

Audio Front-End Decoupling: Why Synthetic-Voice Simulators Fail to Predict Real-World Voice Agent Behavior

Shoutrik Das Suraj Agrawal Aashraya Sachdeva

Observe.AI

{shoutrik.das, suraj.agrawal, aashraya.sachdeva}@observe.ai

Abstract

Large Language Model (LLM)-based voice agents are routinely cleared for deployment using synthetic audio regression suites, with engineering teams treating augmented commercial text-to-speech (TTS) as a safe proxy for real human callers. This paper presents the first systematic approach to quantify the profound behavioral drift that occurs when agents optimized on simulated audio are deployed in the real world. We demonstrate that the assumption of acoustic equivalence fundamentally fails: real production audio triggers 2–13× more VAD misfires than synthetic data across Silero, WebRTC, and TEN-VAD, while simulated environments cause text-grounded turn detectors to artificially flip decisions 33% of the time, compared to just 3% in reality. A paired four-vendor TTS calibration sweep proves this misranking is structural: on every front-end component, optimizing thresholds for synthetic audio actively degrades real-world performance. To prevent teams from flying blind, we introduce a novel set of signal-layer metrics and an open-source framework to audit any simulator prior to adoption. Rather than testing the agent, our method tests the simulator against real target-domain interactions. By identifying where a simulator’s acoustics deviate from reality, teams can definitively know which front-end tuning decisions are safe to trust.

1 Introduction

Voice agents driven by large language models now run inbound and outbound calls at scale across customer service, telesales, healthcare scheduling, and accessibility. Before deploying one, a team needs to know how it will behave in production, because a failure once live costs missed calls, customer churn, and compliance exposure. The standard answer is a sandbox that exercises the agent end-to-end before it goes live.

Today that sandbox almost always simulates the agent’s *decisions*. Whether it calls the right tool, retrieves the right record, and follows the right flow through the call. Zhou et al. (2026) formalise this as the user-simulator-to-real gap, and benchmarks such as τ -bench (Yao et al., 2024) measure the same axis. The shared assumption is that an agent’s failures are reasoning, tool-selection, or policy failures.

Those simulators target the agent’s *decision layer*, what it chooses to do given the conversation. Production exposes a second class of failure, at the *user-engagement layer*, how the agent manages the real-time exchange with the caller. The LLM is not responsible for it, and no chat-style evaluation surfaces it. People talk to a voice agent as they would to a person, and expect it to behave like one: to respond when they finish, to stay quiet while they are still speaking, and to stop the moment they break in. These are turn-taking expectations: the agent’s reasoning can be correct and still fail them, and a decision-layer simulator will not notice.

These conversational-experience failures motivated *audio simulation*: methods that render the user side of a conversation as audio and replay it against the deployed agent in the same streaming loop a real caller would use. τ -Voice (Ray et al., 2026) is one such framework, and read-speech corpora such as VoiceBench (HLT-Lab, 2024) serve a similar role. The LLM core still decides what to say and which tool to call. The signals it consumes, though, come from an upstream *audio front end*: voice activity detector (VAD), turn-detection classifier, noise filter, automatic speech recognition (ASR). When the front end emits the right onset, offset, and end-of-turn signals, the failures above do not occur. Audio simulation is, in effect, a test of that front end.

Audio simulation assumes the synthetic audio it feeds the agent represents the audio it will see in production. This assumption is untested. In this

work, we test it, and find that it fails across all three universal front-end components at once.

We decompose the audio front end into voice activity detector, noise filter, and turn analyzer and measure each on five datasets (two open call-centre corpora, two synthetic suites, and an internal log of 4,999 production calls spanning 6 customer deployments; section 3), three streaming VAD architectures (Silero (Silero Team, 2024), WebRTC (Wise-man, 2016), TEN-VAD (TEN-framework, 2025b)), and an end-to-end voice-to-voice model with no explicit VAD (Moshi (Défossez et al., 2024)). Each component fails on synthetic in its own characteristic way, and across all three the operating point that scores best on synthetic audio is the wrong one to deploy. For the VAD, real production audio misfires at 2–13 $\times$ the rate of synthetic audio across Silero, WebRTC, and TEN-VAD (Table 1): the VAD fires on background noise after the user’s last word and the agent waits in silence for a user already done. For the noise filter, the production internal NF reduces WebRTC misfire 24pp on production but only 6pp on synthetic, and the same filter has the opposite word-error-rate (WER) effect on the two corpora (§A.6). The same direction reproduces in Moshi, which has no explicit VAD to blame. The effect is sharpest when thresholds are tuned directly: a grid-search on the synthetic regression suite (section 3) lands every component’s argmin-loss threshold at a value real audio does not prefer (Figure 1); a team that calibrates or A/B-tests a front end on synthetic audio ships the configuration that fails on real calls (section 5.3).

The same experimentation produces **UASI** (User-Audio Simulator Index), a five-metric signal-layer audit that scores a candidate simulator against a real reference. We calibrate the per-metric pass threshold against a real-vs-real null, with zero false positives in 200 random within-corpus splits.

Contributions.

