ABC SYSTEM DESCRIPTION FOR NIST SRE 2024

Jahangir Alam[3], *Sara Barahona*[4], *Dominik Boboš*[5], *Lukáš Burget*[1], *Sandro Cumani*[2], *Mohamed Dahmane*[3], *Jiangyu Han*[1], *Miroslav Hlavaček*[5], *Martin Kodovsky*[5], *Federico Landini*[1], *Ladislav Mošner*[1], *Petr Palka*[1], *Tomaš Pavliček*[5], *Junyi Peng*[1], *Oldřich Plchot*[1], *Gnana Praveen Rajasekhar*[3], *Johan Rohdin*[1], *Anna Silnova*[1], *Themos Stafylakis*[6], *Lin Zhang*[1]

[1]Brno University of Technology, Speech@FIT, Brno, Czechia
[2]Politecnico di Torino, Turin, Italy
[3]Computer Research Institute of Montreal (CRIM), Montreal (Quebec) Canada
[4]Universidad Autonoma de Madrid, Madrid, Spain
[5]Phonexia, Brno, Czechia
[6]Omilia - Conversational Intelligence, Athens, Greece

ABSTRACT

This paper presents the ABC team's submission to the NIST SRE 2024 evaluation, a collaboration among BUT, Polito, Phonexia, Omilia, UAM, and CRIM. Our team participated in all evaluation tracks (audio-only, visual-only, and audio-visual) under both fixed and open conditions. We developed a variety of frontends, backends, and strategies for calibration and fusion to optimize system performance.

The fixed and open conditions share some solutions. In the audio-only systems, we employed ResNet variants and the newly introduced ReDimNet model as frontends for embedding extraction. Then, we explored various backends including cosine scoring, Probabilistic Linear Discriminant Analysis, and Pairwise Support Vector Machine. For the visual-only systems, we adopted the Insightface framework, utilized ResNet100 and MagFace pre-trained on the MS1MV2 dataset. Cosine scoring under various strategies were applied, with logistic regression used for both calibration and fusion. Finally, scores from audio-only and visual-only systems were fused using logistic regression for submission to the audio-visual track. Building on the fixed condition, the open condition included enhancements such as larger ResNet models, additional training data from the VoxBlink2 dataset, and the pre-trained XLS-R foundation model.

1. INTRODUCTION

This submission to NIST SRE 2024 is a collaborative effort of BUT, Polito, Phonexia, Omilia, UAM, and CRIM. We participated in all tracks (audio-only, audio-visual, and visual-only) under both fixed and open conditions. This document describes the submitted systems. First, we provide a detailed description of the fixed track submission. Then, for the open track, we primarily highlight the additions and modifications done to the fixed systems.

2. FIXED CONDITION

2.1. Audio-only systems

2.1.1. Training data and augmentations

For training the system, we used these databases:

- NIST CTS Superset [1] - used for training the embedding extractors and scoring backends
- NIST SRE 2021 evaluation set [2] - used for training scoring backends

We used Kaldi-style [3] augmentation with MUSAN [4] and RIR [5] database, where we excluded Babble noise and Music from MUSAN since these did not abide by the rules of the fixed track. A Kaldi-style energy-based VAD was used to remove silent parts from the waveform.

When training fixed track models, we downsampled all 16kHz data to 8kHz (both for training and evaluation sets).

2.1.2. Development dataset

For monitoring our performance and for both calibration and fusion, we used the official SRE 2024 [6] development dataset provided by NIST and LDC.

2.1.3. Frontend

XI-ResNet-34: To train all of our embedding extractors, we utilized WeSpeaker toolkit [7, 8]. As a first extractor, we used ResNet34 architecture [9], but in place of the common statistical pooling, we utilized xi-vector approach [10]. The training followed the VoxCeleb recipe of WeSpeaker for ResNets utilizing all of the suggested hyperparameters. The training minimizes AAM objective [11] with the margin set to 0.2 and scale to 32. For the first 20 epochs of the training the margin is set to 0 and then gradually increased from 0 to 0.2 in a course of 20 epochs. Finally, it is fixed for the rest of the training until epoch 150 is reached. The learning rate scheduler is set as in the original recipe - warming up the learning rate from 0 to its highest value (0.1) for 6 epochs and then exponentially decreasing it to 5e-5 for the rest of the training. The training was performed using segments of 2 seconds duration. After 150 epochs, we increased the length of the training segments to 10 seconds and continued training for 10 more epochs with the fixed margin and learning rate.

