

Are audio DeepFake detection models polyglots?

Bartłomiej Marek¹, Piotr Kawa², Piotr Syga²

¹CISPA – Helmholtz Center for Information Security, Germany

²Wrocław University of Science and Technology, Poland

bartlomiej.marek@cispa.de, {piotr.kawa, piotr.syga}@pwr.edu.pl

Abstract

Since the majority of audio DeepFake (DF) detection methods are trained on English-centric datasets, their applicability to non-English languages remains largely unexplored. In this work, we introduce a benchmark for the multilingual audio DF detection challenge by evaluating various adaptation strategies. Our experiments focus on analyzing models trained on English benchmark datasets, as well as intra-linguistic (same-language) and cross-linguistic adaptation approaches. Our results indicate considerable variations in detection efficacy, highlighting the difficulties of multilingual settings. We show that limiting the training dataset to English negatively impacts the efficacy, while using even a small amount of data in the target language proves more beneficial for detection than adding larger volumes of data from multiple non-target languages combined.

Index Terms: Audio DeepFakes, DeepFake detection, multilingual audio DeepFakes

1. Introduction

The rapid growth of generative AI, especially in voice synthesis, has made it easier to create personalized voices. Technologies such as text-to-speech (TTS) and voice cloning (VC) need only seconds of voice input to produce convincing replicas [1]. While useful for personal assistants, these tools can also be misused, such as for audio DeepFakes (DF). Malicious DFs can undermine media credibility and enable harmful manipulation, from political disinformation such as AI-generated Polish campaign ads¹ to financial fraud, such as a \$25.6M scam in Hong Kong using executive impersonation².

Despite extensive research on DeepFake detection, a challenge similar to spoofing countermeasures [2], key issues persist, including limited diverse data and poor generalization [3]. Recent advancements [4, 5, 6] have democratized voice technology, enabling users to generate high-quality speech across languages. A Recorded Future Inc. report³ found 82 DeepFakes of public figures in 38 countries (July 2023–July 2024), with 30 of them holding elections, underscoring DFs’ global impact. Mitigation may require localized strategies, yet most models remain trained primarily on English and Chinese due to dataset availability.

The recent Multi-Language Audio Anti-Spoofing Dataset (MLAAD) [7] enables research on cross-language model generalization. Our work introduces a benchmark to evaluate multilingual DF detection, focusing on adaptation strategies: fine-

tuning, training from scratch, or using English pre-trained models to enhance performance across languages.

Our work extends prior studies [8, 9] by adding intra- and cross-lingual adaptations, language correlations, and investigating language families while limiting potential biases from overlapped synthesizers [10]. Specifically, we take a broader perspective by analyzing languages from three families: Germanic (English, German), Romance (French, Italian, Spanish), and Slavic (Polish, Russian, Ukrainian), with a focus on designing a benchmark that is as unbiased as possible using public data.

We aim to determine whether language-specific data are necessary for accurate detection and how to best adapt English-trained models when only a small dataset in the target language is available. We empirically explore three essential research questions (RQ) in this area. Specifically, our objective is to check to what extent the detection efficacy varies by language, whether English benchmark-trained models are sufficient for effective cross-linguistic detection, and which targeted strategies best support DF detection in specific languages, precisely intra- or cross-lingual adaptations, **even assuming access to very limited resources in a specific language**.

RQ1: To what extent can English-trained detection DFs models generalize to multilingual scenarios?

Current, publicly available benchmarks are English-centric, potentially leaving detection models underprepared for real-world scenarios involving non-English audio. This explores the challenges of relying solely on benchmark-trained models for diverse linguistic contexts, focusing on their performance in specific non-English languages.

RQ2: How does language choice influence DeepFake detection effectiveness?

Despite advances in audio DF detection, there is limited understanding of how language influences detection efficacy. This analyzes how detection effectiveness varies across languages and whether multilingual training data enhance or degrade performance compared to language-specific approaches.

RQ3: Which is the more effective adaptation strategy: language-specific with limited data, or multilingual with larger datasets?

We compare training a language-specific model with limited data, using an English-trained model, and fine-tuning it with language-specific or multilingual data. We aim to determine if a more targeted or diverse language dataset would be more effective for our strategy. In the latter case, it would be necessary to determine which languages are most suitable for adaptation.

The codebase related to our research can be found in⁴.

¹<https://notesfrompoland.com/2023/08/25/opposition-criticised-for-using-ai-generated-deepfake-voice-of-pm-in-polish-election-ad>

²<https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>

³<https://go.recordedfuture.com/hubfs/reports/ta-2024-0924.pdf>

⁴https://github.com/bartlomiejmarek/are_audio_df_polyglots

Table 1: Hours of training data used for W2V+AAISIST XLS-R 300m [11] and Whisper medium [12].

| Model | Languages | | | | | | | |
|----------------|-----------|--------|--------|--------|--------|--------|-------|-----|
| | de | en | es | fr | it | pl | ru | uk |
| W2V XLS-R 300m | 69 493 | 25 378 | 22 258 | 23 973 | 21 943 | 20 912 | 166 | 72 |
| Whisper | 438 218 | 13 344 | 11 100 | 9 752 | 2 585 | 4 278 | 9 761 | 697 |

2. Related works

DeepFake detection has gained attention as TTS and VC algorithms create increasingly realistic audio fakes. A substantial portion of the research in this domain has used ASVspoof datasets [13, 14, 15, 16], considered a gold standard for the anti-spoofing domain. However, these datasets do not fully represent real-world scenarios in terms of language coverage, as they exclusively consist of English samples.

The discrepancy between generation methods in the training and test sets leads to significant performance drops, especially in real-world settings. While [17] proposed a method for dealing with unseen methods of DF generation, [3] showed that detection models have even greater difficulty in the correct classification of real-world samples, publishing the "In the Wild" dataset. Moreover, most published datasets contain only English samples [18, 19, 20], with [18] including some Japanese utterances and [21, 22, 23, 24] providing Chinese ones.

Recent publications indicate that the effectiveness of detection methods significantly degrades across linguistic boundaries [25, 26]. The results suggest that detection systems exhibit substantial bias toward training language characteristics. Specifically, performance drops significantly when evaluated using unfamiliar languages or accents (even within the same language). [25]. On the other hand, the performance of detection methods is also heavily dependent on the overlap between the speech generators seen during training and those used in testing, thus highlighting the inability of existing models to generalize to unseen synthesis techniques [10, 26].

A recent Multi-Language Audio Anti-spoofing Dataset (MLAAD) [7] attempts to address the gap, providing over 76,000 utterances in 23 languages, with DF generated using 54 systems to present samples with varied distributions. This publicly available, large-scale audio fake corpus spans over 160 hours and forms a complete dataset with the M-AILABS Speech Dataset, consisting of authentic audio recordings from public domain books. For this research, we used the third version of the MLAAD dataset ⁵.

Notably, due to the large number of languages in the dataset, the number of samples in each language is limited. Given the problem with the generalization of the models, we cannot be sure if the efficacies of the detection models rely on the target language, i.e., the language in which the utterance is spoken. This motivated this paper so that further research could focus on the most promising way to detect non-English DFs.

3. Experimental setup

Throughout the article, we investigate the efficacy of English-trained detection models in detecting DFs for various languages and provide a strategy to improve the detection efficacy for DF utterances in languages with severely limited datasets that might be used to train or fine-tune the detection model. In this study, we concentrate on the languages represented in the M-AILABS Speech Dataset, which provides a comprehensive sample of au-

thentic language data in these languages. We have selected English (en) and German (de) from the Germanic family, French (fr), Italian (it), and Spanish (es), which represent the Romance languages, and Polish (pl), Russian (ru), and Ukrainian (uk), which represent the Slavic languages.

