Revealing Cross-Lingual Bias in Synthetic Speech Detection under Controlled Conditions

Victor Moreno[1], *João Lima*[1], *Flávio Simões*[2], *Ricardo Violato*[2], *Mário Uliani Neto*[2], *Fernando Runstein*[2], *Paula Costa*[1]

[1]Universidade Estadual de Campinas (UNICAMP), Brazil
[2]CPQD, Brazil

paulad@unicamp.br

Abstract

Speech-based biometric systems have been increasingly deployed in high-stakes domains such as banking, forensics, and authentication. However, these systems remain vulnerable to synthetic speech attacks, such as spoofing and deepfakes. Recent research has focused on developing countermeasures (CMs) capable of detecting manipulated audio. In this work, we investigate whether language identity influences the detectability of synthetic speech in a state-of-the-art CM pipeline. We train a detector on English-only data and evaluate it under controlled conditions using spoofed samples in ten languages synthesized by a standardized text-to-speech system. Despite uniform synthesis settings, we observe significant language-dependent disparities in detection performance. These results suggest that language identity acts as a latent bias factor, challenging the cross-lingual generalization of current CM systems and underscoring the need for fairness-aware multilingual evaluation protocols.

Index Terms: automatic speaker verification, deepfake detection, audio anti-spoofing, speech-based biometrics.

1. Introduction

There is an expanding set of use cases for speech-based biometric systems, including critical application domains such as banking, forensic analysis, and identity authentication. Despite ongoing advances in Automatic Speaker Verification (ASV) systems, they remain vulnerable to attacks based on synthetic speech, including spoofing and deepfake audio. These attacks exploit state-of-the-art (SOTA) text-to-speech (TTS) and voice conversion (VC) technologies to convincingly mimic target speakers and bypass authentication mechanisms. In response, a range of countermeasures have been developed to distinguish bona fide speech from synthetic or manipulated audio. Despite significant advances, a key challenge remains: countermeasure (CM) systems often struggle to generalize across diverse data conditions, particularly when faced with mismatched languages, speakers, or synthesis techniques. Previous work has shown that detection performance can be sensitive to demographic groups, with notable disparities related to gender, accent, and speech impairments [1]. Yet, the role of the specific language spoken itself remains underexplored, not just as a carrier of content, but also as a potential source of systematic detection bias, even under controlled and balanced synthesis conditions.

In this work, we focus on a key generalization question: Can a detector trained exclusively on English speech effectively identify spoofed utterances in other languages? To address this, we conduct a series of carefully controlled cross-lingual experiments using a unified synthesis pipeline and a standardized evaluation protocol. Our analysis spans a diverse set of ten languages, all generated using consistent text-to-speech settings, aiming to isolate the effect of language from other confounding factors. To ensure that observed trends reflect real limitations in today's leading systems, our detection backbone adopts the AASIST+wav2vec2 pipeline, a leading model in recent spoofing detection benchmarks[2].

To develop the baseline model used in our cross-lingual bias analysis, we rely on the ASVspoof 5 corpus, the latest edition of a long-running community challenge designed to advance the development and evaluation of anti-spoofing technologies [3]. This edition introduces a large-scale dataset of English-language utterances collected from crowd-sourced recordings under diverse acoustic conditions, accompanied by spoofed samples generated through modern TTS and VC systems, including adversarial attacks. We use Track 1, which focuses on standalone spoof detection in realistic open-domain scenarios. Its scale, variability, and challenging design provide a solid foundation for assessing cross-lingual generalization and language-related biases in modern detection systems.

For evaluation, we leverage MLAAD [4], a multi-language dataset for the task of audio anti-spoofing. In particular, we focus on a specific subset of MLAAD consisting of spoofed speech generated using Meta's Massively Multilingual Speech (MMS) TTS system [5]. This choice is a key aspect of our study, as it helps mitigate the influence of synthesis quality on cross-lingual comparisons. The MMS subset was created under a controlled setting: spoofed speech for each language is synthesized using the same model architecture, hyperparameters, and training procedure, varying only the target language. This design allows us to isolate the effect of language while minimizing confounding factors related to the generation process.

The obtained results reveal a consistent and systematic language-based performance gap: synthetic speech in Romanian, Russian, French, and Finnish is more likely to be detected as a spoof than English, German, Swahili, and Ukrainian. These disparities emerge despite uniform generation conditions, challenging assumptions of model fairness and exposing language as a latent source of bias in modern spoofing detection pipelines.

This work brings the following contributions:

1. We present what is, to the best of our knowledge, the first cross-linguistic study of spoof detection using controlled TTS attacks in ten languages;
2. We demonstrate that language identity alone can significantly affect detection performance, even when the architecture, training data, and synthesis protocols are held constant;
3. We analyze the implications of these findings for the fairness and generalization of state-of-the-art spoofing detectors;
4. We release a reproducible benchmark and analysis pipeline to

support future multilingual and bias-aware research in audio deepfake detection, along with publicly available code [1].

