The 1997 Speaker Recognition Evaluation Plan

Last Modification: May 9th, 1997
Version: 2.1

Introduction

The 1997 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 1997 evaluation will be conducted May. A follow-up workshop for evaluation participants will be held during June, to discuss research findings. 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. That is, the task is to determine whether a specified target speaker is speaking during a given speech segment.[2] This task is 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.

The Evaluation

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 a test hypotheses. Each of these hypotheses will then be required to be judged as true or false, and the correctness of these decisions will be tallied.[3]

The formal evaluation measure will be a detection cost function, defined as a weighted sum of the miss and false alarm error probabilities:

- $C_{Det} = C_{Miss} * P_{Miss|Target} * P_{Target} + C_{FalseAlarm} * P_{FalseAlarm|NonTarget} * P_{NonTarget}$

The parameters of this cost function are the relative costs of detection errors, CMiss and CFalseAlarm, and the a priori probability of the target, PTarget. The primary evaluation will use the following parameter values:

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

In addition to the (binary) detection decision, a decision ***score*** will also be required for each test hypothesis.[4] This decision score will be used to produce detection error tradeoff curves, in order see how misses may be traded off against false alarms.

Evaluation Conditions

Training

There will be 3 training conditions for each target speaker. All 3 of these conditions will use 2 minutes of training speech data from the target speaker. The 3 conditions are:

- "**One-session**" training. The training data will be 2 minutes of speech data taken from only one conversation. This data will be stored in two files, with 1 minute of speech in each.
- "**One-handset**" training. Equal amounts of training data will be taken from two different conversations collected using the same handset. (The same telephone number, actually.) The training data for this *one-handset* condition will comprise the first of the *one-session* training files, plus an additional file containing 1 minute of speech from a different session (but from the same telephone number).
- "**Two-handset**" training. Equal amounts of training data will be taken from two conversations collected using different handsets. (Different telephone numbers, actually.) The training data for this *two-handset* condition will comprise the first of the *one-session* training files, plus an additional file containing 1 minute of speech from a session collected from a different telephone number.

The actual duration of the training files 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.

Test

Performance will be computed and evaluated separately for female and male target speakers and for the 3 training conditions. For each of these training conditions, there are 2 different test conditions of interest. These are:

- **Test segment duration.** Performance will be computed separately for 3 different test durations. These durations will be nominally 3 seconds, 10 seconds and 30 seconds. Actual duration will vary from nominal so that whole turns may be included whenever possible. Actual durations will be constrained to be within the ranges of 2-4 seconds, 7-13 seconds, and 25-35 seconds, respectively. A single turn will be used for the test segments whenever possible.

- **Same/different handset.** Performance will be computed separately for test segments which use a training handset versus those segments which use a different handset (as determined by the phone number that the speaker was using). The type of handset (same/different) being used in the test segment will be unknown to the system under test.[5]

Development Data

The development data for this evaluation will comprise the DevSet and EvalSet for last year's evaluation. The 1996 DevSet is one CD-ROM labeled **sid96d1** and the 1996 EvalSet is two CD-ROM's labeled **sid96e1f** and **sid96e1m**. Sites intending to perform the evaluation and to submit results to NIST may acquire these development data 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 1 corpus.[6] Both training and test segments will be constructed by concatenating consecutive turns for the desired speaker, similar to what was done last year. Each segment will be stored as a continuous speech signal in a separate SPHERE file. The speech data will be stored in 8-bit mulaw format. The SPHERE headers will include auxiliary information to document to source file, start time and duration of all excerpts which were used to construct the segment.[7]

NIST will manually audit all segments to verify that the selected speech is for the identified speaker and does not include any significant extraneous speech from other speakers. There will be between 400 and 500 speakers that will serve both as target speakers and as non-target (impostor) speakers.[8] There will be additional speakers that will serve only as impostors.

The evaluation corpus will be supplied on 6 CD-ROM's. For convenience, data will be grouped according to sex and stored separately - three discs for female data and three discs for male data. Knowledge of the sex of the target speaker is admissible side information and may be used if desired.

The evaluation data will include both training data and test data. The number of test segments from each target speaker will vary, with an average of about 10 test segments per target speaker and test duration. (For each speaker, each of the test segments of a given duration will be from a unique conversation for that speaker.) This will make a total of about 2500 test segments for each sex and for each of the three test durations.[9]

Evaluation Rules

A total of nine tests constitute the evaluation. These tests are namely a test for each of the three test durations for each of the three training conditions. Every evaluation participant is required to submit all of the results for each test performed.[10] In the event that a participating site does not submit a complete set of results, NIST will not report any results for that site. For all nine tests in this evaluation, there will

be a grand total of about 50,000 target speaker trials and 500,000 non-target speaker trials (see Evaluation Data Set Organization below).

