Data Dependency on Measurement Uncertainties in Speaker Recognition Evaluation

Jin Chu Wu, Alvin F. Martin, Craig S. Greenberg and Raghu N. Kacker
Information Technology Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

The National Institute of Standards and Technology 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 and type II errors. The sampling variability can result in measurement uncertainties. In our prior study, the data independency was assumed in using the nonparametric two-sample bootstrap method to compute the standard errors (SE) of the detection cost function based on our extensive bootstrap variability studies in ROC analysis on large datasets. In this article, the data dependency caused by multiple uses of the same subjects is taken into account. 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, or subsequently on target scores and non-target scores within the sets, respectively. 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. Some suggestions regarding the test design are provided.

Keywords: Data dependency; Measurement uncertainty; Speaker recognition evaluation; Biometrics; ROC analysis; Bootstrap; Standard error.

1 Introduction

The ongoing series of Speaker Recognition Evaluations (SRE) conducted by the National Institute of Standards and Technology (NIST) has had a great impact on the research efforts and the technical capabilities regarding the general problem of text independent speaker recognition [1]. Each test consisted of a sequence of trials. Each trial consisted of a training model speaker and a test speech segment. The speaker recognition system should decide whether the speech of the model speaker occurred in the test speech segment and generate a similarity score. A higher score indicates greater confidence that the test speech is spoken by the model speaker. 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. Each test included about 20,000 target scores and 80,000 non-target scores [1, 2].

The speaker detection performance is measured using a detection cost function, that is defined as a weighted sum of the probabilities of type I error (miss) and type II error (false alarm), which

Active and Passive Signatures III, edited by G. Charmaine Gilbreath, Chadwick T. Hawley
Proc. of SPIE Vol. 8382, 83820D · © 2012 SPIE · CCC code: 0277-786X/12/$18 · doi: 10.1117/12.918467

represent a tradeoff and are negatively correlated [1]. The sampling variability results in uncertainties of any measures [3]. It is hard to calculate the variance of the detection cost function analytically because the two probabilities are correlated.

In our prior study of the SRE, the uncertainties in terms of standard errors (SE) were computed using the nonparametric two-sample bootstrap method based on our extensive bootstrap variability studies in ROC analysis on large datasets, using the assumption of data independency [3-9]. The two samples involved are referred to as a set of target scores and a set of non-target scores, which characterize the speaker recognition system that generates them and usually do not have well defined parametric forms [9-11]. The bootstrap method assumes that the random samples drawn from a population are independent and identically distributed (i.i.d.). With such an assumption, the bootstrap units are scores in the sample.

However, the NIST speaker recognition data do contain dependency. In this article, the data dependency is taken into account while computing the measurement uncertainties in SRE. The data dependency arises basically from multiple uses of the same subjects 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.

In this article, the 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 can preserve the data dependency while the bootstrap resampling takes place. Certainly, how the sample is grouped into sets can impact the bootstrap results. Now 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., one-layer and two-layer bootstrap methods are proposed based on whether the nonparametric two-sample bootstrap resampling takes place randomly with replacement (WR) 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 can cause that the probabilities for target (non-target) scores being selected are not the same and the numbers of scores resampled are different from iteration to iteration when the bootstrap resampling is carried out. To avoid all these, the speaker recognition data is adjusted so that all target sets 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 are kept as large as possible. After data adjustment, the SEs of the detection cost functions computed using the three bootstrap methods can be compared on an equal footing. And it can reduce the variance of the computation as well.

Further, the bootstrap method is a stochastic process. The results will vary for different runs, and thus constitute a probability distribution of SEs. 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 spread of the distributions of SEs. It is found that taking account of the data dependency increases the estimated SEs and the variations of SEs.

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. Through two-sample bootstrap method, the uncertainties of much more complicated measures, such as the detection cost function defined as a weighted sum of two probabilities, can be computed. Further, in this article the probability issues of similarity scores being selected and the numbers of scores resampled at each iteration are investigated, and the hypothesis testing is conducted on the distributions of SEs to reveal their relationships while dealing with different resampling methods in bootstrap.

