FBI Voice Database For
Automated Speaker Recognition Systems

EVALUATION TEST PLAN FOR THE

FORENSIC VOICE DATA SET

October, 1998

Prepared for:

FBI
Engineering Research Facility
Quantico, VA

Prepared by:

Tracor, Inc.
6500 Tracor Lane, MS 1-8
Austin, TX 78725

1.0 INTRODUCTION **1**

2.0 TECHNICAL OBJECTIVE **1**

3.0 DATA SETS **2**

3.1 DEVELOPMENT DATA SET 2
3.2 TRAINING DATA SET 3
3.3 TESTING DATA SET. 5
3.4 TEST DESCRIPTIONS 5
3.5 PREFERRED ORDER OF TESTING 5
3.6 EXAMPLE OF DEVELOPMENT AND EVALUATION TEST 10

4.0 EVALUATION **10**

4.1 EVALUATION MEASURES FOR OPEN SET VERIFICATION TESTS 10
4.2 EVALUATION MEASURES FOR CLOSED SET IDENTIFICATION TESTS 11
4.3 RULES FOR THE OPEN SET VERIFICATION EVALUATION 11
4.4 RULES FOR THE CLOSED SET IDENTIFICATION EVALUATION 12

5.0 FORMAT FOR SUBMISSION OF RESULTS **12**

6.0 ADDITIONAL INFORMATION REQUESTED **13**

7.0 SCHEDULE AND FUNDING **13**

8.0 CONTACTS **14**

1.0 INTRODUCTION

The purpose of this study is to evaluate existing speaker recognition systems using a data set that was designed specifically for forensic applications. The results of the study are intended to help the FBI assess the current technology, and to identify those systems that can be used to assist them in the analysis of recorded speech data. It is becoming increasingly necessary for the FBI to be familiar with this technology, and to understand its limitations. The eventual goal of the FBI is to acquire one and possibly multiple systems that can be used to augment aural and spectrographic analysis for a variety of speaker recognition tasks.

The evaluation results are primarily for internal use only by the FBI. The results for each participant will be made available to that participant, and we anticipate that the FBI will allow publication of those results. Each participant is encouraged to allow anonymous distribution of their results to other participants to serve as a benchmark. The FBI will not distribute or publish results involving a participant's name, unless express permission is granted.

Tracor is an engineering and defense contracting company that is assisting the FBI in the evaluation of speaker recognition systems. Tracor has a long history of experience with automatic recognition systems, and has been involved in the government testing of automatic detection and classification systems over the last ten years. Tracor was selected for the evaluation role for its experience, and because it does not have a conflict of interest – it is not involved in the development or sale of any speaker recognition systems. Additionally, Tracor is willing to safeguard and to respect any information provided by a participant in this evaluation. Tracor will have access to the evaluation results and to specific system-related information that may be requested by the FBI, and is prepared to sign a nondisclosure agreement. If a participating company does not have a nondisclosure agreement and would like Tracor to sign one, we are prepared to provide a standard limited-time nondisclosure agreement specific to this project.

The remainder of this document provides the technical objective of this evaluation, an overview of the data to be used for development and evaluation, the evaluation criteria and rules, and the format for submission of results. Additional reference information includes the following documents:

- User's Manual for Training and Testing for the FBI Voice Database Automated Speaker Recognition Systems, available from Tracor.
- The NIST 1998 Speaker Recognition Evaluation Plan, available on the web at address http://www.itl.nist.gov/div894/894.01/spkrec98.htm.

2.0 TECHNICAL OBJECTIVE

This speaker recognition evaluation consists of two tasks: open set verification and closed set identification. Open set verification is the primary task interest, and consists of determining if a particular speech segment matches (is the same speaker as) a known, or claimant voice sample. Closed set identification consists of comparing a speech segment with a set of known voice samples, and giving a forced choice decision of which known voice best matches the test sample and a rank order of the known voice samples in order of match to the test sample.

