Proceedings of the 1984
Winter Simulation Conference
S. Sheppard, U. Pooch, D. Pegden (eds.)

OPERATIONAL EVALUATION MODELING OF AUTOMATIC SPEAKER VERIFICATION SYSTEMS

David E. Crabbs
Interstate Voice Products
1849 West Sequoia Avenue
Orange, CA 92668

John R. Clymer
Department of Electrical Engineering
California State University, Fullerton
Fullerton, CA 92634

ABSTRACT

This study uses operational evaluation techniques to model a system which processes human speech to verify the identity of persons seeking access to a facility or resource. The system consists of hardware and software for accepting analog speech; extracting time, frequency, and amplitude characteristics; producing compact digital templates containing the features for speaker identification; and cross-referencing the templates with reference patterns to establish the degree of similarity between an utterence and a set of utterences for the person whose identity is being claimed. A decision algorithm is implemented to determine whether the speaker is valid or an imposter based on the degree of similarity observed.

A conceptual model has been tested and used to simulate variations in system attributes in order to optimize system performance. Performance is evaluated in terms of the number of imposters who can defeat the system, and the number of rejected valid speakers.

BACKGROUND

The purpose of this study is to model and study the behavior of a system which processes human speech inputs to verify the identity of persons seeking access to a priviledged remote computer database over a telephone modem link. The system has been prototyped from preexisting speech recognition equipment originally intended to recognize words rather than individuals. A mathematical model of this system has been constructed, tested and verified, and then used to vary system attributes in an effort to optimize performance in terms of the number of rejected valid users, or Type I errors, and the number of accepted imposters, or Type II errors.

An extensive series of physical tests of the system has been performed, with a signiticant population of both male and female speakers acting both as valid speakers and as imposters deliberately misrepresenting themselves in an effort to defeat the system. It is desired to reduce both of these error rates to a minimum by modifying system parameters to improve performance. Since physical data collection is a costly and time consuming process, a simulation model is seen as an appropriate design tool.

Two FORTRAN computer models have been written which will perform the simulation described above. One uses a discrete event simulation approach, and the other uses a flowgraph reduction technique.

OPERATIONAL SCENARIO

The speaker verification system is to be used to screen people seeking telephone access to a remote computer database. The user is required to claim his identity by keying in ID information from a touch tone telephone keypad. The system then verifies this identity claim by means of a voice verification process which involves requiring the user to speak a series of prompted words into the telephone. The voice signal passes through a computer interface where it is digitized and encoded, and the resulting voice template is passed along to the computer for pattern analysis. The template is evaluated against a catalog of reference patterns for the claimed identity. A score is assigned based on the degree of similarity. If the score falls at or below a certain rejection threshold (RTHL), the template is considered not similar enough, and is labeled a miss. Likewise, if the score is above the threshold, it is considered a hit.

In order to gain access the speaker is required to speak a randomly prompted sequence of words, and accumulate a certain number of hits before acquiring a certain number of misses. If he gets too many misses, he is 'conditionally rejected'. When this occurs both the total accumulated number of hits and the total accumulated number of misses are decremented by a fixed amount, and the criteria for positive identification are made more stringent. If the speaker still acquires the maximum number of misses before getting the required number of hits, he is identified as an imposter. Otherwise, he is considered to be a valid user and is accordingly granted access.

This scenerio requires that there be reference patterns available for all authorized speakers. These reference patterns are gathered during an enrollment procedure which is similar to the access approval procedure, except that reference voice inputs must be gathered for the entire list of words or phrases that may be later prompted, and each word or phrase is subject to a multiple number of "training passes". The requirement for multiple samples of each word guarantees that normal fluctuations in the enrollee's speech will be acounted for in subsequent access attempts. The reference patterns are formed by extracting only the information consistant for all of the passes.

System performance was tested under many conditions using a physical prototype. Many factors were considered, such as the choice of the test vocabulary, the associated similiarity measures for proper verification, and the basic decision algorithm. The prototype system had the required number of hits set to 5, the allowed number of misses set to 3, and, for the 'conditional reject' case, the number of hits and number of misses were both decremented by 2 while the word acceptence threshold score was simultaneously incremented by 2. Word scores were assigned between 0 and 128, with 128 being a perfect match. A given speaker had to accumulate at least 5 hits to get in, and was allowed 5 misses in the worst case before being rejected.