Our findings underscore critical limitations in the generalization capability of current CM systems and expose language-induced biases that are not accounted for by conventional evaluation protocols. As synthetic speech generation continues to scale globally, we advocate for the development of language-agnostic or explicitly language-aware countermeasures that can ensure robust and equitable protection across linguistic boundaries.

2. Related Works

The ISO/IEC 30107 standard provides a foundational framework for understanding presentation attacks in biometric systems, including those based on voice [6]. These attacks exploit vulnerabilities at various stages of the biometric pipeline, most critically at the point of data capture, by using synthetic or manipulated input to impersonate legitimate users. In the context of speaker verification, such attacks include both audio spoofing and deepfake speech, typically generated via TTS or VC technologies. These threats remain a significant challenge for voice-based authentication systems. Despite advances in presentation attack detection (PAD) mechanisms, issues such as generalization failures and high false acceptance rates persist, underscoring the need for robust, fair, and adaptable countermeasure strategies.

2.1. State-of-the-Art Audio Anti-Spoofing

Recent advances in spoofing and deepfake detection have been driven by the adoption of end-to-end neural architectures capable of learning directly from raw audio or self-supervised embeddings [7]. Among these, Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks (AASIST) [8] has become one of the most prominent approaches, leveraging spectro-temporal graph attention networks to model discriminative speech artifacts. When integrated with self-supervised front-ends such as wav2vec2 [9], AASIST has demonstrated strong generalization in the detection of synthetic speech under various conditions [2]. This combination has been further explored and refined by the USTC-KXDIGIT [10], whose system achieved top performance in the ASVspoof 5 Challenge [11]. Their approach fused AASIST with wav2vec2 embeddings and introduced architectural enhancements and ensemble techniques to improve robustness under open and closed evaluation conditions. These results reaffirm the effectiveness of combining graph-based models with self-supervised representations in high-stakes detection tasks. Motivated by these findings, we adopt the AASIST+wav2vec2 configuration as the foundation for our experiments, leveraging its strong generalization performance as a baseline to assess whether such robustness holds under cross-lingual evaluation scenarios.

2.2. Bias and Fairness in Spoofing Detection

Fairness has become a central concern in deepfake detection, particularly in the image and video domains, where numerous studies have documented performance disparities across demographic groups such as race, gender, and age. These findings have led to the development of fairness-aware benchmarks and training strategies. In the speech domain, however, efforts to understand and address such biases in synthetic speech detection remain relatively scarce. Despite substantial progress in countermeasure development, most systems are still optimized for average case performance, often neglecting how detection accuracy varies between different user groups.

FairSSD is the first work to systematically address this gap by providing a large-scale analysis of demographic bias in synthetic speech detection systems [1]. Their experiments reveal that current models often misclassify bona fide speech from male speakers, non-native accents, and individuals with speech impairments at significantly higher rates than other groups. These disparities persist across different model architectures and training settings, indicating that bias is not merely incidental, but structurally embedded in existing detection pipelines. Our work extends this line of investigation by shifting the focus from speaker-specific demographic traits to the role of language identity in spoof detection. While FairSSD highlights the risks of unequal treatment across speaker groups, we investigate whether linguistic variation alone, even under matched synthesis protocols and balanced data, can lead to systematic performance gaps. This direction is relevant as detection systems are increasingly deployed in multilingual environments, where fairness between languages becomes essential for both security and inclusiveness.

2.3. Multilingual Dataset

Progress in synthetic speech detection has been largely driven by datasets focused on English or Chinese, such as ASVspoof, ADD, and FakeOrReal [12, 13, 14]. Although these resources have enabled key advances, their monolingual scope limits generalization and restricts the assessment of language-related biases, an increasingly relevant issue as detection systems are deployed globally. The MLAAD dataset [4] addresses this limitation by providing over 420 hours of synthetic speech in 38 languages, created with 91 TTS models, comprising 42 different architectures, which are trained on 14 different source datasets. Built on the M-AILABS corpus [15], which provides high-quality multilingual read speech recorded by native speakers, MLAAD offers a standardized and linguistically diverse benchmark for evaluating spoof detection systems, addressing a critical yet often overlooked axis of variation. Due to its diversity in both language coverage and synthesis architectures, MLAAD constitutes a robust benchmark and a valuable complementary resource for training spoofing detection models. More importantly, it enables fairness-focused research beyond demographic traits. In our work, we use MLAAD to analyze language-based bias, focusing on a subset generated with MMS TTS models [5].

