Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations

Wanying Ge, Jose Patino, Massimiliano Todisco and Nicholas Evans

EURECOM, Sophia Antipolis, France

poster: 4273

Introduction

Motivation

Despite recent success in **developing effective solutions** for spoofing detection, little is known to understand **what information** is being used to influence the classifier output.

In this paper we

- Use SHapley Additive exPlanations (SHAP) to gain insights about how anti-spoofing solutions work.
- Analyse difference between classifiers, also the difference between sub-band features.

SHapley Additive exPlanations

Definition

SHAP value φ_i can be both positive and negative to reflect the relative (un)importance of a particular feature to a classifier output. To obtained φ_i , a classifier *f(x)* is trained twice, with and without the inclusion of a chosen feature i :

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \delta_i(S)$$

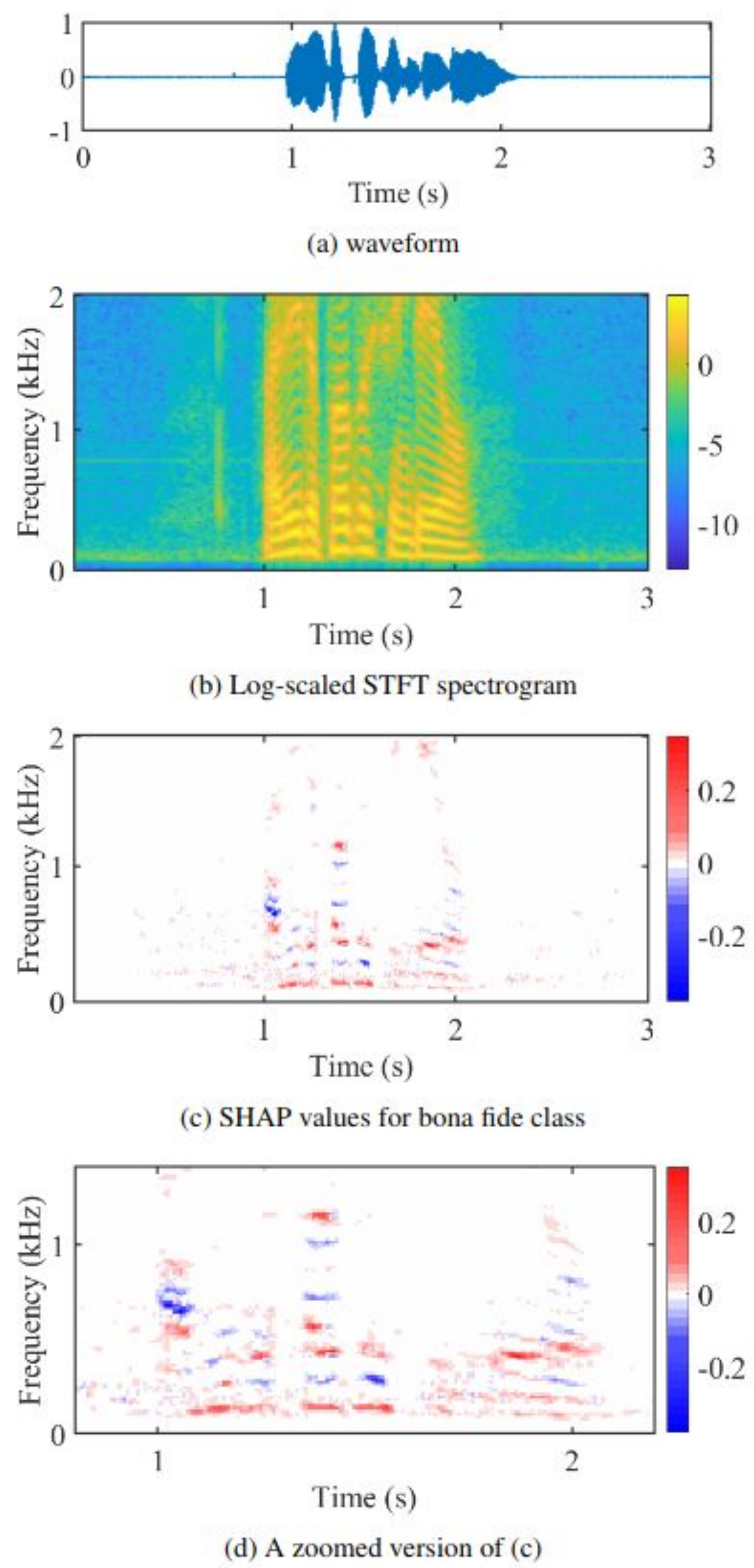
where S is a feature subset of full set of features F, and δ_i is the prediction difference of feature i being presented and absent.

