EcoSpeak: Cost-Efficient Bias Mitigation for Partially Cross-Lingual Speaker Verification

Divya V Sharma
IIIT-Delhi
divyas@iiitd.ac.in

Abstract

Linguistic bias is a critical problem concerning the diversity, equity, and inclusiveness of Natural Language Processing tools. The severity of this problem intensifies in security systems, such as speaker verification, where fairness is paramount. Speaker verification systems are biometric systems that determine whether two speech recordings are of the same speaker. Such user-centric systems should be inclusive to bilingual speakers. However, Deep neural network models are linguistically biased. Linguistic bias can be full or partial. Partially cross-lingual bias occurs when one test trial pair recording is in the training set's language, and the other is in an unseen target language. Such linguistic mismatch influences the speaker verification model's decision, dissuading bilingual speakers from using the system. Domain adaptation can mitigate this problem. However, adapting to each existing language is expensive. This paper explores cost-efficient bias mitigation techniques for partially cross-lingual speaker verification. We study the behavior of five baselines in five partially cross-lingual scenarios. Using our baseline behavioral insights, we propose EcoSpeak, a low-cost solution to partially cross-lingual speaker verification. EcoSpeak incorporates contrastive linguistic (CL) attention. CL attention utilizes linguistic differences in trial pairs to emphasize relevant speaker verification embedding parts. Experimental results demonstrate EcoSpeak's robustness to partially cross-lingual testing.

1 Introduction

Linguistic bias is a crucial problem that harms the diversity, equity, and inclusiveness of Natural Language Processing (NLP) tools. The severity of this problem further increases in security systems, such as speaker verification, where fairness is critical. Speaker verification systems are biometric systems that determine whether two speech recordings are of the same speaker. The two input speech recordings form a trial pair. Positive or negative trial pairs indicate whether the recordings are of the same speaker. Speaker verification systems have applications in forensics, e-commerce, law, and access-control mechanisms (Estevez and Ferrer, 2023). These systems can be text-independent or text-dependent (Wu and Liao, 2021). Text-independent systems verify speakers without any constraint on speech content. Such systems work by analyzing the acoustic differences in trial pairs, consequently saving users from memorizing passphrases. Therefore, text-independent systems offer a better user experience than text-dependent systems.

Deep Neural Network (DNN) models have shown outstanding results in text-independent speaker verification (Chung et al., 2018; Nagrani et al., 2020, 2017). However, the embeddings obtained from DNN models often entangle acoustic and linguistic information (Zhou et al., 2021). Consequently, DNN-based speaker verification models become linguistically biased (Lu et al., 2009; Yang et al., 2022). Linguistic bias makes the model consider irrelevant language information in embeddings while making decisions for speaker verification, leading to performance degradation on unseen target languages. Such a bias can be full or partial. In the fully cross-lingual scenario, both the test trial pair recordings are in the target language t that is different from the source (or the training set) language s. In contrast, partially cross-lingual is another crucial scenario where one of the test trial pair recordings is in s, and the other is in t

Most of the previous works focus on the fully cross-lingual scenario. However, about 40% of the global population is bilingual (Wu and Liao, 2021). Therefore, addressing the partially cross-lingual challenge is essential to enhance the usability of speaker verification models. A viable solution to this problem is domain adaptation (Lee et al., 2020; Zhu and Chen, 2022; Chen et al., 2020; Wang et al.,

Findings of the Association for Computational Linguistics: NAACL 2024, pages 379–394

2019; Rohdin et al., 2019; Tu et al., 2019; Xia et al., 2019). However, adapting the model to each of the 7,000 existing languages is expensive. Alternatively, we can train models on large-scale cross-lingual datasets to enhance their generalizability to unseen languages (Chojnacka et al., 2021). However, this approach incurs enormous computational and storage costs. High computational cost leads to high carbon emissions, which impacts the environment (Schwartz et al., 2020; Xu et al., 2021).

In this work, we investigate cost-efficient techniques to mitigate partially cross-lingual bias in text-independent speaker verification. We propose EcoSpeak, a low-cost solution to mitigate partially cross-lingual bias. The EcoSpeak architecture incorporates a lightweight residual network, novel contrastive linguistic (CL) attention, and the bias corrector. Here, residual connections allow the model to emphasize the low-level acoustic features essential for speaker verification. Our proposed CL attention mechanism utilizes linguistic differences in trial pair recordings to generate attention weights for speaker verification. The bias corrector module penalizes the speaker verification probabilities based on linguistic differences in trial pair recordings. We first study the behavior of five baselines on five partially cross-lingual test sets created using speech recordings in four low-resource languages. Subsequently, we investigate the effectiveness of EcoSpeak on these test sets without domain adaptation. Furthermore, we explore low-cost fine-tuning techniques to enhance the generalizability of EcoSpeak for unseen low-resource languages.

We summarize our main contributions below:

1. We study the behavior of five baseline models on five partially cross-lingual test sets for four low-resource languages.

2. We propose EcoSpeak, a cost-efficient solution for bias mitigation in partially cross-lingual speaker verification.

3. We investigate the effectiveness of EcoSpeak on partially cross-lingual test sets. Furthermore, we explore cost-efficient fine-tuning strategies to enhance the model's generalizability to unseen languages.

2 Related Works

Partially Cross-Lingual Bias: Training on large-scale cross-lingual datasets mitigates partially cross-lingual bias (Wu and Liao, 2021; Qin et al., 2021). However, it is hard to find such cross-lingual labeled datasets (Wu and Liao, 2021). Moreover, this approach incurs enormous computational and storage costs. Another viable option is multi-task learning (Zhou et al., 2021). Multi-task learning can make the model jointly learn speaker identities and reduce the effect of linguistic bias. Furthermore, a fusion of multiple models can mitigate linguistic bias (Qin et al., 2021; Thienpondt et al., 2020). However, fusion would increase the system's inference cost. Notably, residual networks are relatively more robust to linguistic differences than many other models (Qin et al., 2021; Thienpondt et al., 2020). However, the reason still needs to be investigated. In this work, we study the behavior of residual networks in the partially cross-lingual scenario. To our knowledge, Thienpondt et al. (2020) is the most closely related work to our problem. In Thienpondt et al. (2020), the authors address the partially cross-lingual scenario where speakers speak Persian as their first language and English as their second language (Zeinali et al., 2019). They proposed subtracting a language compensation offset if the utterances in the trial pair are in different languages. Nevertheless, they focussed on closed-set speaker verification where the test utterance belongs to the set of known speakers within the training set. In contrast, we focus on the open-set scenario where test trial pair recordings can belong to unknown speakers outside the training set.

Green Speech Processing: The NLP community strives towards developing inclusive and environment-friendly models (Schwartz et al., 2020; Xu et al., 2021). However, speech processing is expensive, requiring enormous computational and storage resources. For instance, the training set of the SpeakerStew consisted of 20,618,000 utterances from 196,000 speakers (Chojnacka et al., 2021). Similarly, the XLS-R model contains about 2B parameters. The training set of XLS-R consisted of nearly half a million hours of speech recordings (Babu et al., 2022). In Qin et al. (2021), authors trained the model on speech recordings from 21,795 virtual speakers and the actual training set speakers for partially cross-lingual bias mitigation. High computational costs lead to high carbon footprints. Therefore, researchers have explored cost-effective bias mitigation techniques for fully cross-lingual speaker verification (Sharma and Buduru, 2022; Li et al., 2022). In this work, we investigate cost-efficient bias mitigation techniques

for partially cross-lingual speaker verification.

