Are Current Fake Audio Detection Language Independent?

Wen Wen
IMS, University of Stuttgart
st186079@stud.uni-stuttgart.de

Abstract

This report investigates the influence of language on the detection of fake audio. There are two generating architectures of fake audio are selected, which are Griffin Lim and VITS. The project employs two detection models, RawNet3 and SpecRNet. Eight languages, spanning three language families, are chosen for evaluation. Results reveal that the impact of language on fake audio detection is not universally consistent across all generation and detection models. Variations in language independence are observed across different generation and detection methodologies.

1 Introduction

As technology for generating fake audio continues to advance, detecting fake audio presents new challenges. What role does language play in fake audio detection? Are there variations across different detection models? This project primarily investigates the impact of language on both raw waveform input model RawNet3 and spectrum feature input model SpecRNet using the MLAAD (Müller et al., 2024) spoof dataset.

2 Description of Code Execution

Please refer to the ReadMe file in Github[1] to execute code.

3 Experimental Setup

3.1 Dataset

The dataset utilized in this project, MLAAD (Multi-Language Audio Anti-Spoof Dataset) [2], is generated from Mailabs and spans across multiple languages, each generated through various TTS architectures. Two distinct architectures for generating fake audio were selected: Griffin Lim, representing a traditional method, and VITS, which utilizes a neural network-based approach. The languages included in this project comprise English, German, French, Russian, Italian, Spanish, Polish, and Ukrainian. Each language of each generation architecture contains 1000 spoof samples and 1000 real samples of the same language are randomly selected to ensure a balanced dataset. Consequently, each generating architecture of each language, such as English fake audio generated using the Griffin Lim architecture, contains a total of 2000 data samples.

To prevent overlap between training and test data, the spoof data is randomly partitioned into three subsets: 70% for training, 15% for validation, and 15% for testing. Similarly, an equal amount of real data is selected in the same way and combined together.

3.2 Models

Two models for detecting deepfakes have been chosen: RawNet3 (weon Jung et al., 2022) and SpecRNet (Kawa et al., 2022). Both models are built upon the foundation of RawNet2. The key distinction lies in their input and architectural design. RawNet3 operates with raw audio input, devoid of spectrogram-like features. Its architecture adopts a hybrid form, merging elements of ECAPA-TDNN with RawNet2, supplemented by additional features such as logarithm and normalization. On the other hand, SpecRNet utilizes LFCC as its Frontend algorithm. It is a novel spectrogram–based model inspired by RawNet2 backbone.

3.3 Experiments

As illustrated in the figure1, experiments are categorized into four groups.

Firstly, Griffin Lim is utilized as the training data architecture, employing it to train the RawNet3 detection model. Then test this model with additional

[1] https://github.com/WenW186079/SpeechTechnology_FakeAudioDetection

[2] https://owncloud.fraunhofer.de/index.php/s/tL2Y1FKrWiX4ZtP#editor

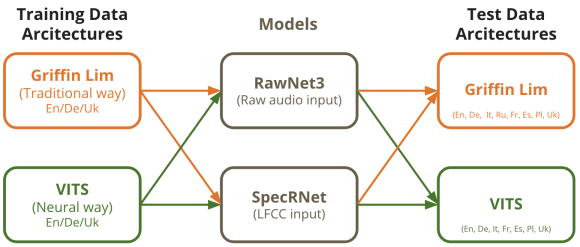

Figure 1: Experiments instruction. Note that the orange route can only lead to orange tests, and likewise, the green route can only lead to green tests.

languages generated by the Griffin Lim architecture. Similarly, we utilize Griffin Lim as the training data architecture for the SpecRNet detection model, followed by testing with more languages generated by Griffin Lim.

Additionally, VITS is employed as the training data architecture, applied to train the RawNet3 detection model and subsequently test it with further languages generated by VITS. Similarly, VITS is used as a training data architecturefor the SpecRNet detection model and test it with additional languages generated by VITS.