As pointed out in Ref. [9], the speaker recognition data had their own characteristics. 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) of real-number similarity scores were preserved while converted to integers. In addition, only a few of similarity scores of the speaker recognition data took the same matching value. Thus, the computation could take an excessive amount of time.

Adjustment of speaker recognition datasets based on the selection probabilities is discussed in Section 2. The formulas for computing the detection cost function are shown in Section 3. The three nonparametric two-sample bootstrap algorithms and how to simulate distributions of the SEs of the detection cost function are presented in Section 4. The resulting measurement uncertainties employing the three bootstrap methods for two speaker recognition systems[1], and comparisons among them are shown in Section 5. The conclusions and discussion can be found in Section 6.

2 Adjustment of speaker recognition datasets

The speaker recognition data dependency is complicated. There are several ways to group data into sets according to different interpretations of data dependency as discussed in Section 1. In this

[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.

article, 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 any dependency implied by the training and test speech id numbers exists.

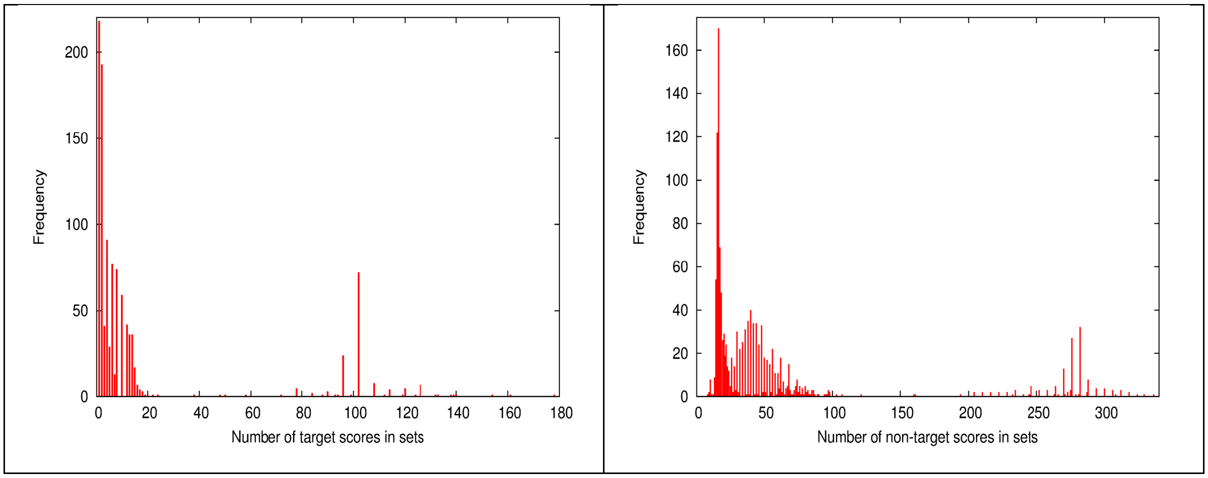

Figure 1 The histograms of the numbers of target scores in sets (Left) and non-target scores in sets (Right).

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. 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 non-target scores in sets, respectively, for the speaker recognition datasets 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.

In the following text, let $\boldsymbol{S}$ denote score sets and α represent similarity scores. The first subscript stands for whether it is referred to as target (T) or non-target (N), the second subscript means the ordinal number of the sets, and the third subscript represents the ordinal number of scores in the set. Thus, the subscript indices are employed to numerate sets and scores as well. For instance, $\boldsymbol{S}_T$ stands for the set of all target sets, $\boldsymbol{S}_{T1}$ denotes the first target set, and $\alpha_{T1\,\mu_{T1}}$ represents the μ_{T1}-th target score in the first target set, which also indicates that the number of target scores in the first target set is μ_{T1}. In other words, μ stands for the number of scores in a set.

Suppose that there are m_T target sets and m_N non-target sets, and the total numbers of target scores and non-target scores are N_T and N_N, respectively. They satisfy,

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

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. The empirical distribution is assumed for each of the observed scores.

If the speaker recognition datasets are assumed to be i.i.d., then the resampling units of the bootstrap method are all similarity scores. With this assumption, it is obvious that 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.

For the one-layer bootstrap method in which the resampling takes place randomly WR only on all score sets, it follows from the Law of Large Numbers that the probabilities for each target score and each non-target score being selected are also 1 / N_T and 1 / N_N , respectively [15].