STUDY OBJECTIVES

Extensive testing was performed on the system prototype with a population of speakers acting as valid speakers and as imposters in order to gather data on the true reject and imposter accept error rates. The collected word score statistics were used in simulation studies to produce sensitivity analyses of the access attempt trial parameters. Among the parameters to be studied were the vocabulary, the required number of hits and allowed number of misses for conclusion of a trial, the conditional reject parameters, and the word rejection thresholds. The goal was to determine a set of parameters which would minimize both the number of valid speakers rejected and the number of imposters accepted. The error mix sought was one in which the errors were equally apportioned between the two error types.

A second objective was to evaluate the relative usefulness of the two models being implemented in terms of their prediction accuracy and in terms of their respective assumptions and limitations. It was also desired to determine the effectiveness of these models when benchmarked against more straightforward statistical approaches.

MODEL DEVELOPMENT

The system operation is represented visually with the directed graph model shown in figures 1 and 2. This visual representation lends itself to model implementation using discrete event simulation techniques which use random number generators to predict probable event sequences. Flowgraph reduction methods can also be implemented.

The various states represented in the figures are defined as follows:

B	- Begin State
SPK	- Selection of Next Speaker
T/I	- True/Imposter Status Determination
TTR	- In-Trial State for True User
ITR	- In-Trial State for Imposter
TAC	- True Speaker Accepted State
TRJ	- True Speaker Rejected State (Type I Error)
IRJ	- Imposter Rejected State
IAC	- Imposter Accepted State (Type II Error)
TST	- End of Trial Set for Selected Speaker
E	- End State

Internal Trial States:

SW	- Word Prompt State
SCR	- Word Scored State
WM	- Word Missed
WH	- Word Hit
CR	- Conditional Reject State
RJ	- Unconditional Reject State
AC	- Trial Successful State
TOV	- Trial Completed State

The input for the two simulation models was a database of true speaker and imposter word scores obtained from physical experiments with a prototype version of the system and a group of adult male and female speakers. A baseline performance was thus obtained during prototype testing. Since the object of studying the behavior of the speaker verification system was to predict its error rate performance, the analysis needs to be concerned with extreme events in the various considered statistical distributions. On account of this, the prediction of error rate performance requires 1) a large amount of data in order to accurately characterize extreme events, and 2) that any assumption regarding the form of any distribution being considered be avoided wherever possible, since the distribution tails will tend to exaggerate these assumptions and perhaps invalidate subsequent predictions.

DISCRETE EVENT SIMULATION MODEL

The discrete event simulation was implemented as a virtually direct translation into FORTRAN code of the directed graph model given in figures 1 and 2. Each event and state represented in the directed graph is also represented in the computer program, and where necessary, branches in the directed graph are determined in the program by means of an integer random number generator. The simulation program performs a specified number of trials for each of the speakers in the database, with each speaker acting both as a true speaker and as an imposter. Each speaker is handled in the order in which he appears in the database. The speaker sequence is fixed since it has no bearing on the accumulated statistics at the end of the simulation. The 50 speakers being used have been judged to be a representative random sampling of adult male and female speakers.

The program begins each utterence by randomly selecting one of the available words, and then it randomly selects a score for that word from the speaker's database of true or imposter scores generated by the prototype. The score is evaluated to determine if it qualifies for acceptance, and then the trial hit and miss counters are evaluated to determine if the trial has arrived at the

