# The NIST Year 2002 Speaker Recognition Evaluation Plan

## 1 INTRODUCTION

The year 2002 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 four 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 and speaker segmentation. These tasks are posed primarily 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 2002 speaker recognition evaluation plan includes the following two tasks[2]:

#### 2.1.1 Speaker detection

This task is NIST's basic speaker recognition task. The task is to determine whether a specified speaker is speaking during a given speech segment.[3]

#### 2.1.2 Speaker segmentation

This task requires identifying the time intervals during which unknown speakers are each speaking in a conversational speech segment. It combines the tasks of speech detection and speaker recognition. No prior knowledge or training data for these speakers is provided, however. Also, the number of different speakers that speak during each segment to be segmented is not specified.

### 2.2 Task Conditions

The year 2002 speaker recognition evaluation plan includes four distinct task conditions for the speaker detection task and one condition for the speaker segmentation task. These are as follows:

#### 2.2.1 One-speaker detection – cellular data

The one-speaker detection task conditions will remain essentially the same as in previous years. The data this year will be taken from the second release of LDC's cellular switchboard corpus (Switchboard Cellular – Part 2).[4] The training data for a target speaker will be two minutes of speech from that speaker, excerpted from a single conversation. Each test segment will be the speech of a single speaker, excerpted from a one-minute segment taken from a (single) conversation. (There will also be some trials using two-speaker training data and one-speaker test segments. See section 4.1.2.)

#### 2.2.2 Two-speaker detection – cellular data

The two-speaker detection task conditions differ from the one-speaker conditions in that no excerpting of the speech of a single speaker is performed, neither in training nor in test. The data will be taken from the second release of LDC's cellular corpus, the same as for the one-speaker detection test. Thus each test segment will be a one-minute segment taken from a (single) conversation, but no excerpting will be performed and the two sides of the conversation will be summed together. Further, the training data for a target speaker will be three whole conversations, with the two sides of the conversation summed together. Thus a major challenge in training will be to discover which of the two speakers in each of the training conversations is the target speaker.[5] (There will also be some trials using one-speaker training data and two-speaker test segments. See section 4.1.2.)

#### 2.2.3 One-speaker detection – extended data

This task condition provides a much larger amount of training data for target speakers – up to an hour of speech per speaker. The intent is to foster new research on improving speaker recognition performance through the discovery and exploitation of higher-level and more complex characteristics of a speaker's speech, such as

---

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

[2] The speaker tracking task that was supported in the NIST 2000 and 2001 speaker recognition evaluations has been eliminated. Factors that led to this task being eliminated include a judgment that there seemed to be less interesting research being performed under this task and less clear application motivation for supporting the task.

[3] In previous evaluation plans, the speaker detection task was divided into a "one-speaker" and a "two-speaker" task. However, this distinction relates to the task conditions rather than the task definition. Therefore in this evaluation plan the one- and two-speaker conditions have been moved to section 2.2, task conditions.

[4] Refer to www.ldc.upenn.edu/Projects/SWB/cellular.

[5] The non-target speakers appearing in a speaker's training data will be controlled so that no non-target speaker appears in more than one training conversation.

idiosyncratic language patterns and nonlinguistic vocalizations. The data will be taken from the LDC's switchboard II corpus, phases 2 and 3. The training data will be all of a target speaker's speech from N whole conversations, with N varying between 1 and 16. The test data will be all of one side of a single conversation. (not both sides summed together)

### 2.2.4 One-speaker detection – multi-modal data

The multi-modal speaker detection condition is a limited simulation of forensic conditions using the FBI Voice Database[6]. This condition will provide a measure of performance when the training and test data are recorded using different input devices and/or channels.

### 2.2.5 Speaker segmentation – various data sources

The speaker segmentation test will involve test segments taken from a variety of sources. These sources will include telephone conversations, broadcast news recordings, and recordings of meetings. The number of speakers will not be specified.

