The Impact of Data Dependence on Speaker Recognition Evaluation

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

***Abstract*—The data dependence due to multiple use of the same subjects has impact on the standard error (SE) of the detection cost function (DCF) in speaker recognition evaluation. The DCF is defined as a weighted sum of the probabilities of type I and type II errors at a given threshold. A two-layer data structure is constructed: Target scores are grouped into target sets based on the dependence, and likewise for non-target scores. On account of the needed equal probabilities for scores being selected when resampling, target sets must contain the same number of target scores, and so must non-target sets. In addition to the bootstrap method with i.i.d. assumption, the nonparametric two-sample one-layer and two-layer bootstrap methods are carried out based on whether the resampling takes place only on sets, or subsequently on scores within the sets. Due to the stochastic nature of the bootstrap, the distributions of the SEs of the DCF estimated using the three different bootstrap methods are created and compared. After performing hypothesis testing, it is found that data dependence increases not only the SE but also the variation of the SE, and the two-layer bootstrap is more conservative than the one-layer bootstrap. The rationale regarding the different impacts of the three bootstrap methods on the estimated SEs is investigated.**

***Index Terms*—Bootstrap, data dependence, multinomial probability, resampling, speaker recognition, standard error (SE).**

I. Introduction

THE National Institute of Standards and Technology (NIST) conducts an ongoing series of Speaker Recognition Evaluations (SREs) [1]. The NIST SREs have made important contributions to the direction of research efforts and the calibration of the technical capabilities of the research community working on the general problem of text independent speaker recognition [2]–[4].

Each test in our SREs consists of a sequence of trials. Each trial consists of a model speech segment defined by the training speech data and spoken by a training model speaker, along with a test speech segment spoken by a test speaker. For each trial, the speaker recognition system generates a similarity score based on the two speech segments. A higher score indicates greater confidence that the test speaker is the training model speaker. Target (non-target) trials are those where the test speaker is (is not) the training model speaker, which is the known ground truth.

Manuscript received December 11, 2015; revised June 16, 2016 and September 8, 2016; accepted September 20, 2016. Date of publication September 30, 2016; date of current version November 28, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomi Kinnunen.

The authors are with the National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: jinchu.wu@nist.gov; alvin.martin@nist.gov; craig.greenberg@nist.gov; raghu.kacker@nist.gov).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2614725

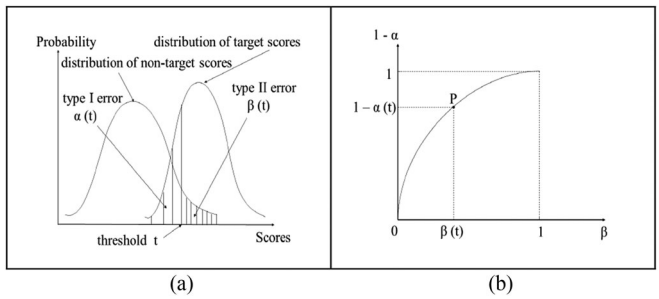


Fig. 1. (A): A schematic diagram of two continuous distributions of target and non-target scores, in which type I error α(t) and type II error β(t) are determined at a given threshold t. (B): A schematic drawing of an ROC curve and the point P (β(t), 1 - α(t)).

The speaker detection performance is measured using a detection cost function (DCF) that is defined as a weighted sum of the probabilities of type I error α(t) (miss) and type II error β(t) (false alarm) at a given threshold t [1]. This relates to ROC analysis [5]–[7]. Fig. 1 (A) shows a schematic diagram of two continuous distributions of target scores and non-target scores in which α(t) and β(t) are determined at a given threshold t. Fig. 1 (B) shows a schematic drawing of an ROC curve and the corresponding point P (β(t), 1 – α(t)).

These two error rates α(t) and β(t) are generally traded off and thus negatively correlated as the threshold t varies. An exception might occur at either end of the two score distributions, where there could be only target scores or only non-target scores causing one of the two rates to be zero. But such regions are very narrow and can hardly be identified in the DET curves as derived from our practices [4]. In addition, it is unreasonable to choose thresholds in these regions.

Notice that the correlation here refers to the relationship between α(t) and β(t). The former is determined solely by the distribution of target scores, and the latter solely by the distribution of non-target scores, regardless of how these scores are generated, which is simply an issue of how the test is designed and is not an issue of how the standard error (SE) of the DCF is estimated.

Measures must be employed in SREs [8]–[11]. However, a measure without an estimated SE is incomplete, because it cannot be used in the practice of evaluating and comparing the performance levels of different systems.

The SE of the DCF is important and may be used to classify speaker recognition systems in terms of their performance accuracies and determine whether the performance differences between systems are statistically significant when evaluating and comparing systems. How to effectively estimate the SE of the DCF for the SREs is a void that needs to be filled.

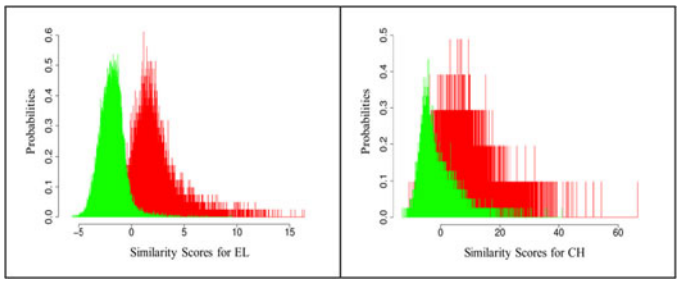


Fig. 2. Distributions of target scores (red) and non-target scores (green) for speaker recognition systems EL (left) and CH (right).

The SE of the DCF was calculated analytically [3], [12]. But it is difficult to do so due to the covariance between α(t) and β(t). Moreover, the analytical method does not take account of how target scores and non-target scores are distributed, which is related to the recognition abilities of speaker verification systems, nor does it take into account the data dependency. All these may cause the analytical method to underestimate the SE of the measure (see Section VIII) [13].

Nevertheless, an upper bound of the analytically computed SE of the DCF can be obtained by setting the negative covariance to zero. This will hold even if the thresholds were at the ends of the two score distributions as described above. The only difference is that if one error rate is equal to zero, the analytical SE might reach its upper bound, depending on how the covariance is defined (the Pearson's correlation coefficient is undefined in this case); and if both error rates are nonzero, the analytical SE of the DCF is definitely smaller than its upper bound because of the negative covariance.

The bootstrap method takes account of all these issues intrinsically and thus estimates the SE of the DCF effectively [5], [13]. For estimating the SE, the bootstrap algorithm was originally designed as follows [14], [15]. If n data values are independent and identically distributed (i.i.d.), the bootstrap algorithm selects randomly with replacement (WR) n data values from these n original data values, and computes a bootstrap replication of the statistic of interest using the n selected data values. Such an iteration is repeated B times. Finally, the SE of the statistic is estimated by the sample standard deviation of the B bootstrap replications of the statistic.

The ROC analysis in SRE generally involves N_T target scores and N_N non-target scores, which characterize the speaker recognition system that generates them and usually do not have well defined parametric distributions as shown in Fig. 2 for systems EL and CH (see Section VIII) with long tails towards higher scores [5], [16]. The parametric bootstrap method is based on a mathematical model [14], [15]. Hence, it cannot be used.

If all scores are i.i.d., the nonparametric two-sample bootstrap algorithm is used to compute the SE of the DCF based on our extensive bootstrap variability studies in ROC analysis on large datasets [5]–[7], [14], [15]. The two samples involved are referred to as a set of target scores and a set of non-target scores. The algorithm selects randomly WR N_T target scores and N_N non-target scores from the two original sets of scores, respectively, and then computes a bootstrap replication of the DCF using these two new sets of scores. After taking B iterations, the SE of the DCF is computed using the sample standard deviation of the B bootstrap replications of the DCF.

However, in reality, the data do contain dependencies due to various reasons, such as the time factor, etc. [15], [17]. It is very often that the data dependency basically arises from multiple use of the same subjects in order to create more target and non-target scores because of limited resources. For instance, a single speaker appears several times within an SRE corpus. The calls from a single speaker are not independent.

Therefore, 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 generated using the same training speaker id number are grouped into a target set, and those non-target scores created using the same training speaker id number are grouped into a non-target set. This can preserve the data dependency while the bootstrap resampling takes place. Thus, a two-layer data structure is constructed: The first layer consists of target sets and non-target sets, and the second layer consists of target scores and non-target scores within these sets.

Based on this two-layer data structure, in addition to a conventional bootstrap method, the nonparametric two-sample one-layer and two-layer bootstrap methods are carried out, depending on whether the resampling takes place randomly WR only on the first layer of the data while the bootstrap units are sets, or subsequently on the second layer while the bootstrap units are the scores within a set, where the similarity scores are assumed to be conditionally independent. This is because scores in the same set are generated by speech segments in which either id numbers of model speech segments or id numbers of test speech segments or both of them are different.

Different sets could have different numbers of similarity scores. This would result in each target (non-target) score not having the same probability of being selected when using the two-layer bootstrap method. It would also mean that the numbers of scores obtained would be different from iteration to iteration when using the one-layer and two-layer bootstrap methods. To avoid all these, the datasets are adjusted in such a way that all target sets contain the same number of target scores, and likewise for the non-target sets. As a result, the three resampling methods are placed on an equal footing, and the variance of the computations can be reduced as well [18].

The SEs of the DCF and the variations of the SEs generated by the three bootstrap methods are compared. Due to the stochastic nature of bootstrap methods, different runs may produce slightly different estimated SEs. Some results may be more probable and others less probable. Thus, a probability distribution of SEs is created. Hence, the comparisons of SEs turn out to be the comparisons of the distributions of SEs.

After performing the hypothesis testing, it is found that the data dependency increases the SE of the DCF and the variation of the SE as well; and the two-layer bootstrap more conservatively estimates the impact of the data dependency on the SE of the

DCF than other two bootstrap methods. So, the impact of data dependency should be taken into account when designing a test and estimating the SE of the DCF thereafter.

