The NIST Year 2008 Speaker Recognition Evaluation Plan

1 INTRODUCTION

The year 2008 speaker recognition evaluation is part of an ongoing series of 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 2008 evaluation will be distinguished from the most recent evaluations, in particular those in 2005 and 2006, by including in the training and test conditions for the core (required) test not only conversational telephone speech data but also conversational speech data (of comparable duration) recorded over a microphone channel involving an interview scenario, and additionally, for the test condition, conversational telephone speech recorded over a microphone channel. Systems will know whether each segment comes from a telephone or a microphone channel, and whether it involves the interview scenario or an ordinary telephone conversation, but will be required to process trials involving all segments of each type. Submitted results will be scored after the fact to determine performance levels for telephone data, for microphone data of different conversational styles and microphone types, and for differing combinations of training and test data.

The optional tests this year will include, in addition to the training and test conditions of recent evaluations, a condition involving longer duration segments of interview data recorded over microphone channels.

The 2008 evaluation will not reuse data from previous evaluations, but some of the target speakers of the 2006 evaluation may reappear in the 2008 evaluation data. Target speakers from evaluations prior to 2006 will not be used in the 2008 evaluation data.

As in recent evaluations, some of the speakers in the telephone conversational data will be bilingual and their evaluation data may include speech in a language other than English as well as speech in English. The microphone recorded interview data will all be in English.

The evaluation will include 13 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 2008. A follow-up workshop for evaluation participants to discuss research findings will be held in June in Montreal, Quebec, Canada. 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 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 2008 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 twelve 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 2008 is divided into 13 distinct and separate tests. Each of these tests involves one of six 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 2008 evaluation will be continuous conversational excerpts. As in recent evaluations, there will be no prior removal of intervals of silence. Also, except for summed channel telephone conversations and long interview segments as described below, two separate conversation channels will be provided (to aid systems in echo cancellation, dialog analysis, etc.). For all such two-channel segments, the primary channel containing the target speaker to be recognized will be identified.

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

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

---

2. ***short2***: One two-channel telephone conversational excerpt, of approximately five minutes total duration[2], with the target speaker channel designated *or* a microphone recorded conversational segment of approximately three minutes total duration involving the target speaker and an interviewer. For the interview segments most of the speech will generally be spoken by the target speaker, and for consistency across the condition, a second zeroed out channel will be included.

3. ***3conv***: Three two-channel telephone conversational excerpts involving the target speaker on their designated sides.

4. ***8conv***: Eight two-channel telephone conversation excerpts involving the target speaker on their designated sides.

5. ***long***: A single channel microphone recorded conversational segment of eight minutes or longer duration involving the target speaker and an interviewer. Most of the speech will generally be spoken by the target speaker.

6. ***3summed***: Three summed-channel telephone conversational excerpts, 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 English word error rates typically in the range of 15-30%. Note, however, that the ASR system will always be run on two separated channels, and run only once for those segments that have been simultaneously recorded over multiple channels, and the ASR transcripts provided may sometimes be superior to what current systems could provide for the actual channel involved. This is viewed as reasonable since ASR systems are expected to improve over time, and this evaluation is not intended to test ASR capabilities.

For the interview segments of the second and fifth conditions described above, the estimated intervals where the target speaker is speaking, as determined by an energy based segmenter utilizing the audio signals from lavalier microphones worn by each of the two speakers, will be provided. Systems may limit their processing to these intervals, or they may choose to process the full segments and do their own speaker separation processing.

2.2.2 Test Segment Conditions

The test segments in the 2008 evaluation will be continuous conversational excerpts. As in recent evaluations, there will be no prior removal of intervals of silence. Also, except for summed channel telephone conversations and long interview segments as described below, two separate conversation channels will be provided (to aid systems in echo cancellation, dialog analysis, etc.).

---

[2] Each conversation side will consist of five minutes of a longer conversation, and will exclude the first minute. This will eliminate from the evaluation data the less-topical introductory dialogue, which is more likely to contain language that identifies the speakers.

For all such two-channel segments, the primary channel containing the putative target speaker to be recognized will be identified.

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

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

2. ***short3***: A two-channel telephone conversational excerpt, of approximately five minutes total duration, with the putative target speaker channel designated *or* a similar such telephone conversation but with the putative target channel being a (simultaneously recorded) microphone channel *or* a microphone recorded conversational segment of approximately three minutes total duration involving the putative target speaker and an interviewer. For the interview segments, most of the speech will generally be spoken by the target speaker, and for consistency across the condition, a second zeroed out channel will be included.

