NISTIR 8025

Measurement Uncertainties of Three Score Distributions and Two Thresholds with Data Dependency

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

http://dx.doi.org/10.6028/NIST.IR.8025

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8025

Measurement Uncertainties of Three Score Distributions and Two Thresholds with Data Dependency

Jin Chu Wu
Alvin F. Martin
Craig S. Greenberg
Information Access Division
Information Technology Laboratory

Raghu N. Kacker
Applied and Computational Mathematics Division
Information Technology Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8025

September 2014

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Acting Under Secretary of Commerce for Standards and Technology and Acting Director

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.IR.8025

Measurement Uncertainties of Three Score Distributions and Two Thresholds with Data Dependency

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). Recently a new paradigm was adopted to evaluate the performance of speaker recognition systems in which three distributions of target, known non-target, and unknown non-target scores, as well as two thresholds were employed. The new detection cost function was defined to be an average of the two weighted sums of the probabilities of type I and type II errors corresponding to the two decision thresholds. In addition, data dependency due to multiple use of the same subjects is also involved. The data were reorganized into a two-layer structure in view of the data dependency and the probability theory. Then, the uncertainties of the detection cost functions were computed using the nonparametric three-sample two-layer bootstrap method. Comparing these results with those calculated by using all the raw data and the nonparametric three-sample bootstrap method with the i.i.d. assumption, the measurement accuracies, i.e., the detection cost functions, have changed little; but the measurement uncertainties, i.e., the standard errors of the detection cost function, have improved as a result of taking account of the data dependency. Forty speaker recognition systems were used as examples.

Index Terms – Metrology, measure, accuracy, uncertainty, bootstrap, data dependency, speaker recognition.

1. 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 technical capabilities of the research community working on the general problem of text independent speaker recognition [2, 3]. In this article an analysis of the overall actual decision cost function results of the SRE12 primary systems for the core condition is presented.

In SRE12 a new paradigm was adopted to evaluate the performance of speaker recognition systems. The three distributions of target, known non-target, and unknown non-target scores, as well as two thresholds were employed. And the new detection cost function was defined to be an average of the two weighted sums of the probabilities of type I (miss) and type II (false alarm) errors corresponding to the two decision thresholds [1].

The probabilities of type I error and type II error are traded off and thus negatively correlated, and it is difficult to calculate the covariance term of two such correlated probabilities analytically. It is also hard to assert that the distributions of the three different kinds of scores be exactly normal with certain mean and standard deviation, respectively (see Section 2.4). The three sets of scores must be resampled simultaneously in order to compute the replications of the detection cost function where the type-I-error probability of the target scores and the two type-II-error probabilities of the known non-target scores and the unknown non-target scores are involved (see Section 3).

Therefore, the uncertainty in terms of the standard error (SE) and the 95 % confidence interval (CI) of the detection cost function is computed using the nonparametric three-sample bootstrap method, where the empirical distribution is assumed for each of the observed scores, based on our extensive bootstrap variability studies in ROC analysis on large datasets [4-9].

In addition, data dependency due to multiple use of the same subjects in order to create more scores is also involved. In this article, data dependency is determined based purely upon whether the same training speaker identification number is used multiple times while generating the datasets for SRE12. The calls from a single speaker are not independent. Under such circumstances, if the data is assumed to be independent and identically distributed (i.i.d.) when the bootstrap method is employed, the uncertainties of the measures will be underestimated [7].

Thus, those target scores, known non-target scores, and unknown non-target scores generated using the same training speaker identification number are grouped into a target set, a known non-target set, and an unknown non-target set, respectively. This can preserve the data dependency while the bootstrap resampling takes place [5, 7, 10].

Different sets may have different numbers of scores. With the bootstrap method, this can result in each target (known non-target, unknown non-target) score not having the same probability of being selected, and the numbers of scores resampled being different from iteration to iteration. In view of the probability theory related to the data dependency, the datasets are adjusted so that all target sets

contain the same number of scores, and likewise for the known non-target sets and the unknown non-target sets [7]. As a result, the variance of the computation can be reduced.

Thus, the speaker recognition data structure has two layers: The first layer consists of target sets, known non-target sets, and unknown non-target sets; and the second layer consists of target scores, known non-target scores, and unknown non-target scores within the sets. Thereafter, the bootstrap resampling can take place randomly with replacement (WR) only on the first layer of the data, or subsequently on the second layer of the data. For resampling on the first layer the bootstrap units are sets, whereas for resampling on the second layer the bootstrap units are the scores within a set, in which the scores are assumed to be conditionally independent.

Nonetheless, based on our prior research, the SE of the detection cost function in SRE12 was computed only using the nonparametric three-sample two-layer bootstrap method. For the sake of comparison, in this article the SE of the cost function was also calculated using the nonparametric three-sample bootstrap with the i.i.d. assumption and using all of the raw data. It was found that the measurement accuracies, i.e., the detection cost functions, have changed little; but the measurement uncertainties, i.e., the standard errors of the detection cost function, have improved as a result of taking account of the data dependency. A total of 47 speaker recognition systems were taken as examples[1, 2].

The bootstrap method on datasets with dependencies was initially studied in the references [5, 10], and applied to other cases later [11, 12]. In this article, the nonparametric three-sample bootstrap rather than the one-sample bootstrap is employed. Through this method, the uncertainties of much more complicated measures in ROC analysis, such as the detection cost function defined as an average of the two weighted sums of two probabilities corresponding to the two decision thresholds, can be computed. More importantly, in this article the probability issues for similarity scores being selected and the numbers of scores being resampled at different iterations are investigated [7].

The way of adjusting the speaker recognition datasets into a two-layer data structure, the notations of data structure, the two-layer resampling method and the related selection probability, and the new distributions of scores after adjustment are presented in Section 2. The detection cost function in SRE12 is shown in Section 3. The nonparametric three-sample two-layer bootstrap algorithm is explored in Section 4. The results with data dependency and the results assuming the data are i.i.d. are presented in Sections 5 and 6, respectively. The comparisons of the two results in terms of performance accuracy and the measurement uncertainty are shown in Section 7. Finally, the conclusions and discussion can be found in Section 8.

2. Adjust the speaker recognition datasets into a two-layer data structure

[1] Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

[2] The speaker recognition systems are in general proprietary. It is NIST policy not to publicly associate speaker evaluation participant site names with system performance results.

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.IR.8025

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

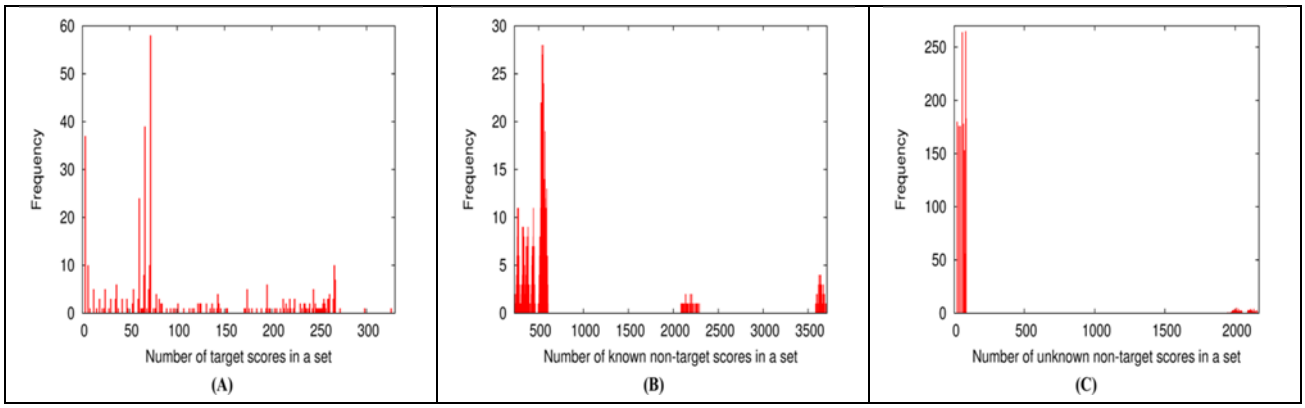

Figure 1 The histograms of the numbers of target scores (A), known non-target scores (B), and unknown non-target scores (C) in a set, respectively.