XI-ResNet-152: As a second embedding extractor we used a bigger version of XI-ResNet-34. The encoder this time was replaced with ResNet152. When training this extractor, we followed the same WeSpeaker recipe that was used for the smaller version with a few

modifications: we trained the model on longer training segments of 3 seconds, speed perturbation was turned off for this experiment, and instead of running the training for 150 epochs followed by 10 additional epochs with 10s training examples, the first stage this time lasted 130 epochs and the second one 5 epochs.

ReDimNet-B3: We also employed Reshape Dimensions Network (ReDimNet) [12] as a speaker embedding extractor. Specifically, we selected the B3 version based on our empirical experience. The final model was trained in two stages, with short and long segments. The hyperparameters for the first one follow those described for XI-ResNet-34. In the second stage (so-called large-margin fine-tuning), the model was further trained for five epochs on six-second segments, optimizing AAM with a margin increased to 0.5. The learning rate decreased exponentially from 1e-4 to 2.5e-5.

2.1.4. Backend

Different backends were employed for embedding classification. We considered cosine (COS) scoring, Probabilistic Linear Discriminant Analysis (PLDA) [13, 14] with its heavy-tailed variant [15], and Pairwise Support Vector Machine (PSVM) [16, 17]. For all backends, we included pre-processing strategies aimed at mitigating mismatch due to language, gender, and channel differences in the recordings. For our primary submission, we adopted the following steps:

- embeddings are centered and length-normalized
- language labels are employed to estimate a language subspace through Linear Discriminant Analysis (LDA). Embeddings are projected in the complement space found by LDA
- a two-components, tied covariance Gaussian mixture model with uniform weights is estimated using gender labels. For each utterance, centered first order statistics are extracted and used in place of the original embedding for further processing (i.e., we compute a soft-re-centering of the embeddings). The same approach is then repeated to compensate for source mismatch, using the source labels of the training data
- embeddings are projected in a speaker LDA subspace. The dimension of the LDA subspace has been optimized for each backend based on the results on the SRE 2024 development set
- projected embeddings are length-normalized. For the xi-vector frontends, the length-normalization step is preceded and followed[1] by Within-Class Covariance Normalization (WCCN). This step was not included for other frontends.

The pre-processing pipeline was trained using long segments obtained by concatenating the CTS superset audio segments coming from the same session, together with enrollment segments of the SRE 2021 evaluation set.

Cosine backend: The cosine backend consists of a simple dot-product of length-normalized embeddings.

PLDA backend: The PLDA backends were trained using the original short segments of the CTS Superset, with the addition of the SRE 2021 evaluation dataset enrollment utterances. We performed (limited) tuning of the PLDA subspace dimension. The models were trained using a standard expectation maximization (EM) algorithm with minimum divergence iterations.

PSVM backend: The PSVM models were trained using the strategy outlined in [17]. The training set consists of long segments obtained by concatenating the CTS Superset audio segments coming from the same session, together with the enrollment segments of the SRE 2021 evaluation set. The PSVM regularizer and the PSVM costs for target and non-target trials were tuned to optimize the performance on the SRE 2024 development data. In addition to [17], we also incorporated side-information at training time. In particular, the PSVM model was trained with embeddings augmented with duration-dependent side-information, encoded as the natural logarithm of the duration of the corresponding utterance. It is worth noting that the resulting scoring function shows strong similarity with the Quality-Measure-aware scoring functional employed in [18] for score calibration. In particular, our approach can be seen as an extension of both the QM_4 method [18] and [19], as that is able to directly estimate and optimize the duration contribution at PSVM training stage, rather than at scoring time.

Despite our efforts in compensating channel mismatches at the embedding level, the verification scores of the different backends show significant intra-condition mismatch that generates a significant intra-condition miscalibration. To address this issue we considered two possible approaches: channel-dependent score normalization and condition-dependent calibration. The former aims at normalizing impostor scores through an adaptive cohort selection that employs source-dependent cohort sets. The latter, on the other hand, employs re-calibration methods to independently re-calibrate the scores of the different conditions with the aim of reducing intra-condition calibration mismatch. For our primary submission, we employed the second strategy (we refer to [20] for an analysis of the limitations of score normalization). Prior-weighted Logistic-Regression (LR) [21] models were trained to estimate an affine transformation for each of the four possible source combinations ("afv-afv", "cts-afv", "afv-cts" and "cts-cts", where afv stands for audio from video and cts stands for conversational telephone speech). To reduce the risk of over-fitting, these models were trained using the SRE 2021 evaluation dataset, allowing us to obtain a significant reduction of minimum primary cost, despite the resulting score not being necessarily globally calibrated for the SRE 2024 development set.