## 3 PERFORMANCE MEASURES

Evaluation will be performed separately for each of the tasks and task conditions in section 2. Evaluation of all these tasks and task conditions will use cost-based performance measures. A cost-based performance measure is used so that the various (application) factors may be weighed and integrated into a single numerical measure of performance. The cost measure of performance is a weighted probabilistic sum of cost over all error conditions. The probabilities involved are both the (application-dependent) probabilities of the various conditions and the (system-dependent) probabilities of error given these conditions. The cost of an error is assumed to be a function of the condition. So, in general, we have:

$$Cost = \sum_{\text{all Cnd}} \{ C_{\text{Error|Cnd}} \times P_{\text{Error|Cnd}} \times P_{\text{Cnd}} \}$$

where

$C_{\text{Error|Cnd}}$ = the cost of an error for condition = Cnd

$P_{\text{Error|Cnd}}$ = the (system) prob. of an error for condition = Cnd

$P_{\text{Cnd}}$ = the (prior) probability of condition Cnd

There will be two basic cost models – one for measuring speaker detection performance and one for measuring speaker segmentation performance.

## 3.1 Speaker Detection Performance

### 3.1.1 The basic speaker detection cost model

The performance measure to be used for all speaker detection tests is the detection cost function, defined as a weighted sum of miss and false alarm error probabilities:

$$C_{\text{Det}} = C_{\text{Miss}} \times P_{\text{Miss|Target}} \times P_{\text{Target}} + C_{\text{FalseAlarm}} \times P_{\text{FalseAlarm|NonTarget}} \times (1-P_{\text{Target}})$$

The parameters of this cost function are the relative costs of detection errors, $C_{\text{Miss}}$ and $C_{\text{FalseAlarm}}$, and the *a priori* probability of the specified target speaker, $P_{\text{Target}}$. The parameter values in Table 1 will be used as the primary evaluation of speaker recognition performance for all speaker detection tasks.

Table 1 Speaker Detection Cost Model Parameters for the primary evaluation decision strategy

| $C_{\text{Miss}}$ | $C_{\text{FalseAlarm}}$ | $P_{\text{Target}}$ |
|---|---|---|
| 10 | 1 | 0.01 |

### 3.1.2 Normalization of speaker detection cost

One of the advantages of using a cost model is that it can be easily applied to different applications simply by changing the model parameters. On the other hand, a potential disadvantage of using cost as a performance measure is that it gives values that often lack intuitive meaning. To improve the intuitive value of the cost measure, we normalize the cost by dividing by $C_{\text{Default}}$, which is defined to be the best cost that could be obtained without processing the input data (i.e., by always making the same decision, namely either to accept or to reject the segment speaker as being the target speaker, whichever gives the lowest cost):

$$C_{\text{Default}} = \min\{ C_{\text{Miss}} \times P_{\text{Target}}, \quad C_{\text{FalseAlarm}} \times P_{\text{NonTarget}} \}$$

and

$$C_{\text{Norm}} = C_{\text{Det}} / C_{\text{Default}}$$

This default normalizing cost represents zero value, because this is the cost for a system that provides no information. The range of values for the normalized cost is:

$$C_{\text{Norm}} \in \{ 0, \ 1 + \max(C_{\text{Ratio}}, C_{\text{Ratio}}^{-1}) \}$$

where

$$C_{\text{Ratio}} = [C_{\text{Miss}} \times P_{\text{Target}}] / [C_{\text{FalseAlarm}} \times (1 - P_{\text{Target}})]$$

For those with more sanguine predilections, it may be desirable to define a normalized system "value":

$$V_{\text{Norm}} = 1 - C_{\text{Norm}}$$

Thus a value of $V_{\text{Norm}} = 1$ represents a "perfect" system – one that incurs no cost (and thus provides maximum value), whereas a value of $V_{\text{Norm}} = 0$ represents a "worthless" system that provides no value. Note that it is possible for $V_{\text{Norm}}$ to be less than 0, but we needn't dwell on that.

### 3.1.3 Speaker detection cost with "no-decision"

In some applications, it may be better that a speaker recognition system not make a decision when the system cannot be confident of that decision. This is particularly relevant for the multi-modal task condition, because this task condition is intended to address forensic applications. Thus, an alternative detection cost model will be defined to accommodate a "no-decision" detection strategy for the "one-speaker detection – multi-modal data task condition".

Clearly, a detection error occurs whenever the system misses the target or falsely detects the target. But a detection error is also deemed to occur whenever a system declines to make a decision. So it is better to make a (correct) decision than to make no decision at all. However, it is also the case that no decision is considered to be better than a wrong decision, especially in forensic applications, and especially when the decision in error is a positive detection.