- An empirical characterisation of audio-front-end decoupling across three universal component types, three VAD architectures, and a voice-to-voice replication.
- A paired voice-clone experiment holding content, speaker, timing, channel, gain, and noise constant, isolating the residual gap to synthetic speech itself (section 5.4).
- **UASI**, a calibrated 5-metric signal-layer audit shipped as `uasi-toolkit` (Das et al., 2026).

The rest of the paper develops these results. Section 2 places the work against decision-layer simulation; Section 3 and section 4 state datasets and decomposition; Section 5 reports the conversational-experience axis; Section 6 reports the acoustic signal-layer axis; Section 7 formalises UASI; Section 8 states deployment recommendations; Section 9 concludes.

2 Related Work

Text-level agent evaluation. Agent evaluation currently treats conversations as pure sequences of natural-language text. A large literature evaluates LLM agents against simulated text-users (Yao et al., 2024; Zhou et al., 2024; Chen et al., 2024). Even efforts to formalize the sim-to-real gap abstract the acoustic substrate away entirely. This is true of work that scores scripted personas (Gromada et al., 2025) and of the LLM-agnostic User-Sim Index (USI) (Zhou et al., 2026). Closest to us, Devanathan et al. (2025) report the same ordering one layer up: synthetic contact-centre transcripts diverge from real ones on a 17-metric diagnostic panel, and prompts tuned on real transcripts outperform prompts tuned on synthetic ones.

Synthetic audio proxies. When evaluation does move to audio, it relies heavily on synthetic proxies. Public benchmarks like VoiceBench (HLT-Lab, 2024) and end-to-end models like Moshi (Défossez et al., 2024) score performance on TTS-rendered test sets, which leaves the synthetic-vs-real gap unquantified. Production frameworks (Daily, 2024; LiveKit, 2025; TEN Framework, 2025) and commercial simulators provide testing environments but lack standardized audits for this divergence. Standard TTS evaluation metrics (Reddy et al., 2021) measure human perceptual quality. That is a different target from ours, which is machine perception: whether a synthetic clip correctly triggers a VAD or a noise filter.

Lessons from robust speech processing. Traditional speech processing has long known that synthetic audio fails to mimic real-world acoustics. Models trained on synthetic noise-plus-speech mixes reliably underperform on real recordings (Reddy et al., 2023; Barker et al., 2018). This prompts standard workarounds such as room-impulse-response augmentation (Ko et al., 2017; Ravanelli et al., 2015) and explicit warnings about corpus mismatch (Snyder et al., 2015). Learned de-

noisers frequently fail via over-suppression on mismatched corpora (Pandey and Wang, 2020; Gonzalez et al., 2023), and speech enhancement artifacts degrade downstream ASR more than raw noise (Iwamoto et al., 2022). This signal-level skepticism has not transferred to voice-agent evaluation.

This work. Isolated studies audit specific pipeline boundaries, such as the ASR/NLU junction (Chen et al., 2024), but front-end audio audits remain rare. Our work connects text-level gap evaluation (Yao et al., 2024; Zhou et al., 2026; Gromada et al., 2025) with the speech mismatch literature (Ko et al., 2017; Reddy et al., 2023; Pandey and Wang, 2020). The point it exposes is concrete: evaluating voice agents on simulated audio does not predict real-world conversational dynamics.

3 Datasets

Five datasets across two groups; details in §A.2.

Real audio.

- **Production:** $n=4,999$ deployed-agent calls across 6 enterprise customer deployments. μ -law 8 kHz mono. Internal data agreement; no audio or transcripts released, aggregate statistics only.
- **HVB:** Harper Valley Bank (Wu et al., 2020), public call-centre corpus; 1, 500 clips.
- **AppTek** (AppTek, 2024): public English call-centre dialogue set; channel-2, first 60 s per call, 1, 500 clips.

Synthetic audio.

- **TAVC** (TTS-Augmented Voice Corpus): a synthetic dataset combining 4 commercial TTS models (ElevenLabs, Hume, Deepgram, OpenAI) across many voices, processed through the τ -Voice augmentation protocol (Ray et al., 2026). $n=6,000$ clips across seven scripted scenario templates (S1–S5); per-template definitions and the augmentation chain in §A.2.
- **VoiceBench** (HLT-Lab, 2024): second synthetic dataset of 5, 076 clips, sampled from the Google-TTS splits of VoiceBench and processed through the same τ -Voice augmentation protocol as TAVC; subset selection and chain detail in §A.2.

4 Method

Production voice agents (Pipecat (Daily, 2024), LiveKit Agents (LiveKit, 2025), TEN (TEN Framework, 2025), and our internal pipeline) vary in architecture but rely on three core components: a **VAD** (mandatory, audio-only), a **noise filter** (optional, varying placement), and a **turn analyzer** (audio- or text-grounded depending on the framework). Because comparing full pipelines confounds component behavior with assembly choice, we evaluate each component independently on synthetic versus real audio.

This isolated evaluation reveals characteristic synthetic-audio failures that whole-assembly tests obscure. We first report the *conversational-experience* impact of these isolated components (section 5), followed by the *acoustic* signal-layer differences that drive it (section 6). Pipeline, dataset, ground-truth, and metric details are deferred to §A.