3 Proposed Approach

The remarkable success of the Deep Convolutional Neural Network (CNN) based speaker recognition models motivates us to investigate linguistic bias in these models. These models consist of multiple CNN layers (Nagrani et al., 2020; weon Jung et al., 2020, 2022). Lower layers capture low-level speech features, whereas higher layers capture high-level speech features. Low-level speech features contain mostly acoustic information, essential for speaker verification (Lesenfants et al., 2019). On the other hand, high-level speech features contain more linguistic information (Nahum et al., 2008). Therefore, deeper models can learn more linguistic details during training and become biased. Thus, to mitigate linguistic bias in speaker verification, we propose EcoSpeak. This section describes the architectural details of EcoSpeak.

Hypothesis: We know that residual connections add lower layer output features to the higher layer output features in a deep CNN model (He et al., 2016). Consequently, low-level acoustic features get added to the higher-level advanced representations. This summation allows the model to emphasize the low-level acoustic features essential for speaker verification. Accordingly, we hypothesize that residual connections help mitigate linguistic bias by making the model focus more on low-level acoustic information.

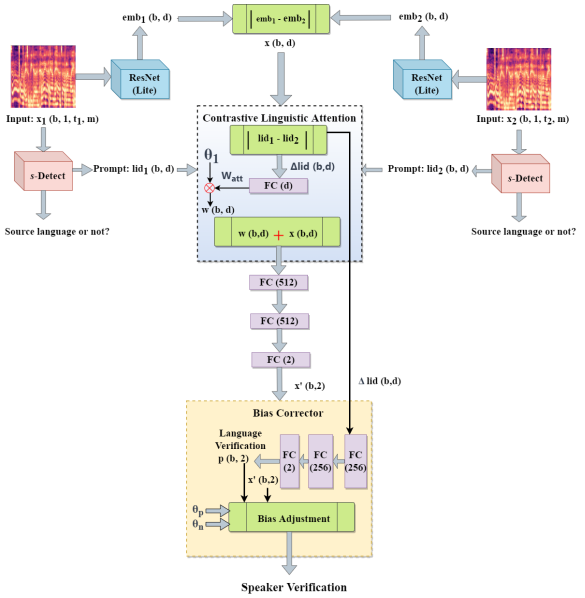


Figure 1: Architecture diagram for EcoSpeak.

Input: Firstly, we preprocess the trial pair recordings to crop silent parts. We then compute 64-dimensional normalized log mel spectrogram features of shape $(b, 1, t_{\text{i}}, m)$ as shown in Figure 1. Here, b denotes the batch size, 1 indicates mono-channel audio, t_{i} denotes the time steps, and m is the number of Mel bands (m=64). Since the duration of input speech recordings may vary during test time, the values of t_1 and t_2 may differ. Next, we pass these features through the ResNet (Lite) and the s-Detect model.

s-Detect: Partially cross-lingual trial pairs contain one speech recording in the source language s and the other speech recording in an unseen target language t. Thus, we use the s-Detect model to determine whether the input speech recording is in s. The model consists of three bidirectional gated recurrent unit (GRU) layers with a hidden size of 128 and a fully connected layer. As shown in Figure 1, it returns the output probability and a 256-dimensional lid_{i} embedding ($d = 256$).

ResNet (Lite): ResNet (Lite) is a lighter variant of the ResNet-34 (He et al., 2016).[1] The model is pre-trained for speaker identification. Speaker identification is a multi-class classification problem where the system accepts a speech recording as input and determines the speaker's identity from the known speakers in the training set. We use the pre-trained ResNet (Lite) in EcoSpeak to extract d-dimensional speaker embeddings as shown in Figure 1. First, we get emb_1 and emb_2 for the trial pair recordings from the avgPool layer of ResNet (Lite).[2] Next, we compute the absolute difference between these embeddings: $x = |emb_1 - emb_2|$. Computing the difference of the trial pair embeddings enables EcoSpeak to focus on the discriminatory information for speaker verification. Furthermore, computing absolute difference ensures that the model's output is unaffected by the input order, as absolute difference is a commutative operation.

Contrastive Linguistic (CL) Attention: Recent works have demonstrated the effectiveness of attention in speaker verification (Desplanques et al., 2020; weon Jung et al., 2020). We propose the contrastive linguistic (CL) attention mechanism for partially cross-lingual speaker verification. CL attention utilizes the linguistic differences between the trial pair recordings to generate attention weights. The attention block receives x as input and lid_1 and lid_2 as prompt inputs. CL attention works as

[1] We compared two ResNet-34 variants and chose a robust and lighter variant. Details are in the ablation study.

[2] Details of ResNet (Lite) layers is in Appendix (A).

follows:

1. **Generate attention weights:** First, we compute the absolute linguistic difference between the trial pair recordings. We then pass the resulting difference embedding through a fully connected layer and apply $ReLU$ to get the CL attention weights, W_{att} as shown below.

   $\Delta lid = |lid_1 - lid_2|$

   $W_{\text{att}} = ReLU(\Delta lid\, W^{\text{T}} + b_{\text{linear}})$

   Here W and b_{linear} are the weights and biases of the fully-connected layer.

2. **Apply attention:** We apply attention weights to the speaker embedding difference and get the output of the CL attention as follows:

   $x' = x + tanh(\theta_1) * W_{\text{att}}$

   Here θ_1 is a learned parameter.

Bias Corrector: To perform speaker verification, we feed x' through fully connected layers, as shown in Figure 1. We pass the resulting speaker verification probabilities to the bias corrector. The bias correction process involves two steps: language verification and bias adjustment.

Language verification: We jointly train EcoSpeak for speaker and language verification. Language verification is the binary classification task of determining whether the trial pair recordings are in the same language. For this task, we pass Δlid through fully connected layers to get the language verification probabilities p, as shown in Figure 1.

Bias adjustment: The decision of speaker verification models is influenced by the linguistic similarity in trial pair recordings. The speaker verification model may favor the positive class if the trial pair recordings are in the same language. On the other hand, if the trial pair recordings are in different languages, the speaker verification model may favor the negative class. Based on this intuition, EcoSpeak incorporates a bias adjustment module, thus ensuring fairness. If the language verification result is positive, the bias corrector prevents EcoSpeak from favoring the positive class while deciding on speaker verification. For this, the bias adjustment process adds a penalty to the negative class as follows: $x'[i,0] = x'[i,0] + |\theta_{\text{n}}|$. If the language verification result is negative, the bias corrector prevents EcoSpeak from favoring the negative class while deciding on speaker verification. For this, the bias adjustment process adds a penalty to the positive class as follows: $x'[i,1] = x'[i,1] + |\theta_{\text{p}}|$. Here θ_{p} and θ_{n} are learned parameters.

4 Experimental Setup

The datasets[3] and baseline models[4] used in this study are publicly available. All sets have an equal number of positive and negative trial pairs. In our experiments, English is the source language s, whereas Tamil, Telugu, Malayalam, and Kannada are the low-resource target languages t.

4.1 Datasets

Pre-train ResNet (Lite): We used the VoxCeleb-2 dev set to train ResNet (Lite) for speaker identification (Nagrani et al., 2020; Chung et al., 2018). The dataset contains 1,092,009 utterances from 5,994 speakers. Furthermore, we evaluated the model performance on the VoxCeleb-1 test set (Nagrani et al., 2017). It contains 37,720 trial pairs. The VoxCeleb datasets contain mostly English speech utterances (Qin et al., 2021). Thus, English is the source language s in our experiments.

Train s-Detect: We trained s-Detect using utterances in English and five Indian languages: Hindi, Tamil, Telugu, Malayalam, and Kannada.[5] We used the Indian-accented English data recorded for the NPTEL 2020 lectures (AI4Bharat, 2020). We obtained Hindi speech recordings from the Multilingual and code-switching ASR Challenge Dataset - sub-task1 (Diwan et al., 2021). In addition to the OpenSLR datasets, we utilized the Tamil and Telugu conversational speech recordings available in the Microsoft Speech Corpus (Microsoft, 2023). We used Malayalam and Kannada speech recordings available in OpenSLR (He et al., 2020).

