Interpreting SSL Representations for Spoof Detection: a WavLM Study

Mohamed Mallat
EURECOM
Sophia Antipolis, France
Mohamed.Mallat@eurecom.fr

Michele Panariello
EURECOM
Sophia Antipolis, France
Michele.Panariello@eurecom.fr

Massimiliano Todisco
EURECOM
Sophia Antipolis, France
Massimiliano.Todisco@eurecom.fr

Nicholas Evans
EURECOM
Sophia Antipolis, France
Nicholas.Evans@eurecom.fr

Anthony Larcher
Université du Mans
Le Mans, France
Anthony.Larcher@univ-lemans.fr

ABSTRACT

Self-supervised learning (SSL) front-ends such as WavLM [1] have become central to modern spoofing countermeasures, achieving strong performance on benchmarks such as ASVspoof 5 [2]. However, these models remain largely opaque: it is still unclear which layers encode spoof-relevant information, how layer behaviour varies across attack types, and whether learned representations can be linked to interpretable acoustic properties. Recent work has also shown that spoofing countermeasures can exploit learning shortcuts [3], further motivating analyses that ground model behaviour in meaningful acoustic properties rather than spurious cues.

We present a layer-wise interpretability study of WavLM Base on the ASVspoof 5 database using the following analyses, illustrated in Figure 1. All representations are extracted from the frozen model without fine-tuning. The frame-level outputs of each layer are summarised by mean and standard deviation pooling into a 1536-dimensional utterance vector.

Per-layer linear probing. An independent linear classifier is trained on each of the 13 layer outputs (L0–L12) using the train split only, with model selection on dev EER and final reporting on the eval split. This provides a transparent measure of linearly accessible discriminative information at each depth. Performance improves consistently with depth up to L11, which achieves a dev EER of 3.33% and an eval EER of 6.07%, before degrading slightly at L12. Per-attack analysis reveals strong layer–attack interactions: some TTS attacks become detectable from mid-layers onward, while others such as ToucanTTS and YourTTS show the opposite trend, with EERs increasing at late layers. Attacks augmented with Malafide [4] and Malacopula [5] adversarial filtering yield consistently high EERs at every layer. This is consistent with the hypothesis that adversarial optimisation suppresses detectable artifacts independently of representation depth, though our setup cannot fully separate artifact suppression from cross-detector transfer of adversarial noise; disentangling the two is left to future work.

CCA-based acoustic grounding. To understand *why* certain layers perform better, we apply Projection-Weighted Canonical Correlation Analysis (PWCCA) [6] between the representations of each layer and 17 handcrafted acoustic features spanning four families: voice quality (jitter, shimmer, HNR), prosodic (F0, energy), spectral (MFCC, centroid, flux), and harmonic/formant features capturing harmonic structure and vocal-tract resonances (harmonic ratio, F1, F2, F3). Two distinct patterns emerge, as shown in Figure 2. Most features peak in alignment with mid-layers (around L7) and then decay, coinciding with the EER improvements observed from early to mid-layers. In contrast, voice-quality features jitter, shimmer and HNR, which vocoders are known to distort, show monotonically increasing correlation all the way to L12. This explains the further EER gains at late layers: they are driven by increasing sensitivity to fine-grained vocal naturalness cues. The slight EER degradation at L12 reflects a loss of spectral information that offsets continued voice-quality gains. t-SNE visualisations corroborate this picture: representations transition from overlapping clusters at L1, to discrete attack-specific groupings at L6, to a continuous horseshoe-shaped manifold at L11 consistent with a shift from categorical spectral encoding to a continuous gradient of vocal naturalness [7].

Layer subset search. The per-layer variation raises a practical question: do we need all 13 layers? We perform an exhaustive search over all $\sum_{k=1}^{13} \binom{13}{k} = 8191$ non-empty subsets. For each, the corresponding representations are concatenated and a linear probe is trained, with selection on dev EER and reporting on eval. The best subset overall (L7–L9–L10–L11–L12) consists of 5 layers and achieves 6.1% eval EER, outperforming full 13-layer pooling (8.08%); the next-best subsets are also 5-layer, indicating that five was the empirically optimal number of layers rather than a fixed choice. All top subsets concentrate in the L6–L12 range, consistent with the CCA findings: early layers encode low-level spectral information with weaker discriminative relevance and do not contribute complementary information in the linear pooling setting.

ACKNOWLEDGMENT

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.

