POLITECNICO DI TORINO
Repository ISTITUZIONALE

Effectiveness of Language Variability Compensation in Speaker Verification

Original
Effectiveness of Language Variability Compensation in Speaker Verification / Mošner, L., Barahona, S., Cumani, S., Rohdin, J., Li, J., Plchot, O.. - ELETTRONICO. - (2026), pp. 6968-6973. (Interspeech 2026 Sidney, Australia 27 September - 1 October 2026).

Availability:
This version is available at: 11583/3016335 since: 2026-10-30T13:52:38Z

Publisher:
ISCA

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

09 November 2026

Interspeech 2026
27 September - 1 October 2026, Sydney, Australia

Effectiveness of Language Variability Compensation in Speaker Verification

Ladislav Mošner[1], *Sara Barahona*[2], *Sandro Cumani*[3], *Johan Rohdin*[1], *Jin Li*[1,4], *Oldřich Plchot*[1]

[1] Faculty of Information Technology, Speech@FIT, Brno University of Technology, Czechia
[2] AUDIAS, Universidad Autónoma de Madrid, Spain
[3] Politecnico di Torino, Italy
[4] Department of EEE, The Hong Kong Polytechnic University, Hong Kong

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

Abstract

We present our submission to the TidyVoice 2026 Challenge along with a systematic study of language-variability compensation across the entire speaker verification pipeline. We investigate multilingual fine-tuning and language compensation techniques at the embedding, back-end, and score levels. Each approach independently yields consistent, comparable gains, confirming that explicitly addressing language effects is crucial. We further show that powerful back-ends such as PSVM and SG-TPSDA perform strongly with language-entangled embeddings, but their advantage decreases when language variability is mitigated at the front-end via adversarial training with gradient reversal. Our submitted system achieves 2.53% EER on tv26_eval-A.

Index Terms: TidyVoice, language variability compensation, gradient reversal, domain invariance

1. Introduction

Research in the speaker verification (SV) field has been mostly fostered by two driving forces: 1) extensive datasets and challenges built around them, and 2) a series of speaker recognition evaluations (SRE) organized by the National Institute of Standards and Technology (NIST). Within the former category, multiple large-scale training and/or evaluation datasets, such as VoxCeleb [1], VoxBlink2 [2], or CN-Celeb [3], have been published. However, these datasets predominantly contain monolingual speakers (English or Chinese) and focus on wideband (16 kHz) audio. In contrast, the NIST SRE series has historically prioritized telephony/narrowband data, with recent iterations moving toward non-mainstream languages. Specifically, SRE21 [4] and SRE24 [5] introduced multilingual trials; however, these evaluations also introduced confounding variables, such as mismatched enrollment/test sources (e.g., telephone speech vs. audio from video), which can obscure the isolated impact of language variation on system performance.

The recently announced TidyVoice 2026 Challenge [6] and dataset [7] focus on the influence of language diversity across speaker recognition trials, complementing the above-mentioned directions. The challenge organizers provided a comprehensive benchmark comprising training and development data from 40 languages. To evaluate model generalization beyond the 40 languages, the evaluation set includes additional 38 unseen languages. Although training data was provided, the challenge allowed for the use of external data sources (excluding Mozilla Common Voice) and pre-trained models.

Speaker and channel variability modeling has been a long-standing research problem, primarily addressed by back-ends [8, 9]. We drew inspiration from recent developments [10] to address the specific channel of interest – language. Language invariance can also be framed as domain invariance and adversarial training has been shown to be effective in promoting it across various speech-related tasks [11, 12, 13, 14]. Enabled by TidyVoiceX, we explore it in the context of SV.

In this paper, we detail our submission to the TidyVoice 2026 Challenge with the following contributions: 1) We show that the language variability compensation in speaker verification leads to consistent improvements in a language-diverse scenario. 2) We analyze the compensation at different levels: embedding extractor, back-end, and scores. We show that similar gains can be achieved independently with different techniques. 3) We also analyze the impact of additional multilingual data during fine-tuning of the embedding extractor.

