Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

The 2024 NIST Speaker Recognition Evaluation

Craig Greenberg[1], *Lukas Diduch*[1], *Audrey Tong*[1], *Elliot Singer*[2], *Trang Nguyen*[2], *Robert Dunn*[2], *Lisa Mason*[3], *Beth Matys*[3]

[1]NIST ITL/IAD/Multimodal Information Group, MD, USA
[2]MIT Lincoln Laboratory, MA, USA
[3]U.S. Department of Defense, MD, USA

sre_poc@nist.gov

Abstract

The 2024 U.S. National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) is the latest in a series of SREs conducted by NIST since 1996. The SRE24 evaluation task was automated person detection and had three evaluation tracks: audio, visual, and audio-visual. New SRE24 features included: variable duration enrollment segments, shorter duration test segments, segments containing multiple persons, and updated cost function parameters. 11 teams consisting of 26 sites participated in SRE24. Evaluation results indicate audio-visual fusion produces significant performance improvements over audio or visual systems. Further audio results analyses indicate: use of unrestricted training data produced significant performance gains; some systems performed better using three 10s segments than three 60s segments for enrollment data, with overall differences being less than expected; and, single model systems performed competitively with the fused multi-model systems.

Index Terms: speaker recognition, speaker recognition evaluation, SRE, NIST SRE

1. Introduction

The U.S. National Institute of Standards and Technology (NIST) has been conducting evaluations of automatic speaker recognition since 1996, and starting in 2019, began conducting evaluations of automatic person recognition by including visual information in addition to audio information [1, 2]. The goal of these evaluations 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 2024 NIST Speaker Recognition Evaluation (SRE24) was conducted in the summer-fall of 2024 and is the 19th in the NIST SRE series [3].

Similar to recent evaluations in the NIST Speaker Recognition Evaluation (SRE) series, SRE24 consisted of three tracks (*audio*, *audio-visual*, and *visual*), distinguished by whether audio, audio-visual, or visual data was available to the systems. Two training conditions were offered that determined whether the data available for system training was limited to a pre-specified set of data (*fixed* training condition) or not (*open* training condition)[1] [5]. Participation in the audio and the audio-visual tracks and fixed training condition was mandatory, while participation in the visual track or the open training condition was optional. Also similar to recent SREs, the data for the audio track in SRE24 focused on conversational telephone speech (CTS) and audio from video (AfV), and included cross-source (i.e., CTS/AfV) and crosslingual trials using data collected outside North America.

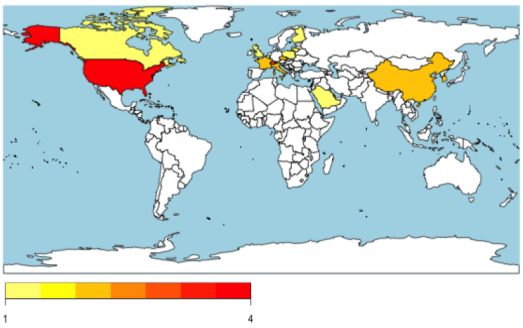


Figure 1: *A heatmap of SRE24 participation by country. Colors denote number of teams per country*

SRE24 also introduced three new features. First, instead of limiting the speech duration of enrollment data to 60 seconds, SRE24 introduced speech duration variability for the enrollment segments using segments of approximately 10, 30, and 60 seconds. Second, SRE24 used test segments in the [5s, 60s] range instead of the normal [10s, 60s]. Lastly, SRE24 also included segments including multiple persons speaking and/or visible for both enrollment and test; however, diarization marks were provided for target speakers for enrollment segments with multiple speakers. A total of 26 sites from academia, industry, and government organizations formed 11 teams that participated in SRE24. A heatmap of SRE24 participation by country can be found in Figure 1. It should be noted that all participant information, including country, was self-reported

2. Task Description

As in recent SREs, SRE24 extended the traditional task from automatic speaker recognition[2], to automatic person detection by including visual data (i.e., still "selfie" images and video recordings without sound) and audio-visual data (i.e., video recordings with sound) as part of the evaluation. The person recognition task in SRE24 is framed as person detection, i.e., given enrollment data and a test recording, determine whether the person in the enrollment data is also present in the test recording. The systems submitted to SRE24 are presented with a series of *trials*, where each trial consists of enrollment data

[1]Unlike other recent SREs, SRE19 consisted of only the *open* training condition [4].

