

ROC Testing of Face Recognition Systems

W. A. Barrett

National Biometrics Test Center

San Jose State University
One Washington Square, Engr 491
San Jose, CA 95192

June, 1998

ROCplan2.doc, vs. 2

Theoretical Basis

An automated face recognition system (AFRS in what follows) is intended to either *identify* a candidate individual out of a database of known persons, or to *verify* that a candidate is who he/she claims to be.

In either case, the AFRS has some *database of enrolled* persons. The database will usually consist of a list of codes, each representing the reduction of a facial image. Each of the persons in the database is assumed by the AFRS to be correctly identified. It's also reasonable to assume that the same person will not appear in the database twice under different aliases. The enrollment process implies that some effort has been made to maintain the validity of the database. Several different images of the same person are permitted, and in some systems, *required*, but these images will be associated with the same identity tag.

In an *identification* process, a candidate image is presented to the AFRS, and the AFRS is expected to produce a list of near-matches from its database, and (preferably) yield some estimate of the closeness of matching of the candidate to each of the selected database images. In a general way, the AFRS is expected to *rank-order* the database members according to their closeness of match to the candidate, and (preferably) yield a closeness measure between the candidate and each of the nearby database members.

In a *verification* process, the person presents some identification, perhaps in the form of a machine-readable badge or card, which supplies a purported identity. The AFRS is expected to accept or reject the person, based on the closeness of matching of the candidate image's code and that found in the database.

Both these processes have these elements in common:

- One or more images of a set of identified individuals are entered in the database through an enrollment process. This amounts to positively identifying the individual in some independent way, capturing an image, then reducing the image to a characteristic *code* or *score* that will be kept in the database associated with the person's unique identity tag.
- The database can be expanded with additional enrollments at any time. The system may engage in a training or retraining process as its database is expanded.
- When a candidate presents him/herself for recognition, the system captures an image, and reduces the image to a *candidate code*.
- The candidate code will in general be compared for *closeness* with all or some of the database member codes so that a decision can be made through a application-specific *threshold level*. Note that identification is more difficult and time-consuming than verification in this respect, yet both have this element in common.

Closeness Measure

The *reduction of an image to a code* and the *measure process by which two codes are to be compared* is the essence of any AFRS.

For example, in Pentland's eigenface method [SIRO87], [PENT91], the code is an N-dimensional vector of real numbers, called the facial *eigenvector*. The whole vector expresses the given face image as a linear combination of *eigenface* images. This vector is obtained from a candidate face by a matrix process that amounts to a least-squares approximation of the candidate face to set of eigenfaces. The eigenface set is usually formed once and for all from a large and diverse set of faces, and can theoretically represent any human face now or in the future living on the planet

The simplest measure of closeness of a candidate vector C_i to a database vector D_i is an Euclidean distance, i.e. the sum of the squares of the differences $C_i - D_i$. The differences might also be weighted,

particularly so if the members of the vector have different physical units, or represent different kinds of biometric measures.

The closeness measure might also be a *Hamming distance*. This was chosen by the developer of the Iriscan system [IRISCAN], [DAUG94] as the most appropriate distance measure for that biometric system.

In any case, most biometric systems, and AFRS in particular, ultimately yield some linear closeness measure by which two candidate measures can be compared. This gives rise to the ROC curve.

The ROC Curve

A tutorial background for the *ROC*, or *Receiver Operating Characteristic*, can be found in several books, in particular, [EGAN75].

The ROC arises from two probability distributions of pair-wise candidate matching codes. One distribution is of all individuals matched against *themselves*, which we'll call the *authentics distribution*. The other is the distribution of all individuals matched against *others*, which we'll call the *imposter distribution*.

Ideally, the authentics distribution will show small code differences. When one image of an individual is matched against another image of the same individual, the code difference should be zero. However, a zero difference is almost never found; instead some distribution of small differences is found. The variations are the result of many factors. In the case of face recognition, the factors include differences in lighting, pose, expression, age, hair styling, glasses, makeup, and more.

Ideally, the imposter distribution will show large code differences. Here, we look at the closeness measure of one individual's image compared to that of another individual. Again, it's possible that the two measures may be relatively small, owing to a natural similarity of the two persons, or makeup, or other factors that tend to make two people resemble each other.

These distributions will resemble those of figure 1.

