What Frozen Speech Embeddings Reveal About Audio Deepfake Detection Difficulty

Yashas Harshad Patil **Krish Rawat** **Anup Shanbhag**
Washington High School James Logan High School Dublin High School
Fremont, California, USA Union City, California, USA Dublin, California, USA
yashaspatil@gmail.com krishrawat0222@gmail.com
anupshanbhag4@gmail.com

Abstract

Audio deepfake detectors perform unevenly across synthesis systems. Can frozen speech representations expose structure behind this variation? We construct a simple direction between genuine and synthetic WavLM embeddings and characterize each system's position and spread along it. Across 61 MLAAD text-to-speech system–language groups, projected spread is strongly associated with mean detector difficulty across three runs ($\rho = 0.60$). It is the only feature surviving correction over 83 recorded tests, and leave-one-system-out ridge prediction gives $R^2 = 0.277$. The genuine–synthetic directions also differ markedly across corpora, revealing domain-dependent representation geometry. As an auxiliary utterance score, the centroid projection reduces MLAAD detector equal error rate from 27.2% to 16.3%, improving on both component scores. Supervised probe comparisons further identify where adaptation gains originate. Together, these results establish a compact representation-space diagnostic of attack-specific detector behavior; the informative statistic and benefits of score adaptation depend on the corpus and detector.

1 Introduction

An audio deepfake detector can achieve a useful average error rate while missing speech from particular synthesis systems. It can also lose accuracy when recordings come from a different dataset (Müller et al., 2022). Understanding this uneven performance is an evaluation problem in its own right: what structure in speech representations helps describe why an existing detector succeeds on some synthesis systems and struggles with others?

We study this question using frozen WavLM representations (Chen et al., 2022). We average frame representations into one vector per utterance and use labeled speech to construct a direction from the genuine centroid to the synthetic centroid. Each utterance then has an inspectable scalar projection. The construction is deliberately simple: it lets us examine detector-relevant structure already present in the representation without modifying the detector architecture or fitting another classifier.

Our central finding concerns the spread of these projections within each synthesis system. Across 61 MLAAD text-to-speech system–language groups (Müller et al., 2024), wider spread is associated with greater WavLM-GAT difficulty, measured by one minus the area under the receiver operating characteristic curve (AUC). The rank correlation is 0.60, and leave-one-system-out ridge prediction gives $R^2 = 0.277$. Spread is the only feature surviving correction across the recorded 83-test screen. A one-dimensional geometric summary thus captures meaningful variation in the behavior of a much richer detector.

Position and spread offer complementary views of that structure. A wide synthetic distribution can place more utterances near genuine speech, while a narrow distribution can also be difficult when centered close to genuine speech. We make this distinction explicit in the method and schematic. Across corpora, the genuine–synthetic direction changes substantially, and different geometric summaries become informative. This provides a representation-space perspective on domain-specific detector behavior.

We also examine the direction as an auxiliary score for individual utterances. On MLAAD, centroid fusion improves on both the detector and the projection alone. Comparisons with supervised probes show which gains come from a simple centroid score and which come from fitting a stronger target-domain classifier. These are secondary applications of the geometric analysis; the primary contribution is understanding variation across synthesis systems.

Proceedings of the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio 2026), pages 1–10
November 10, 2026 ©2026 Association for Computational Linguistics

Our contributions are:

- **Geometric diagnostic.** We link within-system spread along a labeled genuine–synthetic direction to detector-specific synthesis-system difficulty.
- **Statistical evidence.** We characterize the MLAAD relationship across 61 systems and three detector runs using screen-wide correction, resampling, permutation analysis, and leave-one-system-out prediction of the selected feature.
- **Domain geometry.** We document substantial cross-corpus changes in direction, alongside stable within-corpus estimates.
- **Score augmentation and decomposition.** We evaluate when geometric scores complement existing detectors and separate centroid-based gains from supervised probe adaptation.

2 Background and evaluation setting

Self-supervised speech representations support spoofing detection through learned downstream classifiers (Tak et al., 2022; Wang and Yamagishi, 2022). Our primary detector, denoted WavLM-GAT, follows Zhang et al. (2025): WavLM features feed adaptive phoneme pooling, graph attention, a bidirectional LSTM, and a classification head. The WavLM-base encoder has approximately 95 million parameters and is frozen; the downstream network has approximately 13 million trainable parameters. AASIST (Jung et al., 2022) provides a second detector architecture. Its spectro-temporal network does not use WavLM, although the geometric scores added to it in our analysis still do.

