Allowing Good Impostors to Test

John M. Colombi J. Scott Reider
Department of Defense
Ft George G. Meade, MD 20755-6451

Joseph P. Campbell

Abstract

Biometric testing should attempt to report unbiased, real-world system performance, especially when tested on limited databases. Though testing on a standard database, such as the Linguistic Data Consortium's YOHO, allows comparison of speaker verification systems, it is well known that certain procedures bias the results low. One such procedure concerns the use of cohort or reference speakers to perform verification, where the cohort speakers are removed as candidate impostors. A method of testing is proposed to remove this bias by modifying the cohort set for each false acceptance test. Results statistically differ for this modified approach, which tries to "best" model the general population with a fixed random sample. Lastly, three techniques to bound the biometric performance, using both parametric and non-parametric resampling is demonstrated.

1. Introduction

Biometric systems often report their performance based on many different testing methodologies. In ICASSP-95, Campbell [1] recommended test procedures for the LDC YOHO Speaker Verification database. These procedures included intra-gender testing and the use of 5 cohort or reference speakers for score normalization. Campbell described the bias dilemma, hypothesizing that reported results will be biased low (decreased False Acceptance rates) when either excluding or allowing cohort speakers. Cohort speakers normalize the test score to remove various word effects of the utterance and have been shown to greatly improve verification accuracy [2]. This article first examines an alternative testing methodology to remove this bias from the reported Equal Error Rate (EER) or Zero Rejection Rate (ZRR) performance of a hidden Markov model speaker verification system. Next, we attempt to provide methods for bounding the EER with confidence intervals.

2. HMM Speaker Verification System

The Linguistic Data Consortium's (LDC) YOHO Speaker Verification database is the only large scale, scientifically controlled and collected, publically available database for speaker verification which allows testing at high confidence levels. The 106 male subjects are examined in this paper, each having provided 96 enrollment utterances and 40 test utterances over a 3-month interval. The speech material consists of "combination-lock" phrases. An example prompt is: "57 - 26 - 64", pronounced "fifty-seven, twenty-six, sixty-four". The total number of words is sixteen producing 56 possible doublets and a list of 166,320 phrases.

Front end analysis of the raw YOHO audio files are pre-emphasized, framed every 20 msec, then 12th order Melwarped Frequency Cepstral coefficients are extracted. We append normalized energy, then augment this 13 dimension vector with both first and second order transitional coefficients forming a 39 dimensional final representation.

Each YOHO male speaker is modeled by a set of 21 hidden Markov models [3] representing context independent phonemes, Table 1.

Table 1. YOHO phoneme list, with silence (sil) and interword space (sp).

|   |   |   |   |   |
|---|---|---|---|---|
| AH | AO | AX | AY | EH |
| F  | IH | IY | K  | N  |
| T  | TH | UW | V  | W  |
|    |    |    |    | sil |
|    |    |    |    | sp  |
Each hidden Markov model consists of 3 states represented by a mixture of 3 Gaussian densities. A single diagonal covariance is shared amongst all models. Each speaker template is thus modeled by 7788 parameters. An utterance, $U$, is scored against a set of speaker $i$ models, $\lambda_i$, using forced alignment (Viterbi decoding) based on the known transcription. This procedure provides a normalized log likelihood of the utterance given the models, $\log p(U|\lambda_i)$. This method gave best performance over other less-constrained Viterbi decoding methods [3]. Training iterations use Baum-Welch reestimation, with initial model bootstrapped from previously trained, similar TIMIT phoneme HMMs.

3. Cohort Normalization

For a particular speaker, enrollment data is used to choose the closest set of speakers, known as cohorts. Reynolds [9] provides a symmetric distortion measure between two models using enrollment utterances from both the target and the potential cohort to determine similarity. If $U_i$, $\lambda_i$ represent speaker $i$ training observations and model respectively, then a symmetric distortion measure can be defined as

$$d_{sym}(\lambda_i, \lambda_j) \equiv \log \frac{p(U_i|\lambda_i)}{p(U_i|\lambda_j)} + \log \frac{p(U_j|\lambda_j)}{p(U_i|\lambda_j)}$$

Stepping through each target speaker $i$, we examine the closeness of each potential cohort speaker $j$. The end result of this process is a sorted list of cohort speakers, unique for each target speaker, which will provide a beneficial normalization to the claimed speaker's score. Other methods of defining cohort speakers include geometric mean, maximum pick within cohort pool and second order Bhattacharyya distance.