5 Conversational Experience

Acoustic simulation gaps cause two primary user-visible failures: *silence* (agent fails to respond) and *overtalk* (agent interrupts). These failures map onto specific components in the front end: silence is driven by false VAD onsets and by noise filters that fail to clean background noise, leaving the VAD in a SPEAKING state long after the user has stopped; overtalk is driven by turn analyzers that fire COMPLETE during a mid-turn pause, releasing the agent before the user is done. Benchmarking these front-end components reveals that optimizing thresholds on synthetic audio selects incorrect operating points, actively degrading real-world deployment. This representation gap persists even in end-to-end voice-to-voice architectures (§A.11).

5.1 Silence

A false VAD onset on background noise after a user turn ends delays the agent’s response, leaving the user listening to silence while the agent waits for a turn already finished. Table 1 shows that real production audio triggers 2–13 $\times$ more false onsets than TAVC across all tested streaming VAD architectures due to acoustic-layer mismatches (section 6).

Inserting a noise filter before the VAD fixes real-world misfires, but synthetic evaluation hides this gain (Table 2). The internal NF reduces WebRTC misfires by 24.2 percentage points (pp) on Produc-

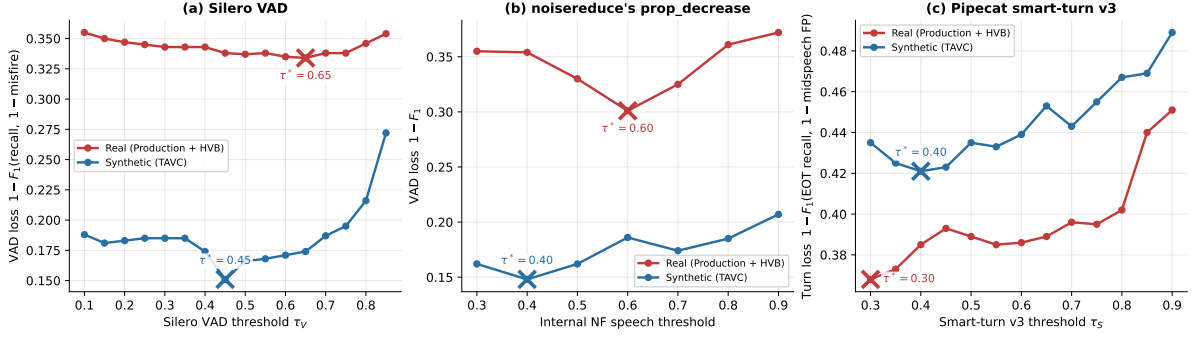


Figure 1: Calibration misranking across the three universal front-end components. Each panel plots $1 - F_1$ loss against the component’s decision threshold for Real (red, Production + HVB mean) and Synthetic (blue, TAVC mean). Crosses mark each line’s argmin-loss threshold. Panel (b) holds the upstream VAD fixed at panel (a)’s synthetic-optimum $\tau_V = 0.50$. Corpus construction, per-panel F_1 definitions, and full numerics in §A.4.

Dataset	Silero	WebRTC	TEN-VAD
Production	30.6%	61.7%	67.6%
HVB	13.9%	53.3%	26.4%
AppTek	22.4%	57.9%	40.6%
VoiceBench*	3.8%	17.7%	11.0%
TAVC*	2.4%	30.0%	17.1%

Table 1: Per-clip VAD misfire rate (§A.2). Real Production audio shows 2–13 $\times$ the misfire rate of TAVC* on every VAD architecture. **Bold** marks the highest misfire rate per VAD architecture; * marks synthetic datasets.

tion but only 6.1pp on TAVC; noisereduce actually increases synthetic misfires (+1.4pp). This happens because synthetic suites lack the complex, time-varying noise structures these filters target (§A.6). A third-party denoiser trained on the DNS corpus, by neither us nor on our data, collapses synthetic misfire from 10.8% to 0.0% while moving Production only from 24.5% to 21.1%. The asymmetry therefore does not depend on our filter’s training data.

Dataset	internal NF	noisereduce
Production	−24.2pp	−12.8pp
HVB	−39.6pp	−28.2pp
AppTek	−38.6pp	−28.3pp
VoiceBench*	−6.8pp	−8.7pp
TAVC*	−6.1pp	+1.4pp

Table 2: Change in pooled WebRTC misfire rate after NF-before-VAD insertion (Placement A). **Bold** marks the cell where NF makes synthetic audio *worse* (regression); * marks synthetic datasets.

5.2 Overtalk

Overtalk occurs when a turn analyzer falsely signals COMPLETE during a mid-turn pause. Frame-

works deploy either *audio-grounded* (Pipecat smart-turn v3 (Pipecat, 2025), Whisper-Tiny over PCM) or *text-grounded* (LiveKit turn-detector (LiveKit, 2025), Qwen2.5-0.5B over STT text) gates. Both inherit synthetic-audio evaluation biases via distinct cascades.

Audio-grounded. As shown in Table 3, TAVC overshoots real Production Turn-End Recall by 14pp and Mid-Turn Pause Recall by 21pp because TTS voices generate clean, artificial prosodic markers (e.g., clear F0 declination) omitted in messy real conversations (section 6).