2.4. Controlled Multilingual Synthesis with MMS

Meta AI's MMS project represents a major step toward inclusive speech technology by supporting more than 1,100 languages for TTS, ASR, and language identification tasks [5]. A key contribution of MMS is its uniform and carefully balanced multilingual training setup, which enables rigorous cross-lingual experimentation under controlled conditions. For TTS, the MMS-lab dataset was constructed from aligned recordings of New Testament readings in various languages, with standardized preprocessing, quality control, and segmentation. Each language model was trained using the same VITS-based architecture, hyperparameters, training schedule, and data curation strategy. This uniformity limits variability unrelated to language identity and enables fair cross-lingual comparison, min-

[1] https://github.com/victorgmoreno/crosslingual_bias_audiodeepfake

imizing confusing variables often present in multilingual corpora. We used the MMS-generated subset of MLAAD to investigate whether language identity alone can impact the performance of spoof detection under tightly controlled synthesis conditions. This allows us to build on a proven architecture while shifting the focus from attack types to language identity, enabling a targeted investigation of cross-lingual detection performance under controlled synthesis conditions.

3. Methods

3.1. Detection Pipeline

We extend a state-of-the-art spoofing detection architecture to support multilingual inference, allowing systematic evaluation of cross-lingual bias under controlled conditions. This architectural adaptation constitutes a core contribution of our work. The system adopts a modular front-end and back-end design, a widely established paradigm in modern countermeasure research [11]. In this framework, the front-end functions as a feature extractor, transforming raw audio waveforms into contextualized embeddings, while the back-end operates as a task-specific classifier that assigns spoofing likelihood scores [2, 16]. We implement this pipeline using wav2vec2 XLS-R 300M as the front-end [9] and AASIST as the back-end [8], resulting in a trainable end-to-end system optimized for robustness and generalization. This configuration builds on the open source baseline proposed by Tak et al. [2] and was subsequently refined in the USTC-KXDIGIT submission, which achieved top performance in the ASVspoof 5 Challenge [10, 3].

The **front-end**, wav2vec2 XLS-R 300M, pretrained in 436K hours of multilingual speech, encodes raw audio waveforms in contextualized frame-level embeddings [9]. These embeddings are then processed by a 2D self-attentive pooling layer, which aggregates the representations across time and frequency dimensions by learning attention weights that emphasize salient speech regions. The resulting pooled feature map serves as input to the **back-end** AASIST module, which models spectro-temporal dependencies using heterogeneous graph attention layers. AASIST constructs separate graphs along the temporal and spectral axes and extracts localized spoof discriminative cues. The final graph representation is passed through a fully connected layer to produce CM scores, which are scalar values that indicate the likelihood of an audio sample being spoofed rather than bona fide. For a given input utterance, the model outputs a score $s \in [0, 1]$, where higher values indicate a higher probability that the audio sample is synthetic.

3.2. Datasets

This study relies on two primary datasets: the ASVspoof 5 Track 01 corpus for CM training and the Multi-Language Audio Anti-Spoofing Dataset (MLAAD) for cross-lingual evaluation.

ASVspoof 5 Track 01: We use this partition to train our spoofing detection model. It contains over 145,000 English-language utterances, evenly split between bona fide and spoofed samples generated using a variety of neural TTS and VC systems, including several unknown to the model during training [11]. Track 01 simulates logical access (LA) attacks in open-domain scenarios and provides balanced speaker gender and diverse recording conditions. To ensure the quality and competitiveness of our model, we validate and evaluate it using the official ASVspoof development and evaluation subsets before applying it to cross-lingual testing on MLAAD.

MLAAD: We use MLAAD [4], a recently introduced multilingual benchmark comprising over 420 hours of synthetic speech in 38 languages, to evaluate whether spoof detection performance is influenced by language identity. Bona fide utterances are sourced from the M-AILABS dataset, while spoofed samples are generated using 91 TTS systems spanning 42 distinct architectures.

MMS Subset: For controlled multilingual evaluation, we select a subset of MLAAD consisting of spoofed utterances synthesized by ten monolingual TTS models from Meta's MMS project [5]. Each model was trained independently in a single language using the same VITS-based architecture, hyperparameters, preprocessing pipeline, training schedule, and corresponding translations of a common textual source (the New Testament). Each TTS system was trained for 100k steps on a single-speaker recording consisting of a reading of the translated base text, ensuring comparable generation quality. The resulting evaluation set comprises 1,000 spoofed utterances per language, totaling 10,000 samples in Finnish, German, Russian, Swahili, Ukrainian, English, French, Dutch, Hungarian, and Romanian. These languages span diverse linguistic families, enabling a systematic investigation of cross-linguistic variation in spoof detectability. Since all generation parameters are kept constant, variation in detection performance can be predominately attributed to language identity, which is the only variable intentionally varied under controlled conditions.