Difference-of-means directions summarize a labeled distinction in an embedding space. Related constructions have been used to analyze word embeddings and language models (Bolukbasi et al., 2016; Arditi et al., 2024). Whereas SSL-based anti-spoofing methods use representations as inputs to learned detectors (Tak et al., 2022; Wang and Yamagishi, 2022), we analyze the frozen representation itself and ask whether a simple labeled direction captures variation in detector difficulty across synthesis systems. Unlike detector logits or trained classifier probes, the direction is estimated directly from class centroids. Unlike generic embedding distances, it separates position and spread along a specified contrast. Its simplicity makes the resulting diagnostic easy to inspect.

The primary analysis uses 61 MLAAD text-to-speech (TTS) system–language groups with at least eight utterances per group, abbreviated as systems below. The cached test subset contains 908 genuine and 938 synthetic clips across 63 such groups; two groups with fewer than eight clips are excluded from the system-level analysis. Additional analyses use ASVspoof 2019 logical access (LA) (Wang et al., 2020), ASVspoof 2021 (Yamagishi et al., 2021), and In-the-Wild (ITW) (Müller et al., 2022). ASVspoof groups utterances by attack identifier, while the ITW analysis groups them by speaker. The ITW results therefore describe speaker-level variation. Appendix A lists the analyzed subsets.

For detector score $d(x)$ oriented so that larger values indicate synthetic speech, system s has difficulty

$$h_s = 1 - \mathrm{AUC}\big(d(B), d(S_s)\big), \tag{1}$$

where B is the shared genuine-speech reference pool and S_s contains synthetic utterances from system s. Perfect ranking gives $h_s = 0$, chance ranking gives $h_s = 0.5$, and reversed rankings can give larger values. This is a property of a detector–dataset pair, not an intrinsic measure of synthesis quality. It does not require choosing a detection threshold.

Three saved WavLM-GAT checkpoints produce similar MLAAD difficulty rankings: pairwise Spearman correlations range from 0.89 to 0.97. The primary regression target is the mean of their three difficulty values for each system. This agreement supports studying a shared pattern of system-level difficulty across these detector runs.

3 Measuring geometry and difficulty

3.1 A labeled direction and scalar projections

For each 16-kHz, three-second clip, let $e_i \in \mathbb{R}^{768}$ be the mean of the frozen WavLM layer-12 frame embeddings. Layers 0, 9, and 12 were examined; layer 12 gave the strongest axis-based result. We use layer 12 following this exploratory comparison.

Let $\mu_{B,j}$ and $a_{B,j}$ be the mean and population standard deviation of coordinate j in the genuine-speech reference data. We standardize coordinates

as

$$\widetilde{e}_{ij} = \frac{e_{ij} - \mu_{B,j}}{a_{B,j} + \epsilon}. \tag{2}$$

with $\epsilon = 10^{-9}$. For a held-out synthesis system s, let S_{-s} contain the synthetic reference utterances from the other systems. The centroid difference and its unit direction are

$$v_{-s} = \text{mean}_{i \in S_{-s}} \widetilde{e}_i - \text{mean}_{i \in B} \widetilde{e}_i, \tag{3}$$

$$u_{-s} = v_{-s} / \|v_{-s}\|_2. \tag{4}$$

The scalar projection of utterance i from system s is $p_i = u_{-s}^{\mathsf{T}} \widetilde{e}_i$. Thus projection means a dot product with a unit vector. Its sign is fixed from genuine toward synthetic speech. The direction uses genuine/synthetic class labels and does not use detector difficulty values h_s.

For the n_s synthetic utterances in system s, define their mean projection and spread as

$$m_{\parallel}(s) = \frac{1}{n_s} \sum_{i \in S_s} p_i, \tag{5}$$

$$\sigma_{\parallel}(s) = \sqrt{\frac{1}{n_s} \sum_{i \in S_s} \big(p_i - m_{\parallel}(s)\big)^2}. \tag{6}$$

The spread is calculated within one synthetic system. It neither pools genuine and synthetic utterances nor combines their standard deviations. A large separation between the class centroids therefore does not itself make $\sigma_{\parallel}$ large. Equation 6 follows the population-SD convention of the audited headline result. The original exploratory table used sample SD; Appendix A records this distinction. For exploratory position comparisons, the saved statistic is the median projection, rather than $m_{\parallel}$.

3.2 Why both position and spread matter

Figure 1 distinguishes location from spread. Holding a synthetic mean on the synthetic side of the genuine distribution, increasing its spread can place more utterances near genuine speech. Conversely, a tightly clustered system can still be difficult if its mean lies inside the genuine distribution. A distant mean with a small spread can be easy to separate. For a distant mean and a large spread, the answer depends on their relative magnitudes and the distribution shapes.

Projected overlap supplies intuition for the diagnostic. WavLM-GAT and AASIST use richer decision functions, so the empirical question is whether spread along this simple direction covaries with their measured difficulty across systems.