[6] "Forensic Automatic Speaker Recognition," Hirotaka Nakasone and Steven D. Beck, Odyssey 2001 Workshop, Chania, Crete, Greece, June 2001

The optimum decision is an extension of that for the simple yes/no detection model where the target is declared whenever the expected cost of a miss is less than the expected cost of a false alarm. In this case, the system should declare a non-target, a target, or make no decision, according to whichever expected cost is lowest – the cost of a miss, a false alarm, or no decision, respectively:

$$E[Cost \mid Miss] = C_{Miss} \times \Pr(\text{Target} \mid score)$$

$$E[Cost \mid FalseAlarm] = C_{FalseAlarm} \times (1 - \Pr(\text{Target} \mid score))$$

$$E[Cost \mid NoDecision] = C_{NoDecision|Target} \times \Pr(\text{Target} \mid score) + C_{NoDecision|NonTarget} \times (1 - \Pr(\text{Target} \mid score))$$

where Pr(Target|*score*) is the (system-judged) probability of the target speaker for a given score (i.e., as a function of score). This probability is the source of the system output decision and takes into consideration the prior probabilities in addition to the speaker recognition information in the speech signal. In terms of the speaker model probability distribution and the prior probability, Pr(Target|*score*) may be expressed as:

$$\Pr(\text{Target} \mid score) = \left[1 + \left(\frac{\Pr(score \mid \text{NonTarget})}{\Pr(score \mid \text{Target})}\right)\left(\frac{1 - P_{Target}}{P_{Target}}\right)\right]^{-1}$$

Pr(Target|*score*) is sometimes called the "confidence", that is, the confidence in making a target decision, as a function of score. This "confidence" measure will be a required output measure for the one-speaker detection – multi-modal task condition.

The parameter values in Table 2 will be used for the alternative decision strategy for the one-speaker detection – multi-modal data condition. $C_{NoDecision|Target}$ is chosen to be 12.5 percent of $C_{FalseAlarm}$, which means that a system should make no positive detection decision unless it is at least 87.5 percent confident that the speaker is the target speaker. Likewise, a system should make no negative detection decision unless it is at least 75 percent confident that the speaker is not the target speaker. (Note that the prior probability of a target plays a major role in determining the confidence of a decision.)

Table 2 Speaker Detection Cost Model Parameters for the "no-decision" alternative decision strategy

| $C_{Miss}$ | $C_{FalseAlarm}$ | $C_{NoDecision\|Target}$ | $C_{NoDecision\|NonTarget}$ | $P_{Target}$ |
|---|---|---|---|---|
| 1 | 2 | 0.25 | 0.25 | 0.5 |

## 3.2 Speaker Segmentation Performance

The performance measure to be used for the speaker segmentation task is the segmentation cost function, defined as a weighted sum of decision errors, weighted by error type and integrated over error duration. The situation for speaker segmentation is more complex than for speaker detection, since the detection task is combined with a recognition task. For speaker segmentation there are five kinds of errors that can occur, all as a function of time. They are:

- Missing a segment of speech (a speaker) when speech is present
- Falsely declaring a segment of speech (a speaker) when there is no speech.
- Assigning a spurious (false alarm) speaker to a segment of speech.
- Assigning a speaker to a segment of speech of an undetected (missed) speaker.
- Assigning an incorrect speaker to a segment of speech.

The speaker segmentation cost function is therefore defined as:

$$C_{Seg} = C_{MissSeg} \times P_{MissSeg} + C_{FASeg} \times P_{FASeg} + C_{MissSpkr} \times P_{MissSpkr} + C_{FASpkr} \times P_{FASpkr} + C_{ErrSpkr} \times P_{ErrSpkr}$$

where the various error probabilities are all prorated durations, i.e., durations that are normalized by (divided by) the total duration of the evaluation segment.

In order to tabulate these errors, and since there is no predefined speaker set, the set of speakers that the speaker segmentation system defines must be reconciled with the set of speakers that the answer key is based on. This reconciliation is performed by finding and using that assignment of speakers that minimizes the speaker segmentation cost function. This "maximally felicitous" mapping preserves speaker integrity, meaning that each system speaker is mapped to at most one reference speaker, and conversely each reference speaker is mapped to at most one system speaker. This results, in general, in unmapped reference speakers (missed speakers) or unmapped system speakers (false alarm speakers), generally depending on whether the system declares fewer or more speakers than the reference. The parameter values in Table 3 will be used as the primary evaluation of speaker segmentation performance.[7]

