The NIST Year 2001 Speaker Recognition Evaluation Plan

1 Introduction

The year 2001 speaker recognition evaluation is part of an ongoing series of yearly evaluations conducted by NIST. These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of text independent speaker recognition. To this end the evaluation was designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to those wishing to participate.

The evaluation will be conducted in the spring. The data will be available in April, with results due to be submitted to NIST about three weeks later. A follow-up workshop for evaluation participants to discuss research findings will be held in June. Specific dates are listed in section 11, Schedule.

Participation in the evaluation is invited for all sites that find the tasks and the evaluation of interest. For more information, and to register to participate in the evaluation, please contact Dr. Alvin Martin at NIST.[1]

2 Technical Objective

This speaker recognition evaluation focuses on the tasks of speaker detection, segmentation, and tracking. These tasks are posed in the context of conversational telephone speech. The evaluation is designed to foster research progress, with the goals of:

- Exploring promising new ideas in speaker recognition.
- Developing advanced technology incorporating these ideas.
- Measuring the performance of this technology.

2.1 Task Definitions

The year 2001 speaker recognition evaluation plan includes the following four tasks:

2.1.1 One-speaker detection

This task is NIST's basic speaker recognition task, as defined in all of NIST's previous annual speaker recognition evaluations. The task is to determine whether a specified speaker is speaking during a given speech segment.

2.1.2 Two-speaker detection

This task is essentially the same as the one-speaker detection task, except that the speech segments include both sides of a telephone call (summed together), rather than being limited to the speech from a single speaker.

2.1.3 Speaker tracking

This task requires determining the time intervals (if any) during which a specified speaker is speaking in a segment of conversational speech. This task is performed on a subset the test segments used for the two-speaker detection task.

2.1.4 Speaker segmentation

This task requires identifying the time intervals during which unknown speakers are each speaking in a conversational speech segment (no speakers are specified). The number of different speakers may or may not be known.

2.2 Task Conditions

The year 2001 speaker recognition evaluation plan includes four distinct combinations of parameters and data. The first of these parts is essentially a repeat of last year's evaluation. This part exercises all of the four speaker recognition tasks. The remaining three parts exercise only the basic one-speaker detection task. NIST's interest remains focused on the one-speaker detection task, this being fundamental to and diagnostic of essentially all application variants of speaker recognition technology.

2.2.1 An expanded version of last year's evaluation

This part of the evaluation test suite includes essentially all of last year's tasks and data. For the one-speaker detection task, however, the number of hypothesized target speakers is being increased for each test segment. This is being done to support more extensive characterization of performance over various conditions.

2.2.2 A test using Spanish language data

The non-conversational Spanish language AHUMADA Corpus will be used to support the one-speaker detection task on Spanish. This was also part of last year's evaluation suite and will be implemented with exactly the same protocol as last year.

2.2.3 A test using cellular telephone data

A cellular telephone corpus, using the SwitchBoard data collection paradigm, has been collected.[2] Part of this corpus will be used to support a separate evaluation of one-speaker detection.

2.2.4 A test to explore idiolectal characteristics

To date the one-speaker detection task has been defined in the context of limited training data. For example, last year the training data was limited to one two-minute session for each target speaker. While this is appropriate for many applications, there are also applications of speaker recognition that admit much greater exposure to target speakers. Therefore, NIST will this year provide an extended training condition for the one-speaker detection task using the SwitchBoard corpus. In addition to the usual acoustical data, both manual and ASR transcriptions will be provided, as special options, to be used as input data. Each target speaker will have up to 1 hour or more of speech data for training. The purpose of this condition is to support the exploration and development of

[1] To contact Dr. Martin, send him email at alvin.martin@nist.gov, or call him at 301/975-3169.

[2] Refer to http://www.ldc.upenn.edu/Projects/SWB/cellular/ for details on this corpus and the SwitchBoard collection paradigm as it was applied to this collection.

idiolectal characteristics for speaker detection. A recent study has found some interesting results in this area.[3]

3 THE EVALUATION

Evaluation will be performed separately for each of the four parts listed in section 2.2. In the case of section 2.2.1, the evaluation will be performed separately for each of the four tasks, of course. In addition, for the speaker segmentation task, two separate evaluations will be supported – one with the number of speakers unknown and an optional one with the number of speakers controlled to be exactly two.