The validation of the nonparametric two-sample bootstrap in ROC analysis on large datasets was studied in Ref. [7]. In this article, large datasets are also assumed [19], [20]. Even if the thresholds were at either end of the two score distributions, as mentioned above, this would have no impact on how the bootstrap algorithm is employed to estimate the SE of the DCF. Moreover, the bootstrap algorithm only resamples target scores and non-target scores separately. It is unrelated to how scores are generated, which is out of the scope of this article.

The bootstrap method on datasets with dependencies was initially studied in Refs. [15], [17], and applied to biometrics later [21], [22]. In this article, however, five important issues are addressed and investigated. First, speaker recognition is taken as the application, in which the statistic of interest is the DCF, defined as a weighted sum of the probabilities of type I and type II errors at a given threshold, which is rather more complicated than the usual measures in biometrics. Second, the nonparametric two-sample, rather than one-sample, bootstrap is employed. The two-sample bootstrap may be used to compute the SEs of all different measures in ROC analysis, as opposed to the one-sample bootstrap that can only be used to calculate the SEs of true acceptance rate and false acceptance rate in biometrics, which are not the statistics of interest in SREs [5]. The SE of the DCF can only be computed using the two-sample bootstrap, not the one-sample bootstrap.

Third, the related probability issues due to the bootstrap resampling nature are taken into account. The nonparametric bootstrap method requires that the objects have equal probabilities of being selected in the random resampling. As a result, the numbers of scores in the target sets and in the non-target sets should be equal, respectively. Fourth, the bootstrap method has a stochastic nature, i.e., different runs can produce different results. Thus, the distributions of the SEs of the DCF estimated using different bootstrap approaches should be created, and then compared by conducting the hypothesis testing to reveal the impact of data dependency on the SRE. The conclusions cannot be obtained based solely on the results from a single random execution of the bootstrap.

Fifth, the rationale for why the three different bootstrap methods have different impacts on the estimated SEs of the DCF is investigated in terms of the multinomial probabilities of selecting bootstrap samples from the original scores as well as the distributions of the bootstrap replications of the DCF.

The issue of data dependency was preliminarily studied in our previous work [23]. Much more comprehensive research is presented in this article. For instance, how the SRE relates to ROC analysis is explicitly shown. The results of evaluating 12 rather than just two speaker recognition systems are presented. Most importantly, it is only in this article that the rationale is investigated for why the SEs computed using three bootstrap methods are different where data dependency is involved.

The DCF is shown in Section II. The analytical approach is presented in Section III. The two-layer data structure is introduced in Section IV. The three resampling methods and the probabilities for a score being selected are explored in Section V. The three nonparametric two-sample bootstrap algorithms are presented in Section VI. A method of generating distributions of the SEs of the DCF is provided in Section VII. The results of 12 speaker recognition systems[1,2] used in SREs are shown in Section VIII. The conclusions and discussion are in Section IX. The proof of the probability for a score being selected for one-layer resampling is presented in Appendix I. The rationale for the different impacts of the three bootstrap methods on the SEs is investigated in Appendix II.

II. The DCF in SREs

In SREs, for the convenience of computing cumulative probabilities for type I error and type II error, for a speaker recognition system, the scores are all converted to integer values if they are not, and expressed inclusively using the set $\mathbf{s} = \{\mathrm{s}_{\min}, \mathrm{s}_{\min}+1, \ldots, \mathrm{s}_{\max}\}$. While converting, as many decimal places of scores as possible are kept. Thus, such a conversion does not result in loss of precision.

Let $\mathrm{f}_i(\mathrm{s})$, $\mathrm{s} \in \mathbf{s}$ and $\mathrm{i} \in \{\mathrm{T}, \mathrm{N}\}$, denote the continuous probability density functions of target scores, and non-target scores. The two corresponding discrete probability distribution functions, denoted by $\mathrm{P}_i(\mathrm{s})$, $\mathrm{s} \in \mathbf{s}$ and $i \in \{\mathrm{T}, \mathrm{N}\}$, are expressed as

$$\boldsymbol{P}_{\mathbf{i}} = \{\mathrm{P}_i(\mathrm{s}) \,|\forall \mathrm{s} \in \mathbf{s} \text{ and } \sum_{\mathrm{s}=\mathrm{Smin}}^{\mathrm{Smax}} \mathrm{P}_i(\mathrm{s}) = 1\}, i \in \{\mathrm{T}, \mathrm{N}\}. \tag{1}$$

The probability of type I error at a given threshold $t \in \mathbf{s}$ for target scores, denoted by $\alpha(t)$, is cumulated from the lowest score $\mathrm{s}_{\min}$. The probability of type II error at t for non-target scores, denoted by $\beta(t)$, is cumulated from the highest score $\mathrm{s}_{\max}$. For discrete probability distributions, when computing $\alpha(t)$ and $\beta(t)$ at t, the probabilities of the scores at this threshold t should be included [24].

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

$$\alpha(t) = \int_{-\infty}^{t} \mathrm{f}_T(\mathrm{s})\mathrm{ds} = \sum_{\mathrm{s}=\mathrm{Smin}}^{t} \mathrm{P}_T(\mathrm{s}) = 1 - \sum_{\mathrm{s}=t+1}^{\mathrm{Smax}} \mathrm{P}_T(\mathrm{s}),$$

$$\beta(t) = \int_{t}^{+\infty} \mathrm{f}_N(\mathrm{s})\mathrm{ds} = \sum_{\mathrm{s}=t}^{\mathrm{Smax}} \mathrm{P}_N(\mathrm{s}), \tag{2}$$

where $\mathrm{P}_{\mathrm{T}}(\mathrm{s}_{\max}+1) = 0$ is assumed and the normalization in Eq. (1) is employed [5], [16]. In practice, these error rates

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

can be obtained by moving the score from the highest score $\mathrm{s}_{\max}$ down to the threshold t one score at a time to cumulate the probabilities of target scores and of non-target scores.

In the SREs, the metric of interest is the DCF defined as a weighted sum of the probabilities of type I and type II errors at a given threshold t [1]

$$CF_{\mathrm{Det}}\ (t) = \mathrm{C_{Miss}} \times \alpha\,(t) \times \mathrm{P_{Target}} + \mathrm{C_{FalseAlarm}} \times \beta\,(t) \times (1 - \mathrm{P_{Target}})\,. \quad (3)$$

The threshold t plays an important role here, which may be determined in many ways. It is a challenging research problem to determine appropriate decision thresholds. 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 [1].

The parameters $\mathrm{C_{Miss}}$ and $\mathrm{C_{FalseAlarm}}$ are the relative costs of detection errors, and the parameter $\mathrm{P_{Target}}$ 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 $\mathrm{C_{Miss}}$, $\mathrm{C_{FalseAlarm}}$, and $\mathrm{P_{Target}}$ were set to be 10, 1, and 0.01, respectively [1].

How to design the DCF, how to choose the threshold, how to set these parameters, and how to generate target scores and non-target scores are all out of the scope of this article. These issues have no impact on how to estimate the SE of the DCF using the bootstrap algorithms described in this article and on how to compute the upper bound of the analytical SE, as pointed out in Section I and described in Section III.

III. The Analytical Approach

Based on Eq. (3), the analytically estimated SE of the DCF is usually expressed as

$$\mathrm{SE(t)}^2{}_{\mathrm{det}} = a^2\mathrm{SE}^2{}_{\alpha(t)} + b^2\mathrm{SE}^2{}_{\beta(t)} + 2ab\mathrm{Cov}\left(\alpha\left(t\right), \beta\left(t\right)\right) \quad (4)$$

where $a = \mathrm{C_{Miss}}\,\mathrm{P_{Target}}$, $b = \mathrm{C_{FalseAlarm}}\,(1 - \mathrm{P_{target}})$, and Cov($\alpha(t)$, $\beta(t)$) is the covariance between α(t) and β(t) [3], [25]. $\mathrm{SE}_{\alpha(\mathrm{t})}$ and $\mathrm{SE}_{\beta(\mathrm{t})}$ may be estimated using SE = sqrt [p (1 – p) / n], but the drawback of this is discussed in Ref. [26].

It is difficult to estimate this covariance. However, as pointed out in Section I, α(t) and β(t) are generally traded off and thus negatively correlated as the threshold t varies. Hence, Cov($\alpha(t)$, $\beta(t)$) is negative [25]. As a result, the upper bound of the analytically computed $\mathrm{SE}(t)_{\mathrm{det}}$ can be obtained by setting the negative Cov($\alpha(t)$, $\beta(t)$) in Eq. (4) to zero [13]. It is noted that the analytically computed $\mathrm{SE}(t)_{\mathrm{det}}$ must be smaller than its upper bound. Due to the stochastic nature of the bootstrap method, such upper bounds will be compared with the 95% CI of the bootstrap estimated SEs in Sections VIII-B and VIII-C.

IV. The Two-Layer Data Structure

The target scores and non-target scores are grouped into target sets and non-target sets, respectively, based on the data dependency. Thus the data structure has two layers: the first layer consists of target sets and non-target sets, and the second layer consists of target scores and non-target scores within these sets. In the following, let $\boldsymbol{S}$ denote score sets, α similarity scores, and μ the number of scores in a set. The first subscript stands for whether it is referred to as target (T) or non-target (N), the second the ordinal number of sets, and the third the ordinal number of scores in a set. In Sections IV and V, scores are denoted by α.

Suppose that there are m_T target sets and m_N non-target sets. Thus, the set $\boldsymbol{S}_{\mathrm{T}}$ of all target sets and the set $\boldsymbol{S}_{\mathrm{N}}$ of all non-target sets are expressed by

$$\boldsymbol{S}_i = \{\boldsymbol{S}_{\mathrm{ij}} | j = 1, \ldots, \mathrm{m}_i\}\,, \mathrm{i} \in \{\mathrm{T}, \mathrm{N}\}\,, \quad (5)$$

where $\boldsymbol{S}_{\mathrm{Tj}}$ are target sets and $\boldsymbol{S}_{\mathrm{Nj}}$ are non-target sets. Each set is expressed in terms of scores by

$$\boldsymbol{S}_{\mathrm{ij}} = \{\alpha_{ijk} | \,\mathrm{k} = 1, \ldots, \mu_{\mathrm{ij}}\},\ \ \mathrm{j} = 1, \ldots \mathrm{m}_i \text{ and } i \in \{\mathrm{T}, \mathrm{N}\}, \quad (6)$$

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.

