The NIST Year 2010 Speaker Recognition Evaluation Plan

1 INTRODUCTION

The year 2010 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 2010 evaluation will be similar to that of 2008 but different from prior evaluations by including in the training and test conditions for the core (required) test not only conversational telephone speech recorded over ordinary (wired or wireless) telephone channels, but also such speech recorded over a room microphone channel, and conversational speech from an interview scenario recorded over a room microphone channel. But unlike in 2008 and prior evaluations, some of the data involving conversational telephone style speech will have been collected in a manner to produce particularly high, or particularly low, vocal effort on the part of the speaker of interest. Unlike 2008, the core test interview segments will be of varying duration, ranging from three to fifteen minutes. 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. Systems will not know a-priori the elicited level of vocal effort in the conversational telephone style speech. 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, for conversational telephone style data of different levels of vocal effort, and for differing combinations of training and test data.

The 2010 evaluation will primarily use recently collected speech data from speakers not included in previous evaluations, but will also include some old and new conversational telephone speech segments from speakers in various past evaluations. Some new speech has recently been collected from speakers appearing in earlier evaluations, and this will support examination in this evaluation of the effect of the time interval between training and test on performance.

Unlike recent evaluations, all of the speech in the 2010 evaluation is expected to be in English, though English may not be the first language of some of the speakers included.

The 2010 evaluation will include an alternative set of parameters for the evaluation performance measure, to be implemented along with the parameter values used in the past evaluations. This is discussed in section 3.

The evaluation will include 9 different speaker detection tests defined by the duration and type of the training and test data. Three-conversation training will not be included this year, and the summed-channel 3-conversation training condition will be replaced with a summed-channel 8-conversation training condition. Because of the changed performance measure, and limited participation in past evaluations, the unsupervised adaptation condition will not be included in the 2010 evaluation.

The 2010 evaluation will also include a Human Assisted Speaker Recognition (HASR) test. This is described in section 11, and will consist of a limited number of trials, which will be a subset of the main evaluation trials. It is intended to test the capabilities of speaker recognition systems involving human expertise, possibly combined with automatic processing. Those participating in this test are not required to also do the core test otherwise required of all evaluation participants.

The evaluation will be conducted from March to May of 2010 (HASR data will be available in February of 2010). A follow-up workshop for evaluation participants to discuss research findings will be held June 24-26 in Brno, the Czech Republic. Specific dates are listed in the Schedule (section 12).

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 (except for HASR-only participants as described in section 11.2). 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 over multiple types of channels. 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 2010 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 fourteen 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 2010 is divided into 9 distinct and separate tests. (The HASR test, discussed in section 11, is not included here.) Each of these tests involves one of 4 training conditions and one of 3 test conditions. One of these tests (see section 2.2.3 below) is designated as the core test. Participants must do the core test (except those doing only the HASR 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.

[1] Send email to speaker_poc@nist.gov, or call 301/975-3605. Each site must complete the registration process by signing and returning the registration form, which is available online at: http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_agreement_v1.pdf .

2.2.1 Training Conditions

The training segments in the 2010 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 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 four 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. ***core***: 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 three to fifteen minutes total duration involving the interviewee (target speaker) and an interviewer. In the former case the designated channel may either be a telephone channel or a room microphone channel; the other channel will always be a telephone one. In the latter case the designated microphone channel will be the A channel, and most of the speech will generally be spoken by the interviewee, while the B channel will be that of the interviewer's head mounted close-talking microphone, with some level of speech spectrum noise added to mask any residual speech of the target speaker in it.

3. ***8conv***: Eight two-channel telephone conversation excerpts involving the target speaker on their designated sides.

4. ***8summed***: Eight summed-channel excerpts from telephone conversations of approximately five minutes total duration formed by sample-by-sample summing of their two sides. Each of these conversations will include both the target speaker and another speaker. These eight non-target speakers will all be distinct.

Word transcripts (always in English), 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, the provision of the interviewer's head-mounted close-talking microphone signal in a time aligned second channel, with speech spectrum noise added to mask any residual speech of the interviewee, is intended to assist systems in doing speaker separation, such as by using a speech detector to determine and remove from processing the time intervals where the interviewer is speaking.

2.2.2 Test Segment Conditions

The test segments in the 2010 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 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 putative target speaker to be recognized will be identified.