Both the verification and identification tasks are important problem areas in forensic speaker recognition for the FBI. In order to make this test applicable to actual situations, the test will be conducted with respect to the following conditions, which have been designed into the test database:

Text Independent – the training and test speech segments are unconstrained.

Text Dependent – the training and test speech segments are constrained to use the same words.

Transmission Mode Independent – The training and test segments are recorded using different input devices, i.e. telephone verses high fidelity microphone.

Transmission Mode Dependent – The training and test segments are recorded using the same input device.

Different Training and Test Segment Lengths – There are (3 and 12), and (30 and 120) segment lengths, however, all segments lengths are not available for all test conditions.

All speaker recognition systems are not designed to perform both verification and identification, or to perform well under all of the conditional tests. Each participant is asked to identify the tests they can complete within the limited time and budget, and to inform Tracor prior to the test.

3.0 DATA SETS

The Forensic Voice Database for Automated Speaker Recognition was developed as a part of project CAVIS during 1985-1989 in cooperation with the Los Angeles County Sheriff's Department and with the NIJ/DOJ grant 85-IJ-CX-0024. All of the speakers are male, and each speech segment contains a single speaker talking with a minimum of dead-time between words. The data was digitized at 16 KHz. and with 16-bit samples. The digitized files are stored as binary PCM data in the Motorola (MSB,LSB) format. Each file has a 1024-byte ASCII header that conforms to the NIST SPHERE format. The file ending is .sph, which conforms to the SPHERE documentation. The data is divided into three different sets, one for **development**, one for **training**, and one for **testing**. A description of each data set, and an example is given in the following sections.

3.1 Development Data Set

The purpose of the development set is to provide a statistical sampling of the acoustics of the transmission modes which are representative of the evaluation data. A transmission mode is an input transmission device, for example a microphone or a telephone handset.

There are multiple transmission modes being evaluated, however, due to the forensic nature of this test, one of the transmission modes will not be identified or represented in the development set, and the call-in telephone samples will not be represented in the development set.

All of the speakers in the development set were recorded simultaneously on three input devices. Two of these devices included a high quality B&K microphone and an in-house telephone of unknown handset type. Both devices are identified and represented in the development set. The third input device is not represented or identified, but has a frequency response similar to that of a telephone. The development set is contained on a separate CD-ROM, and consists of the following:

- The format of the development set is similar to that of the evaluation set.
- The speech modes, transmission modes, and segment lengths in the development set are representative of the evaluation data.
- There are ten separate speakers, all male.
- Two transmission modes from simultaneous recordings are represented, i.e. they are not independent speech segments.

The lists of files and the speaker label information for the development set is provided on a PC-formatted floppy disk, and is separated into training and testing folders. The format of the Training Development Set Labels is as follows:

Filename Spkr_label Trans_mode

The format of the Testing Development Set Labels is as follows:

Filename Spkr_label Trans_mode TRN_set_folder TRN_trans_mode

The Filename is a standard eight character filename similar to that used in the evaluation data set, with a .wav ending. The Spkr_label is a four digit label that is unique to that speaker. The Trans_mode consists of a T for telephone, or M for microphone. In each test file list, the training set folder (TRN_set_folder) and the training transmission mode (TRN_trans_mode) is given, providing information about which set of models to use, and what is being tested. The transmission modes are NOT labeled in the evaluation data set.

The remaining speakers, not represented in the development set, were asked to telephone in from arbitrary locations around the Los Angeles area. These telephone calls were then recorded at the receiving end of the call. There is no information available regarding the telephone handsets or the telephone exchange used in these phone calls. Any publicly available speech corpus can be used as the development set for these speaker files. For example, the NIST 1997 EvalSet can be obtained from the Linguistics Data Consortium[1] at the University of Pennsylvania.

3.2 Training Data Set

The training data is contained on three CD-ROMs, and is divided into 48 training data sets, each representing a different set of test conditions. These test conditions are

[1] http://www.ldc.upenn.edu/

categorized according to four levels of difficulty, two primary training set lengths, two speaking modes, and three transmission modes. These conditions are summarized in the following table.