For the two-layer bootstrap method where the resampling 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 scores in the sets, respectively, which are assumed to be conditionally independent, the probability for a score α_{ijk} in a set $\boldsymbol{S}_{ij}$ being selected is

$$P_{\text{2-layer}}(\alpha_{ijk}) = P(\boldsymbol{S}_{ij}) \text{ x } P(\alpha_{ijk} \mid \boldsymbol{S}_{ij}) = \frac{1}{m_i} \text{ x } \frac{1}{\mu_{ij}}, \tag{2}$$

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

These probabilities are the same for all scores within a set, but different from set to set, regardless of whether it is target or non-target.

The unequal selection probabilities for two-layer resampling must be eliminated, while using the bootstrap to compute the measurement uncertainties of the detection cost function. If the numbers of scores in target sets, i.e., μ_{Tj}, $j = 1, \ldots, m_T$, are set to be equal, then the probability for each target score being selected will be 1 / N_T due to Eq. (1). The same argument holds true for the non-target case. Consequently, this will be the same as those in the previous two resampling methods. Hence, the measurement uncertainties calculated using these three resampling methods can be compared on an equal footing. In addition, this can reduce the variance of the computation.

In the meantime, the new datasets should be restructured in such a way that the total numbers of target scores and non-target scores are kept as large as possible. After adjustment, the new datasets had 132 target sets (130 non-target sets), each of which contained 96 target scores (244 non-target scores) that were randomly selected without replacement from the raw target (non-target) set if the number of scores in the set was greater than or equal to what was required; and thus the total number of target (non-target) scores was 12,672 (31,720).

3 The detection cost function in speaker recognition evaluation

For a speaker recognition system, the scores after converting to integers are expressed inclusively using the integer score set $\{s\} = \{s_{min}, s_{min}+1, \ldots, s_{max}\}$, 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.

Let $P_I(t)$ denote the probability of type I error at a threshold $t \in \{s\}$ for target scores, which is cumulated from the lowest score s_{min}. Let $P_{II}(t)$ denote the probability of type II error at a threshold t for non-target scores, which is cumulated from the highest score s_{max}. The probabilities of target scores and non-target scores at this threshold t must be taken into account, while computing $P_I\ (t)$ and $P_{II}\ (t)$ at a threshold t for discrete probability distribution [16].

Thus, at a threshold value $t \in \{\,s\,\}$, their estimators are expressed, respectively, as

$$\left.\begin{aligned} \hat{P}_I\ (t) &= 1 - C_T\ (t + 1) \\ \hat{P}_{II}\ (t) &= C_N\ (t) \end{aligned}\right. \qquad \text{for } t \in \{\,s\,\}\ , \tag{3}$$

where $C_T\ (s_{max} + 1) = 0$ is assumed [3]. Based on Eq. (3), 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. In this article, the detection cost function is defined as the metric of interest in SRE, which is a weighted sum of the probabilities of type I and type II errors at a threshold t [1]

$$C_{Det}\ (t) = C_{Miss} \times P_I\ (t) \times P_{Target} + C_{FalseAlarm} \times P_{II}\ (t) \times (1 - P_{Target})\ . \tag{4}$$

The threshold t plays an important role here, which can be determined in many ways. It is a challenging research problem to determine appropriate decision thresholds, which is out of the scope of this article. In this article, the thresholds were provided by the tested speaker recognition systems in order to make an explicit speaker detection decision for each trial.

The parameters C_{Miss} and $C_{FalseAlarm}$ 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}, $C_{FalseAlarm}$, and P_{Target} were set to be 10, 1, and 0.01, respectively [1].

4 The three bootstrap methods and the distributions of SEs

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. (4). Thus, the estimate of the uncertainty of the detection cost function at a threshold t in terms of SE is computed using the three different nonparametric two-sample bootstrap methods based on our extensive studies of bootstrap variability in ROC analysis on large datasets [3-8].

