The NIST Year 2004 Speaker Recognition Evaluation Plan

1 INTRODUCTION

The year 2004 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 is designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to those wishing to participate.

The 2004 evaluation will use new conversational speech data collected in the Mixer Project using the Linguistic Data Consortium's new "Fishboard" platform.[1] This data will be mostly conversational telephone speech in English as in previous evaluations, but it is expected to include some speech in languages other than English and may include some microphone data. The evaluation will include twenty-eight different speaker detection tests defined by the duration and type of both the training and the test segments of the individual trials of which these tests are composed. For each such test, an unsupervised adaptation mode will be offered in addition to the basic test.

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

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.[2]

2 TECHNICAL OBJECTIVE

This evaluation focuses on speaker detection, 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 Definition

The year 2004 speaker recognition evaluation is limited to the broadly defined task of **speaker detection**. This has been NIST's basic speaker recognition task over the past eight years. The task is to determine whether a specified speaker is speaking during a given speech segment.[3]

2.2 Task Conditions

Previous evaluations have included both a limited data condition and an extended data condition. Limited data meant that the training and test segment data for each trial consisted of two minutes or less of concatenated segments of speech data, with silence intervals removed, while extended data meant that each of these consisted of an entire conversation side or, for training, multiple conversation sides. It has been decided this year to remove the specific distinction between limited and extended data tests, and to no longer do silence removal, but to offer multiple testing conditions involving the amount and type of data available for both the training and the test segments.

Thus the speaker detection task for 2004 includes tests involving seven distinct training conditions and four distinct (test) segment conditions. There will thus be 28 different combinations of training/segment conditions. A test (sequence of trials) will be offered for each of these combinations. One of these (see section 2.2.3) is designated the core test. Participants must do the core test and may choose to do any subset of the remaining tests. Results must be submitted for all trials included in each test for which any results are submitted. For each test, there will also be an optional unsupervised adaptation condition. A site may do the adaptation condition for a particular test only if it also does the particular test without adaptation.

2.2.1 Training Conditions

The training segments in the 2004 evaluation will be continuous conversational excerpts. Unlike in past years, there will be no prior removal of intervals of silence. For some training conditions the NIST energy-based automatic speech detector will be used to estimate the duration of actual speech in the chosen excerpts.

The seven training conditions to be included involve target speakers defined by the following training data:

1. An excerpt from a single channel conversation side estimated to contain approximately 10 seconds of speech
2. An excerpt from a single channel conversation side estimated to contain approximately 30 seconds of speech
3. A single channel conversation side, of approximately five minutes total duration[4]

---

[1] See http://www.upenn.edu/mixer/

[2] To contact Dr. Martin, send him email at alvin.martin@nist.gov, or call him at 301/975-3169. Each site must return a signed registration form to complete the registration process: http://www.nist.gov/speech/tests/spk/2004/register.pdf

[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, both for training and for test segments, are included under task conditions in section 2.2.

[4] Each conversation side will consist of the last five minutes of a six-minute conversation. This will eliminate from the evaluation data the less-topical introductory dialogue, which is more likely to contain identifying information about the speakers.

4. Three single channel conversation sides involving the same speaker
5. Eight single channel conversation sides involving the same speaker
6. Sixteen single channel conversation sides involving the same speaker
7. Three summed-channel conversations, formed by sample-by-sample summing of the two sides of actual conversations, each including a common speaker (the target of interest) and a second speaker not participating in the other two conversations

Word transcripts derived from an automatic speech recognition (ASR) system[5] will be provided for all English training segments of each condition. These transcripts will, of course, be errorful, perhaps with word error rates in the 20-30% range.

2.2.2 Test Segment Conditions

The test segments in the 2004 evaluation will be continuous conversational excerpts. Unlike in past years, there will be no prior removal of intervals of silence. For some test segment conditions the NIST automatic speech detector will be used to estimate the duration of actual speech in the chosen excerpts.

The four test segment conditions to be included are the following:

1. An excerpt from a single channel conversation side estimated to contain approximately 10 seconds of speech
2. An excerpt from a single channel conversation side estimated to contain approximately 30 seconds of speech
3. A single channel conversation side, of approximately five minutes total duration[4]
4. A summed channel conversation, formed by sample-by-sample summing of the two sides of an actual conversation

Errorful ASR word transcripts derived from an ASR system will be provided for all English-language test segments of each condition.