The three 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. ***core***: One two-channel telephone conversational excerpt, of approximately five minutes total duration, with the target speaker channel designated *or* a microphone recorded conversational segment of three to fifteen minutes total duration involving the interviewee (speaker of interest) and an interviewer. In the former case the designated channel may either be a telephone channel or a room microphone channel; the other channel will always be a telephone one. In the latter case the designated microphone channel will be the A channel, and most of the speech will generally be spoken by the interviewee, while the B channel will be that of the interviewer's head mounted close-talking microphone, with some level of speech spectrum noise added to mask any residual speech of the target speaker in it.

3. ***summed***: A summed-channel telephone conversation of approximately five minutes total duration formed by sample-by-sample summing of its two sides

Word transcripts (always in English), produced using an automatic speech recognition (ASR) system as described in section 2.2.1, will be provided for all test segments of each condition.

For the interview segments, the provision of the interviewer's head mounted close-talking microphone signal in a time aligned second channel, with speech spectrum noise added to mask any residual speech of the interviewee, is intended to assist systems in doing speaker separation, such as by using a speech detector to determine and remove from processing the time intervals where the interviewer is speaking.

2.2.3 Training/Segment Condition Combinations

The matrix of training and test segment condition combinations is shown in **Table 1**. Note that only 9 (out of 12) of the possible condition combinations will be included in this year's evaluation. Each test consists of a sequence of trials, where each trial consists

[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.

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 2010 evaluation. All participants (except those doing HASR only) are required to submit results for this test. Each participant may also choose to submit results for all, some, or none of the other 8 test conditions. For each test for which results are submitted, results for **all** trials must be included.

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

		Test Segment Condition		
		10sec	**core**	**summed**
Training Condition	**10sec**	optional		
	core	optional	**required**	optional
	8conv	optional	optional	optional
	8summed		optional	optional

3 PERFORMANCE MEASURE

Each trial of each test 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]

There will be a single basic cost model for measuring speaker detection performance, to be used for all speaker detection tests. In 2010, however, for two of the test conditions (including the core condition), there will be a new set of parameter values used to compute the detection cost over the test trials. The old parameter values used in previous evaluations will be used for the other conditions and will also be computed for these two conditions, thus supporting historical comparisons with past evaluations.

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

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

The parameters of this cost function are the relative costs of detection errors, C_{Miss} and $C_{\text{FalseAlarm}}$, and the *a priori* probability of the specified target speaker, P_{Target}.

For the core test, and for the "train on 8conv/test on core" condition (see section 2.2.3), the parameter values in **Table 2** will be used as the primary evaluation metric of speaker recognition performance.

Table 2: Speaker Detection Cost Model Parameters for the core and 8conv/core test conditions

C_{Miss}	$C_{\text{FalseAlarm}}$	P_{Target}
1	1	0.001[4]

These parameters differ from those used in prior NIST SRE evaluations. The parameters for the historical cost function, which will be the primary metric in 2010 for the other test conditions, are specified in **Table 3**.

Table 3: Speaker Detection Cost Model Parameters to be computed for all test conditions

C_{Miss}	$C_{\text{FalseAlarm}}$	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):

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

and

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

In addition to the actual detection decision, a score will also be required for each test hypothesis. This score should reflect the system's estimate of the probability that the test segment contains speech from the target speaker. Higher scores should indicate greater estimated probability that the target speaker's speech is present in the segment. The 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 scores across all target speakers.

NIST traditionally reports for each evaluation system the actual normalized C_{Det} score as defined above, and the minimum possible such score based on the DET curve, assuming perfect calibration. For historical continuity with respect to the core test and the train on 8conv/test on core condition NIST will report this minimum score for both the new and historical cost functions for these conditions.

The ordering of the scores is all that matters for computing the detection cost function, which corresponds to a particular application defined by the parameters specified above, 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

[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.

[4] Note that this application/evaluation prior is definitely not the same as the target prior of the evaluation corpus

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.[5]

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.[6]

4 EVALUATION CONDITIONS

Performance will be measured, graphically presented, and analyzed, as discussed in section 3, over all the trials of each of the 9 tests specified in section 2.2.3, 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 extrinsic (channel) factors such as telephone transmission type, and microphone type, will be examined. The effects of intrinsic (speaker) factors such as sex, age, and native English speaking status will also be examined. Speaking style factors, including conversational telephone vs. interview speech and the effects of high or low vocal effort will be investigated. We will also examine performance as a function of the time interval between the recording of training and test (target speaker) segments. 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 four training conditions. NIST is interested in examining how performance varies among these conditions for fixed test segment conditions.