[2]most typically the evaluation task has been automatic speaker detection [1].

10.21437/Interspeech.2025-2018

paired with a test recording. For each trial, systems were required to output a log-likelihood ratio (LLR), defined as

$$LLR(t_i) = \frac{P(t_i|H_0)}{P(t_i|H_1)} \quad (1)$$

where $P(\cdot)$ denotes the probability density function, and H_0/H_1 represent the null/alternative hypothesis (i.e., the enrollment data does/doesn't contain a person in common with the test segment), and t_i is the enrollment data and test segment pair making up the i^{th} trial.

3. Data

3.1. Training set

As mentioned previously, SRE24 consisted of *fixed* and *open* training conditions. The purpose of the *fixed* training condition, which was required for all participants, is to have a common dataset that participants used to train their system in order to decouple changes in system performance that result from the training data versus from the model and/or training algorithm. The SRE24 *fixed* training set consisted of the following datasets, made available to participants by the University of Pennsylvania's Linguistic Data Consortium (LDC):

- SRE CTS Superset (LDC2021E08) [6]
- SRE16 Evaluation Test Set (LDC2019S20) [7]
- SRE21 Evaluation Test and Dev Set (LDC2024E10) [5]
- SRE24 Evaluation Dev Set: Data (LDC2024E12)
- SRE24 Evaluation Dev Set: Annotations (LDC2024E34)
- JANUS Multimedia Dataset (LDC2019E55) [8]

Publicly available, non-speech audio and data could also be used, provided that the data was described appropriately in the final system description (e.g., noise samples, impulse responses, filters).

In the *open* training condition, the use of proprietary and/or publicly available data for system training (beyond the data provided in the *fixed* training condition) was permitted, and participants were required to describe the data used as part of their evaluation submission. Participation in the *open* condition was optional, but highly encouraged.

3.2. Development and test sets

NIST provided a small amount of development (DEV) data, meant to be a representative sample of the SRE24 evaluation (TEST) data, that sites were free to use for system development and training. The SRE24 DEV and TEST datasets were pulled from the TELVID corpus [9], a new multilingual and multimodal corpus collected by the LDC. TELVID consists of telephone calls, video recordings, and selfies that were collected outside North America, and were spoken in English, French, and Tunisian Arabic. All CTS audio segments extracted from TELVID were encoded as a-law sampled at 8 kHz in SPHERE formatted files, while the AfV audio segments were encoded as 16-bit FLAC files sampled at 16 kHz. All video data was encoded as MPEG4. There was no overlap in speakers between the DEV and TEST datasets.

SRE24 had two enrollment conditions, which defined the number of speech segments or images provided to create a target speaker/person model: 1-segment and 3-segment enrollment. Two new enrollment features were also introduced in SRE24: varying duration enrollment segments and enrollment segments that could contain multiple speakers. In the 1-segment condition, only one audio segment, image, or audio segment with an image was provided 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 of speech was provided, while for visual trials, a close-up image of the target individual (e.g., a selfie) was provided. In the 3-segment enrollment condition, which only involved the CTS data, three segments from the same phone number and in the same language were provided to build the model of the target speaker/person, either 10, 30, or 60 seconds each, or a mixture of all three durations. Diarization marks of the target speakers were provided in the cases where multiple speakers were present in an enrollment segment. For both the CTS and AfV data, the speech duration of the test segments was randomly sampled in the 5 seconds to 60 seconds range. Counts and other data statistics for the SRE24 DEV and TEST data can be found in Table 1.

4. Performance Measurement

The primary performance metric used in SRE24 was similar to the metrics used in recent SREs, a detection cost function defined as a weighted linear combination of false-reject (miss) and false-accept (false-alarm) rates. Given a decision threshold θ, the normalized cost function for SRE24 is defined as

$$C_{Norm}(\theta) = P_{Miss} + \beta \times P_{FA}(\theta) \quad (2)$$

where

$$\beta = \frac{C_{FA}}{C_{Miss}} \times \frac{1 - P_{Target}}{P_{Target}}. \quad (3)$$

The parameters C_{Miss} , C_{FA}, and P_{Target} are defined as the cost of a missed detection, the cost of a false-alarm, and the *a priori* probability of the specified target individual, respectively. For all tracks in SRE24, the primary metric, $C_{Primary}$, is the average of C_{Norm} at two operating points along the detection error trade-off (DET) curve [10], and parameters

$$C_{Miss} = C_{FA} = 1; P_{Target} = [0.01, 0.005]. \quad (4)$$

Actual detection costs were computed from the trial scores by applying the Bayes-optimal detection threshold of $\log(\beta)$, where log denotes the natural logarithm. Minimum detection costs were computed by taking the minimum cost over all thresholds. More information on the derivation of metrics can be found in the SRE24 evaluation plan [3].