In this article, the data dependency is determined based purely upon whether the training speaker identification number is used multiple times. Those target scores, known non-target scores, and unknown non-target scores, generated using the same training speaker identification number, are grouped into a target set, a known non-target set, and an unknown non-target set, respectively. In other words, a two-layer data structure is constructed: The first layer consists of target sets, known non-target sets, and unknown non-target sets, and the second layer consists of target scores, known non-target scores, and unknown non-target scores within sets. This structure preserves the data dependency while the bootstrap resampling takes place.

The total numbers of target scores, known non-target scores, and unknown non-target scores are 41 897, 1 291 587, and 407 827, respectively. They were grouped into 394 target sets, 1 918 known non-target sets, and 1 918 unknown non-target sets, respectively. In Figure 1 are depicted the histograms of the numbers of target scores (A), known non-target scores (B), and unknown non-target scores (C) in a set, respectively.

Figure 1 shows that different sets contain quite different numbers of scores, indicating that different training speaker identification numbers were used different numbers of times, and the numbers of sets that contain a particular number of scores are also quite different. Moreover, the three types of scores have very different distributions. Such wide variations of numbers of scores in sets can have impact on the probability for a score to be selected.

2.2 The notations of the two-layer data structure

In the following text, let $\boldsymbol{S}$ denote score sets, α represent similarity scores, and μ be the number of scores in a set. The first subscript stands for whether it is referred to as target (T) or known non-target (K) or unknown non-target (U), the second subscript means the ordinal number of sets, and the third subscript represents the ordinal number of scores in a set.

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.IR.8025

Suppose that there are m_T target sets, m_K known non-target sets, and m_U unknown non-target sets. Thus, the set $\boldsymbol{S}_T$ of all target sets, the set $\boldsymbol{S}_K$ of all known non-target sets, and the set $\boldsymbol{S}_U$ of all unknown non-target sets are expressed by

$$\boldsymbol{S}_i = \{ \boldsymbol{S}_{ij} \mid j = 1, \ldots, m_i \}, i \in \{T, K, U\}, \quad (1)$$

where $\boldsymbol{S}_{Tj}$ are target sets, $\boldsymbol{S}_{Kj}$ are known non-target sets, and $\boldsymbol{S}_{Uj}$ are unknown non-target sets.

target $\boldsymbol{S}_T$	sets	$\boldsymbol{S}_{T1}$	$\boldsymbol{S}_{T2}$	……	$\boldsymbol{S}_{T\,m_T}$
	scores	$\alpha_{T11}, \alpha_{T12}, \ldots, \alpha_{T1\mu_{T1}}$	$\alpha_{T21}, \alpha_{T22}, \ldots, \alpha_{T2\mu_{T2}}$	……	$\alpha_{Tm_T1}, \alpha_{Tm_T2}, \ldots, \alpha_{Tm_T\mu_{Tm_T}}$

Table 1 The m_T target sets, and the target scores contained in each set.

non-target $\boldsymbol{S}_K$	sets	$\boldsymbol{S}_{K1}$	$\boldsymbol{S}_{K2}$	……	$\boldsymbol{S}_{K\,m_K}$
	scores	$\alpha_{K11}, \alpha_{K12}, \ldots, \alpha_{K1\mu_{K1}}$	$\alpha_{K21}, \alpha_{K22}, \ldots, \alpha_{K2\mu_{K2}}$	……	$\alpha_{Km_K1}, \alpha_{Km_K2}, \ldots, \alpha_{Km_K\mu_{Km_K}}$

Table 2 The m_K known non-target sets, and the known non-target scores contained in each set.

non-target $\boldsymbol{S}_U$	sets	$\boldsymbol{S}_{U1}$	$\boldsymbol{S}_{U2}$	……	$\boldsymbol{S}_{U\,m_U}$
	scores	$\alpha_{U11}, \alpha_{U12}, \ldots, \alpha_{U1\mu_{U1}}$	$\alpha_{U21}, \alpha_{U22}, \ldots, \alpha_{U2\mu_{U2}}$	……	$\alpha_{Um_U1}, \alpha_{Um_U2}, \ldots, \alpha_{Um_U\mu_{Um_U}}$

Table 3 The m_U unknown non-target sets, and the unknown non-target scores contained in each set.

And each set is expressed in terms of scores by

$$\boldsymbol{S}_{ij} = \{ \alpha_{ijk} \mid k = 1, \ldots, \mu_{ij} \}, j = 1, \ldots, m_i \text{ and } i \in \{T, K, U\}, \quad (2)$$

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

The target, known non-target, and unknown non-target sets and scores contained in each set are explicitly listed in Tables 1, 2, and 3, respectively. The m_T target sets $\boldsymbol{S}_{T1}, \boldsymbol{S}_{T2}, \ldots, \boldsymbol{S}_{T\,m_T}$, forming the set $\boldsymbol{S}_T$, contain $\mu_{T1}, \mu_{T2}, \ldots, \mu_{T\,m_T}$ target scores, respectively. The m_K known non-target set $\boldsymbol{S}_{K1}, \boldsymbol{S}_{K2}, \ldots, \boldsymbol{S}_{K\,m_K}$, forming the set $\boldsymbol{S}_K$, have $\mu_{K1}, \mu_{K2}, \ldots, \mu_{K\,m_K}$ known non-target scores, respectively. And the m_U unknown non-target set $\boldsymbol{S}_{U1}, \boldsymbol{S}_{U2}, \ldots, \boldsymbol{S}_{U\,m_U}$, forming the set $\boldsymbol{S}_U$, have $\mu_{U1}, \mu_{U2}, \ldots, \mu_{U\,m_U}$ unknown non-target scores, respectively.

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

$$\boldsymbol{T} = \{ \alpha_{Tjk} \mid k = 1, \ldots, \mu_{Tj} \text{ and } j = 1, \ldots, m_T \},$$

$$\boldsymbol{K} = \{ \alpha_{K\,j\,k} \mid k = 1, \ldots, \mu_{K\,j} \text{ and } j = 1, \ldots, m_K \},$$
$$\boldsymbol{U} = \{ \alpha_{U\,j\,k} \mid k = 1, \ldots, \mu_{U\,j} \text{ and } j = 1, \ldots, m_U \}. \quad (3)$$

The sets $\boldsymbol{S}_{i\,j}$, $\boldsymbol{T}$, $\boldsymbol{K}$, and $\boldsymbol{U}$ should all be viewed as multisets, in which members are allowed to appear more than once. All similarity scores are treated as different objects because 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.

Finally, the total numbers of target scores, known non-target scores, and unknown non-target scores, i.e., N_T, N_K, and N_U, satisfy

$$N_i = \sum_{j=1}^{m_i} \mu_{i\,j}, \quad \text{where } i \in \{T, K, U\}. \quad (4)$$

2.3 The two-layer resampling method and the related selection probability

As mentioned in Section I, the nonparametric three-sample two-layer bootstrap method is employed to compute the measurement uncertainty of the detection cost function with the triple distributions of scores and data dependency. The two-layer resampling takes place randomly WR not only on the first layer of the data, i.e., target sets, known non-target sets, and unknown non-target sets, but also on the second layer of the data, i.e., the scores in the sets, respectively. The scores in the sets are assumed to be conditionally independent. Therefore, the resampling units for the first layer are sets and for the second layer are the scores in the sets.

