NIST 2019 Speaker Recognition Evaluation: CTS Challenge

July 22, 2019

1 Introduction

The 2019 speaker recognition evaluation (SRE19) 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) for NIST to effectively measure system-calibrated performance of the current state of technology, (2) to provide a common test bed that enables the research community to explore promising new ideas in speaker recognition, and (3) to support the community in their development of advanced technology incorporating these ideas. 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 evaluations are designed to focus on core technology issues and to be simple and accessible to those wishing to participate.

SRE19 will consist of two separate activities: 1) a leaderboard-style challenge using conversational telephone speech (CTS) extracted from the unexposed portions of the Call My Net 2 (CMN2) corpus, and 2) a regular evaluation using audio-visual material extracted from the unexposed portions of the Video Annotation for Speech Technology (VAST) corpus. This document describes the task, the performance metric, data, and the evaluation protocol as well as rules/requirements for Part 1 (i.e., the CTS Challenge). The evaluation plan for Part 2 will be described in another document. **Note that in order to participate in the regular evaluation (i.e., Part 2), one must first complete Part 1.** The SRE19 will be organized in a similar manner to SRE18, except for this year's evaluation only the *open* training condition will be offered (see Section 2.2).

Participation in the SRE19 CTS Challenge 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 in the SRE19 CTS Challenge and the evaluation web platform, data, and the scoring software will be available free of charge. Participating teams in the SRE19 CTS Challenge will have the option[1] to attend the post-evaluation workshop to be co-located with IEEE ASRU workshop in Sentosa, Singapore, on December 12-13, 2019. Information about evaluation registration can be found on the SRE19 website[2].

2 Task Description

2.1 Task Definition

The task for the SRE19 CTS Challenge 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

[1] Workshop registration is required for attendance.

[2] `https://www.nist.gov/itl/iad/mig/nist-2019-speaker-recognition-evaluation`

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

where $P(\cdot)$ 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.

2.2 Training Condition

The training condition is defined as the amount of data/resources used to build a Speaker Recognition (SR) system. Unlike SRE16 and SRE18, this year's evaluation only offers the open training condition that allows the use of any publicly available and/or proprietary data for system training and development. The motivation behind this decision is twofold. First, results from the most recent NIST SREs (i.e., SRE16 and SRE18) indicate limited performance improvements, if any, from unconstrained training compared to *fixed* training. Participants cited lack of time and/or resources during the evaluation period for not demonstrating significant improvement with *open* vs *fixed* training. Second, the number of publicly available large-scale data resources for speaker recognition has dramatically increased over the past few years (e.g., see VoxCeleb[3] and SITW[4]). Therefore, removing the *fixed* training condition will allow more in-depth exploration into the gains that can be achieved with the availability of unconstrained resources given the success of data-hungry Neural Network based approaches in the most recent evaluation (i.e. SRE18).

For the sake of convenience, in particular for the new and first-time participants, NIST also will provide an *in-domain* Development set that can be used for system training and development purposes:

- 2019 NIST Speaker Recognition Evaluation CTS Challenge Development Set (LDC2019E59)

This Development set simply combines the SRE18 CTS Dev and Test sets into one package. Participants can obtain this dataset through the evaluation web platform (`https://sre.nist.gov`) after they have signed the LDC data license agreement.

Although SRE19 allows unconstrained system training and development, participating teams must provide a sufficient description of speech and non-speech data resources as well as pre-trained models used during the training and development of their systems (see Section 6.4.2).

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 and SRE18, gender labels will not be provided. There are two enrollment conditions for the SRE19 CTS Challenge:

- **One-segment** – in which the system is given only one segment, approximately containing 60 seconds of speech [5], to build the model of the target speaker.
- **Three-segment** – where the system is given three segments, each containing approximately 60 seconds of speech to build the model of the target speaker, all from the same phone number. This condition only involves the Public Switched Telephone Network (PSTN) data.

2.4 Test Conditions

For the SRE19 CTS Challenge, the trials will be divided into two subsets: a progress subset, and an evaluation subset. The progress subset will comprise 30% of the trials and will be used to monitor progress in the

[3] `http://www.robots.ox.ac.uk/~vgg/data/voxceleb/`
[4] `http://www.speech.sri.com/projects/sitw/`
[5] As determined by a speech activity detector (SAD) output.

leaderboard. The remaining 70% of the trials will form the evaluation subset, and will be used to generate the official final results determined at the end of the challenge.