3.1. Models

Audio DF detection methods are based on either direct waveform analysis via end-to-end models [27, 28], or feature-based methods employing front-end extractors, which derive acoustic properties from raw audio for subsequent deep-learning [29, 30, 31]. Our research systematically examines the leading DF detection models [27, 30, 29, 32], assuming limited availability of non-English resources. Specifically, we evaluate the following scenarios: i) English-trained detection model on the target language, ii) training from scratch directly on the target language, iii) fine-tuning pre-trained English models on target languages, iv) fine-tuning with a single related or unrelated language, v) fine-tuning with multiple languages while excluding the target language. For all the experiments, we assume that non-English resources are severely limited. Given that two of our models utilize SSL architectures, we selected a pretrained W2V XLS-R 300m [11] and Whisper medium [12]. In Table 1 we can observe the amount of data used to train the specific language.

3.2. Datasets

We establish a baseline by utilizing models trained on the entire English benchmark dataset ASVspoof2019 LA, consisting of the training, development, and evaluation sets, and treat them as a reference point for language-specific fine-tuning, as most researched audio DFs detection solutions used English or Chinese for training (some of the latter's sets, e.g., ADD [21], were not publicly available during this research).

Several languages of MLAADV3 are generated using four architectures (*XTTS v1.1*, *XTTS v2*, *Griffin-Lim*, *VITS*) [7]. To avoid any overlap between fine-tuning and evaluation architectures, we employ strict constraints. We perform fine-tuning of the models, initially pre-trained on ASVspoof 2019, with audio samples generated using *VITS* and *Griffin-Lim*, and later evaluate on spoof samples synthesized with *XTTS v1.1* and *XTTS v2*. To ensure fairness, we also reverse this setup, (*XTTS* architectures are used for fine-tuning, whereas *VITS* and *Griffin-Lim* for testing). This guarantees no overlap between the methods in fine-tuning and test datasets. As shown recently [10], such overlap could lead to incorrect conclusions on language transferability because the ease or challenge might stem from generator-specific artifacts rather than linguistic properties. This separation across samples is essential to maintain the integrity of our comparisons. For evaluation assessment, we calculate the Equal Error Rate (EER) (in %) for both runs separately and report an average of them. Our approach focuses on conducting standardized comparisons within the available data resources.

Even for this limited number of languages and architectures, we took proactive steps to generate missing data to avoid potential bias. Specifically, we synthesized samples using *XTTS*

⁵<https://owncloud.fraunhofer.de/index.php/s/tL2Y1FKrWiX4ZtP>

Table 2: Edit-distance (ED) means \pm standard deviations for fake and original transcriptions for MLAAD dataset [7].

| | de | en | es | fr | it | pl | ru | uk |
|------------------------|---------------|---------------|---------------|---------------|----------------|----------------|---------------|-----------------|
| ED _{fake} | 2.4 \pm 2.6 | 4.5 \pm 4.9 | 2.5 \pm 3.1 | 4.8 \pm 5.1 | 9.5 \pm 10.4 | 5.8 \pm 14.0 | 4.9 \pm 4.6 | 23.2 \pm 34.0 |
| ED _{original} | 3.5 \pm 7.4 | 2.8 \pm 3.7 | 2.1 \pm 2.5 | 4.0 \pm 4.3 | 2.0 \pm 2.0 | 2.7 \pm 3.1 | 2.9 \pm 3.6 | 6.1 \pm 11.3 |

v1.1 and XTTS v2 for English and VITS samples for Russian, which were not included in the original MLAAD. Moreover, it is important to note that the original MLAAD dataset does not contain Ukrainian samples generated using XTTS v1.1 and XTTS v2 due to the limitations of these generators. This led us to use GlowTTS and Facebook Massively Multilingual Speech instead of excluding this language from our study.

3.3. Hyperparameters

We train and fine-tune the models utilizing distinct hyperparameter configurations depending on the model architecture. For the lightweight architectures (LFCC+AASIST, LFCC+MesoNet, RawGAT-ST), we utilize a learning rate of $5e-03$ and a weight decay of $2.5e-05$. In contrast, for the larger models, W2V+AASIST and Whisper+AASIST, we utilize a learning rate of $5.0e-06$ and a weight decay of $5e-07$. Due to limited data resources, we use an unchanged hyperparameter configuration, as well as a more traditional, lower learning rate and weight decay typically applied during fine-tuning. Thus, for fine-tuning lightweight models, we use a learning rate of $1e-04$, a weight decay of $5.0e-06$, an learning rate of $5.0e-06$, and a weight decay of $5.0e-07$. Similarly, we fine-tune W2V+AASIST and Whisper+AASIST using either a learning rate of $2.5e-6$ or $5e-6$, both with a weight decay of $2.5e-7$. Furthermore, the SSL front-ends (W2V and Whisper) remain fully trainable during both training and fine-tuning, with all weights unfrozen. We do not apply any data augmentation techniques. The evaluation results are the mean of 10 runs of 90% of the test dataset.

4. Results

In this section, we present the evaluation results of the scenario defined in Sect. 3 to determine the efficacies of the model in limited data availability across various languages and adaptation strategies. Firstly, we evaluate English-trained models on the target languages. Then, we train from scratch in the target language to verify the need for large-scale data training in multilingual settings. We further explore fine-tuning with a single related or unrelated language and with multiple languages while omitting the specific language, thus examining the potential of cross- and intra-linguistic adaptations. Intralinguistic strategy, focusing on within-language fine-tuning, demonstrates that even a limited amount of data can meaningfully improve detection accuracy in the target language. Cross-linguistic adaptations, that is the strategy that uses any other language(s) than the target one, offer a promising path for enhancing low-resource language performance through transfer from potential similarities between languages or similar models' interpretations, but the results point out the challenges and limitations, thus leading to using unusual combinations of languages that are not related to each other.

In Tables **bold** values indicate the best performance for a specific language (across all models), while underlined values highlight the best performance for a specific pre-trained model within each language.

4.1. Baseline

The experiment shows that the performance of models trained on the entire ASVspoof2019 LA, comprising 97,168 training samples and 24,293 validation samples, varies very depending on language and model. As shown in Table 3, W2V+AASIST, pre-trained on an English benchmark dataset, achieves the highest performance across nearly all tested languages, thus substantially outperforming the lowest EERs of other models for each language, presenting superior generalization capabilities across languages. While the best LFCC-based model varies across languages, these models trailed behind W2V+AASIST overall and far outperformed alternatives like Whisper+AASIST and RawGAT-ST, demonstrating insufficient cross-lingual generalization capabilities.

The results show widely differing efficacies of audio DF detection across models and languages. Notably, specific languages revealed more significant challenges for the pre-trained models. In particular, the Russian language exhibits the highest challenge, achieving an EER of $15.74 \pm 8.91\%$ for W2V+AASIST and $17.89 \pm 9.79\%$ for LFCC+MesoNet. Analyzing the best-performing models trained on the English benchmark dataset, it is noteworthy that DFs in some languages are even more detectable than in English, despite being trained only with English samples. Specifically, LFCC+AASIST and LFCC+MesoNet detect DFs more effectively in French, Polish, and Ukrainian, while W2V+AASIST shows improved performance in these languages, as well as Italian. While Ukrainian's performance can be explained by poorer sample quality compared to other languages, as visible in Table 2, the superior performance of the remaining languages relative to English requires additional analysis in subsequent works.

4.2. Language training

To investigate the effectiveness of limited training data, we train models from scratch **with a single language** following the dataset partitioning described in Section 3. The aim is to assess whether even a small number of samples for a particular language and training the model could be an alternative to long training on a large dataset. The SSL-based front-end is re-initialized as in previous experiments.

The results presented in Table 4 indicate that models trained from scratch generally perform poorly, especially compared to the pre-trained models, which overreach in every scenario. However, LFCC+AASIST achieves the best results relatively across all models. Nevertheless, only Russian achieves slightly better performance than pre-trained models with the English benchmark. On the other hand, for the remaining languages, we observe a significant drop in detection efficacy. Therefore, we cannot replace large-scale pretraining with small, targeted data. Since W2V+AASIST and LFCC-based models significantly outperform RawGAT-ST and Whisper+AASIST in our experiments, we report only the results for the former models and place the latter in Appendix A.