2.2.3 Training/Segment Condition Combinations

The matrix of training and segment condition combinations is shown in Table 1. A test will be offered for each combination. Each test consists of a sequence of trials, where each trial consists of a target speaker, defined by the training data provided, and a test segment. The system must decide whether speech of the target speaker occurs in the test segment. Both an actual decision ('T' or 'F') and a likelihood score, indicating system confidence in the correctness of this decision, must be submitted for each trial.

The shaded box in Table 1 corresponds to the condition of a single conversation side as training and a single conversation side as test segment. The test for this condition will be defined as the **core test** for the 2004 evaluation. All participants are required to submit results for this core test. Each participant may choose to also submit results for all, some, or none of the other 27 test conditions. For each test for which results are submitted, they must be submitted for all trials included in the test.

[5] All conversations will be processed at BBN using a system derived from their RT-03 conversational telephone speech STT evaluation system.

Table 1: Matrix of training and test segment conditions. The shaded entry is the required core test condition.

		Test Segment Condition			
		10 sec	30 sec	1 side	1 conv
Training Condition	10 sec	X	X	X	X
	30 sec	X	X	X	X
	1 side	X	X	X	X
	3 sides	X	X	X	X
	8 sides	X	X	X	X
	16 sides	X	X	X	X
	3 convs	X	X	X	X

2.2.4 Unsupervised Adaptation Mode

In previous evaluations, adaptive strategies were not allowed and each trial was restricted to use data from a single test segment and a single (static) model. This year, an unsupervised adaptation mode will be supported, allowing models to be updated based on test segments processed in previous trials.

As in previous evaluations, for each trial systems may not in general use information about other evaluation target speakers or other test segments. In unsupervised adaptation mode, however, the trials for each target speaker model must be processed in order, and after each trial the model may optionally be updated based on the test segments in the preceding trials. How this is accomplished should be discussed in the system descriptions (see section 10).

Thus for each test, participants will have the option of doing the test in unsupervised adaptation mode. Unsupervised adaptation results may only be submitted for tests for which (standard) non-adaptive results are also submitted, and the performance results with and without such adaptation will be compared.

3 PERFORMANCE MEASURE

There will be a single basic cost model for measuring speaker detection performance, to be used for all speaker detection tests. This detection cost function is defined as a weighted sum of miss and false alarm error probabilities:

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

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 specified target speaker, P_{Target}. The parameter values in **Table** 2 will be used as the primary evaluation of speaker recognition performance for all speaker detection tests.

Table 2: Speaker Detection Cost Model Parameters for the primary evaluation decision strategy

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

3.1 Normalization

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 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):

$$C_{Default} = \min\left\{\begin{matrix} C_{Miss} \times P_{Target}, \\ C_{FalseAlarm} \times (1 - P_{Target}) \end{matrix}\right\}$$

and

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

4 EVALUATION CONDITIONS

Speaker detection performance will be evaluated in terms of the detection cost function. For each test, the cost function will be computed over the sequence of trials provided and over subsets of these trials of particular evaluation interest. Each trial must be independently judged as "true" (the model speaker speaks in the test segment) or "false" (the model speaker does not speak in the test segment), and the correctness of these decisions will be tallied.[6]

In addition to the actual detection decision, a decision (likelihood) score will also be required for each test hypothesis. Higher scores will be taken to indicate greater confidence that "true" is the correct decision and lesser confidence that "false" is the correct decision. 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.[7]

4.1 Training Data

As discussed in section 2.2.1, there will be seven training conditions. NIST will be interested in examining how performance varies among these conditions for fixed test segment conditions.

Most of the training data will be in English, but some training conversations involving bi-lingual speakers may be collected in Arabic, Mandarin, Russian, and Spanish. Thus it will then be possible to examine how performance is affected by whether or not the training language matches the language, generally English, of the test data. For the training conditions involving multiple conversations, the effect of having a mix of languages in the training may also be examined. The language used in all training data files will be indicated in the file header and available for use.

All training data is expected to be collected over telephone channels.

The sex of each target speaker will be provided to systems.

For all training conditions, errorful ASR transcriptions of all English language data will be provided along with the audio data. Systems may utilize the data provided as they wish. The acoustic data may be used alone, the transcriptions may be used alone, or all data may be used in combination.[8]

4.1.1 Single Channel Excerpts

As discussed in section 2.2.1, there will be training conditions consisting of excerpts with approximately 10-seconds and approximately 30-seconds of estimated speech duration. These estimated durations will vary so that the excerpts may include only whole turns whenever possible, but they will be constrained to lie in the ranges of 8-12 seconds for "10-second" excerpts, and 25-35 seconds for "30-second" excerpts.

