How to Test the Performance of Speech Verifiers and Statistical Evaluation

Jörg Tacke & Andreas Wolf

VOICE.TRUST AG*
Landshuter Allee 12-14
80637 Munich, Germany
{jt|aw}@voicetrust.de

Abstract: Biometric identification and verification technologies, in the past, have promised high performance levels. Such performance statements lead to the assumption, that these biometric systems are highly secure. Field data tests have shown substantial discrepancy compared to specified error rates. In order to reflect target scenario deployments when gathering test data, we suggest to acquire test data from actual deployments, which implies a test population reflecting that of the target group. Four impostor levels are defined. For statistical analysis we suggest Sequential Testing according to Wald, in order to minimize population size and still show the statistical significance of low empirical error rates.

1 Introduction and Related Work

Performance Reports of biometric deployments share one predominant reason as to why biometric systems have not achieved the market penetration and acceptance as predicted by major research institutions: Unacceptable error rates [DPA03]. Currently, there is no standardization in methods for evaluation of biometric system performance. Each vendor develops unique test procedures and uses data sets of varying size, quality and origin to establish performance metrics. Test procedures must cover issues relevant to deployments, clearly provide a basis for objective comparison of biometric system performance based on compliant reports and disclosures. Part of evaluation must be an accepted, feasible statistical analysis to provide statistical significant results. This will allow customers and integrators to accurately assess system cost versus performance based on their particular needs [ISO03].

2 Testing

In order to measure the performance level of a given system and its changes due to design changes or environment changes, we suggest to design a test suite, which allows to execu-

*VOICE.TRUST AG, founded in July 2000, is the leading producer of secure authentication solutions via digital speech verification. VOICE.TRUST solutions lead to a dramatic reduction in operating costs over conventional authentication solutions of up to 80 percent. Areas of use include secure authentication for PIN and Password Reset, Single Sign-On, Remote Access or Two-Factor Authentication in the areas of network security, call center, mobile banking and e-commerce.

te reproducible and comprehensible technical tests regarding authentication performance. As the False Acceptance Rate (FAR) is the only rate primarily effecting security, the False Rejection Rate (FRR) should also be measured to indicate the comfort level of the product and as an indicator of how a user could be motivated to bypass the system due to low comfort level. It must be noted, that the FAR is not the only factor to be considered when evaluating security of a biometric system. [BEM02] gives guidance on which factors to consider when executing a Common Criteria (CC) evaluation for biometric systems and clearly demonstrates, that the FAR is not the only security risk: capture, extraction, template storage, administrator and resource manager threats, just to mention a few security issues of general authentication systems. As described in [MW+02], many indicators may be tested. Most efficiently though are:

1. FAR: Claims of identity are incorrectly confirmed.
2. FRR: Truthful claims of identity that are incorrectly denied.
3. Failure to Enrol Rate (FER): Expected portion of the population for whom the system is unable to generate repeatable templates.
4. Failure to Acquire Template Rate (FATR).

Testing Requirements. In order to avoid suffering the wide spread difference between laboratory results and field test results, we recommend to acquire test data from actual deployments. Hence, we do not follow [BEM02] in as far as tests should be carried out in controlled environments that match the environment defined in the security area. The test user population is required to be large enough to be representative [MW+02] and statistically significant.

Test Design. Data acquisition strategy must follow a pattern matching the normal use of the product. This refers to input devices, transmission devices, background noise, dialogue deployed, and the psychological authentication scenario effect: It is observed from a face recognition deployment at a nuclear power plant, that enrolments not executed at the actual point of authentication cause a FRR of $> 90\%$. To ensure reproducibility and comprehensibility, data sets must be acquired via a Data Acquisition Module (DAM), then processed via the Data Processing Module (DPM). DAM must ensure quality of templates acquired. DPM will encompass the verification process as well as the decision process as to accept or deny identities' claim. The number of test data sets required must be estimated using formulas described later in this paper. The test data acquisition and processing strategy consequently must follow the procedure deployed in the product, and must use the products components. As impostor testing must be a focus of biometric testing due to biometrics connotation to security, we defined four impostor levels:

- Impostor Level 1 equals the zero effort attack.
- Impostor Level 2 equals the informed impostor attack: Impostor has knowledge of correct answers to system prompts.

