE-1	CAVE-2	CAVE-3	Ours
2.56	30.41	30.82	31.63	25.25	21.17	22.03	21.92	20.15
5.12	14.33	14.30	14.54	12.05	10.94	11.54	11.22	10.78
7.68	9.42	9.51	10.21	8.32	7.51	7.83	7.74	6.79
10.24	6.23	6.12	6.81	4.92	3.95	4.38	4.15	3.18

Table 5
Comparative results on the narrow-band set of the KING database for several a priori threshold setting methods at different testing segment lengths

Length (s)	e.e.t.	Gauss	3σ	Furui	CAVE-1	CAVE-2	CAVE-3	Ours
2.56	68.52	68.21	70.80	65.16	62.83	63.77	63.04	62.19
5.12	59.45	60.21	61.87	56.69	53.28	55.66	54.04	52.42
7.68	42.52	42.77	44.18	39.96	37.97	39.16	38.47	36.91
10.24	38.64	38.87	39.22	36.04	33.92	34.84	34.35	32.44

Table 6
Comparative results on the PKU-TD database for several a priori threshold setting methods with pruning abnormal data (cf. Table 1)

Method	S_3	S_4	S_5	Averaging
CAVE-1	9.94	13.15	10.91	11.33
(gain)	(−0.82)	(−0.61)	(−0.09)	(−0.51)
CAVE-3	9.57	12.75	10.82	11.05
(gain)	(−0.41)	(−0.12)	(0.01)	(−0.17)
Ours	8.99	11.80	10.61	10.47
(gain)	(−0.52)	(−0.10)	(−0.28)	(−0.30)

Comparative results show that our method consistently results in the best performance no matter how long the length of testing segments is. Thus, it turns out that the introduction of more reliable statistics in threshold setting yields better generalization in mismatch environments.

4.3.2. Results of pruning abnormal data

As argued in this paper, pruning abnormal data could improve the generalization capability of a statistics-based a priori speaker-dependent threshold setting method. In order to evaluate its performance, we have applied our pruning algorithm to three typical speaker-dependent methods; i.e. CAVE-1, CAVE-3, and ours. Prior to threshold setting, we apply our pruning algorithm to the validation set in order to eliminate the abnormal data.

Table 6 show the HTERs of three methods along with our pruning algorithm on the PKU-TD database. For comparison, we also present the error reduction rate (gain) after pruning in Table 6. It is observed from Table 6 that after pruning the overall performance of a GMM-based speaker verification system is degraded to some extent regardless of methods for threshold setting. This result indicates that our pruning algorithm seems inapplicable to the text-dependent mode, which will be discussed in the next section. By a fixed parameter, $\kappa = 2.0$, in our pruning algorithm, the averaging percentage of pruned data is 9.2%, and the maximal and minimal percentages of pruned data are 18.9% and 5.3%, respectively, for 35 speakers.

Now we report the simulation results through the use of our pruning algorithm on the KING database. Table 7 shows the detailed HTERs of three methods incorporated by our pruning algorithm on the wide-band set in terms of a testing segment of 5.12 s, and Table 8 presents the performance of three methods along with our pruning algorithm on the narrow-band set in terms of a testing segment of 7.68 s. For comparison, we also show the error reduction rate (gain) after pruning in Tables 7 and 8. Based on the simulation results, the application of pruning abnormal data causes the performance of three speaker-dependent threshold setting methods including ours to be improved consistently at different sessions. Moreover, it is observed from Table 8 that the gains on sessions S_{06}–S_{10} is larger than those on sessions S_{04} and S_{05} in general. As described in Section 4.1.1, the narrow-band set contains a large mismatch between S_{01}–S_{05} and S_{06}–S_{10} due to distinctive channels. Thus the simulation results imply that our pruning algorithm performs better for data collected on more mismatch conditions.

Based on the averaging HTERs at different sessions, furthermore, Figs. 4 and 5 illustrate the pruning effects for three methods on the wide-band and narrow-band sets in terms of different testing segment lengths. It is evident from Figs. 4 and 5 that the HTERs corresponding to different testing segments are consistently lowered after pruning. By a fixed

Table 7
Comparative results on the wide-band set of the KING database for several a priori threshold setting methods with pruning abnormal data

Method	S_{04}	S_{05}	S_{06}	S_{07}	S_{08}	S_{09}	S_{10}	Averaging
CAVE-1	10.33	8.07	9.91	11.49	10.31	11.14	9.42	10.10
(gain)	(1.01)	(1.10)	(0.77)	(0.38)	(0.93)	(0.90)	(0.83)	(0.84)
CAVE-3	10.85	8.67	10.21	11.66	10.25	11.38	10.15	10.45
(gain)	(0.58)	(0.96)	(0.76)	(0.27)	(1.08)	(0.79)	(0.92)	(0.77)
Ours	10.31	7.77	9.59	11.09	9.75	10.81	9.23	9.79
(gain)	(0.95)	(1.09)	(0.84)	(0.48)	(1.16)	(0.97)	(1.45)	(0.99)

The testing segment length is 5.12 s, and here the gain refers to the error reduction rate after pruning (cf. Table 2).

