2000 Speaker Recognition Evaluation
Evaluation Plan

Last Modification: January 18, 2000

Version 1.0

1. Introduction

The 2000 speaker recognition evaluation is part of an ongoing series of yearly evaluations conducted by NIST. These evaluations provide 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 was designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible.

The 2000 evaluation will be conducted in the spring. The data will be available in April, with results due to be submitted to NIST about three weeks later. A follow-up workshop for evaluation participants to discuss research findings will be held in June. For the specific dates see section *11. Schedule* below.

Participation in the evaluation is solicited for all sites that find the tasks and the evaluation of interest. For more information, and to register a desire to participate in the evaluation, please contact Dr. Alvin Martin at NIST.

2. Technical Objective

This speaker recognition evaluation focuses on the tasks of speaker detection, segmentation, and tracking. These tasks are posed in the context of conversational telephone speech and for limited training data. *(An optional speaker detection test will use speech from the Spanish AHUMADA Corpus, which is spontaneous but not conversational.)* The evaluation is designed to foster research progress, with the goals of:

1. Exploring promising new ideas in speaker recognition.

2. Developing advanced technology incorporating these ideas.

3. Measuring the performance of this technology.

The 2000 evaluation will include the following four tasks:

1. ***One-speaker detection.*** This task is NIST's basic speaker recognition task, as defined in all of NIST's previous annual speaker recognition evaluations. The . The first will be the one-speaker detection task, which has been the focus of the previous annual speaker evaluations coordinated by NIST. This task is to determine whether a specified target speaker is speaking during a given speech segment. Each speech segment is formed by concatenating consecutive utterances of a given test speaker.

2. ***Two-speaker detection.*** This task is essentially the same as the one-speaker detection task, except that the speech segments include *both* sides of a telephone call (summed together), rather than being limited to the speech from a single speaker.

3. ***Speaker tracking.*** This task requires determining the time intervals (if any) during which a specified target speaker is speaking in a segment of conversational speech. This task is performed on a subset the test segments used for the two-speaker detection task.

4. ***Speaker segmentation.*** This task requires identifying the time intervals during which unknown speakers are each speaking in a conversational speech segment (no specified target speakers are given). The number of different speakers may or may not be known.

3. The Evaluation

Evaluation will be performed separately for each of the speaker recognition tasks:

1. One-speaker detection
2. Two-speaker detection
3. Speaker tracking
4. Speaker segmentation

For the one-speaker detection task, evaluation will be performed separately for the conversational data and for the optional test on the AHUMADA data. For the speaker segmentation task, evaluation will be performed separately for segments with the number of speakers known in advance to be two and for segments with a number of speakers not known in advance. Testing on the segments with a known number of speakers will be optional for this task.

3.1. Speaker Detection Tasks

For the detection tasks the formal evaluation measure is the detection cost function, defined as a weighted sum of the miss and false alarm error probabilities:

$$C_{Det} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times P_{NonTarget}$$

The parameters of this cost function are the relative costs of detection errors, C_{Miss} and $C_{FalseAlarm}$, and the *a priori* probability of the target, P_{Target}. The following parameter values will be used as the primary evaluation of speaker recognition performance for the detection tasks:

$C_{Miss} = 10;$ $C_{FalseAlarm} = 1;$ $P_{Target} = 0.01;$ $P_{NonTarget} = 1 - P_{Target} = 0.99$

Speaker detection performance will be evaluated by measuring the correctness of detection decisions for an ensemble of speech segments in terms of the detection cost function. These segments will represent a statistical sampling of conditions of evaluation interest. For each of these segments a set of target speaker identities will be assigned as test hypotheses. Each of these hypotheses must be independently judged as true or false, and the correctness of these decisions will be tallied.

In addition to the actual detection decision, a decision ***score*** will also be required for each test hypothesis. This decision score will be used to produce detection error tradeoff curves, in order to see how misses may be traded off against false alarms.

