The NIST Year 2006 Speaker Recognition Evaluation Plan

1 INTRODUCTION

The year 2006 speaker recognition evaluation is part of an ongoing series of yearly evaluations conducted by NIST. These evaluations are 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 2006 evaluation will reuse from the 2005 evaluation some of the conversational telephone speech data collected for the Mixer Corpus by the Linguistic Data Consortium using the "Fishboard" platform, and will use some additional unexposed data from this collection and some similar data collected more recently. Some unexposed or newly collected "multi-channel" data collected simultaneously from a number of auxiliary microphones will also be included. The data will be mostly English speech, but will include some speech in four additional languages.

The evaluation will include 15 different speaker detection tests defined by the duration and type of the training and test data. For each such test, an unsupervised adaptation mode will be offered in addition to the basic test.

The evaluation will be conducted in April and May of 2006. A follow-up workshop for evaluation participants to discuss research findings will be held late in June in San Juan, Puerto Rico, preceding the Odyssey 2006 Workshop there. 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. Participating sites must follow the evaluation rules set forth in this plan and must be represented at the evaluation workshop. For more information, and to register to participate in the evaluation, please contact NIST.[1]

2 TECHNICAL OBJECTIVE

This evaluation focuses on speaker detection 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 2006 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 ten years. The task is to determine whether a specified speaker is speaking during a given segment of conversational speech.

2.2 Task Conditions

The speaker detection task for 2006 is divided into 15 distinct and separate tests. Each of these tests involves one of five training conditions and one of four test conditions. One of these tests (see section 2.2.3) is designated as the core test. Participants must do the core test and may choose to do any one or more of the other 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. Sites choosing the adaptation option for a test must also perform the test without adaptation to provide a baseline contrast.

2.2.1 Training Conditions

The training segments in the 2006 evaluation will be continuous conversational excerpts. As in the previous two years, there will be no prior removal of intervals of silence. Also, as in 2005, both sides of all two-channel conversations will be provided (to aid systems in echo cancellation, dialog analysis, etc.). For all two-channel segments, the channel containing the putative target speaker to be recognized will be identified.

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

1. A two-channel (4-wire) excerpt from a conversation estimated to contain approximately 10 seconds of speech of the target on its designated side (The NIST energy-based automatic speech detector will be used to estimate the duration of actual speech in the chosen excerpts.)

2. One two-channel (4-wire) conversation, of approximately five minutes total duration[2], with the target speaker channel designated.

3. Three two-channel (4-wire) conversations involving the target speaker on their designated sides

4. Eight two-channel (4-wire) conversations involving the target speaker on their designated sides

5. Three summed-channel (2-wire) conversations, formed by sample-by-sample summing of their two sides. Each of these conversations will include both the target speaker and another speaker. These three non-target speakers will all be distinct.

English language word transcripts, produced using an automatic speech recognition (ASR) system, will be provided for all training segments of each condition. These transcripts will, of course, be errorful, with word error rates typically in the range of 15-30%.

---

[1] Send email to speaker_poc@nist.gov, or call 301/975-3169. Each site must complete the registration process by signing and returning the registration form, which is available online at: . http://www.nist.gov/speech/tests/spk/sre-06_registration.pdf

[2] 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 language that identifies the speakers.

2.2.2 Test Segment Conditions

The test segments in the 2006 evaluation will be continuous conversational excerpts. As in the past two years, there will be no prior removal of intervals of silence. Also, as in 2005, both sides of all two-channel conversations will be provided (to aid systems in echo cancellation, dialog analysis, etc.). For all two-channel segments, the channel containing the putative target speaker to be recognized will be identified.

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

1. A two-channel (4-wire) excerpt from a conversation estimated to contain approximately 10 seconds of speech of the putative target speaker on its designated side (The NIST energy-based automatic speech detector will be used to estimate the duration of actual speech in the chosen excerpts.)

2. A two-channel (4-wire) conversation, of approximately five minutes total duration, with the putative target speaker channel designated.

3. A summed-channel (2-wire) conversation formed by sample-by-sample summing of its two sides

4. A two-channel (4-wire) conversation, with the usual telephone speech replaced by auxiliary microphone data in the putative target speaker channel. This auxiliary microphone data will be supplied in 8 kHz 8-bit μ-law form.

