Significance Test with Data Dependency in Speaker Recognition Evaluation

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

Abstract

To evaluate the performance of speaker recognition systems, a detection cost function defined as a weighted sum of the probabilities of type I and type II errors is employed. The speaker datasets may have data dependency due to multiple uses of the same subjects. Using the standard errors of the detection cost function computed by means of the two-layer nonparametric two-sample bootstrap method, a significance test is performed to determine whether the difference between the measured performance levels of two speaker recognition algorithms is statistically significant. While conducting the significance test, the correlation coefficient between two systems' detection cost functions is taken into account. Examples are provided.

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

1 Introduction

The Speaker Recognition Evaluation (SRE) is an ongoing project conducted by the National Institute of Standards and Technology (NIST) [1]. It has had a great impact on the research and development of technology in the community of the audio, speech, and language processing. Each trial in an SRE test consists of a training model speaker and a test speech segment. The speaker recognition system must decide whether speech of the model speaker occurs in the test speech segment and generate a score. A higher score indicates greater confidence that the test speech is spoken by the model speaker. Target (non-target) scores are generated by trials in which the test speech segment contains (does not contain) speech of the model speaker defined in the training data.

To evaluate the performance of speaker recognition systems, a detection cost function defined as a weighted sum of the probabilities of type I error (miss) and type II error (false alarm) is employed as a metric [1]. These two error rates represent a tradeoff and are negatively correlated [2]. Further, the NIST speaker recognition data contain dependencies [3], which arise largely from multiple uses of the same subjects in order to generate more target and non-target scores due to limited resources. This data dependency is complicated, due in part to the way the data are collected.

In our test, the data dependency is determined based solely upon the multiple use of the training model speaker identification (ID) number. Target scores and non-target scores generated using the

Active and Passive Signatures IV, edited by G. Charmaine Gilbreath, Chadwick Todd Hawley,
Proc. of SPIE Vol. 8734, 87340I · © 2013 SPIE · CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2014536

same training model speaker ID number are grouped into a target set and a non-target set, respectively, so as to preserve the data dependency. Then the speaker datasets are refined to a two-layer data structure: the first layer consists of target sets and non-target sets, and the second layer consists of target scores and non-target scores within sets.

The sampling variability, including the data dependency, results in uncertainties of the detection cost function in the SRE. The covariance between the type I and II errors and the data dependency make the analytical computation of such uncertainties difficult. Hence, in our prior studies, the standard error (SE) of the detection cost function was estimated using the two-layer nonparametric two-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 [2-9].

In order to make the probabilities for scores being selected equal with respect to different target sets and non-target sets, respectively, while resampling in the two-layer bootstrap, the datasets are adjusted in such a way that all target sets contain the same number of scores and likewise for the non-target sets [3]. In the meantime, the total numbers of selected target and non-target scores are kept as large as possible. The new datasets had 132 target sets (130 non-target sets), each of which contained 96 target scores (244 non-target scores) that were randomly selected without replacement from the raw target (non-target) set if the number of scores in the set was greater than or equal to what was required. The total numbers of target and non-target scores were 12,672 and 31,720.

The two samples involved are referred to as a sample of target scores and a sample of non-target scores, which characterize the speaker recognition system that generates them and usually do not have well defined parametric forms [10, 11]. In the two-layer bootstrap, the nonparametric two-sample resampling takes place randomly with replacement (WR), not only on the first layer of the data while the bootstrap units are sets, but also subsequently on the second layer while the bootstrap units are scores within a set, where the scores are conditionally independent.

In evaluating and comparing the performances of speaker recognition systems, it may be of interest to determine whether the difference between the measured performance level of a specific speaker recognition algorithm and a hypothesized criterion value is real or by chance, or to determine whether the difference between the measured performance levels of two algorithms is statistically significant [2, 12]. The principles stay the same for these two scenarios. In SRE the latter is often of more interest, and thus is explored in this article.

To do so, it is insufficient to only compute the uncertainty of the detection cost function. For instance, comparison issues may be examined intuitively to some extent using the 95 % confidence intervals (CI) derived from the uncertainty. But it is difficult to reach any conclusion when the two 95 % CIs overlap. Moreover, CIs alone cannot provide quantitative information (such as *p*-values) on the statistical significance of the difference. Thus, statistical hypothesis testing is employed.

By examining the relationship between the two types of 95 % CIs, it was found that the one computed using the quantile method matched very well with the one derived using the normality assumption for the distribution of the detection cost function. This suggests that the detection cost

function be regarded as approximately normally distributed. Thereafter, the Z-test may be used to perform significance testing.