In other words, the m_T target sets $\boldsymbol{S}_{T1}, \boldsymbol{S}_{T2}, \ldots, \boldsymbol{S}_{T\mathrm{m}_T}$ contain $\mu_{T1}, \mu_{T2}, \ldots, \mu_{T\mathrm{m}_T}$ target scores $\{\alpha_{T11}, \alpha_{T12}, \ldots, \alpha_{T1\mu_{T1}}\}$, $\{\alpha_{T21}, \alpha_{T22}, \ldots, \alpha_{T2\mu_{T2}}\}, \ldots,$ $\{\alpha_{T\mathrm{m}_T1}, \alpha_{T\mathrm{m}_T2}, \ldots, \alpha_{T\mathrm{m}_T\mu_{T\mathrm{m}_T}}\}$, respectively; and the m_N non-target sets $\boldsymbol{S}_{N1}, \boldsymbol{S}_{N2}, \ldots, \boldsymbol{S}_{N\mathrm{m}_N}$ have $\mu_{N1}, \mu_{N2}, \ldots, \mu_{N\mathrm{m}_N}$ non-target scores $\{\alpha_{N11}, \alpha_{N12}, \ldots, \alpha_{N1\mu_{N1}}\}$, $\{\alpha_{N21}, \alpha_{N22}, \ldots, \alpha_{N2\mu_{N2}}\}$, $\ldots$, $\{\alpha_{N\mathrm{m}_N1}, \alpha_{N\mathrm{m}_N2}, \ldots, \alpha_{N\mathrm{m}_N\mu_{N\mathrm{m}_N}}\}$, respectively.

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

$$\boldsymbol{T} = \{\alpha_{\mathrm{Tjk}} | \mathrm{k} = 1, \ldots, \mu_{\mathrm{Tj}} \text{ and } \mathrm{j} = 1, \ldots, \mathrm{m}_T\}, \quad (7)$$

and

$$\boldsymbol{N} = \{\alpha_{\mathrm{Njk}} | \mathrm{k} = 1, \ldots, \mu_{\mathrm{Nj}} \text{ and } \mathrm{j} = 1, \ldots, \mathrm{m}_N\}. \quad (8)$$

The sets $\boldsymbol{S}_{ij}$, $\boldsymbol{T}$, and $\boldsymbol{N}$ are all viewed in the sense of a multiset, in which members are allowed to appear more than once. All similarity scores are treated as separate objects because they were generated by different trials in the test, even though some of them may have common values. The empirical distribution is assumed for each of the observed scores [15]. That is, the probability of each score is assigned to be the reciprocal of the total number of observed target or non-target scores.

Finally, the total numbers of target scores and non-target scores, i.e., $\mathrm{N_T}$ and $\mathrm{N_N}$, satisfy

$$\mathrm{N}_i = \sum_{\mathrm{j}=1}^{\mathrm{m}_i} \mu_{\mathrm{ij}}, \ \ \text{where i} \in \{\mathrm{T},\ \mathrm{N}\}\,. \quad (9)$$

V. The Three Resampling Methods and the Probabilities for a Score Being Selected

A. Resampling With the i.i.d. Assumption

The first method is that the resampling takes place with the assumption that the data are i.i.d. Then, the resampling units are all of the similarity scores. The probability for a score being selected with respect to the total number of scores selected is

$1/N_T$ equally for each target score and $1/N_N$ equally for each non-target score.

B. One-Layer Resampling

The second method is the one-layer resampling that takes place randomly WR only on the first layer of the data, i.e., target sets and non-target sets, whereas the data dependency is preserved while resampling [15], [17]. Then, the resampling units are all score sets. Hence, if a set is selected, then all scores within the set are selected. It is readily seen that the probability of selecting a score in regard to the total number of selections approaches $1/m_i$, $i \in \{T, N\}$.

However, if the probability of selecting a score is defined with respect to the total number of scores selected, then it follows from the Law of Large Numbers that the probability for a score α_{ijk} being selected is

$$P_{1-\text{layer}}(\alpha_{ijk}) = \frac{1}{N_i}, \quad k = 1, \ldots, \mu_{ij}, j = 1, \ldots, m_i \text{ and } i \in \{T, N\}. \quad (10)$$

The proof of Eq. (10) is presented in Appendix I.

In estimating the SE of the DCF using the bootstrap method, what matters is all scores that have been chosen by the one-layer resampling. In addition, the probability definition leading to Eq. (10) is consistent with the probability definition for the other two resampling methods. Hence, such a definition is adopted in this article. So, the probabilities for each target score and each non-target score being selected are $1/N_T$ and $1/N_N$, respectively.

C. Two-Layer Resampling

The third method is the two-layer resampling. The selection takes place randomly WR not only at the first layer of the data but also at the second layer of the data, i.e., target scores and non-target scores in the sets, which are assumed to be conditionally independent as discussed in Section I. The resampling units for the first layer are sets and for the second layer are scores in the sets.

The probability for a score α_{ijk} in set $\boldsymbol{S}_{ij}$ being selected is

$$P_{2-\text{layer}}(\alpha_{ijk}) = P(\boldsymbol{S}_{ij}) \times P(\alpha_{ijk} \mid \boldsymbol{S}_{ij}) = \frac{1}{m_i} \times \frac{1}{\mu_{ij}}, \quad (11)$$
$$k = 1, \ldots, \mu_{ij}, j = 1, \ldots, m_i \text{ and } i \in \{T, N\}.$$

This is with respect to the total number of scores selected. These probabilities are the same for all scores within a set, but different from set to set due to different numbers of scores in different sets as indicated by μ_{ij} for both target and non-target. This is different from the previous two resampling methods.

D. The Requirement of Datasets Based on the Probability for a Score Being Selected

The nonparametric bootstrap method demands that the objects have equal probabilities to be selected in the random resampling [15]. Thus, it is not appropriate that target scores are selected with unequal probabilities in the random resampling for the two-layer bootstrap, and the same holds for non-target scores. The impact of varied numbers of scores in sets on the probabilities for a score being selected should be eliminated.

In the two-layer resampling, if the numbers of scores in the target sets, i.e., μ_{Tj}, $j = 1, \ldots, m_T$, are all set to be equal to μ_T, then the probability for each target score being selected will be $1/N_T$ due to Eq. (9). Analogously, it will be $1/N_N$ for each non-target score, if all μ_{Nj}, $j = 1, \ldots, m_N$, are set to be equal to μ_N. Thus, the probabilities for each target score and each non-target score being selected in the two-layer resampling will be the same as those in the other two resampling methods.

In addition, this dataset requirement can ensure that the same numbers of target scores and of non-target scores are obtained at different iterations using the three different bootstrap methods (see Sections VI-C and VI-D). Such a structure of datasets can reduce the variance of the computation [18]. Further, the SEs of the DCFs calculated using the three resampling methods can be compared on an equal footing.

Our datasets had 132 target sets (130 non-target sets), each of which contained 96 target scores (244 non-target scores). Thus, the total number of target (non-target) scores was 12,672 (31,720). Thus, there are tens of thousands of scores in our datasets [20].

1: **function** WR_Random_Sampling_Set (N, $\boldsymbol{\Gamma}$, $\boldsymbol{\Theta}$)
2: **for** $i = 1$ **to** N **do**
3: select randomly WR an index $j \in \{1, \ldots, N\}$
4: $\theta_i = \gamma_j$
5: **end for**
6: **end function**

VI. The Three Nonparametric Two-Sample Bootstrap Algorithms

The estimate of the SE of the DCF at a threshold t is computed using the three different nonparametric two-sample bootstrap resampling methods based on our extensive studies of bootstrap variability in ROC analysis on large datasets [5]–[7], [14], [15]. From here on, the superscript indices are used for the numeration of the resampling iterations.

A. A Function WR_Random_Sampling_Set

First of all, a function WR_Random_Sampling_Set is shown above, which will be frequently employed in the following algorithms. In this function, $\boldsymbol{\Gamma}$ stands for a set of sets or a set of scores, N is the cardinality of the set $\boldsymbol{\Gamma}$, $\boldsymbol{\Theta}$ represents a new set of sets or a new set of scores accordingly with the same cardinality, and γ_j and θ_i are members of the sets $\boldsymbol{\Gamma}$ and $\boldsymbol{\Theta}$, 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 $\boldsymbol{\Gamma}$ is randomly selected WR to become a member of a new set $\boldsymbol{\Theta}$, as indicated in Steps 3 and 4. As a result, N members (sets or scores) are randomly selected WR from the set $\boldsymbol{\Gamma}$ to form a new set $\boldsymbol{\Theta}$.

Algorithm I (two-sample bootstrap with i.i.d. assumption)

1: **for** i = 1 **to** B **do**
2: WR_Random_Sampling_Set (N_T, $\boldsymbol{T}$, $\boldsymbol{\Theta}^i$)
3: WR_Random_Sampling_Set (N_N, $\boldsymbol{N}$, $\boldsymbol{\Xi}^i$)
4: $\boldsymbol{\Theta}^i$ and $\boldsymbol{\Xi}^i$ => statistic $\widehat{\mathrm{CF}}^i$
5: **end for**
6: $\{\widehat{\mathrm{CF}}^i \mid i=1,\ldots,B\}$=> $\hat{\mathrm{SE}}$ and $(\hat{Q}(\alpha/2), \hat{Q}(1-\alpha/2))$
7: **end**

Algorithm II (two-sample one-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: WR_Random_Sampling_Set
$(m_N, \boldsymbol{S}_N, \boldsymbol{S}'^{\,i}_N = \{\boldsymbol{S}'^{\,i}_{Nj} \mid j = 1, \ldots, m_N\})$
4: $\boldsymbol{S}'^{\,i}_T$ and $\boldsymbol{S}'^{\,i}_N$ => statistic $\widehat{\mathrm{CF}}^i$
5: **end for**
6: $\{\widehat{\mathrm{CF}}^i \mid i = 1, \ldots, B\}$ => $\hat{\mathrm{SE}}$ and $(\hat{Q}(\alpha/2), \hat{Q}(1-\alpha/2))$
7: **end**

B. An Algorithm With the i.i.d. Assumption

With the i.i.d. assumption for the data, the resampling units are scores. Thus the nonparametric two-sample bootstrap algorithm with the i.i.d. assumption—Algorithm I is shown above. In this algorithm, B is the number of bootstrap replications, $\boldsymbol{T}$ is the set of all N_T original target scores and $\boldsymbol{N}$ is the set of all N_N original non-target scores as shown in Eqs. (7) and (8).