These results confirm that having a large, language-independent amount of data enhances detection more than small, language-specific datasets, thus highlighting and rein-

Table 3: The mean EER scores of baseline models trained with the large English dataset evaluated with the data split procedure described in Section 3. **Bold** values indicate the best performance for a specific language.

| Model | Trained with | Languages | | | | | | | |
|----------------|--------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|---------------------|--------------------|
| | | de | en | es | fr | it | pl | ru | uk |
| W2V+AASIST | ASVspoof2019 | 2.81 ± 0.20 | 1.57 ± 0.14 | 2.48 ± 0.31 | 0.38 ± 0.04 | 0.31 ± 0.09 | 0.36 ± 0.03 | 15.74 ± 8.91 | 0.60 ± 0.39 |
| LFCC+AASIST | | 5.84 ± 0.44 | 3.74 ± 1.74 | 24.67 ± 2.67 | 1.34 ± 0.34 | 4.53 ± 0.95 | 1.11 ± 0.16 | 22.80 ± 5.62 | 1.39 ± 1.18 |
| LFCC+MesoNet | | 6.42 ± 0.40 | 10.15 ± 0.76 | 12.24 ± 2.21 | 2.66 ± 1.83 | 4.54 ± 1.15 | 1.76 ± 0.57 | 17.89 ± 9.79 | 8.02 ± 8.36 |
| RawGAT-ST | | 47.04 ± 4.73 | 43.58 ± 2.04 | 41.74 ± 4.69 | 44.72 ± 12.60 | 43.40 ± 13.80 | 41.87 ± 8.55 | 32.78 ± 10.13 | 35.46 ± 6.54 |
| Whisper+AASIST | | 43.57 ± 1.48 | 42.76 ± 1.31 | 42.73 ± 6.20 | 41.32 ± 4.93 | 35.52 ± 13.32 | 42.49 ± 6.69 | 35.69 ± 10.14 | 31.95 ± 6.43 |

Table 4: The mean EER scores of trained from scratch with a single language.

| Model | Trained with | Languages | | | | | | | |
|--------------|--------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|---------------------|--------------------|
| | | de | en | es | fr | it | pl | ru | uk |
| W2V+AASIST | de | 13.35 ± 5.90 | 16.67 ± 8.19 | 21.22 ± 5.01 | 10.71 ± 4.89 | 13.79 ± 4.45 | 6.80 ± 4.02 | 25.60 ± 6.12 | 16.91 ± 9.58 |
| | en | 20.71 ± 12.37 | 16.59 ± 15.43 | 26.44 ± 17.17 | 21.56 ± 19.09 | 24.94 ± 19.09 | 19.88 ± 18.12 | 36.49 ± 0.83 | 10.73 ± 7.80 |
| | es | 25.26 ± 10.14 | 28.54 ± 16.25 | 29.10 ± 18.37 | 20.36 ± 11.42 | 20.55 ± 12.28 | 20.34 ± 11.90 | 29.49 ± 13.46 | 25.85 ± 19.35 |
| | fr | 8.60 ± 3.37 | 13.87 ± 3.54 | 8.84 ± 1.64 | 11.28 ± 1.44 | 5.38 ± 1.81 | 3.32 ± 1.73 | 17.25 ± 6.33 | 1.33 ± 0.35 |
| | it | 28.66 ± 17.03 | 32.75 ± 8.83 | 29.22 ± 11.65 | 25.59 ± 17.83 | 25.70 ± 17.27 | 23.71 ± 17.25 | 34.85 ± 9.89 | 19.18 ± 15.42 |
| | pl | 16.10 ± 7.47 | 32.85 ± 6.78 | 21.63 ± 4.38 | 20.14 ± 0.28 | 13.54 ± 6.81 | 6.79 ± 2.42 | 20.33 ± 1.24 | 21.06 ± 11.00 |
| | ru | 21.07 ± 7.69 | 27.57 ± 0.29 | 18.04 ± 2.70 | 18.23 ± 10.17 | 14.20 ± 6.58 | 13.70 ± 7.85 | 12.27 ± 2.39 | 9.60 ± 4.41 |
| | uk | 32.23 ± 6.17 | 26.99 ± 3.30 | 28.24 ± 0.69 | 24.18 ± 5.15 | 21.84 ± 5.71 | 24.28 ± 1.53 | 31.95 ± 3.26 | 23.26 ± 6.24 |
| LFCC+AASIST | de | 7.97 ± 3.52 | 4.51 ± 2.41 | 26.63 ± 12.50 | 5.92 ± 0.69 | 3.21 ± 1.24 | 10.43 ± 8.83 | 19.82 ± 18.86 | 2.35 ± 0.79 |
| | en | 7.71 ± 1.66 | 4.93 ± 0.11 | 16.83 ± 4.15 | 5.11 ± 0.09 | 1.43 ± 0.31 | 2.67 ± 0.19 | 14.18 ± 13.75 | 7.43 ± 7.78 |
| | es | 9.23 ± 1.16 | 4.98 ± 0.48 | 20.51 ± 20.83 | 4.36 ± 0.15 | 1.32 ± 0.85 | 1.78 ± 0.65 | 25.49 ± 26.04 | 3.31 ± 0.33 |
| | fr | 9.16 ± 1.75 | 6.74 ± 0.33 | 7.00 ± 2.42 | 5.13 ± 0.08 | 3.77 ± 0.62 | 1.11 ± 0.13 | 16.78 ± 17.49 | 2.69 ± 2.84 |
| | it | 8.21 ± 1.04 | 7.27 ± 0.29 | 5.69 ± 4.67 | 6.06 ± 0.08 | 4.30 ± 0.25 | 2.26 ± 0.28 | 17.82 ± 18.78 | 2.86 ± 0.91 |
| | pl | 10.09 ± 0.59 | 7.77 ± 0.99 | 16.05 ± 5.61 | 6.14 ± 0.15 | 2.02 ± 0.09 | 1.31 ± 0.06 | 16.88 ± 16.12 | 5.00 ± 5.21 |
| | ru | 15.48 ± 8.91 | 16.26 ± 16.09 | 34.01 ± 17.64 | 9.16 ± 9.55 | 9.09 ± 8.61 | 7.55 ± 7.69 | 24.76 ± 19.50 | 7.63 ± 1.03 |
| | uk | 10.46 ± 1.41 | 7.12 ± 1.12 | 12.24 ± 1.94 | 4.59 ± 0.52 | 1.89 ± 0.23 | 4.44 ± 0.21 | 18.27 ± 15.66 | 1.49 ± 0.21 |
| LFCC+MesoNet | de | 13.41 ± 3.64 | 8.84 ± 6.34 | 24.12 ± 0.31 | 5.15 ± 4.76 | 9.01 ± 3.93 | 13.10 ± 12.53 | 30.48 ± 12.81 | 6.74 ± 7.03 |
| | en | 7.78 ± 0.48 | 6.80 ± 0.38 | 13.36 ± 5.47 | 7.16 ± 0.02 | 1.99 ± 0.40 | 3.73 ± 0.29 | 17.82 ± 14.55 | 3.21 ± 0.23 |
| | es | 10.95 ± 0.24 | 10.66 ± 2.85 | 13.91 ± 2.27 | 2.50 ± 2.04 | 7.09 ± 3.05 | 2.33 ± 0.56 | 22.63 ± 18.00 | 4.90 ± 5.05 |
| | fr | 7.12 ± 1.56 | 13.06 ± 0.42 | 16.19 ± 9.08 | 6.20 ± 0.13 | 2.31 ± 0.38 | 5.92 ± 0.46 | 21.90 ± 9.65 | 7.67 ± 0.70 |
| | it | 8.91 ± 1.09 | 5.39 ± 2.46 | 16.52 ± 2.80 | 7.50 ± 0.99 | 2.82 ± 1.66 | 2.84 ± 1.86 | 15.65 ± 14.79 | 2.96 ± 3.08 |
| | pl | 6.81 ± 1.05 | 31.22 ± 1.64 | 17.75 ± 7.01 | 10.38 ± 0.22 | 2.66 ± 0.64 | 4.74 ± 0.09 | 18.91 ± 10.69 | 3.85 ± 4.06 |
| | ru | 17.52 ± 12.41 | 23.43 ± 17.51 | 27.47 ± 11.93 | 12.96 ± 12.97 | 17.41 ± 15.59 | 14.03 ± 13.46 | 25.70 ± 10.51 | 4.08 ± 0.48 |
| | uk | 9.59 ± 2.42 | 6.57 ± 2.28 | 23.14 ± 11.66 | 6.61 ± 0.15 | 4.28 ± 2.31 | 1.55 ± 0.47 | 24.64 ± 2.87 | 2.83 ± 2.98 |

forcing the value of extensive, even language-independent, data for DF detection.

