VOICE OR LENGTH? COUNTERFACTUAL AUDITING OF TIMING SHORTCUTS IN AUDIO DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A benchmark can remove a shortcut from its evaluation data while leaving it in its training data. A detector may then learn a highly predictive nuisance cue that the benchmark neither rewards nor exposes. We study this gap for timing in audio deepfake detection, and separate a shortcut's *availability* in a corpus from a detector's *reliance* on it. Availability: ASVspoof 5 equalises timing across classes, by design, only in the partitions it tests on; on held-out speakers of its training partition, timing alone separates the classes at AUC 95.6, against chance on its evaluation set. Across four corpora the predictive component differs, and on ASVspoof 2019 LA (LA19) the components that survive boundary trimming, the standard control, retain AUC 87.7 of the 88.3 available. Reliance: matched-length evaluation confounds speech amount with boundary context and a shifting population, so we intervene instead, holding the speech byte-identical. Planting real bona fide boundary silence on spoofed ASVspoof 5 clips raises a released XLS-R Conformer from 5.8 to 9.4% EER, against 7.1% when every clip receives it. Incentive: fine-tuning that checkpoint further on its natural view lowers its LA19 benchmark EER from 0.35 to 0.17% while raising its length-matched In-the-Wild EER by 4.7 points, so, in this setup, selecting on the benchmark picks the detector that relies more on the shortcut. Randomly placed training crops reverse this and remove the planted-silence reliance, at a cost in LA19 EER, and turning the intervention into an objective — a consistency loss between each crop and the same speech with label-independent silence added — gives the lowest length-matched In-the-Wild EER of any LA19-trained detector (6.6%, from 9.2%); a duration-aware architecture, tested over twenty seeds, does not help. For timing, removing the cue from a test set does not test for it; intervening on it does. We release the code, ports of five released checkpoints, and per-clip scores for every condition.

1 INTRODUCTION

Synthetic speech from modern text-to-speech (TTS) and voice-conversion systems is now hard for listeners to tell apart from real speech. This makes automatic detection of spoofed audio a practical safety need. On standard benchmarks the problem looks close to solved. Detectors built on self-supervised speech representations reach equal error rates (EER) below 1% on ASVspoof 2021 LA (Tak et al., 2022b; Rosello et al., 2023; Kim et al., 2025). The same systems, however, perform much worse on recordings collected outside the benchmark (Müller et al., 2022). On deepfakes that circulated online in 2024, open-source audio detectors lose about half of their AUC (Chandra et al., 2025).

A common explanation for such gaps is *shortcut learning* (Geirhos et al., 2020): a model can use a feature that predicts the label in the training data but has nothing to do with the task. In audio anti-spoofing, a well-documented example is the length of the leading silence in ASVspoof 2019. A classifier that sees *only* this length already reaches about 85% accuracy (Müller et al., 2021). Timing artefacts are by now a recognised problem: Zhang et al. (2023) show that removing silence with a voice-activity detector sharply degrades detectors, Rubio et al. (2026) find non-speech to be the dominant shortcut for an XLS-R front-end, and ASVspoof 5 is built with a pipeline that equalises the leading and trailing non-speech and the whole utterance duration across the two classes (Wang

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2026b). We therefore take it as established that **how long a clip is**, and how much silence surrounds it, can predict its label, and we do not claim to discover it.

ASVspoof 5 applies that equalisation to its development and evaluation partitions only, leaving the training partition untouched by design, as a test of whether detectors avoid the cue. We find the cue left there is strong: on held-out speakers of the training partition, timing alone separates the classes at AUC 95.6, against chance on the evaluation set. Whether a detector trained there learned it is therefore invisible to the benchmark, by construction. That is one instance of a general gap. "Duration" is not one variable, and the standard response — trim the silence, compare at matched lengths — controls much less than its name suggests. Three things happen at once when a clip is shortened: the detector gets less speech, it loses the non-speech context around that speech, and the set of clips long enough to qualify changes, so the population under evaluation changes as well. A curve of EER against length mixes all three.

The question, in short: *when a detector does well, is it detecting the voice or the length?* What we can do is decompose the second half of it, hold each part fixed while the others move, and test what removes it.

Contributions. Our contribution is not identifying timing as a shortcut, but separating shortcut availability from detector reliance with interventions that preserve the speech evidence.

1. **Availability is not uniform, and evaluation can hide it (§3).** A component-wise timing analysis shows that the corpora do not share one shortcut: boundary silence dominates In-the-Wild (AUC 77.1), total duration dominates FoR (81.6), and on LA19 the components that *survive boundary trimming* still reach 87.7 of the 88.3 available. ASVspoof 5's training partition, left unequalised by design, carries the strongest cue we measure (95.6) while its evaluation partition carries none, so standard evaluation cannot reveal whether a detector learned it.
2. **Counterfactual tests of reliance (§5).** Matched-length curves change speech amount, boundary context and the evaluated population at once — the LA19 attack mix drifts 10.31 percentage points across lengths, and a fixed cohort moves released checkpoints' curves by up to 2.89 EER points. Holding speech byte-identical separates boundary context, and planting real bona fide boundary silence on spoofed ASVspoof 5 clips only raises the released XLS-R Conformer from 5.8 to 9.4% EER, against 7.1% when every clip receives it.
3. **Benchmark optimisation can favour the shortcut; changing the training view removes it (§6).** Further natural-view fine-tuning of a released checkpoint halves its LA19 benchmark EER and worsens its length-matched In-the-Wild EER by 4.7 points. Randomly placed training crops reverse this — most of the gain survives even when the crops keep their silence — and remove the transplant reliance, at a cost in LA19 EER; the same holds when training on ASVspoof 5. Turning the intervention into a training objective, counterfactual timing invariance, gives the lowest length-matched In-the-Wild EER of any LA19-trained detector (6.6%). A duration-aware attention span, built as a probe and tested over twenty seeds, is matched by a duration-independent control.

2 Related work

Detectors and benchmarks. Early end-to-end systems operate on raw waveforms, such as RawNet2 (Tak et al., 2021) and the graph-attention AASIST (Jung et al., 2022). The strongest current systems fine-tune self-supervised front-ends (Baevski et al., 2020; Babu et al., 2022) with AASIST (Tak et al., 2022b), Conformer (Rosello et al., 2023; Gulati et al., 2020), layer-selection (Zhang et al., 2024a), temporal-channel attention (Truong et al., 2024) or Mamba (Xiao & Das, 2025) back-ends, usually with RawBoost (Tak et al., 2022a); curriculum learning on predicted naturalness improves generalisation further (Kim et al., 2025). Lower SSL layers carry most of the useful information (El Kheir et al., 2025), which motivates our frozen, truncated front-end. Training is usually on ASVspoof 2019 LA (Wang et al., 2020) with evaluation on ASVspoof 2021 (Liu et al., 2023); newer corpora include ASVspoof 5 (Wang et al., 2026b;a), WaveFake (Frank & Schönherr, 2021) and FoR (Reimao & Tzerpos, 2019), and real-world sets such as In-the-Wild (Müller et al., 2022) and Deepfake-Eval-2024 (Chandra et al., 2025) report large drops.

Timing shortcuts. That these benchmarks contain artefacts inflating reported accuracy is established: Chettri et al. (2020) traced ASVspoof 2017 results to dataset artefacts, Müller et al. (2021) showed leading-silence length alone predicts the label in ASVspoof 2019, Zhang et al. (2023) studied silence systematically, and Rubio et al. (2026) find non-speech dominant for an XLS-R front-end. ASVspoof 5's designers treat it as settled enough to engineer against, equalising peak amplitude, leading and trailing non-speech and *whole utterance duration* across classes (Wang et al., 2026b) — in the development and evaluation partitions, leaving the training partition unprocessed by design; §3 measures how much of the cue that leaves. We therefore claim no discovery of the shortcut, but ask which timing component a detector reads, whether the standard controls control it, and what removes it.

Three precedents bound what is new. Zhang et al. (2024b) report EER against evaluation duration, sweeping 1–6 s for four systems, though on untrimmed audio, without class balancing, and framed as robustness rather than shortcut removal. Smeu et al. (2025) identify a leading-silence shortcut in audio-visual data and re-evaluate prior detectors on a trimmed test set — the same move, one modality across. Martín-Doñas et al. (2024) diagnosed the ASVspoof 5 duration bias and fixed their input at 6 s so the network could not exploit it, which is consistent with the training-partition cue we measure. Closest in spirit is Rubio et al. (2026), who perturb acoustic properties of a single XLS-R system within a graphical-model framing. We differ in three ways that matter for what can be concluded: their standard inputs are truncated or padded to 64,600 samples rather than resampled in time, so timing cues can survive; we intervene on the boundary while holding the speech byte-identical, which separates non-speech context from speech amount; and we audit five released checkpoints across three front-end families rather than one, which is what makes ranking stability measurable at all. Our contribution is thus not identifying timing as a shortcut, but separating its availability in a corpus from a detector's reliance on it.

Duration as a modelling choice. Several works treat duration as a property to handle rather than a cue to exploit: AASIST2 targets short utterances (Zhang et al., 2024b), Rosello et al. (2023) show training crop length strongly changes EER and process variable-length inputs, and LENS-DF studies long-form noisy recordings (Liu et al., 2025). None takes duration as an input to the forward pass, which is the design we construct (§6.3, Appendix O) to ask what happens when one does; AASIST2 uses it in the training objective, as a margin linear in the training clip's duration, but not at inference.

Shortcuts, consistency and attention biases. Shortcut learning is a general failure mode (Geirhos et al., 2020), and its corpus-level form is documented in vision (Torralba & Efros, 2011) and natural-language inference (Gururangan et al., 2018). Where that literature compares remedies, the data-side ones hold up: balancing groups is competitive with worst-group objectives (Sagawa et al., 2020; Idrissi et al., 2022), and retraining only the last layer on balanced data recovers most of what specialised methods buy (Kirichenko et al., 2023). Our results have the same shape, one domain across: changing the training view helps (§6.2), a duration-aware architecture does not (§6.3). Consistency objectives over augmented views are standard (Hendrycks et al., 2020; Sohn et al., 2020; Tarvainen & Valpola, 2017) including over differently sized crops (Caron et al., 2020; 2021), and appear in this domain across spectral resolutions (Shahriar, 2026); our multi-length consistency is an instance of that family, not a new principle. Distance-proportional attention biases were introduced for length extrapolation (Press et al., 2022), and Sukhbaatar et al. (2019) learn per-head spans including an input-predicted variant. Our duration-adaptive span differs only in conditioning that span on one utterance-level quantity, so we treat it as a controlled probe of *how* duration may enter a detector, not as a new attention mechanism.

3 Timing shortcuts in deepfake corpora

Data. We use four corpora.

- **ASVspoof 2019 LA** (LA19; Wang et al., 2020): the train split (2,580 bona fide / 22,800 spoofed clips), the dev split and the eval split (7,355 / 63,882). Our eval labels match the official protocol for all 71,237 clips.
- **In-the-Wild** (ITW; Müller et al., 2022): 19,963 bona fide and 11,816 spoofed clips of 54 distinct speaker identifiers in the supplied metadata.

- **FoR** (Reimao & Tzerpos, 2019): the `for-norm` train (26,939 / 26,927) and test (2,264 / 2,370) splits.
- **ASVspoof 5** (ASV5; Wang et al., 2026b), track 1: the train split (18,797 / 163,560), the dev split and the uncoded eval trials (7,031 / 26,732).

For every clip we measure four timing features: its duration, the duration of speech after silence trimming (§4), and the lengths of the leading and trailing silence.

Timing distributions differ between classes, and in different directions. Table 8a summarises the per-class distributions, which Figure 2 in Appendix J plots in full.