For the contrastive submission, we utilized the same frontends as the primary one but used different embedding pre-processing and scoring. In all cases, we used Nuisance attribute projection (NAP) [22, 23] to remove the direction corresponding to speaker gender, followed by another NAP to project out the direction the source variability (cts vs. afv, where the mean of cts was estimated on CTS Superset and mean of the afv was estimated on afv part of NIST SRE 2021). NAP was followed by centering the embeddings, reducing the dimensionality by LDA and length normalization. The scoring in all cases was a simple cosine similarity. Finally, the channel normalization [24] for the scores was applied to remove intra-condition mismatch as discussed above.

2.1.5. Calibration and fusion

We addressed global miscalibration using generative calibration based on the Variance-Gamma approach [25], which models the distribution of the verification scores in terms of Variance-Gamma

[1] The second WCCN is not strictly required for the PLDA backend, but it allows for faster convergence. For PSVM it affects regularization, allowing us to further improve performance for the xi-vector frontends.

densities whose parameters represent "effective" variances of the training and evaluation population. To train the parameters, we employed a Quasi-Newton (QN)-accelerated EM algorithm [26]. The parameters were initialized from a linear Variance-Gamma [26] model, also trained using the QN-EM approach. In order to better capture the score dynamics, duration side-information was included in the model [25]. Due to time constraints, the calibration models were only trained on a set of approximately 10,000 target and 10,000 nontarget trials extracted from the SRE 2024 development set.

The final primary system consists of a combination of different frontends and backends. To reduce the number of combinations and the risk of over-fitting, we performed a pre-fusion step that combines the scores of different backends for the same frontend. For each frontend, we selected the optimal backend combination based on the results on the SRE 2024 development set. The fusion is based on prior-weighted logistic regression (LR) [21], with a scalar per system and a single bias term. Our primary submission consists of the score-level fusion of the resulting scores of each frontend, again obtained by means of prior-weighted LR. Fusion weights were estimated on a subset of the SRE 2024 development trials. The LR prior was set to 0.01.

For the contrastive submission, all three systems were precalibrated on SRE 2024 dev using LR and then fused on the same set.

2.1.6. Audio systems results

The results on the SRE 2024 development set for the selected frontend/backend combinations, as well as for the partial backend fusions and the primary submission, are shown in Table 1. Among frontends, two XI-ResNet frontends provide similar results, while the ReDimNet model is slightly less effective. Concerning backends, different approaches tend to provide similar results in terms of primary metric, with PLDA being on average slightly less effective, and PSVM providing significantly lower EER. When focusing on individual frontends, we typically also observe notable improvement from backend fusions, despite the backend models sharing the same input embeddings. Further combination of the scores provided by each frontend provides an additional improvement for all considered metrics. In general, although the results may be biased since the calibration models were trained using the same speakers and a subset of the evaluated scores, we observe that the models demonstrate good calibration.

2.2. Visual-only systems

To develop a visual-only speaker verification system, we leveraged face recognition models pre-trained on the MS1MV2 [11] and Glint360k [27] datasets. The MS1MV2 dataset, widely used for face recognition research, contains approximately 5.8 million facial images from around 87,000 unique identities.

2.2.1. MS1MV2 dataset-based Visual Systems (S1-S5)

In this case, a ResNet100 configuration-based face recognition model, which was pre-trained on the MS1MV2 facial image dataset, was employed for the extraction of visual features (i.e., embeddings) from the SRE 2024 enrollment images and test video recordings. Before the extraction of embeddings, as pre-processing steps, face detection and alignment were performed. For the detection of faces from videos, a pretrained RetinaFace [28] face detection system was used. After detecting faces, we aligned them using the landmarks provided by the Multi-task Cascaded Convolutional Network (MT-CNN) [29]. The face alignment step is crucial to standardize the input faces before they are passed through the pre-trained face recognition model (e.g., ResNet100) for feature extraction. Based on the Insightface framework and adopting different scoring mechanisms we built five visual-only systems denoted as S1, S2, S3, S4, and S5.

- S1: Cosine similarity is performed between median enrollment embeddings and all the test embeddings. The maximum of the scores is taken as the final verification score.
- S2: At first, agglomerative hierarchical clustering is applied on the test embeddings with a stopping threshold of 0.7. The cosine similarity is then performed between median enrollment embeddings and all the cluster centers of the test embeddings. Finally, the maximum of the scores is taken as the final verification score.
- S3: Same as S2 but with a stopping threshold of 0.65.
- S4: Cosine scoring is performed between median enrollment embeddings and the self-attention-based refined embeddings [30]. The maximum of the scores is considered as the final verification score.
- S5: Cosine scoring is performed between average enrollment embeddings and the self-attention-based refined embeddings [30]. The maximum of the scores is considered as the final verification score.