The detection cost functions of the two speaker recognition systems may or may not be correlated, depending on how the test is designed and how the sets of scores are generated. In our SRE tests, all the scores of the different systems were generated on a common set of speakers and speech segments and, therefore, are highly correlated. Thus, the resulting detection cost functions are also correlated. In this article, an algorithm is provided to find the correlated pairs of metrics from the correlated scores, and then the correlation coefficient of two detection cost functions is computed [2, 12].

The notations of sets and scores are provided in Section 2. The formulas for computing the detection cost function are presented in Section 3. The general formulas of hypothesis testing for comparing performance levels of two systems are shown in Section 4. The two-layer nonparametric two-sample bootstrap algorithm is provided in Section 5. An algorithm for computing the correlation coefficient of two cost functions is described in Section 6. The results involving five speaker recognition systems[1] are presented in Section 7. The conclusions and discussion can be found in Section 8.

2 The notations of sets and scores

target $\boldsymbol{S}_T$	sets	$\boldsymbol{S}_{T1}$	$\boldsymbol{S}_{T2}$	……	$\boldsymbol{S}_{T\,m_T}$
	scores	$\alpha_{T11}, \alpha_{T12}, \ldots, \alpha_{T1\,\mu_{T1}}$	$\alpha_{T21}, \alpha_{T22}, \ldots, \alpha_{T2\,\mu_{T2}}$	……	$\alpha_{T\,m_T1}, \alpha_{T\,m_T2}, \ldots, \alpha_{T\,m_T\,\mu_{T\,m_T}}$

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

non-target $\boldsymbol{S}_N$	sets	$\boldsymbol{S}_{N1}$	$\boldsymbol{S}_{N2}$	……	$\boldsymbol{S}_{N\,m_N}$
	scores	$\alpha_{N11}, \alpha_{N12}, \ldots, \alpha_{N1\,\mu_{N1}}$	$\alpha_{N21}, \alpha_{N22}, \ldots, \alpha_{N2\,\mu_{N2}}$	……	$\alpha_{N\,m_N1}, \alpha_{N\,m_N2}, \ldots, \alpha_{N\,m_N\,\mu_{N\,m_N}}$

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

As discussed in Section 1, the speaker datasets are refined to a two-layer data structure: sets and scores in sets. Suppose that the numbers of target and non-target sets are m_T and m_N. Thus, the set $\boldsymbol{S}_T$ of all target sets and the set $\boldsymbol{S}_N$ of all non-target sets are expressed by

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

where $\boldsymbol{S}_{Tj}$ are target sets and $\boldsymbol{S}_{Nj}$ are non-target sets. In terms of scores, each set can be denoted as

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

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

where $\alpha_{T\,j\,k}$ and $\alpha_{N\,j\,k}$ are target and non-target scores, and $\mu_{i\,j}$ stands for the number of scores in sets. They are explicitly listed in Table 1 and Table 2.

As stated in Section 1, the datasets are adjusted in such a way that all target sets contain the same number of scores and likewise for the non-target sets, that is, $\mu_{i\,1} = \mu_{i\,2} = \ldots = \mu_{i\,m_i}$, where $i \in \{T, N\}$. The total number of target scores N_T and the total number of non-target scores N_N are,

$$N_i = \sum_{j=1}^{m_i} \mu_{i\,j}, \quad \text{where } i \in \{T, N\}. \tag{3}$$

Hence, the probabilities for each target score and each non-target score being selected are $1 / N_T$ and $1 / N_N$, respectively [3]. In addition, this can reduce the variance of the computation.

Moreover, since the scores are grouped into sets based on the training model speaker ID numbers, the two scores of any two speaker recognition systems with the same ordinal number of sets and the same ordinal number of scores in sets are generated by the speakers and speech segments with the same ID numbers, for both target and non-target scores. Therefore, these two scores of two systems are correlated.

3 The detection cost function in speaker recognition evaluation

After converting scores to integer, without loss of generality, for a speaker recognition system, the scores are expressed inclusively using the integer score set $\{s\} = \{s_{min}, s_{min}+1, \ldots, s_{max}\}$. Let C_i (s), $i \in \{T, N\}$ denote the cumulative probabilities of target scores and non-target scores from the highest score s_{max} down to an integer score s, respectively. The probability of type I error at a threshold $t \in \{s\}$ for target scores, denoted by $P_I\,(t)$, is cumulated from the lowest score s_{min}. The probability of type II error at a threshold t for non-target scores, denoted by $P_{II}\,(t)$, is cumulated from the highest score s_{max}. For discrete probability distribution, while computing $P_I\,(t)$ and $P_{II}\,(t)$ at a threshold t, the probabilities of target scores and non-target scores at the threshold t must be taken into account [13].