4.3. Language fine-tuning

Fine-tuning pre-trained models with English benchmark data allows us to assess the need for fine-tuning to improve models' detectability and potential cross-language generalization capabilities. Fine-tuning and further evaluation follow the data split described in Section 3.

We first assess whether fine-tuning with single-language data enhances audio DF detection by adding linguistic context, improving performance over English pre-trained models. Based on the results in Table 5, we can distinguish two trends. In the first one, intra-linguistic adaptation is more efficient and thus reduces the EER compared to pre-trained and cross-adaptation models. This group includes better-performing models: LFCC+AASIST and W2V+AASIST. On the other hand, the second group shows a trend that fine-tuning with a specific language is most effective. Specifically, RawGAT-ST, fine-tuned with Polish, and Whisper+AASIST, as well as LFCC+MesoNet, fine-tuned with English, achieve the lowest EER across most languages for this specific model.

A deeper analysis reveals that W2V+AASIST consistently outperforms other architectures, achieving the lowest EER for most languages, confirming the effectiveness of intralinguistic adaptation. Cross-lingual adaptation remains crucial for multilingual models, with W2V+AASIST showing competitive results when the pre-trained model performs well. However, their performance remains comparable to or worse than the baseline for languages like Russian and some Germanic languages. Re-

sults for LFCC+AASIST indicate similar trends with the best results on the diagonal, thus indicating that intra-language adaptations are more effective. Notably, cross-linguistic fine-tuning of the LFCC+AASIST is effective in two scenarios: improving a well-performing pre-trained model (e.g., in the case of French or Polish) or fine-tuning with Italian, which indicates overperforming other even intralinguistic adaptations.

The further investigation focuses on two key aspects of linguistic adaptability in DF detection: the impact of removing a single language from fine-tuning and the trade-off between language-specific and multilingual training. We fine-tune pre-trained models on **all multilingual data except one language** at a time, following the data split detailed in Section 3. During these experiments, we assess whether combining multiple languages for training might provide comparable results, especially in the context of cross-language adaptations. As shown in Table 6, fine-tuning with limited language-specific data generally outperforms a larger multilingual dataset, excluding the target language, with Ukrainian as the only exception. Intriguingly, our analysis reveals that if one language is excluded to optimize the system, German emerges as the most suitable candidate for removal from the training set. This is only relevant for scenarios where German language detection is not a requirement, as its exclusion demonstrated a positive effect on the overall system performance for almost all other languages.

Our results suggest maximizing language coverage in training data whenever feasible. For language-specific deployments, focus on using relevant language data, as even limited amounts demonstrate more usefulness than larger, more linguistically diverse datasets that lack the target language.

Table 5: The mean EER scores of fine-tuned *with* a specific language with the data split procedure described in Section 3.

| Model | Fine-tuned with | Languages | | | | | | | |
|--------------|-----------------|--------------------|--------------------|---------------------|--------------------|--------------------|--------------------|---------------------|--------------------|
| | | de | en | es | fr | it | pl | ru | uk |
| W2V+AASIST | de | 1.65 ± 0.71 | 4.16 ± 2.07 | 5.59 ± 1.56 | 0.70 ± 0.28 | 1.49 ± 0.18 | 0.84 ± 0.05 | 17.80 ± 9.24 | 0.35 ± 0.13 |
| | en | 4.07 ± 1.02 | 0.63 ± 0.10 | 2.29 ± 0.74 | 0.26 ± 0.10 | 0.49 ± 0.28 | 0.24 ± 0.04 | 22.97 ± 5.57 | 0.13 ± 0.10 |
| | es | 4.32 ± 0.65 | 2.80 ± 1.92 | 1.10 ± 0.34 | 0.22 ± 0.08 | 0.23 ± 0.04 | 0.29 ± 0.11 | 16.19 ± 10.34 | 0.16 ± 0.05 |
| | fr | 3.01 ± 0.45 | 3.15 ± 2.10 | 4.62 ± 0.65 | 0.11 ± 0.06 | 0.33 ± 0.07 | 0.24 ± 0.04 | 20.90 ± 7.22 | 0.06 ± 0.06 |
| | it | 3.81 ± 0.58 | 3.74 ± 2.63 | 3.49 ± 0.19 | 0.56 ± 0.44 | 0.22 ± 0.14 | 0.28 ± 0.08 | 19.62 ± 9.27 | 0.11 ± 0.03 |
| | pl | 2.48 ± 0.57 | 4.58 ± 3.87 | 3.36 ± 0.50 | 0.42 ± 0.19 | 0.42 ± 0.21 | 0.19 ± 0.08 | 18.58 ± 8.42 | 0.20 ± 0.08 |
| | ru | 3.73 ± 1.28 | 4.76 ± 2.32 | 1.69 ± 0.17 | 0.90 ± 0.24 | 0.91 ± 0.10 | 0.50 ± 0.26 | 3.59 ± 3.03 | 0.76 ± 0.08 |
| | uk | 2.52 ± 0.57 | 2.50 ± 1.72 | 3.01 ± 0.26 | 0.26 ± 0.06 | 0.28 ± 0.06 | 0.36 ± 0.18 | 18.08 ± 9.92 | 0.42 ± 0.35 |
| LFCC+AASIST | de | 5.52 ± 0.81 | 4.04 ± 2.89 | 20.05 ± 1.57 | 1.15 ± 0.79 | 3.46 ± 0.99 | 2.36 ± 1.72 | 24.00 ± 5.94 | 1.45 ± 1.41 |
| | en | 6.89 ± 1.70 | 2.13 ± 1.05 | 20.46 ± 3.49 | 0.34 ± 0.09 | 2.56 ± 0.26 | 0.88 ± 0.15 | 24.47 ± 5.56 | 1.20 ± 1.05 |
| | es | 8.07 ± 2.13 | 3.71 ± 0.10 | 12.51 ± 1.63 | 0.48 ± 0.35 | 3.44 ± 0.17 | 1.07 ± 0.11 | 16.70 ± 10.07 | 0.98 ± 0.73 |
| | fr | 7.39 ± 0.64 | 1.44 ± 0.62 | 19.93 ± 6.28 | 0.22 ± 0.19 | 2.32 ± 0.35 | 0.44 ± 0.11 | 25.80 ± 5.49 | 0.65 ± 0.61 |
| | it | 7.15 ± 0.56 | 1.43 ± 0.24 | 13.43 ± 1.14 | 0.41 ± 0.07 | 1.22 ± 0.12 | 0.52 ± 0.06 | 18.90 ± 10.63 | 0.77 ± 0.78 |
| | pl | 8.28 ± 1.40 | 3.77 ± 2.63 | 20.55 ± 6.27 | 0.49 ± 0.17 | 2.41 ± 0.56 | 0.69 ± 0.16 | 23.84 ± 5.73 | 1.37 ± 1.26 |
| | ru | 7.70 ± 1.85 | 4.84 ± 3.03 | 22.47 ± 9.20 | 2.80 ± 2.19 | 7.22 ± 4.17 | 2.81 ± 1.90 | 14.07 ± 9.57 | 1.17 ± 0.96 |
| | uk | 7.23 ± 2.01 | 5.31 ± 4.17 | 21.06 ± 2.82 | 0.55 ± 0.08 | 3.11 ± 0.80 | 0.74 ± 0.09 | 24.48 ± 4.27 | 0.36 ± 0.23 |
| LFCC+MesoNet | de | 17.55 ± 0.50 | 17.65 ± 3.89 | 18.75 ± 5.41 | 11.33 ± 3.75 | 13.61 ± 0.39 | 9.15 ± 6.13 | 21.94 ± 11.46 | 23.04 ± 23.91 |
| | en | 6.29 ± 0.43 | 6.55 ± 3.51 | 15.80 ± 4.84 | 0.69 ± 0.11 | 2.92 ± 0.80 | 0.56 ± 0.32 | 18.64 ± 8.79 | 6.98 ± 7.32 |
| | es | 9.58 ± 3.98 | 11.15 ± 6.97 | 18.85 ± 2.19 | 3.51 ± 0.38 | 9.65 ± 0.85 | 4.15 ± 2.58 | 22.96 ± 5.60 | 20.13 ± 21.18 |
| | fr | 7.55 ± 0.53 | 12.51 ± 0.92 | 13.37 ± 2.01 | 2.99 ± 2.00 | 6.01 ± 1.69 | 1.57 ± 0.55 | 18.67 ± 9.34 | 8.15 ± 8.51 |
| | it | 7.53 ± 1.46 | 11.42 ± 2.51 | 14.54 ± 3.94 | 2.34 ± 0.96 | 5.63 ± 0.09 | 2.18 ± 1.37 | 18.31 ± 10.16 | 11.25 ± 11.80 |
| | pl | 8.06 ± 3.76 | 11.47 ± 6.94 | 15.78 ± 4.53 | 2.12 ± 0.65 | 5.53 ± 1.90 | 3.37 ± 3.13 | 19.33 ± 10.35 | 17.98 ± 18.92 |
| | ru | 20.63 ± 12.80 | 23.28 ± 10.78 | 21.04 ± 0.49 | 25.35 ± 25.02 | 21.34 ± 17.05 | 19.00 ± 15.10 | 34.20 ± 4.33 | 9.06 ± 0.49 |
| | uk | 7.82 ± 2.92 | 9.91 ± 5.12 | 15.45 ± 5.74 | 1.58 ± 0.37 | 5.50 ± 2.08 | 2.40 ± 2.12 | 18.80 ± 10.98 | 8.42 ± 8.84 |