3.1 Speaker Detection Tasks

For the detection tasks the formal evaluation measure is the detection cost function, defined as a weighted sum of the miss and false alarm error probabilities:

$$\begin{aligned} C_{Det} &= C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ &+ C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times P_{NonTarget} \end{aligned}$$

The parameters of this cost function are the relative costs of detection errors, C_{Miss} and $C_{FalseAlarm}$, and the *a priori* probability of the target (specified) speaker, P_{Target}. The following parameter values will be used as the primary evaluation of speaker recognition performance for the detection tasks:

C_{Miss}	$C_{FalseAlarm}$	P_{Target}	$P_{NonTarget}$
10	1	0.01	1 - P_{Target}

Speaker detection performance will be evaluated by measuring the correctness of detection decisions for an ensemble of speech segments in terms of the detection cost function. These segments will represent a statistical sampling of conditions of evaluation interest. For each of these segments a set of speaker identities will be assigned as test hypotheses. Each of these hypotheses must be independently judged as true or false, and the correctness of these decisions will be tallied.[4]

In addition to the actual detection decision, a decision score will also be required for each test hypothesis. This decision score will be used to produce detection error tradeoff curves, in order to see how misses may be traded off against false alarms.[5]

3.2 Speaker Tracking

Speaker tracking will be evaluated by comparing the time periods when the target speaker is talking, as determined by the system, with the reference time periods as determined by NIST. NIST will determine the time periods when each speaker is speaking with the aid of an energy based speech detector on the individual speech channels (conversation sides) before summing or, where available, by using time marks determined by human transcribers.

Scoring will be limited to those times where the speech detector indicates that only one of the speakers is speaking. In addition, times within 250 milliseconds of the end of an interval of speech will also be ignored. (Thus, for example, speech segments of less than 0.5 seconds will not be scored.)

The decisions for each hypothesized speaker will be compared with a reference answer key to determine the miss and false alarm rates for speaker tracking, according to the following computation:

$$P_{Miss} = \int_{\text{Target Speech}} \delta(D_t, F)\,dt \Big/ \int_{\text{Target Speech}} dt$$

$$P_{FalseAlarm} = \int_{\text{NonTarget Speech}} \delta(D_t, F)\,dt \Big/ \int_{\text{NonTarget Speech}} dt$$

where

$$D_t = \text{the system output } (T \text{ or } F) \text{, as a function of time}$$

$$\delta(\mathrm{x}, \mathrm{y}) = \begin{cases} 1 \text{ if } \mathrm{x} = \mathrm{y}, \\ 0 \text{ otherwise} \end{cases}$$

The parameter values for the cost function for speaker tracking will be different from those for speaker detection. The tracking parameter values will be:

C_{Miss}	$C_{FalseAlarm}$	P_{Target}	$P_{NonTarget}$
1	1	0.5	1 - P_{Target}

In addition to the actual decisions for each detected target interval, a decision score will also be required. This decision score will be used to produce DET (detection error tradeoff) curves, in order to see how misses may be traded off against false alarms.

3.3 Speaker Segmentation

As in the tracking task, speaker segmentation will be evaluated by comparing system-determined regions where the various speakers are talking with reference regions determined by NIST. NIST will determine the regions where each speaker is speaking using an energy based speech detector on the individual speech channels (conversation sides) before summing or, where available, by using time marks determined by human transcribers.

As in the tracking task, scoring will be limited to those times where the speech detector indicates that only one of the speakers is speaking. Times within 250 milliseconds of the end of an interval of speech will also be ignored.

The decisions for each hypothesized speaker will be compared with a reference answer key to determine the miss and false alarm rates for speaker tracking, according to the following computation:

NIST will score separately the segments where the number of speakers present is known in advance and those where the number is not known. The same scoring procedure will be used in each case, however.