Table 8
Comparative results on the narrow-band set of the KING database for several a priori threshold setting methods with pruning abnormal data

Method	S_{04}	S_{05}	S_{06}	S_{07}	S_{08}	S_{09}	S_{10}	Averaging
CAVE-1	29.17	30.42	44.38	35.42	36.84	35.66	35.21	35.30
(gain)	(1.46)	(1.17)	(3.67)	(3.32)	(2.61)	(4.32)	(2.12)	(2.67)
CAVE-3	29.42	30.01	45.17	36.31	36.12	37.23	36.54	35.82
(gain)	(1.54)	(0.93)	(4.14)	(2.55)	(3.09)	(2.96)	(3.40)	(2.65)
Ours	28.83	29.14	43.51	35.12	35.97	35.45	35.02	34.71
(gain)	(0.81)	(0.78)	(3.81)	(3.27)	(2.91)	(2.46)	(1.27)	(2.20)

The testing segment length is 7.68 s, and here the gain refers to the error reduction rate after pruning (cf. Table 3).

parameter, $\kappa = 2.0$, in our pruning algorithm, the averaging percentage of pruned data is 21.5%, and the maximal and minimal percentages of pruned data are 24.9% and 19.4%, respectively, for 49 speakers on the wide-band set. Similarly, the averaging percentage of pruned data is 27.1%, and the maximal and minimal percentages of pruned data are 32.2% and 21.8%, respectively, for 51 speakers on the narrow-band set.

Given the EERs listed in Section 4.3.1, the simulation results indicate that our pruning algorithm leads to considerable improvements even though there are miscellaneous mismatches among different sessions in the narrow-band set. Thus, simulation results suggest that pruning abnormal data provide an alternative way to further improve the generalization capability of a statistics-based a priori speaker-dependent threshold setting method.

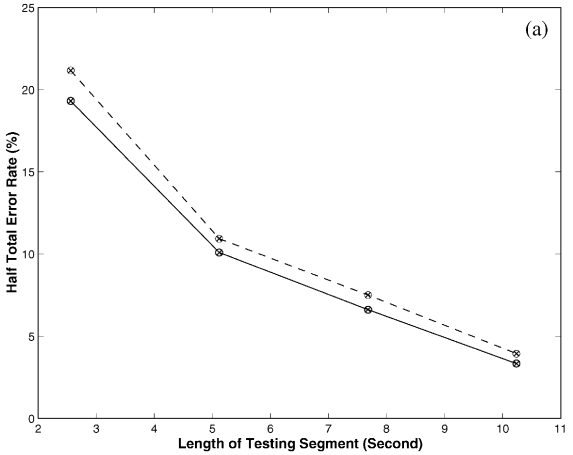
4.3.3. Results of on-line incremental threshold update

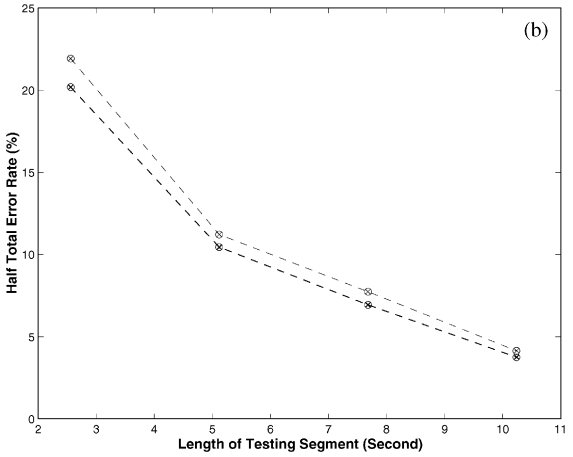
As mentioned above, threshold update would be an effective way to improve the performance as long as new data are available. For performance evaluation of threshold update, we have applied our on-line algorithm to three statistics-based a priori methods; i.e. CAVE-1, CAVE-3, and ours. Note that the performance of our on-line threshold update presented here is on the basis of thresholds resulting from pruning in the text-independent mode, while this procedure works without pruning in the text-dependent mode.

In previous studies, threshold update is regarded as a process of supervised learning [9]. In this circumstance, both a coming datum and its owner are known so that such an update can be an error-free process with external helps. Unfortunately, such sort of methods are impractical for an operational system. In reality, more often, a speaker verification system should work autonomously, thus threshold update without interference is demanded. In our simulations, we have conducted two types of experiments, i.e. supervised and autonomous updates, on the basis of our on-line incremental method. By supervised update, the statistics will be re-estimated based on the current decision-making on the datum and its known ownership. In contrast, the coming datum accepted by a system is always used to re-estimate the statistics in autonomous update no matter whether the system has made a wrong decision or not. Moreover, the coming datum accepted is used to re-estimate not only the claimed speaker's statistics for his/her own threshold update but also the pseudo-impostor's statistics for other speakers' threshold update on-line.