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

```
function WR_Random_Sampling_Set (N, $\mathbf{\Gamma}$, $\mathbf{\Theta}$)
for i = 1 to N do
        select randomly WR an index j $\in$ { 1, …, N }
        $\theta_i = \gamma_j$
end for
end function
```

where **Γ** stands for a set of sets or a set of scores, N is the cardinality of the set **Γ**, **Θ** represents a new set of sets or scores accordingly with the same cardinality, and γ_j and θ_i are members of the sets **Γ** and **Θ**, respectively. Notice 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 **Γ** is randomly selected WR to become a member of a new set **Θ**, as indicated in Steps 3 and 4. As a result, N members (sets or scores) are randomly selected WR from the set **Γ** to form a new set **Θ**.

To simplify the presentation, the two-layer bootstrap algorithm is presented first. From here on, the superscript indices are used for the numeration of the resampling iterations. The two-layer resampling is conducted 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 the two-layer nonparametric two-sample bootstrap is as follows.

Algorithm **(two-layer bootstrap)**

```
1: for i = 1 to B do
2:     WR_Random_Sampling_Set ( $m_T, \boldsymbol{S}_T, \boldsymbol{S}'^{\,i}_{T} = \{ \boldsymbol{S}'^{\,i}_{Tj} \mid j = 1, \ldots, m_T \}$ )
3:     for k = 1 to $m_T$ do
4:         WR_Random_Sampling_Set ( $\mu'^{\,i}_{Tk}, \boldsymbol{S}'^{\,i}_{Tk}, \boldsymbol{S}''^{\,i}_{Tk}$ )
5:     end for
6:     WR_Random_Sampling_Set ( $m_N, \boldsymbol{S}_N, \boldsymbol{S}'^{\,i}_{N} = \{ \boldsymbol{S}'^{\,i}_{Nj} \mid j = 1, \ldots, m_N \}$ )
7:     for k = 1 to $m_N$ do
8:         WR_Random_Sampling_Set ( $\mu'^{\,i}_{Nk}, \boldsymbol{S}'^{\,i}_{Nk}, \boldsymbol{S}''^{\,i}_{Nk}$ )
9:     end for
10:    $\boldsymbol{S}''^{\,i}_{T} = \{ \boldsymbol{S}''^{\,i}_{Tj} \mid j = 1, \ldots, m_T \}$ and $\boldsymbol{S}''^{\,i}_{N} = \{ \boldsymbol{S}''^{\,i}_{Nj} \mid j = 1, \ldots, m_N \}$ => statistic $\hat{C}^i$
11: end for
12: $\{\hat{C}^i \mid i = 1, \ldots, B\} => S\hat{E}$
13: end
```

where B is the number of the two-sample bootstrap replications, i.e., the number of iterations as shown from Step 1 to 11, $\boldsymbol{S}_T$ is the set of all target sets and $\boldsymbol{S}_N$ is the set of all non-target sets, and m_T and m_N are the cardinalities of the set $\boldsymbol{S}_T$ and the set $\boldsymbol{S}_N$, respectively.

In the i-th iteration, as shown in Step 2 and Step 6, the function WR_Random_Sampling_Set is applied to the first layer of datasets, i.e., the target and non-target sets. That is, m_T target sets are randomly selected WR from the set $\boldsymbol{S}_T$ of all original target sets to form a new set $\boldsymbol{S}'^{\,i}_{T} = \{ \boldsymbol{S}'^{\,i}_{Tj} \mid j = 1, \ldots, m_T \}$, and m_N non-target sets are randomly selected WR from the set $\boldsymbol{S}_N$ of all original non-target sets to constitute a new set $\boldsymbol{S}'^{\,i}_{N} = \{ \boldsymbol{S}'^{\,i}_{Nj} \mid j = 1, \ldots, m_N \}$.

Subsequently, the same function is applied to the second layer of datasets, i.e., the similarity scores in sets as well. As shown from Step 3 to 5, m_T iterations take place after the first-layer resampling of the target sets in Step 2. In the k-th iteration, $\mu'^{\,i}_{Tk}$ target scores are randomly selected WR from the target set $\boldsymbol{S}'^{\,i}_{Tk}$, which is the k-th new target set from the first-layer resampling, to form the k-th new target set $\boldsymbol{S}''_{\mathrm{T}\,k}{}^{i}$ of the second-layer resampling. The analogous interpretation can be applied to non-target scores in the non-target set $\boldsymbol{S}'_{\mathrm{N}\,k}{}^{i}$ as shown from Step 7 to 9.

