The 2019 NIST Audio-Visual Speaker Recognition Evaluation

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 the most recent in an ongoing series of speaker recognition evaluations (SRE). There were two components to SRE19: 1) a leaderboard style Challenge using unexposed conversational telephone speech (CTS) data from the Call My Net 2 (CMN2) corpus, and 2) an Audio-Visual (AV) evaluation using video material extracted from the unexposed portions of the Video Annotation for Speech Technologies (VAST) corpus. This paper presents an overview of the Audio-Visual SRE19 activity including the task, the performance metric, data, and the evaluation protocol, results and system performance analyses. The Audio-Visual SRE19 was organized in a similar manner to the audio from video (AfV) track in SRE18, except it offered only the *open* training condition. In addition, instead of extracting and releasing only the AfV data, unexposed multimedia data from the VAST corpus was used to support the Audio-Visual SRE19. It featured two core evaluation tracks, namely audio only and audio-visual, as well as an optional visual only track. A total of 26 organizations (forming 14 teams) from academia and industry participated in the Audio-Visual SRE19 and submitted 102 valid system outputs. Evaluation results indicate: 1) notable performance improvements for the audio only speaker recognition task on the challenging *amateur* online video domain due to the use of more complex neural network architectures (e.g., ResNet) along with soft margin losses, 2) state-of-the-art speaker and face recognition technologies provide comparable person recognition performance on the *amateur* online video domain, and 3) audio-visual fusion results in remarkable performance gains (greater than 85% relative) over the audio only or visual only systems.

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 community to explore promising new ideas in speaker recognition, and 3) to support the community in their development of advanced technology incorporating these ideas.

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 Technologies (VAST) corpus [3], also collected by the LDC. This paper presents an overview of the Audio-Visual SRE19 including the task, the performance metric, data, and the evaluation protocol as well as results and performance analyses of submissions. The SRE19 CTS Challenge overview and results are described in another paper [4]. It is worth noting here that the CTS challenge also served as a prerequisite for the Audio-Visual SRE19, 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 uploads.

The Audio-Visual SRE19 was organized in a similar manner to the audio from video (AfV) track of SRE18 [5], except it only offered the *open* training condition which allowed participants to use any publicly available and/or proprietary data for system training and development purposes. Moreover, in addition to the regular audio-only track, the Audio-Visual SRE19 also introduced audio-visual and visual-only tracks. Addition of these new tracks change the basic task in the Audio-Visual SRE19 to person detection (as opposed to speaker recognition), that is, determining whether a specified target person is present in a given test video recording. System submission was required for the audio and audio-visual tracks, but optional for the vi-

Table 1: *Audio-Visual SRE19 tracks.*

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

†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.

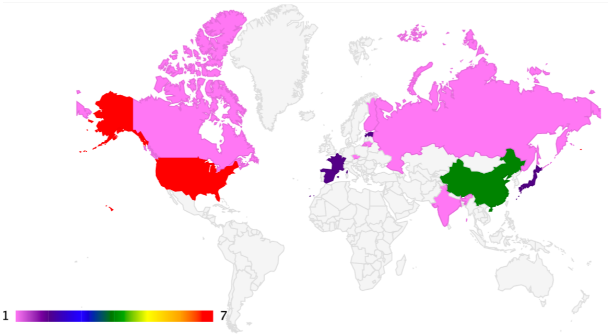

Figure 1: *Heat map of the world countries showing the number of Audio-Visual SRE19 participating sites per country.*

sual track. Table 1 summarizes the tracks for the Audio-Visual SRE19.

In addition, instead of extracting and releasing only the AfV data, unexposed multimedia data (i.e., videos) from the VAST corpus was used to support the Audio-Visual SRE19. Unlike the AfV track in SRE18 for which NIST released a very small in-domain development set containing data from only 10 speakers, SRE19 provided a much larger in-domain development set containing videos from 52 individuals from the VAST portion of SRE18 (i.e., only the videos in which the target individuals' faces were visible). In addition to the VAST development data, LDC also released selected data resources from the IARPA JANUS Benchmark-B [6], namely the JANUS Multimedia Dataset [7] which could also be used for system training and development purposes. The participants could register up to three systems for each track (i.e., audio, audio-visual, and visual), one of which under each track should have been designated as the primary system, and the other two as either contrastive or single best systems. Teams could make an unlimited number of submissions for each of the three systems until the evaluation period was over. Over the course of the evaluation, which ran from August 15, 2019 through October 21, 2019, a total of 14 teams, 8 of which were led by industrial institutions, from 26 sites made 102 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 per track (i.e., audio, visual, and audio-visual) in the Audio-Visual SRE19 is shown in Figure 2.