4.1.2 Single Channel Conversation Sides

As discussed in section 2.2.1, there will be training conditions consisting of one, three, eight, or sixteen single conversation sides of a given speaker. These sides will consist of approximately five minutes from an original six minute conversation side, with an initial segment of around one minute excised. The excision point will be chosen so as not to include a partial speech turn. Areas of silence within the five minutes of conversation chosen will not be excised.

4.1.3 Summed Channel Conversations

As discussed in section 2.2.1, the final training condition will consist of three whole conversations, minus initial segments of about a minute each. In contrast with the other training conditions, however, the two sides of each conversation, in which both the target speaker of interest and another speaker participate, will be summed together. Thus the challenge is to distinguish speech by the intended target speaker from speech by other participating speakers. To make this challenge feasible, the training conversations will be chosen so that each non-target speaker participates in only one conversation, while the target speaker participates in all three.

The difficulty of finding the target speaker's speech in the training data is affected by whether the other speaker in a training conversation is of the same or of the opposite sex as the target. Systems will not be provided with this information, but may use automatic gender detection techniques if they wish. Performance results will be examined as a function of how many of the three training conversations contain same-sex other speakers.

Note that an interesting contrast will exist between this training condition and that consisting of three single conversation sides.

[6] 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.

[7] Decision scores for all of the trials in a given test will be pooled before plotting detection error tradeoff curves. Thus it is necessary to normalize scores across speakers to achieve satisfactory detection performance.

[8] Note, however, that there will be some non-English training data, for which no meaningful ASR transcripts will be available.

4.2 Test data

As discussed in section 2.2.2, there will be four test segment conditions. NIST will be interested in examining how performance varies among these conditions for fixed training conditions.

For a limited number of speakers some test conversations may be collected using non-telephone channels. Several microphone types will be included in this collection. Thus it will be possible to examine how performance is affected by whether or not test data is recorded over a telephone channel, and by the type of microphone used in non-telephone test data. The non-telephone data will include some or all of the following microphone types:

- Ear-bud/lapel mike
- Miniboom mike
- Courtroom mike
- Conference room mike
- Distant mike
- Near-field mike
- PC stand mike
- Microcassette mike

Information on the microphone type used in each non-telephone test segment data will be available to recognition systems.

With rare exceptions, all test data speech is expected to be in English.

For all test segments conditions, errorful ASR transcriptions of the (English language) data will be provided along with the audio data. Systems may utilize the data provided as they wish. The acoustic data may be used alone, the ASR transcriptions may be used alone, or all data may be used in combination.

4.2.1 Single Channel Excerpts

As discussed in section 2.2.2, there will be test segment conditions consisting of excerpts with approximately 10-seconds and approximately 30-seconds of estimated speech duration These estimated durations will vary so that the excerpts may include only whole turns whenever possible, but they will be constrained to lie in the ranges of 8-12 seconds for "10-second" excerpts, and 25-35 seconds for "30-second" excerpts.

4.2.2 Single Channel Conversation Sides

As discussed in section 2.2.2, there will be a test segment condition consisting of a single conversation side of a given speaker. Each such side will consist of approximately five minutes from an original six minute conversation side, with an initial segment of around one minute excised. The excision point will be chosen so as not to include a partial speech turn. Areas of silence within the five minutes of conversation chosen will not be excised.

4.2.3 Summed Channel Conversations

As discussed in section 2.2.2, the there will be a test segment condition consisting of a single whole conversation, minus an initial segment of about a minute. In contrast with the other test segment conditions, however, the two sides of this conversation will be summed together, and both the target speaker and that speaker's conversation partner will be represented in each conversation.

The difficulty of determining whether the target speaker speaks in the test conversation is affected by the sexes of the speakers in the test conversation. For no trials will both speakers be of opposite sex from the target. Systems will not be told whether the two test speakers are of the same or opposite sex, but may use automatic gender detection techniques if they wish. Performance results will be examined with respect to whether one or both target conversation speakers are of the same sex as the target.

Note that an interesting contrast will exist between this condition and that consisting of a single conversation side.

4.3 Factors Affecting Performance

All trials will be same-sex trials. This means that the sex of the test segment speaker, or of at least one test segment speaker when the test segment is a summed channel conversation, will be the same as that of the target speaker model. Performance will be reported separately for males and females and pooled across sex.

