The Speaker and Language Recognition Workshop (Odyssey 2020)
1-5 November 2020, Tokyo, Japan

Analysis of ABC Submission to NIST SRE 2019 CMN and VAST Challenge

Jahangir Alam[5], *Gilles Boulianne*[5], *Lukáš Burget*[1], *Mohamed Dahmane*[5], *Mireia Diez*[1], *Ondřej Glembek*[1], *Marc Lalonde*[5], *Alicia Lozano-Diez*[1,6], *Pavel Matějka*[1,2], *Petr Mizera*[4], *Ladislav Mošner*[1], *Cédric Noiseux*[5], *Joao Monteiro*[5], *Ondřej Novotný*[1], *Oldřich Plchot*[1], *Johan Rohdin*[1], *Anna Silnova*[1], *Josef Slavíček*[2], *Themos Stafylakis*[4], *Pierre-Luc St-Charles*[5], *Shuai Wang*[1,3], *Hossein Zeinali*[1],

[1]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czechia
{matejkap,iplchot,...}@fit.vutbr.cz
[2]Phonexia, Czechia
slavicek@phonexia.com
[3]Speechlab, Shanghai Jiao Tong University, China
feixiang121976@sjtu.edu.cn
[4]Omilia - Conversational Intelligence, Athens, Greece
tstafylakis@omilia.com
[5]CRIM, Montreal (Quebec), Canada
jahangir.alam@crim.ca
[6]Audias-UAM, Universidad Autonoma de Madrid, Madrid, Spain
alicia.lozano@uam.es

Abstract

We present a condensed description and analysis of the joint submission of ABC team for NIST SRE 2019, by BUT, CRIM, Phonexia, Omilia and UAM. We concentrate on challenges that arose during development and we analyze the results obtained on the evaluation data and on our development sets. The conversational telephone speech (CMN2) condition is challenging for current state-of-the-art systems, mainly due to the language mismatch between training and test data. We show that a combination of adversarial domain adaptation, backend adaptation and score normalization can mitigate this mismatch. On the VAST condition, we demonstrate the importance of deploying diarization when dealing with multi-speaker utterances and the drastic improvements that can be obtained by combining audio and visual modalities.

1. Introduction

This paper presents the ABC team submission to the NIST 2019 Speaker Recognition Evaluation (SRE19), resulted from a collaborative effort between Brno University of Technology, Phonexia, Omilia, CRIM and Universidad Autonoma de Madrid. SRE19 consists of two separate activities: The first activity is a leaderboard-style challenge, where participants could daily evaluate three systems on unexposed portions of the Call My Net 2 (CMN) corpus comprising 8kHz conversational telephone speech in Tunisian Arabic. Therefore, similarly to NIST SRE18, the main challenge is to build systems using plentiful (mostly English) out-of-domain data and to adapt them to the target domain using the limited available Tunisian Arabic telephone data. The second activity is a regular NIST evaluation, where submitted systems are evaluated on the unexposed portions of the Video Annotation for Speech Technology (VAST) corpus. Besides the 16kHz speech data, this corpus contains also the corresponding video recordings of the speakers. Therefore, the new challenge in this evaluation is to use also the visual information to improve the speaker recognition performance.

This paper describes our very competitive submissions for both CMN and VAST conditions. For VAST, we also compare audio-only and audio-visual systems. Our primary submission to each condition is a score-level fusion of several individual systems. All individual systems are based on x-vectors – embeddings extracted from audio (or video) recording using Deep Neural Networks (DNN). The individual systems mainly differ in the particular DNN architecture (variants of TDNN and ResNet) and the back-end model/strategy for comparing the embeddings. More technical details on the systems can be also found in [1] and [2]. For both CMN and VAST, we report significant improvements obtained by fusing 4 similarly performing audio systems. For VAST, impressive improvements (up to 70% relative) are obtained when adding a video based system to the fusion, even though the proposed video-only systems yield inferior performance compared to that of the audio-only systems.