- **FoR.** Duration nearly separates the classes. Bona fide clips have a median duration of 4.12 s and spoofed clips 1.54 s (KS distance 0.70). 19.6% of bona fide clips are 7 s or longer, against 4 of the 26,927 spoofed clips (0.015%). Silence has already been removed from FoR. This asymmetry is specific to the `for-norm` training split: in the `for-norm`, `for-original`, `for-2sec` and `for-rerecorded` test splits no clip of either class reaches 7 s, and the median duration ratio falls from 2.67 to 1.63 or below.
- **LA19.** The strongest cue is silence, as reported by Müller et al. (2021). The median leading silence is 0.75 s for bona fide eval clips and 0.04 s for spoofed ones (KS 0.72).
- **ITW.** Both cues point the *other* way. Spoofed clips are longer (median 4.05 vs. 2.99 s; 23.1% vs. 9.5% of clips are 7 s or longer), and they carry *more* leading silence than bona fide clips (0.23 vs. 0.00 s).

A detector that learned "long silence means bona fide" from LA19 is therefore pushed in the wrong direction on ITW.

Timing alone predicts the label. A gradient-boosted classifier that sees *only* the four timing features, and nothing of the audio content, reaches held-out AUCs of 83.3 on LA19 (EER 23.6%), 81.6 on FoR (24.1%) and 78.0 on ITW (28.8%, speaker-disjoint split; Table 8b). Some of this information also transfers across corpora: trained on LA19 and tested on FoR, the classifier still reaches an AUC of 75.0. For LA19 the useful information is mainly silence; for FoR it is duration.

Which timing property, and how much of the detector does it match? An AUC of 83 is hard to read alone: it is only alarming relative to how much separability the audio supports, and it says nothing about *which* timing property carries it. As a normalised descriptor we therefore report a Temporal Shortcut Index (Table 9, Appendix L). The descriptor is split into components — total duration, boundary silence, internal pauses within the trimmed region, speech amount — a weak classifier is fitted to each, and the resulting AUC is rescaled by that of a released acoustic detector on the same clips, $\mathrm{TSI} = (A_{\mathrm{temporal}} - 50)/(A_{\mathrm{audio}} - 50)$, so 1 means timing alone matches the detector and 0 means it carries nothing. It measures availability, not reliance: a high value says a detector *could* use timing, not that it does, which is what §5 tests.

Two things follow. The corpora do not share a shortcut: on ITW it is boundary silence (AUC 77.1) while total duration is nearly uninformative (58.2); on FoR it is entirely total duration (81.6) with boundary silence at chance (49.5), because `for-norm` ships pre-trimmed; on LA19 both contribute. Treating "duration" as one variable mis-describes at least two of the three. And boundary trimming does not remove the shortcut: the components surviving it — speech amount and internal pause structure — still reach AUC 87.7 on LA19 against 88.3 for all timing (TSI 0.75 against 0.77). Trimming helps substantially on ITW, where the shortcut *is* the boundary (77.2 to 68.3), and not at all on FoR (81.5 to 81.5). A trimmed evaluation is not a timing-controlled one.

ASVspoof 5 removes the cue only where it is measured. ASVspoof 5 was built against exactly these artefacts: its pipeline equalises leading and trailing non-speech and whole-utterance duration across the two classes, and its designers apply it to the development and evaluation partitions but deliberately not to the training partition, to encourage detectors that avoid the shortcut (Wang et al., 2026b). We measure what that leaves. On a speaker-disjoint half of the training partition, the same timing-only classifier reaches AUC 95.6 — boundary silence alone 92.1, duration alone 80.1 — well above the 83.3 of LA19. Bona fide training clips carry a median 0.25 s of leading and 0.32 s

of trailing non-speech against 0.04 and 0.03 s for spoofed ones, and are longer (14.9 against 11.3 s). Applied to the development or evaluation partition, the same classifiers fall to AUC 44.9–50.3, at or below chance. The corpus therefore offers a stronger timing shortcut to the detectors trained on it than LA19 does, and its test partitions contain none of it — so a detector that learned the cue is neither rewarded nor exposed there. An absent cue is not a tested one; §5 puts it back.

An earlier duration-gated probe (DADAT) is reported for completeness in Appendix H but excluded from our conclusions, because its pooling gave it a second, padding-mediated duration channel.

4 MATCHED-LENGTH EVALUATION

The natural views are `fixed` (first 64,600 samples, repeat-padded for shorter clips) and `full` (untrimmed audio capped at 10 s). The fixed view has constant tensor length but can still retain timing cues through silence and repetition.

Matched-length evaluation adds three steps. An energy detector trims leading and trailing low-energy regions (25 ms frames, 10 ms hop, threshold 40 dB below the loudest frame). For each $L \in \{1, 2, 3, 4, 6, 8\}$ s we retain clips with at least L seconds of speech after trimming and take their central L seconds. The classes are then balanced with a fixed sampling seed, capped at 4,000 clips per class for LA19 and ITW and 2,264 for FoR; conditions with fewer than 100 clips per class are dropped (Appendix E). We report each $\mathrm{EER}(L)$, its mean $\overline{\mathrm{EER}}_{\mathrm{DC}}$ over available lengths, and its maximum. Available lengths differ by corpus, so those means are not comparisons over a common length distribution.

We call this matched-*length*, not duration-controlled, because it fixes the waveform length and little else. It does not guarantee speech-only content: internal pauses, low-energy speech and recording artefacts survive, and §3 shows the timing components that survive trimming remain nearly as predictive on LA19 as all timing together. It does not hold the evaluated population fixed, and for duration-conditioned models it does not even remove the duration they are told. §5 and §6.3 measure each of these and remove them where we can. A natural-to-matched gap is therefore a sensitivity to a specified preprocessing, not by itself an estimate of shortcut reliance; comparing models *within* one condition holds the clips fixed but still cannot separate timing reliance from short-input robustness.

The FoR variants give a descriptive control: on `for-2sec`, where every clip is cut to two seconds during corpus construction and duration carries no label information, the 1 s condition still costs $+11.6$ EER points against $+13.6$ on `for-norm`, so poor short-crop performance need not imply a duration cue (Appendix I).

Statistics. Intervals are two-level paired bootstraps that pair clips across models and resample training seeds within each; families are Benjamini–Hochberg corrected, and a speaker-cluster bootstrap is reported where identity exists (Appendix D).

5 WHAT MATCHED-LENGTH EVALUATION DOES NOT CONTROL

Three quantities move together when a clip is shortened: the amount of speech, the non-speech context around it, and the set of clips eligible to be shown at all. A curve of EER against L is their sum. This section separates them, and then asks directly whether a detector decision depends on boundary silence.

The evaluated population changes with length, and we can remove it. A clip qualifies for the L-second condition only if it keeps L seconds of speech after trimming, so the cohort shrinks and shifts as L grows. LA19 lets us measure the shift, because it labels attack families: between 1 s and 3 s the spoof mix moves 10.31 percentage points of total variation, the largest single change being A17 at 5.27 pp (Table 11). Families are far from equally hard — for the XLS-R Conformer at $L{=}1$ the pooled EER of 11.14 sits well above the median family (5.52) and far below the worst (31.50) — so a pooled EER is carried by a few of them, and if their share changes with L, part of the curve is composition. The direction is not even shared across architectures: A17 is among the easiest families for the SSL-based detectors and among the harder for AASIST, so the same drift pushes different models different ways, which is one mechanism by which rankings reorder. We therefore select the

cohort once, from clips eligible at the longest length, and draw every shorter view from those same recordings, so that $x_i^{1s} \subset x_i^{2s} \subset \cdots \subset x_i^{L_{\max}}$. Curves keep their shape — every model still degrades sharply as speech is removed — but move by 0.91 EER points on average and up to 2.89 (Table 12): that displacement is the part of the published curve that was the population changing. The cost is power, since four seconds of speech leaves only 478 balanced LA19 clips against 8,000 at L=1, so we report both and treat neither as a substitute for the other.

Separating boundary context from speech amount. Shortening a clip removes its boundary silence along with the speech, so a matched-length gap cannot say which one the detector was using. We hold the speech constant instead. The centre L seconds of trimmed speech are byte-identical across conditions, and P seconds of boundary silence are added back at each end, resampled from a bank pooled over both classes. That pooling is what makes this an intervention rather than a correlation: the added context is drawn from the same distribution whatever the clip's label, so it cannot be informative about that label, and any change in the output is the detector responding to boundary context itself. P=0 is exactly the matched-length view and serves as the paired reference.

Adding one second of uninformative silence to each end of two seconds of unchanged speech moves AASIST from 13.28 to 20.49 EER, changing 43.2% of its decisions (Table 10). Since neither the speech nor the informativeness of the padding changed, the decision is partly a function of non-speech structure, and the deterioration seen under trimming is not only lost speech evidence. The XLS-R-based systems are more telling still: their EER *falls* when a little silence is restored, the Conformer going from 6.02 to 3.43 at 0.25 s before rising again. Their best operating point is therefore not the one with the most speech but the one whose boundary statistics resemble the untrimmed audio they were trained on — which is what a dependence on boundary context looks like. FoR ships pre-trimmed, with no boundary silence to resample, so the intervention is undefined there — consistent with its boundary-silence AUC of 49.5 (Appendix L), which localises its shortcut entirely in total duration.

Planting the cue. Write a clip as $x = (s, z)$, with s the speech and z its timing context. Availability (§3) asks whether z predicts y; *reliance* asks whether $f(s, z) \neq f(s, z')$ for a label-preserving change $z \to z'$ with s byte-identical. We measure it two ways: the EER change when z' is planted on spoofed clips only, net of planting it on every clip, and the counterfactual flip rate — the share of spoofed clips rejected on s alone but accepted as bona fide on (s, z'), at the model's own EER threshold. ASVspoof 5's evaluation set has no boundary silence left to read (§3), which lets us add it back under control. From a balanced set of 2,000 bona fide and 2,000 spoofed evaluation clips with at least 4 s of speech we take the central 4 s, and give it the leading and trailing non-speech of a real bona fide ASVspoof 5 *training* clip, the donor chosen deterministically per clip. In one condition only the spoofed clips receive it; in the control every clip does. The speech is byte-identical across all three views, so a detector merely disturbed by extra non-speech moves in both conditions, while one that reads bona-fide-like silence as evidence of bona fide speech moves in the first only. The released XLS-R Conformer behaves like the latter. Its EER rises from 5.8 to 9.4 when the silence is planted on spoofs, against 7.1 when every clip gets it (difference +2.3, 95% CI [+1.7, +3.0]); its curriculum variant goes from 4.9 to 10.9 against 6.8; 4.8% and 7.9% of their spoofs flip to bona fide (12.2% for SSL-AASIST). The effect is timing, not the donor's acoustics: planting *digital* silence of exactly the same durations moves the Conformer from 5.8 to 9.1% against 6.9% (net +2.2, [+1.6, +2.8]) and CL+DT to 10.9% against 6.7%. The same test tells us, in §6.2, whether a fix removes the reliance or only the opportunity.

6 EXPERIMENTS

Setup. Three kinds of detector appear below. The five released checkpoints of §6.1 are audited as published. Two of them are then fine-tuned end to end under different training views (§6.2). And a set of detectors with a *frozen* XLS-R front-end and about 0.7 M trainable parameters is trained from scratch, on LA19 or on ASVspoof 5, as a controlled testbed in which only the handling of duration changes (§6.2, §6.3); these are not meant to compete on absolute EER. LA19-trained models are selected on LA19 dev, ASVspoof 5-trained ones on ASVspoof 5 dev (Appendix B). Table 1 reports the released checkpoints and the LA19-trained testbed; Table 2 the training-view experiments.

