NIST 2018 Speaker Recognition Evaluation Plan

August 17, 2018

1 Introduction

The 2018 speaker recognition evaluation (SRE18) is the next in an ongoing series of speaker recognition evaluations conducted by the US National Institute of Standards and Technology (NIST) since 1996. The objectives of the evaluation series are (1) to explore promising new ideas in speaker recognition, (2) to support the development of advanced technology incorporating these ideas, and (3) to measure and calibrate the performance of the current state of technology. The evaluations are intended to be of interest to all researchers working on the general problem of text-independent speaker recognition. To this end, the evaluation is designed to focus on core technology issues and to be simple and accessible to those wishing to participate.

SRE18 will be organized in a similar manner to SRE16, focusing on speaker detection over conversational telephone speech (CTS) collected outside North America. Again *fixed* and *open* training conditions will be offered to allow cross-system comparisons and to understand the effect of additional and unconstrained amounts of training data on system performance. There are, however, a few differences between SRE16 and SRE18. In particular, in addition to CTS recorded over a variety of handsets (PSTN), voice over IP (VOIP) data, which is also collected outside North America, as well as audio from video (AfV) will be included as development and test material in SRE18.

Participation in SRE18 is open to all who find the evaluation of interest and are able to comply with the evaluation rules set forth in this plan. There is no cost to participate, but participating teams must be represented at the evaluation workshop to be held in Athens, Greece on December 16-17, 2018. Information about evaluation registration can be found on the SRE18 website[1].

2 Task Description

2.1 Task Definition

The task for SRE18 is *speaker detection*: given a segment of speech and the target speaker enrollment data, automatically determine whether the target speaker is speaking in the segment. A segment of speech (test segment) along with the enrollment speech segment(s) from a designated target speaker constitute a *trial*. The system is required to process each trial independently and to output a log-likelihood ratio (LLR), using natural (base e) logarithm, for that trial. The LLR for a given trial including a test segment u is defined as follows

$$LLR(u) = \log\left(\frac{P\left(u|H_0\right)}{P\left(u|H_1\right)}\right). \tag{1}$$

where $P\left(\cdot\right)$ denotes the probability distribution function (pdf), and H_0 and H_1 represent the null (i.e., u is spoken by the enrollment speaker) and alternative (i.e., u is not spoken by the enrollment speaker) hypotheses, respectively.

[1] `https://www.nist.gov/itl/iad/mig/nist-2018-speaker-recognition-evaluation`

2.2 Training Conditions

The training condition is defined as the amount of data/resources used to build a Speaker Recognition (SR) system. The task described above can be evaluated over a *fixed* (required) or *open* (optional) training condition.

- **Fixed** – The fixed training condition limits the system training to specific *common* data sets which are as follows:
  - 1996–2008 NIST SRE Data (LDC2009E10)
  - 2010 NIST SRE and Follow-up Data (LDC2012E09)
  - 2012 NIST SRE Test Set (LDC2016E45)
  - 2016 NIST SRE Development Set (LDC2018E47)
  - 2016 NIST SRE Test Set (LDC2018E30)
  - Comprehensive Switchboard with transcripts (LDC2018E48)
  - Comprehensive Fisher English with transcripts (LDC2018E49)
  - MIXER 6 (LDC2013S03)[2]
  - 2018 NIST SRE Development (*dev*) Set (LDC2018E46)

  Participants can obtain these data from the Linguistic Data Consortium (LDC) after they have signed the LDC data license agreement. In addition to these, participants may also use VoxCeleb[3] and SITW[4] corpora. For the *fixed* training condition, only the specified speech data may be used for system training and development, to include all sub-systems, e.g., speech activity detection (SAD), and auxiliary systems used for automatic labels/processing (e.g., language recognition). Publicly available, non-speech audio and data (e.g., noise samples, impulse responses, filters) may be used and should be noted in the system description (see Section 6.4.2). Participation in this condition is required.

  Note: The use of pretrained models on data other than what is designated above is not allowed in this condition.

- **Open** – The *open* training condition removes the limitations of the *fixed* condition. In addition to the data listed in the *fixed* condition, participants can use other proprietary and/or publicly available data. LDC will also make selected data from the IARPA Babel Program (LDC2018E50) available to be used in the *open* training condition. Participation in this condition is optional but strongly encouraged to demonstrate the gains that can be achieved with unconstrained amounts of data.