The sex of each target speaker will be provided to systems (see section **Error! Reference source not found.**), but other emographic information will not be.

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.

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 or eight two-channel telephone conversational excerpts of a given speaker. (The first of these conditions will also include 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 Interview Segments

As discussed in section 2.2.1, one of the training conditions involves conversational interview segments (along with single two-channel telephone conversations). These will have varying durations of between three and fifteen minutes from a longer interview session. The effect of longer or shorter segment durations on performance may be examined. Two channels will be provided, the first from a microphone placed somewhere in the interview room, and the other from the interviewer's head mounted close-talking microphone with some level of speech spectrum noise added to mask any interviewee speech. 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.

4.1.4 Summed-channel Conversations

As discussed in section 2.2.1, one of the training conditions will consist of eight 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, are 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 eight.

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 may also be examined as a function of how many of the eight training conversations contain same-sex other speakers.

4.2 Test data

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

[5] The 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.

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

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 10-second 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 an 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 Interview Segments

As discussed in section 2.2.2, one of the test conditions involves conversational interview segments (along with single two-channel telephone conversations). These will have varying durations of between three and fifteen minutes from a longer interview session. The effect of longer or shorter segment durations on performance may be examined. Two channels will be provided, the first from a microphone placed somewhere in the interview room, and the other from the interviewer's head mounted close-talking microphone with some level of speech spectrum noise added to mask any interviewee speech. 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.

4.2.4 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 are 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 or most 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. If some trials are same-number, primary interest will be on results for different-number trials (see section 4.4 below), which may be contrasted with results on same-number trials.

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 Common Evaluation Condition

In each evaluation NIST has specified one or more common evaluation conditions, subsets 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 trial subsets are treated as the basic official evaluation outcomes. Because of the multiple types of training and test conditions in the 2010 core test, and the likely disparity in the numbers of trials of different types, it is not appropriate to simply pool all trials as a primary indicator of overall performance. Rather, the common conditions to be considered in 2010 as primary performance indicators will include the following subsets of all of the core test trials:

1. All trials involving interview speech from the same microphone in training and test
2. All trials involving interview speech from different microphones in training and test

3. All trials involving interview training speech and normal vocal effort conversational telephone test speech
4. All trials involving interview training speech and normal vocal effort conversational telephone test speech recorded over a room microphone channel
5. All different number trials involving normal vocal effort conversational telephone speech in training and test
6. All telephone channel trials involving normal vocal effort conversational telephone speech in training and high vocal effort conversational telephone speech in test
7. All room microphone channel trials involving normal vocal effort conversational telephone speech in training and high vocal effort conversational telephone speech in test
8. All telephone channel trials involving normal vocal effort conversational telephone speech in training and low vocal effort conversational telephone speech in test
9. All room microphone channel trials involving normal vocal effort conversational telephone speech in training and low vocal effort conversational telephone speech in test

4.5 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 2010 evaluation the test conditions involving normal vocal effort English language conversational telephone speech will be essentially identical those used in 2008. Conditions for interview speech in certain channels will also be quite similar. Thus it will be possible to make fairly direct comparisons between 2010 and 2008 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 2010 target speakers will include some used in the earlier evaluations, 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 2010 evaluation that also participated in 2008 are strongly encouraged to submit to NIST results for their (unmodified) 2008 (or earlier year) systems run on the 2010 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 2010 systems for use in similar comparisons in future evaluations.

5 DEVELOPMENT DATA

All of the previous NIST SRE evaluation data, covering evaluation years 1996-2008 may be used as development data for 2010. This includes the additional interview speech used in the follow-up evaluation to the main 2008 evaluation. All of this data, or just the 2008 data not already received, will be sent to prospective evaluation participants by the Linguistic Data Consortium on a hard drive (or DVD's for the 2008 follow-up data only), provided the required license agreement is signed and submitted to the LDC.[7]

A very limited amount of development data representing the high and low vocal effort telephone speech that is new for 2010 will also be made available. This will include three phone conversations for each of five speakers, one high vocal effort conversation, one low vocal effort conversation, and one normal vocal effort conversation. This data will be made available, by the end January of 2010, on a single CD-ROM to all registered 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] or of its earlier conversational telephone collection projects. The conversational telephone collections have invited participating speakers to take part in numerous conversations on specified topics with strangers. The platforms used to collect the data either automatically initiated calls to selected pairs of speakers, or allowed participating speakers to initiate calls themselves, with the collection system contacting other speakers for them to converse with. Speakers were generally encouraged to use different telephone instruments for their initiated calls.