For the speaker segmentation task, a system must produce hypothesized speaker turns (or segments of speech produced from a single speaker) and a generic speaker label for each turn (e.g., speaker 0, speaker 1, ...). All turns from the same speaker should have the same generic speaker label. Unlike the other detection

[3] See G. Doddington, “Some Experiments on Idiolectal Differences among Speakers”, available on the NIST website:

http://www.nist.gov/speech/tests/spk/2001/doc/n-gram_experiments-v06.pdf

[4] This means that an explicit speaker detection decision is required for each trial. Explicit decisions are required because the task of determining appropriate decision thresholds is a necessary part of any speaker detection system and is a challenging research problem in and of itself.

[5] Decision scores from the various target speakers will be pooled before plotting detection error tradeoff curves. Thus it is necessary to normalize scores across speakers to achieve satisfactory detection performance.

tasks, this is a classification task, which can be characterized by a single error rate. (There is only a single error since all speech must be accounted for and a "miss" for one hypothesized speaker label will generate a corresponding "false alarm" for another hypothesized speaker label, so the two errors are no longer independent.) To compute the classification error a search for the best (minimum error) mapping of hypothesized speaker labels to true speakers for each conversation is performed, and then the errors are accumulated over the ensemble of test conversations. The error rate is computed as follows:

Assume a system produces M hypothesized speaker labels for a conversation that actually contains N true speakers. (When the number of speakers is known in advance, we will have $M = N$). When $M < N$ we automatically generate N-M speaker labels with no associated speech segments. We first produce a one-to-one mapping of true speakers $\{i\}$ with hypothesized speaker labels $\{j\}$

$$\mathit{map}(i) \;=\; j; \quad i=1,..,N; \quad 1 <= j <= M$$

Let $\mathit{true\_duration}(i,\ conv)$ be the total duration of speech by true speaker i in a conversation denoted conv, and let $\mathit{hyp\_duration}(i,j,conv)$ be the total duration of speech common to true speaker i and hypothesized speaker label j in this conversation. The error rate is then

$$\mathit{error}(map,\ conv) \;=\; 1 \;-\; (S\ i\ \mathit{hyp\_duration}(i,\ map(i),\ conv)\ /\ S\ i\ \mathit{true\_duration}(i,\ conv))$$

The best one-to-one mapping, $\mathit{map}^*(i)$, used for this conversation is the one producing the minimum $\mathit{error}(map,\ conv)$. For the conversation we then log the values

$$\mathit{hit}(conv) = S\ i\ \mathit{hyp\_duration}(i,\ map^*(i),\ conv)$$

and

$$\mathit{total}(conv) = S\ i\ \mathit{true\_duration}(i,\ conv)$$

The total (weighted) error over an ensemble of conversations is finally computed as

$$\mathit{total\_error} = 1 - S\ conv\ \mathit{hit}(conv)\ /\ S\ conv\ \mathit{total}(conv)$$

4 EVALUATION CONDITIONS

In previous evaluations the use of telephone handset labels (namely either "*electret*" or "*carbon-button*") has provided a significant improvement in speaker recognition performance. This labeling was done automatically by analysis and classification of the speech signal in the trial. This year we will again provide these labels (most will be "*electret*") and their likelihoods.[6] During evaluation, systems will be allowed to use these labels and likelihoods for all one-speaker training and test data, *except for the cellular data*. No label information will be provided for the two-speaker test, however.

4.1 Last year's evaluation + Spanish language

This section reviews the conditions for last year's evaluation, including the Spanish language AHUMADA corpus.

4.1.1 Training

Training data for each speaker will consist of about two minutes of speech from a single conversation, for all speaker detection tasks. The actual duration of the training files used will vary slightly from this nominal value so that whole turns may be included whenever possible. Actual durations will, however, be constrained to lie within the range of 110-130 seconds.

4.1.2 One-Speaker Detection Test

Each test segment will be extracted from a 1-minute excerpt of a single conversation and will be the concatenation of all speech from the subject speaker during the excerpt. The duration of the test segment will therefore vary, depending on how much the segment speaker spoke.

Evaluation will be limited to "different-handset" tests only. This will be done by ensuring that test segments and models are taken from different phone number data.

Most of the training and test segments will be from handsets algorithmically determined to be of electret microphone type, but some data from handsets with carbon-button type microphone will be included for contrast.

