The 2019 NIST Speaker Recognition Evaluation CTS Challenge

Seyed Omid Sadjadi[1], Craig Greenberg[1], Elliot Singer[2,†], Douglas Reynolds[2,†], Lisa Mason[3], Jaime Hernandez-Cordero[3]

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

craig.greenberg@nist.gov

Abstract

In 2019, the U.S. National Institute of Standards and Technology (NIST) conducted a leaderboard style speaker recognition challenge using conversational telephone speech (CTS) data extracted from the unexposed portion of the Call My Net 2 (CMN2) corpus previously used in the 2018 Speaker Recognition Evaluation (SRE). The SRE19 CTS Challenge was organized in a similar manner to SRE18, except it offered only the *open* training condition. In addition, similar to the NIST i-vector challenge, the evaluation set consisted of two subsets: a *progress* subset, and a *test* subset. Trials for the *progress* subset comprised 30% of the target speakers from the unexposed portion of the CMN2 corpus and was used to monitor progress on the leaderboard, while trials from the remaining 70% of the speakers were allocated for the *test* subset, which was used to generate the official final results determined at the end of the challenge. Which subset (i.e., *progress* or *test*) a trial belonged to was unknown to challenge participants, and each system submission had to contain outputs for all of the trials. The SRE19 CTS Challenge also served as a prerequisite for entrance to the main SRE19 whose primary task was audio-visual person recognition. A total of 67 organizations (forming 51 teams) from academia and industry participated in the CTS Challenge and submitted 1347 valid system outputs. This paper presents an overview of the evaluation and several analyses of system performance for all primary conditions in the CTS Challenge. Compared to the CTS track of SRE18, the SRE19 CTS Challenge results indicate remarkable improvements in performance which are mainly attributed to 1) the availability of large amounts of in-domain development data (publicly available and/or proprietary) from a large number of labeled speakers, 2) speaker representations (aka embeddings) extracted using extended and more complex end-to-end neural network frameworks, and 3) effective use of the provided large development set.

1. Introduction

The United States National Institute of Standards and Technology (NIST) organized the 2019 Speaker Recognition Evaluation (SRE19) in the summer–fall of 2019. It was the latest in the ongoing series of speaker recognition technology evaluations conducted by NIST since 1996 [1, 2]. 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 to explore promising new ideas in speaker recognition, and 3) to support the community in their development of advanced technology incorporating these ideas. The basic task in the NIST SREs is speaker detection, that is, determining whether a specified target speaker is talking in a given test speech recording.

SRE19 consisted 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 collected by the Linguistic Data Consortium (LDC), which was also previously used to extract the SRE18 CTS development and test sets, and 2) a regular evaluation using audio-visual material extracted from the unexposed portions of the Video Annotation for Speech Technology (VAST) corpus [3], also collected by the LDC. This paper describes the task, the performance metric, data, and the evaluation protocol as well as results and performance analyses of submissions for the SRE19 CTS Challenge. The Audio-Visual SRE19 overview and results is described in another paper [4]. It is worth noting here that the CTS challenge also served as a prerequisite for the audio-visual evaluation, meaning that in order to participate in the regular evaluation, one must have first completed the challenge (i.e., submitted to NIST valid system outputs along with sufficiently detailed system description reports). SRE19 was coordinated entirely online using a freshly designed web platform[1] deployed on Amazon Web Services (AWS)[2] that supported a variety of evaluation-related services such as registration, data license agreement management, data distribution, system output submission and validation/scoring, and system description/presentation uploads.

The SRE19 CTS Challenge was organized in a similar manner to the CTS track of SRE18 [5], except it only offered the *open* training condition in which participants were allowed to use any publicly available and/or proprietary data for system training and development purposes. In addition, a much larger development set was released for the SRE19 CTS Challenge which contained the entire SRE18 development and test sets including segments from 213 labeled speakers as well as segments from more than 1000 unlabeled speakers. Furthermore, similar to the NIST i-vector speaker recognition challenge [6], the evaluation set consisted of two subsets: a *progress* subset, and a *test* subset. Trials for the *progress* subset comprised 30% of the target speakers from the unexposed portion of the CMN2 corpus

[†]The work of MIT Lincoln Laboratory is sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[1]https://sre.nist.gov
[2]see Disclaimer.