For example, I start with English fake audio generated by Griffin Lim, which is used to train the RawNet3 detection model. Then individually test this model with various languages, including German, Italian, Russian, French, Spanish, Polish, and Ukrainian, all of which contain spoof audio only generated by Griffin Lim. Then take the detection model trained on German Griffin Lim data and cross-test the results with the spoof samples in all previously mentioned languages. This iterative approach is repeated for each language, ensuring comprehensive evaluation.

The main idea is to ensure consistency in the fake audio structure between training and testing data, with variation limited to language. Here, the speaker is also a potential influencing factor. Therefore, more speakers are added to make the model less sensitive to speakers, see as table1.

4 Results and Analysis

Detailed results can be seen from file[3]. Here, the focus lies in analyzing these findings.

From figure 2 we can see, when using Griffin Lim to generate training and test fake audio, whether the results obtained by RawNet3 or SpecRNet detect model, the trends in EER results remain consistent. Specifically, performances in German, Spanish, and Russian have notably declined, each to varying degrees, indicating potential issues with the data quality itself. If language were a influencing factor, we would expect unseen languages and seen one to exhibit distinct behaviors, yet no such trend is observed.

[3] https://github.com/WenW186079/SpeechTechnology_FakeAudioDetection/blob/main/test_results.csv

Language	Gender		Total
	Female	Male	
EN	2	1	3
DE	4	1	5
IT	1	1	2
RU	1	2	3
FR	2	3	5
ES	1	2	3
PL	1	1	2
UK	1	5	6

Table 1: The number of real-sample speakers for each language

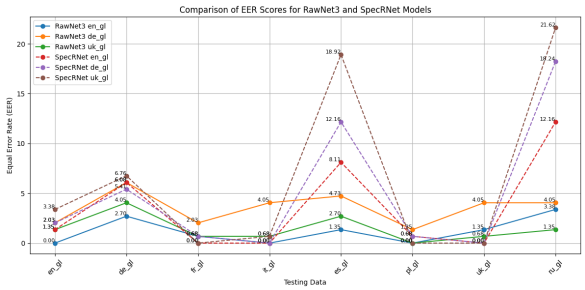

Figure 2: Comparison of EER Scores for RawNet3 and SpecRNet Models

Based on language homology, we assume that Germanic languages such as English and German share similarities, Romance languages like French, Italian, and Spanish exhibit commonalities, and Slavic languages including Polish, Ukrainian, and Russian demonstrate similarities.

However, linguistic homology does not consistently provide clear advantages. For example in figure 2, RawNet3 trained on German performs worse than one trained on Ukrainian data, when testing German and English. Additionally, in all the testing Russian data result, SpecRNet trained on Ukrainian exhibits the poorest performance.

Also seen in figure 3, the score for each language family is calculated as the average of its constituent languages. Although for Ukrainian Griffin Lim data trained in RawNet3 model, Slavic performs the best among the three languages, Ukrainian Griffin

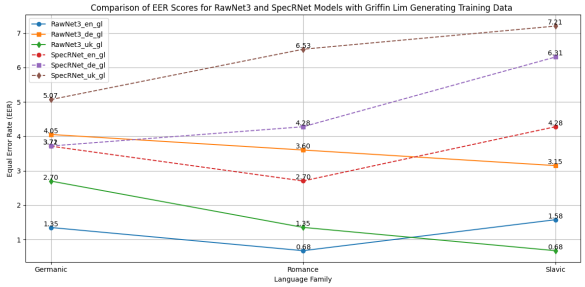

Figure 3: Comparison of EER Scores for RawNet3 and SpecRNet Models with Griffin Lim Generating Training Data.

