The Speaker and Language Recognition Workshop (Odyssey 2022)
28 June - 1 July 2022, Beijing, China

DEVELOPMENT OF ABC SYSTEMS FOR THE 2021 EDITION OF NIST SPEAKER RECOGNITION EVALUATION

Jahangir Alam[3], *Radek Beneš*[4], *Marián Beszédeš*[4], *Lukáš Burget*[1], *Mohamed Dahmane*[3], *Abderrahim Fathan*[3], *Hamed Ghodrati*[3], *Ondřej Glembek*[1], *Woo Hyun Kang*[3], *Pavel Matějka*[1], *Ladislav Mošner*[1], *Oldřich Plchot*[1], *Johan Rohdin*[1], *Anna Silnova*[1], *Themos Stafylakis*[2]

[1]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czechia
[2]Omilia - Conversational Intelligence, Athens, Greece
[3]Computer Research Institute of Montreal (CRIM), Montreal (Quebec) Canada
[4]Innovatrics, Bratislava, Slovakia

ABSTRACT

In this contribution, we provide a description of the ABC team's collaborative efforts toward the development of speaker verification systems for the NIST Speaker Recognition Evaluation 2021 (NIST-SRE2021). Cross-lingual and cross-dataset trials are the two main challenges introduced in the NIST-SRE2021. Submissions of ABC team are the result of active collaboration of researchers from BUT, CRIM, Omilia and Innovatrics. We took part in all three close condition tracks for audio-only, audio-visual and visual-only verification tasks. Our audio-only systems follow deep speaker embeddings (e.g., x-vectors) with a subsequent PLDA scoring paradigm. As embeddings extractor, we select some variants of residual neural network (ResNet), factored time delay neural network (FTDNN) and Hybrid Neural Network (HNN) architectures. The HNN embeddings extractor employs CNN, LSTM and TDNN networks and incorporates a multi-level global-local statistics pooling method in order to aggregate the speaker information within short time-span and utterance-level context. Our visual-only systems are based on pre-trained embeddings extractors employing some variants of ResNet and the scoring is based on cosine distance. When developing an audio-visual system, we simply fuse the outputs of independent audio and visual systems. Our final submitted systems are obtained by performing score level fusion of subsystems followed by score calibration.

1. INTRODUCTION

Speaker verification is the task of verifying the claimed speaker identity based on the given speech recordings and has become a key technology for personal authentication in various applications. Speaker detection is the core task in NIST's speaker recognition evaluations (SREs). Like SRE 2019 [1], the 2021 edition of speaker recognition evaluation (SRE21) focuses on speaker verification over conversational telephone speech (CTS) and audio from video (AfV). Similar to SRE 2019, SRE21 also features three tracks namely audio-only, visual-only and audio-visual tracks involving automatic person verification using audio, image, and audio-visual or video modalities, respectively. But unlike SRE2019, the SRE21 introduces two major challenges and they are (i) cross-lingual trials and (ii) cross-dataset trials [2].

In order to tackle the introduced challenges, in ABC team, we focused on different parts or stages of automatic speaker verification pipeline. One can build separate verification systems for CTS & AfV and then perform score level fusion for taking care of the cross-corpus (i.e., CTS versus AfV) trials. The second way of tackling this issue is to train verification systems on the pooled training recordings from CTS and AfV sources, possibly with data augmentation with various audio codecs simulation. Alternatively, one can use only broadband AfV corpus (e.g., VoxCeleb), generate supplementary data employing diversified audio codecs for mimicking CTS scenarios and then use the augmented data to build verification systems. On our side, we mainly adopted the second approach though one of our systems was developed on the top of CTS [3] data only.

For SRE21 audio-only track, we adopted the deep speaker embedding framework, where a suitable embedding extractor is used for training and extracting the deep speaker embeddings in the first phase. In the second phase, a probabilistic linear discriminant analysis (PLDA) or cosine distance backend is used for verification scoring over the extracted enrollment and test embeddings. The popular x-vector - PLDA backend combination is one such framework which uses a time-delay neural network (TDNN) [4] architecture as embeddings extractor backbone network. In this work, as embeddings extractor backbone network, we explored Factored TDNN (FTDNN) [5], some variants of Residual Neural Network (ResNet) [6] and recently proposed Hybrid Neural Network (HNN) [7, 8] architectures to learn robust speaker embeddings. The HNN architecture employs CNN, LSTM and TDNN networks in cascade for capturing complementary information available among these individual networks and incorporates a multi-level global-local statistics pooling to aggregate the speaker information within short time-span and utterance-level context. Following the extraction of deep speaker embeddings, we used probabilistic linear discriminant analysis (PLDA) back-end for scoring after applying centering, whitening, dimension reduction by LDA, and length normalization. Additionally, as post processing we also adopted nuisance attribute projection (NAP) to compensate for dataset shifts and domain adaptation over the model (e.g., supervised PLDA adaptation) and embedding space.

