Analysis of the ABC classification backends for NIST SRE24

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

Analysis of the ABC classification backends for NIST SRE24

Sandro Cumani[1], *Anna Silnova*[2], *Sara Barahona*[3], *Ladislav Mošner*[2], *Oldřich Plchot*[2], *Johan Rohdin*[2]

[1]Politecnico di Torino, Italy
[2]Brno University of Technology, Czechia
[3]AUDIAS, Universidad Autónoma de Madrid, Spain

sandro.cumani@polito.it,isilnova@fit.vut.cz,sara.barahona@uam.es,
imosner@fit.vut.cz,iplchot@fit.vut.cz,rohdin@fit.vut.cz

Abstract

We present an analysis of the classification backends of the ABC submission for the audio tracks of the NIST 2024 Speaker Recognition Evaluation (SRE24). Our analysis covers embedding pre-processing, classification and score-level normalization, calibration and fusion strategies adopted to cope with the source, language and duration mismatch challenges of SRE24. We show that Pairwise Support Vector Machines provide the best results, which can be further improved, for single frontends, through score-level fusion of additional classifiers. We also show that condition-aware score calibration can mitigate the effects of source mismatch, whereas score normalization methods proved ineffective. Finally, we show that generative calibration is able to achieve competitive results with respect to other approaches.

Index Terms: Speaker verification, Speaker Recognition Evaluation, Classification backend, Pairwise Support Vector Machine, Score calibration

1. Introduction

NIST 2024 Speaker Recognition Evaluation (SRE24) [1] is the latest of a series of evaluations promoted by NIST to assess the performance of speaker verification technologies. SRE24 consists of an audio-only task, involving mismatched, cross-source and cross-language trials, an audio-visual verification problem that requires combining visual (i.e, face recognition) and speaker verification approaches, and a video-only optional task. The main audio track includes variable duration, source (conversational telephone speech (CTS) and audio-from-video), and language (Arabic, English and French) mismatched trials, including cross-language target and non-target trials. The evaluation consists of a primary, fixed set training condition, and an optional open set condition. The former requires all models to be trained with a fixed set of corpora provided by the organizers, the CTS Superset [2], NIST SRE 2016 [3] and 2021 Evaluation [4] and Janus multimedia [5] datasets, with only partial overlap with the languages of SRE24 evaluation set, and a 20-speaker development set (SRE24-Dev) that mimics the evaluation conditions. The audio component of the ABC submission is based on the score-level fusion of different subsystems, consisting of different speaker embedding extractors, pre-processing pipelines, and embedding classifiers. In this work, we analyze the different backends, i.e., classification methods, score normalization, calibration and fusion techniques, and refer the reader to the companion paper [6] for further analysis of our acoustic embedding extractors.

The paper is organized as follows. Sections 2 and 3 describe the embedding pre-processing, backend classifiers, score normalization, calibration and fusion approaches for the open and fixed conditions, respectively. Experimental results are shown in Section 4. Section 5 presents our conclusions.

2. Fixed-training condition

2.1. Frontend embedding extractors

Our submission employs different speaker embedding extractors, trained with the WeSpeaker toolkit [7, 8] on CTS Superset segments:

- *XI-ResNet-34*, *XI-ResNet-152* and *XI-ResNet-221*, ResNet-based models [9] with a xi-vector [10] approach replacing the standard statistical pooling, trained to optimize an Additive Angular Margin (AAM) loss [11] over short segments and fine-tuned using longer segments
- *ReDimNet-B3*, a Reshape Dimensions Network [12] trained to optimize the AAM loss over short segments and fine tuned using longer segments.

Multi-segment enrollment models are computed as the average of the embeddings of the corresponding recordings.

2.2. Embedding pre-processing

Our embedding pre-processing consists of different steps aimed at reducing the impact of duration, source and language mismatch and removing nuisance components:

- Embeddings are centered and length-normalized
- A 5-dimensional language subspace is estimated through Linear Discriminant Analysis (LDA). Embeddings are projected in the LDA complement space.
- 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 (i.e., we implement a soft gender-dependent re-centering). The same approach is then repeated to compensate for source mismatch.
- Embeddings are projected in a speaker LDA subspace, with dimensionality optimized for each backend on SRE24-Dev.
- Projected embeddings are length-normalized. For the ResNet frontends, length normalization is preceded and followed by Within-Class Covariance Normalization [13].

