The NIST Year 2012 Speaker Recognition Evaluation Plan

1 INTRODUCTION

The year 2012 speaker recognition evaluation (SRE12) is the next in an ongoing series of speaker recognition evaluations conducted by NIST. These evaluations serve to support speaker recognition research and to calibrate the performance of speaker recognition systems. 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 basic task in NIST's speaker recognition evaluations is speaker detection, i.e., to determine whether a specified target speaker is speaking during a given segment of speech. While the basic task in SRE12 remains unchanged, SRE12 task conditions represent a significant departure from previous NIST SRE's. In previous evaluations, the evaluation test set, which is released at the beginning of the evaluation period, has contained both the training data and the test data. In SRE12, however, most target speakers will be taken from previous SRE corpora, with the training data being provided to evaluation participants at the time of registration, well in advance of the evaluation period. Furthermore, in SRE12 the training data for each such target speaker comprises all of the data from previous SRE's, both training and test, and will include a fairly large number of speech segments taken from multiple recording sessions. Similar to SRE10, all of the speech in SRE12 is expected to be in English, though English may not be the first language of some of the speakers included.

Participation in the evaluation is invited for all who find the task and evaluation of interest and are able to comply with the evaluation rules set forth in this plan. Further, participants must be represented at the evaluation workshop, to be held in Orlando, Florida, USA on December 11-12, 2012. To register, please fill out and follow the instructions on the registration form.[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 2012 speaker recognition task is **speaker detection**, as described briefly in the introduction. This has been NIST's speaker recognition task over the past sixteen years. The task is to determine whether a specified target speaker is speaking during a given segment of speech. More explicitly, one or more samples of speech data from a speaker (referred to as the "target" speaker) are provided to the speaker recognition system. These samples are the "training" data. The system uses these data to create a "model" of the target speaker's speech. Then a sample of speech data is provided to the speaker recognition system. This sample is referred to as the "test" segment. Performance is judged according to how accurately the test segment is classified as containing (or not containing) speech from the target speaker.

SRE12 includes an optional evaluation of human-assisted speaker recognition (HASR12). The HASR12 task and evaluation is described in section 11.

In previous NIST evaluations the system output consisted of a detection decision and a score representing the system's confidence that the target speaker is speaking in the test segment. NIST has recently encouraged expressing the system output score as the natural logarithm of the estimated likelihood ratio, defined as:

$$\text{LLR} = \log (\text{pdf}(\text{data} \mid \text{target hyp.}) / \text{pdf}(\text{data} \mid \text{non-target hyp.}))$$

Because of the general community acceptance of using the log likelihood ratio as a score, in SRE12 NIST is requiring that the system output score for each trial be the natural logarithm of the likelihood ratio. Further, since the detection threshold may be determined from the likelihood ratio, system output in SRE12 will not include a detection decision.

2.2 Task Conditions

The speaker detection task for 2012 is divided into 9 distinct and separate tests (not counting the HASR test discussed in section 11). Each of these tests involves one of three training conditions and one of five test conditions. One of these tests is designated as the core test which all participants must complete (except for those doing only the HASR test). Participants may also choose to do one or more of the other tests. Results must be submitted for ***all*** trials in each test for which results are submitted.

In SRE12 knowledge of all targets is allowed in computing each trial's detection score. This differs from all previous SRE's. Previously systems were restricted to use only knowledge of the single target speaker that was specified as the trial target. To test the effect of this knowledge on system performance, the SRE12 evaluation data will also include data from new speakers (for the non-target trials), to provide a basis for comparison of performance under the two conditions (of having versus not having knowledge of non-target speakers).

All of the speech in SRE12 will be in English.

2.2.1 Training Conditions

Target speaker training data in SRE12 will comprise all of the speech data associated with the target speakers chosen from the LDC speaker recognition speech corpora used in previous SRE's. There will be no more than 2,250 target speakers. A list of target speakers will be supplied, along with the relevant LDC speech corpora, when participants register to participate in the SRE12 evaluation. In addition, some previously unexposed target speakers, along with their relevant speech training data, will be supplied at evaluation time. Some of these additional speakers may have only one training segment. It should be noted that no

[1]. http://nist.gov/itl/iad/mig/upload/registration_sre12-v0.pdf .

For more information, please send email to speaker_poc@nist.gov additional restrictions are placed upon the use of these previously unexposed target speakers; in particular, knowledge of these targets is allowed in computing each trial's detection score.