All trials involving telephone test segments will be different number trials. This means that the telephone numbers, and presumably the telephone handsets, used in the training and the test data segments will be different from each other.

Past NIST evaluations have shown that the type of telephone handset and the type of telephone transmission channel used can have a great effect on speaker recognition performance. Factors of these types will be examined in this evaluation.

Telephone callers in the Mixer collection (see section 6) are asked to classify the transmission channel as one of the following types:

- Cellular
- Cordless
- Regular (ie., land-line)

Telephone callers in the Mixer collection are asked to classify the instrument used as one of the following types:

- Speaker-phone
- Head-mounted
- Ear-bud
- Regular (ie., hand-held)

Performance will be examined as a function of the telephone transmission channel type and of the telephone instrument type in both the training and the test segment data. The effects of different types of cellular transmission encoding may also be considered.

4.4 Unsupervised Adaptation

As discussed in section 2.2.4, an unsupervised adaptation mode will be supported for each test. Performance with and without such adaptation will be compared for participants attempting tests with unsupervised adaptation.

4.5 Common Evaluation Condition

In each evaluation NIST specifies a common evaluation condition.[9] The performance results on trials satisfying this condition are

[9] In past NIST evaluations this was referred to as the "primary" condition. The term "common evaluation condition" is more

treated as the basic official evaluation outcome. The common evaluation condition in the 2004 evaluation will be regarded as all trials meeting each of the following specifications:

- Part of the core test as defined in section 2.2.3
- All training and test speech in English
- All training and test speech involve a telephone channel
- Male or female target – pooled across sex
- Hand-held telephone instruments used in all training and test speech
- All training and test segment data involves either land-line or cellular (not cordless) telephone transmission channels

4.6 Comparison With Previous Evaluations

In each evaluation it is of interest to compare performance results, particularly of the best performing systems, with those of previous evaluations. This is complicated by the fact that the evaluation conditions change in each successive evaluation.

The 2004 evaluation is no exception in this regard. The concatenation of speech segments with areas of silence removed, as practiced in the past, will not be done this year. There will be multiple durations utilized in the various training and test conditions, and none will be exactly as in the past. And whereas past evaluations have largely focused either primarily on landline data or primarily of cellular data, this evaluation will involve a mixture of both.

Nonetheless, NIST will examine how the results for test conditions most similar to those used in past compare with the past results. The test condition most similar to the "limited data" condition of the recent evaluations will be that involving training on a single conversation side and testing on 30-second excerpts. The condition most similar to the most widely examined "extended data" results of the past will be that involving training on eight conversation sides and testing on a single conversation sides. Participating sites, particularly those that participated in previous evaluations, may wish to consider including one of these tests in the results they submit for this evaluation. NIST will examine, after the fact, results on the subsets of trials of these tests that most resemble the conditions of past evaluation tests to facilitate the most meaningful comparison of performance results achieved over time in the course of the evaluations.

5 DEVELOPMENT DATA

The evaluation data for 2003 evaluation will serve as the development data for this year's evaluation, and will be covered by the LDC license agreement noted in section 6. Please refer to last year's evaluation plan for details.[10]

Note that no development data that is specific to the changed format and collection methods of the 2004 evaluation data (described in section 6) is being provided. Participating sites may use other speech corpora to which they have access for development. Such corpora should be described in the system descriptions. The original Switchboard-1 Corpus may be used, but participating sites are cautioned, particularly with respect to the development of background speaker models, that an effort is being made to recruit a limited number of the speakers in that corpus to participate in the new data collections from which this year's evaluation data will be selected.

6 EVALUATION DATA

The training and test segment data will be all newly collected by the Linguistic Data Consortium (LDC). The Mixer Project invited participating speakers to take part in numerous six-minute conversations on specified topics with people they did not know. The Fishboard platform allowed an automaton to initiate calls to selected pairs of speakers for most of the conversations, while individual speakers initiated some calls themselves, with the automaton contacting other speakers for them to converse with. Speakers initiating calls were encouraged to use unique telephone numbers (and thus generally unique telephone handsets) for their initiated calls.

The conversational data for this evaluation, to be distributed to participants by NIST on CD-ROM's, has not been publicly released. The LDC will provide a license agreement, which non-member participating sites must sign, governing the use of this data for the evaluation. The ASR transcript data, and any other auxiliary data which may also be supplied, will be made available by NIST in electronic form to all registered participants.