2. Language variability compensation

2.1. Embedding extractor level

Conventional speaker embedding extractors are typically trained using a cross-entropy objective with speaker identity labels. In multilingual datasets such as TidyVoiceX, where speakers speak in multiple languages, the network may reduce training loss by exploiting both speaker and language cues. This can result in language information being entangled within the learned embeddings. While a sufficiently powerful back-end might compensate for this effect, we instead promote language-agnostic representations directly at the embedding level.

Inspired by adversarial domain adaptation [15, 16], we utilize the joint availability of speaker and language labels in TidyVoiceX to suppress language-dependent information in the latent space. As shown in Figure 1, in the upper branch supervised with the ArcFace loss ($\mathcal{L}_s$) [17], frame-level features are pooled into an utterance-level representation and projected into speaker embedding space. In parallel, the utterance-level features are fed to a language classification branch without embedding normalization, serving to compute a cross-entropy language loss ($\mathcal{L}_l$). A Gradient Reversal Layer (GRL) is inserted after pooling in the language branch to encourage language-invariant representations in the shared backbone. The network is trained using a multi-task objective: $\mathcal{L} = \mathcal{L}_s + \mathcal{L}_l$. Through this adversarial setup, the language branch is trained to identify the spoken language, while the shared front-end is simultaneously updated to suppress language-discriminative information, effectively "purifying" the speaker embeddings.

2.2. Feature Domain Augmentation with DSU

To improve robustness under cross-lingual mismatch, we adopt a feature domain augmentation strategy based on modeling Domain Shifts with Uncertainty (DSU) [18], which has been effectively used in spoofing detection [19].

10.21437/Interspeech.2026-3367

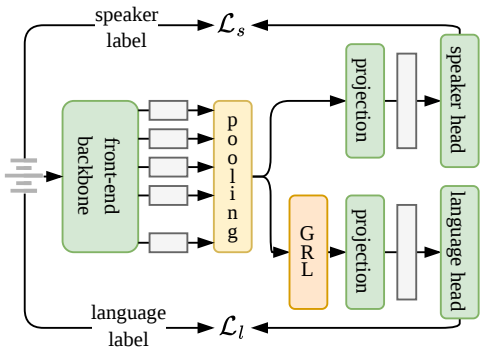


Figure 1: *A scheme of incorporating a gradient reversal layer (GRL) into an embedding extractor with a language classifier.*

DSU perturbs internal feature statistics during training to simulate assumed unseen domain shifts. First, channel-wise mean and standard deviation are computed for each sample by accumulating the feature maps in convolutional networks over the spatial dimensions. These quantities define means of two distributions estimated over the batch, from which samples are drawn to perturb internal features. We refer the reader to [18] for a detailed description.

2.3. Back-end level

To address language variability, we also investigated two back-ends: the Pairwise Support Vector Machine (PSVM) of [20, 21], which has proven robust against channel and language variability [10], and a novel Spherical-Gaussian variation (SG-TPSDA) of the Toroidal Probabilistic Spherical Discriminant Analysis of [9], which replaces the spherical likelihood of the original TPSDA with an isotropic Gaussian likelihood. The back-end models have been chosen for their proven effectiveness in the speaker verification task and their natural ability to decouple speaker and non-speaker variability. PSVM achieves the decoupling implicitly, through an objective function that naturally tends to identify speaker-rich subspaces in its input space. The TPSDA model, on the other hand, can include multiple explicit speaker and channel variability factors, although in our experiments we employed only a single speaker subspace.

To further reduce the impact of language information, we also explored an additional pre-processing step that employs Linear Discriminant Analysis (LDA) to project away the dominant language directions. In particular, we compute a discriminant language subspace, whose dimensionality is tuned on the development set, and project the embedding onto its complement.

