

NISTIR 7810

**Data Dependency
on Measurement Uncertainties
in Speaker Recognition Evaluation**

*Jin Chu Wu
Alvin F. Martin
Craig S. Greenberg
Raghu N. Kacker*

NISTIR 7810

Data Dependency on Measurement Uncertainties in Speaker Recognition Evaluation

Jin Chu Wu
Alvin F. Martin
Craig S. Greenberg
Raghu N. Kacker

October 2011



U.S. Department of Commerce
John E. Bryson, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Director

Data Dependency on Measurement Uncertainties in Speaker Recognition Evaluation

Jin Chu Wu^a, Alvin F. Martin^a, Craig S. Greenberg^a and Raghu N. Kacker^b
^aInformation Access Division, ^bApplied and Computational Mathematics Division,
Information Technology Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

The National Institute of Standards and Technology (NIST) conducts an ongoing series of Speaker Recognition Evaluations (SRE). Speaker detection performance is measured using a detection cost function defined as a weighted sum of the probabilities of type I error and of type II error. The sampling variability can result in measurement uncertainties. Thus, the uncertainties of the detection cost functions must be taken into consideration in SRE. In our prior study, the data independence was assumed while applying the nonparametric two-sample bootstrap methods based on our extensive bootstrap variability studies on large datasets to compute the standard errors (SE) of detection cost functions. In this article, the data dependency caused by multiple usages of the same subjects is taken into account. Hence, the data are grouped into target sets and non-target sets, and each set contains multiple scores. One-layer and two-layer bootstrap methods are proposed based on whether the two-sample bootstrap resampling takes place only on target sets and non-target sets, respectively, or subsequently on target scores and non-target scores within the sets. The SEs of the detection cost function using these two methods along with those with the assumption of data independency are compared. It is found that the data dependency increases both estimated SEs and the variations of SEs. Thus, in order to obtain more accurate measures in SRE, the data should be sampled randomly. Based on our research, some suggestions regarding the test design are provided.

Keywords: Speaker recognition evaluation; Biometrics; Bootstrap; Data dependency; Resampling; Uncertainty; Standard error; Confidence interval.

1 Introduction

The National Institute of Standards and Technology (NIST) conducts an ongoing series of Speaker Recognition Evaluations (SRE) [1]. The NIST SREs have made important contributions to the direction of research efforts and the calibration of technical capabilities of the research community working on the general problem of text independent speaker recognition.

Each test in the NIST SRE has consisted of a sequence of trials, where each trial consists of a model speaker, based on the training data provided, and a test speech segment. For each trial, the speaker recognition system must decide whether the speech of the model speaker occurred in the test speech segment and generate a similarity score. Target (non-target) trials are those where the test speech segment contains (does not contain) speech of the model speaker defined in the training data. A higher similarity score indicates greater confidence that the speech of the model speaker occurs in the test speech segment. In the 2008 NIST SRE, each test generally included about 20,000 target scores and 80,000 non-target scores [1, 2].

In the NIST SRE, the speaker detection performance is measured using a detection cost function, which is defined as a weighted sum of the probabilities of type I error (miss) and of type II error (false alarm) [1]. As is well-known, the sampling variability results in uncertainties of any measures [3]. Even when samples are collected under the same circumstances, the measures in evaluation may fluctuate. Hence, while evaluating and comparing the performances of speaker recognition systems, the uncertainties of measures must be taken into account. A key issue is how to calculate the uncertainties of detection cost functions in terms of standard errors (SE).

The probabilities of type I error and type II error represent a tradeoff. In other words, these two probabilities are negatively correlated. It is hard to compute analytically the covariance term (i.e., the cross term) of such two correlated probabilities, the linear combination of which forms the detection cost function in SRE. As a result, it is difficult to calculate the variance of such a detection cost function analytically. In our prior study of the SRE, the uncertainties of detection cost functions were computed using nonparametric two-sample bootstrap methods based on our extensive bootstrap variability studies on large datasets, using the assumption of data independence [3-9].

The two samples involved are the set of target scores and the set of non-target scores, and they constitute two distributions [3, 6]. A detection cost function is characterized by the relationship between these two distributions [10, 11]. These two distribution functions characterize the speaker recognition system that generates them. In other words, the performance of a system is determined by both target matching and non-target matching effectiveness. All statistics of interest in SRE are influenced by the combined impact of these two samples.

Furthermore, it is known from previous biometric studies that these two distributions 1) usually do not have well defined parametric forms; 2) may be considerably different even for the same system; and 3) may vary substantially from system to system, which differentiates systems in terms of matching accuracy [10]. Similar variations of distributions were also observed in the speaker recognition data. An example can be found in Ref. [9]. This suggests that the nonparametric

statistical analysis is appropriate for evaluating speaker recognition data, and the empirical distribution is assumed for each of the observed scores.

The bootstrap method assumes that an independent and identically distributed (i.i.d.) random sample of size n is drawn from a population with its own probability distribution. With the i.i.d. assumption, the bootstrap units are scores in the sample. However, the NIST speaker recognition data contains dependency. In this article, this data dependency is taken into account.

The data dependency is caused by multiple usages of the same subject in order to create more target and non-target scores. The data dependency is complicated, due in part to the way the data was collected. There are several ways to interpret the dependencies of the data. How the sample is grouped into sets can impact the bootstrap results.

In this article, data dependency is determined based purely upon whether the training speaker identification (id) number is used multiple times. Those target scores and non-target scores generated using the same training speaker id number are grouped into a target set and a non-target set, respectively. This recognizes the data dependency while the bootstrap resampling takes place. Thus, the speaker recognition data structure has two layers: The first layer consists of the target sets and non-target sets, and the second layer consists of the target scores and non-target scores within the sets.

It should be noted that there are other forms of data dependency not taken into account by this procedure. Different non-target trials involving a specific training speaker may involve the same test segment speaker. Further, different trials involve the same or different recording channels (telephone or multiple types of room microphone) of the speakers involved. Also varying are the speech styles (including conversational telephone or face-to-face interview) of the speakers, and the extent of high or low or normal vocal effort encouraged. In some cases different trials involve the same speech of the training speaker, or of the test segment speaker (or both), but recorded over different microphones. Further work will be necessary to address these types of data dependency.

In addition to a bootstrap method in which all data are assumed to be i.i.d., based on the data structure stated above, one-layer and two-layer bootstrap methods are proposed based on whether the nonparametric two-sample bootstrap resampling takes place only on the first layer of the data, i.e., the target sets and non-target sets, or subsequently on the second layer, i.e., the target scores and non-target scores within the sets. Resampling on the first layer indicates that the bootstrap units are sets, while resampling on the second layer means that the bootstrap units are the scores within a set, where the similarity scores are assumed to be conditionally independent.