Table 1: EER (%) under standard and matched-length evaluation. *nat.*: first 4.04 s of the untrimmed clip (the usual input); *trim*: whole clip without leading/trailing silence; *DCE*: mean EER over matched lengths L (LA19: 1–4 s, ITW: 1–8 s, FoR: 1–2 s, limited by the number of long bona fide and spoofed clips); *max*: worst length. Our models: mean±std over seeds (number in brackets). †Released checkpoints, evaluated with our ports (Appendix B). AASIST and RawNet2 are at or near chance on ITW and FoR, so their matched-length numbers there are reported descriptively without interpreting their changes as shortcut effects.

	LA19 eval				In-the-Wild				FoR test			
Model	nat.	trim	DCE	max	nat.	trim	DCE	max	nat.	trim	DCE	max
Released (fine-tuned XLS-R)												
XLS-R Conformer†	0.35	8.2	5.7	11.1	7.5	8.6	9.2	22.6	7.9	9.4	16.8	22.3
+ CL + DT†	0.30	6.8	6.6	14.0	6.4	8.5	11.0	28.7	7.3	10.0	19.2	24.7
SSL-AASIST†	0.30	5.7	7.8	14.0	11.4	13.8	13.9	28.9	7.4	17.3	7.1	9.7
Released (other front-ends)												
AASIST†	0.82	15.5	12.8	19.3	43.6	51.1	50.0	57.6	21.6	22.3	25.5	26.0
RawNet2†	4.8	20.9	23.1	34.2	49.2	47.8	54.3	55.8	48.4	68.0	45.9	52.3
Frozen XLS-R, boundary-trimmed random crops												
RAND (20)	4.5±0.9	4.1±0.3	3.8±0.3	6.7±0.3	13.4±1.0	12.0±1.1	11.3±1.1	20.4±2.1	22.9±2.8	15.9±2.0	22.1±2.4	29.5±2.7
RAND + DAS (5)	4.5±1.1	4.0±0.3	3.7±0.3	6.5±0.3	13.4±0.5	12.1±0.9	11.8±0.9	21.6±1.3	20.6±4.0	16.9±1.9	23.1±2.9	30.6±3.5
RAND + fixed-slope ALiBi + MLC (5)	3.4±0.4	3.8±0.3	3.5±0.2	6.6±0.2	13.9±1.0	12.5±0.6	11.8±0.7	20.1±0.5	18.7±2.6	15.4±2.1	20.2±1.9	27.8±2.0
RAND + MLC (5)	4.6±2.3	4.1±0.9	3.7±0.6	6.6±0.5	13.7±0.9	12.4±1.1	11.9±0.9	20.4±1.2	21.7±1.7	16.0±1.8	20.8±2.1	27.7±1.9
RAND + DAS + MLC (ours) (20)	3.9±0.8	3.8±0.4	3.6±0.4	6.5±0.4	14.1±0.8	12.7±0.8	11.9±0.9	20.0±1.1	19.7±3.0	15.8±2.1	20.3±2.2	27.5±2.4

6.1 RE-EXAMINING RELEASED DETECTORS

Silence and length matter for state-of-the-art checkpoints. We audit five released checkpoints across three front-end families — the two XLS-R Conformers of Kim et al. (2025), SSL-AASIST (Tak et al., 2022b), AASIST (Jung et al., 2022) and RawNet2 (Tak et al., 2021) — with port validation documented per model in Appendix B. Silence removal costs every one of them. Comparing `full` with `trim`, the same clip with and without its leading and trailing silence, LA19 EER rises from 0.20 to 8.21% (XLS-R Conformer), 0.28 to 6.76% (+ CL+DT), 0.20 to 5.73% (SSL-AASIST), 1.38 to 15.53% (AASIST) and 7.35 to 20.85% (RawNet2), for the removal of a median 0.30 s per clip, 10.9% of its duration: $+5.5$ to $+14.1$ EER points. All four systems reporting sub-1% EER on the standard input end between 5.7 and 15.5%, and at one second of matched speech the five score 11.1–34.2%. Only the three XLS-R-based systems transfer off-corpus — AASIST reaches 43.6% EER on ITW and RawNet2 49.2%, at or near chance — so the cross-corpus discussion concerns those three.

Rankings can change under duration control. On ITW with the standard input the three order as CL+DT (6.4) < XLS-R Conformer (7.5) < SSL-AASIST (11.4). That does not survive length matching. The curriculum and dynamic-temperature variant beats the base model by 1.1 EER points on natural clips (paired bootstrap 95% CI $[-1.6, -0.5]$), consistent with Kim et al. (2025); with silence removed the difference vanishes (-0.1, $[-0.7, 0.6]$); at one second it is significantly *worse*, by 6.1 points on ITW ($[5.1, 7.2]$) and 2.9 on LA19 ($[2.1, 3.6]$), and its ITW point estimate closes on SSL-AASIST (28.7 against 28.9), the model it had beaten by five points (Figure 3). An improvement measured on natural clips need not hold for speech of a given length.

6.2 CHANGING THE TRAINING VIEW

If the shortcut is in the data, the lever is the view a detector is trained on, and the place to pull it is a released checkpoint. We fine-tune the whole of the released XLS-R Conformer and SSL-AASIST on LA19 train for 1,500 steps (batch 16, Adam, learning rate $5 \cdot 10^{-6}$, their original two-way cross-entropy) in three arms that differ only in the view: the natural view they were released with (first 4.04 s, untrimmed), silence-free crops ($L \sim \mathcal{U}[1, 6]$ s of boundary-trimmed speech, one length per batch), and a per-batch 50/50 mixture of the two. The natural arm is the control — same clips, steps, optimiser and augmentation — so a difference between arms is the view, not the extra training. We keep the last step: no model selection, which would have to pick a view to select on.

Benchmark optimisation can favour the shortcut. More training on the natural view halves the Conformer's LA19 benchmark EER, 0.35 to 0.17%, and makes its length-matched In-the-Wild EER worse by 4.7 points ([+3.0, +7.4]; Table 2). Silence-free crops at the same budget do the reverse.

Table 2: Changing only the training view. EER (%); *nat.*: first 4.04 s of the untrimmed clip; *DCE*: mean over matched lengths; ASV5: ASVspoof 5 eval. *Silence transplant*: change in ASVspoof 5 EER when the boundary non-speech of real bona fide training clips is added to the spoofed clips only (*spoof*) or to every clip (*both*), with 4 s of speech held byte-identical (§5). A detector that reads boundary silence as evidence of bona fide speech rises in the first column and not the second. Mean±std over seeds (number in brackets). †Released checkpoints.

	Silence transplant		LA19 eval		In-the-Wild		FoR	ASV5 eval	
Model	spoof	both	nat.	DCE	nat.	DCE	DCE	nat.	DCE
Released checkpoint, then fine-tuned on LA19 train (1,500 steps, same budget per arm)									
XLS-R Conformer† (released)	+3.6	+1.2	0.35	5.7	7.5	9.2	16.8	5.1	7.5
fine-tuned, natural view (5)	+2.3	+0.9	0.17±0.05	5.2±1.4	9.8±1.6	13.9±2.6	15.3±4.9	5.4±0.68	7.9±1.1
fine-tuned, 50/50 mix (5)	−0.9	+0.2	0.98±0.50	4.9±1.2	9.3±0.70	9.1±0.79	16.4±4.6	3.6±0.18	5.5±0.33
fine-tuned, silence-free crops (5)	−0.8	+0.3	2.5±2.1	4.2±0.67	8.6±0.40	7.8±0.87	12.6±4.3	3.2±0.15	5.0±0.22
fine-tuned, crops with silence kept (5)	−0.4	+0.4	0.73±0.27	4.2±0.93	9.7±1.9	8.9±1.8	11.4±1.0	3.6±0.23	5.6±0.45
fine-tuned, CTI (silence-free crops + invariance) (3)	−0.2	+0.2	3.0±1.3	4.6±0.41	6.5±0.45	6.6±0.37	17.8±1.6	3.0±0.08	4.9±0.06
SSL-AASIST† (released)	+9.2	+3.7	0.30	7.8	11.4	13.9	7.1	6.7	9.7
fine-tuned, natural view (3)	+2.3	+2.1	0.23±0.05	6.5±0.79	13.8±0.38	15.8±0.93	6.7±0.91	7.9±0.09	10.5±0.13
fine-tuned, 50/50 mix (3)	−1.4	+1.1	4.6±0.25	7.5±0.29	10.7±0.79	10.3±1.2	13.6±2.0	5.4±0.05	7.5±0.14
fine-tuned, silence-free crops (3)	−0.9	+1.1	8.5±2.0	8.1±1.2	10.3±0.47	9.3±0.38	16.4±1.6	5.8±0.36	7.6±0.39
Frozen XLS-R back-end trained from scratch on									
LA19, natural view (3)	+1.4	−1.2	0.17±0.04	4.5±0.45	16.1±1.5	15.4±1.6	20.3±2.9	8.2±0.76	10.7±0.58
LA19, silence-free crops (20)	−0.7	+0.7	4.5±0.87	3.8±0.30	13.4±1.0	11.3±1.1	22.1±2.3	–	–
ASVspoof 5, natural view (6)	+0.8	−0.8	17.0±11.2	14.1±7.6	10.3±7.4	8.8±4.5	20.8±19.7	3.4±1.1	6.2±1.6
ASVspoof 5, silence-free crops (6)	+0.5	+0.2	17.9±3.7	9.4±0.51	7.9±2.6	5.3±0.90	8.7±2.3	2.4±0.64	4.5±0.89

Against that control they lower length-matched EER by 6.0 points on In-the-Wild ([−8.9, −4.1]) and by 2.9 on ASVspoof 5 ([−4.0, −2.2]), a corpus none of the arms has seen; against the released checkpoint itself, by 1.3 ([−2.1, −0.5]) and 2.5 ([−2.8, −2.2]). They also raise LA19 benchmark EER from 0.17 to 2.5%, so a practitioner choosing between the two on LA19 would ship the control. The mixture keeps LA19 at 0.98% while recovering most of the In-the-Wild difference (−4.8, [−7.6, −2.9]), which is the practical choice if the benchmark number must stay low. FoR, with 1–2 s of eligible speech and wide seed variation, does not separate the Conformer's arms. SSL-AASIST is a less clean case. Silence-free crops help it on In-the-Wild (−6.5, [−7.7, −5.3] against its control) and on ASVspoof 5 (−2.0 against the released model, [−2.5, −1.5]), but its LA19 benchmark EER rises to 8.5% and it loses the unusually good length-matched FoR result of the released checkpoint (7.1 to 16.4%): for this model the view trades corpora rather than improving all of them.

Which part of the view matters. The silence-free arm changes two things at once: it removes boundary silence, and it randomises the length and position of the crop. A fourth Conformer arm keeps the random crops but takes them from the untrimmed audio, so silence survives wherever a crop reaches it. It recovers most of the difference: against the control it lowers length-matched EER by 4.9 points on In-the-Wild ([−8.1, −2.5]) and 2.3 on ASVspoof 5 ([−3.5, −1.4]), at an LA19 benchmark EER of only 0.73%. Removing the silence on top adds 0.6 points on ASVspoof 5 ([−1.1, −0.2]) and a non-significant 1.1 on In-the-Wild ([−3.0, +0.5]), for 1.8 points of LA19 benchmark EER. What the control teaches is mostly the natural view itself — the first 4.04 s of every clip, its boundary silence always in the same place — and varying where the crop falls is most of the fix.

The transplant test shows the mechanism rather than the score. Planting bona fide boundary silence on spoofs raises the natural-view control's ASVspoof 5 EER by 2.3 points ([+1.5, +3.2]); after fine-tuning on silence-free crops the same intervention changes it by −0.8 ([−1.2, −0.4]), and on crops that keep the silence by −0.4 ([−0.8, +0.0]). Flip rates fall from 3.9% for the control to 0.2% and 0.6%. Neither model treats planted silence as evidence of bona fide speech any more; if anything they read it weakly the other way.

