

How Fast, How Sure, and Where It Breaks: A Streaming Sequential-Decision Layer for Voice-Deepfake Detection under Real Telephony Codecs

Ilja Semjonovs

`mr.ilja.semyonov@gmail.com`

Research Article

Keywords: voice anti-spoofing, audio deepfake detection, telephony codecs, streaming inference, sequential decision, confidence calibration

Posted Date: August 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10257178/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

How Fast, How Sure, and Where It Breaks: A Streaming Sequential-Decision Layer for Voice-Deepfake Detection under Real Telephony Codecs

Ilja Semjonovs^{1*}

^{1*}Independent Researcher.

Corresponding author(s). E-mail(s): mr.ilya.semyonov@gmail.com;

Abstract

Voice-cloning attacks increasingly target the telephone channel, where a codec has compressed the audio and a live call must be judged from however much has arrived. We wrap AASIST, reused without retraining, to consume audio as a growing prefix, emit a per-duration calibrated confidence, and commit only when confident, deferring otherwise. All audio is ASVspoof 2019 LA passed through four real wideband codecs — Opus, Speex-WB, EVS, AMR-WB — with packet loss, not real telephone calls; the detector is frozen, so what we report is a decision layer and a channel characterization, not an accuracy result.

The channel dominates: the same detector’s equal-error rate (EER) spans 0.8% to 29% across codecs and bitrates, so a robustness claim means little without naming the channel. Recast as a sequential reject option on a matched risk–coverage curve, the streaming layer on a harsh channel (Speex-WB q4 + 20% loss) decides $\sim 53\%$ of calls at 2.9% committed-call error, against $\sim 13\%$ for the full-utterance detector at the same coverage and $\sim 24\%$ if it must decide every call. This is selection, not accuracy: a max-confidence selector credits $\sim 83\%$ of the apparent advantage to short early prefixes repeated to fill the model’s fixed input window (tile-padding), and restricting it to prefixes ≥ 1.5 s collapses that share to $\sim 13\%$ — a collapse that replicates on two further harsh channels and a weaker raw-waveform backbone (RawNet2). The gain is bought by abstaining on $\sim 47\%$ of calls and rests on calibration that does not survive a channel mismatch. Calibration transfer is the actionable result: a clean-audio calibrator yields 3–3.5 $\times$ the matched error on harsh low-bitrate channels, and asymmetrically — over-confidence commits wrongly, over-caution only defers — so calibration should be fit on a representative harsh condition. The channel-independent gain is latency, roughly halved (~ 1.6 s forced-inclusive vs. the 3.26 s grand mean).

Keywords: voice anti-spoofing, audio deepfake detection, telephony codecs, streaming inference, sequential decision, confidence calibration

1 Introduction

Synthetic-speech generators are now cheap and capable enough to drive telephone fraud, where a cloned voice reaches the victim over a phone channel. That threat motivates this work but is not its evaluation target: we study how a frozen detector behaves under real telephony codecs, and how a streaming sequential-decision layer changes *when* and *how confidently* it can decide. All audio is codec-augmented ASVspoof 2019 LA, not real telephone calls. The scope is voice deepfakes over telephony, not image or video deepfakes or non-acoustic phone fraud.

Phone audio is degraded: it has passed through a speech codec, often at low bitrate, sometimes with packet loss, distorting the fine spectro-temporal cues a detector relies on. A live call also streams in: the detector cannot wait for a complete utterance but must decide from whatever has arrived, while reporting how confident it is and abstaining when it is not.

The closest prior work, by Shi et al., established a telephony-degradation benchmark (ADD-C Shi et al. (2025): six codecs and packet loss over spoofed speech, benchmarking AASIST) and an ultra-short-input detector (S-MGAA, “Detection at the First Greeting: Hi!” Shi et al. (2026)) classifying 0.5–2.0s clips under those degradations. We adopt their codec-degradation setting and AASIST backbone.

What that line does not address is the sequential decision process of a streaming call: prior work classifies pre-segmented clips, each an independent decision. We instead treat the utterance as a signal arriving over time and ask the detector to (i) update a calibrated confidence as each new slice of audio arrives, (ii) commit only when that confidence is high enough, under a two-sided early-stopping rule, and (iii) abstain when it never crosses the bar. This turns a one-shot classifier into an operating policy with an explicit latency–confidence–abstention trade-off. The sequential early-decision machinery was developed for general sound events Phan et al. (2018); Zhao et al. (2022); we adapt it to streaming deepfake speech under telephony (Sec. 3.4).

Our contributions:

- **A streaming sequential-decision layer** over a frozen detector: growing-prefix inference, per-duration calibration, and a two-sided commit-when-confident stopping rule with abstention, adding no training and transferring across two raw-waveform backbones (AASIST, RawNet2) (Sec. 3).
- **A real-codec channel characterization** (four wideband codecs plus packet loss): the same frozen detector’s error spans near-clean to severely degraded by codec and bitrate alone (Sec. 4.2).
- **A latency–confidence operating curve** in the selective-prediction tradition, with a controlled risk–coverage analysis separating a genuine latency gain from an apparent accuracy gain that is mostly a tile-padding artifact (Sec. 4.3).

- **A calibration-transfer ablation**, the most actionable result: clean-fit calibration is overconfident, and asymmetrically so, on harsh channels, yielding a concrete deployment rule (Sec. 4.4).

Scope boundaries are detailed in Sec. 5.

2 Related work

2.1 Telephony-degraded and short-input deepfake detection

The closest prior work is Shi et al. on spoofed speech under channel degradation. ADD-C Shi et al. (2025) is a degradation *benchmark* — six wideband codecs (AMR-WB, EVS, IVAS, Opus, Speex-WB, SILK) crossed with packet loss — benchmarking standard detectors including AASIST at a fixed four seconds, documenting a substantial robustness drop. The S-MGAA / “Hi!” paper Shi et al. (2026) extends this to ultra-short inputs, plotting average EER against clip duration. We adopt both the setting and the AASIST backbone. The distinction is the *form of the decision*: there each clip is one independent, pre-segmented decision at fixed length, reported as an EER grid or EER-vs-duration curve. None treats the input as a call that streams in, maintains a confidence trajectory, applies a commit-when-confident stopping rule, or abstains — the sequential decision process we add on their benchmark and backbone.