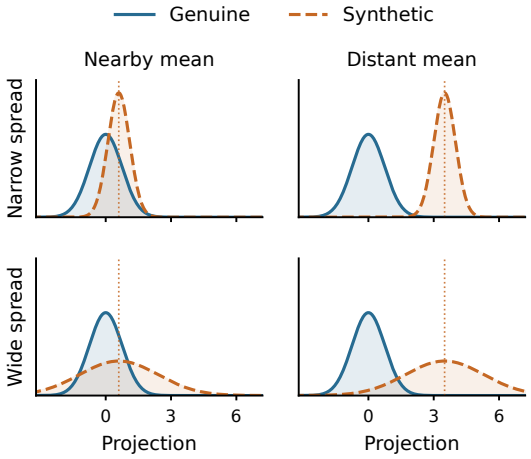


Figure 1: Position and spread both affect projected overlap. Each panel shows illustrative densities of genuine speech (blue) and one synthetic system (orange); these are schematics, not observations. Columns change the synthetic mean and rows change its spread. The same spread can accompany very different overlap, illustrating why position and spread describe distinct properties.

3.3 Association and held-out prediction

We distinguish two uses of leaving a system out. For geometric measurement, its utterances are excluded when constructing its reference direction. For predicting difficulty, ridge regression with regularization $\alpha = 1$ is fitted to the other systems' measured pairs $(\sigma_{\parallel}, h)$ and predicts the held-out system. The scalar predictor is standardized across the 61 systems before these regression folds. The reported prediction statistic is

$$R^2 = 1 - \frac{\sum_s (h_s - \widehat{h}_s)^2}{\sum_s (h_s - \bar{h})^2}. \tag{7}$$

A negative value indicates worse prediction than the overall-mean reference in this definition. The regression evaluates the selected feature using a fixed table of system measurements; feature discovery preceded this analysis. We interpret the result as within-MLAAD evidence for that feature. Appendix A specifies the feature-table construction and scaling.

Spearman's ρ measures rank association across systems. The recorded screen contains 83 predictor–difficulty tests. We report its global

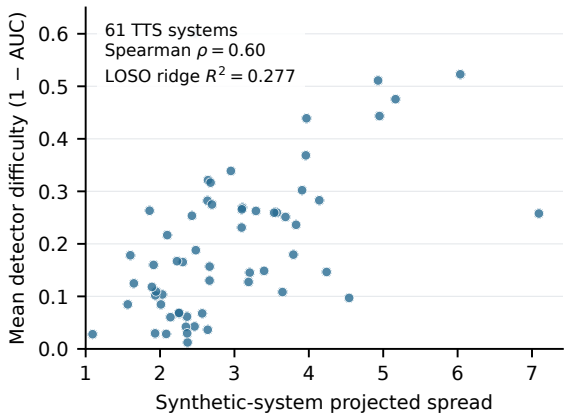


Figure 2: MLAAD system-level geometry and detector difficulty. Each point is one cached synthesis system–language group; difficulty averages three WavLM-GAT runs. Spread is expressed using the population-SD convention. The annotations summarize the audited association and ridge regression defined in Section 3.3.

Benjamini–Hochberg false-discovery-rate correction (Benjamini and Hochberg, 1995), using $q < 0.05$ as the significance threshold. Its rank-correlation p-values and the separate regression permutation test are distinct. The latter compares predictive performance with 4,000 shuffles of system difficulties (Ojala and Garriga, 2010). The permutation test keeps the geometric feature table fixed.

4 Geometric structure of detector difficulty

4.1 Spread predicts MLAAD difficulty

On 61 MLAAD systems, spread correlates positively with WavLM-GAT difficulty: $\rho = 0.60$, with a reported bootstrap 95% interval of $[0.41, 0.73]$ (Figure 2). The leave-one-system-out ridge regression attains $R^2 = 0.277$ with permutation $p = 2.5 \times 10^{-4}$. Table 1 compares representative features from the recorded screen. The spread statistic is its sole global FDR survivor, with $q \approx 3.1 \times 10^{-5}$. Projection position and axis-independent dispersion have smaller absolute correlations and do not pass this correction.

The association is stable across the available sensitivity checks. Removing one system at a time gives correlations from 0.58 to 0.63. Partial rank correlations controlling for root-mean-square signal energy or utterance count are both approximately 0.58.

The association is stronger when only systems with more utterances are retained: $\rho = 0.59$ for 61 systems with at least eight utterances, 0.64 for 51 systems with at least 12, and 0.69 for 28 systems with at least 16. Greater measurement precision is one possible explanation. Because filtering also changes the system population, we interpret this as a sensitivity check.