- Impostor Level 3 equals the trained impostor attack. Impostor has been trained on true identities pronunciation. This may be achieved by playing true identities sound files to impostor.
- Impostor Level 4 corresponds to replay attacks. Impostor attempts must be made under same conditions as genuine attempts. The use of so-called *background databases* of biometric features acquired from different (possibly unknown) environments and population cannot be considered best practice [MW+02].

3 Calibration and Comparative Testing for Future Product Releases

For future calibration and testing purposes with different test scenarios or environments (test population etc.), we suggest to include the acquired test data into the product. This will only be of substantial value, if data acquisition etc. are properly documented, so future sample population and data sets can be acquired in comparable manner. Collected biometric data is referred to as *corpus*, information about them is referred to as *database*. Volunteer consent form must be signed by sample population.

Security of Test Data Integrity. Security is measured by the FAR. Mapping of every utterance must be secured by a authentication procedure. This may be achieved by issuing a PIN for every test user. Analysis of, e. g., CLIP (calling line identification presentation), or the telephone number of the calling party is not sufficient as it authenticates the calling device but not the identity. In order to avoid data collection errors, mechanisms ensuring the quality of data sets as well as proper labelling must be implemented. In speech verification, speech recognition in combination with a. m. PIN procedures can prevent volunteers from using wrong PINs or typing in wrong PINs.

Demographic Analysis. Along with the acquisition of biometric data sets, demographic information of the identities should be collected for future variance analysis (i. e.: Name, gender, age, education, geographic location of place of birth, experience with biometric systems, telephone number, quality of service, quality of device). This enables demographic and variance (S^2) analysis of the test population and its impact on performance, which in turn allows performance estimation of future target test population: No matter what the size of the test population, and no matter how good the match with the standard distribution population is, a prospect customer will always argue that *his* population is different and not representative of his target population. Hence we suggest, to carefully analyze the given test population and its impact on system performance, which in turn allows performance estimates for specific populations.

4 Statistical Evaluation

Undoubtedly, due to unknown correlations and anticipated error rates it is advisable to maximize the test population for statistical significant results [BEM02]. On the other hand, one must not ignore *the law of diminishing returns*: A point will be reached where errors due to bias in environment used, or in volunteer selection, will exceed those due to

size of the crew and number of tests [MW+02]. For correct interpretation of measures, the following information is required: Details of the test scenario, details of the operational application, demographics of the volunteer crew, details of the test environment, time separation between enrolments and test transactions, quality and decision thresholds used during data collection, details of which factors potentially affecting performance were controlled, and how these were controlled, details of test procedure and policies for determining enrolment failures etc.. At the same time demands for field tests in actual deployments, in order to avoid the laboratory testing problem in biometrics [MW+02], must be considered. The number of factors required for correct interpretation of measures and the problems with a large test population indicate, that significance is a function of test size and quality of information regarding the sample acquisition. Hence, in the best case the aim must be to find a method, that is proven in other areas faced with a similar problematic, and allows to minimize sample size in order to prove very small error rates.

The estimation of automated biometric-based verification systems can be formulated as a parameter estimation problem based on the outcomes of repeated Bernoulli experiments. A Bernoulli experiment is a random experiment that has only two classes of outcome: Success or failure [SSK03]. For a biometric system, success means a correct acceptance or denial of a user. Given the independence of the test trials, results will show binominal distribution. Empirical error rates, also called point estimation, need to be substantiated by investigating their significance, which depends on the test sample size and the estimated error rate.

Stochastic Independence of Attempts. If A and B are independent events, A can occur on a given trial regardless of B occurring, and vice versa. In other words, it is possible for A to occur without B occurring, A and B to occur, or B to occur alone [MA03]. The assumption of independent identically distributed attempts may be achieved, if each genuine attempt uses a different volunteer, and if no two impostor attempts involve the same volunteer. With n volunteers, we would have n genuine attempts and $n/2$ impostor attempts [MW+02]. The problem with this approach is, that due to the variance in speech a single independent attempt of n users will not provide information about, e. g., the FRR for a user in n attempts. The logical consequence from this is, that restricting data to a single attempt per volunteer does not contribute to the aim of determining error rates significant to actual deployments.

