NIST 2024 Speaker Recognition Evaluation Plan

June 10, 2024

1 Introduction

The 2024 Speaker Recognition Evaluation (SRE24) 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 to (1) effectively measure system-calibrated performance of the current state of technology, (2) provide a common framework that enables the research community to explore promising new ideas in speaker recognition, and (3) 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. This document describes the task, performance metric, data, evaluation protocol, and rules/requirements for SRE24.

SRE24 will be organized similar to SRE21, focusing on speaker detection over conversational telephone speech (CTS) and audio from video (AfV). It will again offer cross-source (i.e., CTS and AfV) and cross-lingual trials, thanks to a multimodal and multilingual (i.e., with multilingual subjects) corpus collected outside North America. However, it will also introduce two new features as compared to previous SREs, including enrollment segment duration variability and shorter duration test segments.

SRE24 will offer both *fixed* and *open* training conditions to allow uniform cross-system comparisons and to understand the effect of additional and unconstrained amounts of training data on system performance (see Section 2.2). Similar to SRE21, SRE24 will consist of three tracks: audio-only, visual-only, and audio-visual, which involves automatic person detection using audio, image, and video materials. System submission is required for the audio and audio-visual tracks, and optional for the visual track. Table 1 summarizes the tracks for the SRE24.

Track	Enroll	Test	Required
Audio-only	CTS / AfV	CTS / AfV	Yes
Visual-only	Close-up Image	Video	No
Audio-Visual	CTS / AfV + Close-up Image	Video	Yes

Table 1: SRE24 tracks

Participation in the SRE24 is open to all who find the evaluation of interest and are able to comply with the evaluation rules set forth in this plan. **One such rule is that at least one site in the team must have either successfully participated in a previous SRE or submitted a reasonable system output to the CTS challenge[1].** Although there is no cost to participate in SRE24 (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 held in San Juan, Puerto Rico, on December 3-4, 2024. Information about

[1] https://sre.nist.gov/cts-challenge
[2] Workshop registration is required.

evaluation registration can be found on the SRE24 website[3].

2 Task Description

2.1 Task Definition

The task for the SRE24 is *speaker/person detection*: given a test segment and a target individual's enrollment data, automatically determine whether the target individual is present in the test segment. The test segment along with the enrollment segment(s) from a designated target individual constitute a *trial*. The system is required to process each trial independently and 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

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

where $P\left(\cdot\right)$ denotes the probability density 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 a speaker or person recognition system. The task described above can be evaluated over a *fixed* (required) or *open* (optional) training condition.

- **Fixed** – The fixed training condition designates a *common* set to facilitate a uniform algorithmic comparison of systems. The *common* training data for SRE24 are as follows:
  - NIST SRE CTS Superset (LDC2021E08)
  - 2016 NIST SRE Evaluation Test Set (LDC2019S20)
  - 2021 NIST SRE Evaluation Test and Development Set (LDC2024E10)
  - 2024 NIST SRE Evaluation Development Set (LDC2024E12)
  - JANUS Multimedia Dataset (LDC2019E55)

  Participants can obtain these data from the Linguistic Data Consortium (LDC) after signing the LDC data license agreement.

  For the tracks involving audio in the *fixed* training condition, only the specified speech data listed above may be used for system training and development, including all sub-systems (e.g., speech activity detection (SAD)) and auxiliary systems used for automatic labeling/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).

  Note: The use of pre-trained speech models on data other than what is designated above is not allowed in this condition. However, teams may use pre-trained image models for the visual tracks, e.g., for face detection and face encoding extraction. There is no restriction placed on image data used.

  Participation in the *fixed* training condition is required for the audio-only and audio-visual tracks.

- **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. Participation in this condition is optional but strongly encouraged to demonstrate the gains that can be achieved with unconstrained amounts of data.

[3] https://sre.nist.gov

Participating teams **must** provide a sufficient description of audio (speech and non-speech) and visual 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 or images provided to create a target speaker/person model. There are two enrollment conditions in SRE24:

- **One-segment** – in this enrollment condition, the system is given only one audio segment and/or image, depending on the track, to build the model of the target speaker/person. For audio trials (i.e., CTS and AfV), one segment containing approximately 10, 30, or 60 seconds [4] of speech is provided, while for visual trials, a close-up image of the target individual (e.g., a selfie) is provided.
- **Three-segment** – in this enrollment condition, the system is given three segments to build the model of the target speaker, either 10, 30, or 60 seconds each, or a mixture of all three durations. These variations allow participants to explore alternative methods of creating the three-segment models. Note that all segments are from the same phone number and in the same language, and that this condition only involves the CTS data.