Further analysis is provided in this paper in order to provide more insights into our system design choices and to show the effectiveness of individual steps or components in our systems. For CMN system, we analyze the effectiveness of different x-vector and score normalization techniques, as well as supervised and unsupervised adaptation strategies of the back-end, which are necessary to adapt the systems to the target domain with scarce in-domain labeled data. For VAST condition, the recordings may contain speech from multiple speakers and the task is to detect if one of them is the target speaker. We approach the problem by employing a variety of diarization strategies, loss functions for training x-vector extractors, and back-ends for evaluating likelihood ratios. We observe that different choices have to be made for VAST and CMN in order to get

10.21437/Odyssey.2020-41

the optimal performance, reflecting the significant differences between the two tasks.

2. CMN - Call My Net 2

2.1. Data & Experimental setup

For training the networks, we used the following datasets:

- NIST SRE 4, 5, 6, 8, 10
- Fisher Arabic
- All switchboard data
- Voxceleb 1 and 2

We performed the following data augmentations which are the same as in Kaldi apart for the compression recipe[1] :

- Reverberated
- Augment with Musan noise, music and speech
- Compression using ogg and mp3 codecs for Voxceleb

After creating a list of utterances for augmentation, a subset of 500K utterances from this list was selected and added to the training data. Afterwards, utterances with less than 500 frames and also speakers with less than 5 training utterances were removed. Finally, the training set consisted of 17054 speakers.

All of the backends utilized the following data:

- Training set - data from Mixer collection (NIST SRE 2004-2010) from which we kept only telephone recordings, approximately 66k utterances.
- Adaptation set - SRE18 development set and 60% of the data from SRE18 evaluation dataset. The resulting set consisted of 8k utterances coming from 137 speakers.
- S-norm data - part of the adaptation set (5 utterances per speaker) and SRE18 unlabeled data.
- Unseen development set for performance evaluation - remaining 40% of the data from SRE18 evaluation dataset.

2.2. Preprocessing

We used FBANK or MFCC features and energy-based VAD from Kaldi SRE16 recipe without any modification. The features are therefore 23-dimensional MFCC or 40 dimensional FBANK, which are extracted from 25 ms windows with 15 ms overlap. The bandwidth is limited between 20 and 3700 Hz for 8kHz sampling frequency or 20-7600Hz for 16kHz sampling frequency. We also apply short-term mean normalization with a sliding window of 3 seconds.

2.3. General pipeline

Every system is a combination of a different DNN x-vector extractor and a backend. We use 4 different DNN topologies and 2 different backends. All individual systems are described below in different subsections, in which the general scheme is first outlined followed by specific system details.

For all of the systems we generate x-vectors for every utterance and do mean normalization. Evaluation x-vectors are centered using the mean computed on the adaptation set, while the backend training x-vectors are centered using their own mean. Then, we apply feature-distribution adaption (FDA) transformation [3] on the training x-vectors. The goal of the transformation is to modify the out-of-domain training x-vectors so that their variance is not lower than the variance of the in-domain adaptation x-vectors in any direction. Then, Gaussian or heavy-tailed PLDA model is trained using the transformed training x-vectors. Additionally, we train "adaptation" model on the untransformed adaptation x-vectors. The final adapted model is derived from the two PLDA models so that the modeled across-speaker covariance matrix is an average of the covariance matrices from the constituent models. Similarly, within-speaker covariance matrices are also interpolated. Finally, we used adaptive symmetric score normalization (adapt S-norm) which computes an average of normalized scores from Z-norm and T-norm [4, 5]. In its adaptive version [5, 6, 7], only a part of the cohort is selected to compute mean and variance for normalization. The 800-top scoring files are selected from the S-norm set.

2.4. Individual Systems

2.4.1. ResNet_GPLDA

ResNet-based embeddings are extracted from a standard 50-layer ResNet (ResNet50) [8]. This network uses 2-dimensional features as input and processes them using 2-dimensional CNN layers. Inspired by the original DNN architecture for x-vector extraction, both mean and standard deviation are used as statistics. The ResNet is trained using SGD optimizer for 6 epochs.

When training the backend for this system, we add SRE16 evaluation data (10k utterances from 201 speakers) to the training set. Besides the centering and FDA preprocessing mentioned above, we apply LDA that reduces the dimensionality of x-vectors from 256 to 250, and length normalization. Then, we train a GPLDA model for which the size of both speaker and channel subspaces is set to 150. The adaptation model is also a GPLDA, but with smaller speaker and channel subspaces are smaller, namely 50.