English language word transcripts, produced using an automatic speech recognition (ASR) system, will be provided for all test segments of each condition.

2.2.3 Training/Segment Condition Combinations

The matrix of training and test segment condition combinations is shown in **Table 1**. Note that only 15 (out of 20) condition combinations will be included in this year's evaluation. 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. The shaded box labeled "required" in **Table 1** is the **core test** for the 2006 evaluation. All participants are required to submit results for this test. Each participant may also choose to submit results for all, some, or none of the other 14 test conditions. For each test for which results are submitted, results for **all** trials must be included.

2.2.4 Unsupervised Adaptation Mode

The unsupervised adaptation mode allows systems to update themselves based on previous trial segments for the target model involved (up to and including the current trial segment). This is in contrast to the non-adaptive mode in which the system is static and the target (and background) speaker models are a function only of the target speaker training data. (The speaker models of course also benefit from speech data used and knowledge acquired during system development.)

In the unsupervised adaptation mode it is required that the trials for each target model be performed in the order given in the test index file (see section 8.3). The trials for each model will be grouped together, and the test segments for each of these target models will be listed in chronological order. Within the testing for each target model, the target (and background) models may be updated by the system after each trial using the test segment data processed thus

far for that target model. However, the adaptation must be discarded and the system reset to its initial unadapted state whenever a new model is encountered in the test index file.

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

		Test Segment Condition			
		10 sec 2-chan	1 conv 2-chan	1 conv summed-chan	1 conv aux mic
Training Condition	10 seconds 2-channel	optional			
	1 conversation 2-channel	optional	**required**	optional	optional
	3 conversation 2-channel	optional	optional	optional	optional
	8 conversation 2-channel	optional	optional	optional	optional
	3 conversation summed-channel		optional	optional	

For each test performed in unsupervised adaptation mode results must also be submitted for that test in non-adaptive mode in order to provide a contrast between adaptive and non-adaptive performance. The unsupervised adaptation techniques used should be discussed in the system description (see section 10).

3 PERFORMANCE MEASURE

There will be a single basic cost model for measuring speaker detection performance, to be used for all speaker detection tests. For each test, a detection cost function will be computed over the sequence of trials provided. 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.[3]

This detection cost function is defined as a weighted sum of miss and false alarm error probabilities:

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

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.

---

[3] 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.

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

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

To improve the intuitive meaning of C_{Det}, it will be normalized by dividing it by the best cost that could be obtained without processing the input data (i.e., by either always accepting or always rejecting the segment speaker as matching the target speaker, whichever gives the lower cost):

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

and

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

In addition to the actual detection decision, a confidence score will also be required for each test hypothesis. This confidence score should reflect the system's estimate of the probability that the test segment contains speech from the target speaker. Higher confidence scores should indicate greater estimated probability that the target speaker's speech is present in the segment. The confidence scores will be used to produce *Detection Error Tradeoff (DET)* curves, in order to see how misses may be traded off against false alarms. Since these curves will pool all trials in each test for all target speakers, it is necessary to normalize the confidence scores across all target speakers.

The ordering of the confidence scores is all that matters for computing the detection cost function, which corresponds to a particular application defined by the parameters specified in section 3, and for plotting DET curves. But these scores are more informative, and can be used to serve any application, if they represent actual probability estimates. It is suggested that participants provide as scores estimated likelihood ratio values, which do not depend on the application parameters. In terms of the conditional probabilities for the observed data of a given trial relative to the alternative target and non-target hypotheses the likelihood ratio *(LR)* is given by:

LR = prob (data | target hyp.) / prob (data | non-target hyp.)

Sites are asked to specify if their scores may be interpreted as likelihood ratio estimates. If so, floating point format should probably be used for scores to avoid any truncation of small values to zero.