Different target (non-target) sets may have different numbers of target (non-target) scores. This would cause each target (non-target) score to not have the same probability of being selected while the bootstrap resampling is carried out using the above three bootstrap methods. In order to avoid this, the speaker recognition data was adjusted so that all target sets would contain the same number of scores and likewise for the non-target sets. In the meantime, the total numbers of target scores and non-target scores obtained should be kept as large as possible.

After data adjustment, the probability for each target (non-target) score being selected becomes the same for the above three bootstrap methods. Thus, the SEs of the detection cost functions computed using the three bootstrap methods may be compared. Further, the bootstrap method is a stochastic process. The results will vary for different runs, and thus constitute a probability distribution. Some results may be more probable and others less probable. Hence, the comparisons of SEs of the detection cost functions may involve the comparisons of the distributions of SEs. It is found that taking account of the data dependency increases the estimated SEs and the variations of SEs. Thus, in order to obtain more accurate measures in SRE, the data should be sampled randomly.

The bootstrap method on datasets with dependencies was initially studied in the references [5, 12], and applied to other cases later [13, 14]. In this article, the nonparametric two-sample bootstrap, rather than the one-sample bootstrap is employed. The two-sample bootstrap can be used to compute uncertainties of much more complicated measures, such as the detection cost function that is defined as a weighted sum of the probabilities of type I error and of type II error. Further, the probability issues of similarity scores being selected are investigated while dealing with different resampling methods in bootstrap.

All similarity scores of the speaker recognition systems are real numbers. While analyzing the data, all real numbers were converted into integers. Different systems employ different numbers of digits in the integer part. Hence, in order to obtain accurate results, five decimal places (i.e., multiplying by 10^5) or up to seven decimal places (i.e., multiplying by 10^7) were preserved. Notice that if the largest integer score is excessively large, the computation can take an excessive amount of time. This is because it has to cover the range of the highest score down to the threshold provided by a vendor each time while computing thousands of bootstrap replications of the detection cost function. The probability distribution functions of similarity scores are all discrete [3, 10]. It is characteristic for the speaker data that only a few instances occur of different similarity scores taking exactly same matching values [3, 6].

Issues related to notations of sets and scores, the three resampling methods and the related probabilities, and the adjustment of the speaker recognition datasets are presented in Section 2. The formulas for computing the probabilities of type I error and of type II error, and the detection cost function in SRE, are shown in Section 3. The nonparametric two-sample bootstrap algorithms with i.i.d. assumption, with one-layer resampling, and with two-layer resampling are presented in Section 4. A method of generating distributions of the SEs of the detection cost function is provided in Section 5. The resulting uncertainties of the detection cost functions employing the three bootstrap methods for four speaker recognition systems¹, and comparisons among them are shown in Section 6. The conclusions and discussion can be found in Section 7. Finally, the proof of the probability for one-layer resampling is presented in the Appendix.

2 Adjust speaker recognition datasets

¹ Specific hardware and software products identified in this paper were used in order to adequately support the development of technology to conduct the performance evaluations described in this document. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

2.1 The distributions of numbers of target and non-target scores within a set

The speaker recognition data dependency is complicated. There are several ways to group data into sets according to different interpretations of data dependency. In this article, the speaker recognition data are grouped into sets purely based on whether or not the training speaker id number is multiply employed in order to generate more target scores and non-target scores. The target scores generated by the same id number of training speaker and test speaker are grouped into a target set, whereas the non-target scores created by the same id number of training speaker but different id numbers of test speakers are grouped into a non-target set, regardless of whether dependency exists implied by the training and test speech ids.

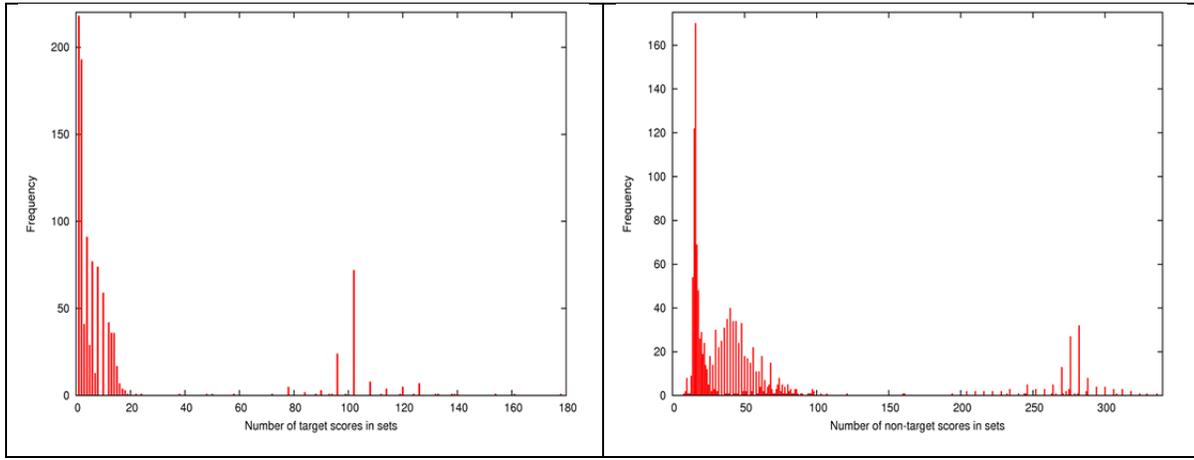


Figure 1 The histograms of the numbers of target scores in sets (Left) and the numbers of non-target scores in sets (Right) of a speaker recognition system labeled as EL.

Hence, 20,449 raw target scores were grouped into 1,093 target sets and 78,327 raw non-target scores were grouped into 1,336 non-target sets. Different sets contain quite different numbers of scores, and the numbers of sets that contain the same number of scores are also quite different. For instance, 218 target sets contain one score each, 193 target sets have two scores each, and so on; only one non-target set contains eight scores, two non-target sets have nine scores each, and so forth. The histograms of the numbers of target scores in sets and the numbers of non-target scores in sets of a speaker recognition system labeled as EL are shown in Figure 1. Such wide variations of numbers of scores in sets can have impact on the probability for a score to be selected.