Training for the invariance directly. The transplant suggests an objective rather than an augmentation. Counterfactual timing invariance (CTI) pairs each silence-free crop $x = (s, \varnothing)$ with $x' = (s, z')$, where z' is 0.1–1 s of leading and trailing non-speech from a training clip of *either* class, so that z' carries no label information, and trains $\mathcal{L} = \frac{1}{2}[\ell_{\mathrm{CE}}(x) + \ell_{\mathrm{CE}}(x')] + \lambda\,\mathrm{KL}_{\mathrm{sym}}\big(p(\cdot \mid x), p(\cdot \mid x')\big)$, λ=1, from the same released weights and at the same budget as every other arm (Appendix C). Over 3 seeds it gives the best In-the-Wild numbers of any fine-tuning arm: 6.5% on the natural view and 6.6% length-matched, against 7.8% for silence-free crops alone (−1.2, [−2.0, −0.3]) and 9.2% for the released checkpoint (−2.5, [−3.0, −2.0]); on ASVspoof 5 it reaches 4.9%

against 5.0%. Where the view-only arms read planted silence weakly as *spoof* evidence, CTI barely responds to it (-0.2, $[-0.6, +0.3]$; flip rate 0.5%): invariance rather than a reversed cue. It still pays on the LA19 benchmark (3.0%), and on FoR it is worse than silence-free crops (17.8% against 12.6%, $[+0.5, +9.0]$), so it is not uniformly better.

Training on ASVspoof 5. The same contrast holds for a detector trained from scratch on ASVspoof 5, whose training split carries the stronger cue. With the frozen-front-end testbed and 6 seeds per view, silence-free crops instead of the natural view lower length-matched EER on ASVspoof 5's own evaluation set by 1.7 points ($[-3.2, -0.3]$), and off-corpus by 3.5 on In-the-Wild ($[-7.3, -0.3]$) and 12.1 on FoR ($[-30.1, -1.2]$). Trained on silence-free ASVspoof 5 crops, a 0.7 M-parameter back-end on a frozen front-end reaches 5.3% length-matched EER on In-the-Wild, below every released checkpoint in Table 1. ASVspoof 5 left its training partition unprocessed to reward detectors that avoid the cue; the natural view does not avoid it, and a silence-free view is a simple way to.

6.3 Can architectural duration awareness replace data control?

The architectural alternative is to tell the detector the duration and let it decide how to use it. We test the least shortcut-prone version we could build (Appendix O): a duration-adaptive span (DAS) in which the clip's trimmed duration d only sets the slope of an ALiBi-style distance penalty per head, $\mathrm{softmax}\big(\boldsymbol{Q}_h\boldsymbol{K}_h^\top/\sqrt{D/H} - m_h(d)\,|t_i - t_j|\big)$ with $m_h(d) = \mathrm{softplus}(\boldsymbol{w}_h^\top \boldsymbol{e}(d) + b_h)$, so it can change how local attention is but has no additive path to the classifier; and multi-length consistency (MLC), in which a long crop's stop-gradient prediction supervises a short crop of the same clip (Chen et al., 2020; Caron et al., 2021). Both are trained on silence-free crops, so the reference is RAND: the same crops with no duration mechanism.

Over twenty seeds they do not deliver. RAND reaches 11.3 ± 1.1 length-matched EER on In-the-Wild and 22.1 ± 2.4 on FoR, against 11.9 ± 0.9 and 20.3 ± 2.2 for DAS+MLC. Of twelve corpus–length cells, Benjamini–Hochberg corrected, four differ and in opposite directions: better on FoR at 1 s by 1.99 points (q=0.036), worse on In-the-Wild at 4, 6 and 8 s by 0.67–0.85 (all q=0.036); none of the ten LA19/ITW cells survives a speaker-cluster bootstrap (Appendix K). The fixed-slope ALiBi control, identical except that its slope cannot depend on d, matches it (20.2 ± 1.9 on FoR, 11.8 ± 0.7 on ITW, five seeds), and destroying its duration input moves no condition by more than 0.07 points (Appendix O). We therefore treat the DAS+MLC comparison as a negative result, not as evidence of an architectural improvement: within this testbed, the training view moves the length-matched numbers and the tested duration mechanism does not.

7 Conclusion

Timing predicts the label in all four corpora, in different ways, and ASVspoof 5 removes it only from its test partitions (Limitations: Appendix efapp:limits). Whether a detector *relies* on such a cue is a separate question that matched-length curves cannot answer; interventions that hold the speech fixed can, and show released detectors' decisions depending on planted silence. Optimising the standard benchmark strengthened that reliance; randomly placed training crops removed it, and an invariance objective built from the same intervention transferred best; a duration-aware architecture did not. For timing in audio deepfake detection, removing the cue from an evaluation set does not test whether a model learned it; interventions that preserve the speech evidence do — and reporting benchmark performance alone can select the less robust detector.

AI use statement

In the final stage of the project we used a generative AI writing assistant to help edit sections of the text, and produce tables and figures from our result files.

Every reported number is re-derived from the saved results by `check_claims.py`. The authors are responsible for the final manuscript, references and experimental claims, including all AI-assisted material.

Ethics statement

This work studies detectors of synthetic speech and aims to make their evaluation more reliable. We use only public corpora under their licences: ASVspoof 2019 and ASVspoof 5 (both ODC-By), In-the-Wild and FoR. In-the-Wild contains recordings of public figures, which we use for detector evaluation and a speaker-disjoint timing-classifier experiment and do not redistribute. We do not create new spoofed audio; the silence transplant only concatenates existing corpus audio for evaluation. Evaluating detectors under duration control could also reveal to attackers which utterance lengths are hardest to detect. We believe the benefit of more honest evaluation outweighs this risk, because the same information is needed to deploy detectors responsibly.

Reproducibility statement

Section 4 and Appendix B describe the data processing, the evaluation lists and the training settings. The anonymous supplementary material contains the code for data preparation, timing statistics, training, matched-length and shared-cohort evaluation, the boundary intervention and the silence transplant, the fine-tuning of released checkpoints, the ports of the five released checkpoints and the counterfactual tests, together with scripts that regenerate every table and figure. It ships 4,238 per-clip score files, one per model and condition behind every reported number, and the cached index files, so the analysis end of the pipeline runs without the corpora and without a GPU: `audit/verify_artifacts.py` recomputes every reported EER from those per-clip scores and checks them against the saved summaries (to stay under the size limit, scores of unreported legacy runs and four index files are omitted and listed in the package README), and `audit/test_regressions.py` runs the regression suite. `check_claims.py` compares manually registered numeric expectations against saved summaries; it does not parse or certify the manuscript, and we do not present it as doing so. What the package does *not* contain is the corpora themselves, the trained back-end checkpoints and the released third-party weights, which must be obtained from their original sources; `data/README.md` records the versions and expected layout. Model selection, training and the released-checkpoint inference therefore cannot be repeated from the package alone. All experiments use public data. Each run used one A100 80 GB. Frozen-front-end training peaks at 5.1 GiB for the measured configuration; fine-tuning a released checkpoint end to end needs far more (we budgeted 24 GB per run), and other GPUs have not been validated. Results are statistically rather than bitwise reproducible: we fix seeds but do not enforce deterministic CUDA kernels, and inference batch shape perturbs scores at the fourth decimal (§B).

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech*, 2022.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Nuria Alina Chandra, Hannah Lee, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Changyeon Lee, Jongwook Choi, et al. Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025.

Wenda Chen, Jonathan Huang, and Tobias Bocklet. Length- and noise-aware training techniques for short-utterance speaker recognition. In *Interspeech*, 2020. arXiv:2008.12218.

Bhusan Chettri, Emmanouil Benetos, and Bob L. T. Sturm. Dataset artefacts in anti-spoofing systems: A case study on the ASVspoof 2017 benchmark. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3018–3028, 2020. doi: 10.1109/TASLP.2020.3036777.

Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, 2017.

Yassine El Kheir, Younes Samih, Suraj Maharjan, Tim Polzehl, and Sebastian Möller. Comprehensive layer-wise analysis of SSL models for audio deepfake detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4070–4082, 2025.

Joel Frank and Lea Schönherr. WaveFake: A data set to facilitate audio deepfake detection. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*, 2020.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *NAACL-HLT*, 2018.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, 2022.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proc. ICASSP*, 2022.

Taewoo Kim, Guisik Kim, Choongsang Cho, and Young Han Lee. Naturalness-aware curriculum learning with dynamic temperature for speech deepfake detection. In *Proc. Interspeech*, 2025.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Xuechen Liu, Wanying Ge, Xin Wang, and Junichi Yamagishi. LENS-DF: Deepfake detection and temporal localization for long-form noisy speech. In *Proc. IEEE International Joint Conference on Biometrics (IJCB)*, 2025.

Juan M. Martín-Doñas, Eros Roselló, Angel M. Gomez, Aitor Álvarez, Iván López-Espejo, and Antonio M. Peinado. ASASVIcomtech: The vicomtech-ugr speech deepfake detection and SASV systems for the ASVspoof 5 challenge. In *ASVspoof Workshop*, 2024. arXiv:2408.10361.

Nicolas M. Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. Speech is silver, silence is golden: What do ASVspoof-trained models really learn? In *Proc. ASVspoof 2021 Workshop*, 2021.

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

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Ricardo Reimao and Vassilios Tzerpos. FoR: A dataset for synthetic speech detection. In *Proc. International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–10, 2019.

Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, and Antonio M. Peinado. A conformer-based classifier for variable-length utterance processing in anti-spoofing. In *Proc. Interspeech*, 2023.

Santiago Rubio, Pilar Bello, Dayana Ribas, Antonio Miguel, Eduardo Lleida, and Alfonso Ortega. An intervention-based framework for shortcut diagnosis in spoofing countermeasures. *arXiv preprint arXiv:2607.03150*, 2026.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

K. A. Shahriar. Lightweight resolution-aware audio deepfake detection via cross-scale attention and consistency learning. *arXiv preprint arXiv:2601.06560*, 2026.

Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18815–18825, 2025. doi: 10.1109/CVPR52734.2025.01753.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020.

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 331–335, 2019.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with RawNet2. In *Proc. ICASSP*, 2021.

Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *Proc. ICASSP*, 2022a.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *Proc. Speaker and Language Recognition Workshop (Odyssey)*, 2022b.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. In *Proc. Interspeech*, 2024.

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

Xin Wang, Héctor Delgado, Nicholas Evans, Xuechen Liu, et al. ASVspoof 5: Evaluation of spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2354–2367, 2026a. doi: 10.1109/TASLPRO.2026.3682962.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, 95:101825, 2026b. doi: 10.1016/j.csl.2025.101825.

Yang Xiao and Rohan Kumar Das. XLSR-Mamba: A dual-column bidirectional state space model for spoofing attack detection. *IEEE Signal Processing Letters*, 2025.

Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised XLS-R and SLS classifier. In *Proc. ACM International Conference on Multimedia*, 2024a.

Yuxiang Zhang, Zhuo Li, Jingze Lu, Hua Hua, Wenchao Wang, and Pengyuan Zhang. The impact of silence on speech anti-spoofing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3374–3389, 2023. doi: 10.1109/TASLP.2023.3306711.

Yuxiang Zhang, Jingze Lu, Zengqiang Shang, Wenchao Wang, and Pengyuan Zhang. Improving short utterance anti-spoofing with AASIST2. In *Proc. ICASSP*, 2024b.