2.4. Score level

Lastly, we attempted language compensation at the score level by using a language comparison score as a quality measure [22]. The motivation is that language similarity may affect the speaker verification score to some extent, which therefore should be reduced in same- or similar-language trials compared to different-language trials. Language comparison scores were obtained by training a ResNet18 [23] language classifier on the TidyVoiceX training data, extracting embeddings from the development data, and scoring the trials using cosine similarity with mean subtraction.

3. Experimental setup

3.1. Training data

Although the TidyVoiceX training data includes examples from 3666 speakers, the total duration is only 370 h. Therefore, we sought to expand the training data while maintaining multilinguality. We opted for two additional datasets: NIST SRE CTS Superset [24] and VoxCeleb2 dev [1]. Despite CTS Superset comprising telephony data, its merit lies in the number of speakers (6867) and duration (6193 h). Moreover, it provides information on 52 spoken languages, including multilingual speakers. To ensure compatibility with our target domain, CTS data was upsampled to 16 kHz. VoxCeleb2 dev is a standard dataset in the SV community. Not only does it contain utterances of 5994 speakers totaling 2369 h of speech, but it also comprises multiple languages. However, since VoxCeleb2 language metadata is not available, we employ language pseudo-labels[1] obtained using an ECAPA-TDNN trained on VoxLingua107 [25] when training under the language compensation strategy.

3.2. Extractor architectures

w2v-BERT 2.0 with MFA As one of the front-ends, we employed w2v-BERT 2.0 [26] with layer adapters [27] and MFA (Multi-scale Feature Aggregation) embedding extractor [28] as presented in [29]. The following reasons motivated our choice. First, the w2v-BERT 2.0 encoder was pre-trained on 4.5 M hours of speech spanning 143 languages in a self-supervised fashion. Second, Li et al. [29] showed that a strong speaker embedding extractor can be built by appending the w2v-BERT 2.0 encoder with layer adapters and the MFA module and eventually fine-tuning with a large-margin strategy (LMFT) [30] on VoxCeleb2 data. The LMFT version, available online[2], served as our starting point, which we adapted.

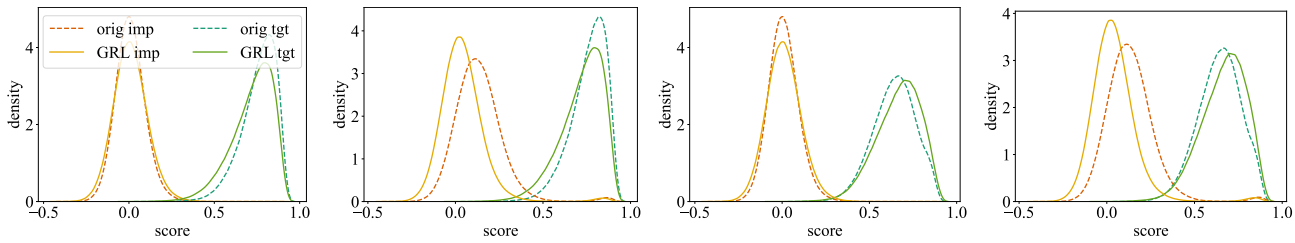
Since MFA was pre-trained on VoxCeleb2 and VoxBlink2, only the weights representing TidyVoice speaker prototypes were initialized from scratch. We updated them with a 5× higher learning rate than the remaining weights, which were optimized with a learning rate scheduled from 10^{-3} to $5 \cdot 10^{-5}$ during 5 epochs using SGD. ArcFace employed a margin of 0.3 and a scale of 32. Input segments were 2 s long.

SimAMResNet We explored ResNet-based architectures as part of our front-ends. In particular, we evaluated SimAMResNet100, a larger version of the baseline SimAMResNet34, in which Simple Attention Modules (SimAM) [31] are inserted into ResNet blocks to enhance feature representation. In addition, Attentive Statistics Pooling (AST) [32] is used instead of conventional temporal statistics pooling to better aggregate speaker information at the utterance level. We used a VoxBlink2 pre-trained model[2] and, as detailed in Section 4.1, explored different data configurations during 6 epochs of LMFT, using a learning rate decaying exponentially from $5 \cdot 10^{-5}$ to $1 \cdot 10^{-5}$. We employed a segment duration of 6 s and exponentially updated the margin from 0 to 0.3 over the first 3 epochs, as using a fixed higher margin led to worse results on the tv26_dev set.