2.3 Enrollment Conditions

The enrollment condition is defined as the number of speech segments provided to create a target speaker model. As in SRE16, gender labels will not be provided. There are two enrollment conditions for SRE18:

- **One-segment** – in which the system is given only one segment to build the model of the target speaker. For CTS (i.e., PSTN and VOIP) data, one segment that approximately contains 60 seconds [5] of speech is provided, while for AfV data, enrollment segments can vary in speech duration from a few seconds to several minutes.
- **Three-segment** – where the system is given three segments containing approximately 60 seconds of speech to build the model of the target speaker, all from the same phone number. This conditions only involves the PSTN data.

[2] It should be noted that portions of this data are included in other packages (e.g., SRE10 and SRE12 test sets), but it is being made available here for the convenience of those who need to train their systems with MIXER 6.

[3] `http://www.robots.ox.ac.uk/~vgg/data/voxceleb/`, and `http://www.robots.ox.ac.uk/~vgg/data/voxceleb2/`

[4] `http://www.speech.sri.com/projects/sitw/`

[5] As determined by SAD output.

2.4 Test Conditions

- For CTS (i.e., PSTN and VOIP) data, the speech duration of the test segments will be uniformly sampled ranging approximately from 10 seconds to 60 seconds. For AfV data, the test segment speech duration may vary from a few seconds to several minutes.
- Trials involving CTS data will be conducted with test segments from both same and different phone numbers as the enrollment segment(s).
- There will be no cross-gender trials.

3 Performance Measurement

3.1 Primary Metric

A basic cost model is used to measure the speaker detection performance and is defined as a weighted sum of false-reject (missed detection) and false-alarm error probabilities for some decision threshold θ as follows

$$\begin{aligned} C_{Det}(\theta) =& C_{Miss} \times P_{Target} \times P_{Miss}(\theta) + \\ & C_{FalseAlarm} \times (1 - P_{Target}) \times P_{FalseAlarm}(\theta), \end{aligned} \tag{2}$$

where the parameters of the cost function are C_{Miss} (cost of a missed detection) and $C_{FalseAlarm}$ (cost of a spurious detection), and P_{Target} (*a priori* probability of the specified target speaker) and are defined to have the following values for CTS and AfV source types:

Source Type	Parameter ID	C_{Miss}	$C_{FalseAlarm}$	P_{Target}
CTS	1	1	1	0.01
	2	1	1	0.005
AfV	3	1	1	0.05

Table 1: SRE18 cost parameters

To improve the interpretability of the cost function C_{Det} in (2), it will be normalized by $C_{Default}$ which is defined as the best cost that could be obtained without processing the input data (i.e., by either always accepting or always rejecting the segment speaker as matching the target speaker, whichever gives the lower cost), as follows

$$C_{Norm}(\theta) = \frac{C_{Det}(\theta)}{C_{Default}}, \tag{3}$$

where $C_{Default}$ is defined as

$$C_{Default} = min \begin{cases} C_{Miss} \times P_{Target}, \\ C_{FalseAlarm} \times (1 - P_{Target}). \end{cases} \tag{4}$$

Substituting either set of parameter values from Table 1 into (4) yields

$$C_{Default} = C_{Miss} \times P_{Target}. \tag{5}$$

Substituting C_{Det} and $C_{Default}$ in (3) with (2) and (5), respectively, along with some algebraic manipulations yields

$$C_{Norm}(\theta) = P_{Miss}(\theta) + \beta \times P_{FalseAlarm}(\theta), \tag{6}$$

where β is defined as

$$\beta = \frac{C_{FalseAlarm}}{C_{Miss}} \times \frac{1 - P_{Target}}{P_{Target}}. \quad (7)$$

Actual detection costs will be computed from the trial scores by applying detection thresholds of $\log(\beta)$, where log denotes the natural logarithm. For trials involving the CTS (i.e., PSTN and VOIP) source type, thresholds will be computed for two values of β, with β_1 for $P_{Target_1} = 0.01$ and β_2 for $P_{Target_2} = 0.005$, while for AfV trials a single threshold, $\log(\beta_3)$, will be computed for $P_{Target_3} = 0.05$. The primary cost measure for SRE18 is then defined as

$$C_{Primary} = \frac{1}{2}\left[\frac{C_{Norm_{\beta_1}} + C_{Norm_{\beta_2}}}{2} + C_{Norm_{\beta_3}}\right]. \quad (8)$$