As indicated in Step 10, all target scores in the new set $\boldsymbol{S}''_{\mathrm{T}}{}^{i} = \{ \boldsymbol{S}''_{\mathrm{T}\,j}{}^{i} \mid j = 1, \ldots, m_{\mathrm{T}} \}$ and all non-target scores in the new set $\boldsymbol{S}''_{\mathrm{N}}{}^{i} = \{ \boldsymbol{S}''_{\mathrm{N}\,j}{}^{i} \mid j = 1, \ldots, m_{\mathrm{N}} \}$ are employed to calculate the estimators of the probabilities of type I and type II errors, i.e., $\hat{P}_I\,(t)$ and $\hat{P}_{II}\,(t)$ using Eq. (3) and then the i-th bootstrap replication of the estimated detection cost function at a given threshold, i.e., $\hat{\mathrm{C}}^{\mathrm{i}}$ using Eq. (4). Finally, as shown in Step 12, from the set $\{ \hat{\mathrm{C}}^{\mathrm{i}} \mid \mathrm{i} = 1, \ldots, \mathrm{B} \}$, the standard error SÊ of the detection cost function is estimated by the sample standard deviation of the B bootstrap replications.

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 corresponding algorithm can be obtained simply by removing Steps 3, 4, 5, 7, 8 and 9 from the above algorithm and modifying Step 10 accordingly.

Further, the algorithm for the bootstrap with i.i.d. assumption for similarity scores can be obtained, if Steps 2, 6 and 10 in the algorithm for the one-layer bootstrap method are replaced by "WR_Random_Sampling_Set (N_{T}, $\boldsymbol{T}$, $\boldsymbol{\Lambda}^{\mathbf{i}}$)", "WR_Random_Sampling_Set (N_{N}, $\boldsymbol{N}$, $\boldsymbol{\Xi}^{\mathbf{i}}$)", and "$\boldsymbol{\Lambda}^{\mathbf{i}}$ and $\boldsymbol{\Xi}^{\mathbf{i}}$ => statistic $\hat{\mathrm{C}}^{\mathrm{i}}$", respectively. That is, in the i-th iteration by calling the function twice, N_{T} target scores are randomly selected WR from the set $\boldsymbol{T}$ of all original target scores to form a new set $\boldsymbol{\Lambda}^{\mathbf{i}}$, N_{N} non-target scores are randomly selected WR from the set $\boldsymbol{N}$ of all original non-target scores to constitute a new set $\boldsymbol{\Xi}^{\mathbf{i}}$, and then all target and non-target scores in these two new sets $\boldsymbol{\Lambda}^{\mathbf{i}}$ and $\boldsymbol{\Xi}^{\mathbf{i}}$ are employed to generate the i-th bootstrap replication of the estimated $\hat{\mathrm{C}}^{\mathrm{i}}$.

With the new data structure described in Section 2, for all three different bootstrap resampling methods, not only does each target (non-target) score have the same probability to be selected, but also the numbers of target scores and the numbers of non-target scores resampled to generate the estimate of the statistic of interest $\hat{\mathrm{C}}^{\mathrm{i}}$ at different iterations in Step 10 are the same, respectively. This can reduce the variance of the computation.

The remaining issue is to determine how many iterations the bootstrap algorithms need to run in order to reduce the bootstrap variance and ensure the accuracy of the computation. 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 like medical decision making, etc.. Moreover, in ROC analysis our 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].

Due to the stochastic nature of the bootstrap method, distributions of SEs need to be generated in order to compare SEs obtained by using different bootstrap algorithms. All three algorithms can only create one estimated SÊ of the detection cost function at a time. However, if such an algorithm runs multiple times, it can generate a distribution of estimated SÊs. 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 three algorithms are executed 500 times each to generate a distribution $\{ \hat{SE}^{i} \mid i = 1, \ldots, 500 \}$. Thereafter, the estimated mean, SE 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. [17] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities.