Metric	Production	TAVC*
Turn-End Recall	0.83	0.97
Mid-Turn Pause Recall	0.71	0.92

Table 3: Pipecat smart-turn v3 (Pipecat, 2025) recall against turn boundaries that are manually annotated on Production and exact by construction on TAVC*. **Bold** marks the inflated synthetic recall scores (14pp / 21pp above real); * marks synthetic datasets.

NFs also destabilize classifiers on synthetic audio: under the internal NF, turn decisions flip on **36.7%** of TAVC clips vs. just **10.0%** on Production. Lacking real noise, the filter misidentifies low-formant TTS frames as noise, suppressing the speech substrate itself (§A.6).

Text-grounded. Text-grounded analyzers process only STT transcripts, so an NF affects them only by changing what STT outputs. Thresholding LiveKit’s per-clip $P(\text{COMPLETE})$ at the no-NF dataset median ($n=30$ each), the internal NF flips $1/30 = 3\%$ of Production decisions, $2/30 = 7\%$ on HVB, and $10/30 = 33\%$ on TAVC, an $11\times$ jump. noisereduce flips 3% on Produc-

tion and 7% on TAVC. At $5\times$ the sample ($n\approx 150$ per dataset) the internal-NF rates are 20.0% [14.2, 27.4] on TAVC* and 4.0% [1.9, 8.5] on Production (Wilson 95%); the intervals do not overlap. The asymmetry confirms the chain: the internal NF strips low-formant synthetic speech mistaken for noise, corrupting the transcript itself (§A.7), whereas noisereduce’s musical-noise residual disrupts audio-grounded readers but transcribes as nothing.

5.3 Optimization on Synthetic Targets the Wrong Threshold

Teams calibrate VAD, NF, and turn-analyzer thresholds by grid-searching on synthetic regression suites. We test that tuning loop directly: sweep each component’s decision threshold on a paired real-and-synthetic corpus and check whether the synthetic-optimal threshold matches the real-optimal one (Figure 1; §A.4 for corpus, F_1 definitions, and per-panel numerics). The argmin-loss thresholds disagree on every component: Silero VAD 0.50 (synth) vs 0.65 (real); internal NF 0.40 vs 0.60; smart-turn v3 0.40 vs 0.30. For VAD and NF the synthetic curve sits well below the real curve at every threshold, so the synthetic audit also *underestimates the absolute production loss* in addition to misplacing the optimum. A team that ships a synthetic-tuned configuration lands at the wrong threshold for production on every component.

5.4 Synthetic Speech, Not the Corpus

The comparisons above audit the simulator as deployed, voices and augmentation chain included, so they establish that the tested simulators mislead rather than that synthetic speech itself causes the gap. To separate the two we rebuilt the synthetic side under maximal matching. For 60 HVB calls (30F/30M) we cloned each caller’s voice, re-rendered that caller’s exact transcripts, and placed each render at the human utterance’s own offset, so content, speaker timbre, pause structure, duration, and ground truth are identical on both sides by construction; both sides then pass through the same μ -law codec. Two further conditions add per-utterance gain matching and the clip’s own measured noise floor (C0–C2, §A.12).

Under full matching the signal panel largely closes: the F0+F1 gap is indistinguishable from zero ($p=0.17$), segment duration closes ($p=0.078$), and spectral flatness falls from +0.51 to +0.058, though it remains significant. Front-end behaviour

does not follow (Table 4). Silero misfires 6.4pp more often on the synthetic side and smart-turn raises mid-turn false-COMPLETE by 15.5pp, while WebRTC misfires 32.7pp more on the *human* side, whose channel carries breaths and transients the clone does not reproduce. TEN-VAD is the informative negative: its gap shrinks monotonically as matching tightens ($-11.2 \rightarrow -5.5 \rightarrow -2.2$ pp) and is not significant under full matching, which is what a working matching pipeline should do to a confound-driven difference.

Every confound named in the reviews is held constant here, so what remains is a property of synthetic speech itself. Its sign is component-dependent, which means no single normalisation turns a TTS corpus into a drop-in proxy for the front end.

Measure (synthetic – human)	Δ	p
F0+F1 energy (M1)	−0.006	0.17
Segment duration (M5)	−0.08 s	0.078
Spectral flatness (M4)	+0.058	< 0.001
Silero misfire	+6.4pp	0.001
smart-turn mid-turn FP	+15.5pp	0.001
WebRTC misfire	−32.7pp	< 0.001
TEN-VAD misfire	−2.2pp	0.11

Table 4: Paired voice-clone experiment under full matching (codec, gain, and noise), $n=60$ calls; median paired delta and Wilcoxon signed-rank p . The signal panel largely closes while three of four detectors still diverge, in component-dependent directions. **Bold** marks the surviving behavioural gaps; TEN-VAD’s gap does not survive matching.

6 Acoustic Analysis

We measure five per-clip signal-layer properties (the Audio Signal Feature Summary, ASFS, panel) across two dataset groups and report the pooled-real-IQR effect size g per cell (Table 5). The three real datasets pass $|g|\leq 1$ on every cell and establish an internally-consistent real reference; each synthetic source is disjoint ($|g|>2$) on two dimensions. A per-frequency-band breakdown of M1’s spectral energy appears in Appendix A.9.