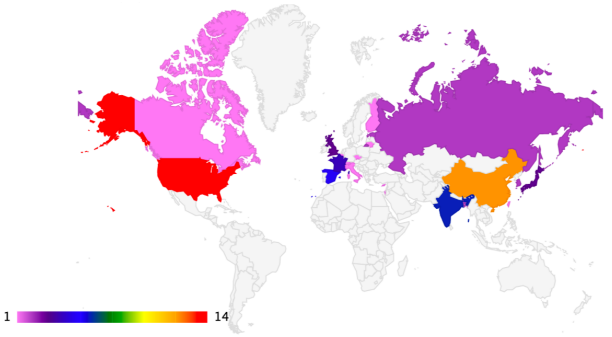


Figure 1: *Heat map of the world countries showing the number of SRE19 CTS Challenge participating sites per country.*

and was used to monitor progress on the leaderboard, while trials from the remaining 70% of the speakers were allocated for the *test* subset, which was used to generate the official final results determined at the end of the challenge. Which subset (i.e., *progress* or *test*) a trial belonged to was unknown to challenge participants, and each system submission had to contain outputs for all of the trials. The participants could make multiple submissions (up to 3 per day), and the leaderboard displayed the best submission performance results thus far received and processed. Over the course of the challenge, which ran from July 15 through October 7, 2019, a total of 51 teams, 23 of which were led by industrial institutions, from 67 sites made 1347 valid submissions (note that the participants processed the data locally and submitted only the output of their systems to NIST for scoring and analysis purposes). Figure 1 displays a heatmap representing the number of participating sites per country. It should be noted that all participant information, including country, was self-reported. The number of submissions per team in the SRE19 CTS Challenge is shown in Figure 2.

Finally, as in SRE18, and in an effort to provide a reproducible state-of-the-art baseline for the SRE19 CTS Challenge, NIST released well in advance of the evaluation period a report [7] containing the baseline speaker recognition system description and results obtained using a state-of-the-art (as of SRE18) deep neural network (DNN) embedding based system (see Section 5 for more details).

2. Task Description

The task for the SRE19 CTS Challenge was *speaker detection*, meaning given a segment of speech and the target speaker enrollment data, automatically determine whether the target speaker is speaking in the segment. A segment of speech (test segment) along with the enrollment speech segment(s) from a designated target speaker constitute a *trial*. The system is required to process each trial independently and to output a log-likelihood ratio (LLR), using natural (base e) logarithm, for that

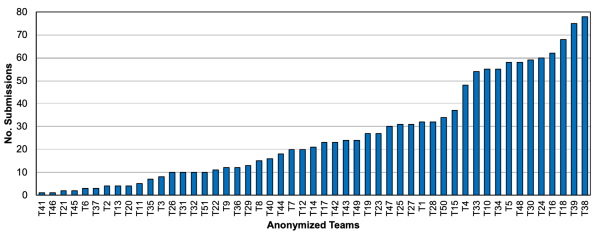


Figure 2: *Submission statistics for the SRE19 CTS Challenge.*

trial. The LLR for a given trial including a test segment s is defined as follows

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

where $P\left(\cdot\right)$ denotes the probability distribution function (pdf), and H_0 and H_1 represent the null (i.e., s is spoken by the enrollment speaker) and alternative (i.e., s is not spoken by the enrollment speaker) hypotheses, respectively.

3. Data

In this section we provide a brief description of the data released in the SRE19 CTS Challenge for system training, development, and test.

3.1. Training set

As noted previously, unlike in SRE18 which offered *fixed* and *open* training conditions, the SRE19 CTS Challenge only offered the *open* training condition that allowed the use of any publicly available and/or proprietary data for system training and development purposes. The motivation behind this decision was twofold. First, results from the most recent NIST SREs (i.e., SRE16 [8] and SRE18) indicated limited performance improvements, if any, from unconstrained training compared to *fixed* training, although participants had cited lack of time and/or resources during the evaluation period for not demonstrating significant improvement with *open* versus *fixed* training. Second, the number of publicly available large-scale data resources for speaker recognition has dramatically increased over the past few years (e.g., see VoxCeleb[3]). Therefore, removing the *fixed* training condition would allow more in-depth exploration into the gains that could 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 [5]). Nevertheless, it is worth noting here that during the discussion sessions at the post-evaluation workshop, which was held in December 2019 in Singapore, several participating teams requested the re-introduction of the *fixed* training condition to facilitate meaningful and fair cross-system comparisons in terms of core speaker recognition algorithms/approaches (as opposed to particular data) used to develop the systems.