MLAAD feature	ρ	Global q
Projected spread, layer 12	$+0.597$	0.000031
Median projection	-0.262	0.148
Centroid norm, layer 12	$+0.289$	0.109
Radius of gyration, layer 0	-0.381	0.059
Radius of gyration, layer 12	-0.308	0.094
Trajectory entropy, layer 9	$+0.352$	0.072

Table 1: Representative entries from the 83-test screen. Radius of gyration measures dispersion without choosing an axis. Only projected spread passes global $q < 0.05$. The global correction spans all 83 recorded tests, including the additional corpora.

The raw $R^2 = 0.277$ is our primary prediction measure. A split-half analysis estimated difficulty reliability at 0.759, with a full-sample correction of 0.863. Their ratio, $0.277/0.863 \approx 32\%$, provides secondary context under that reliability model rather than a causal variance decomposition.

4.2 Other corpora favor different statistics

The exploratory comparisons show that different corpora emphasize different geometric statistics. On ASVspoof 2019 LA, median position along the direction correlates with attack difficulty at $\rho = -0.71$ over 13 attacks ($p = 0.006$). The ITW speaker analysis instead reports a correlation with orthogonal position ($\rho = 0.53$, 29 speakers, $p = 0.003$). Here orthogonal position summarizes the norm of the embedding component perpendicular to the direction. An axis-independent dispersion statistic has a slightly stronger ITW correlation ($\rho = -0.56$); neither has a clear independent association after controlling for the other. These are exploratory associations with different statistics and grouping units. ASVspoof 2021 within-corpus position correlations of 0.60 and 0.64 for the two detectors do not survive the reported family correction.

MLAAD emphasizes spread, whereas the smaller ASVspoof and ITW samples emphasize position-related or axis-independent statistics. This motivates studying which properties of the genuine and synthetic distributions determine the most informative summary. Their means, disper-

Corpus pair	MLAAD frame	Raw
MLAAD / ITW	$+0.103$	$+0.027$
MLAAD / ASVspoof 2021	-0.207	-0.140
ITW / ASVspoof 2021	-0.238	-0.118

Table 2: Audited cross-corpus direction cosines. The middle column uses the same MLAAD genuine-speech scaling for all corpora; the last uses raw embeddings. These values depend on the coordinate metric.

sion, and other structure can all affect projected overlap.

Using WavLM spread to rank fine-tuned AASIST difficulty on MLAAD gives a smaller association, $\rho = 0.35$ ($p = 0.006$, global $q = 0.072$). Its leave-one-system-out regression gives $R^2 = -0.032$ ($p = 0.13$). The weaker result highlights the detector-specific nature of the diagnostic: geometry is shared, but difficulty depends on the decision function.

4.3 Corpus directions reveal domain dependence

For two directions u_a, u_b, cosine similarity is $u_a^\mathsf{T} u_b$ when both are unit normalized. Values near 1 indicate similar orientations, values near 0 indicate weak alignment, and negative values indicate opposing components. Table 2 compares corpus directions in two coordinate metrics. We also assess stability within each corpus by splitting its data in half. Split-half cosines measure how consistently the direction is estimated within the sampled corpus.

The audited pairs have weak or negative alignment, while within-corpus split-half cosines are 0.926 for MLAAD, 0.982 for ITW, and 0.981 for ASVspoof 2021. This contrast is consistent with domain-specific representation geometry. The marked change in direction offers a geometric perspective on why a detector-relevant signal may not transfer unchanged across corpora. The magnitudes depend on the coordinate metric, as the two columns show.

We treat these cosines as descriptive measurements. Transferring the hardness predictor to 13 ASVspoof 2021 attacks gives $\rho = 0.25$ and $p = 0.41$, leaving cross-corpus prediction an open question in this small sample.

5 Using labeled geometry to adapt scores

5.1 Centroid projection and supervised probes

The preceding sections use geometry to describe differences among synthesis systems. Here we ask a separate question: can the same labeled direction supply a score for individual utterances? For centroid fusion, that value is its projection $p(x)$ onto a direction estimated from reference data. If z_d and z_p standardize detector and projection scores using training-partition means and standard deviations, the fused score is

$$f(x) = z_d(d(x)) + \lambda z_p(p(x)). \tag{8}$$

Both components are oriented toward synthetic speech. In the MLAAD study, the coefficient λ maximizes training-partition AUC over $\{0, 0.25, 0.5, 0.75, 1, 1.5, 2, 3\}$. The axis and coefficient use the same training partition, rather than a separate inner validation split. Test labels are excluded from these choices. No per-system difficulty labels enter this fusion rule.

A supervised linear discriminant analysis (LDA) probe is a different score. It estimates a class-separating linear function using the class means and within-class covariance of the embeddings. The implementation fits raw embeddings using the least-squares solver with automatic covariance shrinkage. The logistic comparator fits genuine-standardized embeddings with inverse regularization $C = 0.1$ and a 2,000-iteration limit. These probes use target-domain class labels and estimate more structure than a centroid difference. Their component-only results separate the value of this additional fitting from the centroid construction.