Six methods for investigating the confidence level of the point estimates are considered:

1. Rule of 3,
2. Rule of 30,
3. Confidence Interval,
4. Statistical Hypotheses Testing,
5. Sequential Testing according to Wald,
6. Test procedure proposed in BEM.

Rule of 3. Applying the *rule of 3* with $p \approx 3 \div N$ for a 95% confidence level [JL97], a test of 85 independent trials returning no errors results in an error rate of 4% or less with 95% confidence. At 90% confidence, the error rate calculates to 2% or less.

Rule of 30. Applying the *Rule of 30* as suggested by Doddington [Dg+00], 30 errors must be observed to be 90% confident, that the true error rate lies within ±30% of the observed error rate. To be 90% confident, that the true error rate is within ±10% of the observed error rate, at least 260 errors are required. To prove a 1% FRR and 0.1% FAR, 3000 genuine attempt trials and 30.000 impostor attempts are demanded. Considering independence, 3.000 enrolees and 30.000 impostors are required.

Confidence Interval. To be applicable, the test data size must be sufficiently large according to [Hj87]: $n * p * (1 - p) \geq 9$. with n = number of trials, p = estimated error probability. The value for $u_{1-\alpha/2}$ can be taken from tables [Hj87]. The confidence interval will be calculated using the following formula for lower limit p_1 and upper limit p_2:

$$p_1 = \frac{2m + u^2_{1-\frac{\alpha}{2}} - u_{1-\frac{\alpha}{2}}\sqrt{u^2_{1-\frac{\alpha}{2}} + 4m(1-\frac{m}{n})}}{2(n + u^2_{1-\frac{\alpha}{2}})} \quad \text{and} \quad p_2 = \frac{2m + u^2_{1-\frac{\alpha}{2}} + u_{1-\frac{\alpha}{2}}\sqrt{u^2_{1-\frac{\alpha}{2}} + 4m(1-\frac{m}{n})}}{2(n + u^2_{1-\frac{\alpha}{2}})}$$

Example: Using test data of 85 independent trials with FAR=0, then p=0. Selecting a error probability of 0.005 which results in confidence interval (0;0.057). So the real error probability lies between 0% and 5.7% with a confidence of 99.5%. Now we can check the condition of the above equations for the boundaries of the confidence interval. This condition is not satisfied, so this method cannot be applied [GT02].

Statistical Hypothesis Testing. A given hypothesis (zero hypothesis) is proven on the basis of test data. The result will be confirmation or disavowal of that thesis. This method can only be applied, if a perception of the error rate exists. So the result of this method will be to reach a yes/no conclusion regarding the value of that error rate. Let us assume we are trying to prove, that the FAR (p) is smaller p_0. Zero hypothesis (h0): $p<p_0$. Alternative hypothesis (h1): $p \geq p_1$. The test statistic is calculated using

$$Z = \frac{m - np_0}{\sqrt{np_0(1-p_0)}}$$

With m=number of errors and n=number of trials. In order to calculate the required data test size before beginning data acquisition, the following formula may be used:

$$n \geq = \left(\frac{\sqrt{p_0(1-p_0)}u_{1-\alpha} + \sqrt{p_1(1-p_1)}u_{1-\beta}}{p_1 - p_0} \right)^2$$

Example: The hypothesis is to be proven that the FAR<1%. So the zero hypothesis will be h(0)<p(0), with p(0)=0,01. The supplier selects a error probability α=0.05, error of first kind, that the outcome of the test will be negative. The customer selects an error probability β=0.05, that the outcome of the test will be positive. The customer selects p(1)=0.04.

Thus, the alternative hypothesis will be: h(1): FAR$\geq$p(1) with p(1)=0.04. Using the before mentioned formulas, the minimal amount of test data sets can be calculated: 263. As the available number is 85, this method cannot be applied [GT02].

