Analysis of ABC Frontend Audio Systems for the NIST-SRE24

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Analysis of ABC Frontend Audio Systems for the NIST-SRE24

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

[1]AUDIAS, Universidad Autónoma de Madrid, Spain [2]Brno University of Technology, Czechia
[3]Athens University of Economics and Business | Omilia | Archimedes AI/Athena RC, Greece
[4]Politecnico di Torino, Italy [5]Phonexia, Czechia

sara.barahona@uam.es, {isilnova, imosner, pengjy, iplchot, rohdin}@fit.vut.cz

Abstract

We present a comprehensive analysis of the embedding extractors (frontends) developed by the ABC team for the audio track of NIST SRE 2024. We follow the two scenarios imposed by NIST: using only a provided set of telephone recordings for training (fixed) or adding publicly available data (open condition). Under these constraints, we develop the best possible speaker embedding extractors for the pre-dominant conversational telephone speech (CTS) domain. We explored architectures based on ResNet with different pooling mechanisms, recently introduced ReDimNet architecture, as well as a system based on the XLS-R model, which represents the family of large pre-trained self-supervised models. In open condition, we train on VoxBlink2 dataset, containing 110 thousand speakers across multiple languages. We observed a good performance and robustness of VoxBlink-trained models, and our experiments show practical recipes for developing state-of-the-art frontends for speaker recognition.

Index Terms: speaker recognition, NIST-SRE, embedding extractors, VoxBlink

1. Introduction

The major body of speaker verification (SV) research focuses on 16 kHz datasets with audio excerpts extracted from (e.g., YouTube) video clips. The VoxCeleb datasets [1, 2] represented a great milestone in pushing the verification performance forward by providing training data comprising thousands of speakers. This initiative has been recently followed by other works, releasing large-scale corpora comprising tens of thousands of speakers [3] and fostering research in the aforementioned domain.

On the other hand, speaker verification in telephony (characterized by various codecs and 8 kHz sampling rate), which is crucial in many real-world applications, seems to be somewhat overlooked in the community. Fortunately, NIST Speaker Recognition Evaluations have consistently pushed the boundaries of speaker recognition technology for decades, focusing especially on challenges related to detecting speakers in noisy conversational telephone speech (CTS) (but also audio from video, AfV, and multi-modal data). The latest edition (SRE24) was no exception. It included a mandatory audio-only track, as well as optional visual-only and audio-visual tracks that require the integration of speech, image, and video data for speaker identification. The main audio track deals with cross-source (CTS vs. AfV) and cross-lingual target and non-target trials. They were drawn from a new multilingual corpus, TELVID-Tunis. As the name suggests, it focuses on non-mainstream languages: Tunisian Arabic, North African French, and accented English. On top of non-standard languages, this latest edition included some novelties such as enrollment duration variability (10, 30, or 60 seconds), shorter test segments (ranging approximately from 5 seconds to 60 seconds), and multi-speaker enrollment data with diarization annotations.

The evaluation proposes two training paradigms: fixed and open. The fixed condition constrains participants to use only organizer-provided datasets, including a CTS superset [4], NIST SRE 2016 [5] and 2021 Evaluation [6], and Janus Multimedia set [7]. Such a constraint poses challenges as most of the data comprises CTS only (while the evaluation data contains both CTS and AfV). Moreover, the majority of recordings of the main training data, CTS Superset, are spoken in English (despite containing more than 50 languages), overweighting one of the evaluation languages. In contrast, the open condition allows participants to explore how unlimited additional training data affects system performance.

In this paper, our goal is to build strong frontends for speaker verification following the NIST SRE24 rules. We strive to achieve this by the following means:

- **Variability of frontends**: As per empirical evidence detailed in [8], diverse frontends tend to be complementary, which provides benefits in fusion. Therefore, we intend to explore ResNet architectures [9] that have proven strong for speaker embedding extraction in numerous evaluations [10, 11] and challenge the recent model, ReDimNet [12], that was shown to provide state-of-the-art results on the VoxCeleb trial lists and promised strong generalizability.
- **Alternative pooling method**: Inspired by consistent improvements in results on the previous NIST evaluation data provided by considering the uncertainty of estimates, we use the xi-vector pooling style within embedding extractors [13].
- **Large-scale out-of-domain dataset**: As noted before, advances in speaker verification on audio data from videos lead to collecting large-scale datasets. We aim to explore and exploit one of them, VoxBlink2 [3], in the non-constrained track. Moreover, we will focus on how such out-of-domain datasets can be beneficially used in the context of various applications.
- **SSL model**: Foundational models trained in a self-supervised way also attracted attention in speaker verification by providing strong results while requiring a shorter time to fine-tune (compared to training embedding extractors from scratch) [14]. Therefore, we aim to explore them in the challenging data.