[3] VoxCeleb:https://www.robots.ox.ac.uk/~vgg/data/voxceleb/, Indian-English (NPTEL):https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset, Hindi:http://openslr.org/103/, Tamil:http://openslr.org/65/, Telugu:http://openslr.org/66/, Malayalam:http://openslr.org/63/, Kannada:https://openslr.org/79/, Microsoft speech corpus:https://www.microsoft.com/en-za/download/details.aspx?id=105292, NISP:https://github.com/iiscleap/NISP-Dataset

[4] VGG-M:https://github.com/Derpimort/VGGVox-PyTorch, X-Vector:https://huggingface.co/speechbrain/spkrec-xvect-voxceleb, ECAPA-TDNN:https://huggingface.co/speechbrain/spkrec-ECAPA-voxceleb, RawNet-2:https://github.com/Jungjee/RawNet/tree/master/python/RawNet2, RawNet-3:https://github.com/Jungjee/RawNet/tree/master/python/RawNet3

[5] Details about the training setup are in Appendix (B).

Test Set	Positive Trial Pairs	Negative Trial Pairs
$\boldsymbol{tt-tt}$	both target (tt)	both target (tt)
$\boldsymbol{ts-tt}$	one target (t), one source (s)	both target (tt)
$\boldsymbol{ts-ts}$	one target (t), one source (s)	one target (t), one source (s)
$\boldsymbol{tt-ts}$	both target (tt)	one target (t), one source (s)
$\boldsymbol{ss-ss}$	both source (ss)	both source (ss)
$\boldsymbol{ss-st}$	both source (ss)	one source (s), one target (t)
$\boldsymbol{st-ss}$	one source (s), one target (t)	both source (ss)

Table 1: Brief description of the Low-Resource Language (LRL) Test Sets. Here, s represents the source language (English), and t represents the target language (the speaker's native language). The $tt-tt$ test set is fully cross-lingual, whereas $ss-ss$ is the same language test set. The remaining five test sets are partially cross-lingual.

Cross-lingual Speaker Verification: We used the NISP dataset for cross-lingual speaker verification experiments (Kalluri et al., 2021). The dataset consists of speech recordings from bilingual speakers having Hindi, Tamil, Telugu, Malayalam, or Kannada as their native language. These bilingual speakers use English as their second language. Thus, each speaker in the dataset has contributed recordings in English and their native language.

4.2 Low-Resource Language Test Sets

Our study focuses on Tamil, Telugu, Malayalam, and Kannada as the target low-resource languages (LRL). We used NISP-LRL native speaker data for cross-lingual testing. We consistently employ the following notations to present our experimental results:

1. s: The source language, i.e., English.
2. t: The target language, ie., the speaker's native language.
3. ts or st: Trial pair where one recording is in English s and the other is in the speaker's native language t.

We created seven LRL test sets described by the following notations:

1. $tt-tt$: All trial pair recordings are in the speaker's native language t.
2. $ts-tt$: Positive trial pair recordings are in different languages ts, whereas negative trial pair recordings are in the speaker's native language tt.
3. $ts-ts$: Each trial pair contains speech recordings in different languages ts.
4. $tt-ts$: Positive trial pairs contain both recordings in the speaker's native language tt, whereas negative trial pair recordings are in different languages ts.
5. $ss-ss$: All recordings are in English s.
6. $ss-st$: Positive trial pairs contain both recordings in English ss whereas negative trial pair recordings are in different languages st.
7. $st-ss$: Positive trial pair recordings are in different languages st, whereas negative trial pair recordings are in English ss.

Table 1 presents a compact description of the seven LRL test sets. Each LRL test set contains 100,000 trial pairs. These sets consist of 25,000 trial pairs from native speakers of each low-resource language. Speakers in negative trial pairs have the same gender. Accordingly, we generated a same language test set ($ss-ss$), fully cross-lingual test set ($tt-tt$) and five partially cross-lingual test sets ($ts-tt$, $ts-ts$, $tt-ts$, $ss-st$ and $st-ss$).

4.3 Baselines

We studied the behavior of the following five baselines on the LRL test sets: RawNet-3, ECAPA-TDNN, RawNet-2, X-Vectors, and VGG-M (weon Jung et al., 2022; Desplanques et al., 2020; Ravanelli et al., 2021; weon Jung et al., 2020; Snyder et al., 2018; Nagrani et al., 2020). The baselines were pre-trained for speaker identification. They accept speech recordings as input and return a speaker embedding. For speaker verification, we input each trial pair recording to the baseline. We compute the cosine similarity score from the obtained embeddings to determine if the recordings are of the same speaker. The X-Vector, ECAPA-TDNN, and RawNet-3 models were trained on combined VoxCeleb-1 and VoxCeleb-2 dev. VGG-M and RawNet-2 were trained on VoxCeleb-1 dev and VoxCeleb-2 dev, respectively.

4.4 Evaluation Metric

Equal Error Rate (EER) is the standard evaluation metric for speaker verification systems (Hansen and

Model	$tt-tt$	$ts-tt$	$ts-ts$	$tt-ts$	$ss-ss$	$ss-st$	$st-ss$
VGG-M (Baseline)	11.40	26.15	22.42	9.47	10.35	**7.90**	**28.06**
X-Vector (Baseline)	6.75	20.38	17.43	5.85	6.92	**5.25**	**22.19**
ECAPA-TDNN (Baseline)	12.46	20.93	19.57	11.96	11.40	**9.30**	**22.65**
RawNet-2 (Baseline)	38.24	**41.48**	39.21	**36.87**	37.90	37.00	39.80
RawNet-3 (Baseline)	41.34	**52.17**	46.54	**36.75**	41.71	44.10	43.60
ResNet+ (Hypothesis)	10.72	**13.55**	12.27	9.81	**9.51**	9.55	12.08
EcoSpeak (Scheme-A)	8.54	**13.88**	12.80	7.64	7.70	**7.37**	13.66
EcoSpeak (Scheme-B)	7.70	12.01	**12.65**	8.09	**7.23**	7.61	11.87
EcoSpeak (Scheme-C)	7.31	9.32	**11.16**	9.06	**6.81**	8.18	9.65

Table 2: Table showing the EER (%) of baselines, ResNet+, and EcoSpeak on the LRL test sets. We have represented each model's best and worst-case performance using bold font. Observations: 1.) Baselines performed the worst in ts - tt or st - ss. 2.) ResNet+ is stable compared to baselines. 3.) EcoSpeak (Scheme-C) performs better than other models on most test sets. It performed the worst in ts - ts, which deviates from the worst-case of baselines.

Hasan, 2015). EER is the value of the False Match Rate (FMR) and False Non-Match Rate (FNMR) when they are equal. FMR refers to the proportion of negative trial pairs incorrectly classified as positive by the system. In contrast, FNMR is the proportion of positive trial pairs incorrectly classified as negative by the system. EER is the value of the FMR when it becomes equal to the FNMR at a particular classification threshold. We used the EER to demonstrate the efficacy of this work. A lower EER indicates a better performance.

5 Experiments and Results

5.1 Baseline Behavioral Insights

The first step towards bias mitigation involves understanding the error patterns in baselines (Choe et al., 2022). Therefore, we examined the performance of baselines on the LRL test sets. As illustrated in Table 2, we observed elevated EER values on $ts-tt$ and $st-ss$. It indicates that a high-linguistic similarity in negative trial pair recordings (tt or ss) leads to performance degradation. This observation suggests that high-linguistic similarity makes the model favor the positive class. Likewise, a low-linguistic similarity in positive trial pairs (ts or st) also leads to performance degradation. This observation suggests that low-linguistic similarity makes the model favor the negative class. Furthermore, we observed lower EER values on the $tt-ts$ and $ss-st$ test sets. It indicates that the baselines perform the best when positive trial pair recordings have high linguistic similarity (ss or tt) and negative trial pair recordings have low linguistic similarity (ts or st). These observations indicate that the linguistic similarity in the trial pair influences the decision of baselines.