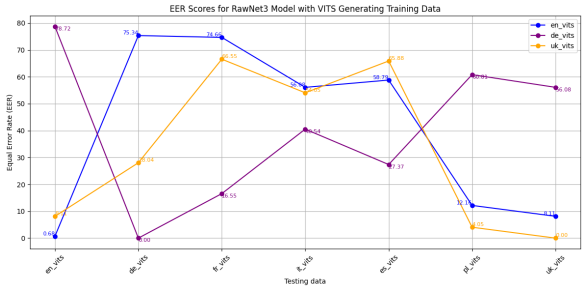

Figure 4: EER Scores for RawNet3 Model with VITS Generating Training Data

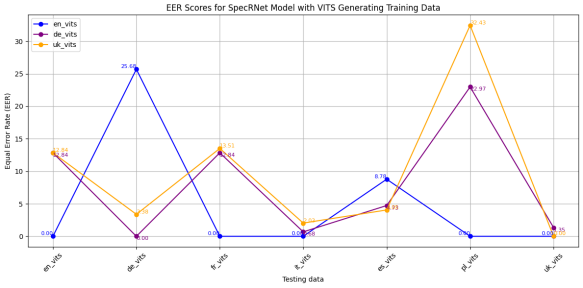

Figure 5: EER Scores for SpecRNet Model with VITS Generating Training Data

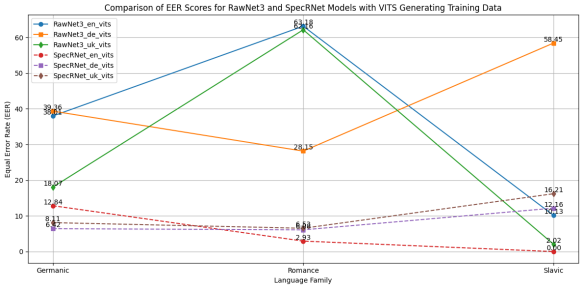

Figure 6: Comparison of EER Scores for RawNet3 and SpecRNet Models with VITS Generating Training Data

Lim data trained in SpecRNet model exhibits the opposite trend, with Slavic performing the worst among the three languages.

In figure 4, for RawNet3 model with VITS Generating training data, it is crucial to see the language when detecting it with RawNet3 detection model. Specifically, RawNet3 trained on German VITS data performs well only when detecting fake audio in the same language(here is German), while its performance is poor in other languages. Similarly, this trend holds for RawNet3 trained on English VITS data and on Ukrainian VITS data.

When fake audio generated by VITS trained in SpecRNet Model, as in figure 5, language is not a significant factor when detected by SpecRNet model. For instance, SpecRNet trained on English VITS data performs well in tests on French, Italian, Polish, and Ukrainian VITS data without prior knowledge of the language. Additional, we can see that SpecRNet trained on German VITS data and SpecRNet trained on Ukrainian VITS data exhibit consistent performance trends across all test results. However, English, which belongs to the same language family of German, does not show consistency.

Figure 6 shows that when it comes to fake audio generated by VITS architecture, the SpecRNet detection model outperforms the RawNet3 detection model. Furthermore, it is speculated that when handling datasets with limited size, employing spectrogram-like features for neural spoof generation tends to result in improved performance.

5 Conclusion

The influence of language varies across different fake audio generating architecture and detection models. According to the MLAAD dataset, for RawNet3 detection model training on fake audio generated by VITS, language plays a crucial role in detection. However, in other cases, language does not exhibit a significant effect.

6 Further research

There are many aspects worthy of further research.

In terms of language similarity, we assume that languages from the same language family exhibit higher similarities, does this hold true for speech? Will there be speech-based language similarity?

In terms of the quality of the dataset itself, given its small size, inevitably influences test outcomes. One potential remedy is to do data argumentation or generate more data.

Moreover, in terms of speakers, can we mitigate their impact on the model by extracting and modifying speaker embedding?

In this project, only English, German, and Ukrainian serve as training data, yet more languages in subsequent training sets may also lead a better insight for the influence of language on the detection of fake audio.

In short, there are still many aspects worthy of further study.

References

Piotr Kawa, Marcin Plata, and Piotr Syga. 2022. Specrnet: Towards faster and more accessible audio deepfake detection. *Preprint*, arXiv:2210.06105.

Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. Mlaad: The multi-language audio anti-spoofing dataset. *International Joint Conference on Neural Networks (IJCNN)*.

Jee weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. 2022. Pushing the limits of raw waveform speaker recognition. *Preprint*, arXiv:2203.08488.