Score adaptation uses labeled genuine and synthetic calibration examples to estimate the direction, standardize scores, and select the coefficient; the detector weights remain fixed. A supervised probe additionally fits a classifier. WavLM-GAT can reuse its existing representations, whereas adding these scores to AASIST requires a WavLM encoder. This distinction determines the added inference cost.

5.2 Centroid fusion benefits detectors with headroom

Table 3 reports equal error rate (EER), the error level where the false-positive and false-negative rates are equal as the threshold varies. All EER

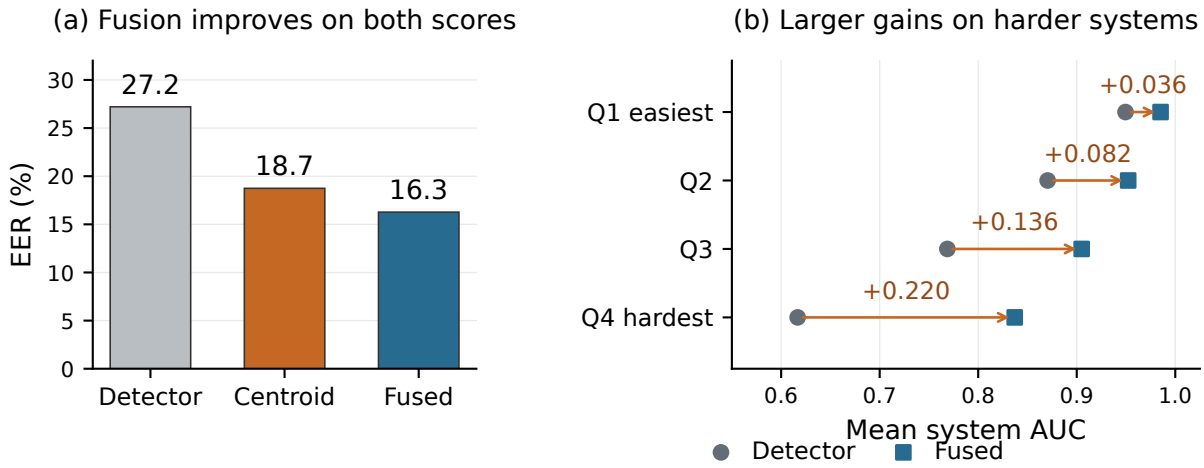


Figure 3: Two views of MLAAD centroid fusion, redrawn from the original saved results. (a) Mean EER across three WavLM-GAT runs: fusion improves on the detector and centroid score alone. (b) Mean system AUC within quartiles defined by baseline detector difficulty; arrows label absolute AUC gains. The largest gains occur in the hardest quartile. This breakdown conditions on baseline difficulty and describes the evaluated systems.

WavLM-GAT setting	Detector	Axis	Fused
ASVspoof 2019, weak	14.4	—	11.4
MLAAD, within corpus	27.2	18.7	16.3

Table 3: Centroid-projection fusion EER (%). The weak detector was trained with 500 utterances. An axis-alone value for that row is not reported here. The MLAAD fusion improves on both component scores.

values in this paper are percentages; differences are percentage points. EER measures the quality of a score ranking and does not supply a deployment threshold.

For the deliberately undertrained ASVspoof 2019 detector, EER decreases by 3.0 points, a relative reduction of approximately 21%. The reported paired-bootstrap interval for the EER difference is $[-4.6, -1.3]$ points, with $p < 10^{-3}$. All five evaluated folds improve. A stronger WavLM-GAT starting at 7.8% EER shows small changes, from approximately -0.04 to $+0.4$ points across the reported variants, with p between 0.56 and 0.80. The stronger detector changes little, with both small improvements and small increases in error.

On MLAAD, the fused EER of 16.3% improves on both the detector's 27.2% and the centroid score's 18.7% (Figure 3a). Five folds hold out synthetic system–language groups and independently partition genuine utterances; this is repeated for three detector runs. Each fold estimates the scaler, axis, score standardizers, and coefficient on its training portion. Improvements occur across all 15 fold-by-run combinations. The selected fusion coefficients are 1.5 or 2.0. Across these settings, larger gains occur for weaker starting detectors. The MLAAD component comparison shows that the centroid score supplies complementary information in that setting.

The original per-system breakdown locates these gains within MLAAD (Figure 3b). Mean AUC gains increase from 0.036 in the easiest baseline-difficulty quartile to 0.220 in the hardest. Quartiles are defined by the three-run mean baseline difficulty, and each point averages system AUCs across systems and runs. This descriptive comparison shows where the observed gain is concentrated, with greater room for improvement in the harder groups.

