

Why Do You Say It Like That?

A Phoneme-level Framework for Explainable Speech Deepfake Detection

Anna Taylor¹, Michele Panariello¹, Massimiliano Todisco¹, Chiara Galdi¹, Nicholas Evans¹, Driss Matrouf²

¹EURECOM, Sophia Antipolis, France

²Laboratoire Informatique d’Avignon, Avignon Université, France

¹firstname.lastname@eurecom.fr

²driss.matrouf@univ-avignon.fr

Abstract—As the accuracy of speech deepfake detection improves with the use of self-supervised representations such as wav2vec 2.0 and HuBERT, understanding why the speech is classified as bona fide or deepfake remains an open challenge. In pursuit of more trustworthy and interpretable AI, we introduce a phoneme-level analysis framework that connects model predictions to phonetic units. Our post-hoc explainability method is generally applicable to a variety of speech deepfake detection systems based on convolutional neural networks since it leverages Gradient-weighted Class Activation Mapping in conjunction with speech recognition to generate saliency maps aligned with phonemes and pauses. This pipeline reveals statistically significant attack- and speaker-dependent phonetic cues associated with spoofed speech in terms that humans can understand. Experiments using ASVspoof 5 show comparable detection performance to similar architectures while providing linguistic interpretations across speakers and spoofing conditions.

Index Terms—deepfake detection, speech processing, explainable artificial intelligence

I. INTRODUCTION

As progress in the field of generative speech technology has pushed the deceptive capabilities of voice conversion and text-to-speech to greater levels, it is consequently ever more important to distinguish between bona fide and spoofed speech. Modern speech deepfake detection systems manage to achieve strong performance by leveraging self-supervised (SSL) speech representations extracted using large foundation models such as wav2vec 2.0 [1], HuBERT [2], and WavLM [3]. Despite these advances, the predictions of such systems remain largely difficult to understand with human-like reasoning [4] and they lack human-perceptible cues that would provide natural explanations for bona fide or spoof classifications. Even though a deepfake detector may successfully classify an utterance, it is often unclear which characteristics of the speech signal contributed to that prediction.

This lack of interpretability presents a challenge for the trustworthiness and transparency of speech deepfake detection systems. In applications where security and fairness are

critical, understanding why a system reached a particular classification is crucial not only for faith in the decision itself, but also for future adjustment of the system. However, explanations based solely on low-level acoustic representations are often difficult to interpret, particularly for speech signals where the structure is more naturally described in linguistic terms, such as phonemes.

Phonemes provide a useful intermediate representation between raw acoustics and human interpretation of spoken language. In accordance with the International Phonetic Association and leading scholars, the definition of a phoneme is strictly limited to the assigned categorical representation of an uttered sound in a given language which does not incorporate its acoustic realization or varied pronunciation across dialects and spontaneous speech [5]. Accordingly, we use ‘phoneme’ only when referring to the language-specific categorical label and ‘phone’ when referring to measurable acoustic properties or realizations.

Prior research has shown that spoofed speech exhibits deviations in a range of phonetic and prosodic characteristics, including segment durations [6], non-speech regions [7], and timing structure [8]. Nevertheless, it remains unclear to what extent a modern deepfake detector relies on these cues and whether particular phonemes (e.g. /o/, /s/, etc.) or phonetic categories (e.g. vowels, fricatives, etc.) contribute more strongly than others to model predictions. Furthermore, the contributions of such patterns across different spoofing attacks or different speakers could be better elucidated.

We introduce a phoneme-level framework for explainable speech deepfake detection alongside analysis of trends exposed by application of our framework to spoof/deepfake utterances in the ASVspoof 5 database [9]. We combine a self-supervised front-end, a convolutional neural network (CNN) classifier, and a post-hoc explanatory module based on Gradient-weighted Class Activation Mapping (Grad-CAM) [10]. By aligning saliency maps with phoneme labels obtained through automatic speech recognition and forced alignment, model activations can be analyzed with respect to linguistically meaningful units. Through the aggregation of phoneme-level attributions across a large evaluation corpus, we investigate

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.

how detector behavior varies across spoofing attacks, speakers, and phonetic features and categories.

Experiments are conducted using the ASVspoof 5 corpus and a WavLM-based [3] detector that achieves a competitive pooled equal error rate for the evaluation set [9]. Statistical analysis reveals significant attack-dependent and speaker-dependent differences in phoneme-level activation patterns, with particularly strong effects observed for several vowels, fricatives, and silence or non-speech regions. These findings suggest that phoneme-level explanations can provide insight into the linguistic cues exploited by the analyzed deepfake detector.