Table 6: The mean EER scores of fine-tuned *without* a single language.

| Model | Fine-tuned without | Languages | | | | | | | |
|--------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| | | de | en | es | fr | it | pl | ru | uk |
| W2V+AASIST | de | 3.61 ± 0.25 | 0.45 ± 0.09 | 0.67 ± 0.53 | 0.04 ± 0.05 | 0.12 ± 0.03 | 0.06 ± 0.03 | 10.81 ± 11.18 | 0.02 ± 0.03 |
| | en | 1.09 ± 0.50 | 1.94 ± 1.12 | 1.18 ± 0.82 | 0.15 ± 0.12 | 0.41 ± 0.11 | 0.08 ± 0.03 | 7.95 ± 8.05 | 0.08 ± 0.03 |
| | es | 1.36 ± 0.29 | 0.52 ± 0.05 | 2.60 ± 0.68 | 0.07 ± 0.04 | 0.51 ± 0.14 | 0.13 ± 0.03 | 12.50 ± 12.78 | 0.06 ± 0.06 |
| | fr | 1.86 ± 0.21 | 0.76 ± 0.05 | 1.37 ± 0.79 | 0.47 ± 0.45 | 0.52 ± 0.11 | 0.16 ± 0.05 | 10.92 ± 10.95 | 0.08 ± 0.03 |
| | it | 1.48 ± 0.21 | 0.70 ± 0.08 | 1.74 ± 1.44 | 0.11 ± 0.12 | 0.66 ± 0.06 | 0.11 ± 0.05 | 10.60 ± 10.72 | 0.06 ± 0.06 |
| | pl | 1.43 ± 0.46 | 0.55 ± 0.07 | 1.26 ± 1.00 | 0.05 ± 0.02 | 0.26 ± 0.12 | 0.17 ± 0.12 | 11.61 ± 11.92 | 0.06 ± 0.06 |
| | ru | 1.51 ± 0.30 | 0.45 ± 0.24 | 0.87 ± 0.54 | 0.13 ± 0.14 | 0.21 ± 0.14 | 0.12 ± 0.04 | 22.62 ± 8.35 | 0.08 ± 0.03 |
| | uk | 1.56 ± 0.60 | 0.62 ± 0.24 | 1.47 ± 1.20 | 0.18 ± 0.14 | 0.38 ± 0.25 | 0.22 ± 0.07 | 7.50 ± 7.59 | 0.03 ± 0.03 |
| LFCC+AASIST | de | 7.99 ± 1.57 | 2.05 ± 0.85 | 7.95 ± 3.22 | 0.15 ± 0.05 | 1.13 ± 0.19 | 0.46 ± 0.06 | 18.53 ± 18.58 | 0.17 ± 0.19 |
| | en | 4.51 ± 0.30 | 3.84 ± 1.18 | 8.79 ± 3.50 | 0.35 ± 0.10 | 1.60 ± 0.18 | 0.50 ± 0.28 | 15.40 ± 15.05 | 0.27 ± 0.29 |
| | es | 4.58 ± 0.47 | 3.55 ± 0.22 | 16.21 ± 5.88 | 0.25 ± 0.13 | 1.72 ± 0.10 | 0.51 ± 0.26 | 14.53 ± 13.94 | 0.25 ± 0.26 |
| | fr | 4.84 ± 0.23 | 3.47 ± 1.79 | 9.04 ± 5.84 | 0.51 ± 0.10 | 1.67 ± 0.65 | 0.82 ± 0.61 | 15.23 ± 15.23 | 0.51 ± 0.53 |
| | it | 5.37 ± 1.19 | 3.94 ± 1.11 | 9.19 ± 4.11 | 0.37 ± 0.17 | 2.08 ± 0.37 | 0.84 ± 0.56 | 15.48 ± 15.28 | 0.32 ± 0.27 |
| | pl | 5.18 ± 0.93 | 3.19 ± 0.87 | 8.77 ± 3.80 | 0.38 ± 0.08 | 1.43 ± 0.45 | 0.52 ± 0.26 | 15.44 ± 15.12 | 0.31 ± 0.33 |
| | ru | 3.58 ± 0.75 | 2.69 ± 2.23 | 8.01 ± 7.63 | 0.31 ± 0.23 | 1.17 ± 0.85 | 0.42 ± 0.32 | 15.44 ± 15.56 | 0.47 ± 0.49 |
| | uk | 6.45 ± 1.74 | 3.86 ± 1.25 | 10.20 ± 3.40 | 0.48 ± 0.03 | 1.82 ± 0.34 | 0.66 ± 0.22 | 14.93 ± 14.59 | 0.53 ± 0.53 |
| LFCC+MesoNet | de | 5.51 ± 0.35 | 7.46 ± 2.24 | 10.54 ± 1.82 | 1.95 ± 1.33 | 4.05 ± 1.62 | 0.88 ± 0.29 | 22.23 ± 19.57 | 2.46 ± 2.56 |
| | en | 11.62 ± 2.64 | 13.48 ± 6.20 | 16.30 ± 7.07 | 6.73 ± 1.20 | 9.92 ± 0.49 | 4.39 ± 3.26 | 23.93 ± 17.59 | 14.68 ± 15.39 |
| | es | 11.23 ± 1.78 | 12.68 ± 5.89 | 15.63 ± 6.20 | 6.27 ± 2.07 | 9.14 ± 0.86 | 3.97 ± 2.95 | 25.30 ± 18.83 | 9.83 ± 10.28 |
| | fr | 13.00 ± 5.29 | 14.14 ± 7.70 | 16.64 ± 8.66 | 7.87 ± 2.10 | 11.08 ± 3.44 | 7.08 ± 6.46 | 25.32 ± 20.48 | 15.77 ± 16.58 |
| | it | 12.76 ± 2.44 | 13.14 ± 6.76 | 17.21 ± 6.90 | 8.09 ± 1.42 | 10.61 ± 1.25 | 5.45 ± 4.05 | 26.15 ± 19.00 | 12.19 ± 12.77 |
| | pl | 12.75 ± 3.19 | 13.05 ± 6.96 | 17.42 ± 6.58 | 7.85 ± 0.67 | 10.73 ± 1.51 | 5.37 ± 4.12 | 26.05 ± 19.30 | 11.88 ± 12.44 |
| | ru | 7.52 ± 2.64 | 8.50 ± 7.63 | 16.49 ± 2.14 | 1.48 ± 1.33 | 4.79 ± 2.84 | 1.97 ± 1.81 | 20.33 ± 12.74 | 8.48 ± 8.94 |
| | uk | 12.36 ± 5.56 | 13.99 ± 7.55 | 17.51 ± 7.96 | 7.04 ± 2.07 | 10.21 ± 4.01 | 6.68 ± 6.11 | 24.44 ± 20.52 | 14.75 ± 15.46 |