The pre-processing pipeline was trained over embeddings obtained by weighted-by-speech-duration averages of CTS Superset embeddings belonging to the same session, and SRE21-Eval enrollment segments. Despite the SRE24 evaluation containing segments with multiple speakers, our preliminary analysis on the SRE24-Dev set showed very limited benefits in employing diarization systems for multi-speaker segments. For this reason, we did not take specific actions to address this issue.

10.21437/Interspeech.2025-146

2.3. Classifiers

Since our embedding extractors optimize the AAM loss, cosine scoring (COS) appears as a natural choice for classification. However, raw cosine scoring often requires significant embedding pre-processing to be effective, and can be outperformed by more robust backend classifiers, as confirmed by our experimental results. For the evaluation, we therefore also investigated Probabilistic Linear Discriminant Analysis (PLDA) [14] and its heavy-tailed [15] variant (HT-PLDA) of [16], and Pairwise Support Vector Machines (PSVM) [17, 18, 19, 20].

Cosine scoring: the cosine backend computes the dot-product of pre-processed, length-normalized embeddings.

PLDA: we employed a PLDA model that represents a D-dimensional embedding $\boldsymbol{\phi}$ as $\boldsymbol{\phi} = \mathbf{U}\mathbf{y} + \mathbf{x}$, where $\mathbf{y}$ is an M-dimensional, a-priori standard normal distributed speaker factor, $\mathbf{x}$ is a normal-distributed residual term and $\mathbf{U}$ is a $D \times M$ matrix representing a speaker subspace. The PLDA models were trained with the CTS Superset original short and the SRE21-Eval enrollment segments.

PSVM: PSVM models the score of an enrollment-test embedding pair $(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t)$ as a quadratic function [18]

$$s(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t) = \boldsymbol{\phi}_e^T \boldsymbol{\Lambda} \boldsymbol{\phi}_t + \boldsymbol{\phi}_e^T \boldsymbol{\Gamma} \boldsymbol{\phi}_e + \boldsymbol{\phi}_t^T \boldsymbol{\Gamma} \boldsymbol{\phi}_t + (\boldsymbol{\phi}_t + \boldsymbol{\phi}_e)^T \mathbf{c} + k \tag{1}$$

whose parameters $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \mathbf{c}, k)$ are trained to optimize a regularized hinge-loss over the whole set of training utterance *pairs*. The PSVM model was trained using the approach of [17, 19] with weighted-by-speech-duration average embeddings of same-session CTS Superset recordings. The regularization coefficient was optimized on the SRE24-Dev set.

Duration-aware PSVM: Duration variability is often a relevant source of accuracy degradation [21, 22, 23, 24]. Typically, the issue is addressed either at score calibration level [21, 22, 24, 25], or, for PLDA classifiers, at model level [23, 26, 27]. In contrast with PLDA, PSVM does not provide a mechanism to account for duration variability. For this reason, we propose an extension of the PSVM model able to explicitly account for duration at training and scoring time, rather than at calibration time. The proposed approach is motivated by the effectiveness of quality measure-based duration-aware score calibration [21, 22, 28], and consists in training a PSVM model using extended embedding vectors $\widehat{\boldsymbol{\phi}}_j = \begin{bmatrix} \boldsymbol{\phi}_j \\ \alpha \log d_j \end{bmatrix}$, where $\boldsymbol{\phi}_j$ is the original embedding, d_j is the corresponding speech duration (total speech duration for multi-enrollment models), and α is a tunable parameter, optimized through cross-validation on SRE24-Dev, that allows for indirect control over the regularization of the model parameters corresponding to duration terms. Stacking embedding and log-duration leads to the scoring function