The speech data for this evaluation (other than that for the HASR test, described in section 11) 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 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 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 whether or not the data was recorded over a telephone line, and whether or not the data is from an interview session.

---

[7] Find link at http://www.nist.gov/speech/tests/sre/2010/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

[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

The header will not contain information on the type of telephone transmission channel or the type of telephone instrument involved. Nor will the room microphone type be identified for the interview data.

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. The two-channel conversational telephone excerpts, both training and test, will all be approximately five minutes in duration. The interview segments, however, will be of varying duration between three and fifteen minutes. The primary channel of interest will be specified, and this will always be the A channel for interview segments. The second, non-primary, channel will contain the interlocutor's speech for telephone segments, and will contain the signal of the interviewer's head mounted, close-talking microphone with some level of speech spectrum noise added for interview segments. The header of each segment will indicate whether it comes from a telephone conversation or from an interview.

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

Table 4: Upper bounds on numbers of models by training condition

Training Condition	Max Models
10sec	3,000
core	6,000
8conv	1,000
8summed	1,000

6.2 Numbers of Test Segments

Table 5 provides estimated upper bounds on the numbers of segments to be included in the evaluation for each test condition.

Table 5: Upper bounds on numbers of segments by test condition

Test Conditions	Max Segments
10sec	6,000
core	25,000
summed	6,000

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 750,000.

7 EVALUATION RULES

Note that rules for HASR-only participants are specified in section 11.

In order to participate in the 2008 speaker recognition evaluation a site must submit complete results for the core test condition 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 (modified rules for the HASR test are specified in section 11.2):

- Each decision is to be based only upon the specified test segment and target speaker model. Use of information about other test segments and/or other target speakers is **not** allowed.[13] For example:
  - Normalization over multiple test segments is **not** allowed.
  - Normalization over multiple target speakers is **not** allowed.
  - Use of evaluation data for impostor modeling is **not** allowed.
  - 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 2010 evaluation.
- 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 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

[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 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.

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 earlier year systems run on 2010 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, but these charts will not 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. Furthermore, publicly claiming to "win" the evaluation is **strictly prohibited**. Participants violating this rule will be excluded from future evaluations.

8 EVALUATION DATA SET ORGANIZATION

This section describes the organization of the evaluation data other than the HASR data, which will be provided separately to those doing this test.

The organization of the evaluation data will be:

- A top level directory used as a unique label for the disk: "**sp10-NN**" where NN is a digit pair identifying the disk
- Under which there will be four sub-directories: "**data**", "**test**", "**trials**", and "**doc**"

8.1 data Sub-directory

This directory will contain all of the speech data files to be used as model training or test segments. Its organization will not be explicitly described. Rather the files in it referenced in other sub-directories will include path-names as well as file names.

8.2 train Sub-directory

The "**train**" directory will contain four training files that define the models for each of the four training conditions. Each will have one record per line containing three or more fields. The first field is the model identifier. The second field identifies the gender of the model, either "m" of "f". The remaining fields specify the speech files in the **data** directory to be used to train the model. These each consist of the file path-name, specifying the subdirectories under the data directory, but not including "data/", and the file name itself, including the ".sph" extension. For the two channel training conditions, each list item also has appended a ":" and a character that specifies whether the target speaker's speech is on the "A" or the "B" channel of the speech file.

The four training files are named:

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

  `32324 f path-name/mrpvc.sph:B`
- "***core.trn***" for the 1 conversation/interview two channel training condition; an example record looks like:

  `42403 m path-name/mrpzt.sph:A`
- "***8conv.trn***" for the 8 conversation two channel training condition; each record includes eight training files each with an appended channel specifier
- "8***summed.trn***" for the 8 conversation summed-channel training condition; each record includes eight training files without channel specifiers

This directory may not be included on the hard drives distributed to evaluation participants, but rather distributed electronically.