M1: formant-band energy. Production parks 70.5% of energy in 200–800 Hz (F0+F1); HVB 73.4%, AppTek 74.1%. TAVC parks only 13.5% there and inverts the rest into the 800–1500 Hz presence band. The deficit is an artefact of voice-specific TTS rendering rather than of the augmentation chain (per voice, M1 spans 0.05 to 0.50

Dataset	Pooled-real-IQR effect size $g = (\text{med}_S - \text{med}_R)/\text{IQR}_R$				
	F0+F1 energy fraction (M1)	voicing fraction (M2)	frame RMS mean (M3)	spectral flatness mean (M4)	seg. duration mean (M5)
Production	-0.17	-0.31	-0.19	+0.39	-0.16
HVB	+0.01	-0.08	-0.05	-0.35	-0.18
AppTek	+0.06	+0.53	+0.18	-0.00	+0.61
VoiceBench*	-0.03	+ 3.56	+ 3.92	-1.15	+1.77
TAVC*	- 3.73	+1.69	+ 3.19	-1.12	+0.41

Table 5: Per-clip ASFS effect size g for each dataset against pooled-real reference $R=\text{Production}+\text{HVB}+\text{AppTek}$. **Bold** marks $|g|>2$ (synthetic-real disjoint); * marks synthetic datasets. Each synthetic source is disjoint on two different dimensions: TAVC* on F0+F1 energy (M1) and frame RMS (M3); VoiceBench* on voicing fraction (M2) and frame RMS (M3).

against a real reference of 0.73, while the chain moves M1 by less than 0.1σ ; §A.10): VoiceBench (wideband read speech) lands at 72.7%, inside the real range, and the inverted pattern reproduces across ElevenLabs, Deepgram, and Hume (§A.10). Learned speech-vs-noise classifiers trained on real wideband speech misclassify low-F0+F1 frames as noise (Pandey and Wang, 2020; Gonzalez et al., 2023; Reddy et al., 2023), the mechanism behind the section 5 STT WER measurements.

M2: voicing structure. Production median voicing fraction is 0.099, HVB 0.127, AppTek 0.202. VoiceBench is **0.573** ($5\times$ overshoot, $+3.56\sigma$), TAVC 0.343 ($3.4\times$, $+1.69\sigma$). TTS produces near-continuous voicing within utterances, omitting the unvoiced/silence interleaving Brady (1968) documents on real telephony, the read-vs-conversational gap of Bijwadia et al. (2022).

M3: dynamic range. The μ -law codec (ITU-T, 1988) and handset preamps compress real frame-RMS into a narrow low band (Production 0.010, HVB 0.012, AppTek 0.015). Both synthetic sources sit $5\text{--}7\times$ higher (TAVC 0.059, VoiceBench 0.070) at $+3\sigma$ effect size. M3 is the most uniformly elevated metric across every TTS source we measured (ElevenLabs, Deepgram, Hume; §A.10). Dynamic-range matching is therefore the cheapest first-pass fix for any TTS-driven simulator. The cause is a calibration shift, not a classifier failure (Martin, 2001).

M4: spectral flatness. Real spectra are broader (Production 0.48, HVB 0.26, AppTek 0.36) than synthetic (~ 0.04 , $g \approx -1.15\sigma$) because of multi-source non-stationary background (Dafna et al., 2015; Ivry et al., 2019; Park et al., 2014; Ivry et al., 2020). Additive noise overlay does not close the gap; joint speech-noise processing via

room-impulse-response augmentation is the standard route (Ko et al., 2017; Ravanelli et al., 2015).

M5: segment-duration structure. Production median Silero segment duration is 1.41 s, HVB 1.39 s, AppTek 2.21 s. VoiceBench overshoots at 3.40 s ($+1.77\sigma$); TAVC at 2.00 s ($+0.41\sigma$, partially closed by explicit pause scenarios). The read-speech to conversational-speech transfer gap (Bijwadia et al., 2022; Skantze, 2021).

7 UASI: A Deployment-Time Simulator Audit

The five-metric ASFS panel of section 6 is also the operational interface of the **User-Audio Simulator Index (UASI)**, a pre-adoption audit a product team can run on any candidate simulator. UASI is the audio analog of the text-level User-Sim Index (Zhou et al., 2026).

Pass threshold. A simulator *passes* metric M_i iff $|g_i|$ falls inside a real-vs-real null. We calibrate that null out of sample. Auditing each real corpus against the other two, as if it were a candidate simulator, gives worst $|g|$ of 0.82 (Production), 0.51 (HVB), and 1.30 (AppTek); the same rotation over five of the six production deployments gives 0.75 to 1.20. Held-out real audio never exceeds $|g|=1.30$, and only M5 (segment duration) crosses 1.0, so the flat $|g|\leq 1$ bound is too strict for M5 and we recalibrate per metric on this measured spread. M1 and M3, which carry the VAD and noise-filter decisions, are unaffected. The failing metrics of the synthetic suites score 1.69 to 3.92, far outside the null (§A.8). efsec:appendix:controlled). Each metric is computed per-clip ($n\geq 30$) and takes seconds of CPU: M1/M3/M4 are FFT-based, M2 uses Praat, M5 uses Silero. A full UASI run against the HVB reference takes under five minutes on a laptop, fast

enough to serve as a pre-adoption gate rather than an offline analysis.

