NISTIR 7722

Measurement Uncertainties in Speaker Recognition Evaluation

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

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

NISTIR 7722

Measurement Uncertainties in Speaker Recognition Evaluation

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

September 2010

U.S. Department of Commerce
Gary Locke, Secretary

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

Measurement Uncertainties in Speaker Recognition Evaluation

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

Abstract – The National Institute of Standards and Technology (NIST) Speaker Recognition Evaluations (SRE) are an ongoing series of projects conducted by NIST. In the NIST SRE, speaker detection performance is measured using a detection cost function, which is defined as a weighted sum of probabilities of type I error and type II error. The sampling variability results in measurement uncertainties of the detection cost function. Hence, while evaluating and comparing the performances of speaker recognition systems, the measurement uncertainties must be taken into account. In this article, the uncertainties of detection cost functions in terms of standard errors (SE) and confidence intervals are computed using the nonparametric two-sample bootstrap methods based on our extensive bootstrap variability studies on large datasets conducted before. The data independence is assumed because the bootstrap results of SEs matched very well with the analytical results of SEs using the Mann-Whitney statistic for independent and identically distributed samples if the metric of area under a receiver operating characteristic curve is employed. Examples are provided.

Index Terms -- speaker recognition evaluation, biometrics, bootstrap, uncertainty, standard error, confidence interval.

1 Introduction

The National Institute of Standards and Technology (NIST) Speaker Recognition Evaluations (SRE) are an ongoing series of projects conducted by NIST [1]. These evaluations have been making 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.

The 2008 NIST SRE consisted of 13 tests. Each test consisted of a sequence of trials, where each trial consisted of a target speaker, defined by the training data provided, and a test segment. For each trial, the system to be evaluated needed to decide whether speech of the target speaker occurred in the test segment, and generate a similarity score, where a higher similarity score indicated greater confidence that the speech of target speaker occurred in the test segment.

Among the 13 tests given in the 2008 NIST SRE, there was a single core test formed by short2 of training conditions and short3 of test segment conditions, for which all participants were required to submit results. The details of the evaluation plan can be found in Ref. [1]. Therefore, this core test was of interest, and in this article all speaker recognition data described were taken from this core test.

In the 2008 NIST SRE, speaker detection performance is measured using a detection cost function. It is defined as a weighted sum of probabilities of type I error (miss) and type II error (false alarm) [1].

As is well-known, the sampling variability results in uncertainties of any measures [2]. That is to say, if sets of samples are collected under the same circumstances, the measures in evaluation may fluctuate. This happens in SRE as well. Hence, while evaluating and comparing the performances of speaker recognition systems, the measurement uncertainties must be taken into account. Now, the key issue is how to calculate the uncertainties of detection cost functions in terms of standard errors (SE) and confidence intervals (CI).

It is hard to compute analytically the covariance term (i.e., the cross term) of correlated probabilities of type I error and type II error, the linear combination of which forms the detection cost function in SRE. As a result, it is difficult to calculate the variance of such a detection cost function analytically.

In the evaluation and comparison of matching algorithms in biometrics in general and in fingerprint technology in particular, the receiver operating characteristic (ROC) analysis is an important statistical technique. In the operational ROC analysis, the uncertainties of measures, such as the true accept rate and the false accept rate in different circumstances, as well as the equal error rate, etc., can be computed using the nonparametric two-sample bootstrap methods based on extensive bootstrap variability studies with large datasets [3-8].

The two samples are referred to as a set of target (i.e., genuine) scores and a set of non-target (i.e., impostor) scores, and they constitute two distributions [7, 8]. An ROC curve is characterized by the relative relationship between these two distributions [9]. These two distribution functions are indeed interrelated by the algorithm that generates them. In other words, the performance of a matching algorithm is affected not only by target matching but also by non-target matching. All statistics of interest in ROC analysis in general as well as in SRE in particular are influenced under the combined impact of these two samples.

Furthermore, it is known from previous studies that these two distributions 1) usually do not have well defined parametric forms; 2) may be considerably different even for the same algorithm; and 3) may vary substantially from algorithm to algorithm, which differentiates algorithms in terms of matching accuracy [10]. The same variations of distributions were observed in the speaker recognition data. This suggests that the nonparametric statistical analysis is appropriate for evaluating speaker recognition data, namely, the empirical distribution is assumed for each of the observed scores.

Therefore, in this article, the uncertainties of detection cost functions, in terms of SEs and CIs, are also computed using the nonparametric two-sample bootstrap methods. The bootstrap method assumes that an independent and identically distributed (i.i.d.) random sample of size *n* is drawn from a population with its own probability distribution. With the i.i.d. assumption, the units of nonparametric two-sample bootstrap are scores in the sample. In case of data dependency, the bootstrap units are the subsets of the sample, into which the sample is regrouped based on data dependencies caused by multiple biometric acquisitions [6, 7, 11, 12]. This way can preserve the dependencies among the data. However, everything else in the bootstrap method remains intact. It is most likely that how to regroup the sample into subsets could have impact on the bootstrap results.

An ROC curve can be measured by the area under the ROC curve (AURC) [10, 13-16]. If the trapezoidal rule is employed, the AURC is equivalent to the Mann-Whitney statistic formed by target and non-target scores. The SE of the Mann-Whitney statistic can be computed analytically and utilized as the SE of AURC. Alternatively, the SE of AURC can be calculated using the nonparametric two-sample bootstrap. Using this metric, the bootstrap results of SEs matched very well with the analytical results of SEs using the Mann-Whitney statistic for i.i.d. samples.