$$\begin{aligned} \widehat{s}(\widehat{\boldsymbol{\phi}}_e, \widehat{\boldsymbol{\phi}}_t) = \;& \widehat{\boldsymbol{\phi}}_e^T \widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{\phi}}_t + \widehat{\boldsymbol{\phi}}_e^T \widehat{\boldsymbol{\Gamma}} \widehat{\boldsymbol{\phi}}_e + \widehat{\boldsymbol{\phi}}_t^T \widehat{\boldsymbol{\Gamma}} \widehat{\boldsymbol{\phi}}_t \\ & + (\widehat{\boldsymbol{\phi}}_t + \widehat{\boldsymbol{\phi}}_e)^T \widehat{\mathbf{c}} + k \\ = \;& s(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t) + s_{qm}(d_e, d_t) \\ & + \alpha \log d_e (\boldsymbol{\lambda}^T \boldsymbol{\phi}_t + 2\boldsymbol{\gamma}^T \boldsymbol{\phi}_e) \\ & + \alpha \log d_t (\boldsymbol{\lambda}^T \boldsymbol{\phi}_e + 2\boldsymbol{\gamma}^T \boldsymbol{\phi}_t) \end{aligned} \tag{2}$$

where $\widehat{\boldsymbol{\Lambda}} = \begin{bmatrix} \boldsymbol{\Lambda} & \boldsymbol{\lambda} \\ \boldsymbol{\lambda}^T & \lambda_k \end{bmatrix}$, $\widehat{\boldsymbol{\Gamma}} = \begin{bmatrix} \boldsymbol{\Gamma} & \boldsymbol{\gamma} \\ \boldsymbol{\gamma}^T & \gamma_k \end{bmatrix}$, and $\mathbf{c} = \begin{bmatrix} \mathbf{c} \\ c_k \end{bmatrix}$ are the model parameters, and

$$\begin{aligned} s_{qm}(d_e, d_t) = \;& \alpha^2 \lambda_k \log d_e \log d_t + \alpha^2 \gamma_k (\log^2 d_e + \log^2 d_t) \\ & + \alpha c_k (\log d_e + \log d_t) \; . \end{aligned} \tag{3}$$

We can observe that (3) has the same formal expression of the quality measure additive term Q_4 of [21]. The scoring function (2) can thus be interpreted as the scoring function of a PSVM model that incorporates quality-measure derived terms. In the following, unless otherwise specified, PSVM will refer to duration-aware PSVM models.

2.4. Score normalization, calibration and fusion

The multi-source and multi-lingual characteristics of the SRE24 evaluation are responsible for significant intra-trial miscalibration, due to differences in the scores distribution for the different conditions. To address this issue, we investigated both score normalization and calibration-based approaches, employing SRE21 Evaluation data as normalization set. Despite our efforts, score normalization approaches such as AS-norm [29, 30] or AD-norm [31] did not provide benefits for our systems, resulting in a performance decrease both on SRE24-Dev and SRE24-Eval. Since score normalization can be interpreted as a (sub-optimal) intra-trial calibration approach [32], we replaced score normalization with a source-dependent recalibration step. For each enrollment-test source combination we estimate a linear calibration transformation using prior-weighted logistic regression (LR) [33, 34] on SRE21-Eval, and apply these transformations to SRE24 trials depending on the source of the trial segments. As shown in Section 4, this approach is effective in reducing intra-trial miscalibration. However, since SRE21-Eval and SRE24-Eval present relevant differences due to different languages, it is not sufficient to obtain globally calibrated scores. To this end we employed a generative duration-aware [24, 25] Variance-Gamma [35, 36] model, trained on a subset of the SRE24-Dev trials.