Then, the probability for a score $\alpha_{i\,j\,k}$ in a set $\boldsymbol{S}_{i\,j}$ being selected in the two-layer data resampling is

$$P_{2\text{-layer}}(\alpha_{i\,j\,k}) = P(\boldsymbol{S}_{i\,j}) \times P(\alpha_{i\,j\,k} \mid \boldsymbol{S}_{i\,j}) = \frac{1}{m_i} \times \frac{1}{\mu_{i\,j}}, \quad (5)$$
$$\text{where } k = 1, \ldots, \mu_{i\,j},\ j = 1, \ldots, m_i \text{ and } i \in \{T, K, U\}.$$

This selection probability is the same for all scores within a set, regardless of whether it is a target set, a known non-target set, or an unknown non-target set. However, the probabilities for scores being selected are set dependent because of different score numbers in different sets represented by the $\mu_{i\,j}$. In other words, target (known non-target and unknown non-target) scores in different sets do not have the same probability of being selected while using the two-layer bootstrap method.

It is clearly inappropriate that target scores, known non-target scores, and/or unknown non-target scores be selected with unequal probabilities for the two-layer data resampling. The impact of varied numbers of scores within a set on the probabilities for a score being selected must be eliminated. The datasets must be adjusted so that all target sets contain the same number of target scores and likewise for the known non-target sets and the unknown non-target sets.

If all $\mu_{T\,j}$, $j = 1, \ldots, m_T$, have equal value μ_T, then the probability for each target score being selected is $1 / N_T$ due to Eq. (4). Hence, each target score can have equal probability to be selected. So is each known non-target score if all $\mu_{K\,j}$, $j = 1, \ldots, m_K$, have equal value μ_K; and so is each unknown non-target score if all $\mu_{U\,j}$, $j = 1, \ldots, m_U$, have equal value μ_U. The probabilities for each known non-

target score and each unknown non-target score being selected are 1 / N_K and 1 / N_U, respectively, due to Eq. (4).

In addition, this dataset adjustment can ensure that the same numbers of target scores, known non-target scores, and unknown non-target scores, respectively, are resampled at different iterations using the bootstrap method as shown in Section 4. Hence, such a structure of datasets can reduce the variance of the computation of the measurement uncertainty of the detection cost function.

2.4 The new distributions of scores after adjustment

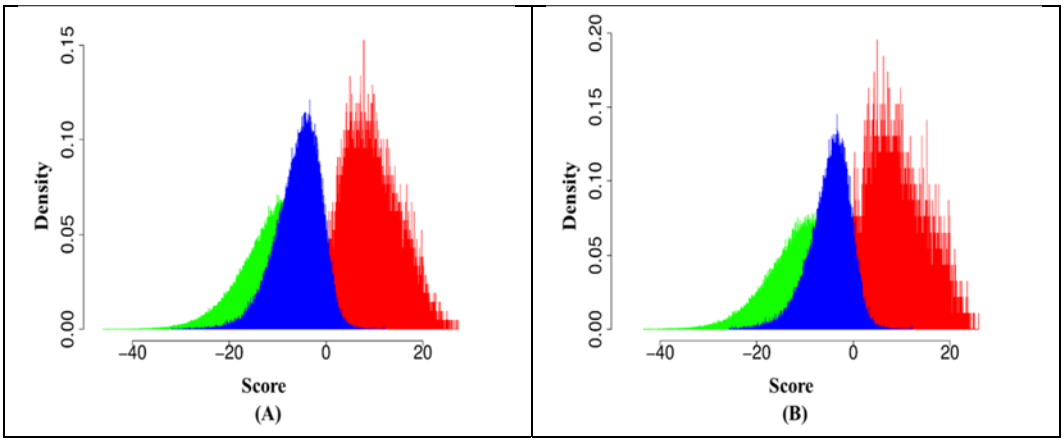

Figure 2 The three distributions of the target scores (red), known non-target scores (green), and unknown non-target scores (blue) of System 1 before adjustment (A) and after adjustment (B).

The datasets should not only be reorganized so that the probabilities for each target score, each known non-target score, and each unknown non-target score being selected will be equal, respectively; but also be chosen to maximize the numbers of similarity scores of each type, which determines the optimal numbers of sets and scores per set for the target sets, known non-target sets, and unknown non-target sets.

As presented in Section 2.1, the total numbers of the raw target scores, known non-target scores, and unknown non-target scores were 41 897, 1 291 587, and 407 827, respectively. These were grouped into 394 target sets, 1 918 known non-target sets, and 1 918 unknown non-target sets, respectively, based on whether the training speaker identification number is used multiple times. The score sets that had fewer scores than the number needed were discarded, while for those score sets that had more scores than needed, the number of scores was reduced by random selection (without replacement) to the number needed. Otherwise, the scores sets were chosen without any selection. Such random selection has little impact on the results (see Section 7.1).

As a result, 95 target score sets, 1 192 known non-target score sets, and 146 unknown non-target score sets were selected, containing 194 target scores, 511 known non-target scores, and 1 967 unknown non-target scores, respectively. Thus, the total numbers of resulting target scores, known non-target scores, and unknown non-target scores were 18 430, 609 112, and 287 182, respectively. Hence, there are still tens of thousands of scores in the new datasets.

This adjustment approximately halves the total number of scores in each category. Figure 2 shows the three distributions of the target scores (red), known non-target scores (green), and unknown non-target scores (blue) of System 1 before adjustment (A) and after adjustment (B). It can be seen that the relative positions of the three distributions of scores before adjustment and after adjustment remains almost the same, except the tails after adjustment are shorter than those before adjustment. The matching process in speaker recognition may be simply regarded as a dichotomous response (i.e., yes or no) with respect of a specified threshold, and thus what really matters is the relative positions of the three distributions of scores [13]. As a result, this indicates that the adjustment of the three datasets has little impact on the performance accuracies of speaker recognition systems, when the data size reaches to over tens and hundreds of thousands in ROC analysis [14].

3. The detection cost function in SRE12

After converting to integer scores for implementation purpose, without loss of generality, for a speaker recognition system, the scores are expressed inclusively using the integer score set $\{s\} = \{s_{min}, s_{min}+1, \ldots, s_{max}\}$, running consecutively from the lowest score s_{min} up to the highest score s_{max}. Let $f_i(s)$, $i \in \{ T, K, U \}$, denote the continuous probability density functions of target scores, known non-target scores, and unknown non-target scores. The three corresponding discrete probability distribution functions, denoted by $P_i(s)$, $s \in \{s\}$ and $i \in \{T, K, U\}$, are expressed as

$$\boldsymbol{P_i} = \{ P_i(s) \mid \forall\, s \in \{s\} \text{ and } \sum_{s=S_{min}}^{S_{max}} P_i(s) = 1 \}, \; i \in \{T, K, U\} . \tag{6}$$

In SRE12, the detection cost function involves the probability of type I error of the target scores, and the probabilities of type II errors of the known non-target scores and the unknown non-target scores, evaluated at two thresholds $t_i \in \{s\}$, $i = 1, 2$, respectively, assuming $t_1 < t_2$. The probabilities of type I errors at thresholds t_i for target scores, denoted by $\alpha_T(t_i)$, are cumulated from the lowest score s_{min}. The probabilities of type II errors at thresholds t_i for known non-target scores and unknown non-target scores, denoted by $\beta_j(t_i)$, $j \in \{K, U\}$, are cumulated from the highest score s_{max}. For discrete probability distribution, while computing $\alpha_T(t_i)$, $\beta_K(t_i)$, and $\beta_U(t_i)$ at thresholds t_i, the probabilities of target scores, known non-target scores, and unknown non-target scores at the thresholds t_i must be taken into account [15].

Hence, the probabilities of type I errors $\alpha_T(t_i)$ and type II errors $\beta_j(t_i)$, where $j \in \{K, U\}$ and $i = 1, 2$, at the two thresholds can be expressed by

$$\alpha_T(t_i) = \int_{-\infty}^{t_i} f_T(s)\, ds = \sum_{s=S_{min}}^{t_i} P_T(s) = 1 - \sum_{s=t_i+1}^{S_{max}} P_T(s), \qquad i = 1, 2, \tag{7}$$

and

$$\beta_j(t_i) = \int_{t_i}^{+\infty} f_j(s)\, ds = \sum_{s=t_i}^{S_{max}} P_j(s), \qquad j \in \{K, U\} \text{ and } i = 1, 2, \tag{8}$$

where $P_T(s_{max} + 1) = 0$ is assumed and the normalization in Eq. (6) is employed. In practice these error rates can be obtained by moving the score from the highest score s_{max} down to the thresholds t_i one score at a time to cumulate the probabilities of target scores, known non-target scores, and unknown non-target scores, respectively.