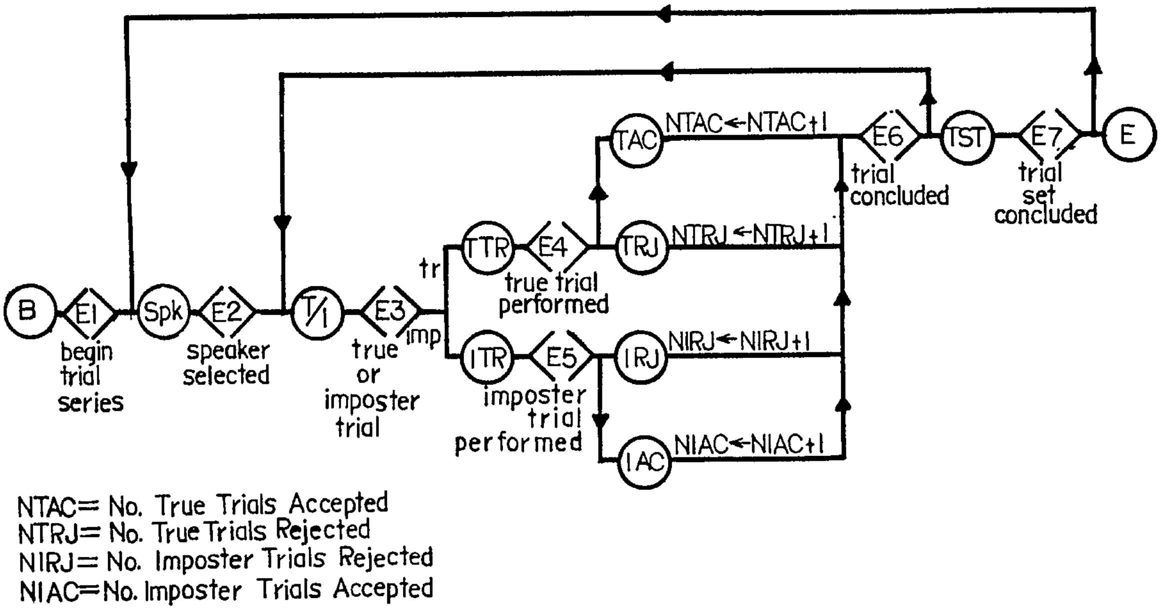

Figure 1: Directed Graph Model

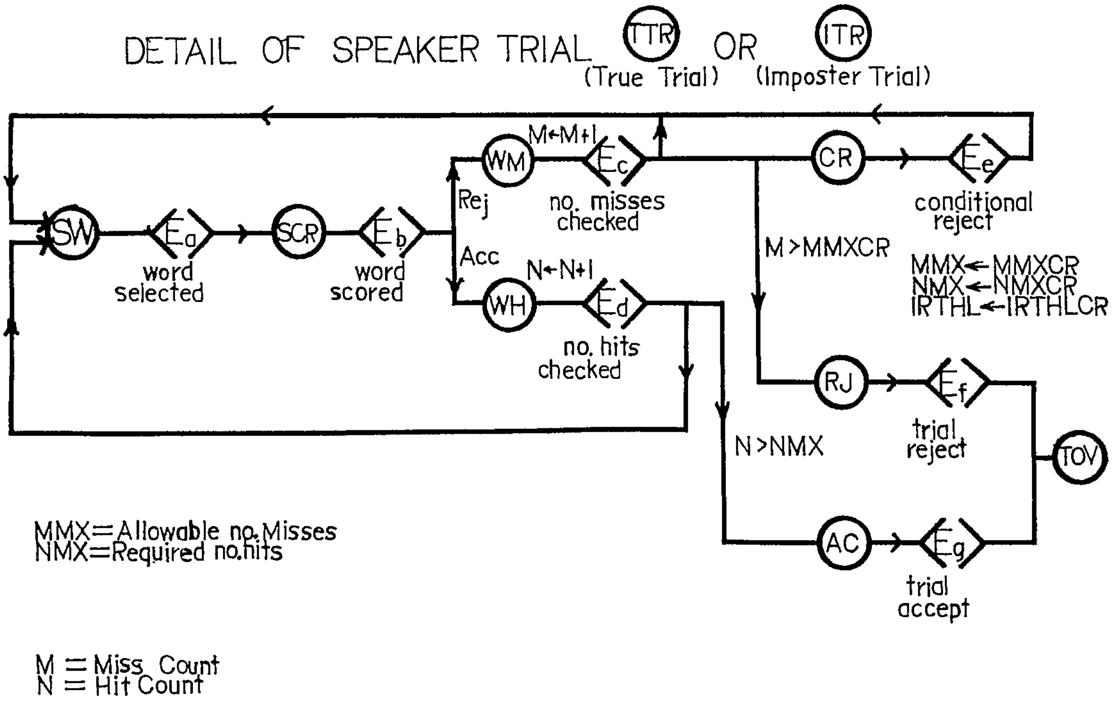

Figure 2: Detail of Directed Graph Model

trial acceptance state, the trial conditional reject state, or the unconditional reject state. In the conditional reject state, the trial parameters are changed to the conditional reject parameters, and the trial resumes. If one of the end states, either acceptance or rejection, is reached, the appropriate state variable counter is incremented. It should be noted that the method of scoring does not involve selection of a score from a theoretical probality distribution. Rather, only scores that were actually obtained during testing are used. Thus, the approach uses histograms of real data instead of a fitted distribution.

