The XXX Submission for NIST SRE24: Lessons Learned Adapting Speaker Recognition to Low-Resource Tunisian Data

Jesús Villalba[1,2]*, Jonas Borgstrom*[3]*, Prabhav Singh*[1]*, Leibny Paola García*[1,2]*, Pedro A. Torres-Carrasquillo*[3]*, Najim Dehak*[1,2]

[1]Center for Language and Speech Processing, Johns Hopkins University, MD, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, MD, USA
[3]MIT Lincoln Laboratory, MA, USA

{jvillalba, psingh54, lgarci27, ndehak3}@AAA.edu, {jonas.borgstrom, ptorres}@ll.mit.edu

Abstract

We present the XXX submission for NIST SRE24 Audio-Visual together with key insights we gained from this evaluation. In the audio fixed conditions, the system utilized Res2Net50 and ResNet100 embeddings, while the open condition additionally included an ECAPA-TDNN with a Multilingual Wav2Vec2 front-end, which emerged as the best single system. The audio back-ends consisted of either PLDA adapted to SRE24 Dev or a mixture of PLDA models tuned to different subconditions. To avoid overfitting, we optimized back-end hyperparameters using two-fold cross-validation. For the visual conditions, we leveraged pre-trained ResNet100 face embeddings. Agglomerative clustering grouped speaker and face identities in multi-speaker test videos. The primary audio fixed system achieved Act. Cp=0.574, while the open condition reached Cp=0.366 on SRE24 Eval. The visual system had Cp=0.169, while the Audio-Visual fusion significantly improved performance, achieving Cp=0.101 (fixed) and Cp=0.087 (open).

Index Terms: speaker verification, nist-sre, evaluation, x-vector, PLDA, audio-visual

1. Introduction

The National Institute of Standards and Technology (NIST) periodically organizes speaker recognition evaluations (SRE) to benchmark the latest advancements in the field [1]. These evaluations center on the speaker detection task, i.e., determining whether the speaker in a test recording matches the speaker in one or more enrollment recordings. Over time, SRE has progressed from telephone speech [2], to far-field mics [3, 4], then to non-English telephone speech [5, 6, 7], and multi-modal evaluations on internet videos [6, 7]. NIST SRE21 [8] featured a multi-modal, multi-language, multi-source evaluation with conversational telephone speech (CTS) and audio from videos (AfV). It introduced new challenges, including cross-source (CTS enrollment, AfV test) and cross-language trials (English, Cantonese, and Mandarin). SRE24[1] follows the same setup as SRE21, incorporating audio, visual (face recognition), and multi-modal tracks. However, it introduces Tunisian-accented Arabic, French, and English as target languages and allows multi-speaker tests, necessitating speaker and face diarization.

This paper presents the XXX[2] submission to NIST SRE24, highlighting key lessons learned. It is a collaboration between AAA and BBB, building on expertise gained from previous evaluations [9, 10, 11, 12]. For audio, we had pipelines based on ResNet100 [13], Res2Net [14], and Multi-lingual Wav2Vec2-ECAPA-TDNN embeddings [15]. Since in-domain data was unavailable for training, the embeddings were processed using PLDA-based back-ends, carefully adapted to 20 in-domain development speakers through two-fold cross-validation. To handle varying source and language conditions, either condition-dependent preprocessing of embeddings or a mixture of condition-dependent PLDA back-ends was applied. For video, we utilized pre-trained face detectors and Subcenter-ArcFace embeddings [16]. Additionally, we implemented a method to discard low-quality face embeddings, improving our results w.r.t. previous evaluations [11, 12].

2. Datasets

2.1. Train Datasets

The audio track proposed fixed and open training conditions. The fixed condition data consisted of 638k recordings from 7,251 speakers combining NIST SRE-CTS Superset [17] (large-scale CTS data compiling SRE 1996-2012), NIST SRE16 Eval [5] (CTS data from 101 Cantonese and 100 Tagalog speakers) and NIST SRE21 [8] (CTS and AfV data from 183 bilingual speakers of English, Mandarin, and Cantonese speakers).

For the open condition, we added VoxCeleb 1+2 [18] (7365 AfV speakers), NIST SRE18 [6] (CTS data from 210 Tunisian Arabic speakers), NIST SRE19 [7] (CTS data from 196 Tunisian Arabic speakers) resulting in a total of 836k recordings from 14,903 speakers. We also reused models from NIST SRE21 fixed [12]. The SRE21 setup excluded SRE21, SRE18, and SRE19 from training and held out a few speakers from SRE-CTS Superset and SRE16 for development.