In SRE12, the two weighted sums of the probabilities of type I and type II errors at thresholds t_1 and t_2, respectively, are defined as [1],

$$W(t_i) = C_{Miss} \times P_{Target\,i} \times \alpha_T(t_i) + C_{FalseAlarm} \times (1 - P_{Target\,i}) \times [P_{Known} \times \beta_K(t_i) + (1 - P_{Known}) \times \beta_U(t_i)], \quad i = 1, 2. \tag{9}$$

Notice that $P_{Target\ 1}$ corresponds to the smaller threshold t_1, and $P_{Target\ 2}$ corresponds to the larger threshold t_2 [1]. The detection cost function is defined as the average of these two weighted sums,

$$C_{Det} = \frac{W(t_1) + W(t_2)}{2}. \tag{10}$$

The parameter C_{Miss} is the cost of a miss, $C_{FalseAlarm}$ is the cost of a false alarm, $P_{Target\ 1}$ and $P_{Target\ 2}$ are the *a priori* probabilities that the segment speaker is the target speaker depending on the thresholds, and P_{known} is the *a priori* probability that the non-target speaker is one of the evaluation target speakers. For this evaluation of speaker recognition performance for all speaker detection tests, the parameters C_{Miss}, $C_{FalseAlarm}$, $P_{Target\ 1}$, $P_{Target\ 2}$, and P_{known} were set to be 1.0, 1.0, 0.01, 0.001, and 0.5, respectively [1].

In this evaluation, systems were required to provide for each trial a score that could be interpreted as a log-likelihood ratio (LLR). Thus, the thresholds t_1 and t_2 were fixed values for all systems corresponding to the specified target trial prior probabilities of 0.01 and 0.001. Specifically, t_1 = ln (99) and t_2 = ln (999) [1].

4. The nonparametric three-sample two-layer bootstrap algorithm

It is difficult to compute analytically the covariance terms of the correlated probabilities shown in Eq. (9). Thus, the estimate of the SE of the detection cost function with data dependency is computed using the nonparametric three-sample two-layer bootstrap resampling methods based on our extensive studies of bootstrap variability in ROC analysis on large datasets [4-9]. From here on, the superscript indices are used for the numeration of the resampling iterations.

Here is a function WR_Random_Sampling_Set that will be frequently employed in the following algorithm,

1: **function** WR_Random_Sampling_Set (N, **Γ**, **Θ**)
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**

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**.

The nonparametric three-sample two-layer bootstrap method is carried out not only on the first layer of the new data structure where the resampling units are target sets, known non-target sets, and unknown non-target sets, but also on the second layer of the data in which the resampling units are target scores, known non-target scores, and unknown non-target scores in sets. Hence, the algorithm is shown as follows.

Algorithm **(the nonparametric three-sample two-layer bootstrap)**

1: **for** i = 1 **to** B **do**
2: WR_Random_Sampling_Set (m_T, $\boldsymbol{S}_T$, $\boldsymbol{S}'_T{}^i = \{ \boldsymbol{S}'_{Tj}{}^i \mid j = 1, \ldots, m_T \}$)
3: **for** k = 1 **to** m_T **do**
4: WR_Random_Sampling_Set (μ_T, $\boldsymbol{S}'_{Tk}{}^i$, $\boldsymbol{S}''_{Tk}{}^i$)
5: **end for**

6: WR_Random_Sampling_Set (m_K, $\boldsymbol{S}_K$, $\boldsymbol{S}'_K{}^i = \{ \boldsymbol{S}'_{Kj}{}^i \mid j = 1, \ldots, m_K \}$)
7: **for** k = 1 **to** m_K **do**
8: WR_Random_Sampling_Set (μ_K, $\boldsymbol{S}'_{Kk}{}^i$, $\boldsymbol{S}''_{Kk}{}^i$)
9: **end for**

10: WR_Random_Sampling_Set (m_U, $\boldsymbol{S}_U$, $\boldsymbol{S}'_U{}^i = \{ \boldsymbol{S}'_{Uj}{}^i \mid j = 1, \ldots, m_U \}$)
11: **for** k = 1 **to** m_U **do**
12: WR_Random_Sampling_Set (μ_U, $\boldsymbol{S}'_{Uk}{}^i$, $\boldsymbol{S}''_{Uk}{}^i$)
13: **end for**

14: $\boldsymbol{S}''_T{}^i = \{ \boldsymbol{S}''_{Tj}{}^i \mid j = 1, \ldots, m_T \}$ and $\boldsymbol{S}''_K{}^i = \{ \boldsymbol{S}''_{Kj}{}^i \mid j = 1, \ldots, m_K \}$ and $\boldsymbol{S}''_U{}^i = \{ \boldsymbol{S}''_{Uj}{}^i \mid j = 1, \ldots, m_U \} \Rightarrow \hat{W}^i(t_1)$ and $\hat{W}^i(t_2) \Rightarrow \hat{C}^i$

15: **end for**
16: $\{ \hat{C}^i \mid i = 1, ..., B \} \Rightarrow S\hat{E}$ and $(\hat{Q}(\alpha/2), \hat{Q}(1-\alpha/2))$
17: **end**

where B is the number of bootstrap replications, the set $\boldsymbol{S}_T$ of all target sets, the set $\boldsymbol{S}_K$ of all known non-target sets, and the set $\boldsymbol{S}_U$ of all unknown non-target sets are expressed in Eq. (1), and m_T, m_K, and m_U are the cardinalities of the sets $\boldsymbol{S}_T$, $\boldsymbol{S}_K$, and $\boldsymbol{S}_U$, respectively.

This algorithm runs B times, as shown from Step 1 to 15. In the i-th iteration, as shown in Steps 2, 6, and 10, the function WR_Random_Sampling_Set is applied three times to sets rather than scores, which are the first layer of datasets. That is, m_T target sets are randomly selected WR from the set $\boldsymbol{S}_T$ to constitute a new set $\boldsymbol{S}'_T{}^i = \{ \boldsymbol{S}'_{Tj}{}^i \mid j = 1, \ldots, m_T \}$, m_K known non-target sets are randomly

selected WR from the set $\boldsymbol{S}_K$ to form a new set $\boldsymbol{S}'_{K}{}^{i} = \{ \boldsymbol{S}'_{K\,j}{}^{i} \mid j = 1, \ldots, m_K \}$, and m_U unknown non-target sets are randomly selected WR from the set $\boldsymbol{S}_U$ to form a new set $\boldsymbol{S}'_{U}{}^{i} = \{ \boldsymbol{S}'_{U\,j}{}^{i} \mid j = 1, \ldots, m_U \}$.

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, μ_T target scores are randomly selected WR from the target set $\boldsymbol{S}'_{T\,k}{}^{i}$, which is the k-th new target set from the first-layer resampling, to form the k-th new target set $\boldsymbol{S}''_{T\,k}{}^{i}$ of the second-layer resampling. The analogous interpretation can be applied to known non-target scores in the known non-target set $\boldsymbol{S}'_{K\,k}{}^{i}$ as shown from Step 7 to 9, and unknown non-target scores in the unknown non-target set $\boldsymbol{S}'_{U\,k}{}^{i}$ as shown from Step 11 to 13.

As shown in Step 14, all target scores in the new set $\boldsymbol{S}''_{T}{}^{i} = \{ \boldsymbol{S}''_{T\,j}{}^{i} \mid j = 1, \ldots, m_T \}$, all known non-target scores in the new set $\boldsymbol{S}''_{K}{}^{i} = \{ \boldsymbol{S}''_{K\,j}{}^{i} \mid j = 1, \ldots, m_K \}$, and all unknown non-target scores in the new set $\boldsymbol{S}''_{U}{}^{i} = \{ \boldsymbol{S}''_{U\,j}{}^{i} \mid j = 1, \ldots, m_U \}$ are employed to generate the i-th estimates $\hat{W}^{i}(t_1)$ and $\hat{W}^{i}(t_2)$, by moving one integer score at a time from the highest score s_{max} down to the larger threshold t_2 and then to the smaller threshold t_1 to calculate the corresponding cumulative probabilities shown in Eqs. (7) and (8), and then using Eq. (9). Thereafter, the i-th bootstrap replication of the estimated detection cost function $\hat{C}^{i}$ can be computed using Eq. (10).