3. ***long***: A single channel microphone recorded conversational segment of eight minutes or longer duration involving the putative target speaker and an interviewer. Most of the speech will generally be spoken by the target speaker.

4. ***summed***: A summed-channel telephone conversation formed by sample-by-sample summing of its two sides

English language word transcripts, produced using an automatic speech recognition (ASR) system as described in section 2.2.2, will be provided for all test segments of each condition.

For the interview segments of the second and third conditions described above, the estimated intervals where the target speaker is speaking, as described in section 2.2.2, will be provided.

2.2.3 Training/Segment Condition Combinations

The matrix of training and test segment condition combinations is shown in **Table 1**. Note that only 13 (out of 24) 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 2008 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 12 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 processed in the order given in the test index file (see section 8.3x). 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. No reprocessing of earlier trials is permitted. The adaptation, however, 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			
		10sec	**short3**	**long**	**summed**
Training Condition	**10sec**	optional			
	short2	optional	**required**		optional
	3conv		optional		optional
	8conv	optional	optional		optional
	long		optional	optional	
	3summed		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:

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

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):

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

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 log likelihood ratio values (using natural logarithms), 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 log likelihood ratio estimates.

A further type of scoring and graphical presentation will be performed on submissions whose scores are declared to represent log 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:

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

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.[4]

---

[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

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 13 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. Several common evaluation conditions of interest, each a subset of the core test, will be defined. And relevant 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 six 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 telephone training conversations involving bi-lingual speakers will be collected in a number of other languages. 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.

Another performance factor of interest for English telephone conversations will be whether or not the speaker is a native U.S. English speaker. Information on this will not, however, be provided to systems.

The sex of each target speaker will be provided to systems in the test index file (see section 8.33).

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 10-second Excerpts

As discussed in section 2.2.1, one of the training conditions is an excerpt of a telephone 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.

4.1.2 Two-channel Conversations

As discussed in section 2.2.1, there will be training conditions consisting of one, three, and eight two-channel telephone conversational excerpts of a given speaker. (The first of these conditions will also include short interview segments.) These will each consist of approximately five minutes from a longer original conversation. The excision points will be chosen so as not to include partial speech turns.

4.1.3 Short Interview Segments

As discussed in section 2.2.1, one of the training conditions involves short conversational interview segments (along with single two-channel telephone conversations). These will each consist of approximately three minutes from a longer interview session. The excision points will be chosen so as not to include partial speech turns. Three minute segments are expected on average to include about as much speech from the speaker of interest as do five minute segments from telephone conversations. Two channels will be provided, the first from a microphone placed somewhere in the interview room, and the other a zero channel provided for consistency across the training condition. Information on the microphone type of the first channel will not be available to systems.

The microphone data will be provided in 8-bit μ-law form that matches the telephone data provided.

The speech of the short interview segments will all be in English.

4.1.4 Long Interview Segments

As discussed in section 2.2.1, one of the training conditions involves long conversational interview segments. These will each consist of eight minutes or more from an interview session. Any excision points will be chosen so as not to include partial speech turns. Only a single channels will be provided from a microphone placed somewhere in the interview room. Information on the microphone type will not be available to systems.

The microphone data will be provided in 8-bit μ-law form that matches the telephone data provided.

The speech of the long interview segments will all be in English.

4.1.5 Summed-channel Conversations

As discussed in section 2.2.1, one of the training conditions will consist of three summed-channel telephone conversation segments of about five minutes 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.

& 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.

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 telephone speech will be in other languages, generally involving bilingual speakers who also speak English. 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.

4.2.1 Excerpts

As discussed in section 2.2.2, one of the test conditions is an excerpt of a telephone 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 involves single two-channel telephone conversational excerpts (or a short interview segment). Each excerpt will consist of approximately five minutes from a longer original conversation. The excision points will be chosen so as not to include partial speech turns.

4.2.3 Short Interview Segments

As discussed in section 2.2.2, one of the test conditions involves short conversational interview segments (along with two-channel telephone conversations). These will each consist of approximately three minutes from a longer interview session. The excision points will be chosen so as not to include partial speech turns. Three minute segments are expected on average to include about as much speech from the speaker of interest as do five minute segments from telephone conversations. Two channels will be provided, the first from a microphone placed somewhere in the interview room, and the other a zero channel provided for consistency across the test condition. Information on the microphone type of the first channel will not be available to systems.

The microphone data will be provided in 8-bit μ-law form that matches the telephone data provided.