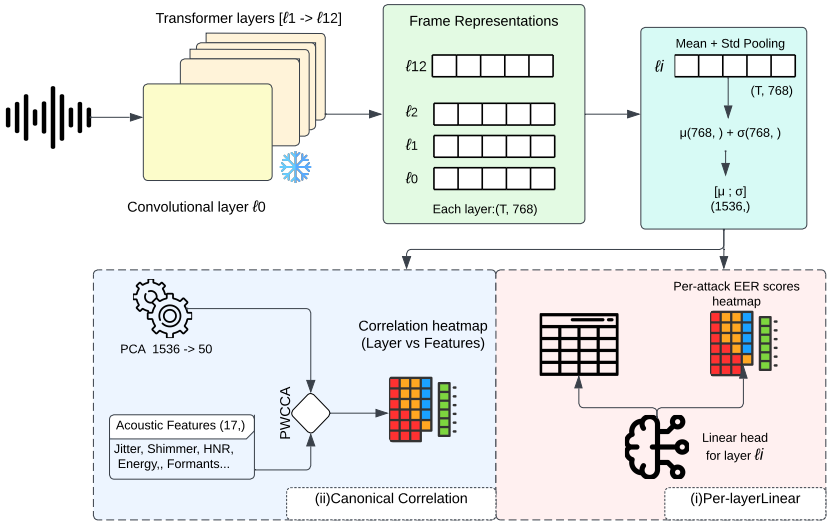

Fig. 1. Overview of the analysis pipeline. WavLM layer representations are summarised by mean+std pooling. Two analyses are performed: per-layer linear probing (right) and PWCCA-based acoustic grounding (left).

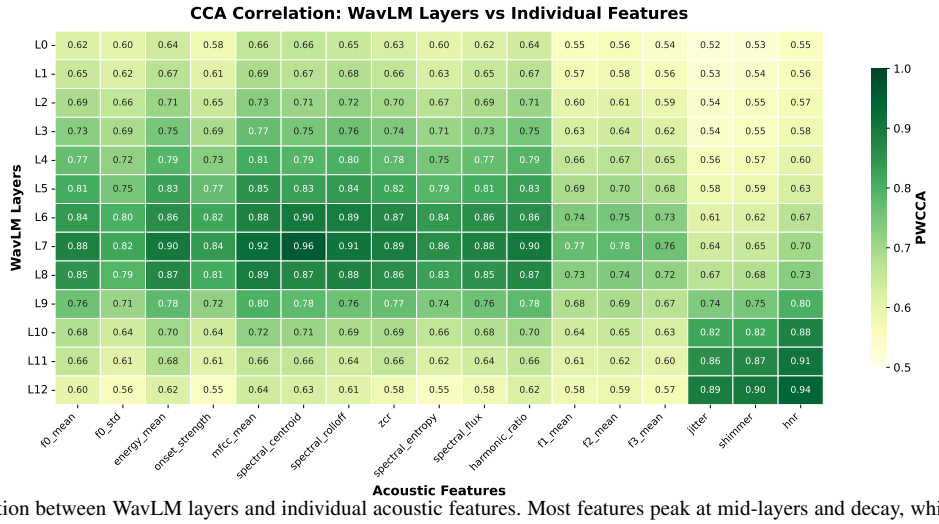

Fig. 2. PWCCA correlation between WavLM layers and individual acoustic features. Most features peak at mid-layers and decay, while voice quality measures (jitter, shimmer, HNR) increase monotonically with depth.

REFERENCES

[1] S. Chen, C. Yu, Y. Wu *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[2] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. 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 *Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 2024, pp. 1–8.

[3] N. Müller, P. Czempin, F. Dieckmann, R. Canals, K. Böttinger, and J. Williams, "Speech is silver, silence is golden: What do ASVspoof-trained models really learn?" in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 55–60.

[4] 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 *Proc. Interspeech*, 2023, pp. 2868–2872.

[5] M. Todisco, M. Panariello, W. Ge, H. Tak, and N. Evans, "Malacopula: Adversarial automatic speaker verification attacks using a neural-based generalised Hammerstein model," in *Proc. ASVspoof 2024 Workshop*, 2024.

[6] A. Morcos, M. Raghu, and S. Bengio, "Insights on representational similarity in neural networks with canonical correlation," in *Proc. NeurIPS*, 2018.

[7] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. IEEE ASRU*, 2021.