The primary evaluation conditions are:

1. The handsets are electret (as determined by the MIT labeler algorithm).
2. The speech duration is 15-45 seconds.

Results will be tabulated separately for male and female model speakers. There will be no cross-sex tests.

4.1.3 Two-Speaker Detection Test

Each test segment will have a duration of nominally 60 seconds and will be the sum of the two sides of a conversation. The actual duration will vary from nominal, so that the test segment begins and ends on a speaker turn boundary. Actual test segment duration will, however, be constrained to lie within the range of 59-61 seconds. Note that the duty cycle of a segment speaker may vary from 0% to 100%.

There are three possible cases with respect to gender for each test segment: both speakers are male; both are female; or one is male and one female. Performance will be computed and evaluated separately for each of these three cases, but the system will not be given prior knowledge detailing the gender mix of the test segment. (Automatic gender detection may be used, of course.)

The primary evaluation conditions are:

1. Both handsets are electret (as determined by the MIT labeler algorithm).
2. The speech duration is 15-45 seconds for both speakers.
3. Both speakers are of the same sex.

4.1.4 Speaker Tracking Test

Test segments for the speaker tracking test will be a subset of the test segments for the two-speaker detection test.

The primary conditions of interest for the speaker tracking task will be the same as for the two-speaker detection task. In addition, it is given that the model speaker is one of the segment speakers.

[6] The labeling will be performed using MIT Lincoln Lab's handset type labeler software.

4.1.5 Speaker Segmentation Test

There will be two distinct conditions for the speaker segmentation task, each with its own distinct type of test segment.

4.1.5.1 The 2-speaker segmentation condition

The test segments for the 2-speaker segmentation condition will be same as the test segments for the speaker tracking task. These test segments have a duration of 1 minute and are conversations between exactly two speakers.

4.1.5.2 The N-speaker segmentation condition

The test segments for the N-speaker segmentation condition will be of longer duration, up to 10 minutes long. These test segments will also be taken from different languages. The number of speakers in each test segment will not be provided. It is given, however, that this number will be no more than 10.

The primary evaluation condition for the speaker segmentation task will be the N-speaker segmentation condition.

4.2 Evaluation of cellular telephone data

This section specifies the training and test conditions for the one-speaker detection task on cellular data.

4.2.1 Training

Training data for each speaker will consist of about two minutes of speech from a single conversation. The actual duration of the training files used will vary slightly from this nominal value so that whole turns may be included whenever possible. Actual durations will, however, be constrained to lie within the range of 110-130 seconds.

There will be no automatically generated handset labels for this data.

4.2.2 Test

The test segments for the cellular data test are defined in the same way as for other one-speaker detection tests, with the following exceptions:

- Evaluation trials for cellular data will include both "same number" and "different number" tests.
- There will be no automatically generated handset labels for the cellular data.

The primary evaluation conditions are:

1. "different number" tests *(subject to change depending on how many different number and same number trials make it into the evaluation set)*
2. The duration of the test segment is 15-45 seconds.

Results will be tabulated separately for male and female model speakers. There will be no cross-sex tests.

4.3 Evaluation of extended training and test

This section outlines the conditions for the one-speaker detection task with extended training and test data. The entire SwitchBoard-I corpus will be used for this evaluation. Since this is a first-time exploration of extended training and test, and because of the need for large amounts of data, the evaluation is defined in a developmental framework rather than as a final independent and unbiased test of any resulting technology.

4.3.1 Training

Speaker training data will comprise all of one or more conversation sides for a given model speaker. A jackknife scheme that rotates training and test data will be used in order to provide an adequate number of tests. In order to provide unbiased results, models must exclude test conversation-sides from target speakers and all data from test impostors. This information will be provided in index files that must be used to control the evaluation. Instructions are given in section 8.3 for the use of this index file information.

Various training options exist. The acoustical data may be used alone, the transcriptions (either manual or ASR) may be used alone, or they may be used in combination. Note that the conversation sides and the transcriptions (both manual and ASR) are presented in their entirety, without excision or deletion.