For embedding training, we augmented speech on-the-fly with MUSAN noise [3], AIR[4] reverberations, and simulated telephone channel (only for 25% of AfV recordings). The telephone simulation involved downsampling to 8 kHz, applying a random bandpass filter (100–300 Hz low-cut, 3400–3700 Hz high-cut), encoding with A-law, mu-law, G723.1, or G726 codecs from torchaudio [5], and upsampling back to 16 kHz. No augmentation was used for back-end training.

2.2. Development datasets

We used two datasets for development:

- **NIST SRE21 Dev**: 20 speakers with 193k audio trials and 38.9k audio-visual trials used for performance monitoring and calibration.
- **NIST SRE24 Dev**: Provided by the organizers, it includes 20 speakers with 1.17M audio trials and 258k audio-visual trials.

[1] https://www.nist.gov/system/files/documents/2024/06/11/NIST_2024_Speaker_Recognition_Evaluation_Plan.pdf

[2] Missing Team names and cites will be added in camera ready to comply with the anonymity requirement

[3] http://www.openslr.org/resources/17

[4] http://www.openslr.org/resources/28

[5] https://pytorch.org/audio/

It was used for back-end adaptation, performance monitoring, calibration, and fusion. We split it into two folds to tune adaptation hyperparameters and prevent overfitting, ensuring each fold had 10 gender-balanced speakers. When splitting, inter-fold non-target trials had to be discarded, but target trials remained unchanged from the original SRE24 Dev.

3. Audio Systems

3.1. ResNet and ECAPA-TDNN

We used log-Mel-filter-bank features with 16 kHz inputs and two configurations: Wideband (80 filters, 20-7600 Hz) and Narrowband (64 filters, 64-3700 Hz). Features were short-time mean-normalized over 3-seconds windows, with silence removed using Kaldi energy VAD or provided time marks.

The embedding networks consisted of an encoder that extracts frame-level discriminant embeddings, a pooling mechanism, and a classification head [19]. As encoder, we used ResNet100 [13] or Res2Net50 [14, 12]. We added frequency-wise squeeze-excitation (FwSE) [20] to the output of each ResNet/Res2Net block. We used channel-wise attentive statistics pooling [21], 192 dim. embeddings, and subcenter additive angular margin softmax loss [22] with two subcenters per class. The networks were trained on 2-second chunks with a margin of 0.2 using Adam optimizer, learning-rate=0.1, halved every 40k(fixed)/50k(open) steps. After training, we performed a large-margin (margin=0.3) fine-tuning on 4 second-chunks (SGD optimizer, learning-rate=0.01 with cosine schedule with a period of 2500 steps, momentum=0.9) where we added hard-prototype mining (8 hard-prototypes) InterTop-K penalty [23] margin (K=5, penalty=0.1). We had fixed/open and Narrowband/Wideband versions of these networks. Additionally, we had a Res2Net50 (without FwSE) and an ECAPA-TDNN (4 layers of 2048 dim) from NIST SRE21 fixed condition [12].

3.2. Wav2Vec2+ECAPA-TDNN

This network uses Multilingual Wav2Vec2 Large, trained on 128 languages[6] [24], as a feature extractor. A weighted average of its hidden layers is then fed into an ECAPA-TDNN embedding network with three 1024-dim. Res2Net layers, following [25]. This was trained in three stages. First, the ECAPA-TDNN and weighted average coefficients were trained with frozen Wav2Vec2 (margin=0.2, SGD optimizer, learning rate=0.4 warmed up 3.5k steps and halved every 10k steps, momentum=0.9, batch-size=1024) on 3-second chunks for 68k steps. Second, it was fine-tuned by unfreezing Wav2Vec2 (margin=0.2, InterTop-K penaly=0.1, learning rate 5e-5 warmed up for 6k steps and halved every 5k steps), for 33k steps. Third, hard-prototype mining fine-tuning (margin=0.4, learning-rate=1.3 with cosine period 2.5k steps) was applied on 8-second chunks for 2.5k steps. We conducted several experiments to identify hyperparameters that would prevent the network from overfitting to the out-of-domain training data and perform well on the SRE24 dev data.