2.2 The notations of sets and scores

target S_T	sets	S_{T1}	S_{T2}	$S_{T m_T}$
	scores	$\alpha_{T11}, \alpha_{T12}, \dots,$ $\alpha_{T1} \mu_{T1}$	$\alpha_{T21}, \alpha_{T22}, \dots,$ $\alpha_{T2} \mu_{T2}$	$\alpha_{T m_T 1}, \alpha_{T m_T 2}, \dots,$ $\alpha_{T m_T} \mu_{T m_T}$

Table 1 The target sets, the number of which is m_T , and the target scores contained in each set.

non-target \mathcal{S}_N	sets	\mathcal{S}_{N1}	\mathcal{S}_{N2}	\mathcal{S}_{Nm_N}
	scores	$\alpha_{N11}, \alpha_{N12}, \dots,$ $\alpha_{N1} \mu_{N1}$	$\alpha_{N21}, \alpha_{N22}, \dots,$ $\alpha_{N2} \mu_{N2}$	$\alpha_{Nm_N1}, \alpha_{Nm_N2},$ $\dots, \alpha_{Nm_N} \mu_{Nm_N}$

Table 2 The non-target sets, the number of which is m_N , and the non-target scores contained in each set.

Assume that the target scores and non-target scores are grouped into target sets and non-target sets, respectively. Suppose that the number of the target sets is m_T , and the number of the non-target sets is m_N . Thus, the set \mathcal{S}_T of all target sets and the set \mathcal{S}_N of all non-target sets are expressed, respectively, as follows,

$$\mathcal{S}_i = \{ \mathcal{S}_{ij} \mid j = 1, \dots, m_i \}, i \in \{T, N\}, \quad (1)$$

where \mathcal{S}_{Tj} are target sets and \mathcal{S}_{Nj} are non-target sets.

In terms of its scores, each set can be expressed as

$$\mathcal{S}_{ij} = \{ \alpha_{ijk} \mid k = 1, \dots, \mu_{ij} \}, j = 1, \dots, m_i \text{ and } i \in \{T, N\}, \quad (2)$$

where α_{Tjk} are target scores, α_{Njk} are non-target scores, and μ_{ij} stands for the number of scores in the corresponding set.

Hence, the set of all target scores and the set of all non-target scores can be denoted, respectively, as

$$\mathbf{T} = \{ \alpha_{Tjk} \mid k = 1, \dots, \mu_{Tj} \text{ and } j = 1, \dots, m_T \}, \quad (3)$$

and

$$\mathbf{N} = \{ \alpha_{Njk} \mid k = 1, \dots, \mu_{Nj} \text{ and } j = 1, \dots, m_N \}. \quad (4)$$

The sets \mathcal{S}_{ij} , \mathbf{T} , and \mathbf{N} are all in the sense of multiset, in which members are allowed to appear more than once. Indeed score can occur multiple times within a set. Finally, the total number of target scores N_T and the total number of non-target scores N_N are, respectively,

$$N_i = \sum_{j=1}^{m_i} \mu_{ij}, \text{ where } i \in \{T, N\}. \quad (5)$$

The target and non-target sets and scores contained in each set are explicitly listed in Table 1 and Table 2, respectively. There are m_T target sets and m_N non-target sets. The target sets $\mathcal{S}_{T1}, \mathcal{S}_{T2}, \dots, \mathcal{S}_{Tm_T}$ contain $\mu_{T1}, \mu_{T2}, \dots, \mu_{Tm_T}$ target scores, respectively; and the non-target sets $\mathcal{S}_{N1}, \mathcal{S}_{N2}, \dots, \mathcal{S}_{Nm_N}$ have $\mu_{N1}, \mu_{N2}, \dots, \mu_{Nm_N}$ non-target scores, respectively.

2.3 The three resampling methods and the related probabilities

In this section, all similarity scores are treated as different objects in the sense that they were generated by different trials in the test, even though some of them have the same value.

2.3.1 Resampling assuming the data is i.i.d.

The first method is that the resampling is taken place on all similarity scores with the i.i.d. assumption for the speaker recognition datasets. In other words, the resampling units are all

similarity scores. With this assumption, the probability for a score being selected is $1 / N_T$ equally for each target score and $1 / N_N$ equally for each non-target score based on the empirical distribution for each of the observed similarity scores.

2.3.2 One-layer resampling

The second method is one-layer resampling which takes place randomly with replacement (WR) only on the first layer of the data, i.e., target sets and non-target sets, respectively. In other words, the resampling units are all score sets. It follows from the Law of Large Numbers that the probability for a score $\alpha_{i j k}$ being selected is

$$P_{1\text{-layer}}(\alpha_{i j k}) = \frac{1}{N_i}, \quad \text{where } k = 1, \dots, \mu_{i j}, j = 1, \dots, m_i \text{ and } i \in \{T, N\}. \quad (6)$$

The proof of Eq. (6) is presented in the Appendix. Hence, the probabilities for each target score and each non-target score being selected are $1 / N_T$ and $1 / N_N$, respectively. They are the same as those in the scenario where the i.i.d assumption is made for the data as stated in Section 2.3.1.

2.3.3 Two-layer resampling

The third method is two-layer resampling that takes place randomly WR not only on the first layer of the data but also on the second layer of the data, i.e., target scores and non-target score in the sets, respectively, which are assumed to be conditionally independent. In other words, the resampling units for the first layer are sets and for the second layer are scores in the sets.

Then, the probability for a score $\alpha_{i j k}$ in a set $S_{i j}$ being selected is

$$P_{2\text{-layer}}(\alpha_{i j k}) = P(S_{i j}) \times P(\alpha_{i j k} | S_{i j}) = \frac{1}{m_i} \times \frac{1}{\mu_{i j}}, \quad (7)$$

where $k = 1, \dots, \mu_{i j}, j = 1, \dots, m_i$ and $i \in \{T, N\}$.

This probability is the same for all scores within a set, regardless of whether it is a target set or a non-target set. However, it is noticed that the probabilities for scores being selected are different from set to set due different score numbers in different sets indicated by $\mu_{i j}$. This situation is quite different from those using the previous two methods.

If all $\mu_{T j}, j = 1, \dots, m_T$, are set to be the same, then each target score can have equal probability to be selected. The probability for each target score being selected is $1 / N_T$. And so is each non-target score if all $\mu_{N j}, j = 1, \dots, m_N$, are set to be equal. The probability for each non-target score being selected is $1 / N_N$.

2.4 The new data structure after adjustment

It is absolutely not appropriate that target scores and/or non-target scores are selected with unequal probabilities for two-layer resampling. The impact of varied numbers of scores within a set on the probabilities for a score being selected must be eliminated. In other words, the numbers of scores in target sets must be equal and likewise for the numbers of scores in non-target sets.

Thereafter, the probabilities for each target score and each non-target score being selected will be the same, respectively, among the three resampling methods as stated in Section 2.3 while using the bootstrap to compute the measurement uncertainties of the detection cost function. As a result, the measurement uncertainties calculated using these three resampling methods can be compared on an equal footing.