Sequential Testing According to Wald. This method of testing was developed prior to 1945 and found wide acceptance in fields where two requirements were demanded: very small sample size and very small error rates had to be proven with statistical significance. Hence, the Wald scheme found wide acceptance in material testing, where the object to be tested will be destroyed by that test. Thus, sequential testing was developed. By verifying the test hypothesis after each experiment, the required sample size can be substantially reduced. This is explained by the graph in Figure 1. After each trial, test data will be plotted into that graph: number of errors on y-axis, number of trials on x-axis. The coordinate of the last result defines the next step: Is it above the top line, the hypothesis must be dismissed, is it below the lower line, the hypothesis is accepted. Is it located between the lines, the test must be carried on.

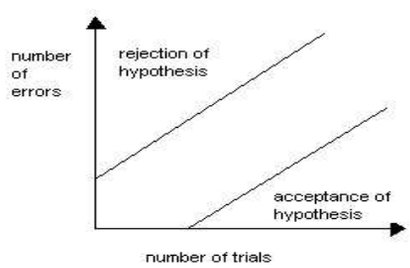

Figure 1: Sequential Testing according to Wald.

The procedure is to be carried out in three steps: 1. Define test hypothesis and error tolerances. H0: FAR$<p_1$ with p_1=0.01 against H1: FAR$>p_1$ with p_1=0.01. 2. As before, the parameters α=0.01, β=0.01, and p_2=0.05 must be selected. The lines in above graph must be calculated: Upper line: $y_1 = cx + b$ and lower line $y_2 = cx - a$.

$$a = \frac{-\ln\frac{\beta}{1-\alpha}}{\ln\frac{p_2}{p_1} * \frac{1-p_1}{1-p_2}} \qquad \text{a=b, if } \alpha=\beta \qquad c = \frac{-\ln\frac{1-p_1}{1-p_2}}{\ln\frac{p_2}{p_1} * \frac{1-p_1}{1-p_2}}$$

With above values we get a=2.74 and c=0.025. The third step comprises the actual testing. With 85 trials with no false acceptance, the data points will move along the x-axis. Consequently, the hypothesis can be accepted, once the lower line intercepts with the x-axis. With the values c and a, $y_2 = 0.025x - 2.74$. Selecting y_2=0, x calculates to 109.6. So if up to trial number 109 no false acceptance occurs, the test hypothesis can be accepted [GT02]

BEM Test Procedure. This procedure may be explained using Figure 2 [BEM02]. The data point will be plotted into the graph as a function of error rate observed and error rate claimed in n independent comparisons. Depending on the coordinates, the test is

concluded if either the claim is supported or rejected. The test must be continued, if the claim is not supported.

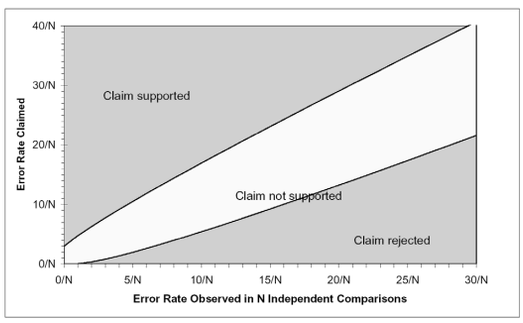

Figure 2: Sequential Testing according to BEM.

Example: A claimed FAR of 1% and sample size of 85 with no errors observed, x=0 and $y = 0.01 * 85 = 0.85$. Plotting that data point into above graph leads to claim not supported. Following this procedure, 300 trials with no errors observed lead to support the claim of an FAR of 1%. Alternatively, BEM suggests a cross-comparison approach, which does not ensure statistical independence, following the Gaussian Elimination. With P people, cross-comparison of attempts/templates for each unordered pair, may exhibit a low degree of correlation. The correlation within these P(P-1)/2 false match attempts will reduce the confidence level for supporting an FMR claim compared with the same number of completely independent attempts. BEM suggests to support a claimed FAR of 1 in 10000 with a crew of 250 individuals. N=P(P-1)/2 with P=250 results to N=31.125 attempts. It must be noted, that [BEM02] does not provide any references regarding mathematical or statistical sources, nor is the procedure presented in detail and therefore not comprehensible [GT02].

Assessment of Statistical Evaluations. Sequential Testing according to Wald requires the lowest sample size required for proving statistical significance at a given error rate. The scheme demands a change in test design, as the test hypothesis has to be verified after each experiment. Adopting the Wald scheme, the pressing question on how to deal with biometrics can be secluded with several advantages: 1. The approach is widely tested and accepted. 2. For very low error rates it is the only feasible approach available. 3. It ensures avoiding the trap of *diminishing returns*.