10.21437/Interspeech.2025-2737

2. Proposed method

2.1. Training data and augmentations

For the fixed condition, we used the NIST CTS Superset [4] to train the embedding extractors. To enhance the robustness of the model, we implemented acoustic data enhancement using the Kaldi toolkit [15], incorporating noise from the MUSAN database [16] and room impulse responses from the RIR database [17]. However, we specifically excluded MUSAN's Babble noise and Music components to comply with fixed track regulations. Prior to model training, non-speech segments were removed through a Kaldi-style energy-based VAD Voice Activity Detection (VAD) system.

The open condition, free from data restrictions, presents an opportunity to leverage large out-of-domain datasets. The largest publicly available speaker verification dataset to date is VoxBlink2 [3], which consists of audios from YouTube videos belonging to 111,284 speakers, significantly surpassing the widely used VoxCeleb [2]. However, since this data does not inherently match the CTS domain, we experimented with downsampling the audio to 8kHz while applying GSM codec to 50% of the data via Sox[1] aiming to simulate the telephone channel.

2.2. Frontend systems

In this section, we present the different embedding extractors explored for facing both fixed and open conditions. For reproducibility, we also describe the training setup, implemented using the WeSpeaker toolkit [18, 19]. Most of our embedding extractors follow the VoxCeleb recipe, employing all the suggested hyperparameters for the training, which consists of two stages, both aimed at minimizing the AAM-Softmax loss [20]. The first stage involves training for 150 epochs with a 2-second segment length and employing a scale of 32 for the AAM loss. Initially, no margin is applied, but between epochs 20 and 40, the margin is gradually increased from 0 to 0.2, and this value is maintained for the remainder of the training. The learning rate scheduler uses a 6-epoch warm-up, linearly increasing the rate from 0 to its highest value (0.1), followed by an exponential decrease to 5e-5 for the rest of the training. The second stage, so-called large-margin fine-tuning, involves further training for 10 more epochs, employing a larger segment length (10 seconds) and also a larger margin value of 0.5, that remains fixed.

XI-ResNet: For the fixed condition, we explored a set of ResNet models [9], which have shown a strong performance in speaker recognition under challenging conditions. Specifically, we explored ResNet34, ResNet152 and ResNet221 architectures. Our key innovation lies in replacing the standard temporal statistic pooling (TSTP) layer with the xi-vector approach [13]. This method integrates uncertainty estimation by incorporating the Bayesian formulation of the linear Gaussian model (i-vector) directly into the pooling layer of the speaker-embedding neural network.

For the XI-ResNet152, we experimented with different modifications on the aforementioned setup including 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: As an alternative to ResNet models, we explore the recently-proposed Reshape Dimensions Network (ReDimNet) [12] architecture under fixed conditions. Although ReDimNet has achieved state-of-the-art results in the VoxCeleb benchmark, it has not yet been applied to the NIST domain. ReDimNet integrates 1D and 2D convolutional blocks by reshaping dimensionality between feature map representations in a single model. We hypothesize that by combining these two blocks, it could better capture the complex temporal and spectral variations present in telephone speech, including those introduced by limited bandwidth and channel noise, compared to ResNets, which predominantly use 2D convolutions. Specifically, we selected the B3 version based on our empirical experience, as it will be shown in Section 4. During the large-margin fine-tuning, we trained for 5 epochs employing six-second segments.

ResNet-152-VB: To leverage the unconstrained data paradigm of the open condition, we explored the VoxBlink2 dataset by training a ResNet152 model. While the xi-approach demonstrated notable efficacy in the fixed condition, we employed the conventional temporal statistic pooling for this model to systematically isolate and evaluate the impact of incorporating the VoxBlink2 data corpus. For feature extraction, we computed 80-dimensional log Mel-filterbank energy features. Following the initial training on VoxBlink2, we performed large-margin fine-tuning on the CTS Superset. As detailed in Section 4, during this stage, we explored varying segment durations to assess their influence on performance across the enrollment and test conditions introduced in SRE24.