In the meantime, the new datasets should be restructured in such a way that the total numbers of target scores and non-target scores must be obtained as large as possible. As described in Section 1, the raw speaker recognition datasets contain about 20,000 target scores and 80,000 non-target scores.

After adjustment based on the interpretation of data dependency implied in the raw datasets as discussed in Section 2.1, the new datasets have 132 target sets, each of which contains 96 target scores that are randomly selected without replacement from the raw target scores in the set if their number is greater than or equal to 96; and thus the total number of target scores is 12,672.

The new datasets also contain 130 non-target sets, each of which has 244 non-target scores that are randomly selected without replacement from the raw target scores in the set if their number is greater than or equal to 244; and thus the total number of non-target scores is 31,720.

Certainly, if the numbers of target scores and non-target scores in the sets are equal to 96 and 244, respectively, no random selection is needed and all scores are selected. On the other hand, the histograms of numbers of raw scores in sets depicted in Figure 1 show that most target sets and non-target sets which contain less numbers of scores are discarded.

3 The detection cost function in speaker recognition evaluation

After converting to integer scores as mentioned in Section 1, without loss of generality, for a speaker recognition system, the scores are expressed inclusively using the integer score set $\{s\} = \{s_{\min}, s_{\min}+1, \dots, s_{\max}\}$, running consecutively from the lowest score s_{\min} to the highest score s_{\max} . Let $C_i(s)$, $i \in \{T, N\}$ denote the cumulative probabilities of target scores and non-target scores from the highest score s_{\max} down to an integer score s , respectively.

The probability of type I error at a threshold $t \in \{s\}$ for target scores, denoted by $P_I(t)$, is cumulated from the lowest score s_{\min} . The probability of type II error at a threshold t for non-target scores, denoted by $P_{II}(t)$, is cumulated from the highest score s_{\max} . For discrete probability distribution, while computing $P_I(t)$ and $P_{II}(t)$ at a threshold t , the probabilities of target scores and non-target scores at this threshold t must be taken into account [15].

Hence, at a threshold value $t \in \{s\}$, the estimators of the probabilities of type I error and type II error are expressed, respectively, as

$$\begin{aligned} \hat{P}_I(t) &= 1 - C_T(t+1) \\ \hat{P}_{II}(t) &= C_N(t) \end{aligned} \quad \text{for } t \in \{s\}, \quad (8)$$

where $C_T (s_{\max} + 1) = 0$ is assumed [3]. Based on Eq. (8), in practice, the estimators $\hat{P}_I (t)$ and $\hat{P}_{II} (t)$ can be obtained by moving the score from the highest score s_{\max} down to the threshold t one score at a time to cumulate the probabilities of target scores and non-target scores, respectively.

A number of metrics exist for measuring the performance of a speaker recognition system [1]. In this article, the detection cost function at a threshold for the primary evaluation of speaker detection performance is employed as the metric of interest. Certainly, the same method of computing the uncertainties of the detection cost functions can be used to compute uncertainties for other metrics in SRE.

The detection cost function at a threshold t is defined as a weighted sum of the probabilities of type I error and of type II error at the threshold t [1]

$$C_{\text{Det}} (t) = C_{\text{Miss}} \times P_I (t) \times P_{\text{Target}} + C_{\text{FalseAlarm}} \times P_{II} (t) \times (1 - P_{\text{Target}}) . \quad (9)$$

Hence, it is a function of the threshold t . It was required that the thresholds be provided by speaker recognition systems in order to make an explicit speaker detection decision for each trial. The thresholds can also be determined in other ways. It is a challenging research problem to determine appropriate decision thresholds, which is out of the scope of this article. Therefore, the thresholds used in this article are those provided by the tested systems.

The parameters C_{Miss} and $C_{\text{FalseAlarm}}$ are the relative costs of detection errors, and the parameter P_{Target} is the *a priori* probability of the specified model speaker. For the primary evaluation of speaker recognition performance for all speaker detection tests, the parameters C_{Miss} , $C_{\text{FalseAlarm}}$, and P_{Target} were set to be 10, 1, and 0.01, respectively [1].

4 The nonparametric two-sample bootstrap using three resampling methods

It is difficult to compute analytically the covariance term of the correlated probabilities of type I error $P_I (t)$ and type II error $P_{II} (t)$ at a threshold t in Eq. (9). Thus, the estimate of the uncertainty of the detection cost function at a threshold t in terms of SE is computed using three different resampling methods described in Section 2.3 in the nonparametric two-sample bootstrap based on our extensive studies of bootstrap variability on large datasets [3-8].

4.0 A function WR_Random_Sampling_Set

First of all, here is a function `WR_Random_Sampling_Set` that will be frequently employed in the following algorithms,

```

1: function WR_Random_Sampling_Set (N, S, Γ)
2: for i = 1 to N do
3:   select randomly WR an index  $j \in \{ 1, \dots, N \}$ 
4:    $\gamma_i = S_j$ 
5: end for
6: end function

```

where \mathbf{S} stands for a set of sets or a set of scores, N is the cardinality of the set \mathbf{S} , $\mathbf{\Gamma}$ represents a new set of sets or scores accordingly with the same cardinality, and s_j and γ_i are members of the sets \mathbf{S} and $\mathbf{\Gamma}$, respectively. Note that this function can be applied to either a set of sets or a set of scores. It runs N iterations as shown from Step 2 to Step 5. In the i -th iteration, a member of the set \mathbf{S} is randomly selected WR to become a member of a new set $\mathbf{\Gamma}$, as indicated in Steps 3 and 4. As a result of this function, N members (sets or scores) are randomly selected WR from the set \mathbf{S} and constitute a new set $\mathbf{\Gamma}$.

4.1 An algorithm assuming the data is i.i.d.

As shown in Eq. (3) through Eq. (5), \mathbf{T} is the set of all N_T target scores and \mathbf{N} is the set of all N_N non-target scores. Then, if the i.i.d. assumption is made for the data, the two-sample bootstrap resampling units are target scores and non-target scores, respectively. Hence, the algorithm of nonparametric two-sample bootstrap is as follows.

Algorithm 1 (i.i.d. bootstrap)

```

1: for  $i = 1$  to  $B$  do
2:   WR_Random_Sampling_Set ( $N_T, \mathbf{T}, \Theta^i$ )
3:   WR_Random_Sampling_Set ( $N_N, \mathbf{N}, \Xi^i$ )
4:    $\Theta^i$  and  $\Xi^i \Rightarrow$  statistic  $\hat{C}^i$ 
5: end for
6:  $\{\hat{C}^i | i=1, \dots, B\} \Rightarrow \hat{SE}$ 
7: end

```

where B is the number of two-sample bootstrap replications. In other words, this algorithm runs B times. As shown from Step 1 to 5, in the i -th iteration, by calling the function in Section 4.0 twice, N_T target scores are randomly selected WR from the set \mathbf{T} to form a new set Θ^i of N_T target scores, N_N non-target scores are randomly selected WR from the set \mathbf{N} to constitute a new set Ξ^i of N_N non-target scores, and then all N_T target scores in the new set Θ^i and all N_N non-target scores in the new set Ξ^i generate the i -th bootstrap replication of the estimated statistic of interest, i.e., \hat{C}^i . Note that from here on in this article, the superscript indices are used for the numeration of the resampling iterations. As always, the subscript indices are employed to indicate target or non-target, and numerate sets and scores.