Key Observations:

1. We observed elevated EER values on $ts-tt$ and $st-ss$. Thus, baselines performed the worst on these test sets. It indicates that linguistic mismatch (ts or st) in positive trial pair recordings and linguistic match (tt or ss) in negative trial pair recordings causes performance degradation. Accordingly, we classify Positive-ts, Positive-st, Negative-tt, and Negative-ss as complex trial pair types.
2. We observed lower EER values on $tt-ts$ and $ss-st$. Thus, baselines performed the best on these test sets. It indicates that linguistic match (tt or ss) in positive trial pair recordings and linguistic mismatch (ts or st) in negative trial pair recordings leads to better baseline performance. Accordingly, we classify Positive-tt, Positive-ss, Negative-ts, and Negative-st as simple trial pair types.

5.2 Behavior of Residual Connections

Next, we investigated the impact of residual connections on cross-lingual testing. For this, we evaluated ResNet+ on LRL test sets. ResNet+ contains 64, 128, 256, and 512 channels for its first, second, third, and fourth layers, whereas ResNet (Lite) contains 32, 64, 128, and 256 channels. We compared the absolute difference between the models' highest and lowest EER scores on the LRL test sets. Table 2 illustrates that we achieved an EER difference of 4.04% (i.e., 13.55-9.51) using ResNet+. This difference is significantly less than that in most baselines. The EER differences for VGG-M, X-vector, ECAPA-TDNN, and RawNet-3 are 20.16%, 16.94%, 13.35% and 15.42%. It

indicates that the ResNet+ is more stable than baselines on LRL test sets. Next, we compared the EER values achieved using the RawNet models on the VoxCeleb-1 and LRL test sets. We achieved EER values of 3.67% and 1.11% on the VoxCeleb-1 test set using RawNet-2 and RawNet-3. However, the performance of RawNet models significantly degraded on the NISP-LRL test sets with more than a 30% increase in EER values. In contrast, using ResNet+, we achieved an EER score of 9.97% on the VoxCeleb-1 test set. This value is closer to the ResNet+ results on LRL test sets. However, ECAPA-TDNN, RawNet-2, and RawNet-3 also incorporate residual connections in their architectures yet have demonstrated high linguistic bias on LRL test sets. It indicates that residual connections alone are not sufficient for bias mitigation.

Summary of Findings: ResNet+ is less linguistically biased than baselines. Thus, residual connections can help mitigate linguistic bias. However, residual connections alone are insufficient for bias mitigation.

5.3 Data Balancing Schemes

Focusing on the quality of training data rather than quantity can help mitigate linguistic bias cost-efficiently (Swayamdipta et al., 2020). Therefore, to explore the impact of data balancing in partially cross-lingual speaker verification, we investigated three data balancing schemes for fine-tuning EcoSpeak. These schemes involve training sets having different distributions of simple and complex trial pairs.

Methodology We created six trial pair types for fine-tuning EcoSpeak: Positive-ts, Positive-tt, Positive-ss, Negative-ts, Negative-tt, Negative-ss. Here, positive and negative indicate whether trial pair recordings are of the same speaker. The notations tt, ss, and ts indicate whether the trial pair recordings are in the same (tt or ss) or different languages (ts). Furthermore, our baseline behavioral insights reveal that Positive-ts, Negative-tt, and Negative-ss are complex trial pair types. In contrast, Positive-tt, Positive-ss, and Negative-ts are the simpler trial pair types. Accordingly, we investigated the following data balancing schemes:

1. *Scheme-A:* In Scheme-A, we generate 200,000 examples for each trial pair type.
2. *Scheme-B:* In Scheme-B, we generate 250,000 and 150,000 examples for each complex and easy trial pair type.
3. *Scheme-C:* In Scheme-C, we generate 300,000 and 100,000 examples for each complex and easy trial pair type.

Accordingly, we created 1,200,000 trial pairs for each scheme, thus obtaining separate training sets for each scheme. We fine-tuned EcoSpeak on the NISP-Hindi speaker data using these scheme-specific training sets. Consequently, we got three EcoSpeak models, one for each scheme. Furthermore, we evaluated the performance of these EcoSpeak models on the LRL test sets without domain adaptation. LRL test sets contain speech recordings of native speakers of Tamil, Telugu, Malayalam, and Kannada.

Observations: We compared the absolute differences between the best-worst case EER values of the three scheme-specific EcoSpeak models. As illustrated in Table 2, we noticed absolute differences of 6.51% (13.88-7.37), 5.42% (12.65-7.23), and 4.35% (11.16-6.81) using Scheme-A, Scheme-B, and Scheme-C. Thus, we achieved the most stable results using Scheme-C. The training set for Scheme-C contains more examples from the complex trial pair type. It suggests that appropriate data balancing schemes can cost-efficiently aid bias mitigation. Furthermore, contrary to what we observed in baselines, EcoSpeak (Scheme-C) performed the worst in $ts-ts$ (and not in $ts-tt$ or $st-ss$). This observation indicates that the performance trend of EcoSpeak deviates from baselines.

5.4 Dataset for fine-tuning EcoSpeak

Due to data scarcity in low-resource target languages, finding appropriate datasets for fine-tuning models is challenging. Therefore, we explored two fine-tuning options for EcoSpeak:
1.) Fine-tuning on weakly related but diverse data.
2.) Fine-tuning on strongly related but small data.

Methodology: For this experiment, we chose Tamil as the low-resource target language (t). We utilized the LRL test sets to create Tamil-LRL test sets. Tamil-LRL test sets include those LRL test set trial pairs that contain speech recordings of only Tamil native speakers. Thus, we got seven Tamil-LRL test sets containing 25,000 trial pairs each.

NISP-Hindi is a diverse dataset (103 speakers), but Hindi is weakly related to Tamil. NISP-Telugu, NISP-Malayalam, and NISP-Kannada are small (fewer speakers) datasets with speech recordings from 60 speakers each. However, these LRLs are strongly related to Tamil. Subsequently, we fine-

Test Set	EcoSpeak-Hindi	EcoSpeak-Telugu	EcoSpeak-Malayalam	EcoSpeak-Kannada
$tt - tt$	**8.31**	9.70	9.98	10.36
$ts - tt$	**10.25**	14.57	13.44	12.97
$ts - ts$	**11.42**	15.94	15.78	14.34
$tt - ts$	**8.86**	10.51	11.61	12.43
$ss - ss$	**6.26**	8.18	9.05	9.85
$ss - st$	**7.42**	10.26	11.04	12.46
$st - ss$	**8.94**	12.63	12.49	11.17

Table 3: Table showing the EER values (%) on Tamil-LRL test sets. The EcoSpeak model fine-tuned on NISP-Hindi native speaker data performed the best. NISP-Hindi is a diverse dataset, but Hindi is weakly related to Tamil.