2.4.2. FTDNN_HTPLDA

For this system we use the factorized TDNN (F-TDNN) architecture proposed in [9]. We train it with the Kaldi toolkit [10] with similar settings to those described in `sre16/v2` recipe. We used the datasets described in Sections 2.1, and features and VAD described in Section 2.2. Moreover, we trained the model for six epochs instead of three and used a modified example generation. Finally, we used 300 frames in all training segments and instead of random segment selection we used almost all available speech from all training speakers.

We trained an HTPLDA model -with the size of the speaker subspace set to 200- on the length-normalized embeddings from the training set, enlarged by adding 4 different augmentations for each training utterance. Then, a smaller HTPLDA model (speaker space of size 100) is trained on the adaptation data. The degrees of freedom parameter was set to 2 for both models.

2.4.3. Res-E-TDNN_DENOI_HTPLDA

For this system, we use a modified version of the Extended-TDNN (E-TDNN) architecture [9] where TDNN layers are interleaved with fully connected linear layers. In all the layers before the pooling layer 768 outputs (channels) are used instead of 512. Also, based on our experience for VoxCeleb challenge [11], we add a few residual connections to these layers. The input of each linear layer before the pooling layer is a summation of the output of all previous TDNN layers. So, the first linear layer receives the input from one TDNN layer, the second one receives it from summation of two TDNN layers and

[1] `https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2`

so on. We train it with the Kaldi toolkit and setup described in Section 2.4.2 with the exception that we trained the model for three epochs.

During extraction, each audio file is pre-processed with denoising based on neural-network autoencoder. The autoencoder, training data, and augmentation are described in [12] and [13]. On top of the x-vectors we train a HTPLDA model with the size of the speaker subspace set to 200 on the 66k embeddings from the training set. SRE16 evaluation data are added to the adaptation set. Adaptation HTPLDA of the same size as the main one is trained on the resulting data. The degrees of freedom parameter was again set to 2 for both models.

2.4.4. Res-E-TDNN_GAN_HTPLDA

In this approach we begin with a Res-E-TDNN network (see Section 2.4.3) trained on SRE English telephone data. We then apply adversarial domain adaptation. Similarly to our approach described in [14], we attach a domain discriminator (a feed-forward neural network with 3 hidden layers and Leaky ReLU) which aims at discriminating between source and target domains (English and Arabic, respectively). The x-vector extractor tries to fool the discriminator by maximizing the binary cross entropy loss of the discriminator. The adversarial loss encourages the extractor to encode utterances into x-vectors that are hard to distinguish in terms of domain (i.e. language). Moreover, as the divergence between the two marginal distributions becomes smaller, a PLDA model trained solely on source x-vectors can perform fairly well on the target domain without any adaptation. Note that different from [14], we use a standard GAN (as opposed to Wasserstein GAN) and we do not augment input features or hidden representations with domain labels. During adaptation, the number of source training speakers is 4254 selected from SRE 2004-2010 (English telephone only), while we augment the target domain data with Fisher Arabic, resulting in 2251 Arabic speakers. Two different softmax layers are used (one for each domain) and their associated cross entropy losses are added, yielding the overall speaker classification loss.

The backend training of this model practically replicates what we did for the F-TDNN (see backend description in Section 2.4.2). The only difference is that we do not use length normalization in this case.

2.5. Calibration and Fusion

The final submission strategy is one common fusion trained on the labeled development set created by holding out 40% of the NIST SRE2018 CMN2 evaluation data. Each system provides log-likelihood ratio scores that undergo score normalization. These scores are first pre-calibrated and then passed to the fusion model. The output of the fusion is then again calibrated.

Both calibration and fusion are trained with logistic regression optimizing the cross-entropy between the hypothesized and true labels on a corresponding development set. Our objective is to improve the error rate on the development set. We observed very similar error rates on our development set and NIST's progress set during submitting our intermediate systems to the NIST leaderboard.