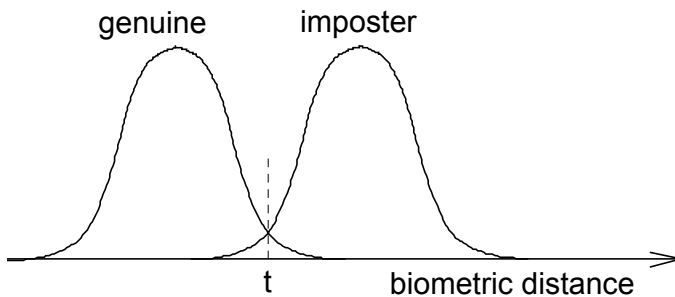


Figure 1. Genuine and imposter distributions

A small biometric distance should correspond to comparisons of a person's image against other images of the same person—the distribution of these distances comprises the *genuine* distribution shown in figure 1. A large biometric distance should correspond to comparisons of a person's image against images of other persons—the distribution of these distances comprises the *imposter* distribution shown in figure 1.

In general, the two distributions will always overlap to some extent. A high quality biometric system will have little or no overlap—this is the case for many fingerprinting systems and the iris scanning system. For face recognition, there will likely be an appreciable overlap.

Regardless of the degree of overlap, we can estimate the performance of a biometric system from these two distributions, for any selection criterion. In figure 1, the distance t provides an approximately equal rate of accepting imposters and of rejecting authentics. The relative rates can be controlled by choosing an appropriate distance threshold level.

If the threshold in figure 1 were set lower than t , the rate of rejecting authentics would fall to a very low level, but at the cost of accepting more imposters. This might be an acceptable threshold level for a bank credit card verification system, in which the credit card is the primary identification tool, the bank is anxious to not reject its customers, but would like some automated means of keeping cards suspected of being stolen. A potential card thief might know the person's PIN, but runs an additional risk of being caught through an image verification test. On the other hand, the level of stolen cards is sufficiently low

and the risk of loss in a transaction is low enough that the bank can afford some loss to card thieves who both know the PIN code and resemble the customer sufficiently well to pass an AFRS test.

By setting the threshold higher than t , more of the authentications would be rejected, but most imposters would be rejected. This might be an appropriate threshold for a prison access system, in which it's much more important that imposters not be accepted than authentications being rejected. Rejection of an authentication means that some other identification means must be employed when an authentication is rejected.

From these distributions, it's clear that the rate of accepting imposters and of accepting authentications can be plotted against each other, using the threshold level as a parameter. The result is an ROC curve, illustrated in figure 2. This concisely shows the tradeoff between these two rates in any application. A low imposter acceptance rate, say 0.1, comes at the expense of a small authentication acceptance rate (approx. 0.3 in figure 2). This would be a *conservative* strategy. For a high authentication acceptance rate, say 0.9, we must accept a high imposter acceptance rate, approx. 0.7. This would be a *liberal* strategy.

Now consider figure 3, taken from J. Daugman's patent [DAUG94] on iris-based recognition. The authentications data (light bars) was constructed from 1228 pairs of images taken of the same person, over a period of time. The imposter data (dark bars) was constructed from 2,064 unrelated iris images. Daugman found an excellent fit for both empirical data to binomial distributions. No experimental overlap was found at Hamming distance 0.32, but the binomial distribution suggests that at this optimal criterion, the odds of a false accept are 1 in 151,000 and of a false reject 1 in 128,000. Daugman properly points out that these odds are *theoretical* in nature, being based on an assumed binomial distribution of the data, and not on empirical data. He also reviews various physical and biological factors that supports the assumption of a binomial distribution, with a high statistical independence of the iris texture measurements.

Developing an ROC Curve

Unless an instrument can be fully characterized through some physical theory, which is usually never the case in biometrics, its ROC curve must be obtained empirically, with appropriate attention paid to statistical experimental design.

A biometric experiment requires a large pool of candidate persons, each of whom should be measured repeatedly by the instrument over some time span of interest. In the case of face recognition, evidence from Phillip's studies [PHIL94] strongly suggests that a year or two of age difference is sufficient to increase the authentication distance measure appreciably.

There must also be some independent objective determination of the person's identity that is at least an order of magnitude better than the system under study. This is especially important in dealing with a database of hundreds or thousands of ordinary people "off the street", so to speak. If even a small reward is offered to the individuals consenting to the measurement, a motive exists to falsify their identity in some way. We propose using an Iriscan system for this determination.

