NIST Baseline Systems for the 2019 Multimedia Speaker Recognition Evaluation

Omid Sadjadi

September 2019

1 Introduction

This document provides a brief description of the NIST baseline speaker and face recognition systems as well as their fusion for the 2019 Multimedia Speaker Recognition Evaluation (SRE19). Specifically, we present speaker and face embedding/encoding based systems that represent the current state of technology in speaker/face recognition (as of SRE18). The speaker recognition system uses the recently developed deep neural network (DNN) speaker embeddings termed x-vectors [1], while the face recognition system uses face encodings extracted using a DNN model with the Inception-ResNet V1 architecture [2]. We begin by giving a summary of the data used to train the various components of the systems, followed by a description of the system components along with their hyper-parameter configurations. Finally, we report experimental results obtained with the baseline systems on the Eval portion of the CORE subset of JANUS Multimedia Dataset (LDC2019E55). Results on the SRE19 Audio-Visual Test set will be released after the evaluation period.

2 JANUS Multimedia Dataset

We use the video data from LDC2019E55 (JANUS Multimedia Dataset) for system development and testing purposes. The JANUS Multimedia Dataset, which has been extracted from the IARPA JANUS Benchmark-B (IJB-B) datatset [3], is described in detail in [4]. It consists of two subsets, namely CORE and FULL, each with a DEV and EVAL split. We only consider the CORE subset in our experiments, because it better reflects the data conditions in the SRE19 Multimedia development and test sets where target speakers are visible. Table 1 summarizes the statistics for the CORE subset of the JANUS Multimedia Dataset. The Dev split is used for system hyperparameter tuning and score normalization, while the Eval split is used to evaluate the performance.

Condition	Split	#enroll videos	#test videos	#target	#nontarget
CORE	Dev	102	319	244	32,294
	Eval	258	914	681	235,131

Table 1: Data statistics for the CORE subset of the JANUS Multimedia Dataset.

3 Speaker Recognition 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 experiments. Figure 1 shows a block diagram of the x-vector baseline system. The x-vector system is built using Kaldi [5] (for x-vector extractor training and extraction) and the NIST SLRE toolkit for back-end scoring.

3.1 Data

The x-vector baseline system is trained using the data recipe available at `https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2`. The x-vector extractor is trained entirely using speech data extracted from combined VoxCeleb 1 and 2 corpora. These datasets contain speech extracted from celebrity interview videos available on YouTube, spanning a wide range of different ethnicities, accents, professions, and ages. For training the x-vector extractor, we use 1,276,888 segments from 7323 speakers selected from VoxCeleb 1 (dev and test), and VoxCeleb 2 (dev), while the remaining segments from the test portion of VoxCeleb 2 are used for debugging purposes. There are on average 174 segments per speaker, with a maximum of 1002 and a minimum of 21 segments per speaker. Note, however, that the segments are not necessarily from unique videos/sessions. In fact, each celebrity video has been split into multiple shorter excerpts to form speaker homogeneous segments, which is determined based on the presence of that celebrity's face in the video. Accordingly, we also create a concatenated copy of VoxCeleb data, where all speech segments belonging to each video are merged to form a single recording. This results in a total of 172,300 recordings from 7323 speakers, which will be used to train the back-end models.

In order to increase the diversity of the acoustic conditions in the training set, a 5-fold augmentation strategy is used that adds four corrupted copies of the original recordings to the training list. The recordings are corrupted by either digitally adding noise (i.e., babble, general noise, music) or convolving with simulated and measured room impulse responses

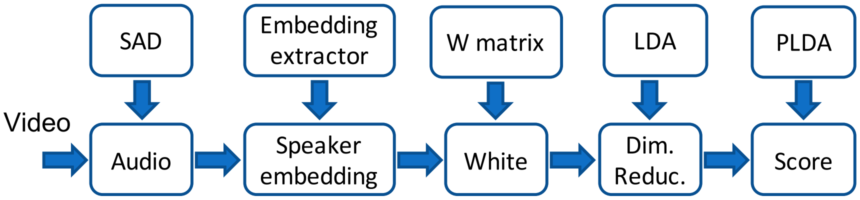

Figure 1: Block diagram of the SRE19 baseline speaker recognition system.

(RIR). The noise and RIR samples are freely available from `http://www.openslr.org` (see [1] for more details). Augmenting the original list with the corrupted copies gives rise to 6,384,440 training segments for the combined VoxCeleb set.

For system development and evaluation, we use the CORE subset of the JANUS Multimedia Dataset described in Section 2. In our experiments, single-channel audio recordings (16-bit PCM) are extracted from the videos at a 16 kHz sample rate using `ffmpeg`.

3.2 Configuration

For speech parameterization, we extract 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 is 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 is trained to discriminate among the 7323 speakers in the training set with 6,384,440 segments. The first 10 hidden layers operate at frame-level, while the last 2 operate at segment-level. There is a 1500-dimensional statistics pooling layer with between the frame-level and segment-level layers that accumulates all frame-level outputs from the 10th layer and computes the mean and standard deviation over all frames for an input segment. After training, embeddings are extracted from the 512-dimensional affine component of the 11th 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 [6].

Prior to dimensionality reduction through LDA (to 250), 512-dimensional x-vectors are centered, whitened, and unit-length normalized. The centering statistics are computed using the in-domain Dev portion of the JANUS Multimedia Dataset. For backend scoring, a Gaussian PLDA model with a full-rank Eignevoice subspace is trained using the x-vectors extracted from all 172,300 concatenated speech segments from the combined VoxCeleb sets, as well as one corrupted version randomly selected from {babble, noise, music, reverb}. Finally, the PLDA verification scores are post-processed using an adaptive score normalization (AS-Norm) scheme proposed in [7]. We use the Dev portion of the JANUS Multimedia Dataset as the cohort set, and select the top 10% of sorted cohort scores for calculating the normalization statistics.