This model used a discrete random number generator to select the sequence of events given the word scores available in the database. This allowed the results to be drawn from an essentially infinite universe of possible event sequences, thereby allowing the assessment of the performance variance and convergence properties of the physical data. The only modelling assumption introduced was that the word scores in a given access attempt were not highly correlated with time (i.e., high or low scores did not cluster together within access attempts).

REDUCTION OF TRIALS BY FLOW GRAPH TECHNIQUE

It is possible to obtain a closed form relationship between the probability of a single utterence being accepted and the probability of a trial being successful, if the sequence of utterences is assumed to be a Markov chain. It must be assumed that each word prompt is independent of all preceeding ones (i.e., the word chosen is not a function of previous selections), and the scores obtained are not correlated with location in the trial sequence (i.e., the trial outcome is unaffected if the speaker perceives that he is likely or unlikely to be accepted).

In order to obtain a closed form, it is necessary to expand the directed flow graph in figure 3.

DIRECTED FLOW GRAPH FOR TRIAL
WITH (NxM) = (5x3) AND (Ncr x Mcr) = (7x5)

```
      P       P       P       P       P
[0,0]---[1,0]---[2,0]---[3,0]---[4,0]---[5,0]
  | 1     | 1     | 1     | 1     | 1  SUCCESS
  Q       Q       Q       Q       Q
  |   P   |   P   |   P   |   P   |   P
[0,1]---[1,1]---[2,1]---[3,1]---[4,1]---[5,1]
  | 1     | 2     | 3     | 4     | 5  SUCCESS
  Q       Q       Q       Q       Q
  |   P   |   P   |   P   |   P   |   P
[0,2]---[1,2]---[2,2]---[3,2]---[4,2]---[5,2]
  | 1     | 3     | 6     | 10    | 15 SUCCESS
  Q       Q       Q       Q       Q
  |  Pcr  |  Pcr  |  Pcr  |  Pcr  |  Pcr    Pcr     Pcr
[0,3]---[1,3]---[2,3]---[3,3]---[4,3]---[5,3]---[6,3]---[7,3]
  | 1     | 4     | 10    | 20    | 35    | 35    | 35 SUCCESS
 Qcr     Qcr     Qcr     Qcr     Qcr     Qcr     Qcr
  |  Pcr  |  Pcr  |  Pcr  |  Pcr  |  Pcr  |  Pcr  |  Pcr
[0,4]---[1,4]---[2,4]---[3,4]---[4,4]---[5,4]---[6,4]---[7,4]
  | 1     | 5     | 15    | 35    | 70    |105    |140 SUCCESS
 Qcr     Qcr     Qcr     Qcr     Qcr     Qcr     Qcr
  |       |       |       |       |       |       |
[0,5]   [1,5]   [2,5]   [3,5]   [4,5]   [5,5]   [6,5]
FAILURE FAILURE FAILURE FAILURE FAILURE FAILURE FAILURE
```

where P = probability of word acceptance
      Q = 1.-P
     Pcr= probability of word acceptance after conditional reject
     Qcr= 1.-Pcr

Begin State = [0,0]
Success End States = [5,0], [5,1], [5,2], [7,3], [7,4]
Failure End States = [0,5], [1,5], [2,5], [3,5], [4,5],
                     [5,5], [6,5]
Begin Conditional Reject States = [0,3], [1,3], [2,3],
                                  [3,3], [4,3]

```
               |
KEY:    ---[A,B]---
               | n        A = Number of Hits
                          B = Number of Misses
                          n = Number of Paths Leading to [A,B]
```

Figure 3: Directed Flow Graph

The probability of being in any state is equal to the probability of having been in the state to the left times the path transmittance (P or Pcr) plus the probability of having been in the state above times the path transmittance (Q or Qcr).

For the (NxM)=(5x3) case shown, the probability of trial success is equal to the sum of the probabilities of the states [5,0], [5,1], [5,2], [7,3], and [7,4]. The probabilities of the first three states can be obtained by multiplying the number of paths to the state by the appropriate number of P's and Q's.

$$P([5,0]+[5,1]+[5,2])=P^{5}*(1.+5Q+15Q^{2})$$

The probability of state [7,3] is

$$P([7,3])=Q^{3}*(Pcr^{7}+3*P*Pcr^{6}+6*P^{2}*Pcr^{5}+10*P^{3}*Pcr^{4}+15*P^{4}*Pcr^{3})$$

and,

$$P([7,4])=Q^{3}*Qcr*(7*Pcr^{7}+18*P*Pcr^{6}+30*P^{2}*Pcr^{5}+40*P^{3}*Pcr^{4}+45*P^{4}*Pcr^{3})$$

