NIST 2019 Speaker Recognition Evaluation Plan

August 16, 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 (AV) 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 the regular evaluation (referred to as SRE19 hereafter). The evaluation plan for the SRE19 CTS Challenge can be found on the SRE19 website[1]. **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 the SRE18, except that for this year's evaluation only the *open* training condition will be offered (see Section 2.2). Moreover, in addition to the regular audio-only track, the SRE19 will also introduce audio-visual and visual-only tracks. System submission is required for the audio and audio-visual tracks, and optional for the visual track. Table 1 summarizes the tracks for the SRE19.

Track	Input	Core
Audio	Audio from Video	Yes
Audio-Visual	Audio and Frames from Video	Yes
Visual	Frames from Video	No

Table 1: The SRE19 tracks

Participation in the SRE19 is open to all who find the evaluation of interest and are able to comply with the evaluation rules set forth in this plan. Although there is no cost to participate in SRE19 (i.e., the evaluation data, web platform, and scoring software will be available free of charge), **participating teams must be represented at the post-evaluation workshop**[2] 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[1].

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

[2] Workshop registration is required.

2 Task Description

2.1 Task Definition

The task for the SRE19 is *individual/person detection*: given a test video segment and a target individual's enrollment video, automatically determine whether the target individual is present in the test segment. The test segment along with the enrollment segment from a designated target individual 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 s is defined as follows

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

where $P\left(\cdot\right)$ denotes the probability distribution function (pdf), and H_0 and H_1 represent the null (i.e., the target individual is present in s) and alternative (i.e., the target individual is not present in s) hypotheses, respectively.

2.2 Training Condition

The training condition is defined as the amount of data/resources used to build an individual/person recognition 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. We note, however, that 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 that can be used for speaker and individual/person 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 audio-visual and visual-only tracks, NIST will also provide two Development sets that can be used for system training and development purposes:

- JANUS Multimedia Dataset (LDC2019E55)
- 2019 NIST Speaker Recognition Evaluation Audio-Visual Development Set (LDC2019E56)

The LDC2019E55, which has been extracted from the IARPA JANUS Benchmark-C datatset, is available from the Linguistic Data Consortium (LDC), subject to approval of the LDC data license agreement. The LDC2019E56 contains the original videos from which the VAST portion of SRE18 Dev/Test sets were compiled. 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 video segments provided to create a target speaker model. There is only one enrollment condition for the SRE19:

[3] `http://www.robots.ox.ac.uk/~vgg/data/voxceleb/`
[4] `http://www.speech.sri.com/projects/sitw/`

- **One-segment** – in which the system is given only one video segment, that can vary in duration from a few seconds to several minutes, to build the model of the target individual.

Note that for the audio track, speech extracted from the enrollment video serves as enrollment data, while for the visual track, face frame(s) (i.e., frames in which the face of the target individual is visible) extracted from the video serve that purpose. Since NIST will only be releasing video files for SRE19, participants are responsible for extracting the relevant data (i.e., speech or face frames) for subsequent processing.

As in the most recent evaluations, gender labels will not be provided for the enrollment segments in the test set.

2.4 Test Conditions

The test conditions for the SRE19 are as follows:

- The test segment video duration may vary from a few seconds to several minutes.
- The test video can contain audio-visual data from potentially multiple individuals.
- There will be both same-gender and cross-gender trials.

3 Performance Measurement

3.1 Primary Metric

A basic cost model is used to measure the individual/person detection performance in SRE19, which 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 individual) and are defined to have the following values:

Source Type	**Parameter ID**	C_{Miss}	$C_{FalseAlarm}$	P_{Target}
AV	1	1	1	0.05

Table 2: The SRE19 detection 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 individual(s) as matching the target individual, 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 the set of parameter values from Table 2 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

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

The actual detection cost will be computed from the trial scores by applying a detection threshold of $\log(\beta)$, where log denotes the natural logarithm. The detection threshold will be computed for β_1 with $P_{Target_1} = 0.05$. The primary cost measure for the SRE19 is then defined as

$$C_{Primary} = C_{Norm}(\log(\beta_1)). \tag{8}$$

In addition to $C_{Primary}$, a minimum detection cost will also be computed by using the detection threshold that minimizes the detection cost. 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 Video Annotation for Speech Technology (VAST) corpus to support speaker recognition research will be used to compile the SRE19 development and test sets.

The VAST corpus contains amateur video recordings (such as video blogs) collected by the LDC from various online media hosting services. The videos vary in duration from a few seconds to several minutes and include speech spoken in English. Each video may contain audio-visual data from potentially multiple individuals who may or may not be visible in the recording, therefore manually produced diarization labels (i.e., speaker time marks), as well as *key* face frames[5] and bounding boxes (that mark an individual's face in the video) will be provided for both the *dev* set and *test* set enrollment videos (but not for the test videos in either set). All video data will be encoded as MPEG4.