3.3. Language Identification

We automatically labeled the evaluation data and used ground truth labels for other datasets. For the fixed condition, we trained a FwSE-ResNet34 LID network on the fixed data. For the open, we used the Res2Net50 trained on the LRE22 Open condition in [26]. Then, we trained Gaussian back-ends on SRE24 dev, to classify between English, Arabic, and French.

[6] https://huggingface.co/facebook/wav2vec2-xls-r-300m

3.4. Speaker Diarization

We performed speaker diarization on AfV test recordings using Agglomerative Hierarchical Clustering (AHC) of speaker embeddings. Each system performed its own diarization with its corresponding embeddings computed from 3-second windows (1-second shift). A PLDA adapted to SRE24 Dev generated the self-similarity matrix for AHC, with score calibration also trained on SRE24 Dev. The AHC stopping threshold was set to 0, with a maximum limit of four speakers. Diarization time marks served as VAD to extract an embedding per diarized speaker. The back-end, then, scored enrollment embeddings against all detected speakers, selecting the highest-scoring match.

Ultimately, speaker diarization improved the single systems' Min DCF by less than 3% relative on the SRE24 Eval, which was not a significant gain. Visual inspection of the videos suggests that most contain only one or two speakers, with the target speaker dominating the audio. This could explain the limited impact of diarization.

3.5. AAA Back-end

AAA back-end pipeline followed AAA-v2 in [12], applying condition-dependent centering, global PCA, Whitening, length normalization, and PLDA adapted to in-domain speakers (Mandarin(CMN)/Cantonese(YUE) for SRE21 and Arabic(ARA)/French(FRA) for SRE24). First, we computed separate means and covariances for CTS and AfV, then adapted per in-domain language. For SRE24, in-domain data included SRE24 Dev, adding SRE18-19 in the open condition. This resulted in six adapted means used for in-domain centering, while the CTS and AfV means were used for out-of-domain. Next, joint PCA dim. reduction/whitening was computed from the average adapted covariances, ensuring balanced condition weighting. The same projection was applied to all data, followed by length normalization. Finally, we trained an SPLDA model in all out-of-domain data and adapted it to the in-domain data.

Back-end hyperparameters were tuned based on the test set (SRE21 or SRE24) and training setup (SRE24 fixed/open or SRE21). SRE24 Dev was split into two folds, with back-end adaptation on one fold and evaluation on the other, yielding *fold* scores. These tuned hyperparameters were then used to train a final back-end on the full SRE24 Dev, evaluated on SRE24 Dev (*cheat* scores) and SRE24 Eval. This resulted in three back-ends (*fold0*, *fold1*, and *cheat*). SRE24 back-end hyperparameters were selected based on *fold* scores.

3.6. BBB Back-end

To handle diverse evaluation conditions, the BBB back-end used an ensemble of scoring pipelines. Each included LDA (150 dim.), global centering, whitening, length normalization, and SPLDA (100 speaker dim.). The pipelines were adapted to each of the following conditions: Gender (Male, Female), Source (CTS, AfV), Active Speech Duration (Short ($<$ 15s), Long ($>$ 15s)) and Language (ENG, ARA, or FRA). For each pipeline, Centering/whitening and PLDA were first trained on out-of-domain data and adapted to in-domain subsets. Fixed-condition systems used NIST SRE21 Eval and SRE24 Dev, while open-condition added SRE18-CTS and SRE19-CTS.

The ensemble of scoring pipelines generated a 9-dimensional score vector $\mathbf{x}$. Target/non-target scores were modeled by Gaussian mixture models with 2-3 components and shared covariances. Trial scores were computed as the log-likelihood ratio $s = \log \frac{\sum_i w_{T,i} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{T,i}, \boldsymbol{\Sigma}_i)}{\sum_i w_{N,i} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{N,i}, \boldsymbol{\Sigma}_i)}$ where weights

Table 1: *Audio systems results on SRE21 Dev, SRE24 Dev Folds, SRE24 Dev Full (Cheating) and SRE24 Eval*