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

$$\begin{aligned} \hat{P}_I\,(t) &= 1 - C_T\,(t+1) \\ \hat{P}_{II}\,(t) &= C_N\,(t) \end{aligned} \qquad \text{for } t \in \{s\}, \tag{4}$$

where $C_T\,(s_{max} + 1) = 0$ is assumed [2, 10]. Based on Eq. (4), in practice, the estimators $\hat{P}_I\,(t)$ and $\hat{P}_{II}\,(t)$ can be obtained by moving the score from the highest score s_{max} down to the threshold t one score at a time to cumulate the probabilities of target scores and non-target scores, respectively. The detection cost function at a threshold t is defined as a weighted sum of the probabilities of type I error and type II error at the threshold t [1]

$$C_{Det}\,(t) = C_{Miss} \times P_I\,(t) \times P_{Target} + C_{FalseAlarm} \times P_{II}\,(t) \times (1 - P_{Target}). \tag{5}$$

Determining an appropriate decision threshold is a challenging research problem, which is outside the scope of this article. Therefore, the thresholds used in this article are those provided by tested speaker recognition systems in order to make an explicit speaker detection decision for each trial.

The parameters C_{Miss} and $C_{FalseAlarm}$ are the relative costs of detection errors, and P_{Target} is the *a priori* probability of the specified model speaker. In our SREs for all speaker detection tests, the parameters C_{Miss}, $C_{FalseAlarm}$, and P_{Target} were set to be 10, 1, and 0.01, respectively [1].

4 Two-algorithm hypothesis testing for comparisons

Let C_1 and C_2 denote the two detection cost functions for any two speaker recognition systems at respective thresholds. Then, the null and alternative hypotheses are

$$\begin{aligned} H_o &: C_1 = C_2 \\ H_a &: C_1 \neq C_2 \,. \end{aligned} \tag{6}$$

If the statistic of interest is normally distributed, the general *Z* statistic for two-algorithm hypothesis testing is expressed as

$$Z = \frac{\hat{C}_1 - \hat{C}_2}{\sqrt{SE^2(\hat{C}_1) + SE^2(\hat{C}_2) - 2r\,SE(\hat{C}_1)\,SE(\hat{C}_2)}} \tag{7}$$

where $\hat{C}_1$ and $\hat{C}_2$ are estimators of two detection cost functions, SE ($\hat{C}_1$) and SE ($\hat{C}_2$) stand for their SEs, and *r* is the correlation coefficient between $\hat{C}_1$ and $\hat{C}_2$.

The Z statistic is distributed as the standard normal distribution with zero expectation and unit variance. The standard error of the detection cost function with data dependency can be computed using the two-layer nonparametric two-sample bootstrap (see Section 5). If the two statistics of interest are positively correlated and the correlation coefficient *r* is not taken into account, it can leave the denominator of Eq. (7) larger and the Z score smaller; thereby reduce the chance of detecting a performance difference between two algorithms.

There is no reason to believe *a priori* that the performance of one algorithm is likely to be better than the performance of the other algorithm. Further, the two-tailed test is generally more conservative than the one-tailed test in the sense that the former is more difficult to reject the null hypothesis for a given significance level [14]. Thus, the two-tailed Z-test is used in this article.

5 An algorithm for the two-layer nonparametric two-sample bootstrap

As stated in Section 1, the two-layer nonparametric two-sample bootstrap method is proposed to compute the estimate of the uncertainty of the detection cost function at a threshold *t*, based on our prior studies of bootstrap variability in ROC analysis on large datasets [2-9]. The two-layer bootstrap is carried out not only on the first layer of the new data structure where the resampling units are target sets and non-target sets, but also on the second layer of the data in which the resampling units are target scores and non-target scores in sets. From here on, the superscript indices are used for the numeration of the resampling iterations. The algorithm is shown as follows.

Algorithm I **(two-layer nonparametric two-sample bootstrap)**

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

3: **for** k = 1 **to** m_T **do**
4: WR_Random_Sampling_Set ($\mu'_{T\,k}{}^{i}$, $\boldsymbol{S}'_{T\,k}{}^{i}$, $\boldsymbol{S}''_{T\,k}{}^{i}$)
5: **end for**

6: WR_Random_Sampling_Set (m_N, $\boldsymbol{S}_N$, $\boldsymbol{S}'_N{}^{i} = \{ \boldsymbol{S}'_{N\,j}{}^{i} \mid j = 1, \ldots, m_N \}$)
7: **for** k = 1 **to** m_N **do**
8: WR_Random_Sampling_Set ($\mu'_{N\,k}{}^{i}$, $\boldsymbol{S}'_{N\,k}{}^{i}$, $\boldsymbol{S}''_{N\,k}{}^{i}$)
9: **end for**