This indicates that while computing the uncertainties of the detection cost function in SRE, the i.i.d. assumption could be made for speaker recognition data. In the 2008 NIST SRE, generally speaking, the total number of target scores was about 20 000 and the total number of non-target scores was about 80 000 [17]. The large size of speaker datasets might have reduced the effect of the dependencies among the data.

All similarity scores of the speaker recognition systems are real numbers. While analyzing the data, all real numbers were converted into integers. Different systems employ different numbers of digits in the integer part. Hence, in order to obtain results as accurate as possible, five decimal places (i.e., multiplying 10^5) or up to seven decimal places (i.e., multiplying 10^7) were preserved. Notice that if the largest integer score is too large, the computation can take too much time. This is because it has to go from the highest score down to the threshold provided by a system every time while computing thousands of bootstrap replications of the detection cost function. The probability distribution functions of similarity scores are all discrete [10]. The characteristic of the speaker data is that only a few of similarity scores take the same value [7, 8].

The methods are presented in Section 2, including the formulations of discrete distribution functions of target and non-target scores, the formulas for computing the probabilities of type I error and type II error, the detection cost function in SRE, and the algorithm of the nonparametric two-sample bootstrap for calculating SEs. In Section 3 are provided the results regarding measurement uncertainties, in terms of SEs and 95% CIs, of the detection cost function involving 12 different speaker recognition systems[1] with the i.i.d. assumption for speaker recognition data while performing bootstrap. Conclusions and discussion can be found in Section 4. In the Appendix, in order to support the i.i.d. assumption, the SE of AURC is computed analytically using the Mann-Whitney statistic as well as numerically using the bootstrap method, and the comparison between these two results is carried out.

2 Methods

2.1 Discrete distribution functions of target and non-target scores

After converting to integer scores as mentioned in Section 1, without loss of generality, for a speaker recognition system, the similarity 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} to the highest score s_{max}. Hence, the target score set is denoted as

$$\mathbf{T} = \{ m_i \mid m_i \in \{s\} \text{ and } i = 1, \ldots, M_T\} , \tag{1}$$

where M_T is the total number of target scores. And the non-target score set is expressed as

$$\mathbf{N} = \{ n_i \mid n_i \in \{s\} \text{ and } i = 1, \ldots, M_N\} , \tag{2}$$

where M_N is the total number of non-target scores.

These two sets of similarity scores constitute two discrete probability distribution functions, respectively. Let P_i (s), where $s \in \{s\}$ and $i \in \{T, N\}$, denote the empirical probabilities of the target scores and the non-target scores at a score s, respectively. It may very well be that some of them are zeroes at some scores in the set {s}. Nonetheless, the two distribution functions can be expressed, respectively, as

$$\boldsymbol{P}_{\mathbf{i}} = \{ P_i (s) \mid \forall\, s \in \{s\} \text{ and } \sum_{\tau=s\min}^{s\max} P_i (\tau) = 1 \} , i \in \{T, N\} . \tag{3}$$

The cumulative discrete probability distribution functions of target scores and non-target scores are defined in this article to be the probabilities cumulated from the highest score s_{max} down to the integer score s, and are expressed as

$$\boldsymbol{C}_{\mathbf{i}} = \{ C_i (s) = \sum_{\tau=s}^{s\max} P_i (\tau) \mid \forall\, s \in \{s\} \} , i \in \{T, N\} \tag{4}$$

where C_i (s), $i \in \{T, N\}$, are the cumulative probabilities of target scores and non-target scores at a score s, respectively.

[1] Specific hardware and software products identified in this report 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.

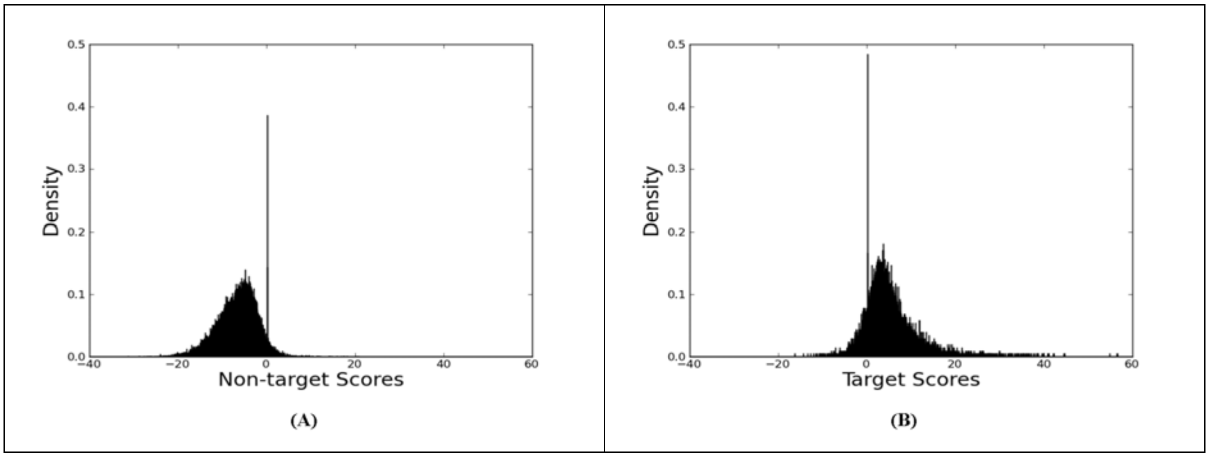

Figure 1 (A): The probability distribution function of non-target scores. (B): The probability distribution function of target scores. Both of them were generated by the speaker recognition system DL. Each of them has a stand-alone peak.