In the SRE, the estimated statistic of interest \hat{C}^i is the i -th estimator of the detection cost function at a given threshold. This estimator can be derived using Eq. (9). In this equation, the estimators of the probabilities of type I error and type II error, i.e., $\hat{P}_I(t)$ and $\hat{P}_{II}(t)$, can be calculated from the two new sets of similarity scores using Eq. (8). Finally, as indicated in Step 6, from the set $\{\hat{C}^i | i = 1, \dots, B\}$, the standard error \hat{SE} of the detection cost function is estimated by the sample standard deviation of the B replications.

The remaining issue is to determine how many iterations this bootstrap algorithm needs to run in order to reduce the bootstrap variance and ensure the accuracy of the computation. In other words, what is an appropriate number B of the nonparametric two-sample bootstrap replications? In our applications, such as biometrics and the evaluation of speaker recognition, etc., the sizes of datasets are tens or hundreds of thousands of similarity scores, which are much larger than those in some other applications of bootstrap methods, such as medical decision making, etc.. Moreover, in our applications, the statistics of interest are mostly probabilities or a weighted sum of probabilities, etc. rather than a simple sample mean. And our data samples of similarity scores have no parametric model to fit. Therefore, the bootstrap variability was re-studied empirically, and the appropriate number of bootstrap replications B for our applications was determined to be 2,000 [3, 6, 7].

4.2 An algorithm for one-layer nonparametric two-sample bootstrap

As discussed in Section 2, the one-layer resampling takes place only on the first layer of the new data structure, namely, the nonparametric two-sample bootstrap units are target sets and non-target sets, respectively. Thus, the algorithm for one-layer nonparametric two-sample bootstrap is as follows.

Algorithm II (one-layer bootstrap)

```

1: for  $i = 1$  to  $B$  do
2:   WR_Random_Sampling_Set ( $m_T, S_T, S_T'^i = \{ S_{Tj}'^i | j = 1, \dots, m_T \}$ )
3:   WR_Random_Sampling_Set ( $m_N, S_N, S_N'^i = \{ S_{Nj}'^i | j = 1, \dots, m_N \}$ )
4:    $S_T'^i$  and  $S_N'^i \Rightarrow$  statistic  $\hat{C}^i$ 
5: end for
6:  $\{ \hat{C}^i | i = 1, \dots, B \} \Rightarrow \hat{S}\hat{E}$ 
7: end

```

where B is the number of two-sample bootstrap replications, the set S_T of all target sets and the set S_N of all non-target sets are expressed in Eq. (1), and m_T and m_N are the cardinalities of the sets S_T and S_N , respectively. In the i -th iteration, as indicated in Step 2 and Step 3, the function in Section 4.0 is applied twice to sets rather than scores. That is, m_T target sets are randomly selected WR from the set S_T to form a new set $S_T'^i = \{ S_{Tj}'^i | j = 1, \dots, m_T \}$, and m_N non-target sets are randomly selected WR from the set S_N to constitute a new set $S_N'^i = \{ S_{Nj}'^i | j = 1, \dots, m_N \}$. As noted in Step 4, all target scores in the new set $S_T'^i$ and all non-target scores in the new set $S_N'^i$ generate the i -th bootstrap replication of the estimated statistic of interest, i.e., \hat{C}^i using Eq. (9). Everything else in this algorithm is the same as those in the algorithm shown in Section 4.1.

It is worth mentioning that with the new data structure as shown in Section 2.4, in each i -th resampling of this one-layer nonparametric two-sample bootstrap, the same number of target scores and the same number of non-target scores are obtained to generate the i -th estimate of the statistic of interest. Otherwise, different numbers of similarity scores would be obtained during different iterations, which could cause more variance in the computation.

4.3 An algorithm for two-layer nonparametric two-sample bootstrap

As described in Section 2, the two-layer resampling is carried on not only on the first layer of the new data structure in which the two-sample bootstrap units are target sets and non-target sets, but also on the second layer of the data in which the bootstrap units are target scores and non-target scores in sets. Hence, the algorithm for two-layer nonparametric two-sample bootstrap is as follows.

Algorithm III (two-layer bootstrap)

```

1: for  $i = 1$  to  $B$  do
2:   WR_Random_Sampling_Set ( $m_T, S_T, S_T''^i = \{ S_{Tj}''^i | j = 1, \dots, m_T \}$ )
3:     for  $k = 1$  to  $m_T$  do
4:       WR_Random_Sampling_Set ( $\mu_{Tk}, S_{Tk}''^i, S_{Tk}''^i$ )

5:   WR_Random_Sampling_Set ( $m_N, S_N, S_N''^i = \{ S_{Nj}''^i | j = 1, \dots, m_N \}$ )
6:     for  $k = 1$  to  $m_N$  do
7:       WR_Random_Sampling_Set ( $\mu_{Nk}, S_{Nk}''^i, S_{Nk}''^i$ )

8:    $S_T''^i = \{ S_{Tj}''^i | j = 1, \dots, m_T \}$  and  $S_N''^i = \{ S_{Nj}''^i | j = 1, \dots, m_N \} \Rightarrow$  statistic  $\hat{C}^i$ 
9: end for
10:  $\{ \hat{C}^i | i = 1, \dots, B \} \Rightarrow \hat{SE}$ 
11: end

```

where B is the number of two-sample bootstrap replications, the set S_T of all target sets and the set S_N of all non-target sets are expressed in Eq. (1), and m_T and m_N are the cardinalities of the set S_T and the set S_N , respectively. In the i -th iteration, as shown in Step 2 and Step 5 the function in Section 4.0 is applied twice to sets, which is the same as in the one-layer bootstrap *Algorithm II* in Section 4.2. Subsequently, the same function is applied to similarity scores in sets as well.

As shown in Steps 3 and 4, m_T iterations take place after the first-layer resampling of the target sets in Step 2. In the k -th iteration, μ_{Tk} target scores are randomly selected WR from the target set $S_{Tk}''^i$, which is the k -th new target set from the first-layer resampling, to form the k -th new target set $S_{Tk}''^i$ of the second-layer resampling.