Recent studies indicated that the chronological order of client and impostor tests critically determines the performance of an autonomous on-line update [14]. Since we have no prior knowledge on such a chronological order,[3] we adopt the following multi-trial method in our simulations. First, we label all the speakers in our database. Then, a random number generator subject to a uniform distribution is employed to randomly select a speaker at a moment. As a result, an utterance belonging to this speaker will be randomly

[3] For an operational system, the statistics underlying such a chronological order could be estimated during real use.





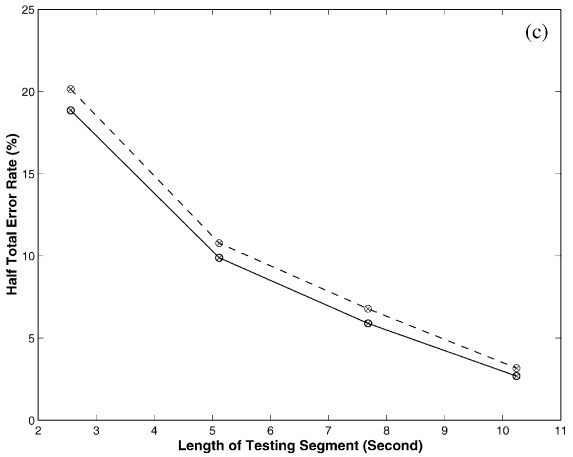
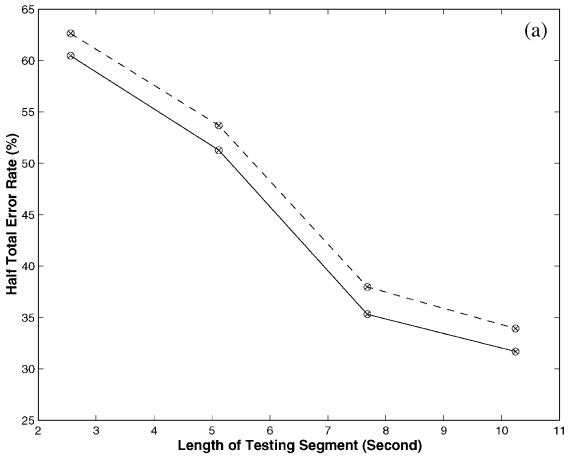
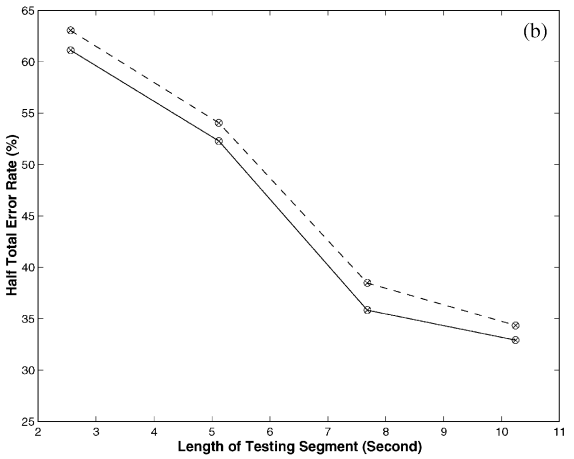


Fig. 4. Comparative results on the wide-band set of the KING database for two a priori threshold setting methods and ours at different testing segment lengths: without pruning (dashed line) vs. with pruning (solid line). (a) Results of CAVE-1; (b) results of CAVE-3; (c) results of ours.





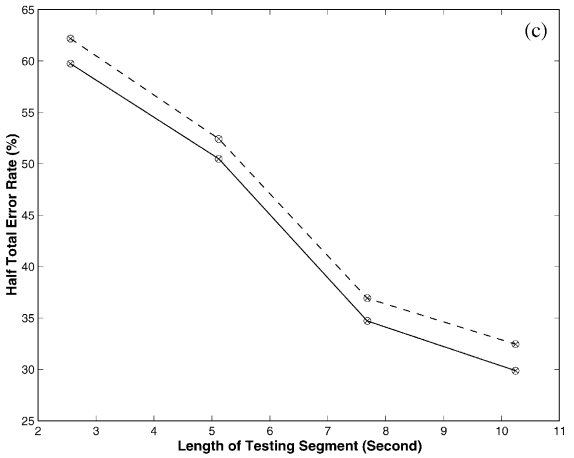


Fig. 5. Comparative results on the narrow-band set of the KING database for two a priori threshold setting methods and ours at different testing segment lengths: without pruning (dashed line) vs. with pruning (solid line). (a) Results of CAVE-1; (b) results of CAVE-3; (c) results of ours.