The speech of the short interview segments will all be in English.

4.2.4 Long Interview Segments

As discussed in section 2.2.2, one of the training conditions involves long conversational interview segments. These will each consist of eight minutes or more from an interview session. Any excision points will be chosen so as not to include partial speech turns. Only a single channel will be provided from a microphone placed somewhere in the interview room. Information on the microphone type will not be available to systems.

The microphone data will be provided in 8-bit μ-law form that matches the telephone data provided.

The speech of the long interview segments will all be in English.

4.2.5 Summed-channel Conversations

As discussed in section 2.2.2, one of the test conditions is a single summed-channel conversational excerpt of about five minutes. Here the two sides of the conversation will be summed together, and 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.

4.3 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.

This evaluation will focus on examining the effects of channel on recognition performance. This will include in particular the comparison of performance involving telephone segments with that involving microphone segments. Since each trial has a training and a test segment, four combinations may be examined here. For test segments only, performance on telephone channel telephone conversations will be compared with performance on microphone channel telephone conversations and with performance on microphone interview segments.

For trials involving microphone segments, it will be of interest to examine the effect of the different microphone types tested on performance, and the significance on performance of the match or mismatch of the training and test microphone types.

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 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 has specified 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. Because of the broader scope of the 2008 evaluation and the multiple types of audio data included in the core test, several common evaluation conditions will be specified. At the same time, it will not to appropriate to examine performance results over all trials of the core test lumped together. The common conditions to be considered will include the following subsets of all of the core test trials:

- All trials involving only interview speech in training and test
- All trials involving interview speech from the same microphone type in training and test
- All trials involving interview speech from different microphones types in training and test
- All trials involving interview training speech and telephone test speech
- All trials involving telephone training speech and non-interview microphone test speech
- All trials involving only telephone speech in training and test
- All trials involving only English language telephone speech in training and test
- All trials involving only English language telephone speech spoken by a native U.S. English speaker in training and test

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 2008 evaluation the test conditions involving conversational telephone speech (with test segments possibly recorded over a microphone channel) are essentially identical those used in 2006. Thus it will be possible to make fairly direct comparisons between 2008 and 2006 for these conditions. 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 different. The 2008 target speakers may include some used in the 2006 evaluation, but most will not have appeared previously. 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 2008 evaluation that also participated in 2006 are strongly encouraged to submit to NIST results for their (unmodified) 2006 (or earlier year) systems run on the 2008 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 2008 systems for use in similar comparisons in future evaluations.

5   DEVELOPMENT DATA

All of the previous NIST NRE evaluation data, covering evaluation years 1996-2006 may be used as development data for 2008. This data will be sent to prospective evaluation participants by the Linguistic Data Consortium on a hard drive provided the required license agreement is signed and submitted to the LDC.[7]

A limited amount of development data representing the interview scenario that is new for 2008 will also be made available. This will include interview sessions involving six speakers, which speakers will not be targets in the 2008 evaluation data. This data will be provided on DVD by request to all sites that have submitted the LDC license agreement described above.

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 (section 10).

6   EVALUATION DATA

Both the target speaker training data and the test segment data, including the interview data, will have been collected by the Linguistic Data Consortium (LDC) as part of the various phases of its Mixer project.[8] The telephone collection part of the Mixer project invited participating speakers to take part in numerous 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 speech data for this evaluation will be distributed to evaluation participants by NIST on a firewire drive. The LDC license agreement described in section 5, which non-member sites must sign to participate in the evaluation, will govern the use of this data for the evaluation. The ASR transcript data, the estimated speech intervals of interview target speakers, and any other auxiliary data which may be supplied, will be made available by NIST in electronic form to all registered participants.

Since both channels of all telephone conversational data are provided, this data will not be processed through echo canceling

---

[7] Find link at http://www.nist.gov/speech/tests/sre/2008/index.html

[8] A description of the recent Mixer collections may be found at: http://papers.ldc.upenn.edu/Interspeech2007/Interspeech_2007_Mixer_345.pdf

software. Participants may choose to do such processing on their own.[9]

All training and test segments will be stored as 8-bit μ-law speech signals in separate SPHERE[10] 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 whether or not the data is from an interview session.

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 interview data, as noted in section 4.

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 two-channel conversational excerpts to be used as training data or as test segments will be approximately five minutes in duration, while all the short interview segments will be approximately three minutes in duration. The primary channel of interest will be specified. Note that for the short interview segments the second, non-primary, channel contain all zeroes. Each segment will be identified as coming either from a telephone conversation or from an interview.