Each evaluation track (audio, visual, audio-visual) was partitioned based on data properties, and C_{Norm} was computed for each partition prior to being averaged. Each partition is a combination of the following factors:

- audio (8 partitions)
  - + speaker/person sex (male vs female)
  - + data source match (Y vs N)
  - + language match (Y vs N)
- visual (2 partitions)
  - + speaker/person sex (male vs female)
- audio-visual (4 partitions)
  - + speaker/person sex (male vs female)
  - + language match (Y vs N)

Table 1: *Data statistics for SRE24 development (DEV) and TEST subsets per track. M and F denote male and female, respectively.*

Track	Subset	#speakers (M / F)	#Enroll (1 / 3 segment(s))	#Test segments	#Target trials	#Non-target trials
Audio	DEV	10 / 10	1023 / 116	2077	110,738	1,064,760
	TEST	124 / 163	14,408 / 1590	29,487	1,565,121	6,930,082
Visual	DEV	10 / 10	20 / -	455	455	4095
	TEST	124 / 163	287 / -	6848	6848	783,560
Audio-Visual	DEV	10 / 10	1023 / 116	455	25,356	233,216
	TEST	124 / 163	14,408 / 1590	6848	382,274	1,607,411

5. Baseline Systems

In order to assist with test set creation and setting evaluation parameters, person recognition baseline systems were developed, one for audio and one for video. The intention was not to carefully optimize the systems for best performance on the test data, but instead for the baseline systems performance to be consistent with recent, off-the-shelf person recognition systems.

5.1. Audio (Speaker Recognition) System

The audio baseline consisted of a neural embedding based system, specifically a Res2Net50 model trained with additive angular margin loss. The data used to train the model came from the CTS Superset, SRE16, and SRE21, all of which are contained within the SRE fixed training dataset (see Sec. 3.1), and included data augmentation using MUSAN [11] and simulated room impulse responses (RIRs). The CTS data was upsampled from 8 kHz to 16 kHz, and no hyperparameter tuning or score calibration took place. See Figure 2 for a diagram illustrating the components of the audio system pipeline.

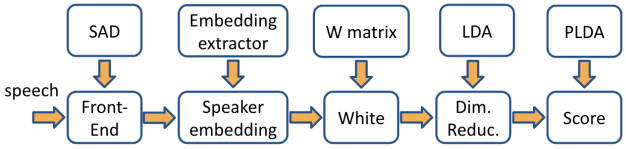


Figure 2: *An illustration of the SRE24 audio baseline system.*

5.2. Video (Face Recognition) System

The video baseline also used a neural embedding based system, in this case a ResNet101 model trained with angular loss based on InsightFace [12], using RetinaFace [13] for face detection. This was followed by agglomerative clustering of the face embeddings using the maximum cluster score as the linkage function, followed by score normalization (s-norm). An illustration of the components making up the visual system pipeline can be found in Figure 3.

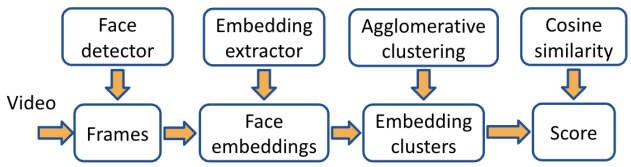


Figure 3: *An illustration of the SRE24 visual baseline system.*

6. Results and Discussion

Figure 4 displays the actual and minimum $C_{Primary}$ scores for the primary systems submitted to the fixed training condition of the audio evaluation track. The calibration error (which can be understood as the difference between the minimum and actual scores) in general was quite low, and approximately half of the systems submitted to SRE24 exhibited better performance on the SRE24 test set than the SRE24 audio baseline system.