Here is an example regarding the distributions of similarity scores. The probability distribution functions of non-target scores and target scores, which were generated by the speaker recognition system DL in the core test as specified in Section 1, are depicted in Figure 1 (A) and (B), respectively. It was found from the figures that each of these two probability densities has a stand-alone peak near score zero. While we have not yet determined the cause of the peaks in score for system DL, we suspect that it may be the result of the way the system handles anomalous evaluation segments. It is difficult to do parametric data modeling for such distributions.

2.2 Probabilities of type I error and type II error

The probability of type I error at a threshold $t \in \{\mathrm{s}\}$, denoted by $P_I\ (t)$, is cumulated from the lowest similarity score $\mathrm{s}_{\min}$ in our application. The probability of type II error at a threshold t, denoted by $P_{II}\ (t)$, is cumulated from the highest score $\mathrm{s}_{\max}$. For discrete probability distribution functions, while computing $P_I\ (t)$ and $P_{II}\ (t)$ at a threshold t, the probabilities of target scores and non-target scores at this threshold t must be taken into account, respectively [18].

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

$$\begin{aligned} \hat{P}_I\ (t) &= 1 - \mathrm{C_T}\ (t + 1) \\ \hat{P}_{II}\ (t) &= \mathrm{C_N}\ (t) \end{aligned} \qquad \text{for } t \in \{\mathrm{s}\}\ , \tag{5}$$

where $\mathrm{C_T}\ (\mathrm{s}_{\max} + 1) = 0$ is assumed [8]. Based on Eq. (5), in practice, the estimators $\hat{P}_I\ (t)$ and $\hat{P}_{II}\ (t)$ 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 non-target scores, respectively.

2.3 The detection cost function in speaker recognition evaluation

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

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

$$\mathrm{C_{Det}}\ (t) = \mathrm{C_{Miss}} \times P_I\ (t) \times \mathrm{P_{Target}} + \mathrm{C_{FalseAlarm}} \times P_{II}\ (t) \times (1 - \mathrm{P_{Target}})\ . \qquad (6)$$

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

The parameters $\mathrm{C_{Miss}}$ 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 target 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].

2.4 Nonparametric two-sample bootstrap

It is difficult to compute analytically the covariance term of the correlated probabilities of type I error $P_I\ (t)$ and type II error $P_{II}\ (t)$ at a threshold *t* in Eq. (6). It is proposed that the estimates of the uncertainty of the detection cost function at a threshold *t* in terms of SE and 95% CI be computed using the nonparametric two-sample bootstrap [3-8]. The algorithm is as follows.

Algorithm **(Nonparametric two-sample bootstrap)**

```
1: for i = 1 to B do
2:     select $M_T$ scores randomly WR from T to form a set {new $M_T$ target scores}$_i$
3:     select $M_N$ scores randomly WR from N to form a set {new $M_N$ non-target scores}$_i$
4:     {new $M_T$ target scores}$_i$ & {new $M_N$ non-targe scores}$_i$ => statistic $\hat{T}_i$
5: end for
6: $\{ \hat{T}_i \mid i = 1, ..., B \} \Rightarrow \hat{SE}_B$ and $(\hat{Q}_B\ (\alpha/2), \hat{Q}_B\ (1-\alpha/2))$
7: end
```

where B is the number of two-sample bootstrap replications and WR stands for “with replacement”. The original target score set **T** with $\mathrm{M_T}$ scores shown in Eq. (1) and the original non-target score set **N** with $\mathrm{M_N}$ scores shown in Eq. (2) are generated by a speaker recognition system. As shown from Step 1 to 5, this algorithm runs B times. In the i-th iteration, $\mathrm{M_T}$ scores are randomly selected WR from the original target score set **T** to form a new set of $\mathrm{M_T}$ target scores, $\mathrm{M_N}$ scores are randomly selected WR from the original non-target score set **N** to form a new set of M_N non-target scores, and then from these two new sets of similarity scores the i-th bootstrap replication of the estimated statistic of interest, i.e., $\hat{T}_i$, is generated.

In the SRE, the estimated statistic of interest $\hat{T}_i$ is the i-th estimator of the detection cost function at a given threshold. This estimator can be derived using Eq. (6). In this equation, the estimators of the probabilities of type I error and type II error, i.e., $\hat{P}_I\ (t)$ and $\hat{P}_{II}\ (t)$, can be calculated from the two new sets of similarity scores using Eq. (5).