Answer to RQ1: The detection of audio DF shows significant variability across languages. The evaluation of English pre-trained models without additional linguistic adaptation reveals that models like W2V+AASIST show cross-lingual generalization capabilities, yet their performance varied significantly across languages. The efficacy of some languages is even better than that of the only language *known to the model*, English. On the other hand, the Russian language reveals limitations in this approach, thus highlighting the critical necessity for adapting even the best-performing models to other languages.

Answer to RQ2: Evaluation using English pre-trained models without additional adaptation reveals notable differences in performance when applied to other languages. Languages such as French, Polish, Italian, and Ukrainian show better detection performance than English, even though the models are trained exclusively on English datasets. Conversely, Russian or Spanish seems more challenging, with pre-trained models yielding higher EERs. Nevertheless, for these languages, the intralinguistic adaptations significantly improve audio DF detection.

Answer to RQ3: The experiments indicate that intralinguistic adaptation is the most effective targeted strategy for improving audio DF detection within specific languages. Our results show that the optimal strategy combines English-pre-trained data with multilingual fine-tuning that includes the target language. Even limited target language data significantly improves detection accuracy in the best-performing models. However, fine-tuning with non-target languages shows varying results, and using either unrelated languages or multilingual datasets without the target language fails to improve performance. Furthermore, our results indicate that fine-tuning on a small corpus of target language data consistently outperforms augmenting the training set with a broader but non-target multilingual mix. In practice, improving detection requires a small set of samples in the language you are interested in, rather than a larger set containing non-target language samples.

5. Conclusion

This research aims to introduce a benchmark and answer the question of the effectiveness of pre-trained and adapted audio DF detection models in multilingual settings. Up to now, the influence of the utterances' language on audio DF detection has not been examined unbiasedly. The experiments, especially for best-performed W2V+AASIST and LFCC+AASIST, demonstrate a key finding across most of the tested languages: a small amount of language-specific data often yields greater improvements in detection than much more multilingual data, but without the target language. Although our study is limited to eight languages, samples were generated using a selective range of methods, which affected the confidence of the results. Despite these limitations, we believe that it is an important starting point for tackling the problems and challenges posed by multilingual DFs. Future research should explore more languages and language families, as well as generators and models, to better understand cross-lingual fine-tuning and transfer effects. Additionally, a comparison between tonal and non-tonal languages would be interesting, as well as exploring emotionally varied samples in various languages.

6. Acknowledgment

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: WCSS, ACK Cyfronet AGH) for providing computer facilities and support

within computational grant no. PLG/2024/017432.

7. References

- [1] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon, "A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI," 2023. [Online]. Available: <https://arxiv.org/abs/2303.13336>
- [2] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022.
- [3] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does Audio Deepfake Detection Generalize?" in *Interspeech 2022*, 2022, pp. 2783–2787.
- [4] E. Casanova, K. Davis, E. Gölge, G. Gökner, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: a massively multilingual zero-shot text-to-speech model," in *Interspeech 2024*, 2024, pp. 4978–4982.
- [5] A. Sharma, P. Kumar, V. Maddukuri, N. Madamshetti, K. G. Kishore, S. S. S. Kavuru, B. Raman, and P. P. Roy, "Fast griffin lim based waveform generation strategy for text-to-speech synthesis," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30 205–30 233, Nov 2020. [Online]. Available: <https://doi.org/10.1007/s11042-020-09321-7>
- [6] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," 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. 5530–5540.
- [7] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "Mlaad: The multi-language audio anti-spoofing dataset," *International Joint Conference on Neural Networks (IJCNN)*, 2024.
- [8] O. Chetia Phukan, G. Kashyap, A. B. Buduru, and R. Sharma, "Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 2496–2506. [Online]. Available: <https://aclanthology.org/2024.findings-naacl.160/>
- [9] T. Liu, I. Kukanov, Z. Pan, Q. Wang, H. B. Sailor, and K. A. Lee, "Towards quantifying and reducing language mismatch effects in cross-lingual speech anti-spoofing," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 1185–1192.
- [10] T.-P. Doan, H. Dinh-Xuan, T. Ryu, I. Kim, W. Lee, K. Hong, and S. Jung, "Trident of poseidon: A generalized approach for detecting deepfake voices," in *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*, 2024.
- [11] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.
- [12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.
- [13] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Haniłçi, M. Sahidullah, and A. Sizov, "ASvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech 2015*, 2015, pp. 2037–2041.
- [14] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

- [15] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.
- [16] X. Wang, H. Delgado, H. Tak, J. weon Jung, H. jin Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. H. Kinnunen, N. Evans, K. A. Lee, and J. Yamagishi, "ASVspoof 5: crowd-sourced speech data, deepfakes, and adversarial attacks at scale," in *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 2024, pp. 1–8.
- [17] P. Kawa, M. Plata, and P. Syga, "Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection," in *INTERSPEECH*, 2022, pp. 4023–4027. [Online]. Available: <https://doi.org/10.21437/Interspeech.2022-10078>
- [18] J. Frank and L. Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," in *35th Conf. on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.
- [19] H. Khalid, S. Tariq, and S. S. Woo, "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset," 2021.
- [20] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, T. Gedeon, and K. Stefanov, "Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7414–7423.
- [21] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, S. Nie, and H. Li, "ADD 2023: the Second Audio Deepfake Detection Challenge," 2023. [Online]. Available: <https://arxiv.org/abs/2305.13774>
- [22] Z. Zhang, Y. Gu, X. Yi, and X. Zhao, "'FMFCC-A: A Challenging Mandarin Dataset for Synthetic Speech Detection,'" in *Digital Forensics and Watermarking*, X. Zhao, A. Piva, and P. Comesana-Alfaro, Eds. Cham: Springer International Publishing, 2022, pp. 117–131.
- [23] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-Truth: A Partially Fake Audio Detection Dataset," in *Proc. Interspeech 2021*, 2021, pp. 1654–1658.
- [24] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, and R. Fu, "CFAD: A Chinese dataset for fake audio detection," *Speech Communication*, vol. 164, p. 103122, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167639324000931>
- [25] R. Ranjan, B. Dutta, M. Vatsa, and R. Singh, "Faking fluent: Unveiling the achilles' heel of multilingual deepfake detection," in *2024 IEEE International Joint Conference on Biometrics (IJCB)*, 2024, pp. 1–10.
- [26] W. Huang, Y. Gu, Z. Wang, H. Zhu, and Y. Qian, "Speechfake: A large-scale multilingual speech deepfake dataset toward cutting-edge speech generation methods."
- [27] H. Tak, J. weon Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 1–8.
- [28] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved RawNet with Feature Map Scaling for Text-independent Speaker Verification using Raw Waveforms," *Proc. Interspeech*, pp. 3583–3587, 2020.
- [29] H. Tak, M. Todisco, X. Wang, J. Jung, J. Yamagishi, and N. W. D. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Odyssey 2022: The Speaker and Language Recognition Workshop, 28 June - 1 July 2022, Beijing, China*, T. F. Zheng, Ed. ISCA, 2022, pp. 112–119. [Online]. Available: <https://doi.org/10.21437/Odyssey.2022-16>
- [30] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security, WIFS 2018, Hong Kong, China, December 11-13, 2018*. IEEE, 2018, pp. 1–7. [Online]. Available: <https://doi.org/10.1109/WIFS.2018.8630761>
- [31] P. Kawa, M. Plata, M. Czuba, P. Szymanski, and P. Syga, "Improved DeepFake Detection Using Whisper Features," in *INTERSPEECH*, 2023, pp. 4009–4013. [Online]. Available: <https://doi.org/10.21437/Interspeech.2023-1537>
- [32] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

A. Additional results

The following tables (Table 7 - Table 9) present the experimental results for RawGAT-ST and Whisper-AASIST models. These models consistently underperformed relative to W2V+AASIST and LFCC-based approaches.