2.2 Other “real-time” framings, and why they differ

Other work attaches “real-time” framings. RTCFake Xue et al. (2026) contributes a dataset of speech through real-time communication platforms, but “real time” describes the *data source*: its detector is an offline batch classifier reporting EER. Two works do stream over a call — an RVC-attack detector with segment-level decisions and call-level aggregation Chinchmalatpure et al. (2025), and a sliding-window CNN with probabilistic aggregation Jose et al. (2026) — but each reports a single operating point with no measured latency, no latency–confidence curve, no commit-when-confident rule, and no codec $\times$ packet-loss matrix. The streaming framing is wanted; none supplies the sequential decision process we develop.

2.3 Sequential early decision in adjacent domains

We adapt the sequential early-decision idea rather than originate it. Phan et al. Phan et al. (2018) introduced early detection for general audio events with a monotonicity constraint for reliable early decision; our persistence guard (Sec. 3.4) is a lightweight realization of it. SEED Zhao et al. (2022) used evidential uncertainty to decide when enough evidence has accrued, of which our confidence-thresholded rule is a simpler instance, and the sequential-test tradition (e.g. SPRT) predates both. What is new is the *application*: none of these addresses speech, telephony codec degradation, or per-duration calibration under a channel, and to our knowledge sequential early decision has not been brought to streaming deepfake-speech detection. Cross-attack-generalization work — e.g. information-bottleneck adversarial alignment Huang et al.

(2025) — targets *unseen attacks*, a problem orthogonal to and composable with ours (Sec. 5.2): such a detector drops into our streaming layer unchanged.

3 Method

3.1 Frozen detector and measurement spine

We use AASIST Jung et al. (2022), a 297,866-parameter raw-waveform graph-attention anti-spoofing network and the backbone the ADD-C benchmark Shi et al. (2025) adopts. We load the released weights and run eval-only, never fine-tuning. On the ASVspoof 2019 LA evaluation set Wang et al. (2020) (71,237 trials) the reproduced model attains 0.83% equal-error rate (EER) (0.8295% before rounding; published 0.83%). The 0.0275 minimum tandem detection cost (t-DCF) is AASIST’s published reference value from the official repository, which we check the reproduced EER against; we do not recompute t-DCF. An independent EER implementation agrees to 1.4×10^{-4} pp.

3.2 Growing-prefix streaming

A phone call streams in rather than arriving as a clip, so we recast detection as a **growing-prefix** problem: at decision time t the system has observed only the first t seconds and must act on that prefix. We evaluate on a fixed grid $t \in \{0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 4.04\}$ s, the last being AASIST’s native 64,600-sample window (≈ 4.04 s at 16 kHz). AASIST expects fixed-length input, so each prefix is **tile-padded** — repeated to fill the window and truncated — reproducing the padding AASIST applies at evaluation Jung et al. (2022) (a full-length prefix is bit-identical to its standard input).

3.3 Per-duration confidence calibration

AASIST’s raw score is not a probability, and a score off 0.25 s does not mean what the same score off 2 s means. We fit a separate one-dimensional Platt (logistic) calibrator per prefix length, mapping the raw score to a calibrated $P(\text{spoof})$; each is fit on held-out development data and applied to the evaluation subsample, never on test. This lets a single call’s confidence trajectory be read consistently as audio arrives; Sec. 4.3 reports per-duration expected calibration error (ECE).

3.4 Two-sided commit-when-confident stopping rule

Given the calibrated trajectory $P(\text{spoof})$ over the grid, the layer chooses *when* to commit, under a two-sided symmetric rule with one patience parameter $\theta \in (0.5, 1]$. At step i it commits

- **spoof** if $P(\text{spoof}) \geq \theta$ for the most recent k steps, or
- **bona fide** if $P(\text{spoof}) \leq 1 - \theta$ for the most recent k steps,

whichever fires first. A fraud-cost-weighted asymmetric threshold is a direct extension we do not implement. The k -consecutive requirement is a **persistence guard** ($k = 2$)

against acting on a single transient window, operationalizing the monotonicity insight of Phan et al. [Phan et al. \(2018\)](#). If neither threshold is reached by the final prefix, the system forces a decision from the full-length score ($P(\text{spoofer}) \geq 0.5$) and flags the call **forced**; its error is still counted. The forced fraction — the count of calls on which the detector never became confident — is the **reject signal** of a selective classifier: in deployment such a call would be deferred to a human reviewer or a second authentication factor.

Sweeping θ traces a **latency-vs-confidence operating curve**: small θ commits early and errs more; large θ waits for stronger evidence at the cost of latency and a rising forced fraction. This operating point, rather than accuracy at a fixed clip length, distinguishes our formulation from the fixed-duration EER grids of prior telephony work [Shi et al. \(2025, 2026\)](#). Latency is capped at the audio available.

3.5 Real-codec telephony pipeline

We evaluate under realistic phone-channel degradation. Following ADD-C [Shi et al. \(2025\)](#) (codec compression plus packet loss), every utterance is round-tripped through a standards-compliant encoder and decoder before the detector sees it, using **four real wideband codecs**:

- **Opus** (libopus), the WebRTC/VoIP default;
- **Speex-WB** (Xiph), a royalty-free wideband codec;
- **EVS** (3GPP TS 26.443 floating-point reference C), the modern VoLTE/VoNR codec;
- **AMR-WB** (3GPP TS 26.173 / ITU-T G.722.2 reference C), the classic 3G/VoLTE wideband codec.

Each codec gets a high/mid/low bitrate ladder, with frame-level packet loss at 0/10/20% on a representative mid-bitrate setting — 21 conditions (clean plus five per codec), with a G.711 narrowband DSP approximation added as a legacy public switched telephone network (PSTN) bracket. This is not ADD-C parity: the gaps (IVAS, SILK) are detailed in [Sec. 5.3](#).

Round-trips are deterministic and run in parallel byte-identically to a serial reference, and packet loss consumes its random stream in fixed file order, so a fixed seed reproduces the exact degraded audio. The AMR-WB reference encoder’s constant 95-sample (5.94 ms) algorithmic delay we measure by cross-correlation, verify mode-independent, and strip, leaving compression distortion as the only codec-attributable difference. (The EVS and AMR-WB reference binaries are patent- and ETSI-licensed, for in-lab use only.)