The likelihood ratio test is a useful tool based on Bayesian analysis for performing speaker verification. The Bayes error rate, a statistical upper bound on performance of any pattern classifier, is achieved by applying the Bayes decision rule. This log likelihood ratio, using probability density functions, either known or approximated, is as follows:

$$\log \frac{p(U|\lambda_c)}{p(U|\lambda_o)} \geq T, \quad \text{where} \quad T \equiv \log \frac{p(\lambda_c)}{p(\lambda_o)}$$

Speaker verification systems are then based on this log-likelihood ratio $L$ of the utterance (or set of utterances) by applying the concept to a claimed model ($\lambda_c$) against not the claimant or impostor ($\lambda_o$),

$$L(U) = \log p(U|\lambda_c) - \log p(U|\lambda_o)$$

(1)

If the above quantity is greater than the threshold $T$, which accounts for the unknown $a$ priori, the maximum likelihood decision is to accept the utterance $U$ as the claimed speaker. We seek to approximate this last term using a set of "close" reference speakers, as suggested by Higgins [7]. Campbell establishes methods for testing on YOHO by calling these reference speakers "cohorts."

Researchers, such as Rosenberg [10] and Furui [11] present several measures for cohort normalization, each an approximation to the last expression of the log likelihood (Equation 1). Specifically, define a set of cohort speakers $S$ of size $|S|$. Then, using the geometric mean over the set of cohort speakers, for example, the log-likelihood ratio is given by

$$L(U) \approx \log \frac{\prod_{j \in S} p(U_j|\lambda_c)}{\prod_{j \in S} p(U_j|\lambda_j)^{|S|}}$$

(2)

Thus, this cohort normalization simply becomes the average of the log likelihood scores over the cohort set.

The overall speaker verification system reported here operates as follows [3], shown in Figure 3. Each utterance is scored, using forced alignment, by the likelihood the utterance was generated by the claimed speaker's model. To remove word effects, this score will be normalized by 5 or 10 reference speakers for the claimed speaker. Lastly, this normalized score will be compared to a global, speaker-independent threshold. Equal error rates are often reported which vary this threshold until the false acceptance error rate equals the false rejection error rate. The zero rejection rate occurs when the threshold is set to have a false rejection rate equal to zero.

4. Cohort Replacement Results

Typically, results are shown for tests excluding cohorts as impostors. We examine the statistical difference when cohorts are allowed to test, as well as when we apply the Modified cohort set methodology. This new method includes a cohort speaker as an impostor, but removes their score in the log-ratio normalization, while adding the next closest cohort. This insures the cohort set size remains constant. This approach attempts to model good impostors in the general population which are not known or available for inclusion in the cohort or reference pool. For each of the 106 targets, the Modified method iterates through the testing of all 105 possible impostors. Equal error rates are shown in Table 2 with zero rejection rates provided in Table 3.

Table 3 depicts the same testing reported at a false reject rate of zero, or the ZRR. This reporting allows a comparison of the statistical difference between the test methods and their corresponding results. Given the number of impostor testing trials performed, the Removed and Allowed testing methods in all cases were statistically equivalent, based on a
95% confidence interval, while the Modified testing method produced results that differed significantly from both methods.

In Figure 2, we examine the cumulative percentage of the False Acceptance errors attributed to each of the 20 closest cohort speakers. This procedure demonstrates to what degree the cohort speakers contribute to False Acceptance errors. For example, using the typical cohort set size of 5 and testing with 4 combination lock phrases, the first 5 cohorts account for only 2% of the total False Accept errors when tested using the Allowed method. This test confirms Campbell’s hypothesis that a speaker would be rejected when their model contributes to the reference score. When allowed to test with the Modified approach, these first 5 cohorts account for 32% of the total False Accept errors.

5. Bounding the Error Rate

In addition to using a fixed database wisely, it is often useful to provide bounds on the performance. The question of independence of biometric samples is often doubted and scrutinized. This section examines reporting equal error rate (EER) with confidence intervals, using both parametric and non-parametric techniques.

5.1. Parametric Strategies

It should first be noted that a data set provides not only an estimate of the EER but also the accuracy of EER, namely the standard error of estimation [6]. This error of estimation can be applied to statistical inference to infer acceptability of a biometric system or to compare the statistical difference between two systems - an important tool when many biometric systems are beginning to hit the market. Parametric approaches to confidence intervals attempt to uncover the standard error of an estimate. For example, letting the number of correct false accept (impostor) errors be \( k \) under \( N \) independent tests, the point estimate of the error rate \( p \) is \( \hat{p} = k/N \). For large \( N \), we use the DeMoivre large sample approximation to the binomial resulting in a \( (1 - \alpha) \) confidence interval expressed by

\[
\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{N}}
\]

using the unit normal \( z \) percentile. Similarly, this same procedure could be applied to the false reject test. Combining both FA and FR confidence results could be done simultaneously on a receiving operating characteristic (ROC) curve. This method remains valid whenever the normal density assumption holds true for the statistic under considera-
Figure 2. Percent of False Accept errors explained by cohorts. Curves indicate the three methods of testing.

Figure 3. Bootstrap samples from the log ratio scores taken from two distributions of true (FR) and impostor (FA) scores.

5.2. Resampling Strategies

Resampling, or bootstrapping, is a method of computational statistics which often avoid many parametric assumptions concerning distributions of the data[6,5,4]. For the verification problem reporting EER, a two distribution method of resampling must be accomplished [6,4].