Although SRE19 allowed unconstrained system training and development, participating teams were required to provide a sufficient description of speech and non-speech (e.g., noise samples, room impulse responses, and filters) data resources as well as pre-trained models used during the training and development of their systems.

3.2. Development and evaluation sets

For the sake of convenience, in particular for new SRE participants, NIST provided an *in-domain* development set that could be used for both system training and development purposes. This Development set simply combined the SRE18 CTS development and test sets into one package (i.e. LDC2019E59). Participants could obtain this dataset through the evaluation web platform (`https://sre.nist.gov`) after signing the LDC data license agreement. The first three rows in Table 1 summarize the statistics for this development set.

[3] `http://www.robots.ox.ac.uk/~vgg/data/voxceleb/`

Table 1: *Statistics for the SRE19 CTS Challenge development (DEV) and evaluation (EVAL), i.e., progress and test sets*

Set	**Dev/Test**	**#speakers (M / F)**	**#1-segment enrollment**	**#3-segment enrollment**	**#Test segments**	**#target/non-target trials**
CTS'18 (DEV)	Dev-labeled	9 / 16	100	25	1566	7830 / 100,265
	Dev-unlabeled	–	–	–	2332	–
	Test	70 / 118	752	188	12,135	19,298 / 2,002,332
CTS'19 (EVAL)	Progress	21 / 37	232	58	4066	20,330 / 618,360
	Test	49 / 88	547	137	9515	47,518 / 2,000,000

Table 2: *Primary partitions in the CTS Challenge* ***progress*** *set*

Partition	**Elements**	**#target**	**#non-target**
Gender	male	7095	141,900
	female	13,235	476,460
#enrollment segments	1	16,264	494,688
	3	4066	123,672
Phone# match	Y	9452	0
	N	10,878	618,320
CTS type	PSTN	15,935	484,700
	VoIP	4395	133,660

Table 3: *Primary partitions in the CTS Challenge* ***test*** *set*

Partition	**Elements**	**#target**	**#non-target**
Gender	male	15,843	433,078
	female	31,675	1,569,090
#enrollment segments	1	38,003	1,600,661
	3	9515	401,507
Phone# match	Y	24,456	0
	N	23,062	2,000,000
CTS type	PSTN	36,308	1,536,768
	VoIP	11,210	465,400

The speech segments in the SRE19 CTS Challenge development (*DEV*) and evaluation (*EVAL*) sets were extracted from the CMN2 corpus collected by the LDC to support speech technology evaluations. The CMN2 corpus consists of CTS recordings spoken in Tunisian Arabic, which were collected over the traditional Public Switched Telephone Network (PSTN) and the more recent Voice over IP (VOIP) platforms outside North America. For CMN2 data collection, the LDC recruited a few hundred speakers called *claques* who made multiple calls to people in their social network (e.g., family, friends). Claques were encouraged to use different telephone instruments (e.g., cell phone, landline) in a variety of settings (e.g., noisy cafe, quiet office) for their initiated calls and were instructed to talk for at least 8–10 minutes on a topic of their choice. All CMN2 recordings are encoded as a-law sampled at 8 kHz in SPHERE [9] formatted files.

Similar to the most recent SREs (i.e., SRE16 and SRE18), there were two enrollment scenarios for the SRE19 CTS Challenge, namely 1-segment and 3-segment conditions. As the names imply, in the 1-segment condition only one approximately 60 s speech segment was given for enrollment, while in the 3-segment condition three approximately 60 s speech segments (from the same phone number) were provided to build the model of the target speaker. It is worth noting that the 3-segment condition only involved the PSTN data, because the number of VoIP calls per *claque* was limited. As part of the *dev* set for the SRE19 CTS Challenge, an *unlabeled* set of 2332 segments (with speech duration uniformly distributed in 10 s to 60 s range) was also made available by NIST. The *unlabeled* segments were extracted from the non-*claque* side of the PSTN/VoIP calls.