4 Experiments and results

4.1 Setup

All experiments use ASVspoof 2019 LA [Wang et al. \(2020\)](#) through the pipeline of [Sec. 3.5](#). Across 21 conditions and three seeds, every sweep result is computed on a

Table 1 Full 21-condition telephony sweep (mean $\pm$ std over 3 seeds, $n_{\text{dev}} = n_{\text{eval}} = 4000$). Full-utterance EER is over all utterances; streaming error/latency/forced are at $\theta = 0.9$. A call that never crosses θ is forced to a full-length decision and its error is *counted* in the streaming error (not excluded); the forced column is the share of such calls. Mean full-clip latency 3.26 s.

Cond.	Description	Full EER (%)	Stream err (%)	Stream lat. (s)	Forced (%)
C0	clean (reference)	0.83 $\pm$ 0.12	2.43 $\pm$ 0.11	1.65 $\pm$ 0.00	1.75 $\pm$ 0.11
O1	Opus 24 kbps, 0% loss	0.93 $\pm$ 0.09	2.42 $\pm$ 0.13	1.67 $\pm$ 0.01	1.77 $\pm$ 0.10
O2	Opus 12 kbps, 0% loss	1.80 $\pm$ 0.11	2.64 $\pm$ 0.09	1.75 $\pm$ 0.01	2.75 $\pm$ 0.09
O3	Opus 6 kbps, 0% loss	15.68 $\pm$ 0.19	7.43 $\pm$ 0.31	2.04 $\pm$ 0.03	22.79 $\pm$ 0.74
O4	Opus 12 kbps, 10% loss	7.82 $\pm$ 0.15	7.29 $\pm$ 0.21	1.95 $\pm$ 0.01	23.18 $\pm$ 1.00
O5	Opus 12 kbps, 20% loss	12.07 $\pm$ 0.67	9.83 $\pm$ 0.90	2.08 $\pm$ 0.01	35.32 $\pm$ 0.84
S1	Speex-WB q8 ($\sim$ 28 kbps), 0% loss	3.23 $\pm$ 0.24	3.23 $\pm$ 0.10	1.76 $\pm$ 0.01	3.62 $\pm$ 0.27
S2	Speex-WB q4 ($\sim$ 15 kbps), 0% loss	9.50 $\pm$ 0.63	5.27 $\pm$ 0.41	1.93 $\pm$ 0.00	10.82 $\pm$ 0.35
S3	Speex-WB q2 ($\sim$ 10 kbps), 0% loss	29.45 $\pm$ 0.48	12.62 $\pm$ 0.46	2.13 $\pm$ 0.01	39.88 $\pm$ 0.46
S4	Speex-WB q4, 10% loss	18.92 $\pm$ 0.41	12.56 $\pm$ 0.61	2.10 $\pm$ 0.01	35.34 $\pm$ 1.53
S5	Speex-WB q4, 20% loss	24.43 $\pm$ 0.25	15.88 $\pm$ 0.14	2.13 $\pm$ 0.01	47.24 $\pm$ 0.41
E1	EVS 24.4 kbps, 0% loss	1.38 $\pm$ 0.10	2.89 $\pm$ 0.15	1.72 $\pm$ 0.00	2.61 $\pm$ 0.14
E2	EVS 13.2 kbps, 0% loss	2.48 $\pm$ 0.06	3.68 $\pm$ 0.17	1.82 $\pm$ 0.00	5.22 $\pm$ 0.20
E3	EVS 7.2 kbps, 0% loss	4.88 $\pm$ 0.35	5.42 $\pm$ 0.39	1.91 $\pm$ 0.01	8.87 $\pm$ 0.58
E4	EVS 13.2 kbps, 10% loss	10.08 $\pm$ 0.43	10.32 $\pm$ 0.07	2.01 $\pm$ 0.02	28.18 $\pm$ 0.79
E5	EVS 13.2 kbps, 20% loss	13.65 $\pm$ 0.32	13.01 $\pm$ 0.17	2.11 $\pm$ 0.00	40.58 $\pm$ 0.27
A1	AMR-WB 23.85 kbps (m8), 0% loss	2.13 $\pm$ 0.05	2.90 $\pm$ 0.11	1.75 $\pm$ 0.00	2.83 $\pm$ 0.14
A2	AMR-WB 12.65 kbps (m2), 0% loss	3.28 $\pm$ 0.29	3.50 $\pm$ 0.18	1.80 $\pm$ 0.00	4.16 $\pm$ 0.20
A3	AMR-WB 6.60 kbps (m0), 0% loss	13.72 $\pm$ 0.37	7.44 $\pm$ 0.33	2.00 $\pm$ 0.01	17.02 $\pm$ 0.56
A4	AMR-WB 12.65 kbps (m2), 10% loss	10.35 $\pm$ 0.24	9.22 $\pm$ 0.37	1.97 $\pm$ 0.01	24.95 $\pm$ 0.98
A5	AMR-WB 12.65 kbps (m2), 20% loss	15.23 $\pm$ 0.52	12.61 $\pm$ 0.63	2.10 $\pm$ 0.01	38.11 $\pm$ 0.55

class-balanced subsample of 4,000 development and 4,000 evaluation utterances (calibrators fit on development, all numbers on the held-out evaluation subsample). Unless stated otherwise we report mean $\pm$ std over seeds {1234, 5678, 9012}, each redrawing both the subsample and the packet-loss pattern. The persistence guard is $k = 2$, and the mean full-clip duration (the latency of a full-utterance decision) is 3.26 s.

Three clean-condition EER figures appear — the same frozen model on the same clean data, differing by sampling not modeling: the 0.83% of Sec. 3.1 (reproduction on the full 71,237-trial LA set); the clean reference C0 = 0.83 $\pm$ 0.12% of Sec. 4.2 (full-utterance EER on the balanced 4k subsample over three seeds, coinciding with the full-set number and so confirming the subsample is representative); and the 1.00% of Sec. 4.3 (the same metric on the same subsample for the single seed 1234 on which the per-prefix and operating curves are reported, just above the 0.83 $\pm$ 0.12% spread).