Deployment-decision-conditional rule. Because each metric maps to a specific component failure mode (section 6), UASI is not a single pass/fail verdict: it is conditional on the decision a team intends to make on the simulator (Table 6). A team that only tunes the VAD on a simulator needs M3 and M4 to pass; a team that A/Bs noise filters needs M1 and M4.

Deployment decision on simulator	Required passes
VAD threshold tuning	M3, M4
NF A/B (e.g., internal vs. nr)	M1, M4
Audio-grounded turn-tuning	M2, M5
Text-grounded turn-tuning	M1 (via STT)
STT WER A/B	M1, M3
End-to-end overtalk-rate prediction	M2, M5
End-to-end silence-rate prediction	M3, M4

Table 6: UASI rules: metrics that must pass for a simulator to give deployment-correct guidance on each task. Mappings ground in the section 6 mechanism citations.

Report card. Table 7 details UASI verdicts. Both TAVC and VoiceBench score 1/5, failing prerequisites for all front-end decisions (Table 6) along distinct axes: TAVC fails M1 via TTS-rendering artifacts, while VoiceBench passes M1 but fails M2 and M5 by omitting essential conversational structure.

Metric	TAVC*	VoiceBench*
M1 F0+F1 energy	-3.73 (fail)	-0.03 (pass)
M2 voicing	+1.69 (fail)	+3.56 (fail)
M3 frame RMS	+3.19 (fail)	+3.92 (fail)
M4 spectral flat.	-1.12 (fail)	-1.15 (fail)
M5 seg. duration	+0.41 (pass)	+1.77 (fail)
NF A/B safe?	no (M1, M4)	no (M4)
VAD-tune safe?	no (M3, M4)	no (M3, M4)
Turn-tune safe?	no (M2)	no (M2, M5)

Table 7: UASI report card: both simulators fail at least one required metric for every decision in Table 6. **Bold** marks failing metrics ($|g| > 1$); * marks synthetic datasets (both columns are synthetic, audited against the real reference).

Fail rejects, pass screens. The two directions of the audit carry different weight, and we measure both. Across 5 synthetic suites $\times$ 3 front-end decisions swept under the losses of §A.4, 14 of the 15 cells that failed a required metric selected a different operating point than the real reference; the one

exception is a tie. Failing is therefore a measured reason to reject a simulator for that decision. Passing is weaker: it removes the structural reasons to expect misranking without certifying that absolute deployment numbers transfer, and the paired corpus of section 5.4 still diverges behaviourally once matched. Behavioural-level closure remains a separate question (§C). Pass means "no structural reason to reject"; failure means "named, mechanism-grounded reason to reject."

Open-source. The audit ships as a pip-installable CLI (Das et al., 2026) that auto-downloads HVB as the default real reference; source, manifests, and reports for every dataset in this paper are included.

8 Industrial Applications

Our findings dictate three strict deployment rules (pipeline integration detailed in §B). First, mandate the UASI audit before adopting any synthetic regression suite. Second, pair component A/B testing with its corresponding UASI dimension (Table 6) to avoid synthetic-driven misranking (section 5.1). Finally, calibrate front-end thresholds exclusively on production holdouts, since synthetic data artificially depresses optimal settings (section 5.3), and reserve simulators solely for post-calibration regression testing.

9 Conclusion

Because industry frameworks often overlook the audio front end, optimizing against synthetic data consistently yields broken production configurations. Across three component types, three VAD architectures, and an end-to-end model, the synthetic optimum always fails in real-world deployment.

This divergence persists despite stringent matching. Even when providing the synthetic system with the caller’s cloned voice, identical timings, and matched noise profiles, the front end still diverges (Section 5.4). Thus, this gap stems from the inherent properties of synthetic speech itself, not corpora mismatches.

To address this, we introduce the User-Audio Simulator Index (UASI), a five-metric signal-layer audit to evaluate candidate simulators against real references before tuning. A failed metric justifies rejecting the simulator, while a pass acts as a screen without guaranteeing transferability. Validated on commercial TTS in English telephony, we have released this toolkit (Das et al., 2026) so teams can audit their own deployments.

Limitations

We list the limitations a reader should keep in mind when generalising the findings.

Production diversity. The production dataset is $n=4,999$ calls spanning 6 enterprise voice-agent customer deployments inside one platform, and the two open call-centre corpora (Harper Valley Bank and AppTek, 1,500 calls each) carry the accent and domain diversity it lacks. We do not test production audio from other agent verticals such as sales bots, IVR navigation, or voice copilots, so generalisation beyond call-centre telephony is conjecture. All five datasets are English, and the formant-band and voicing-fraction signatures the ASFS panel reads will move with the phonetic inventory of other languages, so its thresholds need re-grounding on target-language reference audio before the panel is applied outside English.