8.3 trials Sub-directory

The "**trials**" directory will contain 9 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, core, 8conv, and 8summed. "*TestCondition*" refers to the test segment condition. Possible values for *TestCondition* are: 10sec, core, and summed.

Each record in a *TrainCondition-TestCondition.ndx* file contains three fields and defines a single trial. The first field is the model identifier. The second field identifies the gender of the model, either "*m*" or "*f*". The third field specifies the test segment under evaluation, located in the **data** directory. It consists of the file path-name, specifying the subdirectories under the data directory, but not including "data/", and the file name itself, including the ".sph" extension. This test segment name will not include the .sph extension. For the two channel test conditions this field also has appended a ":" and a character that specifies whether the speech of interest is on the "A" or the "B" channel of the speech file. (This will always be "A" for interview test segments.) Example records might look like:

- `72116 m path-name/nrbrw:B  or`
- `50773 f path-name/kpdpm`

This directory may not be included on the hard drives distributed to evaluation participants, but rather distributed electronically.

8.4 doc Sub-directory

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

This section does not apply to the HASR test, whose submission requirements are described separately (section 11.4).

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 9 test index files) must be provided to NIST in a separate file for each test condition using a standard ASCII format, with one record for each trial decision. The file name should be intuitively mnemonic and should be constructed as "SSSsss_N_traincondition_testcondition_isprimary_isllr", where

- SSSsss (3-6 characters) identifies the site
- N identifies the system.
- traincondition refers to the training condition and whose models are defined in the corresponding training file. Possible values for t*raincondition* are: 10sec, core, 8conv, and 8summed.
- testcondition refers to the test segment condition. Possible values for t*estcondition* are: 10sec, core, and summed.
- isprimary refers to the whether the submission is for the primary system in the test condition. Possible values for isprimary are: primary and alternate.
- isllr refers to whether the system scores can correctly be interpreted as log likelihood ratio values. Possible values are: llr and other

9.1 Format for Results

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

1. The training type of the test – **10sec**, **core**, **8conv**, or 8**summed**
2. The segment type of the test – **10sec**, **core**, or **summed**
3. The sex of the target speaker – **m** or **f**
4. The target model identifier
5. The test segment identifier
6. The test segment channel of interest, either "a" or "b"
7. The decision – **t** or **f** (whether or not the target speaker is judged to match the speaker in the test segment)
8. The 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.

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 results for earlier year systems run on the 2010 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 as part of the submission. 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 should be reported as a multiple of real-time for the data processed. This may be reported separately for each test. Sites must also describe the CPU(s) utilized and the amounts of memory used.

11 HASR TEST

The Human Assisted Speaker Recognition (HASR) test will contain a subset of the core test trials of SRE10 to be performed by systems involving, in part or in whole, human judgment to make trial decisions. The systems doing this test may include large amounts of automatic processing, with human involvement in certain key aspects, or may be solely based on human listening. The humans involved in a system's decisions may be a single person or a panel or team of people. These people may be professionals or experts in any type of speech or audio processing, or they may be simply "naïve" listeners. The required system descriptions (section 11.4) must include a description of the system's human element.

Forensic applications are among the applications that the HASR test serves to inform, but the HASR test should not be considered to be a true or representative "forensic" test. This is because many of the factors that influence speaker recognition performance and that are at play in forensic applications are controlled in the HASR test data, which are collected by the LDC following their collection protocols.

While fully automatic systems are expected to run the full core test (section 2.2.3), they may, if desired, also be run on the HASR trials in accordance with the procedures given here. While the HASR trials are a subset of the core trials, note the scoring procedure (see section 11.3 below) will be different (and simpler) in HASR than in the core test.

HASR is clearly a new type of test for NIST evaluations, and accordingly it should be viewed as a pilot test. If response to it in 2010 is favorable, it will be continued, and refined, in future evaluations.

11.1 Trials and Data

To accommodate different interests and levels of effort, two test sets will be offered, one with 15 trials (HASR1), and one with 150 trials (HASR2). HASR participants may choose to perform either test.

Because of the small numbers of trials in the HASR test set, the difficulty of the test will be increased by selection of difficult trials. Objective criteria will be used to select dissimilar test conditions for target trials and similar speakers for non-target trials.