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

1. ***Core***: All speech data, including microphone and telephone channel recordings, available for each target speaker.

2. ***Telephone***: All telephone channel speech data available for each target speaker. This condition prohibits the use in any way of the microphone data from any of the designated target speakers. Microphone data from speakers other than those specified as target speakers may be used, for example, for background models, speech activity detection models, etc.

3. ***Microphone***: All microphone channel speech data available for each target speaker. This condition prohibits the use in any way of the telephone data from any of the designated target speakers. Telephone data from speakers other than those specified as target speakers may be used, for example, for background models, speech activity detection models, etc.

2.2.2 Test Segment Conditions

The test segments in the 2012 evaluation will be mostly excerpts of conversational telephone speech but may contain interviews. There will be one required and four optional test segment conditions:

1. ***Core***: One two-channel excerpt from a telephone conversation or interview, containing nominally between 20 and 160 seconds of target speaker speech. Some of these test segments will have additive noise imposed.

2. ***Extended***: The test segments will be the same as those used in ***Core***. The number of trials in Extended tests will exceed the number of trials in Core tests.

3. ***Summed***: A summed-channel excerpt from a telephone conversation or interview, containing nominally between 20 and 160 seconds of target speaker speech formed by sample-by-sample summing of its two sides.

4. ***Known***: The trial list for the known test segment condition will be the same as in ***Extended.*** The system should presume that all of the non-target trials are by known speakers.

5. ***Unknown***: The trial list for the unknown test segment condition will be the same as in ***Extended.*** The system should presume that all of the non-target trials are by unknown speakers.

2.2.3 Training/Test Segment Condition Combinations

The matrix of training and test segment condition combinations is shown in Table 1. Note that only 9 (out of 15) of the possible condition combinations will be included in the 2012 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 highlighted text labeled "required (Core Test)" in Table 1 is the **Core test** for the 2012 evaluation, and all participants (except those completing HASR only) are required to complete the core test. Participants are encouraged, but not required, to submit results for one or more of the other eight optional tests. 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

		Training Condition		
		Core	**Microphone**	**Telephone**
Test Segment Condition	**Core**	**required (Core test)**	optional	optional
	Extended	optional	optional	optional
	Summed	optional		
	Known	optional		
	Unknown	optional		

3 PERFORMANCE MEASURES

The primary performance measure for SRE12 will be a detection cost, defined as a weighted sum of miss and false alarm error probabilities. There are two significant changes from past practice regarding how this primary cost measure will be computed in SRE12:

- First, no detection decision output is needed because trial scores are required to be log likelihood ratios. Thus the detection threshold is a known function of the cost parameters, and so the trial detection decisions are determined simply by applying this threshold to the trials' log likelihood scores.
- Second, the primary cost measure in SRE12 will be a combination of two costs, one using the cost parameters from SRE10 and one using a greater target prior. This is intended to add to the stability of the cost measure and to increase the importance of good score calibration over a wider range of log likelihoods.

The cost function used in SRE12 to compute costs accounts separately for known and unknown non-target speakers:

$$\begin{aligned} C_{\mathrm{Det}} \;=\; & \mathrm{C}_{\mathrm{Miss}} \times \mathrm{P}_{\mathrm{Target}} \times P_{\mathrm{Miss|Target}} \\ & + \mathrm{C}_{\mathrm{FalseAlarm}} \times (1\text{-}\mathrm{P}_{\mathrm{Target}}) \\ & \quad \times (P_{\mathrm{FalseAlarm|KnownNonTarget}} \times \mathrm{P}_{\mathrm{Known}} \\ & \quad\quad + \mathrm{P}_{\mathrm{FalseAlarm|UnknownNonTarget}} \times (1\text{-}\mathrm{P}_{\mathrm{Known}})) \end{aligned}$$

The parameters of this performance measure are:

- $\mathrm{C}_{\mathrm{Miss}}$, the cost of a miss,
- $\mathrm{C}_{\mathrm{FalseAlarm}}$, the cost of a false alarm,
- $\mathrm{P}_{\mathrm{Target}}$, the *a priori* probability that the segment speaker is the target speaker[2], and
- $\mathrm{P}_{\mathrm{Known}}$, the *a priori* probability that the non-target speaker is one of the evaluation target speakers[3].