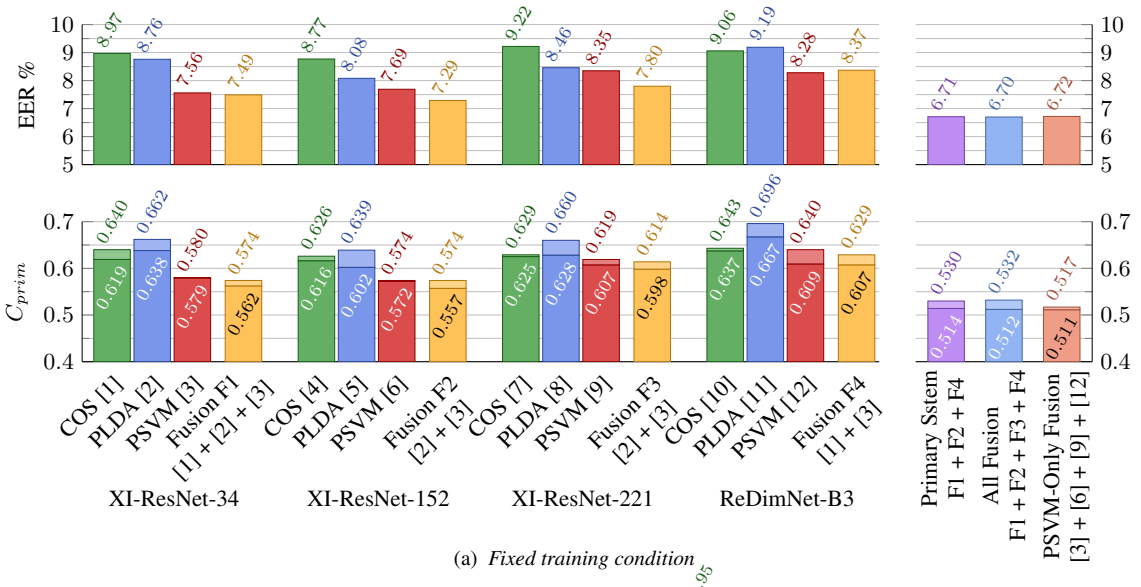
The primary submission combines the scores of several frontend / backend combinations. To reduce the risk of overfitting, for each frontend, we pre-computed a single score, obtained as a linear combination of the scores of a subset of the considered backends. The subset was chosen for each frontend based on the results on the SRE24-Dev set. The fusion weights were estimated from the SRE24-Dev set through prior-weighted LR with the target prior set to 0.01. The final score is obtained by a second linear model, also trained with prior-weighted LR over the pre-fused scores of each frontend.

3. Open-training condition

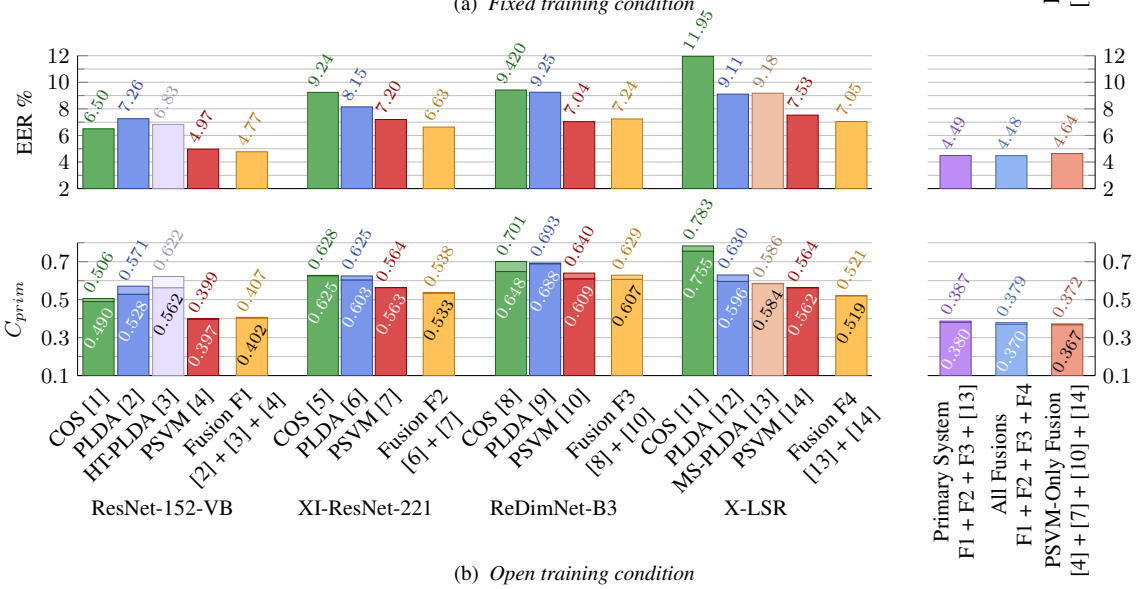
The open-training submission employs a subset of the frontends presented in the previous section, XI-ResNet-221 and ReDimNet-B3, and additionally:

- *ResNet-152-VB*, a ResNet-based model trained with 8KHz-resampled short segments, half of which were encoded through a GSM codec, from the VoxBlink2 [37] dataset, and fine-tuned with 10 s segments on the CTS Superset.
- A *XLS-R* pre-trained model [38], combined with a multi-head factorized attention (MHFA) [39] module, both fine-tuned using upsampled CTS Superset recordings.