Table 1 – Summary of the Training Data Set

Level	Trial Numbers	Length (sec)	Speaking Mode	# Speakers
I	1-3	29	Spontaneous	40-41
I	4-6	29	Reading	8-9
I	7-9	4*29	Spontaneous	40-41
I	10-12	4*29	Reading	8-9
II	1-3	3	Prescribed	18-19
II	4-6	29	Prescr. Reading	20-21
II	7-9	4*3	Prescribed	18-19
II	10-12	4*29	Prescr. Reading	20-21
III	1-3	29	Spontaneous	40-87
III	4-6	29	Reading	8-9
III	7-9	4*29	Spontaneous	40-87
III	10-12	4*29	Reading	8-9
IV	1-3	3	Prescribed	18-19
IV	4-6	29	Prescr. Reading	20-21
IV	7-9	4*3	Prescribed	18-19
IV	10-12	4*29	Prescr. Reading	20-21

The Level in column 1 refers to the level of difficulty, corresponding to the following criteria:

Level I = Text Independent, Transmission Mode Independent
Level II = Text Dependent, Transmission Mode Independent
Level III = Text Independent, Transmission Mode Dependent
Level IV = Text Dependent, Transmission Mode Dependent

There are 12 trials per level of difficulty, which are organized according to the different conditions on the test. The length is the individual file length in seconds. In trials 7-12, there are four files from the same speaker; these files can be appended together, providing a four times increase in the segment length. The speaking mode refers to the relationship of the words and phrases among the speech samples. For spontaneous mode, the speaker was asked to talk about a given topic, and all words and phrases are random among files. For reading mode, the speaker was asked to read a passage, and all passages were different. For prescribed reading mode, each speaker was asked to read the same passage. For prescribed mode, each speaker was asked to repeat the same short phrase.

The training filenames for each of the 48 tests are stored in separate folders on a training set floppy disk. The speaker labels for each file is also listed along with the filename. A more complete description of the training data CD-ROMs and of the organization of the files can be found in the User's Manual for Training and Testing.

3.3 Testing Data Set.

The testing data is contained on three CD-ROMs, and is divided into 48 testing data sets, each required for a different set of test conditions. These test conditions are categorized similarly to the training data set. The testing filenames for each of the 48 tests are stored in separate folders on a testing set floppy disk. An additional description of the test data CD-ROMs and of the organization of the files can be found in the User's Manual for Training and Testing.

3.4 Test Descriptions

A description of each of the 48 tests is provided in Tables 2-5 below. These tables, along with Table 1 provide a complete description of the tests and test conditions. These tables provide the following information:

> **Level-Trial** refers to the Level of difficulty of the test, and the conditional trial number.
>
> **Eval Folders TRN/TST** refers to the folder names containing the training and testing files for the evaluation test.
>
> **Trans_Mode TRN/TST** refers to the transmission modes in the training and test data. M is microphone, T is telephone, and O is other (similar to telephone).
>
> **# Eval Speakers** refers to the number of speakers in the evaluation for that particular trial.
>
> **Dev Folders TRN/TST** refers to the training and test folders containing the development data for that particular trial – to be used for setting thresholds.
>
> **# Dev Speakers** refers to the number of speakers in the development data for that particular trial.

3.5 Preferred Order of Testing

Because of the large number of tests and lack of available funding, we realize that the participants may not be able to perform all tests. The FBI's preferred order of testing on level of difficulty is I, III, II, and then IV. Within each of the 12 trials per level of difficulty, the FBI is most interested in trails 1-2-3, then 4-5-6, then 7-8-9, and lastly 10-11-12.