System			SRE21 Dev			SRE24 Dev Folds			SRE24 Dev Full			SRE24 Eval		
Idx	Embed.	BE	EER	Min Cp	Act Cp	EER	Min Cp	Act Cp	EER	Min Cp	Act Cp	EER	Min Cp	Act Cp
Single Fixed														
1f	FwSE-Res2Net50-WB	AAA	3.97	0.351	0.405	8.31	0.617	0.638	3.01	0.372	0.520	7.81	0.632	0.681
2f	FwSE-Res2Net50-NB	AAA	**2.95**	**0.297**	0.404	8.26	0.641	0.655	3.30	0.477	0.593	7.94	0.631	0.652
3f	FwSE-ReNet100-WB	AAA	4.18	0.369	0.402	**7.55**	**0.556**	**0.568**	**2.25**	0.322	**0.421**	7.13	0.612	0.709
4f	FwSE-ResNet100-NB	AAA	3.47	0.326	**0.368**	7.58	0.568	0.589	2.33	**0.319**	0.446	**6.90**	**0.569**	**0.613**
5f	FwSE-ReNet100-WB	BBB	5.01	0.389	0.959	10.95	0.637	0.644	4.07	0.388	0.544	9.69	0.716	0.784
6f	FwSE-ResNet100-NB	BBB	4.87	0.366	0.934	9.41	0.643	0.657	3.76	0.403	0.522	8.69	0.652	0.681
7f	FwSE-Res2Net50-WB	BBB	6.89	0.453	1.715	8.78	0.594	0.601	3.42	0.372	0.516	8.81	0.671	0.686
8f	FwSE-Res2Net50-NB	BBB	7.78	0.459	0.716	9.29	0.632	0.646	3.78	0.419	0.577	8.38	0.626	0.628
Single Open														
1o	W2V2-ECAPA-TDNN	AAA	3.50	0.322	0.404	**5.08**	**0.319**	**0.321**	2.56	0.244	0.376	**4.42**	**0.377**	**0.457**
2o	FwSE-Res2Net50-WB-Std	AAA	2.67	0.314	0.384	6.15	0.416	0.419	1.61	**0.198**	0.322	5.33	0.500	0.668
3o	FwSE-Res2Net50-WB-NoCodec	AAA	2.64	0.312	0.384	6.13	0.428	0.434	**1.51**	**0.198**	0.321	5.26	0.499	0.678
4o	FwSE-ResNet100-WB	AAA	3.55	0.379	0.423	6.27	0.419	0.422	2.45	0.302	0.544	5.04	0.446	0.473
5o	Res2Net50-SRE21	AAA	1.92	0.260	0.332	6.03	0.433	0.434	2.68	0.331	0.599	5.29	0.450	**0.456**
6o	ECAPA-TDNN-SRE21	AAA	2.64	0.329	0.386	7.60	0.505	0.507	4.37	0.407	0.644	6.65	0.555	0.559
7o	FwSE-ResNet100-WB	BBB	1.73	0.267	0.640	9.35	0.611	0.639	4.13	0.329	0.489	6.63	0.577	0.623
8o	W2V2-ECAPA-TDNN	BBB	**1.36**	**0.237**	**0.301**	7.41	0.434	0.439	1.75	0.210	**0.243**	5.81	0.483	0.484
Submissions Fixed														
Primary: 3f+4f+5f+1f+6f						6.30	0.486	0.490	1.49	0.237	0.408	5.96	0.547	0.574
Contrastive: 3f+4f+5f+1f+6f+2f+8f+7f						6.19	0.483	0.483	1.50	0.238	0.413	**5.93**	**0.542**	**0.568**
Single: 3f						7.55	0.556	0.568	2.25	0.322	0.421	7.13	0.612	0.709
Submissions Open														
Primary: 1o+2o+5o+4o+7o						4.40	0.249	0.252	1.22	0.161	0.381	**3.60**	**0.318**	**0.366**
Contrastive: 1o+2o+5o+4o+7o+6o+3o+8o						4.31	0.251	0.254	1.36	0.170	0.399	3.67	0.324	0.387
Single: 1o						5.08	0.319	0.321	2.56	0.244	0.376	4.42	0.377	0.457

means, and covariances were the Maximum Likelihood estimates across the in-domain datasets.

3.7. Calibration and Fusion

As the BBB back-end (Sec. 3.6) produced well-calibrated scores, no explicit calibration was applied. AAA trained a condition-dependent calibration of scores s into log-likelihood ratios as $\text{LLR} = as + b + \mathbf{w}_l^{\text{T}}\mathbf{l} + \mathbf{w}_c^{\text{T}}\mathbf{c}$ where a and b are condition-independent scaling and bias; $\mathbf{l}$, $\mathbf{c}$,

$$\mathbf{l} = \begin{bmatrix}\text{language-match=Y}\\ \text{language-match=N}\end{bmatrix} \quad \mathbf{c} = \begin{bmatrix}\text{source-match=Y}\\ \text{source-match=N}\end{bmatrix} \qquad (1)$$

are 1-hot vectors that indicate the language, and source conditions; and $\mathbf{w}_l$, $\mathbf{w}_c$ are trainable weights representing condition-dependent biases. This calibration improved Act Cp of individual systems on SRE24 Eval between 2 and 10% relative w.r.t. condition independent.