XLS-R: In the open condition, we also made use of a foundation model pre-trained in a self-supervised way, trying to confirm/contradict the benefits shown in the context of the VoxCeleb data [14]. Specifically, we opted for XLS-R [21] as it was pre-trained on 436K hours of multilingual data comprising (at least dialects of) languages in the evaluation set. 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 [14] 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 is a lightweight attention-based embedding extractor compatible with various backbones comprising transformer encoder blocks since it employs per-frame representations at various levels of models (arguably rich in different information, e.g., phonetic or speaker-related). 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. Experimental Results

Focusing on comparing and analyzing the effects of different embedding extractors, we employed cosine scoring as our classifier, a natural choice given our optimization of the AAM loss. To isolate the effects of different frontend systems, we applied a consistent preprocessing pipeline consisting of centering, dimensionality reduction using Linear Discriminant Analysis (LDA) and length normalization of embeddings. Specifically, we intentionally omit additional pre-processing or calibration techniques to focus on the embedding extractor itself. However, for models trained on the VoxBlink2 dataset, we found it beneficial to exclude the LDA step. Performance is evaluated using Equal Error Rate (EER) and $minC_{primary}$,

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

Table 1: *Comparison of single frontend systems employing cosine scoring on the SRE 2024 development and evaluation sets for both fixed and open train conditions.*

Condition	Frontend	SRE24 dev		SRE24 eval		
		$minC_{primary}$	**EER (%)**	$minC_{primary}$	**EER (%)**	**FLOPs (G)**
Fixed	XI-ResNet-34	0.688	13.91	0.747	14.40	551
	XI-ResNet-152	0.615	**10.16**	0.695	10.41	176
	XI-ResNet-221	**0.597**	10.26	**0.683**	**10.18**	254
	ReDimNet-B3	0.728	14.70	0.784	14.33	**71**
Open	ResNet-152-VB	**0.522**	**9.31**	**0.562**	**7.59**	**219**
	XLS-R	0.666	12.02	0.681	11.69	270

as defined by the SRE24 evaluation plan [22]. Computational complexity is assessed using Floating Point Operations per second (FLOPs).

3.1. Fixed systems

In Table 1, we show the results for the different frontends explored for the fixed condition. While ReDimNet has achieved state-of-the-art results on the VoxCeleb benchmark, ResNet-based backbones consistently outperformed it across both the development and evaluation sets of the SRE24 challenge. This performance advantage of ResNet was observed despite ReDimNet's lower computational complexity, achieving the lowest FLOP value. As expected, among the ResNet models, the larger XI-ResNet-221 achieves the best performance. However, its gains over the mid-sized XI-ResNet-152 are not drastic, indicating that the benefits of scaling up the model size may be limited beyond a certain point.

The development of the ReDimNet frontend involved exploring different model sizes, as detailed in Table 2. Specifically, we evaluated configurations B0, B2, B3, and B6, on the development set to identify the optimal model size. The B3 configuration yielded the best performance, suggesting that further increases in model complexity yielded minimal gains in our domain. Given that even this optimized ReDimNet configuration did not surpass the performance of even the smallest ResNet model, only the B3 configuration was included in our final submission.

3.2. Open systems

Results for the open condition system are also shown in Table 1. The use of VoxBlink2 dataset during the first stage of the training has an enormous positive impact on the results, considerably improving performance metrics. This improvement is primarily attributed to the dataset's extensive speaker diversity, which strengthens the model's generalization capabilities and results in a notably lower EER on the evaluation set.

In contrast, our attempt to develop a robust cross-lingual system using the pre-trained multi-lingual XLS-R model did not yield the expected improvements, obtaining a performance degradation compared to the fixed condition systems. Given this model was trained with a set of 8 kHz data and contained telephone speech, results suggest that further fine-tuning strategies should be studied.

3.3. Effects of resampling Voxblink2 dataset

When training our ResNet152 over the VoxBlink2 dataset, we initially downsampled the audio to 8 kHz to align with the characteristics of our fine-tuning domain, the CTS Superset.

Table 2: *Ablation study of different ReDimNet configurations over the SRE development set.*

Frontend	$minC_{primary}$	**EER (%)**
ReDimNet-B0	0.943	27.47
ReDimNet-B2	0.783	15.48
ReDimNet-B3	**0.728**	**14.70**
ReDimNet-B6	0.777	18.76

Pre-training on this substantial corpus of domain-adapted data yielded our best-performing system, demonstrating strong generalization on the SRE24 evaluation set. However, this 8 kHz pre-trained model exhibited limited generalization to other domains, as evidenced by the VoxCeleb1 results presented in Table 3.