3.2. Speaker Tracking

The following describes the formal evaluation measure for the tracking task:

As last year, NIST will determine the regions where each speaker is speaking using an energy based speech detector on the individual speech channels (conversation sides) before summing. This year, however, scoring will be done with respect to the time intervals where the speech detector indicates only one of the speakers to be speaking. In addition, buffers of 250 milliseconds from the ends of each interval will be ignored for scoring purposes. Thus only speech segments of half a second or more will be used for scoring.

The decisions for each hypothesized target speaker will be compared with a reference answer key to determine the miss and false alarm rates for speaker tracking, according to the following computation:

$$P_{Miss} = \int_{\text{Target speech}} \delta(D_t, F)\,dt \Big/ \int_{\text{Target speech}} dt$$

$$P_{FalseAlarm} = \int_{\text{NonTarget speech}} \delta(D_t, T)\,dt \Big/ \int_{\text{NonTarget speech}} dt$$

where

D_t = the system output (T or F), as a function of time

$$\delta(x, y) = \begin{cases} 1 \text{ if } x = y, \\ 0 \text{ otherwise.} \end{cases}$$

The parameter values for the cost function for speaker tracking will be different from those for speaker detection. The tracking parameter values will be:

$C_{Miss} = 1$; $C_{FalseAlarm} = 1$; $P_{Target} = 0.5$; $P_{NonTarget} = 1 - P_{Target} = 0.5$

In addition to the actual decisions for each detected target interval, a decision ***score*** will also be required. This decision score will be used to produce DET (detection error tradeoff) curves, in order to see how misses may be traded off against false alarms.

3.3. Speaker Segmentation

As in the tracking task, NIST will determine the regions where each speaker is speaking, by using an energy-based speech detector on the individual channels (conversation sides) before summing or, where available, by using the time marks determined by human transcribers. As in the tracking task, scoring will be done with respect to those time intervals where only one speaker is speaking, and will exclude buffers of 250 milliseconds from the ends of each interval. Thus only speech segments of half a second or more will be used for scoring.

NIST will score separately the segments where the number of speakers present is known in advance and those where this is not known. The same scoring procedure will be used in each case, however.

For the speaker segmentation task, a system must produce hypothesized speaker turns (or segments of speech produced from a single speaker) and a generic speaker label for each turn (e.g., speaker 0,

speaker 1, ...). All turns from the same speaker should have the same generic speaker label. Unlike the other detection tasks, this is a classification task, which can be characterized by a single error rate. (There is only a single error since all speech must be accounted for and a "miss" for one hypothesized speaker label will generate a corresponding "false alarm" for another hypothesized speaker label, so the two errors are no longer independent.) To compute the classification error a search for the best (minimum error) mapping of hypothesized speaker labels to true speakers for each conversation is performed, and then the errors are accumulated over the ensemble of test conversations. The error rate is computed as follows:

Assume a system produces *M* hypothesized speaker labels for a conversation that actually contains *N* true speakers. (When the number of speakers is known in advance, we will have M = N). When M < N we automatically generate N-M speaker labels with no associated speech segments. We first produce a putative one-to-one mapping of true speakers *{i}* with hypothesized speaker labels *{j}*

$map(i) = j;\ i=1,..,N,\ 1 <= j <= M$

Letting *true_duration(i, conv)* = the total duration of speech by true speaker i in a conversation denoted *conv*, and *hyp_duration(i ,j ,conv)* = the total duration of speech common to true speaker *i* and hypothesized speaker label *j* in this conversation, we then compute the error rate for this mapping as

$error(map, conv) = 1 - S_i\ hyp\_duration(i, map(i), conv) / S_i\ true\_duration(i, conv)$

The best one-to-one mapping, *map*(i)*, used for this conversation is the one producing the minimum error(map, conv). For the conversation we then log the values