5 Results

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
PB	0.098744	0.001119 0.182849 x 10^{-4} (0.001084, 0.001155)	0.004150 0.677488 x 10^{-4} (0.004012, 0.004286)	0.004288 0.675055 x 10^{-4} (0.004149, 0.004420)
CH	0.236771	0.002294 0.367097 x 10^{-4} (0.002224, 0.002367)	0.004646 0.759175 x 10^{-4} (0.004509, 0.004790)	0.005172 0.819121 x 10^{-4} (0.005020, 0.005345)

Table 1 The estimated detection cost functions, the estimated means, SÊs and 95 % CÎs of distributions of SEs of the detection cost functions generated using the three bootstrap methods for two speaker recognition systems.

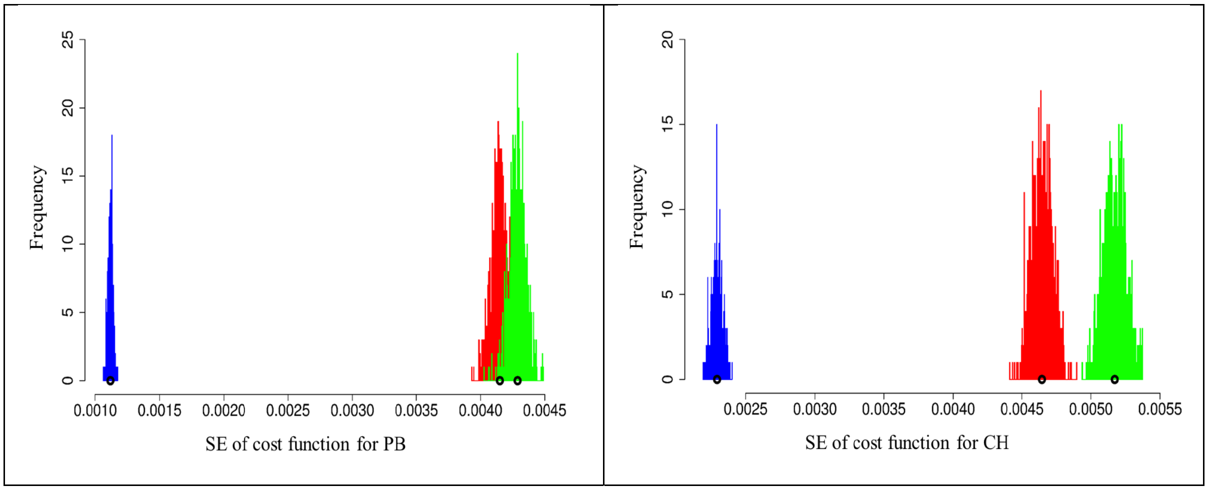

Figure 2 The histograms of SEs of the detection cost functions generated using the i.i.d. bootstrap (blue), the one-layer bootstrap (red), and the two-layer bootstrap (green), respectively, for two speaker recognition systems PB and CH. The black circle stands for the estimated mean of the distribution.

Two speaker recognition systems, labeled as PB and CH, are employed as examples. Their estimated detection cost functions, and estimated means, SÊs and 95 % CÎs of the distributions of SEs of the detection cost functions generated using the three different bootstrap methods are all shown in Table 1. System PB with a smaller value of detection cost function is more accurate than System CH, and thus has smaller uncertainty. The corresponding distributions of SEs of the detection cost functions are depicted in Figure 2, where the estimated means are represented by black circles.

The distributions of SEs shown in Table 1 and Figure 2 have two important features. The first feature is concerning the variances of the distributions of SEs. The one generated using the i.i.d. bootstrap is less than those created using the other two bootstrap methods. This feature can be seen from Table 1 and the widths of the histograms in Figure 2. Different runs of bootstrap method might produce different results of SEs. Hence, the bootstrap method with the i.i.d. assumption creates less variation of SEs than the other two bootstrap methods do on the datasets with dependency.

The second feature is regarding the positions of the distributions of SEs. It follows from Table 1 and Figure 2 that 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. The former two distributions overlap for System PB, and are separated for System CH, as shown in Figure 2. Then, what is the significant relationship between these two distributions?