4.2 The channel dominates: from near-clean to severely degraded

The *same* frozen detector on the *same* utterances ranges from near-transparent to severely degraded depending only on the codec (Fig. 1, Table 1). Full-utterance EER spans the clean reference C0 = 0.83 $\pm$ 0.12% to $\sim$ 29% (Speex-WB at 10 kbps; S3, 0% loss). The ordering is consistent at every bitrate: EVS is the most transparent codec,

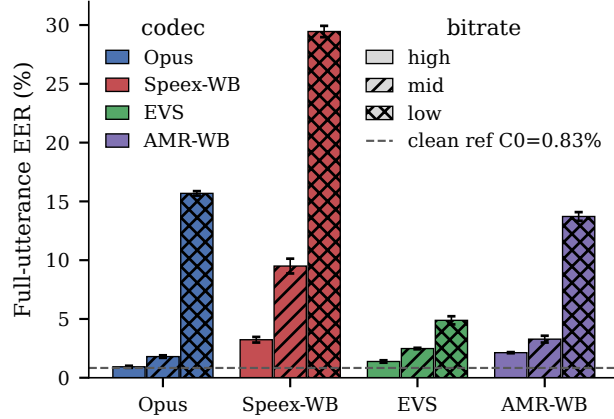


Fig. 1 Full-utterance EER by codec and bitrate (0% loss), the same frozen detector on the same utterances (mean $\pm$ std over seeds {1234, 5678, 9012}). Colour encodes codec, hatch the high/mid/low bitrate tier; the dashed line is the clean reference C0. EVS is the most transparent codec at every tier and Speex-WB the harshest; Opus drops sharply only at its lowest bitrate. Per-condition values, including the packet-loss conditions, are in Table 1.

Speex-WB the harshest, with Opus and AMR-WB between. Opus shows a bitrate cliff rather than gradual decay — near-clean at 24 and 12 kbps, then a $\sim 9\times$ jump to 15.68% at 6 kbps — while AMR-WB degrades gracefully to 13.72% at its 6.6 kbps floor. Effective error is thus a property of the channel, not the detector alone.

Packet loss compounds the codec effect for every codec (Table 1). The G.711 legacy-PSTN bracket gives 3.20% EER, between Opus 24 kbps and Speex-WB 10 kbps; as a single channel it would have hidden the codec ranking, which is why “which codec” is a first-class variable.

4.3 Streaming: latency, coverage, and the forced fraction

At $\theta = 0.9$ the streaming rule commits at a mean 1.6–2.1 s across all 21 conditions, against 3.26 s full-utterance — roughly half (Table 1). That figure is forced-inclusive (forced calls run to full length); committed-only latency is lower, as the risk-coverage points show.

Error must be read against coverage. On a benign channel the benefit is latency: at $\theta = 0.9$ the rule commits at ~ 1.6 s with a small forced fraction ($\sim 1.8\%$) and a risk-coverage curve close to the full-utterance system’s (Fig. 2); deciding from the whole utterance is more accurate only when every call must be decided.

Under harsh degradation the streaming error at $\theta = 0.9$ sits below the full-utterance EER: Speex-WB 10 kbps 0% loss (S3) 12.62% vs. 29.45%, Opus 6 kbps 7.43% vs. 15.68%, AMR-WB 6.6 kbps 7.44% vs. 13.72%. These are different metrics: the full-utterance figure is threshold-free EER, the streaming figure is error at a fixed calibrated operating point over decisions committed from a shorter prefix. Every call is still decided — one that never reaches confidence is forced to a full-length decision and counted — while the forced fraction rises: Speex-WB 10 kbps 0% loss 39.9%,

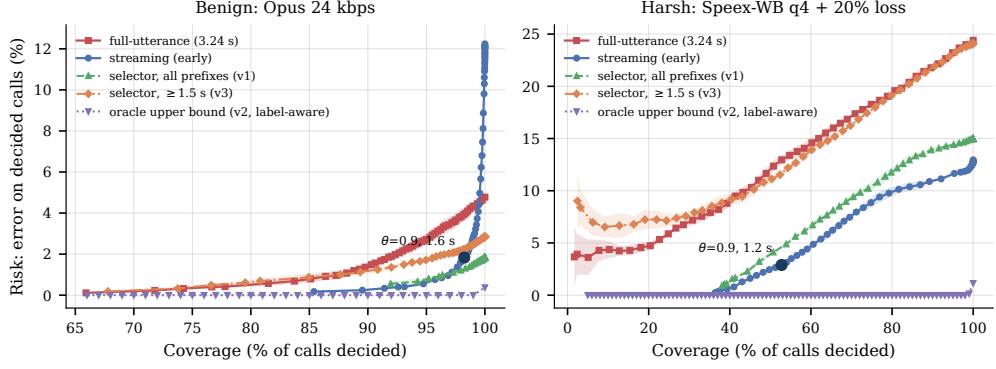


Fig. 2 Matched risk-coverage comparison of the streaming layer and the full-utterance detector as selective classifiers: error on decided calls vs. fraction decided, sweeping the confidence threshold (mean $\pm$ std over seeds {1234, 5678, 9012}; bands show the std). Both use the identical per-duration calibration. *Left*, a benign channel (Opus 24 kbps); *right*, a harsh channel (Speex-WB q4 + 20% loss). The marked point is $\theta = 0.9$ at its mean committed latency. Also shown are three non-sequential selectors that pick each call’s most-decisive prefix: the label-free max-confidence variants v1 (all prefixes, dashed) and v3 (≥ 1.5 s, dash-dot) and the label-aware oracle v2 (dotted); v1/v2/v3 are defined and interpreted in the text.

Speex-WB mid + 20% loss 47.2%, Opus 12 kbps + 20% loss 35.3%, EVS mid + 20% loss 40.6%. Under heavy degradation only about half the calls commit early.

The same event — a call whose confidence never crosses θ before the audio runs out — is accounted two ways. In the operating-point table (Table 1) it is forced-and-counted: decided by the full-length fallback, its error in the streaming error. In the risk-coverage curve (Fig. 2) it is abstained-and-excluded: a deferral removed from the risk, which by construction is the error on decided calls only. The two answer different questions — the error if every call must be decided, versus the error on the calls the system chooses to decide.

