Why Do You Say It Like That?
A Phoneme-Level Framework for Explainable Speech Deepfake Detection

Anna Taylor[1]*, Michele Panariello*[1]*, Massimiliano Todisco*[1]*, Chiara Galdi*[1]*, Nicholas Evans*[1]

[1]Digital Security, EURECOM, France

firstname.lastname@eurecom.fr

Abstract

As speech deepfake detection increasingly relies on foundation models and self-supervised speech representations, such as wav2vec 2.0 [1] and HuBERT [2], explaining *why* an utterance is classified as bona fide or deepfake remains an open challenge. In pursuit of more trustworthy and interpretable artificial intelligence, we focus on phonetic cues that deviate in synthetic speech, since phonemes and pauses provide a natural representation of speech structure that humans can understand. However, it is still unclear which phonetic features are exploited by these systems, whether certain phoneme classes are more informative than others for detecting synthetic speech, and whether the resulting evidence generalizes across speakers or instead depends on speaker-specific traits. We introduce a phoneme-level analysis framework that connects model decisions to interpretable phonetic units. By applying our explainability technique, we identify which phonemes, including pauses, contribute most strongly to deepfake detection within a particular utterance. Although further refinement is needed to disentangle the saliency of identified phonemes from causality and to account for global artifacts such as unnatural prosody, this research provides a novel approach to phonetically-driven explainability, while maintaining performance on par with similar approaches on the ASVspoof 5 dataset [3].

Our phoneme-level explanatory framework provides post-hoc interpretability for deepfake detection systems based on convolutional neural networks by linking model activations to linguistically meaningful units. As illustrated in Figure 1, our pipeline consists of three components: a front-end feature extractor, a back-end binary classifier, and a phoneme-level explanatory module. Frame-level features are extracted using a frozen WavLM Base+ model [4], although the proposed explainability method is compatible with other front-end encoders as well. These acoustic representations are then passed to a temporal convolutional classifier with masked temporal average pooling, which produces two logits corresponding to the bona fide and spoof classes. The classifier is trained on the ASVspoof 5 training set using a cross-entropy loss and evaluated in terms of pooled and per-attack Equal Error Rate (EER). For explainability, Gradient-weighted Class Activation Mapping (Grad-CAM) [5] is applied to the final convolutional layer to generate class-specific temporal activation maps that separately highlight the regions contributing most to bona fide and spoof predictions. These saliency maps are overlaid on a mel spectrogram aligned with the phonemes of the input utterance using the Whisper Turbo model [6] for ASR and the Bournemouth Forced Aligner English model [7]. As shown in Figure 2, computing activations separately for the bona fide and spoof logits enables direct comparison of model responses across phonemes and pauses, thereby facilitating analysis of how latent acoustic cues relate to higher-level linguistic structure.

Experiments on the ASVspoof 5 evaluation set yield a pooled EER in line with prior work on this architecture, yet reveal substantial variability across both speakers and attack conditions. While some cases achieve low EERs (e.g., below 4%), others exceed 20–40%, highlighting heterogeneous spoofing characteristics and pronounced speaker-dependent effects that motivate both attack- and speaker-specific analysis.

Phoneme-level attribution analysis reveals strong statistical differences in classifier attention across the ASVspoof 5 spoofing attacks. Kruskal-Wallis statistical tests [8] show that multiple phonemes (e.g., /l/, /s/, /ɪ/, /æ/, /ʊ/, /f/, /z/) exhibit significantly different importance distributions across attacks. Similarly, phonetic category-level analysis indicates that vowels, fricatives, nasals, and stops all vary significantly across attacks, with a strong association between attack type and dominant phonetic category (χ^2 test, $p < 10^{-9}$). Silence or pause regions also display strong attack-dependent behavior, with significant differences observed across numerous attacks, notably between text-to-speech and voice conversion-based attacks. Furthermore, there is a substantial spread in per-speaker vulnerability, with strong speaker effects on decision confidence and meaningful variation in where Grad-CAM activates, suggesting that model reliance is driven by phonetic content and how it is shaped by speaker identity.

It is important to note here that Grad-CAM provides saliency rather than causal attribution and depends on intermediate convolutional representations. As such, the resulting explanations may reflect correlated features, such as segment durations or pitch contours, and can be sensitive to model architecture [9], misleading in certain settings [10] and susceptible to adversarial manipulation [11]. Additionally, the reliance on ASR and forced alignment introduces potential transcription errors, particularly in degraded or highly synthetic speech, that could propagate through the pipeline. Nevertheless, aggregating attribution across phoneme-aligned segments in a large dataset reveals consistent, statistically significant attack- and speaker-dependent patterns in model behavior.