The same set of target scores, known non-target scores, and unknown non-target scores is employed to compute the estimate of the detection cost function, using Eq. (7) through Eq. (10) where the two thresholds t_1 and t_2 are involved. Therefore, here the same set, rather than two different sets, of resampled target scores, known non-target scores, and unknown non-target scores is used to calculate the i-th estimates $\hat{W}^{i}(t_1)$ and $\hat{W}^{i}(t_2)$, and thus the i-th bootstrap replication of the detection cost function, $\hat{C}^{i}$, in Step 14.

Finally, as indicated in Step 16, from the set $\{\hat{C}^{i} \mid i = 1, \ldots, B\}$, the standard error $S\hat{E}$ of the detection cost function is estimated by the sample standard deviation of the B replications, and the $(1 - \alpha)\times 100$ % confidence interval (CI) $(\hat{Q}(\alpha/2), \hat{Q}(1-\alpha/2))$ at the significance level α is estimated by the $\alpha/2\times 100$ % and $(1 - \alpha/2) \times 100$ % quantiles of the bootstrap distribution [5]. Definition 2 of quantile in Ref. [16] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities. If 95 % $C\hat{I}$ is of interest, α is set to be 0.05.

With the adjusted data structure shown in Section 2.4, the same number of target scores, the same number of known non-target scores, and the same number of unknown non-target scores are obtained, respectively, in Step 14 to compute the estimate of the bootstrap replication of the detection cost function at different iterations of the nonparametric three-sample two-layer bootstrap. This can reduce the variance of the computation.

This algorithm can be easily modified so that it can be applied to the case where the data are assumed to be i.i.d., simply replacing Steps 2 to 13 by

WR_Random_Sampling_Set (N_T, $\boldsymbol{T}$, $\boldsymbol{\Theta^i}$)
WR_Random_Sampling_Set (N_K, $\boldsymbol{K}$, $\boldsymbol{\Xi^i}$)
WR_Random_Sampling_Set (N_U, $\boldsymbol{U}$, $\boldsymbol{\Psi^i}$)

where $\boldsymbol{T}$ is the set of all N_T original target scores, $\boldsymbol{K}$ is the set of all N_K original known non-target scores, and $\boldsymbol{U}$ is the set of all N_U original unknown non-target scores as shown in Eq. (3). That is, in the i-th iteration by calling the function WR_Random_Sampling_Set three times, N_T target scores are randomly selected WR from the set $\boldsymbol{T}$ to form a new set $\boldsymbol{\Theta^i}$, N_K known non-target scores are randomly selected WR from the set $\boldsymbol{K}$ to constitute a new set $\boldsymbol{\Xi^i}$, and N_U unknown non-target scores are randomly selected WR from the set $\boldsymbol{U}$ to constitute a new set $\boldsymbol{\Psi^i}$. In the meantime, Step 14 is replaced by

$\boldsymbol{\Theta^i}$, $\boldsymbol{\Xi^i}$, and $\boldsymbol{\Psi^i}$ => $\hat{W}^i(t_1)$ and $\hat{W}^i(t_2)$ => $\hat{C}^i$.

That is, all target scores, known non-target scores, and unknown non-target scores in these three new sets $\boldsymbol{\Theta^i}$, $\boldsymbol{\Xi^i}$, and $\boldsymbol{\Psi^i}$ are employed to generate the i-th bootstrap replication of the estimated detection cost function at two given thresholds, $\hat{C}^i$, using Eqs. (7) through (10).

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

5. Results with data dependency

A total of 47 speaker recognition primary core systems were evaluated as examples. While scores are generated for all trials, the matching process may be simply regarded as a dichotomous response (i.e., yes or no) with respect to a specified threshold.

Five systems had all scores for all trials below both thresholds. This indicates that these systems declared all trials to involve different speakers by "just saying no." In this case, the two probabilities of type I error corresponding to the two thresholds equal 1, and all the probabilities of type II error equal 0. Further, given the formulation and the parameter settings specified in Eq. (9) and Eq. (10) of Section 3, the detection cost function is fixed at 0.0055.

One system had all scores above both thresholds. This implies that this system claimed all trials to involve the same speakers by "just saying yes." Hence, the two probabilities of type I error equal 0, and all the probabilities of type II error equal 1. Moreover, by Eq. (9) and Eq. (10), the detection cost function is fixed at 0.9945.

sys	cost function SÊ (relative error) 95 % CÎ	sys	cost function SÊ (relative error) 95 % CÎ	sys	cost function SÊ (relative error) 95 % CÎ	sys	cost function SÊ (relative error) 95 % CÎ
1	0.001949 0.000172 (17.33 %) (0.001622, 0.002298)	11	0.002884 0.000179 (12.18 %) (0.002539, 0.003260)	21	0.004515 0.000113 (4.90 %) (0.004286, 0.004735)	31	0.005517 0.000137 (4.86 %) (0.005006, 0.005533)
2	0.001958 0.000169 (16.96 %) (0.001630, 0.002289)	12	0.003094 0.000206 (13.08 %) (0.002677, 0.003483)	22	0.004630 0.000175 (7.43 %) (0.004296, 0.004992)	32	0.005954 0.000296 (9.74 %) (0.005519, 0.006612)
3	0.002278 0.000194 (16.72 %) (0.001894, 0.002663)	13	0.003170 0.000208 (12.86 %) (0.002761, 0.003564)	23	0.004645 0.000099 (4.20 %) (0.004449, 0.004825)	33	0.007756 0.000795 (20.09 %) (0.006379, 0.009505)
4	0.002408 0.000160 (12.98 %) (0.002108, 0.002736)	14	0.003468 0.000208 (11.76 %) (0.003052, 0.003886)	24	0.004648 0.000088 (3.72 %) (0.004474, 0.004811)	34	0.010172 0.000907 (17.48 %) (0.008656, 0.012209)
5	0.002571 0.000190 (14.46 %) (0.002219, 0.002944)	15	0.003591 0.000227 (12.40 %) (0.003154, 0.004023)	25	0.004883 0.000100 (4.00 %) (0.004689, 0.005072)	35	0.010612 0.000699 (12.90 %) (0.009345, 0.011984)
6	0.002602 0.000220 (16.54 %) (0.002186, 0.003035)	16	0.003827 0.000200 (10.24 %) (0.003439, 0.004222)	26	0.004884 0.000071 (2.86 %) (0.004742, 0.005022)	36	0.011227 0.000412 (7.19 %) (0.010451, 0.012091)
7	0.002675 0.000207 (15.19 %) (0.002278, 0.003082)	17	0.004004 0.000174 (8.52 %) (0.003661, 0.004346)	27	0.005096 0.000075 (2.87 %) (0.004936, 0.005226)	37	0.045313 0.001152 (4.98 %) (0.042995, 0.047557)
8	0.002678 0.000220 (16.11 %) (0.002248, 0.003133)	18	0.004049 0.000236 (11.43 %) (0.003590, 0.004509)	28	0.005369 0.000023 (0.83 %) (0.005321, 0.005409)	38	0.052869 0.001372 (5.09 %) (0.050327, 0.055650)
9	0.002800 0.000165 (11.57 %) (0.002490, 0.003143)	19	0.004087 0.000246 (11.81 %) (0.003596, 0.004561)	29	0.005410 0.000084 (3.03 %) (0.005257, 0.005577)	39	0.075391 0.005255 (13.66 %) (0.065085, 0.085735)
10	0.002847 0.000184 (12.67 %) (0.002484, 0.003198)	20	0.004170 0.000131 (6.17 %) (0.003903, 0.004421)	30	0.005491 0.000223 (7.97 %) (0.004989, 0.005494)	40	0.676352 0.008992 (2.61 %) (0.657825, 0.693410)

Table 4 The estimated cost functions, SÊs (relative errors), and 95 % CÎs of 40 speaker recognition systems numbered according to their performance levels in descending order after taking account of the data dependency where the uncertainties were computed using the nonparametric three-sample two-layer bootstrap method and the relative error was approximately estimated by 1.96 times SÊ divided by the cost function.