For the SRE19 CTS Challenge, the evaluation trials were divided into two subsets: a *progress* subset, and a *test* subset. Trials for the *progress* subset comprised 30% of the target speakers from the unexposed portion of the CMN2 corpus and was used to monitor progress on the leaderboard, while trials from the remaining 70% of the speakers were allocated for the *test* subset which was used to generate the official final results determined at the end of the challenge. The challenge test conditions were as follows:

- The speech durations of the test segments were uniformly sampled ranging approximately from 10 seconds to 60 seconds.
- Trials were conducted with test segments from both same and different phone numbers as the enrollment segment(s).
- There were no cross-gender trials.

The last two rows of Table 1 show the statistics for the SRE19 CTS Challenge *progress* and *test* subsets.

4. Performance Measurement

Similar to the past SREs, the primary performance measure for the SRE19 CTS Challenge was a detection cost defined as a weighted sum of false-reject (miss) and false-accept (false-alarm) error probabilities. Equation (2) specifies the CTS Challenge primary normalized cost function for some decision threshold θ,

$$C_{norm}(\theta) = P_{miss}(\theta) + \beta \times P_{fa}(\theta), \tag{2}$$

where β is defined as

$$\beta = \frac{C_{fa}}{C_{miss}} \times \frac{1 - P_{target}}{P_{target}}. \tag{3}$$

The parameters C_{miss} and C_{fa} are the cost of a missed detection and cost of a false-alarm, respectively, and P_{target} is the *a priori* probability that the test segment speaker is the specified target speaker. The primary cost metric, $C_{primary}$ for the CTS Challenge was the average of normalized costs calculated at two points along the detection error trade-off (DET) curve [10], with $C_{miss} = C_{fa} = 1$, $P_{target} = 0.01$ and $P_{target} = 0.005$. Here, $\log(\beta)$ was applied as the detection

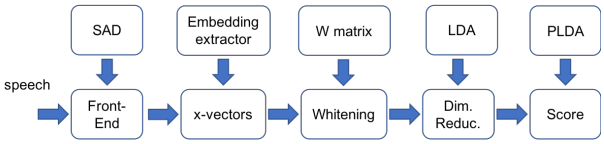


Figure 3: *A simplified block diagram of the baseline speaker recognition system for the SRE19 CTS Challenge.*

threshold θ for computing the actual detection costs. Additional details can be found in the SRE19 CTS Challenge evaluation plan [11].

Similar to the recent SREs (i.e., SRE16 and SRE18), the test data was divided into 16 partitions. Each partition is defined as a combination of: speaker gender (male vs female), number of enrollment segments (1 vs 3), enrollment-test phone number match (Yes vs No), and CTS source type (PSTN vs VoIP). However, because no actual "phone number" metadata was available for either enrollment or test segments extracted from the VoIP calls, the phone number match field only contained "N" for those calls, thereby reducing the effective number of partitions to 12. Also, all non-target trials are from the different (as opposed to the same) phone number partition, assuming each phone number would be only used by one individual. More information about the various partitions in the SRE19 CTS Challenge *progress* and *test* subsets can be found in Tables 2 and 3. $C_{primary}$ was calculated for each partition, and the final result was the average of all the partitions' $C_{primary}$'s.

Also, a minimum detection cost was computed by using the detection thresholds that minimized the detection cost. Note that for minimum cost calculations, the counts for each condition set was equalized before pooling and cost calculation, that is, the minimum cost was computed using a single threshold not one per condition set.

5. Baseline system

In this section, we describe the x-vector baseline system setup including speech and non-speech data used for training the system components as well as the hyper-parameter configurations used in our evaluations. Figure 3 shows a block diagram of the x-vector baseline system. The x-vector system is built using Kaldi [12] (for x-vector extractor training) and the NIST SLRE toolkit for back-end scoring.

5.1. Data

The x-vector baseline system was developed using the data recipes available at `https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2` as well as `https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2`. The x-vector extractor for the *progress* set was trained entirely using speech data extracted from combined VoxCeleb 1 and 2 corpora, while the x-vector extractor for the *test* set used the prior SRE data (i.e., SRE04-10 as in the Kaldi `sre16` recipe) in addition to the combined VoxCeleb. This was done to ensure the baseline results would serve as a fair comparison point for the first time participants who might only have access to the VoxCeleb data, but not to the prior SRE data. In order to increase the diversity of the acoustic conditions in the training set, a 5-fold augmentation strategy was used that added four corrupted copies of the original recordings to the training list. The recordings were corrupted by either digitally adding noise (i.e., babble, general noise, music) or convolving with simulated and measured room impulse responses (RIR). The noise and RIR samples are freely available from `http://www.openslr.org` (see [13] for more details).