$hit(conv) = S_i\ hyp\_duration(i, map^*(i), conv)$

and

$total(conv) = S_i\ true\_duration(i, conv)$

The total (weighted) error over an ensemble of conversations is finally computed as

$total\_error = 1 - S_{conv}\ hit(conv) / S_{conv}\ total(conv)$

4. Evaluation Conditions

In previous evaluations the use of telephone handset labels (one of two types, either *electret* or *carbon-button*) has been demonstrated to provide a significant improvement in speaker recognition performance. This labeling was determined automatically by analysis and classification of the same speech signal used for the speaker recognition evaluation. This year we continue to provide the labeler's hard decision of handset type (most of which will be electret) and the *likelihoods* used to determine the hard decision. During evaluation, systems will be allowed to use these hard decisions and likelihoods for all one-speaker training and test data. No label information will be provided for the two-speaker data.

4.1. Training

In past evaluations it was shown that the use of more data and varied data for training (involving multiple conversations and/or handsets) gives better performance.

In order to simplify this aspect of the evaluation in 1999, data was provided for only one training condition, namely "*two-session*." This year data will be provided for only the "*one-session*" training condition. Training data for each speaker will consist of two minutes of speech from a single conversation. The actual duration of the training files used will vary from the nominal value of 2 minutes, so that whole turns may be included whenever possible. Actual durations will be constrained to be within the range of 110-130 seconds.

Neither side of a conversation used for training will be used to create test files.

4.2. One-Speaker Detection Test

This year's evaluation will again use test-segments of varying durations with no distinct groupings by duration. No more than one test segment will be selected from each conversation side. Male and female segments will be grouped separately, and performance results for each sex will be noted. There will be no cross-gender trials.

This year all test segments will be from a different phone number, and presumably different handset, than that used in the segment speaker's training data. Most of the training and test segments will be from handsets algorithmically determined to be of electret microphone type, but some data from handsets with carbon-button type microphone will be included for contrast.

The primary measure of performance will be for trials where the target speaker training and the test segment are both of the electret type, and the segment duration is between 15 and 45 seconds.

4.3. Two-Speaker Detection Test

For this test, the duration of each test segment will be nominally 60 seconds, with the actual duration varying between 59 and 61 seconds. Actual duration will vary from nominal so that whole turns may be included whenever possible. The duty cycle of a segment speaker will vary from close to zero up to 100%.

There are three possible cases with respect to gender for each test segment; both speakers are male, both are female, or one is male and one female. Performance will be computed and evaluated separately for each of these three cases, but the system will **not** be given prior knowledge detailing the gender mix of the test segment. (Automatic gender detection may be used, of course.)

The primary measure of performance will be for trials where the target speaker training and both sides of the summed test segment are of the electret type, the durations of speech by each of the two segment speakers is between 15 and 45 seconds, and the two segment speakers are of the same sex.

4.4. Tracking Test

For this test, the evaluation conditions will be exactly as for the two-speaker detection test, with segments of approximately one minute in duration, but the number of test segments to process will be reduced, and the number of target hypothesized speakers per test segment will be reduced.

The primary measure of performance will be for trials satisfying the primary condition for the two-speaker detection task and in addition, the target speaker is one of the segment speakers.

4.5. Speaker Segmentation Test

This test will use two rather different types of test segments. The first type will consist of the summed

two-channel segments of the tracking task, which will be known to contain exactly two speakers. The second type will consist of summed two-channel segments, which will be in multiple languages, will generally be of longer duration, up to a maximum of 10 minutes, and may include as many as 10 speakers in a segment. The exact number of speakers in each of these segments will **not** be given to systems as prior knowledge.

The primary measure of performance will be for the test segments with an unknown number of speakers.