As indicated in Steps 6 and 7, m_N iterations take place after the first-layer resampling of the non-target sets in Step 5. In the k -th iteration, μ_{Nk} non-target scores are randomly selected WR from the non-target set $S_{Nk}''^i$, which is the k -th new non-target set from the first-layer resampling, to constitute the k -th new non-target set $S_{Nk}''^i$ of the second-layer resampling.

As shown in Step 8, all target scores in the new set $S_T''^i = \{ S_{Tj}''^i | j = 1, \dots, m_T \}$ and all non-target scores in the new set $S_N''^i = \{ S_{Nj}''^i | j = 1, \dots, m_N \}$ generate the i -th bootstrap replication of the estimated statistic of interest, i.e., \hat{C}^i using Eq. (9). Everything else in this algorithm is the same as in the algorithm shown in Section 4.1.

With the new data structure described in Section 2.4, the same numbers of similarity scores are selected during different iterations of the two-layer nonparametric two-sample bootstrap. It can reduce the variance of the computation. This is the same as in the one-layer bootstrap as stated in Section 4.2.

5 A method of generating distributions of SEs of the detection cost function

If SEs obtained by using different bootstrap algorithms need to be compared, one SE is far less enough, due to the stochastic nature of the bootstrap method as discussed in Section 1. In other words, a distribution of SEs generated by using an algorithm needs to be investigated.

All three algorithms shown in Section 4 can only create one estimated $\hat{S}\hat{E}$ of the detection cost function, respectively. However, if such an algorithm runs multiple times, it can generate a distribution of estimated $\hat{S}\hat{E}$ s. Based on our previous studies, in order to create a stable distribution, it is enough that the algorithm be executed 500 times [3, 6-8].

Hence, the algorithms shown in Section 4.1, 4.2, and 4.3 are executed 500 times each to generate a distribution $\{ \hat{S}\hat{E}^i \mid i = 1, \dots, 500 \}$. Thereafter, the estimated mean and 95% confidence interval (CI) of such a distribution can be calculated. While computing the estimated 95% CI of a distribution, the Definition 2 of quantile in Ref. [16] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities.

6 Results

Four speaker recognition systems, labeled as EL, LZ, PB, and CH, are employed as examples². They have different matching accuracies as indicated by the values of their detection cost functions – the smaller the detection cost functions, the more accurate the speaker recognition systems. The estimated detection cost functions of the four systems, and the estimated means, $\hat{S}\hat{E}$ s and 95 % CIs of the distributions of SEs of their detection cost functions generated using three different bootstrap methods, namely, the i.i.d. bootstrap, the one-layer bootstrap, and the two-layer bootstrap, are all shown in Table 3, where the speaker recognition systems are listed in the ascending order of the value of the cost function. It shows in Table 3 that generally the smaller the detection cost functions, the smaller the uncertainties. The corresponding distributions of SEs of the detection cost functions along with the estimated means represented by black circles are depicted in Figure 2.

The distributions of SEs shown in Table 3 and Figure 2 have two important features. The first feature is that the variance of the distribution of SEs generated using the i.i.d. bootstrap is less than the variances of other two distributions of SEs for each speaker recognition system, as shown in Table 3. This feature is reflected by the widths of the histograms depicted in Figure 2. As is known, the bootstrap method is a stochastic process, namely, different runs of bootstrap method should produce different results of SEs. Hence, this feature indicates that the bootstrap method applied on the i.i.d. datasets can create less variation of SEs than the bootstrap method conducted on the datasets with dependency does.

² The speaker recognition systems are proprietary. Hence, they cannot be disclosed.

Systems	Cost Function	Mean, SE and 95% CI of distribution of SEs of cost function		
		i.i.d. Bootstrap	One-Layer Bootstrap	Two-Layer Bootstrap
EL	0.022199	0.000687 0.105594×10^{-4} (0.000666, 0.000706)	0.001859 0.292555×10^{-4} (0.001806, 0.001920)	0.001975 0.329929×10^{-4} (0.001916, 0.002043)
LZ	0.040098	0.000888 0.133725×10^{-4} (0.000863, 0.000917)	0.002730 0.446791×10^{-4} (0.002646, 0.002817)	0.002870 0.460217×10^{-4} (0.002781, 0.002956)
PB	0.098744	0.001119 0.182849×10^{-4} (0.001084, 0.001155)	0.004150 0.677488×10^{-4} (0.004012, 0.004286)	0.004288 0.675055×10^{-4} (0.004149, 0.004420)
CH	0.236771	0.002294 0.367097×10^{-4} (0.002224, 0.002367)	0.004646 0.759175×10^{-4} (0.004509, 0.004790)	0.005172 0.819121×10^{-4} (0.005020, 0.005345)

Table 3 The estimated detection cost functions, the means, $\hat{S}\hat{E}$ s and 95 % $\hat{C}\hat{I}$ s of distributions of SEs of the detection cost functions generated using the i.i.d. bootstrap, the one-layer bootstrap, and the two-layer bootstrap, respectively, for four speaker recognition systems labeled as EL, LZ, PB, and CH.

