MULTI-FINGER PENETRATION RATE AND ROC VARIABILITY FOR AUTOMATIC FINGERPRINT IDENTIFICATION SYSTEMS

James L. Wayman, Director
U.S. National Biometric Test Center
College of Engineering
San Jose State University

I. Introduction

In previous papers [1-7], we considered performance estimation of biometric systems based on assumptions of measurement independence between fingers. We noted in those papers that such assumptions are generally incorrect, but lacking any data on measure correlations, no quantitative estimates of the effect on system performance were offered. Although measurement correlations effect error rates and throughput of all biometric systems, it is the performance of large-scale identification systems that is most critically effected by data correlations because of the large number of measurement comparisons generally made.

Currently operational, large-scale biometric identification is restricted to Automatic Fingerprint Identification Systems (AFIS). In this paper, we will estimate various measure correlations for AFIS from new fingerprint test data. The multi-finger test data is available for both false match/false non-match comparison errors and binning error/penetration rate estimation. Specifically, in this paper we will estimate penetration rates for single finger systems based on thumb, index, middle and ring fingers, and multi-finger systems for two thumb, two index finger and combined four thumb-index finger systems. Penetration rates calculated from test data are compared to theoretical calculations based on recent finger-dependent pattern classification statistics published by the FBI [8].

We will show Receiver Operating Characteristic (ROC) curves computed with non-matching comparisons differentiated between fingers in communicating and non-communicating bins. Further, we will develop different ROC curves for thumb, index, middle and ring fingers of right and left hands. Finally, the variability of the "impostor" distribution across test samples will be discussed.

II. Test Data

The electronically "live" scanned Philippine fingerprint test data base [3] was used in this test. The data consisted of two sets, enrollment or "training", and "test" data. The training set, consisting of 4080 distinct fingerprints, was taken from 510 individual adult volunteers, each giving eight fingerprints (thumb through ring fingers on both hands). All volunteers were employees of the Social Security System of the Republic of

the Philippines. Most were office and administrative workers and 55% were women. The test set of 4128 prints was collected one to six weeks after the training set from 506 individual volunteers. Of these 506 volunteers, 409 were common to both test and training data sets. Ten volunteers in the test set donated two sets of 8 prints each. 97 volunteers in the training set were not represented in the test set.

A third "practice" set of 80 images from 10 volunteers, whose images were in both test and training sets, was taken 6 weeks after the test database was completed.

Prints were imaged with an Identicator DF-90 "flat" scanner, believed to be "Appendix G" compliant and an "MRT" frame grabber in a lap-top computer. Front-end quality control software from Identicator was employed. The Identicator "Biometric Enrollment System" collection and database management software was used for this project. The prints were stored, using loss-less compression, as "TIFF" images. Some image quality loss, attributable to frame-grabber noise during collection, was noticed in the upper right hand quadrant of most images.

III. Vendor Testing

To date, six AFIS vendors have had their algorithms evaluated against this data. The current test procedure is to send any requesting vendor training, test and practice data sets. The ordering of the test data image files has been randomly scrambled, but the practice images are clearly linked to their corresponding training set images. These practice images allow the vendors to tune any internal parameters required by our data quality or format. Any vendor can request testing of matching and/or binning algorithms.

For the matching test, the vendor returns a 4128x4080 matrix of comparison scores for all test prints compared to all training prints. For the binning test, the vendor returns the bin assignments for all test and training prints, and the rules by which bins are determined to be "communicating" or "non-communicating". In large-scale AFIS system, prints in "communicating" bins are similar enough that they must be compared for possible matching. Upon receipt of all of this data, we release to the vendor the "key" linking the test and training sets.

In this analysis, we used the score matrix from the "best" matching vendor tested to date, meaning the score matrix that produced the generally lowest ROC. We used the binning results from the "best" binning vendor tested to date, meaning the data that we judged presented the best trade-off between penetration and bin error rates. Binning and matching data used here was not from the same vendor. Precise matching values and binning assignments are not discussed here to protect the identity of the vendors.

IV. Finger Dependency of Penetration Rate

It is well known that print classification statistics are finger-dependent. Table 1 shows classification statistics by finger from a recent FBI-authored report [8].