5 VOICE.TRUST Server Testing

The VOICE.TRUST Server is a multi-level-speech-authentication-system. A test suite was designed following the requirements layed out in chapter 2: Using the product's components, DAM and DPM were implemented. Data acquisition strategy followed Impostor Level 2: Informed impostor attack. Impostor Level 4 (replay attack) was not tested in 2003 and is now being tested after implementation of a challange response procedure, using the product's same verification components. Security of test data integrity was ensured by semantically checking the test callers name and user ID as well as the required PIN before allowing data collection.

Volunteers were acquired and managed by a seperate, independant organization. 126 volunteers participated, mentioning four times their user ID, first name and last name and pass phrase for initial enrolment. For false rejection testing volunteers mentioned six times the previously listed utterances. For imposter testing, volunteers were asked to imposter the previously listed utterances for two different individuals. Due to not all volunteers completing their sessions, complete data sets of 64 volunteers were extracted for error rate determination.

Demographic data was collected and analyzed: First name, last name, gender, age, education, geographic location of place of birth, telephone number, quality of telephone connection, make of telephony device used, e-mail address. Calls were made from office and home office environments, using digital and analogue fixed telephone lines as well as cellular phones, thus fulfilling the requirement for actual deployments.

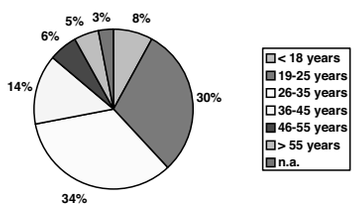

Figure 3: Test participants by age.

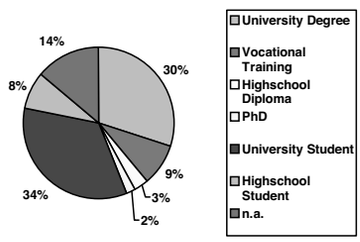

Figure 4: Test participants by education.

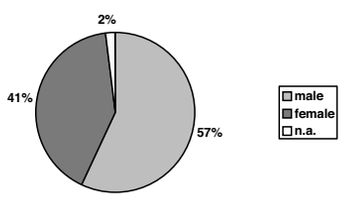

Figure 5: Test participants by gender.

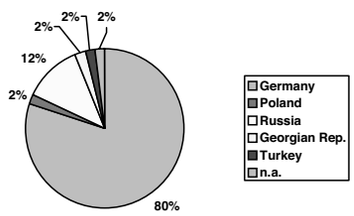

Figure 6: Test participants by origin.

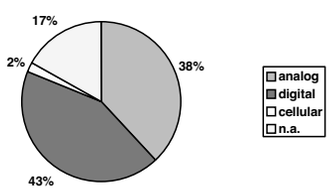

Figure 7: Test participants by line quality.

The identity is asked several times for utterances. As the content of these utterances differs

vastly, minimal correlation could be observed. Future research by the authors will prove this more in detail. Below, the FAR and FRR as a function of decision threshold for each level are shown.

For statistical evaluation of this test please see section 4: Examples used stem from this test.

Total authentication quality can be derived from the following graph, which shows FAR and FRR as a function of authentication session, comprising 3 verification trials each. Please note, that in spite of acceptance and rejection errors at each verification level, the overall authentication performance shows several settings, at which no FA nor FR were observed for all volunteers.

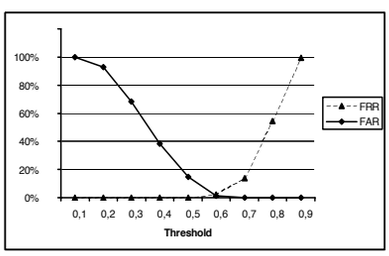

Figure 8: Error rates verification level 1.

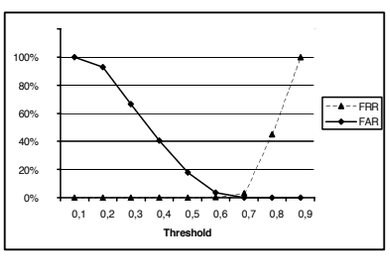

Figure 9: Error rates verification level 2.