The single channel long interview segments to be used as training data or as test segments will be eight minutes or longer in duration.

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

6.1 Numbers of Models

Table 3 provides estimated upper bounds on the numbers of models (target speakers) to be included in the evaluation for each training condition.

Table 3:   Upper bounds on numbers of models by training condition

Training Condition	Max Models
10sec	2000
short2	4000
3conv	2000
8conv	2000
long	2000
3summed	2000

6.2 Numbers of Test Segments

Table 4 provides estimated upper bounds on the numbers of segment to be included in the evaluation for each test condition.

Table 4:   Upper bounds on numbers of segments by test condition

Test Conditions	Max Segments
10sec	5000
short3	8000
long	2000
summed	5000

6.3 Numbers 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 for each test condition is expected not to exceed *100,000*.

7 EVALUATION RULES

In order to participate in the 2008 speaker recognition evaluation a site must submit complete results for the core test condition (without unsupervised adaptation) as specified in section 2.2.3.[11] Results for other tests are optional but strongly encouraged.

Participating sites, particularly those with limited internal resources, may utilize publicly available software designed to support the development of speaker detection algorithms.[12] The software used should be specified in the system description (section 10).

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.[13] 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.

---

[9] One publicly available   source of such software is http://www.ece.msstate.edu/research/isip/projects/speech/software/legacy/fir_echo_canceller/

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

---

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

[12] One publicly available source is the Mistral software for biometric applications developed at the University of Avignon along with other European sites: http://mistral.univ-avignon.fr/en/

[13] 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.

- Use of evaluation data for impostor modeling is **not** allowed, except as permitted for the unsupervised adaptation mode condition.
- Speech data from past evaluations may be used for general algorithm development and for impostor modeling, but may not be used directly for modeling target speakers of the 2008 evaluation.

- 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-created information 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 condition for official scoring by NIST. Results for systems using unsupervised adaptation and results for earlier year systems run on 2008 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 will be excluded from future evaluation participation.
- Dissemination of results
  - Participants may publish or otherwise disseminate their own results.
  - NIST will generate and place on its web site charts of all system results for conditions of interest and, unlike past practice, these charts may contain the site names of the systems involved. Participants may publish or otherwise disseminate these charts, unaltered and with appropriate reference to their source.

- Participants may not publish or otherwise disseminate their own comparisons of their performance results with those of other participants without the explicit written permission of each such participant. Participants violating this rule will 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: "**sp08-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 6 training conditions.
- **female**: Contains 6 training files that define the *female* models for each of the 6 training conditions. (The format of these files is defined below.)
- **male**: Contains 6 training files that define the *male* models for each of the 6 training conditions. (The format of these files is defined below.)

The 6 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 two 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 6 training files in each gender directory are named:

- "*10sec.trn*" for the 10 second two channel training condition; an example record looks like:

  `3232 mrpv.sph:B`
- "*short2.trn*" for the 1 conversation/short interview two channel training condition; an example record looks like:

  `4240 mrpz.sph:A`
- "*3conv.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`
- "*8conv.trn*" for the 8 conversation two channel training condition.
- "*long.trn*" for the long interview two channel training condition; an example record looks like:

  `3310 nrkw.sph:A`
- "*3summed.trn*" for the 3 conversation summed-channel training condition; an example record looks like:

  `5047 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 13 index files, one for each of the evaluation tests. 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, short2, 3conv, 8conv, long, and 3summed. "*TestCondition*" refers to the test segment condition. Possible values for *TestCondition* are: 10sec, short3, long, and summed.

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 condition and for interview test segments.) For example, for the train on three conversations two channel and test on one conversation/short interview index file "3conv-short3.ndx" a record looks like: "7211 m nrbw B".

The records in these 13 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 13 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 – **10sec**, **short2**, **3conv**, **8conv**, **long**, or **3summed**
2. Adaptation mode. "**n**" for no adaptation and "**u**" for unsupervised adaptation.
3. The segment type of the test – **10sec**, **short3**, **long**, or **summed**
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 log 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 earlier year systems run on the 2008 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 31, 2008.

The evaluation data set will be distributed by NIST so as to arrive at participating sites on April 7, 2008.

The deadline for submission of evaluation results to NIST is May 8, 2008 at 11:59 PM, Washington time.

Initial evaluation results will be released to each site by NIST on May 19, 2008.

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 17-18, 2008 at McGill University in Montreal, Quebec, Canada. 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.