TABLE 1: SINGLE FINGER CLASSIFICATION STATISTICS

Pattern Type	Finger Position										
	1	2	3	4	5	6	7	8	9	10	Ave
Arch	3.01%	6.09%	4.43%	1.24%	0.86%	5.19%	6.29%	5.88%	1.78%	1.15%	3.59%
Tented Arch	0.40%	7.72%	3.20%	1.03%	0.72%	0.58%	7.96%	4.53%	1.45%	1.10%	2.87%
Right Loop	**51.26%**	**36.41%**	**73.38%**	**51.20%**	**83.03%**	0.63%	**16.48%**	1.66%	0.51%	0.12%	**31.47%**
Left Loop	0.46%	**16.96%**	1.47%	1.10%	0.26%	**58.44%**	**39.00%**	**70.30%**	**61.47%**	**86.11%**	**33.56%**
Whorl	**44.77%**	**32.45%**	**17.21%**	**45.24%**	**14.96%**	**35.04%**	**29.93%**	**17.30%**	**34.57%**	**11.33%**	**28.28%**
Scar	0.03%	0.17%	0.13%	0.06%	0.06%	0.04%	0.14%	0.12%	0.06%	0.06%	0.09%
Amp	0.07%	0.20%	0.18%	0.14%	0.12%	0.09%	0.20%	0.20%	0.16%	0.13%	0.15%
Sum	100.00%	100.00%	100.00%	100.01%	100.01%	100.01%	100.00%	99.99%	100.00%	100.00%	100.00%

When each print can be classified only into a single bin, the equation for calculating penetration rate from classification statistics is given in [1] as

$$Pn = p_K + \sum_{i=1}^{K-1} (p_i + p_K) p_i \qquad (1)$$

where Pn is the penetration rate, p_i is the probability that the print is of the i^{th} classification and the k^{th} classification is considered as "unknown". This equation was applied to the data of Table 1. Scarred fingers where considered of "unknown" classification and the data was re-normalized after removal of the amputated finger statistics. Table 2 shows the resulting penetration rates for this approach when fingers in each position are compared to corresponding fingers, right to right, left to left, right to left (or left to right), or all to all.

TABLE 2: SINGLE FINGER PENETRATION RATES FROM FBI STATISTICS

Finger	Penetration Rate			
	Right-> Right	Left-> Left	Right->Left	All -> All
Thumb	0.54	0.56	0.20	0.37
Index	0.44	0.44	0.37	0.40
Middle	0.85	0.83	0.09	0.47
Ring	0.63	0.70	0.23	0.45
Little	0.92	1.0	0.03	0.49

By equation (1), penetration rate will generally decrease with increasing number of classifications of non-zero probability. The 5-type classification system of Table 1 does not represent an optimal approach by any measure and AFIS classification algorithms do not generally use this system. Further, AFIS can place prints in multiple classifications, so penetration rate cannot be determined from classification probabilities using equation (1). The values in Tables 1 and 2 simply make for an interesting comparison when testing AFIS classification algorithms.

To test AFIS penetration rate, we compared the classifications of each training print to those of all other training prints. Using the vendor's rules of "communication", we calculated the percentage of all comparisons that showed communicating bins. Results were differentiated by finger. As mentioned, 409 volunteers were represented in both training and test data sets. Because of errors in the data collection process, there were only about 404 training-test pairs for any particular finger. All comparisons are symmetric. Therefore, there were about 404x403/2= 81,406 non-independent comparisons made for penetration rate.

The penetration rate benefits of fingerprint classification come at the cost of classification errors. If the individual test and training prints of a matching pair are placed in non-communicating bins, the prints will not be matched. To test bin error using the AFIS binning algorithm, we compared binning assignments for each training-test pair based on the bin communication rules. There were about 404 matching pairs for each finger.

Table 3 shows the bin error and penetration rates individually for thumb, index middle and ring fingers. The binning error rate is best for thumbs and left index fingers and worst for right middle and ring fingers. None of the error rate differences between fingers is statistically significant at even the 90% confidence level[1].