The challenge test conditons are as follows:

- The speech duration of the test segments will be uniformly sampled ranging approximately from 10 seconds to 60 seconds.
- Trials 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:

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

Table 1: The SRE19 CTS Challenge 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. Thresholds will be computed for two values of β, with β_1 for $P_{Target_1} = 0.01$ and β_2 for $P_{Target_2} = 0.005$. The primary cost measure for the SRE19 CTS Challenge is then defined as

$$C_{Primary} = \frac{C_{Norm_{\beta_1}} + C_{Norm_{\beta_2}}}{2}. \quad (8)$$

Similar to SRE18 (CTS trials), 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, on the evaluation web platform.

4 Data Description

The data collected by the LDC as part of the CMN2 corpus will be used to compile the SRE19 CTS Challenge 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.

The Development and Test sets 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 development set only)
docs/

4.2 Trial File

The trial file, named `sre19_cts_challenge_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[6]

For example:

```
modelid segmentid side
1001_sre19 dtadhlw_sre19 a
1001_sre19 dtaekaz_sre19 a
1001_sre19 dtaekbb_sre19 a
```

4.3 Development Set

Participants in the SRE19 CTS Challenge will receive data for development experiments that will mirror the evaluation conditions. The development data will simply combine the SRE18 CTS Dev and Test sets, and will include:

- 213 speakers from the CMN2 portion of SRE18
- Associated metadata which will be listed in the file `sre18_{dev|eval}_segment_key.tsv` located in the `docs` directory as outlined in section 4.1. The file contains information about the segments and speakers from the CMN2 portion of SRE18, and 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)

As part of the SRE19 CTS Challenge *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* (i.e., callee) 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 the number 9.

The development data may be used for any purpose.

4.4 Training Set

Section 2.2 describes the training condition for the SRE19 CTS Challenge (i.e., *open* training condition). Participants are allowed to use any publicly available and/or proprietary data they have available for system training and development purposes. The SRE19 CTS Challenge participants will also receive a Development set (described in previous Section) that they can use for system training. To obtain this Development data, participants must sign the LDC data license agreement which outlines the terms of the data usage.

[6]SRE19 CTS Challenge segments will be single channel so this field is always "a"

5 Evaluation Rules and Requirements

The SRE19 CTS Challenge is conducted as an open evaluation where the test data is sent to the participants to process 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 make at least one **valid** submission for the open 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 manual transcript of the speech.
- The participants are allowed to use any automatically derived information for training, development, enrollment, or test segments.
- The participants are allowed to use information available in the SPHERE header.
- The participants may make multiple challenge submissions (up to 3 per day). A leaderboard will be maintained by NIST indicating the best submission performance results thus far received and processed.

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

- 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[7]: *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.

Sites failing to meet the above noted rules and requirements, will be excluded from future evaluation participation, and their registrations will not be accepted until they are committed to fully participate.

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

6 Evaluation Protocol

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

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, as well as 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 participant is defined as a member or representative of a site who takes part in the evaluation (e.g., John Doe)
- 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)

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 development/training data for the SRE19 CTS Challenge.

6.4 Submission Requirements

Each team must make at least one valid submission for the challenge, processing all test segments. Submissions with missing test segments will not pass the validation step, and hence will be rejected.

Each team is required to submit a system description at the designated time (see Section 7). The final evaluation results (on the 70% evaluation subset) will be made available only after the system description report has been received by NIST 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 SRE19(since the data is single channel)
`LLR` - The log-likelihood ratio

For example:

```
modelid segmentid side LLR
1001_sre19 dtadhlw_sre19 a 0.79402
1001_sre19 dtaekaz_sre19 a 0.24256
1001_sre19 dtaekbb_sre19 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 SRE19 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	July 15, 2019
Registration period	July 15 - September 09, 2019
Training data available	July 15, 2019
Evaluation data available to participants	July 15, 2019
System output due to NIST	July 15 - October 07, 2019
Final official results released	October 28, 2019
Post-evaluation workshop	December 12–13, 2019