Table 9
Comparative results on the PKU-TD database for two a priori threshold setting methods and ours by the supervised update

Method	S_3	S_4	S_5	Averaging
CAVE-1	8.21	12.14	10.02	10.12
(gain)	(1.13)	(0.40)	(0.80)	(0.71)
CAVE-3	8.36	12.34	10.37	10.35
(gain)	(0.80)	(0.29)	(0.47)	(0.52)
Ours	7.86	11.18	9.82	9.62
(gain)	(0.61)	(0.52)	(0.51)	(0.55)

Here the gain refers to the error reduction rate after supervised update (cf. Table 1).

Table 12
Comparative results on the PKU-TD database for two a priori threshold setting methods and ours by the autonomous update

Method	S_3	S_4	S_5	Averaging
CAVE-1	8.43	12.35	10.29	10.36
(gain)	(0.69)	(0.19)	(0.53)	(0.47)
CAVE-3	8.51	12.49	10.41	10.47
(gain)	(0.65)	(0.14)	(0.43)	(0.40)
Ours	7.98	11.29	10.01	9.76
(gain)	(0.49)	(0.41)	(0.32)	(0.41)

Here the gain refers to the error reduction rate after the autonomous update (cf. Table 1).

chosen and then used as the current testing data. Once the current test is done, the chosen utterance is ruled out from the candidate data. In each trial, this procedure is repeated until all the testing data are used up. In the above way, our simulations contain ten trials for reliability and the averaging results are reported here.

Table 9 shows the text-dependent performance of supervised update on the PKU-TD database. From Tables 9, it is evident that the performance is significantly improved by our threshold update method for all three threshold setting methods tested. In particular, the gain for CAVE-1 is up to 0.71% in average. In Tables 10 and 11, we show the performance of the supervised update on the KING database. Apparently, our update method leads to significant improvements in the text-independent mode. After update, in particular, an averaging HTER based on our threshold setting method is even lower than the overall EER of all seven wide-band testing sets (7.95%), and the gain for CAVE-3 is up to 2.30% in average on the wide-band set. Similarly, the supervised update also leads to improvements on the narrow-band set; an averaging HTER produced by our threshold setting method is very close to the overall EER of all seven narrow-band testing sets (32.37%) and the gain for CAVE-1 is up to 2.24%. Thus simulation results indicate that our on-line threshold update method considerably improves the performance in a supervised learning way.

Furthermore, we show the performance of autonomous update on the PKU-TD database in Table 12 and those of

Table 10
Comparative results on the wide-band set of the KING database for two a priori threshold setting methods and ours by the supervised update

Method	S_{04}	S_{05}	S_{06}	S_{07}	S_{08}	S_{09}	S_{10}	Averaging
CAVE-1	9.12	7.14	7.93	9.56	9.02	9.24	8.04	8.58
(gain)	(1.21)	(0.93)	(1.98)	(1.93)	(1.29)	(1.90)	(1.38)	(1.52)
CAVE-3	9.38	7.27	7.88	9.46	9.15	9.42	8.13	8.67
(gain)	(1.47)	(1.40)	(2.33)	(2.20)	(1.10)	(1.86)	(2.02)	(2.30)
Ours	8.34	6.15	6.76	9.22	8.08	9.04	7.29	7.84
(gain)	(1.97)	(1.62)	(2.83)	(1.87)	(1.67)	(1.77)	(1.94)	(1.95)

The testing segment length is 5.12 s, and here the gain refers to the error reduction rate after the supervised update (cf. Table 7).

Table 11
Comparative results on the narrow-band set of the KING database for two a priori threshold setting methods and ours by the supervised update

Method	S_{04}	S_{05}	S_{06}	S_{07}	S_{08}	S_{09}	S_{10}	Averaging
CAVE-1	27.43	27.92	41.25	33.91	33.43	34.12	33.38	33.06
(gain)	(1.74)	(2.50)	(3.13)	(1.51)	(3.41)	(1.54)	(1.83)	(2.24)
CAVE-3	27.89	28.97	42.92	34.42	34.21	34.37	34.18	33.85
(gain)	(1.53)	(1.04)	(2.25)	(1.89)	(1.91)	(2.86)	(2.36)	(1.97)
Ours	26.92	27.29	40.87	33.38	32.86	33.57	32.67	32.51
(gain)	(1.91)	(1.85)	(2.64)	(1.74)	(3.11)	(1.88)	(2.35)	(2.20)

The testing segment length is 7.68 s, and here the gain refers to the error reduction rate after the supervised update (cf. Table 8).