10: $\boldsymbol{S}''_T{}^{i} = \{ \boldsymbol{S}''_{T\,j}{}^{i} \mid j = 1, \ldots, m_T \}$ and $\boldsymbol{S}''_N{}^{i} = \{ \boldsymbol{S}''_{N\,j}{}^{i} \mid j = 1, \ldots, m_N \}$ => statistic $\hat{C}^i$
11: **end for**

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

1.1: **function** WR_Random_Sampling_Set (L, **Γ**, **Θ**)
1.2: **for** i = 1 **to** L **do**
1.3: select randomly WR an index $j \in \{ 1, \ldots, L \}$
1.4: $\theta_i = \gamma_j$
1.5: **end for**
1.6: **end function**

The bootstrap calls the function WR_Random_Sampling_Set. In this function, **Γ** stands for a set of sets or a set of scores, L 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. This function can be applied to either a set of sets or a set of scores. It runs L iterations as shown from Step 1.2 to 1.5. 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 1.3 and 1.4. As a result, L members (sets or scores) are randomly selected WR from the set **Γ** to form a new set **Θ**.

In Algorithm I, B is the number of the bootstrap replications, i.e., the number of iterations as shown from Step 1 to 11, $\boldsymbol{S}_T$ is the set of all target sets and $\boldsymbol{S}_N$ is the set of all non-target sets as expressed in Eq. (1), and m_T and m_N are the cardinalities of the set $\boldsymbol{S}_T$ and the set $\boldsymbol{S}_N$, respectively.

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

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

As indicated in Step 10, all target scores in the new set $\boldsymbol{S}''_{T}{}^{i} = \{ \boldsymbol{S}''_{T\,j}{}^{i} \mid j = 1, \ldots, m_T \}$ and all non-target scores in the new set $\boldsymbol{S}''_{N}{}^{i} = \{ \boldsymbol{S}''_{N\,j}{}^{i} \mid j = 1, \ldots, m_N \}$ are employed to calculate the estimators of the probabilities of type I and type II errors, i.e., $\hat{P}_I(t)$ and $\hat{P}_{II}(t)$ using Eq. (4) and then the i-th bootstrap replication of the estimated cost function at a given threshold, i.e., $\hat{C}^i$ using Eq. (5).

With the new data structure described in Section 1, not only does each target (non-target) score have the same probability to be selected, but also the same numbers of target scores and the same numbers of non-target scores, respectively, are resampled in Step 10 at different iterations of the two-layer nonparametric two-sample bootstrap. All these can reduce the computation variance.

Finally, as shown in Step 12, 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 bootstrap replications, and the estimators of the $\alpha/2$ 100 % and (1 - $\alpha/2$) 100 % quantiles of the bootstrap distribution, denoted by $\hat{Q}(\alpha/2)$ and $\hat{Q}(1-\alpha/2)$, at the significance level α can be calculated [5]. Definition 2 of quantile in Ref. [15] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities. Thus, $(\hat{Q}(\alpha/2), \hat{Q}(1-\alpha/2))$ stands for the estimated bootstrap (1 - α) 100 % $C\hat{I}$. If 95 % $C\hat{I}$ is of interest, then α is set to be 0.05.

The remaining issue is to determine how many iterations the bootstrap algorithms need to run in order to reduce the bootstrap variance and ensure the accuracy of the computation. In our applications, such as biometrics and the evaluation of speaker recognition, etc., the sizes of datasets are tens to hundreds of thousands of scores, which are much larger than those in some other applications of bootstrap methods like medical decision making, etc. [5]. Moreover, in ROC analysis our statistics of interest are mostly probabilities or a weighted sum of probabilities, etc. rather than a simple sample mean. And most importantly our data samples of 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 2000 [2, 8, 9].

6 An algorithm for computing the correlation coefficient

As discussed in Sections 1 and 2, the two detection cost functions for any two speaker recognition systems are correlated. For example, consider two speaker recognition systems denoted by A and B. They have the same two-layer data structures, and generate two scores with the same ordinal number of sets and the same ordinal number of scores in sets by matching the speakers and speech segments with the same ID numbers, for both target scores and non-target scores. Therefore, these two scores corresponding to the two systems co-vary. Consequently the detection cost functions of any two systems computed using Eqs. (4) and (5) are also correlated [2, 12]. An algorithm that picks up the correlated scores and then computes the correlation coefficient of two cost functions is as follows.