We trained three calibrations on separate score sets (*fold0*, *fold1*, *cheat*) but found they did not generalize well across sets. Uncertain about the best calibration for evaluation, we implemented a mixture of calibration functions. We trained a six-component GMM on each non-calibrated score set, reserving two Gaussians for target scores and four for non-targets. Denoting these GMMs as $p(s|\text{fold0})$, $p(s|\text{fold1})$ and $p(s|\text{cheat})$, the final score was $\text{LLR} = \sum_{t\in\{\text{fold0,fold1,cheat}\}} p(t|s) f_t(s)$ where s is uncalibrated score, and f_t are the calibration functions trained on each score set. In the post-evaluation analysis, we found that the SRE24 Dev *cheat* score distribution did not align with Eval, and using the calibration mixture helped prevent excessively high Actual Cp. However, the best approach would have been calibrating solely on the *fold* scores. For instance, in the Wav2Vec2+ECAPA-TDNN model, calibrating on *cheat* scores resulted in an Act Cp of 1 on SRE24, while calibrating on *folds* reduced it to 0.405, and the calibration mixture yielded 0.457. Nonetheless, system fusion mitigated the calibration error, ultimately ensuring no impact on the primary fusions.

A single fusion was trained on calibrated SRE24 Dev *fold0* and *fold1* scores using a greedy fusion approach [9, 12]. First, we calibrated all systems and selected the best one based on the lowest actual cost. Then, we iteratively added systems, choosing the best combination at each step. Fusion was trained at $P_{\mathcal{T}} = 0.01$ basing system selection on the average of ActDCF at $P_{\mathcal{T}} = 0.01$ and $P_{\mathcal{T}} = 0.005$.

3.8. Audio Submissions and Results

Table 1 summarizes the results for our single systems and submissions under fixed and open conditions. The Primary and Contrastive submissions indicate the systems included in the fusion and the order in which they were selected by the greedy algorithm. In the fixed condition, narrowband (NB) models outperformed wideband (WB) models on average, with an Act Cp of 0.63 compared to 0.7. Additionally, Res2Net performed better than ResNet100, with an Act Cp of 0.66 versus 0.69. However, the best single system on Eval was ResNet100-NB, which outperformed ResNet100-WB—the best on Dev—by 14% relative. Primary and Contrastive fusions further improved Act Cp by 19% and 20%, respectively, mitigating single-system miscalibration and narrowing the gap between Min and Act Cp.

In the open condition, models trained on SRE21 performed as well as or better than newer models. No significant differences were observed between networks with and without codec augmentation. Wav2Vec2+ECAPA-TDNN achieved the best Min Cp, though some miscalibration placed it close to ResNet100 and Res2Net50. The Primary fusion improved Act Cp by 20% w.r.t. to the best single system. The primary open system improved by 36% compared to the primary fixed system. The fixed condition's performance was hindered by the limited availability of AfV data, which only came from SRE21. This issue was resolved in the open condition by reintroducing VoxCeleb data.

Table 2: *Visual systems results on SRE21 and SRE24 Visual*

System	SRE 21 Visual dev			SRE21 Visual eval			SRE 24 Visual dev			SRE 24 Visual Eval		
	EER	Min Cp	Act Cp	EER	Min Cp	Act Cp	EER	Min Cp	Act Cp	EER	Min Cp	Act Cp
Submissions												
Primary	2.38	0.082	0.123	**2.03**	**0.114**	**0.119**	**1.47**	**0.050**	**0.105**	**2.07**	**0.149**	**0.169**
Contrastive	2.43	0.082	**0.122**	2.05	**0.114**	**0.119**	1.56	0.063	0.106	2.14	0.152	0.170
Single	**2.35**	**0.079**	**0.122**	2.10	**0.114**	**0.119**	1.56	0.058	**0.105**	2.10	0.152	0.170

4. Visual Systems

4.1. Pre-Trained Detector and Embedding Extractor

We sampled video frames at 3 frames-per-second (FPS), which allowed us to compensate for the posterior removal of low-quality frames. Face detection was performed using a pre-trained RetinaFace R50[7] model with a decreasing detection threshold, ensuring that faces are detected across varying quality levels. Detected faces were aligned with the facial landmarks and, then, embedded using a ResNet-100 trained with Subcenter ArcFace loss on the WiderFace dataset[8].