This algorithm runs B times. As shown from Step 1 to 5, in the i-th iteration by calling the function in Section VI-A twice, N_T target scores are randomly selected WR from the set $\boldsymbol{T}$ to form a new set $\boldsymbol{\Theta}^i$, N_N non-target scores are randomly selected WR from the set $\boldsymbol{N}$ to constitute a new set $\boldsymbol{\Xi}^i$, and then all target and non-target scores in these two new sets $\boldsymbol{\Theta}^i$ and $\boldsymbol{\Xi}^i$ are employed to generate the i-th bootstrap replication of the DCF at a given threshold, $\widehat{\mathrm{CF}}^i$, using Eqs. (2) and (3).

Finally, as indicated in Step 6, from the set $\{\widehat{\mathrm{CF}}^i \mid i = 1, \ldots, B\}$, the standard error $\hat{\mathrm{SE}}$ of the DCF is estimated by the sample standard deviation of the B replications, and the (1 - α) 100% confidence interval (CI) $(\hat{Q}(\alpha/2), \hat{Q}(1-\alpha/2))$ at the significance level α is estimated by the α/2 100% and (1 - α/2) 100% quantiles of the bootstrap distribution [15]. Definition 2 of quantile in Ref. [27] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities [28]. If the 95% $\hat{\mathrm{CI}}$ is of interest, then α is set to be 0.05.

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 computation accuracy for our applications. That is, what is an appropriate number B for the nonparametric two-sample bootstrap replications? In our applications, such as biometrics and speaker recognition, the sizes of datasets are tens or hundreds of thousands of scores, which are much larger than those in some other applications of bootstrap methods, such as medical decision making. In ROC analysis, the statistics of interest are mostly probabilities or a weighted sum of probabilities rather than a simple sample mean. And our data samples of scores have no parametric model to fit. Thus, the bootstrap variability was re-studied empirically. The appropriate number of bootstrap replications B for our applications was determined to be 2,000 [5]–[7].

C. An Algorithm for the Nonparametric Two-Sample One-Layer Bootstrap

As discussed in Section V, the one-layer resampling takes place only on the first layer of the new data structure, namely, the resampling units are target sets and non-target sets, respectively. Thus, the nonparametric two-sample one-layer bootstrap algorithm – Algorithm II is shown above. In this algorithm, B is the number of bootstrap replications, the set $\boldsymbol{S}_T$ of all target sets and the set $\boldsymbol{S}_N$ of all non-target sets are expressed in Eq. (5), and m_T and m_N are the cardinalities of the sets $\boldsymbol{S}_T$ and $\boldsymbol{S}_N$, respectively.

In the i-th iteration, as indicated in Step 2 and Step 3, the function in Section VI-A is applied twice to sets rather than scores. That is, m_T target sets are randomly selected WR from the set $\boldsymbol{S}_T$ to constitute 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 $\boldsymbol{S}_N$ to form a new set $\boldsymbol{S}'^{\,i}_N = \{\boldsymbol{S}'^{\,i}_{Nj} \mid j = 1, \ldots, m_N\}$. As noted in Step 4, all target scores in the new set $\boldsymbol{S}'^{\,i}_T$ and all non-target scores in the new set $\boldsymbol{S}'^{\,i}_N$ are employed to generate the i-th bootstrap replication of the estimated DCF $\widehat{\mathrm{CF}}^i$. Everything else here is the same as in the algorithm shown in Section VI-B.

With the data structure shown in Section V-D, the same numbers of target scores and the same numbers of non-target scores are always obtained in Step 4 to compute the estimate of the statistic of interest at different iterations of the nonparametric two-sample one-layer bootstrap. This can reduce the variance of the computation [18].

D. An Algorithm for the Nonparametric Two-Sample Two-Layer Bootstrap

As described in Section V, the two-layer resampling is carried out not only on the first layer of the new data structure where the resampling units are target sets and non-target sets, but also on the second layer of the data where the resampling units are target scores and non-target scores in the sets. Hence, the nonparametric two-sample two-layer bootstrap algorithm—Algorithm III is presented. In this algorithm, B is the number of bootstrap replications, the set $\boldsymbol{S}_T$ of all target sets and the set $\boldsymbol{S}_N$ of all non-target sets are in Eq. (5), and m_T and m_N are the cardinalities of the sets $\boldsymbol{S}_T$ and $\boldsymbol{S}_N$. In the i-th iteration, as shown in Step 2 and Step 6, the function in Section VI-A is applied to the first layer of datasets twice, which is the same as in the one-layer bootstrap Algorithm II in Section VI-C.

Subsequently, the same function is applied to the second layer of datasets, i.e., to the similarity scores in the sets as well. As shown from Step 3 to 5, m_T iterations take place after the

Algorithm III (two-sample 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 ($\mu_T, \boldsymbol{S}'^{\,i}_{Tk}, \boldsymbol{S}''^{\,i}_{Tk}$)
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 ($\mu_N, \boldsymbol{S}'^{\,i}_{Nk}, \boldsymbol{S}''^{\,i}_{Nk}$)
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 $\widehat{CF}^i$
11: **end for**
12: $\{\widehat{CF}^i \mid i = 1, \ldots, B\} =>$ $S\hat{E}$ and $(\hat{Q}(\alpha/2), \hat{Q}(1-\alpha/2))$
13: **end**

first-layer resampling of the target sets in Step 2. In the k-th iteration, μ_T 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 $\boldsymbol{S}''^{\,i}_{Tk}$ of the second-layer resampling. The analogous interpretation can be applied to non-target scores in the non-target set $\boldsymbol{S}'^{\,i}_{Nk}$ as shown from Step 7 to 9.

As shown in Step 10, all target scores in the new set $\boldsymbol{S}''^{\,i}_T = \{\boldsymbol{S}''^{\,i}_{Tj} \mid j = 1, \ldots, m_T\}$ and all non-target scores in the new set $\boldsymbol{S}''^{\,i}_N = \{\boldsymbol{S}''^{\,i}_{Nj} \mid j = 1, \ldots, m_N\}$ are employed to generate the i-th bootstrap replication of the estimated DCF $\widehat{CF}^i$. Everything else in this algorithm is the same as in the algorithm shown in Section VI-B.

With the data structure described in Section V-D, not only does each target (non-target) score have the same probability to be selected, but also the same numbers of target scores and of non-target scores are obtained in Step 10 to estimate the DCF at different iterations of the nonparametric two-sample two-layer bootstrap. These can reduce the variance of computation [18].

VII. A Method of Generating Distributions of SEs of the Statistic of Interest

If the SEs obtained by using different bootstrap algorithms need to be compared, using just one estimate of SE is far from enough due to the stochastic nature of the bootstrap method. Hence, a distribution of the $S\hat{E}$s of the DCF, which are each estimated to be the sample standard deviation of 2,000 bootstrap replications of the DCF, needs to be investigated.

All three algorithms in Section VI create one estimated $S\hat{E}$ of the DCF at a time. Running an algorithm multiple times can generate a distribution of estimated $S\hat{E}$s of DCF. Based on our studies, to create a stable distribution, it is enough that the algorithm be executed 500 times [5]–[7].

Hence, the algorithms shown in Sections VI-B, VI-C, and VI-D are executed 500 times each to create a distribution $\{S\hat{E}^{\,i} \mid i = 1, \ldots, 500\}$. Then, the estimated mean, SE and 95% CI of such a distribution are calculated, where Definition 2 of quantile in Ref. [27] is adopted as shown in Section VI-B.

VIII. Results

Twelve speaker recognition systems were tested. It is time consuming to generate a distribution of estimated $S\hat{E}$s of the DCF as described in Section VII. To make the presentation clear, the four systems EL, LZ, PB, and CH were arbitrarily chosen from the 12 systems. However, these four systems do represent a range of performance levels (see the DCFs shown in Tables I and III below). The results of the four systems are described in detail. And the results of the other eight systems UJ, BK, DL, AF, FI, PM, CO, and DG are briefly discussed. The distributions of the $S\hat{E}$s computed using the three different bootstrap approaches are explored.

A. The $S\hat{E}$ and $C\hat{I}$ of the DCF for the Four Systems

The estimated DCFs, and the estimated $S\hat{E}$s (relative error) and 95% $C\hat{I}$s of the DCFs for the four systems computed using the three different bootstrap approaches are shown in Table I. They have different matching accuracies—the smaller the DCF, the more accurate the speaker recognition system. The relative errors defined by (1.96 x $S\hat{E}$ / DCF) show the significance of the estimated $S\hat{E}$s with respect to the DCFs.

The estimated 95% $C\hat{I}$s shown in Table I were all calculated using the quantile method as described in Section VI-B. They can also be computed by multiplying 1.96 by the estimated $S\hat{E}$, assuming that the distribution of 2000 bootstrap replications of the DCF is normal.

These two types of 95% $C\hat{I}$s are matched up to the third or fourth decimal place in all cases. For instance, for System EL using the two-layer resampling, the 95% $C\hat{I}$ derived from the quantile method is (0.018384, 0.026084) as shown in Table I, while it is (0.018374, 0.026024) based on the normality assumption.

In addition, the Shapiro-Wilk normality test [28] was conducted on the distributions of the DCFs. As many as seven p-values were between 7% and 68%. Two p-values were about 1%, and three were less than 1%. All these indicate that the DCF may be regarded as approximately normally distributed.

B. The Distributions of the $S\hat{E}$s Estimated Using the Three Different Bootstrap Algorithms and the Analytically Computed $S\hat{E}$s for the Four Systems

The estimated means, $S\hat{E}$s (relative error) and 95% $C\hat{I}$s of the distributions of SEs of the DCFs for the four systems estimated using the i.i.d. bootstrap, the one-layer bootstrap, and the two-layer bootstrap, respectively, are all presented in Table II. The relative errors defined by (1.96 x $S\hat{E}$ / mean) show the significance of the estimated $S\hat{E}$s with respect to the means. The corresponding distributions of the SEs of the DCFs along with the estimated means represented by black circles are depicted in Fig. 3.