To compare on equal terms we read both as selective classifiers and plot the **risk-coverage curve**: error on decided calls (risk) against fraction decided (coverage) Chow (1970); El-Yaniv and Wiener (2010); Geifman and El-Yaniv (2017). We sweep θ for the streaming rule and the same two-sided threshold at full length for the baseline, on one benign channel (Opus 24 kbps) and one harsh channel. The harsh exemplar is Speex-WB q4 + 20% loss, distinct from the 0%-loss Speex-WB 10 kbps (S3) of the operating-point table. Both share the identical per-duration calibration and read the identical per-prefix calibrated scores — two post-processings of one cached score array (the baseline reads each call’s final-prefix score, streaming reads the growing prefix under the stopping rule), so the comparison differs only in when a decision is allowed, and the baseline is never disadvantaged. Curves are recomputed from these scores for all three seeds, bands showing the spread.

On the benign channel the curves are close: the full-utterance system is more accurate only near full coverage (4.8% error deciding every call), while streaming reaches comparable or lower risk at slightly reduced coverage ~ 1.6 s sooner. On the harsh channel (Speex-WB q4 + 20% loss) the streaming curve sits below the full-utterance reject option at every coverage. At $\theta = 0.9$ it decides $\sim 53\%$ of calls (i.e.

abstains on the complementary $\sim 47\%$) at 2.9% error (3-seed mean, $\text{std} \leq 0.13$ pp), where the full-utterance *baseline* errs $\sim 13\%$ at the same (matched $\sim 53\%$) coverage and $\sim 24\%$ deciding every call.

Four error figures describe this condition, each over a different population: streaming’s committed-only 2.9% (decided calls only); the full-utterance baseline’s $\sim 13\%$ at the matched $\sim 53\%$ coverage; the full-utterance $\sim 24\%$ deciding every call; and streaming’s own decide-every-call figure, the $\theta = 0.9$ forced-inclusive 15.88% (abstentions forced to full length and counted). Streaming decisions arrive at a mean 1.2 s rather than 3.24 s.¹ The effect is consistent across all three seeds (Fig. 2 bands are tight). Full-curve dominance is shown only for this condition; the selector decomposition below we confirm on two further harsh channels spanning a different mechanism and two more codecs, and we do not claim full-curve dominance for every channel.

Two effects drive the gap. The reject option is matched — both may abstain under the same calibration — but temporal selection is not: streaming commits on an early prefix, while the baseline must decide from the whole degraded utterance. To isolate this, we add a **max-confidence selector** that post-processes the same calibrated scores (no re-inference): for each call it picks the single most decisive prefix (farthest from $P(\text{spoo}) = 0.5$ over the grid), the best a non-sequential, label-free selector could choose, then applies the two-sided threshold with no persistence guard and no forcing.

That argmax can land on a very short, tile-padded prefix — the windows the per-prefix EER flags as least reliable (Fig. 4). We therefore run the selector twice: over all prefixes (**v1**), and restricted to prefixes ≥ 1.5 s (**v3**). The 1.5 s cut is the length by which the per-prefix EER (Fig. 4) has fallen close to its full-length floor; below it the detector sits in the unreliable short-window regime the selector exploits. At the $\theta = 0.9$ operating point (matched coverage $c^* = 52.8 \pm 0.4\%$) on Speex-WB q4 + 20% loss the two diverge sharply (3-seed mean $\pm$ std): the full-utterance system errs $12.99 \pm 0.47\%$ and streaming reaches $2.89 \pm 0.13\%$, a ~ 10.1 pp gap. The all-prefix selector v1 cuts risk to $4.66 \pm 0.19\%$, crediting $82.5 \pm 1.5\%$ of the gap (8.3 of 10.1 pp) to temporal selection. Restricting to ≥ 1.5 s (v3) raises risk to $11.64 \pm 0.41\%$ — almost back to full-utterance — collapsing that share to $13.3 \pm 3.7\%$ (1.3 of 10.1 pp). The conclusion is not knife-edge on the 1.5 s cut: restricting instead to ≥ 1.0 s or ≥ 2.0 s gives temporal-selection shares of $8.5 \pm 4.9\%$ and $9.6 \pm 3.2\%$ (oracle risk 12.11 and 12.01%), both below the ≥ 1.5 s value and an order of magnitude under the all-prefix 82.5% — if anything 1.5 s is the most generous of the three to temporal selection. The flip is stable across seeds (the v1-to-v3 gap never below 62 pp). We take v3 as the estimate: the apparent temporal-selection advantage is overwhelmingly the selector exploiting sub-1.5 s tile-padded prefixes rather than genuine early classification.

With the artifact controlled, the sequential rule and reject option together account for only $\sim 13\%$ of the harsh-channel advantage. The two surviving gains are not accuracy: latency (committed-only ~ 1.2 s vs. the 3.24 s full-utterance mean for this condition) and the calibrated deferral signal that lets the system abstain. The label-aware oracle **v2** sits at $\sim 0\%$: selecting by labels rather than confidence, it commits on the easiest coverage-matched fraction and so marks the lowest risk any selector

¹The 1.2 s is the mean latency of the *committed* calls on this condition; the 1.6–2.1 s range above averages in forced calls, which run to full length. The 3.24 s full-utterance latency is the mean full-clip duration on this condition’s subsample, matching the 3.26 s grand mean of Sec. 4.1 to within per-condition sampling.

Table 2 Selector decomposition across the secondary harsh channels and the second backbone (RawNet2), 3-seed means. The headline Speex-WB q4+20% (AASIST) row is repeated from the text for reference. “gap” is the full-utterance minus streaming error at matched coverage; v1 and v3 are the share of that gap credited to the all-prefix and ≥ 1.5 s max-confidence selectors; “collapse” is the v1→v3 share drop (pp) where the gap is large enough to quantify a reliable share, and ✓ where it is not. In every channel v3 lands at or above the full-utterance curve.

Backbone	Channel	gap (pp)	v1 (%)	v3 (%)	collapse (pp)
AASIST	Speex-WB q4+20%*	10.1	82.5	13.3	69
AASIST	Opus 6 kbps	9.9	95.7	4.9	91
AASIST	AMR-WB 12.65k+20% [‡]	2.7	n/a	n/a	✓
RawNet2	Speex-WB q4+20%	5.9	n/a	n/a	✓
RawNet2	Opus 6 kbps	1.3	n/a	n/a	✓

*Headline; full decomposition in text. *n/a* / ✓: the matched gap is too small (or the label-free v1 selector beats the causal rule) for a reliable share, so v1, v3, and their collapse are omitted; ✓ records that v3 still lands at or above the full-utterance curve. [‡]Weak test: only a 2.7 pp gap to explain.

could reach at this coverage (since $\sim 98.8\%$ of calls have at least one correct prefix). Being label-dependent, it is unattainable and bounds the headroom rather than setting the bar; the operative baseline remains the full-utterance system. The wide gap between that floor and streaming’s $\sim 2.9\%$ shows confidence-based window selection leaves most headroom unclaimed.