4.2. Post-Processing for Low-Quality Image Removal

Improving over previous evaluations, we ensure that only high-quality embeddings are retained by estimating a *quality vector* ($\mathbf{q} = [d_{\text{eye}}, \text{black-ratio}]$) for each detected face, improving relative improvements of around 7%. The *Eye Distance* d_{eye} reflects the relative size of the detected face, with larger and well-proportioned faces generally yielding higher values. First, we discard faces with eye distance lower than the maximum eye distance in the video divided by two.

The *Black Pixel Ratio* is the proportion of black pixels in the cropped image. Pixels with near-zero intensity indicate occlusions, poor lighting, or low exposure. A higher black pixel ratio signifies lower quality. A second filtering stage selects face embeddings with black-ratio $< t_{\text{thr}}$ by iteratively applying thresholds $t_{\text{thr}} \in \{0.1, 0.25, 0.5\}$ until the number of valid embeddings is larger than a minimum (set to 3). Typically, more than 5 valid embeddings are found.

4.3. AHC+Cosine Back-end

We used cosine similarity as the metric for comparing face embeddings. The embeddings from the test video were clustered using agglomerative clustering (AHC) with a stopping threshold t_{AHC}, expecting the clusters to represent different individuals or face orientations. Finally, we scored the enrollment embedding against all test cluster centers and selected the maximum score.

4.4. Calibration and Fusion

Visual systems were calibrated and fused using linear logistic regression on SRE21 Visual Dev+Eval and SRE24 Dev. The single system used a single AHC+cosine back-end with $t_{\text{AHC}} = 0.7$ on ResNet100 face embeddings. The primary fusion combined three back-ends with $t_{\text{AHC}} \in \{0.5, 0.6, 0.7\}$, while the contrastive included back-ends with $t_{\text{AHC}} \in \{0.6, 0.7\}$.

4.5. Visual Submissions and Results

Table 2 shows the results of the visual systems on SRE21 Visual Dev and Eval, and SRE24 Visual Dev. The primary fusion did not provide a significant gain over the single system.

5. Audio-Visual Submissions and Results

Assuming well-calibrated log-likelihood ratios and independence between audio and visual modalities, the audio-visual fusion log-likelihood ratio was obtained by summing the audio and visual scores. The primary and contrastive submissions fused their respective primary or contrastive audio systems with the visual primary systems, while single submissions combined audio and visual single systems. Table 3 presents the results. In the fixed/open condition, AV Single improved Act Cp by 85%/82% over Audio-only Single and 34%/42% over Visual-only Single. AV Primary improved Act Cp by 83%/82% over Audio-only Primary and 40%/42% over Visual-only Primary. These results highlight the complementarity of both modalities. Despite the visual modality outperforming audio, video still benefits from integrating audio information.

Table 3: *Audio-Visual systems results on SRE24 Audio-Visual*

System	SRE24 AV Dev			SRE24 AV Eval		
	EER	Min Cp	Act Cp	EER	Min Cp	Act Cp
Fixed						
Primary	0.59	0.014	0.030	**1.13**	**0.100**	0.101
Contrastive	0.59	0.015	0.030	**1.13**	**0.100**	**0.100**
Single	0.63	0.025	0.037	1.32	0.112	0.113
Open						
Primary	0.27	0.010	0.061	**0.83**	**0.086**	**0.087**
Contrastive	0.26	0.011	0.069	0.84	0.087	0.089
Single	0.42	0.028	0.069	1.00	0.095	0.098

6. Conclusion and Discussion

We presented the XXX systems for NIST SRE24. For the audio fixed conditions, the system used Res2Net50 and ResNet100 embeddings, while the open condition also included an ECAPA-TDNN with a Multilingual Wav2Vec2 front-end, which was the best single system. The audio back-ends were either PLDA adapted to SRE24 Dev or a mixture of PLDA models adapted to different evaluation conditions. To prevent overfitting, we used two-fold cross-validation to tune back-end adaptation hyperparameters. For the visual conditions, we employed pre-trained ResNet100 face embeddings with cosine scoring back-ends. Agglomerative clustering was applied to group speaker and face identities in multi-speaker test videos.

