Plan for Follow-up Evaluation to SRE08

1 INTRODUCTION

This evaluation is a follow-up to the NIST 2008 Speaker Recognition Evaluation (SRE08). It is intended to explore further one of the new test conditions included in SRE08. This is the test condition involving training and test on short conversational interview segments, where short means segments of approximately three minutes in duration and includes mainly speech from an interview subject of interest, as well as some speech of an interviewer. The recording channels include a variety of microphone types placed in the interview room.

The microphones included in the SRE08 interview data were all ones for which a small amount of development was made available prior to that evaluation. This evaluation will include test segments recorded on additional microphones included in the Mixer 5 collection for which no development data has been released. A key aim is to examine performance on speech recorded over these heretofore unexposed channels.

This evaluation will reuse some of the interview training data of SRE08, which participating sites should already have. The same model identifiers as in SRE08 will be utilized. The test segment data will be newly supplied to sites. These test segments will involve the same interview target speakers and interview sessions used in the earlier evaluation. Some will involve the same microphone channels as used in SRE08; others will be from microphones not used previously.

The evaluation will be conducted in August and September of 2008. Specific dates are listed in the Schedule (section 11).

Participation in the evaluation is invited from all sites that participated in SRE08 and find the task in this evaluation of interest. Participating sites must submit results generated by running their unaltered SRE08 primary systems on the evaluation data, and may optionally submit results for additional systems. Each site must follow the evaluation rules set forth in this plan (section 7). For more information, and to register to participate in the evaluation, please contact NIST.[1]

2 TECHNICAL OBJECTIVE

This follow-up evaluation focuses on speaker detection in the context of conversational interview type speech. It is designed in particular to measure the performance of SRE08 systems in previously unexposed test segment channel conditions.

2.1 Task Definition

The task is to determine whether a specified speaker is speaking during a given segment of conversational interview speech.

2.2 Task Condition

The single test in this follow-up evaluation involves training and test on conversational interview segments of about three minutes in duration. These segments involve primarily speech of a target interview subject along with some speech of the person conducting the interview. Results must be submitted for all trials included in this test.

2.2.1 Training Condition

Each training segment will consist of a conversational excerpt of approximately three minutes total duration involving the target speaker and an interviewer. Most of the speech will generally be spoken by the target speaker. There will be no prior removal of any intervals of silence, and the segment will be single channel and 8-bit mu-law encoded.

2.2.2 Test Segment Condition

Each test segment will consist of a conversational excerpt of approximately three minutes total duration involving the target speaker and an interviewer. Most of the speech will generally be spoken by the target speaker. There will be no prior removal of any intervals of silence, and the segment will be single channel and 8-bit mu-law encoded.

2.2.3 Additional Metadata Provided

English language word transcripts, produced using an automatic speech recognition (ASR) system, will be provided for all training and test segments. These transcripts will, of course, be errorful, with English word error rates typically in the range of 15-30%. ASR output for two different recording channels will be provided, and these will in general be different recording channels from that used in the segment. The two ASR channels will be from the lavalier microphones worn by the target and by the interviewer. The ASR transcripts provided may well 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.

Also provided for each training and test segment will be files giving the estimated intervals where the target speaker is speaking, as determined by an energy-based segmenter utilizing the audio signals from the lavalier microphones worn by the two speakers. Systems may limit their processing to these intervals, or they may choose to process the full segments and do their own speaker separation processing.

3 PERFORMANCE MEASURE

The performance measure is unchanged from SRE08. This section is unchanged from the corresponding section of the SRE08 plan.

There will be a single basic cost model for measuring speaker detection performance, to be used for all speaker detection tests. For each test, a detection cost function will be computed over the sequence of trials provided. Each trial must be independently judged as "true" (the model speaker speaks in the test segment) or "false" (the model speaker does not speak in the test segment), and the correctness of these decisions will be tallied.[2]

[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.nist.gov/speech/tests/sre/2008/sre08_registration.pdf

[2] This means that an explicit speaker detection decision is required for each trial. Explicit decisions are required because the task of

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

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

The parameters of this cost function are the relative costs of detection errors, C_{Miss} and $\text{C}_{\text{FalseAlarm}}$, and the *a priori* probability of the specified target speaker, P_{Target}. The parameter values in **Table 1** will be used as the primary evaluation of speaker recognition performance for all speaker detection tests.