When experimenting with DSU, we trained six model variants by inserting the DSU module at different positions, including after the initial convolution and after each ResNet stage. During inference, the DSU module is disabled. We selected the variant that performed best on the development set.

[1] https://huggingface.co/datasets/johbac/voxceleb-language-metadata

[2] https://github.com/wenet-e2e/wespeaker/blob/master/docs/pretrained.md



(a) *tgt* =, *imp* ≠ (b) *tgt* =, *imp* = (c) *tgt* ≠, *imp* ≠ (d) *tgt* ≠, *imp* =

Figure 2: *Score distributions for tv26_dev sublists obtained with w2v-BERT 2.0 w/ MFA (w/ or w/o GRL) and cosine similarity scoring.*

Table 1: *Results of the language variability compensation at the embedding extractor level on the tv26_dev trial list and its sublists: "tgt =" and "tgt ≠" denote target trials with the same and different languages, respectively. The analogy holds for languages of non-target/impostor (imp) trials.*

	Front-end	tv26_dev		tgt =, imp ≠		tgt =, imp =		tgt ≠, imp ≠		tgt ≠, imp =	
		EER	MinDCF	EER	MinDCF	EER	MinDCF	EER	MinDCF	EER	MinDCF
1	SimAM ResNet100	1.50	0.654	0.30	0.041	1.61	0.997	0.59	0.084	2.69	1.000
2	SimAM ResNet100 (DSU)	1.45	0.647	0.30	0.040	1.56	0.997	0.58	0.083	2.60	1.000
3	SimAM ResNet100 (GRL)	1.30	0.629	0.55	0.065	1.60	0.998	0.81	0.108	2.08	1.000
4	SimAM ResNet100 (GRL+DSU)	**1.20**	**0.614**	0.52	0.064	1.48	0.998	0.79	0.103	1.91	1.000
5	w2v-BERT 2.0 w/ MFA	1.50	0.669	0.14	0.026	1.41	1.000	0.38	0.065	2.83	1.000
6	w2v-BERT 2.0 w/ MFA (GRL)	**0.99**	**0.583**	0.47	0.055	1.35	0.999	0.62	0.079	1.58	1.000

Table 2: *Impact of fine-tuning data on the performance measured on tv26_dev and Vox1-H.* T *stands for TidyVoiceX training set,* V *for VoxCeleb2 dev,* C *for CTS Superset.*

Front-end	fine-tuning data	tv26_dev		Vox1-H	
		EER	MinDCF	EER	MinDCF
SimAM ResNet100	-	3.22	0.798	1.76	0.175
	T	2.82	0.785	5.32	0.443
	T+V	1.94	0.731	1.06	0.101
	T+V+C	**1.50**	**0.654**	1.00	0.093
w2v-BERT 2.0 with MFA	-	3.51	0.846	0.73	0.071
	T	**1.50**	**0.669**	1.61	0.152
	T+V	1.85	0.706	1.01	0.099
	T+V+C	1.75	0.698	0.96	0.093

3.3. Back-end model

The PSVM, Spherical-Gaussian TPSDA (SG-TPSDA), and the LDA complement projection models were trained on the TidyVoiceX training dataset. Both the LDA and SG-TPSDA subspace dimensions were tuned on the development set using the embeddings from networks that do not include language compensation. The LDA space dimension was set to 15, whereas the SG-TPSDA speaker subspace dimension was set to 120. For PSVM, we scaled the "default" regularizer employed in [20] by a factor of 5, based on the results on the development set. PSVM embeddings were length-normalized, whereas the SG-TPSDA model did not use length normalization.