We learned several lessons in this evaluation. VoxCeleb data remains essential for strong AfV performance, as its absence severely degraded fixed-condition results. Large-margin and Wav2Vec2 fine-tuning tended to overfit to out-of-domain data, reducing performance on Tunisian data compared to networks without fine-tuning. Optimizing fine-tuning hyperparameters, such as learning rates and early stopping, required extensive experimentation. Diarization had minimal impact on system performance, while source- and language-dependent calibration proved beneficial. Cross-validation scores provided more reliable calibration than *cheating* (full-dev trained) back-end scores. Audio fusion improved results by approximately 20% and mitigated the effects of suboptimal calibration. For visual systems, filtering out low-quality frames enhanced performance. Audio-visual fusion yielded substantial gains, improving over audio-only by 85% and video-only by 34-40%.

[7] https://github.com/deepinsight/insightface/tree/master/model_zoo

[8] http://shuoyang1213.me/WIDERFACE/WiderFace_Results.html

7. References

[1] G. R. Doddington, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2-3, pp. 225–254, jun 2000. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(99)00080-1

[2] M. Przybocki, A. F. Martin, and A. N. Le, "NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora - 2004, 2005, 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1951–1959, sep 2007. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=4291612

[3] L. Brandschain, D. Graff, C. Cieri, K. Walker, and C. Caruso, "The Mixer 6 Corpus: Resources for Cross-Channel and Text Independent Speaker Recognition," in *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC10*, Valletta, Malta, may 2010, pp. 2441–2444.

[4] C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey, and J. Hernandez-Cordero, "The 2012 NIST speaker recognition evaluation," in *Interspeech 2013*. ISCA: ISCA, aug 2013, pp. 1971–1975. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2013/greenberg13_interspeech.html

[5] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. Greenberg, D. Reynolds, E. Singer, L. Mason, and J. Hernandez-Cordero, "The 2016 NIST Speaker Recognition Evaluation," in *Interspeech 2017*. ISCA: ISCA, aug 2017, pp. 1353–1357. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2017/sadjadi17_interspeech.html

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

[7] S. O. Sadjadi, C. Greenberg, E. Singer, D. A. Reynolds, L. Mason, and J. Hernandez-cordero, "The 2019 NIST Speaker Recognition Evaluation CTS Challenge," in *Proceedings of Odyssey 2020-The Speaker and Language Recognition Workshop*, Tokyo, Japan, 2020.

[8] S. O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "The 2021 nist speaker recognition evaluation," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 322–329.

[9] Anonymous, "Anonymous title," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH 2019*, Graz, Austria, sep 2019.

[10] ——, "Anonymous title," *Computer Speech & Language*, oct 2019.

[11] ——, "Anonymous title," in *Odyssey 2020 The Speaker and Language Recognition Workshop*. Tokyo, Japan: ISCA, nov 2020.

[12] ——, "Anonymous title," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022.

[13] N. Torgashov, R. Makarov, I. Yakovlev, P. Malov, A. Balykin, and A. Okhotnikov, "The id r&d voxceleb speaker recognition challenge 2023 system description," 2023. [Online]. Available: https://arxiv.org/abs/2308.08294

[14] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A New Multi-Scale Backbone Architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, p. 652–662, Feb 2021.

[15] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[16] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, "Sub-center arcface: Boosting face recognition by large-scale noisy web faces," in *European Conference on Computer Vision*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221341463

[17] S. O. Sadjadi, "NIST SRE CTS Superset: A large-scale dataset for telephony speaker recognition," aug 2021. [Online]. Available: http://arxiv.org/abs/2108.07118

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

[19] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association, INTERSPEECH 2017*. Stockholm, Sweden: ISCA, aug 2017, pp. 999–1003. [Online]. Available: http://www.danielpovey.com/files/2017{_}interspeech{_}embeddings.pdf

[20] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification," in *Interspeech 2021*. ISCA, Aug. 2021, p. 2302–2306. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2021-1570

[21] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech2020*, 2020, pp. 1–5.

[22] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[23] M. Zhao, Y. Ma, Y. Ding, Y. Zheng, M. Liu, and M. Xu, "Multi-query multi-head attention pooling and inter-topk penalty for speaker verification," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6737–6741.

[24] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[25] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[26] Anonymous, "Anonymous title," in *INTERSPEECH 2023*, 2023.