Table 7: The mean EER scores of trained from scratch with a single language.

| Model | Trained with | Languages | | | | | | | |
|----------------|--------------|-------------------|-------------------|------------------|-------------------|-------------------|-------------------|------------------|-------------------|
| | | de | en | es | fr | it | pl | ru | uk |
| RawGAT-ST | de | 50.96 \pm 4.94 | 53.41 \pm 11.67 | 60.26 \pm 0.63 | 50.33 \pm 2.76 | 50.37 \pm 8.11 | 46.44 \pm 6.24 | 54.03 \pm 8.84 | 52.84 \pm 6.01 |
| | en | 44.99 \pm 1.58 | 37.74 \pm 9.38 | 49.87 \pm 4.15 | 54.00 \pm 2.05 | 44.29 \pm 2.06 | 50.58 \pm 6.72 | 51.70 \pm 9.92 | 53.37 \pm 10.28 |
| | es | 51.62 \pm 10.80 | 74.85 \pm 2.36 | 59.63 \pm 0.46 | 55.15 \pm 10.88 | 56.11 \pm 14.28 | 61.83 \pm 21.78 | 57.70 \pm 1.57 | 52.04 \pm 17.29 |
| | fr | 49.69 \pm 8.56 | 52.09 \pm 4.88 | 58.38 \pm 0.59 | 48.07 \pm 4.47 | 50.17 \pm 6.10 | 54.79 \pm 11.27 | 58.34 \pm 4.26 | 47.60 \pm 1.73 |
| | it | 48.59 \pm 11.34 | 59.06 \pm 8.02 | 56.33 \pm 0.36 | 55.26 \pm 8.43 | 60.68 \pm 8.01 | 57.41 \pm 13.93 | 49.38 \pm 7.12 | 40.91 \pm 5.44 |
| | pl | 54.51 \pm 2.00 | 45.62 \pm 2.60 | 52.42 \pm 8.18 | 48.13 \pm 3.55 | 54.56 \pm 3.35 | 34.80 \pm 10.42 | 53.27 \pm 7.74 | 67.03 \pm 2.62 |
| | ru | 36.54 \pm 9.92 | 57.12 \pm 0.51 | 53.45 \pm 3.46 | 45.12 \pm 0.34 | 39.79 \pm 15.07 | 41.99 \pm 2.34 | 36.72 \pm 0.72 | 22.63 \pm 9.60 |
| | uk | 58.55 \pm 5.02 | 68.47 \pm 3.28 | 55.68 \pm 3.38 | 51.25 \pm 12.93 | 52.60 \pm 12.05 | 64.26 \pm 15.54 | 45.23 \pm 2.73 | 39.66 \pm 2.64 |
| Whisper+AASIST | de | 53.55 \pm 0.23 | 45.75 \pm 5.32 | 56.27 \pm 1.93 | 46.28 \pm 6.83 | 51.21 \pm 1.82 | 54.00 \pm 4.17 | 51.35 \pm 1.90 | 45.34 \pm 12.18 |
| | en | 46.20 \pm 5.11 | 34.78 \pm 1.54 | 48.82 \pm 5.32 | 44.20 \pm 1.76 | 45.52 \pm 6.28 | 50.87 \pm 12.14 | 48.31 \pm 6.86 | 49.81 \pm 9.56 |
| | es | 55.39 \pm 1.11 | 49.36 \pm 5.64 | 54.89 \pm 0.63 | 48.19 \pm 6.19 | 57.61 \pm 0.28 | 58.28 \pm 2.00 | 54.17 \pm 2.36 | 49.46 \pm 7.90 |
| | fr | 48.13 \pm 4.79 | 43.10 \pm 1.18 | 50.11 \pm 3.77 | 40.00 \pm 0.21 | 48.06 \pm 5.84 | 52.90 \pm 8.07 | 44.38 \pm 6.18 | 41.78 \pm 14.32 |
| | it | 50.16 \pm 0.31 | 45.04 \pm 3.49 | 52.52 \pm 1.88 | 44.22 \pm 5.14 | 51.32 \pm 0.62 | 54.98 \pm 5.73 | 48.57 \pm 2.87 | 42.45 \pm 14.66 |
| | pl | 54.58 \pm 0.96 | 42.62 \pm 5.01 | 57.16 \pm 2.46 | 47.76 \pm 6.37 | 52.42 \pm 3.28 | 53.71 \pm 2.42 | 52.61 \pm 0.33 | 48.32 \pm 15.14 |
| | ru | 51.77 \pm 1.68 | 46.05 \pm 3.02 | 52.03 \pm 1.86 | 42.96 \pm 6.62 | 51.22 \pm 2.65 | 53.27 \pm 3.72 | 43.02 \pm 0.90 | 39.01 \pm 11.99 |
| | uk | 48.64 \pm 4.94 | 49.39 \pm 3.97 | 49.34 \pm 6.08 | 46.68 \pm 6.95 | 43.53 \pm 13.28 | 44.05 \pm 10.94 | 42.37 \pm 8.91 | 34.30 \pm 2.03 |

Table 8: The mean EER scores of fine-tuned **with** a specific language with the data split procedure described in Section 3.