Table 1: Speaker Detection Cost Model Parameters for the primary evaluation decision strategy

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

$$\text{C}_{\text{Default}} = \min \left\{ \begin{array}{l} \text{C}_{\text{Miss}} \times \text{P}_{\text{Target}}\ , \\ \text{C}_{\text{FalseAlarm}} \times (1 - \text{P}_{\text{Target}}) \end{array} \right\}$$

and

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

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

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

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

Sites are asked to specify if their scores may be interpreted as log likelihood ratio estimates.

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

$$\text{C}_{\text{llr}} = 1 \ / \ (2 * \log 2) * (\sum \log(1+1/s)/\text{N}_{\text{TT}}) + (\sum \log(1+s))/\text{N}_{\text{NT}})$$

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.[3]

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.[4]

4 EVALUATION CONDITIONS

Performance will be measured, graphically presented, and analyzed, as discussed in section 3, over all the trials and over subsets of these trials of particular evaluation interest. In particular, the effects of microphone type on performance will be examined.

4.1 Training Data

As noted above, there will be a single training condition involving interview segments approximately three minutes in duration, selected from longer interview sessions. The excision points will be chosen so as not to include partial speech turns. The single channel of audio provided for each segment will be from a microphone placed somewhere in the interview room. Information on the microphone type being utilized in each segment will not be available to systems.

The sex of each target speaker will be provided to systems. All speech will be in English.

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.

Time estimates of the intervals where the interview subject is speaking, as determined by an energy-based segmenter, will also be provided for all audio data. Systems may utilize this information or choose not to utilize it as they wish.

4.2 Test data

As noted above, there will be a single test condition involving interview segments approximately three minutes in duration, selected from longer interview sessions. The excision points will be chosen so as not to include partial speech turns. The single channel of audio provided for each segment will be from a microphone placed somewhere in the interview room. Information on the

determining appropriate decision thresholds is a necessary part of any speaker detection system and is a challenging research problem in and of itself.

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

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

microphone type being utilized in each segment will not be available to systems.

All speech will be in English.

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.

Time estimates of the intervals where the interview subject is speaking, as determined by an energy-based segmenter, will also be provided for all audio data. Systems may utilize this information or choose not to utilize it as they wish.

4.3 Factors Affecting Performance

All trials will be *same-sex* trials. This means that the sex of the test segment speaker 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.

It will be of interest to examine the effect of the different microphone types tested on performance, and most particularly, the effect on performance of the use of previously unseen test segment microphone types.

5 DEVELOPMENT DATA

All of the previous NIST NRE evaluation data, covering evaluation years 1996-2006 may be used as development data.

Note that no development data is being provided that corresponds to the test segment microphones to be included in this evaluation that were included in SRE08.

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

The interview data used in this evaluation was collected by the Linguistic Data Consortium (LDC) as part of its Mixer 5 project.[5] The LDC license agreement that sites were required to sign to participate in SRE08 will govern the use of this data for the evaluation.

The training data was provided to participants previously in SRE08, while the test segment data will be distributed to evaluation participants by NIST on a firewire drive. It will also include the corresponding ASR transcript data and the files of estimated speech intervals of interview target speakers, in the same form as supplied previously for SRE08.

All training and test segments will be stored as 8-bit μ-law speech signals in separate SPHERE[6] files. The SPHERE header of each such file will contain some auxiliary information as well as the standard SPHERE header fields. This auxiliary information will include the language of the interview, which will always be English.

6.1 Number of Models

The models used will be the short Mixer 5 models used in SRE08. The number of such models is on the order of 1,500.

6.2 Number of Test Segments

The total number of test segments included in the evaluation will not exceed 10,000.

6.3 Number of Trials

Separate files will list all of the male model and male test segment identifiers included in the evaluation, and all of the female model and female test segment identifiers included in the evaluation. The trials to be processed will then be all same gender combinations of a model and test segment (full matrix). The total number of trials will not exceed 10,000,000.

7 EVALUATION RULES

Each system for which results are submitted must include decisions and scores for all trials of the evaluation. Each site must submit results for its SRE08 primary system without any alterations to this system using the models created previously. Each site may also, submit full results for other systems, which may include secondary SRE08 systems, modified versions of SRE08 systems, or newly developed systems. This is optional, but encouraged.

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

- Each decision is to be based only upon the specified test segment and target speaker model. Use of information about other test segments and/or other target speakers is **not** allowed.[7] For example:
  - Normalization over multiple test segments is **not** allowed, except as permitted for the unsupervised adaptation mode condition.
  - Normalization over multiple target speakers is **not** allowed.
  - Use of evaluation data for impostor modeling is **not** allowed, except as permitted for the unsupervised adaptation mode condition.
  - Speech data from past evaluations may be used for general algorithm development and for impostor modeling, but may not be used directly for modeling target speakers of the 2008 evaluation.
- 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. There will be no cross-sex trials.
- Listening to the evaluation data, or any other human interaction with the data, is **not** allowed before all test results have been submitted. This applies to training data as well as test segments.