Model	#Parameters	Size (MB)	Time (sec)	CO_2 (kgCO2eq)	Electricity (kWh)
RawNet-3	16,280,322	62.30	4000	0.46	0.73
ECAPA-TDNN	22,150,912	85.00	2195	0.23	0.36
RawNet-2	13,379,378	51.10	1360	0.13	0.20
VGG-M	17,909,219	68.40	1252	0.11	0.18
X-Vector	8,172,473	31.50	**1014**	**0.09**	**0.14**
EcoSpeak	**6,660,233**	**25.50**	1165	0.10	0.16

Table 4: Table comparing the cost of EcoSpeak with baselines. The model size and number of parameters reported for EcoSpeak include the size and parameters of s-Detect. The time, carbon emissions, and electricity consumption statistics reported in the table represent the inference cost on the $tt - tt$ LRL test set.

tuned EcoSpeak on these datasets to get EcoSpeak-Hindi, EcoSpeak-Telugu, EcoSpeak-Malayalam, and EcoSpeak-Kannada models.

Observations: Table 3 illustrates the EER values we achieved on the Tamil-LRL test sets using EcoSpeak-Hindi, EcoSpeak-Telugu, EcoSpeak-Malayalam, and EcoSpeak-Kannada. We observed lower EER using the EcoSpeak-Hindi model compared to other EcoSpeak models. Thus, the EcoSpeak-Hindi model performed the best on Tamil-LRL test sets. It indicates that fine-tuning on a weakly related diverse dataset can be better than fine-tuning on a strongly related limited dataset. Overfitting on small datasets can reduce the model's generalization ability.

5.5 Cost Analysis

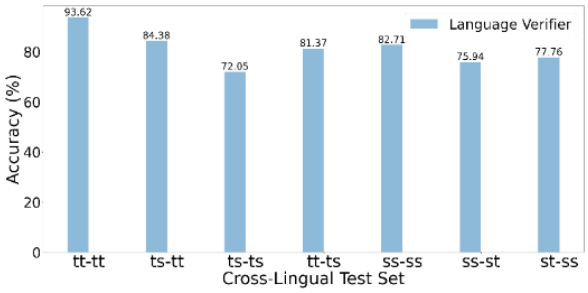
This work aims to investigate cost-efficient solutions to partially cross-lingual speaker verification. Therefore, we compared the costs associated with the baselines and our proposed EcoSpeak. We focussed on the model size, number of parameters, and the inference costs (time, carbon emission, and electricity consumption) of these models. Table 4 illustrates that EcoSpeak has a lesser model size and number of parameters than the baselines. Furthermore, we compared the inference costs of the EcoSpeak-Hindi model with the baselines on the $tt - tt$ LRL test set. Table 4 demonstrates that EcoSpeak-Hindi takes less inference time than most baselines. Additionally, we observed lower carbon emissions and electricity consumption from EcoSpeak-Hindi compared to most baselines when tested on the $tt - tt$ LRL test set. Table 4 shows that EcoSpeak's inference cost is comparable to the X-Vector model. However, Table 2 demonstrates that EcoSpeak is more stable than X-vector on the LRL test sets. It is because EcoSpeak (Scheme-C) shows an EER variation of 4.35% (i.e., 11.16-6.81) on the LRL test sets. In contrast, X-Vector shows an EER difference of 16.94% (i.e., 22.19-5.25) on the LRL test sets. Therefore, our findings indicate that EcoSpeak is a cost-efficient solution to partially cross-lingual speaker verification.

5.6 Ablation Study

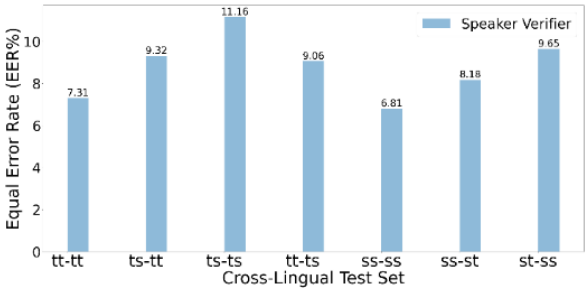
To analyze EcoSpeak results, we did an ablation study, as shown in Table 5. Firstly, we observed that *ResNet (Lite)* performs better than ResNet+ on the LRL test sets. Furthermore, it is lighter than ResNet+. In EcoSpeak, we chose ResNet (Lite) to extract speaker embeddings from the trial pair recordings. Next, instead of cosine similarity, we used fully connected layers for speaker verification in *ResNet (Lite)+fc*. We fed the absolute difference of the trial pair ResNet (Lite) embeddings

Model	$tt-tt$	$ts-tt$	$ts-ts$	$tt-ts$	$ss-ss$	$ss-st$	$st-ss$
ResNet+	10.72	**13.55**	12.27	9.81	**9.51**	9.55	12.08
ResNet (Lite)	9.52	**12.14**	10.96	**8.33**	8.54	9.13	10.58
ResNet (Lite)+fc	11.16	**14.72**	13.87	10.67	10.57	**10.29**	13.76
CL Attention	7.47	9.29	**11.70**	9.67	**6.94**	8.48	9.88
EcoSpeak	7.31	9.32	**11.16**	9.06	**6.81**	8.18	9.65

Table 5: Ablation study results for EcoSpeak. Observation: CL attention mitigates linguistic bias.



a) Language Verification



b) Speaker Verification

Figure 2: Figure shows a negative correlation between EcoSpeak's language and speaker verification performance. High language verification accuracy causes a low EER in speaker verification and vice-versa.

to the fully connected layer. The fully-connected layers were fine-tuned on the NISP-Hindi native speaker data using Scheme-C. The poor performance of ResNet (Lite)+fc indicates that data balancing alone is insufficient for bias mitigation. Still, we fine-tuned the CL attention model and EcoSpeak using Scheme-C, described in Section 5.3.

The *CL attention* model outperformed ResNet (Lite) on most LRL test sets. Interestingly, we observed significant improvements in the two most challenging partially cross-lingual scenarios, $ts-tt$ and $st-ss$. It suggests that the CL attention effectively emphasizes or de-emphasizes speaker verification embeddings based on linguistic differences in the trial pair recordings. *EcoSpeak* incorporates the CL attention and the bias corrector. It performed better than the CL attention model on most LRL test sets. EcoSpeak performed the worst on $ts-ts$. The reason is that EcoSpeak's performance in language verification affects its performance in speaker verification, as evidenced by Figure 2. EcoSpeak's higher language verification accuracy causes a lower EER score in speaker verification and vice-versa. The model performed the worst for language verification on $ts-ts$. It justifies EcoSpeak's worst-case speaker verification performance on $ts-ts$.

6 Conclusions and Future Work

This paper investigates the behavior of five baseline speaker verification models on five partially cross-lingual test sets. Empirical results demonstrate that a high linguistic similarity in negative trial pair recordings and a low linguistic similarity in positive trial pair recordings causes performance degradation. Furthermore, residual networks are relatively robust to cross-lingual testing. Using these insights, we proposed EcoSpeak, a low-cost solution to mitigate bias in partially cross-lingual speaker verification. EcoSpeak incorporates residual connections, contrastive linguistic attention, and the bias corrector. Empirical results demonstrate the robustness of our proposed model on partially cross-lingual test sets. EcoSpeak's performance trend deviates from the baselines. It turns out that utilizing linguistic differences to emphasize and de-emphasize relevant speaker verification embedding parts can mitigate partially cross-lingual bias.

Our insights can contribute to the development of more robust domain-invariant architectures. Furthermore, this work encourages the community to explore greener approaches to expensive speech processing. For instance, based on our empirical results, we recommend leveraging diverse datasets in a weakly related language for bias mitigation in an unseen low-resource target language. Additionally, our proposed data balancing schemes can save the cost of training on large-scale datasets. We also recommend a detailed cost analysis to develop environment-friendly and inclusive models.

7 Limitations

This work explores cost-efficient techniques for bias mitigation in partially cross-lingual speaker verification. Our proposed approach has the following limitations:

1. **Correlation between speaker and language verification performance:** EcoSpeak's performance on the language verification task affects its performance on the speaker verification task. One way to address this limitation is to use a more robust s-Detect model. It is because EcoSpeak accepts s-Detect embeddings as prompt inputs for CL attention and language verification. Therefore, having a more robust s-Detect can enhance the speaker verification performance of EcoSpeak.