When the classifier *f(x)* is a complex model, such as a deep neural network, to avoid repetitive retraining of the network, the calculation of SHAP value is simplified to:

$$f(\mathbf{x}) \approx g(\mathbf{x}') = \phi_0 + \sum_{i=1}^{D} \phi_i x'_i$$

where $g(x)$ is the approximated explanation model of , and $\mathbf{x}'$ is the simplified feature that only contains 0 (absence of feature) or 1 (presence of feature).

The obtained SHAP values are of the same size as the input feature, and can be visualised in a similar manner to the spectrogram.



An audio waveform is shown in (a), and the corresponding temporal-spectral spectrogram is shown in (b). SHAP values for bona fide class are shown in (c) and (d), each with positive value being red and negative value being blue.

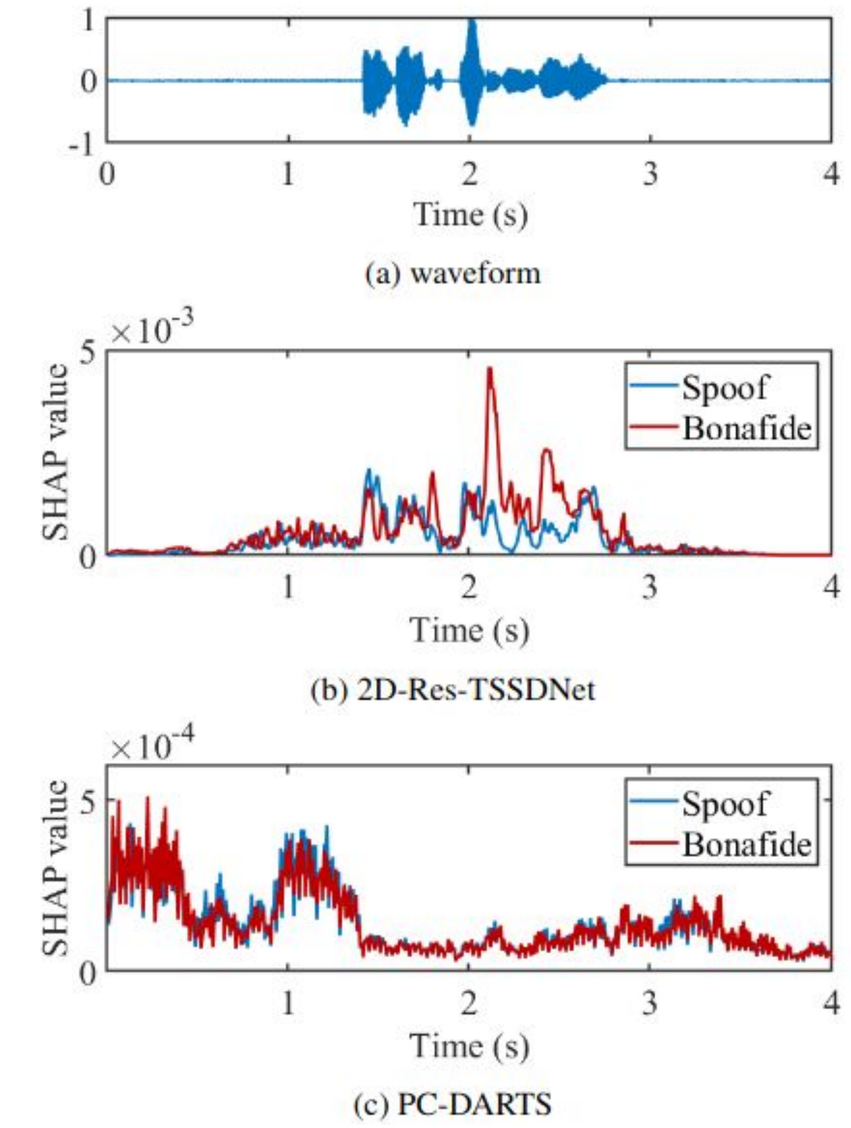
Explainability for spoofing detection

Visualisation

- Only positive SHAP values for both classes are shown; negative values are approximately symmetric to the positive values.
- Average SHAP values are shown across the full spectrum, focusing on either temporal axis or spectral axis.

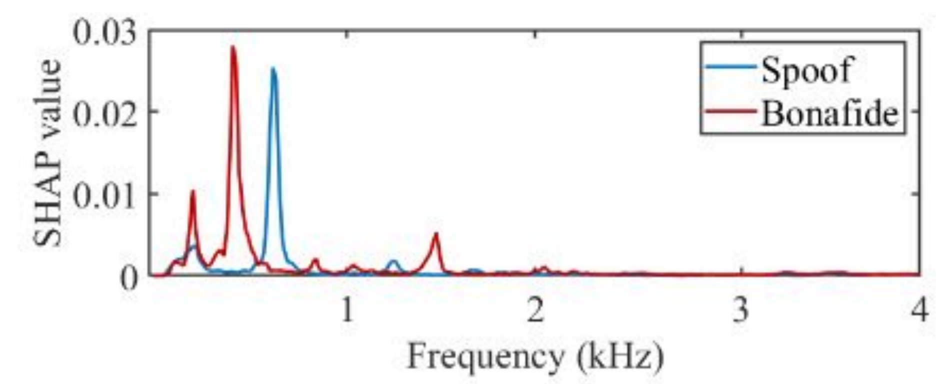
Temporal analysis

- A spoofed audio file from ASVspoof2019 LA Evaluation set, LA_E_4428024.
- In subfigure (b), the averaged SHAP values for both bonafide and spoof are higher in the speech interval, while in (c), values are higher in the non-speech interval.
- The 2D-Res-TSSDNet model detects artefacts in speech intervals, while the PC-DARTS model uses information mostly in non-speech intervals.



Spectral analysis