4.3.2 Test

The task is one-speaker detection. One whole conversation side will serve as the test segment. As in training, the acoustical data may be used alone, the transcriptions (either manual or ASR) may be used alone, or they may be used in combination. And as in training, the data are presented in their entirety for the whole conversation side, without excision or deletion.

Results will be evaluated as a function of the amount target speaker training data, the handset types, and speaker sex. For some but not all true-speaker trials, the test handset will be among those included in the target speaker training data. Some cross-sex trials will also be included.

5 DEVELOPMENT DATA

5.1 Last year's evaluation

The development data for last year's evaluation will remain the same as last year. Please refer to last year's evaluation plan for details.[7]

5.2 The cellular telephone evaluation

The development data for the cellular telephone evaluation comes from the same corpus as the evaluation data, namely SwitchBoard-II Phase 4.

NIST has created the development set (NIST speech disc R71_1_1) using the speakers who participated in 4 or fewer conversations. More details are specified on the CD.

Training data is provided for 22 females and 39 males. There are 34 female test segments and 44 male test segments.

Note, the development set makes use of only a single side of each conversation, unlike in past evaluations, NIST will use the other side for evaluation data.

5.3 The extended data evaluation

The extended data evaluation will be treated as an exploratory R&D activity. Thus the test corpus will also serve a dual role as a development data set.

[7] The year 2000 speaker recognition evaluation plan may be accessed from http://www.nist.gov/speech/tests/spk/2000/doc/

6 EVALUATION DATA

6.1 Last year's evaluation

The evaluation data for last year's evaluation, with the exception of the index files for one-speaker detection, will remain the same as last year. Please refer to last year's evaluation plan for details.[7]

6.2 The cellular telephone evaluation

The evaluation data for one-speaker detection using cellular data will be drawn from the SwitchBoard-II Corpus, Phase 4. All conversations will be processed through echo canceling software before being used to create training and test segments.

Training and test segments will be constructed by concatenating consecutive turns of the desired speaker. Each such segment will be stored as an 8-bit μ-law continuous speech signal in a separate SPHERE file. The SPHERE header of each such file will contain some auxiliary information as well as the standard SPHERE header fields.

There will be about 190 target speakers and about 2,200 test segments. Each test segment will be evaluated against 11 hypothesized speakers of the same sex as the segment speaker.

6.3 The extended data evaluation

All of the SwitchBoard-I corpus will serve as both the development and the evaluation data for the extended data evaluation. The recently updated and corrected versions of the manual transcriptions are available from the Institute for Signal and Information Processing (ISIP)[8]. Dragon Systems has generously provided ASR transcriptions of the complete SwitchBoard corpus that participants may access from NIST's web site.[9] The audio corpus itself must be obtained from the Linguistic Data Consortium (LDC)[10], which makes it available for sale to non-members.

7 EVALUATION RULES

In order to participate in the 2001 speaker recognition evaluation, a site must complete, in its entirety, at least one complete evaluation of one of the four evaluation tasks. This may be done for one or more tasks and for one or more of the different parts of the evaluation.[11]

For the segmentation task participants may, if they choose, submit only results for the CALLHOME data, for which the number of speakers present will be unknown. This option is intended to accommodate those wishing to do speaker segmentation but not speaker tracking.

All participants must observe the following evaluation rules and restrictions:

- Each decision is to be based only upon the specified test segment and target speaker. Use of information about other test segments and/or other target speakers is **not** allowed.[12] For example:
  - Normalization over multiple test segments is **not** allowed.
  - Normalization over multiple target speakers is **not** allowed.
  - Use of evaluation data for impostor modeling is **not** allowed (except for the extended data test as indicated in the index files).
- The use of transcripts for training is **not** allowed (except for the extended data test, of course).
- Knowledge of the sex of the *target* speaker (implied by data set directory structure as indicated below) **is** allowed.
- Knowledge of the sex of the speaker(s) in the *test segment* is **not** allowed, except as determined by automatic means.
- Knowledge of handset type information (found in the SPHERE header under the field name "handset_type" -- see SPHERE Header Information, below) **is** allowed for the one-speaker test segments (because this information was determined by automatic means).
- Knowledge of the handset type mixture of the two-speaker test segments is **not** allowed, except as determined by automatic means.
- For the segmentation task, knowledge of the number of speakers present is **not** allowed, except as determined by automatic means.
- Listening to the evaluation data, or any other experimental interaction with the data, is **not** allowed before all test results have been submitted. This applies to training data as well as test segments.
- The corpora from which the evaluation data are taken, namely the SwitchBoard-II Phases 1 and 2, CALLHOME, and AHUMADA Corpora, may **not** be used for any system training or R&D activities related to this evaluation, except for that CALLHOME data specifically provided for development, and except for the extended data exploratory R&D task.
- Knowledge of the "*start*" and "*ending*" times that were used to construct the test segments (found in the SPHERE header -- see SPHERE Header Information, below) **is** allowed.
- Knowledge of any information available in the SPHERE header **is** allowed.

8 EVALUATION DATA SET ORGANIZATION

8.1 Last year's evaluation

Please refer to last year's evaluation plan for details.[7]

8.2 The cellular telephone evaluation

The evaluation data set organization of the cellular telephone data will be:

[8] Transcriptions may be downloaded from the ISIP website: http://www.isip.msstate.edu/projects/switchboard/

[9] ASR transcriptions for the SwitchBoard corpus may be accessed from http://www.nist.gov/speech/tests/spkr/2001/extended-data/

[10] Information about the SwitchBoard corpus may be found on the LDC website: http://morph.ldc.upenn.edu/Catalog/LDC93S7.html

[11] Participants are encouraged to do as many tests as possible. However, it is absolutely imperative that results for all of the test segments and target speakers in a test be submitted in order for that test to be considered valid and for the results to be accepted.

[12] This means that the technology is viewed as being "application-ready". Thus a system must be able to perform speaker detection simply by being trained on a specific target speaker and then performing the detection task on whatever speech segment is presented, without the (artificial) knowledge of other test data.

- A single top level directory used as a unique label for the disk: "**sid01cel**"
- Under which there will be three directories "**train**" "**test**" and "**doc**"
- "**train**" and "**test**" will both contain "**male**" and "**female**" subdirectories which in turn will contain the appropriate data
- Training data will be a SPHERE formatted file. The file name will be a four-digit speaker ID with a ".sph" extension.
- Test data will be pseudo random names consisting of four characters followed by a ".sph" extension.
- Each test subdirectory will contain an index file (detect.ndx) that identifies the evaluation trials to be performed.

8.3 The extended data evaluation

The SwitchBoard-I corpus (current version 2 from the LDC) will serve as the primary data set for the extended data evaluation. In addition, NIST will provide a speaker-conversation table and an evaluation control file to support system development and to define the evaluation test. These two files are available via web access.[13]

8.3.1 The speaker-conversation table

The speaker-conversation table is a file that gives the conversation-side filenames for each speaker in the corpus.[14] The format for these records is:

> **speaker** = SPKR-ID, sex = S, **conversation-sides** = {CNV-SIDE}

where:

> SPKR-ID is the speaker identifier,
>
> S is either **M** (for male) for **F** (for female), and
>
> CNV-SIDE is a conversation side identifier. CNV-SIDE is defined to be the identification number of a conversation followed by either A (for the caller) or B (for the callee). For example, "1234A". {CNV-SIDE} is the set of all conversation sides in the SwitchBoard corpus for which speaker SPKR-ID is the speaker, whitespace separated.

8.3.2 The evaluation control file

A single evaluation control file will be used to supervise the evaluation. This file will control the creation of models and define the testing of those models. The structure of the control file will accommodate systems that create a background model in addition to the obligatory target model. This control of background model creation is necessary to ensure unbiased testing, because of the jackknifing of training and test data within the test corpus.

Because of the jackknifing of training and test data, multiple background models will need to be created during the course of the extended data test. Recognizing that background model creation can be the most time consuming part of system development, the evaluation control file will be structured to reduce the number of background models needed.

The evaluation control file will contain records of three different types. The first type will be the background model specification record. Then, for each background model specification there will be one or more target model specification records. Finally, for each target model there will be one or more trial specification records.