2. **More experimental validation for contrastive linguistic (CL) Attention:** Our proposed CL attention mechanism relies on the intuition that the learned CL attention weights shall correlate with the speaker verification embeddings. Therefore, modulating the CL attention weights with the speaker verification embedding emphasizes those embedding parts that are more influenced by linguistic variations. However, CL attention is a relatively new approach. In this study, we experimented on five partially cross-lingual test sets created for four low-resource languages. Still, extensive experimental validation in more languages is required to validate the effectiveness of CL attention.

3. **Low-resource language datasets used to train s-Detect**: We did not explicitly fine-tune EcoSpeak on the target low-resource languages (Tamil, Telugu, Malayalam, and Kannada). However, we used speech recordings from different datasets in the target low-resource languages to train the s-Detect model. Nevertheless, this approach is practical as it is easier to find language identification datasets than cross-lingual datasets of bilingual speakers for speaker verification.

Linguistic bias is a complex problem to address using a single bias mitigation technique. This work offers a combination of low-cost bias mitigation techniques in the form of EcoSpeak. In the future, combining our proposed techniques with other novel ideas can further aid bias mitigation.

References

AI4Bharat. 2020. Nptel2020-indian-english-speech-dataset. https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset.git.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Zhengyang Chen, Shuai Wang, and Yanmin Qian. 2020. Adversarial Domain Adaptation for Speaker Verification Using Partially Shared Network. In *Proc. Interspeech 2020*, pages 3017–3021.

June Choe, Yiran Chen, May Pik Yu Chan, Aini Li, Xin Gao, and Nicole Holliday. 2022. Language-specific effects on automatic speech recognition errors for world englishes. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7177–7186, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Roza Chojnacka, Jason Pelecanos, Quan Wang, and Ignacio Lopez Moreno. 2021. SpeakerStew: Scaling to Many Languages with a Triaged Multilingual Text-Dependent and Text-Independent Speaker Verification System. In *Proc. Interspeech 2021*, pages 1064–1068.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proc. Interspeech 2018*, pages 1086–1090.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, pages 3830–3834.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, and Karthik Sankaranarayanan. 2021. MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In *Proc. Interspeech 2021*, pages 2446–2450.

Mariel Estevez and Luciana Ferrer. 2023. Study on the fairness of speaker verification systems across accent and gender groups. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

John H.L. Hansen and Taufiq Hasan. 2015. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6):74–99.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu speech synthesis systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6494–6503, Marseille, France. European Language Resources Association.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Shareef Babu Kalluri, Deepu Vijayasenan, Sriram Ganapathy, Ragesh Rajan M, and Prashant Krishnan. 2021. Nisp: A multi-lingual multi-accent dataset for speaker profiling. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6953–6957.

Kong Aik Lee, Koji Okabe, Hitoshi Yamamoto, Qiongqiong Wang, Ling Guo, Takafumi Koshinaka, Jiacen Zhang, Keisuke Ishikawa, and Koichi Shinoda. 2020. NEC-TT Speaker Verification System for SRE'19 CTS Challenge. In *Proc. Interspeech 2020*, pages 2227–2231.

Damien Lesenfants, J. Vanthornhout, Eline Verschueren, Lien Decruy, and Tom Francart. 2019. Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations. *Hearing Research*, 380.

Jingyu Li, Wei Liu, and Tan Lee. 2022. EDITnet: A Lightweight Network for Unsupervised Domain Adaptation in Speaker Verification. In *Proc. Interspeech 2022*, pages 3694–3698.

Liang Lu, Yuan Dong, Xianyu Zhao, Jiqing Liu, and Haila Wang. 2009. The effect of language factors for robust speaker recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4217–4220.

Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. pages 18–24.

Microsoft. 2023. Microsoft speech corpus (indian languages). https://www.microsoft.com/en-za/download/details.aspx?id=105292.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech Language*, 60:101027.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proc. Interspeech 2017*, pages 2616–2620.

Mor Nahum, Israel Nelken, and Merav Ahissar. 2008. Low-level information and high-level perception: The case of speech in noise. *PLoS biology*, 6:e126.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA.

Xiaoyi Qin, Chao Wang, Yong Ma, Min Liu, Shilei Zhang, and Ming Li. 2021. Our Learned Lessons from Cross-Lingual Speaker Verification: The CRMI-DKU System Description for the Short-Duration Speaker Verification Challenge 2021. In *Proc. Interspeech 2021*, pages 2317–2321.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

Johan Rohdin, Themos Stafylakis, Anna Silnova, Hossein Zeinali, Lukáš Burget, and Oldřich Plchot. 2019. Speaker verification using end-to-end adversarial language adaptation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6006–6010.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM*, 63(12):54–63.

Divya Sharma and Arun Balaji Buduru. 2022. FAt-Net: Cost-effective approach towards mitigating the linguistic bias in speaker verification systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1247–1258, Seattle, United States. Association for Computational Linguistics.

David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. Spoken Language Recognition using X-vectors. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, pages 105–111.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck. 2020. Cross-Lingual Speaker Verification with Domain-Balanced Hard Prototype Mining and Language-Dependent Score Normalization. In *Proc. Interspeech 2020*, pages 756–760.

Youzhi Tu, Man-Wai Mak, and Jen-Tzung Chien. 2019. Variational Domain Adversarial Learning for Speaker Verification. In *Proc. Interspeech 2019*, pages 4315–4319.

Xueyi Wang, Lantian Li, and Dong Wang. 2019. Vae-based domain adaptation for speaker verification. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 535–539.

Jee weon Jung, Seung bin Kim, Hye jin Shim, Ju ho Kim, and Ha-Jin Yu. 2020. Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms. In *Proc. Interspeech 2020*, pages 1496–1500.

Jee weon Jung, Youjin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. 2022. Pushing the limits of raw waveform speaker recognition. In *Proc. Interspeech 2022*, pages 2228–2232.

Yi-Chieh Wu and Wen-Hung Liao. 2021. Toward text-independent cross-lingual speaker recognition using english-mandarin-taiwanese dataset. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8515–8522.

Wei Xia, Jing Huang, and John H.L. Hansen. 2019. Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5816–5820.

Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. 2021. A survey on green deep learning. *ArXiv*, abs/2111.05193.

Seunghan Yang, Debasmit Das, Janghoon Cho, Hyoungwoo Park, and Sungrack Yun. 2022. Domain Agnostic Few-shot Learning for Speaker Verification. In *Proc. Interspeech 2022*, pages 595–599.

Hossein Zeinali, Kong-Aik Lee, Jahangir Alam, and Luká Burget. 2019. Short-duration speaker verification (sdsv) challenge 2020: the challenge evaluation plan. *ArXiv*, abs/1912.06311.

Yi Zhou, Xiaohai Tian, and Haizhou Li. 2021. Language agnostic speaker embedding for cross-lingual personalized speech generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3427–3439.

Donghui Zhu and Ning Chen. 2022. Multi-source domain adaptation and fusion for speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2103–2116.

A ResNet (Lite) Architecture

Table 6 shows the details of ResNet (Lite).

Layer	Input shape	Output shape
conv1	[b, 1, 301, 64]	[b, 32, 297, 60]
maxpool1	[b, 32, 297, 60]	[b, 32, 149, 30]
layer1	[b, 32, 149, 30]	[b, 32, 149, 30]
layer2	[b, 32, 149, 30]	[b, 64, 38, 8]
layer3	[b, 64, 38, 8]	[b, 128, 10, 2]
layer4	[b, 128, 10, 2]	[b, 256, 3, 1]
avgpool	[b, 256, 3, 1]	[b, 256, 1, 1]
fc1	[b, 256]	[b, 512]
fc2	[b, 512]	[b, num_speakers]

Table 6: Architecture details of the ResNet (Lite) speaker identification model.