5.3 AASIST separates probe information from fusion gains

The AASIST comparison reveals where the adaptation gain comes from. For the unadapted checkpoint evaluated outside its training domain, fusion reduces EER by 26.0 points on MLAAD and 32.5 points on ITW (Table 4). The supervised probe alone is stronger still, outperforming fusion by 1.6 and 6.0 points, respectively. Target-domain WavLM embeddings therefore contain substantial discriminative information that the transferred detector does not exploit; in these rows the probe captures all of the gain.

The original ITW protocol permits genuine

Target corpus / split	AASIST	LDA	Fused
MLAAD	37.6	**10.0**	11.6
ITW, original	48.6	**10.1**	16.1
ITW, speaker-disjoint	—	**14.3**	19.3
ASVspoof 2021	7.3	11.5	7.6

Table 4: Supervised LDA adaptation of AASIST, EER (%). Original ITW folds share genuine speakers across training and test. The saved stricter audit separates genuine speakers as well. A separately reported AASIST value for that audit is unavailable. The LDA probe outperforms fusion on MLAAD and under both ITW protocols.

Corpus	Added score	5	25	250
MLAAD	Centroid	28.3	32.7	16.6
	LDA	—	32.3	13.6
ITW	Centroid	34.9	31.2	28.7
	LDA	—	28.7	19.2

Table 5: Calibration sweep: fused EER (%) at the indicated number of labeled examples per class. LDA is evaluated only from 25 examples per class. Calibration groups and test subsets change across draws and sizes; the entries are not a controlled learning curve on a fixed test set.

speakers to appear in both training and test folds. Separating genuine speakers raises LDA-alone EER to 14.3% and fused EER to 19.3%. Table 4 retains both protocols to show the effect of speaker separation.

The MLAAD probe comparison provides further context: LDA alone achieves 10.0% EER and logistic regression 11.2%, compared with the centroid score's 18.7%. Their mean EER reductions when fused with WavLM-GAT are approximately 15.7 and 15.4 points, respectively, versus 10.9 points for centroid fusion. The probe fusion uses fixed $\lambda = 1.5$, whereas centroid fusion selects its coefficient. A comparison under matched label and compute budgets would further clarify these tradeoffs.

5.4 Calibration-size evidence

A calibration sweep evaluates 5, 10, 25, 50, 100, and 250 labeled examples per class, with 20 attempted group-aware draws at each size. It fits the direction and fusion coefficient on the calibration observations and excludes their groups from evaluation. ITW uses speaker groups; MLAAD uses synthesis groups for synthetic speech. This sweep adapts WavLM-GAT, so it is separate from the AASIST comparison.

Table 5 gives selected sizes. Small calibration sets yield variable errors, and centroid results are nonmonotonic. The saved detector baseline uses the full subset, while adapted scores exclude calibration groups; we report adapted EERs only. The sweep describes the evaluated calibration settings, with changes in group coverage and test composition limiting label-efficiency conclusions.

6 Discussion

The main finding is that a simple direction in frozen speech representations captures meaningful variation in detector difficulty across MLAAD synthesis systems. The diagnostic makes this variation inspectable: systems can be compared by how their synthetic utterances occupy the genuine–synthetic direction, alongside their observed errors. Direct AUC measures performance; projected spread supplies a compact representation-space description of the failure structure. In an evaluation with labeled genuine and synthetic examples, one can report detector AUC and projected spread by synthesis group, inspect high-spread groups for concentrated errors, and prioritize them for further testing. Because the informative statistic changes across corpora, this ranking should be derived within the target evaluation domain.

The cross-corpus results extend this perspective. Directions are stable within the sampled corpora yet differ markedly between them, and the informative summary changes with the dataset. Together, these observations motivate studying the distributional conditions under which spread, position, or other geometry best describes difficulty. As expected for a one-dimensional diagnostic, substantial variation remains for content, speaker, channel, synthesis characteristics, and other representation features to explain.

The score experiments show a related practical use. Centroid fusion can improve on both components when a detector has headroom, as on MLAAD. Stronger supervised probes capture additional target-domain information, and the AASIST decomposition identifies the probe as the source of its largest gains. Reporting component-only scores makes the contribution of each source of information visible. These findings support representation-space analysis as a tool for understanding detector behavior and evaluating adaptation choices.

7 Conclusion

Simple structure in frozen WavLM representations captures meaningful variation in audio deepfake detection difficulty. Across 61 MLAAD synthesis system–language groups, projected spread provides a compact hardness signal, supported by the recorded feature screen and leave-one-system-out ridge analysis. Cross-corpus direction changes reveal substantial domain dependence. Labeled geometric scores can also complement weaker detectors, while supervised probes expose additional information in target-domain embeddings. Together, these findings motivate representation-space diagnostics for understanding and auditing detector failures across synthesis systems. Further evaluation across encoders, synthesis families, and independently held-out corpora would clarify their reach.