A further type of scoring and graphical presentation will be performed on submissions whose scores are declared to represent likelihood ratios. A log likelihood ratio *(llr)* based cost function, which is not application specific and may be given an information theoretic interpretation, is defined as follows:

C_{llr} = 1 / (2 * log2) * ($\sum$log(1+1/s)/N_{TT}+ ($\sum$log(1+s))/N_{NT})

where the first summation is over all target trials, the second is over all non-target trials, N_{TT} and N_{NT} are the total numbers of target

and non-target trials, respectively, and *s* represents a trial's likelihood ratio score.[4] i

Graphs based on this cost function, somewhat analogous to DET curves, will also be included. These may serve to indicate the ranges of possible applications for which a system is or is not well calibrated.[5]

4 EVALUATION CONDITIONS

Performance will be measured, graphically presented, and analyzed, as discussed in section 3, over all the trials of each of the 15 tests specified in section 2, and over subsets of these trials of particular evaluation interest. Comparisons will be made of performance variation across the different training conditions and the different test segment conditions which define these tests. The effects of factors such as language, telephone transmission type, and microphone type, will be examined. The possible performance benefit of unsupervised adaptation will be considered. As in previous years, a common evaluation condition (a subset of the core test) will be defined. And comparisons will be made between this year's evaluation results and those of recent past years.

4.1 Training Data

As discussed in section 2.2.1, there will be five training conditions. NIST is 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 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 will have been collected over telephone channels.

The sex of each target speaker will be provided to systems in the test index file (see section 8.3).

For all training conditions, English language ASR transcriptions of all data will be provided along with the audio data. Systems may utilize this data as they wish. The acoustic data may be used alone, the transcriptions may be used alone, or all data may be used in combination.[6]

4.1.1 Excerpts

As discussed in section 2.2.1, one of the training conditions is an excerpt of a conversation containing approximately 10 seconds of

---

[4] This reasons for choosing this cost function, and its possible interpretations, are described in detail in the paper "Application-independent evaluation of speaker detection" in Computer Speech & Language, volume 20, issues 2-3, April-July 2006, pages 230-275, by Niko Brummer and Johan du Preez.

[5] See the discussion of *Applied Probability of Error (APE)* curves in the reference cited in the preceding footnote.

[6] Note, however, that the ASR transcripts will all be generated by an English language recognizer, regardless of the actual language being spoken.

estimated speech duration in the channel of interest. The actual duration of target speech will vary (so that the excerpts include only whole turns whenever possible) but the target speech duration will be constrained to lie in the range of 8-12 seconds.

4.1.2 Two-channel Conversations

As discussed in section 2.2.1, there will be training conditions consisting of one, three, and eight two-channel conversations of a given speaker. These will consist of approximately five minutes from an original six minute conversation, with an initial segment of about one minute excised. The excision point will be chosen so as not to include a partial speech turn.

4.1.3 Summed-channel Conversations

As discussed in section 2.2.1, one of the training conditions will consist of three summed-channel conversations, minus initial segments of about a minute each. Here the two sides of each conversation, in which both the target speaker 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 two-channel conversations.

4.2 Test data

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

Most of the test data will be in English, but some may be in Arabic, Mandarin, Russian, or Spanish. The language used in all test data files will be indicated in the file header and available for use.

For all test conditions, English language ASR transcriptions of the data will be provided along with the audio data. Systems may utilize this data 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.[6]

4.2.1 Excerpts

As discussed in section 2.2.2, one of the test conditions is an excerpt of a conversation containing approximately 10 seconds of estimated speech duration in the channel of interest. The actual duration of target speech will vary (so that the excerpts include only whole turns whenever possible) but the target speech duration will be constrained to lie in the range of 8-12 seconds.

4.2.2 Two-channel Conversations

As discussed in section 2.2.2, one of the test conditions is a single two-channel conversation. (The channel of interest will be designated in the test index file – see section 8.3.) Each conversation will consist of approximately five minutes from an original six minute conversation, with an initial segment of about

one minute excised. The excision point will be chosen so as not to include a partial speech turn.

4.2.3 Summed-channel Conversations

As discussed in section 2.2.2, one of the test conditions is a single summed-channel conversation, minus an initial segment of about a minute. Here the two sides of the conversation will be summed together, and only one of the two speakers included may match a target speaker specified in a trial.

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. Systems will not be told whether the two test speakers are of the same or opposite sex, but automatic gender detection techniques may be used. Performance results will be examined with respect to whether one or both of the test speakers are of the same sex as the target. (For all trials there will be at least one speaker who is of the same sex as the target speaker.)