The main contributions of this work are threefold:

- we introduce a phoneme-level explainability framework that links speech deepfake detector predictions to linguistically interpretable units;
- we provide a large-scale statistical analysis of phoneme-level attributions across spoofing attacks and speakers;
- we demonstrate that meaningful phonetic patterns can be extracted from a self-supervised speech deepfake detection system without sacrificing detection performance.

II. RELATED WORK

We review related work in three areas relevant to the proposed framework: speech deepfake detection, explainability methods for speech processing, and phonetic analyses of deepfake speech. Together, these areas motivate the need for explanations that connect detector behavior to linguistically meaningful speech units.

A. Speech Deepfake Detection

The danger posed by neural speech synthesis and voice conversion systems has led to growing interest in and reliance on speech deepfake detection. Early approaches utilized hand-crafted acoustic features designed to capture artifacts introduced by synthesis and conversion algorithms [11], [12]. More recently, deep learning systems based on CNNs and transformer architectures have become dominant [12], particularly when paired with self-supervised speech representations learned from large quantities of unlabeled speech data [9], [13]–[15].

Among these approaches, representations extracted from models such as wav2vec 2.0 [1], HuBERT [2], and WavLM [3] have demonstrated strong performance across a range of spoof/deepfake detection benchmarks [15], [16]. These representations capture rich acoustic and linguistic information and often outperform systems trained using hand-crafted features [9], [12]. While detection accuracy has improved substantially, understanding what information these systems exploit remains an open challenge. In many cases, modern detectors achieve strong performance but provide little insight into which aspects of speech distinguish bona fide from spoofed utterances.

B. Explainability in Speech Processing

Explainable artificial intelligence aims to make model predictions more transparent by identifying characteristics or regions of an input that contribute to the output, i.e. a bona fide vs spoof prediction. In computer vision, saliency-based techniques [17] such as Grad-CAM [10] have become widely used for interpreting CNNs [18]. Similar approaches have been applied to text [19] and speech processing tasks including speech recognition [20], speaker verification [21], emotion recognition [22], and deepfake detection [23].

Despite their popularity, saliency methods present several challenges [18]. Highlighted regions indicate *where* a model is sensitive but cannot directly explain *what linguistic evidence* supports a model’s prediction in meaningful terms that humans naturally use to interpret speech. Furthermore, Grad-CAM 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 [24], misleading in certain settings [25], and susceptible to adversarial manipulation [26]. While existing work focuses on visualizing important time-frequency regions, relatively little effort has been devoted to translating these activations into linguistic units that can be readily understood by human analysts [27], [28].

C. Phonetic Analysis of Deepfake Speech

Generative speech technologies have long been known to exhibit phonetic and prosodic deviations relative to natural speech. Prior studies have reported differences in segment duration, coarticulation, speaking rate, pause structure, and articulatory consistency [29]–[31]. Although many of these artifacts have become less pronounced as synthesis quality has improved [32], they continue to provide valuable information for both forensic analysis and automatic detection systems [31].

Several studies have investigated phoneme-level characteristics of spoofed speech, often focusing on acoustic measures such as phone duration or spectral properties [27], [33]–[36]. However, relatively little work has examined whether modern deepfake detectors rely on these phonetic cues to help distinguish bona fide from spoofed speech. Consequently, there remains a gap between phonetic analyses of spoofed speech and explainability methods for deepfake detection. With the framework proposed in this work, we seek to bridge this gap by linking detector activations to phoneme-aligned segments and analyze them at scale.

III. PHONEME-LEVEL EXPLAINABILITY FRAMEWORK

Our framework comprises three components: an SSL-based front-end, a back-end deepfake detection classifier, and a phoneme-level explanatory module. Detection components produce scores for both bona fide and spoof classes, while the explanatory module aligns class-specific attribution signals with phonemes, enabling the analysis of detector behavior in linguistically interpretable terms.

A. Deepfake Detection System

Input speech is first processed using an SSL-based front-end feature extractor, which produces frame-level representations encoding acoustic and linguistic information. These representations are then passed to a temporal CNN, which acts as the back-end classifier. The temporal frame-level outputs are aggregated using masked temporal average pooling, which averages the representations over speech frames while excluding padded frames introduced for batching. This produces a fixed-dimension, utterance-level representation. A final classification layer produces two logits corresponding to the bona fide and spoof classes.

While the detection architecture itself is not the primary contribution of this work, it provides a suitable deepfake detection system whose internal representations can be analyzed through the proposed explanatory framework. This allows us to investigate which characteristics of the speech signal contribute to the bona fide/spoof class scores and how these attribution patterns vary across speakers and spoofing attacks.

B. Grad-CAM-Based Attribution

To identify which components of an utterance provide support for each class hypothesis ('bona fide' or 'spoof'), Grad-CAM [10] is applied to the final convolutional layer of the classifier. Grad-CAM weights the activation maps of a convolutional layer by the gradients of a target output logit, thereby highlighting temporal regions of the input representation that are most associated with that target prediction. In this work, we compute separate Grad-CAM maps for the bona fide and spoof logits and refer to these class-specific attribution maps as *bona fide-CAM* and *spoof-CAM*, respectively.

