IDIAP RESEARCH REPORT

TEAM SWITZERLAND SUBMISSION TO NIST SRE24 SPEAKER RECOGNITION EVALUATION

Amrutha Prasad Hatef Otroshi Shahreza
Andrés Carofilis Aref Farhadipour Shiran Liu
Srikanth Madikeri Anjith George[a] Petr Motlicek
Sébastien Marcel Masoumeh Chapariniya
Valeriia Perepelytsia Teodora Vukovic
Volker Dellwo

Idiap-RR-10-2025

OCTOBER 2025

[a]Idiap Research Institute

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

TEAM SWITZERLAND SUBMISSION TO NIST SRE24 SPEAKER RECOGNITION EVALUATION

Amrutha Prasad[1,2], *Hatef Otroshi Shahreza*[1], *Andres Carofilis*[1], *Aref Farhadipour*[3], *Shiran Liu*[4], *Srikanth Madikeri*[3], *Anjith George*[1], *Petr Motlicek*[1,2], *Sébastien Marcel*[1] *Masoumeh Chapariniya*[3], *Valeriia Perepelytsia*[3], *Teodora Vukovic*[3], *Volker Dellwo*[3]

[1] Idiap Research Institute, Martigny, Switzerland
[2] Brno University of Technology, Brno, Czechia
[3] Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
[4] Department of Informatics, University of Zurich, Zurich, Switzerland

ABSTRACT

This report describes our systems submitted to the NIST Speaker Recognition Evaluation (SRE) 2024 evaluation, for audio, audio-visual and visual verifications across both fixed and open condition scenarios. For the audio track under fixed conditions, we employ a ResNet-based embedding extraction system. The embeddings are classified using Probabilistic Linear Discriminant Analysis (PLDA) which is submitted as a primary system. We explore normalizing flows as an alternative backend classifier which allow using non-linear transformations and still obtain a model that can explicitly represent a probability density. The model makes no assumption about the distribution of the observations. For the visual track, we utilize the EdgeFace face recognition model with cosine similarity scoring, which also achieves the best Equal Error Rate (EER) in audio-visual track. In the open condition scenario, we deploy a pretrained ResNet-293 model, trained on the VoxBlink dataset, which is then fused with the fixed condition ResNet + PLDA system in the audio track.

***Index Terms*—** speaker identification, ResNet, Normalizing flow, Face recognition

1. INTRODUCTION

Given a speech segment and a target speaker, the goal of speaker recognition is to automatically determine if the target speaker was in the segment.

In the audio modality, we applied energy-based speech activity detection for the front-end system. Our submission mainly consists of the ResNet34 back-end systems with (i) Probabilistic Linear Discriminant Analysis (PLDA) [1] and (ii) Normalizing flows [2] classifiers. Normalizing flows are a general class of machine algorithms that use the transformation theorem [2, Chapter 1.2.1] in order to model a probability distribution. This is done through the application of invertible transformations on latent variables that belong to a known distribution [3, 4]. Although they gained their popularity within the field of generative modeling [5], the fact that they represent a probability distribution makes it theoretically possible to use them for classification and regression tasks. In this work, we apply normalizing flows for speaker recognition which is presented in [6].

For the visual modality, we adopted a compact EdgeFace [7] face recognition backbone for visual feature extraction.

2. AUDIO TRACK

The **CTS Superset** [8] training set which has a total duration of approximately 6'000 h and more than 6'800 speakers collected from a large number of telephony speech segments is used to train the embedding extractor and the subsequent classifiers.

2.1. Fixed Training Condition

2.1.1. Primary System

The backend model is a ResNet34 trained using WeSpeaker [9, 10] recipe[1]. The architecture follows the XI-vector [11] to pool the temporal information compared to the standard and mean and standard deviation pooling. The hyperparameters followed the recipe of Wespeaker for Voxceleb [12]. The model is further trained for an additional ten epochs with 10s training segments keeping the margin and learning rate fixed.

The embeddings extracted from this model is then used for PLDA classifier after dimensionality reduction with LDA to 100 dimensions using the CTS Superset.

2.1.2. Alternate System

In addition to PLDA, we evaluated normalizing flows [6] as an alternative backend for speaker verification. Flows offer the possibility to compose many invertible non-linear transformations to model probability distributions, allowing for more flexibility compared to the linear transformations used in PLDA. Specifically, flows do not assume a Gaussian distribution for the input embeddings, which alleviates the dependency on length normalization, a common preprocessing step in PLDA.

For the training of the normalizing flows we experimented with different combinations of hyperparameters. Among the hyperparameters tested, we evaluated learning rates of 0.001 and 0.005, batch sizes of 64 and 128, and a "number of blocks" of 2, 3, and 4. The "number of blocks" parameter controls the number of blocks in the normalizing flow architecture, each comprising a series of invertible transformations such as inverse coupling layers with channel splitting and masking [6]. Once the best checkpoint for the development set was found, it was used to generate the evaluation set

[1]We wish to acknowledge the team from Brno University of Technology for sharing the model.

scores. These scores are then calibrated using Linear and Isotonic regressions.

2.2. Open Training Condition

2.2.1. Voxblink system

The Voxblink system consists of a ResNet293 network trained on Voxblink and Voxceleb datasets [12]. WebRTC VAD[2] is applied to remove non-speech content, and diarization information is utilized wherever provided. We use this model to extract one embedding for every audio file. Cosine similarity is used to compare embeddings and then score calibration is applied.

2.2.2. Primary System

We fuse two systems: (i) the ReNet34-PLDA system used in the Fixed condition (see Section 2.1.1), and (ii) the Voxblink system mentioned in the previous subsection. The systems are fused at the score level post calibration. Both systems were provided equal weights of 0.5.