5. Development Data

The primary development data for this evaluation will be the 1999 Speaker Recognition Evaluation kit (NIST Speech Disc R55-1.1 - R55-5.1). This data, drawn from the Switchboard-2 Corpus, Phase 3, was used in 1999 for the one and two-speaker detection and the tracking tasks. The tracking data may also be used as development data for the segmentation task. In addition, NIST will make available on a single CD-ROM some conversational speech segments from the CALLHOME Corpus, along with associated speaker turn time markings, for development work for the segmentation task with an unknown number of speakers. Some of these will be segments containing more than two speakers. There will be no AHUMADA development data as such. However, some read text telephone data from a number of male speakers, one session per speaker, collected around the same time and in a manner similar to the AHUMADA speakers, will be made available to sites wishing to use it to estimate channel conditions. Sites intending to perform the evaluation and to submit results to NIST may acquire any of this data set and associated documentation from NIST free of charge by contacting Dr. Martin. The Switchboard-1 Corpus, including the 1996 Speaker Recognition Evaluation kit may also be used for development work if sites wish. Note that the Switchboard-2 Corpus, Phases 1 and 2, including in particular the test data of the 1997 and 1998 NIST evaluations, **may not** be used.

6. Evaluation Data

The conversational data for the one-speaker detection, two-speaker detection and speaker tracking tasks will be drawn from the SwitchBoard-2 Corpus, Phases 1 and 2. This is a recycling of the data used in previous evaluations. All conversations have been processed through echo canceling software before being used to create training and test segments.

The optional one-speaker detection test of this evaluation will use the spontaneous telephone speech portion of the AHUMADA Corpus, which consists of three telephone sessions for each of the 103 male speakers.

Training and one-speaker test segments from Switchboard-2 will be constructed by concatenating consecutive turns of the desired speaker. All speech data will be stored as an 8-bit m-law continuous speech signal in a separate SPHERE file. The SPHERE header of each such file will contain some auxiliary information as well as the standard SHERE header fields.

NIST will seek to maximize the number of target speakers with electret training data, but some carbon-button training data will be included as well. There will be several hundreds of target speakers of each sex. These target speakers will also serve as non-target (impostor) speakers. Not every test segment will come from a target speaker and not every target speaker will have test segments.

The two-speaker detection and tracking test data will consist of the summation of the two sides of a given conversation leaving in areas of silence. Each test segment will be bounded by silence, as determined by the speech detector used to create this data set. No handset information will be provided

for this data.

The speaker segmentation data will consist of the speaker tracking segments, and of segments, which are again the summation of the two sides of a conversation, drawn from the CALLHOME Corpus.

6.1. One-Speaker Detection Test

There will be an average of between five and six test segments per target speaker. There will be several thousands of test segments. For each test segment, there will be trials against 11 putative speakers.

Each **one-speaker test** segment will be the speech from a single-side of the summed speech segment used for the two-speaker task. That is, for every test segment used in the two-speaker task, there will be two corresponding segments in the one-speaker task, one from each original channel. This will allow comparison of two-speaker performance with one-speaker performance.

6.2. Two-Speaker Detection Test

The average number of test segments containing a given target speaker will be about 6. The total number of test segments will be half of the total number of *one-speaker* test segments. Each segment may contain 0, 1, or 2 target speakers. For each test segment there will be trials against 22 putative speakers. Test segments that contain:

- **Two Males** - will have 22 male putative speakers
- **Two Females -** will have 22 female putative speakers
- **One Male and One Female** - will have 11 male and 11 female putative speakers

6.3. Speaker Tracking Test

The speaker tracking test segments will be a subset of the data used for the two-speaker detection test. These will be limited to segments where each conversation side is algorithmically determined to be of electret handset type. To the extent possible, equal numbers of segments will be included which contain two male speakers, two female speakers, and one male and one female speaker. The total number of segments to be processed will be no greater than 1000.