For building visual-only verification systems, we mainly relied on the pre-trained face verification model based on InsightFace [9] that employs ResNet101 architecture as face or visual embeddings extractor. Cosine similarity was used as a backend (or classifier) for verification scoring between enrollment and test embeddings.

In addition to the regular audio-only track, for the first time, the NIST introduced audio-visual track in SRE 2019 and provided a common framework that enabled the speaker recognition research

10.21437/Odyssey.2022-48

community to explore promising new ideas on multimodal biometrics. In SRE 2019, developed audio-visual systems based on simple score level fusion strategies yielded remarkable performance gains over the unimodal (i.e., audio-only or visual-only) systems, providing a strong evidence about the complementarity of these two modalities [1, 10]. Hence, for SRE21, we formed our audio-visual systems by simply fusing the scores of independent audio and visual systems.

2. AUDIO-ONLY VERIFICATION SYSTEMS

In this section, we describe speaker verification systems developed using voice/audio biometric trait alone. This includes data preparation, training embeddings extractor, backend and lastly, calibration and fusion.

2.1. ResNet34 Embeddings Extractor

2.1.1. Training data and data augmentations

For training the system, we used following databases:

- NIST CTS Superset [3]
- Voxceleb 1 & 2 [11]

There are in total 14096 speakers. We used Kaldi style data augmentation with MUSAN database [12].

- 8k dataset: When training 8kHz sampling frequency-based systems, we downsample all broadband (i.e., Voxceleb 1 & 2) data to 8kHz.
- 16k dataset: When building 16kHz sampling frequency-based models, we use upsampled 8kHz data, original 16kHz data, and 16kHz data downsampled to 8kHz, passed through GSM codec and upsampled back to 16kHz.

Following Mel-frequency filterbank features are fed to the input of the network:

- 8k dataset: 64-dimensional Mel-filterbanks with frequency band limited to 20-3800Hz
- 16k dataset: 80-dimensional Mel-filterbanks with frequency band limited to 20-7600Hz

2.1.2. Development dataset

For monitoring our performance and for both calibration and fusion, we used the official SRE2021 development dataset [2] provided by the NIST and LDC.

2.1.3. ResNet34 as Backbone Network

The backbone of these embedding extractors is a 34-layer ResNet. All convolutional kernels are 3×3, and the number of channels is (64,128,256,256) and the first convolutional layer also outputs 64 channels. The number of convolutional layers per block is (3, 4, 6, 3). The input features are 64-dimensional Mel filterbanks, extracted from 8kHz audio files, and the training segments contain 350 frames.

There are three main differences between the proposed extractors and other ResNet architectures typically used in speaker recognition. First, for all three ResNet systems from our submission, only standard deviation features are included in the statistics pooling layer. The approach was examined in [6] and appears to generalize better, at least in cases where there exists dataset-shifts between training and test settings. Second, for two systems in our primary fusion (Omilia_BUT-RN34_str4_stat and Omilia_BUT-RN34_str4_embd), a reduced temporal stride is applied, which seems helpful for generalizing to new languages. The temporal stride per ResNet stage is set to (1,2,1,2) (i.e. a cumulative stride equal to 4 instead of the standard 8) while the frequency stride is the typically used (1,2,2,2). The motivation is to reduce the receptive field by a factor of 2 in order to model shorter speech patterns, which should be more language-independent. Finally, we experimented with extracting statistics instead of embeddings, as the latter representations are susceptible to overfitting the training speakers and languages. Statistics allows us to experiment with unsupervised dimensionality reduction methods (e.g. PCA) and possibly retain directions that are more discriminative for languages and domains not included in the training set. As a result, two of ResNet systems from our primary fusion use the output of the statistics pooling layer in place of embeddings, and one uses traditional embeddings.

The networks are trained using multi-speaker classification and with Additive Angular Margin (also known as ArcFace [9]) loss with 30 and 0.3 scale and margin, respectively. As optimizer, we use stochastic gradient descent with momentum equal to 0.9. The minibatch size is 256, however to fit it in a single GPU we split the minibatch into 16 "microbatches" of 16 examples each and use gradient accumulation. The initial learning rate (LR) is equal to 0.2 which we divide by 2 when the loss does not improve for more than 3000 model updates in the held-out set (the final LR is 0.2/64).

2.1.4. Backend

Four audio systems from the primary submission share the same approach to train the backend:

First, nuisance attribute projection (NAP) is used to remove the direction corresponding to speaker gender [13, 14]. Then, we proceed with centering the data, LDA reducing dimensionality to 75 or 100, and length normalization. In case the input vectors were the outputs of statistics pooling layer from the embedding network, before LDA, we apply PCA reducing dimensionality of the input from 2048 (2047 after NAP) to 256.