To further analyze the impact of resampling data in the different domains, we explored training with a combined dataset comprising both the original 16 kHz audio and the resampled 8 kHz data. The previously downsampled 8 kHz audio, which also had the GSM codec applied randomly to 50% of segments, was upsampled back to 16 kHz. This approach exposed the model to both original and simulated telephone speech. We trained ResNet152 with exactly the same parameters as detailed in Section 2.2. However, due to the doubled dataset size and that each epoch iterates over the whole dataset, we trained for 80 epochs to approximate the total number of training iterations used previously. As shown in Table 3, this combined-data model achieved comparable performance to the 8 kHz version on the SRE24 setup, even outperforming it in terms of EER. Additionally, incorporating the original 16 kHz data significantly improved generalization to other real domains, resulting in a performance gain on the VoxCeleb dataset.

For a comprehensive comparison, we also included available pre-trained VoxBlink2 models from WeSpeaker in Table 3. While these models are also ResNet-based, they are not directly comparable due to architectural differences: they incorporate Simple Attention Modules (SimAM) [23] within the ResNet blocks and utilize Attentive Statistics Pooling (AST) [24] as aggregation function. While these 16kHz systems demonstrated strong performance, they did not outperform our ResNet152-VB models on the SRE24 evaluation set. Nevertheless, using state-of-the-art pre-trained embedding extractors such as SimAM-ResNet100-ASP and its fine-tuned version on the CTS Superset is a very good alternative to our training approaches with 8kHz or 16kHz hybrid data. These systems are very compelling when we aim for a universal system capable of operating both in the telephone and wideband domains.

Table 3: *Effect of Voxblink2 sampling rate on SRE24 and VoxCeleb1 datasets. VoxCeleb1 results are only shown for models trained during the first stage, employing no margin-finetuning on a different dataset. The impact of fine-tuning ResNet152-VB on different segment lengths is also shown.*

			SRE24 dev		SRE24 eval		VoxCeleb1		
Data	**Frontend**	**FT length**	$minC_{p.}$	**EER (%)**	$minC_{p.}$	**EER (%)**	**O**	**E**	**H**
16 kHz	SimAM-ResNet34-ASP[2]	10s	0.536	9.11	0.611	8.10	1.11	1.17	2.24
16 kHz	SimAM-ResNet100-ASP[2]	10s	0.590	9.75	0.634	8.06	**0.76**	**0.89**	**1.76**
16 kHz + 8 KHz ↑	ResNet152-VB[3]	10s	0.541	**8.99**	0.582	**7.53**	1.65	1.37	2.73
8 kHz	ResNet152-VB[4]	10s	**0.522**	9.31	**0.562**	7.59	2.42	2.15	4.32
		6s	0.544	10.06	0.641	8.51	-	-	-
8kHz	ResNet152-VB	20s	0.519	8.81	0.534	7.25	-	-	-
		40s	**0.482**	**7.93**	**0.491**	**6.47**	-	-	-

[2] https://github.com/wenet-e2e/wespeaker/blob/master/docs/pretrained.md.
[3] https://huggingface.co/sarabarahona/voxblink2-ResNet152-16k
[4] https://huggingface.co/sarabarahona/voxblink2-ResNet152-8k

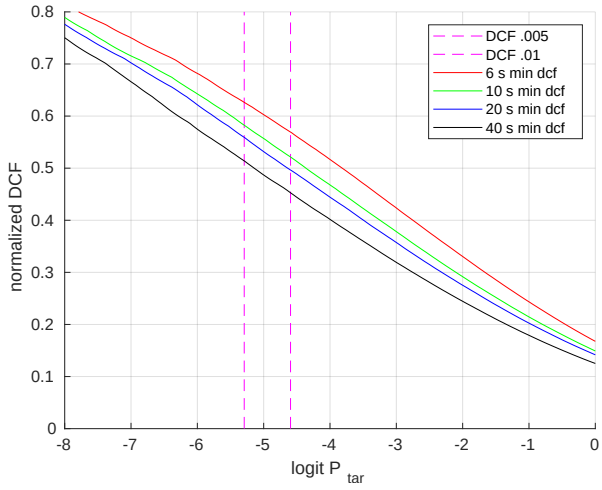

Figure 1: *DCF plots for ResNet152-VB fine-tuned on segments of different lengths.*

3.4. Impact of segment length in the fine-tuning process

Given that one of the key challenges introduced in SRE24 was the presence of shorter test segments and variability in enrollment durations, we investigated the impact of different segment lengths during the fine-tuning stage on the CTS Superset using our ResNet152-VB model. We systematically increased the segment length up to 40 seconds, observing a consistent improvement in performance. While both the development and evaluation sets exhibited gains, the evaluation set benefited the most, achieving a 23.98% reduction in EER. In terms of $minC_{primary}$, the extension of the segment length also contributed to a better generalization, closing the gap between the evaluation and development results.