Synthetic-TTS coverage. Our synthetic evaluation rests on VoiceBench and TAVC, spanning four commercial TTS vendors (ElevenLabs (ElevenLabs, 2025), Hume, Deepgram, OpenAI) processed through the τ -Voice augmentation protocol. Open-source neural TTS, parametric synthesisers, and vendors outside this pool are untested, so whether the gap holds at the same magnitude for them is open. The ASFS panel is constructed to be vendor-agnostic and the deployment rule does not assume our particular gap signs.

Open VAD and turn-taking models. Our pipeline uses the open-source Silero VAD (Silero Team, 2024) and the publicly distributed Pipecat smart-turn v3 classifier (Pipecat, 2025), with cross-VAD evidence on WebRTC (Wiseman, 2016) and TEN-VAD (TEN-framework, 2025b). TEN Turn Detection (TEN-framework, 2025a) is the Qwen2.5-7B sibling of LiveKit’s Qwen2.5-0.5B and we expect the same mechanism, but we do not measure it. The end-to-end reproduction in Moshi (Défossez et al., 2024) runs on a small number of clips per source (§A.11); extending the paired-prompt design of section 5.4 to Moshi at scale is left to future work.

Per-component sample sizes. Every placement experiment in section 5 (VAD misfire, $NF \rightarrow VAD$, $NF \rightarrow STT$ WER, $NF \rightarrow$ turn-analyzer flip) is 7,999 clips for real group and 11,076 clips for synthetic group. The calibration sweep in §A.4 runs on a paired $n_R=500$ real (500 Production +

500 HVB) and $n_S=500$ TAVC (4 vendors $\times$ 40 prompts) corpus.

Residual confounds. The primary real-versus-synthetic comparison is observational: the corpora differ in speaker, content, channel, and noise. Section 5.4 removes those confounds by construction on 60 calls, and the per-clip matched-SNR and matched-gender re-aggregations in §A.8 rule out SNR mismatch and gender skew for the noise-filter result. We do *not* test the full Cartesian product of SNR, gender, accent, and noise type, and the paired corpus covers one vendor and one public source.

Ethical Considerations

The production-call corpus ($n=4,999$ calls across 6 customer deployments) consists of real customer-to-agent telephone interactions captured by an operational voice-agent deployment. We treat this data with the following controls.

Use basis. The calls were collected under the operational consent and data-use terms governing the deployment, which permit analysis of recordings for service evaluation and improvement. No separate ethics review was conducted. The study is non-interventional analysis of already-collected operational data, and the controls in this section are the operative safeguard. We release no recordings, no transcripts, no caller-identifying metadata, and no per-clip features, and every production number reported in this paper is an aggregate computed internally over the corpus.

Aggregation and de-identification. Only aggregate, clip-level distributional statistics computed from the audio (section 6) and aggregate event counts derived from the frame trails (section 5) appear in this paper. Calls are referred to internally by hashed short identifiers, and even those do not appear here.

Third-party and synthetic material. The Harper Valley Bank (Wu et al., 2020) and AppTek (AppTek, 2024) corpora are released by their authors under their own terms; we use only the public splits, unmodified beyond clipping each file to its first 60 s for tractable smart-turn inference. The synthetic-persona audio is generated through a commercial TTS service (ElevenLabs, 2025) the authors have licensed access to. Persona scripts and voice handles are research-internal, and voice-handle behaviour is reported only in aggregate.

