TECHNICAL WHITE PAPER · V1.0 · 27 AUGUST 2026

Analyzing voice calls for fraud from a pass-through SIP hop

> Open Voice Shield is a SIP pass-through service. Your switch sends the call through it as an ordinary hop, and the media is relayed and captured. After the call clears, the audio is transcribed and scored, and the platform returns **block**, **review** or **allow** with the reasons behind it. This paper covers the threat model, where it sits in the call path, what each pipeline stage does, what we measured, and what it does not do.

DOCUMENT	ANALYSIS STARTS
`v1.0`	`after BYE`
27 Aug 2026	nothing is injected into live audio

SIGNALLING	ANALYZED AUDIO
`SIP`	`G.711`
UDP, TCP; TLS on onward legs	A-law and μ-law today

SECTION 1

Scope, and who this is for

This paper is written for the people who will deploy it: NOC engineers, call-centre operations managers, and carrier or wholesale voice teams. It assumes you read SIP ladders, know what a `pcap` is, and will not accept a claim without a method behind it.

Two rules govern every number on this site. We publish only figures we measured, with sample size, date and hardware attached. And we publish no fraud-detection accuracy, precision or recall figure, because we have no labelled corpus of real fraudulent and legitimate calls. What we can show — latency, transcription error against a known script, and how far independent analysis tiers agree on the same calls — is on the benchmark page (/benchmark), with the raw data.

SECTION 2

Threat model

What the service is designed to catch, what the adversary controls, and where the boundary of the design sits.

The adversary

A human or scripted caller placing outbound calls through a trunk you operate or terminate. They control the caller ID they present, the script they read and how long they stay on the line. They do not control your switch, and do not know the call is being analyzed.

The harm

Impersonation of a bank, card issuer, government agency, utility or tech-support desk. Harvesting of card numbers, security codes, bank details or one-time codes. Prize and debt-collection pressure. Telemarketing that ignores identification, purpose and do-not-call rules. For the operator carrying the traffic, the harm is regulatory exposure and terminating-carrier complaints as much as the victim's loss.

The observation point

A pass-through hop sees the signalling and, because the media is relayed, both audio directions on separate channels. It sees nothing the parties did outside the call. It decides after the call clears, so it protects the *next* call, not the one it just heard.

Explicitly out of scope

NOT ADDRESSED	WHY, AND WHAT AN OPERATOR SHOULD DO INSTEAD
Stopping the call in progress	The verdict exists after the call clears. Enforcement applies to later calls: an automation rule can block or divert an ANI for a chosen window as soon as a verdict lands.
Caller-ID attestation	The platform records the ANI presented and reasons about the audio. It does not validate or issue a signed attestation. Keep your existing attestation controls.
Synthetic-voice and voice-clone detection	Not implemented. Analysis reads what was said, not how the voice was produced.
Adversarial evasion	A caller who knows they are being transcribed can code-switch, whisper, use jargon or push details to SMS. We have not measured how far that degrades a verdict.
Codecs beyond G.711	G.711 A-law and μ-law are decoded natively. Media the decoder cannot read is reported as an unsupported-codec error on that call rather than guessed at.
Content the call never carried	No enrichment from external reputation feeds, credit data or social sources. The verdict rests on this call's audio and its metadata.
Design boundaries as of 27 August 2026. Anything not listed as shipped is not claimed.	