3.3. Monolingual Training Setup

We train the model end-to-end on the ASVspoof 5 Track 01 training partition for 100 epochs. These samples include diverse synthesis artifacts from a wide range of TTS and VC models, including unknown attack types, reflecting contemporary deepfake speech pipelines [3]. The balance of genders is preserved within each class to avoid demographic skew during training.

The complete model, including the wav2vec2 encoder, 2D self-attentive pooling, and the AASIST back end, is jointly optimized using binary cross-entropy loss and the Adam optimizer (learning rate $1e^{-4}$, batch size 32).

At this stage, no multilingual data or augmentation is used. Following prior work [2], we allow fine-tuning of the wav2vec2 transformer layers during training, enabling domain adaptation to the spoof detection task and enhancing robustness against both seen and unseen attacks. The model is exposed only to English during training, allowing us to later probe whether its treatment of spoofed speech generalizes across languages or exhibits systematic bias. This training regime reflects real-world operational settings where multilingual-labeled spoofed data are scarce or unavailable.

3.4. Evaluation Protocols

We assess the impact of language identity on spoof detection by evaluating our model on a controlled subset of the MLAAD dataset [4], consisting of spoofed utterances from ten MMS TTS models [5]. Each model was trained under identical conditions, with fixed speaker, architecture, training schedule, and input text, ensuring that any performance differences are due to language alone. The selected languages span diverse language families, providing a representative yet controlled multilingual testbed to analyze language-specific variation in detection performance.

Bona fide references are excluded for two key reasons. First, we chose to prioritize broader linguistic diversity over constrained pairwise comparisons. Since not all languages present in the MMS subset have corresponding bona fide ref-

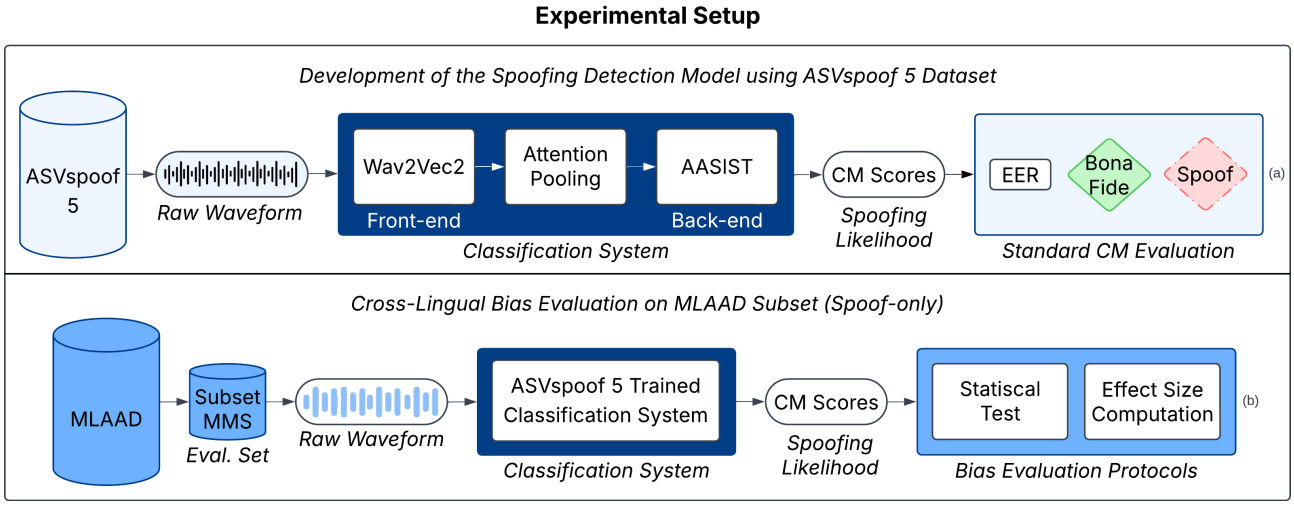

Figure 1: *Overview of the experimental setup and evaluation framework. (a) The AASIST+Wav2Vec2 model is trained and evaluated on the ASVspoof 5 Track 01 Corpus, which includes both bona fide and spoofed English speech. The model operates as a front-end/back-end pipeline that converts raw audio waveforms into countermeasure (CM) scores. Its performance is measured using the Equal Error Rate (EER). (b) The trained model is then applied to a controlled, spoof-only subset of the MLAAD dataset, synthesized using Meta's Massively Multilingual Speech (MMS) TTS system. This subset includes ten languages generated under identical synthesis conditions, allowing language identity to be isolated as the variable of interest. The resulting CM scores are analyzed using statistical hypothesis testing (Mann–Whitney U test) and effect size estimation (Common Language Effect Size, CLES) to assess cross-lingual detection bias.*

erences in M-AILABS, restricting the study to those that do would significantly reduce language coverage and, in turn, weaken the generalization of our findings. Second, instead of computing traditional binary classification metrics such as EER or t-DCF, we choose to examine whether the model's output scores for spoofed speech vary systematically with language. This spoof-only evaluation aligns with real-world deployment priorities, where undetected synthetic speech poses the most critical vulnerability. To assess potential bias, we analyze the CM scores produced by the detector.