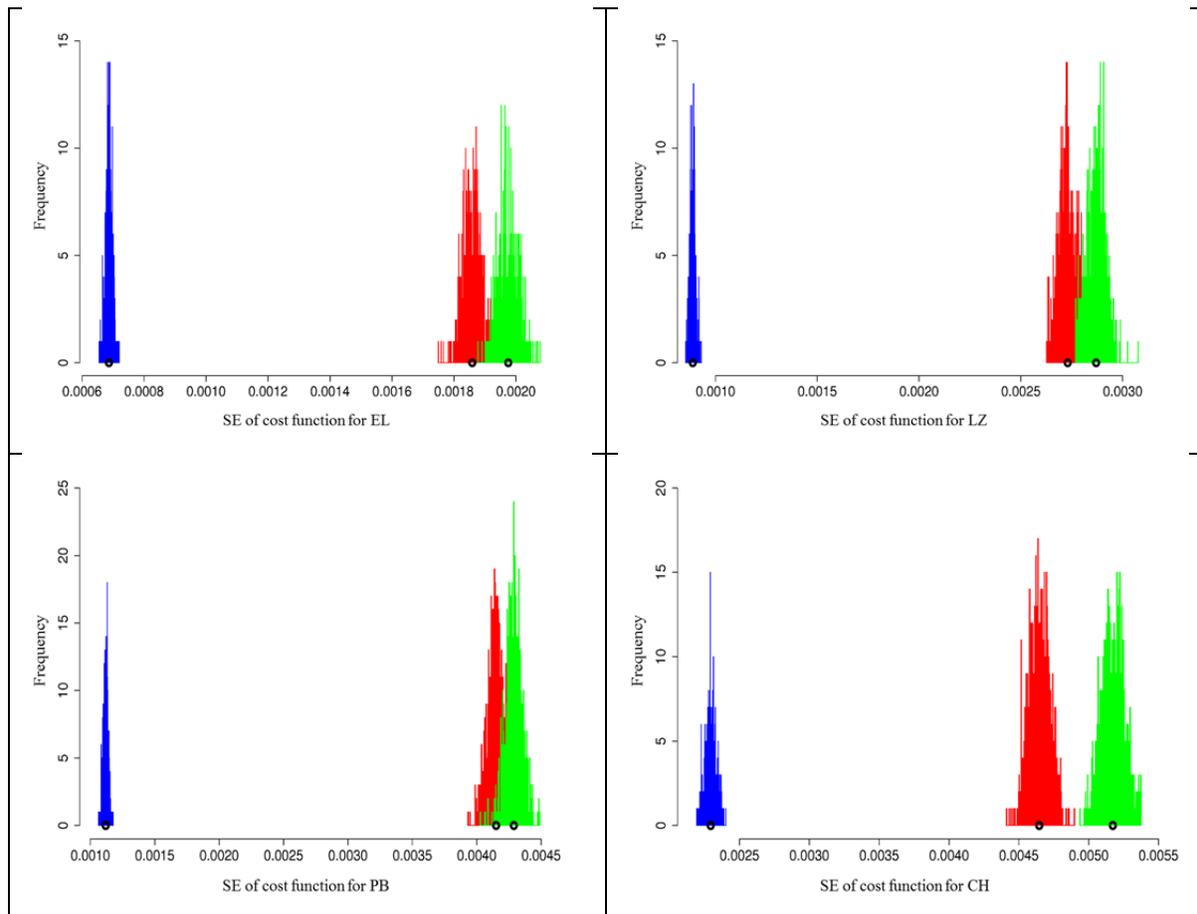


Figure 2 The histograms of SEs of the detection cost functions generated using the i.i.d. bootstrap (left - blue), the one-layer bootstrap (middle - red), and the two-layer bootstrap (right - green), respectively, for four speaker recognition systems EL, LZ, PB, and CH. The black circle stands for the estimated mean of the distribution.

The second feature is regarding the relationship among the positions of three distributions of SEs. The two distributions of SEs created using the one-layer bootstrap and the two-layer bootstrap are well separated, towards larger SEs, from the distribution of SEs generated using the i.i.d. bootstrap for each of four systems. Now the question is: What is the relationship between the former two distributions on the right side for each system shown in Figure 2?

First, some preliminary observations regarding estimated means and variances of distributions are made. These two distributions overlap to some extent. The estimated 95% $\hat{C}I$ of one distribution overlaps the estimated 95% $\hat{C}I$ of the other distribution for each system except System CH, as shown in Table 3. However, the estimated mean of one distribution is at the outside or just at the border of the estimated 95% $\hat{C}I$ of the other distribution. In addition, the ratio of the variance of the distribution of SEs computed using the one-layer bootstrap method to the variance of the distribution of SEs calculated using the two-layer bootstrap method is between 0.79 and 1.01 for all four systems, which can be obtained by the estimated $\hat{S}E$ s of the distributions presented in Table 3.

Second, the hypothesis testing is conducted on both estimated means and variances of distributions. As stated in Section 5, the estimated 95% $\hat{C}I$ s shown in Table 3 were all calculated using the quantile method. They do match up to the fourth to fifth decimal place the estimated 95% $\hat{C}I$ s computed by assuming the distribution of SEs of the detection cost function is normal, i.e., multiplying 1.96 by the estimated $\hat{S}E$ of such a distribution. For instance, for System EL using the two-layer bootstrap method, the estimated 95% $\hat{C}I$ s computed using the quantile method is (0.001916, 0.002043) as shown in Table 3 and the one calculated using the normality assumption is (0.001911, 0.002040). Moreover, the Shapiro-Wilk normality test [17] was conducted on the 12 distributions of SEs (three resampling methods for each of the four systems), and it was observed that nine p-values were between 14% and 88% which were much greater than 5% and three p-values were 1.7%, 0.5%, and 0.5%, respectively. These analyses suggest that the estimated $\hat{S}E$ s of the detection cost function calculated using three different resampling methods in bootstrap may be regarded as normally distributed.

Hence, the Z-test for comparing the means and the F-test for comparing the variances can be carried out on the distribution of SEs of the detection cost function computed using the one-layer bootstrap method and the distribution of SEs calculated using the two-layer bootstrap method for each speaker recognition system [3, 15, 17, 18]. It indicates in Table 3 and Figure 2 that the mean of the former distribution is less than the mean of the latter distribution. Hence, the one-tailed Z-test is applied. All p-values for the four systems are close to one, which strongly suggests the null hypothesis is accepted. In other words, the mean of the distribution created by the one-layer bootstrap is significantly below the mean of the distribution generated by the two-layer bootstrap regardless of the speaker recognition systems. Further, the p-values of the two-tailed F-test are all greater than 5% except for System EL, which is 0.73%. It suggests that the null hypothesis, i.e., the ratio of the variances of these two distributions is equal to one, cannot be rejected.

Combining the results from these two hypothesis tests plus the preliminary observations, it can be concluded that the distribution of SEs of the cost function computed using the two-layer bootstrap method for the datasets with inherent data dependencies is significantly on the right side of the

distribution of SEs calculated using the one-layer bootstrap method, whereas the latter is well separated, towards larger SEs, from the distribution of SEs computed using the bootstrap method assuming the data are i.i.d..

Therefore, the data dependency increases the measurement uncertainties. The two-layer bootstrap method more conservatively estimates the impact of the data dependency on the measurement uncertainties than the one-layer bootstrap method does, since the second-layer resampling increases the variations of the measure. Both methods can increase the variations of SEs with respect to the method with the i.i.d. assumption. If the dependencies do exist in the datasets, the bootstrap method with the i.i.d. assumption can underestimate the measurement uncertainties. In order to obtain more accurate measures and less variation of SEs, the best way is to randomly select i.i.d. data samples. If the data dependency is inevitable for increasing the size of datasets due to limited resources, then the i.i.d. assumption cannot be made and the two-layer bootstrap method is recommended while using the bootstrap methods to compute the measurement uncertainties.

7 Conclusions and discussion

In many applications, it is hard to compute the SE of a measure analytically. This happens in the speaker recognition evaluation, in which the statistic of interest is a detection cost function, which is defined as a weighted sum of the probabilities of type I error and of type II error. These two probabilities are traded off each other and thus negatively correlated. Indeed, it is difficult to compute such a correlation coefficient analytically.