TABLE I
THE ESTIMATED DCFS, SÊS (RELATIVE ERROR) AND 95% CÎS OF THE DCFS COMPUTED USING THE I.I.D. BOOTSTRAP, THE ONE-LAYER BOOTSTRAP, AND THE TWO-LAYER BOOTSTRAP FOR FOUR SPEAKER RECOGNITION SYSTEMS LABELED AS EL, LZ, PB, AND CH

System	DCF	SÊ (relative error) and 95% CÎ of DCF		
		i.i.d. Bootstrap	One-Layer Bootstrap	Two-Layer Bootstrap
EL	0.022199	0.000696 (6.15%) (0.020855, 0.023575)	0.001909 (16.85%) (0.018549, 0.026151)	0.001952 (17.23%) (0.018384, 0.026084)
LZ	0.040098	0.000894 (4.37%) (0.038227, 0.041792)	0.002810 (13.74%) (0.034972, 0.045794)	0.002897 (14.16%) (0.034641, 0.045880)
PB	0.098744	0.001115 (2.21%) (0.096640, 0.101104)	0.004226 (8.39%) (0.090678, 0.107155)	0.004301 (8.54%) (0.090357, 0.107294)
CH	0.236771	0.002318 (1.92%) (0.232121, 0.240992)	0.004669 (3.87%) (0.227487, 0.246115)	0.005092 (4.22%) (0.226647, 0.247300)

TABLE II
THE UPPER BOUND OF THE ANALYTICAL SÊS, AND THE ESTIMATED MEANS, SÊS (RELATIVE ERROR) AND 95% CÎS OF DISTRIBUTIONS OF SES OF THE DCFS COMPUTED USING THE I.I.D. BOOTSTRAP, THE ONE-LAYER BOOTSTRAP, AND THE TWO-LAYER BOOTSTRAP FOR FOUR SPEAKER RECOGNITION SYSTEMS EL, LZ, PB, AND CH

System	Upper bound of the analytical SÊ	Mean, SÊ (relative error) and 95% CÎ of the distribution of the SEs of the DCF		
		i.i.d. Bootstrap	One-Layer Bootstrap	Two-Layer Bootstrap
EL	0.000686	0.000687 0.105594×10^{-4} (3.01%) (0.000666, 0.000706)	0.001859 0.292555×10^{-4} (3.08%) (0.001806, 0.001920)	0.001975 0.329929×10^{-4} (3.27%) (0.001916, 0.002043)
LZ	0.000888	0.000888 0.133725×10^{-4} (2.95%) (0.000863, 0.000917)	0.002730 0.446791×10^{-4} (3.21%) (0.002646, 0.002817)	0.002870 0.460217×10^{-4} (3.14%) (0.002781, 0.002956)
PB	0.001118	0.001119 0.182849×10^{-4} (3.20%) (0.001084, 0.001155)	0.004150 0.677488×10^{-4} (3.20%) (0.004012, 0.004286)	0.004288 0.675055×10^{-4} (3.09%) (0.004149, 0.004420)
CH	0.002294	0.002294 0.367097×10^{-4} (3.14%) (0.002224, 0.002367)	0.004646 0.759175×10^{-4} (3.20%) (0.004509, 0.004790)	0.005172 0.819121×10^{-4} (3.10%) (0.005020, 0.005345)

TABLE III
THE DCFS, THE UPPER BOUND OF THE ANALYTICAL SÊS, AND THE ESTIMATED MEANS (RELATIVE ERRORS) AND 95% CÎS OF THE DISTRIBUTIONS OF THE SES OF THE DCFS COMPUTED USING THE I.I.D. BOOTSTRAP, THE ONE-LAYER BOOTSTRAP, AND THE TWO-LAYER BOOTSTRAP FOR EIGHT SYSTEMS

System	DCF	Upper bound of the analytical SÊ	Mean (relative error) and 95% CÎ of the distribution of the SEs of the DCF		
			i.i.d. Bootstrap	One-Layer Bootstrap	Two-Layer Bootstrap
UJ	0.028996	0.000502	0.000502 (3.39%) (0.000486, 0.000517)	0.001977 (13.36%) (0.001919, 0.002033)	0.002031 (13.73%) (0.001961, 0.002093)
BK	0.031588	0.000520	0.000521 (3.32%) (0.000503, 0.000536)	0.001814 (11.26%) (0.001759, 0.001874)	0.001876 (11.64%) (0.001818, 0.001934)
DL	0.040880	0.000571	0.000571 (2.74%) (0.000555, 0.000588)	0.001735 (8.32%) (0.001682, 0.001793)	0.001819 (8.72%) (0.001756, 0.001878)
AF	0.073500	0.000502	0.000502 (1.34%) (0.000487, 0.000517)	0.001667 (4.45%) (0.001621, 0.001718)	0.001735 (4.63%) (0.001683, 0.001788)
FI	0.096988	0.000346	0.000346 (0.70%) (0.000336, 0.000357)	0.000757 (1.53%) (0.000733, 0.000783)	0.000829 (1.68%) (0.000805, 0.000856)
PM	0.161254	0.001886	0.001884 (2.29%) (0.001824, 0.001948)	0.004871 (5.92%) (0.004729, 0.005013)	0.005221 (6.35%) (0.005055, 0.005381)
CO	0.223263	0.002194	0.002195 (1.93%) (0.002125, 0.002264)	0.006476 (5.69%) (0.006255, 0.006689)	0.006822 (5.99%) (0.006623, 0.007026)
DG	0.455384	0.002777	0.002780 (1.20%) (0.002694, 0.002870)	0.009267 (3.99%) (0.008990, 0.009520)	0.009645 (4.15%) (0.009339, 0.009926)

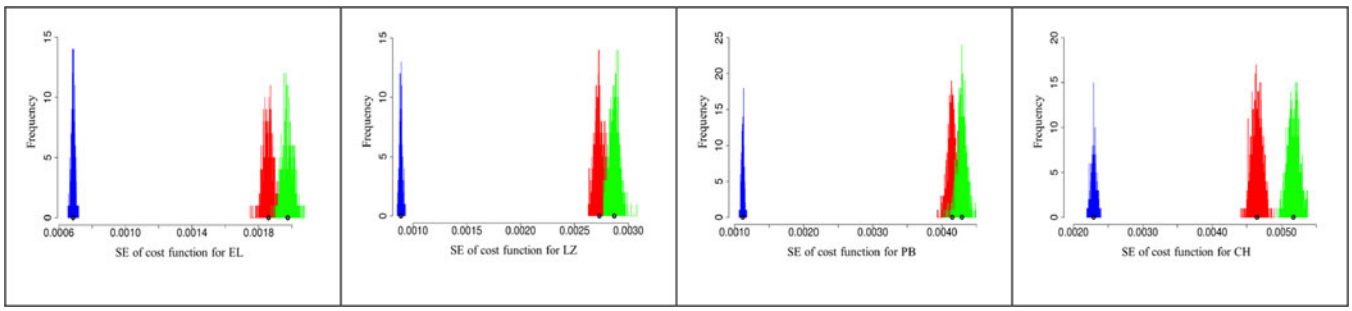


Fig. 3. The histograms of the SEs of the DCFs generated using the i.i.d. bootstrap (left - blue), the one-layer bootstrap (middle - red), and the two-layer bootstrap (right - green) for four systems EL, LZ, PB, and CH. The black circle stands for the estimated mean of the distribution.

As stated in Section III, the covariance is negative and is hard to be estimated. Hence, only the upper bounds of the analytically computed SEs of the DCF are listed in Table II. To take account of the stochastic nature of the bootstrap method, these upper bounds are compared with the 95% CI of the bootstrap estimated SEs of the DCF. Table II shows that the upper bounds of the analytical SEs without negative covariance are around the midpoints of the 95% CIs of the SEs estimated using the i.i.d. bootstrap. It is noted that the analytically computed SEs are smaller than their upper bounds because of the negative covariance. And it is also noted that the i.i.d. bootstrap does not take data dependency into account.

It is worth mentioning that in Table I, all estimated SÊs of the DCF calculated using different bootstrap methods were obtained by a random execution of a stochastic process. However, they all fall within the estimated 95% CÎ of the bootstrap estimated SÊs of the DCF that were shown in Table II. For example, for System EL using the two-layer resampling, the estimated SÊ 0.001952 in Table I falls within the 95% CÎ of the SE (0.001916, 0.002043) in Table II.

The distributions of SEs have three important features. The first feature regards the positions of the distributions of the SEs. As shown in Table II and depicted in Fig. 3, the estimated means imply that the two distributions of SEs created using the one-layer bootstrap and the two-layer bootstrap on the datasets with dependency are well separated, towards larger SEs, from the distribution of SEs generated using the i.i.d. bootstrap for each of the four systems. This indicates that the data dependency increases the SEs of the DCF.

The second feature concerns the variances of the distributions of SEs. Different runs of the bootstrap method might produce different results of SEs due to the stochastic nature of the bootstrap method. As evidenced by the estimated SÊs and 95% CÎs in Table II as well as by the widths of the histograms depicted in Fig. 3, the one-layer and the two-layer bootstrap methods create larger variation of the SEs than the i.i.d. bootstrap method does for each system. This indicates that the data dependency increases the variation of SEs.

The third feature is about what is the statistically significant relationship between the two distributions of the SEs of the DCF generated using the one-layer bootstrap and the two-layer bootstrap, respectively. To explore this, here are some preliminary observations from Table II. For each system, the estimated 95% CÎs of these two distributions overlap to some extent, but the estimated mean of one distribution generally does not fall inside the estimated 95% CÎ of the other distribution. Further, the ratios of the variances generated using the one-layer bootstrap to those created using the two-layer bootstrap are between 0.79 and 1.01 for all four systems.

Then, hypothesis testing may be conducted on both the estimated means and variances of the distributions. In order to do so, the types of the distributions need to be examined. The estimated 95% CÎs shown in Table II were all calculated using the quantile method as discussed in Section VII. They do match, up to the fourth or fifth decimal place, the 95% CÎs computed by assuming that the distribution of the SEs of the DCF is normal, i.e., by multiplying 1.96 by the estimated SÊ of such a distribution. For example, for System EL using the two-layer resampling, the 95% CÎ derived from the quantile method is (0.001916, 0.002043) as shown in Table II, while it is (0.001911, 0.002040) by assuming normality.