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

4.2.4 Auxiliary Microphone Conversations

As discussed in section 2.2.2, one of the test conditions is a two-channel conversation in which the channel of interest is an auxiliary microphone channel. The other channel will contain normal telephone data. As with the normal two-channel conversation test condition, about five minutes from an original six-minute conversation will be provided. The microphone data will be provided in single byte 8-bit μ-law form that matches the telephone data provided.

Several types of auxiliary microphones will be included in this data. 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
- Mini-boom mike
- Courtroom mike
- Conference room mike
- Distant mike
- Near-field mike
- PC stand mike
- Micro-cassette mike

Information on the microphone type used in each non-telephone test segment data will not be available to recognition systems.

These auxiliary microphone conversations will all be in English. ASR transcriptions will be provided as they are for all other calls. Note, however, that the ASR transcript will be produced using telephone data input rather than the auxiliary microphone signal.

4.3 Factors Affecting Performance

All trials will be *same-sex* trials. This means that the sex of the test segment speaker in the channel of interest (two-channel), or of at least one test segment speaker (summed-channel), will be the same as that of the target speaker model. Performance will be reported separately for males and females and also for both sexes pooled.

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. (For some telephone conversational data collected at the sites collecting the auxiliary microphone data, information other than phone numbers may be used to establish that different handsets are used.)

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 to the extent that information of this type is available.

Telephone callers are generally asked to classify the transmission channel as one of the following types:

- Cellular
- Cordless
- Regular (i.e., land-line)

Telephone callers are generally also asked to classify the instrument used as one of the following types:

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

Performance will be examined, to the extent the information is available and the data sizes are sufficient, as a function of the telephone transmission channel type and of the telephone instrument type in both the training and the test segment data.

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 (a subset of trials in the core test that satisfy additional constraints) in order to better foster technical interactions and technology comparisons among sites. The performance results on these trials are treated as the basic official evaluation outcome. The common evaluation condition for the 2006 evaluation will be the subset of the trials in the core test that satisfy the following condition:

- The test segment and all of the training data for the target model are in English.

Note that all transmission and instrument types will be included in the common evaluation condition this year.

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 generally complicated by the fact that the evaluation conditions change in each successive evaluation. For the 2006 evaluation the test conditions are essentially identical to ones used in 2005, and most are similar to ones used in 2004. Thus it will be possible to make fairly direct comparisons between 2006 and 2005 and even 2004 tests. Comparisons may also be made with the results of earlier evaluations for conditions most similar to those in this evaluation.

While the test conditions will match those used previously, the test data will be partially different. The 2006 target speakers will all be different from those of the 2004 evaluation, but will include many of the same speakers as in 2005. The question always arises of to what extent are the performance differences due to random differences in the test data sets. For example, are the new target speakers in the current evaluation easier, or harder, on the average to recognize? To help address this question, sites participating in the 2006 evaluation that also participated in 2004 or 2005 are strongly encouraged to submit to NIST results for their (unmodified) 2004 or 2005 systems run on the 2006 data for the same test conditions as previously. Such results will not count against the limit of three submissions per test condition (see section 7). Sites are also encouraged to "mothball" their 2006 systems for use in similar comparisons in future evaluations.

5 DEVELOPMENT DATA

The evaluation data for the 2004 evaluation will serve as the primary development data for this year's evaluation. This data is covered by the LDC license agreement noted in section 6. Please refer to the 2004 evaluation plan for details.[7]

All of the cross-channel microphone speech data used in the 2005 evaluation, and all of the telephone data involving the speakers of this microphone data, will also be available as development data for the 2006 evaluation. NIST will be making this data available as a package, and it will be covered by the LDC license agreement as well.

Participating sites may use other speech corpora to which they have access for development. Such corpora should be described in the site's system description. The original Switchboard-1 corpus may be used, but note that an effort is being made to recruit a limited number of the speakers from that corpus to participate in the new data collection from which this year's evaluation data will be selected.

6 EVALUATION DATA

Both the target speaker training data and the test segment data will have been collected by the Linguistic Data Consortium (LDC) as part of the Mixer project or in more recent similar collections. The Mixer project invited participating speakers to take part in numerous six-minute conversations on specified topics with strangers. The Fishboard platform used to collect this data automatically initiated calls to selected pairs of speakers for most of the conversations, while participating speakers also initiated some calls themselves, with the collection system contacting other speakers for them to converse with. Speakers were encouraged to use different telephone instruments for their initiated calls.