Limitations

All geometric measurements use WavLM; generality to encoders such as HuBERT, wav2vec 2.0, and Whisper remains untested. The primary finding concerns text-to-speech systems, while the available voice-conversion results do not show the same association. Projected spread (`sd_along`) and layer 12 emerged from exploratory analysis within MLAAD, motivating independent replication on additional corpora. The analyzed MLAAD subset is English-focused, so the evidence should not be read as broad multilingual validation. Cross-corpus analyses suggest that the most informative geometric statistic is domain dependent, and transfer to new datasets remains an open question. Target-domain geometric adaptation requires labeled bona-fide and synthetic examples. The adaptation experiments compare useful alternatives, but their label budgets, coefficient-selection procedures, and compute costs are not fully matched across methods; comparisons under identical budgets would clarify their relative value.

Ethical considerations

This work uses existing research datasets to characterize detector weaknesses and improve robustness evaluation. False accusations of manipulation can harm speakers, while missed synthetic speech can enable impersonation. Aggregate AUC and EER are insufficient evidence of authenticity for an individual recording, and detector outputs should not be used alone for such judgments.

Speech can reveal identity and personal content; use and redistribution remain subject to the recordings' original terms and consent conditions. Representation-space diagnostics may also reveal properties that an adaptive adversary could exploit. Our purpose is defensive evaluation of detector failures.

Acknowledgments

We thank Jathin Korrapati, Ryan Lagasse, Archana Vaidheeswaran, Arjun Bahuguna, and Kevin Zhu for their mentorship, guidance, and support throughout this research.

OpenAI Codex assisted with manuscript organization, citation and formatting checks, and simulated reviewer feedback. It did not run experiments or generate empirical results.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6367–6371.

Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does audio deepfake detection generalize? In *Interspeech 2022*, pages 2783–2787.

Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. MLAAD: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Markus Ojala and Gemma C Garriga. 2010. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jeeweon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 112–119.

Xin Wang and Junichi Yamagishi. 2022. Investigating self-supervised front ends for speech spoofing countermeasures. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 100–106.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, et al. 2020. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 47–54.

Kuiyuan Zhang, Zhongyun Hua, Rushi Lan, Yushu Zhang, and Yifang Guo. 2025. Phoneme-level feature discrepancies: A key to detecting sophisticated speech deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1066–1074.

A Reproducibility details

A.1 Analyzed subsets and units

The MLAAD cache contains 1,846 clips (908 genuine, 938 synthetic) in 63 synthetic system–language groups. Two groups have five and six clips, leaving 61 after the minimum-eight filter. Geometry and detector scores reuse these cached representations, a subset of the published dataset.

ASVspoof 2019 position analysis uses 800 genuine and 800 synthetic clips from attacks A07–A19. ASVspoof 2021 transfer uses 400 genuine plus 40 synthetic clips per attack across 13 attacks (920 total), with codec `none`. ITW geometry uses 1,500 clips per class; its table has 53 speakers, of whom 29 have at least ten synthetic clips for difficulty analysis. These corpora differ in sampling and grouping.

A.2 Encoder and detector artifacts

The geometry encoder is the Hugging Face `microsoft/wavlm-base` checkpoint, evaluated without gradient updates. The loader explicitly restores positional-convolution weight-normalization parameters from the pretrained checkpoint when the installed Transformers/PyTorch combination exposes the newer parametrization interface. This prevents legacy `weight_g`/`weight_v` tensors from being left unmatched. Layer-12 hidden states are averaged over the time dimension. Cached clips have 48,000 samples at 16 kHz; the MLAAD preprocessing uses center crops, and the ITW helper center-crops long recordings and pads short ones with zeros at the end.

The primary MLAAD score files correspond to saved checkpoints labeled `main`, `s42`, and `s1024`. The main detector was trained on the training split of MLAAD-tiny and selected using its validation split. Its training record lists 10,348 training and 2,218 validation clips, seven epochs, batch size 20, and best validation epoch 5. Its augmentation probability is 0.35, with noise/pitch/reverberation sampling weights of 0.5/0.3/0.2. The saved settings use noise SNRs from 8 to 40 dB, upward pitch shifts from 1 to 5 semitones, and reverberation times from 0.2 to 0.9 seconds. The replication checkpoints are selected at validation epochs 5 and 3. Evaluation is within MLAAD; generator, speaker, and text overlap with detector training is unverified.