Moreover, the Shapiro-Wilk normality test [28] was conducted on the distributions of the SEs generated by the three bootstrap methods for the four systems. 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%. This suggests that the estimated SÊs of the DCF 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 [5], [23], [29]. It is observed in Table II and Fig. 3 that the mean of the distribution of the one-layer bootstrap is less than the mean of the distribution of the two-layer bootstrap. Hence, the one-tailed Z-test is applied. All p-values for the four systems are close to one, which strongly suggests that the above observation regarding the means hold true significantly.

Further, the p-values of the two-tailed F-test are greater than 5% except for System EL, where it is 0.73%. It indicates that the null hypothesis, i.e., the ratio of the variances of these two distributions is equal to one, cannot be rejected [24]. Although the SEs computed using the two-layer bootstrap are generally larger than those calculated using the one-layer bootstrap as shown in Table II, the difference is not significant.

Combining the results from these two hypothesis tests plus the preliminary observations, it can be concluded that the distribution of SEs of the DCF computed using the two-layer bootstrap for the datasets with data dependency is significantly to the right side of the distribution of SEs calculated using the one-layer bootstrap. This indicates that the two-layer bootstrap can better deal with the issue of data dependency and thus is more conservative than the one-layer bootstrap.

C. The Results of the Eight Systems

For the eight speaker recognition systems labeled as UJ, BK, DL, AF, FI, PM, CO, and DG, the estimated DCFs, means (relative errors) and 95% CÎs of distributions of SEs of the DCFs computed using the i.i.d. bootstrap, the one-layer bootstrap, and the two-layer bootstrap are shown in Table III. The estimated DCFs show that examples include systems with different performance levels. The relative error, defined by 1.96 x mean / DCF, shows the significance of the mean of the distribution of SEs of the DCF with respect to the DCF.

As done in Section VIII-B, only the upper bounds of the analytically computed SEs of the DCF are listed in Table III. The same conclusion can be drawn here regarding the upper bounds of the analytical SEs without negative covariance and the 95% CIs of SEs estimated using the i.i.d. bootstrap without considering data dependency.

For each system, the means of the three distributions increase in the order given, the widths of the estimated 95% CÎs of the

one-layer and two-layer bootstrap methods are fairly equal but larger than the one for the i.i.d. bootstrap, and the 95% CÎs also move toward larger SEs although for some systems the two CÎs computed using the one-layer and two-layer bootstrap methods overlap.

Thus, the features regarding the distribution of SEs of the DCF shown in Table III are exactly the same as those presented in Table II. As a result, the conclusions drawn from these eight systems are the same as those reached from the statistical analysis of the results of the previous four systems.

The three bootstrap methods have different impacts on the estimated SÊs of the DCF. The results for the 12 systems are conclusive. Indeed, they are backed theoretically by the rationale investigated in Appendix II.

IX. Conclusions and Discussion

As stated in Section I, using the analytical method to compute the SE of DCF in ROC analysis with data dependency is inappropriate. As shown in Sections VIII-B and VIII-C, the upper bounds of the analytical SEs are around the midpoints of the 95% CIs of the SEs estimated using the i.i.d. bootstrap that does not take account of data dependency. And the latter are all on the left side of the 95% CIs of the SEs estimated using the one-layer and the two-layer bootstrap methods. All these indicate that the analytically computed SEs of the DCF are smaller than those computed using one-layer and two-layer bootstrap methods. The larger one should be taken as the estimate of the SE of the DCF [30].

Because of the resampling process and its way of estimating the SE of the DCF, the nonparametric two-sample bootstrap algorithm can take account of the score distributions and the covariance intrinsically. Due to the two-layer data structure, the one-layer and two-layer bootstrap methods can take data dependency into consideration. As a result, the bootstrap method can estimate the SE of the DCF more effectively than the analytical approach.

Hence, the SE of the DCF was computed using the nonparametric two-sample bootstrap method. A void in the SREs was filled. To further address this problem, many issues will be involved.

As is known, the bootstrap method assumes that the random samples are i.i.d. In reality, data dependency may be inevitable in order to increase the size of datasets due to limited resources. In this article, the data dependency due to multiple use of the same subjects, i.e., the same training speaker id numbers, was taken into consideration, and its impact on the SEs of the DCF was studied.

To preserve such data dependency while the bootstrap resampling takes place, 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. As a result, a two-layer data structure is constructed.

Thereafter, besides the conventional nonparametric two-sample bootstrap method with the i.i.d. assumption, the two-sample one-layer bootstrap method resampling only on sets, and the two-sample two-layer bootstrap method resampling on sets and subsequently on scores in the sets are carried out.

To make scores have equal probabilities of being selected in the random resampling when the nonparametric two-sample two-layer bootstrap algorithm is employed, it is suggested that the numbers of target scores in target sets be the same and likewise for the numbers of non-target scores in non-target sets. Moreover, the same numbers of target (non-target) scores can be obtained at different iterations while the bootstrap resampling takes place. All these can reduce the variance of the computation.

Our research shows that data dependency does have an impact on the estimates of the SEs of measures. So, when designing a test and constructing a dataset, if data dependency is involved, the structure of the datasets should be chosen in accordance with the above suggestion. Otherwise, resources may be wasted and the SEs of the measures may not be able to be computed appropriately.

Equal-number-of-score sets can be easily chosen from not-equal-number-of-score sets by randomly selecting without replacement scores from those sets, in which the numbers of scores are larger than a specified number. The specified numbers for target sets and for non-target sets are determined, respectively, by the trial-and-error optimization so that the total numbers of target scores and of non-target scores are as large as possible.

Due to the stochastic nature of the bootstrap methods, the distributions of the estimated SEs of the DCF created by using the three different bootstrap methods were investigated and compared. In our studies, all bootstrap estimated SEs of the DCF obtained by random executions of the bootstrap fell in the 95% CI of the bootstrap estimated SEs of the DCF.

After performing the hypothesis testing and investigating the rationale theoretically for why the three bootstrap methods can have different impacts on the estimated SEs of the DCF, it is revealed that the data dependency can increase the SE of the DCF as well as the variation of SEs. The i.i.d. bootstrap does not take data dependency into account and thus underestimates the SEs of the DCF. And the two-layer bootstrap can better deal with the issue of data dependency and thus more conservatively estimates the SEs of the DCF (i.e., larger SEs) than the one-layer bootstrap method does. Indeed, these are all supported by the rationale investigated in Appendix II.

Indeed, it is a common practice to be more conservative when estimating the SE of a measure, i.e., obtaining a larger SE among all reasonable approaches [30]. In conclusion, the nonparametric two-sample two-layer bootstrap method is recommended when data dependency is involved.

In our studies, tens of thousands of scores were involved, as shown in Section V-D. It seems that the large size of datasets cannot reduce the impact of data dependency on the SEs of measures in ROC analysis.

The conclusions reached in this article are drawn from testing 12 speaker recognition systems in SREs. Recently, we used the bootstrap method developed in this article to compute SEs of measures using a completely different paradigm in which three score distributions and two thresholds were involved, and thus a different formulation of the DCF was employed with data dependency tested on 40 speaker recognition systems; and we observed the same characteristics [31]. Thus, not taking account of data dependency and using the bootstrap method with i.i.d.

assumption can underestimate the SEs of the DCF even with a doubled data size.

Different measures, such as equal error rate (EER), etc. have been used in SREs [8]–[11]. The methods proposed in this article, along with the methods presented in Ref. [5], can be employed to estimate the SEs of those measures.

The error bars of the DCF displaying the SE and 95% CI can be used to evaluate the performance level of a speaker recognition system against a hypothesized value. This is related to one-algorithm hypothesis testing, which can simply be judged by observing whether the 95% CI of the DCF contains, or lies below, or lies above the hypothesized value [5]. Further, the error bars can also be used to classify speaker recognition systems into different classes in terms of performance accuracies.

When the error bars overlap, two-algorithm hypothesis testing can be employed to compare two speaker recognition systems and determine the statistical significance of their performance difference. To do so, the Z-test can be carried out, because the DCF may be regarded as approximately normally distributed as discussed in Section VIII-A. Certainly, the correlation coefficient between the two DCFs of two systems involved in the Z-test should be taken into account. An algorithm for computing such correlation coefficients can be found in Refs. [5], [29].

APPENDIX I

Proof of Eq. (10): The proof here 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, $\ldots$, 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, $\ldots$, m, where $\sum_{i=1}^{m} n_i = n$. Thus, the relative frequency f of occurrence of a score in the i-th set with respect to the total number of all scores that have been chosen is

$$f = \frac{n_i}{\sum_{j=1}^{m} n_j \times \mu_j} = \frac{\frac{n_i}{n}}{\sum_{j=1}^{m} \frac{n_j}{n} \times \mu_j}. \tag{A.1}$$

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

$$\lim_{n \to \infty} \frac{n_i}{n} = \frac{1}{m},\ i = 1, \ldots, m. \tag{A.2}$$

Therefore, by the Law of Large Numbers, as the number of selections n goes to infinity, the relative frequency f in Eq. (A.1) approaches the probability p for a score being selected with respect to the total number of all scores that have been chosen, that is

$$p = \frac{\frac{1}{m}}{\sum_{j=1}^{m} \frac{1}{m} \times \mu_j} = \frac{1}{\sum_{j=1}^{m} \mu_j} = \frac{1}{N}. \tag{A.3}$$

APPENDIX II

The rationale for why the SEs computed using the three bootstrap methods are different is explored by investigating bootstrap samples of the original scores, bootstrap replications of the DCF, and the estimated SÊs and their distributions.

A. The Three Multinomial Probabilities (MP) of Selecting Bootstrap Samples

As stated in Section VI, the resampling takes place randomly WR on all scores for the bootstrap with i.i.d. assumption, only on sets for the one-layer bootstrap, and not only on sets but subsequently also on all scores in the selected sets for the two-layer bootstrap.