Algorithm II **(Correlation coefficient)**

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

6: Synchronized_WR_Random_Sampling_Set
(m_N, $\mathbf{S}^A{}_N$, $\mathbf{S}^A{}'_N{}^i = \{ \mathbf{S}^A{}'_{Nj}{}^i \mid j = 1, \ldots, m_N \}$, $\mathbf{S}^B{}_N$, $\mathbf{S}^B{}'_N{}^i = \{ \mathbf{S}^B{}'_{Nj}{}^i \mid j = 1, \ldots, m_N \}$)
7: **for** k = 1 **to** m_N **do**
8: Synchronized_WR_Random_Sampling_Set ($\mu'_{Nk}{}^i$, $\mathbf{S}^A{}'_{Nk}{}^i$, $\mathbf{S}^A{}''_{Nk}{}^i$, $\mathbf{S}^B{}'_{Nk}{}^i$, $\mathbf{S}^B{}''_{Nk}{}^i$)
9: **end for**

10: $\mathbf{S}^A{}''_T{}^i = \{ \mathbf{S}^A{}''_{Tj}{}^i \mid j = 1, \ldots, m_T \}$ and $\mathbf{S}^A{}''_N{}^i = \{ \mathbf{S}^A{}''_{Nj}{}^i \mid j = 1, \ldots, m_N \}$ => statistic $\hat{C}^{Ai}$
11: $\mathbf{S}^B{}''_T{}^i = \{ \mathbf{S}^B{}''_{Tj}{}^i \mid j = 1, \ldots, m_T \}$ and $\mathbf{S}^B{}''_N{}^i = \{ \mathbf{S}^B{}''_{Nj}{}^i \mid j = 1, \ldots, m_N \}$ => statistic $\hat{C}^{Bi}$
12: **end for**

13: $\{\hat{C}^{Ai} \mid i = 1, ..., M\}$ and $\{\hat{C}^{Bi} \mid i = 1, ..., M\}$ => the correlation coefficient $\hat{r}^{AB}{}_C$
14: **end**

2.1: **function** Synchronized_WR_Random_Sampling_Set (L, $\mathbf{\Gamma}^A$, $\mathbf{\Theta}^A$, $\mathbf{\Gamma}^B$, $\mathbf{\Theta}^B$)
2.2: **for** j = 1 **to** L **do**
2.3: select randomly WR an index $k \in \{ 1, \ldots, L \}$
2.4: $\theta^A{}_j = \gamma^A{}_k$
2.5: $\theta^B{}_j = \gamma^B{}_k$
2.6: **end for**
2.7: **end function**

Algorithm II is similar to Algorithm I, except that the Algorithm II is applied to two speaker recognition systems simultaneously. Based on our bootstrap variability studies, the number of iterations M is set to be 2000 [2, 8, 9].

The function Synchronized_WR_Random_Sampling_Set can be applied to either sets or scores. $\mathbf{\Gamma}^A$, $\mathbf{\Theta}^A$, $\mathbf{\Gamma}^B$, and $\mathbf{\Theta}^B$ stand for a set of sets or a set of scores generated by Systems A and B, respectively. L is their cardinalities. $\gamma^A{}_k$, $\theta^A{}_j$, $\gamma^B{}_k$, and $\theta^B{}_j$ are their members. The two sets or scores $\gamma^A{}_k$ and $\gamma^B{}_k$ of two Systems A and B with the same ordinal number k co-vary. Therefore, this function synchronizes the sampling in a set $\mathbf{\Gamma}^A$ created by System A and the selection in a set $\mathbf{\Gamma}^B$ generated by System B so that the sets or the scores with the same ordinal number k are chosen to form two new sets $\mathbf{\Theta}^A$ and $\mathbf{\Theta}^B$, respectively. In other words, the correlated scores generated by these two systems are selected.

From Step 1 to 12, Algorithm II runs M iterations. In the i-th iteration, in Step 2, the function is applied simultaneously to the first layer of the datasets, i.e., the set $\mathbf{S}^A{}_T$ of target sets generated by System A and the set $\mathbf{S}^B{}_T$ of target sets created by System B so that the two target sets with the same

ordinal number in the two systems' datasets are randomly selected WR and form two new sets $\mathbf{S}^{A}{}'_{T}{}^{i}$ and $\mathbf{S}^{B}{}'_{T}{}^{i}$ of target sets.

Then, from Step 3 to 5, this function is applied simultaneously to the second layer of the datasets m_T times created by the two systems. Hence, the target scores in set $\mathbf{S}^{A}{}'_{Tk}{}^{i}$ generated by System A and the target scores in set $\mathbf{S}^{B}{}'_{Tk}{}^{i}$ created by System B with the same ordinal number in the two systems' datasets are randomly chosen WR. These correlated scores constitute two new sets of target scores $\mathbf{S}^{A}{}''_{Tk}{}^{i}$ and $\mathbf{S}^{B}{}''_{Tk}{}^{i}$, respectively.