To guide our investigation, we formulate the null hypothesis that there is no significant difference between score distributions across languages. The alternative hypothesis is that at least one language's score distribution differs from the rest. We test these hypotheses by performing pairwise Mann-Whitney U tests [17]. All tests were conducted as two-sided analyses to detect any significant difference (positive or negative) in the score distributions, and the raw p-values were adjusted using the conservative Bonferroni method to mitigate the risk of Type I errors. For each test, we calculated the corresponding Common Language Effect Size (CLES) [18]. Let X and Y be two populations of detector CM scores assigned to a pair of languages with sample sizes m and n. CLES is calculated according to Equation 1. CLES(X, Y) may be interpreted as the probability that a randomly selected CM score in language X will be greater than a randomly selected score in language Y.

$$\text{CLES}(X,Y) = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} I(x_i, y_j) \qquad (1)$$

$$\text{where } I(x_i, y_j) = \begin{cases} 1 & \text{if } x_i > y_j \\ 0.5 & \text{if } x_i = y_j \\ 0 & \text{if } x_i < y_j \end{cases}$$

These evaluation methods serve our purpose and fit our data, since they do not rely on normality assumptions. While the Mann-Whitney U test provides measures of statistical significance to address our hypotheses, CLES scores provide an easily interpretable understanding of the magnitude and direction of differences between CM scores across languages.

4. Results

We begin by validating our detection model by calculating a standard evaluation metric on the ASVspoof 5 Track 01 evaluation set, which comprises around 680k audio samples. This metric, widely adopted in the literature, involves binary classification of bona fide and spoofed speech in English, measuring performance via the Equal Error Rate (EER). EER reflects the point at which the false acceptance and false rejection rates are equal. Our model achieves an EER of 5.16%, aligning with the top-performing systems in the ASVspoof 5 Challenge [3]. This result confirms the reliability of the selected architecture in a standard monolingual setting and establishes it as a strong baseline for investigating performance under controlled cross-lingual evaluation.

Table 1 presents summary statistics of the CM scores obtained for each language. The first observation is that detector scores are generally high for most languages, which is expected since all samples are synthetic, and we are working with a SOTA detector model capable of identifying synthetic speech in challenging scenarios with high confidence. However, there are clear deviations from this behavior. Although the system exhibits near-perfect detection output for Romanian, its performance can be highly variable, as in the case of Swahili and Hungarian, or even extremely degraded, as evidenced by Ukrainian.

Figure 2 shows an overview of the score distributions alongside individual observations of data points. The score distributions exhibit a bimodal pattern, with values concentrated near the extremes of the range. Scores close to 1 correspond to high confidence in classifying the input as spoofed, whereas