2.2.3. Contrastive System

For the contrastive system, we fuse the ResNet34 model that uses normalizing flows (see Section 2.1.2) with the Voxblink system. However, we used the SRE24 dev set to train the Normalizing flow classifier. The Voxblink system is used as the other system for score fusion with a weight of 0.6, while a weight of 0.4 was applied on the ResNet34 model's scores.

3. VISUAL TRACK

We used a compact EdgeFace [7] face recognition backbone for visual feature extraction. EdgeFace is a hybrid architecture that integrates the strengths of CNNs and transformers optimized for computational efficiency and minimal resource usage. Specifically, we used the *edgeface-base* variant, which was trained on the WebFace12M dataset [13]. The model processes aligned and cropped images of 112×112 pixels, and is trained with a CosFace [14] classification head. After training, the model produces embeddings with a dimensionality of 512.

We used MTCNN [15] to detect and align faces from enrollment images, and then extracted EdgeFace embeddings. We also processed all frames of each video in the test set; we detected all faces with MTCNN and extracted embeddings with the pretrained EdgeFace model. To reduce the error introduced by alignment, we also used a pretrained pose detection model [16] and considered faces in the test set which had poses less than 45 degrees in yaw and pitch orientations.

For our submission, we aggregated embeddings extracted from each frame for a video and then compared them to embeddings of the enrollment set. To this end, in each frame, we compared extracted embeddings with embeddings from previous frames. If the similarity was greater than a predefined threshold, we consider the same identity and update the embedding of that identity in that video with a weighted summation of new embedding (0.1) and previous embedding (0.9). The dimensions of the embeddings generated is 512.

[2] https://webrtc.org/

4. AUDIO-VISUAL TRACK

4.1. Fixed Training Condition

We use the output scores from two single best systems.

The primary system utilizes the EdgeFace model with cosine similarity distance, as described in Section 3. The alternate system employs ResNet-34 for embedding extraction with a PLDA backend described in Section 2.1.1

4.2. Open Training Condition

In the open condition, we similarly use a single best system along with an additional system, which are fused at the score level.

The primary system for the open condition is the same EdgeFace + cosine similarity setup used in the fixed condition. The alternate system, however, combines ResNet-34 with PLDA and ResNet-293 through score fusion to improve the performance.

5. PROCESSING TIME

5.1. Audio Track

5.1.1. Fixed Training Condition

For the ResNet34 system, we used an NVIDIA 3090 GPU with 24 GB VRAM which takes an average of 1 s to generate an embedding for one utterance. The PLDA training and EER computation takes around 31s.

5.1.2. Open Training Condition

For the Voxblink system, we used an NVIDIA 4090 GPU with 24 GB of VRAM, enabling processing at approximately 15 utterances per second.

5.2. Visual track

The face detection along face recognition takes an average of 730 ms per image/frame on an RTX3090 GPU.

6. ACKNOWLEDGEMENTS

This work was partially supported by Armasuisse Science and Technology and the Swiss Center for Biometrics Research and Testing at Idiap.

7. REFERENCES

[1] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[2] A. Gut, *An Intermediate Course in Probability*. Springer Publishing Company, Incorporated, 2009.

[3] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing Flows: An Introduction and Review of Current Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, 2021.

[4] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing Flows for Probabilistic Modeling and Inference," *Journal of Machine Learning Research*, vol. 22, no. 57, pp. 1–64, 2021. [Online]. Available: http://jmlr.org/papers/v22/19-1028.html

Table 1. Results on the SRE 24 development set of NIST SRE 2024 using the NIST toolkit. EER: Equal Error Rate, min C: minimum Decision Cost Function, act C: actual Decision Cost Function. †: Flow model trained with the CTS Superset and ‡: Flow model trained with SRE24 dev set.

System	Classifier	SRE24 dev set		
		EER	min.C	act.C
Fixed Condition (Audio)				
ResNet34	PLDA	10.0	0.630	0.720
ResNet34	Flow†	12.0	0.641	0.871
ResNet34	Flow‡	10.9	0.458	0.465
Visual				
EdgeFace	Cosine similarity	1.3	0.031	0.040
Fixed Condition (Audio-Visual)				
EdgeFace	Cosine similarity	1.1	0.039	0.097
ResNet34	PLDA	11.9	0.749	0.864
Open Condition (Audio)				
ResNet293 (Voxblink2)	Cosine similiarity	10.7	0.715	0.804
Fusion (ResNet34-PLDA+Resnet293)	-	7.3	0.518	0.784
Fusion (ResNet34-NF+Resnet293)	-	5.3	0.409	0.437
Open Condition (Audio-Visual)				
ResNet293 (Voxblink2)	Cosine similiarity	9.5	0.701	0.771
Fusion (ResNet34-PLDA+Resnet293)	-	7.7	0.537	0.924

[5] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *International Conference on Learning Representations*, April 2016.

[6] A. Espuña, A. Prasad, P. Motlicek, S. Madikeri, and C. Schuepbach, "Normalizing flows for speaker and language recognition backend," in *Proc. odyssey 2024*, 2024, pp. 74–80.

[7] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel, "EdgeFace: Efficient face recognition model for edge devices," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[8] O. Sadjadi, "Nist sre cts superset: A large-scale dataset for telephony speaker recognition," 2021-08-16 04:08:00 2021. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933116

[9] 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.

[10] 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, p. 103104, 2024.

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

[12] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[13] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du *et al.*, "WebFace260M: A benchmark unveiling the power of million-scale deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 492–10 502.

[14] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[15] 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.

[16] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 2074–2083.