2.6. Analysis

Table 1 shows the results of our fusion of 4 systems submitted to the NIST SRE 2019 CMN evaluation together with the results of the individual systems. The best single system is F-TDNN with HTPLDA, but all 4 systems are very close in performance with close to perfect calibration. We observe 20% relative gain from the fusion on the SRE19 evaluation data.

Based on the lessons learned from Interspeech 2019 Voxceleb challenge, where conventional PLDA backends did not provide optimal performance, we ran an analysis of different backends for F-TDNN and ResNet based x-vectors. Table 2 summarizes these results showing that the HTPLDA is the best choice for both architectures in this domain. The results in Table 4, which will be analyzed below in more detail, shows instead that cosine distance scoring together with domain adaptation techniques provides the best results on the VAST domain.

The key element of last NIST evaluations is the adaptation of the system to the new domain. Table 3 shows the analysis of different adaptation and normalization strategies. Overall, the best attained by applying mean subtraction with length normalization, FDA and supervised PLDA adaptation followed by adaptive snorm.

3. VAST - Audio Systems

3.1. Data & Experimental setup

We train the backend on approximately 145k utterances from VoxCeleb 2 (original speech segments corresponding to the same session are concatenated together). For the adaptation, we used 37 utterances of SRE18 VAST development data.

3.2. Preprocessing

We used FBANK or PLP features and energy-based VAD from Kaldi SRE16 recipe without any modification. The features are 40 dimensional FBANK or 23 dimensional PLP, which are extracted from 25 ms windows with 15 ms overlap. The bandwidth is limited to 20-7600Hz for 16kHz sampling frequency. We apply also short-term mean normalization with a sliding window of 3 seconds.

3.3. General info

Different architectures of DNN extractors were trained and used to extract x-vectors for each utterance.

In the backend part, all training, adaptation and evaluation x-vectors are centered using the training x-vectors mean. Then, we apply feature-distribution adaption (FDA) transformation [3] on the training x-vectors. After FDA, we apply length normalization and LDA dimensionality reduction followed by another length normalization.

After the preprocessing described above we either train Gaussian PLDA model or use simple cosine scoring to compare the x-vectors. In all cases, we used adapt S-norm. As a cohort, we used a subset of the PLDA training data.

In fact, the scoring procedure for VAST condition is more intricate as test utterances can contain speech from multiple speakers. The task is to detect if one of these speakers is the same as in the enrollment utterance. For this purpose, all test files are processed by a diarization system based on Agglomerative Hierarchical Clustering (AHC) of x-vectors, which are extracted from input recordings every 0.25 seconds (see Section 3.6.2 and [15] for more details). The diarization systems are run to produce 4 different outputs with 1,2,3 and 4 speakers. Then, an x-vector is extracted for each speaker suggested by the 4 diarization outputs resulting in 10 x-vectors per test file. All 10 test x-vectors were compared with the enrollment x-vector using the backend described in the next sections and maximum

Table 1: Results of single systems and submitted fusions for NIST SRE2019 CMN.

#	System	Backend	SRE18 CMN eval			SRE19 CMN eval		
			minDCF	actDCF	EER (%)	minDCF	actDCF	EER (%)
1	ResNet	GPLDA	0.281	0.285	3.77	0.341	0.352	3.29
2	F-TDNN	HTPLDA	0.255	0.257	3.37	0.311	0.321	3.01
3	Res-E-TDNN_DENOI	HTPLDA	0.312	0.314	4.30	0.355	0.359	3.77
4	Res-E-TDNN_GAN	HTPLDA	0.273	0.275	3.76	0.312	0.317	3.18
	Fusion 1+2+3+4		0.216	0.218	3.00	0.268	0.277	2.43

Table 2: Results of the different backend strategies for different types of x-vectors. Results are reported on NIST SRE 2019 CMN2.

	ResNet		F-TDNN	
	minDCF	EER [%]	minDCF	EER [%]
GPLDA	0.344	3.37	0.328	3.51
HTPLDA	0.346	3.37	0.305	2.93
cos. dist.	0.468	5.54	0.425	5.21

Table 3: Results of the x-vector normalization and backend adaptation techniques on the telephone data CMN2

	SRE18 eval EER (%)	SRE19 eval EER (%)
no mean subtraction	7.36	7.26
mean subtraction	5.89	5.80
+ FDA	4.84	4.73
+ supervised PLDA adaptation	3.85	3.52
+ snorm	5.15	4.90
+ FDA + snorm	3.97	3.86
+ PLDA adapt + snorm	3.53	3.13
+ FDA + PLDA adapt + snorm	3.29	2.91

score was selected as the representative score for the given trial.