All recordings are downsampled to 8 kHz using `sox`.

5.2. Configuration

For speech parameterization, we extracted 23-dimensional MFCCs (including c0) from 25 ms frames every 10 ms using a 23-channel mel-scale filterbank spanning the frequency range 20 Hz–3700 Hz. Before dropping the non-speech frames using an energy based SAD, a short-time cepstral mean subtraction was applied over a 3-second sliding window.

For x-vector extraction, an extended TDNN with 12 hidden layers and rectified linear unit (RELU) non-linearities was trained to discriminate among the speakers in the training set. After training, embeddings were extracted from the 512-dimensional affine component of the 11^{th} layer (i.e., the first segment-level layer). More details regarding the DNN architecture (e.g., the number of hidden units per layer) and the training process can be found in [14].

Prior to dimensionality reduction through LDA (to 250), 512-dimensional x-vectors were centered, whitened, and unit-length normalized. The centering and whitening statistics were computed using the in-domain development data (i.e., LDC2019E59). For backend scoring, a Gaussian PLDA model with a full-rank Eigenvoice subspace was trained using the x-vectors extracted from either 170 k concatenated speech segments from the combined VoxCeleb sets (for the *progress* set), or 50 k speech segments from prior SRE data (for the *test* set), as well as one corrupted version randomly selected from {babble, noise, music, reverb}. The PLDA parameters were then adapted on the in-domain development data (i.e., LDC2019E59) using Bayesian maximum *a posteriori* (MAP) estimation.

Finally, the PLDA verification scores were post-processed using an adaptive score normalization (AS-Norm) scheme proposed in [15]. We used LDC2019E59 as the cohort set, and selected the top 10% of sorted cohort scores for calculating the normalization statistics.

It is worth emphasizing that the configuration parameters employed to build the baseline system are commonly used by the speaker recognition community, and no attempt was made to tune the hyperparameters or data lists utilized to train the models.

6. Results and Discussion

In this section we present some key results and analyses for SRE19 CTS Challenge submissions, in terms of minimum and actual costs as well as DET performance curves.

Figure 4 shows performance of the best submissions per team per subset as well as performance of the baseline systems [7] in terms of the actual and minimum costs, for the SRE19 CTS Challenge *progress* and *test* subsets, respectively. Baseline 1 and 2 denote the baseline speaker recognition systems trained without and with prior SRE data, respectively (see Section 5 for more details). Here, the y-axis limit is set to 0.5 to facilitate cross-system comparisons in the lower cost region. Several observations can be made from the two plots. First, performance trends on the two subsets are generally similar, although slightly better results are observed on the *progress* subset compared to the *test* subset, a phenomenon which is speculated to primarily result from overtuning/overfitting of the submission systems on the *progress* set. Second, nearly half of

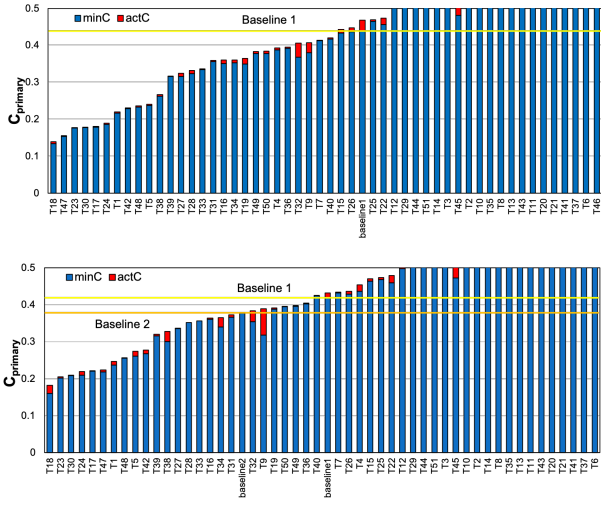


Figure 4: *Performance of the SRE19 CTS Challenge submissions in terms of actual (red) and minimum (blue) costs for the progress (top) and test (bottom) subsets.*