The CTS portion of the evaluation data will be divided into 16 partitions. Each partition is defined as a combination of the number of enrollment segments (1 vs 3), speaker gender (male vs female), data source (PSTN vs VOIP), and phone number match (Y vs N). However, because no actual "phone number" metadata is available for the VOIP calls, the phone number match field only contains "N" for those calls, thereby reducing the effective number of partitions to 12. $C_{Primary}$ will be calculated for each partition, and the final result is the average of all the partitions' $C_{Primary}$'s.

Also, a minimum detection cost will be computed by using the detection thresholds that minimize the detection cost. Note that for minimum cost calculations, the counts for each condition set will be equalized before pooling and cost calculation (i.e., minimum cost will be computed using a single threshold not one per condition set).

NIST will make available the script that calculates the primary metric.

3.2 Alternative Metric

In addition to the primary metric, an alternative, information theoretic measure may be computed that considers how well all scores represent the likelihood ratio and that penalizes for errors in score calibration. This performance measure is defined as

$$C_{llr} = \frac{1}{2 \times \log(2)} \times \left(\frac{\sum \log(1 + \frac{1}{s})}{N_{TT}} + \frac{\sum \log(1 + s)}{N_{NT}}\right), \quad (9)$$

where the first summation is over all target trials N_{TT}, the second is over all non-target trials N_{NT}, and s represents a trial's likelihood ratio[6].

4 Data Description

The data collected by the LDC as part of the Call My Net 2 (CMN2) and Video Annotation for Speech Technology (VAST) corpora to support speaker recognition research will be used to compile the SRE18 development and test sets.

The CMN2 data are composed of PSTN and VOIP data collected outside North America, spoken in Tunisian Arabic. Recruited speakers (called *claque* speakers) made multiple calls to people in their social network (e.g., family, friends). Claque speakers were encouraged to use different telephone instruments (e.g., cell phone, landline) in a variety of settings (e.g., noisy cafe, quiet office) for their initiated calls and were instructed to talk for at least 8 minutes on a topic of their choice. All CMN2 segments will be encoded as a-law sampled at 8 kHz in SPHERE formatted files.

[6]The reasons for choosing this cost function, and its possible interpretations, are described in detail in the following paper: N. Brümmer and J. du Preez "Application-independent evaluation of speaker detection" *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

The VAST data are composed of audio extracted from YouTube[7,8] videos that vary in duration from a few seconds to several minutes and include speech spoken in English. Each audio recording may contain speech from multiple talkers, therefore manually produced diarization labels (i.e., speaker time marks) will be provided for both the *dev* and *eval* enrollment cuts (but not for the test cuts). All VAST data will be encoded as 16-bit FLAC files sampled at 44 kHz.

The test set will be distributed by NIST via the online evaluation platform (`https://sre.nist.gov`).

4.1 Data Organization

The development and test sets follow a similar directory structure:

<base_directory>/

README.txt
data/
enrollment/
test/
unlabeled/ (in training set only)
docs/

4.2 Trial File

The trial file, named `sre18_{dev|eval}_trials.tsv` and located in the `docs` directory, is composed of a header and a set of records where each record describes a given trial. Each record is a single line containing three fields separated by a tab character and in the following format:

`modelid<TAB>segmentid<TAB>side<NEWLINE>`

where
`modelid` - The enrollment identifier
`segmentid` - The test segment identifier
`side` - The channel[9]

For example:

```
modelid segmentid side
1001_sre18 dtadhlw_sre18 a
1001_sre18 dtaekaz_sre18 a
1001_sre18 dtaekbb_sre18 a
```

4.3 Development Set

Participants in the SRE18 evaluation will receive data for development experiments that will mirror the evaluation conditions. The development data will be drawn from CMN2 and VAST and will include:

- 25 speakers from CMN2 ($\sim$10 segments per speaker)
- 10 speakers from VAST (2 to 4 segments per speaker, manually produced diarization marks will be provided for the enrollment cuts which can be found under the `docs` directory in `sre18_dev_enrollment_diarization.tsv`)

[7] YouTube is a trademark of Google LLC.