3.4. Individual Systems

3.4.1. ResNet_GPLDA and ResNet_COS

The topology and training scheme is described in Section 2.4.1 and the training data and preprocessing in Section 3.1 and 3.2.

For this system we used additive angular margin loss (denoted as 'AAM loss') which was proposed for face recognition [16] and introduced to speaker verification in [17]. Instead of training the AAM loss from scratch, we only fine-tune a well-trained NN using conventional softmax loss. The parameters controlling the AAM loss [17] are set as follows in all our experiments: s is set to 30 and m is set to 0.2..

We used the development part of VOXCELEB-2 dataset [18] for DNN training. This set has 5994 speakers spread over 145 thousand sessions (distributed in approx. 1.2 million speech segments). We used original speech segments together with their augmentations. The augmentation process was based on the Kaldi recipe (see Section 2.1) and it resulted in additional 5 million segments.

We used two backend strategies with this system:

- GPLDA - we reduce the x-vector dimensionality to 200 using LDA and a Gaussian PLDA model is trained with the size of the speaker and channel subspace set to 200 (i.e full-rank). We use 100 top scoring files from the cohort (5k x-vectors from the training set) for S-norm.
- COS - we reduce the x-vector dimensionality to 100 using LDA and we perform cosine similarity scoring. We use 100 top scoring files from the cohort (5k x-vectors from the training set) for S-norm.

3.4.2. TDNN_GPLDA

This system is the well-known TDNN based x-vector topology trained with Kaldi toolkit [10] using SRE16 recipe with the following modifications:

- Training networks with 6 epochs (instead of 3). We did not see any considerable difference with more epochs.
- We use 200 frames for all training segments (instead of random durations between 200 and 400 frames).
- We sample the training segments in such a way that all regions of a recording are equally used (instead selecting the segments completely at random).
- the amount of training data is increased 5 times by including the 4 copies of the data with different augmentations, same as for ResNet in Section 3.4.1.

This system is trained on the VoxCeleb 1 and 2 development sets (1152 + 5994 speakers respectively), 2338 speakers from LibriSpeech dataset [19] and 1735 speakers from DeepMine dataset [20]. For all training data, we first discard utterances with less than 400 frames (measured after applying the VAD). After that, all speakers with less than 8 utterances (including augmentation data) are removed.

The LDA dimensionality is 150. Gaussian PLDA model is trained with the size of the speaker and channel subspace set to 150 (i.e full-rank). We used 150 top scoring files from the cohort (25k x-vectors from the training set) for snorm.

3.4.3. 8kHz_PLP_TDNN_GPLDA

This system is trained on 8kHz CMN telephone data on 23 dimensional PLP features. Embeddings are adapted to VAST by using only unsupervised adaptation employing CORAL [21]. LDA dimensionality is set to 200. We do not use S-norm with this system.

3.5. Calibration and Fusion

The final submission strategy is one common fusion trained on the labeled development set. Each system provides log-likelihood ratio scores.These scores are first pre-calibrated and then passed into the fusion. The output of the fusion is then re-calibrated.

Both calibration and fusion are trained with logistic regression optimizing the cross-entropy between the hypothesized and true labels on a corresponding development set. Our objective is to improve the error rates on the NIST SRE 2019 VAST development set.

3.5.1. Data

We use the labeled NIST SRE2018 VAST evaluation set to train the calibration and fusion as we are using NIST SRE 2018 VAST development data in some cases to do the system adaptation and we want to avoid the overlap which exists between NIST 2018 and NIST 2019 VAST development sets. We have not split the datasets in any way and we take the risk of having an overlap which exists between NIST 2018 VAST evaluation set and NIST 2019 VAST development set.

