Are Current Fake Audio Detection Language Independent?

Project of Speech Technology

Wen Wen

IMS, University of Stuttgart

st186079@stud.uni-stuttgart.de

Institut für **Maschinelle Sprachverarbeitung**

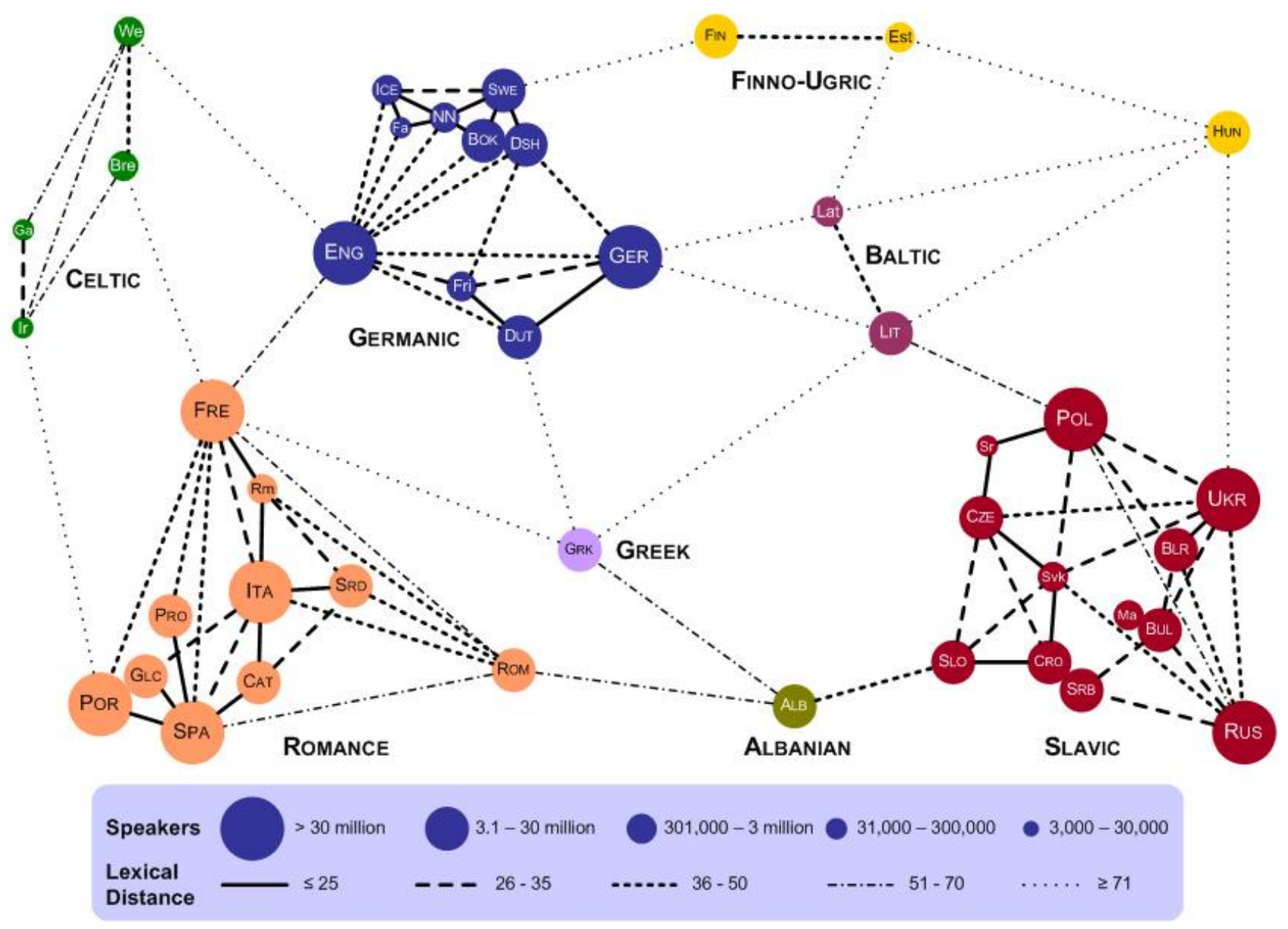
Research questions

Does language play a significant role in detecting fake audio using a raw waveform input model? What about when employing a spectrogram features input model?