The resulting attribution maps provide temporal evidence in support of each class hypothesis; both bona fide-CAM and spoof-CAM can exhibit non-zero attribution values for the same utterance, regardless of its ground-truth label. Rather than treating these maps as the final explanation, we use them as intermediate attribution signals that are subsequently mapped onto phoneme-labeled time spans.

C. Phoneme Alignment and Attribution Aggregation

To associate class-specific attribution scores with speech units, each utterance is first transcribed using Whisper [37] Turbo¹. The resulting transcript is then aligned to the acoustic signal using the Bournemouth Forced Aligner [38], [39]² English model, which provides phoneme-level labels and corresponding phone boundaries.

Figure 1 shows an example of the output bona fide-CAM and spoof-CAM aligned with phonemes. The figure illustrates how the two class-specific attribution signals can highlight different temporal regions of the same utterance and how these regions can be associated with phoneme-labeled segments.

Given a phone or non-speech segment spanning frames t_1 through t_2 , an importance score is computed by averaging

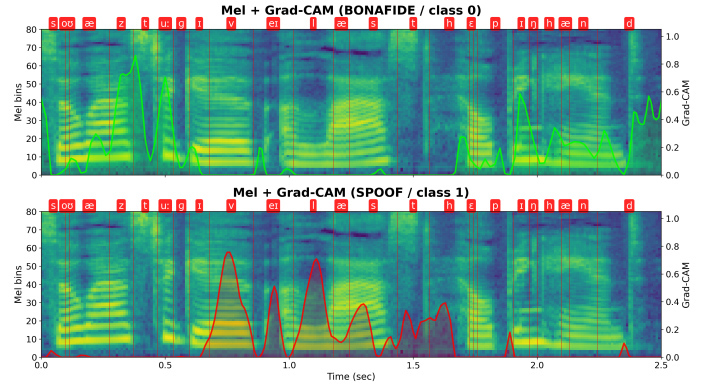


Fig. 1. Example of phoneme-aligned Grad-CAM attribution maps for the bona fide and spoof class hypotheses. The mel spectrogram is shown in the background, while the overlaid curves indicate the temporal Grad-CAM attribution scores for the bona fide logit (top) and spoof logit (bottom). Phoneme labels obtained from forced alignment are shown above each spectrogram. The two maps provide class-specific attribution signals and are therefore not mutually exclusive.

the corresponding Grad-CAM attribution values within the segment. This produces one bona fide-CAM and one spoof-CAM importance score for each phone or non-speech segment in the utterance. By aggregating these scores across the corpus, we can examine how support for bona fide or spoof predictions varies across phoneme identities, phonetic categories, speakers, and spoofing attacks.

D. Statistical Analysis

Statistical analyses are performed on the aggregated phoneme-level attribution scores, which correspond to the average bona fide-CAM or spoof-CAM value within the aligned segment. We then analyze how the resulting score distributions vary across spoofing attacks, speakers, and phonetic categories.

For each phoneme, Kruskal-Wallis tests [40] are used to assess whether the corresponding attribution scores differ significantly between spoofing attacks. This non-parametric test is used because the attribution scores are not assumed to follow a normal distribution. When a significant effect is observed, post-hoc pairwise comparisons are performed between attack pairs, with Benjamini-Hochberg correction applied to control for multiple comparisons [41].

In addition to statistical significance, we report epsilon-squared effect sizes, which range from 0 to 1 where $\epsilon^2 \geq 0.01$ is small in magnitude, $\epsilon^2 \geq 0.06$ is medium, and $\epsilon^2 \geq 0.14$ is large [42]. This is important because the large number of phoneme instances in the evaluation set can make even very small differences statistically significant. Effect sizes therefore help distinguish differences that are merely detectable due to sample size from differences that explain a meaningful proportion of the variation in attribution scores.

Additional analyses are performed at the phonetic-category level by grouping phonemes into vowels, fricatives, stops, nasals, affricates, approximants, and non-speech regions. For each utterance or attack condition, we identify which phonetic

¹<https://github.com/openai/whisper>

²<https://huggingface.co/Tabahi/CUPE-2i>

categories receive the largest attribution scores and use chi-square tests to examine whether the dominant categories are associated with the spoofing attack. This allows us to determine whether different attacks tend to concentrate class-specific support on different types of speech units.

IV. EXPERIMENTAL SETUP

Here, we describe the experimental protocol used to evaluate the proposed framework. We first outline our use of the ASVspoof 5 dataset and the partitions used for training and evaluation. We then detail our implementation, which is available as open-source for reproducibility under the same GPU environment³, followed by the metrics used to assess detection performance and phoneme-level explainability.