| Model | Fine-tuned with | Languages | | | | | | | |
|----------------|-----------------|-------------------|-------------------|------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| | | de | en | es | fr | it | pl | ru | uk |
| RawGAT-ST | de | 48.61 \pm 6.53 | 52.45 \pm 10.74 | 49.51 \pm 4.06 | 45.32 \pm 11.54 | 47.33 \pm 15.91 | 47.81 \pm 2.12 | 49.03 \pm 9.52 | 51.04 \pm 5.89 |
| | en | 46.65 \pm 6.03 | 35.35 \pm 3.29 | 45.11 \pm 5.11 | 42.01 \pm 12.08 | 41.27 \pm 14.44 | 45.68 \pm 7.57 | 48.06 \pm 4.37 | 43.90 \pm 4.70 |
| | es | 54.11 \pm 14.84 | 56.92 \pm 13.41 | 43.79 \pm 9.29 | 51.05 \pm 20.87 | 49.16 \pm 18.67 | 51.14 \pm 19.09 | 47.85 \pm 9.40 | 41.40 \pm 12.45 |
| | fr | 45.84 \pm 6.34 | 47.28 \pm 0.48 | 49.56 \pm 0.84 | 46.25 \pm 9.77 | 47.66 \pm 12.59 | 45.46 \pm 8.90 | 46.01 \pm 9.93 | 39.09 \pm 0.70 |
| | it | 41.30 \pm 7.22 | 44.56 \pm 4.15 | 38.72 \pm 2.12 | 42.35 \pm 10.60 | 39.67 \pm 10.17 | 44.16 \pm 3.94 | 40.99 \pm 14.75 | 39.75 \pm 15.76 |
| | pl | 41.16 \pm 0.60 | 26.49 \pm 16.40 | 47.83 \pm 7.36 | 43.82 \pm 0.59 | 44.09 \pm 4.48 | 40.25 \pm 3.44 | 47.77 \pm 13.70 | 58.93 \pm 10.43 |
| | ru | 49.86 \pm 5.95 | 46.31 \pm 5.93 | 36.53 \pm 3.74 | 45.02 \pm 15.46 | 35.56 \pm 18.29 | 51.82 \pm 0.94 | 22.38 \pm 3.22 | 23.19 \pm 4.15 |
| | uk | 43.14 \pm 3.21 | 53.74 \pm 7.27 | 40.79 \pm 3.24 | 42.36 \pm 5.86 | 41.47 \pm 4.57 | 43.32 \pm 1.94 | 33.36 \pm 1.25 | 37.92 \pm 5.01 |
| Whisper+AASIST | de | 46.85 \pm 1.42 | 39.95 \pm 1.71 | 45.74 \pm 3.79 | 41.81 \pm 6.28 | 36.80 \pm 12.11 | 46.11 \pm 4.15 | 38.16 \pm 8.95 | 35.83 \pm 9.76 |
| | en | 42.51 \pm 0.77 | 32.28 \pm 1.97 | 43.41 \pm 1.65 | 40.84 \pm 4.81 | 34.24 \pm 10.64 | 43.13 \pm 1.39 | 35.71 \pm 5.39 | 34.82 \pm 10.98 |
| | es | 46.60 \pm 2.12 | 42.59 \pm 0.79 | 46.66 \pm 3.87 | 43.62 \pm 6.28 | 37.67 \pm 12.97 | 47.49 \pm 5.29 | 38.71 \pm 8.82 | 36.08 \pm 9.80 |
| | fr | 44.09 \pm 0.43 | 40.89 \pm 1.92 | 43.66 \pm 3.10 | 39.52 \pm 5.38 | 35.61 \pm 11.58 | 43.87 \pm 1.91 | 35.05 \pm 5.94 | 33.56 \pm 10.80 |
| | it | 44.20 \pm 0.66 | 39.63 \pm 0.46 | 45.13 \pm 2.01 | 41.79 \pm 5.31 | 36.02 \pm 10.92 | 45.89 \pm 1.37 | 36.90 \pm 5.85 | 35.46 \pm 11.50 |
| | pl | 46.89 \pm 1.79 | 41.69 \pm 4.06 | 47.89 \pm 2.84 | 42.25 \pm 8.08 | 38.26 \pm 14.36 | 46.54 \pm 4.59 | 39.32 \pm 10.17 | 36.55 \pm 11.59 |
| | ru | 44.41 \pm 0.30 | 40.17 \pm 1.12 | 43.90 \pm 3.24 | 39.23 \pm 7.79 | 36.17 \pm 12.51 | 45.09 \pm 3.45 | 33.89 \pm 7.71 | 32.58 \pm 9.97 |
| | uk | 44.90 \pm 3.58 | 43.16 \pm 0.56 | 45.20 \pm 6.05 | 40.72 \pm 8.34 | 37.61 \pm 13.55 | 44.86 \pm 8.34 | 36.81 \pm 11.12 | 30.26 \pm 6.00 |

Table 9: The mean EER scores of fine-tuned **without** a single language.

| Model | Fine-tuned without | Languages | | | | | | | |
|----------------|--------------------|------------------|-------------------|------------------|------------------|-------------------|------------------|------------------|-------------------|
| | | de | en | es | fr | it | pl | ru | uk |
| RawGAT-ST | de | 38.58 \pm 7.93 | 46.93 \pm 9.90 | 36.21 \pm 1.20 | 36.94 \pm 8.58 | 33.60 \pm 7.11 | 29.66 \pm 1.59 | 31.04 \pm 1.66 | 19.90 \pm 9.52 |
| | en | 37.62 \pm 6.24 | 51.36 \pm 5.58 | 37.09 \pm 0.40 | 38.51 \pm 7.52 | 34.50 \pm 5.70 | 31.06 \pm 1.14 | 35.71 \pm 5.82 | 25.21 \pm 15.50 |
| | es | 29.37 \pm 0.17 | 54.01 \pm 0.18 | 37.19 \pm 0.29 | 28.00 \pm 0.31 | 25.14 \pm 0.19 | 27.66 \pm 0.21 | 25.59 \pm 0.20 | 22.82 \pm 0.23 |
| | fr | 36.29 \pm 7.19 | 42.81 \pm 10.04 | 35.23 \pm 0.35 | 37.28 \pm 6.84 | 32.29 \pm 6.78 | 28.74 \pm 1.52 | 36.75 \pm 0.60 | 22.30 \pm 9.00 |
| | it | 40.85 \pm 8.32 | 47.11 \pm 5.69 | 38.42 \pm 1.15 | 42.04 \pm 9.62 | 36.21 \pm 6.54 | 33.93 \pm 1.61 | 36.40 \pm 1.46 | 20.11 \pm 11.56 |
| | pl | 40.17 \pm 8.68 | 50.31 \pm 7.09 | 37.56 \pm 4.42 | 38.74 \pm 9.91 | 35.47 \pm 7.82 | 40.27 \pm 7.74 | 35.92 \pm 4.40 | 20.26 \pm 9.85 |
| | ru | 40.89 \pm 5.58 | 49.67 \pm 8.83 | 36.51 \pm 1.68 | 39.68 \pm 6.49 | 36.48 \pm 5.10 | 37.88 \pm 1.55 | 41.49 \pm 6.33 | 25.23 \pm 8.83 |
| | uk | 35.82 \pm 7.66 | 45.03 \pm 10.93 | 36.10 \pm 0.44 | 40.27 \pm 6.52 | 31.24 \pm 5.99 | 26.36 \pm 0.77 | 35.46 \pm 0.59 | 27.11 \pm 14.97 |
| Whisper+AASIST | de | 42.97 \pm 0.59 | 32.95 \pm 1.83 | 43.70 \pm 4.20 | 38.39 \pm 5.84 | 35.97 \pm 9.60 | 45.63 \pm 0.90 | 35.29 \pm 6.54 | 29.44 \pm 11.26 |
| | en | 45.05 \pm 1.03 | 36.83 \pm 2.30 | 45.08 \pm 5.32 | 39.02 \pm 7.38 | 37.00 \pm 11.23 | 46.35 \pm 3.33 | 36.21 \pm 8.62 | 30.06 \pm 10.76 |
| | es | 43.13 \pm 0.19 | 31.66 \pm 0.97 | 44.47 \pm 3.06 | 38.13 \pm 6.48 | 35.57 \pm 9.97 | 44.18 \pm 0.95 | 35.23 \pm 7.85 | 29.30 \pm 10.50 |
| | fr | 44.25 \pm 0.45 | 32.29 \pm 0.99 | 44.85 \pm 4.52 | 40.31 \pm 7.14 | 36.30 \pm 10.62 | 45.12 \pm 1.91 | 36.21 \pm 8.19 | 29.80 \pm 10.30 |
| | it | 43.55 \pm 0.27 | 33.11 \pm 0.71 | 43.96 \pm 4.92 | 38.33 \pm 6.95 | 37.09 \pm 11.02 | 44.74 \pm 1.94 | 35.36 \pm 7.70 | 29.65 \pm 10.61 |
| | pl | 43.14 \pm 0.34 | 32.21 \pm 2.39 | 43.69 \pm 5.03 | 37.18 \pm 6.02 | 35.85 \pm 10.03 | 45.30 \pm 1.27 | 34.30 \pm 7.11 | 29.56 \pm 11.15 |
| | ru | 43.23 \pm 0.63 | 32.04 \pm 1.22 | 44.62 \pm 4.47 | 38.74 \pm 6.22 | 36.49 \pm 10.27 | 45.06 \pm 1.30 | 36.89 \pm 7.93 | 29.84 \pm 11.01 |
| | uk | 43.55 \pm 0.28 | 32.51 \pm 1.64 | 44.80 \pm 3.60 | 38.44 \pm 5.35 | 36.34 \pm 9.40 | 45.89 \pm 0.61 | 36.06 \pm 6.16 | 31.77 \pm 12.43 |