[8] Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

[9] SRE18 segments will be single channel so this field is always "a"

- Associated metadata which will be listed in the following file located in the `docs` directory as outlined in section 4.1.

  - `sre18_dev_segment_key.tsv`: information about the segments and speakers from CMN2 and VAST, which includes the following fields: `segmentid` (segment identifier), `subjectid` (LDC speaker id), `gender` (male or female), `partition` (enrollment, test, or unlabeled), `phone_number` (anonymized phone number), `speech_duration` (segment speech duration), `data_source` (CMN2 or VAST)

As part of the SRE18 *dev* set, an *unlabeled* (i.e., no speaker ID, gender, or language labels) set of 2332 segments (with speech duration uniformly distributed in 10 s to 60 s range) from the CMN2 collection will also be made available. The segments are extracted from the *non-claque* side of the PSTN/VOIP calls. NIST will provide phone number metadata for the *unlabeled* segments, with the caveat that the phone numbers for these segments are unaudited and may not necessarily be reliable indications of speaker IDs, because one phone number may be associated with multiple callees, and one callee may be associated with multiple phone numbers. Also, note that for the *unlabeled* cuts, the `subjectid` field in the segment key file simply provides call IDs (not speaker IDs) prepended with 9 (the number 9).

The development data may be used for any purpose.

4.4 Training Set

Section 2.2 describes the two training conditions: Fixed (required) and Open (optional). For the *fixed* training condition, SRE18 participants will receive from LDC a "common" set of data resources. This includes all previous SRE data sets, a *dev* set selected from the new CMN2 and VAST speech collections, and Switchboard and Fisher corpora that contain transcripts. To obtain these data, participants must sign the LDC data license agreement which outlines the terms of the data usage. In addition to this "common" set, participants are also allowed to use VoxCeleb and SITW corpora for system training and development purposes in the *fixed* condition.

For the *open* training condition, in addition to the data noted above, LDC will also release selected data resources from the IARPA Babel Program. All training sets will be available directly from the LDC[10]. Participants are encouraged to submit results for the contrastive *open* training condition to demonstrate the value of additional data.

5 Evaluation Rules and Requirements

SRE18 is conducted as an open evaluation where the test data is sent to the participants who process the data locally and submit the output of their systems to NIST for scoring. As such, the participants have agreed to process the data in accordance with the following rules:

- The participants agree to abide by the terms guiding the training conditions (fixed or open).
- The participants agree to process at least the fixed training condition.
- The participants agree to process each trial independently. That is, each decision for a trial is to be based only upon the specified test segment and target speaker enrollment data. The use of information about other test segments and/or other target speaker data is not allowed.
- The participants agree not to probe the enrollment or test segments via manual/human means such as listening to the data or producing the transcript of the speech.

[10] `https://www.ldc.upenn.edu`

- The participants agree not to produce manual/human annotations of the unlabeled training data, such as employing a service like Amazon's Mechanical Turk. Informal listening and spectral analysis of subsets of the audio are acceptable.
- The participants are allowed to use any automatically derived information for training, development, enrollment, or test segments, provided that the automatic system used conforms to the training data condition (fixed or open) for which it is used.
- The participants are allowed to use information available in the SPHERE header.
- The participants can submit up to three systems per training condition. Bug-fix does not count toward this limit.

In addition to the above data processing rules, participants agree to comply with the following general requirements:

- The participants agree to have one or more representatives at the evaluation workshop to present a meaningful description of their system(s). Evaluation participants failing to do so will be excluded from future evaluation participation.
- The participants agree to the guidelines governing the publication of the results:
  - Participants are free to publish results for their own system but must not publicly compare their results with other participants (ranking, score differences, etc.) without explicit written consent from the other participants.
  - While participants may report their own results, participants may not make advertising claims about their standing in the evaluation, regardless of rank, or winning the evaluation, or claim NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (15 C.F.R. § 200.113) shall be respected[11]: *NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.*
  - At the conclusion of the evaluation NIST generates a report summarizing the system results for conditions of interest, but these results/charts do not contain the participant names of the systems involved. Participants may publish or otherwise disseminate these charts, unaltered and with appropriate reference to their source.
  - The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

6 Evaluation Protocol

To facilitate information exchange between the participants and NIST, all evaluation activities are conducted over a web-interface.