For the open condition, we employed mostly the same pre-processing strategy and the same backends employed for the fixed training condition. In this case, however, the training lists were expanded to incorporate both VoxCeleb2 [40] development data and SRE 2018 [41] evaluation data. The lists were optimized over SRE24-Dev on a frontend / backend combination basis. For the ResNet-152-VB frontend we also employed a HT-PLDA model [16], whereas for XLS-R, due to time constraints, a single Multi-Style PLDA (MS-PLDA) model was



(a) *Fixed training condition*



(b) *Open training condition*

Figure 1: *Percent Equal Error Rate (EER) and primary cost C_{prim} for the different components of our submission, SRE24 evaluation set. For C_{prim}, we report both the minimum cost (darker part of the bars, values inside the each bar) and calibration loss / actual cost (lighter upper part of the bars, values above each bar). For each frontend, we report the results of the considered back-end, and the score-level fusion (yellow bars) of the backends that were selected for that frontend, based on the results on SRE24-Dev data. The last three columns correspond to (i) our primary submission, composed of the fusion of the XI-ResNet-34, XI-ResNet-152, and ReDimNet-B3 frontends, (ii) a fusion that includes all the considered frontends, and (iii) a fusion of all the PSVM backends.*

trained, consisting of the model interpolation [42] of two PLDA models, trained over the CTS Superset and VoxCeleb2 data, respectively.

4. Experimental results

The different components and primary submission performance is reported in Figures 1a (fixed condition) and 1b (open condition). For all embeddings and training conditions, the PSVM backend (red bars) provides the best results. Cosine scoring (green bars) proves effective in terms of primary cost C_{prim} [1], but is the worst or close to worst in terms of Equal Error Rate (EER). Fusion of different backends (yellow bars), chosen to optimize the SRE24-Dev C_{prim}, allows for further improvement with respect to PSVM models. The last three bars of each figure correspond to the primary submission (purple), consisting of a SRE24-Dev-optimized selection of a subset of the considered frontends, the fusion of all frontends (light blue) and the fusion of PSVM backends (light red). We can observe that our primary fusion is close to optimal. However, while fusion of different backends improves results for single frontends, it does not provide better results than a PSVM-only fusion.

Table 1 analyzes the effects of the different pre-processing modules for cosine and PSVM backends with XI-ResNet-152 and ResNet-152-VB embeddings. We can observe that, without any form of pre-processing, both classifiers obtain significantly worse results, with cosine scoring almost doubling the EER for

Table 1: *Contribution of different pre-processing modules for cosine and PSVM backends, XI-ResNet-152 (fixed condition) and ResNet-152-VB (open condition). The first row of each block corresponds to classification of the raw embeddings. The following rows incrementally show the effects of the different modules. These results do not include global score calibration. The PSVM results, except for the last row, refer to a PSVM without duration modeling.*

	Fixed cond.		Open cond.	
	C_{prim}^{min}	% EER	C_{prim}^{min}	% EER
Cosine	0.788	16.24	0.561	7.57
+ LDA	0.656	9.64	0.508	6.59
+ nuisance comp.	0.647	9.20	0.507	6.57
+ cond. dep. calibration	**0.614**	**8.75**	**0.498**	**6.54**
PSVM (no duration)	0.670	9.60	0.450	5.48
+ LDA	0.640	8.98	0.436	5.28
+ nuisance comp.	0.620	8.56	0.436	5.28
+ cond. dep. calibration	0.580	7.82	0.402	5.09
+ log-duration	**0.572**	**7.69**	**0.397**	**4.97**

Table 2: *Comparison of score normalization and condition-dependent calibration, cosine and PSVM backends, pre-processed embedding, XI-ResNet-152 (fixed condition) and ResNet-152-VB (open condition) frontends.*

		Fixed cond.		Open cond.	
		C_{prim}^{min}	% EER	C_{prim}^{min}	% EER
Cosine	No normalization	0.647	9.20	0.500	**6.42**
	AS-Norm†	0.788	9.92	0.669	7.34
	AD-Norm†	0.648	9.28	0.507	6.76
	Cond. dep. cal.	**0.614**	**8.75**	**0.488**	6.45
PSVM	No normalization	0.606	8.29	0.430	5.14
	AS-Norm†	0.720	9.24	0.521	5.91
	AD-Norm†	0.616	8.45	0.440	5.36
	Cond. dep. cal.	**0.572**	**7.69**	**0.397**	**4.97**

†Cohort size optimized for *SRE24-Eval* set for these experiments

the fixed training condition, whereas PSVM models tend to be more robust [20]. LDA proves fundamental for cosine, and, to a lesser degree, beneficial also for the more robust PSVM model. The additional modules, language, gender and source compensation, provide smaller improvement, especially for the open condition. The condition-dependent calibration, instead, proves again relevant, improving results for both backends. The inclusion of duration information in PSVM models allows for further, consistent, although small, improvement.