The hypothesis testing is conducted on both estimated means and variances of distributions. The Shapiro-Wilk normality test [18] was conducted on the distributions of SEs generated by the three bootstrap methods for Systems PB and CH, respectively. It was observed that five p-values were between 37% and 88% which were much greater than 5%, and only one p-value was 1.7%. This suggests that the estimated $\widehat{\text{SE}}$s of the detection cost function calculated using the three different resampling methods be regarded as approximately normally distributed.

Hence, the Z-test for comparing the means and the F-test for comparing the variances can be carried out on the two distributions generated using the one-layer bootstrap and the two-layer bootstrap, respectively, for each system [3, 16, 18, 19]. It is observed in Table 1 and Figure 2 that the mean of the former is less than the mean of the latter. Hence, the one-tailed Z-test is applied. The p-values for the two systems are all close to one. This suggests that the above observation hold true significantly. Further, the p-values of the two-tailed F-test are all greater than 5%. It indicates that the null hypothesis, i.e., the ratio of the variances of these two distributions equals one, cannot be rejected.

Thus, the distribution of SEs of the cost function computed using the two-layer bootstrap for the datasets with inherent data dependencies is significantly on the right side of the distribution of SEs calculated using the one-layer bootstrap. Certainly, both of them are well separated, towards larger SEs, from the distribution of SEs computed using the bootstrap method with the i.i.d. assumption.

6 Conclusions and discussion

It is difficult to compute the SE of the measure in the SRE analytically, where the statistic of interest is a detection cost function defined as a weighted sum of the probabilities of type I and type II errors. These two probabilities are traded off each other and thus negatively correlated. It is hard to determine their correlation coefficient analytically.

To calculate the uncertainties of such a measure, the nonparametric two-sample bootstrap method based on our extensive bootstrap variability studies in ROC analysis on large datasets is employed. The premise of the bootstrap method is that the datasets must be i.i.d.. If the datasets are i.i.d., it can be employed without any modification.

In the test design, from the statistical point of view, the data should be randomly sampled to be i.i.d.. However, the datasets may contain data dependencies due to multiple uses of the same subjects in order to increase the size of datasets because of limited resources. The data dependency may be interpreted in different ways because of the complicated nature. Different interpretations of data dependencies can cause different ways of grouping similarity scores into sets to preserve the data dependency while the bootstrap resampling takes place, and subsequently will have impact on the bootstrap results.

In this article, the impact of the data dependency on the uncertainties of measures in SRE was studied. The two bootstrap methods with one-layer resampling and two-layer resampling, respectively, are proposed. In the test design, for properly employing 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. Moreover, based on such new data structure, the same numbers of target (non-target) scores can be resampled at different iterations to generate the estimator of the statistic of interest for all three different resampling methods. All these can reduce the variance of the computation.

The data dependency involved in the datasets acts like "noise" behind "signal". After performing the hypothesis testing on the distributions of SEs of the statistic of interest, it reveals that the data dependency can increase the measurement uncertainties and the variation of uncertainties as well. For the datasets with dependency, the bootstrap method with the i.i.d. assumption underestimates the uncertainties of measures. And the two-layer bootstrap estimates more conservatively the impact of the data dependency on the measurement uncertainties than the one-layer bootstrap does. Further, the bootstrap methods with one-layer resampling and two-layer resampling may produce different SEs from different runs, which differ larger than what the i.i.d. bootstrap does.

Hence, in order to reduce the sampling variability 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, then the i.i.d. assumption cannot be made and the two-layer bootstrap method rather than the one-layer bootstrap method is recommended while using the bootstrap methods to compute the measurement uncertainties.

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].

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. J.C. Wu, A.F. Martin, C.S. Greenberg and R.N. Kacker, Data dependency on measurement uncertainties in speaker recognition evaluation, NISTIR 7810, National Institute of Standards and Technology, October, (2011).
16. B. Ostle, and L.C. Malone, Statistics in Research: Basic Concepts and Techniques for Research Workers, fourth ed., Iowa State University Press, Ames, (1988).
17. R.J. Hyndman, and Y. Fan, Sample quantiles in statistical packages, American Statistician 50, 361-365 (1996).
18. R: A Language and Environment for Statistical Computing, The R Development Core Team, Version 2.8.0, 2008, at http://www.r-project.org/.
19. 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).