A. Dataset

Experiments are conducted using the ASVspoof 5 [9] corpus, which contains bona fide speech and spoofed speech generated using a diverse collection of text-to-speech (TTS) and voice conversion (VC) systems. The dataset includes substantial variability in speaker characteristics, recording conditions, and spoofing attacks, making it suitable for both detection evaluation and interpretability analysis.

The official training partition is used for model optimization, while the ASVspoof 5 evaluation set is used for inference and statistical analysis. The training partition contains utterances collected from 400 speakers, spoofed utterances generated using 8 different attack algorithms, and a total of 18,797 bona fide and 163,560 spoofed utterances. The speaker-disjoint evaluation set contains data collected from 737 previously unseen speakers, spoofed utterances generated using 16 different attack algorithms, and a total of 680,774 utterances, of which 138,688 are bona fide and 542,086 are spoofed.

B. Implementation Details

In our experimental implementation, the SSL-based front-end is instantiated using WavLM [3] Base⁺⁴, which is kept frozen during training. The temporal convolutional back-end consists of a single one-dimensional convolutional block with 512 channels, followed by masked temporal average pooling and a linear classification layer. The detector is trained using cross-entropy loss on the ASVspoof 5 training partition. Optimization is performed using the Adam optimizer [43] for 200 epochs with a cosine annealing learning-rate schedule and warmup. The trained model is then applied to the ASVspoof 5 evaluation set, where both detection scores and intermediate activations are extracted for the subsequent explainability analysis.

C. Evaluation Metrics

Detection performance is measured using the Equal Error Rate (EER). Results are reported both as pooled EER and on a per-attack basis in order to characterize variability across spoofing conditions.

For explainability analysis, we evaluate the phoneme-level attribution scores by testing whether they vary systematically across spoofing attacks, speakers, and phonetic categories. Statistical significance tests are used to determine whether observed differences in attribution scores are unlikely to arise by chance, while effect-size measures are used to quantify the magnitude of these differences. This allows us to identify which phonemes and phonetic categories receive consistently different levels of class-specific support across attacks and speakers.

V. RESULTS AND ANALYSIS

In this section, we report the detection performance of the proposed system and analyze class-specific phoneme-level attribution scores obtained from bona fide-CAM and spoof-CAM. We examine variation across spoofing attacks, temporal locations, non-speech regions, and speakers.

A. Detection Performance

Our detector achieves a pooled EER of 8.52% for the ASVspoof 5 evaluation set. Rather than proposing a new detection architecture and comparing with other systems, this work instead focuses on explaining model predictions through our phoneme-level interpretability framework. While pooled performance is competitive [9], substantial variability is observed across spoofing attacks. Most attacks yield EERs below 8%, with some EERs even below 4%; however, attacks A24 (in-house ASR-based VC [9]) and A28 (pre-trained YourTTS [44]) remain considerably more challenging, with EERs exceeding 20%. This variability suggests that **different generative speech systems introduce distinct artifacts** and motivates the attack specific phoneme-level analysis presented in the following sections.

B. Attack-Dependent Phoneme Importance

To investigate whether the detection of different spoofing attacks relies on distinct phonetic cues, we analyze frequency-normalized, phoneme-level spoof-CAM attribution scores across attacks. Frequency normalization was applied to remove the confounding effect of phoneme occurrence rates in the English language, which naturally follow the frequency distribution of function words in English regardless of whether the calculations come from spoken or written language [45]. Therefore, we ensure that attribution differences reflect the relative importance of a phoneme to the detector rather than its frequency in the corpus.

For each phoneme, we performed a Kruskal-Wallis [40] test across attack conditions, followed by Benjamini-Hochberg [46] correction for multiple comparisons. Effect sizes were quantified using epsilon-squared (ϵ^2) [42], which estimates the proportion of attribution variance explained by attack identity.

Of the top 25 phonemes with the highest spoof-CAM scores after frequency normalization (including non-speech represented by the SIL token), all exhibit statistically significant differences across attacks ($p_{BH} < 0.001$), indicating that

³<https://github.com/eurecom-asip/PhonemeXAI>

⁴<https://huggingface.co/microsoft/wavlm-base-plus>

TABLE I
PHONEMES EXHIBITING THE STRONGEST ATTACK-DEPENDENT VARIATION IN FREQUENCY-NORMALIZED GRAD-CAM ATTRIBUTION SCORES FOR BOTH SPOOF AND BONA FIDE PREDICTIONS, RANKED BY KRUSKAL-WALLIS EFFECT SIZE (ϵ^2).