A Limitations

The fine-tuning experiment uses one budget, one learning rate and five seeds (three for SSL-AASIST and for CTI, whose λ we did not tune), with no model selection; the mixture arm is one point on the trade-off, not its frontier. The transplant plants one kind of cue, boundary silence taken from bona fide training clips, and its size depends on that donor distribution. The shared-cohort check is low-powered (478 balanced LA19 clips at 4 s), the boundary intervention is undefined on pre-trimmed FoR, and FoR's 1–2 s of eligible speech is too little to separate the fine-tuning arms. Clip-level intervals ignore shared speakers: clustering leaves no LA19/ITW cell of the duration-conditioning comparison significant, and FoR lacks speaker labels. Our ASVspoof 5-trained detectors use a frozen front-end. CTI is tested on one architecture, the XLS-R Conformer, with three seeds; whether it transfers to other detectors is untested.

B Implementation details

Data preparation. All audio is decoded once to 16 kHz mono 16-bit PCM and stored in a memory-mapped file together with an index of labels, speakers, attacks and silence boundaries.

- **ASVspoof 2019 LA.** We use a public Parquet copy of the official release. Its eval labels, attack IDs and speaker IDs match the official CM protocol for all 71,237 trials.
- **In-the-Wild.** We read the official release, including its `meta.csv`.
- **FoR.** We use the `for-norm` version. Two training files could not be decoded and were skipped.

Silence boundaries are computed as described in §4: frame RMS over 25 ms frames with a 10 ms hop, with a threshold of 40 dB below the loudest frame.

Detectors trained in this work.

- **Front-end.** The first 12 transformer layers of `facebook/wav2vec2-xls-r-300m`, frozen and run in half precision. Its 13 hidden states are mixed with learned softmax weights.

- **Back-end.** A linear projection to 144 dimensions with LayerNorm and GELU, followed by two pre-norm transformer blocks (4 heads, feed-forward size 576, dropout 0.1), attentive statistics pooling and a two-layer MLP. The DAS slopes of Eq. 1 are computed per clip and per head, with distances measured in seconds at 50 frames per second.
- **Optimisation.** AdamW (learning rate $3{\cdot}10^{-4}$, weight decay 10^{-2}) with 5% linear warm-up and cosine decay, batch size 32 and 8 epochs with 25,380 class-weighted draws each (sampling with replacement, with the final incomplete batch dropped). Gradients are clipped at norm 1, and we use mixed precision.
- **Augmentation.** A random gain of ± 6 dB, plus white noise at an SNR between 15 and 40 dB with probability 0.5. No RawBoost.
- **Model selection.** We keep the checkpoint with the lowest EER on a balanced LA19 dev subset (all 2,548 bona fide clips and 2,548 random spoofed clips). The dev view is `fixed`, `full` or whole `trim` audio for the corresponding training regime; crop-trained models are selected on whole trimmed clips. Ties retain the latest epoch.

One run takes 0.3–2.3 h on a single A100 80 GB, with a median of 1.0 h over the 96 runs reported here. Peak memory for the heaviest configuration (MLC, two views, batch 32, a 6 s long view) is 5.1 GiB, for the measured configuration; portability to a 12 GB card is not verified.

Released checkpoints. We evaluate the LA models of Kim et al. (2025) (`XLSR_Conformer` and `XLSR_Conformer_CL_DT`, file `best_model_avg.pth`) without fairseq.

- **Porting.** The fine-tuned XLS-R weights are mapped onto `transformers.Wav2Vec2Model`, and the Conformer back-end is the original code (`conformer==0.3.2`).
- **Check.** On 600 random ITW clips per model, our scores correlate with the authors' released score files at Pearson $r > 0.9999$, agree in sign on $\geq$99.8% of clips, and differ by 0.013–0.016 on average. From the released score files, we recompute ITW EERs of 7.29% and 6.60%, as reported by the authors.
- **Batching.** These models have no padding mask, so variable-length views (`full`, `trim`) are scored one clip at a time.

We port three further released checkpoints the same way, and document architecture, score or benchmark checks as available. Only the AASIST check below reproduces its cited EER on the full LA19 evaluation protocol.

- **AASIST** (Jung et al., 2022), weights `AASIST.pth` from the authors' repository. Loading is strict, with no missing or unexpected keys. On the full LA19 eval protocol (7,355 bona fide, 63,882 spoofed) our port gives 0.830% EER against the published 0.83%.
- **RawNet2** (Tak et al., 2021), the ASVspoof 2021 baseline weights. The LA and DF releases contain byte-identical checkpoints (both md5 `863a3982...`), so there is only one pre-trained RawNet2; both baselines are trained on the ASVspoof 2019 LA train partition. The released `SincConv` keeps its Mel filter bank as a plain attribute rather than a buffer, so it never enters the checkpoint; we rebuild it with the original code and confirm the reconstruction is bit-identical, after which loading is strict. A full forward pass agrees with the released implementation to 0.0 absolute error, and on a balanced 1,500-clip-per-class subsample the port gives 4.20% EER against the published 4.66% on the full eval set.
- **SSL-AASIST** (Tak et al., 2022b), the XLS-R front-end with an AASIST back-end. As with Kim et al. (2025), the fairseq XLS-R weights are remapped onto `transformers.Wav2Vec2Model` so that fairseq is not required; 422 of 429 front-end keys map, and the seven that do not are the pre-training-only quantiser and projection heads, which the inference path never uses. The back-end is the authors' own code, loaded strictly. Unlike the Kim et al. (2025) port we cannot check this one against a released score file, because the scores distributed with it are for ASVspoof 2021; the evidence that the remapping is faithful is that it obtains 0.30% EER on a balanced 4,000-per-class LA19 subset (not the full official protocol) and that the same mapping is verified to $r > 0.9999$ on the Kim et al. (2025) checkpoints.

For every port we also check the sign of the score on held-out data: reversing it would move LA19 EER to 95–99%, so the bona fide direction is not in doubt.

DADAT. DADAT uses a 128-band log-mel spectrogram (FFT size 1024, hop 256, audio zero-padded to 15 s, dB relative to the maximum with an 80 dB floor). Its encoder has 4 transformer layers with d=256, 8 heads, a feed-forward size of 1024 and dropout 0.3. The attention output of every layer is multiplied by $\sigma(\boldsymbol{W}\boldsymbol{e}(d))$, a padding mask is derived from the duration, and the output is pooled with average and max pooling. We use the weights after 50 epochs. For the counterfactual test we recompute the features exactly as in its preprocessing and evaluate on balanced random subsets (1,500 clips per class for LA19 eval, 2,264 for FoR test).

Evaluation subsets. The `fixed`, `full` and `trim` views use one balanced subset per corpus (fixed seed): 4,000 clips per class for LA19 eval and ITW, and all 2,264 bona fide FoR test clips with 2,264 random spoofed ones. For each DCE length, the balanced subset is drawn from the clips that have enough trimmed speech (Appendix E). EER is computed at every distinct score threshold, which makes it robust to ties.

C ASVspoof 5, the silence transplant and fine-tuning

ASVspoof 5 data. We use track 1 of the official release (Zenodo record 14498691, ODC-By 1.0), with every archive checked against the published MD5 sums. The train partition has 18,797 bona fide and 163,560 spoofed clips and the dev partition 10,700 and 36,700. For evaluation we use the eval trials without codec processing (codec label "-", 7,031 bona fide and 26,732 spoofed); the codec-processed trials enter only the timing statistics. Audio is decoded with `soundfile` and silence boundaries are computed exactly as for the other corpora (§4). Evaluation subsets are balanced at 4,000 clips per class, as for LA19 and ITW; the 8 s condition keeps 2,384 per class.

Timing statistics. The training-partition figures in §3 use a speaker-disjoint 50/50 split of the train partition (grouped by speaker, fixed seed) and the same gradient-boosted classifier and feature sets as Table 8. The dev and eval figures train on the whole train partition. Medians: bona fide / spoofed duration 14.9 / 11.3 s on train and 7.1 / 7.1 s on eval; leading non-speech 0.25 / 0.04 s on train (KS 0.49) and 0.01 / 0.01 s on eval (KS 0.04).

Detectors trained on ASVspoof 5. The frozen-front-end testbed of Appendix O without a duration input, with the optimiser, augmentation and budget of Appendix B (8 epochs of 793 class-balanced batches of 32), trained on the train partition and selected on a balanced dev subset of 2,548 clips per class, scored on the `fixed` view for the natural-view model and on whole trimmed clips for the crop-trained one. Seeds 0–5 for each view.

Silence transplant. From the uncoded eval trials with at least 4 s of trimmed speech (6,894 bona fide, 25,851 spoofed) we draw a balanced subset of 2,000 per class and take the central 4 s. Donors are the 11,902 bona fide training clips whose leading and trailing non-speech are both at least 50 ms; for each eval clip one donor is drawn by a generator seeded with the clip's index, so every model sees identical audio. The transplanted non-speech has a median of 0.36 s leading and 0.43 s trailing. Every view is scored at its full length, one clip at a time. Intervals in Table 2 and the text are the two-level paired bootstrap of §4, applied to the difference between the planted and control changes.

Digital-silence control. The same clips and donors, with each donor's leading and trailing segments replaced by zeros of the same length, so only the amount and position of silence change (`silence_transplant.py --zero`). Net effect (spoof-only minus every-clip change in EER, 95% interval): released XLS-R Conformer +2.2 [+1.6, +2.8], + CL + DT +4.2 [+3.3, +5.1], SSL-AASIST +4.7 [+3.9, +5.6]; fine-tuned natural view +1.6 [+0.8, +2.5], silence-free crops −1.0 [−1.4, −0.7], CTI −0.2 [−0.4, +0.1] (seeds 0–2 for the fine-tuned models).

Fine-tuning. The whole model is trained, starting from the released weights: Adam (learning rate $5 \cdot 10^{-6}$, weight decay 10^{-4}), 75 steps of linear warm-up and then a constant rate, 1,500 steps of 16 class-balanced clips drawn with replacement (about one pass over LA19 train), gradient clipping at

norm 1, half-precision autocast, and the augmentation of Appendix B. The loss is cross-entropy on the released two-way output layer. Neither model has a padding mask, so each batch has one length: silence-free crops shorter than L are repeat-padded from their own trimmed speech, so no silence is reintroduced, and the rare clip with less than 10 ms of trimmed speech falls back to its untrimmed audio. LA19 dev EER is logged every 500 steps for monitoring only; the final step is evaluated. A run takes 8–43 minutes on one A100 80 GB. Released and fine-tuned models are scored with the same code and on the same clips.

Counterfactual timing invariance. Same optimiser, budget, augmentation and batch of 16 clips as the other arms, but each clip is presented twice: a silence-free crop of $L \sim \mathcal{U}[1, 6]$ s, and the same samples with $P \sim \mathcal{U}[0.1, 1]$ s (one draw per batch) of the leading and trailing non-speech of a random LA19 training clip, tiled to length P, added at the ends. Donors are the training clips of both classes with at least 50 ms of non-speech at each end. Gain and noise augmentation are drawn independently for the two views. The consistency term is the symmetric KL divergence between the two posteriors, each side with a stop-gradient on its target. The ASVspoof 5 transplant uses ASVspoof 5 donors, so the test context is never one seen in training. Seeds 0–2.

D STATISTICAL PROCEDURE

For released checkpoints we report `full`-to-`trim` changes on the same LA19 clips and clip-paired differences within conditions. For DAS+MLC versus RAND the family is every eligible corpus–length cell — four on LA19, six on ITW, two on FoR — with Benjamini–Hochberg correction across all twelve. This is retrospective; no preregistration is claimed. The bootstrap pairs clips across methods and resamples training seeds independently within each method, so its intervals concern mean EER across training runs and assume independent clips. Speaker and source dependence is a real limitation: where identity is available we repeat the analysis clustered by speaker and report it as a sensitivity family (Appendix K), and FoR carries neither speaker nor generator identity, so it cannot be clustered at all. Five-seed ablations and FoR-variant comparisons are exploratory.

