1998 Speaker Recognition Evaluation

Last Modification: March 26, 1998
Version 2.0

Introduction

The 1998 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 1998 evaluation will be conducted in February. A follow-up workshop for evaluation participants will be held during March, 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.

In 1996 and 1997 handset variation was featured as a prominent technical challenge to be addressed. While handset variation remains a formidable challenge, the 1998 evaluation will direct greatest attention toward speaker recognition performance for the case in which both training and test data are from the same source.

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, C_{Miss} and $C_{FalseAlarm}$, and the a priori probability of the target, P_{Target}. The primary evaluation will use the following parameter values:

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

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 to see how misses may be traded off against false alarms.

Evaluation Conditions

In 1997, the use of telephone handsets labels (one of two types, either *electret* or *carbon-button*) was 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. Because of the benefit that this labeling process provides, the speech data to be provided in the 1998 evaluation will also be so labeled.[5] During evaluation, the speaker recognition systems will be allowed access to this information, for all training and test data.

Training

There will be 3 training conditions for each target speaker. Two of these conditions will use 2 minutes of training speech data from the target speaker, while the other training condition will use more than 2 minutes of training speech data. The 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.
- "**Two-session**" training. Equal amounts of training data will be taken from two different conversations collected from the same phone number. The training data for this *two-session* 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 phone number).[6]
- "**Two-session-full**" training. All available speech data taken from two different conversations collected from the same phone number. The training data for this *two-session-full* condition will comprise all the data used for the *one-session* and *two-session* training conditions, plus two additional files which will contain additional speech data available from each of the two sessions used for the other two training conditions.

The actual duration of the training files used for the 1-session and 2-session training conditions will vary from the nominal value of 1 minute, so that whole turns may be included whenever possible. Actual durations for these training conditions will be constrained to be within the range of 55-65 seconds. The duration of the training files for the two-session-full condition will not be constrained and will be variable over speakers.

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 3 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 phone number.**[7] Performance will be computed separately for test segments from the training phone number versus those segments from different phone numbers. For this test, the handset type label (*electret* or *carbon-button*) will be the same as that used in training, and it will be provided as side information to the system under test.
- **Same/different handset type.** Performance will be computed separately for test segments with the same handset type label as training, versus segments with a different handset label. For this test, all test segments will be from phone numbers different from the training number.

Development Data

The development data for this evaluation will be the EvalSet for the 1997 evaluation. The 1997 EvalSet exists on six CD-ROM's labeled **sid97e***. Sites intending to perform the evaluation and to submit results to NIST may acquire these data and associated documentation from NIST free of charge by contacting Dr. Martin.[8]

Evaluation Data

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

NIST will verify that the speech segments are properly labeled and fairly represent the speaker. There will be about 250 female and 250 male speakers. These speakers will serve both as target speakers and as non-target (impostor) speakers.[11]

The evaluation corpus will be supplied on 6 CD-ROM's, by necessity. 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. There will be an average of about 10 test segments for each target speaker and for each test duration. (Each of these test segments will be from a different conversation.) This will make a total of about 2500 test segments for each sex and for each of the three test durations. (Note, however, that some of the test segments will be from speakers other than the target speakers.) For each test segment, there will be about ten trials. Thus, for all of the tests in this evaluation, there will be a total of about 45,000 target speaker trials and about 405,000 non-target speaker trials, giving a grand-total of about 450,000 tests [2500 files/duration/sex * 2 sexes * 3 durations * 3 training conditions * 10 tests per file]

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.[12] In the event that a participating site does not submit a complete set of results, NIST will not report any results for that site.

The following evaluation rules and restrictions 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.[13] 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 (implied by data set directory structure as indicated below) **is** allowed.
- Knowledge of the sex of the target speaker (implied by data set directory structure as indicated below) **is** allowed.
- Knowledge of the handset type (found in the SPHERE header under the field name "handset_type" -- see SPHERE Header Information, below) **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 2 corpus, may not be used for any system training or R&D activities related to this evaluation.[14]
- 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 the test condition: same phone number versus different phone number test (given in the index files, as described below), **is** allowed.
- Knowledge of any information available in the SPHERE header **is** allowed.