Finally, as indicated in Step 6, from the set $\{\hat{T}_i \mid i = 1, ..., B\}$, the estimator of the SE, i.e., the unbiased standard deviation $\hat{SE}_B$, and the estimators of the α/2 100% and (1 - α/2) 100% quantiles of the bootstrap distribution, denoted by $\hat{Q}_B\ (\alpha / 2)$ and $\hat{Q}_B\ (1 - \alpha / 2)$, respectively, at the significance level α can be calculated. The Definition 2 of quantile in Ref. [19] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities. Thus, $(\hat{Q}_B\ (\alpha / 2), \hat{Q}_B\ (1 - \alpha / 2))$ stands for the estimated bootstrap (1 - α) 100% CÎ. If 95% CÎ 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 accuracy of the computation. In other words, what is the number of the nonparametric two-sample bootstrap replications?

In our applications, such as biometrics and the evaluation of speaker recognition, etc., the sizes of data sets are tens or hundreds of thousands of similarity scores, which are much larger than those in some other applications of bootstrap methods, such as medical decision making, etc.. Moreover, in our applications, the statistics of interest are mostly probabilities or a weighted sum of probabilities, etc. rather than a simple sample mean. And our data samples of similarity scores have no parametric model to fit. Therefore, the bootstrap variability was re-studied, and the further study is underway. Based on our prior empirical bootstrap variability studies, the appropriate number of bootstrap replications B for our applications was determined to be 2000 [4-7].

3 Results

The estimated uncertainties of the detection cost functions in SRE, in terms of SEs and 95 % CIs, were all computed using the algorithm of the nonparametric two-sample bootstrap. In this article, while performing bootstrap, the speaker recognition data are assumed to be i.i.d.. This assumption is supported by the evidences shown in the Appendix. With this assumption, the bootstrap units are similarity scores in the data set. Hence, the nonparametric two-sample bootstrap algorithm in Section 2.4 can be employed without any modification.

In Table 1 are shown the estimated detection cost functions, their estimated SÊs, and 95 % CÎs of 12 speaker recognition systems, named as UJ, EL, etc., in the core test short2-short3 for primary actual decision [1]. The estimated detection cost functions were derived using Eq. (6), in which all parameters were set in Section 2.3 and the thresholds were all provided by speaker recognition systems.

Systems	Cost Functions	SÊs	95% CÎs
UJ	0.030810	0.000436	(0.029987, 0.031657)
EL	0.033668	0.000594	(0.032523, 0.034853)
BK	0.036482	0.000484	(0.035502, 0.037447)
DL	0.039645	0.000457	(0.038792, 0.040574)
LZ	0.052373	0.000694	(0.050977, 0.053707)
AF	0.066043	0.000436	(0.065190, 0.066864)
FI	0.093903	0.000233	(0.093445, 0.094345)
PB	0.103623	0.000789	(0.102058, 0.105197)
PM	0.110816	0.001025	(0.108707, 0.112812)
CH	0.144010	0.001150	(0.141677, 0.146164)
CO	0.146433	0.001155	(0.144172, 0.148688)
DG	0.328201	0.001650	(0.325022, 0.331460)

Table 1 The estimated detection cost functions, SÊs, and 95 % CÎs of 12 speaker recognition systems in the core test short2-short3 for primary actual decision.

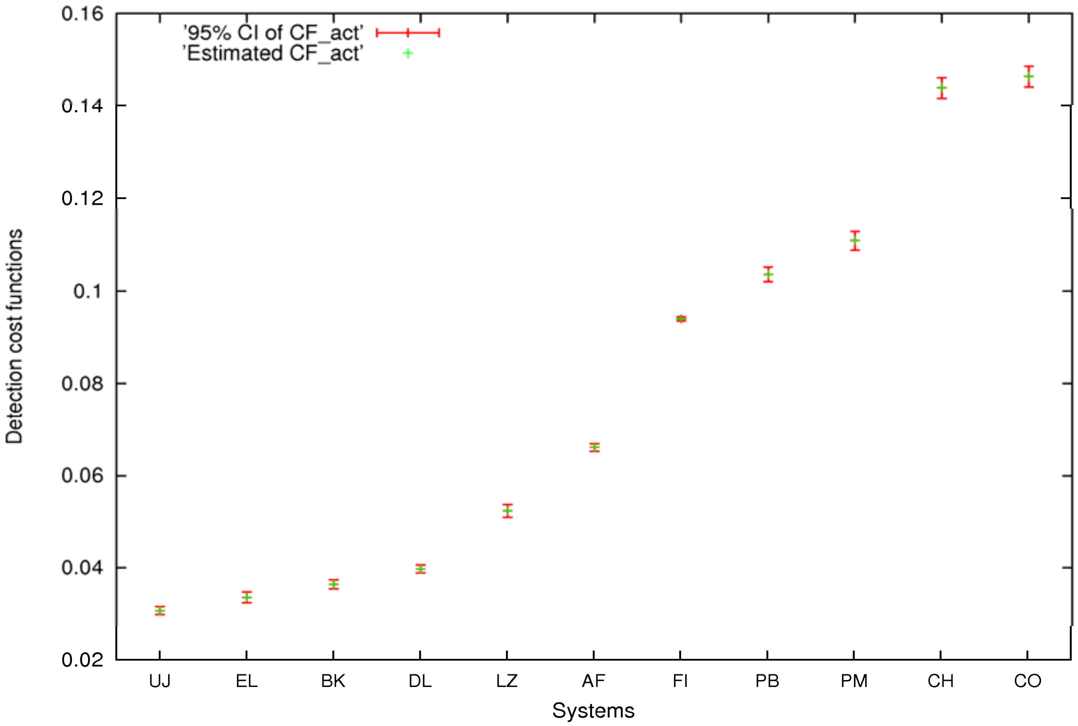

Figure 2 The estimated detection cost functions, and 95 % CÎs of 11 speaker recognition systems in the core test short2-short3 for primary actual decision.