[2] Note that $\mathrm{P}_{\mathrm{Target}}$, the target prior used to compute system performance, is not the same as the prior probability of target trials in the corpus.

Table 2: Speaker Detection Cost Model Parameters

		C_{Miss}	C_{FA}	$P_{Target\text{-}A1}$	$P_{Target\text{-}A2}$	P_{Known}
Test Segment Condition	**Core** **Extended** **Summed**	1	1	0.01	0.001	0.5
	Known					1
	Unknown					0

To improve the intuitive meaning of C_{Det}, it will be normalized by dividing it by the best cost that could be obtained without knowledge of the input data:

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

where $C_{Default} = C_{Miss} \times P_{Target}$

Thus

$$C_{Norm} = P_{Miss|Target} +$$

$$\beta \times \begin{matrix} P_{Known} \times P_{FalseAlarm|KnownNontarget} + \\ 1 - P_{Known} \times P_{FalseAlarm|UnknownNontarget} \end{matrix}$$

where $\beta = \frac{C_{FalseAlarm}}{C_{Miss}} \; \frac{1-P_{target}}{P_{target}}$

Actual detection costs will be computed from the trial scores by applying detection thresholds of log(β) for the two values of β, with β_{A1} (for $P_{Target\text{-}A1}$) being 99 and β_{A2} (for $P_{Target\text{-}A2}$) being 999.

The primary cost measure for SRE12 is defined as:

$$C_{primary} = \frac{C_{Norm}\ \beta_{A1} + C_{Norm}\ \beta_{A2}}{2}$$

Also, a minimum detection cost will be computed by using the detection thresholds that minimize the detection cost.

In addition to the primary performance measure, an alternative, information theoretic measure will be computed that considers how well all scores represent the likelihood ratio and that penalizes for errors in score calibration. This performance measure is defined as:

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

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]

A useful variant of C_{llr} is to limit evaluation to the low false alarm region. The motivation for doing this is to improve the informative power of C_{llr} in the low false alarm region. This is important because the large majority of non-target scores, which are of no interest (since they are correctly rejected), nonetheless have a major influence on the computed value of C_{llr}. A simple way of focusing the low false alarm region is to limit the trials in the calculation of C_{llr} to only those for which P_{Miss} is greater than the minimum over the range of interest. A reasonable minimum value of P_{Miss}, given the current state of technology, is 10%. Using this value, this variant of C_{llr} may be called $C_{llr\text{-}M10}$.

In order to foster interest in speaker recognition performance measurement, NIST would like to encourage participants to propose additional performance measures for use in future NIST SRE's. Sites wishing to submit proposals should send email to speaker_poc@nist.gov for details.

4 EVALUATION CONDITIONS

Performance will be measured, graphically presented, and analyzed, as a function of various conditions of interest. These will include the training and test conditions.

For all training conditions, English language ASR transcriptions of all data will **NOT** be provided along with the audio data. This is a change from recent SRE's, where ASR transcripts were provided.

4.1.1 Two-channel Conversations

As mentioned in section 2.2.2, there will be test segments each consisting of an excerpt from a two-channel telephone conversation. These will vary in duration and amount of speech. The effect of longer or shorter segment durations on performance may be examined. The excision points will be chosen to minimize the likelihood of including partial speech turns.

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

4.1.2 Interview Segments

As mentioned in section 2.2.2, there will be test segments each consisting of an excerpt from an interview. These will vary in duration and amount of speech. 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. Information on the microphone type of the first channel will not be available to systems.

The microphone channel data will be provided in 16-bit linear-pcm form that differs from the telephone data provided.

4.1.3 Summed test segment condition

As mentioned in section 2.2.2, there will be test segments each consisting of an excerpt from a telephone conversation where the two sides of each conversation, in which both the target speaker and another speaker participate, are summed together. Thus the challenge is to be able to correctly detect the target speaker despite the presence of speech from another speaker.

4.2 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 (or of at least one test segment speaker for the summed test segment condition), will be the same as that of the target speaker model. Performance will be

[3] Note that P_{Known}, the known non-target prior used to compute system performance, is not the same as the prior probability of known non-target trials in the corpus.

[4] 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.

reported separately for males and females and also for both sexes pooled.

This evaluation will include an examination of the effects of channel on recognition performance. This will include in particular the comparison of performance involving telephone segments with that involving microphone segments.

For trials involving microphone test 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 presence of the test microphone in the training data.

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, which may be contrasted with results on same-number trials.