Finally, as in SRE18, and in an effort to provide reproducible state-of-the-art baselines for the Audio-Visual SRE19, NIST released well in advance of the evaluation period a report [8] containing descriptions of speaker and face recognition baseline systems and results obtained using these standalone state-of-the-art (as of SRE18) deep neural network (DNN) embedding based systems as well as their fusion (see Section 5 for more details).

2. Task Description

The primary task for the Audio-Visual SRE19 was *person detection*, meaning that given a test video segment and a target individual's enrollment video, automatically determine whether the target person is present in the test segment. The test segment along with the enrollment segment from a designated target individual constitute a *trial*. The system is required to process each trial independently and to output a log-likelihood ratio (LLR), using natural (base e) logarithm, for that trial. The LLR for a given trial including a test segment s is defined as follows

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

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

3. Data

In this section we provide a brief description of the data released for the Audio-Visual SRE19 for system training, development, and test.

3.1. Training set

As noted previously, unlike in SRE18 which offered both *fixed* and *open* training conditions, the Audio-Visual SRE19 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 [9] 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 and face recognition has dramatically increased over the past few years (e.g., 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.

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

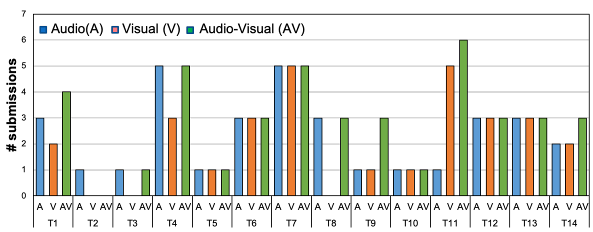

Figure 2: *Submission statistics for the Audio-Visual SRE19.*

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

Table 2: *Statistics for the JANUS Multimedia Dataset (CORE) and the Audio-Visual SRE19 development (DEV) and TEST sets.*

Set	**DEV/TEST**	**#speakers (M / F)**	**#Enroll segments**	**#Test segments**	**#Target**	**#Non-target**
JANUS (CORE)	DEV	102*	102	319	244	32,294
	TEST	258*	258	914	681	235,131
SRE19 (AV)	DEV	15 / 37	52	108	108	5508
	TEST	47 / 102	149	452	452	66,896

*gender information not available

3.2. Development and test sets

For the sake of convenience, in particular for the audio-visual and visual-only tracks, NIST provided two *in-domain* development (DEV) sets that could be used for both system training and development purposes. The Audio-Visual SRE19 *DEV* sets were as follows:

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

The JANUS Multimedia Dataset (LDC2019E55) [7], which was extracted from the IARPA JANUS Benchmark-B datatset [6], was available from the LDC, subject to approval of the LDC data license agreement. It consists of two subsets, namely CORE and FULL, each with a *DEV* and *TEST* split. We only consider the CORE subset in this paper, because it better reflects the data conditions in the Audio-Visual SRE19 *DEV* and *TEST* sets where target speakers are assumed visible. The first two rows in Table 2 summarize the statistics for the JANUS Multimedia Dataset CORE subset.

The SRE19 Audio-Visual Development (*DEV*) Set (LDC2019E56), on the other hand, contained the original videos from which the VAST portion of the SRE18 *DEV* and *TEST* sets were compiled. Participants could obtain this dataset through the evaluation web platform (`https://sre.nist.gov`) after signing the LDC data license agreement. Unexposed portions of the VAST corpus were used to compile the Audio-Visual SRE19 *TEST* set. The second two rows in Table 2 summarize the statistics for the Audio-Visual SRE19 *DEV* and *TEST* sets.