In Table 1, the speaker recognition systems are listed in the ascending order of the cost function. The smaller the detection cost functions, the more accurate the speaker recognition systems. As shown in Table 1, generally speaking, the smaller the detection cost functions, the smaller the uncertainties. This is consistent with what was observed in previous studies [7-10, 17].

It is worth pointing out that the estimated 95 % CÎs shown in Table 1 were calculated using the Definition 2 of quantile as indicated in Section 2.4. The estimated 95 % CÎs could also be computed by assuming that the distribution of 2 000 bootstrap replications of the detection cost function was normal, namely, obtained by multiplying 1.96 and the estimated SÊ. It was found that these two types of 95 % CÎs were matched up to the third to fourth decimal place for all 12 systems shown in Table 1. For instance, for system UJ, the 95 % CÎ derived from the quantile method was (0.029987, 0.031657) as shown in Table 1, and the 95 % CÎ with normality assumption was (0.029955, 0.031665).

In Figure 2 are depicted the estimated detection cost functions, and their estimated 95 % CÎs of 11 speaker recognition systems in the core test short2-short3 for primary actual decision. The estimated detection cost function of the speaker recognition system DG is 0.328201, which is much larger than all others. In order to show the scales of the estimated 95 % CÎ of all other 11 systems, the system DG is not shown in Figure 2. In Figure 2, it also shows that the estimated 95 % CÎ of system CH overlaps the one of system CO. This can happen very often while comparing and evaluating the performances of speaker recognition systems.

4 Conclusions and discussion

Like the applications of ROC analysis in biometrics [2, 7, 8], the uncertainties of the detection cost function, in terms of SEs and 95 % CIs, in the 2008 NIST SRE were successfully computed using the nonparametric two-sample bootstrap. Such a detection cost function is defined as a weighted sum of probabilities of type I error and type II error. Thus, it is hard to compute its variance analytically.

In this article, the bootstrap method was carried out with the i.i.d. assumption for the speaker recognition datasets. Hence, the bootstrap units are similarity scores in the samples rather than subsets into which the data were regrouped according to the dependencies inside the data.

Such an assumption was supported by the fact as shown in the Appendix. The sizes of target scores and non-target scores in speaker recognition data are tens of thousands. It seems that the large size of the datasets could reduce the effect of the dependencies that existed among the data.

Nonetheless, our investigation is underway to apply the nonparametric two-sample bootstrap method without the i.i.d. assumption to computing the uncertainties of the detection cost function in SRE, and then compare the measurement uncertainties derived from different ways. As a matter of fact, from the statistical point of view, the sample should be collected as randomly as possible in test design.

As discussed in Section 3, the estimated 95 % CÎ in SRE were computed using two approaches: one was derived from the quantile method and the other was computed with the assumption of the normal distribution of detection cost functions. These two types of 95 % CÎs were matched up to the third to fourth decimal place for all 12 systems. Moreover, the Shapiro-Wilk normality test [20] was conducted on the 2000 bootstrap replications of the detection cost functions for all

12 systems, and it was found that the majority of p-values were greater than 5 %. It indicates that the detection cost function is normally distributed.

As a consequence, the hypothesis testing can be used to evaluate and compare the performances of speaker recognition systems [2]. Further, the 95 % CIs of two systems may very well overlap, as shown in examples in Section 3. Hence, the hypothesis testing is an important approach.

Appendix – the SE of AURC

As discussed in Section 1, an ROC curve can be characterized by AURC. The SE of AURC for the speaker data can be computed in two ways. One is analytical way using the SE of the Mann-Whitney statistic for i.i.d. samples; the other is numerical way using nonparametric two-sample bootstrap method in which the bootstrap units are scores in the data set. The two results matched very well. It indicates that the i.i.d. assumption for these speaker data could be made while using the two-sample bootstrap method to calculate the uncertainties of the detection cost function.

A.1 Analytical computation of SE of AURC

It is assumed that the trapezoidal rule is employed while computing AURC. Then, the AURC is equivalent to the Mann-Whitney statistic directly formed from the discrete target and non-target scores. This method of computing AURC is widely used [21]. Further, the variance of the Mann-Whitney statistic can be computed analytically. Thus, it can be utilized as the variance of AURC. All related formulas for analytically computing the SE of AURC can be found in the references [10, 13-16]. For convenience, they are also listed in this Appendix.

A.1.1 Compute AURC

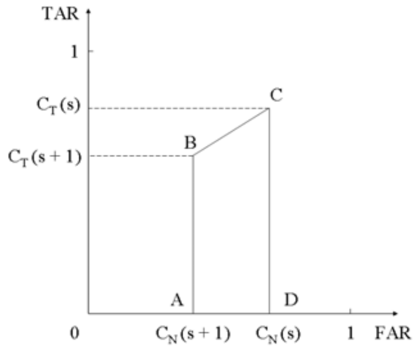

Figure 3 A schematic drawing of four points A, B, C, and D along with their coordinates in the FAR-and-TAR coordinate system. They form a trapezoid at a score s, and BC is a segment of an ROC curve.