- A spoofed audio file from ASVspoof2019 LA Evaluation set, LA_E_2634822.
- A greater support for bona fide class can be noticed at 0.5kHz, while for spoofed class, it's 0.6kHz.
- This may imply that the artefacts for detecting two classes are located at different frequency regions.



Conclusion

- SHAP values can reveal the influence of individual features upon classifier behaviour.
- For a given classifier, SHAP values can be used to highlight the attention of the classifier at low-level spectro-temporal level.
- DNN models can use different temporal or spectral intervals from the same waveform input for decision making.
- Future work includes using SHAP to explore differences between spoofing attack algorithms, and to explain the performance difference among well-trained classifiers.

Reference

- S. M. Lundberg, S.-i. Lee and D. Fohr, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.
- S. Sivasankaran, E. Vincent and D. Fohr, "Explaining deep learning models for speech enhancement," in Proc. INTERSPEECH, 2021, pp. 696–700.
- W. Ge, M. Panariello, J. Patino et al., "Partially-connected differentiable architecture search for deepfake and spoofing detection," in Proc. INTERSPEECH, 2021, pp. 4319–4323.
- G. Hua, A. B.-j. Teoh and H. Zhang, "Towards end-to-end synthetic speech detection," IEEE Signal Processing Letters, vol. 28, pp. 1265–1269, 2021.

Acknowledgements

The first author is supported by the TReSPAsS-ETN project funded bythe European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813. The second author is supported by the ExTENSoR project funded by the French Agence Nationale de la Recherche (ANR).