The decomposition reproduces on two further harsh channels — Opus 6 kbps (a low-bitrate cliff, not packet loss) and AMR-WB 12.65 kbps + 20% loss — and on a $\sim 5\times$ weaker backbone, RawNet2 (clean full-utterance EER 4.585% vs. AASIST’s 0.83%, within the published single-system band 4.66–5.64% Tak et al. (2021)). Table 2 collects them: in every channel the ≥ 1.5 s v3 selector lands at or above the full-utterance curve, so the apparent early-window edge is a tile-padding artifact, not early classification. In several channels the label-free v1 selector beats the causal streaming rule, making its share degenerate ($>100\%$ or <0); we therefore rest the claim on v3’s position relative to the full-utterance curve, not the share magnitude, as for the headline channel.

The 2.9% is bought by abstaining on $\sim 47\%$ of calls, useful only where deferral is cheap — a deferral-tolerant screening stage backed by a human reviewer or a second authentication factor. Handing off nearly half of all calls would be operationally prohibitive for a hard real-time path with no fallback. The deferral signal also depends entirely on the confidence being calibrated Guo et al. (2017), which Sec. 4.4 shows does not survive a channel mismatch. What holds channel-independently is that the layer moves an axis the baseline cannot, latency, and emits a per-duration-calibrated signal for *when* the detector is out of its depth.

Sweeping θ traces the full operating curve. On clean LA it is a monotone Pareto front: error falls from 12.1% at 0.52 s ($\theta = 0.5$) to 1.6% at 2.11 s ($\theta \approx 0.995$), the forced fraction rising 0 to 8.5% (Fig. 3). Early decisions are hard before any channel effect: the clean per-prefix EER (Fig. 4) falls from 16.75% at 0.25 s to 1.00% at full

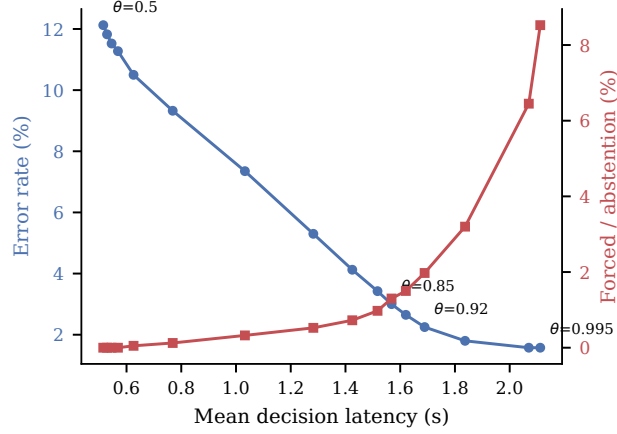


Fig. 3 Clean streaming operating curve as θ is swept: error rate (left axis) and forced/abstention fraction (right axis) versus mean decision latency. More patience buys lower error at the cost of latency and rising abstention.

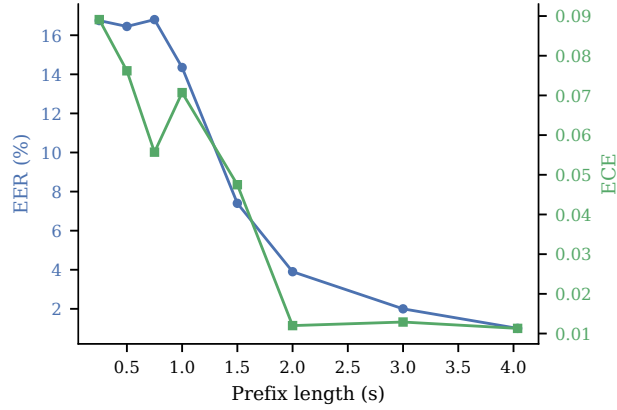


Fig. 4 Clean per-prefix EER and ECE versus prefix length (clean C0, seed 1234). Both fall as more audio arrives; below ~ 1.5 s the detector is unreliable even on clean audio, which the max-confidence selector of Sec. 4.3 exploits through tile-padded short windows.

length (seed 1234), calibration improving in step. Part of this is the same tile-padding confound, part intrinsic difficulty; either way AASIST decides far less reliably below ~ 1.5 s even on clean audio, a floor on how early any stopping rule on this backbone can safely commit.

4.4 Calibration does not transfer from clean to harsh channels

The stopping rule acts on calibrated confidence, so does one calibrator suffice or must it be re-fit per channel? We run a source $\times$ target ablation over the 0%-loss ladder (clean plus high/mid/low per codec): fit one per-duration calibrator on a source’s development data, apply it to every target’s evaluation data, and measure ECE and

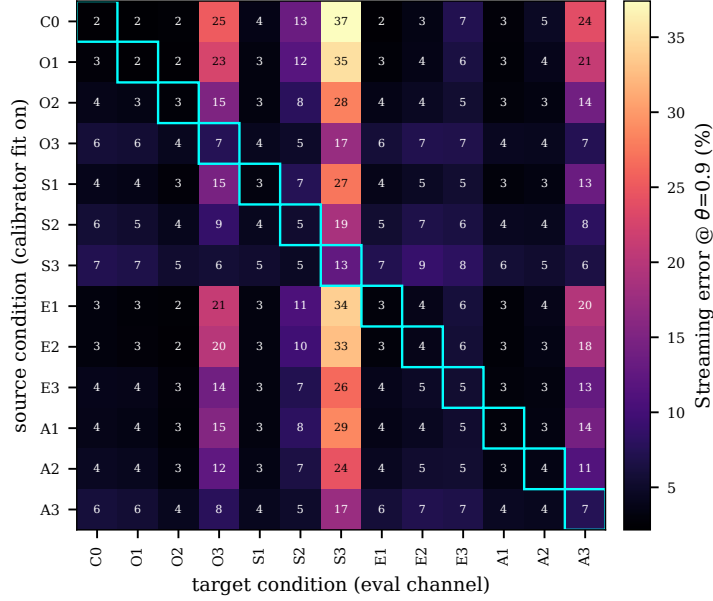


Fig. 5 Calibration transfer: streaming error (%) at $\theta = 0.9$ when a calibrator fit on the source condition (row) is applied to the target condition (column). Condition codes: O = Opus, S = Speex-WB, E = EVS, A = AMR-WB, with 1/2/3 = high/mid/low bitrate (e.g. S3 = Speex-WB 10 kbps, O3 = Opus 6 kbps, A3 = AMR-WB 6.6 kbps). The matched diagonal (cyan) is well-calibrated everywhere; a clean (C0) calibrator deployed onto harsh low-bitrate targets is overconfident by up to $3.5\times$.