Similarly, Figure 5 and Figure 6 display the actual and minimum $C_{Primary}$ scores for the primary systems submitted to the fixed training condition of the visual and audio-visual evaluation tracks, respectively. Calibration error seemed most prevalent in the audio visual evaluation track, and the teams submitting the systems exhibiting best performance in the visual track were not entirely the same as those teams that submited the systems exhibiting best performance in the audio track.

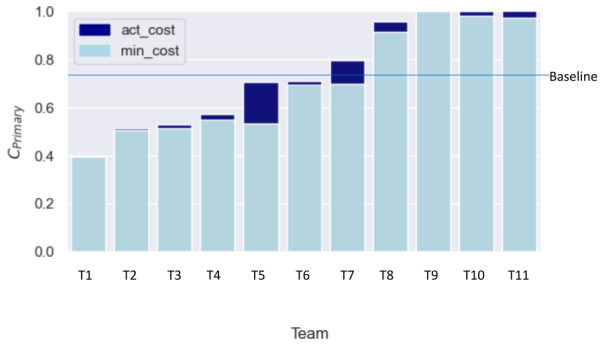


Figure 4: *The minimum and actual $C_{Primary}$ results for primary systems in the audio track, fixed training condition*

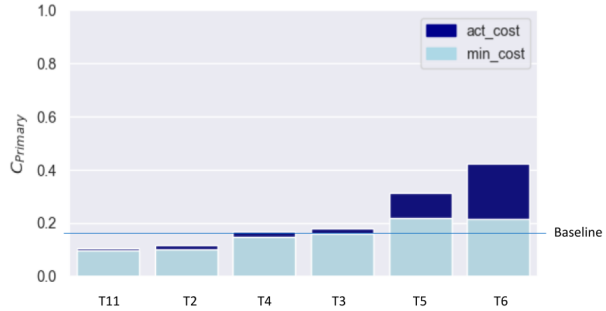


Figure 5: *The minimum and actual $C_{Primary}$ results for primary systems in the visual track, fixed training condition*

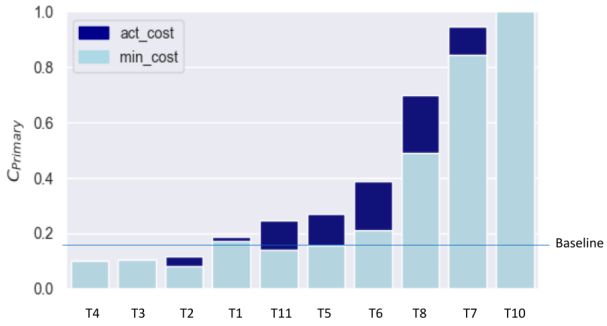


Figure 6: *The minimum and actual $C_{Primary}$ results for primary systems in the audio-visual track, fixed training condition*

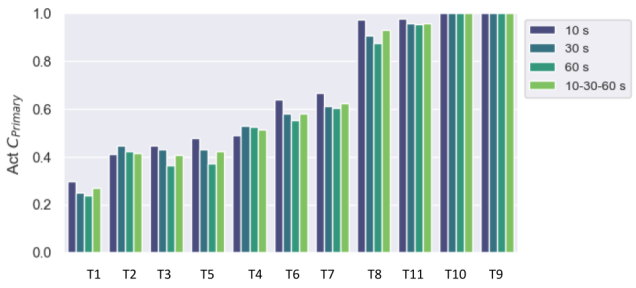


Figure 8: *The actual $C_{Primary}$ results for primary systems in the audio track for 3 segment enrollment comparing the duration of the enrollment segments.*

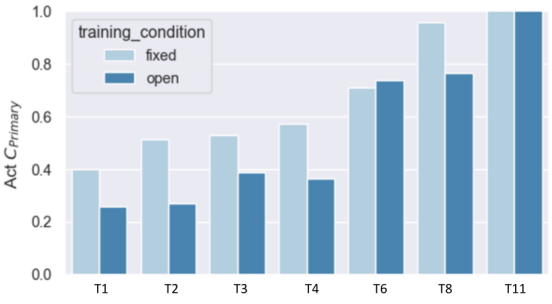


Figure 7: *The actual $C_{Primary}$ results for primary systems in the audio track, comparing the fixed vs open training condition*