The conversational data for this evaluation will be distributed to evaluation participants by NIST on a firewire drive. The LDC provides a license agreement[8], 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 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.

---

[7] www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf

[8] Available online at http://www.nist.gov/speech/tests/spk/2006/

All training and test segments will be stored as 8-bit μ-law speech signals in separate SPHERE[9] 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 and whether or not the data was recorded over a telephone line.

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. Nor will the microphone type be identified for the auxiliary microphone test, as noted in section 4.2.4.

6.1 Excerpts

The 10-second two-channel excerpts to be used as training data or as test segments will be continuous segments from single conversations that are estimated to contain approximately 10 seconds of actual speech in the channel of interest. When both channels are channels of interest for different trials, then each will contain approximately 10 seconds of actual speech.

The number of training segments is expected *not to exceed 2000*. The number of test segments is expected *not to exceed 4000*.

6.2 Two-channel Conversations

The two-channel conversations to be used as training data or as test segments will be approximately five minutes in duration.

The number of conversations to be used for training is expected *not to exceed 10,000*. The number of speaker models based on a single conversation, and the numbers of models specified by 3 or by 8 conversations are each expected *not to exceed 2000*.

The number of test segments is expected not to exceed *4000*.

6.3 Summed-channel Conversations

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

The number of summed channel training conversations is expected *not to exceed 2400*. These will be used to specify *no more than 800* target speaker models. The number of summed-channel conversation test segments is expected *not to exceed 4000*.

6.4 Auxiliary Microphone Conversations

These two-channel conversations to be used as test segments will be approximately five minutes in duration.

The number of test segments is expected *not to exceed 2000*.

6.5 Number of Trials

The trials for each of the 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 *75,000*.

7 EVALUATION RULES

In order to participate in the 2006 speaker recognition evaluation a site must submit complete results for the core test condition

(without unsupervised adaptation) as specified in section 2.2.3.[10] Results for other tests are optional but strongly encouraged.

All participants must observe the following evaluation rules and restrictions in their processing of the evaluation data:

- 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.[11] 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, except as permitted for the unsupervised adaptation mode condition.
- If an unsupervised adaptation condition is included, the test segments for each model must be processed in the order specified.
- The use of manually produced transcripts or other human-produced 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 no cross-sex trials are planned, but that summed-channel segments may involve either same sex or opposite sex speakers.
- 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 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 human 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.

The following general rules about evaluation participation procedures will also apply for all participating sites:

- Access to past presentations – Each new participant that has signed up for, and thus committed itself to take part in, the upcoming evaluation and workshop will be able to receive, upon request, the CD of presentations that were presented at the preceding workshop.
- Limitation on submissions – Each participating site may submit results for up to three different systems per evaluation

---

[9] ftp://jaguar.ncsl.nist.gov/pub/sphere_2.6a.tar.Z

[10] It is imperative that results be complete for every test submission. A test submission is complete if and only if it includes a result for every trial in the test.

[11] 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 the training data for a specific target speaker and then performing the detection task on whatever speech segment is presented, without the (artificial) knowledge of other test data.

condition for official scoring by NIST. Results for systems using unsupervised adaptation and results for earlier year systems run on 2006 data will not count against this limit. Note that the answer keys will be distributed to sites by NIST shortly after the submission deadline. Thus each site may score for itself as many additional systems and/or parameter settings as desired.

- Attendance at workshop – Each evaluation participant is required to have one or more representatives at the evaluation workshop who must present there a meaningful description of its system(s). Evaluation participants failing to do so may be excluded from future evaluation participation.

- Dissemination of results
    - Participants may publish and otherwise disseminate their own results.
    - Participants may publish and otherwise disseminate anonymous charts, produced by NIST, of all system results for a condition.
    - Participants may not publish or otherwise disseminate the names or results of other participants without the explicit written permission of each such participant. Participants violating this rule may be excluded from future evaluations.