the $\theta = 0.9$ streaming operating point, over three seeds. Target EER is calibrator-invariant (it depends only on raw scores) — a check the target EERs pass, reproducing the Sec. 4.2 means exactly (C0 $0.83 \pm 0.12\%$, Opus 6 kbps $15.68 \pm 0.19\%$, Speex-WB 10 kbps (S3) $29.45 \pm 0.48\%$).

A matched (diagonal) calibrator is well-calibrated everywhere: mean ECE 0.0246 ± 0.0014 , streaming error $4.76 \pm 0.12\%$. Mismatched transfer is worse on average (ECE 0.0752 ± 0.0010 , error $7.94 \pm 0.11\%$), but the average hides the real risk, which lives in one corner. A clean-fit calibrator on a harsh low-bitrate target is severely overconfident: streaming error rises to $25.28 \pm 0.59\%$ on Opus 6 kbps (matched 7.43%), $37.41 \pm 0.57\%$ on Speex-WB 10 kbps (matched 12.62%), and $23.71 \pm 0.79\%$ on AMR-WB 6.6 kbps (matched 7.44%) — 3 to $3.5\times$ the matched error, ECE up to 0.34, dwarfing seed variance (all bars ≤ 0.8 pp). EVS is the lone exception: transparent enough that even its 7.2 kbps calibrator is nearly interchangeable with the clean one (C0 $\rightarrow$ EVS 7.2 kbps costs $+0.009$ ECE).

The failure is asymmetric, which makes it actionable. A gentle-channel calibrator on a harsh channel makes the system overconfident — wrong commits, the dangerous direction. A harsh-channel calibrator on a gentle channel makes it over-timid, which under the two-sided rule becomes more abstentions with only mild added error — the safe direction. So a single clean-trained calibrator is unsafe to ship into low-bitrate telephony; if only one can be deployed, fit it on a representative harsh condition. The

matched diagonal is uniformly well-calibrated ($ECE \leq 0.10$), so re-fitting per channel where it is known is the safe default. Fig. 5 shows the full source $\times$ target matrix.

5 Limitations

What the work does not establish, in order of weight: the single-dataset external-validity gap (Sec. 5.1) and the unseen-attack limit (Sec. 5.2) are load-bearing; codec coverage (Sec. 5.3) and smaller methodological choices (Sec. 5.4) are secondary.

5.1 A single dataset

Every utterance originates from ASVspoof 2019 LA Wang et al. (2020), a TTS and voice-conversion corpus recorded clean, to which we apply codecs and packet loss. This is codec-augmented benchmark speech, not telephone-call audio.

Two factors separate it from genuine call audio. First, real phone audio carries room and handset acoustics, background noise, automatic gain control, network jitter, and packet concealment beyond uniform frame loss — none reproduced by a codec round-trip on clean studio audio, any of which could shift the operating curve. Second, the ASVspoof classes are produced under controlled conditions, while fraud calls differ in provenance, language, and synthesis systems. Evaluating on genuine phone-channel audio is the most valuable next step. The runs also use balanced 4,000/4,000 subsamples; the subsample’s clean EER ($0.83 \pm 0.12\%$) matches the full-eval reproduction (0.83%), so it is representative but a subsample.

5.2 Unseen attacks: the field’s open problem, and how we scope around it

Generalization to *unseen* synthesis attacks is the field’s central open problem, not one we claim to solve; nothing here is evidence of robustness to novel generators.

ASVspoof 2019 LA’s evaluation set is by design built around attacks disjoint from its training set Wang et al. (2020), so the frozen detector is already tested against attacks unseen in training, and our channel results inherit that. But those attacks predate the corpus: modern neural-codec and diffusion generators are absent, and a detector frozen on 2019-era attacks may fare very differently — most plausibly worse — against them. Our findings are thus *conditional on a given detector*: how a frozen detector behaves as audio streams in under telephony, not whether it catches tomorrow’s clones. Because the layer is composable with the backbone (no training, black-box scorer), a more generalizable detector — including the cross-attack-generalization line Huang et al. (2025) — substitutes under the same rule, calibration, and measurement.

5.3 Codec coverage is partial

We encode four of the six ADD-C codecs as real encoders (Opus, Speex-WB, EVS, AMR-WB) and claim no ADD-C parity. IVAS and SILK lack a true encoder here — IVAS overlaps EVS in mono telephony with no standalone encoder available to us, SILK only approximable through Opus-internal SILK — so we report them as gaps

rather than ship an approximation under a real codec’s name. G.711 is an explicit DSP approximation for a legacy-PSTN channel. Our ladders sample three operating points per codec plus loss at three rates on one mid-bitrate setting: the shape of the degradation, not its whole surface. This does not undermine Sec. 4.2, which only needs the channels we ran to span near-clean to $\sim 29\%$, but the codec ranking covers the four studied.

5.4 Other methodological choices

(i) Short prefixes are tile-padded to AASIST’s window (Sec. 3.2); a variable-length streaming model could behave differently, and as Sec. 4.3 shows, this padding accounts for most of the apparent harsh-channel temporal-selection advantage. (ii) We *trace* the operating curve by sweeping θ ; a single deployed θ must be chosen on development data, never test. (iii) The stopping rule is symmetric; a fraud-cost-weighted asymmetric operating point is a direct extension left to future work. (iv) The calibration-transfer ablation runs on the 0%-loss substrate, so the packet-loss $\times$ calibrator-mismatch interaction is untested; since loss compounds error we expect transfer to be at least as fragile under loss, but did not test it. (v) The methodology is demonstrated on *two* frozen backbones, AASIST and RawNet2 Tak et al. (2021): all four phenomena replicate on a detector $\sim 5\times$ weaker, with the tile-padding artifact *larger* there — evidence these are properties of the method, not one model. Still, AASIST and RawNet2 are both raw-waveform networks from a related anti-spoofing lineage; a genuinely out-of-family test would use an SSL or spectrogram backbone (e.g. a wav2vec 2.0-based detector), and absolute magnitudes remain backbone-specific.

