1999 Speaker Recognition Evaluation
Evaluation Plan

Last Modification: February 5th, 1999

Version 6.03

Introduction

The 1999 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 1999 evaluation will be conducted in the spring. The data will be available late in March, with results due to be submitted to NIST about three weeks after this. A follow-up workshop for evaluation participants to discuss research findings will be held toward the end of May. For the specific dates see Schedule below.

Participation in the evaluation is solicited for all sites that find the task 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[1].

Technical Objective

The current speaker recognition evaluation focuses on the task of speaker detection and tracking. These tasks are posed in the context of conversational telephone speech and for limited training data. 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, and
3. measuring the performance of this technology.

NIST has expanded the technical challenges in 1999 to include the following three distinct 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 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. ***Multi-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, rather than being limited to the speech from a single speaker. Each speech segment is formed by adding together the two (separately collected) sides of a telephone conversation.

3. ***Speaker tracking.*** This task is to perform speaker detection as a function of time. The task is to identify the times when the specified target speaker is speaking during a given telephone conversation. The speech segments to be processed are taken from the multi-speaker detection task.

The Evaluation

Evaluation will be performed separately for the three speaker recognition tasks: one-speaker detection, two-speaker detection, and speaker tracking. For each of these tasks the formal evaluation measure will be a detection cost function, defined as a weighted sum of the miss and false alarm error probabilities:

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

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

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

Performance will be evaluated in terms of the detection cost function for an ensemble of speech segments. These segments will represent a statistical sampling of conditions of evaluation interest, and a set of target speaker identities will be assigned as test hypotheses for each segment. Each hypothesis is to be processed in turn, independently of all others.

Speaker Detection

Speaker detection performance will be evaluated by measuring the correctness of detection decisions for an ensemble of speech segments. 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 then be judged as true or false, and the correctness of these decisions will be tallied.[2]

In addition to the (binary) detection decision, a decision ***score*** will also be required for each test hypothesis.[3] 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.

Speaker Tracking

For the speaker tracking task, those intervals of each test speech segment belonging to the target speaker must be detected, for each given target speaker hypothesis. System output decisions will be compared with a reference answer key[4] 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

$$D_t = \text{the system output } (T \text{ or } F), \text{ as a function of time}$$

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

In addition to each (binary) decision output for the detected target intervals, a decision ***score*** will also be required for each output interval. 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.

Evaluation Conditions

In previous evaluations the use of telephone handsets 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 produced automatically by analysis and classification of the same speech signal used for speaker recognition. As in the 1998 evaluation, handset type labels will be provided with the one-speaker data this year.[5] During evaluation, the speaker recognition systems will be allowed to use these labels, for all one-speaker training and test data. No labels will be provided for the two-speaker (which is also two-channel) data. Moreover, as noted, the emphasis will be on results involving Electret data only.

Training

In past evaluations multiple training conditions, involving training data from different combinations of conversations, were considered for each speaker. As expected, more data and varied data (involving multiple conversations and/or handsets) was found to give better performance. In order to simplify this aspect of the evaluation this year, there will be only one training condition in this year's evaluation. This will be what was referred to previously as "*two-session*" training. The training data for each speaker will consist of one minute of speech from each of two different conversations collected from the same phone number.[6] The actual duration of the training files used will vary from the nominal value of 1 minute, so that whole turns may be included whenever possible. Actual durations will be constrained to be within the range of 55-65 seconds.

One-Speaker Detection Test

In past evaluations test segments of different durations were provided and evaluated separately. As expected, longer durations always gave better performance than shorter durations. In order to simplify this aspect of the evaluation this year, segments of different durations will not be grouped into separate tests. While durations will vary, there will be no more than one test segment from a conversation used for each evaluation task. Male and female segments will be grouped separately for the one-speaker detection task, and performance results for each sex will be noted.

The segments for the one-speaker tests will consist of essentially the same data as in the two-speaker tests (described below), but with each of the two single-sided channels in a two-speaker test segment being tested as a separate segment, and having periods of silence removed. Thus, the length of these test segments will vary from close to zero up to a full minute.