The analogous interpretation can be applied to non-target sets and non-target scores in sets from Step 6 to 9. In Step 10, the target scores in set $\mathbf{S}^{A}{}''_{T}{}^{i}$ and the non-target scores in set $\mathbf{S}^{A}{}''_{N}{}^{i}$ for System A produce the i-th bootstrap replication of the estimated detection cost function $\hat{C}^{Ai}$ for System A. In Step 11, the correlated target scores in set $\mathbf{S}^{B}{}''_{T}{}^{i}$ and the correlated non-target scores in set $\mathbf{S}^{B}{}''_{N}{}^{i}$ for System B produce the i-th bootstrap replication $\hat{C}^{Bi}$ for System B. Thus, the correlated pairs of bootstrap replications of estimated cost functions are calculated from the correlated scores. After M iterations, finally in Step 13, the estimated correlation coefficient of the detection cost functions, $\hat{r}^{AB}{}_{C}$, is computed from these two sets of correlated bootstrap replications of estimated cost functions [13].

7 Results

Five speaker recognition systems, labeled as EL, UJ, BK, LZ and DL[2], are used for illustration. In Table 3 are shown their estimated detection cost functions, and the estimated SÊs and 95 % CÎs computed using the two-layer nonparametric two-sample bootstrap method by taking account of the data dependency.

The estimated 95 % CÎs shown in Table 3 were all calculated using the quantile method as described in Section 5. They can also be computed by multiplying 1.96 by the estimated SÊ, assuming that the distribution of 2000 bootstrap replications of the detection cost function is normal. These two types of 95 % CÎs are matched up to the third or fourth decimal place for all five systems. For instance, for System EL, the 95 % CÎ derived from the quantile method is (0.018384, 0.026084) as shown in Table 3, while it is (0.018374, 0.026024) based on the normality assumption. This suggests that the detection cost function may be regarded as normally distributed.

Systems	Cost functions	SÊs	95% CÎs
EL	0.022199	0.001952	(0.018384, 0.026084)
UJ	0.028996	0.002026	(0.025082, 0.033150)
BK	0.031588	0.001883	(0.028046, 0.035311)
LZ	0.040098	0.002897	(0.034641, 0.045880)
DL	0.040880	0.001841	(0.037185, 0.044511)

Table 3 The estimated detection cost functions, SÊs, and 95 % CÎs of five speaker recognition systems.

[2] It is the policy of NIST and the evaluation sponsors not to publicly associate specific SRE participants with their evaluation performance results, and therefore system names are encoded in this article.

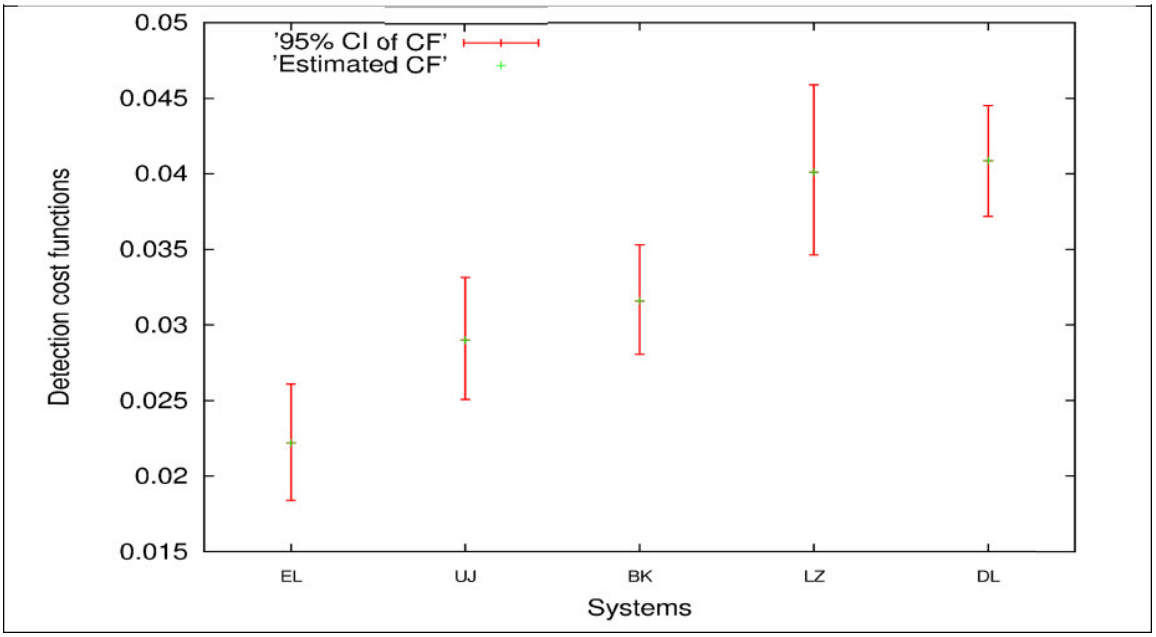

Figure 1 The estimated detection cost functions, and 95 % CÎs of five speaker recognition systems.