After data pre-processing, we train a mixture of 3 PLDA models [15]: each component of the mixture is a PLDA trained on the data coming from one of three languages: English, Cantonese, and Mandarin. At test time, we estimate the log-likelihood of the enrollment and test segments for each of the models: $\log P(\mathbf{R} \mid M_i)$, where $\mathbf{R}$ is a single embedding for single-session enrollment and test data, and set of three embeddings for multi-session enrollment models, M_i is one of the three PLDAs. Passing these quantities through softmax, we obtain weights that are used to scale the LLR speaker verification scores obtained for each of PLDAs in the mixture. We compute two matrices of weighted scores: one corresponds to weights computed for the enrollment models and another one corresponds to weights based on the test segments. Final score matrix is the average of these two.

The backend models were trained on English, Cantonese, and Mandarin data from CTS-superset. Backend training used the embeddings extracted from original data and one random augmentation per recording.

2.2. Factored TDNN (FTDNN) Embeddings Extractor

2.2.1. FTDNN

In this system, we use the factorized TDNN architecture proposed in [5]. We train it with the Kaldi toolkit [16] with the settings in the sre16/v2 recipe except that we used 16k dataset, features, development test set described in Section 2.1.1. PLDA backend was used for scoring.

2.3. Hybrid Neural Networks (HNN) Embeddings Extractor

This section provides an overview of the speaker verification systems based on Hybrid Neural Networks (HNN) Embeddings Extractor [7, 8] for the NIST SRE 2021 audio-only track.

2.3.1. *Training data*

For training speaker discriminant neural network, we used augmented version of CTS-superset data [3]. In order to generate supplementary data on the top of original CTS-superset training data, we use offline data augmentation using RIRs and MUSAN noises [12]. We also performed on the fly data augmentation using SpecAugment [17] technique.

In order to train PLDA model, we used augmented version of all mixer data from CTS-superset [3].

2.3.2. *Features & SAD*

23-dimensional Mel-frequency cepstral coefficients (MFCC) were extracted using an analysis window of 25 msec with a frame shift of 10 msec. Features are normalized using cepstral mean normalization over a window of 300 frames. Energy-based speech activity detector (SAD) was used to get rid of non-speech frames.

2.3.3. *HNN as Backbone Network*

As speaker discriminant deep embeddings extractor, we employed a hybrid deep learning architecture, introduced in [7, 8], that employs CNN, LSTM and TDNN networks for learning more discriminative local descriptors by capturing the complementarity of CNN, LSTM and DNN/TDNN networks.

The hybrid speaker embeddings extractor, as depicted in Fig. 1, also uses a multi-level global-local statistics pooling, which not only considers the global statistics of each network module, but also extracts the local statistics to take the speaker information within the local variability into account.

The 23-dimensional MFCCs are used as input features to this hybrid model, which are passed through the IDCT-layer (inverse discrete cosine transform) to obtain 23-dimensional Mel-Filterbank (MFB) features. The MFCCs, being decorrelated, are more easily compressible without any information loss and therefore, take less storage space than MFB coefficients.

Over the MFB features, SpecAugment is applied on the fly, where both time and frequency masking are performed. The augmented spectral features are then passed through 5 2-dimensional CNN (2D-CNN) layers to capture the local spectral characteristics.

The 2D-CNN module is then followed by a frame-level network which is composed of TDNN and LSTM layers, to extract local descriptors with sufficient temporal information for speaker discrimination.

The multi-level statistics pooling (MLSP) [18] is used for aggregating first- and second- order statistics from the last layers of CNN, LSTM and TDNN blocks. However, unlike the conventional x-vector, the hybrid architecture extracts the statistics not only globally, but also locally to exploit the short-duration correlation.

As depicted in Fig. 1, each module (i.e., TDNN, LSTM) takes both the frame-level outputs from the previous layer, and the local statistics extracted from them as input. During the local statistics pooling operation, the input sequences are resampled and the pooling window shift rates are adjusted to match the sequence length with the frame-level features.

After propagating the input features to the frame-level network, a global statistics pooling is performed to aggregate the local descriptors obtained from the CNN, LSTM and TDNN blocks. The global first- and second- order statistics are concatenated to a fixed-dimensional utterance-level representation.

The pooled statistics are then projected into a 512-dimensional embedding vector via two fully-connected layers. Once the training is completed, the embeddings are extracted from the fully-connected layer close to the global statistics pooling layer.

Table 1: System performances on the Audio-only tracks of NIST SRE21 Development set.

Track	**System**	**EER**	**min_C**
Audio-only	HNN	12.40	0.648
	HNN (lda_dim=75)	10.89	0.517
	HNN(supPLDA)	8.89	0.487

Table 2: System performances on the Audio-only track of NIST SRE21 Evaluation set.