Table 2 compares the condition-dependent calibration approach to AS-norm and AD-norm. The calibration / normalization set consists of a subset of SRE21-Eval segments. We observe that score normalization is ineffective, with methods like AS-norm decreasing significantly the performance, whereas AD-norm incurs in a smaller but consistent performance degradation. These results are consistent with previous analyses of score normalization [32, 31], considering also that the lack of a large number of segments from different speakers matching the SRE24 evaluation condition impairs the quality of score normalization approaches.

Figure 1b allows also evaluating the quality of our calibration and fusion methods. Our approach employs generative calibration followed by discriminative fusion and results in a final score that is well calibrated, despite slight calibration losses for some frontend / backend combinations (mainly PLDA models). Figure 2 compares our generative calibration, with and without duration information, with LR and Quality Measure-based calibration in terms C_{prim} and C_{llr} [43, 44] for PLDA (chosen because of its bad calibration) and PSVM models. We observe that the generative models provide slightly better overall calibration (lower C_{llr}), although LR models allow for better control of the calibration loss for a specific application. LR achieves slightly better C_{prim} for PLDA models, but is more sensitive to the choice of the training prior. Duration information does not significantly improve performance for both approaches. While for PSVM this is expected, since duration has been already accounted for, the lack of improvement for PLDA models was somewhat surprising, although the results in Table 1 also suggest that duration mismatch may be less significant than we expected, and both embedding extractors and pre-processing (in particular length normalization) already mitigated most of the duration mismatch.

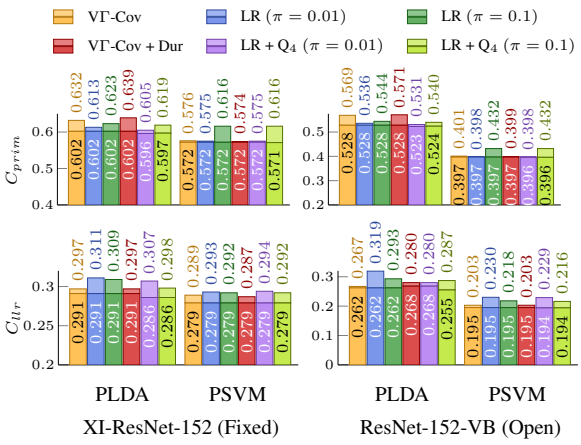


Figure 2: *Minimum (darker color, values inside the bars) and actual (lighter color, values above the bars) C_{prim} and C_{llr}, PLDA and PSVM backends. VΓ-Cov and VΓ-Cov-Dur refer to the Variance-Gamma model without and with duration information. LR and LR + Q_4 correspond to logistic regression without and with quality measures, trained with different priors π.*

5. Conclusions

We have presented an analysis of the backend classifiers of the ABC submission to NIST SRE24. Our analysis has shown that, despite cosine scoring being somewhat effective when embedding extractors have been trained with data matching a specific use case, proper backend modeling still provides superior performance. In our submission, PSVM classifiers have shown to provide better accuracy than other state-of-the-art approaches. We have also shown that generative calibration models are a viable alternative to discriminative models, providing better overall calibration, despite a limited control over the performance for a specific operating point. Furthermore, we have shown that calibration-based score normalization can provide significant improvement over traditional score normalization approaches for use cases where labeled target trials are available for normalization purposes.

6. Acknowledgements

The work was supported by Czech MoE projects OP JAK CZ.02.01.01/00/23_020/0008518 and e-INFRA CZ 90254.