2.2.2. MagFace Visual Systems (S8) pretrained on MS1MV2 dataset

The first step toward building the MagFace-based visual system (S8) involves the detection and alignment of faces from SRE 2024 enrollment images and test videos. Following this, embeddings from the detected faces are extracted. More specifically, for all our experiments, a residual network architecture (ResNet101) [31] was considered for face detection. Then, all the aligned face images underwent spatial and color jittering as a data augmentation step. For frames where no faces were detected, a second pass was performed as a redetection phase after applying sophisticated image enhancing techniques on such images. Afterwards, the embeddings were extracted using a residual neural network architecture based on the MagFace paradigm [32], pretrained on the MS1MV2 dataset.

Each image to be recognized (from the gallery) is represented by a group of centroids obtained by clustering the embeddings of the corresponding augmented images. Each frame of the video undergoes the same pre-processing techniques, resulting in a set of centroids. The cardinality of the frame and image centroids is not necessarily the same. Finally, a cross-group cosine similarity is calculated between the pairs of centroids to determine the similarity between the face in the image and the face present in the frame. For a broader view of the face recognition system, refer to Figure 1. It is worth mentioning that there is no restriction on the number of faces that should be present both in the image and in the video frames.

2.2.3. Calibration and fusion

When submitting individual systems, we calibrated using LR on the video parts of NIST SRE 2024 development set and NIST SRE 2021 evaluation set - the trials from the two sets were pooled together. For fusion, we have opted to use two strategies: for the primary visual submission we used LR fusion of the pre-calibrated scores. For the contrastive submission, we used exactly the same pre-calibrated scores as the primary one, but instead of the trained fusion, we applied a simple average of the scores coming from the two fused systems. The motivation for the latter is our belief that such a fusion

Table 1: Comparison of selected frontends, backends and primary submission on the SRE 2024 development and evaluation sets (fixed condition). The results are given in terms of Equal Error Rate (EER), and NIST-defined primary costs, as detailed in the NIST SRE 2024 evaluation plan [6].

		SRE24 dev			SRE24 eval				
		$C_{primary}$			$C_{primary}$			C_{llr}	
frontend	backend	min.	act.	EER (%)	min.	act.	EER (%)	min.	act.
XI-ResNet-152	PLDA (U 100)	0.523	0.525	8.45	0.602	0.639	8.08	0.291	0.297
	PSVM	0.524	0.532	7.73	0.572	0.574	7.69	0.279	0.287
	Fusion	0.496	0.499	7.52	0.557	0.574	7.29	0.268	0.274
	COS	0.507	0.527	8.18	0.616	0.626	8.77	0.313	0.321
XI-ResNet-34	COS	0.534	0.539	8.26	0.619	0.640	8.97	0.319	0.326
	PLDA (U 100)	0.564	0.568	8.77	0.638	0.662	8.76	0.311	0.313
	PSVM	0.539	0.543	7.47	0.579	0.580	7.56	0.276	0.283
	Fusion	0.503	0.506	7.40	0.562	0.574	7.49	0.274	0.278
ReDimNet-B3	COS	0.524	0.602	9.11	0.637	0.643	9.06	0.326	0.337
	PSVM	0.536	0.539	8.28	0.609	0.640	8.28	0.301	0.311
	Fusion	0.523	0.528	8.36	0.607	0.629	8.37	0.304	0.313
Primary system		0.440	0.446	6.92	0.514	0.530	6.71	0.250	0.256
PSVM sub-systems fusion		0.457	0.464	6.89	0.515	0.521	6.73	0.250	0.257
XI-ResNet-221	PLDA (U 100)	0.536	0.542	8.41	0.628	0.660	8.46	0.305	0.311
	PSVM	0.582	0.604	7.99	0.607	0.619	8.35	0.300	0.308
	Fusion	0.509	0.519	7.77	0.598	0.614	7.80	0.285	0.292
	COS	0.592	0.592	8.89	0.625	0.629	9.22	0.328	0.330
Primary system + XI-ResNet-221		0.426	0.432	6.87	0.512	0.532	6.70	0.249	0.255

strategy is less prone to overfitting and can be safer in a situation of extremely low amount of errors made by the visual systems.

2.3. Audio-visual systems

The audio-visual primary system is based on the LR fusion of the same audio sub-systems and visual sub-systems that were employed for the primary audio-only and visual-only fusions. The fusion weights were estimated using a subset of the SRE 2024 development set audio-visual trials. For the two single systems, we used pre-calibrated scores fused through LR. In this case, calibration was trained on the respective part of SRE 2024 dev (i.e., these are exactly the systems used in audio-only and visual-only tasks). The fusion is trained on the audio-visual part of the development set.