Overall, this work demonstrates that a phoneme-level framework can transform low-level saliency into interpretable insights, which enable systematic analysis of how deepfake detectors behave across speakers and spoofing conditions. Future work could investigate causal attribution methods, incorporate prosodic features such as duration, pitch, and intensity, and extend this framework to model-agnostic, multilingual, and cross-architecture comparisons.

Index Terms: speech deepfake detection, explainable AI, phoneme-level analysis, phonetic features.

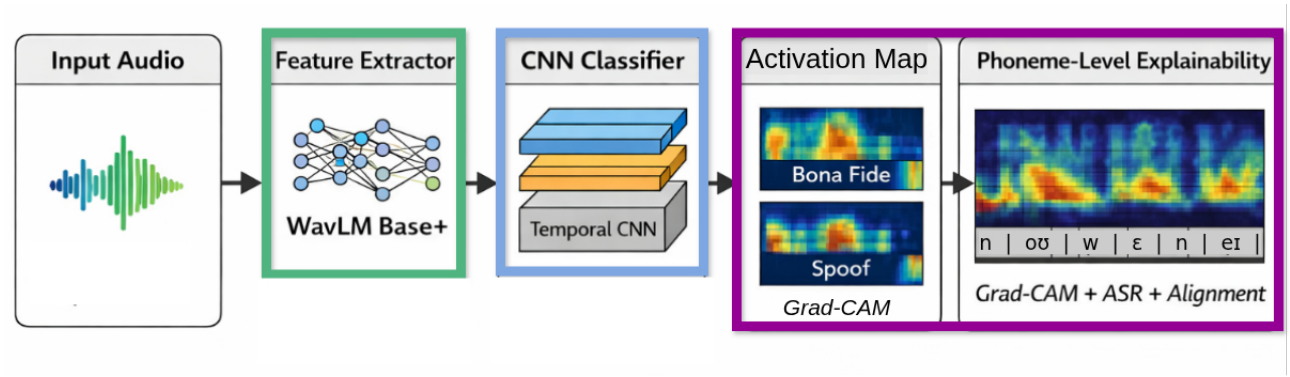

Figure 1: *Phoneme-level explainable speech deepfake detection pipeline: audio is encoded using WavLM Base+ (in green box) and classified by a temporal CNN into bona fide or spoof (in blue box). Grad-CAM produces class-specific saliency maps aligned to phoneme boundaries via ASR and forced alignment (in purple box), enabling analysis of model activations over phonetic units.*

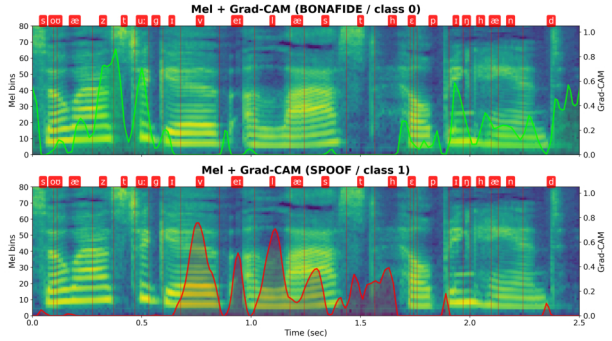

Figure 2: *Phoneme-aligned Grad-CAM attributions for bona fide (top) and spoof (bottom) overlaid on a mel spectrogram representing the utterance "so as to give a last helping hand" from a bona fide sample. Differences in activation highlight phoneme-level regions that contribute to model decisions.*

1. Acknowledgments

This work was supported by the COMPROMIS project (ANR22-PECY-0011) funded by a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program.

2. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, Oct. 2021.

[3] 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.

[4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:239885872

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[7] A. Rehman, J. Cai, J.-J. Zhang, and X. Yang, "BFA: Real-time Multilingual Text-to-speech Forced Alignment," 2025. [Online]. Available: https://arxiv.org/abs/2509.23147

[8] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. [Online]. Available: https://doi.org/10.1080/01621459.1952.10483441

[9] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[10] S. Jain and B. C. Wallace, "Attention is not Explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1357/

[11] T. Viering, Z. Wang, M. Loog, and E. Eisemann, "How to Manipulate CNNs to Make Them Lie: the GradCAM Case," 2019. [Online]. Available: https://arxiv.org/abs/1907.10901