3.5.2. Results

The first part of table 6 shows the 4 audio-based individual speaker verification systems which are used in the fusion. The first 3 are the 16kHz systems and are very close in performance. The 4th system is a 8kHz PLP based system trained on telephone data. Although it performs poorly on its own, it still provided gains in the fusion.

3.6. Analysis

3.6.1. PLDA vs Cosine distance scoring for ResNet

The purpose of this experiment is to compare backends for the ResNet system with AAM finetuning. The results are reported in Table 4 where the first part analyzes the results without adaptation and the second part analyzes results with FDA transformation as preprocessing. The general trend for the ResNet architecture in this domain is that we have better results with cosine distance as backend. Adaptive S-norm boosts the performance of both - PLDA and cosine distance.

Table 4: Results of the different backend strategies for ResNet with AAM.

	SRE18 VAST eval		SRE19 AV eval	
	minDCF	EER [%]	minDCF	EER [%]
No adaptation				
PLDA	0.464	11.44	0.144	2.89
PLDA+snorm	0.455	10.68	0.124	2.74
COS	0.459	10.60	0.152	3.01
COS+snorm	0.465	9.46	0.133	2.61
FDA transform of PLDA training and snorm data				
PLDA	0.423	12.68	0.138	2.62
PLDA+snorm	0.373	10.08	0.126	2.54
COS	0.459	10.60	0.152	3.01
COS+snorm	0.351	7.77	0.129	2.41

3.6.2. Impact of diarization

Given that in this condition test utterances can contain speech from multiple speakers, we consider the different strategies for scoring. Results comparing these strategies are shown in Table 5. For the first line no diarization is used and a single x-vector is extracted from the whole test utterance and scored against the enrollment x-vector. BHMM shows the results when using the DIHARD II winning system [15] for diarization. This system is based on Bayesian HMM clustering of x-vectors, which can automatically determine the number of speakers. An x-vector is extracted from the speech of each identified speaker and scored against the enrollment x-vector. Out of these, the maximum score is selected. AHC 1-4 spk is the approach used in our submission, as described in section 3.3, where AHC is used to perform clustering into 1,2,3 and 4 speakers and the corresponding 10 x-vectors are extracted and scored against the enrollment x-vector. The remaining lines correspond to the same strategy with a different number of maximum speakers.

Table 5: Effect of diarization in VAST.

	SRE18 VAST eval		SRE19 AV eval	
	minDCF	EER [%]	minDCF	EER [%]
no diarization	0.364	9.38	0.143	4.41
BHMM	0.332	9.66	0.127	3.06
AHC 1-2 spks	0.323	8.91	0.110	2.54
AHC 1-3 spks	0.321	8.37	0.102	2.47
AHC 1-4 spks	**0.319**	8.43	**0.100**	2.39
AHC 1-7 spks	0.332	**7.85**	0.101	**2.30**

We can see that the strategies making use of diarization provides significant better results than the naive approach with a single x-vector per test utterance. Inspecting the diarization results we observed that although the BHMM provides very good diarization output, the simpler AHC strategy provides better resutls in terms of speaker recognition performance.

4. VAST – Audio Visual

4.1. Individual Visual Systems

4.1.1. CRIM_V_S1PL

The video baseline system is inspired by the definition of the Face Recognition System in the SRE19 Multimedia Baseline description document (Section 4). First, embeddings are extracted for the enrollment videos using the facial bounding boxes and frame indices provided in the dataset. The corresponding image regions are cropped, normalized, and passed to a Squeeze-Excitation variation of a ResNet-50 [22] pre-trained on VGGFace2 [23] to produce a set of facial embeddings. For each enrollment video, the embeddings are averaged to create a single feature vector that corresponds to a subject. Next, we use the Single-shot Scale-invariant Face Detector (S3FD) of [24] to detect roughly one face per second in the test videos. With those detections' bounding boxes, we extract new facial embeddings using the same approach as before. Finally, in each trial, we compute the cosine similarity between the (averaged) subject embedding and the automatically extracted embeddings. The output score for each video is the maximum similarity found between embedding pairs in that video. No score normalization is performed.