2.4. Results

Table 2 summarizes the results of the systems submitted for the fixed condition. The first part of the table corresponds to the performance of individual systems used in the submitted fusion. The systems are evaluated on NIST SRE 2024 development set using the official performance metrics of the evaluation.

3. OPEN CONDITION

3.1. Audio systems

3.1.1. Training data

In addition to databases used in the fixed condition, for the open condition, we utilize:

- VoxBlink2 dataset [33]
- VoxCeleb2 development set [34]
- NIST SRE 2018 development and evaluation sets [35]

All data were downsampled to 8kHz (both for training and evaluation sets).

3.1.2. Frontend

ResNet-152-VB: We explored training with the recently released VoxBlink2 [33] dataset, which consists of audios from YouTube videos belonging to 111,284 speakers. The original 16kHz data is downsampled to 8kHz, and each recording has a probability of 0.5 of being subjected to GSM codec using Sox[2]. The training procedure follows the VoxCeleb recipe implemented in the WeSpeaker toolkit. For feature extraction, we computed 80-dimensional log Mel-filterbank energy features. The embedding extractor consists

[2]https://sourceforge.net/projects/sox/

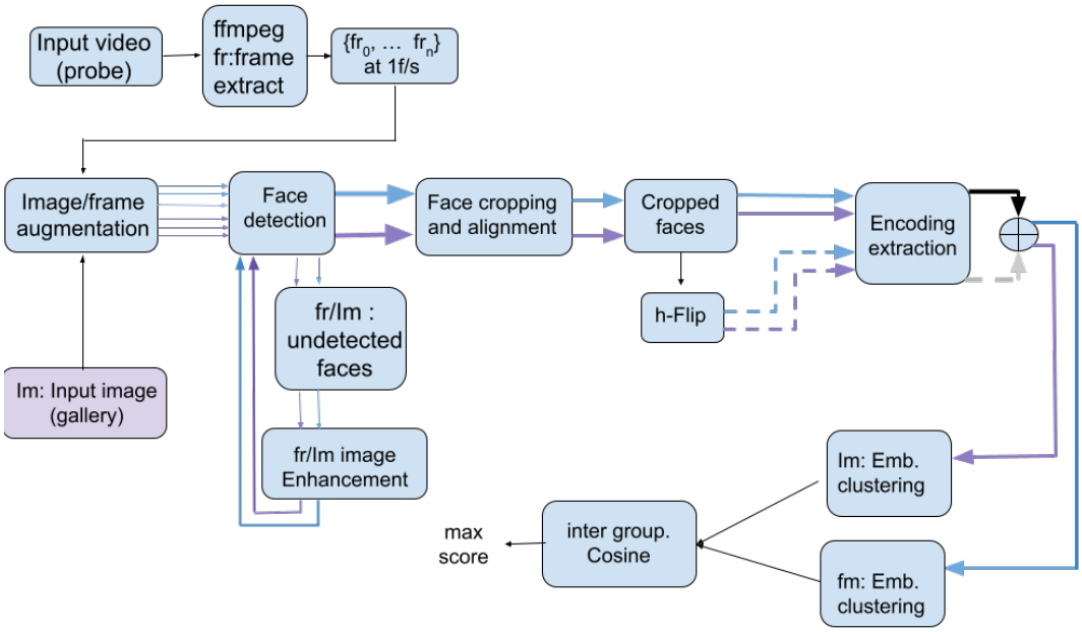

Fig. 1: An overview of the MagFace-based visual-only speaker verification framework.

of a ResNet152 with statistical pooling. It was trained on 2-second segments for 150 epochs using AAM-Softmax loss with a scale factor of 32. The margin parameter is gradually increased from 0 to 0.2 between epochs 20 and 40. We applied an exponential decay to the learning rate after a 6-epoch warm-up, with a maximum rate of 0.1 and a final rate of 5e-5.

Following training on VoxBlink2, we performed Large-Margin Fine-Tuning on the CTS Superset. Specifically, the model was further trained for 10 epochs employing lower learning rates than in the previous phase, starting from 1e-4 and exponentially decreasing it to 2.5e-5. During this phase, we experimented with longer segment durations (6 and 10 seconds) and increase the AAM-Softmax margin to 0.5.

XI-ResNet-221: This embedding extractor is a scaled-up version of the XI-ResNets used in the fixed condition submission. It is trained on the same CTS superset (i.e., this extractor could be used in the fixed track). In this case, the training optimizes AAM-Softmax loss with the same hyperparameters as before, training segments are 3 seconds long, no speed perturbation was used, and, unlike the models used in the fixed condition, there was no fine-tuning of the network on the longer segments.