The probability of trial success is

$$Psucc=P([5,0]+[5,1]+[5,2])+P([7,3]+[7,4])$$

Similar methods apply for other [N*M] and conditional reject configurations. A program was written to obtain simulation results from the utterence success probabilities using the method outlined below. The utterence success probabilities were obtained by histogramming the scores in the database. This method required the assumption that the probabitility densities for true speaker word scores and imposter word scores were each homogeneous across the individual words in the vocabulary, in addition to the Markovian requirement of the other method above. Note also that no variance or convergence information was directly available.

PROTOTYPE TEST RESULTS

Physical data collection consisted of tests with a population of 50 adult speakers, with 25 male and 25 female. All of the templates were saved, and the voice input was taken over a telephone line. Post experimental analysis was performed to obtain error rates and optimum RTHL's for each word in the vocabulary. The vocabulary included twenty five words which were selected to give an indication of performance as a result of various hypothesized word characteristics, such as nasalization, numbers of syllables, and selections of vowels represented. Templates were collected for two types of spectrum encoding, a spectral slope coding and a binary sonogram. During subsequent simulation studies the templates were rescored using various techniques to try to enhance the speaker-discriminating information contained in the templates and reference patterns. The effectiveness of the binary sonogram was studied by scoring the templates both with and without (ie, using only slope encoding) it. The results are shown for these cases in Table 1.

TABLE 1

BEST RTHL'S FOR PROTOTYPE SYSTEM DATA SETS

		NO SONOGRAMS			WITH SONOGRAMS		
WORD	WORD	BEST	P(7x5)		BEST	P(7x5)	
#		RTHL	%	RANK	RTHL	%	RANK
1	EIGHT	112	2.57	10	100	4.08	21
2	NINE	111	0.79	2	94	0.65	8
3	FOUR	116	2.56	9	110	1.98	14
4	ZERO	113	1.27	3	102	0.56	3
5	SIX	108	5.47	18	91	3.31	19
6	MANUAL	108	1.91	6	92	0.59	5
7	POINT	109	6.12	24	94	7.33	24
8	MEGA	112	2.83	11	100	0.57	4
9	HUNDRED	111	4.37	16	98	0.61	7
10	TWO	110	1.80	5	96	0.59	6
11	HIGH	116	4.60	17	106	1.03	13
12	HAMMER	111	1.57	4	94	0.06	1
13	CALIFORNIA	111	5.66	19	96	4.53	22
14	ALABAMA	113	0.78	1	100	0.39	2
15	NUMBER	113	5.73	20	101	2.62	16
16	NOVEMBER	115	2.97	12	100	0.66	9
17	MANY	108	2.56	8	93	0.80	12
18	ZEBRA	112	5.90	21	100	2.30	15
19	XYLOPHONE	114	3.43	15	104	0.68	11
20	TOMATO	110	3.35	14	94	0.67	10
21	INCLUDE	109	13.18	25	97	7.49	25
22	HUMAN	109	3.04	13	95	2.82	18
23	INFORMATION	108	5.99	23	90	4.91	23
24	PEPPERMINT	111	5.94	22	96	3.69	20
25	COLONIAL	112	2.53	7	100	2.69	17
MEANS		111.3	3.877		97.7	2.224	

The test results indicated a number of factors significantly affected system performance. As anticipated, system performance was strongly influenced by word and word specific RTHL selection. The vocabulary used included a number of words that yielded very little discrimination. However, the range of word error rates obtained across the vocabulary indicated that very good performance is possible given the right choice of words. For the speaker population tested, projected trial error rates were calculated for each word at the optimum threshold based on a binomial expansion of a 7 by 5 trial decision matrix.

The vocabulary was rank ordered in terms of its estimated error rate performance based upon the physical tests. The rank ordering was done by histogramming the word scores for all speakers, and finding the RTHL which would have produced the smallest degree of overlap between the true speaker distribution and the imposter distribution. The estimated error rate for this RTHL for each word was determined on the assumption that the vocabulary consisted of only the chosen word, and that the decision matrix was 5 misses by 7 hits with no conditional reject allowance. Note that is assumption relates the probability of trial success to the probability of word score above RTHL by a simple binomial expansion. Specifically,

$$P([7,5])=P^{7}*(1.+7Q+28Q^{2}+84Q^{3}+210Q^{4}).$$

The range of word error rate statistics indicated in the tables suggests that removal of the worst half of the vocabulary would yield substantial improvement. Access error rates of less than 3% should be attained in this manner for the combined frequency coding and binary sonogram scoring technique after editing out noise corrupted data.