In order to calculate the uncertainties of such a measure, the alternative way is to employ the nonparametric two-sample bootstrap method in our applications. The premise of using the bootstrap method is that the datasets must be i.i.d.. If the datasets are i.i.d., the nonparametric two-sample bootstrap method based on our extensive bootstrap variability studies on large datasets can be employed in our applications without any modification.

In this article, the impact of the data dependency on the uncertainties of measures in speaker recognition evaluation by using the nonparametric two-sample bootstrap was studied. If the datasets contain data dependencies due to multiple usages of the same subject in order to increase the size of datasets due to limited resources, the bootstrap method with i.i.d. assumption, i.e., the bootstrap units are scores without grouping them into sets, can underestimate the uncertainties of measures.

To compute the uncertainties of measures on the datasets with dependencies, the two-layer bootstrap method rather than the one-layer bootstrap method is recommended, since the former can more conservatively estimate the impact of the data dependency on the measurement uncertainties than the latter. In order to properly employ the two-layer bootstrap method, after grouping the data into sets, the numbers of target scores in target sets must be the same and likewise for the numbers of non-target scores in non-target sets. In such a way, the probability for each target score being selected will be equal, and so is the probability for each non-target score being selected.

The data dependency involved in the datasets, acting like “noise” behind “signal”, can increase the estimated measurement uncertainties as well as the variations of SEs, and thus the accurate measure

cannot be achieved. Hence, in order to obtain more accurate measures, the data dependency must be avoided. Indeed, from the statistical point of view, the sample should be collected randomly in test design. Further, the data dependency may be interpreted in different ways because of the complications of multiple usages of the same subject. Different interpretations of data dependencies can cause different ways of grouping similarity scores into sets and subsequently will have impact on the bootstrap results.

It seems that the large size of datasets cannot reduce the impact of data dependency on the measurement uncertainties. The reason why the SEs of the area under ROC curve computed using the bootstrap method for the large size of datasets containing data dependency but with the i.i.d. assumption are very close to those calculated analytically using the Mann-Whitney statistics is that the methods are blind to the data dependency implied in the datasets [9]. Nonetheless, on the other side, running on the i.i.d. datasets, the fact that these two types of results are very close validates the nonparametric two-sample bootstrap methods in our applications [8].

Appendix – the Probability for One-Layer Resampling

The proof of the probability for one-layer resampling holds good for both target sets and non-target sets. Suppose that N scores are grouped into m score sets and the i -th set contains μ_i scores, $i = 1, \dots, m$, where $\sum_{i=1}^m \mu_i = N$. Suppose that n selections take place and the i -th set is selected n_i times, $i = 1, \dots, m$, where $\sum_{i=1}^m n_i = n$. Thus, the relative frequency f of occurrence of a score in the i -th set is

$$f = \frac{n_i}{\sum_{j=1}^m n_j \times \mu_j} = \frac{\frac{n_i}{n}}{\sum_{j=1}^m \frac{n_j}{n} \times \mu_j}. \quad (\text{A.1})$$

In the meantime, because the sets are equally likely to be selected for the one-layer resampling,

$$\lim_{n \rightarrow \infty} \frac{n_i}{n} = \frac{1}{m}, \quad i = 1, \dots, m. \quad (\text{A.2})$$

Therefore, due to the Law of Large Numbers, as the number of selections n goes to infinity, the relative frequency in Eq. (A.1) approaches to the probability for a score being selected, that is,

$$p = \frac{\frac{1}{m}}{\sum_{j=1}^m \frac{1}{m} \times \mu_j} = \frac{1}{\sum_{j=1}^m \mu_j} = \frac{1}{N}. \quad (\text{A.3})$$

References

1. "The NIST Year 2008 Speaker Recognition Evaluation Plan", the URL of the website is at http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf, (2008).
2. J.C. Wu, and C.L. Wilson, An empirical study of sample size in ROC-curve analysis of fingerprint data, in Biometric Technology for Human Identification III, Proc. SPIE 6202, 620207 (2006).
3. J.C. Wu, A.F. Martin and R.N. Kacker, Measures, uncertainties, and significance test in operational ROC analysis, J. Res. Natl. Inst. Stand. Technol. 116 (1), 517-537 (2011).
4. B. Efron, Bootstrap methods: Another look at the Jackknife, Ann. Statistics 7, 1-26 (1979).
5. B. Efron, and R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, (1993).
6. J.C. Wu, Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap, NISTIR 7449, National Institute of Standards and Technology, September, (2007).
7. J.C. Wu, A.F. Martin and R.N. Kacker, Further studies of bootstrap variability for ROC analysis on large datasets, NISTIR 7730, National Institute of Standards and Technology, October, (2010).
8. J.C. Wu, A.F. Martin and R.N. Kacker, Validation of two-sample bootstrap in ROC analysis on large datasets using AURC, NISTIR 7733, National Institute of Standards and Technology, October, (2010).
9. J.C. Wu, A.F. Martin, C.S. Greenberg and R.N. Kacker, Uncertainties of measures in speaker recognition evaluation, in Active and Passive Signatures II, Proc. SPIE 8040, 804008 (2011).
10. J.C. Wu, and C.L. Wilson, Nonparametric analysis of fingerprint data on large data sets, Pattern Recognition 40 (9), 2574-2584 (2007).
11. J.C. Wu, and M.D. Garris, Nonparametric statistical data analysis of fingerprint minutiae exchange with two-finger fusion, in Biometric Technology for Human Identification IV, Proc. SPIE 6539, 65390N (2007).
12. R.Y. Liu, and K. Singh, Moving blocks jackknife and bootstrap capture weak dependence, in Exploring the limits of bootstrap, ed. by LePage and Billard. John Wiley, New York, (1992).
13. R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha and A.W. Senior, Guide to Biometrics, Springer, New York, 269-292 (2003).
14. N. Poh, A.F. Martin and S. Bengio, Performance generalization in biometric authentication using joint user-specific and sample bootstraps, IEEE Trans. Pattern Analysis and Machine Intelligence, 29(3), 492-498 (2007).
15. B. Ostle, and L.C. Malone, Statistics in Research: Basic Concepts and Techniques for Research Workers, fourth ed., Iowa State University Press, Ames, (1988).
16. R.J. Hyndman, and Y. Fan, Sample quantiles in statistical packages, American Statistician 50, 361-365 (1996).
17. R: A Language and Environment for Statistical Computing, The R Development Core Team, Version 2.8.0, 2008, at <http://www.r-project.org/>.
18. J.C. Wu, A.F. Martin, R.N. Kacker and C.R. Hagwood, Significance test in operational ROC analysis, in Biometric Technology for Human Identification VII, Proc. SPIE 7667, 76670I (2010).