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

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

[7] 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.

- Knowledge of any information available in the SPHERE header **is** allowed.

The following general rules about dissemination of results will also apply for all participating sites:

- Participants may publish or otherwise disseminate their own results.
- NIST will generate and place on its web site charts of all system results for conditions of interest and, unlike past practice, these charts may contain the site names of the systems involved. Participants may publish or otherwise disseminate these charts, unaltered and with appropriate reference to their source
- Participants may not publish or otherwise disseminate their own comparisons of their performance results with those of other participants without the explicit written permission of each such participant. Participants violating this rule will be excluded from future evaluations

8 EVALUATION DATA SET ORGANIZATION

The organization of the evaluation data on the firewire drive will be:

- A top level directory used as a unique label for the disk: "**sre08_followup-1**"
- Under which there will be five sub-directories: "**test**", "**asr**", "**vad** " "**trials**", and "**doc**"

8.1 test Subdirectory

The "**test**" directory will contain two subdirectories denoted "**male**" and "**female**". Each will contain single-channel short interview segments involving speakers of the indicated gender. The file names will be arbitrary ones of five characters along with a ".sph" extension.

8.2 asr trials Subdirectory

The "**asr**" directory will contain two subdirectories denoted "**male**" and "**female**". Each will contain asr transcript files for the test segments of the corresponding gender. The file names will be of five characters and correspond to those of the test subdirectories along with a ".cfm" extension.

8.3 vad Subdirectory

The "**vad**" directory will contain two subdirectories denoted "**male**" and "**female**". Each will contain files of estimated target speaker speech intervals, based on voice activity detection software provided to NIST, for the test segments of the corresponding gender. The file names will be of five characters and correspond to those of the test subdirectories along with a ".vad" extension.

8.4 trials Subdirectory

The "**trials**" subdirectory will contain two subdirectories denoted "**male**" and "**female**". Each will contain two text files denoted "**models**" and "**test_segments.**

The "**models**" files will contain lists of model identifiers, one per record. These identifiers will be a subset of the short-2 Mixer 5 model identifiers of the SRE08 evaluation.

The "**test_segments**" files will contain lists of the test segments found in the **test** subdirectory for the corresponding gender, one per record.

The trials for this test will consist of all pairings of a model from the **models** file and a test segment from the **test_segments** file for each gender (male and female).

8.5 doc Subdirectory

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

9 SUBMISSION OF RESULTS

Participating must report results for the test in its entirety. These results must be provided to NIST in a single file using a standard ASCII format, with one record for each trial decision. The file name should be intuitively mnemonic and should be constructed as "sitename_N", where

- sitename identifies the site (6 characters maximum)
- N identifies the system ("1" for the required SRE08 primary system)

9.1 Format for Results

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

1. The sex of the target speaker – **m** or **f**
2. The target model identifier
3. The test segment identifier
4. The decision – **t** or **f** (whether or not the target speaker is judged to match the speaker in the test segment)
5. The confidence score (where larger scores indicate greater likelihood that the test segment contains speech from the target speaker)

9.2 Means of Submission

Submissions should be made via ftp. The appropriate addresses for submissions will be supplied to participants receiving evaluation data. Sites should also indicate if it is the case that the confidence scores in a submission are to be interpreted as log likelihood ratios.

10 SYSTEM DESCRIPTION

No new system description is expected for the required SRE08 primary system. If results from additional systems are also submitted, a brief description of algorithms used in each such system must be submitted along with the results.

11 SCHEDULE

The deadline for signing up to participate in the evaluation is August 4, 2008.

The evaluation data set will be distributed by NIST so as to arrive at participating sites on August 11, 2008.

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

Evaluation results will be released to the participating sites by NIST on September 22, 2008.

12 GLOSSARY

Test – A collection of trials constituting an evaluation component.

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

Target (model) speaker – The hypothesized speaker of a test segment, one for whom a model has been created from training data.

Non-target (impostor) speaker – A hypothesized speaker of a test segment who is in fact not the actual speaker.

Segment speaker – The actual speaker in a test segment.

Target (true speaker) trial – A trial in which the actual speaker of the test segment *is in fact* the target (hypothesized) speaker of the test segment.

Non-target (impostor) trial – A trial in which the actual speaker of the test segment *is in fact not* the target (hypothesized) speaker of the test segment.