the submissions outperform the baseline system trained on VoxCeleb (i.e., baseline 1), while the number is smaller when compared to the baseline that utilizes the prior SRE data. Third, a majority of the systems achieve relatively small calibration errors, in particular on the *progress* subset. This is in line with the calibration performance of the submitted systems observed for the SRE18 CTS domain. Finally, it can be seen from the figures that, except for the top performing team, the performance gap among the next top-5 teams is not remarkable. A statistical analysis of performance (e.g., confidence intervals for the cost estimates) that sheds more light on actual performance differences among the top performing systems follows later in this section.

Compared to the most recent SRE (i.e., SRE18), there is a notable improvement in speaker recognition performance. Figure 5 presents a performance comparison of SRE18 versus SRE19 CTS submissions for several top performing systems, in terms of actual and minimum detection costs. Performance improvements as large as 70% are achieved by some leading systems, while for others more moderate, but consistent, improvements are observed. These performance improvements are largely attributed to 1) the availability of large amounts of in-domain development data from a large number of labeled speakers (e.g., the entire SRE18 CTS development and test data, or other proprietary in-domain data), and 2) the use of extended and more complex end-to-end neural network frameworks for speaker embedding extraction that can effectively exploit vast

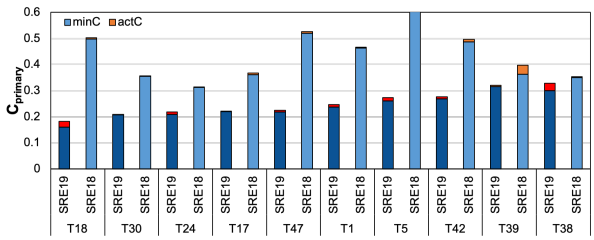


Figure 5: *Performance comparison of SRE18 vs SRE19 CTS submissions for several top performing systems.*

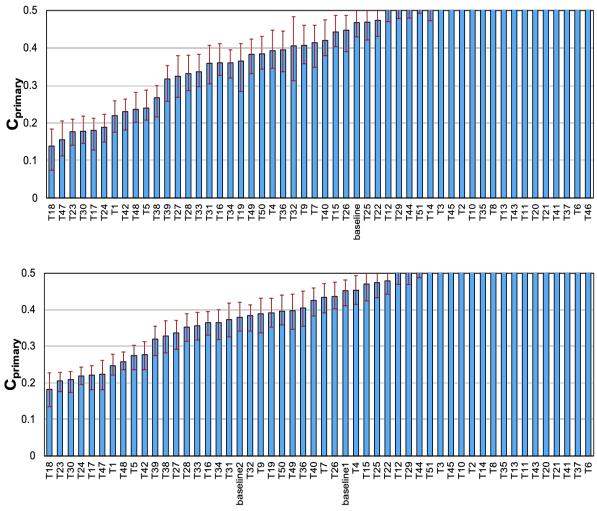
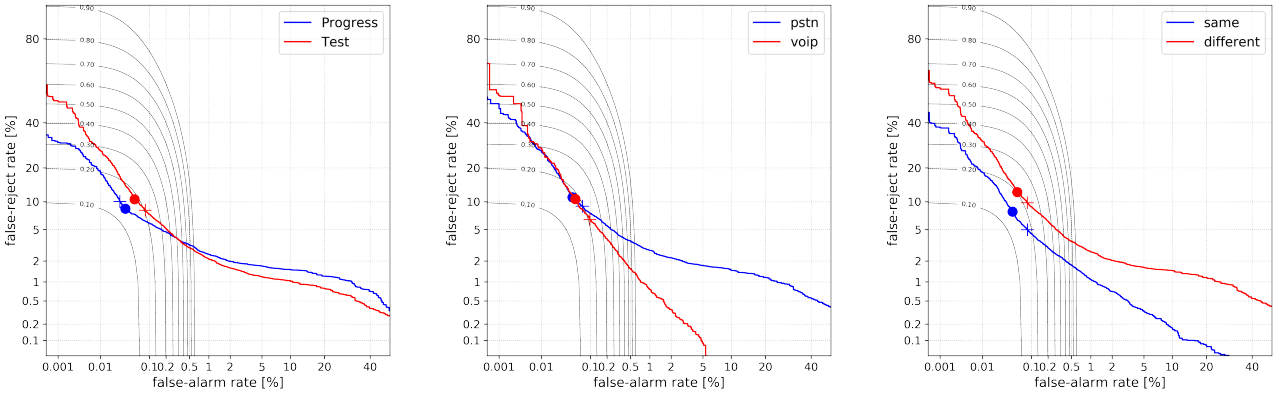