For duration-conditioned models a counterfactual test keeps the `trim` waveform fixed and supplies a constant duration (1.5, 3 or 8 s) or permutes durations across clips; we report EER and decision flips at logit zero. This measures sensitivity to the explicit scalar, not invariance to timing cues encoded in the waveform.

E NUMBER OF CLIPS PER DURATION CONDITION

Table 3: Clips available for each matched-length condition: bona fide / spoof clips with at least L seconds of speech after silence trimming, and in brackets the number kept per class after balancing (capped at 4,000 for LA19 eval and In-the-Wild, 2,264 for FoR test). A condition is skipped when fewer than 100 clips per class remain.

	matched length L (s)					
Set	1	2	3	4	6	8
LA19 eval	7,081 / 61,471 (4,000)	3,645 / 38,819 (3,645)	1,076 / 19,889 (1,076)	239 / 9,879 (239)	–	–
In-the-Wild	18,963 / 11,302 (4,000)	14,081 / 9,220 (4,000)	9,253 / 7,110 (4,000)	5,885 / 5,506 (4,000)	2,550 / 3,277 (2,550)	1,255 / 1,948 (1,255)
FoR test	2,242 / 2,095 (2,095)	1,677 / 539 (539)	–	–	–	–

F LABEL INFORMATION IN THE TRAINING VIEWS

Table 4 reports how well the *duration of a training view* alone separates the classes on LA19 train. We report the direction-free AUC, $\max(\mathrm{AUC}, 1 - \mathrm{AUC})$.

- **Natural views.** Leading silence is informative (AUC 65.3), but clip duration is only weakly informative (AUC 52.1).
- **Silence-free random crops.** The view duration is at chance level (AUC 51.2 for RAND and 50.5 for the short MLC view).

Table 4: Direction-free AUC of a single timing quantity of the training view for bona fide vs. spoof (LA19 train; 50 = no information).

Training view: quantity	AUC
natural, whole clip: clip duration	52.1
natural, whole clip: leading silence	65.3
natural, first 4.04 s: content duration	53.0
trimmed, whole clip: speech duration	53.5
RAND crop ($L \sim \mathcal{U}[1, 6]$ s)	51.2
MLC short view ($L_s \sim \mathcal{U}[1, 3]$ s)	50.5
MLC long view ($\leq$6 s)	53.5

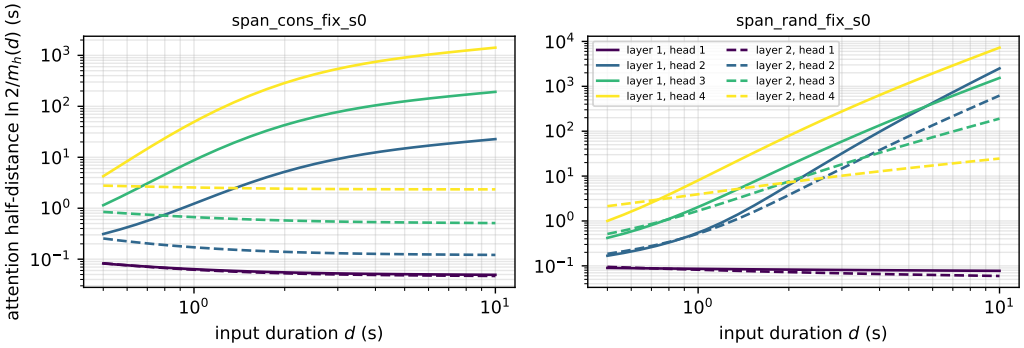

Figure 1: Learned attention half-distance $\ln 2/m_h(d)$ against the duration input, per layer and head, for the two seed-0 models that use it. Training moves the spans away from their ALiBi initialisation, but by very different amounts across heads.

G Learned attention spans

Figure 1 plots the learned slopes $m_h(d)$ as attention half-distances $\ln 2/m_h(d)$, for the seed-0 model of each variant that uses them.

DAS is not inert. $\boldsymbol{W}$ starts at zero, so training begins from fixed ALiBi, and it moves away in both directions at once. In DAS+MLC the two sharpest heads get *steeper* as the clip lengthens ($11.1 \rightarrow 13.9$ and $11.4 \rightarrow 14.7\,\mathrm{s}^{-1}$ between 1 s and 8 s of input, half-distances of two to three frames throughout), while one head flattens from 0.50 to $0.035\,\mathrm{s}^{-1}$, taking its half-distance from 1.4 s — local inside the clip — to 20 s, which over an 8 s input is no decay at all. Counting a head as duration-dependent when $m_h(d)$ varies by more than $1\,\mathrm{s}^{-1}$ over $d \in [0.5, 10]$ s, four of eight heads qualify in DAS+MLC and six of eight in DAS alone; per-head variation runs from 0.05 to $6.48\,\mathrm{s}^{-1}$ and the learned slopes themselves span 0.0001–$14.8\,\mathrm{s}^{-1}$ across the two models.

We avoid summarising this as a ratio of half-distances, which is what the figure plots: a head whose slope approaches zero has an arbitrarily large half-distance while doing nothing at all, and on that scale the most extreme head changes by a factor of $1{,}700$ without the attention pattern changing in any way that matters. The slopes are the honest units. What they show is that the mechanism is active in most heads, unevenly, and largely in heads that were already near the ends of the ALiBi initialisation.

H Counterfactual duration inputs for DADAT

We examine DADAT, a duration-gated mel-spectrogram transformer (3.7 M parameters) built for this study, as a probe of explicit duration conditioning. It gates every attention output with an embedding of the clip duration, and was trained on a duration-stratified mixture of the four FoR variants and LA19 train. Holding the spectrogram fixed and changing the duration input (which changes both gating and the attention mask) (Table 5): permuting the durations across clips raises

Table 5: Duration-input interventions for DADAT. EER (%), decisions flipped at logit zero (%), and share labelled spoof (%). The duration input changes both gating and the attention mask; this is not a gate-only intervention. The last two rows replace all clip-specific features with constants. With per-clip maximum normalization, digital silence maps to 0 dB, so -80 dB is a constant floor-feature probe rather than the preprocessing of silent audio. Balanced subsets: 1,500 (LA19) and 2,264 (FoR) clips per class.

	LA19 eval			FoR test		
Duration input	EER	flips	spoof	EER	flips	spoof
true duration	14.5	–	33.6	15.1	–	33.8
permuted across clips	34.1	34.5	62.3	18.1	13.5	40.0
constant 1.5 s	16.1	12.6	33.5	18.0	12.3	42.3
constant 3 s	31.5	30.7	58.8	18.0	13.3	36.3
constant 8 s	43.5	66.3	98.4	27.6	22.3	25.0
constant -80 dB features, true duration	42.5	41.2	42.4	28.5	50.3	81.1
constant 0 dB features, true duration	46.0	33.6	0.0	54.1	33.8	0.0

LA19 EER from 14.5% to 34.1% and flips 34.5% of decisions, and telling the model every clip is 8 s long makes it call 98.4% of clips spoofed. FoR withstands the permutation better, 15.1% to 18.1%, but not the constant: 8 s costs it 12.5 points, to 27.6%. Replacing the spectrogram with a constant -80 dB feature map while keeping the true duration still leaves 28.5% EER on FoR and 42.5% on LA19: on FoR the duration input alone ranks clips well above chance with no clip-specific acoustic features. This intervention tests the combined duration path, not the gate in isolation. It is also, we found during this audit, not only the duration path: the probe's global pooling ran over every time step including padded ones, which were masked as attention keys but still acted as queries, so the amount of padding reached the classifier as a second duration channel. We report the probe as trained, and correct the pooling in the released code for anyone retraining it. We treat this model as a legacy probe throughout and place no weight on it in the main text. The model has clearly learned that the duration input predicts the label. This is harmful whenever the relation between duration and label changes, as it does between LA19, FoR and ITW.

I SHORTCUT REMOVAL VERSUS SHORTER AUDIO

The FoR variants give a descriptive control. Trimmed duration has direction-free AUC 51.9 on `for-2sec`, where every clip is truncated to two seconds during corpus construction, against 81.8 on `for-norm` and 83.7 on `for-original`. On `for-2sec` the L=1 s condition still costs the silence-free-crop baseline $+11.6$ EER points, against $+13.6$ on `for-norm` (Appendix I). Poor short-crop performance therefore need not imply a strong monotone duration cue; the two-point difference is descriptive, since the variants differ in preprocessing and composition and are not randomised interventions on a common population.

A caveat on how these deltas are measured. The `trim` view is scored on one balanced subset per corpus, while each `crop_L` condition is balanced separately among the clips long enough to supply L seconds of speech, so the two are not drawn from the same clips. On the FoR variants this barely matters — `for-2sec` and `for-rerecorded` share 100% of their clips between the two views and `for-norm` and `for-original` 97.7% and 97.8% — and restricting every delta below to the shared clips moves it by at most 0.27 EER points, with the `for-2sec` and `for-norm` figures the argument rests on moving by 0.00 and -0.09. The same comparison would not be safe on LA19 or In-the-Wild, where only 30% and 28% of clips are shared, which is why we make it only on the FoR family. The `full`-against- `trim` comparison in §6.1 is unaffected: both views are scored on the same subset. The rows also rest on different seed counts, twenty for `for-norm` and five for the other variants, since only `for-norm` is part of the headline comparison. Restricting `for-norm` to the same five seeds moves its figure from $+13.6$ to $+13.3$, so the gap against the chance-cue `for-2sec` baseline is 1.7 rather than 2.0 EER points and the conclusion is unchanged.

Table 6: Separating shortcut removal from shorter audio. For each FoR variant: the direction-free AUC of a classifier that sees only the trimmed speech duration, the KS distance between the class duration distributions, and the change in EER from the `trim` view to the L=1 s condition. `for-2sec` truncates every clip to two seconds during corpus construction, so its monotone duration score is near chance. Differences between variants are descriptive and do not identify a causal shortcut contribution.

	duration cue		ΔEER, `trim` → 1 s			
Variant	AUC	KS	RAND	DAS+MLC	Conformer	+CL+DT
`for-2sec`	51.9	0.04	+11.6	+10.7	+12.5	+9.9
`for-rerecorded`	57.6	0.16	+2.9	+3.3	+16.8	+12.5
`for-norm`	81.8	0.51	+13.6	+11.7	+12.9	+14.7
`for-original`	83.7	0.55	+17.0	+16.9	+17.3	+17.8

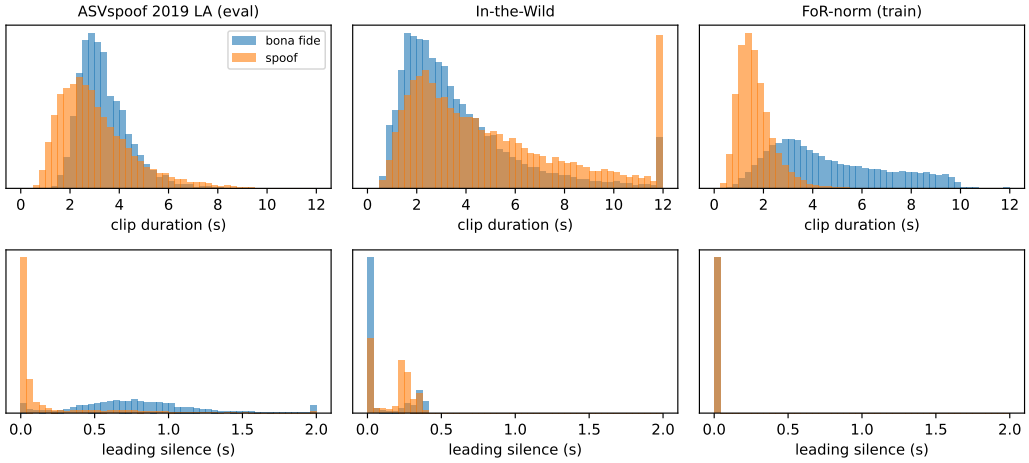

Figure 2: Per-class distributions of clip duration (top, clipped at 12 s) and leading silence (bottom, clipped at 2 s). The dominant timing cue differs between corpora and can change direction.