Dataset

- Spoof Dataset : MLAAD
- Equal number of spoof and bona-fide samples (1000+1000). As the MLAAD is generated based on Mailabs, the duration of each label is similar.
- Selected fake audio generating architectures: Griffin Lim as the traditional method and VITS as the neural network-based method.
- Selected languages: English, German, French, Russian, Italian, Spanish, Polish, Ukrainian.

Language similarity

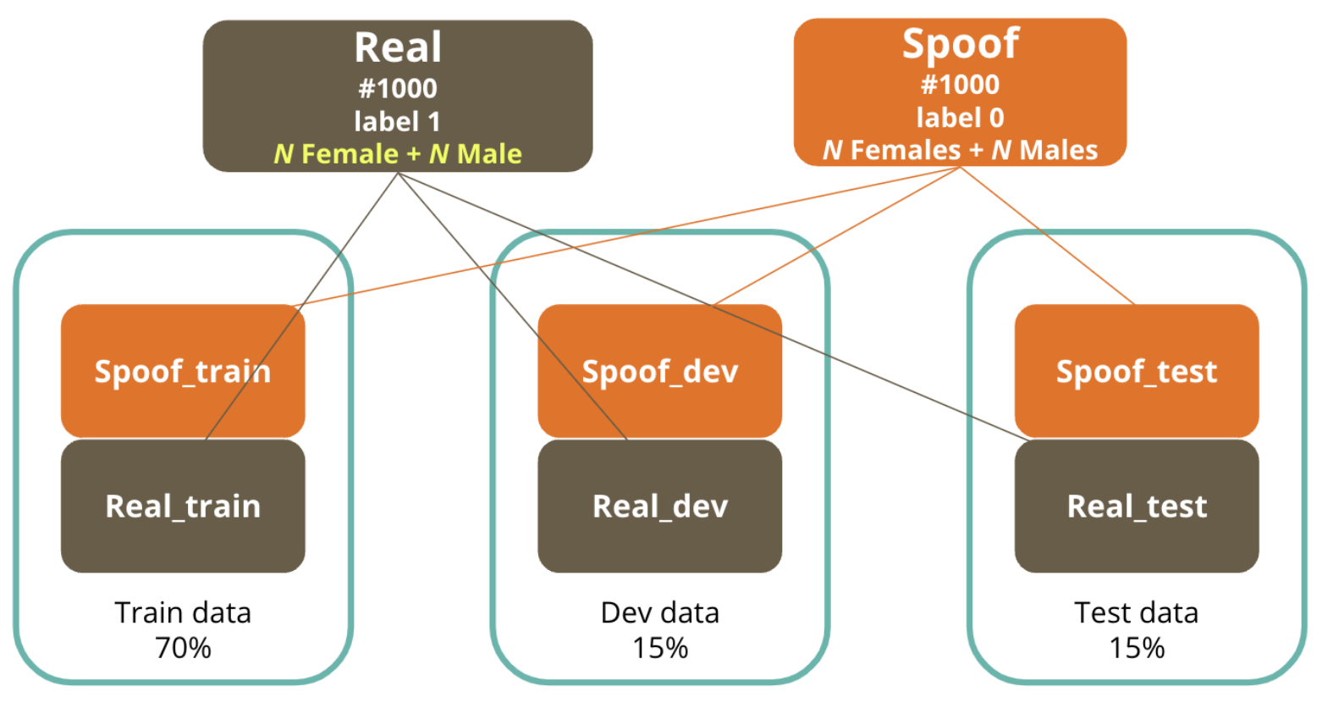


- Germanic: English(EN), German(DE)
- Romance: French(FR), Italian(IT), Spanish(ES)
- Slavic: Polish(PL), Ukrainian(UK), Russian(RU)

Models

- RawNet3
  - Input: raw audio
  - No spectrogram-like features
  - The architecture is in a hybrid form of the ECAPA-TDNN and the RawNet2 with additional features including logarithm and normalisation.
- SpecRNet
  - Frontend algorithm: LFCC
  - A novel spectrogram–based model inspired by RawNet2 backbone.