Track	**System**	**EER**	**min_C**
Audio-only	HNN	13.46	0.657
	HNN (lda_dim=75)	10.28	0.563
	HNN(supPLDA)	8.71	0.507

2.3.4. *Results on the SRE21 Development and Evaluation Test Sets*

We build 3 audio-based speaker verification systems using hybrid neural network (HNN) - based embeddings extractor:

- HNN: In this case LDA is used to reduce embeddings dimension to 200 and then PLDA scoring is applied. No score normalization or any other post-processing is applied. This system is denoted as CRIM-HNN in Table 5.
- HNN (lda_dim=75): In this case LDA is used to reduce embeddings dimension to 75 then PLDA scoring is applied. No score normalization or any other post-processing is applied.
- HNN (supPLDA): LDA is applied to reduce embeddings dimension to 80 and then scoring is performed using adapted PLDA model which was adapted using supervised PLDA adaptation (supPLDA) technique. In this case, we trained two PLDA models, one on mixer data from CTS-superset and another on SRE16 evaluation data [19]. After that, the PLDA parameters (i.e. across-class and within-class covariances) were adapted by doing interpolation. Optimal interpolation parameter was set to $\alpha = 0.70$. No score normalization or any other post-processing is applied.

Tables 1 and 2 present results attained by all three systems mentioned above on the development and evaluation test sets of NIST SRE 2021 audio-only track. Significant improvements in performance were achieved by reducing the dimension of embeddings to 75 or 80 and applying supervised PLDA adaptation.

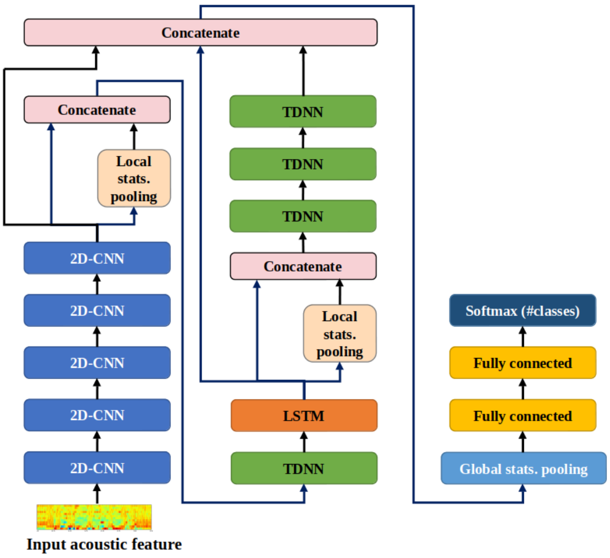


Fig. 1: A schematic diagram of the CNN-LSTM-TDNN - based hybrid neural networks (HNN) embeddings extractor used for the development of voice biometric system for NIST-SRE 2021.

2.4. Calibration and Fusion

All systems were first pre-calibrated and then passed into the fusion. The output of the fusion was then again re-calibrated.

Both calibration and fusion were trained with logistic regression optimizing the cross-entropy between the hypothesized and true labels on a corresponding development set. Our objective was to improve the error rate on the development set.

As we observed disproportionate results on female and male subsets of development trials, we have decided to perform our primary fusion and calibration of our single best system in a gender dependent way. We have calibrated and fused twice – once on female and once on male development trials and to obtain final scores on evaluation set, we have multiplied scores of corresponding fusions with a gender posterior and summed them. Our secondary audio fusion was performed traditionally in a gender independent way on the whole development set.

2.4.1. Gender ID

Gender labels for the evaluation data were inferred as follows. Similar to Section 2.1.4, we train a mixture of PLDAs, only this time instead of using languages we use gender labels to define the mixture components. Unlike in Section 2.1.4, we do not project away the gender direction from the data before training the mixture. Then, normalized likelihoods of each enrollment model or test segment for female model $P(\mathbf{r} \mid M_f)$ are used as soft gender labels (it is 1 for hard female label and 0 for a male label). The final label for a trial is an average of two labels for different sides of that trial.

3. VISUAL-ONLY VERIFICATION SYSTEMS

3.1. Systems developed at CRIM

Three visual-only verification systems were built at CRIM based on pre-trained models. In this section, we provide description of all three face verification systems.

3.1.1. CRIM_AF

For the visual task, embeddings were extracted from every enrollement (speaker) image, and every single video frame of test segments where a face is detected. The ffmpeg tool was used to extract frames from videos every second.

To extract embeddings, we employ insightface tensorflow deep face analysis toolbox [20] with onnxruntime-gpu as inference backend to produce 512-length face embeddings which are then used to compute cosine similarities between a trial's enrollment embedding and test segment frames' embeddings, extracted every second. For each trial, the maximum score of cosine similarities is chosen. In our configuration, we use a pretrained RetinaFace [21] model for face detection, and an antelopev2 model of model size 407MB for face recognition. The latter is based on ResNet100 model trained on the Glint360K [22] training dataset. This configuration, throughout our experiments, has shown to overperform by far all other tried models (Insightface's buffalo_l, and arcface, facenet, and vggface pretrained face recognition models available in the DeepFace [23] framework). The results of this system, denoted as **CRIM_AF**, on the NIST SRE 2021 visual-only track development and evaluation sets are included in Tables 3 and 4, respectively.

