The JHU-MIT System for NIST SRE24: Post-Evaluation Analysis

Jesús Villalba*†, Jonas Borgstrom‡, Prabhav Singh*,
Leibny Paola García*†, Pedro A. Torres-Carrasquillo‡, Najim Dehak*†
**Center for Language and Speech Processing*, *Johns Hopkins University*, Baltimore, MD, USA
†*Human Language Technology Center of Excellence*, *Johns Hopkins University*, Baltimore, MD, USA
‡*MIT Lincoln Laboratory*, MA, USA

***Abstract*—We present the JHU-MIT submission for NIST SRE24, along with post-evaluation analysis and key insights. In the audio fixed condition, our system used Res2Net50 and ResNet100 embeddings; 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 via two-fold cross-validation. For the visual condition, we leveraged pre-trained ResNet100-Subcenter-ArcFace embeddings. Agglomerative clustering was used to diarize speaker and face identities in multi-speaker 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 yielded Cp=0.169, and audio-visual fusion further improved performance, achieving Cp=0.101 (fixed) and Cp=0.087 (open).**

***Index Terms*—Speaker Recognition, NIST, Evaluation, Speaker Embeddings, PLDA, Calibration**

I. 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 focus 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 evolved from telephone speech [2], to far-field microphones [3], [4], then to non-English telephone speech [5]–[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[2] follows a similar setup to SRE21, incorporating audio, visual (face recognition), and multi-modal tracks. However, it introduces new target languages–Tunisian-accented Arabic, French, and English–and permits multi-speaker test segments, thereby requiring speaker and face diarization.

This paper presents the JHU-MIT submission to NIST SRE24. This is a joint effort between JHU and MIT-LL, building on expertise gained from previous evaluations [9]–[12]. We describe the diverse systems developed for this evaluation and provide post-evaluation analyses. These include the impact of different embedding architectures, back-end models, and calibration strategies, as well as how results vary with gender, source, and language conditions. We also examine the complementarity of audio and visual modalities.

For the audio track, we developed systems based on ResNet100 [13], Res2Net [14], and multilingual Wav2Vec2-ECAPA-TDNN embeddings [15]. Due to the lack of in-domain training data, all embeddings were processed using PLDA-based back-ends, which were carefully adapted to the 20 in-domain development speakers via two-fold cross-validation. To address variability across source and language conditions, we employed either condition-dependent preprocessing of embeddings or ensembles of condition-specific PLDA back-ends. For the video track, we used pre-trained face detectors alongside Subcenter-ArcFace embeddings [16]. Additionally, we implemented a quality control mechanism to discard low-quality face embeddings, which led to improved performance compared to previous evaluations [11], [12].

Post-evaluation analysis showed that Wav2Vec2-ECAPA-TDNN was the top-performing single system, with ResNet100 and Res2Net yielding comparable results. Fusion of multiple systems led to moderate performance gains, often mitigating the effects of miscalibration present in individual systems. The analysis also underscored the importance of calibrating using cross-validation scores. Moreover, it highlighted the value of including VoxCeleb in training to prevent performance degradation on AfV conditions. Despite having both AfV and CTS data, we observed substantial degradation in cross-source trials compared to source-matched trials, while performance loss due to cross-language trials was relatively minor.

II. Datasets

A. Train Datasets

The audio track proposed fixed and open training conditions. The fixed data consisted of 638k recordings from 7,251 speakers combining NIST SRE-CTS Superset [17] (large-scale

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Departments, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of Defense.

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

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).

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. The telephone simulation was applied to 25% of AfV recordings and involved downsampling to 8 kHz, applying a bandpass filter with random cut-off frequencies (100–300 Hz low-cut, 3400–3700 Hz high-cut), encoding with one of the following torchaudio[5] codecs: A-law, mu-law, G723.1, or G726, and then upsampling back to 16 kHz. No augmentation was used for back-end training.

B. 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. 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 full trial list.

III. Audio Systems

A. 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 the encoder, we used ResNet100 [13] or Res2Net50 [12], [14]. 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].

B. 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]. The model 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 penalty=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 schedule with a period 2.5k steps) was applied on 8-second chunks for 2.5k steps. This network quickly overfitted to the training data, making it ineffective on the SRE24 set. We conducted several experiments to identify hyperparameters—such as margins, learning rates, and early stopping criteria—that would mitigate overfitting and improve performance on the SRE24 dev set.