Some of the test segments will include additive noise (noise added as a post-processing step after recording) or will be recorded in an intentionally noisy environment or both. The impact of noise on performance will be examined in this evaluation.

The Core test will include relatively large amounts of training data distributed in advance of the evaluation period as well as limited training data distributed at the start of the evaluation period. NIST will compare performance of speakers in these training conditions.

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)

4.3 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 test conditions in the 2012 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 2012 as primary performance indicators will include the following subsets of all of the core test trials:

1. All trials involving multiple segment training and interview speech in test without added noise in test
2. All trials involving multiple segment training and phone call speech in test without added noise in test
3. All trials involving multiple segment training and interview speech with added noise in test
4. All trials involving multiple segment training and phone call speech with added noise in test
5. All trials involving multiple segment training and phone call speech intentionally collected in a noisy environment in test

4.4 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. This is particularly problematic for SRE12, given the change in task conditions as discussed in section 1. For the 2012 evaluation the training condition released at evaluation time and consisting of a single segment will be similar to the task condition in 2010. Thus it will be possible to make relatively direct comparisons between 2012 and 2010 in this limited circumstance.

To help address the desire to make comparison with previous efforts, sites participating in the 2012 evaluation that also participated in 2010 are encouraged to submit to NIST results for their (unmodified) 2010 (or earlier year) systems run on the 2012 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 2012 systems for use in similar comparisons in future evaluations.

5 DEVELOPMENT DATA

All of the previous NIST SRE evaluation data, covering evaluation years 1996-2010, may be used as development data for 2012. This includes the additional interview speech used in the follow-up evaluation to the main 2008 evaluation. All of this data, or just the data not already received, will be sent to prospective evaluation participants by the Linguistic Data Consortium on one or more hard drives or DVD’s, provided the required license agreement is signed and submitted to the LDC.[5] This development data includes the SRE12 training data for most of the target speakers (training data for some target speakers will be released at the beginning of the evaluation period along with the test data).

Participating sites may use other speech corpora to which they have access for development. Such corpora must be described in the site’s system description (section 10).

6 EVALUATION DATA

The test 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 USB hard drive. The LDC license agreement described in section 5, which all sites must sign to participate in the evaluation, will govern the use of this data for the evaluation.

[5] http://nist.gov/itl/iad/mig/upload/2012_NIST_SRE_Data_Agreement-v3.pdf

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.[6]

All telephone channel test data will be encoded as 8-bit μ-law speech samples and all microphone channel data will be encoded as 16-bit linear pcm. All test data will be stored in separate SPHERE[7] files. In addition to the information that is contained in a standard SPHERE header, evaluation data will include in the header entries for channel (mic or tel) and speaking style (interview or phonecall). The SPHERE header will not contain information on the type of telephone transmission channel or the type of telephone instrument or microphone involved.

6.1 Numbers of Test Segments

Table 3 provides upper bounds on the numbers of segments[8] to be included in the evaluation for each test condition.

Table 3 Upper bounds on the number of test segments

Test Data	Max Segments
Core/Extended	100,000
Summed	100,000

6.2 Numbers of Trials

Table 4 gives upper bounds on the numbers of trials to be included in the evaluation for each test condition.

The trials for each of the speaker detection tests will be specified in separate index files. These will be text files in which each record specifies the target speaker id, the test segment, and the side for a particular trial.

Table 4 Upper bounds on the number of trials

Test Conditions	Max Trials
Core	1,000,000
Extended (optional)	100,000,000
Summed (optional)	1,000,000

7 EVALUATION RULES[9]

In order to participate in the 2012 speaker recognition evaluation a site must submit complete results for the required test condition as specified in section 2.2. A test submission is complete if and only if it includes a score for every trial in the test.

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 score is to be based only upon the training data and the specified test segment. Information about other test segments (including for example normalization of scores over multiple test segments) is **not** allowed.[10]
- The use of manually produced transcripts or other human-created information is **not** allowed.
- Knowledge of the sex of the *target* speaker **is** allowed. Note that no cross-sex trials are planned, but that summed-channel segments may include speech from an opposite sex speaker.
- Listening to the evaluation test data, or any other human interaction with the test data, is **not** allowed. It should be noted, however, that human interaction with the evaluation **training data** is permitted.
- 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 2012 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

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

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

[8] A segment is a single unique audio file and includes both sides of the conversation, either as two separate channels or a single summed channel.