For each test segment there will be trials against 4 putative speakers. These 4 will be a subset of the 22 putative speakers for the segment in the two-speaker detection task. Conversations that contain:

- **Two Males** - will have 4 male putative speakers
- **Two Females -** will have 4 female putative speakers
- **One Male and One Female –** will have 2 male and 2 female putative speakers

6.4. Speaker Segmentation Test

The first part of the test data for this task will be the test segments of the speaker tracking task. The second part will consist of test segments of varying length, up to a maximum duration of 10 minutes. They will contain a varying number of speakers, and will be in a variety of languages. The speakers in these segments should all be different from those appearing in the detection and tracking segments, and each such speaker should not appear in more than one segment. The total number of these segments to be processed will be no greater than 500.

6.5. One-speaker detection - Spanish Test

The AHUMADA Corpus, collected in Spain, consists of data from 103 male speakers. It includes data from three telephone sessions (calls) by each speaker. Included in each session is at least one minute of "spontaneous" speech on a topic chosen by the speaker. One session used a common handset for all speakers who each made an internal line phone call. A second session used nine different handsets, with six to fourteen of the speakers sharing each handset. The calls in this session were external. A third session involved each speaker making an external call from his home using a unique handset.

The AHUMADA task will be viewed as a one-speaker detection task similar to that described above for conversational data. The training and test segments will consist of the entire spontaneous speech portion of one of a speaker's telephone sessions. Each such portion will be a minute or more in duration. Thus there will be training data for approximately 103 speakers. There will be approximately 206 test segments. One factor to be examined is the effect of non-target (impostor) speakers using (or not using) the same handset for training as the true segment speaker. There will be approximately 20 hypothesized target speakers for each test segment.

7. Evaluation Rules

In order to participate in the 2000 speaker recognition evaluation, a site must complete one (or more) of the following tests that use conversational data, in its entirety:

- One-speaker detection
- Two-speaker detection
- Speaker segmentation
- Speaker tracking

The one-speaker detection *AHUMADA* test is optional, given the completion of one of the above tests. As with all tests, participants must submit results for the entire test as defined.

For the segmentation task participants may, if they choose, submit only results for the CALLHOME data, for which the number of speakers present will be unknown. This option is intended to accommodate those wishing to do speaker segmentation but not speaker tracking.

All participants must observe the following evaluation rules and restrictions:

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

- Knowledge of the sex of the target speaker (implied by data set directory structure as indicated below) **is** allowed, but knowledge of the sex mixture of the two-speaker test segments is **not** allowed, except as determined by automatic means.

- Knowledge of handset type information (found in the SPHERE header under the field name "handset_type" -- see SPHERE Header Information, below) **is** allowed for the one-speaker test segments. Knowledge of the handset type mixture of the two-speaker test segments is **not** allowed, except as determined by automatic means.

- For the segmentation task, knowledge of the number of speakers present, except as determined by automatic means, is **not** allowed.

- Listening to the evaluation data, or any other experimental interaction with the data, is **not** allowed before all test results have been submitted. This applies to target speaker training data as well as test segments.

- The corpora from which the evaluation data are taken, namely the SwitchBoard-2 Phases 1 and 2, CALLHOME, and AHUMADA Corpora, may not be used for any system training or R&D activities related to this evaluation, except for that CALLHOME data specifically provided for development.

- Knowledge of the "*start and ending*" times that were used to construct the test segments (found in the SPHERE header -- see SPHERE Header Information, below) **is** allowed.

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

8. Evaluation Data Set Organization

It is intended that this year's evaluation kit will be distributed on **8** CD-ROM's. Each will contain a single top-level directory used as a unique label for the disk:

- **sid00etr** - disc containing the training data
- **sid00e1m** - disc containing the male 1-speaker detection test segments
- **sid00e1f** - disc containing the female 1-speaker detection test segments.
- **sid00e2a** - disc 1 of the 2-speaker detection, tracking, and segmentation test segments
- **sid00e2b** - disc 2 of the 2-speaker detection, tracking, and segmentation test segments
- **sid00sg1** - disc 1 of the additional speaker segmentation test segments
- **sid00sg2** - disc 2 of the additional speaker segmentation test segments
- **sid00ah1** - disc containing the optional AHUMADA test segments

The disk of training data will have two top-level directories denoted **male** and **female** each containing training data for speakers of the corresponding sex. Each directory will contain one SPHERE-formatted speech data file for each target speaker. The names of these files will be the ID of the target speaker followed by a ".sph" extension.