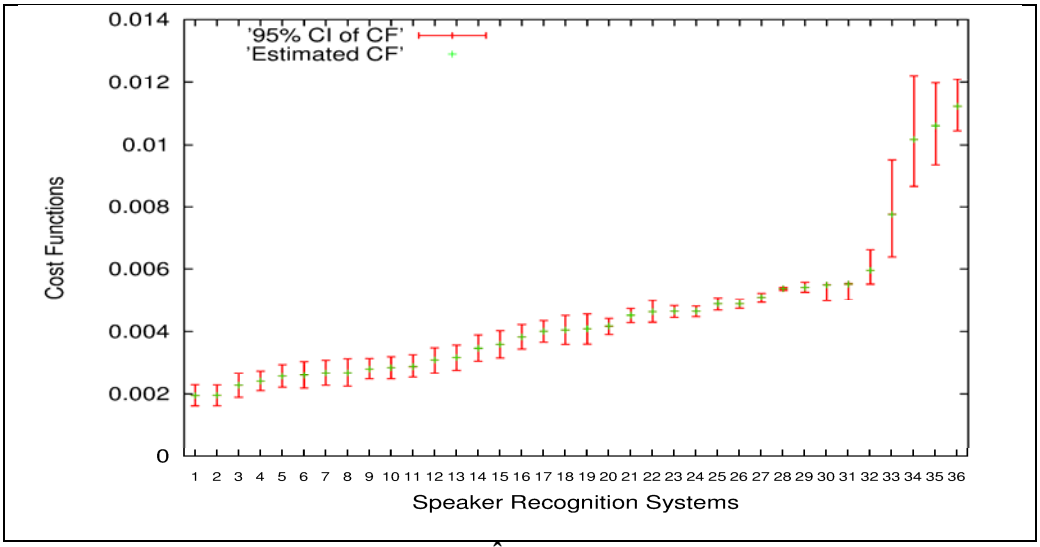

Figure 3 The estimated cost functions and 95 % CÎs of the first 36 speaker recognition systems after taking account of the data dependency where the uncertainties were computed using the nonparametric three-sample two-layer bootstrap method.

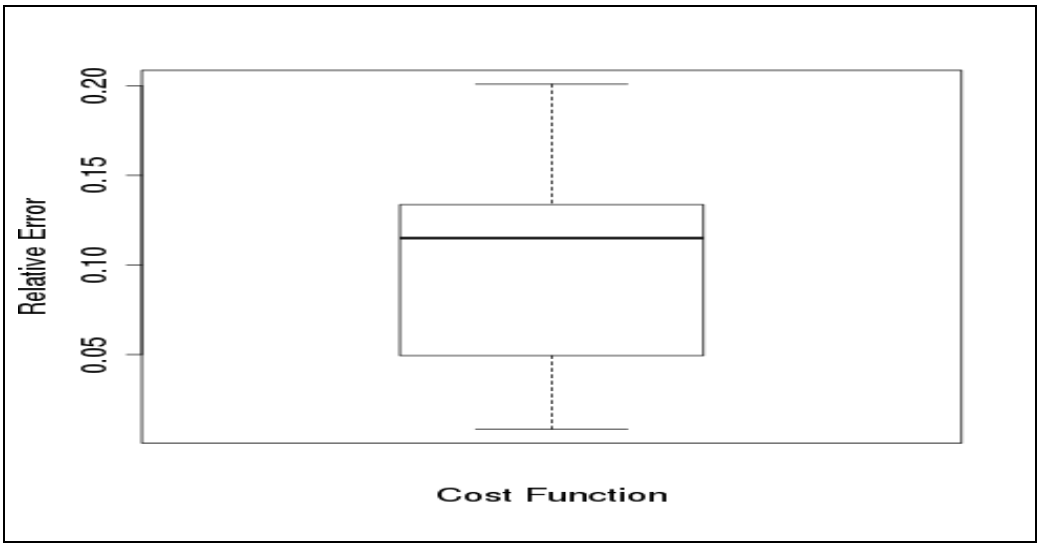

Figure 4 A box diagram of relative errors of the cost functions for the 40 speaker recognition systems.

As a result, for all six of the above systems, no SE estimate associated with the performance accuracy could be obtained. Moreover, one out of 47 systems contained illegitimate scores such as "inf" and "nan." Therefore, these seven speaker recognition systems were not analyzed further.

The performance accuracies and uncertainties of the detection cost function of the remaining 40 speaker recognition systems were investigated. In Table 4 are shown the estimated detection cost functions, SÊs (relative errors), and 95 % CÎs of these 40 speaker recognition systems numbered according to their performance levels in descending order after taking account of the data dependency. The uncertainties were computed using the nonparametric three-sample two-layer bootstrap method and the relative error was approximately estimated by 1.96 times SÊ divided by the cost function.

In Figure 3 are depicted the estimated cost functions and 95 % CÎs of the first 36 speaker recognition systems after taking account of the data dependency while the uncertainties were computed using the nonparametric three-sample two-layer bootstrap method. If the other four systems were also included in this figure, the estimated 95 % CÎs of the first 32 systems would be too small to be seen clearly.

Only the top 30 speaker recognition systems achieved performance accuracies better than that of the "just saying no" system, whose detection cost function is 0.0055. However, the other ten systems performed even worse than the "just saying no" system.

Figure 4 depicts a box diagram of the relative errors of these cost functions for the 40 speaker recognition systems. The 25th percentile, the mean, and the 75th percentile of the 40 relative errors are 4.95 %, 11.50 %, and 13.37 %, respectively. The mean is 10.03 %. The largest is 20.09 %, the second smallest is 2.61 %, and the smallest is 0.83 %. The difference between the second smallest and the smallest is quite large.

The smallest relative error is for System 28. Its estimated 95 % CÎ is very narrow. This system had all known and unknown non-target scores below both thresholds, i.e., always saying "NO" for each non-target trial. Thus, no matter how these non-target scores are resampled while the bootstrap method takes place as described in Section 4, the corresponding four probabilities of type II error, $\beta_j\ (t_i)$, $j \in \{K, U\}$ and $i = 1, 2$, in Eq. (8), are always zero. In other words, all these non-target scores have no impact on the variance of the detection cost function.