[9] Rules for the HASR evaluation are specified in section 11.

[10] 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.

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 the HASR test.

The organization of the evaluation data will be:

- A top level directory used as a unique label for the disk: "**sp12-NN**" where NN is a digit pair identifying the disk
- Under which there will be three sub-directories: "**data**", "**test**", and "**doc**"

8.1 data Sub-directory

The "**data**" directory will contain all of the speech test segments as well as any training segments not previously released. Its organization will not be explicitly described. Rather the files in it will be referenced in other sub-directories.

8.2 train Sub-directory

The "**train**" directory will contain a table of all target speakers that provides links to their speech files located either in the data directory or in the training data distributed by the LDC. This table is a superset of the information that was also provided to evaluation participants at the time of registration.

8.3 trials specification

There will be three index files, named **core**.ndx, **summed**.ndx, and **extended**.ndx, to be used for the identically named test conditions. (The extended.ndx file will also be used for the **known** and **unknown** test conditions.)

Each record in the index files will correspond to one trial and will contain three comma separated fields:

1. The first field is a target speaker identification string.
2. The second is the file name of a test segment within the data directory.
3. The third is the channel designator (either "A" or "B").

These index files will be distributed to evaluation participants via FTP.

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).

Results for each test must be provided to NIST in a single separate file using standard ASCII format, with one record for each trial.

Each file record must document its trial output with 4 comma separated fields:

1. The target speaker identification string
2. The test segment file name
3. The channel designator
4. The score. In SRE12 the score is required to represent the system's estimate of the log likelihood ratio (i.e., the natural logarithm of the target/non-target likelihood ratio).

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

New to SRE12, NIST will be releasing software that verifies a submission's validity. More information on the submission checker software will be made available to participants prior to the start of the evaluation.

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. This should include a description of any human interaction with the evaluation training data.

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 2012 data. Please note that a "primary" system for each test completed must be identified as part of the submission. Sites are welcome to present descriptions of and performance results for additional systems beyond those submitted 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 SRE12 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 11.2) 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.

11.1 Trials and Data

To accommodate different interests and levels of effort, two test sets will be offered, one with 20 trials (HASR1), and one with 200 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.

Data used in the 2010 HASR pilot evaluation will be made available upon request to any site participating in the 2012 HASR evaluation.

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:

- Use of information about other trials 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 mid-2012, but will basically work as follows:

- NIST will release the first trial as a three-field record as specified in section 8.3 for the core index file.
- 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 remains a pilot evaluation with an unknown level of participation, participating sites will not in general be expected to be represented at the SRE12 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 ("**same**" if the segment speaker is judged to be the target speaker, otherwise "**different**") will be required. 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 for 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 "same" decisions and -1.0 corresponding to "different" decisions. This is acceptable. The scores will be used to produce *Detection Error Tradeoff (DET)* curves[11], or a discrete set of DET points, and compared with the performance of automatic systems on the same trial set.

For each submission, the system description (section 11.2) 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. [12]

11.4 Submissions

HASR trial submissions should use the following record format:

1. The test condition – "HASR1" or "HASR2"
2. The trial index number (1 through 20 for HASR1, 1 through 200 for HASR2)
3. The decision as specified above in section 11.3
4. The score as specified above in section 11.3

12 SCHEDULE

The deadline for signing up to participate in the evaluation is August 1, 2012.

The HASR data set will become available for sequential distribution of trial data to registered participants in this test beginning on August 1, 2012

The evaluation data (other than the HASR data) set will be distributed by NIST so as to arrive at participating sites on September 24, 2012.

The deadline for submission of evaluation results (including all HASR trial results) to NIST is October 15, 2012 at 11:59 PM, Washington, DC time (EDT or GMT-5).

Initial evaluation results will be released to each site by NIST on November 5, 2012.

The deadline for site workshop presentations to be supplied to NIST in electronic form for inclusion in the workshop proceedings is December 3rd, 2012.

[11] For details regarding DET curves, see: http://www.itl.nist.gov/iad/mig/publications/storage_paper/det.pdf

[12] 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 deadline for registration and room reservations for the workshop is to be determined.

The follow-up workshop will be held December 11th-December 12th, 2012 in Orlando, Florida, USA. All sites participating in the main evaluation 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 speaker – The hypothesized speaker of a test segment, one for whom a model has been created from training data.

Non-target 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 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 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.