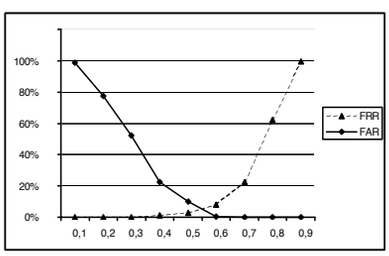

Figure 10: Error rates verification level 3.

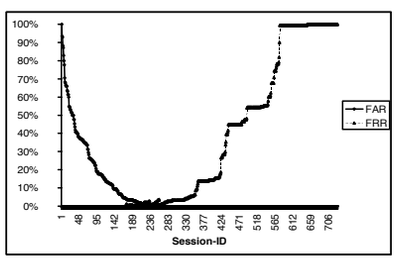

Figure 11: Sorted error rates per session.

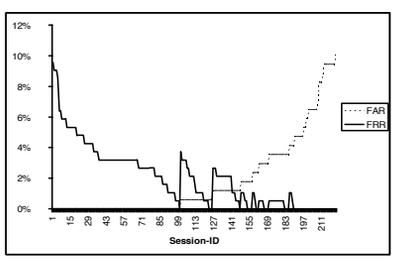

Figure 12: Sorted error rates near EER zone.

6 Conclusion

In face of cumulating news about poor performance results of biometric systems, the most urgent question next to performance improvement is how this performance can be tested in a relevant and feasible kind, and how the test results can be evaluated relevantly. We suggest data acquisition from actual deployments and sequential testing to minimize sample size. The fear of reducing the chance of selling their product, vendors do not clearly supply customers with decision matrixes to define their demands regarding a biometric product beyond the basics. This leads to performance tests with products not designed for the tested scenario. For example, most products do not provide for a live test or only with additional sensors etc. On the other hand, many applications do not require a live test, as a supervisor is present. If, e. g., a finger print system is tested with a silicon finger, of course it will fail causing bad press for that product and the entire technology. Performance requirements must be defined, then the appropriate product selected, embedded and calibrated for the desired application. Vendors must supply customers with the appropriate decision matrixes or consulting services. In any case the customers attention must be directed to this fact. Design of a test suite including features for calibration and comparative testing for future product releases are introduced and test results are presented.

References

[BEM02] Common Criteria Biometric Evaluation Methodology Working Group Biometric Evaluation Methodology, Release 1.0, August 2002, p. 19–21, 23, table 12.

[Dg+00] Doddington, G. R. et al.: The NIST speaker recognition evaluation: Overview methodology, systems, results, perspective. Speech Communication, 2000.

[DHP99] Deller, J. R., Hansen, J. H. L., and Proakis, J. G.: Discrete Time Processing of Speech Signals, pp. 100. IEEE Press Classic Reissue, Wiley Interscience, 1999.

[DPA03] DPA/API, Network World http://www.networkworld.de/index.cfm?pageid=156 & id=90541 & type=detail, 2003.

[GT02] Grans, K., Tekampe, N.: Dokumentation der Testumgebung und Testergebnisse Voice.Trust Server, TV-IT, 2002.

[Hj87] Hartung, J.: Statistik, Lehr- und Handbuch der angewandten Statistik; 1987, Oldenbourg.

[ISO03] ISO/IEC JTC 1/SC 37 N 49, 2003.

[JL97] Jovanovic, B. D., Levy, P.: A look at the rule of three. The American Statistican, 1997, pp137–139.

[MA03] http://www.chance.math.neu.edu/book/independant.html, April 11, 2003

[MW+02] Mansfield, A. J. , Wayman, J. L.: Best Practice in Testing and Reporting, Performance of Biometric Devices; Biometric Working Group, Version 2.01, p. 19, 2002.

[SSK03] Shen, W. et al: Evaluation of Automated Biometrics.Based Indentification and Verification Systems, Stanford University, ccrma-www.stanford.edu/∼jhw/bioauth/general/00628719.pdf, March 25th, 2003.

[Wa45] Wald, A.: Sequential test for statistical hypothese, Annals of Mathematical Statistics, vol. 16, pp.117–186, 1945.