Spoof-CAM		Bona fide-CAM	
Phoneme	ϵ^2	Phoneme	ϵ^2
/oʊ/	0.135	/ʌ/	0.113
/ɛ/	0.115	/ɛ/	0.110
/aɪ/	0.113	/ɪ/	0.108
/s/	0.101	/oʊ/	0.107
/f/	0.098	/i:/	0.102
/z/	0.092	/ɪ/	0.099
/et/	0.091	/h/	0.097
/i:/	0.088	/v/	0.097
/h/	0.086	/et/	0.096
/w/	0.072	/aɪ/	0.091

the importance assigned to a phoneme by the spoof detector depends on the spoof generation method. However, effect sizes varied substantially between individual phonemes.

The strongest attack-dependent effects were observed for /oʊ/ ($\epsilon^2 = 0.135$), /ɛ/ ($\epsilon^2 = 0.115$), /aɪ/ ($\epsilon^2 = 0.113$), /s/ ($\epsilon^2 = 0.101$), and /f/ ($\epsilon^2 = 0.099$) with the top 10 phonemes ranked by attack-dependent variation shown in Table I. For /oʊ/ as much as 13.5% of the variation in frequency-normalized importance is explained by the spoofing attack used to generate the speech. Rather than concentrating on a single universal artifact, **statistically significant attack-dependent variation is distributed across multiple phonemes and phonetic categories in the evaluated detector.**

Interestingly, the largest effects were concentrated among vowels such as /oʊ/, /ɛ/, and /aɪ/, as well as fricatives, particularly /s/, /f/, and /h/. Here the evidence suggests that attack-specific artifacts arise most strongly in acoustically complex speech regions, which are known to be challenging for TTS and VC systems [31].

Furthermore, analysis of bona fide-CAM activations revealed that many phonemes show substantial attack-dependent variation in their bona fide importance scores. In particular, /ʌ/ ($\epsilon^2 = 0.113$), /ɛ/ ($\epsilon^2 = 0.110$), /ɪ/ ($\epsilon^2 = 0.108$), and /oʊ/ ($\epsilon^2 = 0.107$) emerged as especially attack-dependent; more results are shown in Table I. Considering the prominence of vowels here, the detector is likely sensitive to speaker-dependent articulatory variability arising more strongly in vowels and lightly constricted sounds like approximants when attempting to identify natural speech. Moreover, the large attack effects indicate that different spoofing systems distort these phonetic cues in distinct ways, resulting in attack-specific attribution patterns.

C. Temporal Localization

Beyond identifying which phonemes contribute to detector predictions, we also investigated the exact location within a phone and within the entire utterance where the highest Grad-CAM activations occur for both spoof and bona fide predictions. For each activation peak, we computed the temporal distance to the nearest phone boundary designated by the

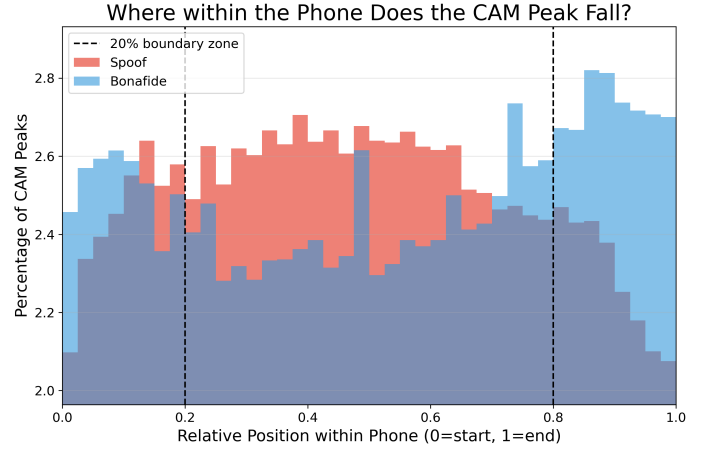


Fig. 2. Grad-CAM activation peaks differ in their relative location within the phone depending on the model’s prediction as spoof or bona fide, representing over 2% of the CAM peaks.

Bournemouth Forced Aligner and to the first and last 20% of the utterance duration, and we find significant differences between human speech and the various spoofing attacks only for phone boundary distance.

Across all attacks, bona fide activation peaks were consistently concentrated near phone boundaries, representing the first or last 20% of the duration, while the spoof activations concentrated within the phones, as seen in Figure 2. Median boundary distances ranged from approximately 10-15 ms, whereas median distances within phone interiors ranged from approximately 37-56 ms. This finding indicates that the evaluated detector relies on acoustic transitions between phonetic units for determining natural human speech and on steady-state phonetic realizations for detecting spoofed speech.