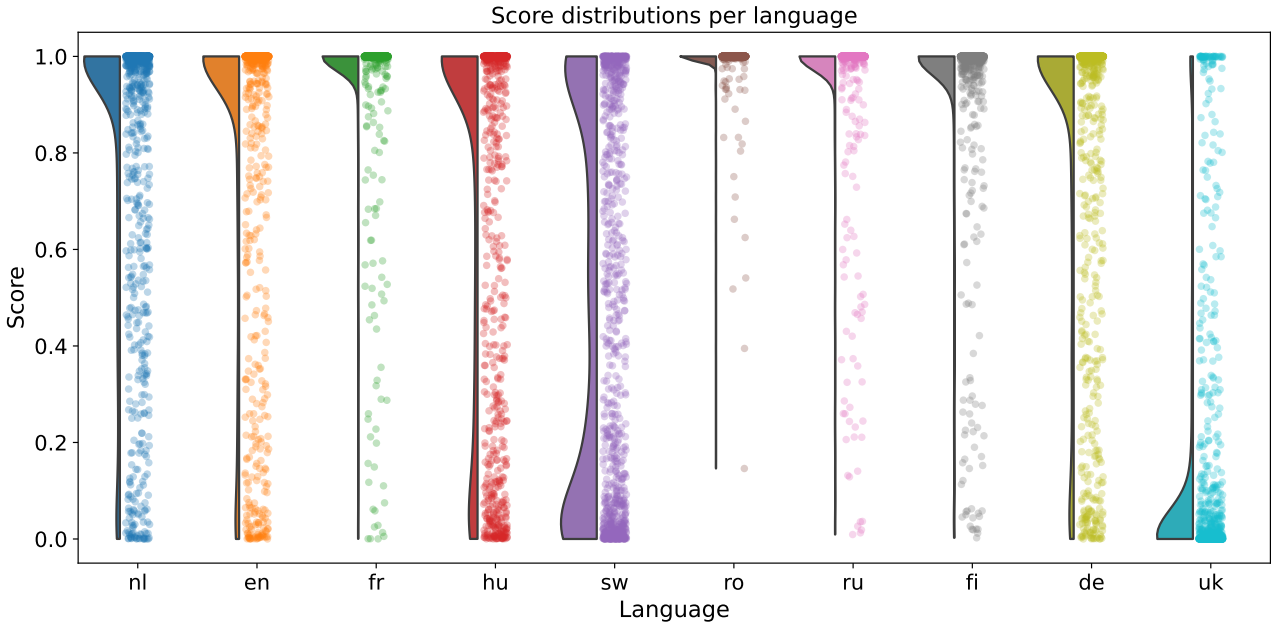

Figure 2: *CM score distributions across the ten MMS languages. Each half violin represents the full distribution of scores assigned to the corresponding language's samples. Individual observations of all scores are shown beside each violin. Data points are jittered along the categorical axis, and darker regions indicate sample concentration. Higher values indicate a higher probability that the audio sample is synthetic. Note that Romanian (ro) and Ukrainian (uk) have opposite distributions, with scores concentrated at the highest and lowest extremes, respectively. We also observe that Swahili (sw) and Hungarian (hu) have highly variable distributions, with scores concentrated at both ends.*

Table 1: *Mean and SD of CM scores by language, computed from balanced populations of 1,000 samples per language.*

Language	ID	**Mean Score**	**Std**
Romanian	ro	0.99	0.05
French	fr	0.97	0.15
Russian	ru	0.97	0.14
Finnish	fi	0.95	0.18
English	en	0.84	0.31
German	de	0.82	0.32
Dutch	nl	0.82	0.30
Hungarian	hu	0.74	0.38
Swahili	sw	0.48	0.41
Ukrainian	uk	0.12	0.27

scores near 0 indicate low spoof likelihood as assessed by the model. The scarcity of mid-range scores suggests that the detector rarely produces uncertain or ambiguous predictions, i.e., scores around 0.5 that would reflect indecision or low separability between spoofed and bona fide-like characteristics, instead favoring highly confident outputs. Romanian, French, Finnish, and Russian exhibit an overwhelming majority of scores clustered tightly at the highest extreme and have lower standard deviations, while German, Hungarian, Swahili, and English, despite also showing a concentration of high scores, also have a significant proportion around the lowest extreme. Ukrainian noticeably has the most distinct scores, as they are clustered around the lowest extreme, implying that the detector struggles to identify these samples as synthetic speech.

Figure 3 presents pairwise comparisons between languages, expanding our analysis beyond descriptive visualizations to quantify the statistical significance of cross-linguistic differences in score distributions. The P-values for each test are reported with respect to three confidence levels, and the CLES values provide effect sizes. There are significant differences for most language pairs, and most of the comparisons yielded $p < 0.001$. Ukrainian and Swahili, which have the two most distinct distributions, are consistently and significantly assigned lower scores when compared to all other languages.

Some apparently similar pairs of score distributions in Figure 2, such as (Finnish, Russian) and (English, Dutch), are significantly different from each other, however, effect sizes vary in these cases. Pairs with significant associated p-values but low effect sizes ($CLES \approx 0.5$) indicate that the observed difference, although consistent, has little magnitude, as in the case of (German, Finnish) and (Romanian, Russian).

A brief auditory analysis was conducted on synthetic audio samples from the MMS-lab subset used in this study to identify acoustic cues that might indicate the presence of detectable artifacts. This analysis focused on broad perceptual aspects of synthesis quality, rather than language-specific features, which would require a dedicated perceptual evaluation conducted with native speakers. The primary motivation was to examine whether noticeable differences in synthesis quality could explain the variation in detector scores observed across languages. Overall, the perceived audio quality of the synthetic speech appeared relatively consistent, with only minor deviations. Romanian samples, those that are detected more consistently by our system, often exhibit audible artifacts, such as buzzing, metallic sounds, or unnatural sibilance of frica-