XLS-R: In the open condition, we also made use of a foundation model pre-trained in a self-supervised way. Considering the multilinguality of pre-training data[3], we opted for XLS-R [36]. A notable advantage of this model is that a subset of pre-training examples is sampled at 8 kHz and contains telephone speech. In the fine-tuning stage, we appended a multi-head factorized attention (MHFA) backend [37] to the pre-trained XLS-R 300M and fine-tuned both components on upsampled CTS Superset recordings, optimizing an AAM-Softmax loss (with a scale of 32 and a margin of 0.2). MHFA comprised 64 heads and produced 256-dimensional embeddings. The learning rate decreased exponentially from 1e-2 to 4.4e-3 over the course of 30 epochs. The pre-trained weights of XLS-R were updated using a learning rate scaled down by a factor of 0.08 compared to MHFA.

[3]Pre-training data comprised VoxPopuli, Multilingual Librispeech, CommonVoice, VoxLingua107, BABEL.

3.1.3. Backend

As a backend for the XLS-R based system, we followed the same embedding preprocessing steps used in the contrastive submission for the fixed condition: NAP to remove gender and source variability, centering, LDA and length normalization. We trained two PLDA models on the preprocessed embeddings: one on the CTS Superset and another one on the SRE 2018 evaluation set. The final scoring was performed with the PLDA which was the interpolation of the two. No score normalization was applied in this case.

For the other frontends, we employed the same strategies as detailed in the previous section. In this case, however, we also employed additional VoxCeleb data for training the backend classifiers and the last stage of our embeddings preprocessing (Section 2.1.4).

3.1.4. Calibration and fusion

We follow the same strategy as in the fixed condition. Details are given in Section 2.2.3. Results for individual systems and frontend-level fusion are given in Table 3.

3.2. Visual and audio-visual systems

For the open condition, we have used an additional visual system that was not ready before the fixed track deadline, system S8 described in Section 2.2.2 - which we now used as "Single Best" submission. The fusion follows the same strategy as for the fixed condition.

Table 2: Results of the systems for the NIST SRE 2024 Fixed Condition. AUDIO, VISUAL, and AV systems were evaluated on the respective trial lists. Audio systems 1-3 each are the fusion of multiple backends with a single frontend. The details on the individual frontend-backend combinations are given in Table 1. Systems 4-6 are the same frontend models, with different pre-processing and a single cosine scoring backend. Numbers in the brackets correspond to the post-evaluation analysis when a bug in scoring eval trials was fixed.

	System	SRE24 dev			SRE24 eval		
		$C_{primary}$ min.	$C_{primary}$ act.	EER (%)	$C_{primary}$ min.	$C_{primary}$ act.	EER (%)
	AUDIO						
1	XI-ResNet-34 fusion	0.503	0.506	7.40	0.562	0.574	7.49
2	XI-ResNet-152 fusion	0.496	0.499	7.52	0.557	0.574	7.29
3	ReDimNet-B3 fusion	0.523	0.528	8.36	0.607	0.629	8.37
4	XI-ResNet-34 cos	0.558	0.577	10.14	0.705 (0.685)	1.0 (0.721)	11.12 (10.48)
5	XI-ResNet-152 cos	0.540	0.558	9.30	0.679 (0.652)	1.0 (0.679)	10.08 (9.43)
6	ReDimNet-B3 cos	0.587	0.598	9.13	0.689 (0.674)	1.0 (0.695)	9.89 (9.48)
	VISUAL						
7	S1	0.080	0.153	1.56	0.163	0.208	2.56
8	S2	0.077	0.130	1.60	0.156	0.182	2.25
9	S3	0.050	0.125	1.56	0.156	0.180	2.17
10	S4	0.037	0.149	2.23	0.156	0.190	2.51
11	S5	0.059	0.179	2.38	0.169	0.211	2.99
	Primary AUDIO Fusion = LR 1+2+3	0.440	0.446	6.92	0.514	0.530	6.71
	Contrastive AUDIO Fusion = LR 4+5+6	0.491	0.502	8.61	0.637 (0.600)	1.0 (0.636)	8.95 (8.39)
	Single Best AUDIO = XI-ResNet-152 PSVM	0.524	0.532	7.73	0.572	0.574	7.69
	Primary VISUAL Fusion = LR 7+9	0.059	0.122	1.76	0.158	0.179	2.20
	Contrastive VISUAL Fusion = AVG 7+9	0.054	0.137	1.56	0.157	0.189	2.36
	Single Best VISUAL = 9	0.050	0.125	1.56	0.156	0.180	2.17
	Primary AV Fusion = LR 1+2+3+7+...+11	0.035	0.038	0.72	0.105	0.106	1.20
	Single Best AV = LR 5+9	0.047	0.051	0.82	0.136 (0.113)	0.137 (0.121)	2.06 (1.50)
	Single Best AV = LR 2+9	0.054	0.063	0.87	0.109	0.111	1.35

3.3. Results

Table 4 shows the results of the individual subsystems and fusions submitted to the open condition.