J Timing distributions

K Speaker-cluster sensitivity analysis

We reran the headline comparison on the ten LA19/ITW conditions using 2,000 bootstrap replicates that sample speaker identifiers with replacement and retain all selected clips from each sampled speaker. Seeds are independently resampled within each method. This sensitivity analysis uses the available metadata (54 ITW speaker identifiers; FoR has no supplied speaker IDs) and is exploratory. It does not account for all possible recording-source or attack clusters. None of the ten conditions survives BH correction within this sensitivity family; the originally reported clip-level significance should therefore not be interpreted as speaker-population evidence. Point estimates are unchanged.

L Timing distributions and the Temporal Shortcut Index

Table 8 gives the per-class timing distributions and the held-out performance of classifiers that see only timing features. Table 9 gives the component-wise Temporal Shortcut Index discussed in §3, including the *survives trimming* row.

Table 7: Speaker-cluster bootstrap of DAS+MLC minus RAND EER (percentage points).

Condition	Difference	95% interval	q (10 cells)
LA19 1 s	-0.19	[-0.45, +0.07]	0.210
LA19 2 s	-0.15	[-0.37, +0.07]	0.259
LA19 3 s	-0.17	[-0.49, +0.17]	0.314
LA19 4 s	-0.36	[-0.95, +0.13]	0.230
ITW 1 s	-0.39	[-1.77, +0.73]	0.464
ITW 2 s	+0.65	[-0.21, +1.45]	0.210
ITW 3 s	+0.75	[-0.02, +1.65]	0.145
ITW 4 s	+0.85	[+0.08, +1.92]	0.145
ITW 6 s	+0.71	[+0.04, +1.89]	0.145
ITW 8 s	+0.67	[-0.00, +1.73]	0.145

Table 8: Timing statistics per class (bona fide / spoof). (a) Median clip duration and leading silence (s), share of clips $\geq$7 s (%) and Kolmogorov–Smirnov distance between the class distributions. (b) AUC / EER (%) of a gradient-boosted classifier that sees only timing features; ITW (A/B) is a speaker-disjoint split.

(a) Distributions

Set	duration	lead. silence	$\geq$7 s	$\mathrm{KS}_{\mathrm{dur}}$	$\mathrm{KS}_{\mathrm{lead}}$
LA19 eval	3.20 / 2.75	0.75 / 0.04	0.9 / 2.6	0.25	0.72
ITW	2.99 / 4.05	0.00 / 0.23	9.5 / 23.1	0.19	0.36
FoR train	4.12 / 1.54	0.00 / 0.00	19.6 / 0.0	0.70	0.00

(b) Timing-only classifiers

	duration		silence		all timing	
Train $\rightarrow$ test	AUC	EER	AUC	EER	AUC	EER
LA19 train $\rightarrow$ LA19 eval	79.3	28.0	80.1	24.0	83.3	23.6
ITW (A) $\rightarrow$ ITW (B)	58.2	44.8	77.1	28.0	78.0	28.8
FoR train $\rightarrow$ FoR test	81.6	24.3	49.5	50.8	81.6	24.1
LA19 train $\rightarrow$ ITW	49.7	48.6	65.7	39.4	59.6	43.4
LA19 train $\rightarrow$ FoR test	69.6	36.7	49.3	50.1	75.0	28.6
FoR train $\rightarrow$ LA19 eval	61.4	42.2	52.1	46.6	48.8	51.0

M ATTACK COMPOSITION, SHARED COHORTS AND THE DURATION-METADATA CONTROL

These tables carry the detail behind §5; the quantities quoted there are taken from them. Table 11 gives the attack-stratified EERs on LA19 and the spoof-mix drift across crop lengths. Table 12 gives the shared-cohort duration curves for the five released checkpoints, beside the per-length-cohort curves of Table 1. Table 13 gives the per-condition effect of removing the source-duration input from the duration-conditioned models, and Table 10 the boundary intervention at fixed speech content.

N DURATION-CONDITIONED DETECTORS TRAINED ON NATURAL CLIPS

Telling a detector the duration costs it cross-corpus transfer, though the training view matters more. All four variants score 0.14–0.17% EER on the standard LA19 input over three seeds, so the benchmark number reveals nothing; off-corpus they separate, CONCAT reaching $26.1 \pm 5.6\%$ on FoR against $19.3 \pm 6.3\%$ without the duration input and GATE $22.7 \pm 3.6\%$ — 3 to 7 EER points of transfer given up for an input that is free in domain. Under duration control, however, the dominant failure is the untrimmed training clips rather than the duration input: GATE has the worst matched-length EER of any detector we train on LA19 and ITW (11.3, 19.6), but with seed spreads of ± 8.5 and ± 10.7 it cannot be told apart from the unconditioned full-clip baseline, whose own worst cell ($42.1 \pm 17.1\%$, ITW at L=1 s) is the worst we produce. Both train on whole untrimmed clips and

Table 9: Temporal Shortcut Index. A gradient-boosted classifier sees only the named timing quantities – never the speech – and is scored on held-out data; ITW uses a speaker-disjoint split. TSI rescales its AUC by that of a released acoustic detector on the same clips, TSI $= (A_{\text{temporal}} - 50)/(A_{\text{audio}} - 50)$, so 1 means timing alone matches the detector and 0 means timing carries nothing. *Survives trimming* is the subset of quantities that boundary trimming leaves intact; on LA19 it retains almost all of the shortcut, so trimming is a much weaker control than it appears. Reference detector AUC — LA19 eval: 100.0 In-the-Wild: 98.1 FoR test: 98.2.

	LA19 eval		In-the-Wild		FoR test	
Timing quantities available	AUC	TSI	AUC	TSI	AUC	TSI
total duration T_{total}	79.3	0.58	58.2	0.17	81.6	0.66
boundary silence $T_{\text{lead}}, T_{\text{trail}}$	80.1	0.60	77.1	0.56	49.5	-0.01
internal pauses $T_{\text{int}}, N_{\text{pause}}$	64.2	0.28	54.3	0.09	59.9	0.21
speech amount $T_{\text{speech}}, R_{\text{speech}}$	87.3	0.75	64.5	0.30	81.5	0.65
all temporal	88.3	0.77	77.2	0.56	81.5	0.65
survives trimming	87.7	0.75	68.3	0.38	81.5	0.65

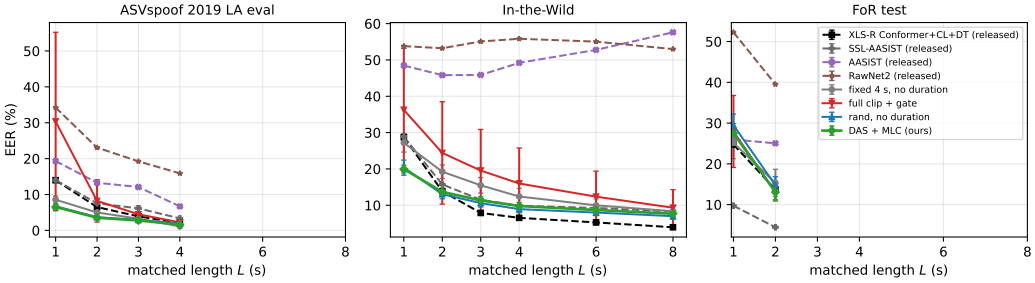

Figure 3: EER as a function of the matched speech length L under matched-length evaluation. Each condition uses L seconds of boundary-trimmed audio. Internal pauses remain, and eligible clips differ across lengths. The released checkpoint (black, dashed) is the strongest system at the standard input but among the weakest at L=1 s, and its ordering relative to our frozen-front-end baselines reverses between the two ends of the x-axis. Error bars are ± 1 standard deviation over seeds.

both are unstable under length matching, whether or not the model is told the length. Their counterfactual flip rates stay an order of magnitude below the legacy probe's — 0.85% for CONCAT and 2.48% for GATE against 12–66% — so a detector on a strong self-supervised front-end does not lean on the scalar the way a mel-spectrogram model does. What it loses is transfer.

O THE DURATION-CONDITIONED PROBE

To ask whether a detector can be told the duration without being led by it, we need detectors that receive it. We build three, on a common frozen backbone, so that the only difference between them is how duration enters. This is a controlled probe, not a bid for state of the art; the backbone, optimiser and augmentation are held fixed throughout and are given in Appendix B.

Backbone. The first K=12 layers of a frozen XLS-R 300M (Babu et al., 2022) encode the waveform; hidden states are combined with learned softmax weights, projected to D=144, and passed through two four-head pre-norm transformer blocks, then attentive statistics pooling and an MLP. About 0.7 M parameters are trained. For crop-trained models d is the source clip's boundary-trimmed duration, not the crop length — a leak we remove below — and $\boldsymbol{e}(d) = \text{MLP}(\log d) \in \mathbb{R}^D$ is its embedding.

Three ways to condition. CONCAT appends $\boldsymbol{e}(d)$ to the pooled vector: an explicit shortcut, included as a reference point. GATE multiplies every attention output by $\sigma(\boldsymbol{W}\boldsymbol{e}(d))$, FiLM-style

Table 10: Boundary intervention at fixed speech content. The centre L seconds of trimmed speech are held byte-identical across columns; only P seconds of boundary silence, resampled from a bank pooled over both classes, are added at each end. Because the added context has the same distribution whatever the clip's label, a change here cannot be the padding being informative – it is the detector responding to boundary context. *shift* is the mean absolute change in bona fide score from $P = 0$, *flip* the percentage of decisions that change, both paired per clip. Silence bank — `la19_eval`: 66,283 segments.

	EER at added boundary P (s)				at $P = 1$	
Model	0	0.25	0.5	1.0	shift	flip %
AASIST	13.3	18.3	19.6	20.5	4.19	43.2
RawNet2	23.1	20.9	23.2	24.3	2.51	0.0
SSL-AASIST	7.6	5.7	7.2	7.4	3.30	37.7
XLS-R Conformer	6.0	3.4	4.4	6.4	3.12	42.7
+ CL + DT	6.5	3.5	3.4	5.8	3.80	49.3

Table 11: Attack-stratified EER on LA19 (all bona fide clips of the view against one attack family's spoofs). Pooled EER sits far above the median family throughout, so it is carried by a few hard families rather than describing a typical one. This matters for duration curves because the eligible population is not attack-balanced: between 1 s and 3 s the spoof mix moves by 10.31 percentage points of total variation (largest single shift A17, 5.27 pp). Part of any per-length-cohort curve is therefore attack composition, which the shared cohort of Table 12 holds fixed by construction.