SECTION 3

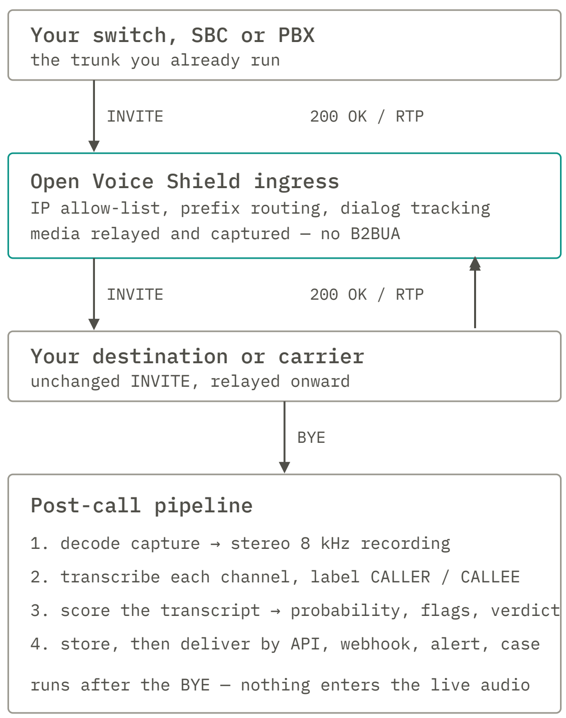
Where it sits in the call path

The ingress is an ordinary SIP hop. Your switch sends it the INVITE. The source IP is matched against the per-tenant allow-list, routing is by prefix so a tech prefix or a dialled range selects the destination, and the INVITE is relayed to the destination you configured — usually your own equipment or your onward carrier. No back-to-back user agent invents a new dialogue: the far end sees your call, not ours.

Media is relayed so it can be captured, one capture per stream. That is what makes a two-channel recording — caller left, callee right — possible without conferencing or injection. Nothing is played into the call, and no audio is added, removed or transcoded on the way through.

Two operational behaviours matter to a NOC.

- The ingress **fails closed**. If the authorization API cannot be reached, or returns something it cannot parse, the INVITE is answered `503 Service Unavailable`, so your switch takes its normal alternate route instead of losing calls silently.
- A prepaid balance is enforced as a per-call maximum duration, returned at authorization time. The dialogue ends at the cut-off rather than running up an unbounded debt.



Signalling and media pass through; the analysis happens afterwards, off the call path.

SECTION 4

The pipeline, stage by stage

Five stages run after the dialogue ends. Each is a queued job, so a burst of hangups queues rather than dropping work. A transient failure is retried with backoff; a configuration error fails the call's analysis at once, and says why.

STAGE	INPUT → OUTPUT	FAILURE BEHAVIOUR
1. Capture	Relayed RTP → one capture per media stream, kept with the call	No capture found after three retries marks the call failed with `recording_not_found`, rather than analyzed from nothing.
2. Decode	Capture(s) → stereo 8 kHz 16-bit recording, caller left, callee right	Timing is rebuilt from RTP timestamps, so loss and silence suppression become silence rather than drift; unreadable payload types stop the call with `unsupported_codec`.
3. Transcribe	Recording → speaker-labelled transcript with per-segment timestamps	Each channel is transcribed separately, which gives reliable `CALLER:` / `CALLEE:` attribution instead of guessed diarization. Silent channels are skipped.
4. Analyze	Transcript + call metadata → structured verdict (schema below)	The declared JSON schema is required. A refusal or unparseable answer becomes a `REVIEW` verdict, never a silent `ALLOW`.
5. Deliver	Verdict → stored record, signed webhook, alert email, case, share link	Webhook delivery is signed and retried. A webhook that cannot be delivered never blocks the verdict from being stored.

Pipeline stages. Measured stage latencies are on the benchmark page (/benchmark); this table describes what each stage does and how it fails.

Why the analysis is fixed work, not per-minute work

Transcription scales with call length. Analysis does not: it reads a transcript, not audio, and most of its input is fixed instructions. In the benchmark run, analysis latency across calls of 40 to 73 seconds stayed inside one band per tier, while transcription time tracked call length directly. That matters when you size a queue. Per-tier figures are on the benchmark page (/benchmark).

SECTION 5

What a verdict contains

Every analyzed call produces the same structured record, returned by the API, sent to your webhook and rendered in the console. Nothing in it is free-form except the fields typed as prose.

FIELD	TYPE	MEANING
`recommendation`	`block | review | allow`	The action the platform recommends for this ANI.
`probability`	`integer 0–100`	How likely the call is a scam or unlawful telemarketing.
`category`	`enum`	Card scam, tech-support scam, government or bank impersonation, debt collection, prize or lottery, utility, compliant or non-compliant telemarketing, robocall, legitimate business, personal, unknown.
`entity`	`string`	The organization the caller claimed to represent; empty if they named none.
`caller_identity`	`string`	The identity the caller stated for themselves.
`red_flags`	`string[]`	Numbered observations, each tied to something said on the call.
`tcpa_comment`	`prose`	Identification, purpose, opt-out handling and misrepresentation, in TCPA / TSR terms.
`summary`	`prose`	What happened on the call, in a few sentences.
`title`	`string`	A one-line description usable as a case title.
`language`	`string`	Language code detected for the conversation.
`keywords`	`string[]`	Terms that drove the classification, for search and rules.