Table 3 Speaker Segmentation Cost Model Parameters

| $C_{MissSeg}$ | $C_{FASeg}$ | $C_{MissSpkr}$ | $C_{FASpkr}$ | $C_{ErrSpkr}$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |

### 3.2.1 Normalization of speaker segmentation cost

The speaker segmentation cost function will be normalized in the same spirit as the speaker detection cost function. $C_{SegDefault}$ is thus defined to be the best cost that could obtain without processing the input data. Since the regions of actual speech are provided, this is defined by always hypothesizing, for each segment, a single speaker speaking wherever speech is present. Then let

$$C_{SegNorm} = C_{Seg} / C_{SegDefault}$$

# 4 EVALUATION CONDITIONS

## 4.1 Evaluation of Speaker Detection

Speaker detection performance will be evaluated in terms of the detection cost function. The cost function will be computed over an ensemble of speech segments selected to represent a statistical sampling of conditions of 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"

[7] For the 2002 evaluation, speech segmentation will be provided. This makes it possible for a system to manipulate its speaker segmentation output so as to eliminate speech segment miss and false alarm errors.

(or "no decision" for the multimodal data task condition), and the correctness of these decisions will be tallied.[8]

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.[9]

As discussed in section 3.1.3, a confidence score, Pr(Target|*score*), is required when no-decisions are an option, as they will be with the multi-modal data. The confidence score will be optional, and not used for scoring purposes, in the detection conditions not involving the multi-modal data.

### 4.1.1 Handset Labels

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, however, handset labels will be provided only for the extended data task condition.[10]

### 4.1.2 Cross-Condition Cellular Training Data

There are two task conditions that will be evaluated using cellular data – one-speaker detection and two-speaker detection. These two task conditions are different in both training and test. Therefore, to better understand performance characteristics and to help attribute differences in performance correctly to training versus test, there will be included some one-speaker test segment trials that use two-speaker models, and some two-speaker test segment trials that use one-speaker models.

### 4.1.3 One-Speaker Detection – cellular data

#### 4.1.3.1 Training Data

Training data for each speaker will consist of about two minutes of speech from a single conversation. The actual duration of the training data 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.

#### 4.1.3.2 Test Data

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 trials for cellular data will include both "same number" and "different number" tests. Evaluation trials may also include some cross-sex conversations, in which the model speaker and test speaker are of opposite sexes.

The primary evaluation conditions are:

1. "different number" tests *(unless there is an insufficient quantity of "different number" tests).*
2. The model is a one-speaker model.
3. The speech duration is between 15 and 45 seconds.
4. Both speakers are of the same sex.

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

### 4.1.4 Two-Speaker Detection – cellular data

#### 4.1.4.1 Training Data

Three whole conversations (minus some introductory comments) will be used for training. In contrast with the one-speaker training condition, however, the two sides of each conversation will be summed together, and both the model speaker and that speaker's conversation partner will be represented in this conversation. Thus the challenge is to separate the speech of the two speakers and then to decide (correctly) which is the model speaker. To make this challenge feasible, the training conversations will be chosen so that all speakers other than the model speaker are represented in only one conversation. Thus the model speaker, who is represented in all three conversations, is the only speaker to be represented in more than one.

#### 4.1.4.2 Test Data

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. "different number" tests *(unless there is an insufficient quantity of "different number" tests).*
2. The model is a two-speaker model.
3. The speech duration is 15-45 seconds for both speakers.
4. Both speakers are of the same sex.

### 4.1.5 One-Speaker Detection – extended data

This section outlines the conditions for the one-speaker detection task with extended training and test data. The entire SwitchBoard-II corpus phases 2 and 3 will be used for this evaluation. In addition to the acoustical data, automatically generated and time

[8] 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.

[9] 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.

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

marked (ASR) transcriptions will be made available to those who wish to use them.

#### 4.1.5.1 Training Data

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 impostor speakers. This information will be provided in index files that must be used to control the evaluation. Instructions are given in section 8.2.2 for the use of this index file information.

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

#### 4.1.5.2 Test Data

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 (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.

### 4.1.6 One-Speaker Detection – multi-modal data

The task is one-speaker detection. The data is recorded over three types of inputs: body microphone, tabletop microphone and telephone (both internal same handset and external different handsets). The training and test data will be text-independent and come from one of the three input types. Systems will be told the input type for each train and test file. There are different training and test segment lengths and all speakers are male.