6 Conclusion

We studied telephone-channel voice-deepfake detection not as clip classification but as a streaming sequential decision: a frozen detector wrapped to consume audio as a growing prefix, report a per-duration calibrated confidence, and commit only when confident, abstaining otherwise. It adds no training, so the object of study is the decision policy and its measurement.

The channel dominates: under four wideband codecs the same detector’s EER ranges from near-clean ($\sim 0.8\%$) to severely degraded ($\sim 29\%$) — EVS most transparent, Speex-WB harshest, an Opus cliff between, packet loss compounding all — so “telephony-robust” means little without naming the codec. *Streaming buys latency and deferral:* it roughly halves decision latency, and under harsh channels its error at a fixed operating point falls below the full-utterance EER as an operating-point-and-selection effect — a pattern holding across three harsh channels (two failure mechanisms, three codecs) and a weaker backbone (RawNet2) (Sec. 4.3). What holds channel-independently is latency and a calibrated deferral signal. *Calibration must match the channel:* a clean-trained calibrator is $3\text{--}3.5\times$ overconfident on harsh low-bitrate channels, and because the failure is asymmetric (overconfidence dangerous, over-timidity merely abstaining), fitting calibration on a representative harsh condition is the safe default.

The analysis uses two frozen raw-waveform detectors but a single dataset of codec-augmented benchmark speech, not real calls, and does not address robustness to unseen attacks — the scope within which these measurements hold. The most valuable next step is to repeat it on genuine phone-channel audio and with more attack-robust backbones; because the layer is composable with the detector, such a substitution needs no change to the method.

Declarations

Funding. This research received no external funding.

Competing interests. The author declares no competing interests.

Ethics approval. Not applicable; the study uses a public benchmark corpus and no human subjects.

Availability of data and materials. The source code and experimental result artifacts supporting the conclusions of this article are available in the Zenodo repository, <https://doi.org/10.5281/zenodo.21108623>.

Project name: Reproducibility archive for “How Fast, How Sure, and Where It Breaks: A Streaming Sequential-Decision Layer for Voice-Deepfake Detection under Real Telephony Codecs”

Project home page: <https://doi.org/10.5281/zenodo.21108623>

Archived version: DOI 10.5281/zenodo.21108623

Operating system: Platform independent

Programming language: Python

Other requirements: see repository README (PyTorch, CUDA; environment specified therein)

License: MIT

The ASVspoof 2019 LA corpus and the EVS (3GPP TS 26.443) and AMR-WB (3GPP TS 26.173) reference codec binaries are not redistributed; they are available from their respective sources (ASVspoof; 3GPP/ETSI) as noted in the repository README.

Authors’ contributions. IS conceived the study, designed and performed all experiments, analyzed the results, and wrote the manuscript.

Acknowledgements. The author used large language model assistants (Anthropic Claude) to assist with code review and manuscript editing. All experimental design, implementation, analysis, and scientific claims are the author’s own, and all reported numbers were verified against source experiment data.

References

- Shi, H., Shi, X., Dogan, S., Alzubi, S., Huang, T., Zhang, Y.: Benchmarking audio deepfake detection robustness in real-world communication scenarios. In: Proc. 33rd European Signal Processing Conference (EUSIPCO) (2025). arXiv:2504.12423
- Shi, H., Shi, X., Dogan, S., Huang, T., Zhang, Y.: Audio deepfake detection at the first greeting: “Hi!”. In: Proc. IEEE ICASSP (2026). arXiv:2601.19573

- Phan, H., Koch, P., McLoughlin, I., Mertins, A.: Enabling early audio event detection with neural networks. In: Proc. IEEE ICASSP, pp. 141–145 (2018). arXiv:1712.02116
- Zhao, X., Zhang, X., Cheng, W., Yu, W., Chen, Y., Chen, H., Chen, F.: SEED: Sound event early detection via evidential uncertainty. In: Proc. IEEE ICASSP (2022). arXiv:2202.02441
- Xue, J., Yi, Z., Huang, Y., Ren, Y., Chen, Y., Fan, C., Su, Z., Zhang, Y., Cai, B.: RTCFake: Speech Deepfake Detection in Real-Time Communication. arXiv preprint arXiv:2604.23742 (2026)
- Chinchmalatpure, P., Chinchmalatpure, S., Chavan, S.: Defense against synthetic speech: Real-time detection of RVC voice conversion attacks. Int. J. Res. Anal. Rev. (IJRAR) **12**(4), 102–109 (2025). arXiv:2601.04227.
- Jose, E.J., S., F.F.N., Babu, G., Latheef, L., M., S.N.: AudioNyx: Real-time detection of audio deepfakes in phone calls. International Research Journal of Modernization in Engineering Technology and Science (IRJMETS) **8**(3) (2026) <https://doi.org/10.56726/IRJMETS92631>
- Huang, P., Wang, S., Yao, S., Zhou, M.: Generalizable Speech Deepfake Detection via Information Bottleneck Enhanced Adversarial Alignment. arXiv preprint arXiv:2509.23618 (2025)
- Jung, J.-w., Heo, H.-S., Tak, H., Shim, H.-j., Chung, J.S., Lee, B.-J., Yu, H.-J., Evans, N.: AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In: Proc. IEEE ICASSP (2022). arXiv:2110.01200
- Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K.A., *et al.*: ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language **64**, 101114 (2020). arXiv:1911.01601
- Chow, C.K.: On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory **16**(1), 41–46 (1970)
- El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. Journal of Machine Learning Research **11**, 1605–1641 (2010)
- Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2017). arXiv:1705.08500
- Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., Larcher, A.: End-to-end anti-spoofing with RawNet2. In: Proc. IEEE ICASSP (2021). arXiv:2011.01108

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proc. 34th International Conference on Machine Learning (ICML), pp. 1321–1330 (2017). [arXiv:1706.04599](https://arxiv.org/abs/1706.04599)