The following evaluation rules and restrictions on system development must be observed by all participants:

- 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.[11] 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 training conditions **is** allowed.
- Knowledge of the sex of the target speaker (and thus also the test segment) **is** 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 corpus from which the evaluation data are taken, namely the SwitchBoard-2 phase 1 corpus, may not be used for any system training or R&D activities related to this evaluation.[12]

Evaluation Data Set Organization

All of the six disks in the EvalSet will have the same organization. Each disk's directory structure will organize the data according to information admissible to the speaker recognition system. This directory structure will be as follows:

- There will be a single top-level directory on each disk, used as a unique label for the disk. These directories will be named **sid97e1f, sid97e2f** and **sid97e3f** for the female data, and **sid97e1m, sid97e2m** and **sid97e3m** for the male data.
- Under each top-level directory there will a subdirectory, namely **train** for storing the training data (on **sid97e1f** and **sid97e1m** only) and **test** for storing the test data on the other four CD's.
  - Under the **train** directory there will be two subdirectories, namely **female** for the female speakers (on **sid97e1f** only) and **male** for the male speakers (on **sid97e1m** only).
    - Under the **female** and the **male** directories there will be four subdirectories, namely **hs1_s1a** (for the first *one-session* training segment), **hs1_s1b** (for the second *one-session* training segment), **hs1_s2** (for the second *one-handset* training session), and **hs2** (for the segment from the second training handset).
      - In each of the **hs*** directories there will be one SPHERE-format speech data file for each of the speakers, containing approximately one minute of speech in mulaw format. The name of this file will be the ID of the target speaker, followed by "**.wav**".
  - Under the **test** directory there will be three subdirectories, namely "**30**" (for the 30 second test segments), "**10**" (for the 10 second test segments), and "**3**" (for the 3 second test segments).
    - In each of the **30**, **10** and **3** segment duration directories will be the test segments of that duration, one SPHERE-format -law speech data file for each test segment. The

name of these files will be pseudo-random alphanumeric strings, followed by "**.wav**".

- Also in each of the segment duration directories will be three index files which specify the tests to be performed on the segments for each of the three training conditions. Each record in these files will contain the name of a test segment file (in the corresponding test segment directory) followed by a number of target speaker ID's, separated by white space. The target speakers listed will all be of the same sex as the segment speakers. The evaluation test for a given training condition and duration will be to process each record's test segment against each of the target speaker ID's listed in that record, for all records in the index file on each of the four disks corresponding to the given training condition and duration. There will be about 10 of these target ID's in each record.[13] The three index files (contained in each of the three segment duration directories) will be named:
  - **1s.ndx** (for targets using the *one-session* training condition)
  - **1h.ndx** (for targets using the *one-handset* training condition)
  - **2h.ndx** (for targets using the *two-session* training condition)

Format for Submission of Results

Sites participating in the evaluation must report test results for all of the tests. These results must be provided to NIST in results files using a standard ASCII record 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 thus contain seven fields, separated by white space and in the following order:

1. The sex of the target speaker - **M** or **F**.
2. The training condition - **1S**, **1H** or **2H**. *(one-session, one-handset or two-handset)*
3. The target speaker ID. *(a 4-digit number)*
4. The test segment duration - **30**, **10** or **3**.
5. The test segment file name. *(excluding the directory and file type)*
6. The decision - **T** or **F**. *(Is the target speaker the same as the test segment speaker?)*
7. The score. *(where the more positive the score, the more likely the target speaker)*

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 05/09/1997.
- The evaluation data set CD-ROM's will be distributed by NIST on 16 May 1997.
- The deadline for submission of evaluation results to NIST is 2 June 1997.
- The follow-up workshop will be held at the Maritime Institute on 25-26 June 1997.[14] Participants in the evaluation will be expected to attend this workshop and to present and discuss their research at it.

---

[1] To contact Dr. Martin, you may send him email at alvin@jaguar.ncsl.nist.gov, or you may call him at (301 975-3169.

[2] Speaker ***detection*** is chosen as the task in order to focus research on core technical issues and thus improve research efficiency and maximize progress. Although important application-level issues suggest more complex tasks, such as simultaneous recognition of multiple speakers, these issues are purposely avoided. This is because these application-level challenges are believed to be more readily solvable, if only the performance of the underlying core technology were adequate, and it is believed that the R&D effort will be better spent in trying to solve the basic but daunting core problems in speaker recognition.

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

[6] The SwitchBoard-2 phase 1 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@upenn.edu).

[7] For information about NIST's SPHERE utilities (including instructions to download SPHERE utilities) visit the NIST, Spoken Natural Language Processing Group's (http://www.nist.gov/speech) website. The source time-marks are documented in each test segement's SPHERE header. The field *segment_origin* lists the information used in constructing the test segment. A *segment_origin* record is of the type: **segment_origin=[conversation_id,channel,start_time,end_time]...**

[8] 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 - the results from last year's evaluation demonstrated that performance was insensitive to whether the impostor was a target speaker or a non-target speaker.

[9] For 1997 Evaluation the number of test segments will be limited to 2500, but there will be more then 2500 segments on the evaluation CDs. These extra segments may be useful in future development work.

[10] 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. If a participant anticipates being unable to complete all the tests, NIST should be consulted for preferences about which tests to perform. Each participant must negotiate its test commitments with NIST before NIST ships the evaluation CD-ROM's to that site.

[11] 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 be 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.

[12] This is a nominal requirement, because the LDC have not yet made the SwitchBoard-2 phase 1 corpus publicly available.

[13] Ten 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.

[14] The Maritime Institute is in the Baltimore-Washington area, not far from BWI International airport.