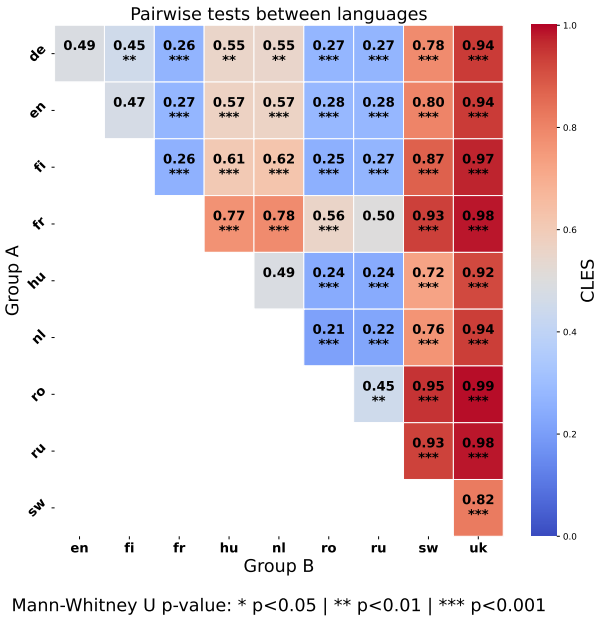

Figure 3: *P-value significance of the pairwise Mann-Whitney U tests and corresponding effect sizes for all languages. A and B are the row and column groups, respectively. We only show* $CLES(A, B)$ *values for simpler visualization, but* $CLES(B, A)$ *is easily obtained by* $1 - CLES(A, B)$.

tives, which may motivate higher detection scores. However, Ukrainian and Swahili, which seem to pose major challenges to the detector, do not seem to deviate from the standard synthesis quality observed for the other languages.

To further assess the influence of various factors on our analysis, we also investigated the potential role of individual speaker identities and speaker gender on the observed detector's performance, but found no significant results. Grouping our data by these factors had no observable effect on the resulting CM score distributions.

5. Discussions

The presented results demonstrate that the detector performance is not uniform across languages, thus allowing us to accept our alternative hypothesis that there is a difference between language groups. Our findings strongly suggest that the detector displays an inherent language bias when trained exclusively with one language.

Unexpectedly, English, the only language seen during the detector's training, did not have the most consistent CM scores. Its standard deviation was relatively high compared to other languages. This is a counterintuitive observation that deserves careful consideration, as it challenges the straightforward assumption that spoofing detection models will exhibit better performance on the precise linguistic data they were exposed to during training. This phenomenon might suggest that our model learned to identify certain potentially language-agnostic artificiality acoustic markers from English, and these markers might stand out more saliently or consistently in the synthesized samples of other languages. It may be the case that linguistic bias is not simply a matter of seen versus unseen languages, but also involves how generalizable the learned features are and how the specific nature of test data in unseen languages interacts with those learned features. This bias likely stems from a combination of factors, potentially including the detector's inability to learn representations that capture certain language-specific phonetic attributes and its over-reliance on specific spectral features or temporal patterns not present in a language's acoustic inventory.

Recent work in cross-lingual TTS systems suggests that the acoustic artifacts introduced during synthesis are not uniform across languages but instead reflect interactions between the phonological structure of the target language and the training regime of the TTS model. Studies show that mismatches in phoneme inventories and prosodic patterns can lead to language-specific distortions, particularly when TTS systems are trained in high-resource languages and applied to phonetically diverse or low-resource targets [19, 20, 21]. These findings offer a plausible explanation for our results: languages like Romanian or Russian may expose clearer or more stereotypical synthesis artifacts that align with the learned decision boundaries of our English-trained CM model, while languages such as Ukrainian or Swahili may produce subtler or structurally unfamiliar artifacts that evade detection. This supports the hypothesis that cross-lingual differences in spoof detection performance may arise not only from the detector's own biases but also from how synthesis artifacts manifest differently across languages.

Detection performance differences between languages belonging to the same linguistic family (e.g. Russian and Ukrainian, both in the East Slavic family), raise questions concerning whether these differences are strictly due to factors such as phoneme realizations and phonotactic patterns (i.e., the possible ways that sounds can be arranged in a language) or if they're amplified by more granular acoustic characteristics of speech samples which may escape the human auditory perception. Although this work may shed light on the source of bias, further investigation is needed to analyze all relevant factors and potentially provide solutions to develop mitigation strategies, as this goes beyond our current scope.

Our findings include occurrences of statistically significant differences between languages, even when the corresponding effect sizes were small. Such outcomes reveal consistent, although subtle, non-random variations. While these differences are minor in magnitude, their relevance becomes critical in the context of voice biometric systems. In high-stakes security applications, small but systematic CM score shifts can lead to disparate False Acceptance or Rejection Rates when universal decision thresholds are applied. This underscores the serious implications of linguistic bias for the overall robustness, fairness, and trustworthiness of spoof detection systems.