6. Results assuming the data are i.i.d.

sys	cost function SÊ 95 % CÎ	sys	cost function SÊ 95 % CÎ	sys	cost function SÊ 95 % CÎ	sys	cost function SÊ 95 % CÎ
1	0.001724 0.000018 (0.001688, 0.001760)	11	0.002635 0.000019 (0.002597, 0.002674)	21	0.004460 0.000012 (0.004436, 0.004483)	31	0.005514 0.000040 (0.005509, 0.005519)
2	0.001710 0.000015 (0.001681, 0.001740)	12	0.002623 0.000016 (0.002590, 0.002653)	22	0.004311 0.000026 (0.004259, 0.004360)	32	0.005927 0.000020 (0.005888, 0.005964)
3	0.002066 0.000018 (0.002031, 0.002101)	13	0.002925 0.000022 (0.002882, 0.002969)	23	0.004665 0.000011 (0.004644, 0.004686)	33	0.007542 0.000034 (0.007474, 0.007609)
4	0.002411 0.000016 (0.002380, 0.002444)	14	0.003196 0.000025 (0.003146, 0.003246)	24	0.004458 0.000010 (0.004437, 0.004477)	34	0.011320 0.000054 (0.011215, 0.011434)
5	0.002432 0.000016 (0.002401, 0.002463)	15	0.002990 0.000020 (0.002951, 0.003030)	25	0.004869 0.000008 (0.004852, 0.004885)	35	0.011764 0.000085 (0.011598, 0.011932)
6	0.002297 0.000024 (0.002252, 0.002346)	16	0.003369 0.000023 (0.003322, 0.003416)	26	0.004785 0.000009 (0.004767, 0.004802)	36	0.011198 0.000059 (0.011083, 0.011316)
7	0.002240 0.000017 (0.002206, 0.002271)	17	0.003816 0.000012 (0.003791, 0.003841)	27	0.005143 0.000007 (0.005129, 0.005156)	37	0.039986 0.000154 (0.039685, 0.040295)
8	0.002225 0.000021 (0.002183, 0.002267)	18	0.003476 0.000021 (0.003436, 0.003518)	28	0.005381 0.000004 (0.005373, 0.005389)	38	0.046666 0.000148 (0.046395, 0.046966)
9	0.002667 0.000018 (0.002631, 0.002703)	19	0.003499 0.000023 (0.003456, 0.003543)	29	0.005384 0.000013 (0.005358, 0.005412)	39	0.068341 0.000212 (0.067934, 0.068774)
10	0.002396 0.000015 (0.002366, 0.002427)	20	0.004136 0.000015 (0.004106, 0.004164)	30	0.005492 0.000174 (0.004991, 0.005494)	40	0.669390 0.000399 (0.668628, 0.670186)

Table 5 The cost functions, SÊs, and 95 % CÎ of 40 speaker recognition systems using all the raw data while the uncertainties were computed using the nonparametric three-sample bootstrap method with the assumption that the data were i.i.d.. The systems were numbered in the same order as in Table 4.

In Table 5 are shown the estimated cost functions, SÊs, and 95 % CÎ of 40 speaker recognition systems using all the raw data while the uncertainties were computed using the nonparametric three-sample bootstrap method with the assumption that the data were i.i.d.. The systems were numbered in the same order as in Table 4.

7. Comparisons of the two results

The results obtained in Section 5 are compared with those obtained in Section 6 in terms of both performance accuracy and measurement uncertainty.

7.1 The aspect of the performance accuracy – the detection cost functions

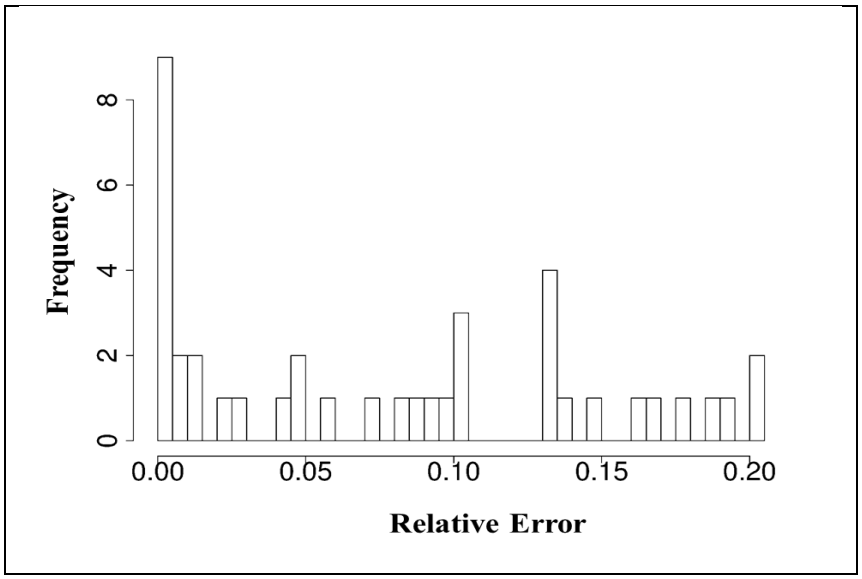

Figure 5 Histogram of the absolute values of the relative changes in performance accuracy as a result of adjusting the dataset for 40 systems.

In Table 5, the total numbers of target scores, known non-target scores, and unknown non-target scores employed were 41 897, 1 291 587, and 407 827, respectively. In Table 4, the total numbers of target scores, known non-target scores, and unknown non-target scores used were 18 430, 609 112, 287 182, respectively. The numbers of all three types of scores were approximately cut in half in each category as a result of grouping the scores into sets based on the data dependency in consideration of the probabilities of scores being selected equally and randomly, as discussed in Sections 2.3 and 2.4.

The relative changes in performance accuracy as a result of adjusting the dataset is defined as $(C_{Det\ after} - C_{Det\ before}) / C_{Det\ before}$. The histogram of their absolute values for 40 systems is depicted in Figure 5.

It shows that only 17.5 % of all the cases, i.e., 7 cases out of 40, had absolute relative changes between 15 % and around 20 %. Among those with the highest changes around 20 % are Systems 8 and 15, whose cost functions are 0.002225 and 0.002990 before adjusting the dataset, and 0.002678 and 0.003591 after adjusting the dataset, respectively.

On the other hand, 82.5 % of cases had absolute relative changes below 15 %. Indeed, 67.5 % of cases had absolute relative changes less than about 10 %, and 45 % of cases had absolute relative changes less than 5 %.

All these indicate that reducing the size in about half due to adjustment of datasets has little impact on the performance accuracies of the speaker recognition systems, i.e., the estimated detection cost function. This is consistent with the conclusion reached in our prior research regarding the sample

size using Chebyshev's inequality when the data size reaches to over tens and hundreds of thousands in ROC analysis [14]. It also suggests that the random selection while reducing the dataset size (See Section 2.4) has little impact on the results.

This is consistent with the observation made in Section 2.4, where Figure 2 shows that the three distributions of scores before adjustment and after adjustment remains almost the same, except that the tails after adjustment are shorter than those before adjustment. This is because the measures in ROC analysis generally depend on the relative positions among different distributions of scores [13].

7.2 The aspect of the measurement uncertainty – the SEs

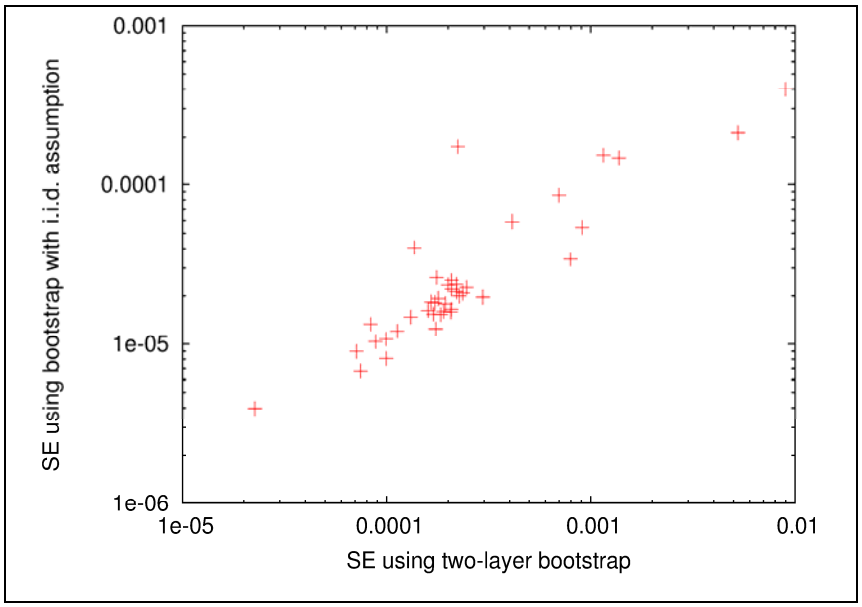

Figure 6 The logarithmic-scale scatterplot of SEs computed using the nonparametric three-sample two-layer bootstrap method with about half of the raw data grouped into score sets against SEs calculated using the nonparametric three-sample bootstrap with i.i.d. assumption with all raw data for 40 systems.