Enough individuals must be measured enough times to establish reasonable error bounds on the ROC. The error bounds can be estimated by standard statistical methods. These are particularly important in comparing one method with another, since the differences may be masked by statistical variation.

Note that *none* of the following experimental approaches are suitable for an ROC evaluation:

- *A database of faces in which the identities are highly suspect.* If someone just photographed lots of people and later tried to match them up by eye, an unknown *a priori* error level would be introduced, which could be larger than the instrumental errors. The same objection applies to images of people for which a high incentive to counterfeit identification exists, for example, those attempting to enter a country illegally, or those wanted by the police.
- *A database of faces consisting of one or two images each of each person.* For example, consider a file of driver's license photographs. Even if the identities were well established (which is not necessarily the case with driver's licenses, given their legal implications), such a database could at best provide an imposter distribution, not an authentications distribution.
- *A database of a few faces captured many times under different circumstances.* Here, an authentications distribution could be established, but not an imposter distribution. Such would be the case of an experiment done with a few people in an office setting.
- *A measurement device in which a threshold must be set a priori, and which only indicates Pass/Fail.* This is the typical situation with a commercial product designed for installation and application. Obtaining an ROC could be done, but only at the very great expense of noting the pass/fail statistics at

a variety of threshold levels and with many trials. Obtaining a code that can later be compared with others for distance is far more effective as an experimental strategy.

- *A measurement system which can only yield Pass/Fail through some hidden training process.* The problem here is again the large number of experiments necessary to characterize the instrument. Note that many simple neural network classifiers fall into this category. They require frequent retraining against a database, provide no distance measure of any consequence, and develop an internal weighting structure that may appear to be very effective, but which cannot be analyzed.

Our Testing Strategy

San Jose State University has an extremely diverse student and faculty, with an approximately equal mix of Caucasian, Asian, Chicano nationalities. The student age range is fairly broad, with the bulk between 18 and 22, but with an appreciable number of older working students.

- The test environment would be set up (to start with) in the student union, perhaps in the cafeteria. We'd collect data over the busy lunch hour and dinner hour on a regular basis. The idea is to get repeat "customers", of which there should be plenty, given that the same bunch of students are likely to eat there regularly, and many obviously have some time to kill.
- A small reward will be offered to induce students to go through the process, for example, a bag of M&Ms, a soda, a trinket or a candy bar. I think most students will be intrigued by our operation, and will want to try it for the fun of it. We'll probably also attract a few of the local street people, but there's no harm in that, providing they don't make a nuisance of themselves.
- We'll use Iriscan to positively identify each subject, and require that the subject either be accepted by Iriscan, or be enrolled. I'm assuming that the false accept/false reject rate of this system is much lower than any other positive identification device that we could use (for example, student ID cards, driver's license, specially printed card, etc.). We'll also use Iriscan to stop students from going through too often (to try to collect more prizes), although some repeat performances wouldn't hurt, and the operators will be able to control any abuses.
- As an additional check, we'll ask for some form of identification, i.e. driver's license number, and/or student ID card, manually checking the photograph against the individual. These will also be recorded and checked against our growing database as we pass students through our system.
- Once a student has passed the Iriscan test, i.e. is now positively identified, we use each of the vendor AFR products *as they are intended to be used in an application* on the student. The idea is that if the person is not previously enrolled, the system should not accept the face, while if previously enrolled, the system should accept the face. *In fact*, we don't care what the product claims. We want to (1) grab the face image for future reference, and also (2) grab the characteristic face code, if one is available. The face code and the positive identification of the Iriscan will make possible the generation of a reliable ROC curve for each product from the whole database of collected image codes that we will have in time.
- The operator will also record any unusual problems experienced by the student with any product, for example, a gross failure of the device to function at all (unlikely), difficulty with adjusting a mirror, etc. These problems are considered distinct from the device's claim of rejection or acceptance.
- Note that we will also be testing Iriscan in a small way since it may be the case that our name/code system is more prone to failure than Iriscan. Having the images and names as backup can help us sort out any obvious discrepancies between our hand-recorded records and the Iriscan claim. My reading of the Iriscan approach to identification suggests that this is at least two orders of magnitude more accurate than any face recognition system currently on the market. It is probably more accurate than our manual identity checking process, which is subject to human error and possible fraud.
- Other biometric instruments (hand scanning, fingerprinting, etc.) might be incorporated over time.
- Of course, each student will be told that his/her personal identity will be kept confidential and not used in any way beyond our statistical research interests, that they have certain rights as an experimental subject, etc. They will each sign a document that they agree to these terms. I have such a document approved by our provost.