The probability of selecting a bootstrap sample is the MP [15]. The three bootstrap methods select bootstrap samples with three different MPs. Though the same bootstrap sample may be selected by different bootstrap methods, the corresponding MPs are different.

Using the notations in Sections IV and V, let N_i, m_i and μ_i, i $\in \{\mathrm{T}, \mathrm{N}\}$ be the total numbers of scores, the numbers of sets, and the numbers of scores in sets for target and non-target, respectively. The discussion here holds good for both the two-sample bootstrap and the one-sample bootstrap.

For the bootstrap with i.i.d. assumption, suppose that a bootstrap sample of size N_i, i $\in \{\mathrm{T}, \mathrm{N}\}$ is randomly drawn WR from the original target sample $\boldsymbol{T}$ in Eq. (7) or non-target sample $\boldsymbol{N}$ in Eq. (8), and it contains $\mathrm{k}_{i\alpha}$ copies of the α-th target or non-target score, subject to $0 \le \mathrm{k}_{i\alpha} \le \mathrm{N}_i$, $\alpha = 1, \ldots, \mathrm{N}_i$, and $\sum_{\alpha=1}^{\mathrm{N}_i} \mathrm{k}_{i\,\alpha} = \mathrm{N}_i$. That is, the number of repetitions of the α-th score in this bootstrap sample is $\mathrm{k}_{i\alpha}$. The summation of MPs for obtaining all possible bootstrap samples with different combinations of the numbers of repetitions of scores $(\mathrm{k}_{i\,1},\ \mathrm{k}_{i\,2},\ \ldots,\ \mathrm{k}_{i\,\mathrm{N}_i})$ is

$$\sum_{\substack{0 \le \mathrm{k}_{i\,\alpha} \le \mathrm{N}_i \\ \sum_{\alpha=1}^{\mathrm{N}_i} \mathrm{k}_{i\,\alpha} = \mathrm{N}_i}} \frac{\mathrm{N}_i!}{\mathrm{k}_{i\,1}!\mathrm{k}_{i\,2}! \ldots \mathrm{k}_{i\,\mathrm{N}_i}!} \times \frac{1}{{\mathrm{N}_i}^{\mathrm{N}_i}},\ i \in \{\mathrm{T}, \mathrm{N}\}. \tag{A.4}$$

For the one-layer bootstrap, the summation of MPs for obtaining all possible bootstrap samples with different combinations of the numbers of repetitions of sets $(\mathrm{k}_{i\,1},\ \mathrm{k}_{i\,2},\ \ldots,\ \mathrm{k}_{i\,\mathrm{m}_i})$ is

$$\sum_{\substack{0 \le \mathrm{k}_{i\,\alpha} \le \mathrm{m}_i \\ \sum_{\alpha=1}^{\mathrm{m}_i} \mathrm{k}_{i\alpha} = \mathrm{m}_i}} \frac{\mathrm{m}_i!}{\mathrm{k}_{i\,1}!\mathrm{k}_{i\,2}! \ldots \mathrm{k}_{i\,\mathrm{m}_i}!} \times \frac{1}{{\mathrm{m}_i}^{\mathrm{m}_i}},\ i \in \{\mathrm{T}, \mathrm{N}\}. \tag{A.5}$$

For the two-layer bootstrap, the summation of MPs for selecting all possible bootstrap samples with different combinations of the numbers of repetitions of sets $(\mathrm{k}_{i\,1},\ \mathrm{k}_{i\,2},\ \ldots,\ \mathrm{k}_{i\,\mathrm{m}_i})$, and different combinations of the numbers of repetitions of scores $(\mathrm{j}_{i\,\beta\,1},\ \mathrm{j}_{i\,\beta\,2},\ \ldots,\ \mathrm{j}_{i\,\beta\,\mu_i})$, $\beta = 1, 2, \ldots, \mathrm{m}_i$, within each of the m_i selected sets with respect to a selected partition $(\mathrm{k}_{i\,1},\ \mathrm{k}_{i\,2}, \ldots,\ \mathrm{k}_{i\,\mathrm{m}_i})$ is

$$\sum_{\substack{0 \le \mathrm{k}_{i\alpha} \le \mathrm{m}_i \\ \sum_{\alpha=1}^{\mathrm{m}_i} \mathrm{k}_{i\alpha} = \mathrm{m}_i}} \frac{\mathrm{m}_i!}{\mathrm{k}_{i\,1}!\mathrm{k}_{i\,2}! \ldots \mathrm{k}_{i\,\mathrm{m}_i}!} \times \frac{1}{{\mathrm{m}_i}^{\mathrm{m}_i}} \times \prod_{\beta=1}^{\mathrm{m}_i} \sum_{\substack{0 \le \mathrm{j}_{i\,\beta\gamma} \le \mu_i \\ \sum_{\gamma=1}^{\mu_i} \mathrm{j}_{i\beta\gamma} = \mu_i}} \frac{\mu_i!}{\mathrm{j}_{i\,\beta 1}!\mathrm{j}_{i\,\beta 2}! \ldots \mathrm{j}_{i\,\beta\mu_i}!} \times \frac{1}{{\mu_i}^{\mu_i}},$$

$$i \in \{\mathrm{T}, \mathrm{N}\}. \tag{A.6}$$

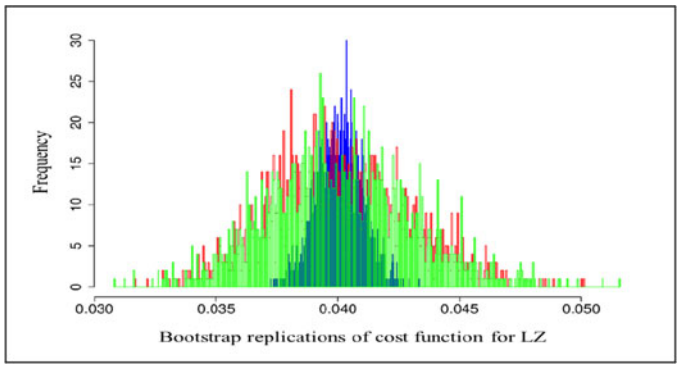


Fig. 4. The three distributions of 2,000 bootstrap replications of the DCF for System LZ, generated using the nonparametric two-sample bootstrap with i.i.d. assumption (blue), the one-layer bootstrap (red), and the two-layer bootstrap (green), respectively.

All total probabilities shown in Eqs. (A.4), (A.5) and (A.6) are normalized to 1. It is known that for all positive integers n

$$\mathrm{e}\sqrt{n}\left(\frac{\mathrm{n}}{\mathrm{e}}\right)^{\mathrm{n}} \geq \mathrm{n}! > \sqrt{2\pi\mathrm{n}}\left(\frac{\mathrm{n}}{\mathrm{e}}\right)^{\mathrm{n}}. \tag{A.7}$$

For large n, the right two items are approximately equal, which is Stirling's formula.

Here are three features. (1) The MP of selecting the original sample is the largest one. For instance, in Eq. (A.4) it is $\mathrm{N}_i! \,/\, \mathrm{N}_i^{\mathrm{N}_i}$, when all $\mathrm{k}_{i\,\alpha}$ are equal to 1. (2) Because $\mathrm{n}!/\mathrm{n}^{\mathrm{n}}$ decreases exponentially as n increases due to Eq. (A.7), the MP decreases very fast as the size of the original sample increases. (3) Because k! in the denominators increases very fast as k grows, the MPs of selecting bootstrap samples other than the original sample decrease very fast as the numbers of repetitions of scores or sets in the bootstrap samples increase.

The impact of these features is as follows. The numbers of scores N_i, $\mathrm{i} \in \{\mathrm{T}, \mathrm{N}\}$ are in the tens of thousands as shown in Section V-D. So, for the bootstrap with i.i.d. assumption, the MP of selecting the original sample is very small, and thus the MPs of selecting bootstrap samples with large numbers of repetitions of scores are even much smaller.

On the contrary, the numbers of sets, m_i, $\mathrm{i} \in \{\mathrm{T}, \mathrm{N}\}$, are in the hundreds, considerably smaller than the total numbers of scores. Hence, for the one-layer bootstrap, the MPs of selecting bootstrap samples with large numbers of repetitions of sets can increase tremendously.

The numbers of scores in sets, μ_i, $\mathrm{i} \in \{\mathrm{T}, \mathrm{N}\}$, are also in the hundreds. Thus, for the two-layer bootstrap, the MPs of selecting bootstrap samples with large numbers of repetitions of scores in sets can also increase compared to the bootstrap with i.i.d. assumption. If a bootstrap sample can be selected by both one-layer and two-layer bootstrap, the MP of the former is larger than the one of the latter due to Eqs. (A.5) and (A.6).

It is clear that all possible bootstrap samples with different combinations of the numbers of score repetitions selected using the one-layer bootstrap form a proper subset of those chosen using the two-layer bootstrap, which in turn constitute a proper subset of those selected using the bootstrap with i.i.d. assumption.

B. The Three Distributions of Bootstrap Replications

It may very well be that different bootstrap samples of the original data produce the same bootstrap replication of a measure. In other words, their relationship is many to one. This is true for the DCF, because it is determined only by the cumulative probabilities of target and non-target scores. Thus, it follows that the set of the bootstrap replications created using the one-layer bootstrap is a proper subset of those generated using the two-layer bootstrap; and the latter is a proper subset of those created using the bootstrap with i.i.d. assumption.

Further, besides the original estimator of the DCF created by the original sample, generally speaking, there are three categories of bootstrap replications of the DCF: (1) those created by the bootstrap samples with small numbers of repetitions and thus close to the original estimator of the DCF; (2) those generated by the bootstrap samples with large numbers of repetitions but nevertheless still close to the original estimator due to the reason stated above; (3) those created by the bootstrap samples with large numbers of repetitions and not close to the original estimator.

For all three bootstrap methods, the MPs of obtaining bootstrap replications become smaller as the numbers of repetitions of scores or sets in the bootstrap samples become larger. This occurs much more rapidly for the i.i.d. bootstrap than the one-layer bootstrap and the two-layer bootstrap.

Regarding the third category of bootstrap replications, if the i.i.d. bootstrap is employed, the MPs are extremely small relative to the probability of obtaining the original estimator, which is already very small. If they can be obtained using the two-layer bootstrap, the probabilities become considerably larger. And if they can be obtained using the one-layer bootstrap, the probabilities can be even much larger.