The speech segments in the Audio-Visual SRE19 *DEV* and *TEST* sets were extracted from the VAST corpus collected by the LDC to support speech technology evaluations. Unlike existing publicly available datasets derived from online "red carpet" and interview style videos featuring celebrities (e.g., VoxCeleb[3]), the VAST corpus contains *amateur* video recordings such as video blogs (Vlogs) extracted from various online media hosting services. The videos are mostly shot using personal recording devices such as cell phones in extremely diverse acoustic backgrounds, illuminations, facial poses and expressions. The videos vary in duration from a few seconds to several minutes and include speech spoken in English. Each video may contain audio-visual data from potentially multiple individuals who may or may not be visible in the recording, therefore manually produced diarization labels (i.e., speaker time marks) and *keyframe* indices[4] along with bounding boxes that mark an individual's face in the video were provided for both the *DEV* set and *TEST* set enrollment videos (but not for the test videos in either set). All video data were encoded as MPEG4. Figure 3 shows speech duration histograms for the enrollment and test segments in the Audio-Visual SRE19 *DEV* (left) and *TEST* (right) sets. Note that enrollment segment speech durations are calculated after applying diarization, while no diarization has been applied to test segments. Nevertheless, the enrollment and test histograms both appear to follow log-normal distributions, and overall they are consistent across the *DEV* and *TEST* sets.

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

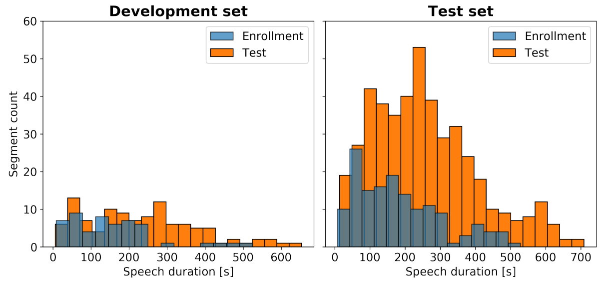

Figure 3: *Distributions of speech duration for the enrollment and test segments in the Audio-Visual SRE19 DEV and TEST sets.*

Similar to the AfV track in SRE18, there was only a 1-segment enrollment condition for the Audio-Visual SRE19 in which the system was given one video segment, that could vary in duration from a few seconds to several minutes, to build the model of the target individual. Note that for the audio track of the Audio-Visual SRE19, speech extracted from the enrollment video served as enrollment data, while for the visual track, face frame(s) (i.e., frames in which the face of the target individual was visible) extracted from the video served that purpose. Since NIST only released video files for the Audio-Visual SRE19, participants were responsible for extracting the relevant data (i.e., speech or face frames) for subsequent processing.

As in the most recent evaluations, gender labels were not provided for the enrollment segments in the *TEST* set. The test conditions for the SRE19 were as follows:

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

4. Performance Measurement

Similar to past SREs, the primary performance measure for the Audio-Visual SRE19 was a detection cost defined as a weighted sum of false-reject (miss) and false-accept (false-alarm) error probabilities. Equation (2) specifies the Audio-Visual SRE19 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 Audio-Visual evaluation was the normalized cost calculated at one operating point along the detection error trade-off (DET) curve [10], with $C_{miss} = C_{fa} = 1$, $P_{target} = 0.05$. Here, $\log(\beta)$ was applied as the detection threshold θ where log denotes the natural logarithm. Additional details can be found in the Audio-Visual SRE19 evaluation plan [11].

In addition to $C_{Primary}$, a minimum detection cost was also computed by using the detection threshold that minimized the detection cost.

5. Baseline systems

5.1. Speaker Recognition

In this section we describe the x-vector baseline speaker recognition 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 4 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.

5.1.1. Data

The x-vector baseline system was developed using the data recipe available at `https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2`. The x-vector extractor was trained entirely using speech data extracted from combined VoxCeleb 1 and 2 corpora. 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).

5.1.2. Configuration

For speech parameterization, we extracted 30-dimensional MFCCs (including c0) from 25 ms frames every 10 ms using a 30-channel mel-scale filterbank spanning the frequency range 20 Hz–7600 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$^{\text{th}}$ 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., LDC2019E56). For backend scoring, a Gaussian PLDA model with a full-rank Eigenvoice subspace was trained using the x-vectors extracted from 170 k concatenated speech segments from the combined VoxCeleb sets as well as one corrupted version randomly selected from {babble, noise, music, reverb}. The PLDA parameters were then adapted to the in-domain development data (i.e., LDC2019E56) 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 LDC2019E56 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.