What We Expect of an AFRS

The capture system will *preferably* consist of a single PC running NT, with a large disk and some kind of port switching arrangement so that it can function for *all* the vendor's systems as well as the Iriscan. Some software might be needed to guide the operator functions, and to pull together the files generated by each of the vendor's software. An alternative is described later.

Each AFRS should be capable of the following, beyond the obvious operations of segmentation, enrollment and recognition:

- Capture the face image as a *digital frame* in some format that we can download and convert. This may or may not be grayscale, depending on the system.
- Capture the time stamp of the image. This is particularly important if we must work with several different machines rather than a single one (see below).
- Capture a characteristic *face code* (supplied by the vendor software) for each. Associated with this would be a *measure function* (also supplied by the vendor) by which the engine decides on the "closeness" of one code relative to another. This step is vital to the development of an experimental ROC, as explained above.
- The face image and face code of a vendor system will be no different whether the system decides to "accept" or "reject" the person against its internal database. An exception to this rule is the Iriscan station, which we use for student identification. In it, we very much care about a rejection, which implies that the student has not been enrolled, and an acceptance, which means the student has been through our system before.

If the vendor system can provide these, then there's no issue of acceptance or rejection on any particular face capture. As our collection increases, so will our ability to construct a good approximation to the ROC for the instrument, and to judge the reliability of the ROC.

Alternatives to a Single Machine

We can manage with several machines if the students are consistently run through the test series in the same order, i.e. like an assembly line. The captures will of course be time-stamped.

Suppose the Iriscan is always first, system A second, system B third, etc. Then at the end of a day's operation, we pull out all the transactions for the day from the database, sorting them by capture time (if not already so sorted). Each station should then have the same number of transactions, the times for them should be ordered across stations, and the correspondence between an Iriscan identification and any one station's capture is thereby assured.

Confidentiality and Conflict of Interest

We recognize that face recognition has become a competitive business, with many vendors offering a variety of products. Most of these systems contain proprietary algorithms or methods which the vendor would prefer be kept secret. Often, a patent application is in process, and that would be jeopardized by a prior publication of the process, even if the publication were by the patent applicant.

Our policy on proprietary information is as follows:

- We do not necessarily require full disclosure of the product's engineering, for example, source code. If the vendor is willing to provide appropriate object modules or DLL software that meets our needs of capturing a raw image, capturing a characteristic image code, and comparing two codes through a distance measurement, then we have no need for further details.
- Should we require proprietary information, or obtain such information in the course of our work, those involved agree to sign nondisclosure agreements as provided by the vendor. We expect the vendor to identify those papers, ideas, software or parts that fall under the scope of a nondisclosure agreement.
- With regard to publication of the results, we invite suggestions from the vendors in this regard. Under no circumstances will we publish results of a particular product or series of products without an appropriate response from the vendor. Of course, our sponsors are entitled to detailed results of our experiments, but we also expect them also to not publish the results without appropriate vendor response.
- As a university agency, with government sponsorship, we have no pecuniary interest in any of the biometric products under study. Our research people have agreed to accept no payments or payments-

in-kind from vendors. For example, no center staff person involved in testing a product will accept consulting fees from that product's vendor.

- To the extent that our grant funding permits, all experimental expenses, including equipment, facility rental, and the payroll of the center staff, will be borne from our client grants, and not from vendor contributions. We make one exception to this rule—we will gratefully accept a donation of a vendor system and/or special software work required to adapt the vendor equipment to our experimental needs.
- We invite comments and criticisms of our experimental approach, calculations and findings with the affected vendors.
- Each vendor has the right to withdraw his equipment from consideration an experimental study series, in which case, we shall destroy all experimental data related to that equipment. We will also have to assume that the product is not ready for the marketplace.

Appendix 1 -- The d' Measure

A single measure has been proposed [DAUG96] that can roughly characterize a biometric system, called the d' measure.