4.1.2. CRIM_V_S2MD

The multitask CNN (MTCNN) was used for the detection of faces. Here, we used only the detected bounding boxes (BB) around the faces. The SENet50 [22] architecture trained on VGG Face2 [23] was then used to produce a set of facial embeddings. Since the videos contain more than one persons, we used Kalman filter to track the extracted BB from frame to frame. The tracking leads to several groups of facial attributes corresponding to different tracklets. The track of one person could be possibly represented by a number of tracklets (groups) if the tracking is broken for any reason (e.g. occlusion, leaving the scene, etc.). The Chinese Whispers algorithm which do not need any prior information about the number of clusters was then used for clustering the embeddings. In the test phase, for each trial we compute the cosine similarity between the (averaged) subject embedding and the automatically extracted embeddings. The output score for each video is the maximum sim-

Table 6: Results of single systems and submnitted fusions, * denotes single best system.

#	System	SRE19 AV dev			SRE19 AV eval		
		minDCF	actDCF	EER (%)	minDCF	actDCF	EER (%)
	Audio systems						
1	ResNet_GPLDA	0.23	0.25	4.14	0.11	0.12	2.01
2	ResNet_COS	0.21	0.23	5.08	0.11	0.14	2.13
3	* TDNN_GPLDA	0.19	0.20	4.25	0.097	0.11	2.19
4	8kHz_PLP_TDNN_GPLDA	0.49	0.49	11.2	0.31	0.32	6.35
	FUSION (PRIMARY) 1+2+3+4	0.15	0.16	3.91	0.092	0.119	1.64
	Visual systems						
5	CRIM_V_S1PL	0.77	0.79	9.17	0.20	0.42	3.91
6	* CRIM_V_S2MD	0.46	0.49	7.24	0.36	0.46	8.62
	FUSION (PRIMARY) 5+6	0.44	0.45	6.66	0.225	0.325	4.29
	Audio-visual systems						
	* FUSION 3+6	0.10	0.10	1.71	0.054	0.060	0.99
	FUSION (PRIMARY) 2+3+4+6	0.07	0.08	1.54	0.050	0.052	0.77
	FUSION (posteval) 2+3+4+5	0.09	0.11	1.82	0.035	0.043	0.50

ilarity found between embedding pairs in that video. No score normalization is performed.

4.1.3. Calibration and Fusion

For audio-visual challenge we used audio systems described in Section 3 and 2 visual systems described in this section. The calibration and fusion strategy was same as was described for audio only systems in Section 3.5 with 2 more visual systems.

The lower part of the Table 6 shows the fusion of single best audio and video system and also fusion of 3 audio systems with one video system. There is a huge improvement (more than 50% relative) in fusion observed by many sites when fusing audio and video modalities. Our video systems were not robust enough and showed different behavior on our development and evaluation set. Therefore we have run also new fusion marked as "posteval" which shows even another 20% relative gain from fusing audio and video systems.

5. Conclusion

In 2019 edition of the NIST speaker recognition evaluation, we have again observed that fusion of several systems yields significant improvements even if all the subsystems are quite similar. In CMN2 condition, we demonstrated that different x-vectors/score normalizations, together with unsupervised/supervised and generative adversarial network-based domain adaptation strategies are important to obtain good results in the target domain with very limited resources. On the VAST condition, we showed that when dealing with multi-speaker test utterances, the use of diarization is necessary to obtain good results. The new challenge in this evaluation was the audio-visual speaker verification on VAST data, which revealed that impressive improvements can be attained by fusing information from the two modalities. We have also observed that different training and scoring strategies have to be used to obtain optimal results for the CMN and VAST conditions. This calls for further analysis which would bring us closer to building truly domain independent speaker verification systems.

6. Acknowledgment

BUT researchers were supported by Czech Ministry of Interior projects Nos. VI20152020025 "DRAPAK" and VI20192022169 "AI v TiV", Czech National Science Foundation (GACR) project "NEUREM3" No. 19-26934X, European Union's Marie Sklodowska-Curie grant agreement No. 843627, European Union's Horizon 2020 grant agreement no. 833635 "ROXANNE" and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science" - LQ1602. CRIM researchers wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant RGPIN-2019-05381 and Ministry of Economy and Innovation (MEI) of the Government of Quebec for the continued support.