5.2. Face Recognition

In this section, we describe the baseline face recognition system setup including the visual data used for training the system components as well as the hyper-parameter configurations used in our experiments. Figure 5 shows a block diagram of the baseline face recognition system which was built using open-source TensorFlow based implementations [16, 17] of 1) a face detector termed MultiTask Cascaded Convolutional Networks (MTCNN) [18], and 2) a face recognizer termed FaceNet [19] (for face encoding extraction). We use the NIST SLRE toolkit for back-end scoring.

5.2.1. Data

The baseline face recognition system utilized a pre-trained model available at `https://github.com/davidsandberg/facenet` (model name: 20180402-114759) which was trained on the VGGFace 2 dataset [20] using the Inception ResNet V1 architecture [21].

5.2.2. Configuration

We began processing by extracting one frame per second from the videos using `ffmpeg`. Then, we applied the MTCNN based face detector on the extracted frames to 1) filter out frames with no faces, and 2) compute the bounding box for the face

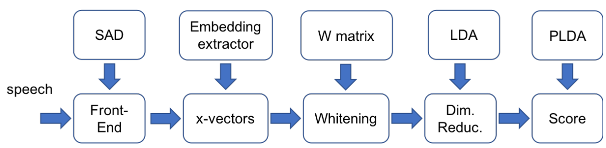

Figure 4: *A simplified block diagram of the baseline speaker recognition system for the Audio-Visual SRE19.*

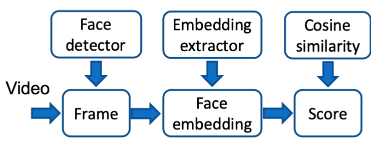

Figure 5: *A simplified block diagram of the baseline face recognition system for the Audio-Visual SRE19.*

that is closest to the center of the frame (as in [17]). Next, the face images were cropped using the bounding box coordinates, whitened (mean and variance normalized), and resized to 160×160 pixels. Finally, FaceNet was used to extract face encodings from the cropped, whitened and resized images.

For enrollment, we used the average of face encodings extracted from the enrollment video for each target individual to build a model for that individual. We only retained the face encodings that scored the highest (greater than 0.5 using cosine similarity) against the average of face encodings obtained using the manually produced bounding box coordinates for the enrollment videos. For test, we kept all face encodings extracted for each test video. In order to compute a single score for each trial involving an enrollment video and a test video, we computed the maximum of the cosine similarity scores obtained by comparing the enrollment encoding and test encodings. Finally, the scores were post-processed using the AS-Norm. We used the *DEV* set as the cohort set, and selected the top 10% of sorted cohort scores for calculating the normalization statistics.

6. Results and Discussion

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

Figure 6 shows the performance of the primary submissions per team per track, as well as performance of the baseline systems (see Section 5), in terms of the actual and minimum costs for the Audio-Visual SRE19 *TEST* set. 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 this figure. First, compared to the most recent SRE (i.e., SRE18), there seem to be notable improvements in audio only speaker recognition performance (see Figure 2b in [5]), which are largely attributed to the use of extended and more complex end-to-end neural network architectures (e.g., ResNet) along with soft margin loss functions (e.g., angular softmax) for speaker embedding extraction that can effectively exploit vast amounts of training data made available through data augmentation and/or large-scale datasets such as VoxCeleb[3]. Second, performance trends of the top 4 teams are generally similar, where the actual detection costs for the audio only submissions are larger than those for the visual only submissions, and the audio-visual fusion (i.e., the combination of speaker and face recognition system outputs) results in substantial gains in person recognition performance (i.e., greater than 85% relative in terms of the minimum detection cost for the leading system compared to their

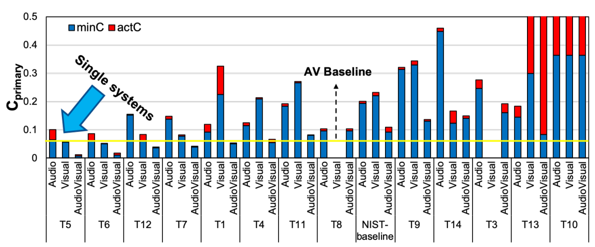

Figure 6: *Performance of the primary submissions for all three tracks (i.e., audio, visual, and audio-visual tracks) of the Audio-Visual SRE19 in terms of the minimum (in blue) and actual (in red) detection costs. The top performing audio and visual systems are both single systems (i.e., no fusion).*