All conversations will have been processed through echo canceling software before being used to create the evaluation training and test segments.

All training and test segments will be stored as 8-bit mu-law continuous speech signals in separate SPHERE files. The SPHERE header of each such file will contain some auxiliary information as well as the standard SPHERE header fields. This auxiliary information will include the language of the conversation, whether or not the data was recorded over a telephone line, and the microphone type for non-telephone data. Most segments will be in English and recorded over a telephone line. The header will not contain information on the type of telephone transmission channel or the type of telephone instrument involved.

6.1 Single Channel Excerpts

The 10-second and 30-second excerpts to be used as training or as test segments will be continuous segments from single conversation sides that are estimated to contain approximately 10 or 30 seconds of actual speech.

The number of single channel excerpt training segments both for the 10-second and for the 30-second training conditions is expected to be around 600. The number of single channel excerpt test segments for each of the two durations is expected to be around 2000.

6.2 Single Conversation Sides

The single conversation sides to be used as training data or as test segments will all be approximately five minutes in total signal duration.

The number of single conversation training sides is expected not to exceed 6400. The number of these to be used to create speaker models based on a single conversation side is expected to be around

---

appropriate in the sense that this condition is used to officially rank system performance, and is not necessarily the condition that is most important to the evaluation.

[10] The year 2003 speaker recognition evaluation plan may be accessed from http://www.nist.gov/speech/tests/spk/2003/doc/

600. The numbers of models specified by 3, 8, or 16 sides are each expected to be around 400 or fewer.

The number of single conversation side test segments is expected to be around 2000.

6.3 Summed Channel Conversations

The summed-channel conversations to be used as training data or as test segments will all be approximately five minutes in total signal duration

The number of summed channel training conversations is expected to be around 1200. These will be used to specify around 400 target speaker models. The number of summed-channel conversation test segments is expected to be around 2000.

6.4 Trials to be included

The trials for each of the 28 speaker detection tests offered will be specified in separate index files. These will be text files in which each record specifies the model and a test segment for a particular trial. The number of trials in each test is expected not to exceed 25,000.

7 EVALUATION RULES

In order to participate in the 2004 speaker recognition evaluation, a site must complete, in its entirety, the core test condition (without unsupervised adaptation) as specified in section 2.2.3.[11] Any other test conditions included must be completed in their entirety.

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 model. Use of information about other test segments (except as permitted for the unsupervised adaptation mode condition) and/or other target speakers is **not** allowed.[12] For example:
  - Normalization over multiple test segments is **not** allowed, except as permitted for the unsupervised adaptation mode condition.
  - Normalization over multiple target speakers is **not** allowed.
  - Use of evaluation data for impostor modeling is **not** allowed.
- If an unsupervised adaptation condition is included, the test segments must be processed in the order specified.
- 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. Note that there will be no cross-sex trials.
- Knowledge of the language used in all segments, which will be provided, is allowed.
- Knowledge of whether or not a segment involves telephone channel transmission, and of the non-telephone microphone type used, which will be provided, is allowed.
- Knowledge of the telephone transmission channel type and of the telephone instrument type used in all segments 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.
- Knowledge of any information available in the SPHERE header **is** allowed.

8 EVALUATION DATA SET ORGANIZATION

The organization of the evaluation data will be:

- A top level directory used as a unique label for the disk: "**sp04-NN**" where NN is a digit pair identifying the disk
- Under which there will be four sub-directories: "**train**", "**test**", "**trials**", and "**doc**"

8.1 train Subdirectory

The "**train**" directory contains three subdirectories:

- **data**: Contains all the SPHERE formatted speech data used for training in each of the seven training conditions.
- **female**: Contains seven training files that defines the *female* models for each of the seven training conditions. (The format of these files is defined below.)
- **male**: Contains seven training files that defines the *male* models for each of the seven training conditions. (The format of these files is defined below.)

The seven training files for both male and female models have the same structure. There is one record per line, and each record contains two fields. The first field is the model identifier and the second field is a comma separated list of speech files (located in the "**data**" directory) that are to be used to train the model.