B Training Setup

Firstly, we trained the ResNet (Lite) for speaker identification on VoxCeleb-2 dev. It took about 40 minutes for the completion of one epoch. One epoch caused 0.18 kgCO2eq carbon emissions and consumed 0.61 kWh of electricity. We trained the model for ten epochs. Secondly, we trained the s-Detect model to detect the source language (English). As described in Section 4.1, we combined speech recordings from different datasets to train s-Detect. We collected 23856, 24884, 20207, 1983, 3633, and 74563 speech recordings in Hindi, Tamil, Telugu, Malayalam, Kannada, and English. Next, we fine-tuned the s-Detect on the NISP-Hindi speaker data so that the EcoSpeak-Hindi model could adapt to the dataset-specific variations of NISP. We combined the NISP-Hindi speaker data with the s-Detect training set and used mixed training to fine-tune s-Detect. Finally, we used this adapted s-Detect to train the EcoSpeak-Hindi model on the NISP-Hindi speaker data. We fine-tuned EcoSpeak-Hindi for four Epochs to prevent overfitting due to data limitations. We froze the EcoSpeak's ResNet (Lite) weights during fine-tuning. We used the CrossEntropyLoss, Adam optimizer, and a learning rate 0.0005. We followed the same procedure to train the EcoSpeak-Tamil, EcoSpeak-Telugu, EcoSpeak-Malayalam, and EcoSpeak-Kannada models. We used one NVIDIA A100 GPU. We also used Librosa for pre-processing and feature extraction, Pytorch for model training, and CodeCarbon for tracking carbon emissions and electricity consumption (McFee

et al., 2015; Paszke et al., 2019).[6]

Dataset	Test Set	RC	RAD
NISP-LRL	$tt-tt$	10.11	**8.16**
	$ts-tt$	13.26	**12.91**
	$ts-ts$	**11.96**	12.08
	$tt-ts$	9.15	**7.78**
	$ss-ss$	8.70	**7.88**
	$ss-st$	9.96	**8.32**
	$st-ss$	11.30	11.30
NISP-Tamil	$tt-tt$	12.38	**10.72**
	$ts-tt$	**14.98**	15.29
	$ts-ts$	**13.47**	14.69
	$tt-ts$	11.14	**10.24**
	$ss-ss$	8.94	**8.11**
	$ss-st$	9.69	**8.88**
	$st-ss$	**12.39**	12.66
NISP-Telugu	$tt-tt$	9.61	**7.41**
	$ts-tt$	10.81	**9.83**
	$ts-ts$	10.07	**9.58**
	$tt-ts$	8.98	**7.15**
	$ss-ss$	10.10	**9.22**
	$ss-st$	10.62	**8.04**
	$st-ss$	8.71	**8.46**
NISP-Malayalam	$tt-tt$	7.24	**6.36**
	$ts-tt$	11.21	**11.12**
	$ts-ts$	**9.28**	9.69
	$tt-ts$	6.17	**5.61**
	$ss-ss$	**6.56**	6.79
	$ss-st$	7.68	**7.58**
	$st-ss$	**8.99**	9.44
NISP-Kannada	$tt-tt$	10.64	**7.97**
	$ts-tt$	15.61	**15.26**
	$ts-ts$	14.05	**13.97**
	$tt-ts$	9.78	**7.98**
	$ss-ss$	8.12	**7.32**
	$ss-st$	10.55	**8.50**
	$st-ss$	**14.13**	14.39

Table 7: EER (%) values on evaluating the RC and RAD on different LRL test sets. The RAD model outperformed RC on most test sets. It justifies our use of the absolute difference in EcoSpeak.

C Absolute Difference in EcoSpeak

EcoSpeak uses the absolute difference operation to compare trial pair embeddings (emb_1, emb_2) for speaker verification. This section describes the experiment that motivated us to use the absolute difference. We compared the following models:

ResNet (Lite)-Concat: In the ResNet (Lite)-Concat model (RC), we feed each trial pair recording through the ResNet (Lite) to get 256-dimensional embeddings (emb_1, emb_2). We concatenate these embeddings to get a 512-dimensional embedding. We feed this concatenated embedding through two fully connected layers having 512 units. Finally, we pass the resulting embedding through a fully connected layer consisting of two units for speaker verification. This model occupies 22.4 MB of disk space.

ResNet(Lite)-AbsoluteDifference: In the ResNet(Lite)-AbsoluteDifference (RAD) model, we feed each trial pair recording through the ResNet (Lite) to get 256-dimensional embeddings (emb_1, emb_2). We compute the absolute difference of these embeddings to get a 256-dimensional embedding. We feed this difference embedding through two fully connected layers having 256 units. Finally, we pass the resulting embedding through a fully connected layer consisting of two units for speaker verification. This model occupies 20.9 MB of disk space.

Language-Specific LRL test sets: We created separate test sets for each LRL under consideration. Thus, we got Tamil-LRL, Telugu-LRL, Malayalam-LRL, and Kannada-LRL test sets. These test sets are the subsets of the original LRL test sets described in Section 4.1. They include trial pairs of native speakers of these languages. Thus, each test set in Tamil-LRL, Telugu-LRL, Malayalam-LRL, and Kannada-LRL contains 25,000 trial pairs.

Observation: We compared the performance of RC and RAD on the LRL test sets (described in Section 4.2) and the language-specific LRLs. Table 7 illustrates that we achieved lower EER values using RAD than RC on most test sets. This observation motivated us to use the absolute difference operation in EcoSpeak.

D Extensive Experimental Validation

Table 8, Table 9, Table 10 and Table 11 illustrates the performance of the baselines, ResNet+ and EcoSpeak on the Tamil, Telugu, Malayalam, and Kannada LRL test sets described in Section C. These tables validate our observations in Sections 5.1, 5.2, and 5.3. Table 12, Table 13 and Table 14 demonstrate the result of the experiment described in Section 5.4 on the Telugu, Malayalam, and Kannada LRL test sets.

[6]https://pypi.org/project/codecarbon/

Model	$tt-tt$	$ts-tt$	$ts-ts$	$tt-ts$	$ss-ss$	$ss-st$	$st-ss$
VGG-M (Baseline)	11.25	31.65	24.68	8.35	8.99	**6.72**	**31.72**
X-Vector (Baseline)	6.18	**26.18**	21.13	4.86	5.18	**4.20**	24.61
ECAPA-TDNN (Baseline)	12.11	**24.73**	22.00	10.75	11.64	**10.17**	24.10
RawNet-2 (Baseline)	38.50	**39.86**	37.78	37.09	34.37	**33.75**	38.63
RawNet-3 (Baseline)	41.13	**54.61**	46.62	**33.92**	41.86	47.33	40.99
ResNet+ (Hypothesis)	12.90	**17.47**	14.26	10.10	**9.24**	10.08	14.00
EcoSpeak (Scheme-A)	9.82	**16.35**	13.94	7.40	6.80	**6.37**	13.18
EcoSpeak (Scheme-B)	8.76	**14.41**	14.09	8.06	**6.69**	7.14	12.44
EcoSpeak (Scheme-C)	8.31	10.25	**11.42**	8.86	**6.26**	7.42	8.94

Table 8: Table showing the EER (%) of baselines, ResNet+, and EcoSpeak on the Tamil-LRL test sets. We have represented each model's best and worst-case performance using bold font. Observations: 1.) Baselines performed the worst in ts - tt or st - ss. 2.) ResNet+ is stable compared to baselines. 3.) EcoSpeak (Scheme-C) performs better than other models on most test sets. It performed the worst in ts - ts, which deviates from the worst-case of baselines.