Furthermore, the difference in terms of the measurement uncertainty was explored. In Table 5, the data are assumed to be i.i.d., and thus the uncertainty of the detection cost function is calculated using the nonparametric three-sample bootstrap algorithm with all raw data. In Table 4, the data dependency is taken into account, and hence the uncertainty of the measure is computed using the nonparametric three-sample two-layer bootstrap algorithm with about half of the raw data grouped into score sets.

By comparing those estimated SÊs shown in both Table 4 and Table 5, it can be seen that the measurement uncertainties, i.e., the estimated SÊs, using all scores under the i.i.d. assumption are much smaller than those taking account of data dependency, for all systems. For instance, for System 1, the estimated SÊ is 0.000018 in Table 5, but 0.000172 in Table 4. This can also clearly be shown in Figure 6, the logarithmic-scale scatterplot of SEs computed in the two different scenarios as stated above for 40 systems.

The estimated $\hat{SE}$ is inversely proportional to the square root of the data size n, i.e., $\hat{SE} \sim 1 / \sqrt{n}$. The data size used in Table 4 is about half of the data size employed in Table 5. Even taking account of this effect of data size, i.e., multiplying the estimated $\hat{SE}$s in Table 5 by a factor of $\sqrt{2}$ (i.e., 1.414214), the resultant $\hat{SE}$s are still far smaller than those in Table 4. For instance, for System 1, the resultant $\hat{SE}$ is 0.000018 x 1.414214 = 0.000026, which is about 6.66 times smaller than 0.000172.

Therefore, the measurement uncertainties will be underestimated without taking account of the data dependency. This is consistent with the conclusions reached in our previous research [7]. If the same subjects are used multiple times (except for generating target trial scores) in creating datasets, to compute the measurement uncertainty, this data dependency must be taken into account.

8. Conclusions and discussion

In SRE12, to evaluate the performance of speaker recognition systems, a new paradigm was employed in three aspects. First, three distributions of target scores, known non-target scores, and unknown non-target scores were created. Second, dichotomous responses were determined at two thresholds. Third, the new detection cost function was defined to be an average of the two weighted sums of the probabilities of type I (miss) and type II (false alarm) errors corresponding to the two decision thresholds.

In reality, data dependency may be inevitable. This is due to the need for multiple use of the same training speakers in order to generate more scores. The three different types of scores are grouped into sets based on the training speakers, and in each the numbers of scores for different training speakers are kept the same, respectively. The former is due to the bootstrap method, and the latter is because of the probability theory. The new datasets have a two-layer structure, and the sizes of the raw data are approximately cut in half.

Thus, scores in each type can be selected with equal probability, and the numbers of target scores, known non-target scores, and unknown non-target scores resampled at each iteration can be the same. This reduces the variance of the computation and ensures the accuracy of the calculation. Moreover, this preserves the data dependency while the bootstrap resampling takes place. The uncertainty in terms of the SE and the 95 % CI of the detection cost function is computed using the nonparametric three-sample two-layer bootstrap method, where the empirical distribution is assumed for each of the observed scores, based on our extensive bootstrap variability studies in ROC analysis on large datasets.

The matching process in the speaker recognition may be simply regarded as a dichotomous response (i.e., yes or no) with respect to a specified threshold. Five out of 47 speaker systems declared all trials to involve different speakers by "just saying no" and one system claimed all trials to involve the same speakers by "just saying yes." Because of the formulation and the parameter settings in Eq. (9) and Eq. (10), the corresponding detection cost functions were fixed at 0.0055 and 0.9945, respectively. In these two cases, no SE estimate associated with the performance accuracy could be obtained. Furthermore, by comparing with other systems as illustrated in Table 4 and Figure 3, there were ten systems whose performance was worse than a "just saying no" system.

Figure 2 shows the three distributions of the target scores, known non-target scores, and unknown non-target scores of System 1 before adjustment and after adjustment. System 1 is the best among the 40 systems. It indicates that good systems can try to separate the distribution of target scores from the distributions of known and unknown non-target scores as apart as possible, and thus significantly outperform "just saying no" systems. This is consistent with our prior results [13].

The relative changes in the detection cost functions for 40 speaker recognition systems vary between 2.61 % and 20.09 %, around a median of 11.50 %, except for the smallest one that is 0.83 % for System 28. System 28 had all known and unknown non-target scores below both thresholds, i.e., always saying "NO" for each non-target trial. Thus, its four probabilities of type II errors are all zero. For this system, only the target scores have impact on the variance of the detection cost function. This is consistent with the consequence of using the bootstrap algorithm as presented in Section 4.

The results of the estimated detection cost functions of the speaker recognition systems and their uncertainties computed using the nonparametric three-sample two-layer bootstrap method using approximately half the raw data due to adjusting the original raw datasets into a new two-layer data structure were compared with those calculated using the nonparametric three-sample bootstrap method under the i.i.d. assumption using all the raw data in terms of performance accuracy and measurement uncertainty.

Cutting the size of the datasets by approximately half has little impact on the relative positions among the three distributions of scores, and thus on the performance accuracies of the speaker recognition systems, namely, the estimated detection cost functions. This is because the measures in ROC analysis generally depend on the relative positions among different distributions of scores [13]. It is consistent with the conclusions obtained in our prior research regarding the sample size using Chebyshev's inequality [14]. When the data size reaches to over tens and hundreds of thousands in ROC analysis, the measured accuracy will improve little. It also suggests that the random selection while reducing the dataset size has little impact on the results.

However, the measured uncertainty after taking account of data dependency is clearly larger than that obtained without taking the data dependency into account even in view of the reduction of data size. These studies involved tens of thousands of target scores, known non-target scores, and unknown non-target scores. The large size of these datasets does little to reduce the impact of data dependency on the uncertainties of measures in ROC analysis. If data dependency is involved, the bootstrap method with the i.i.d. assumption underestimates the uncertainties of measures.

It may be concluded that if subjects are employed multiple times in generating datasets (except for creating target trial scores), issues of data dependency should be taken into account, the dataset needs to be adjusted in a way such as that suggested in this article, and then the nonparametric three-sample two-layer bootstrap method may be implemented to compute the measurement uncertainties. This is consistent with the conclusions reached in our previous research [7].

References

1. "The NIST Speaker Recognition Evaluation", http://www.itl.nist.gov/iad/mig/tests/spk/ (2012).
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, Lang. Process.*, vol. 15, no. 7, pp. 1951-1959, Sep. 2007.
3. C.S. Greenberg, A.F. Martin, G.R. Doddington, and J.J. Godfrey, "Including human expertise in speaker recognition systems: report on a pilot evaluation," in *Proc. ICASSP*, pp. 5896-5899, 2011.
4. B. Efron, "Bootstrap methods: Another look at the Jackknife," *Ann. Statistics*, vol. 7, no. 1, pp. 1-26, 1979.
5. B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.
6. J. C. Wu, A. F. Martin, and R. N. Kacker, "Bootstrap variability studies in ROC analysis on large datasets," *Communications in Statistics - Simulation and Computation*, vol. 43, no. 1, pp. 225–236, 2014.
7. J.C. Wu, A.F. Martin, C.S. Greenberg, and R.N. Kacker, "Data dependency on measurement uncertainties in speaker recognition evaluation," in Active and Passive Signatures III, Proc. SPIE 8382, 83820D, 2012.
8. 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*., vol. 116, no. 1, pp. 517–537, 2011.
9. 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, Sep. 2007.
10. 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.
11. R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, and A.W. Senior, *Guide to Biometrics*. Springer, New York, pp. 269-292, 2003.
12. 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*, vol. 29, no. 3, pp. 492-498, 2007.
13. J.C. Wu and C.L. Wilson, "Nonparametric analysis of fingerprint data on large data sets," *Pattern Recognition*, vol. 40, no. 9, pp. 2574-2584, 2007.
14. 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.
15. B. Ostle and L.C. Malone, *Statistics in Research: Basic Concepts and Techniques for Research Workers, fourth ed.*. Iowa State University Press, Ames, 1988.
16. R.J. Hyndman and Y. Fan, "Sample quantiles in statistical packages," *American Statistician*, vol. 50, pp. 361-365, 1996.