4. Experiments

4.1. Front-end training data

We present the results obtained with different fine-tuning data in Table 2. We primarily focus on tv26_dev, but we also complement the analysis with VoxCeleb1-hard (Vox1-H) for better insights. Intuitively, fine-tuning on the TidyVoiceX training set improves the performance of both our front-ends on tv26_dev while degrading results on Vox1-H due to a mismatch in languages and domain (read vs. spontaneous speech) compared to the pre-trained versions. The introduction of VoxCeleb2 dev caused mixed outcomes. It substantially improved SimAMResNet100's performance on tv26_dev but degraded that of w2v-BERT 2.0 with MFA. We hypothesize that the former benefited from the increase in the number of training speakers, whereas the latter model, pre-trained on massively multilingual data, degraded due to fine-tuning on predominantly English data. The results on Vox1-H are consistent, showing the benefit of in-domain fine-tuning data. CTS Superset, although belonging to a different domain, further improved SimAMResNet100 on tv26_dev, highlighting the importance of both its extended duration and the diversity of multilingual speakers. For w2v-BERT 2.0 with MFA, its multilingualism just helped the model compared to an unfavorable T+V combination. Despite the size of CTS Superset, we observe only a negligible but consistent difference on Vox1-H.

4.2. Front-end-level language compensation

In Table 1, we present the impact of the front-end-level techniques targeting language agnosticity described in Section 2. Based on the results on the tv26_dev set, we conclude that both gradient reversal and DSU provide a performance boost for SimAMResNet100, with GRL being overall superior. Importantly, gradient reversal and DSU can be applied simultaneously. Our results reveal the complementarity of the two techniques, as the combination outperforms either method alone.

The tv26_dev trial list comprises a mixture of all available trials. In order to show the effect of our techniques, it is insightful to split the list into four subsets containing 1) target trials with the same language (tgt =) and impostor trials with a different language (imp ≠), 2) target trials with the same language and impostor trials with the same language, 3) target trials with a different language and impostor trials with a different language, and 4) target trials with a different language and impostor trials with the same language. The results on the subsets reveal the behavior of our initial models (without language com-

Table 3: *Back-end-level language variability compensation and combination with front-end-level techniques, along with the system fusion.* (NO)LN *stands for (no) length normalization.*

	Front-end	Back-end	tv26_dev		tv26_eval-A		tv26_eval-U	
			EER	MinDCF	EER	MinDCF	EER	MinDCF
	Baseline [6]	cosine similarity	3.07	0.82	9.06	0.65	11.59	0.60
1	SimAM ResNet100	PSVM-LN (LDA 15 dims.)	1.05	0.609	3.51	0.275	3.94	0.245
2	w2v-BERT 2.0 w/ MFA	SGTPSDA-NOLN (LDA 15 dims.)	0.93	0.596	2.99	0.229	4.57	0.289
3	1+2		0.81	0.579	2.53	0.190	3.40	0.196
4	SimAM ResNet100 (GRL+DSU)	PSVM-LN (LDA 15 dims.)	1.19	0.613	3.40	0.254	4.05	0.291
5	SimAM ResNet100 (GRL+DSU)	PSVM-LN (LDA 5 dims.)	1.07	0.603	3.23	0.252	3.64	0.278
6	w2v-BERT 2.0 w/ MFA (GRL)	SGTPSDA-NOLN (LDA 15 dims.)	1.09	0.591	3.60	0.241	4.15	0.262
7	w2v-BERT 2.0 w/ MFA (GRL)	SGTPSDA-NOLN (LDA 5 dims.)	1.03	0.585	3.35	0.233	3.49	0.229

pensation) that agrees with intuition. For case 1, the parity of languages in target trials makes embeddings more similar. In contrast, differences in language in impostor trials make embeddings more dissimilar, resulting in dramatically better performance than on tv26_dev. On the other hand, the most challenging trials are those described by case 4. When imposing language invariance via the gradient reversal method, the trends across both our front-end models are the same: Most importantly, we observe a significant improvement for case 4, which is the most indicative of whether the technique's empirical outcomes meet the hypothesis. It is worth noting that the performance degrades in the easiest case 1, which might seem unintuitive. However, these results show that language information becomes less pronounced in the embeddings, as it was significantly helping the initial models achieve very low EER.