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: "**sp06-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 the SPHERE formatted speech data used for training in each of the seven training conditions.
- **female**: Contains five training files that define the *female* models for each of the seven training conditions. (The format of these files is defined below.)
- **male**: Contains five training files that define the *male* models for each of the seven training conditions. (The format of these files is defined below.)

The five training files for both male and female models have similar structures. Each has one record per line, and each record contains two fields. The first field is the model identifier. The second includes a comma separated list of speech files (located in the "**data**" directory) that are to be used to train the model. For the 2-channel training conditions, each list item also specifies whether the target speaker's speech is on the "A" or the "B" channel of the speech file.

The five training files in each gender directory are named:

- "*10sec4w.trn*" for the 10 second two-channel training condition, an example record looks like:
    ```
    3232 mrpv.sph:B
    ```
- "*1conv4w.trn*" for the 1 conversation two-channel training condition, an example record looks like:
    ```
    4240 mrpz.sph:A
    ```
- "*3conv4w.trn*" for the 3 conversation two-channel training condition, an example record for this training condition looks like:
    ```
    7211 mrpz.sph:B,hrtz.sph:A,nost.sph:B
    ```
- "*8conv4w.trn*" for the 8 conversation training condition.
- "*3conv2w.trn*" for the 3 conversation summed-channel training condition, an example record looks like:
    ```
    3310 nrfs.sph,irts.sph,poow.sph
    ```

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 index files, one for each of the possible combinations of the five 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: 10sec4w, 1conv4w, 3conv4w, 8conv4w, and 3conv2w. "*TestCondition*" refers to the test segment condition. Possible values for *TestCondition* are: 10sec4w, 1conv4w, 1conv2w, and 1convmic.

Each record in a *TrainCondition-TestCondition.ndx* file contains four 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. The fourth field specifies the channel of the test segment speech of interest, either "A" or "B". (This will always be "A" for the summed channel test.) An example for the train on three conversations two-channel and test on one conversation two-channel index file "3conv4w-1conv2w.ndx" looks like: "7211 m nrbw B".

The records in these 20 files are ordered numerically by model identifier, and within each model's tests, chronologically by the recording dates of the test segments. Thus each index file specifies the processing order of the trials for each model. (This order of processing is mandatory 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 xx 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 nine fields, separated by white space and in the following order:

1. The training type of the test – **10sec4w**, **1conv4w**, **3conv4w**, **8conv4w**, or **3convs2w**

2. Adaptation mode. "**n**" for no adaptation and "**u**" for unsupervised adaptation.

3. The segment type of the test – **10sec4w**, **1conv4w**, **1conv2w**, or **1convmic**

4. The sex of the target speaker – **m** or **f**

5. The target model identifier

6. The test segment identifier

7. The test segment channel of interest, either "a" or "b"

8. The decision – **t** or **f** (whether or not the target speaker is judged to match the speaker in the test segment)

9. The confidence score (where larger scores indicate greater likelihood that the test segment contains speech from the target speaker)

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. Sites should also indicate if it is the case that the confidence scores in a submission are to be interpreted as likelihood ratios.

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. A single site may submit the results for up to three separate systems for evaluation for each particular test, not counting test results using unsupervised adaptation and not counting results for 2004 or 2005 systems run on the 2006 data. If results for more than one system are submitted for a test, however, the site must identify one system as the "primary" system for the test prior to performing the evaluation. Sites are welcome to present descriptions of and results for additional systems at the evaluation workshop.

For each system for which results are submitted, 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 27, 2006.

The evaluation data set will be distributed by NIST so as to arrive at participating sites on April 24, 2006.

The deadline for submission of evaluation results to NIST is May 14, 2006 at 11:59 PM, Washington time.

Initial evaluation results will be released to each site by NIST on May 22, 2006.

The deadline for site workshop presentations to be supplied to NIST in electronic form for inclusion in the workshop CD-ROM is (a date to be determined).

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

The follow-up workshop will be held June 25-27, 2006 at the Ritz Carlton Hotel in San Juan, Puerto Rico in conjunction with the IEEE Odyssey 2006 Speaker and Language Recognition Workshop. All sites participating in the evaluation must have one or more representatives in attendance to discuss their systems and results.

12 GLOSSARY

Test – A collection of trials constituting an evaluation component.

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

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.

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.

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