The evaluation conditions will be examining performance on same and cross input conditions and be conditioned on training and testing length. The original user's manual and original evaluation test plan for this data are available for more detailed information. This evaluation, however, will involve only the "Level 1" and "Level III" testing of the original evaluation plan, and the data names and organization and submission formats (see sections 8 and 9) will be different.

## 4.2 Evaluation of Speaker Segmentation

Data for the speaker segmentation task will be drawn from a variety of different sources, including telephone conversations, broadcast news, and meetings. The duration of each segment will be approximately one to two minutes. The number of speakers in a test segment will not be given. However, the source type (being "telephone conversations", "broadcast news", or "meetings"), will be given. All speech data will be limited to the English language. Furthermore, time marks will be provided to indicate speech and silence segments.

# 5 DEVELOPMENT DATA

The evaluation data for the different parts of last year's evaluation will serve as the development data for corresponding parts of this year's evaluation. Please refer to last year's evaluation plan for details.[11]

The FBI Voice Database contains a small development data subset, described in the original evaluation test plan. This development data is available on request from NIST for any site planning to participate in the multi-modal portion of this year's evaluation.

# 6 EVALUATION DATA

## 6.1 One-speaker detection – cellular data

The evaluation data will be drawn from the new Switchboard Cellular Corpus, Part 2. All conversations will have been 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 (but note section 4.1.2). Each such segment will be store as an 8-bit mu-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 400 target speakers and about 3500 test segments. Each test segment will be evaluated against 11 hypothesized speakers of the same sex as the segment speaker.

## 6.2 Two-speaker detection – cellular data

The same data will be used as in one-speaker detection with cellular data. The training and test data will have the two sides of the conversation summed together (but note section 4.1.2). The training segments will consist of whole conversation sides. The test segments will each be approximately one minute in duration. The one-speaker test segments (section 6.1) will be concatenated one-speaker excerpts from these segments.

There will be about 400 target speakers and about 1500 test segments. Each test segment will be evaluated against 22 hypothesized speakers.

## 6.3 One-speaker detection – extended data

The Switchboard-II Corpus, Phases 2 and 3, will serve as the evaluation data for the extended data evaluation. Speech recognition output of this data will also be made available. The audio data must be obtained from the Linguistic Data Consortium (LDC)[12], which makes it available for sale to non-members. (Phase 2 is currently available, and Phase 3 will be available early in 2002.)

## 6.4 One-speaker detection – multi-modal data

The data is digitized at 16 KHz. with 16-bit pcm samples. It is further described in the original user's manual and the original evaluation test plan documents noted previously.

---

[11] The year 2001 speaker recognition evaluation plan may be accessed from http://www.nist.gov/speech/tests/spk/2001/doc/

[12] Corpus information may be found on the LDC website: http://www.ldc.upenn.edu/Catalog/by_type.html - speech.telephone

## 6.5 Speaker segmentation – various data sources

The number of test segments will be approximately 600, no more than 200 from each data source. They will all be in English and have a duration of one to two minutes. The number of speakers in each segment will vary and will not the specified. The telephone conversation segments will be drawn from the various parts of Switchboard (land-line and cellular) and/or from the LDC's CallHome and CallFriend Corpora. The broadcast news segments will be drawn from the various Broadcast News Corpora collected by the LDC. The meeting segments will be drawn from meetings collected by NIST in its Meeting Data Collection Laboratory. The broadcast news and meeting segments will have a 16 KHz. sampling rate with 16-bit pcm samples.

# 7 EVALUATION RULES

In order to participate in the 2002 speaker recognition evaluation, a site must complete, in its entirety, at least one complete evaluation of one of the four evaluation task conditions.[13]

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.[14] 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 manually produced transcripts or other information for training is **not** allowed.
- 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, of course).
- For the segmentation task, knowledge of the number of speakers present is **not** allowed, except as determined by automatic means.
- For the segmentation task, knowledge of the source type ("telephone conversations", "broadcast news", or "meetings"), **is** allowed, and will be supplied.
- 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.
- 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 One-speaker detection – cellular data

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

- A single top level directory used as a unique label for the disk: "**sid01clN**" where N is a digit identifying the disc.
- 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 (detectN.ndx) that identifies the evaluation trials to be performed, where "N" is a digit. (This will provide different index names even if the male test data is spread out across more than 1 disc.)