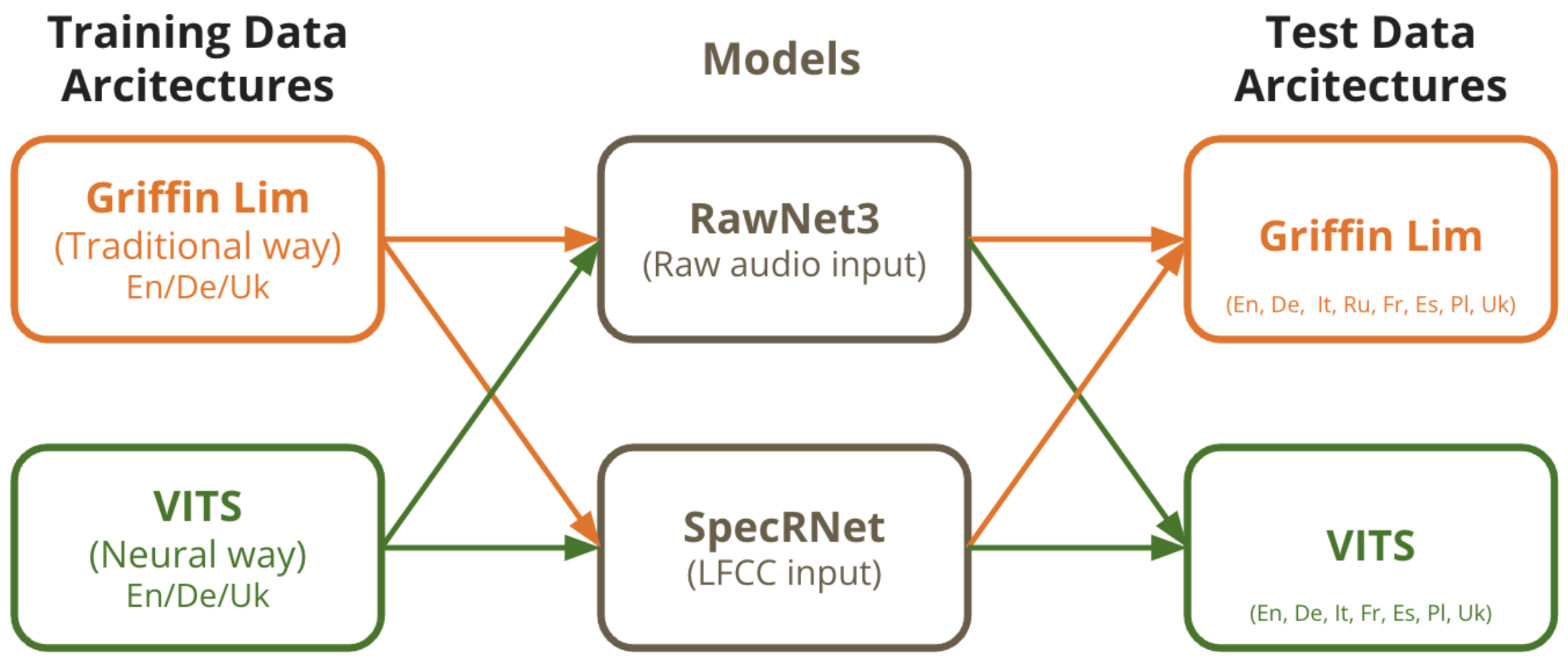
Meta Files Generation



The number of real-sample speakers:

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

Experiments



Training Data Arcitectures

Griffin Lim (Traditional way) En/De/Uk

VITS (Neural way) En/De/Uk

Models

RawNet3 (Raw audio input)

SpecRNet (LFCC input)

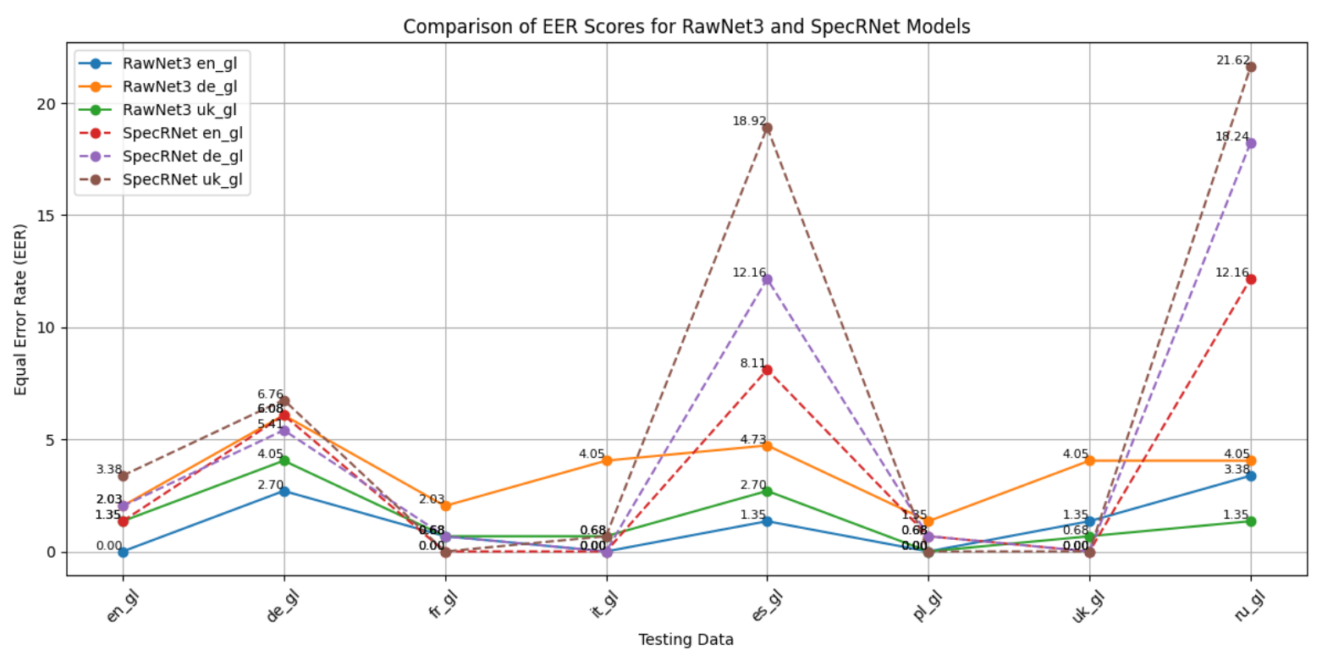
Test Data Arcitectures

Griffin Lim (En, De, It, Ru, Fr, Es, Pl, Uk)