After conversion of similarity scores to integers, the distributions of target scores and non-target scores are all discrete. As a result, the ROC curve is no longer a smooth curve. While cumulating probabilities of target scores and non-target scores from the highest similarity score, respectively, an ROC curve can go horizontally, vertically, inclined toward upper right, or stay where it is for each decrement of score, depending on whether $P_N(s)$ and/or $P_T(s)$ are nonzero or not. Thus, the AURC consists of a set of trapezoids, each of which is built by a rectangle and a triangle in general. The trapezoid can be reduced to a rectangle, a vertical line, or a point.

Without loss of generality, a trapezoid is shown in Figure 3. In the FAR (false accept rate)-and-TAR (true accept rate) coordinate system, at a score $s \in \{s\}$, by including zero-frequency scores, a trapezoid is constructed by four points: A (C_N (s + 1), 0), B (C_N (s + 1), C_T (s + 1)), C (C_N (s), C_T (s)), and D (C_N (s), 0), in clockwise direction, assuming $C_N (s_{max} + 1) = C_T (s_{max} + 1) = 0$. This boundary condition corresponds to the origin of the FAR-and-TAR coordinate system, and will be applied throughout the following discussion. The lengths ($C_N (s) - C_N (s + 1)$) (i.e., P_N (s)) and ($C_T (s) - C_T (s + 1)$) (i.e., P_T (s)) form a triangle, and the lengths ($C_N (s) - C_N (s + 1)$) (i.e., P_N (s)) and C_T (s + 1) (i.e., $\sum_{\tau=s+1}^{s\max} P_T(\tau)$) create a rectangle. As a consequence, the estimated AURC can be calculated as,

$$\begin{aligned}\hat{A} &= \sum_{s=s\max}^{s\min} \text{trapezoid}(s) \\ &= \sum_{s=s\max}^{s\min} \text{triangle}(s) + \sum_{s=s\max}^{s\min} \text{rectangle}(s) \\ &= \sum_{s=s\max}^{s\min} P_N(s) \times [\, \frac{1}{2} \times P_T(s) + \sum_{\tau=s+1}^{s\max} P_T(\tau)\,]\end{aligned} \tag{7}$$

Note that the summation runs consecutively in the descending order from s_{max} to s_{min}, including zero-frequency scores, and $\sum_{\tau=s\max+1}^{s\max} = 0$ is assumed according to the above boundary condition. This notation will be applied throughout the following discussion.

A.1.2 Relate AURC to the Mann-Whitney statistic

In order to relate AURC to the Mann-Whitney statistic, the order relations among similarity scores are established as follows. All the M_N scores in the non-target score set **N** in Eq. (2) are compared with all the M_T scores in the target score set **T** in Eq. (1). It counts 1, ½, or zero depending whether a non-target score s_N is less than, equal to, or greater than a target score s_T. This rule can be expressed as

$$\mathbf{R}(s_T, s_N) = \begin{cases} 1 & \text{if } s_N < s_T \\ \frac{1}{2} & \text{if } s_N = s_T \\ 0 & \text{if } s_N > s_T \end{cases} \tag{8}$$

After converting probabilities of target and non-target scores in Eq. (7) back to frequencies and by including zero-frequency scores, the first term in Eq. (7) shows the total number of score pairs in which the non-target score is equal to the target score, weighted by ½ and divided by $M_T M_N$. And the second term in Eq. (7) represents the total number of score pairs in which the non-target score is less than the target score, weighted by 1 and divided by $M_T M_N$. This term is the so called “the number of inversions” in a sequence formed by non-target and target scores [22].

Finally, the estimated AURC can be re-written as

$$\hat{A} = \frac{1}{M_T M_N} \times \sum_{s_T=1}^{M_T} \sum_{s_N=1}^{M_N} \mathbf{R}\,(s_T, s_N) \tag{9}$$

Except for the coefficient, this is exactly the Mann-Whitney statistic formed by the target and non-target scores. As a consequence, the variance of AURC can be obtained by computing the variance of the Mann-Whitney statistic.

A.1.3 Compute SE of AURC

The variance of the Mann-Whitney statistic can be computed analytically and it is utilized as the variance of AURC. To do so, two more cumulative probability distribution functions are required. One is

$$\boldsymbol{Q_T} = \{\, Q_T(s) = \sum_{\tau=s+1}^{s\max} P_T(\tau) \mid \forall\, s \in \{s\} \,\} . \tag{10}$$

The other one is

$$\boldsymbol{Q_N} = \{\, Q_N(s) = \sum_{\tau=s\min}^{s-1} P_N(\tau) \mid \forall\, s \in \{s\} \,\} \tag{11}$$

where another boundary condition $\sum_{\tau=s\min}^{s\min-1} = 0$ is assumed. Note that the cumulation of probabilities is taken place from s_{max} down to s + 1 with respect to target scores in Eq. (10), and from s_{min} up to s – 1 on non-target scores in Eq. (11).