Performance will be computed separately for test segments and target speakers where the training and test phone numbers are the same[7], where they are different but of the same type (both carbon button or both Electret), and where they are of different types. Furthermore, the primary measure of performance will be for the case where the phone numbers are different, but the training and test segment are both of the Electret type, and the duration is between 25 and 35 seconds.

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 the target 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.)

Speaker Tracking Test

For this test, the evaluation conditions will be exactly as for the ***two-speaker detection test***, although the number of test segments to process will be reduced, and the number of target hypotheses per test segment will be reduced.

Development Data

The development data for this evaluation will be the evaluation data sets for the main 1998 evaluation plus the evaluation set for the summer 1998 development evaluation. The main 1998 EvalSet exists on six CD-ROM's labeled **sid98e***. The summer 1998 development EvalSet exists on six CD-ROM's labeled ss98de (three CD's) and ms98de (three CD's). Sites intending to perform the evaluation and to submit results to NIST may acquire these data sets and associated documentation from NIST free of charge by contacting Dr. Martin.

Evaluation Data

The evaluation data will be drawn from the SwitchBoard-2 phase 3 corpus.[8] The training and one-speaker test segments will be constructed by concatenating consecutive turns for the desired speaker, as was done for the main 1998 evaluation (including processing each conversation through echo canceling software[9] before each test segment is created). Each such segment will be stored as a continuous speech signal in a separate SPHERE file.[10] The speech data will be stored in 8-bit m-law format. The handset type label will be supplied in the SPHERE header. The SPHERE header will also include auxiliary information to document the source file, start time and duration of all excerpts that were used to construct the segment.

NIST will verify that the speech segments are properly labeled. All speakers with at least two conversations from a single phone number will serve as target speakers and will have training data supplied. (For some of these target speakers there may be no test segments.) There will be about 250 target speakers of each sex. These target speakers will also serve as non-target (impostor) speakers.[11]

There will also be some test segments spoken by speakers who had insufficient data to serve as targets.

The two-speaker detection test data will consist of the summation of the two sides of a given conversation. The conversation will be processed through an echo-canceller before summing the sides. The data will be stored in 8-bit m-law format. Unlike the training data and the one-speaker test data, areas of silence will *not* be removed. Each test segment will be bounded by silence, as determined by the speech detector used to create this data set.[12] No handset labels will be provided for this data.

The evaluation corpus will be supplied on 5 CD-ROM's. For convenience, one CD will contain the training data, two will contain the one-speaker test segments (one each for males and females), and two will contain the two-speaker test segments. (See Evaluation Data Set Organization below.)

One-Speaker Detection Test

There will be an average of between five and six test segments per target speaker. Each test segment will be from a different conversation side, and all will be from conversations that are not used for training data. The total number of test segments will be no greater than 4000. (Note that some of the test segments will be from speakers other than the target speakers.) For each test segment, there will be trials against 11 putative speakers[13], all of the same sex as the speaker of the segment.

The **one-speaker test** data will comprise all of the single-sided segments used in the two-speaker test. That is, corresponding to each summed speech segment in the two-speaker test there will be two segments in the one-speaker test, one from each original channel. The non-speech segments *will* be removed from this data. This will enable comparison of two-speaker performance with one-speaker performance.

Two-Speaker Detection Test

There will be a single test segment from each conversation in the corpus that is not used for training data. No test segments will be from conversations used for training. The total number of such segments will be no greater than 2000. Each segment may contain 0, 1, or 2 target speakers. The average number of test segments containing a given target speaker will be about 6.

For each test segment there will be trials against 22 putative speakers for the two-speaker detection task. 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

Speaker Tracking Test

The evaluation data for the speaker-tracking test will be a subset of the segments used for the two-speaker detection test. These will be limited to segments where each conversation side is of the same handset type. Wherever possible, each target speaker will be included in at least two segments. 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

Evaluation Rules

A single test for each of the three evaluation tasks will constitute the evaluation. These tasks are namely one-speaker detection, two-speaker detection, and speaker tracking. Every evaluation participant is required to undertake the full test for at least one of the tasks. The additional tests are then optional for each participant, but participants must submit all of the results for each test performed.[14]

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.[15] 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 the handset type (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.
- 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 corpus from which the evaluation data are taken, namely the SwitchBoard-2 phase 3 corpus, may not be used for any system training or R&D activities related to this evaluation.[16]
- 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.