Most biometric systems operate along the following principles:

- The device accepts some reading or *measurement* of the human subject. The reading may be a fingerprint image, a facial image, or a hand contour. The measurement is typically a video image frame in greyscale and may be several hundred kilobytes in size.
- The device applies a *recognition algorithm* to the measurement, which usually includes noise filtering, extraction of significant features, etc. The algorithm is expected to yield a *biometric code* that contains the significant attributes of the measurement. The trick in designing a suitable algorithm is to find a uniform way of reducing the large size of the measurement to a relatively small code. A goal in many biometric systems is to develop a code that can be fit on a "smart credit card", which may have space for only 64 bytes of code.
- One biometric code is compared to another with a second algorithm, using what is called a *distance metric*. The distance metric may be especially designed for the particular biometric system. Typical metrics include the Euclidean distance and the Hamming Distance. An Euclidean distance might apply to a code that can be interpreted as vector in an n -dimensional space. A Hamming distance might apply to a code in which each of the bits are statistically independent.

The distance between two measurements $M1$ and $M2$ can be summarized by examining the statistics of two possible cases S and D , as follows:

case S: $M1$ and $M2$ are of the same person, taken at different times under possibly different circumstances.

case D: $M1$ and $M2$ are of different persons.

The intent is that the distance $d(S)$ should be small, while the distance $d(D)$ should be large. In fact, with a large population of measurement samples, these two distances each exhibit a distribution, which can be described by a *mean* and a *standard deviation*. Figure 1 is an example of a pair of such distributions. The *genuine* distribution is the distribution of $d(S)$, while the *imposter* distribution is that of $d(D)$.

The overall *quality* of a biometric system can be judged by a single measure, d' , as discussed by Daugman and Williams [DAUG96].

d' is the ratio of the distance between the means of the imposter and genuine distributions divided by the conjoint measure of their standard deviations, given in Equation 1:

Equation 1

$$d' = \frac{\|M_{imposter} - M_{genuine}\|}{\sqrt{(SD_{imposter}^2 + SD_{genuine}^2)}/2}$$

This measures how much the genuine and imposter distributions overlap, assuming that each one approximates a normal distribution. (They are sometimes not normally distributed). A large d' implies very little overlap—the distribution functions are widely separated compared to their width. A small d' implies considerable overlap.

The extent of the overlap can be estimated from a table of the accumulated normal distribution function. d' is just the coordinate of this function. A few values are given in the table below:

d'	probability of false rejection (and false acceptance)
0	0.5
1	0.15865526
2	0.022750062
3	0.001349967
4	3.1686E-05
5	2.87105E-07
6	9.90122E-10
7	1.28808E-12
8	6.66134E-16

As d' increases, the probability of a false rejection decreases exponentially.

For comparison purposes, the d' of an *automated* fingerprint system developed at the University of Michigan is approximately 2.1 [JAIN97]. This low value may come as a surprise to those who were brought up with the idea that "no two individuals have the same fingerprints", which was a favorite motto of J. Edgar Hoover.

However, note that this result applies to a *single* print, not a set taken with all ten fingers. Since the prints for each finger appear to be statistically independent, the probability of two individuals possessing the identical print characteristics in all ten fingers is approximately $(0.02)^{10} = 10^{-17}$. The population of the earth is approximately 10^{10} persons, so one could argue that 10 million generations would have to pass before two individuals would have the same *set* of fingerprints. Nevertheless, out of all the individuals on the earth, there will be a large number with a matching print from one finger.

Also note that this result is for an *automated* fingerprint system, not that for a skilled individual trained in fingerprint comparison. The statistics of human comparisons are not known, but fingerprint specialists are generally believed to be much better at distinguishing prints than machine methods, at present. We speculate that a skilled forensic fingerprint specialist might achieve a discrimination level of 10^{-5} .

The use of a *single* fingerprint as a way of convicting a person of a crime, given no other supporting evidence, is clearly questionable. This issue will surely arise as the use of AFIS on a large scale to track down wanted persons through fingerprint records increases.

References

[DAUG94] J. Daugman, *Biometric personal Identification System Based on Iris Analysis*, U. S. Patent 5,291,560, issued Mar, 1, 1994.

[DAUG96] J. G. Daugman and G. O. Williams, *A proposed standard for biometric decidability*, in *Proc. CardTech/SecureTech Conf.*, Atlanta, GA, 1996, pp. 223-234

[EGAN75] Egan, James P. *Signal detection theory and ROC-analysis*, New York : Academic Press, 1975.

[IRISCAN] IriScan, 133-Q Gaither Drive, Mt. Laurel, NJ 08054-1701, 1-800-333-6777.

[JAIN97] A. K. Jain, L. Hong, S. Pankanti and R. Bolle, *An Identity-Authentication System Using Fingerprints*, *Proc. IEEE*, vol. 85, No. 9, Sept. 1997, pp 1365-1388.