2.4 Test Conditions

The test conditions in SRE24 are as follows:

- The speech duration of the test segments (whether audio or video) will be randomly sampled ranging approximately from 5 seconds to 60 seconds.
- Trials involving CTS data will be conducted with test segments from both same and different phone numbers as the enrollment segment(s).
- Trials (target and non-target) involving audio data will be conducted with test segments spoken both in the same and different languages as the enrollment segment(s).
- Trials (target and non-target) involving audio data will be conducted with test segments originating from both same and different source type (i.e., CTS vs AfV) as the enrollment segment(s).
- There will be no cross-gender trials.
- Each test video contains audio data from only a single individual, although there is no guarantee that the individual is visible in the video.

For AfV trials in the audio-only track, NIST will extract and release audio segments from videos; however, participants are responsible for extracting the relevant audio-visual data (i.e., speech or face frames) from videos for the visual-only and audio-visual tracks.

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

3 Performance Measurement

3.1 Primary Metric

A basic cost model is used to measure the speaker/person detection performance in SRE24, which is defined as a weighted sum of missed detection and false-alarm error probabilities for some decision threshold θ as follows

[4] As determined by SAD output. 60/30/10 second enrollment segments are nested.

$$C_{Det}(\theta) = C_{Miss} \times P_{Target} \times P_{Miss}(\theta) + C_{FalseAlarm} \times (1 - P_{Target}) \times P_{FalseAlarm}(\theta), \quad (2)$$

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

Track	Parameter ID	C_{Miss}	$C_{FalseAlarm}$	P_{Target}
Audio	1	1	1	0.01
	2	1	1	0.005
Visual	1	1	1	0.01
	2	1	1	0.005
Audio-Visual	1	1	1	0.01
	2	1	1	0.005

Table 2: SRE24 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

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

where $C_{Default}$ is defined as

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

Substituting the set of parameter values from Table 2 into (4) yields

$$C_{Default} = C_{Miss} \times P_{Target}. \quad (5)$$

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

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

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. The detection 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 each track in SRE24 is then defined as

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

Each track (audio, visual, audio-visual) will be divided into a number of partitions. Each partition is a combination of the following factors:

– audio (8 partitions)

    + speaker/person gender (male vs female)
    + data source match (Y vs N)
    + language match (Y vs N)
- visual (2 partitions)
    + speaker/person gender (male vs female)
- audio-visual (4 partitions)
    + speaker/person gender (male vs female)
    + language match (Y vs N)

Accordingly, the $C_{Primary}$ will be calculated for each partition, and the final result is the average of all $C_{Primary}$'s across the various partitions.

Note that although the audio-visual task will include trials involving either AfV enrollment segments or CTS enrollment segments, only cross-source trials (e.g., CTS enroll and AfV test) will be used for computing the official metric for SRE24. Accordingly, the data source and phone number match will always be set to "No" for the audio-visual trials. No partitioning/equalization will be applied to the visual trials.

In addition to $C_{Primary}$, 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.

4 Data Description

A new multimodal and multilingual corpus collected by the LDC, TELVID, will be used to compile the development and test sets for SRE24.

The TELVID corpus is composed of phone calls and video recordings collected outside North America, spoken in Tunisian Arabic, French, and English. Recruited subjects (aka *claques*) made multiple calls to people in their social network (e.g., family, friends), and recorded videos of themselves talking alone or sometimes with other people. They also supplied close-up images of their faces (i.e., selfies). The CTS segments extracted from TELVID will be encoded as a-law sampled at 8 kHz in SPHERE formatted files, while the AfV segments will be encoded as 16-bit FLAC files sampled at 16 kHz. All video data will be encoded as MPEG4.

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/
    docs/
```

4.2 Trial File

The trial files, named `sre24_{audio|visual|audio-visual}_{dev|eval}_trials.tsv` and located in the `docs` directory, are composed of a header and a set of records where each record describes a given trial. Each record is a single line with tab-separated fields in the following format:

4.2.1 audio track

`modelid<TAB>segmentid<NEWLINE>`

where

`modelid`: The enrollment identifier

`segmentid`: The test segment identifier

For example

```
modelid         segmentid
mabfihrgh_sre24 adneoysfm_sre24.sph
mabfihrgh_sre24 amrsbwpsm_sre24.sph
```

4.2.2 visual track

`imageid<TAB>segmentid<NEWLINE>`

where

`imageid`: The enrollment identifier

`segmentid`: The test segment identifier

For example