The six disks of test segment data will each contain multiple test segments of the appropriate type, one SPHERE-format file for each test segment. The names of these files will be pseudo-random alphanumeric strings, followed by ".sph".

Each of these six disks will also contain an index file, which will list the trials for each test.

The file that denotes 1-speaker detection trials will be **detect1.ndx**, and the file that denotes the 2-speaker detection trials will be **detect2.ndx**. Each record in these files will contain the name of a test segment file on the disk followed by a number of target speaker ID's, separated by white space. The corresponding speaker detection tests will consist of processing each record's test segment against each of the target speaker ID's listed in that record. There will be 11 of these target ID's in each record for the disks of one-speaker test segments, and 22 for the disks of two-speaker test segments.

The file that denotes speaker tracking trials will be **tracking.ndx**. Each record in these files will contain the name of a test segment file on the disk followed by 4 target speaker ID's, separated by white space. The speaker tracking test will consist of processing each record's test segment against each of the target speaker ID's listed in that record.

The file that denotes segmentation trials will be **segment.ndx**. Each record in these files will contain a test segment file on the disk. The segmentation test will consist of processing each record's test segment.

9. Format for Submission of Results

Results for each task must be stored in a single file, according to the formats defined in this section. The file name should be intuitively mnemonic and should be constructed as "SSS_N_TTT", where

SSS := identifies the site,

N := identifies the system, and

TTT := identifies the task (*1sp, 2sp, trk, sg2 or sgn*)

9.1. One-Speaker Test Results

Sites participating in the one-speaker evaluation tests must report results for whole tests, including all of the test segments. These results must be provided to NIST in a single results file using a standard ASCII format, with one record for each decision. Each record must document its decision with the target identification, test segment identification, and decision information. Each record must contain six fields, separated by white space and in the following order:

1. The sex of the target speaker - **M** or **F** (always **M** for the AHUMADA test)
2. The target speaker ID (*a four digit number*)
3. The test - (**1** *for one-speaker detection,* ***A*** *for AHUMADA test*)
4. The test segment file name (*excluding the directory and file type*)
5. The decision - **T** or **F** (*is the target speaker the same as the speaker in the test segment?*)
6. The score (*where the more positive the score, the more likely the target speaker*)

9.2. Two-Speaker Test Results

Sites participating in the two-speaker evaluation must report results for the whole test, including all of the test segments. These results must be provided to NIST in a single results file using standard ASCII format, with one record for each decision. Each record must document its decision with target

identification, test segment identification, and decision information. Each record must contain six fields, separated by white space and in the following order:

1. The sex of the target speaker - **M** or **F**
2. The target speaker ID (*a four--digit number*)
3. The test - **2** (*for two-speaker*)
4. The test segment file name (*excluding the directory and file type*)
5. The decision - **T** or **F** (*is the target speaker an actual speaker in the test segment*?)
6. The score (*where the more positive the score, the more likely the target speaker*)

9.3. Speaker Tracking Test Results

Sites participating in the speaker tracking evaluation must report results for the whole test, including all of the test segments. These results must be provided to NIST in a single results file using standard ASCII format , with one SGML (Standard Generalized Markup Language)-tagged data set for each test segment/target speaker hypothesis. Each data set will contain all tracking results for a test segment/target speaker combination.

Each tracking results data set will contain data for all decision intervals within a test segment. These data are namely the interval's start time, decision and score. (Intervals must be contiguous, so that the end time of an interval is implicitly specified by the start time of the following interval.) The format of the results for a single test segment/target speaker combination is defined to be:

<track segment=SEGMENT_NAME **target**=TARGET_ID>

TIME DECISION SCORE

TIME DECISION SCORE

…

</track>

WHERE:

<track …> Identifies the beginning of tracking results for the segment.