[PENT91] Pentland, A, and Turk, M. *Eigenfaces for Recognition*, *Journal of Cognitive Neuroscience*, vol 3, No. 1, pp 71-86.

[PHIL94] Phillips, P. Johnathon, Rauss, P.J., and Der, S. Z. *FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results*, Army Research Laboratory ARL-TR-995, October 1996.

[SIRO87] Sirovich, L, and Kirby, M. *Low-dimensional procedure for the characterization of human faces*, *J. Opt. Soc. Am A*, Vol. 4, No. 3, March 1987, pp 519-524.

Additional material on face recognition may be found in the excellent web site

Figure 2

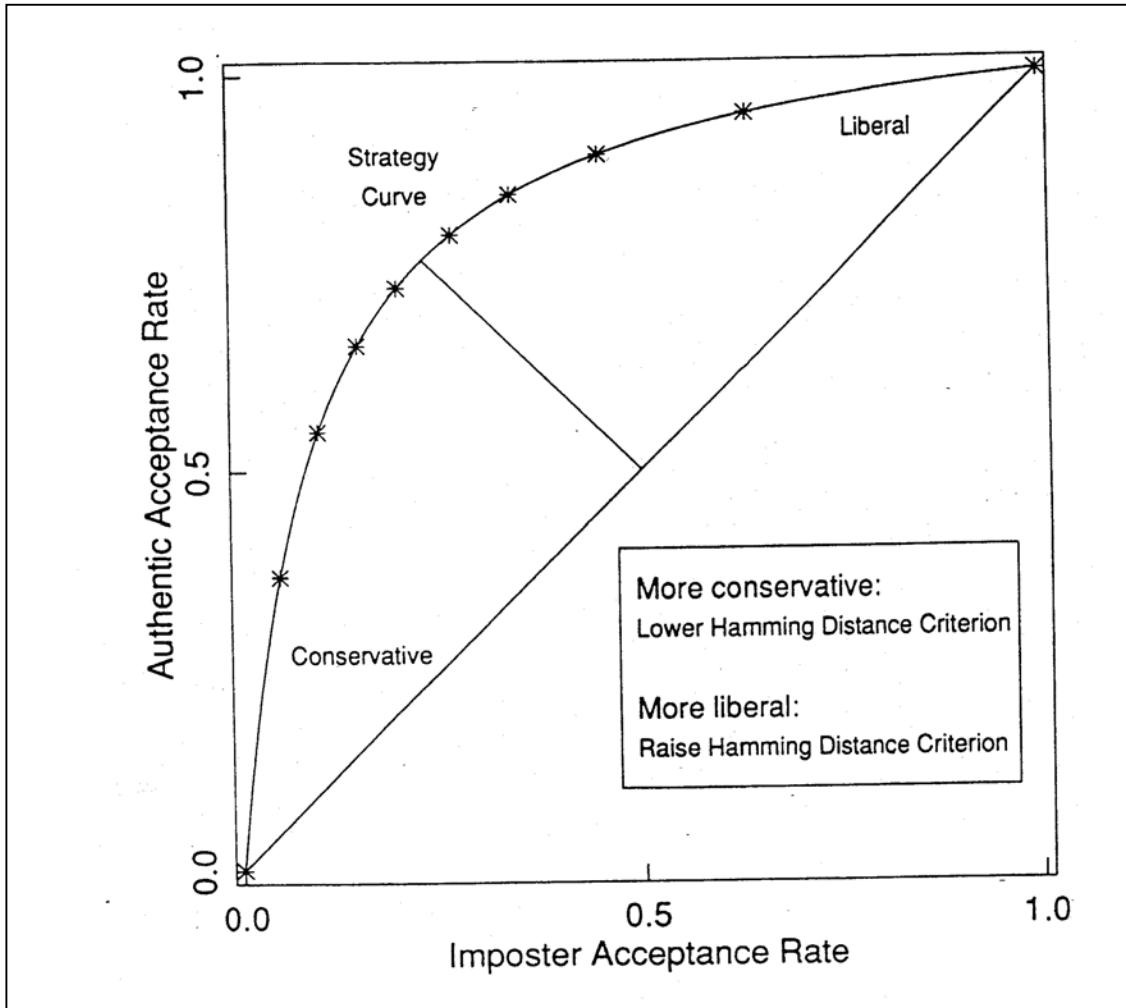


Figure 3