6. Conclusion

This work presents a systematic investigation into the role of language identity as a source of bias in a SOTA synthetic speech detection architecture. We trained the model exclusively on English speech and evaluated it under a controlled experimental setup on a multilingual dataset of spoofed speech samples generated by a standardized TTS architecture. Our findings reveal significant language-dependent shifts in detector performance. While spoofed speech in some languages (e.g., Romanian, Russian) is detected with near-perfect reliability, others (e.g., Ukrainian, Swahili) pose substantial challenges to the detector, despite similar synthesis protocols and model configurations.

Remarkably, the model performs better for several unseen languages than it does for English, the only language on which

it was trained. This finding suggests that language bias in spoofing detection is not simply a function of exposure but likely arises from complex interactions between learned representations and language-specific phonetic or acoustic characteristics. This language bias exposes limitations in the cross-lingual generalization capability of current spoofing countermeasures, and future work is needed to identify its underlying causes, potentially by examining how specific phonetic inventories, prosodic patterns, or acoustic manifestations of synthesis artifacts differ across languages and how these variations are processed by current detection models.

To support future research in this area, we release a fully reproducible codebase that allows researchers to replicate our experiments and extend them to new settings. The repository provides modular support for swapping front-end and back-end architectures, changing evaluation datasets, and configuring different bias-sensitive metrics. This framework is designed to facilitate broader investigations into cross-lingual bias in spoofing detection, enabling easy adaptation to other countermeasure systems and multilingual corpora.

As synthetic speech technologies continue to evolve and proliferate globally, achieving extreme natural-sounding quality, it becomes urgent to develop detection models that are either robustly language-agnostic or explicitly account for linguistic diversity. Our observations highlight the need for a paradigm shift in the development and evaluation of synthetic speech detectors, underscoring the importance of cross-lingual testing and bias mitigation strategies to ensure robust, fair, and secure deployment across diverse linguistic communities.

7. Acknowledgements

This study is partially funded by CAPES – Finance Code 001. We thank the São Paulo Research Foundation (Fapesp) Horus project, Grant #2023/12865-8 for the partial funding. We also thank the São Paulo Research Foundation (Fapesp) BI0S project, Grant #2020/09838-0 for the partial funding. João Lima, Paula Costa, and Victor Moreno are affiliated to the Dept of Computer Engineering and Automation (DCA), Faculdade de Engenharia Elétrica e de Computação, and are part of the AI Lab, Recod.ai, Institute of Computing, UNICAMP.

8. References

[1] A. K. S. Yadav, K. Bhagtani, D. Salvi, P. Bestagini, and E. J. Delp, "Fairssd: Understanding bias in synthetic speech detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4418–4428.

[2] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[3] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, N. Evans, K. A. Lee, J. Yamagishi, M. Jeong, G. Zhu, Y. Zang, Y. Zhang, S. Maiti, F. Lux, N. Müller, W. Zhang, C. Sun, S. Hou, S. Lyu, S. Le Maguer, C. Gong, H. Guo, L. Chen, and V. Singh, "ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech," *Computer Speech & Language*, vol. 95, p. 101825, 2026.

[4] 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," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–7.

[5] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[6] "ISO/IEC 30107-1:2023 information technology — biometric presentation attack detection — part 1: Framework," International Organization for Standardization, ISO/IEC, Geneva, Switzerland, 2023, available at: https://www.iso.org/standard/83828.html.

[7] H. Tak, "End-to-end modeling for speech spoofing and deepfake detection," Ph.D. dissertation, Sorbonne Université, 2023.

[8] 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.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[10] Y. Chen, H. Wu, N. Jiang, X. Xia, Q. Gu, Y. Hao, P. Cai, Y. Guan, J. Wang, W.-L. Xie, L. Fang, S. Fang, Y. Song, W. Guo, L. Liu, and M. Xu, "Ustc-kxdigit system description for asvspoof5 challenge," in *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 2024, pp. 109–115.

[11] 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: crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 2024, pp. 1–8.

[12] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech 2015*, 2015, pp. 2037–2041.

[13] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[14] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2019, pp. 1–10.

[15] I. Celeste, "M-ailabs speech dataset," https://github.com/imdatceleste/m-ailabs-dataset, 2025, accessed: 2025-05-28.

[16] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 100–106.

[17] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[18] K. O. McGraw, "A common language effect size statistic." *Psychological bulletin*, vol. 111, no. 2, p. 361, 1992.

[19] K. Zhou, S. Zhao, Y. Ma, C. Zhang, H. Wang, D. Ng, C. Ni, T. H. Nguyen, J. Q. Yip, and B. Ma, "Phonetic enhanced language modeling for text-to-speech synthesis," in *Interspeech 2024*, 2024, pp. 3440–3444.

[20] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," in *Interspeech 2019*, 2019, pp. 2080–2084.

[21] J. Ye, H. Zhou, Z. Su, W. He, K. Ren, L. Li, and H. Lu, "Improving cross-lingual speech synthesis with triplet training scheme," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6072–6076.