11.2 Rules

The rules on data interaction as specified in section 7 not allowing human listening or transcript generation or other interaction with the data, do not apply, but the requirement for processing each trial separately and making decisions independently for each trial remains in effect. Specifically:

- Each decision is to be based only upon the specified test segment and target speaker model. Use of information about other test segments and/or other target speakers is **not** allowed.

This presents a dilemma for human interactions, however, because humans inherently carry forward information from prior experience. To help minimize the impact of this prior exposure on human judgments, the trials will be released sequentially via an online automatic procedure. The protocol for this sequential testing will be specified in greater detail in early 2010, but will basically work as follows:

- NIST will release the first trial for download to each participant.
- The participant will process that trial and submit the result to NIST in the format specified in section 11.4.
- NIST will verify the submission format, and then make the next trial available for download to the participant.

The training and test speech data for each trial may be listened to by the human(s) involved in the processing as many times and in any order as may be desired. The human processing time involved must be reported in the system descriptions (see section 11.4 below).

The rules on dissemination of results as specified in section 7 will apply to HASR participants,

System descriptions are required as specified in section 10. They may be sent to NIST at any time during the processing of the HASR trials, or shortly after the final trial is processed. They should also describe the human(s) involved in the processing, how human expertise was applied, what automatic processing algorithms (if any) were included, and how human and automatic processing were merged to reach decisions. Execution time should be reported separately for human effort and for machine processing (if relevant).

Because HASR is a pilot evaluation with an unknown level of participation, participating sites will not in general be expected to be represented at the SRE10 workshop. NIST will review the submissions, and most particularly the system descriptions, and will then invite representatives from those systems that appear to be of particular interest to the speaker recognition research community to attend the workshop and offer a presentation on their system and results. One workshop session will be devoted to the HASR test and to comparison with automatic system results on the HASR trials.

HASR is open to all individuals and organizations who wish to participate in accordance with these rules.

11.3 Scoring

Scoring for HASR will be very simple. Trial decisions ("true" or "false") will be required as in the automatic system evaluation. In light of the limited numbers of trials involved in HASR, we will simply report for each system the overall number of correct detections ($N_{correct}$ detections on N_{target} trials) and the overall number of correct rejections ($N_{correct}$ rejections on $N_{non\text{-}target}$ trials).

Scores for each trial will be required as in the automatic system evaluation, with higher scores indicating greater confidence that the test speaker is the target speaker. It is recognized, however, that when human judgments are involved there may only be a discrete and limited set of possible score values. In the extreme, there might only be two; e.g., 1.0 corresponding to "true" decisions and -1.0 corresponding to "false" decisions. This is acceptable. DET curves, or a discrete set of DET points will be generated, and compared with the performance of automatic systems on the same trial set.

For each submission, the system description (section 11.4) should specify how scores were determined. Where this is a discrete set, the meaning of each possible score should be explained. It should also be indicated whether the scores may be interpreted as log likelihood ratios. [14]

11.4 Submissions

HASR trial submissions should use the following record format, which is a somewhat shortened version of that specified in section 9.1:

1. The test condition – "HASR1" or "HASR2"
2. The target model identifier
3. The test segment identifier
4. The test segment channel of interest, either "a" or "b"
5. The decision as specified above in section 11.3
6. The score

12 SCHEDULE

The deadline for signing up to participate in the evaluation is March 1, 2010.

The HASR data set will become available for sequential distribution of trial data to registered participants in this test beginning on February 15, 2010

The evaluation data (other than the HASR data) set will be distributed by NIST so as to arrive at participating sites on March 29, 2010.

The deadline for submission of evaluation results (including all HASR trial results) to NIST is April 29, 2010 at 11:59 PM, Washington, DC time (EDT or GMT-4).

Initial evaluation results will be released to each site by NIST on May 14, 2010.

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).

[14] A possible description of multiple scoring classes, and how they might be viewed as corresponding to log likelihood ratios, is offered in "Forensic Speaker Identification", Taylor & Francis, 2002, by Philip Rose, on page 62.

The follow-up workshop will be held June 24-26, 2010 in Brno, the Czech Republic. This workshop will precede the Odyssey 2010 international workshop at this location.[15] All sites participating in the main evaluation (core test) must have one or more representatives in attendance to discuss their systems and results.

13 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.

[15] See http://www.speakerodyssey.com/