BASELINE SIMULATION RESULTS

The baseline case was defined to be the conditions under which the physical tests were carried out, except that the optimum RTHL's were used instead of the nominal thresholds, and the decision matrix was expanded to (7x5) with no conditional reject (the reason for this is given in the section below on decision matrix sensitivity). The nominal thresholds were discarded once the optimum thresholds were obtained from the tests. The baseline conditions, then, were as follows:

1) all 25 words of the test vocabulary were used with their optimum RTHL's
2) all 50 test speakers were used
3) scoring was based upon the slope-encoding only
4) the decision matrix was 5 misses by 7 hits, with no allowance for conditional rejects.

The purpose of these runs was two-fold. First, it was necessary to find the degree of correspondence between each of the two models and the physical tests. Second, it was necessary to determine the degree of convergence that could be provided by the discrete event simulation model, and, in addition, the amount of variance in the performance error rates obtained during the physical data collection.

A series of runs were performed for the discrete event simulation. This series consisted of a set of runs with the trial multiplication factor set to one (i.e., each run corresponded to the equivalent of one complete set of prototype tests), and another set with the trial multiplication factor varied over a range. The purpose of the first set of runs was to determine the amount of variance in the performance results predicted by a run involving the same number of events that were recorded during the physical tests. The second set was intended to provide an indication of the convergence properties of the simulation. The results are tabulated below.

	True Rejects (%)	Imposter Accepts (%)
Physical Tests:	13.2	11.4
Flow Graph Model:	10.24	9.45
Discrete Event Simulation:		
Variance Run # 1:	10.70	10.29
Variance Run # 2:	10.14	9.14
Variance Run # 3:	9.30	8.86
Variance Run # 4:	10.14	9.71
Variance Run # 5:	12.39	11.14
Variance Run # 6:	10.42	8.86
Variance Run # 7:	11.27	8.57
Variance Run # 8:	12.96	9.71
Variance Run # 9:	10.42	9.14
Variance Run # 10:	10.14	11.71
Average Variance Run:	10.79	9.71
Standard Deviation:	1.058	1.021
Convergence Runs:		
(TMF=10):	10.51	9.46
(TMF=30):	10.48	9.46

From the tabulated results it can be seen that the flow graph model and the discrete event simulation give almost identical results. This is a useful observation since the discrete event simulation is very time consuming when the trial multiplication factor is much greater than ten. However, it should also be noted that the two models differ from the physical test results by a statistically significant amount. Both models underpredict both error rates by an amount that is within the 99% confidence band (plus or minus three standard deviations), but is outside of the 95% (two standard deviations) confidence region. This difference can not be ignored, but is not too large to permit sensitivity analyses using the models.

The reason for the difference between the tests and the models must lie in the assumptions common to both, since the two models agree so well. The major common assumption is that there is no correlation process occurring between trials, and this assumption is hypothesized to be the culprit. Clearly, if a group of low (or high) word scores are found clustered together within a single trial, the outcome of that trial is much more likely to be a reject decision (or, in the converse case, an accept). Neither model considers this possibility. Furthermore, this correlation process has in fact been observed under real conditions.

Several reasons may exist for such a correlation, but two are of special note. The first has to do with the nature of noise on the telephone line. This is clearly a condition that is established at the outset of the telephone call, and will remain in effect for the duration. If the line is disconnected and a new call is put in, then it is very reasonable to expect a different noise environment for the next access attempt.

The second reason is in the nature of the human voice. The voice will exhibit variations in pitch and formant structure for the same word if it is spoken by the same speaker on different days. This is obvious if the speaker has a cold, or is anxious or upset. Clearly, emotional state and health are variables that are sustained across a verification attempt, but not from day to day.

A correlation process based on either of the above phenomena would have the effect of increasing both true reject and imposter accept error types for an RTHL balanced vocabulary (balanced on the assumption that there is no correlation). This is in agreement with results in the tabulation.