Figure 6: *Performance confidence intervals (95%) of the SRE19 CTS Challenge submissions for the progress (top) and test (bottom) subsets.*

amounts of training data made available through data augmentation and/or large-scale datasets such as VoxCeleb[3].

It is common practice in the machine learning community to perform statistical significance tests to facilitate a more meaningful cross-system performance comparison. Accordingly, to encourage the speaker recognition community to consider significance testing while comparing systems or performing model selection, we computed bootstrapping-based 95% confidence intervals using the approach described in [16]. To achieve this, we sampled, with repetition, the unique speaker model space along with the associated test segments 1000 times, which resulted in 1000 actual detection costs, based on which we calculated the quantiles corresponding to the 95% confidence margin. Figure 6 shows the performance confidence intervals (around the actual detection costs) for each submission for both the *progress* (top) and *test* (bottom) subsets. It can be seen that, in general, the *progress* subset exhibits a wider confidence margin than the *test* subset, which is expected because it has a relatively smaller number of trials. Also, notice that a majority of the top systems may perform comparably under different samplings of the trial space. Another interesting observation that can be made from the figure is that systems with larger error bars may be less robust than systems with roughly comparable performance but smaller error bars. For instance, although T_{18} achieves the lowest detection cost, it exhibits a much wider confidence margin compared to the second top system. These observations further highlight the importance of statistical significance tests while reporting performance results or in the model selection stage during system development, in particular when the number of trials is relatively small.

Figures 7a, 7b, and 7c show speaker recognition performance for the top performing submission in terms of DET curves as a function of: evaluation subset (i.e., *progress* vs *test*), CTS type (i.e., PSTN vs VoIP), and enrollment-test phone number match for PSTN calls (same vs different), respectively. The solid black curves in Figures 7a, 7b, and 7c represent equi-cost contours, meaning that all points on a given contour correspond to the same detection cost value. Firstly, consistent with our observations from Figure 4, the detection errors (i.e., false-alarm and false-reject errors) across the operating points of interest (i.e., the low false-alarm region) for the *test* subset are



(a) (b) (c)

Figure 7: *DET performance curves for the leading system by (a) data source (**progress** vs **test**), (b) CTS type (PSTN vs VoIP), and (c) enrollment-test phone number match (**same** vs **different**). Filled circles and crosses represent minimum and actual costs, respectively.*

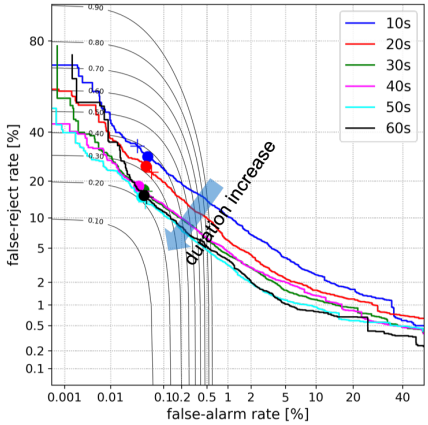


Figure 8: *DET curve performances of a top performing system for the various segment speech durations (10 s–60 s) in the test set.*

greater than those for the *progress* subset. In addition, the calibration error for the *test* subset is relatively larger. As noted previously, we speculate that these primarily result from overtuning/overfitting of the submission systems on the *progress* set. Secondly, contrary to the results observed on the SRE18 CTS domain where performance on the PSTN data was better than that on the VOIP data across all operating points, it seems from Figure 7b that for the operating points of interest (i.e., the low false-alarm region) the performance on the PSTN data is comparable to that on the VoIP data. We speculate this is due to the large amounts of VOIP data available for system development in SRE19 compared to SRE18 where only a small amount of VOIP development data was supplied. Finally, as expected, better performance is observed when speech segments from the same phone number are used in trials. Nevertheless, the error rates still remain relatively high even for the same phone number condition. This indicates that there are factors other than the channel (phone microphone) that may adversely impact speaker recognition performance. These include both intrinsic (variations in speaker's voice) and extrinsic (variations in background acoustic environment) variabilities.