However, some of such bootstrap replications of the DCF cannot be generated using the one-layer bootstrap; but they can be created using the two-layer bootstrap. This is because resampling scores can take place on scores within sets only for the two-layer bootstrap. In other words, the two-layer bootstrap can generate more such bootstrap replications than the one-layer bootstrap does.

C. Three Bootstrap Estimated SEs and Their Distributions

In practice, for each of the three bootstrap methods, only a finite number of bootstrap samples are chosen to compute the bootstrap replications of the DCF. It is 2,000 based on our bootstrap variability studies [5]–[7]. So, for each bootstrap method, only those bootstrap samples and thus those bootstrap replications selected with relatively large MPs can be created.

Hence, for the bootstrap with i.i.d. assumption, only those bootstrap replications of the DCF with small numbers of repetitions of scores, and thus located close to the original estimator, can be generated. For the one-layer bootstrap, not only can some bootstrap replications located near the original estimator be created, but also some with large numbers of repetitions of sets and not located close to the original estimator can be generated. For the two-layer bootstrap, even more such bootstrap replications can get created.

Thus, the dispersion of the distribution of bootstrap replications of the DCF gets larger, when the bootstrap with i.i.d. assumption, the one-layer bootstrap, and the two-layer bootstrap are employed in turn. For the bootstrap method, the SE of a measure is estimated from the distribution of the bootstrap replications of the measure [15]. Hence, the SE of the DCF computed using the bootstrap with i.i.d. assumption is far smaller than the one calculated using the one-layer bootstrap, which is smaller than the one computed using the two-layer bootstrap.

As a result, the distributions of SEs of the DCF generated using the one-layer and the two-layer bootstrap are well separated towards larger SEs from the one created using the i.i.d. bootstrap; and the distribution of SEs of the DCF using the two-layer bootstrap lies above the one using the one-layer bootstrap. The bootstrap with i.i.d. assumption cannot pick up those bootstrap replications in Category 3, but the one-layer and two-layer bootstrap can. This narrows the variation of the SE of the DCF computed using the bootstrap with i.i.d. assumption relative to the other two bootstrap methods.

Fig. 4 shows the three distributions of the 2,000 bootstrap replications of the DCF for System LZ, which are generated using the nonparametric two-sample bootstrap with the i.i.d. assumption (blue), the one-layer bootstrap (red), and the two-layer bootstrap (green), respectively. They have the same relationship as described here.

ACKNOWLEDGMENT

The authors would like to thank Dr. G. A. Sanders at the National Institute of Standards and Technology for his valuable comments.

REFERENCES

[1] "The NIST Speaker Recognition Evaluation," 2012. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/

[2] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the mixer corpora –2004, 2005, 2006," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 7, pp. 1951–1959, Sep. 2007.

[3] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "NIST and NFI-TNO evaluations of automatic speaker recognition," *Comput. Speech Language*, vol. 20, pp. 128–158, 2006.

[4] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1895–1898.

[5] J. C. Wu, A. F. Martin, and R. N. Kacker, "Measures, uncertainties, and significance test in operational ROC analysis," *J. Res. Nat. Inst. Standards Technol.*, vol. 116, no. 1, pp. 517–537, 2011.

[6] J. C. Wu, A. F. Martin, and R. N. Kacker, "Bootstrap variability studies in ROC analysis on large datasets," *Commun. Statist.—Simulation Comput.*, vol. 43, no. 1, pp. 225–236, 2014.

[7] J. C. Wu, A. F. Martin, and R. N. Kacker, "Validation of nonparametric two-sample bootstrap in ROC analysis on large datasets," *Commun. Statist.—Simul. Comput.*, vol. 45, no. 5, pp. 1689–1703, 2016.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[9] V. Hautamäki *et al.*, "Sparse classifier fusion for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 8, pp. 1622–1631, Aug. 2013.

[10] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 11, pp. 2425–2438, Nov. 2013.

[11] T. Hasan and J. H. L. Hansen, "Maximum likelihood acoustic factor analysis models for robust speaker verification in noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 381–391, Feb. 2014.

[12] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. Speaker Language Recognit. Workshop (Odyssey)*, Toledo, 2004, pp. 237–244.

[13] J. C. Wu, A. F. Martin, G. A. Sanders, and R. N. Kacker, "Bootstrap method versus analytical approach for estimating uncertainty of measure in ROC analysis on large datasets," in preparation.

[14] B. Efron, "Bootstrap methods: Another look at the Jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.

[15] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.

[16] J. C. Wu and C. L. Wilson, "Nonparametric analysis of fingerprint data on large data sets," *Pattern Recognit.*, vol. 40, no. 9, pp. 2574–2584, 2007.

[17] R. Y. Liu and K. Singh, "Moving blocks jackknife and bootstrap capture weak dependence," in *Exploring the Limits of Bootstrap*, R. LePage and L. Billard Eds. New York, NY, USA: Wiley, 1992.

[18] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. New York, NY, USA: Cambridge Univ. Press, 2007, ch. 7.

[19] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective," *Speech Commun.*, vol. 31, no. 2–3, pp. 225–254, Jun. 2000.

[20] J. C. Wu and C. L. Wilson, "An empirical study of sample size in ROC-curve analysis of fingerprint data," *Biometric Technology for Human Identification III, Proc. SPIE*, 6202, 620207, 2006.

[21] R. M. Bolle, N. K. Ratha, and S. Pankanti, "Error analysis of pattern recognition systems–the subsets bootstrap," *Comput. Vis. Image Understanding*, vol. 93, pp. 1–33, 2004.

[22] N. Poh, A. F. Martin, and S. Bengio, "Performance generalization in biometric authentication using joint user-specific and sample bootstraps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 492–498, Mar. 2007.

[23] J. C. Wu, A. F. Martin, C. S. Greenberg, and R. N. Kacker, "Data dependency on measurement uncertainties in speaker recognition evaluation," *Active and Passive Signatures III, Proc. SPIE*, 8382, 83820D, 2012.

[24] B. Ostle and L. C. Malone, *Statistics in Research: Basic Concepts and Techniques for Research Workers*, 4th ed. Ames, IA, USA: Iowa State Univ. Press, 1988.

[25] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 8th ed. Ames, IA, USA: Iowa State Univ. Press, 1989.

[26] K. Linnet, "Comparison of quantitative diagnostic tests: type I error, power, and sample size," *Statist. Med.*, vol. 6, pp. 147–158, 1987.

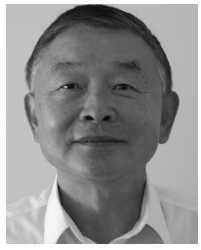
[27] R. J. Hyndman and Y. Fan, "Sample quantiles in statistical packages," *Amer. Statist.*, vol. 50, pp. 361–365, 1996.

[28] *R: A Language and Environment for Statistical Computing*, The R Development Core Team, Version 2.8.0, 2012. [Online]. Available: http://www.r-project.org/

[29] J. C. Wu, A. F. Martin, C. S. Greenberg, R. N. Kacker, and V. M. Stanford, "Significance test with data dependency in speaker recognition evaluation," *Active and Passive Signatures IV, Proc. SPIE*, 8734, 87340I, 2013.

[30] K. O. Hajian-Tilaki and J. A. Hanley, "Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data," *Academic Radiol.*, vol. 9, no. 11, pp. 1278–1285, 2002.

[31] J. C. Wu, A. F. Martin, C. S. Greenberg, and R. N. Kacker, "Measurement uncertainties of three score distributions and two thresholds with data dependency," *Nat. Inst. Standards Technol.*, NISTIR 8025, Sep. 2014.



Jin Chu Wu was accepted into the Graduate School, Fudan University, Shanghai, China, in 1978. He received the Ph.D. degree in theoretical high energy physics from the University of Pittsburgh, Pittsburgh, PA, USA, in 1985. His research focused on grand unification theories (GUTs) and lattice gauge theory. He joined the Superconducting Super Collider Laboratory in Dallas, TX, USA. Now he is a Researcher at the National Institute of Standards and Technology, Gaithersburg, MD, USA. His current research interests include nonparametric statistics and its applications.

Alvin F. Martin received the Ph.D. degree in mathematics from Yale University, New Haven, CT, USA, in 1977. From 1991 and until his recent retirement, he has worked as a Mathematician in the Multimodal Information Group, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, where he contributed to its evaluations of large vocabulary continuous speech recognition. From 1996 through 2011, he coordinated the world-wide series of NIST Speaker Recognition Evaluations and NIST Language Recognition Evaluations of automatic systems. He has also served on the Speaker Recognition Subcommittee of NIST's Organization of Scientific Area Committees for Forensic Science. He is a member of the Mathematical Association of America.

Craig S. Greenberg received the B.M. degree from Vanderbilt University, Nashville, TN, USA, in 2003, the B.A. (Hons.) degree in logic, information, and computation from the University of Pennsylvania, Philadelphia, PA, USA, in 2007, and the M.S. degree in applied mathematics from Johns Hopkins University, Baltimore, MD, USA, in 2012. He is currently working toward the Ph.D. degree in computer science from the University of Massachusetts, Amherst, MA, USA. He works as a Mathematician with the National Institute of Standards and Technology, Gaithersburg, MD, USA, in speaker recognition and language recognition as well as data science. He is currently a Research Assistant with the University of Pennsylvania, a Programmer with the Linguistic Data Consortium, and an English Language Annotator with the Institute for Research in Cognitive Science. He has been a member of the International Speech Communication Association since 2008. He has received two official letters of recognition for his contribution to speaker recognition evaluation.

Raghu N. Kacker received the Ph.D. degree in statistics from Iowa State University, Ames, IA, USA, in 1979. He is a Mathematical Statistician at the National Institute of Standards and Technology, Gaithersburg, MD, USA. His current interests include software testing, statistical methods, and evaluation of uncertainty in outputs of computational models and physical measurements. He is a Fellow of the American Statistical Association and a Fellow of the American Society for Quality. He has coauthored more than 125 refereed publications and one book. He has received the Distinguished Technical Staff Award from Bell Labs, and Bronze medal and Silver medal from the US Department of Commerce.