Figure 1 depicts the estimated detection cost functions, and their estimated 95 % CÎs, for the five speaker recognition systems. The estimated 95 % CÎs overlap in some cases. For instance, the 95 % CÎ of System EL somewhat overlaps the one of System UJ; and the latter considerably overlaps that of System BK. If two speaker recognition systems need to be compared to determine whether the difference between their performances is statistically significant, no conclusion can be reached merely based on the relationship of these confidence intervals if they overlap, unless the hypothesis testing is carried out. With the above normality assumption for the distribution of the cost function, the two-algorithm hypothesis testing provided in Section 4 can be employed.

The correlation coefficient of the detection cost functions of two systems appearing in Eq. (7) is estimated using Algorithm II presented in Section 6. This algorithm involves a synchronized random resampling. Due to the stochastic nature of such resampling, in this article, the algorithm was run 20 times, and the average out of these runs was taken to be the resultant correlation coefficient, in order to reduce the computational fluctuation. In practice, if the *p*-value is considerably different from the critical value of interest such as 5 %, 1 %, etc., then Algorithm II only needs to run once.

Systems	EL	UJ	BK	LZ	DL
EL	1.000000	0.233958	0.433872	0.620300	0.388808
UJ		1.000000	0.347396	0.196418	0.425286
BK			1.000000	0.437193	0.640776
LZ				1.000000	0.426599
DL					1.000000

Table 4 The average correlation coefficients of two detection cost functions out of 20 runs of five speaker recognition systems.

Systems	EL	UJ	BK	LZ	DL
EL	1.0000	0.0058	0.0000	0.0000	0.0000
UJ		1.0000	0.2463	0.0005	0.0000
BK			1.0000	0.0015	0.0000
LZ				1.0000	0.7713
DL					1.0000

Table 5 The two-tailed *p*-values of two speaker recognition systems, where the correlation coefficients were taken into account.

All correlation coefficients are shown in Table 4. They are all positive. It indicates as expected that all systems tend to assign higher or lower scores to particular trials. These results also provide evidence that the synchronized Algorithm II is quite reasonable for computing the correlation coefficient.

All two-tailed *p*-values of system pair among the five speaker recognition systems are presented in Table 5. In this table, only two *p*-values are greater than 5 %. They are 24.63 % for Systems UJ and BK and 77.13 % for Systems LZ and DL. This indicates that the null hypothesis cannot be rejected, i.e., the performance differences between UJ and BK and between LZ and DL are not significant, even though the estimated detection cost functions of UJ and LZ are smaller than those of BK and DL, respectively. This conclusion is consistent with the observation in Figure 1, where the estimated 95 % CÎs of the cost functions for UJ and LZ overlap those for BK and DL considerably.

All other off-diagonal *p*-values in Table 5 are much less than 5 %. This suggests that the null hypothesis of no difference be strongly rejected. That is, the performance difference between the corresponding two systems is real. For instance, comparing Systems EL and UJ, the two-tailed *p*-value is 0.58 %. Thus, the performance of System EL is significantly better than the performance of System UJ, although their estimated 95 % CÎs slightly overlap as shown in Table 3 and Figure 1.

In addition, the magnitudes of the *p*-values in Table 5 suggest, to some extent, how much the corresponding 95 % CÎs overlap. Thus, they describe quantitatively how significant the differences are between the performances of the two systems. The statistical hypothesis testing provides quantitative information (such as *p*-values) regarding the statistical significance of differences.

8 Conclusions and discussion

The SRE involves evaluation and comparison of speaker recognition systems. It can be important to determine whether the difference between the performance level of one speaker recognition system and a performance criterion value, or the difference between the performance levels of two systems is statistically significant. In this article, the latter case was investigated, but the principle involved in the former case is similar.

To evaluate the performance of speaker recognition systems, a detection cost function defined as a weighted sum of the probabilities of type I error (miss) and type II error (false alarm) is employed as

a metric. The NIST speaker recognition data contain dependencies due to multiple uses of the same subjects. Thus, the scores are grouped into sets to preserve the data dependency, and the speaker datasets are refined into a two-layer data structure.

The sampling variability, including this data dependency, results in uncertainty for the value of the detection cost function. The uncertainties of the detection cost function in terms of SE and 95 % CI were computed using the two-layer nonparametric two-sample bootstrap method with 2000 bootstrap replications based on our prior variability studies of bootstraps in ROC analysis on large datasets.