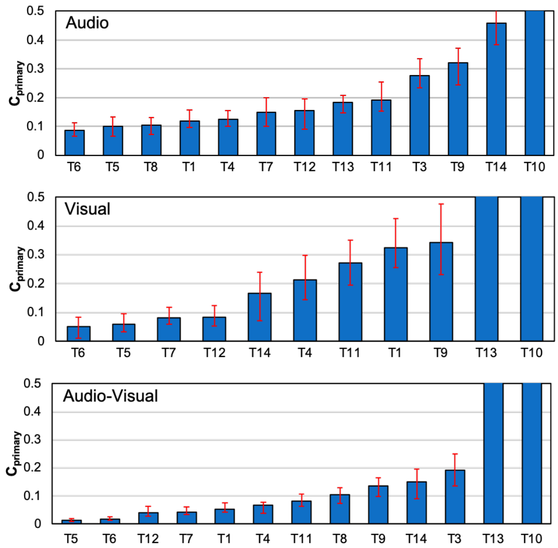

Figure 7: *Performance confidence intervals (95%) of the Audio-Visual SRE19 submissions for the audio (top), visual (middle), and audio-visual (bottom) tracks.*

speaker- or face-recognition system alone). Third, more than half of the submissions outperform the baseline audio-visual system, with the leading system achieving larger than 90% improvement over the baseline. Fourth, in terms of calibration performance, mixed results are observed; for some teams (e.g., the top 2 teams) the calibration errors (i.e., the absolute different between the maximum and minimum costs) for speaker recognition systems are larger than those for face recognition systems, while for some others the opposite is true. Finally, in terms of the minimum detection cost, the two top performing speaker and face recognition systems achieve comparable results, which is a very promising outcome of this evaluation for the speaker recognition community, given the results reported in prior studies (e.g., see [7] where face recognition is shown to outperform speaker recognition by a large margin). It is worth emphasizing here that the top performing speaker and face recognition systems (i.e., team T_4) are both single systems (i.e., no fusion).

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 [22]. To achieve this, we sampled, with repetition, the unique speaker model space along with the associated test segments 1,000 times, which resulted in 1,000 actual detection costs, based on which we calculated the quantiles corresponding to the 95% confidence margin. Figure 7 shows the performance confidence intervals (around the actual detection costs) for each team for the audio (top), visual (middle), and audio-visual (bottom) tracks. It can be seen that, in general, the audio systems exhibit narrower confidence margins than their visual counterparts. This could be partly due to the fact that the majority of the participants, who are from the speaker recognition community, used off-the-shelf face recognition systems along with pretrained models not necessarily optimized for the task at hand in SRE19. Also, notice that several leading systems may perform comparably under different samplings of the trial space. An-

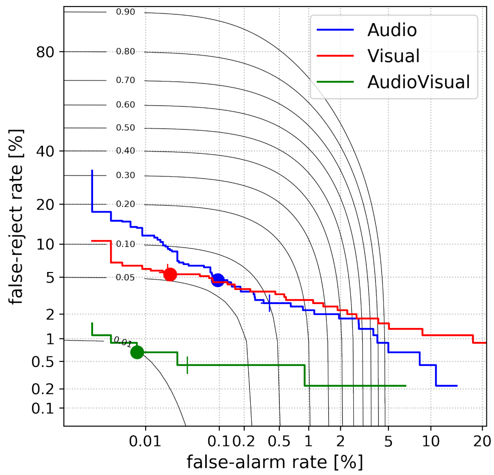

Figure 8: *DET curve performance of the top performing system for the* ***audio****,* ***visual****, and* ***audio-visual*** *tracks. Filled circles and crosses represent minimum and actual costs, respectively.*

other interesting observation that can be made from the figure is that audio-visual fusion seems to boost the decision making confidence of the systems by a significant margin, to the point where the two leading systems statistically significantly outperform the other systems. 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.

Figure 8 shows DET performance curves from the leading system for the audio, visual, and audio-visual tracks. The solid black curves in the figure 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 6 1) the audio-visual fusion provides remarkable improvements in performance across all operating points on the DET curve, which is expected given the complementarity of the two modalities (i.e., audio and visual), and 2) for a wide range of operating points, the speaker and face recognition systems provide comparable performance. Hence, the DET curves in Figure 8 confirm that the operating point dependent results in Figure 6 are consistent across a wider range of operating points, if not all of them.