The VAST Development and Test sets will be distributed by NIST via the online evaluation platform (`https://sre.nist.gov`), while the JANUS Multimedia Dataset will be released by the LDC.

4.1 Data Organization

The Development and Test sets follow a similar directory structure:

<base_directory>/
- README.txt
- data/
  - enrollment/
  - test/
- docs/

4.2 Trial File

The trial file, named `sre19_av_{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

[5]Note that only a few (out of potentially many) target face frames per enrollment video have been manually annotated.

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 will receive data for development experiments that will mirror the evaluation conditions, and will include:

- videos from 52 individuals from the VAST portion of SRE18
- Associated metadata which will be located in the `docs` directory as outlined in section 4.1:
  - `sre19_av_dev_segment_key.tsv` contains information about the video segments as well as the individuals within them, and includes the following fields:
    - `segmentid` (segment identifier)
    - `subjectid` (LDC speaker id)
    - `gender` (male or female)
    - `partition` (enrollment or test)
  - `sre19_av_dev_enrollment_diarization.tsv` contains manually produced time marks for target speakers, and includes the following fields:
    - `segmentid` (segment identifier)
    - `speaker_type` (speaker type, always "target")
    - `start` (start of target speaker segment time mark in seconds)
    - `end` (end of target speaker segment time mark in seconds)
  - `sre19_av_dev_enrollment_boundingbox.tsv` contains manually produced information about target individuals' faces (e.g., coordinates) in videos, and includes the following fields:
    - `segmentid` (segment identifier)
    - `speaker_type` (speaker type, always "target")
    - `face_frame_sec` (the frame in which a target individual's face is visible in seconds)
    - `bounding_box` (coordinates for a target individual's face in a specified frame as [x1,y1,x2,y2])
    - `face_covered` (whether the a target individual's face is covered)
    - `eyewear` (whether the a target individual is wearing glasses)
    - `facial_hair` (whether the a target individual has facial hair)

[6] SRE19 segments will be assumed single channel, therefore this field is always "a"

In addition to the data noted above, LDC will also release selected data resources from the IARPA JANUS Benchmark-C, namely the JANUS Multimedia Dataset[7] (LDC2019E55).

These development data may be used for any purpose.

4.4 Training Set

Section 2.2 describes the training condition for the SRE19 (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 participants will also receive two Dev sets (i.e., LDC2019E55 and LDC2019E56) that they can use for system training. To obtain these Development data, participants must sign the LDC data license agreement which outlines the terms of the data usage.

5 Evaluation Rules and Requirements

The SRE19 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 or watching the data, or producing the manual transcript of the speech, or producing the manual face coordinates.
- 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 header of the video files.
- The participants can register up to three systems for each track (i.e., audio, audio-visual, and visual) of the *open* training condition, one of which under each track should be designated as the primary system. Bug-fix does not count toward this limit. Teams can make unlimited number of submissions for each of the three systems until the evaluation period is over.

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

- The participants agree to submit reports to NIST that describe in sufficient length details of their systems and submissions. The system description reports should comply with guidelines described in Section 6.4.2.
- The participants agree to have one or more representatives at the post-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:

[7]The data is described in detail in the following paper: G. Sell, K. Duh, D. Snyder, D. Etter, D. Garcia-Romero, "Audio-Visual Person Recognition in Multimedia Data From the IARPA Janus Program," in *Proc. IEEE ICASSP*, pp. 3031-3035, 2018.

- 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[8]: *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 must not publicly publish or otherwise disseminate these charts.
- 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 future registrations will not be accepted until they commit to fully comply with the rules.

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)

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

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.

6.4 Submission Requirements

Each team must make at least one valid submission for the audio-only and the audio-visual tracks, processing all test segments. Submissions with missing test segments will not pass the validation step, and hence will be rejected. Submission for the visual-only track is optional but highly encouraged to gain insights into how the face recognition technology compares with the speaker recognition technology on the same data.

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 SRE19 since the data is assumed 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 track 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, diarization, face detection, face tracking, features, normalization) and back-end (e.g., background models, speaker/face 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,
- a complete description of the system combination strategy (e.g., score normalization/calibration for fusion) used for audio-visual individual/person recognition,
- 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	August 14, 2019
Registration period	August 15 - September 16, 2019
Training data available	August 15, 2019
Evaluation data available to participants	August 15, 2019
System output and system description due to NIST	October 21, 2019
Final official results released	October 28, 2019
Post-evaluation workshop	December 12–13, 2019