Normalization, face alignment (warping based on facial landmarks), and image cropping to produce 112×112 face images were applied exactly as originally employed in insightface. We use a single NVidia GTX1050Ti GPU which took around 20 hours to process all video frames to produce face embeddings. Embeddings were normalized by first subtracting the mean of each embedding, then subtracting value of the global mean of the development set embeddings. When the model is unable to detect any face from a video, and thus face recognition is not possible, an arbitrary neutral score of 0.5 is returned for cosine similarity. Face detection model was perfectly effective at detecting faces from videos at the exception of only 2 videos and 9 videos (because of over-exposure issues) respectively in the development set and evaluation set.

3.1.2. *CRIM_MD_1p66*

The design of the **CRIM_MD_1p66** face recognition model is based on the InsightFace (pytorch-based) recognition system [9]. First, an embedding is generated for each image and each sampled frame, respectively, from the enrollment and the test sets. Then models are created for each image and each video. A single identity is considered, and it corresponds to the first detected face when multiple faces are found. Mainly, the recognition pipeline integrates 3 components, feature generation, model definition, and scoring strategy.

3.1.2.1. *Embedding generation*

First, from each image/frame, the face region, if indeed exists, is delimited using the RetinaFace detector [21]. A post processing permits then to warp and crop the face using 5 facial fiducial points. After that, the obtained face image is resized to a standard size of 112×112 pixels. The RetinaFace detector has also been adopted as a baseline of the NIST visual speaker recognition system [24]. The face image is then fed to the InsightFace model [9] to generate the corresponding embeddings. The feature generator adds and normalizes the embeddings of the images/frames themselves to those of their flipped versions.

3.1.2.2. *Target model generation*

The features are generated for each image from the enrollment set as specified above. Hence, each target model is represented by a single embedding vector. Then, for each video from the test set the frames sequence is sampled at a frame rate of 1/FPS [1]. In this case, the number of embedding vectors depends on the length of the video sequence. In each video there is a limited number of emotion/pose variation modes, and often the frames are almost similar. The **CRIM_MD_1p66** system exploits the Chinese Whispers clustering algorithm [25] to properly group the embeddings of the test video. An appropriate threshold allows to merge the most similar embeddings and keep those that are less similar, as they are, even if they are from the same identity. The idea is to keep only their restrained representative modes that well explain the face variations in the video. The algorithm only needs a threshold parameter, contrary to the k-means++ clustering algorithm that is suggested in the NIST's baseline [24] where the number of modes (k) must be fixed in advance. Clearly, there is no impetus suggesting that all video sequences should have the same number of centroids. An example of face clustering using the Chinese Whispers algorithm is given in [26, 27].

3.1.2.3. *Scoring strategy*

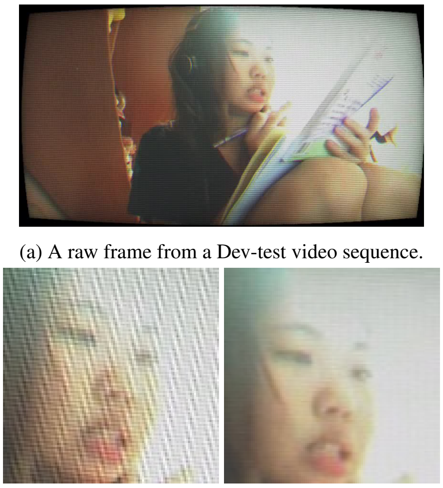
A pairwise cosine similarity is calculated between the enrollment image model and each one of the embeddings' centroid of the test video for a given trial pair i.e. (single enrollment embedding, test video embeddings' centroids). The maximum of the pairwise scores gives a final score for the current trial.

3.1.2.4. *Missing faces*

In the development set, all the faces of the enrollment and the test sets were detected. However, in the evaluation set there were $23/3\,177$ test videos where the detector does not detect any face; either because the detector failed or there was no face in the frame at all. The total number of trials with missing faces reached $2\,263$ pairs. Our strategy for inferring an arbitrary score for these videos is to choose a scoring threshold with a smaller EER from the development set. The arbitrary score corresponds to the average of all the scores obtained by removing at each time one test video. The inferred value defines the ambiguous area around the decision boundary separating the target and non-target classes.