Table 13
Comparative results on the set of wide-band set of the KING database for two a priori threshold setting methods and ours by the autonomous update

Method	S_{04}	S_{05}	S_{06}	S_{07}	S_{08}	S_{09}	S_{10}	Averaging
CAVE-1	9.34	7.55	8.25	9.97	9.33	9.01	8.41	8.84
(gain)	(0.99)	(0.52)	(1.66)	(1.52)	(0.98)	(1.67)	(1.01)	(1.26)
CAVE-3	9.88	7.79	8.90	10.27	9.47	10.23	8.94	9.35
(gain)	(0.97)	(0.88)	(1.31)	(1.39)	(0.78)	(1.05)	(1.11)	(1.10)
Ours	9.04	7.11	7.46	10.08	9.35	9.43	7.18	8.52
(gain)	(1.27)	(0.66)	(2.03)	(1.11)	(0.40)	(1.38)	(2.03)	(1.27)

The testing segment length is 5.12 s, and here the gain refers to the error reduction rate after the autonomous update (cf. Table 7).

Table 14
Comparative results on the narrow-band set of the KING database for two a priori threshold setting methods and ours by the autonomous update

Method	S_{04}	S_{05}	S_{06}	S_{07}	S_{08}	S_{09}	S_{10}	Averaging
CAVE-1	28.24	28.67	42.42	34.61	35.41	35.04	34.11	34.07
(gain)	(0.93)	(1.75)	(1.96)	(0.81)	(1.43)	(0.64)	(1.10)	(1.23)
CAVE-3	28.64	30.01	43.74	35.47	35.35	36.04	35.86	35.82
(gain)	(0.78)	(0.79)	(1.43)	(0.84)	(0.77)	(1.19)	(0.68)	(0.92)
Ours	27.86	28.42	42.11	34.80	33.94	34.55	34.07	33.68
(gain)	(0.97)	(0.62)	(1.40)	(0.32)	(2.03)	(0.90)	(0.95)	(1.03)

The testing segment length is 7.68 s, and here the gain refers to the error reduction rate after the autonomous update (cf. Table 8).

the KING database in Tables 13 and 14. Although the performance of autonomous update is degraded in comparison with that of supervised update, our threshold update method yields fair improvements regardless of operating modes. From Table 13, it is observed that the gain for any of three threshold setting methods is not lower than 1.10% in average on the wide-band set after such an update and the averaging HTER by our threshold setting method is quite close to the overall EER of all seven wide-band testing sets. Similarly, it is evident from Table 14 that the autonomous update leads to improvements on the narrow-band set for all the three threshold setting methods. For CAVE-1, in particular, an error reduction gain of 1.23% is achieved on the narrow-band set. Thus, simulation results in autonomous update demonstrate that our threshold update method provides a promising way for real use.

5. Discussion

As a novel data selection procedure, pruning abnormal data provides an alternative way to improve generalization of a statistics-based a priori speaker-dependent threshold setting method. However, our empirical studies indicate that it does not work in the text-dependent circumstance (cf. Table 6). To a great extent, a statistical model tends to capture verbal information of a specific text itself while we try to employ it for modeling a speaker's characteristics. Unlike the text-independent mode, no mismatch occurs in verbal information for the text-dependent mode. In other words, the same text content used for speaker verification leads to similar utterances (speech waves) even for different speakers, which causes the deviation of intra-speaker's and inter-speaker's scores to be small. Thus, the use of our pruning manipulation may introduce a larger bias to the original data in contrast to that by the mismatch environments. For use in the text-dependent mode, our pruning method demands robust speaker features that can differentiate speakers by scatting those scores corresponding to utterances of different speakers in decision space. On the other hand, the parameter κ was fixed as the termination condition in our simulations. Although doing so yields improvements in generalization, we point out that the parameter κ is adjustable for different speakers. Thus, how to give a proper termination condition is still an open problem to be studied in the future. Our empirical studies indicate that the termination condition highly depends upon the deviation of intra-speaker's scores. Furthermore, we find that the number of abnormal data that need eliminating seems to be in proportion to the value of variance of all the training data prior to pruning; i.e. the more numbers of abnormal data should be eliminated, the larger the variance is, and vice versa. Based on the heuristics, we expect that an automatic termination condition can be achieved by exploring the relationship between

the number of abnormal data pruned and the deviation of intra-speaker's scores.

Now we attempt to relate our methods to other sophisticated threshold setting and update methods. First, as a statistics-based a priori method, our threshold setting method may be viewed as a direct extension to the Furui's and CAVE-1 methods. Furui's threshold setting is performed by the use of only the first- and the second-order statistics of inter-speaker's scores, while the first-order statistics are merely employed for threshold setting in CAVE-1. In contrast, ours employs both the first-order statistics of intra-speaker's scores and the first- and the second-order statistics of inter-speaker's. In terms of the performance, ours can be viewed as an improved version of Furui's and CAVE-1 by introducing more reliable statistics. Next, a speaker-independent shift is applied to the estimate of first-order intra-speaker's statistics in CAVE-2 against mismatches. Although such a correction is optimized on a registration population, the same correction is used for different speakers. As well known, changes of different speakers' voice may be involved in different mismatch environments. This fact incorporated by our empirical studies shows that a fixed correction could not compensate mismatch sufficiently for all the speakers. Unlike CAVE-2, our pruning method provides a way to correct bias of estimates in a speaker-dependent way. Note that the number of abnormal data eliminated for different speakers is distinct even though the parameter κ is fixed in our algorithm. Although CAVE-2 is proposed only for the Gaussian model of the utterance log-likelihood ratio distribution, its correction method can be viewed as a special case of ours in the sense of pruning abnormal data.