The probability B_{TTN}, that two randomly chosen target matches will obtain higher similarity scores than one randomly chosen non-target match, can be written as

$$B_{TTN} = \sum_{s=s\min}^{s\max} P_N(s) \times [Q_T^2(s) + Q_T(s) \times P_T(s) + \frac{1}{3} \times P_T^2(s)] \tag{12}$$

And the probability B_{NNT}, that one randomly chosen target match will get higher similarity score than two randomly chosen non-target matches, can be expressed as

$$B_{NNT} = \sum_{s=s\min}^{s\max} P_T(s) \times [Q_N^2(s) + Q_N(s) \times P_N(s) + \frac{1}{3} \times P_N^2(s)] \tag{13}$$

Finally, the analytical estimator of SE of AURC can be computed as

$$\hat{SE}_A(A) = \text{square root}\ \{ \frac{1}{M_T M_N} \times [\hat{A}(1-\hat{A}) + (M_T - 1)(B_{TTN} - \hat{A}^2) + (M_N - 1)(B_{NNT} - \hat{A}^2)]\} \tag{14}$$

A.2 Bootstrap computation of SE of AURC

The estimated $S\hat{E}$ of AURC can also be calculated using the nonparametric two-sample bootstrap method. When the data set is assumed to be i.i.d., the bootstrap units are scores in the data set rather than subsets of the sample into which the sample data are regrouped according to data dependencies.

With such an assumption, the algorithm of the nonparametric two-sample bootstrap shown in Section 2.4 can be employed. In Step 4 of the algorithm, after randomly resampling WR the two original score sets **T** and **N**, the i-th bootstrap replication of the estimated $A\hat{U}RC$, i.e., $\hat{T}_i$ = $A\hat{U}RC_i$, can be calculated from the two new sets of target scores and non-target scores using Eq. (7).

After B iterations, these B bootstrap replications of the estimated $A\hat{U}RC$ constitute a bootstrap distribution. Finally, the bootstrap estimator of SE of AURC denoted by $S\hat{E}_B(A)$ is obtained from such a bootstrap distribution, as indicated in Step 6 of the algorithm.

A.3 Comparisons between analytical results and bootstrap results

While comparing the two estimators of SE of AURC, a relative error η is employed and defined as

$$\eta = |\ S\hat{E}_B(A) - S\hat{E}_A(A)\ | / S\hat{E}_A(A) \times 100\ \% \qquad (15)$$

where $S\hat{E}_A(A)$ is the analytical estimator of SE of AURC computed using Eq. (14), and $S\hat{E}_B(A)$ is the bootstrap estimator calculated in Section A.2.

Systems	$A\hat{U}RC$	$S\hat{E}_A(A)$	$S\hat{E}_B(A)$	Relative Errors (%)
UJ	0.986781	0.000376	0.000367	2.64
DL	0.979069	0.000491	0.000489	0.37
BK	0.979061	0.000542	0.000545	0.62
EL	0.978651	0.000635	0.000642	1.12
LZ	0.965603	0.000734	0.000747	1.83
AF	0.904570	0.001284	0.001304	1.50
PM	0.904184	0.001162	0.001139	1.96
CH	0.901069	0.001133	0.001141	0.69
DG	0.892396	0.001245	0.001212	2.66
CO	0.857253	0.001532	0.001495	2.40
FI	0.856774	0.001445	0.001538	6.41
PB	0.800933	0.001860	0.001888	1.51

Table 2 The estimated $A\hat{U}RC$s, analytical $S\hat{E}_A(A)$s, bootstrap $S\hat{E}_B(A)$s with the i.i.d. assumption, and the relative errors of 12 speaker recognition systems.

In Table 2 are listed the estimated $A\hat{U}RC$s, the analytical estimators $S\hat{E}_A(A)$, the bootstrap estimators $S\hat{E}_B(A)$ with the i.i.d. assumption, and the relative errors η of 12 speaker recognition systems. These 12 systems were randomly selected from those who participated in the core test short2-short3.

The analytical result of SE of AURC derived from the target and non-target scores of any speaker recognition system is unique. Thus, it could be treated as a reference. However, the bootstrap result of SE of AURC for a system is stochastic. In other words, the result fluctuates for different runs. Nonetheless, the bootstrap results with the i.i.d. assumption shown in Table 2 for different systems were obtained by a random run, respectively. If these results were very close to the reference results, then it supported that the i.i.d. assumption were quite reasonable for the speaker recognition data.