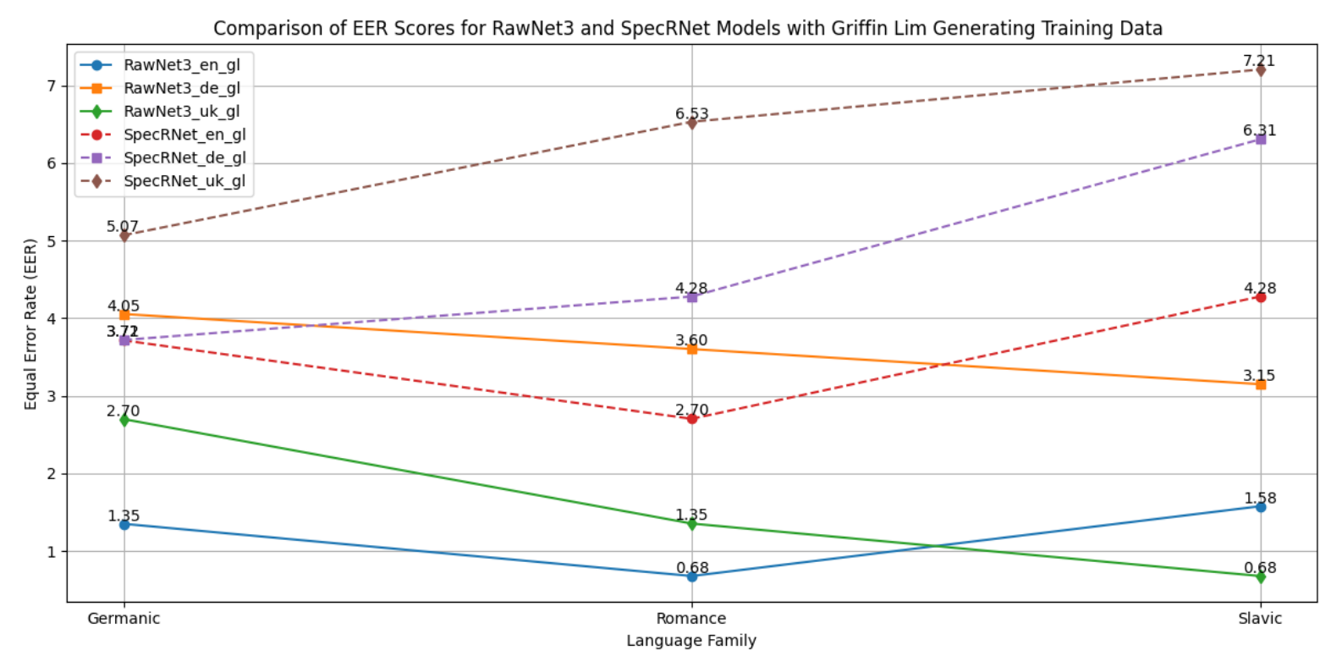
VITS (En, De, It, Fr, Es, Pl, Uk)

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.