For most of threshold update methods in speaker verification, the new data are pooled to update the parameters off-line in a speaker model. Then a threshold is indirectly updated on the basis of the re-trained speaker model. Recently, some researchers [9,14] proposed a method that can update the parameters of a hidden Markov model (HMM) on-line, and the threshold update is performed based on the new HMM accordingly. In contrast, our method gives an on-line incremental threshold update in output space of speaker models, and thus, allows a speaker verification system to perform threshold update regardless of types of speaker models. This salient feature distinguishes between other existing threshold update methods and ours.

6. Concluding remarks

In this paper, we have presented our methods towards better making a decision in speaker verification. The proposed methods include the use of more reliable statistics for threshold setting, elimination of abnormal data for better estimation of underlying statistics, and on-line incremental threshold update in decision space. Comparative results in different operating modes show that our methods yield satisfactory performance even for data collected on mismatch conditions and, in particular, the joint use of our methods leads to considerable improvements in comparison with recent threshold setting methods. In addition, our pruning and on-line threshold update methods may be directly applied to some statistics-based a priori threshold setting methods as done in this paper, which also results in the improved performance. All of our methods tend to give insight into creating a real high-performance speaker verification system, although the performance of our methods by incorporating other effective technologies, e.g. score normalization by a reference model, still needs to be investigated.

For better making a decision in speaker verification, there are several open issues to be addressed in our ongoing work. First, most of a priori methods set a threshold by building a linear mapping between the estimate of some certain statistics of scores and a threshold in a heuristic way. In terms of their performance, we highly believe that there might be alternative mapping forms for better threshold setting. Therefore, the development of such a non-linear mapping will be one of our ongoing research topics. Second, there should be a underlying relationship between an estimate of statistics and a proper threshold that can maximize the generalization in terms of a given data set. To our knowledge, such a relationship still keeps unknown and, therefore, the exploration of such a underlying relations in a systematic way is another topic in our ongoing studies. Next, for most of speaker verification systems, the component modules, i.e. feature extraction, creation of speaker models, and threshold setting, perform independently in the learning phase. However, the system performance highly depends upon all the components. Therefore, we suggest exploring a global optimization strategy such that threshold setting is performed along with the creation of other component modules. Finally, combination of different threshold setting methods provides an effective way to utilize complementary information, and thus, becomes one of our ongoing research topics towards better making a decision in speaker verification.

Acknowledgements

The author would like to thank J.H. Liu for constructive discussions and his help in simulations as well as an anonymous reviewer whose comments improved the presentation in the paper. The partial work reported in this paper was done while the author worked at Peking University and visited The Hong Kong Polytechnic University.

Appendix A

In this appendix, we adopt an incremental way to estimate the first- and the second-order moments for a given data set. Suppose that the current data set contains n ($n > 1$)

samples. Thus, the instantaneous mean $\tilde{\mu}_n$ and variance $\tilde{\sigma}_n$ are estimated as

$$\tilde{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i, \qquad \tilde{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \tilde{\mu}_n)^2. \tag{A.1}$$

When a new sample, $\tilde{x}_{n+1}$, is appended, we can update the instantaneous estimate of mean and variance in an incremental way without the direct use of historical data. For re-estimating the mean and variance, the application of Eq. (A.1) leads to

$$\begin{aligned}\tilde{\mu}_{n+1} &= \frac{1}{n+1}\sum_{i=1}^{n+1}\tilde{x}_i \\ &= \frac{1}{n+1}\left(\sum_{i=1}^{n}\tilde{x}_i + \tilde{x}_{n+1}\right) \\ &= \frac{n}{n+1}\,\frac{\sum_{i=1}^{n}\tilde{x}_i}{n} + \frac{\tilde{x}_{n+1}}{n+1} \\ &= \frac{n\tilde{\mu}_n + \tilde{x}_{n+1}}{n+1}\end{aligned} \tag{A.2}$$