Table 2 – Level I Evaluations: Text Independent, Transmission Independent

Level-Trial	Eval Folders TRN/TST	Trans_Mode TRN/TST	# Eval Speakers	Dev Folders TRN/TST	# Dev Speakers
I-1	L13TRN01 L01TST01	M T,O	40	D13TRN01 D01TST01	10
I-2	L13TRN02 L01TST02	T M,O	41	D13TRN02 D01TST02	10
I-3	L13TRN03 L01TST03	O M,T	41	Use I-2	
I-4	L13TRN04 L01TST04	M T,O	8	Use I-1	
I-5	L13TRN05 L01TST05	T M,O	9	Use I-2	
I-6	L13TRN06 L01TST06	O M,T	9	Use I-2	
I-7	L13TRN07 L01TST07	M T,O	40	D13TRN07 D01TST07	10
I-8	L13TRN08 L01TST08	T M,O	41	D13TRN08 D01TST08	10
I-9	L13TRN09 L01TST09	O M,T	41	Use I-8	
I-10	L13TRN10 L01TST10	M T,O	8	D13TRN10 D01TST10 Or I-7	2
I-11	L13TRN11 L01TST11	T M,O	9	D13TRN11 D01TST11 Or I-8	2
I-12	L13TRN12 L01TST12	O M,T	9	Use I-11	

Table 3 – Level II Evaluations: Text Dependent, Transmission Independent

Level-Trial	Eval Folders TRN/TST	Trans_Mode TRN/TST	# Eval Speakers	Dev Folders TRN/TST	# Dev Speakers
II-1	L24TRN01 L02TST01	M T,O	20	D24TRN01 D02TST01	4
II-2	L24TRN02 L02TST02	T M,O	18	D24TRN02 D02TST02	4
II-3	L24TRN03 L02TST03	O M,T	19	Use II-2	
II-4	L24TRN04 L02TST04	M T,O	21	D24TRN01 D02TST01	6
II-5	L24TRN05 L02TST05	T M,O	21	D24TRN02 D02TST02	6
II-6	L24TRN06 L02TST06	O M,T	20	Use II-5	
II-7	L24TRN07 L02TST07	M T,O	20	D24TRN07 D02TST07	4
II-8	L24TRN08 L02TST08	T M,O	18	D24TRN08 D02TST08	4
II-9	L24TRN09 L02TST09	O M,T	19	Use II-8	
II-10	L24TRN10 L02TST10	M T,O	21	D24TRN10 D02TST10	6
II-11	L24TRN11 L02TST11	T M,O	21	D24TRN11 D02TST11	6
II-12	L24TRN12 L02TST12	O M,T	20	Use II-11	

Table 4 – Level III Evaluations: Text Independent, Transmission Dependent

Level-Trial	Eval Folders TRN/TST	Trans_Mode TRN/TST	# Eval Speakers	Dev Folders TRN/TST	# Dev Speakers
III-1	L13TRN01 L03TST01	M M	40	D13TRN01 D03TST01	10
III-2	L03TRN02 L03TST02	T T	87	D13TRN02 D03TST02	10
III-3	L13TRN03 L03TST03	O O	41	Use III-2	
III-4	L13TRN04 L03TST04	M M	8	Use III-1	
III-5	L13TRN05 L03TST05	T T	9	Use III-2	
III-6	L13TRN06 L03TST06	O O	9	Use III-2	
III-7	L13TRN07 L03TST07	M M	40	D13TRN07 D03TST07	10
III-8	L03TRN08 L03TST08	T T	87	D13TRN08 D03TST08	10
III-9	L13TRN09 L03TST09	O O	41	Use III-8	
III-10	L13TRN10 L03TST10	M M	8	D13TRN10 D03TST10 Or III-7	2
III-11	L13TRN11 L03TST11	T T	9	D13TRN11 D03TST11 Or III-8	2
III-12	L13TRN12 L03TST12	O O	9	Use III-11	