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. Some directories on some disks may be empty. 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 **sid98e1f, sid98e2f,** and **sid98e3f** for the female data, and **sid98e1m, sid98e2m,** and **sid98e3m** for the male data.
- Under each top-level directory there will be two data subdirectories, namely **train** for storing the training data and **test** for storing the test data, and one documenation directory **doc**.
    - Under the train directory there will be five subdirectories, namely **s1a** (for the first *one-session* training segment), **s1b** (for the second *one-session* training segment), **s2a** (for the second *two-session* training segment) (on the e1" discs only), **s1r** (for the additional *two-session-full* training data from the first session) and **s2r** (for the additional *two-session-full* training data from the second session) (on the "e2" discs only)
        - In each of these directories there will be one SPHERE-format speech data file for each of the speakers, containing the appropriate amount speech in mu-law format. The name of this file will be the ID of the target speaker, followed by ".sph".
    - 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 mu-law speech data file for each test segment. The name of these files will be pseudo-random alphanumeric strings, followed by ".sph".

          Also in each of the segment duration directories will be three index files which specify the tests to be performed (one file 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 "same_num" or "diff_num" (which identifies the test condition as same or different phone number) followed by a number of target speaker ID's, separated by white space. The evaluation test will be to process each record's test segment against each of the target speaker ID's listed in that record. There will be 10 of these target ID's in each record.[15] 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)
- **2s.ndx** (for targets using the *two-session* training condition)
- **2f.ndx (**for targets using the two-session-full 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** or **2S or 2F**. *(one-session or two-session or two-session-full)*
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 16 January 1998.
- The evaluation data set CD-ROM's will be distributed by NIST on 9 February 1998.
- The deadline for submission of evaluation results to NIST is 23 February 1998.
- Room reservations for the follow-up workshop (see below) must be received by March 9, 1998.
- 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 31 March and 1 April 1998[16]. Participants in the evaluation will be expected to attend this workshop and to present and discuss their research at it.

SPHERE Header Information

<table>
<tr><td><b>Field Name:</b></td><td>segment_origin</td></tr>
<tr><td><b>Type</b>:</td><td>STRING</td></tr>
<tr><td><b>Purpose</b>:</td><td>Provides data excising information, for recreating the speech waveform.</td></tr>
<tr><td><b>Contents</b>:</td><td>For each identified speech segment we will provide the following:
<message_number>,<channel>,<begin_time>,<end_time>

where
<table>
<tr><td>message_number</td><td>The conversation ID. as it will appear on the released corpus</td></tr>
<tr><td>channel</td><td>1 digit value of {1,2}</td></tr>
<tr><td>begin_time</td><td>Float, time in seconds</td></tr>
<tr><td>end_time</td><td>Float, time in seconds</td></tr>
</table>
The information for Multiple speech segments will be separated by colons.</td></tr>
</table>

Field Name:	handset_type
Type:	STRING
Purpose:	Classifies the handset used in recording the speech segments as either "electret" or "carbon_ button"
Contents:	electret | carbon_button

Field Name:	handset_prob_carbon
Type:	FLOAT
Purpose:	Identifies the probability that this test segment came from a carbon-button handset
Contents:	Float, range 0.0 - 1.0

Field Name:	handset_prob_electret
Type:	FLOAT
Purpose:	Identifies the probability that this test segment came from an electret handset
Contents:	Float, range 0.0 - 1.0

Field Name:	handset_log_like_carbon
Type:	FLOAT

Purpose:	Identifies the log likelihood probability that this test segment came from a carbon-button handset
Contents:	Float

Field Name:	handset_log_like_electret
Type:	FLOAT
Purpose:	Identifies the log likelihood probability that this test segment came from an electret handset
Contents:	Float

END NOTES

[1] To contact Dr. Martin, send him email at Alvin.Martin@nist.gov, or 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 believe 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 labeling will be performed using MIT Lincoln Lab's handset type labeler software.

[6] The two session used for training will also be constrained to exhibit the same handset type label. That is, the two sessions must be labeled both 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] Handset type labels for last year's data were not part of the 1997 EvalSet, nor was the answer key. These will be made available via ftp access. Contact Dr. Martin for details.

[9] The SwitchBoard-2 phase 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 be obtained by contacting the LDC (by telephone at

215/898-0464 or via email at ldc@ldc.upenn.edu).

[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] 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 of 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.

[13] 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.

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

[15] 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.

[16] 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.