The format for the background model specification record is:

> **BM: excluded-speakers** = {SPKR-ID}

where:

> SPKR-ID is a speaker identifier, and {SPKR-ID} is the set of speakers that must be **excluded** from the background model, whitespace separated. (These speakers are those from whom test data will be drawn.)

The format for the target model specification record is:

> **TM:** MODEL-ID **target-sides** = {CNV-SIDE}

where:

> MODEL-ID is a unique model identifier. This is required for the extended data task because there are multiple models for each speaker in this task. Having a unique model ID is therefore needed in order to uniquely associate a particular detection output with the model that produced it.
>
> {CNV-SIDE} is the set of conversation sides (spoken by the target speaker) from which the target model is to be created, whitespace separated. These conversation sides are the **only** data that may be used to create the target speaker model. There will be no more than 30 of these conversation sides per model.

The format for the trial specification record is:

> **test-sides** = {CNV-SIDE}

where:

> {CNV-SIDE} is the set of conversation sides to be used as test segments, whitespace separated, with one trial per test segment. {CNV-SIDE} contains data for both the target speaker and impostors.

The evaluation control file will specify no more than 10 different background models, no more than a total of 5,000 different target models, and no more than an aggregate total of 60,000 different trials. Some cross-sex trials will be included in the evaluation.

9 FORMAT FOR SUBMISSION OF RESULTS

Results for each test must be stored in a single file, according to the formats defined in this section. The file name should be intuitively mnemonic and should be constructed as "SSS_N_TTT", where

- SSS identifies the site,
- N identifies the system, and
- TTT identifies the task (**1sp**, **2sp**, **trk**, **sg2** or **sgn**).

[13] At http://www.nist.gov/speech/tests/spk/2001/extended-data/

[14] It is important to use the information in this table. Although the SwitchBoard corpus (version 2 distributed by LDC) contains tables that convey this information, unfortunately many of the speaker assignments are incorrect. The speaker-conversation table provided here contains the correct assignments.

9.1 One-Speaker Test Results

Sites participating in the one-speaker evaluation tests must report results for whole tests, including all of the test segments. These results must be provided to NIST in a single results file using a standard ASCII format, with one record for each decision. Each record must document its decision with the target identification, test segment identification, and decision information. Each record must contain six fields, separated by white space and in the following order:

1. The sex of the target speaker – **M** or **F** (always **M** for the AHUMADA test)
2. The target model ID[15] (*a four digit number*)
3. The test – (**1** for one-speaker detection, **A** for AHUMADA, **C** for the cellular telephone evaluation, and **E** for the extended training/test evaluation.)
4. The test segment identifier. This is the test segment file name (*excluding directory and file type*) for all of the tasks except the extended data task, in which case it is the conversation-side ID.
5. The decision – **T** or **F** (*is the target speaker judged to be the same as the speaker in the test segment*)
6. The score (*where the more positive the score, the more likely the target speaker*)

9.2 Two-Speaker Test Results

Sites participating in the two-speaker evaluation must report results for the whole test, including all of the test segments. These results must be provided to NIST in a single results file using standard ASCII format, with one record for each decision. Each record must document its decision with target identification, test segment identification, and decision information. Each record must contain six fields, separated by white space and in the following order:

1. The sex of the target speaker - **M** or **F**
2. The target speaker ID (*a four--digit number*)
3. The test - **2** (*for two-speaker*)
4. The test segment file name (*excluding directory and file type*)
5. The decision - **T** or **F** (*whether the target speaker judged to be the same as the speaker in the test segment*)
6. The score (*where the more positive the score, the more likely the target speaker*)

9.3 Speaker Tracking Test Results

Sites participating in the speaker tracking evaluation must report results for the whole test, including all of the test segments. These results must be provided to NIST in a single results file using standard ASCII format, with one SGML (Standard Generalized Markup Language)-tagged data set for each test segment/target speaker hypothesis. Each data set will contain all tracking results for a test segment/target speaker combination.