C. Language Identification

We used automatic language identification (LID) to label the evaluation data, while ground-truth labels were used for all other datasets. In the fixed condition, we trained an FwSE-ResNet34 LID model using only the fixed training data. For the open condition, we employed a Res2Net50 model trained on the LRE22 open condition, as described in [26]. To classify utterances as English, Arabic, or French, we trained three-class linear Gaussian back-ends using the LID network embeddings extracted from the SRE24 development set.

D. Speaker Diarization

We performed speaker diarization on AfV test recordings using Agglomerative Hierarchical Clustering (AHC) of speaker embeddings, computed from 3-second windows (1-second shift). A PLDA adapted to SRE24 Dev generated the AHC self-similarity matrix, 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' Act DCF by between 0 and 5% relative on the SRE24 Eval (see Table I), which was not a significant gain. Visual inspection

[3] http://www.openslr.org/resources/17
[4] http://www.openslr.org/resources/28
[5] https://pytorch.org/audio
[6] https://huggingface.co/facebook/wav2vec2-xls-r-300m

TABLE I
ABLATION ON DIARIZATION AND CALIBRATION ON SRE24 EVAL

Cond		ResNet100-WB Fixed		W2V2-ECAPA Open	
Diar.	Calib.	Min Cp	Act Cp	Min Cp	Act Cp
N	N	0.638	0.926	0.467	0.471
N	CD-Folds	0.609	0.620	0.389	0.413
Y	N	0.635	0.930	0.453	0.457
Y	CI-Folds	0.635	0.637	0.453	0.457
Y	CI-Cheat	0.635	1.030	0.453	0.746
Y	CI-Mix	0.635	0.745	0.453	0.522
Y	CD-Folds	**0.608**	**0.623**	**0.374**	**0.395**
Y	CD-Cheat	0.612	1.120	0.379	1.04
Y	CD-Mix	0.612	0.709	0.377	0.457

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.

E. JHU Back-end

JHU back-end pipeline followed JHU-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 ($\boldsymbol{\mu}_{\mathrm{CTS}}, \mathbf{S}_{\mathrm{CTS}}$) and AfV ($\boldsymbol{\mu}_{\mathrm{AFV}}, \mathbf{S}_{\mathrm{AFV}}$), then adapted them per in-domain language. For SRE21, the in-domain data consisted of NIST SRE21 Eval and CMN/YUE speakers in the SRE-CTS Superset. For SRE24, in-domain data included SRE24 Dev, adding SRE18-19 in the open condition. Thus, we obtained six adapted mean–covariance pairs $\{\boldsymbol{\mu}_{a-b}, \mathbf{S}_{a-b} | a \in \{\mathrm{AFV}, \mathrm{CTS}\}, b \in \{\mathrm{ENG}, \mathrm{CMN}, \mathrm{YUE}\}/\{\mathrm{ENG}, \mathrm{ARA}, \mathrm{FRA}\}\}$. The adapted means were used to center the in-domain data. The rest of the training data was centered using $\boldsymbol{\mu}_{\mathrm{CTS}}$ and $\boldsymbol{\mu}_{\mathrm{AFV}}$. Next, joint PCA dimensionality reduction/whitening was computed from the average of adapted covariances,

$$\mathbf{S} = \frac{1}{6} \sum_{a \in \{\mathrm{AFV}, \mathrm{CTS}\}} \sum_{\substack{b \in \{\mathrm{ENG}, \mathrm{CMN}, \mathrm{YUE}\}/ \\ \{\mathrm{ENG}, \mathrm{ARA}, \mathrm{FRA}\}}} \mathbf{S}_{a-b} \;, \tag{1}$$

ensuring balanced condition weighting. The same projection was applied to all data, followed by length normalization. Finally, we trained an SPLDA model on all out-of-domain data and adapted it to the in-domain data.

Back-end hyperparameters were tuned based on the test (SRE21 or SRE24) and training setups (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 hyperparameters were then used to train a final back-end on the full SRE24 Dev, which we also evaluated on SRE24 Dev (denoted as *cheat* scores) and SRE24 Eval. This resulted in three back-ends (*fold0*, *fold1*, and *cheat*). SRE24 hyperparameters were selected based on *fold* scores.