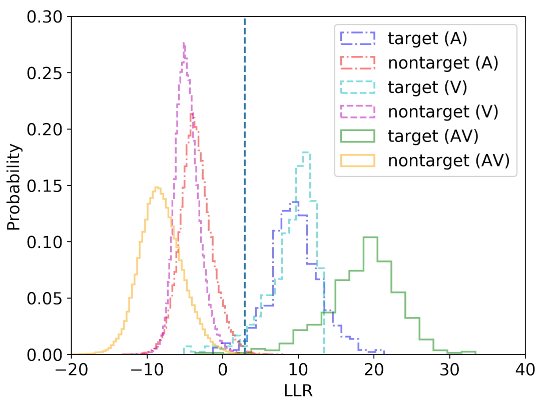

Figure 9: *Normalized target and non-target score distributions from the leading system for the* ***audio (A)****,* ***visual (V)****, and* ***audio-visual (AV)*** *tracks. The vertical dashed line represents the detection threshold.*

Motivated by the relatively low person recognition error rates achieved by the leading audio-visual system, i.e., 0.44% equal error rate (EER), we also conducted an error analysis of low scoring target and high scoring non-target trials, to gain insights regarding the nature of the issues associated with the remaining system errors on the Audio-Visual SRE19 *TEST* set. We found that, out of a total of 452 and 66,896 target and non-target trials, respectively, the system only made 2 false-reject (miss), and 27 false-accept (false-alarm) errors. A manual inspection of the trials (i.e., both enrollment and test videos) associated with these errors suggests that the majority of these trials indeed represent challenging conditions for even humans (non-expert) due to the diversity of the acoustic backgrounds, illuminations, poses, facial expressions, and appearances (e.g., facial hair, glasses, caps/hats).

Figure 9 shows normalized target and non-target score distributions from the leading system for all tracks. The vertical dashed line represents the detection threshold. It can be seen that the score distributions from the audio only and face only systems roughly align, with the target and non-target distributions exhibiting some overlap at the threshold point. However, after the audio-visual fusion, the target and non-target classes are well separated with minimal overlap at the threshold, thereby significantly reducing the detection errors, in particular the false-rejects (misses).

7. Conclusion

Given the observed performance challenges presented by the AfV data in SRE18 and the growing interest of the speaker recognition research community in applying speaker recognition to more realistic multimedia applications, in 2019, NIST organized the first audio-visual SRE to 1) facilitate further exploration of speaker recognition technology in the AfV data domain, and 2) provide participants the opportunity to explore the possibility of fusing face and speaker recognition technologies. In this paper, we presented an overview of the Audio-Visual SRE19 activty including the task, data, the performance metric, the baseline system, as well as results and performance analyses. Compared to SRE18, the evaluation results indicate great progress in audio-only speaker recognition on the challenging AfV domain which is mainly attributed to the use of more complex neural network architectures (e.g., ResNet) along with soft margin losses. In addition, the audio-visual fusion was found to result in remarkable performance gains (greater than 85% relative) over the audio only or face only systems. Finally, state-of-the-art speaker and face recognition technologies were found to provide comparable person recognition performance on the challenging *amateur* online video domain.

8. Disclaimer

These 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 speaker recognition evaluation CTS challenge," 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] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother, "IARPA Janus benchmark-B face dataset," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 592–600.

[7] 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 *Proc. IEEE ICASSP*, April 2018, pp. 3031–3035.

[8] S. O. Sadjadi, "NIST baseline systems for the 2019 audio-visual speaker recognition evaluation," NIST, Tech. Rep., 2019.

[9] 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.

[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 Plan," https://www.nist.gov/document/2019nistmultimediaspeakerrecognitionevaluationplanv3pdf, 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] I. de Paz Centeno, "MTCNN," https://github.com/ipazc/mtcnn, [Online; accessed 2-January-2020].

[17] D. Sandberg, "Face recognition using TensorFlow," https://github.com/davidsandberg/facenet, [Online; accessed 2-January-2020].

[18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE CVPR*, June 2015, pp. 815–823.

[20] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 67–74.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, pp. 4278–4284.

[22] 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.