Regarding the convergence runs, it can be seen that the average variance run error rates approach the TMF=10 convergence run quite closely. This is to be expected since the 10 trial average is essentially another TMF=10 convergence run. It is also to be expected that the ten trial case will be reasonably converged, since the variance should be approximately inversely proportional to the number of trials. Thus, if the standard deviation for a single run is about 1.0, then the standard deviation for ten trials should be about 0.3, and about 0.2 for 30 trials.

SENSITIVITY RESULTS

The first set of sensitivity runs was aimed at determining the effect of the vocabulary. Physical test results suggested a strong relationship between system performance and the selected vocabulary. This also makes intuitive sense since one would expect some spoken sounds to convey a good deal more speaker dependent information than others. Another reason why a strong relationship should be expected is that the conversion of the analog input to a digital template is imperfect at preserving the information in the original signal, and it is reasonable to expect that the degree of information loss will be highly dependent upon the type of information in the original speech. It would be difficult to sort out these two effects as they relate to speaker verification, but it should at least be possible to empirically select a set of verification words that perform generally well.

A sensitivity run was performed using the flow graph model with the best 15 words (RTHLs optimal, slope encoding only) for all speakers from Table 1. In addition, a run was made using only those of the best 15 which were also among the best 15 for male speakers alone and female speakers alone. Consequently, this run used only nine of the original 25 words. The results are shown below:

		True Rejects (%)	Imposter Accepts (%)
A	Baseline Run:	10.54	9.45
B	Best 15:	8.95	7.75
C	Best 15 M&F:	8.72	9.01

Both of these results are disappointing in terms of the results suggested by the physical data. The top 15 words in Table 1 all had projected error rates, using the method described above, of under 4% for both error types. This is not only disappointing, but also apparently inconsistent.

The resolution of this inconsistency must lay with the modelling assumptions. The first suspect assumption is probably the one used to predict individual word performance in the rank ordering described above. It is obvious that considering each word separately is not equivalent to predicting performance based on an aggregate probability distribution for all 15. Nonetheless, if all of the words were to perform individually with error rates below 4%, the aggregate would also be expected to be in this range. Consequently, the inconsistancy must lay elsewhere.

The most likely suspect was deemed to be the assumption that the word scores were not correlated with the trial. If this assumption was in fact invalid, then there should be evidence in the distribution of trial errors among the different speakers in the population, since clustering of scores within a particular trial would probably result in visible score clustering within speakers as well. Speakers contributing much more than their statistical share of the true trial errors across the population, whom we shall refer to as 'goats', could be responsible for much of the apparent performance loss. Clustered imposter errors could also be a factor, but as these are due to an unexpected success instead of an unexpected failure, it seems logical that the true errors could more easily be traced to some specific problem with the system or the speaker (i.e., telephone noise or speaker health and emotional state changes).

A series of runs using the discrete event simulation were performed with intermediate results printed out for each speaker in an effort to identify the goats if any were present. The runs used the conditions of the nine word case for which results were obtained above. In these runs three goats were found among the 50 speaker population who were responsible for more than 60% of the true reject errors. The results for one run, using a trial multiplication factor of 100, are shown below:

Speaker	Total TR (% of)	Total IA (% of)
Louise A.	26	0.3
Norman B.	14	0
Lilia S.	24	0
Total Errors	64	0.3

Removing these three speakers from the test data would thus yield a substantial true speaker performance improvement. In fact, this would put the true speaker performance within the 4% maximum suggested by the binomial expansion of the individual word distributions.

Although some speakers were also found who contributed more than their share of imposter errors, that is, there were also wolves and eels in the population, these speakers were a larger subset of the imposter population and had the imposter errors more evenly distributed among them. No one had more than 15% of the total imposter errors. Also, as these errors could not be pinned on noise or other abnormal variations from the test conditions, there appeared to be no justification for separating them from the other speakers (except if the reference patterns were polluted, which would lead to a reduction of information available for decision making and might cause all the scores to become artifically inflated).

With the three goats removed (i.e., using 94% of the test population), the flow graph model gave the following results for the RTHLs adjusted around the values shown:

	True Rejects (%)	Imposter Accepts (%)
RTHL=115	4.28	9.54
RTHL=116	8.45	6.20

EFFECT OF THE BINARY SONOGRAM

A program was used to create scores for the same test data both with and without the binary sonogram. The flowgraph model was used to evaluate the baseline case and cases B and C with the three goats removed. The results are shown below:

		True Rejects (%)	Imposter Accepts (%)
Baseline run	100	6.56	6.34
B, Goats Out	101	4.41	4.57
C, Goats Out	101	3.44	4.62

These error rates represent substantial improvements over the equivalent slope encoding only cases. The improvements are supported by the physical test results discussed above (see particularly Table 1).