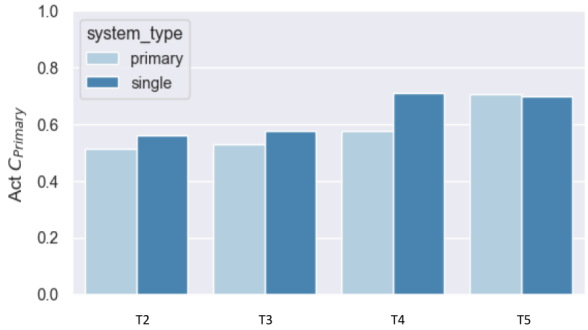


Figure 9: *The actual $C_{Primary}$ results for primary systems in the audio track, comparing single system vs primary system, which could include multiple models.*

Figure 7 compares the actual $C_{Primary}$ values for the systems submitted to the fixed vs open training conditions in the SRE24 audio track. For many of the teams, the systems submitted to the open training condition exhibited substantially better actual $C_{Primary}$ scores than the corresponding system submitted to the fixed training condition. This phenomenon seemed to be of greater magnitude in SRE24 than was the case for SRE21 [5].

In Figure 8, the actual $C_{Primary}$ scores for the three segment enrollment trials in audio track can be seen for each team's primary system. For some systems, enrollment segments with longer duration did not necessarily lead to improved performance.

A comparison between the primary systems submitted to SRE24, which typically involved the fusion of several models per system, versus the team-chosen "best" single-model systems can be seen in Figure 9. The difference in actual $C_{Primary}$ values on the SRE24 test data between the primary and single-model system appears to be relatively modest, suggesting that some single-model systems might be a reasonable choice in deployment environments that focus on speed of operation and ease of maintenance.

7. Conclusion

The 2024 NIST speaker recognition evaluation is the latest in the SRE series, and it focused on variable duration enrollment recordings, shorter duration test recordings, recordings containing multiple persons, and a lower false alarm region of performance than in previous SREs. Similar to recent SREs, the evaluated single model systems performed competitively with the evaluated multi-model systems, and, more so than other SREs, for the audio track, systems trained with an unconstrained dataset outperformed systems trained using a smaller fixed dataset. SRE24 was the first SRE to have multiple enrollment recordings of variable duration, and some systems performed better when enrolling with shorter duration segments. This paper lays out the general information about and most basic results of SRE24. Many more results and analysis have and will be obtained, which we hope to publish in the future.

8. Disclaimer

The results presented in this paper are not to be construed or represented as endorsements of any participants system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

The work of MIT Lincoln Laboratory (MITLL) is sponsored by the Department of Defense under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

9. References

[1] C. S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, "Two Decades of Speaker Recognition Evaluation at the National Institute of Standards and Technology," *Computer Speech & Language*, vol. 60, p. 101032, 2020.

[2] O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2019 NIST Audio-Visual Speaker Recognition Evaluation," in *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 259–265.

[3] NIST SRE Planning Team, "NIST 2024 Speaker Recognition Evaluation Plan," https://www.nist.gov/system/files/documents/2024/06/11/NIST_2024_Speaker_Recognition_Evaluation_Plan.pdf, accessed: 2025-01-29.

[4] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "The 2019 NIST Speaker Recognition Evaluation CTS Challenge." in *Odyssey*, 2020, pp. 266–272.

[5] S. O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "The 2021 NIST Speaker Recognition Evaluation," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 322–329.

[6] S. O. Sadjadi, "NIST SRE CTS Superset: A large-scale dataset for telephony speaker recognition," *arXiv preprint arXiv:2108.07118*, 2021.

[7] S. O. Sadjadi, C. S. Greenberg, T. Kheyrkhah, K. Jones, K. Walker, S. Strassel, and D. Graff, "2016 NIST Speaker Recognition Evaluation Test Set," https://https://catalog.ldc.upenn.edu/LDC2019S20, accessed: 2025-01-29.

[8] G. Sell, K. Duh, D. Snyder, D. Etter, and D. Garcia-Romero, "Audio-visual person recognition in multimedia data from the IARPA Janus program," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3031–3035.

[9] Anonymous, "Telvid: A multilingual multi-modal corpus for speaker recognition," 2025, submitted to Interspeech 2025.

[10] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1997, pp. 1895–1898.

[11] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[12] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, p. 5962–5979, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2021.3087709

[13] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," 2019. [Online]. Available: https://arxiv.org/abs/1905.00641