4. CPU AND MEMORY USAGE DETAILS OF THE SUBMISSIONS

5. REFERENCES

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

[2] 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.

[3] 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.

[4] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[5] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[6] "NIST 2024 Speaker Recognition Evaluation Plan," 2024.

[7] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] S. Wang, Z. Chen, B. Han, H. Wang, C. Liang, B. Zhang, X. Xiang, W. Ding, J. Rohdin, A. Silnova, et al., "Advancing speaker embedding learning: Wespeaker toolkit for research and production," *Speech Communication*, vol. 162, pp. 103104, 2024.

[9] 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.

[10] K. A. Lee, Q. Wang, and T. Koshinaka, "Xi-vector embedding for speaker recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 1385–1389, 2021.

[11] 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.

[12] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov,

Table 3: Comparison of selected frontends, backends and primary submission on the SRE 2024 development and evaluation sets (open condition). The results are given in terms of Equal Error Rate (EER), and NIST-defined primary costs, as detailed in the NIST SRE 2024 evaluation plan [6].

		SRE24 dev			SRE24 eval				
		$C_{primary}$			$C_{primary}$			C_{llr}	
frontend	backend	min.	act.	EER (%)	min.	act.	EER (%)	min.	act.
ResNet--152-VB	PLDA (U 100)	0.470	0.479	7.46	0.528	0.571	7.26	0.268	0.280
	HT-PLDA	0.466	0.468	7.53	0.562	0.622	6.83	0.256	0.267
	PSVM	0.372	0.373	6.02	0.397	0.399	4.97	0.195	0.203
	Fusion	0.358	0.361	5.90	0.402	0.407	4.77	0.190	0.195
	COS	0.473	0.478	7.12	0.490	0.506	6.50	0.246	0.254
XI-ResNet-221	PLDA (U 100)	0.522	0.526	7.74	0.603	0.625	8.15	0.295	0.300
	PSVM	0.512	0.515	7.52	0.563	0.564	7.20	0.266	0.272
	Fusion	0.469	0.471	6.84	0.533	0.538	6.63	0.251	0.255
ReDimNet-B3	COS	0.536	0.538	8.99	0.648	0.701	9.42	0.336	0.346
	PSVM	0.507	0.509	7.92	0.557	0.565	7.04	0.263	0.273
	Fusion	0.479	482	7.72	0.543	0.557	7.24	0.269	0.276
XLS-R	Multi-PLDA	0.579	0.580	10.48	0.584	0.586	9.18	0.324	0.326
Primary system		0.321	0.324	5.60	0.380	0.387	4.49	0.180	0.185
PSVM sub-systems fusion		0.341	0.345	5.84	0.384	0.388	4.79	0.189	0.196
PSVM sub-systems fusion + XLS-R PLDA		0.323	0.327	5.70	0.369	0.371	4.53	0.179	0.185
XLS-R	PSVM	0.502	0.508	8.02	0.562	0.566	7.53	0.276	0.281
	Multi-PLDA + PSVM	0.478	0.483	7.92	0.519	0.521	7.05	0.259	0.262
Primary system with XLS-R fusion		0.309	0.312	5.46	0.370	0.379	4.48	0.177	0.183
ResNet--152-VB larger cal. set	PLDA (U 100)	0.470	0.472	7.47	0.528	0.552	7.29	0.269	0.280
	HT-PLDA	0.466	0.470	7.53	0.562	0.595	6.83	0.256	0.267
	PSVM	0.372	0.377	6.02	0.397	0.397	4.97	0.195	0.203
	Fusion	0.360	0.361	5.90	0.402	0.407	4.81	0.190	0.195

and N. Torgashov, "Reshape Dimensions Network for Speaker Recognition," in *Interspeech 2024*, 2024, pp. 3235–3239.

[13] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of the 9th European Conference on Computer Vision*, 2006, vol. Part IV of *ECCV'06*, pp. 531–542.

[14] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *in Proceedings of Odyssey*, June 2010.

[15] N. Brummer, A. Silnova, L. Burget, and T. Stafylakis, "Gaussian meta-embeddings for efficient scoring of a heavy-tailed plda model," in *Proceedings of Odyssey 2018*, 06 2018, pp. 349–356.