TABLE 3: SINGLE FINGER BINNING ERROR AND PENETRATION RATES FROM TEST DATA

Finger	Error Rate		Penetration Rate			
	Right	Left	Right-> Right	Left-> Left	Right->Left	All -> All
Thumb	0.002	0.002	0.70	0.67	0.26	0.47
Index	0.005	0.002	0.46	0.43	0.40	0.42
Middle	0.012	0.007	0.74	0.66	0.29	0.49
Ring	0.010	0.007	0.74	0.66	0.40	0.55

V. Penetration Rates of Multi-Finger Systems

In Reference [1], prediction of penetration and bin error rate performance for systems using multiple fingerprints was discussed under the assumption that the errors

[1] This is established by testing with a cumulative binomial distribution the null hypothesis that observed errors for each finger could have come from the same error probability.

and penetration rates are independent. The general equation for multiple-finger penetration rate can be written as

$$P_{ensemble} = \prod_{i=1}^{T} P_i \tag{2}$$

where P_i is the penetration rate of the i^{th} finger and $P_{ensemble}$ is the total penetration rate of the multi-finger "ensemble". In reality, the binning assignments for thumb, index, midddle, or ring fingers of a person are not independent, but are highly positively correlated. Therefore, we would expect a true penetration rate less than that calculated from equation (2).

Binning error rate for the multi-finger case, again under the assumption of error independence, is given in [1] by

$$1 - \varepsilon_{ensemble} = \prod_{i=1}^{T} (1 - \varepsilon_i) \tag{3}$$

where ε_i is the bin error rate of the i^{th} finger and $\varepsilon_{ensemble}$ is the total error rate for the ensemble. If errors are positively correlated, the value $\varepsilon_{ensemble}$ of will be smaller than calculated using (3).

Using the same AFIS binning algorithm, we tested about 404 finger pairs for left-right thumb, index, middle and ring fingers with every other similar pair in the training data set. Again, these were symmetric comparisons, so there were about 81,406 non-independent comparisons. Both binning errors and penetration rates were measured and are given as Table 4. Included in Table 4 are the error rates calculated from the test data in Table 3 by equation (3) under the assumption of error independence. Test and calculated error rates are identical except for the case of middle fingers. The middle finger test error rate is slightly smaller than that calculated by (3). In the test data of about 404 pairs, there were two instances of classification errors occurring on both left and right middle fingers of the same volunteer. Again, the error rate differences between fingers is not statistically significant.

Also included in Table 4 are the penetration rates calculated from both test and FBI data in Tables 2 and 3 by equation (2) under the assumption of classification independence. Test penetration rates are somewhat (10-20%) higher for all fingers than those calculated using equation (2) from the test data of Table 3, indicating some positive classification correlations between left and right fingers. Test penetration rates are also higher than calculated using (2) with the FBI data from Table 2, except for the middle finger.

Table 5 shows error and penetration rates for four-finger (both thumbs and both index) and eight-finger binning systems. While binning error rates behave as though independent, penetration rates do not. The penetration rate on the four-finger system was

found to be15%, while an assumption of finger classification independence would have lead to a 9% penetration rate based on the single-finger values. The eight-finger system showed a penetration rate of 8%, with a predicted value of 2%.

TABLE 4: TWO-FINGER BINNING STATISTICS

Finger	Error Rate	Error if independent	Penetration Rate	Penetration if independent	
				FBI Data	Test Data
Thumb	0.005	0.005	0.52	0.30	0.47
Index	0.007	0.007	0.25	0.19	0.20
Middle	0.015	0.019	0.55	0.71	0.49
Ring	0.017	0.017	0.55	0.44	0.49

TABLE 5: MULTIPLE-FINGER BINNING STATISTICS

Fingers	Error Rate	Error if independent	Penetration Rate	Penetration if independent	
				FBI Data	Test Data
Four: Thumb and index	0.012	0.012	0.15	0.059	0.093
Eight: Thumb index, middle, ring	0.040	0.048	0.08	0.018	0.022

VI. ROC Curves for Communicating and Non-Communicating Impostor Comparisons

In an AFIS system, submitted fingerprints are binned, then compared only to enrolled prints placed in similar (communicating) bins. We might hypothesize that there is a greater probability for prints in communicating bins to be falsely matched than for prints in non-communicating bins. We computed the ROC for the test fingerprints in three ways: comparing communicating impostors only, comparing non-communicating impostors only, and comparing all impostors. Figures 1 and 2 show three ROCs each for right and left thumb comparisons. We note that the false match rate for the communicating comparisons is almost an order of magnitude greater than for the non-communicating comparisons at some points in the ROC.