This method estimates the underlying distribution \( G \) from the empirical probability distribution of the actual data, using a Monte Carlo method. The bootstrap sample is created by making \( N \) independent draws from the data, with replacement. Then, the bootstrap statistic is evaluated on the bootstrap sample, \( e = e(g_1, g_2, \ldots, g_N) \). After a large number of bootstrap estimates, the standard error can be evaluated or the nonparametric confidence intervals from the bootstrap histogram. The \( 1 - 2\alpha \) confidence interval simply uses the 100\( \alpha \) and 100(1-\( \alpha \)) percentiles of the bootstrap histogram.

Specifically for the statistic EER, a sampling from two distributions, \( G \) and \( H \), is required. Refer to Figure 3. Let \( e \) be the estimate of EER, a statistic of both the random samples from the false acceptance scores, \( g_1, g_2, \ldots, g_N \) and the the false rejection scores, \( h_1, h_2, \ldots, h_M \). Using Ephron's notation, create a bootstrap sample,

\[
e^*_b = e(g_1^*, g_2^*, \ldots, g_N^*, h_1^*, h_2^*, \ldots, h_M^*)
\]

where, for example, the \( g_i^* \) are sampled with replacement from the original false accept scores. Each bootstrap sample consists of randomly sampling \( N \) FA scores and \( M \) FR scores, then estimate the desired statistic, such as EER.

As suggested by Campbell in ICASSP 95 [1], one way of removing bias from the reported EER statistic (biased low), was to resample the speaker pool. Instead of fixing the scores and performing a resampling of those EER scores, as recently suggested by Diegert [4], one fixes the speaker population and resamples a target pool, a cohort pool and an impostor pool. Refer to Figure 5.2. For example, each bootstrap sample could consists of selecting target speakers, cohort speakers and impostor speakers. Subsequently, for the targets, a sorted list of cohorts will be calculated, based on enrollment data. Using test data, each of the targets will attempt to gain access as themselves resulting in a set of FR scores. Then, the set of impostors will attempt to gain access as each of the targets, resulting in a set of FA scores. The EER for this bootstrap sample is found be varying a global threshold. This procedure is repeated for \( B \) bootstrap samples.

Figure 5 compares each of the three methods of bounding Equal Error rate: resampling the log-likelihood ratio scores, resampling speakers and the parametric confidence interval using Equation 3 in Figure 5. Confidence Interval estimates are the tightest for the the resampling, with replacement of the log ratio scores, and appear widest for the resampling of speakers.
6. Discussion

This research first examined three test procedures for biometric systems. The goal is to unbiased the impostor testing to ensure that reliable and realistic statistics are reported. We report our equal error rates with two different cohort set sizes, two variations on combinations/test, and three impostor testing methods. We analyze the confidence interval for the difference between the Modified method and either the Removed or Allowed methods. In all cases of Table 3, we conclude there is a significant difference, using a 95% confidence interval, between the Modified ZRR and both the Removed and Allowed ZRRs. We have proposed a method where results are less biased than standard reporting procedures. Next, we then review three methods of bounding the true system performance of a speaker verification system.

By far the simplest method involves assuming the true EER is within a symmetric confidence interval about the point estimate of FER. The other two techniques use computational statistics to repeatedly resample either the log ratio scores or speakers to provide an non-parametric estimate of the EER histogram. All three techniques provide similar results, reported here for 1, 2 and 4 combinations/test.

References

[1] J. P. Campbell, Jr. Testing with the YOHO CD-ROM voice verification corpus. In Proc. of the 1995 ICASSP, pages 541-545, 1995.
[2] D.-S. Y. ChiWei Che, Qiguang Lin. An HMM approach to text-prompted speaker recognition. In Proc. of the 1996 ICASSP, volume I, pages 673-676, 1996.
[3] J. M. Colombi, T. R. Anderson, D. W. Ruck, S. K. Rogers, and M. Oxley. Cohort selection and word grammar effects for speaker recognition. Proc. of the 1996 ICASSP, 1:85-88, May 1996.
[4] K. V. Diegert. Estimating performance characteristics of biometric verifiers. In The Eight Biometric Consortium, San Jose State College, CA, June 1996.
[5] B. Efron. Computer and the theory of statistics: Thinking the unthinkable. SIAM Review, 21(4):460-480, October 1979.
[6] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37(1):36-48, February 1983.
[7] A. Higgins, L. Bahler, and J. Porter. Speaker verification using randomized phrase prompting. Digital Signal Processing, 1:89-106, 1991.
[8] P. G. Hoel. Introduction to Mathematical Statistics. John Wiley & Sons, Inc., New York, NY, fourth edition, 1971.
[9] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, 17(1-2):91-108, 1995.
[10] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong. The use of cohort normalized scores for speaker verification. In Proceedings for the 1992 International Conference on Spoken Language Processing, volume I, pages 599-602, 1992.
[11] F. Sadaoki. An overview of speaker recognition technology. In Proc. of the 1994 ECSA Workshop on Speaker Recog. Ident. and Ver., 1994.