Focusing on DSU, we observe that augmenting the internal representations during training consistently improves performance, or at least does not degrade it, across all trial lists. DSU does not make use of language information during training. Therefore, it is expected that the trends in the results will differ from those with the GRL technique. We note that we experimented with DSU incorporated into both versions of w2v-BERT 2.0. Due to a negligible effect, we do not present the results here. We intend to investigate it further in the future. We also stress that our analysis was performed using the tv26_dev trials comprising languages seen in training (due to the unavailability of labels for the evaluation trial lists at the time of submission). Therefore, the behavior under more difficult language conditions remains to be explored.

Table 1 also shows that MinDCF with a target trial probability of 0.01 defined by the challenge approaches 1 for trial sublists that contain impostor trials with the same language. To explain this behavior, we show the distribution of scores across all trial sublists in Figure 2. Subfigures 2b and 2d depict the second mode of the non-target score distribution, which affects the MinDCF values, suggesting potential issues with some of the labels. Figure 2 also shows favorable changes in distributions after incorporating GRL, which are more pronounced for the non-target trials.

4.3. Back-end-level language compensation and submission

Table 3 presents, among other results, those for back-end-level language compensation. System 1 comprises SimAM ResNet100 + PSVM (with length-normalized embeddings), and system 2 combines w2v-BERT 2.0 w/ MFA + SG-TPSDA (without length normalization). The two models can be compared with models 1 and 5 (with no compensation) in Table 1, respectively. The comparison shows the value of the back-ends.

Table 3 also shows systems that combine both embedding-level and back-end-level language compensation. Rows with models 4 through 7 present results for the same back-ends trained with front-end-level language-compensated embeddings. We can observe that the back-end-level compensation is, in this case, much less effective — for both front-ends, good results are achieved only when the number of LDA dimensions is drastically reduced, and for the w2v-BERT 2.0 embeddings, we still observe a degradation relative to cosine scoring on the development set. On the other hand, we can observe that language-compensated back-ends trained on the original, non-language-compensated embeddings still achieve the best results on the development set. This suggests that while front-end-level compensation is indeed effective, it may still incur a loss of speaker information that could be effectively extracted by ad-hoc back-end models.

Finally, Table 3 summarizes the results of our final submission to the challenge. It includes back-end level language compensation and consists of the fusion of systems 1 and 2. The fusion weights have been estimated on the development set by means of prior-weighted logistic regression [33, 34], with a target prior set to 0.5. The best single system (2) and the fusion (3) consistently outperform the baseline on all the trial lists.

4.4. Score-level language compensation

To analyze the impact of score-level language compensation, we divided the development set randomly into two parts with non-overlapping speakers, each having 100k trials. We used the first part for mean subtraction before cosine scoring and to train a logistic regression calibration model with language-comparison scores as quality measures. For system 5 in Table 1 (which does not include language compensation), this improved the EER on the second part from 1.28% to 1.10%, while for system 6 (which uses adversarial language compensation) the improvement was negligible (from 0.79% to 0.78% EER).

5. Conclusion

This paper focused on addressing the harmful effects of language variability in SV trials, a problem brought to the community's attention by the TidyVoice 2026 Challenge.

We evaluated multiple pre-trained front-ends and highlighted the importance of appropriate fine-tuning data. Enforcing language-invariant embeddings via a gradient reversal in a language classifier significantly improved performance, while comparable gains were achieved by combining language-entangled embeddings with suitable back-end models. We also demonstrated improvements using score-level compensation by introducing language-comparison scores.