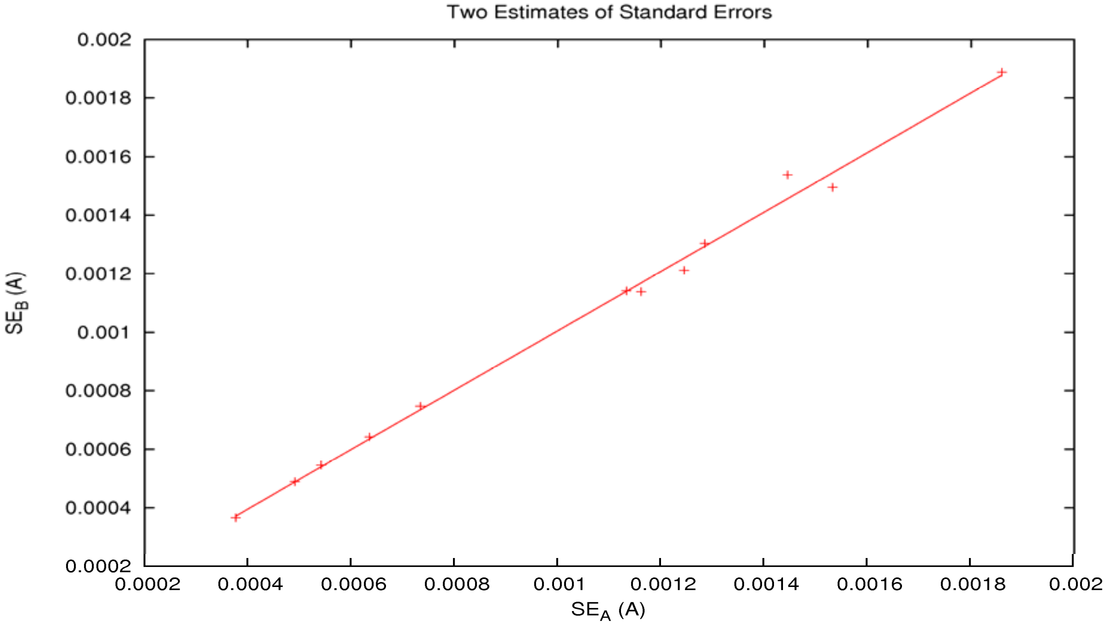

Figure 4 The scatter plot of the estimated bootstrap $\hat{SE}_B$ (A)s with the i.i.d. assumption versus the estimated analytical $\hat{SE}_A$ (A)s along with the best-fit straight line, the slope of which is close to 1 and the intercept of which is close to zero.

Indeed, all relative errors that quantify the difference, except for one that is 6.41 % for system FI, are not larger than 2.66 %. Including this outlier, the median of the relative errors is 1.67 % and the mean is 1.98 %. Excluding this outlier, the median is 1.51 % and the mean is 1.57 %. All these relative errors are quite small. In other words, the two results matched very well. This is also evidenced by the scatter plot of the estimated bootstrap $\hat{SE}_B$ (A)s with the i.i.d. assumption versus the estimated analytical $\hat{SE}_A$ (A)s along with the best-fit straight line, the slope of which is close to 1 and the intercept of which is close to zero, as shown in Figure 4.

As a consequence, the speaker recognition system data can be assumed to be i.i.d.. As mentioned in Section 1, in the 2008 NIST SRE, generally speaking, the sizes of target scores and non-target scores were about 20 000 and 80 000, respectively. The large size of speaker datasets might have reduced the effect of the dependencies among the data.

References

1. "The NIST Year 2008 Speaker Recognition Evaluation Plan", the URL of the website is at http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf
2. J.C. Wu, A.F. Martin, R.N. Kacker, C.R. Hagwood, Significance test in operational ROC analysis, in Biometric Technology for Human Identification VII, Proceedings of SPIE Vol. 7667, 76670I (2010).
3. B. Efron, Bootstrap methods: Another look at the Jackknife. Ann. Statistics, 7:1-26, 1979.
4. P. Hall, On the number of bootstrap simulations required to construct a confidence interval, Ann. Statist. 14 (4) (1986) 1453-1462.
5. B. Efron, Better bootstrap confidence intervals, J. Amer. Statist. Assoc. 82 (397) (1987) 171-185.
6. B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.
7. J.C. Wu, Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap, NISTIR 7449, National Institute of Standards and Technology, September, 2007.
8. J.C. Wu, Operational measures and accuracies of ROC Curve on large fingerprint data Sets, NISTIR 7495, National Institute of Standards and Technology, May, 2008.
9. J.C. Wu, M.D. Garris, Nonparametric statistical data analysis of fingerprint minutiae exchange with two-finger fusion, in Biometric Technology for Human Identification IV, Proceedings of SPIE Vol. 6539, 65390N (2007).
10. J.C. Wu, C.L. Wilson, Nonparametric analysis of fingerprint data on large data sets, Pattern Recognition 40 (9) (2007) 2574-2584.
11. R.Y. Liu, K. Singh, Moving blocks jackknife and bootstrap capture weak dependence. Exploring the limits of bootstrap, ed. by LePage and Billard. John Wiley, New York, 1992.
12. R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, A.W. Senior, Guide to Biometrics, Springer, New York, 2003 pp. 269-292.
13. J.A. Hanley, B.J. McNeil, A method of comparing the area under two ROC curves derived from the same cases, Radiology 148 (1983) 839-843.
14. J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29-36.
15. D. Bamber, The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, J. Math Psych 12 (1975) 387-415.
16. G.E. Noether, Elements of nonparametric statistics, John Wiley and Sons, New York, 1967, pp. 31-32.
17. J.C. Wu, C.L. Wilson, An empirical study of sample size in ROC-curve analysis of fingerprint data, in Biometric Technology for Human Identification III, Proceedings of SPIE Vol. 6202, 620207 (2006).
18. B. Ostle, L.C. Malone, Statistics in Research: Basic Concepts and Techniques for Research Workers, fourth ed., Iowa State University Press, Ames, 1988.
19. R.J. Hyndman, Y. Fan, Sample quantiles in statistical packages, American Statistician 50 (1996) 361-365.

20. R: A Language and Environment for Statistical Computing, The R Development Core Team, Version 2.8.0, 2008, at http://www.r-project.org/.
21. T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, The ROCR Package: Visualizing the performance of scoring classifiers, Version 1.0-2, 2007, at http://rocr.bioinf.mpi-sb.mpg.de/ROCR.pdf.
22. B.L. van der Waerden, Mathematical Statistics, Springer, Berlin, 1969 p. 274 and pp. 333–335.