Model	view	pooled	median	min	max	IQR
XLS-R Conformer	trim	8.21	2.26	0.01	24.82	8.07
XLS-R Conformer	$L = 1$	11.14	5.52	1.12	31.50	8.48
XLS-R Conformer	$L = 2$	6.06	1.79	0.00	20.77	4.99
+ CL + DT	trim	6.76	2.27	0.21	21.93	5.37
+ CL + DT	$L = 1$	13.99	8.93	1.31	33.08	12.82
+ CL + DT	$L = 2$	6.50	2.72	0.00	22.36	4.58
SSL-AASIST	trim	5.72	2.27	0.00	20.00	4.57
SSL-AASIST	$L = 1$	14.00	6.58	0.00	41.23	11.70
SSL-AASIST	$L = 2$	7.56	2.68	0.00	27.69	6.36
AASIST	trim	15.53	6.13	0.33	33.33	15.76
AASIST	$L = 1$	19.32	14.60	0.33	44.30	18.66
AASIST	$L = 2$	13.28	4.71	0.00	32.31	12.65
RawNet2	trim	20.85	22.82	5.43	30.40	13.84
RawNet2	$L = 1$	34.23	36.05	6.67	52.99	23.73
RawNet2	$L = 2$	23.07	25.63	3.00	34.35	18.69

(Perez et al., 2018; de Vries et al., 2017); we know of no published anti-spoofing detector that conditions its forward pass this way, but it is what a practitioner would reach for first. DAS instead lets duration set the slope of a distance penalty, so that it can decide how local each head is without gaining an additive path to the classifier:

$$\mathrm{Attn}_h(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_h \boldsymbol{K}_h^\top}{\sqrt{D/H}} - m_h(d)\,|t_i - t_j| \right) \boldsymbol{V}_h, \qquad m_h(d) = \mathrm{softplus}\big(\boldsymbol{w}_h^\top \boldsymbol{e}(d) + b_h\big), \tag{1}$$

with t_i the time of frame i. Biases start at slopes spaced geometrically between 8 and $0.25\,\mathrm{s}^{-1}$, as in ALiBi (Press et al., 2022), of which equation 1 is the duration-conditioned case; fixing $\boldsymbol{w}_h{=}\mathbf{0}$ gives a learned but duration-*independent* distance bias, and is the control that matters most in §6.3.

Training views. Architecture cannot remove a shortcut that is in the data, so the training view is the other half of the design. Under RAND each clip is trimmed as in §4 and a random L-second crop is taken, $L \sim \mathcal{U}[1, 6]$ s drawn per clip, with padding masked throughout; crop lengths are then independent of the label, though the supplied source-duration scalar is not (Appendix F). Multi-

Table 12: Shared-cohort duration curves for released checkpoints. One balanced cohort is selected once, from clips with at least C seconds of speech, and *every* matched length is then taken from those same recordings, so the population no longer changes with L. *drift* is the mean absolute difference from the per-length-cohort curve of Table 1: it is the part of the published curve attributable to the evaluated population changing rather than to duration. Columns are matched length L in seconds.

		shared-cohort EER at L (s)						
Model	trim	1	2	3	4	6	8	drift
LA19 eval, cohort eligible at 4 s ($n = 478$ clips, balanced)								
XLS-R Conformer	1.3	8.4	3.3	2.5	1.3	–	–	1.81
+ CL + DT	1.3	12.6	4.6	3.3	2.1	–	–	0.97
SSL-AASIST	1.3	12.6	6.3	4.2	3.3	–	–	1.17
AASIST	4.6	17.6	10.9	9.6	6.7	–	–	1.65
RawNet2	14.2	33.5	23.8	18.8	15.9	–	–	0.48
In-the-Wild, cohort eligible at 4 s ($n = 8000$ clips, balanced)								
SSL-AASIST	8.8	30.3	15.4	11.1	9.9	–	–	0.55
AASIST	51.1	47.5	44.5	46.5	49.2	–	–	0.72
RawNet2	55.4	52.2	53.5	55.5	55.8	–	–	0.59
FoR test, cohort eligible at 2 s ($n = 1078$ clips, balanced)								
XLS-R Conformer	9.3	22.4	11.1	–	–	–	–	0.10
+ CL + DT	10.0	27.6	13.6	–	–	–	–	1.49
SSL-AASIST	3.3	9.6	4.5	–	–	–	–	0.04
AASIST	23.9	25.0	25.0	–	–	–	–	0.46
RawNet2	51.8	49.9	39.5	–	–	–	–	1.18

Table 13: Removing the source-duration input at evaluation (DAS+MLC, 6 seeds). Under a matched length every waveform is the same length, but a model trained with `dur_source=clip` is still told the speech duration of the recording the crop came from, which remains label-predictive; this is the strongest remaining duration cue inside a matched-length condition. *as trained* supplies it; *removed* overrides it with the crop length, so the duration input is constant within a condition and carries no information. Differences are paired by seed, with a 95% interval on the paired difference.

Set	L (s)	as trained	removed	Δ	95% CI
LA19 eval	1	6.36	6.36	+0.01	[-0.01, +0.03]
LA19 eval	2	3.38	3.37	-0.00	[-0.02, +0.01]
LA19 eval	3	2.51	2.52	+0.02	[-0.01, +0.05]
LA19 eval	4	1.39	1.32	-0.07	[-0.21, +0.07]
In-the-Wild	1	20.08	20.09	+0.01	[-0.01, +0.02]
In-the-Wild	2	13.79	13.78	-0.01	[-0.04, +0.02]
In-the-Wild	3	11.19	11.21	+0.02	[+0.00, +0.04]
In-the-Wild	4	9.64	9.64	-0.00	[-0.01, +0.00]
In-the-Wild	6	8.58	8.59	+0.01	[-0.00, +0.02]
In-the-Wild	8	7.55	7.59	+0.04	[-0.03, +0.11]
FoR test	1	26.68	26.73	+0.05	[-0.08, +0.18]
FoR test	2	12.39	12.43	+0.04	[-0.04, +0.11]

length consistency (MLC) adds a short view $\boldsymbol{x}_s$, $L_s \sim \mathcal{U}[1, 3]$ s, and a long view $\boldsymbol{x}_\ell$ of up to 6 s:

$$\mathcal{L} = \tfrac{1}{2}\big[\ell_{\mathrm{BCE}}(z_s, y) + \ell_{\mathrm{BCE}}(z_\ell, y)\big] + \lambda\, \ell_{\mathrm{BCE}}\big(z_s,\ \mathrm{sg}[\sigma(z_\ell)]\big), \tag{2}$$

where sg stops the gradient and λ=1: the long view teaches the short one, so the model is pushed towards one decision however much of the utterance it hears. This is cross-view self-distillation (Caron et al., 2021) and length-variability training (Chen et al., 2020) applied to this setting, not a new objective family. For comparison we also train on natural views, FIXED (first 4.04 s, repeat-padded, after Tak et al., 2022b) and FULL (whole untrimmed clip up to 8 s, given its true duration).

Table 14: EER (%) under standard and matched-length evaluation. *nat.*: first 4.04 s of the untrimmed clip (the usual input); *trim*: whole clip without leading/trailing silence; *DCE*: mean EER over matched lengths L (LA19: 1–4 s, ITW: 1–8 s, FoR: 1–2 s, limited by the number of long bona fide and spoofed clips); *max*: worst length. Our models: mean±std over seeds (number in brackets). †Released checkpoints, evaluated with our ports (Appendix B). AASIST and RawNet2 are at or near chance on ITW and FoR, so their matched-length numbers there are reported descriptively without interpreting their changes as shortcut effects. These detectors train on natural (untrimmed) views and are discussed in Appendix N.

	LA19 eval				In-the-Wild				FoR test			
Model	nat.	trim	DCE	max	nat.	trim	DCE	max	nat.	trim	DCE	max
Frozen XLS-R, natural views												
fixed 4 s (3)	0.17±0.05	4.2±0.7	4.5±0.6	8.6±1.0	16.1±1.9	15.8±3.0	15.4±2.0	27.2±2.6	17.7±5.4	24.4±5.5	20.3±3.6	25.2±3.9
full clip (3)	0.15±0.04	5.9±3.0	8.8±5.6	25.5±19.8	12.9±0.4	14.2±1.5	18.5±6.3	42.1±17.1	19.3±6.3	12.2±3.4	18.3±3.9	24.0±4.1
full clip + CONCAT (3)	0.17±0.04	5.1±0.8	4.7±0.7	9.9±3.5	14.3±2.2	15.1±1.6	13.2±1.9	25.8±5.6	26.1±5.6	13.6±1.3	19.5±0.3	25.5±0.4
full clip + GATE (3)	0.14±0.01	9.6±6.3	11.3±8.5	30.4±24.9	14.3±1.0	14.9±2.1	19.6±10.7	36.3±16.9	22.7±3.6	12.7±1.5	20.5±5.3	27.9±8.8

FULL RESULTS

Duration-conditioned detectors trained on *natural* clips are reported separately in Appendix N: the duration input costs them 3–7 EER points of cross-corpus transfer, but under length matching their dominant failure is the untrimmed training view.

The reference point is RAND: silence-free random crops, no duration mechanism. Against it, DAS and MLC do not deliver a consistent improvement. Over twenty seeds RAND reaches 11.3 ± 1.1 matched-length EER on ITW and 22.1 ± 2.4 on FoR, against 11.9 ± 0.9 and 20.3 ± 2.2 for DAS+MLC.

A clip-paired bootstrap with independent seed resampling, Benjamini–Hochberg corrected across all twelve matched-length conditions, separates them in four places and in opposite directions: DAS+MLC is better on FoR at L=1 s by 1.99 points (95% CI $[-3.59, -0.47]$, q=0.036) and worse on ITW at every crop of four seconds or more ($+0.85$, $+0.71$, $+0.67$ at L=4, 6, 8 s, all q=0.036). ITW at 3 s and FoR at 2 s are nominally significant but miss the corrected threshold, and the four LA19 conditions differ by 0.15–0.36 points, none significant (Appendix K tabulates all twelve). The pattern is consistent rather than scattered — it helps on short FoR speech and hurts on longer ITW speech — but it is not an improvement. These clip-level results are also sensitive to the resampling unit: an exploratory speaker-cluster bootstrap widens ITW 8 s to a near-zero-crossing interval, and none of its ten LA19/ITW cells then survives correction.

Two further results bound the mechanism's contribution. The fixed-slope ALiBi control is identical to DAS+MLC except that its slope cannot depend on d; it reaches 20.2 ± 1.9 on FoR and 11.8 ± 0.7 on ITW against 20.3 ± 2.2 and 11.9 ± 0.9, so at five seeds the duration-*independent* control is indistinguishable from the duration-adaptive one. That is not an equivalence test, but it is the comparison that would have to come out differently for the conditioning to be doing the work. And at five seeds the ITW L=1 comparison favoured DAS+MLC by 1.43 points with an interval excluding zero; twenty seeds reduce it to 0.39 with $[-1.35, +0.63]$ — a caution about seed budgets in this literature, ours included, rather than proof the earlier result was a false positive.

Removing the duration a model is told. One leak remains, for duration-conditioned models only: under a matched length every waveform has the same length, but our crop-trained models are told the *source* recording's speech duration, which still varies and stays label-predictive within the condition. We re-score every seed with that input overridden by the crop length, making it uninformative, with the result below.

The duration these models are still told. Every result above evaluates DAS+MLC as trained, so inside a matched-length condition it still receives the source recording's speech duration (see above) — and since that scalar stays label-predictive there, the one favourable result is exactly where it would be worrying. It is not the explanation. Re-scoring all 6 seeds with the input replaced by the crop length, constant within a condition and so uninformative, moves FoR at L=1 s from 26.68 to 26.73 EER, a paired change of 0.05 points (95% CI $[-0.08, 0.18]$), and no condition on any corpus moves by more than 0.07 points (Table 13). The FoR result survives removing the cue — but the same measurement limits the mechanism, since a model whose output barely moves when its

duration input is destroyed is not visibly using it. Direct measurement agrees: changing that input flips 0.18% of DAS's decisions on average and 0.99% at worst across twenty seeds, three corpora and four counterfactual inputs, against 2.48% for GATE, while the fixed-slope control flips 0.00%. Part of this is structural, as $m_h(d)$ reweights attention within a clip with no additive path to the classifier; but it is not inertia, because the learned slopes do move with duration, in four of eight heads by more than $1\,\mathrm{s}^{-1}$ (Appendix G). On this evidence the invariance does not buy accuracy.