3.1.3. *CRIM_MD_NEW2*

The face recognition system **CRIM_MD_NEW2** is a nuanced variant of the **CRIM_MD_1p66** system. The difference between them lies in the fact that we have added a preprocessing layer that is in charge of denoising all the pipelined raw images/frames. The layer consists of a filtering operation in the frequency domain using the fast Fourier transform to attenuate the high frequency signal. An example of a filtered image is given in figure 2c. The filtering operation led to an increase in the number of missing faces, 26 against 23 for the previous model, bringing up the number of trial pairs without faces to $2\,579/283\,011$. It results in a slightly different ambiguity threshold (0.71) against (0.70). These findings are in almost perfect agreement with the results from [9] where the authors have managed to show a clear separation between the angle distribution of the positive pairs and the negative ones; their optimal threshold value, which can be referred to as an ambiguity angle, was around 65° for ArcFace and about 70° in the case of Triplet-Loss; please refer to [9, Fig. 6]. Comparatively, the ambiguity scores (0.7) and (0.71) of our two models are well aligned with the ambiguity angles from [9] as $\arccos(0.70 \times 2 - 1) = 66.4^\circ$, notice that the scores have been scaled to $[0, 1]$.



(a) A raw frame from a Dev-test video sequence.

(b) Without filtering

(c) With filtering

Fig. 2: Face detection, cropping, and alignment: Filtered vs. non filterd image/frame.

In figure 3, we show the impact of the extra processing step in **CRIM_MD_NEW2** on the score distribution. Many samples have been moved to the correct class. The relative order of the samples' scores is also an important factor with regard to the evaluation metric (EER). The figure reveals clear evidence of the target/non-target discrimination afforded by the proposed systems.

3.1.4. *Results on the SRE21 Development and Evaluation Test Sets*

Tables 3 and 4 present results obtained by all three face verification systems mentioned above on the development and evaluation test

[1] The videos are of different frames-per-second (FPS) rate.

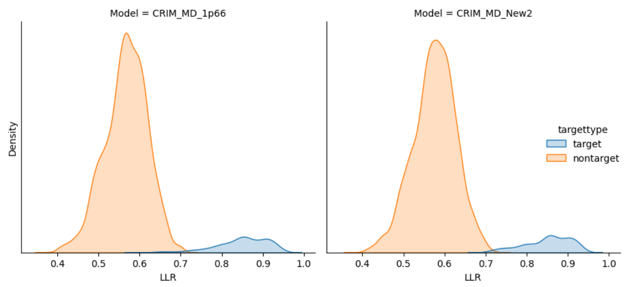


Fig. 3: Score distributions of target and non-target pairs in the development test set.

Table 3: System performances on Visual-only tracks of the NIST SRE21 Development set.

Track	System	EER	min_C
Visual-only	CRIM_AF	2.15	0.04
	CRIM_MD_1p66	1.66	0.027
	CRIM_MD_NEW2	0.17	0.017

sets of NIST SRE 2021 visual-only track. On the development set best performance was achieved by the **CRIM_MD_NEW2** system whereas on the evaluation set system **CRIM_MD_1p66** yielded the best face verification results.

3.2. Systems Developed at Innovatrics

Innovatrics provided their toolkit with API to be incorporated to the system at BUT. Here are the main components of the Innovatrics face recognition system:

- **Face detection**: accuracy / speed trade-off tuned version of approach described in [28]
- **Facial landmark detection**: Precise detection of facial landmarks is necessary for proper alignment of face before face embedding can be extracted. The modified version of the method described in [29] was used for facial landmark detection.
- **Face embedding**: face embedding with size of 512B was generated by neural net with ResNet100 architecture [30]. The network was trained using modified ArcFace loss [9] on more than 10M images (80% google searched images of celebrities, 20% internal) of several 100k identities.

Table 4: System performances on Visual-only track of the NIST SRE21 Evaluation set.

Track	System	EER	min_C
Visual-only	CRIM_AF	5.41	0.099
	CRIM_MD_1p66	1.49	0.067
	CRIM_MD_NEW2	1.62	0.081

3.3. Calibration and Fusion

When submitting individual systems, we just calibrated with LR on the whole video development set. When fusing, we were struggling with extremely low amounts of errors on the DEV set and we resorted to averaging the scores of individual calibrated systems and a subsequent post-calibration with LR.

4. AUDIO-VISUAL VERIFICATION SYSTEMS

During SRE 2019 and subsequently it became evident that the audio and visual modalities pose complementary information and score level fusion audio- and visual-only systems lead to significant gain in performance over the unimodal systems. Inspired by this findings from SRE 2019 [1, 10], when submitting to the audio-visual track, we used the same approach of averaging and post-calibration of scores as in the visual-only track. For both tracks, we used the corresponding DEV set to calibrate.

5. RESULTS AND DISCUSSIONS

In this section, we report results of the ABC team's competitive individual and fused speaker/person verification systems developed for the NIST SRE21 for audio-only, visual-only and audio-visual fixed condition tracks on the SRE21 development and evaluation test sets. We also report post-evaluation results obtained by performing fusion of additional combination of submitted systems. The equal error rate (EER), minimum primary cost (min_C) and actual primary cost (act_C) are used as metrics for evaluating the verification performances.