DECISION MATRIX SENSITIVITY

Two sets of decision matrix sensitivity runs were performed using the discrete event simulation. The first set of runs was intended to examine the trend when (NxM) is varied by adding a constant to each term (i.e., (N+K x M+K)). The second set was to explore the trend when N was fixed at 7 and M varied. It appeared that both the imposter and true reject errors could be reduced almost indefinitely by simultaneously increasing both N and M. However, there also appeared to be a point of diminishing returns beyond about (7x5). In any case, going much beyond (7x5) would have been impractical because it would have made the verification procedure too long. Using a (7x5) matrix instead of a (5x3) matrix would evidently have yielded about the same imposter error rate, but a substantially reduced true reject error rate (imposter rate goes up as M goes up, and true errors increase with decreasing M). Hence, (7x5) was selected as the best available configuration, and this was used throughout instead of the original (5x3) matrix.

The baseline case of the previous section is compared for the (7x5) without conditional reject and (5x3) with conditional reject cases below.

	True Rejects (%)	Imposter Accepts (%)
(7x5) Baseline run (no CR)	6.56	6.34
(5x3) Baseline run (w/ CR)	7.10	7.12

In both cases the optimum RTHL was around 100. Both cases allow a maximum of five misses before an unconditional reject is determined. However, the (5x3) case allows the user to be accepted in only five hits if they are acquired before three misses. It can be seen that dropping the conditional reject allowance and requiring seven hits regardless yields slightly reduced error rates.

The (7x5) matrix was not used with conditional reject because this could have resulted in as many as

(7x5) + (2x2) = (9x7)

hits and misses, and this was felt to be too large.

CONCLUSIONS AND RECOMMENDATIONS

Comparison runs with the two models show very good agreement with each other. However, both tend to underestimate overall error rate performance because they both fail to consider the effects of word score correlations within access attempts. However, they underestimate by amounts that are considered tractable for the purposes of sensitivity analyses and general system performance assessment.

The following is a tentative algorithm specification for the speaker verification system based upon results from prototype testing of the lab system and simulation results.

1) The decision matrix is 5 word rejections by 7 required word acceptances with no allowance for conditional rejects.

2) The scoring algorithm includes both the frequency spectral coding and the binary sonogram.

3) The following is the recommended vocabulary:

	WORD	RTHL	P(7x5)
1	HAMMER	94	.06
2	ALABAMA	100	.39
3	ZERO	102	.56
4	MEGA	100	.57
5	MANUAL	92	.59
6	TWO	96	.59
7	HUNDRED	98	.61
8	NINE	94	.65
9	NOVEMBER	100	.66
10	TOMATO	94	.67
11	XYLOPHONE	104	.68
12	MANY	93	.80
13	HIGH	106	1.03
14	FOUR	110	1.98
15	ZEBRA	100	2.30

The system performance obtained for this set of specifications using the models described earlier is on the order of 3.5% to 4.5% for both Type I and Type II errors.

REFERENCES

1) Atal, B.S., "Automatic Recognition of Speakers from Their Voices," Proceedings of the IEEE, Vol. 64, No. 4, April 1976, pp. 460-475.

2) Clymer, J.R., OPERATIONAL EVALUATION MODELING. Copyright by John R. Clymer, 1980.

3) Crabbs, D.E., and Conrad, D.P., "Practical Experience with a Prototype Speaker Verification System," Proceedings of MIDCON/84, September 11-13, 1984.

4) Doddington, G.R. "Personal Identity Verification Using Voice," Proceedings of ELECTRO/76, May 11-14, 1976.

5) Doddington, G.R. "Voice Authentication Gets the Go-Ahead for Security Systems," SPEECH TECHNOLOGY, Vol. 2, No. 1, Sep/Oct 1983.

6) Ford, W. "Speaker Identity Verification System," Interstate Voice Product User's Group Conference, Itasca, IL, June 20-22, 1982.

7) Rabiner and Schafer, DIGITAL PROCESSING OF SPEECH SIGNALS, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1978.

8) Rosenberg, A.E. "Automatic Speaker Verification: A Review," Proceedings of the IEEE, Vol. 64, No. 4, April 1976, pp. 475-487.

9) Viglione, S.S. "Low Cost Voice Recognition Systems," Proceedings of MIDCON/81, November 10-12, 1981.