F. MIT-LL Back-end

To handle diverse evaluation conditions, the MIT-LL 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 score vectors were modeled by Gaussian mixture models with 2-3 components and shared covariances. The final trial scores were computed as the log-likelihood ratio

$$s = \log \frac{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_T, \boldsymbol{\Sigma})}{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_N, \boldsymbol{\Sigma})} \tag{2}$$

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

G. Calibration and Fusion

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

$$\mathbf{l} = \begin{bmatrix} \text{language-match=Y} \\ \text{language-match=N} \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} \text{source-match=Y} \\ \text{source-match=N} \end{bmatrix} \tag{3}$$

are 1-hot vectors that indicate the language, and source conditions, respectively; and $\mathbf{w}_l$, $\mathbf{w}_c$ are trainable weights representing condition-dependent biases.

We trained three of these calibrations on separate score sets (*fold0*, *fold1*, and *cheat*) but found they did not generalize well across sets, e.g., *fold0* calibration was not good for *fold1* or *cheat*, so we were uncertain about the best calibration set for the eval. Uncertain about the best calibration for evaluation, we implemented a mixture of calibration functions. To this end, we trained a six-component GMM on each non-calibrated score set, reserving two Gaussians to model the target score distribution and four for non-target. Denoting these GMMs as $p(s|\text{fold0})$, $p(s|\text{fold1})$ and $p(s|\text{cheat})$, the final score was

$$\mathrm{LLR} = \sum_{t \in \{\text{fold0}, \text{fold1}, \text{cheat}\}} p(t|s) f_t(s) \;, \tag{4}$$

where s is the uncalibrated score, and f_t are the calibration functions trained on each score set.

In the post-evaluation analysis, we found that the Dev *cheat* score distribution did not align with the Eval, leading to very high Act. Cp. The proposed calibration mixture allowed us to keep a reasonable Actual Cp. However, the best approach would have been calibrating solely on the pooled *fold* scores. Table I presents a comparison of different calibration strategies on the SRE24 Eval set, including condition-independent and condition-dependent variants, using *cheat*, *fold*, or mixed calibration scores. For the Wav2Vec2+ECAPA-TDNN system,

TABLE II
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	Calib.	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	JHU	Eval	3.97	0.351	0.405	8.31	0.617	0.638	3.01	0.372	0.520	7.81	0.632	0.681
1f-p			Post				8.23	0.610	0.617	3.09	0.381	0.587	7.76	0.632	0.635
2f	FwSE-Res2Net50-NB	JHU	Eval	**2.95**	**0.297**	0.404	8.26	0.641	0.655	3.30	0.477	0.593	7.94	0.631	0.652
2f-p			Post				8.32	0.648	0.650	3.41	0.489	0.658	7.94	0.628	0.629
3f	FwSE-ReNet100-WB	JHU	Eval	4.18	0.369	0.402	**7.55**	**0.556**	**0.568**	**2.25**	0.322	**0.421**	7.13	0.612	0.709
3f-p			Post				7.53	**0.563**	**0.567**	**2.26**	0.326	**0.485**	7.09	0.608	0.623
4f	FwSE-ResNet100-NB	JHU	Eval	3.47	0.326	**0.368**	7.58	0.568	0.589	2.33	**0.319**	0.446	**6.90**	**0.569**	**0.613**
4f-p			Post				**7.51**	0.571	0.575	2.37	**0.324**	0.518	**6.85**	**0.565**	**0.566**
5f	FwSE-ReNet100-WB	MIT-LL	Eval	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	MIT-LL	Eval	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	MIT-LL	Eval	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	MIT-LL	Eval	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	JHU	Eval	3.50	0.322	0.404	**5.08**	**0.319**	**0.321**	2.56	0.244	0.376	**4.42**	**0.377**	**0.457**
1o-p			Post				**5.03**	**0.327**	**0.333**	2.59	**0.264**	**0.416**	**4.31**	**0.374**	**0.395**
2o	FwSE-Res2Net50-WB-Std	JHU	Eval	2.67	0.314	0.384	6.15	0.416	0.419	1.61	**0.198**	0.322	5.33	0.500	0.668
2o-p			Post				6.46	0.471	0.484	2.86	0.352	0.624	5.22	0.485	0.486
3o	FwSE-Res2Net50-WB-NoCodec	JHU	Eval	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	JHU	Eval	3.55	0.379	0.423	6.27	0.419	0.422	2.45	0.302	0.544	5.04	0.446	0.473
4o-p			Post				6.17	0.424	0.442	**2.53**	0.309	0.600	4.88	0.444	0.445
5o	Res2Net50-SRE21	JHU	Eval	1.92	0.260	0.332	6.03	0.433	0.434	2.68	0.331	0.599	5.29	0.450	**0.456**
5o-p			Post				5.73	0.428	0.441	3.01	0.357	0.628	5.04	0.444	0.445
6o	ECAPA-TDNN-SRE21	JHU	Eval	2.64	0.329	0.386	7.60	0.505	0.507	4.37	0.407	0.644	6.65	0.555	0.559
6o-p			Post				7.46	0.511	0.524	4.73	0.433	0.671	6.51	0.551	0.553
7o	FwSE-ResNet100-WB	MIT-LL	Eval	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	MIT-LL	Eval	**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
Primary-Post: 1f-p+2f-p+3f-p+4f-p							6.19	0.499	0.505	1.58	0.257	0.436	**5.85**	**0.530**	**0.541**
Single-Post: 3f-p							7.53	0.563	0.567	2.26	0.326	0.485	7.09	0.608	0.623
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
Primary-Post: 1o-p+2o-p+4o-p+5o-p							4.16	0.260	0.264	1.50	0.187	0.418	**3.37**	**0.308**	**0.331**
Single-Post: 1o-p							5.03	0.327	0.333	2.59	0.264	0.416	4.31	0.374	0.395