Evaluation Data Set Organization

There will be five disks in the EvalSet. Each will contain a single top-level directory used as a unique label for the disk. The disk of training data will have the name **sid99etr**. The two disks of one-speaker test segments will have the names **sid99e1m** and **sid99e1f** and contain male and female data respectively. T he two disks to two-speaker test segments will have the names **sid99e2a** and **sid99e2b**.

The disk of training data will have two top-level directory denoted **male** and **female** and containing training data for speakers of the corresponding sex. Each directory will contain two SPHERE-format speech data files for each target speaker, each file containing one minute of speech in m-law format from a different conversation side. The names of these files will be the ID of the target speaker followed by "a.sph" or "b.sph".

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

Each of these four disks will also contain a file denoted **detect.ndx** that specifies the detection tests to be performed. 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 two two-speaker disks will each contain a file denoted **tracking.ndx** that specifies the tracking tests to be performed. 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.

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", or "trk").

One-Speaker Test Results

Sites participating in the *one-speaker* evaluation tests must report results for whole tests, including ***all*** of the test segments, for each duration being reported. 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 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 - **1** *(for one-speaker)*
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)

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 results file, with only one test per file, using a 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 the same as speaker in the test segment?)*
6. The score *(where the more positive the score, the more likely the target speaker)*

Speaker Tracking 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 results file, with only one test per file, using a standard ASCII format, with one sgml-tagged data set for each test segment for each 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>**

<u>WHERE:</u>

"**<track** … **>**" identifies the beginning of tracking results for the segment.

SEGMENT NAME: The test segment file name *(4 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 result files must be less than 100MB in size, uncompressed.[17]

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.

Execution Time

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.

Schedule

- The deadline for signing up to participate in the evaluation is 5 March 1999.
- The evaluation data set CD-ROM's will be distributed by NIST on 29 March 1999.
- The deadline for submission of evaluation results to NIST is 19 April 1999.
- Room reservations for the follow-up workshop (see below) must be received by 3 May 1999.
- The follow-up workshop will be held at The Inn and Conference Center of the University of Maryland University College in College Park, Maryland[18] on 3 and 4 June 1999. Participants in the evaluation will be expected to attend this workshop and to present and discuss their research at it.

End Notes

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

[2] 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.

[3] 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.

[4] 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).

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

[6] The two sessions used for training will also be constrained to exhibit the same handset type label. That is, both sessions must be labeled as either Electret or carbon-button.

[7] The "same phone number" condition in this evaluation is not really a fair test. This is because virtually all of the impostor data is collected from phone numbers that are different from those used for the target speakers' training data. Thus it is only the target speakers who use the same phone number. The impostors use different phone numbers and thus are more easily discriminated against.

[8] The SwitchBoard-2 phase 3 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 be obtained by contacting the LDC (by telephone at 215/898-0464 or via email at ldc@ldc.upenn.edu).

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

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

[11] In the 1996 evaluation, speakers were identified as either "target" or "non-target". This distinction (namely of being a target or a 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 the 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.

[12] The speech detector used, "speech", is packaged in the NIST SPQA (Speech Quality Assurance) software package, available at: www.nist.gov/speech/software.htm. 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.

[13]Eleven target ID's per test segment was chosen to maximize the efficiency of the evaluation for a given level of statistical significance. This 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 numbers 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 tests 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.

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

[15] 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 (artificial) knowledge of the speech of other speakers and other segments in the test set.

[16] This is a nominal requirement, because the LDC have not yet made the SwitchBoard-2 phase 3 corpus publicly available.

[17] Note: If a system makes a decision every 1/100th of a second, it will on average make 6000 decisions per segment/trial. If the system outputs a score that contains 5 characters, the resulting file will be approximately 310M. If the score were reduced to 3 characters, the resulting file would be approximately 264M. If a system makes a decision every 1/10th of a second, it will on average make 600 decisions per segment/trial. If the system outputs a score that contains 5 characters, the resulting file will be approximately 31M.

[18] The UMUC Conference Center is located on the University of Maryland College Park campus close to Washington, D.C. Access is available from the BWI and National Airports.