The detection cost function may be regarded as approximately normally distributed regardless of the distributions of target scores and non-target scores. This assumption is supported by the matches between two types of 95 % CIs. One is computed using the definition of quantile, while the other is calculated based on the assumption that the distribution of 2000 bootstrap replications of the statistic of interest is normal. As a consequence, it seems reasonable to apply the Z-test.

As shown in Sections 1 and 2, the scores of any two speaker recognition systems in SRE are correlated. Hence, the detection cost functions of two systems are also correlated. If the two statistics of interest are indeed positively correlated and the correlation coefficient is not taken into account, the likelihood of detecting a difference between the performance levels of two systems is reduced. In this article, a synchronized algorithm is provided to calculate such correlation coefficients.

This algorithm is a stochastic process, since it involves a synchronized sampling. In practice, if the *p*-value is not considerably different from the critical value of interest, such as 5 %, 1 %, etc., then this algorithm needs to run numerous times (20 in our case) in order to reduce the computational fluctuation. The average correlation coefficient from these is taken to be the resultant correlation coefficient for the significance test.

When conducting comparisons, the 95% CIs can be examined intuitively. It is difficult, however, to reach any conclusion when the two 95 % CIs overlap. Determining whether the difference is real or by chance may be addressed using a significance test. As presented in Section 7, although the 95 % CIs of Systems EL and UJ did slightly overlap, the hypothesis testing showed that the difference in performance levels between these two algorithms was statistically significant.

The pairwise comparison conducted after obtaining *a priori* knowledge from the relationship among the 95% CIs as described in Section 7 is to show how crucial the significance test is if the two 95% CIs overlap while determining whether the difference between the performances of the two algorithms is statistically significant. If the confidence intervals for any combinations of algorithms are of interest, for instance, then some multiple comparison procedures, such as Tukey's method, Scheffe's method, Bonferroni's method and so on, might need to be employed [16, 17, and references therein].

Conventionally, if the two-tailed *p*-value is greater than or equal to 5 %, the null hypothesis is not rejected; if it is less than 5 %, the null hypothesis is rejected in favor of the alternative hypothesis. In the literature [5], it is alternatively suggested: If the *p*-value is less than 0.10, borderline evidence is

against H_o; if the *p*-value is less than 0.05, reasonably strong evidence is against H_o; if the *p*-value is less than 0.025, strong evidence is against H_o; if the *p*-value is less than 0.01, very strong evidence is against H_o.

References

1. "The NIST Speaker Recognition Evaluation", the URL of the website is at http://www.itl.nist.gov/iad/mig/tests/spk/ (2012).
2. J.C. Wu, A.F. Martin and R.N. Kacker, Measures, uncertainties, and significance test in operational ROC analysis, J. Res. Natl. Inst. Stand. Technol. 116 (1), 517-537 (2011).
3. 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).
4. B. Efron, Bootstrap methods: Another look at the Jackknife, Ann. Statistics 7, 1-26 (1979).
5. B. Efron and R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, (1993).
6. 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).
7. R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha and A.W. Senior, Guide to Biometrics, Springer, New York, 269-292 (2003).
8. 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).
9. 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, in press (2013).
10. J.C. Wu and C.L. Wilson, Nonparametric analysis of fingerprint data on large data sets, Pattern Recognition 40 (9), 2574-2584 (2007).
11. J.C. Wu, A.F. Martin, C.S. Greenberg and R.N. Kacker, Uncertainties of measures in speaker recognition evaluation, in Active and Passive Signatures II, Proc. SPIE 8040, 804008 (2011).
12. J.C. Wu, A.F. Martin, R.N. Kacker and C.R. Hagwood, Significance test in operational ROC analysis, in Biometric Technology for Human Identification VII, Proc. SPIE 7667, 76670I (2010).
13. B. Ostle and L.C. Malone, Statistics in Research: Basic Concepts and Techniques for Research Workers, fourth ed., Iowa State University Press, Ames, (1988).
14. G. E. P. Box, J. S. Hunter and W. G. Hunter, Statistics for experimenters: design, innovation, and discovery, second ed., John Wiley & Sons, Inc., New York, (2005).
15. R.J. Hyndman and Y. Fan, Sample quantiles in statistical packages, American Statistician 50, 361-365 (1996).
16. H. Abdi, The Bonferroni and Sidak corrections for multiple comparisons, in Encyclopedia of Measurement and Statistics, ed. by N. Salkind, Thousand Oaks (CA): Sage, (2007).
17. Engineering Statistics Handbook, by the National Institute of Standards and Technology, at http://www.itl.nist.gov/div898/handbook/ (2012).