The aforementioned $minC_{primary}$ results depend on the operating points chosen by the organizers. They correspond to a specific application of the verification systems. In order to provide insight into the expected performance in various applications, we show DCF plots in Figure 1. Not only do we observe improvement stemming from longer training segments for both operating points of interest (marked by vertical dashed lines), but it is consistent across a wide range of operating points. We note that we did not partition scores in any way when computing minDCF metric here.

4. Conclusion

This paper targeted speaker verification in a challenging mixture of telephony and audio-from-video speech. With the aim of building a strong frontend, we explored various architectures, pooling methods, and models pre-trained in a self-supervised way.

What we consider the most prominent contribution is the analysis of using large-scale (not necessarily in-domain) datasets to obtain noteworthy performance and generalizability. First, we showed substantial improvements from models pre-trained on (potentially downsampled and GSM-augmented) VoxBlink2 and fine-tuned on CTS Superset. Bearing in mind the length of enrollment and test recordings, we gradually increased the duration of fine-tuning segments while consistently improving performance across a wide range of operating points. Finally, our ResNet152-VB model targeted the domain of the NIST SRE24 evaluation data. However, it lacks strong generalization ability, which was tested on the VoxCeleb data. We showed that a comparable performance on the evaluation data and considerable generalization ability can be achieved when pre-training the model on a mixture of original 16 kHz data and its copy, which was subject to downsampling with the optional application of GSM codec and followed by upsampling back to 16kHz.

Our pre-trained models on VoxBlink2 have been released on HuggingFace, and we plan to release them also on WeSpeaker website to complement already available models. Using such pre-trained models can save the research community valuable computing resources.

5. Acknowledgements

This work was partly supported by FPI PRE2022-104808 and project PID2021-125943OB-I00, funded by FSE+ and MCIN/AEI/10.13039/501100011033/FEDER, UE, from the Spanish Ministerio de Ciencia e Innovación, Agencia, and the Fondo Europeo de Desarrollo Regional. It was also supported by the European Horizon 2020 Marie Skłodowska-Curie grant ESPERANTO (No. 101007666), and by the Ministry of Education, Youth and Sports of the Czech Republic (MoE) through the OP JAK project "Linguistics, Artificial Intelligence and Language and Speech Technologies: from Research to Applica-

tions" (ID: CZ.02.01.01/00/23_020/0008518). Computing on the IT4I supercomputer was supported by MoE through the e-INFRA CZ project (ID: 90254).

6. References

[1] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017*, 2017, pp. 2616–2620.

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

[3] 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.

[4] 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

[5] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 nist speaker recognition evaluation," in *Interspeech*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:263893047

[6] 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.

[7] G. Sell, K. Duh, D. Snyder, D. Etter, and D. Garcia-Romero, "Audio-visual person recognition in multimedia data from the iarpa janus program," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3031–3035.

[8] S. Cumani, A. Silnova, S. Barahona, L. Mošner, O. Plchot, and J. Rohdin, "Analysis of the ABC classification backends for nist sre24," in *Interspeech 2025*, 2025.

[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] J. Huh, J. S. Chung, A. Nagrani, A. Brown, J.-w. Jung, D. Garcia-Romero, and A. Zisserman, "The VoxCeleb Speaker Recognition Challenge: A Retrospective," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, p. 3850–3866, 2024. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2024.3444456

[11] 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.

[12] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov, and N. Torgashov, "Reshape Dimensions Network for Speaker Recognition," in *Interspeech 2024*, 2024, pp. 3235–3239.

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

[14] 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.

[15] 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.

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

[17] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] "NIST 2024 Speaker Recognition Evaluation Plan," Online, 2024. [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-2024-speaker-recognition-evaluation-sre24

[23] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 11 863–11 874.

[24] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech 2018*, 2018, pp. 2252–2256.