FIGURE 1:

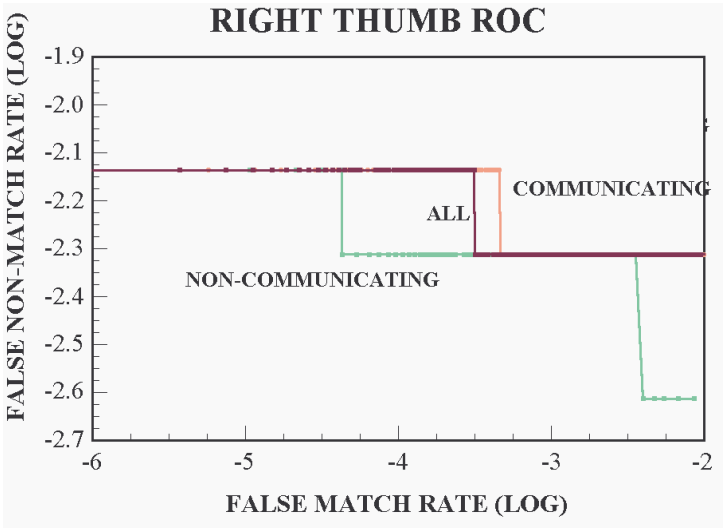
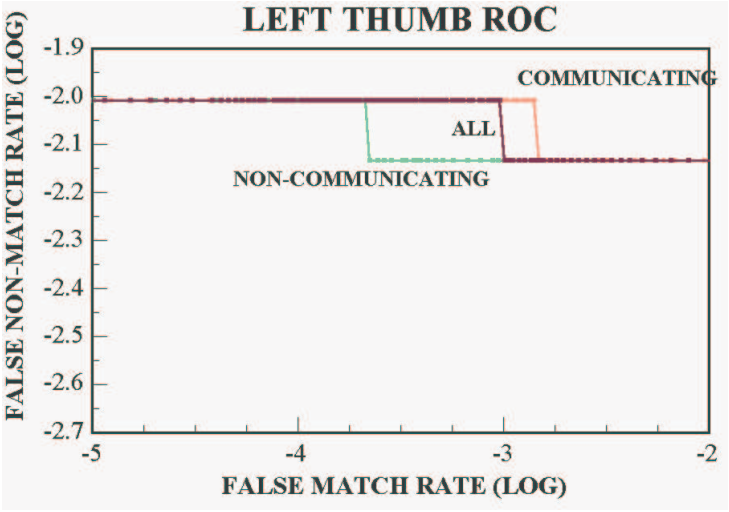


FIGURE 2:



VII. Finger Dependency of ROC

Does the ROC vary depending upon which finger is used? We calculated the ROC for thumbs, index, middle and ring fingers using impostor comparisons only with the same fingers from communicating bins. For example, impostor scores for thumbs were developed by comparing right thumbs only to other right thumbs, and left thumbs only to other left thumbs, with communicating classifications. In all, about 410 genuine comparisons and between 100,000 and 200,000 impostor comparisons were made for each finger. Figures 3 and 4 show right and left hand ROCs for each finger position.

Both graphs show generally increasing error rates as we move from thumbs through ring fingers.

FIGURE 3:

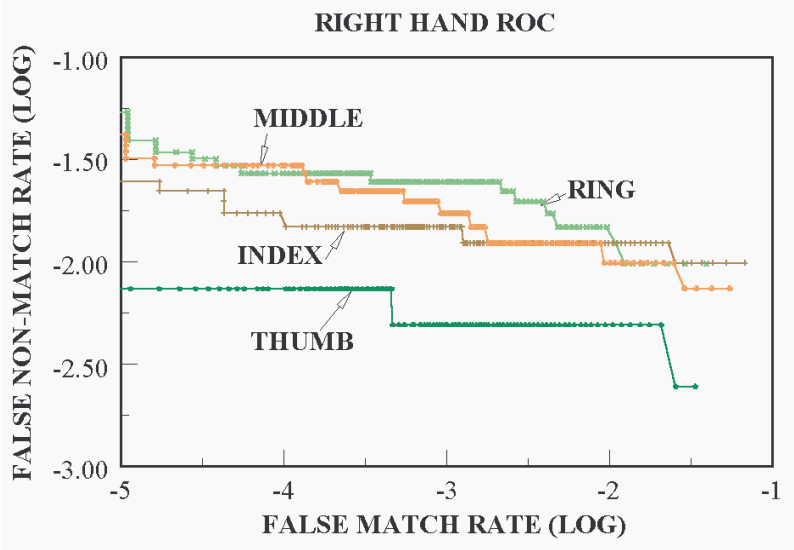
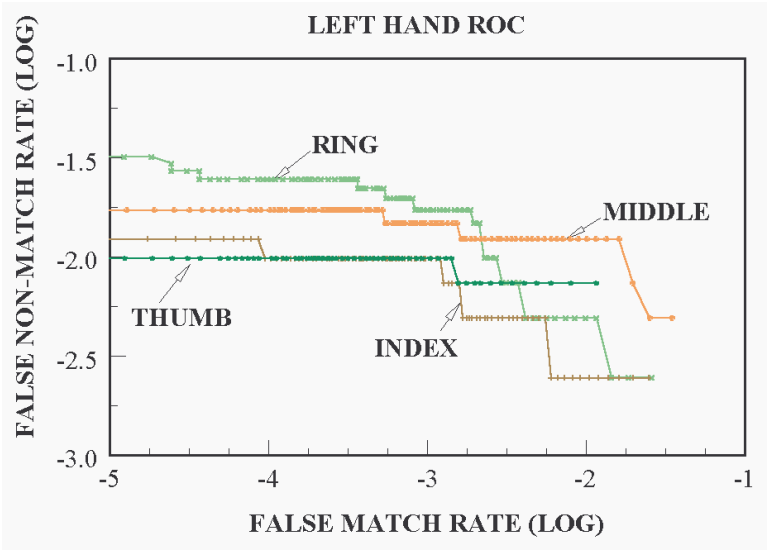
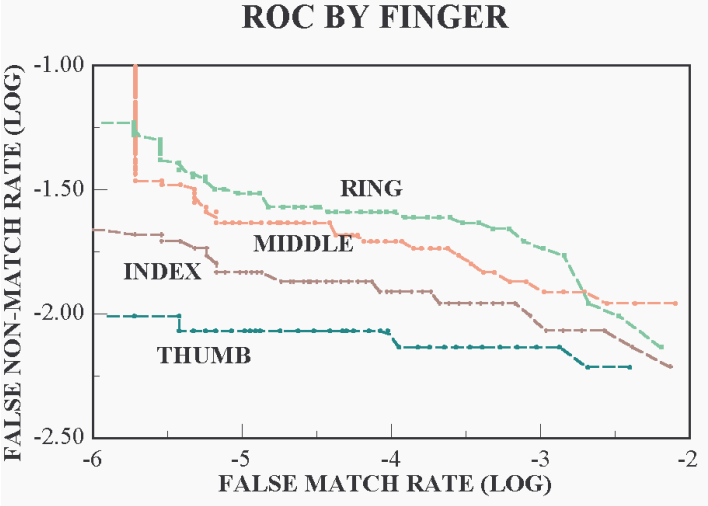


FIGURE 4:



The most notable difference between the right and left hand ROC curves is the difference in thumb error rates, with left thumbs showing worse performance than right thumbs. Figure 5 combines ROCs of both left and right for each finger position and clearly shows increasing errors as we move from thumbs through ring fingers.

FIGURE 5:



We also tested to see if a correlation exists between left and right finger scores for thumb and index fingers of the individual users. Using the non-parametric Kendall's Tau test [9] over about 409 volunteers with eight fingers in both enrollment and test sets, τ = 0.33 and 0.26 for thumbs and index fingers respectively. Comparing ranks of right thumbs to right index fingers, τ=0.28. None of these measures is statistically significant at any significance level, indicating that individual users do not generally have correlated finger scores.

VIII. Impostor Distribution Variation Across Test Samples

Researchers in biometric identification talk about "sheep", "goats", "wolves" and "lambs" to indicate the variability of error rates of a specific biometric system across various users [10]. Most users are "sheep" who can use the system consistently well and are not easily impersonated. "Goats" are those users who cannot consistently be identified. "Wolves" are users who can be easily mistaken for another user in a "zero effort"[2] attack. "Lambs" are users easily preyed upon by "wolves".