The seven training files in each gender directory are named:

- "***10sec.trn***" for the 10 second training condition, an example record looks like: "`3232 mrpv.sph`"
- "***30sec.trn***" for the 30 second training condition, an example record looks like: "`5241 mrpw.sph`"
- "***1side.trn***" for the 1 side training condition, an example record looks like: "`4240 mrpz.sph`"
- "***3sides.trn***" for the 3 sides training condition, an example record for this training condition looks like: "`7211 mrpz.sph,hrtz.sph,nost.sph`"
- "***8sides.trn***" for the 8 sides training condition.
- "***16sides.trn***" for the 16 sides training condition.
- "***3convs.trn***" for the 3 conversations (summed sides) training condition, an example record looks like: "`3310 nrfs.sph,irts.sph,poow.sph`"

[11] Participants are encouraged to do as many tests as possible. However, it is absolutely imperative that results for all of the trials 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.

8.2 test Subdirectory

The "**test**" directory contains one subdirectory:

- **data**: This directory contains all the SPHERE formatted speech test data to be used for each of the four test segment conditions. The file names will be arbitrary ones of four characters along with a ".sph" extension.

8.3 trials Subdirectory

The "**trials**" directory contains twenty-eight index files, one for each of the possible combinations of the seven training conditions and four test segment types. These index files define the various evaluation tests. The naming convention for these index files will be "*TrainCondition-TestCondition*.ndx" where *TrainCondition*, refers to the training condition and whose models are defined in the corresponding training file. Possible values for *TrainCondition* are: 10sec, 30sec, 1side, 3sides, 8sides, 16sides, and 3convs. "*TestCondition*" refers to the test segment condition. Possible values for *TestCondition* are: 10sec, 30sec, 1side, and 1conv.

Each record in a *TrainCondition-TestCondition.ndx* file contains exactly three fields and defines a single trial. The first field is the model identifier. The second field identifies the gender of the model, either "*m*" or "*f*". The third field is the test segment under evaluation, located in the **test/data** directory. This test segment name will not include the .sph extension. An example for the train on 3-sides and test on 1side index file "3sides-1side.ndx" looks like: "`7211 m nrbw`".

The records in these 28 files are ordered numerically by model identifier, and within each model's tests, alphabetically by the test segments. Each index file orders the trials as they are to be processed when unsupervised adaptation is used

8.4 doc Subdirectory

This will contain text files that document the evaluation and the organization of the evaluation data. This evaluation plan document will be included.

9 SUBMISSION OF RESULTS

Sites participating in one or more of the speaker detection evaluation tests must report results for each test in its entirety. These results for each test condition (1 of the 28 test index files) must be provided to NIST in a single file using a standard ASCII format, with one record for each trial decision. The file name should be intuitively mnemonic and should be constructed as "SSS_N", where

- SSS identifies the site, and
- N identifies the system.

9.1 Format for Results

Each file record must document its decision with the target model identification, test segment identification, and decision information. Each record must contain eight fields, separated by white space and in the following order:

1. The training type of the test – **10sec**, **30sec**, **1side**, **3sides**, **8sides**, **16sides**, or **3convs**
2. Adaptation mode. "**n**" for no adaptation and "**u**" for unsupervised adaptation.
3. The segment type of the test – **10sec**, **30sec**, **1side**, or **1conv**
4. The sex of the target speaker – **m** or **f**
5. The target model identifier
6. The test segment (minus the ".sph" extention).
7. The decision – **t** or **f** (whether or not the target speaker is judged to match the speaker in the test segment)
8. The likelihood score (where the more positive this score, the more likely the target and segment speakers are judged to match)

9.2 Means of Submission

Submissions may be made via email or via ftp. The appropriate addresses for submissions will be supplied to participants receiving evaluation data.

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 evaluation data, as if the test were run on a single CPU. This should be reported separately for creating models from the training data and for processing the test segments, and may be reported either as absolute processing time or as a multiple of real-time for the data processed. The additional time required for unsupervised adaptation should be reported where relevant. 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 March 14, 2004.

The evaluation data set CD-ROM's will be distributed by NIST on March 29, 2004.

The deadline for submission of evaluation results to NIST is April 22, 2004.

Evaluation results will be released to each site by NIST on April 29, 2004.

The deadline for site workshop presentations to be supplied to NIST in electronic form for inclusion in the workshop CD-ROM is May 27, 2004.

Registration and room reservations for the workshop must be received by (a date to be determined).

The follow-up workshop will be held on June 3-4, 2004 at the Hotel Beatriz in Toledo, Spain in conjunction with the 2004: A Speaker Odyssey workshop on speaker and language recognition. Those participating in the evaluation are expected to present and discuss their findings at this NIST portion of the workshop.

12 GLOSSARY

Trial – The individual evaluation unit involving a test segment and 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.

Test – A collection of trials constituting an evaluation component.

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