Figure 8 shows DET curves for the various test segment speech durations (10 s–60 s) in the SRE19 CTS Challenge. Results are shown for a top performing submission. Limited performance difference is observed for durations longer than 40 s. However, there is a rapid drop in performance when the speech duration decreases from 30 s to 20 s, and similarly from 20 s to 10 s. This indicates that additional speech in the test recording helps improve the performance when the test segment speech duration is relatively short (below 30 seconds), but does not make a noticeable difference when there is at least 30 seconds of speech in the test segment. It is also worth noting that the calibration error (i.e., the gap between filled circles and crosses) increases as the test segment duration decreases.

7. Conclusion

In 2019, NIST organized the first leaderboard style SRE activity where raw CTS data (as opposed to embeddings) were provided as input to the systems. In this paper, we presented a summary of the SRE19 CTS Challenge (including the task, data, performance metric, the baseline system, as well as results and performance analyses) whose primary objectives were to systematically measure the recent progress in speaker recognition technology, in particular in the CTS domain, and to stimulate new ideas and collaborations. In addition, the CTS Challenge served as a prerequisite for the Audio-Visual SRE19. Results and analyses presented in this paper indicate great progress in speaker recognition technology compared to SRE18, with relative performance improvements as large as 70% for the leading system. Nevertheless, the performance gap on certain data partitions (e.g., PSTN vs VOIP or same vs different phone number) remains relatively large, at least for certain operating regions. This motivates further research towards developing a more robust technology that can maintain performance across a wide range of operating points and conditions (e.g., new data sources, languages, and channels).

8. Disclaimer

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

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

9. References

[1] NIST, "NIST Speaker Recognition Evaluation," https://www.nist.gov/itl/iad/mig/speaker-recognition, [Online; accessed 28-December-2019].

[2] 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, 2020.

[3] J. Tracey and S. Strassel, "VAST: A corpus of video annotation for speech technologies," in *Proc. LREC*, Miyazaki, Japan, May 2018.

[4] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "The 2019 NIST audio-visual speaker recognition evaluation," in *Proc. Speaker Odyssey (submitted)*, Tokyo, Japan, May 2020.

[5] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "The 2018 NIST speaker recognition evaluation," in *Proc. INTERSPEECH*, Graz, Austria, September 2019, pp. 1483–1487.

[6] D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, C. S. Greenberg, A. F. Martin, A. McCree, M. A. Przybocki, and D. A. Reynolds, "Summary and initial results of the 2013-2014 speaker recognition i-vector machine learning challenge," in *Proc. INTERSPEECH*, Singapore, Singapore, September 2014, pp. 368–372.

[7] S. O. Sadjadi, "NIST baseline system for the 2019 speaker recognition evaluation CTS challenge," NIST, Tech. Rep., 2019.

[8] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," in *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 1353–1357.

[9] NIST, "Speech file manipulation software (SPHERE) package version 2.7," ftp://jaguar.ncsl.nist.gov/pub/sphere-2.7-20120312-1513.tar.bz2, 2012, [Online; accessed 28-December-2019].

[10] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. EUROSPEECH*, Rhodes, Greece, September 1997, pp. 1899–1903.

[11] NIST, "NIST 2019 Speaker Recognition Evaluation: CTS Challenge," https://www.nist.gov/document/2019nistspeakerrecognitionchallengev8pdf, 2019, [Online; accessed 27-December-2019].

[12] D. Povey *et al.*, "Kaldi Speech Recognition Toolkit," https://github.com/kaldi-asr/kaldi, [Online; accessed 01-March-2018].

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*. Calgary, AB: IEEE, April 2018, pp. 5329–5333.

[14] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. IEEE ICASSP*, May 2019, pp. 5796–5800.

[15] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *Proc. INTERSPEECH*, August 2011, pp. 2365–2368.

[16] N. Poh and S. Bengio, "Estimating the confidence interval of expected performance curve in biometric authentication using joint bootstrap," in *Proc. IEEE ICASSP*, vol. 2, April 2007, pp. II–137–II–140.