SEGMENT NAME: The test segment file name (*four alphanumeric characters*.)

TARGET ID The target speaker ID (*a four-digit number*).

TIME: The starting time of the decision interval, in seconds.

DECISION: The decision (T or F) applied to the interval.

SCORE: The score for the decision interval.

</track> Identifies the end of tracking results for the segment.

Due to the theoretically unlimited size of the results file for the tracking task, a practical limit will be imposed on the size of a single results file. All results files must be less than 100MB in size, uncompressed.

9.4. Segmentation Test Results

Sites participating in the segmentation evaluation must report results for a whole test, either the condition where it is known there are exactly 2 speakers in the test segment (SG2) or for the condition where it is unknown how many speakers are in the test segment (SGN). Each of these results must be provided to NIST in a single results file using a standard ASCII format. This file should be a concatenation of all segment records. A segment record should be created as follows.:

**

START_TIME END_TIME SPEAKER_ID

START_TIME END_TIME SPEAKER_ID

…

****

WHERE:

<segment …> Identifies the beginning of segmentation record.

SEGMENT NAME: The test segment file name (*four alphanumeric characters*.)

START_TIME: The starting interval time (to the hundredth of a second).

END_TIME: The ending interval time (to the hundredth of a second).

SPEAKER_ID: The speaker cluster this segment belongs to [0-9].

 Identifies the end of a segmentation record.

There will be no more than 10 unique speakers per test segment. Each segment record should make use of the 10 digits zero through 9 to represent a speaker cluster, beginning with zero and incrementing by one for each new speaker.

Due to the theoretically unlimited size of the results file for the segmentation task, a practical limit will be imposed on the size of a single results file. All results files must be less than 100MB in size, uncompressed.

10. System Description

A brief description of the system (the algorithms) used to produce the results must be submitted along with the results, for each system evaluated. It is permissible for a single site to submit multiple systems for evaluation. In this case, however, the submitting site must identify one system as the "primary" system prior to performing the evaluation.

Sites must report the CPU execution time that was required to process the test data, as if the test were run on a single CPU. Sites must also describe the CPU and the amount of memory used.

11. Schedule

The deadline for signing up to participate in the evaluation is 31 March 2000.

The evaluation data set CD-ROM's will be distributed by NIST on 17 April 2000.

The deadline for submission of evaluation results to NIST is 8 May 2000.

Room reservations for the follow-up workshop (see below) must be received by (to be determined) May 2000.

The follow-up workshop will be held at the Inn and Conference Center of the University of Maryland University College in College Park, Maryland on (to be determined) and (to be determined) June 2000. Those participating in the evaluation are expected to present and discuss their findings at the workshop.

END-NOTES

[1] To contact Dr. Martin, send him email at alvin.martin@nist.gov, or call him at 301/975-3169.

[2] See Ortegae-Garcia, J., et al, 'AHUMADA': A Large Speech Corpus in Spanish for Speaker Characterization and Identification, to appear in Speech Communication Journal.

[3] Note that explicit speaker detection decisions are required. 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 decision scores from the various target speakers will be pooled before plotting detection error tradeoff curves. Thus it is important to normalize scores across speakers to achieve satisfactory detection performance.

[5] The reference answer key will be produced using a NIST speech detector applied to the target side of the conversation (before combining the two sides).

[6] The labeling will be performed using MIT Lincoln Lab's handset type labeler software.

[7] The 1999 results suggested that performance is poorer for segments of less than 15 seconds, but invariant with duration for segments that are longer than 15 seconds.