6. Acknowledgments

The work was supported by 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 Applications" (ID:CZ.02.01.01/00/23_020/0008518), by "LLMs4EU", co-funded by the Digital Europe Programme under GA 101198470 and MoE under No. 8Y26003, by FPI PRE2022-104808 funded by MICIU/AEI/10.13039/501100011033 and FSE+, by project PID2024-160789OB-I00 funded by MICIU/AEI/10.13039/501100011033/FEDER,UE, and by project SI4/PJI/2024-00237 (COSER-IA) funded by Comunidad de Madrid. Computing on IT4I supercomputer was supported by MoE through the e-INFRA CZ (ID:90254).

7. Generative AI Use Disclosure

The authors disclose that parts of the paper were polished using Gemini 3. The LLM was used solely to edit the original text, not to generate it from scratch. The authors did the necessary corrections to the edited text.

8. References

[1] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech 2018*, 2018, pp. 1086–1090.

[2] 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.

[3] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "CN-Celeb: A challenging Chinese speaker recognition dataset," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[4] O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "NIST 2021 speaker recognition evaluation plan," National Institute of Standards and Technology (NIST), Tech. Rep., 2021. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=932697

[5] National Institute of Standards and Technology, "NIST 2024 speaker recognition evaluation plan," U.S. Department of Commerce, National Institute of Standards and Technology, Tech. Rep., 2024. [Online]. Available: https://www.nist.gov/document/sre24-evaluation-plan

[6] A. Farhadipour, J. Marquenie, S. Madikeri, T. Vukovic, V. Dellwo, K. Reid, F. M. Tyers, I. Siegert, and E. Chodroff, "TidyVoice 2026 Challenge evaluation plan," *arXiv preprint*, 2026, arXiv:2601.21960 [cs.SD].

[7] A. Farhadipour, J. Marquenie, and E. C. Srikanth Madikeri, "TidyVoice: A curated multilingual dataset for speaker verification derived from common voice," in *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 19 047–19 051.

[8] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[9] A. Silnova, N. Brümmer, A. Swart, and L. Burget, "Toroidal probabilistic spherical discriminant analysis," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] 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, pp. 3978–3982.

[11] A. Tripathi, A. Mohan, S. Anand, and M. Singh, "Adversarial learning of raw speech features for domain invariant speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5959–5963.

[12] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4889–4893.

[13] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, L. Burget, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6006–6010.

[14] X. Cai, Z. Wu, K. Zhong, B. Su, D. Dai, and H. Meng, "Unsupervised cross-lingual speech emotion recognition using domain adversarial neural network," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[15] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, p. 1180–1189.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2096–2030, Jan. 2016.

[17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 4690–4699.

[18] X. Li, Y. Dai, Y. Ge, J. Liu, Y. Shan, and L. Duan, "Uncertainty modeling for out-of-distribution generalization," in *International Conference on Learning Representations*, 2022, pp. 1–16.

[19] J. Peng, J. Li, J. Rohdin, L. Zhang, M. Hlaváček, and O. Plchot, "BUT systems for WildSpoof Challenge: SASV in the wild," *arXiv preprint*, 2025, arXiv:2512.12851 [eess.AS].

[20] 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.

[21] 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.

[22] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

[23] 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.

[24] 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

[25] J. Valk and T. Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Proc. IEEE SLT Workshop*, 2021.

[26] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, and et al., "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint*, 2023, arXiv:2312.05187 [cs.CL].

[27] D. Cai and M. Li, "Leveraging ASR pretrained conformers for speaker verification through transfer learning and knowledge distillation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3532–3545, 2024.

[28] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, and H. Meng, "MFA-Conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Interspeech 2022*, 2022, pp. 306–310.

[29] Z. Li, M. Cheng, and M. Li, "Enhancing speaker verification with w2v-BERT 2.0 and knowledge distillation guided structured pruning," *arXiv preprint*, 2025, arXiv:2510.04213 [eess.AS].

[30] J. Thienpondt, B. Desplanques, and K. Demuynck, "The IdLab VoxSRC-20 submission: Large margin fine-tuning and quality-aware score calibration in DNN based speaker verification," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5814–5818.

[31] 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.

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

[33] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. A. van Leeuwen, P. Matějka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Transactions on 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 *Interspeech 2013*, 2013.