Table 5 – Level IV Evaluations: Text Dependent, Transmission Dependent

Level-Trial	Eval Folders TRN/TST	Trans_Mode TRN/TST	# Eval Speakers	Dev Folders TRN/TST	# Dev Speakers
IV-1	L24TRN01 L04TST01	M M	20	D24TRN01 D04TST01	4
IV-2	L24TRN02 L04TST02	T T	18	D24TRN02 D04TST02	4
IV-3	L24TRN03 L04TST03	O O	19	Use IV-2	
IV-4	L24TRN04 L04TST04	M M	21	D24TRN01 D04TST01	6
IV-5	L24TRN05 L04TST05	T T	21	D24TRN02 D04TST02	6
IV-6	L24TRN06 L04TST06	O O	20	Use IV-5	
IV-7	L24TRN07 L04TST07	M M	20	D24TRN07 D04TST07	4
IV-8	L24TRN08 L04TST08	T T	18	D24TRN08 D04TST08	4
IV-9	L24TRN09 L04TST09	O O	19	Use IV-8	
IV-10	L24TRN10 L04TST10	M M	21	D24TRN10 D04TST10	6
IV-11	L24TRN11 L04TST11	T T	21	D24TRN11 D04TST11	6
IV-12	L24TRN12 L04TST12	O O	20	Use IV-11	

3.6 Example of Development and Evaluation Test

An example of training and testing for a development set and an evaluation set is given for Level I, trial 1.

The Development Set floppy disk contains two folders: Trainfiles-d for the training file lists, and Testfiles-d for the test file lists. Inside the folder Trainfiles-d, the file D13TRN01.LST contains a list of the filenames and their associated ground truth for developing speaker models under the conditions of this particular test. There are ten files, one for each speaker, and each containing 30 seconds of spontaneous speech recorded using a high fidelity microphone. Inside the folder Testfiles-d, the file D01TST01.LST contains a list of the test filenames and their associated ground truth. There are test files of different lengths, including 3, 12, 30, and 120 seconds in length for each of the ten speakers. The development set is to be used for quick tuning of the models, and setting thresholds according the evaluation criteria described in the next section. All of the training and testing files for the development set are on one CD-ROM in the folder Fv1_dev.

There are two Evaluation Set floppy disks, one for training file lists and one for testing file lists. The training set floppy disk contains the folder named Trainfiles, and the test set floppy disk contains the folder Testfiles. Inside the folder Trainfiles, the file L13TRN01.LST contains a list of the filenames and their associated ground truth speaker labels. The evaluation set list format is slightly different from the development set list format. The training files are located on one of three CD-ROMS, in a folder with the same name as the training file lists (L13TRN01). Inside the folder Testfiles on the testing floppy disk, the file L01TST01.LST contains a list of the test filenames with no ground truth information. The testing files are located on one of three testing CD-ROMS, in a folder with the same name as the test file lists (L01TST01).

4.0 EVALUATION

This section defines the evaluation measures that will be used to determine the performance of the systems for both the open set verification and the closed set identification tests. The rules and restrictions for each test are also defined.

4.1 Evaluation Measures for Open Set Verification Tests

The performance for open set verification will be evaluated by measuring the correct detection decisions for an ensemble of target speech segments. For each test segment, a set of target speaker identities from the training set will be assigned as the test hypotheses. The true (same speaker) or false (different speakers) binary decisions will be made to minimize a detection cost function and the minimum detection cost achieved will be the primary performance measure. In addition to the binary decision, a decision score will be necessary in order to plot the Receiver Operating Characteristic (ROC) performance curves.

The decision cost function is defined as the weighted sum of the miss and false alarm probabilities, where:

$$C_{Det} = C_{Miss} * P_{Miss/Target} * P_{Target} + C_{FalseAlarm} * P_{FalseAlarm/NonTarget} * P_{NonTarget}$$

The parameters of the cost function are defined below:

$$C_{Miss} = 4; \quad C_{FalseAlarm} = 1; \quad P_{Target} = 0.01; \quad P_{NonTarget} = 1 - P_{Target} = 0.99$$

The thresholds used to meet this cost function are to be calculated from the development data sets. Scores should be produced so that a numerically larger score means that the target and test files are more likely to be the same speaker.

4.2 Evaluation Measures for Closed Set Identification Tests

The performance for closed set identification will be evaluated by measuring the error rate in identifying the speaker of the test segment from the set of target speakers. In addition to classification error rate, the average rank of the true speaker in the rank ordered list of target speakers will be used as a performance measure.