[8] The SwitchBoard-2 Corpus was created by the University of Pennsylvania's Linguistic Data Consortium (LDC) for the purpose of supporting research in speaker recognition. Information about this corpus and other related research resources may obtained by contacting the LDC (by telephone at 215/898-0464 or via email at ldc@ldc.upenn.edu).

[9] This corpus was created by the University of Pennsylvania's Linguistic Data Consortium (LDC) for the pupose of supporting research in multi-lingual conversational speech recognition. Information about this corpus and other related research resouces may be obtained by contacting the LDC (by telephone at 215/898-0464 or via email at ldc@ldc.upenn.edu).

[10] To obtain the echo canceling software and associated documentation, visit the Mississippi State web-site at: www.isip.mmstate.edu/resources/technology/software/1996/fir_echo_canceller/ec_v2_5.tar.Z

[11] Documentation on the SPHERE file format and SPHERE software may be obtained via ftp from: ftp://jaguar.ncsl.nist.gov/pub/sphere_2.6a.tar.Z

[12] The handset type label and likelihood value will be supplied in the SPHERE header. It will also include information to document the source file, start time, and duration of all excerpts that were used to construct the segment.

[13] In the 1996 evaluation, speakers were identified as either "target" or "non-target". This distinction (namely of being a target or non-target) was associated with the speaker. This year, the appellation of "target" or "non-target" is associated with the speaker's role rather than the speaker's identity. The reason for this change is that the distinction made no difference in the results from 1996 and it was demonstrated that performance was insensitive to whether the impostor was a target speaker or a non-target speaker.

[14] The speech detector used, "speech", is packaged in the NIST SPQA (*Speech Quality Assurance*) software package, available at: www.nist.gov/speech/tools. It is an energy detector, operating on a single channel of PCM speech data. Identified speech segments (reported in 10 millisecond frames) are bounded by 20 milliseconds of silence.

[15] Eleven target ID's per test segment was chosen to maximize the efficiency of the evaluation for a given level of statistical significance. The results from the performance design goal - given a false alarm probability 10 times lower than the miss probability, it takes ten times more impostor trials to produce equal number of miss and false alarm errors.

Participating sites may also choose to do an optional extension of the one-speaker detection test. This will consist of doing a trial of each test segment against **all** same-sex target speakers instead of just the 11 specified putative speakers. Doing this 'complete matrix' of test is not likely to affect the overall qualitative performance results, but it will support the study of the effects of speaker differences on speaker recognition systems. For some systems the additional computational costs of testing each test segment against all targets may not be excessive. Sites interested in pursuing this option should contact Dr. Martin about procedures to be followed for submitting the extended results.

[16] As with the main one-speaker detection task, interested sites are invited to provide scores for all (presumably 103) speakers (all of whom are male) for each test segment.

[17] Participants are encouraged to do as many tests as possible. However, it is absolutely imperative that results for all of the test segements and target speakers in a test be submitted in order for that test to be considered valid and for the results to be accepted.

[18] The reason for this rule is that the technology is viewed as being "application-ready". This means that the technology must be ready to perform speaker detection simply by being trained on a specific target speaker and then performing the detection task on whatever speech segments are presented, without the (artifical) knowledge of the speech of other speakers and other segments in the test set.

---

12. GLOSSARY

Trial - The individual evaluation unit for each task involving a test segment and (except for segmentation) a hypothesized speaker.

Target trial - A trial where the hypothesized speaker is a true speaker in the test segment

Non-target (impostor) trial - A trial where the hypothesized speaker is not a true speaker in the test segment

Target speaker - A speaker for whom training data is provided and who may therefore be used as a hypothesized speaker in trials

Non-target (impostor) speaker - A hypothesized speaker in a trial who is not an actual speaker in the test segment

Segment speaker - An actual peaker in a test segment

One-session training - Training data for a target speaker consisting of speech extracted from a single conversation

Two-session training - Training data for a target speaker consisting of speech extracted from two different conversations

Turn - A maximal interval within a two-person conversation during which only one participant speaks while the other remains silent