7. References

[1] Jahangir Alam, Gilles Boulianne, Ondřej Glembek, Alicia Díez Lozano, Pavel Matějka, Petr Mizera, Joao Monteiro, Ladislav Mošner, Ondřej Novotný, Oldřich Plchot, A. Johan Rohdin, Anna Silnova, Josef Slavíček, Themos Stafylakis, Shuai Wang, and Hossein Zeinali, "Abc nist sre 2019 cts system description," in *Proceedings of NIST*. 2019, pp. 1–6, National Institute of Standards and Technology.

[2] Jahangir Alam, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Alicia Díez Lozano, Pavel Matějka, Petr Mizera, Ladislav Mošner, Ondřej Novotný, Oldřich Plchot, A. Johan Rohdin, Anna Silnova, Josef Slavíček, Themos Stafylakis, Shuai Wang, Hossein Zeinali, Mohamed Dahmane, Pierre-Luc St-Charles, Marc Lalonde, Cédric Noiseux, and Joao Monteiro, "Abc system description for nist multimedia speaker recognition evaluation 2019," in *Proceedings of NIST 2019 SRE Workshop*. 2019, pp. 1–7, National Institute of Standards and Technology.

[3] Pierre-Michel Bousquet and Mickael Rouvier, "On Robustness of Unsupervised Domain Adaptation for Speaker Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2958–2962.

[4] P. Kenny, "Bayesian speaker verification with heavy–

tailed priors," keynote presentation, Proc. of Odyssey 2010, June 2010.

[5] Pavel Matějka, Ondřej Novotný, Oldřich Plchot, Lukáš Burget, Mireia Sánchez Diez, and Jan Černocký, "Analysis of score normalization in multilingual speaker recognition," in *Proceedings of Interspeech 2017*. 2017, pp. 1567–1571, International Speech Communication Association.

[6] D. E. Sturim and Douglas A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *ICASSP*, 2005, pp. 741–744.

[7] Yaniv Zigel and Moshe Wasserblat, "How to deal with multiple-targets in speaker identification systems?," in *Proceedings of the Speaker and Language Recognition Workshop (IEEE-Odyssey 2006)*, San Juan, Puerto Rico, June 2006.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Fred Richardson, Suwon Shon, François Grondin, Réda Dehak, Leibny Paola García-Perera, Daniel Povey, Pedro A. Torres-Carrasquillo, Sanjeev Khudanpur, and Najim Dehak, "State-of-the-Art Speaker Recognition for Telephone and Video Speech: The JHU-MIT Submission for NIST SRE18," in *Proc. Interspeech 2019*, 2019, pp. 1488–1492.

[10] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[11] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, "BUT System Description to VoxCeleb Speaker Recognition Challenge 2019," in *VoxCeleb 2019 Workshop*, 2019.

[12] Ondřej Novotnỳ, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, et al., "Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition," *Computer Speech & Language*, 2019.

[13] Ondřej Novotný, Pavel Matějka, Oldřich Plchot, and Ondřej Glembek, "On the use of DNN Autoencoder for Robust Speaker Recognition," Tech. Rep., 2018.

[14] Johan Rohdin, Themos Stafylakis, Anna Silnova, Hossein Zeinali, Lukáš Burget, and Oldřich Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6006–6010.

[15] Federico Landini, Shuai Wang, Mireia Diez, Lukáš Burget, Pavel Matějka, Kateřina Žmolíková, Ladislav Mošner, Oldřich Plchot, Ondřej Novotný, Hossein Zeinali, and Johan Rohdin, "BUT System Description for DIHARD Speech Diarization Challenge 2019," *arXiv preprint arXiv:1910.08847*, 2019.

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[17] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," *arXiv preprint arXiv:1906.07317*, 2019.

[18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 1086–1090.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[20] Hossein Zeinali, Hossein Sameti, and Themos Stafylakis, "Deepmine speech processing database: Text-dependent and independent speaker verification and speech recognition in persian and english," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 386–392.

[21] Jahangir Alam, Gautam Bhattacharya, and Patrick Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 176–180.

[22] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," 2018.

[23] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, 2018, pp. 67–74.

[24] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li, "S^3fd: Single shot scale-invariant face detector," *CoRR*, vol. abs/1708.05237, 2017.