4.3 Rules for the Open Set Verification Evaluation

The following rules and restrictions are defined for the open set verification tests, and must be observed:

- Each decision is to be based only on the information from the specified test segment and target speaker. Use of additional information about the test segment or the target speaker, such as normalization over multiple test segments or multiple target speakers, is not allowed.
- Listening to, or the use of transcripts for the short 3-second prescribed text files will be allowed. However, we are not allowing transcripts of the longer 30-second files, which are primarily spontaneous speech.
- Listening to the evaluation files is not allowed prior to submitting results. However, listening to the development data is allowed.
- Knowledge of the training conditions, as derived from this document or the Training and Testing User’s Manual is allowed.

4.4 Rules for the Closed Set Identification Evaluation

The rules and restrictions for the closed set identification tests are identical to the rules for open set verification, except for the first rule. Identification requires knowledge of the entire set of target speakers:

- Each decision is to be based only on the information from the specified test segment and the training set of target speakers.

5.0 FORMAT FOR SUBMISSION OF RESULTS

All evaluation results must be provided to Tracor in results using a standard ASCII record format, with one record for each decision. Separate results files will be output for the verification test results and the identification results.

The output results file for the **open set verification tes**t will consist of one record per decision, and each record containing 7 columns on one row, separated by white spaces. This output format is identical to that for the NIST 1998 Speaker Recognition Evaluation Test Plan. A description of each column in the output results file is given below:

Column	Description
1	Sex (all M)
2	Training list filename from floppy disk (e.g. L13TRN01)
3	Target Speaker ID number (4-digit label)
4	Test list filename from floppy disk (e.g. L03TST01)
5	Test segment filename
6	Decision (T/F)
7	Score (larger number means target and test speaker more likely to be the same)

The output results for each test file in the **closed set identification test** will consist of an output for each model in the training set. Tracor will be responsible for rank ordering the scores and using the information to calculate various performance measures. Each record should start on a new line with carriage return. A description of each column in the results file is given below:

Column	Description
1	Sex (all M)
2	Training list filename from floppy disk (e.g. L13TRN01)
3	Target Speaker ID number (4-digit label)
4	Test list filename from floppy disk (e.g. L03TST01)
5	Test segment filename
6	Decision (T/F)
7	Score (larger number means target and test speaker more likely to be the same)[2]

The preferred means of submitting results is by sending the files to the Tracor anonymous FTP site. Filenames and directories can be verbally agreed on prior to submitting the results. The current directory for these results can be accessed as follows:

Host name: ftp.aard.tracor.com
Login: anonymous
Password your_name@your_address
ftp> cd speakerid
ftp> put your_files
ftp> bye

6.0 ADDITIONAL INFORMATION REQUESTED

Additional information that is requested along with the results includes a high-level description of the algorithm that was used, a description of the computer system that the test was run on, the name of any additional speech corpuses that were used for development , and the average CPU time for a single CPU to process the test data.

7.0 SCHEDULE AND FUNDING

We are interested in getting the results as soon possible, and are requesting that all test results be completed by the beginning of November, 1998. We understand that there may be scheduling conflicts, and are willing to work with each participant.

There is limited funding available to perform this evaluation, and there is **no** funding available this year to pursue speaker recognition development work. Participants that require funding are requested to contact Steve Beck or Larry Deuser at Tracor to discuss this issue. We are attempting to treat all participants in a fair manner with regard to this evaluation test.

[2] Systems that do not produce a match score for closed identification should use (-1*model_rank) as the match score.

8.0 CONTACTS

The contacts for this evaluation are:

Name	Affiliation	Contact Information
Steven D. Beck	Tracor	6500 Tracor Ln. MS 1-8 Austin, TX 78725 Phone: (512) 929-2034 FAX: (512) 929-2055 Email: steve_beck@tracor.com
Dr. Larry M. Deuser	Tracor	6500 Tracor Ln. MS 1-8 Austin, TX 78725 Phone: (512) 929-2047 FAX: (512) 929-2055 Email: larry_deuser@tracor.com
Dr. Hiro Nakasone	FBI	Phone: (703) 630-6488 FAX: (703) 630-6693