7. References

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

[2] O. Sadjadi, "NIST SRE CTS Superset: A large-scale dataset for telephony speaker recognition," 2021. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933116

[3] "NIST 2016 Speaker Recognition Evaluation plan," 2016. [Online]. Available: https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016

[4] "NIST 2021 Speaker Recognition Evaluation plan," 2021. [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-2021-speaker-recognition-evaluation-sre21

[5] G. Sell *et al.*, "Audio-visual person recognition in multimedia data from the Iarpa Janus program," in *Proc. ICASSP 2018*, 2018.

[6] S. Barahona *et al.*, "Analysis of ABC frontend audio systems for the NIST-SRE24," in *Proc. Interspeech 2025*, 2025.

[7] H. Wang *et al.*, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *Proc. ICASSP 2023*. IEEE, 2023.

[8] S. Wang *et al.*, "Advancing speaker embedding learning: Wespeaker toolkit for research and production," *Speech Communication*, vol. 162, 2024.

[9] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. CVPR 2016*. IEEE, 2016.

[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, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *Proc. CVPR 2019*, 2019.

[12] I. Yakovlev *et al.*, "Reshape Dimensions Network for speaker recognition," in *Proc. Interspeech 2024*, 2024.

[13] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. ICSLP 2006*, 2006.

[14] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. ECCV'06*, 2006.

[15] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Keynote presentation, Proc. Odyssey 2010*, 2010.

[16] N. Brümmer *et al.*, "Gaussian meta-embeddings for efficient scoring of a heavy-tailed plda model," in *Proc. Odyssey 2018*, 2018.

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

[18] S. Cumani *et al.*, "Pairwise discriminative speaker verification in the i-vector space," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[19] S. Cumani and P. Laface, "Training pairwise support vector machines with large scale datasets," in *Proc. ICASSP 2014*, 2014, pp. 1645–1649.

[20] S. Cumani *et al.*, "Gender independent discriminative speaker recognition in i-vector space," in *Proc. of ICASSP 2012*, 2012.

[21] M. I. Mandasari *et al.*, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

[22] 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.

[23] S. Cumani, "Fast scoring of full posterior PLDA models," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2036–2045, 2015.

[24] S. Cumani and S. Sarni, "The distributions of uncalibrated speaker verification scores: a generative model for domain mismatch and trial-dependent calibration," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, pp. 1–16, 01 2023.

[25] S. Cumani and S. Sarni, "A generative model for duration-dependent score calibration," in *Proc. Interspeech 2021*, 2021.

[26] S. Cumani, O. Plchot, and P. Laface, "Probabilistic Linear Discriminant Analysis of i–vector posterior distributions," in *Proc. of ICASSP 2013*, 2013.

[27] T. Stafylakis *et al.*, "Text-dependent speaker recognition using PLDA with uncertainty propagation," in *Proc. Interspeech 2013*, 2013.

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

[29] S. Cumani *et al.*, "Comparison of speaker recognition approaches for real applications," in *Proc. Interspeech 2011*, 2011.

[30] Z. Karam, W. Campbell, and N. Dehak, "Towards reduced false-alarms using cohorts," in *Proc. of ICASSP*, 2011, pp. 4512–4515.

[31] S. Cumani and S. Sarni, "From adaptive score normalization to adaptive data normalization for speaker verification systems," in *Proc. Interspeech 2023*, 2023.

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

[33] N. Brummer *et al.*, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[34] N. Brümmer and G. R. Doddington, "Likelihood-ratio calibration using prior-weighted proper scoring rules," in *Proc. of Interspeech 2013*, 2013.

[35] S. Cumani, "On the distribution of speaker verification scores: Generative models for unsupervised calibration," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 547–562, 2021.

[36] S. Cumani, "Normal variance-mean mixtures for unsupervised score calibration," in *Proc. Interspeech 2019*, 2019.

[37] Y. Lin *et al.*, "Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark," in *Proc. Interspeech 2024*, 2024.

[38] A. Babu *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech 2022*, 2022.

[39] J. Peng *et al.*, "An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification," in *Proc. IEEE SLT 2022*, 2023, pp. 555–562.

[40] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018.

[41] "NIST 2018 Speaker Recognition Evaluation plan," 2021. [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-2018-speaker-recognition-evaluation

[42] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Proc. ICASSP 2014*, 2014, pp. 4047–4051.

[43] 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.

[44] D. Van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," *Lecture Notes in Computer Science*, vol. 4343, pp. 330–353, 01 2007.