Towards a Transparent and Interpretable Strategy for Spoofed Speech Detection

Carolina Lins Machado[1], Xin Wang[2], and Junichi Yamagishi[2]
c.machado@nfi.nl; {wangxin|jyamagis}@nii.ac.jp
[1]*Netherlands Forensic Institute, The Hague, The Netherlands* [2]*National Institute of Informatics, Tokyo, Japan*

Introduction

- Artificially-generated (spoofed) speech poses unprecedented challenges for forensic investigators and legal systems [1,2].
- Many detection systems are "black-boxes" and in forensic contexts the interpretability of conclusions is crucial [3,4].
- A fair justice outcome requires decision outputs understandable and justifiable to all parties involved in the process [4,5].

Can acoustic-phonetic features and explainable machine learning approaches provide clarity on the process of spoofed speech detection?

Goals:

(i) Understand how acoustic-phonetic features perform in various spoofing types.

(ii) Provide a baseline against which future state-of-the-art attacks can be compared to.

Method

Datasets: ASVspoof 2015 [6], 2019 [7], 2021 [8], 5 [9] and Deepfake-Eval-2024 [10].

Features (extracted in Praat [11]):

Local		*Global*	
Feature	***Measurement***	***Feature***	***Measurement***
Formants	F1; F2; F3	Harmonic-to-noise ratio	Mean
Spectral tilt	H1-H2; H1-A1; H1-H2; H1-A3		Standard Deviation
	A1-A2; A1-A3; A2-A3	Peaks-per-second	
Jitter	Local	Intensity slopes	Mean
	Absolute		Standard Deviation
	Relative average perturbation	Signal periodicity	2kHz-4 kHz
	Difference of difference of periods		4 kHz-6 kHz
	Five-point period perturbation quotient		6 kHz-8 kHz
Shimmer	Local	F0 wiggliness	
	Three-point amplitude perturbation quotient	F0 spaciousness	
	Five-point amplitude perturbation quotient	F0 slopes	Mean
	Average absolute difference		Standard Deviation
		Spectral flatness	
		Spectral centroid	

Experiment 1: Understand the decision process

Binary Classification with Decision Trees (sklearn)

- Balanced datasets (train, dev, eval) divided into seen and unseen attacks
- Hyperparameter tuning with grid search/10-fold CV
- Three full models (different data partitions) subsequently pruned.

Experiment 2: Assess the relationship between features and ML algorithms

Binary Classification with AutoML pipeline (LazyPredict)

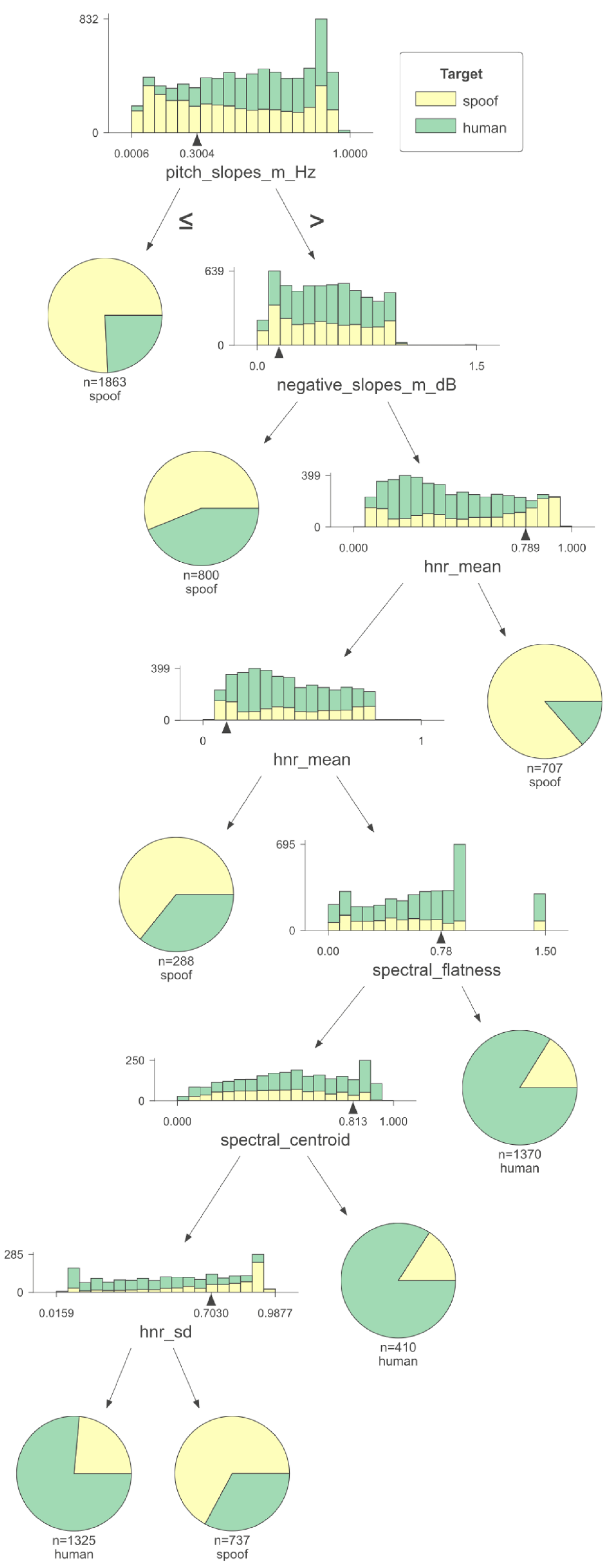
- 26 classifiers, including linear and tree-based models; ensemble methods; SVM; Naïve Bayes; Discriminant Analysis algorithms; K-NNs; Multi-Layer Perceptron; Nearest Centroid; Calibration- and Propagation-based models; Dummy Classifier.