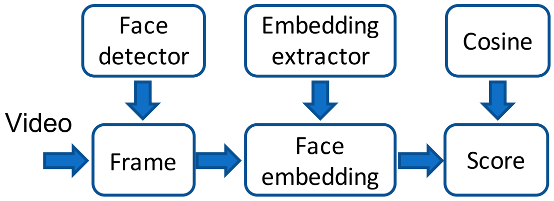

Figure 2: Block diagram of the SRE19 baseline face recognition system.

4 Face Recognition System

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 2 shows a block diagram of the baseline face recognition system which is built using open-source TensorFlow based implementations [8, 9] of 1) a face detector termed MultiTask Cascaded Convolutional Networks (MTCNN) [10], and 2) a face recognizer termed FaceNet [11] (for face encoding extraction). We use the NIST SLRE toolkit for back-end scoring.

4.1 Data

The baseline face recognition system utilizes a pre-trained model available at `https://github.com/davidsandberg/facenet` (model name: 20180402-114759) which has been trained on VGGFace 2 dataset [12] using the Inception ResNet V1 architecture [2]. VGGFace 2 is a large-scale dataset containing 3.31 million images from 9,131 subjects spanning a wide range of different ethnicities, accents, professions and ages.

As in the speaker recognition system, we use the CORE subset of the JANUS Multimedia Dataset (see Section 2) for system development and testing purposes.

4.2 Configuration

We begin processing by extracting one frame per second from the CORE videos using `ffmpeg`. Then, we apply 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 that is closest to the center of the frame (as in [9]). Next, the face images are cropped using the bounding box coordinates, whitened (mean and variance normalized), and resized to 160×160 pixels. Finally, the FaceNet is used to extract face encodings from the cropped, whitened and resized images.

For enrollment, we use the average of face encodings extracted from a video to build a model for each target individual[1], while for test we keep all face encodings. In order to compute a single score for each trial involving an enrollment video and a test video, we compute the maximum of the cosine similarity scores[2] obtained by comparing the enrollment encoding and test encodings. Finally, the scores are post-processed using the AS-Norm. We use the Dev portion of the JANUS Multimedia Dataset as the cohort set, and select the top 10% of sorted cohort scores for calculating the normalization statistics.

[1] One could also use the manually produced bounding box coordinates for the enrollment videos in JANUS, however we found that they were not always reliably marking the target faces. One reason could be that the bounding boxes for JANUS were generated using Amazon Mechanical Turk (AMT) crowd-sourcing service. We note that bounding box coordinates for the SRE19 Development and Test sets have been manually marked by professional annotators.

[2] We note that unlike in embedding based speaker recognition, whitening and LDA/PLDA have not proven beneficial for embedding based face recognition and hence are not used here.

Table 2: NIST baseline system performances on the JANUS Multimedia Dataset.

System	Approach	Training Data	Set	EER [%]	min_C	act_C
Speaker Rec.	x-vector	VoxCeleb	JANUS Eval	6.11	0.241	0.245
Face Rec.	FaceNet	VGGFace2	JANUS Eval	6.31	0.217	0.226
Fusion	average	–	JANUS Eval	2.63	0.125	0.245

5 Results

In this section, we present the experimental results on the Eval portion of the JANUS Multimedia Dataset obtained using the speaker and face recognition baseline systems. Results are reported in terms of the equal error rate (EER) as well as the minimum and actual primary costs (denoted as min_C and act_C, respectively) defined in the SRE19 Multimedia evaluation plan [13]. To compute these performance measures, we use the official SRE19 scoring software available at `https://sre.nist.gov`.

Table 2 summarizes results for the baseline speaker and face recognition systems (first and second rows, respectively), as well as their linear fusion with equal weights (last row). Note that no calibration is applied to the baseline system outputs. It is also worth emphasizing here that one could potentially 1) use publicly available and/or proprietary data in addition to VoxCeleb for speaker recognition, 2) apply speaker diarization and/or face tracking to the audiovisual data, and 3) exploit the bounding box information for enrollment videos for face recognition, to further improve the performance. Nevertheless, these are beyond the scope of the baseline system, and therefore not considered in this report.

6 Disclaimer

The NIST SRE19 baseline systems were developed to support speaker recognition research. Comparison of systems and results against these systems and their results are not to be construed or represented as endorsements of any participant's system(s) or commercial product(s), or as official findings on the part of NIST or the U.S. Government. The reader of this report acknowledges that changes in the data domain and system configurations, or changes in the amount of data used to build a system, can greatly influence system performance.

Because of the above reasons, these systems should not be used for commercial product testing and the results should not be used to make conclusions regarding which commercial products are best for a particular application.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] D. Povey *et al.*, "Kaldi Speech Recognition Toolkit," https://github.com/kaldi-asr/kaldi, [Online; accessed 26-July-2018].

[6] 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.

[7] 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.

[8] I. de Paz Centeno, "MTCNN," https://github.com/ipazc/mtcnn, [Online; accessed 26-August-2019].

[9] D. Sandberg, "Face recognition using TensorFlow," https://github.com/davidsandberg/facenet, [Online; accessed 26-August-2019].

[10] 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.

[11] 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.

[12] 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.

[13] NIST, "NIST 2019 Speaker Recognition Evaluation Plan," https://www.nist.gov/document/sre19-multimedia-evaluation-plan, 2019, [Online; accessed 26-August-2019].