and

$$\begin{aligned}\tilde{\sigma}_{n+1} &= \frac{1}{n+1}\sum_{i=1}^{n+1}(\tilde{x}_i - \tilde{\mu}_{n+1})^2 \\ &= \frac{1}{n+1}\left\{\sum_{i=1}^{n}(\tilde{x}_i - \tilde{\mu}_{n+1})^2 + (\tilde{x}_{n+1} - \tilde{\mu}_{n+1})^2\right\} \\ &= \frac{1}{n+1}\left\{\sum_{i=1}^{n}[(\tilde{x}_i - \tilde{\mu}_n) + (\tilde{\mu}_n - \tilde{\mu}_{n+1})]^2 \right. \\ &\quad \left. + (\tilde{x}_{n+1} - \tilde{\mu}_{n+1})^2\right\} \\ &= \frac{1}{n+1}\left\{\sum_{i=1}^{n}[(\tilde{x}_i - \tilde{\mu}_n)^2 + 2(\tilde{x}_i - \tilde{\mu}_n)(\tilde{\mu}_n - \tilde{\mu}_{n+1}) \right. \\ &\quad \left. + (\tilde{\mu}_n - \tilde{\mu}_{n+1})^2] + (\tilde{x}_{n+1} - \tilde{\mu}_{n+1})^2\right\} \\ &= \frac{1}{n+1}\left\{\sum_{i=1}^{n}(\tilde{x}_i - \tilde{\mu}_n)^2 + \sum_{i=1}^{n}(\tilde{\mu}_n - \tilde{\mu}_{n+1})^2 \right. \\ &\quad \left. + (\tilde{x}_{n+1} - \tilde{\mu}_{n+1})^2\right\} \\ &= \frac{1}{n+1}\left\{n\tilde{\sigma}^2 + \frac{n}{(n+1)^2}(\tilde{x}_{n+1} - \tilde{\mu}_n)^2 \right. \\ &\quad \left. + \frac{n^2}{(n+1)^2}(\tilde{x}_{n+1} - \tilde{\mu}_n)^2\right\} \\ &= \frac{n(n+1)\tilde{\sigma}^2 + n(\tilde{x}_{n+1} - \tilde{\mu}_n)^2}{(n+1)^2}.\end{aligned} \tag{A.3}$$

For the last four steps in Eq. (A.3), we apply the following facts achieved from Eqs. (A.1) and (A.2):

$$\sum_{i=1}^{n}(\tilde{x}_i - \tilde{\mu}_n) = 0,$$

$$\tilde{\mu}_{n+1} - \tilde{\mu}_n = \frac{\tilde{x}_{n+1} - \tilde{\mu}_n}{n+1}$$

and

$$\tilde{x}_{n+1} - \tilde{\mu}_{n+1} = \frac{n}{n+1}(\tilde{x}_{n+1} - \tilde{\mu}_n).$$

Without loss of generality we can denote $\tilde{x}_{n+1}$ as the most abnormal datum in terms of our pruning method. Once $\tilde{x}_{n+1}$ is given, Eq. (15) can be immediately achieved from Eq. (A.2) to incrementally re-estimate the new mean after pruning.

References

[1] G.R. Doddington, M.A. Przybocki, A.F. Martin, D.A. Reynolds, The NIST speaker recognition evaluation—overview, methodology, systems, results, perspective, Speech Commun. 31 (2000) 225–254.

[2] F. Bimbot, M. Blomberg, L. Boves, D. Genoud, H.P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, J.B. Pierrot, An overview of the CAVE project research activities in speaker verification, Speech Commun. 31 (2000) 1437–1462.

[3] H. Meng et al., ISIS: a multilingual spoken dialog system developed with CORBA and KQML agents, Proceedings of the ICSLP, 2000, pp. II150–II153.

[4] S. Furui, Recent advances in speaker recognition, Pattern Recognition Lett. 18 (1997) 859–872.

[5] S. Furui, Cepstral analysis technique for automatic verification, IEEE Trans. Acoustics Speech Signal Process. 29 (1981) 254–272.

[6] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, B.H. Juang, A vector quantization approach to speaker recognition, AT & T Tech. J. 66 (1987) 14–26.

[7] A.E. Rosenberg, C.H. Lee, F.K. Soong, Sub-word unit talker verification using hidden Markov models, Proceedings of the ICASSP, 1990, pp. 269–272.

[8] N. Tishby, On the application of mixture AR hidden Markov models to text independent speaker recognition, IEEE Trans. Acoustics Speech Signal Process. 39 (1991) 563–570.

[9] T. Matsui, T. Nishitani, S. Furui, Robust methods of updating model and a priori threshold in speaker verification, Proceedings of the ICASSP, 1996, pp. 97–100.

[10] F. Bimbot, H.P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, J.B. Pierrot, Speaker verification in the telephone network: research activities in the CAVE project, Proceedings of the Eurospeech, 1997, pp. 971–974.

[11] J. Lindberg, J. Koolwaaij, H.P. Hutter, D. Genoud, J.B. Pierrot, M. Blomberg, F. Bimbot, Techniques for a priori decision threshold estimation in speaker verification, Proceedings of the RLA2C Workshop, 1998, pp. 89–92.