[11] See `http://www.ecfr.gov/cgi-bin/ECFR?page=browse`

6.1 Evaluation Account

Participants must sign up for an evaluation account where they can perform various activities such as registering for the evaluation, signing the data license agreement, uploading the submission and system description. To sign up for an evaluation account, go to `https://sre.nist.gov`. The password must be at least 12 characters long and must contain a mix of upper and lowercase letters, numbers, and symbols. After the evaluation account is confirmed, the participant is asked to join a site or create one if it does not exist. The participant is also asked to associate his site to a team or create one if it does not exist. This allows multiple members with their individual accounts to perform activities on behalf of their site and/or team (e.g., make a submission) in addition to performing their own activities (e.g., requesting workshop invitation letter).

- A site is defined as a single organization (e.g., NIST)
- A team is defined as a group of organizations collaborating on a task (e.g., Team1 consisting of NIST and LDC)
- A participant is defined as a member or representative of a site who takes part in the evaluation (e.g., John Doe)

6.2 Evaluation Registration

One participant from a site must formally register his site to participate in the evaluation by agreeing to the terms of participation. For more information about the terms of participation, see Section 5.

6.3 Data License Agreement

One participant from each site must sign the LDC data license agreement to obtain the training data for the fixed training condition and Babel data for the open training condition.

6.4 Submission Requirements

Each team must participate in the *fixed* training condition. Teams are encouraged to participate in the *open* training condition to demonstrate the gains that can be achieved leveraging unconstrained amounts of data. For each training condition, a team can submit up to three systems and must designate one as the *primary* system that NIST uses for cross-team comparisons.

There should be one output file for each training condition per system. Teams must process all test segments. Submission with missing test segments will not pass validation and will be rejected.

Each team is required to submit a system description at the designated time (see Section 7). The evaluation results are made available only after the system description report is received and confirmed to comply with guidelines described in Section 6.4.2.

6.4.1 System Output Format

The system output file is composed of a header and a set of records where each record contains a trial given in the trial file (see Section 4.2) and a log likelihood ratio output by the system for the trial. The order of the trials in the system output file must follow the same order as the trial list. Each record is a single line containing 4 fields separated by tab character in the following format:

`modelid<TAB>segment<TAB>side<TAB>LLR<NEWLINE>`

where

`modelid` - The enrollment identifier

`segmentid` - The test segment identifier
`side` - The channel (always "a" for SRE18 since the data is single channel)
`LLR` - The log-likelihood ratio

For example:

```
modelid segmentid side LLR
1001_sre18 dtadhlw_sre18 a 0.79402
1001_sre18 dtaekaz_sre18 a 0.24256
1001_sre18 dtaekbb_sre18 a 0.01038
```

There should be one output file for each training condition for each system. NIST will make available the script that validates the system output.

6.4.2 System Description Format

Each team is required to submit a system description. The system description must include the following items:

- a complete description of the system components, including front-end (e.g., speech activity detection, features, normalization) and back-end (e.g., background models, i-vector/embedding extractor, LDA/PLDA) modules along with their configurations (i.e., filterbank configuration, dimensionality and type of the acoustic feature parameters, as well as the acoustic model and the backend model configurations),
- a complete description of the data partitions used to train the various models (as mentioned above). Teams are encouraged to report how having access to the development set (labeled and unlabeled) impacted the performance,
- performance of the submission systems (primary and secondary) on the SRE18 development set (or a derivative/custom dev set), using the scoring software provided via the web platform (`https://sre.nist.gov`). Teams are encouraged to quantify the contribution of their major system components that they believe resulted in significant performance gains,
- a report of the CPU (single threaded) and GPU execution times as well as the amount of memory used to process a single trial (i.e., the time and memory used for creating a speaker model from enrollment data as well as processing a test segment to compute the LLR).

The system description should follow the latest IEEE ICASSP conference proceeding template.

7 Schedule

Milestone	Date
Evaluation plan published	April 2018
Registration period	May - September, 2018
Training data available	May, 2018
Evaluation data available to participants	September 10, 2018
System output due to NIST	October 10, 2018
Preliminary results released	October 29, 2018
Post evaluation workshop co-located with SLT in Athens, Greece	December 16-17, 2018