Model	$tt-tt$	$ts-tt$	$ts-ts$	$tt-ts$	$ss-ss$	$ss-st$	$st-ss$
VGG-M (Baseline)	9.90	23.53	20.94	**8.51**	12.56	10.31	**25.67**
X-Vector (Baseline)	6.90	17.54	15.46	6.22	7.92	**5.27**	**21.80**
ECAPA-TDNN (Baseline)	12.50	19.22	19.14	12.10	11.73	**8.79**	**22.54**
RawNet-2 (Baseline)	37.22	**40.63**	38.94	**35.98**	38.14	36.48	39.02
RawNet-3 (Baseline)	40.83	**49.94**	43.78	**34.90**	42.98	44.50	40.75
ResNet+ (Hypothesis)	10.45	**10.41**	11.61	10.63	**12.40**	10.61	10.52
EcoSpeak (Scheme-A)	7.82	**11.43**	11.20	**6.82**	8.32	7.42	11.25
EcoSpeak (Scheme-B)	6.45	9.18	**10.34**	**6.27**	7.59	6.63	9.24
EcoSpeak (Scheme-C)	**6.42**	7.09	**9.30**	7.82	7.40	7.38	7.76

Table 9: Table showing the EER (%) of baselines, ResNet+, and EcoSpeak on the Telugu-LRL test sets. We have represented each model's best and worst-case performance using bold font. Observations: 1.) Baselines performed the worst in ts - tt or st - ss. 2.) ResNet+ is stable compared to baselines. 3.) EcoSpeak (Scheme-C) performs better than other models on most test sets. It performed the worst in ts - ts, which deviates from the worst-case of baselines.

Model	$tt-tt$	$ts-tt$	$ts-ts$	$tt-ts$	$ss-ss$	$ss-st$	$st-ss$
VGG-M (Baseline)	10.09	22.63	18.90	8.18	9.92	**7.29**	**24.18**
X-Vector (Baseline)	6.83	16.76	14.30	**5.71**	7.34	5.85	**18.88**
ECAPA-TDNN (Baseline)	12.05	17.02	16.58	11.91	10.95	**9.28**	**19.29**
RawNet-2 (Baseline)	37.37	**41.28**	38.94	**35.50**	38.44	37.86	39.69
RawNet-3 (Baseline)	42.99	**54.00**	49.98	**39.00**	39.73	40.69	48.83
ResNet+ (Hypothesis)	8.88	**12.42**	11.06	8.74	**8.19**	8.74	11.14
EcoSpeak (Scheme-A)	7.97	12.93	12.17	8.10	8.06	**7.70**	**13.41**
EcoSpeak (Scheme-B)	**6.98**	11.38	**12.08**	8.12	8.26	8.23	11.90
EcoSpeak (Scheme-C)	**6.78**	9.89	**11.51**	8.90	6.96	8.27	10.10

Table 10: Table showing the EER (%) of baselines, ResNet+, and EcoSpeak on the Malayalam-LRL test sets. We have represented each model's best and worst-case performance using bold font. Observations: 1.) Baselines performed the worst in ts - tt or st - ss. 2.) ResNet+ is stable compared to baselines. 3.) EcoSpeak (Scheme-C) performs better than other models on most test sets. It performed the worst in ts - ts, which deviates from the worst-case of baselines.

Model	$tt-tt$	$ts-tt$	$ts-ts$	$tt-ts$	$ss-ss$	$ss-st$	$st-ss$
VGG-M (Baseline)	11.63	23.70	20.02	10.12	8.68	**5.46**	**27.74**
X-Vector (Baseline)	6.38	21.14	16.80	5.76	5.92	**5.10**	**21.67**
ECAPA-TDNN (Baseline)	13.02	22.75	20.13	13.10	11.17	**8.82**	**24.62**
RawNet-2 (Baseline)	40.00	**44.26**	41.16	**38.90**	40.79	39.97	41.87
RawNet-3 (Baseline)	40.30	**49.98**	45.68	**39.22**	42.86	43.82	44.13
ResNet+ (Hypothesis)	10.49	**13.19**	11.65	9.63	**8.06**	8.73	12.06
EcoSpeak (Scheme-A)	8.65	14.31	13.21	8.07	**7.71**	7.90	**15.68**
EcoSpeak (Scheme-B)	8.43	12.67	**13.54**	9.78	**6.49**	8.50	13.31
EcoSpeak (Scheme-C)	7.73	9.94	**12.04**	10.58	**6.51**	9.65	11.58

Table 11: Table showing the EER (%) of baselines, ResNet+, and EcoSpeak on the Kannada-LRL test sets. We have represented each model's best and worst-case performance using bold font. Observations: 1.) Baselines performed the worst in ts - tt or st - ss. 2.) ResNet+ is stable compared to baselines. 3.) EcoSpeak (Scheme-C) performs better than other models on most test sets. It performed the worst in ts - ts, which deviates from that of baselines.

Test Set	EcoSpeak-Hindi	EcoSpeak-Tamil	EcoSpeak-Malayalam	EcoSpeak-Kannada
$tt-tt$	**6.42**	6.73	7.30	8.41
$ts-tt$	**7.09**	7.99	9.09	9.41
$ts-ts$	**9.30**	10.02	10.45	12.22
$tt-ts$	7.82	8.19	**7.70**	10.46
$ss-ss$	**7.40**	8.26	7.81	9.51
$ss-st$	7.38	**7.20**	7.87	10.02
$st-ss$	**7.76**	8.47	9.02	10.09

Table 12: Table showing the EER values (%) on Telugu-LRL test sets. The EcoSpeak model fine-tuned on NISP-Hindi native speaker data performed the best on most test sets. NISP-Hindi is a diverse dataset, but Hindi is weakly related to Telugu.

Test Set	EcoSpeak-Hindi	EcoSpeak-Tamil	EcoSpeak-Telugu	EcoSpeak-Kannada
$tt-tt$	**6.78**	8.06	7.22	8.56
$ts-tt$	**9.89**	10.62	10.43	10.56
$ts-ts$	**11.51**	12.42	12.98	12.93
$tt-ts$	**8.90**	9.91	9.90	11.34
$ss-ss$	**6.96**	8.47	8.02	8.43
$ss-st$	**8.27**	10.63	9.11	11.30
$st-ss$	**10.10**	11.50	12.41	10.63

Table 13: Table showing the EER values (%) on Malayalam-LRL test sets. The EcoSpeak model fine-tuned on NISP-Hindi native speaker data performed the best. NISP-Hindi is a diverse dataset, but Hindi is weakly related to Malayalam.

Test Set	EcoSpeak-Hindi	EcoSpeak-Tamil	EcoSpeak-Malayalam	EcoSpeak-Telugu
$tt-tt$	**7.73**	8.20	8.73	8.79
$ts-tt$	9.94	10.81	11.94	**8.60**
$ts-ts$	12.04	13.02	14.08	**11.11**
$tt-ts$	10.58	**9.46**	11.03	10.62
$ss-ss$	6.51	**6.30**	6.98	7.51
$ss-st$	9.65	**8.33**	8.94	9.49
$st-ss$	11.58	12.98	12.78	**10.84**

Table 14: Table showing the EER values (%) on Kannada-LRL test sets. EcoSpeak-Hindi's poor performance on Kannada-LRL test sets is due to a lack of Kannada data for training s-Detect. Tamil, Telugu, and Kannada belong to the Dravidian language family and hence have similarities. Therefore, EcoSpeak-Tamil and EcoSpeak-Telugu performed better than EcoSpeak-Hindi on the Kannada-LRL test sets.