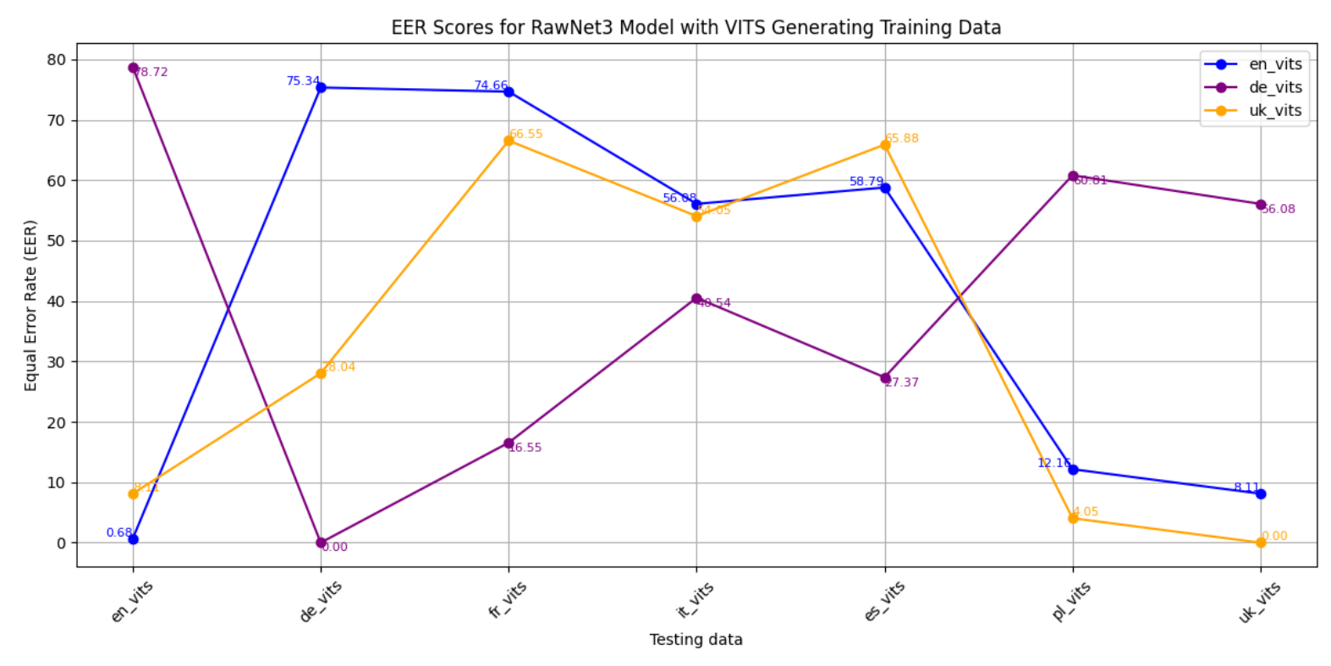
Results



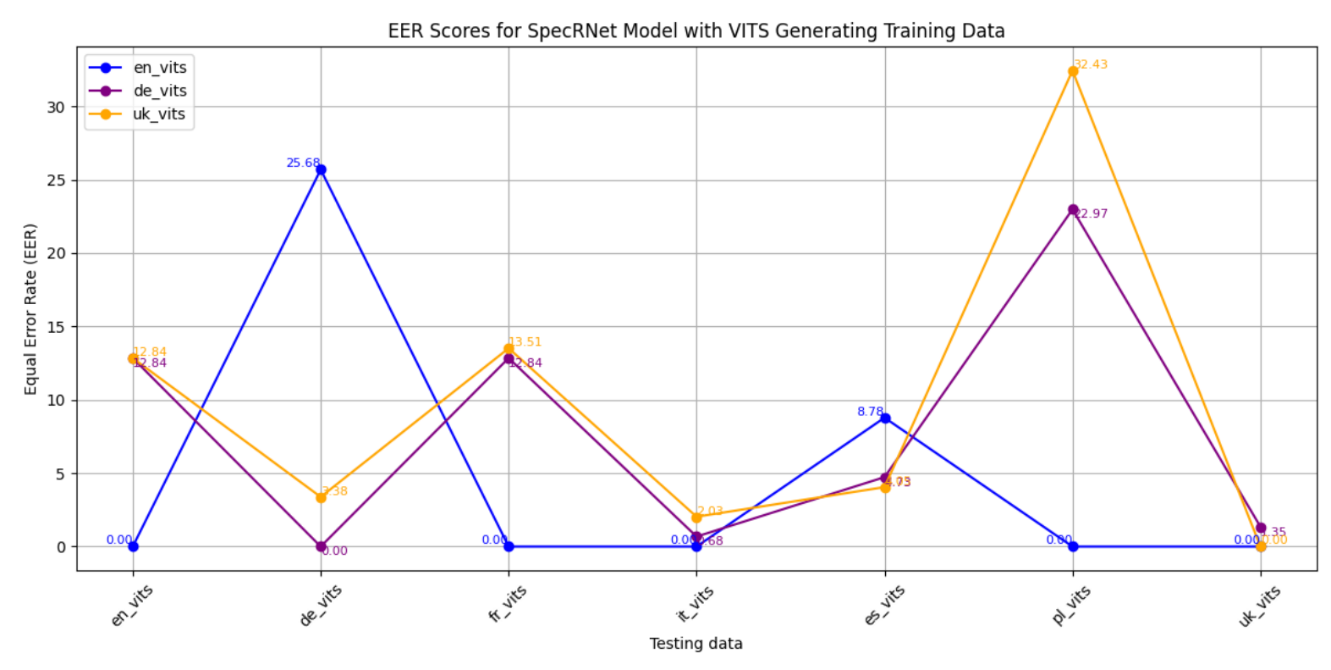
- When using Griffin Lim to generate training and test fake audio, whether the results obtained by RawNet3 or SpecRNet detect model, the trends in EER results remain consistent. Specifically, performances in German, Spanish, and Russian have notably declined, each to varying degrees.
- Linguistic homology doesn't confer clear advantages. For instance, the performance of *RawNet3(Train(de_gl))* in *Test(en_gl)* and *Test(de_gl)* is not as good as *RawNet3(Train(uk_gl))*. Additionally, *SpecRNet(Train(uk_gl))* exhibits the poorest performance in *Test(ru_gl)*.