calibration with *cheat* scores resulted in an Actual Cp of 1.04, whereas calibrating on *fold* scores reduced it substantially to 0.395, and the mixture approach yielded 0.457. The results also show that condition-dependent calibration improved Actual Cp across individual systems on SRE24 Eval by 2–13% relative, compared to condition-independent calibration.

Fusion was trained using calibrated scores from SRE24 Dev *fold0* and *fold1* through a greedy selection strategy [9], [12]. We first calibrated all individual systems and selected the one with the lowest actual cost. Then, additional systems were iteratively added, each time choosing the combination that yielded the best performance. Fusion training was performed at $P_{\mathcal{T}} = 0.01$, with system selection guided by the average Actual DCF computed at $P_{\mathcal{T}} = 0.01$ and $P_{\mathcal{T}} = 0.005$.

H. Audio Submissions

Table II summarizes the results for our single systems and submissions under fixed and open conditions. EER and Cprimary are equalized across common conditions (Gender(M/F), Source-Match (Y/N), Language-Match (Y/N)) following the NIST SRE primary metric. This ensures all conditions contribute equally to the metrics. At evaluation time, the systems based on JHU back-end used the Mixture of calibrations described in Section III-G. At post-evaluation (systems denoted as xx-p), we calibrated the system only on the *fold* cross-validation scores. This post-eval calibration consistently improved Act Cp for all single systems in both fixed and open conditions. The Primary and Contrastive submissions indicate which systems were included in the fusion and the order in which they were selected by the greedy algorithm. We report both the original submissions and the hypothetical ones using the post-eval systems.

In the fixed condition, narrowband (NB) models slightly outperformed wideband (WB) models on average, with an Act Cp of 0.626 compared to 0.677 (7% relative difference). ResNet100 models performed comparably to Res2Nets, showing only a 2% relative difference. Notably, the best single system on Eval was ResNet100-NB, outperforming ResNet100-WB—the best on Dev—by 14%. Primary and Contrastive fusions further reduced Act Cp by 19% and 20%, respectively,