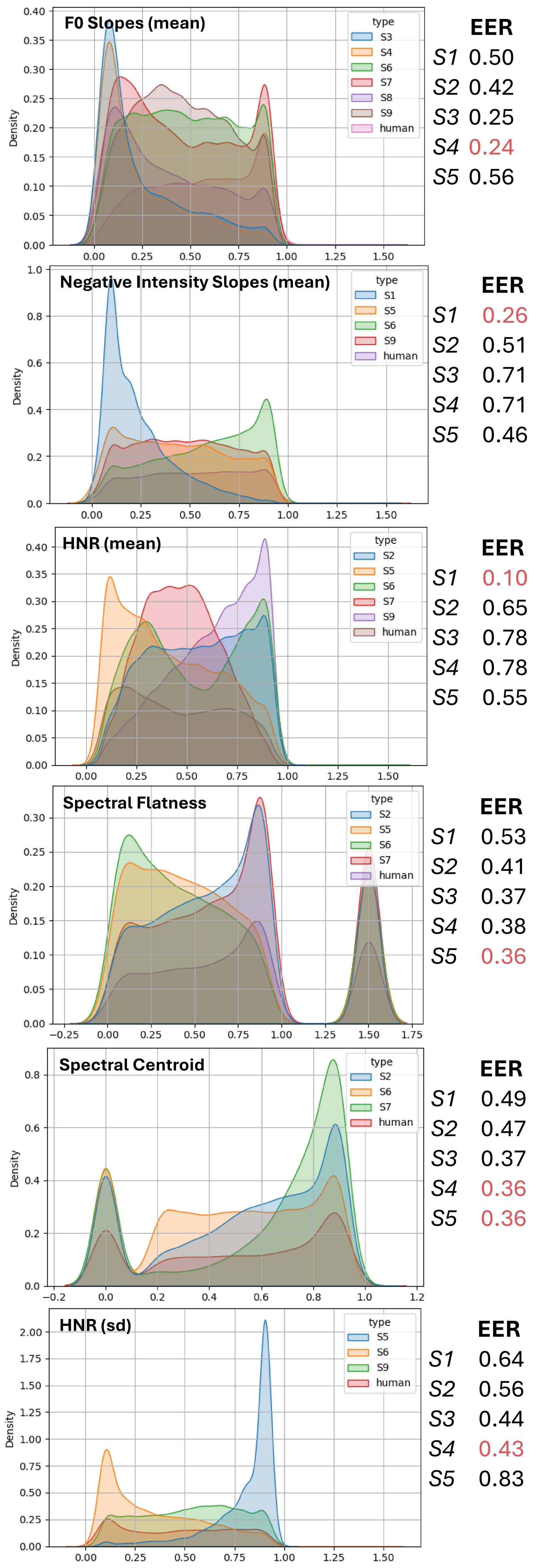
References: **[1]** Gambín, Á. F., Yazidi, A., Vasilakos, A. V., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: Current and future trends. *Artificial Intelligence Review*, *57*(3). **[2]** Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 910–932. **[3]** Mitchell, F. (2014). The use of Artificial Intelligence in digital forensics: An introduction. *Digital Evidence and Electronic Signature Law Review*, *7*(0). https://doi.org/10.14296/deeslr.v7i0.1922. **[4]** Hall, S. W., Sakzad, A., & Choo, K.-K. R. (2022). Explainable artificial intelligence for digital forensics. *WIREs Forensic Science*, *4*(2), e1434. **[5]** Siegel, D., Kraetzer, C., Seidlitz, S., & Dittmann, J. (2024). Media Forensic Considerations of the Usage of Artificial Intelligence Using the Example of DeepFake Detection. *Journal of Imaging*, *10*(2). **[6]** Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., & Sizov, A. (2015). ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. *Interspeech 2015*, 2037–2041. **[7]** Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., Juvela, L., Alku, P., Peng, Y.-H., Hwang, H.-T., Tsao, Y., Wang, H.-M., Maguer, S. L., Becker, M., Henderson, F., ... Ling, Z.-H. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, *64*, 101114. **[8]** Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., & Delgado, H. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 47–54. **[9]** Wang, X., Delgado, H., Tak, H., Jung, J., Shim, H., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T. H., Evans, N., Lee, K. A., & Yamagishi, J. (2024). ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 1–8. **[10]** Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). *Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024*. **[11]** Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.4.08) [Computer software]. http://www.praat.org/

Preliminary Results

Experiment 1:

- Decision trees allow a visualization of the feature space and model decisions.
- Some features performed better in detecting certain spoof types than others.





Experiment 2:

Results (averaged over 3 subsets) revealed an interplay between features and ML algorithms.

- Tree-based ensemble models performed better on seen attacks.
- Nearest Centroid, QDA, Naïve Bayes performed better on unseen attacks.

Seen attacks			Unseen attacks		
Model	***Balanced Accuracy***	***F1 Score***	***Model***	***Balanced Accuracy***	***F1 Score***
Light GBM	0.69	0.71	Nearest Centroid	0.92	0.66
Random Forest	0.85	0.92	Quadratic Discriminant Analysis	0.49	0.66
SVC	0.56	0.49	Naïve Bayes (Bernoulli)	0.49	0.66
Extra Trees Classifier	0.56	0.49	Naïve Bayes (Gaussian)	0.49	0.66
Bagging Classifier	0.56	0.49	Light GBM	0.68	0.64

IAFPA
July 20 – 23 2025
The Hague, The Netherlands