## 8.2 Two-speaker detection – cellular data

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

- A single top level directory used as a unique label for the disk: "**sid02clN**" where N is a digit identifying the disc.
- Under which there will be three directories "**train**" "**test**" and "**doc**"
- The "**train**" directory will contain all the speech data to be used for training the various models. There will also be two lists ("m_train.lst" and a "f_train.lst"). Each list will contain one record per line, with a record consisting of a model id followed by three sphere waveforms to be used to create the model. Each field in the training lists will be separated by white space.
- The "**test**" will contain the two-speaker evaluation test data and will be pseudo random names consisting of four characters followed by a ".sph" extension.
- The **test** subdirectory will contain an index file (detectN.ndx) that identifies the evaluation trials to be performed.

## 8.3 One-speaker detection – extended data

The SwitchBoard-II phase 2 corpus 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

[13] 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.

[14] 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.

support system development and to define the evaluation test. These two files are available via web access.[15]

### 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.[16] 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.

## 8.4 One-speaker detection – multi-modal data

The evaluation data set organization of the multi-modal data will be:

- A single top level directory used as a unique label for each of the five multi-modal evaluation discs: "**sid02mm1**", "**sid02mm2**", "**sid02mm3**", "**sid02mm4**" and "**sid02mm5**".
- Under which there will be three directories "**train**" (on the first two discs only), "**test**" (on discs two through five), and "**doc**" (same data on all five discs).
- The "**train**" directory will contain one sub-directory for each of the 388 models to be created. The names of these subdirectories will be a four-digit ID followed by a letter representing the input device, followed by a two-digit number. (Sample: 1313M02) In each of these "model" subdirectories there will be either one or four SPHERE formatted files to be used to train that model.
- The "**test**" directory will contain the test data in SPHERE formatted files. The naming convention will be: fv1_XXXX.sph, where XXXX is a four-digit number. There will be one index file (**detectX.ndx,** *where X corresponds to the disc*) that identifies the evaluation trials to be performed. The first field of the index file will be the *test-segment*, followed by a *list of model*s to be evaluated.
- The "**doc**" directory will contain three text files.
  - "**trainlst.txt**" will contain two colon separated fields. The first field will identify the model and the second will be a comma-separated list of the SPHERE formatted files available for training the model.
  - "**trnlabel.txt**" will contain two colon separated fields. The first field will identify the model

[15] At http://www.nist.gov/speech/tests/spk/2002/extended-data/

[16] It is mandatory to use the information in this table.

and the second will identify the input device. Possible input devices include:

- **M** – Microphone
- **B** – Body microphone
- **T** – Telephone.

- "**tstlabel.txt**" will contain two colon separated fields. The first field will identify the test segment and the second will identify the input device (M, B or T).

## 8.5 Speaker Segmentation – various data sources

The speaker segmentation task will make use of data from various sources, including Broadcast News, telephone conversations, and meetings. More detailed information will be included in future releases of this evaluation plan. This information will also be made available from the 2002 NIST Speaker Recognition evaluation website: http://www.nist.gov/speech/tests/spk/2002/

# 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**, or **seg**).

## 9.1 Speaker Detection 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 seven fields[17], separated by white space and in the following order:

1. The sex of the target speaker – **M** or **F**
2. The target model ID[18] (*a four digit number*)
3. The test – (**1C** for one-speaker detection – cellular data, **2C** for two-speaker detection – cellular data, **1E** for one-speaker detection – extended data, **1M** for one-speaker detection – multi-modal data.)
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*)
7. The confidence of the target speaker, as defined in section 3.1.3.[19] Note this confidence score is required for the multi-model condition of one speaker detection.

## 9.2 Segmentation Test Results

Sites participating in the segmentation evaluation must report results for the whole test for each system tested. 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].

 Identifies the end of a segmentation record.

Evaluation for the three sources of data will be performed separately, but since each site is required to process all three data sources, the system output should be submitted in one file.

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.

---

[17] The seventh field is optional except for the multi-modal data condition.

[18] 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.

[19] The confidence of the target speaker is required in order to allow NIST to evaluate performance for different application parameters for decision strategies that include a no-decision option.

## 11 SCHEDULE

The deadline for signing up to participate in the evaluation is March 1, 2002.

The evaluation data set CD-ROM's will be distributed by NIST on March 18, 2002.

The deadline for submission of evaluation results to NIST is April 15, 2002.

Room reservations for the follow-up workshop must be received by (a date to be determined).

The follow-up workshop will be held on May 20-21 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.