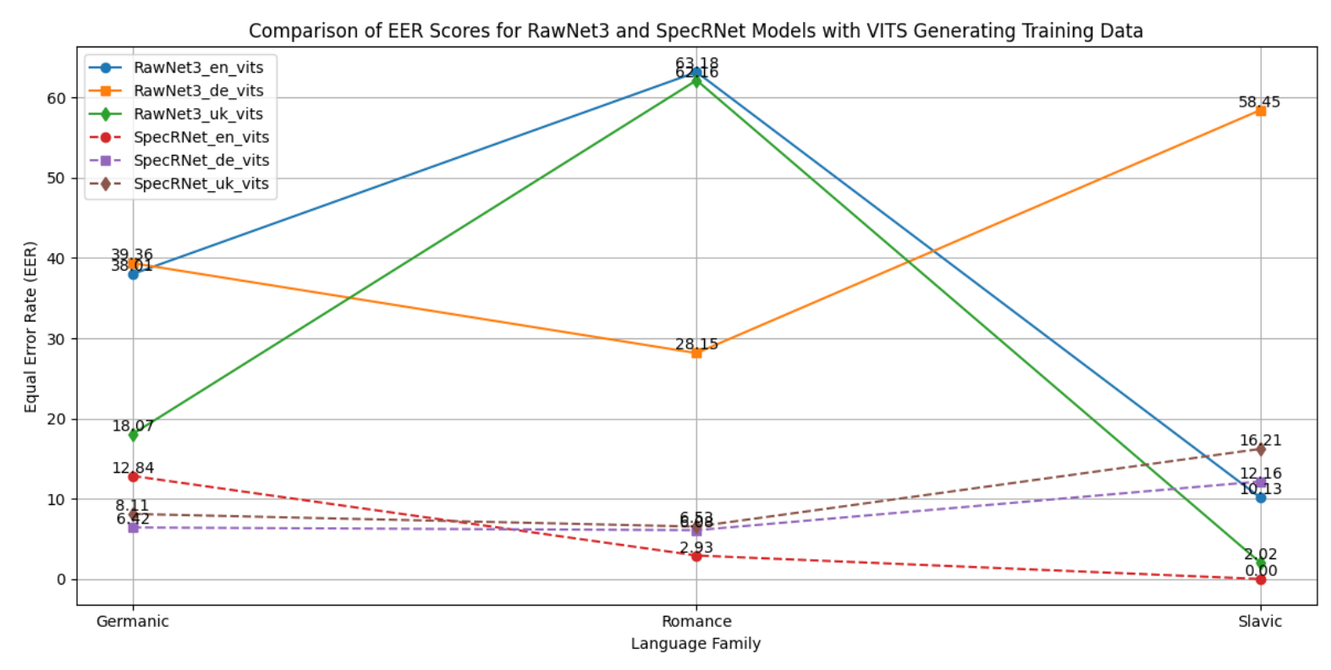
- The score for each language family is calculated as the average of its constituent languages. For example, Germanic is computed as the average score of English and German.
- Although for *RawNet3(Train(uk_gl))*, Slavic performs the best among the three languages, *SpecRNet(Train(uk_gl))* exhibits the opposite trend, with Slavic performing the worst among the three languages.



- For the fake audio generated by VITS, 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.



- For fake audio generated by VITS, 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.
- 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.



- 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.

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.