[16] S. Cumani, N. Brümmer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[17] S. Cumani and P. Laface, "Large scale training of Pairwise Support Vector Machines for speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1590–1600, 2014.

[18] M. I. Mandasari, R. Saeidi, and D. A. van Leeuwen, "Quality measures based calibration with duration and noise dependency for speaker recognition," *Speech Communication*, vol. 72, pp. 126–137, 2015.

[19] L. Ferrer, M. McLaren, and N. Brümmer, "A speaker verification backend with robust performance across conditions," *Computer Speech & Language*, vol. 71, pp. 101258, 2022.

[20] S. Cumani and S. Sarni, "Impostor score statistics as quality measures for the calibration of speaker verification systems," in *Proceedings of Odyssey 2022*, 2022.

[21] N. Brümmer and J. A. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[22] 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.

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

Table 4: Results of the systems for the NIST SRE 2024 open condition. AUDIO, VISUAL, and AV systems were evaluated on the respective trial lists. Audio systems 1-3 each are the fusion of multiple backends with a single frontend. The details on the individual frontend-backend combinations are given in Table 3. Visual systems 5-9 are exactly the same ones as used in the fixed condition.

	System	SRE24 dev			SRE24 eval		
		$C_{primary}$ min.	$C_{primary}$ act.	EER (%)	$C_{primary}$ min.	$C_{primary}$ act.	EER (%)
	AUDIO						
1	XI-ResNet-221 open fusion	0.469	0.471	6.84	0.533	0.538	6.63
2	ResNet-152-VB open fusion	0.364	0.367	5.94	0.402	0.407	4.77
3	ReDimNet-B3 open fusion	0.479	0.482	7.72	0.543	0.557	7.24
4	XLS-R Multi-PLDA	0.579	0.580	10.48	0.584	0.586	9.18
	VISUAL						
5	S1	0.080	0.153	1.56	0.163	0.208	2.56
6	S2	0.077	0.130	1.60	0.156	0.182	2.25
7	S3	0.050	0.125	1.56	0.156	0.180	2.17
8	S4	0.037	0.149	2.23	0.156	0.190	2.51
9	S5	0.059	0.179	2.38	0.169	0.211	2.99
10	S8	0.032	0.125	1.11	0.157	0.179	1.86
	Primary AUDIO Fusion = LR 1+2+3+4	0.321	0.324	5.60	0.380	0.387	4.49
	Contrastive AUDIO Fusion = LR 1+2+3	0.333	0.337	5.76	0.388	0.395	4.71
	Single Best AUDIO = ResNet-152-VB PSVM	0.372	0.373	6.02	0.397	0.399	4.97
	Primary VISUAL Fusion = LR 7+10	0.026	0.112	1.11	0.148	0.164	1.80
	Contrastive VISUAL Fusion = AVG 7+10	0.035	0.099	1.11	0.138	0.152	1.79
	Single Best VISUAL = 10	0.032	0.125	1.11	0.157	0.179	1.86
	Primary AV Fusion =LR 1+...+10	0.010	0.011	0.14	0.107	0.266	1.07
	Contrastive AV = AVG Primary AU +Primary VI	0.010	0.011	0.27	0.093	0.173	0.97
	Single Best AV = LR 2 +10	0.011	0.012	0.26	0.095	0.216	1.03

[24] G. Lavrentyeva, S. Novoselov, V. Volokhov, A. Avdeeva, A. Gusev, A. Vinogradova, I. Korsunov, A. Kozlov, T. Pekhovsky, A. Shulipa, et al., "Stc speaker recognition system for the nist sre 2021.," in *Odyssey*, 2022, pp. 354–361.

[25] S. Cumani and S. Sarni, "The distributions of uncalibrated speaker verification scores: A generative model for domain mismatch and trial-dependent calibration," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2204–2219, 2023.

[26] S. Cumani, "On the distribution of speaker verification scores: Generative models for unsupervised calibration," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 547–562, 2021.

[27] 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.

[28] 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.

[29] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[30] J. A. Villalba Lopez, 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 *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 273–280.

[31] 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.

[32] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14225–14234.

[33] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, "Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark," in *Interspeech 2024*, 2024, pp. 4263–4267.

[34] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[35] O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2018 nist speaker recognition evaluation," 2019-09-15 00:09:00 2019, INTERSPEECH, Graz, AT.

[36] 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.

[37] J. Peng, O. Plchot, T. Stafylakis, L. Mošner, L. Burget, and J. Černocký, "An Attention-Based Backend Allowing Efficient Fine-Tuning of Transformer Models for Speaker Verification," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 555–562.