Attack-dependent differences were also substantial. Pairwise comparisons revealed significant differences for 124 of 136 attack pairs (16 spoofing attacks plus bona fide) after Benjamini-Hochberg correction ($p < 0.05$), indicating that **spoofing attacks differ not only in the phonemes that attract model attention but also in the temporal location of the salient evidence.**

D. Non-Speech and Pause Regions

Non-speech and pause regions are consistently found to be informative for spoof predictions, and even more so for bona fide. On average, between 7% and 25% of total Grad-CAM spoof and bona fide attribution mass was assigned to non-speech segments with significant differences across ASVspoof 5 attacks. Attack identity had a particularly strong effect on bona fide-CAM non-speech attribution ($\epsilon^2 = 0.102$), substantially exceeding the corresponding spoof-CAM effect ($\epsilon^2 = 0.043$). This is in line with other findings [47]–[49] that have focused specifically on non-speech regions to detect natural human speech phonation and pause patterns which generative speech technologies struggle to accurately replicate.

The most extreme differences can be seen between attacks A20 (in-house unit-select TTS [9]) and Malafide adversarial

attack [50]) and A21 (ToucanTTS [51] and BigVGAN [52]). Despite both systems using TTS with similar detection performance (5.5% EER vs 6.5%), the former system has approximately 25% of the bona fide attribution mass and 21% of the spoof-CAM concentrated in non-speech regions, whereas, for the latter, non-speech content represents only 7% of bona fide attributions and 8% of spoof. This contrast suggests that **our deepfake detector unevenly exploits attack-specific characteristics that may vary widely** such as pause duration, pause placement, turn-taking timing, voicing onset/offset behavior, and transitions between speech and non-speech segments.

E. Speaker-Dependent Effects

In addition to attack-dependent variation, substantial speaker-dependent effects are also observed. Detector behavior varies considerably at the speaker level, both in terms of decision confidence and the phonetic evidence used to support spoof predictions. Figure 3 illustrates the highest-ranking discriminative phonemes for a subset of three speakers, revealing that the phonemes receiving the greatest Grad-CAM attribution differ markedly across individuals. For example, /l/, /a:/, and /b/ are among the most discriminative phonemes for speaker E_3699, whereas /oʊ/, /ʃ/, and /z:/ dominate for speaker E_1389, and /f/, /aʊ/, and /ɑ:/ for speaker E_4315. These differences indicate that **the analyzed detector does not rely on a fixed set of universally informative phonemes but instead adapts to speaker-specific acoustic characteristics that reveal substantial variation in per-speaker vulnerability**.

Speaker characteristics also influence broader phonetic attribution patterns. Significant gender-dependent differences in spoof probability were observed for 15 of the 16 spoofing attacks after Benjamini-Hochberg correction, indicating that the detector’s confidence depends on speaker gender for nearly all synthesis conditions except MaryTTS [53]. Likewise, the proportion of spoof-CAM assigned to voiced phones differed significantly between genders for 12 attacks, while non-speech and pause regions exhibited significant gender effects for 13 attacks. Bona fide speech also showed significant gender effects for both voiced and non-speech attribution distributions. Together, these findings suggest that detector predictions are affected not only by synthesis artifacts themselves but also by how those artifacts are shaped by speaker identity. Consequently, **attribution patterns vary jointly by spoofing attack and speaker characteristics, with the phonetic evidence for a given attack varying systematically across speakers**, although our speaker analysis reflects the evaluation corpus rather than a balanced speaker population.

VI. DISCUSSION

Our analysis reveals significant variation in class-specific attribution patterns across spoofing attacks, speakers, and phonetic information. Vowels, fricatives, approximants, non-speech regions, voicing, and speaker gender emerge as particularly informative, while the relative importance of individual

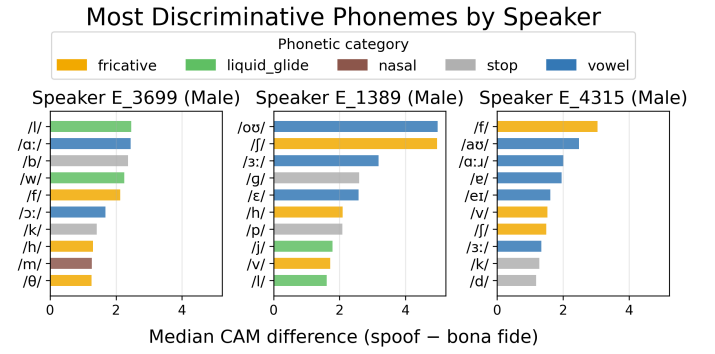


Fig. 3. Example of the most discriminative phonemes for three representative speakers. Bars show the median difference between spoof-CAM and bona fide-CAM attribution scores for each phoneme, grouped by phonetic category. The relative importance of phonemes differs substantially across speakers, indicating that no single phoneme consistently dominates across all speakers.

phonemes and the temporal location within phones varies substantially across attacks and speakers.