[12] J.B. Pierrot, J. Lindberg, J. Koolwaaij, H.P. Hutter, D. Genoud, M. Bolmberg, F. Bimbot, A comparison of a priori threshold setting procedures for speaker verification in the CAVE project, Proceedings of the ICASSP, 1998, pp. 331–334.
[13] A.E. Rosenberg, Automatic speaker verification: review. Proc. IEEE 64 (1976) 475–487.
[14] C. Fredouille, J. Mariethoz, C. Jaboulet, J. Hennebert, J.F. Bonastre, C. Mokbel, F. Bimbot, Behavior of a Bayesian adaptation method for incremental enrollment in speaker verification, Proceedings of the ICASSP, 2000, pp. 1197–1200.
[15] J.M. Naik, Speaker verification—a tutorial, IEEE Commun. Mag. 28 (1990) 42–48.
[16] H.L. Higgins, L.G. Bahler, J.E. Porter, Speaker verification using randomized phrase prompting, Digital Signal Process. 1 (1991) 89–106.
[17] M.J. Carey, E.S. Parris, J.S. Bridle, A speaker verification system using alpha-nets, Proceedings of the ICASSP, 1991, pp. 396–399.
[18] A.E. Rosenberg, J. Delong, C.H. Lee, B.H. Juang, F.K. Soong, The use of cohort normalized scores for speaker recognition, Proceedings of the ICSLP, 1992, pp. 599–602.
[19] T. Matsui, S. Furui, Likelihood normalization for speaker verification using a phoneme-independent and speaker-independent model, Speech Commun. 17 (1995) 109–116.
[20] C.S. Liu, H.C. Wang, C.H. Lee, Speaker verification using normalized log-likelihood scores, IEEE Trans. Speech Audio Process. 4 (1996) 56–60.
[21] M.A. Lund, C.C. Lee, A robust sequential test for text-independent speaker verification, J. Acoust. Soc. Am. 99 (1996) 609–621.
[22] A.E. Rosenberg, P.S. Parthasarathy, Speaker background models for connected digit password speaker verification, Proceedings of the ICASSP, 1996, pp. 81–84.
[23] R.A. Finan, A.T. Sapeluk, R.I. Damper, Impostor cohort selection for score normalization in speaker verification, Pattern Recognition Lett. 18 (1997) 881–888.
[24] J.M. Colombi, Cohort selection and word grammar effects for speaker verification, Proceedings of the ICASSP, 1998, pp. 85–88.
[25] R. Auchenthaler, M. Carey, H. Lloyd-Thomas, Score normalization for text-independent speaker verification systems, Digital Signal Process. 10 (2000) 42–54.
[26] G. Gravier, J. Kharroubi, G. Chollet, On the use of prior knowledge in normalization schemes for speaker verification, Digital Signal Process. 10 (2000) 213–225.
[27] Y. Zhang, D. Zhang, X. Zhu, A novel text-independent speaker verification method based on the global speaker model, IEEE Trans. Systems Man Cybernet. A 30 (2000) 598–602.
[28] D.A. Reynolds, A Gaussian mixture modeling approach to text-independent speaker identification, Ph.D. Thesis, Georgia Institute of Technology, 1992.
[29] X.K. Qing, K. Chen, On use of Gaussian mixture model for multilingual speaker verification: an empirical study, Proceedings of the ISCSLP, 2000, pp. 263–266.
[30] ITT/LDC, The KING Speaker Recognition Corpus, Produced and Distributed by the LDC, CD-ROM, 1994.
[31] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[32] L.L. Scharf, Statistical Signal Processing Detection, Estimation, and Time Analysis, Addison-Wesley, Reading, MA, 1991.
[33] M.H. DeGroot, Optimal Statistical Decisions, McGraw-Hill, New York, 1970.
[34] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B 39 (1977) 1–38.

About the Author—KE CHEN received his B.S. and M.S. degrees from Nanjing University in 1984 and 1987, respectively, and his Ph.D. from Harbin Institute of Technology in 1990, all in Computer Science.

He has joined the Faculty of The University of Birmingham since 2001. From 1990 to 1992, he was a postdoctoral researcher at Tsinghua University. During 1992–1993, he was a postdoctoral fellow of Japan Society for Promotion of Sciences and worked at Kyushu Institute of Technology. From 1996 to 1998, he was a visiting professor at The Ohio State University. During 2000–2001, he held a visiting researcher position in Microsoft Research Asia and a visiting professorship at Hong Kong Polytechnic University. Since 1994, he has been in the Faculty of Peking University, where he is an adjunct professor. He has published over 90 technical articles in refereed journals, invited book chapters, and international conferences. His current research interest includes statistical pattern recognition, machine learning with an emphasis on nature inspired learning, and their applications to machine perception. He was a recipient of the TOP Award for Progress of Science and Technology from the National Education Ministry in China, a recipient of NSFC Distinguished Young Principal Investigator Award and a recipient of "Trans-Century Talented Professional" Award from the China National Education Ministry. He is now a senior member of IEEE, a member of IEEE Computer Society, a member of IEEE Systems, Man, and Cybernetics Society as well as a member of International Neural Networks Society.