Table 5 shows results of our individual audio-only systems developed using ResNet34, FTDNN and Hybrid Neural Network (HNN)-based deep embeddings extractors. The primary and secondary audio-only fused systems as well as the single best system's results are also reported in the same table. One can observe from the reported results on the development and evaluation sets that the difference in performances is very narrow between the single best and fused audio-only verification systems.

Table 5 also presents results of our individual and fused visual-only systems. On the SRE21 evaluation set, in terms of EER, **CRIM_MD_1p66** system outperformed the other two systems. On the other hand, **Innovatrics** system achieved the best results in terms of min_C and act_C on the same test set. We can see from this table that not much improvement in performances were attained from the fusion of visual-only systems. As the visual-only systems were built following almost identical techniques they pose very little complementary information.

Additionally, the audio-visual systems results are reported in Table 5. Compared to the performance of audio-only systems audio-visual systems, obtained by score-level fusion of audio- & visual-only systems, demonstrated remarkable gain in performances. Again, compared to visual-only systems' results performance improvements secured by the audio-visual systems is small. This is due to the design simplicity of visual-only track. For visual-only track, NIST provided a close-up image of the target speaker for enrollment and single speaker test video or image for test which makes visual-only track an extremely easy task.

Our observations or findings throughout the participation of SRE21 can be summarized as follows:

- In audio-only track, PLDA backend performed better than cosine scoring.

Table 5: Results of the systems for the NIST SRE 2021 Fixed Condition. AUDIO systems were evaluated on the whole development set, VISUAL and AV ones were evaluated on the audio-visual subset of the development set. * Notice the difference between audio system 2 and Single Best Audio system. The only difference between the two is the calibration approach used. In the latter system, gender-dependent calibration is used while in the former one not. **This system was not submitted during the evaluation.

	System	**2021 dev set**			**2021 evl set**		
		min_C	**act_C**	**EER**	**min_C**	**act_C**	**EER**
	AUDIO						
1	Omilia_BUT-RN34_str8_stat	0.580	0.631	10.00 %	0.559	0.563	9.63 %
2	*Omilia_BUT-RN34_str4_stat	0.482	0.494	7.67 %	0.468	0.473	7.90 %
3	Omilia_BUT-RN34_str4_embd	0.482	0.507	8.58 %	0.473	0.480	8.38 %
4	BUT-FTDNN	0.504	0.523	11.24 %	0.508	0.517	8.82 %
5	CRIM-HNN	0.648	0.656	12.40 %	0.657	0.668	13.46 %
	VISUAL						
6	CRIM_MD_1p66	0.027	0.067	1.66 %	0.067	0.314	1.49 %
7	CRIM_MD_NEW2	0.017	0.021	0.17 %	0.081	0.292	1.62 %
8	Innovatrics	0.018	0.018	1.82 %	0.037	0.099	1.80 %
	Primary AUDIO Fusion = 1+2+3+4	0.437	0.441	6.66 %	0.446	0.447	7.63 %
	Secondary AUDIO Fusion = 2+5	0.481	0.492	7.67 %	0.466	0.472	7.89 %
	*Single Best AUDIO = 2	0.471	0.478	6.70 %	0.454	0.459	7.98 %
	Primary VISUAL Fusion = 6+7+8	0.004	0.005	0.33 %	0.043	0.433	1.25 %
	Single Best VISUAL = 7	0.017	0.021	0.17 %	0.081	0.292	1.62 %
	Single Best VISUAL = 8	0.018	0.018	1.82 %	0.037	0.099	1.80 %
	Primary AV Fusion = 2+6+7+8	0.001	0.001	0.07 %	0.040	0.387	1.10 %
	**Secondary AV Fusion = 2+5+8	0.011	0.025	1.10 %	0.029	0.036	1.97 %
	Single AV Best = 2+7	0.004	0.005	0.03 %	0.057	0.220	1.31 %

- Dimensionality reduction by LDA played a key role on boosting audio-only speaker verification performance. Better results were provided by PLDA backend when embeddings' dimension were reduced to either 80 or 100 by LDA.
- Fine-tuning the ResNet embeddings extractor on longer duration segments brought interestingly important improvement in performances. Decreasing the stride in ResNet was proved to be conducive.
- Supervised PLDA adaptation by treating SRE 2016 evaluation data [19] as in-domain data was helpful and has lead in gains in the verification performance.
- Visual-only verification task becomes much simpler if a close-up image of the target individual is used for enrollment and single speaker test video or image for test.
- Audio and visual biometric traits pose significant complementary information and therefore, lead to improved performance when fused together.