A. Limitations and Future Work

The proposed framework extends conventional saliency analysis by transforming Grad-CAM activations into statistically-validated, phoneme-level interpretations, providing meaningful insight into the factors influencing the analyzed detector’s predictions. Notably, Grad-CAM itself does not establish direct causal attribution, and it is possible that highlighted phonemes may reflect correlated acoustic properties. In addition, explanations are conditioned on the underlying network architecture and may differ across alternative detector designs.

The framework also depends on automatic transcription and forced alignment. Errors introduced during either stage may affect phone boundaries and attribution estimates, particularly for highly degraded or strongly synthetic speech.

Future work will investigate alternative attribution techniques, incorporate prosodic measures, and extend the analysis to multilingual settings, cross-architecture validation, human evaluation, and phoneme-ablation tests. An important direction will be determining whether the phonemes identified as salient correspond to measurable acoustic deviations in deepfake speech, thereby strengthening the connection between model explanations and phonetic phenomena.

VII. CONCLUSIONS

After applying our framework to ASVspoof 5 data, the evidence shows significant speaker-dependent effects as well as attack-dependent patterns across both phonetic features (voicing, pauses, etc.) and phoneme classes (vowels, fricatives, etc.). These findings reinforce the claim that the evaluated detector can be systematically analyzed in linguistically interpretable terms without compromising performance. Overall, this work demonstrates that a phoneme-level explainability framework can transform low-level saliency into interpretable insights for systematic analysis of how a deepfake detector behaves across speakers and spoofing conditions.

VIII. AI USE DISCLOSURE

During this research, Generative AI tools (OpenAI ChatGPT and Anthropic Claude) were used to assist with manuscript formatting and editing as well as development of analysis code. The authors reviewed, revised, and verified all generated content and accept full responsibility for the accuracy and originality of the work.

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.
- [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] 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>
- [4] Z. C. Lipton, “The mythos of model interpretability,” *Commun. ACM*, vol. 61, no. 10, p. 36–43, Sep. 2018. [Online]. Available: <https://doi.org/10.1145/3233231>
- [5] R. K. Moore and L. Skidmore, “On the use/misuse of the term ‘phoneme’,” *CoRR*, vol. abs/1907.11640, 2019. [Online]. Available: <http://arxiv.org/abs/1907.11640>
- [6] K. Fujita, A. Ando, and Y. Ijima, “Phoneme duration modeling using speech rhythm-based speaker embeddings for multi-speaker speech synthesis,” in *Proceedings of Interspeech 2021*. ISCA, 2021, pp. 3141–3145.
- [7] Y. Zhang, W. Wang, and P. Zhang, “The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System,” in *Interspeech 2021*, 2021, pp. 4279–4283.
- [8] Y. Zhu, S. Powar, and T. H. Falk, “Characterizing the temporal dynamics of universal speech representations for generalizable deepfake detection,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.08099>
- [9] 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.
- [10] 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>
- [11] M. Sahidullah, T. Kinnunen, and C. Hanilçi, “A comparison of features for synthetic speech detection,” in *Interspeech 2015*, 2015, pp. 2087–2091.
- [12] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, “Audio deepfake detection: What has been achieved and what lies ahead,” *Sensors*, vol. 25, no. 7, p. 1989, 2025. [Online]. Available: <https://doi.org/10.3390/s25071989>
- [13] X. Wang and J. Yamagishi, “Investigating self-supervised front ends for speech spoofing countermeasures,” in *The Speaker and Language Recognition Workshop (Odyssey 2022)*. ISCA, 2022.
- [14] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, “Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation,” in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.
- [15] K. Schäfer and M. Neu, “Real-world audio deepfake detection using ssl-based speech models and diverse training data,” in *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2025, pp. 1465–1469.
- [16] S. Dowerah, A. Kulkarni, A. Kulkarni, H. M. Tran, J. Kalda, A. Fedorchenko, B. Fauve, D. Lolive, T. Alumäe, and M. Magimai-Doss, “Speech df arena: A leaderboard for speech deepfake detection models,” *IEEE Open Journal of Signal Processing*, vol. 7, pp. 73–81, 2026.
- [17] V. Petsiuk, A. Das, and K. Saenko, “RISE: randomized input sampling for explanation of black-box models,” in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 151. [Online]. Available: <http://bmvc2018.org/contents/papers/1064.pdf>
- [18] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, “Explainable artificial intelligence: A comprehensive review,” *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3503–3568, 2022.
- [19] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘‘why should i trust you?’’: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: <https://doi.org/10.1145/2939672.2939778>
- [20] X. Ian Wu, P. Bell, and A. Rajan, “Explanations for automatic speech recognition,” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257232550>
- [21] X. Wu, C. Luu, P. Bell, and A. Rajan, “Explainable attribute-based speaker verification,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.19796>
- [22] H. M. Jayasinghe, K. W. Wong, and A. Nugaliyadde, “A systematic review of interpretability and explainability for speech emotion features in automatic speech emotion recognition,” *Pattern Recognition*, vol. 171, p. 112122, 2026.
- [23] W. Ge, J. Patino, M. Todisco, and N. Evans, “Explaining deep learning models for spoofing and deepfake detection with shapley additive explanations,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6387–6391.
- [24] 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.
- [25] 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/>
- [26] T. Vierung, Z. Wang, M. Loog, and E. Eiseemann, “How to Manipulate CNNs to Make Them Lie: the GradCAM Case,” 2019. [Online]. Available: <https://arxiv.org/abs/1907.10901>
- [27] M. Chhibber, J. Mishra, and T. H. Kinnunen, “Phonetically explainable speech deepfake detection,” 2026. [Online]. Available: <https://arxiv.org/abs/2606.15454>
- [28] T. Liu, L. Zhang, R. K. Das, Y. Ma, R. Tao, and H. Li, “How Do Neural Spoofing Countermeasures Detect Partially Spoofed Audio?” in *Interspeech 2024*, 2024, pp. 1105–1109.
- [29] S. A. Duffy and D. B. Pisoni, “Comprehension of synthetic speech produced by rule: A review and theoretical interpretation,” *Language and Speech*, vol. 35, no. 4, pp. 351–389, 1992.
- [30] D. H. Klatt, “Review of text-to-speech conversion for english,” *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.
- [31] T. Yang, C. Sun, S. Lyu, and P. Rose, “Forensic deepfake audio detection using segmental speech features,” *Forensic Science International*, vol. 379, p. 112768, 2026.
- [32] D. Sityaev, K. Knill, and T. Burrows, “Comparison of the ITU-t p.85 standard to other methods for the evaluation of text-to-speech systems,” in *Interspeech 2006*, 2006, pp. paper 1233–Tue2WeO.3.
- [33] “Concatenated phoneme models for text-variable speaker recognition,” in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1993, pp. 391–394.
- [34] A. Leemann, M.-J. Kolly, and V. Dellwo, “Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison,” *Forensic Science International*, vol. 238, pp. 59–67, 2014. [Online]. Available: <https://doi.org/10.1016/j.forsciint.2014.02.019>
- [35] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, and E. Ambikairajah, “Phoneme specific modelling and scoring techniques for anti spoofing system,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6106–6110.

