NIST 2020 CTS Speaker Recognition Challenge Evaluation Plan

August 7, 2020

1 Introduction

Following the success of the 2019 Conversational Telephone Speech (CTS) Speaker Recognition Challenge, which received 1347 submissions from 67 academic and industrial organizations, the US National Institute of Standards and Technology (NIST) will be organizing a 2020 CTS Challenge, the next iteration of an ongoing series of speaker recognition evaluations conducted by 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.

Similar to the 2019 CTS Challenge, the 2020 evaluation will feature a leaderboard-style challenge offering an *open/unconstrained* training condition (see Section 2.2), but using CTS recordings extracted from multiple data sources containing multilingual speech. In addition, unlike the 2019 CTS Challenge, no Development set will be released for the 2020 CTS challenge, and it will feature *public* leaderboards with live results on the *progress* subset as well as periodically updated results from the *test* subset.

This document describes the task, the performance metric, data, and the evaluation protocol as well as rules/requirements for the 2020 CTS Challenge. A regular speaker recognition evaluation (SRE) has also been planned for late 2020. The evaluation plan for the regular SRE20 will be described in another document. **Note that in order to participate in the regular evaluation, one must first complete the CTS Challenge.**

Participation in the 2020 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 challenge and the evaluation web platform, data, and the scoring software will be available free of charge. Participating teams in the 2020 CTS Challenge will have the option[1] to attend the post-evaluation workshop to be held in June 2021. Information about evaluation registration can be found on the announcement website[2].

2 Task Description

2.1 Task Definition

The task for the 2020 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

[1] Workshop registration is required for attendance.

[2] `https://www.nist.gov/itl/iad/mig/nist-2020-cts-speaker-recognition-challenge`

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 density 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. Similar to the 2019 CTS challenge, this year's evaluation only offers the open/unconstrained training condition that allows the use of any publicly available (e.g., see VoxCeleb[3] and SITW[4]) and/or proprietary data for system training and development. Unlike the past few SREs, no training/Development set will be initially available for the 2020 CTS Challenge.

Although the 2020 CTS Challenge 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 the most recent SREs, gender labels will not be provided. There are two enrollment conditions in the 2020 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.

2.4 Test Conditions

For the 2020 CTS Challenge, the trials will be divided into two subsets: a progress subset, and a test subset. The progress subset will comprise 30% of the trials and will be used to monitor progress in the leaderboard. The remaining 70% of the trials will form the test subset, and will be used to generate the official results which will be periodically published by NIST to a publicly accessible leaderboard (e.g., every few months).

The challenge test conditions are as follows:

- The speech duration of the test segments will be uniformly sampled ranging approximately from 10 seconds to 60 seconds.
- Trials will include test segments from both same and different phone numbers as the enrollment segment(s).
- There will be no cross-gender or cross-lingual trials.

[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.

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	C_{Miss}	$C_{FalseAlarm}$	P_{Target}
CTS	1	1	0.05

Table 1: The 2020 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

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

Actual detection costs will be computed from the trial scores by applying a detection threshold of $\log(\beta)$, where log denotes the natural logarithm. The threshold will be computed for β with $P_{Target} = 0.05$. The primary cost measure for the 2020 CTS Challenge is then defined as

$$C_{Primary} = C_{Norm_\beta} \tag{8}$$

Similar to 2019 CTS Challenge, the evaluation trials will be divided into several, but fewer, partitions. Each partition is defined as a combination of the number of enrollment segments (1 vs 3) and speaker gender (male vs female). A $C_{Primary}$ will be calculated for each partition, and the final result is a weighted

average of $C_{Primary}$'s across the various data sources.

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).

4 Data Description

The challenge dataset will comprise CTS recordings extracted from multiple data sources containing multilingual speech. All segments will be encoded as a-law sampled at 8 kHz in SPHERE formatted files. The challenge dataset will be distributed by NIST via the online evaluation platform (`https://sre.nist.gov`).

4.1 Data Organization

The challenge dataset follows a similar directory structure as in the previous evaluations:

<base_directory>/

README.txt
data/
enrollment/
test/
docs/

4.2 Trial File

The trial file, named `sre20_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_sre20 dtadhlw_sre20 a
1001_sre20 dtaekaz_sre20 a
1001_sre20 dtaekbb_sre20 a
```

The segmentid(s)-to-modelid mapping file, named `sre20_cts_challenge_enrollment.tsv`, will be located under the `docs/` directory.

4.3 Development Set

No Development set will be available for the 2020 CTS Challenge. Teams are responsible for harvesting, collecting, or creating their own Development set from the existing datasets or the world wide web.

[6] 2020 CTS Challenge segments will be single channel so this field is always "a"

4.4 Training Set

Section 2.2 describes the training condition for the 2020 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.

5 Evaluation Rules and Requirements

The 2020 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. Teams will be provided with the option to use anonymized names for the purpose of leaderboard display.

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, and these results/charts will contain the names of the participating teams involved with their consent. Participants may publish or otherwise disseminate these charts, unaltered and with appropriate reference to their source.

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

- 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.

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). Teams will be provided with the option to use anonymized names for the purpose of leaderboard display.

- 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 2020 CTS Challenge, when/if it becomes available.

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). Results on the test 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 the CTS Challenge since the data is single channel)
`LLR` - The log-likelihood ratio

For example:

```
modelid segmentid side LLR
1001_sre20 dtadhlw_sre20 a 0.79402
1001_sre20 dtaekaz_sre20 a 0.24256
1001_sre20 dtaekbb_sre20 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

To allow sites to learn from each other and increase the collective scientific knowledge, each team is required to produce and share a clear system description, covering training, tuning and inference, such that other researchers could reasonably reproduce their work. System descriptions can be uploaded and shared via the challenge web platform (`https://sre.nist.gov`).

In order for NIST to receive comparable and informative system descriptions, the following information is recommended to be included:

- Abstract
- Notable highlights (novel aspects)
- Data resources (training and development datasets)
- Algorithmic description
- Experimental results
- Hardware description and timing report (CPU/GPU resources and run-times, memory footage)

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, 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),

- performance of the submission systems on a development set, using the performance metric parameters defined in Section 3. 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, 2020
Registration period	ongoing
Evaluation data available to participants	August, 2020
System output due to NIST	ongoing
Final official results released	Periodically
Post-evaluation workshop	June, 2021