TABLE III
GENDER, SOURCE AND LANGUAGE ABLATION ON SRE24 EVAL

	Primary-Post Fixed			Primary-Post Open		
Cond.	EER	Min Cp	Act Cp	EER	Min Cp	Act Cp
Global	5.85	0.530	0.541	3.37	0.308	0.331
Male	**5.60**	0.517	0.588	**3.31**	**0.301**	0.341
Female	**5.60**	**0.471**	**0.495**	3.44	0.305	**0.320**
Source-Match	**4.68**	**0.388**	**0.439**	**2.98**	**0.237**	**0.269**
Source-Mismatch	6.51	0.599	0.644	3.77	0.369	0.393
Lang-Match	**5.37**	**0.471**	**0.527**	**3.16**	**0.279**	**0.311**
Lang-Mismatch	5.83	0.517	0.556	3.59	0.327	0.350

helping correct miscalibration in single systems and narrowing the gap between Min and Act Cp. The post-eval Primary achieved a 5% relative improvement over the original Primary, benefiting from better-calibrated single systems.

In the open condition, models trained on SRE21 performed as well as, or better than, newer models. No significant differences were observed between networks trained with or without codec augmentation. Wav2Vec2+ECAPA-TDNN achieved the lowest Min Cp, though slight miscalibration placed its Act Cp close to that of ResNet100 and Res2Net50. However, with post-eval calibration, Wav2Vec2+ECAPA-TDNN outperformed the best ResNet100 and Res2Net models by 11% relative. The Primary fusion improved Act Cp by 20% over the best single system. After post-eval calibration, this gain was reduced to 16%. The Primary open system outperformed the Primary fixed system by 36% (Eval) to 38% (Post-eval). This result suggests that fixed condition performance was limited by the scarcity of AfV data, which was available only from SRE21. These findings underscore our continued reliance on VoxCeleb data to achieve strong performance in AfV scenarios.

I. *Analysis of Gender, Source and Language*

Table III presents an ablation study on the impact of Gender, Source, and Language on the results of the post-eval Primary systems. Metrics are equalized across SRE common conditions, e.g., to compute the gender results, we equalize the weights of source-match/mismatch and language-match/mismatch to ensure the result is independent of the proportion of trials of type. The table shows that, despite improved post-eval calibration, calibration gaps remain across common conditions. These results suggest that further gains could be achieved by narrowing these gaps through more effective condition-dependent calibration.

Regarding gender, male and female had similar EER. However, female trials were significantly better in terms of Act Cp (15% in fixed, 6% in open), partly due to a larger calibration gap in male trials. As for source, the performance gap between source-match and mismatch trials was the largest among all factors, exceeding the differences between genders and between language-match/mismatch. This trend appears not only in the fixed condition, which has limited AfV training data, but also in the open condition, which includes balanced CTS and AfV training. In the fixed setup, source-mismatch trials had an Act Cp 47% higher than source-match; in the open, the gap was 46%. Thus, adding VoxCeleb data improved performance for both source conditions but did not reduce the gap. Regarding language, mismatched trials increased Act Cp by only 5–12%, which is substantially smaller than the source mismatch effect and comparable to the gender-based difference.

IV. VISUAL SYSTEMS

A. *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].

B. *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, resulting in relative gains 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* refers to the proportion of black pixels in the cropped image. These black pixels typically appear along the image borders when the user is not facing the camera or when facial landmarks are poorly detected. 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.

C. *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}, assuming clusters correspond to different individuals or face orientations. Finally, we scored each enrollment embedding against all cluster centers in the test video and selected the maximum score.

D. *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 $t_{\text{AHC}} \in \{0.6, 0.7\}$.

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

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

TABLE IV
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

TABLE V
SYSTEMS RESULTS ON SRE24 AUDIO-VISUAL

System	SRE24 AV Dev			SRE24 AV Eval		
	EER	Min Cp	Act Cp	EER	Min Cp	Act Cp
Fixed Single						
Audio Single	2.77	0.322	0.483	8.96	0.738	0.772
Visual Single	1.56	0.058	0.105	2.10	0.152	0.170
AV Single	0.63	0.025	0.037	**1.32**	**0.112**	**0.113**
Fixed Primary						
Audio Primary	1.99	0.217	0.429	6.99	0.638	0.654
Visual Primary	1.47	0.050	0.105	2.07	0.149	0.169
AV Primary	0.59	0.014	0.030	**1.13**	**0.100**	0.101
AV Contrastive	0.59	0.015	0.030	**1.13**	**0.100**	**0.100**
Open Single						
Audio Single	4.36	0.395	0.55	6.15	0.508	0.540
Visual Single	1.56	0.058	0.105	2.10	0.152	0.170
AV Single	0.42	0.028	0.069	**1.00**	**0.095**	**0.098**
Open Primary						
Audio Primary	2.07	0.257	0.571	4.18	0.389	0.415
Visual Primary	1.47	0.050	0.105	2.07	0.149	0.169
AV Primary	0.27	0.010	0.061	**0.83**	**0.086**	**0.087**
AV Contrastive	0.26	0.011	0.069	0.84	0.087	0.089