- [36] S. Gupta, M. Ravanelli, P. Germain, and C. Subakan, "Phoneme Discretized Saliency Maps for Explainable Detection of AI-Generated Voice," in *Interspeech 2024*, 2024, pp. 3295–3299.
- [37] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.
- [38] 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>
- [39] A. Rehman, J.-J. Zhang, and X. Yang, "Cupe: Contextless universal phoneme encoder for language-agnostic speech processing," in *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP 2025)*, ICNLSP. International Conference on Natural Language and Speech Processing, 2025.
- [40] 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>
- [41] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.
- [42] F. Fiel Peres, "Effect sizes for nonparametric tests," *Biochemia Medica*, vol. 36, no. 1, p. 010101, 2026. [Online]. Available: <https://doi.org/10.11613/BM.2026.010101>
- [43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015, poster. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [44] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International conference on machine learning*. PMLR, 2022, pp. 2709–2720.
- [45] J. V. Tobias, "Relative occurrence of phonemes in american english," *The Journal of the Acoustical Society of America*, vol. 31, no. 5, pp. 631–631, 1959.
- [46] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 01 1995. [Online]. Available: <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- [47] N. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is silver, silence is golden: What do asvspoof-trained models really learn?" *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.
- [48] Y. Zhang, Z. Li, J. Lu, H. Hua, W. Wang, and P. Zhang, "The impact of silence on speech anti-spoofing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3374–3389, 2023.
- [49] B. Chettri and B. L. Sturm, "A deeper look at gaussian mixture model based anti-spoofing systems," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5159–5163.
- [50] M. Panariello, W. Ge, H. Tak, M. Todisco, and N. Evans, "Malafide: a novel adversarial convolutive noise attack against deepfake and spoofing detection systems," in *Interspeech 2023*, 2023, pp. 2868–2872.
- [51] F. Lux, J. Koch, and N. T. Vu, "Low-resource multilingual and zero-shot multispeaker tts," in *Proceedings of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, 2022, pp. 741–751.
- [52] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigV-GAN: A universal neural vocoder with large-scale training," in *International Conference on Learning Representations (ICLR)*, 2023.
- [53] I. Steiner and S. Le Maguer, "Creating new language and voice components for the updated MaryTTS text-to-speech synthesis platform," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: <https://aclanthology.org/L18-1501/>