Each tracking results data set will contain data for all decision intervals within a test segment. These data are namely the interval's start time, decision and score. (Intervals must be contiguous, so that the end time of an interval is implicitly specified by the start time of the following interval.) The format of the results for a single test segment/target speaker combination is defined to be:

<**track segment** = SEGMENT_NAME
target = TARGET_ID>

TIME DECISION SCORE

TIME DECISION SCORE

…

</**track**>

where:

<track …> Identifies the beginning of tracking results for the segment.

SEGMENT NAME: The test segment file name (*four alphanumeric characters.*)

TARGET ID The target speaker ID (*a four-digit number*).

TIME: The starting time of the decision interval, in seconds.

DECISION: The decision (T or F) applied to the interval.

SCORE: The score for the decision interval.

</**track**> Identifies the end of tracking results for the segment.

Due to the theoretically unlimited size of the results file for the tracking task, a practical limit will be imposed on the size of a single results file. All results files must be less than 100MB in size, uncompressed.

9.4 Segmentation Test Results

Sites participating in the segmentation evaluation must report results for a whole test, either the condition where it is known there are exactly 2 speakers in the test segment (SG2) or for the condition where it is unknown how many speakers are in the test segment (SGN). Each of these results must be provided to NIST in a single results file using a standard ASCII format. This file should be a concatenation of all segment records. A segment record should be created as follows:

<**segment filename**=SEGMENT_NAME>

START_TIME END_TIME SPEAKER_ID

START_TIME END_TIME SPEAKER_ID

…

</**segment**>

where:

<segment …> Identifies the beginning of segmentation record.

SEGMENT NAME: The test segment file name (*four alphanumeric characters.*)

START_TIME: The starting interval time (*to the hundredth of a second*).

END_TIME: The ending interval time (*to the hundredth of a second*).

SPEAKER_ID: The speaker cluster this segment belongs to [0-9].

</segment> Identifies the end of a segmentation record.

[15] The target model ID is simply the speaker ID, except for the extended data task. For the extended data task, detection trials are performed for multiple models for each speaker, and therefore a target model ID is required to uniquely identify the trials.

There will be no more than 10 unique speakers per test segment. Each segment record should make use of the ten digits 0-9 to represent a speaker cluster, beginning with 0 and incrementing by 1 for each new speaker.

Due to the theoretically unlimited size of the results file for the segmentation task, a practical limit will be imposed on the size of a single results file. All results files must be less than 100MB in size, uncompressed.

10 SYSTEM DESCRIPTION

A brief description of the system(s) (the algorithms) used to produce the results must be submitted along with the results, for each system evaluated. It is permissible for a single site to submit multiple systems for evaluation for a particular test. In this case, however, the submitting site must identify one system as the "primary" system for the test prior to performing the evaluation.

Sites must report the CPU execution time that was required to process the test data, as if the test were run on a single CPU. Sites must also describe the CPU and the amount of memory used.

11 SCHEDULE

The deadline for signing up to participate in the evaluation is 24 February 2001.

The evaluation data set CD-ROM's will be distributed by NIST on 12 March 2001.

The deadline for submission of evaluation results to NIST is 9 April 2001.

Room reservations for the follow-up workshop (see below) must be received by (to be determined) May 2001.

The follow-up workshop will be held on 14-15 May 2001 at a location yet to be determined. Those participating in the evaluation are expected to present and discuss their findings at the workshop.

12 GLOSSARY

Trial – The individual evaluation unit for each task involving a test segment and (except for segmentation) a hypothesized speaker.

Target (true speaker) trial – A trial in which the actual speaker of the test segment *is in fact* the target (hypothesized) speaker of the test segment.

Non-target (impostor) trial – A trial in which the actual speaker of the test segment *is in fact not* the target (hypothesized) speaker of the test segment.

Target (model) speaker – The hypothesized speaker of a test segment, one for whom a model has been created from training data.

Non-target (impostor) speaker – A hypothesized speaker of a test segment who is in fact not the actual speaker.

Segment speaker – The actual speaker in a test segment.

One-session training – Training data for a target speaker consisting of speech extracted from a single conversation.

Two-session training – Training data for a target speaker consisting of speech extracted from two different conversations.

Turn – The interval during a conversation during when one participant speaks while the other remains silent.