Verdict fields, as returned per call.

Alongside the verdict, the call keeps its recording, its speaker-labelled transcript with timestamps, its SIP ladder and its capture, so an analyst can check the reading against the evidence. All of it downloads as one offline bundle.

SECTION 6

Acting on a verdict

Automation rules

Match on the verdict and act on the ANI: block it, or divert it to a destination you choose, for a window you set. This is where a verdict becomes enforcement.

Cases

Group calls into an investigation with an assignee, comments and e-mail notification, so a fraud team works a pattern, not a list of rows.

Alerts and webhooks

Alert rules deliver by e-mail on conditions you define. Outbound webhooks are signed so your receiver can verify them, and go to a pre-validated address.

Sharing evidence

A read-only link with a per-link password shows a call to someone without an account. The export bundle packages report, recording, capture and SIP log for offline review.

Voice AI forwarding BETA

Forwarding a call to a voice AI destination is in beta. It is not part of the measured pipeline described in this paper.

SECTION 7

What we measured

The headline results are here. Method, corpus and raw data are on the benchmark page (/benchmark).

CORPUS **11** calls synthetic, scripted, 40–73 s each	DECODE **151** ms median, 73 s two-leg capture
TRANSCRIBE **0.15** × RT on-premise standard tier, median	ANALYZE **13.8** s highest tier, median per call
TIER AGREEMENT **90.9** % lowest pairwise agreement on the recommendation, n=11 calls	

Measured 27 August 2026 on a development machine, with the tiers reached over a normal internet connection. Hardware, method, sample sizes, the per-call table and the raw data file are on the benchmark page (/benchmark). These are that machine's numbers, not a production service level.

SECTION 8

What we have not measured yet

Stated plainly, because for this audience an honest gap is worth more than a confident number nobody can reproduce.

Detection accuracy on real fraud. No accuracy, precision, recall or false-positive rate is published, because we do not have a labelled corpus of real fraudulent and legitimate calls. Our corpus is synthetic and written by us, so any rate computed on it would measure our own script writing.

Multilingual performance. Everything measured so far is English. Transcription and analysis both accept other languages, and neither has been benchmarked in one.

Adversarial evasion. We have not tested a caller who deliberately evades transcription or scoring — code switching, deliberate mumbling, moving the ask to SMS, or splitting a script across several short calls.

Throughput at scale. Stage latencies here are single-call measurements on one machine. Sustained throughput, queue behaviour under a hangup burst and concurrency limits are not yet benchmarked.

Audio quality effects. Packet loss, jitter, low-bitrate transcoding upstream, hold music and IVR audio all affect transcription. The corpus is clean synthesized speech, so it does not exercise any of that.

Long-run verdict stability. Analysis models are updated by their vendors. We have not yet run the same corpus over time to quantify how much a verdict drifts when a model version changes.

SECTION 9

Handling of recordings and evidence

Recordings and captures are stored on the operator's own disk, or in a bucket the operator controls with server-side encryption and public access blocked. Retention is configurable, and expiry is real deletion. Every endpoint scopes a call to its own tenant, and security-relevant actions go to an audit log against the acting user.

The full control list is on the security page `(/security)`: ingress authentication, credential handling, webhook signing and SSRF protection, share-link protection, rate limiting.

SECTION 10

Reproducing this

The harness, the corpus and the raw result file are published so these numbers can be checked rather than believed. The benchmark page `(/benchmark)` lists the corpus with a SHA-256 per file, the exact command, the hardware, and the per-call results behind every aggregate.

Document v1.0, published 27 August 2026 by Open Voice Shield. Figures dated where they appear. Corrections: support@telecomsxchange.com.