In the comparison matrix, the fingerprints in the rows can be considered as attempted attacks on the fingerprints of the columns. Because we have only two samples of each finger, we cannot test for "goats", those consistently not matched to their own enrollment template. We can, however, test for "wolves" and "lambs". Because of the lack of score correlation between prints of an individual user, we have chosen to test for "wolves" and "lambs" at the single print level. A "wolf" row will have consistently higher scores across the columns of enrollment prints, not considering, of course, the genuinely matching enrollment image, while a "lamb" column will have higher scores across the rows. Again, we limited our comparisons to prints in communicating bins. Therefore, for each row we summed the scores across all columns that communicated

[2] The term "zero effort attack" means that the attack is passive and does not involve active efforts at impersonation.

with the row print and for each column we summed scores across the rows. Because the number of communicating comparisons will vary, these results must be normalized against the number of comparison scores used for each "wolf" row or "lamb" column. This produces the mean communicating impostor score.

If the comparison matrix were symmetric, each "wolf" row mean would be identical to the matching print's "lamb" column mean. The comparison matrix is not symmetric, however, for two reasons. Firstly, the prints represented in the columns are images acquired at a different time from the prints represented in the rows. Secondly, fingerprint comparison scores are not symmetric. The score of the comparison of print A to print B is not generally equal to score of the comparison of print B to print A. Therefore, we computed both the row and the column sums.

Using a one-way analysis of variance [11], we tested the null hypothesis that all the communicating scores in the matrix came from the same distribution against the opposing hypothesis that the distribution was row dependent. Combining results for right and left thumbs, 1420 thumbs were in about 336,000 communicating comparisons . The "F" statistic was calculated at 6.7, which is much larger than the critical value of nearly 1 for this number of "degrees of freedom". Thus, the alternate hypothesis was accepted. This shows that there are "wolves".

Then we repeated this test for column dependency in the thumb data, calculating the "F" statistic as 9.0, with 1437 columns in about 278,000 comparisons. We again accept the alternate hypothesis that the data is column dependent, showing that there are also "lamb" fingerprints.

Figure 6 shows a histogram of the mean row impostor thumb scores. Also graphed is the histogram of the mean column thumb scores. Because these distributions are nearly identical, they are not individually labeled. If all the means were nearly identical, Figure 6 would show a sharp spike. If there were strictly "sheep" and "wolves", there would be two spikes, one at a low and one at a high score value. Figure 6 shows both "lamb" and "wolf" distributions to be smoothly spread. This indicates that there are "sheep" and "wolves", and "sheep" and "lambs", but the boundary between them is not well defined.

Figure 7 shows the same study done on index fingerprints. Results are seen to be the same. Analysis of variance of the index finger rows gave an "F" statistic of 7.5. The "F" statistic for index finger columns was 10.1. With the 1420 relevant rows or columns and the 232,000 communicating comparisons, both of these "F" statistics are significant at all reasonable significance levels.

The existence of lambs and wolves calls into question the suitability of system false-match error rate equations [1] based on the assumption that all stored templates have the same probability of being falsely matched. Equations of the type

$$\mathrm{FMR}_{\mathrm{sys}} = 1 - (1 - \mathrm{FM})^{N} \tag{4}$$

where FMR_{sys} is the system false match rate, FM is the false match rate of a single comparison (assumed to be uniform) and N is the number of stored templates, should be more reasonably replaced with the form

$$FMR_{sys} = 1 - \prod_{i=1}^{N} (1 - FM_i) \qquad (5)$$

yielding higher estimates for the system false match rate, FMR_{sys}, if $FM_i \neq \text{constant}$.

FIGURE 6:

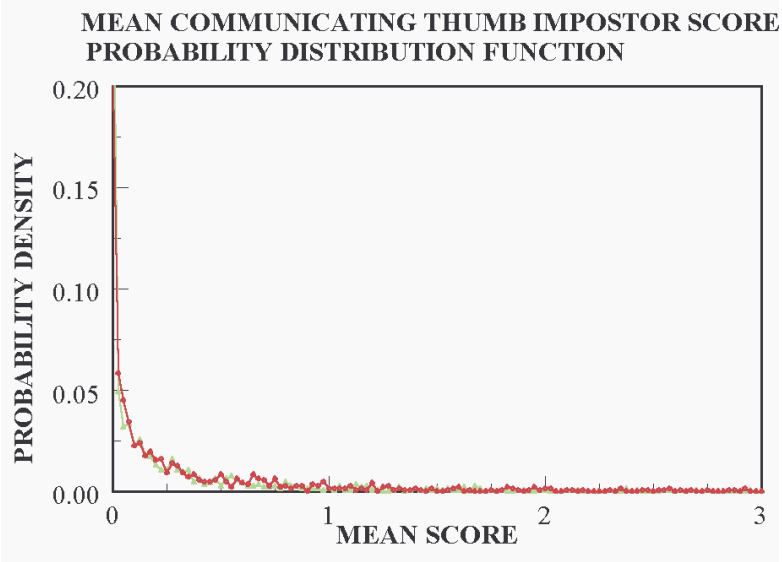
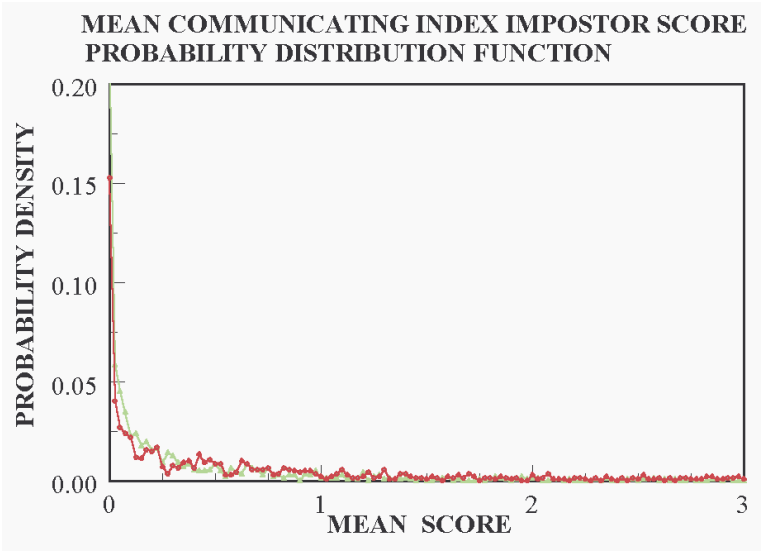


FIGURE 7:



IX. Conclusions

We can make the following conclusions:

1) ROCs developed from images in communicating bins show worse performance than those developed without consideration of the binning.
2) Thumbs have lower binning and comparison error rates, but index fingers have better penetration rate.
3) Both binning and comparison error rates increase as we move from thumb, through index to ring fingers.
4) Because of pattern correlations across individual users, penetration rates for multiple finger systems cannot be accurately estimated from single finger penetration rates.
5) Matching scores and binning errors are not correlated across the fingers in the general individual user.
6) "Wolves" and "lambs" exist, but there is a gradual transition between sheep and these populations.
7) The existence of population variability in error rates calls into question the validity of system false match rate equations based upon the assumption that error probabilities are consistent across the population.

X. References

[1] J.L. Wayman, "Error Rate Equations for the General Biometric System", IEEE Robotics and Automation Magazine, March 1999.
[2] J.L. Wayman, "A Scientific Approach to Evaluating Biometric Systems Using a Mathematical Methodology", Proc. CTST'97, pg. 477-492
[3] J.L. Wayman, "Benchmarking Large-Scale Biometric System: Issues and Feasibility", Proc. CTST Government'97, Sept. 1997
[4] J.L. Wayman, "The Science of Biometric Technologies: Testing, Classifying, Evaluating", Proc. CTST'97, pg. 385-394
[5] J.L. Wayman, "Testing and Evaluating Biometric Technologies: What the Customer Needs to Know", Proc. CTST'98, pg. 385-394
[6] J.L. Wayman, "A Generalized Biometric Identification System Model", Proc. of the IEEE Asilomar Conference on Signals, Systems, and Computers, Nov.,1997
[7] J. L. Wayman, "Technical Testing and Evaluation of Biometric Devices" in A. Jain, etal, eds. Biometrics: Information Security in a Networked Society, (Kluwer Academic Press, 1999)
[8] FBI numbers from where????
[9] W.H. Press, etal Numerical Recipes in C, 2nd ed, (Cambridge University Press, New York, 1992)
[10] G. Doddington, etal "Sheep, Goats, Lambs and Wolves: An Analysis of Individual Differences in Speaker Recognition Performance", ICSLP'98, Sidney, Australia, November 1998

[11] A.L.Edwards, Experimental Design in Psychological Research, 4th ed. (Holt, Reinhart, and Winston, New York, 1972)