The fine-tuned AASIST checkpoint supplies the rank-association comparison in Section 4.2. The transferred, unadapted checkpoint supplies Table 4. The anonymous companion tables contain the audited per-system values, counts, and all 83 recorded feature tests. They support inspection of the reported analysis, but do not replace the raw-audio split manifests, training code, and checkpoints needed for end-to-end reproduction.

A.3 Geometry conventions and regression

Coordinate standard deviations use the population convention and add 10^{-9} to avoid division by zero. The axis uses a clip-weighted synthetic centroid: systems with more reference clips contribute more to that mean. The audited projection is $(\widetilde{e}_i -$

$\widetilde{\mu}_B)^\mathsf{T} u_{-s}$; centering by $\widetilde{\mu}_B$ is equivalent to Equation 4's projection up to numerical roundoff, because genuine-centered coordinates have zero genuine mean. This extra centering does not change within-system spread.

The exploratory table computed sample standard deviations with denominator $n_s - 1$. The audited headline recomputation used population standard deviations with denominator n_s, yielding $R^2 = 0.276647$ and $\rho = 0.597462$. Figure 2 directly uses the saved audited feature table and three-run mean difficulty, matching the convention in Equation 6. The audit reconstructed these features from cached embeddings. We retain its population convention in the formal definition and report its rounded statistics.

The ridge predictor includes an intercept and uses $\alpha = 1$. Standardizing its input once across all systems precedes the regression leave-one-out loop. System difficulties average the three detector runs; labels are synthetic-positive. The feature table is constructed once: each system has its own leave-one-system-out reference direction, and those measurements are reused across the regression folds. The permutation analysis also reuses this table and does not repeat the preceding feature screen.

The headline audit uses 4,000 difficulty permutations with random seed 1, reporting $(1 + k)/(4000 + 1)$ when k null scores meet or exceed the observed R^2. The resulting $p = 0.00024994$ is rounded in the text. Its rank-correlation confidence interval uses 10,000 bootstrap resamples with seed 2, sampling the 61 system-level $(\sigma_\parallel, h)$ pairs with replacement and keeping their measured spread and AUC values fixed. The uncorrected rank-correlation p-value is 3.71×10^{-7}; it is not the permutation p-value for R^2.

A.4 MLAAD score fusion and probes

The score-fusion implementation shuffles sorted synthetic group identifiers with seed 42 and divides them into five folds. It separately permutes genuine utterance indices into five folds. In each round, four folds estimate the genuine coordinate scaler and centroid direction. Projection and detector-score standardizers use the training observations from both classes, with population standard deviations. The coefficient grid and training-AUC criterion are specified in Section 5. Predictions are concatenated across test folds before calculating each run's EER; the headline averages the three run-level EERs.

For the three detector runs, centroid fusion selects λ values $[1.5, 1.5, 1.5, 1.5, 2]$, $[1.5, 2, 2, 1.5, 2]$, and $[1.5, 1.5, 1.5, 1.5, 2]$. The unrounded mean EERs are 0.272113 for the detector, 0.187429 for the centroid score, and 0.162693 for fusion. The paired utterance bootstrap uses 2,000 resamples and seed 42; the mean EER-difference interval is approximately $[-0.1286, -0.0908]$. These are utterance-resampling intervals, not uncertainty intervals over new corpora or synthesis families.

The LDA comparator uses `solver=lsqr` and `shrinkage=auto` on raw 768-dimensional embeddings. Logistic regression uses genuine-standardized embeddings, $C = 0.1$, and at most 2,000 iterations. Both use the same five synthetic-group/genuine-utterance folds as the centroid comparison. Their fusion comparison fixes $\lambda = 1.5$ and fits score standardizers on training observations. This differs from the centroid coefficient search and should be considered when comparing their gains.

The AASIST adaptation audit uses the same shrinkage-LDA specification. Its original protocol holds out synthetic groups but randomly splits genuine utterances; the stricter ITW audit also separates genuine speakers. Table 4 retains both ITW rows to show the effect of this choice.

A.5 Calibration-size sweep

For size n and repetition r, calibration uses seed $1000r + n$. The script permutes class-specific group IDs, collects at least n utterances, and takes the first n; selected groups are excluded from testing. Draws with fewer than 200 test clips or one class are skipped. MLAAD genuine clips use five deterministic index-based groups, not speakers; ITW groups both classes by speaker. Independence therefore differs by corpus.

Calibration AUC selects the coefficient from $\{0.5, 1, 1.5, 2\}$, with no detector-only option. The same LDA specification is attempted for $n \geq 25$. Means average retained draws, then detector runs (three for MLAAD, one for ITW). Adapted EER uses changing test subsets; detector-only EER uses the full corpus subset, so saved improvements are unmatched. Table 5 reports adapted EERs only, not performance on a fixed external test set.