6. CONCLUSION

In this work, we presented an overview of the ABC team's competitive efforts toward the development of automatic person verification systems for NIST speaker Recognition Evaluation 2021 (SRE21) based on audio, visual and audio-visual biometric traits. Introduction of cross-lingual and cross-dataset trials made audio-only verification track much more challenging compared to previous SREs. On the other hand, instead of providing video recording as in the SRE 2019, in SRE21, a close-up image (e.g., selfie) of the target individual was provided for enrollment which turned visual-only track so simpler that it yielded almost zero EER (equal error rate) on the visual development test set. Once again multi-modal fusion of audio-only and visual-only systems demonstrated the best performance compared to unimodal systems.

7. ACKNOWLEDGMENT

BUT researchers were supported by Czech Ministry of Interior project No. VJ01010108 ”ROZKAZ”, Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, Czech Ministry of Education, Youth and Sports project no. LTAIN19087 ”Multi-linguality in speech technologies”, European Union's Horizon 2020 project No. 833635 ROXANNE and Horizon 2020 Marie Sklodowska-Curie grant ESPERANTO, No. 101007666. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project ”e-Infrastructure CZ – LM2018140”. CRIM researchers wish to acknowledge funding from the Government of Canada's New Frontiers in Research Fund (NFRF) through grant NFRFR-2021-00338, 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.

8. REFERENCES

[1] Seyed, C. Greenberg, E. Singer, D. Olson, L. Mason, and J. Hernandez-Cordero, "The 2019 nist audio-visual speaker

recognition evaluation," 2020-05-18 2020, The Speaker and Language Recognition Workshop: Odyssey 2020, Tokyo, -1.

[2] O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "Nist 2021 speaker recognition evaluation plan," 2021-07-12 04:07:00 2021.

[3] O. Sadjadi, "Nist sre cts superset: A large-scale dataset for telephony speaker recognition," 2021-08-16 04:08:00 2021.

[4] Y. Cai, L. Li, D. Wang, and A. Abel, "Deep speaker vector normalization with maximum gaussianality training," 2020.

[5] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, R. Dehak, L. P. García-Perera, D. Povey, P. A. Torres-Carrasquillo, S. Khudanpur, and N. 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.

[6] S. Wang, Y. Yang, Y. Qian, and K. Yu, "Revisiting the statistics pooling layer in deep speaker embedding learning," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[7] J. Alam, A. Fathan, and W. H. Kang, "Text-independent speaker verification employing cnn-lstm-tdnn hybrid networks," in *23rd International Conference on Speech and Computer (SPECOM), Lecture Notes in Computer Science, Springer, Cham*, 2021, vol. 12997, pp. 1–13.

[8] W. H. Kang, J. Alam, and A. Fathan, "Hybrid network with multi-level global-local statistics pooling for robust text-independent speaker recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, vol. accepted for publication.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[10] J. Alam, G. Boulianne, L. Burget, M. Dahmane, M. Diez Sánchez, A. Lozano-Diez, O. Glembek, P.-L. St-Charles, M. Lalonde, P. Matejka, P. Mizera, J. Monteiro, L. Mosner, C. Noiseux, O. Novotný, O. Plchot, J. Rohdin, A. Silnova, J. Slavicek, T. Stafylakis, S. Wang, and H. Zeinali, "Analysis of ABC Submission to NIST SRE 2019 CMN and VAST Challenge," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 289–295.

[11] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech Language*, vol. 60, pp. 101027, 2020.

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

[13] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in *Proceedings of ICASSP*. IEEE, 2014, pp. 4002–4006.

[14] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for svm speaker recognition," in *Proceedings of ICASSP*, 2005, vol. 1, pp. I/629–I/632 Vol. 1.

[15] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multi-condition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4257–4260.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[18] Y. Tang, G. Ding, J. Huang, X. He, and B. Zhou, "Deep speaker embedding learning with multi-level pooling for text-independent speaker verification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6116–6120.

[19] Seyed, T. Kheyrkhah, A. Tong, C. Greenberg, D. Olson, E. Singer, L. Mason, and J. Hernandez-Cordero, "The 2016 nist speaker recognition evaluation," 2017-08-20 2017, Interspeech 2017, Stockholm, -1.

[20] "An open source 2d&3d deep face analysis library," https://github.com/deepinsight/insightface/tree/master/python-package, Accessed: 2021-10-28.

[21] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[22] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, et al., "Partial fc: Training 10 million identities on a single machine," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1445–1449.

[23] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020, pp. 23–27.

[24] G. C. S. E. M. L. Sadjadi, O. and D. Reynolds, "NIST 2021 Speaker Recognition Evaluation Plan, NIST SRE," 2021.

[25] C. Biemann, "Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems," in *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, New York City, June 2006, pp. 73–80, Association for Computational Linguistics.

[26] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[27] D. E. King, "dlib Python API," 2021.

[28] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "Ssh: Single stage headless face detector," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4875–4884.

[29] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 88–97.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.