E. Visual Submissions and Results

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

V. 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 Primary visual systems, while the Single submissions combined Single audio and visual systems.

Table V presents results comparing single-modality Post-Eval Single and Primary systems to Audio-Visual Single and Primary fusions. Note that in the Audio-only modality, NIST primary metrics were calculated using equalized source-Match and mismatch Trials. In contrast, for the Audio-Visual modality, the NIST primary metric includes only source-mismatch trials and excludes source-matched trials. This explains why the Audio system results shown in Table V appear worse than those reported in Table II for the Audio modality.

In the fixed and open conditions, AV Single improved Act Cp by 85% and 82% over the Audio-only Single, and by 34% and 42% over Visual-only Single. Similarly, AV Primary improved Act Cp by 83% and 82% over Audio-only Primary, and by 40% and 42% over Visual-only Primary. These results highlight the complementarity between modalities. Despite the visual modality outperforming the audio modality, it still benefited significantly from integrating audio information.

VI. CONCLUSION AND DISCUSSION

We presented the JHU-MIT systems for NIST SRE24. For the audio fixed condition, 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 various evaluation sub-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 key lessons from this evaluation. ResNet100 and Res2Net performed comparably, with narrowband models slightly outperforming wideband in the fixed condition. VoxCeleb data remained essential for strong AfV performance–its absence severely degraded fixed-condition results. Source-mismatch trials caused the most severe performance drop, even when AfV data was included in training, while gender and language mismatches had a smaller impact. Although AfV data in the open condition improved both source-matched and mismatched trials, the performance gap between them remained nearly unchanged. 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 performance, while source- and language-dependent calibration proved beneficial. Cross-validation scores provided more reliable calibration than *cheating* scores (from back-end trained on full SRE24 Dev) or mixture of calibrations, with condition-dependent strategies reducing Act Cp by up to 13% relative. Audio fusion improved results by approximately 20% and mitigated the effects of sub-optimal calibration. Visual performance was further improved by filtering low-quality frames. Audio-visual fusion yielded substantial gains, improving over the audio-only system by 85% and the video-only system by 34-40%.

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] 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. Garcia-Perera, D. Povey, P. 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 *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH 2019*, Graz, Austria, sep 2019.

[10] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, "State-of-the-art Speaker Recognition with Neural Network Embeddings in NIST SRE18 and Speakers In The Wild Evaluations," *Computer Speech & Language*, p. 101026, oct 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0885230819302700

[11] J. Villalba, D. Garcia-Romero, N. Chen, G. Sell, J. Borgstrom, A. McCree, L. P. Garcia Perera, S. Kataria, P. S. Nidadavolu, P. Torres-Carrasquiilo, and N. Dehak, "Advances in Speaker Recognition for Telephone and Audio-Visual Data: the JHU-MIT Submission for NIST SRE19," in *Odyssey 2020 The Speaker and Language Recognition Workshop*. Tokyo, Japan: ISCA, nov 2020, pp. 273–280. [Online]. Available: http://www.isca-speech.org/archive/Odyssey_2020/abstracts/88.html

[12] J. Villalba, B. J. Borgstrom, S. Kataria, M. Rybicka, C. D. Castillo, J. Cho, L. P. García-Perera, P. A. Torres-Carrasquillo, and N. Dehak, "Advances in cross-lingual and cross-source audio-visual speaker recognition: The jhu-mit system for nist sre21," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 213–220.

[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] J. Villalba, J. Borgstrom, M. Jahan, S. Kataria, L. P. Garcia, P. Torres-Carrasquillo, and N. Dehak, "Advances in language recognition in low resource african languages: The jhu-mit submission for nist lre22," in *INTERSPEECH 2023*, 2023, pp. 521–525.