```
imageid             segmentid
ibcocorvk_sre24.jpg vabfyelzl_sre24.mp4
ibcocorvk_sre24.jpg vafrwafdt_sre24.mp4
```

4.2.3 audio-visual track

`modelid<TAB>imageid<TAB>segmentid<NEWLINE>`

where

`modelid`: The audio enrollment identifier

`imageid`: The image enrollment identifier

`segmentid`: The test segment identifier

For example

```
modelid         imageid             segmentid
mabfihrgh_sre24 iwovojqic_sre24.jpg vabfyelzl_sre24.mp4
mabfihrgh_sre24 iwovojqic_sre24.jpg vafrwafdt_sre24.mp4
```

4.3 Development Set

Participants in SRE24 will receive data for development experiments that will mirror the evaluation conditions. The data will be organized as outlined in section 4.1 and will include:

- Audio and video segments as well as close-up images from 20 individuals in TELVID, located in the `data` directory.
- Associated trial and key files, located in the `docs` directory:

- `sre24_{audio|visual|audio-visual}_dev_segment_key.tsv` contains information about the audio/video segments and images as well as the individuals within them and includes the following fields:
  - `segmentid` (segment identifier)
  - `subjectid` (LDC speaker id)
  - `gender` (male or female)
  - `source_type` (CTS or AfV)
  - `language` (Arabic, French or English)
  - `partition` (enrollment-10, enrollment-30, enrollment-60, enrollment-selfie, or test)
  - `conversationid` (conversation identifier)
- `sre24_{audio|visual|audio-visual}_dev_trials.tsv` contains information about the trials as outlined in section 4.2
- `sre24_{audio|visual|audio-visual}_dev_trial_key.tsv` contains information about the trial key. It follows the same format as the trial list with one new field:
  - `targettype` (target or nontarget)

4.4 Training Set

Section 2.2 describes the two training conditions: Fixed (required) and Open (optional). For the *fixed* training condition, SRE24 participants will receive from LDC a "common" set of data resources. To obtain these data, participants must sign the LDC data license agreement which outlines the terms of the data usage.

For the *open* training condition, in addition to the data noted above, participants can use other proprietary and/or publicly available data, provided that a sufficient description of data resources used will be included in their system description reports (see Section 6.4.2). Participants are encouraged to submit results for the contrastive *open* training condition to demonstrate the value of additional data.

All training sets for the *fixed* condition will be available directly from the LDC[5].

5 Evaluation Rules and Requirements

SRE24 is conducted as an open evaluation where the test data is sent to the participants to process locally and then participants submit their system output to NIST for scoring. As such, the participants must agree 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 make at least one **valid** submission for both the audio-only and audio-visual tracks under 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/person enrollment data. The use of information about other test segments and/or other target speaker/person 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, 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, provided that the automatic system used conforms to the training data condition (fixed or open) for which it is used.

[5] https://www.ldc.upenn.edu

- The participants are allowed to use information available in the header of SPHERE or video files.

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:
  - Participants are free to publish results for their own system but **must not publicly compare their results with other named 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:[6] *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. Participants may remove the anonymity from their own results in these charts, but **must not reveal the identities of other anonymous participants** without their explicit written consent.
  - 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 submissions and

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

system descriptions. 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 defined below or create one if it does not exist. The participant is also asked to associate their 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 their 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 SRE24.

6.4 Submission Requirements

Each team must make at least one valid submission for the audio-only and the audio-visual tracks in the fixed condition, 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 can complement the speaker recognition technology.

The participants can register up to three systems per track (audio, audio-visual, and visual) per training condition (open and fixed). This results in a maximum of eighteen total registered systems, one submission per system. Under each track and training condition, one system will be designated as the primary system. Bug-fixes do not count toward this limit.

Each team is required to submit system descriptions 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 3 fields separated by tab character in the following format:

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

where
`modelid` - The enrollment identifier
`segmentid` - The test segment identifier

`LLR` - The log-likelihood ratio

For example:

```
modelid segmentid LLR
1001_sre24 dtadhlw_sre24 0.79402
1001_sre24 tewdfaz_sre24 0.24256
1001_sre24 daaekbb_sre24 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 for each system (all primary and alternative systems). A 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 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 submissions for that system on the SRE24 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, and
- 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/face 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 Tentative Schedule

Milestone	**Date**
Evaluation plan published	June 11, 2024
Evaluation registration period	June - September 2024
Training and development data available to participants	June 2024
Evaluation data available to participants	August 2024
System output and system description due to NIST	October 2024
Final official results released	November 2024
Workshop registration period Post-evaluation workshop	November 2024 December 3 - 4, 2024