References

- AppTek. 2024. AppTek multi-accent call center dialogues corpus. Hugging Face Dataset. https://huggingface.co/datasets/apptek-com/apptek_callcenter_dialogues.
- J. Barker, S. Watanabe, E. Vincent, and J. Trmal. 2018. The fifth ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines. In *Proc. Interspeech*, pages 1561–1565.
- S. Bijwadia, S.-Y. Chang, T. N. Sainath, B. Li, Y. Zhang, and Y. He. 2022. Unified end-to-end speech recognition and endpointing for fast and efficient speech systems. In *Proc. IEEE SLT*.
- Paul Boersma. 2001. Praat, a system for doing phonetics by computer. In *Glott International*, volume 5, pages 341–345.
- S. F. Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120.
- P. T. Brady. 1968. A statistical analysis of on–off patterns in 16 conversations. *Bell System Technical Journal*, 47(1):73–91.
- O. Cappé. 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, 2(2):345–349.
- Jiangning Chen, Ziyun Zhang, and Qianli Hu. 2024. Optimizing entity resolution in voice interfaces: An ASR-aware entity reference expansion approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1–7.
- I. Cohen. 2003. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5):466–475.
- E. Dafna, T. Rosenwein, A. Tarasiuk, and Y. Zigel. 2015. Voice activity detection in the presence of breathing noise using neural network and hidden Markov model. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.
- Daily. 2024. Pipecat: open-source framework for real-time voice and multimodal ai agents. <https://github.com/pipecat-ai/pipecat>.
- Shoutrik Das, Suraj Agrawal, and Aashraya Sachdeva. 2026. uasi-toolkit: open-source signal-layer audit toolkit for voice-agent simulators. <https://github.com/shoutrik-oai/uasi-toolkit>.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. <https://arxiv.org/abs/2410.00037>. Preprint, arXiv:2410.00037.
- Rishikesh Devanathan, Varun Nathan, and Ayush Kumar. 2025. Why Synthetic Isn’t Real Yet: A Diagnostic Framework for Contact Center Dialogue Generation. <https://arxiv.org/abs/2508.18210>. Preprint, arXiv:2508.18210.
- ElevenLabs. 2025. ElevenLabs v3: text-to-speech voice models. <https://elevenlabs.io>.
- P. Gonzalez, T. S. Alstrøm, and T. May. 2023. Assessing the generalization gap of learning-based speech enhancement systems in noisy and reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Justyna Gromada, Alicja Kasicka, Ewa Komkowska, Łukasz Krajewski, Natalia Krawczyk, Bartosz Przybył, Morgan Veyret, Lina Rojas-Barahona, and Michał K. Szczerbak. 2025. Evaluating conversational agents with persona-driven user simulations based on large language models: A sales bot case study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 230–245.
- HLT-Lab. 2024. VoiceBench: a multilingual read-speech benchmark for speech understanding. <https://huggingface.co/datasets/hlt-lab/voicebench>.
- ITU-T. 1988. ITU-T Recommendation G.711: Pulse Code Modulation (PCM) of Voice Frequencies. Technical report, International Telecommunication Union.
- A. Ivry, B. Berdugo, and I. Cohen. 2019. Voice activity detection for transient noisy environment based on diffusion nets. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):254–264.
- A. Ivry, B. Berdugo, and I. Cohen. 2020. Evaluation of deep-learning-based voice activity detectors and room impulse response models in reverberant environments. In *Proc. IEEE ICASSP*.
- K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri. 2022. How bad are artifacts? Analyzing the impact of speech enhancement errors on ASR. In *Proc. Interspeech*.
- T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *Proc. IEEE ICASSP*, pages 5220–5224.
- LiveKit. 2025. LiveKit Agents: a framework for building realtime voice AI agents. <https://github.com/livekit/agents>.
- R. Martin. 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5):504–512.

- A. Pandey and D. Wang. 2020. On cross-corpus generalization of deep learning based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- J. Park, W. Kim, D. K. Han, and H. Ko. 2014. Voice activity detection in noisy environments based on double-combined Fourier transform and line fitting. *The Scientific World Journal*, 2014:146040.
- Pipecat. 2025. smart-turn v3. GitHub repository. <https://github.com/pipecat-ai/smart-turn>.
- M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo. 2015. The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 275–282.
- Soham Ray, Keshav Dhandhanian, Victor Barres, and Karthik Narasimhan. 2026. τ -Voice: Benchmarking full-duplex voice agents on real-world domains. arXiv preprint arXiv:2603.13686. <https://arxiv.org/abs/2603.13686>.
- C. K. A. Reddy, H. Dubey, K. Koishida, R. Cutler, and 1 others. 2023. ICASSP 2023 deep noise suppression challenge. Microsoft Research. <https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-icassp-2023/>
- Chandan K.A. Reddy, Vishak Gopal, and Ross Cutler. 2021. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6493–6497.
- Tim Sainburg. 2019–2025. noisereduce: noise reduction in Python using spectral gating. GitHub repository. <https://github.com/timsainb/noisereduce>.
- Silero Team. 2024. Silero VAD: pre-trained enterprise-grade voice activity detector. <https://github.com/snakers4/silero-vad>.
- G. Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.
- D. Snyder, G. Chen, and D. Povey. 2015. MUSAN: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484.
- TEN Framework. 2025. TEN: Open-source framework for building real-time conversational ai agents. GitHub repository. https://github.com/TEN-framework/ten_framework.
- TEN-framework. 2025a. TEN Turn Detection: semantic turn detection with Qwen2.5-7B for full-duplex dialogue. <https://github.com/TEN-framework/ten-turn-detection>.
- TEN-framework. 2025b. TEN-VAD: a low-latency streaming voice activity detector. <https://huggingface.co/TEN-framework/ten-vad>.
- John Wiseman. 2016. py-webrtcvad: Python interface to the WebRTC voice activity detector. <https://github.com/wiseman/py-webrtcvad>.
- Mike Wu, Jonathan Nafziger, Anthony Davis, Noah Goodman, and Chris Piech. 2020. Harper valley bank: a labeled spoken dialog dataset. <https://github.com/cricketclub/gridspace-stanford-harper-valley>.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ -bench: A benchmark for tool-agent-user interaction in real-world domains. Preprint, arXiv:2406.12045.
- X. Zhou, H. Zhu, L. Mathur, R. Zhang, H. Yu, Z. Qi, L.-P. Morency, Y. Bisk, D. Fried, G. Neubig, and M. Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *Proc. ICLR*.
- Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, and Maarten Sap. 2026. Mind the sim2real gap in user simulation for agentic tasks. <https://arxiv.org/abs/2603.11245>. Preprint, arXiv:2603.11245.

A Method details (deferred from section 4)

A.1 Pipeline under test

We replay each audio file through a real-time voice-agent pipeline built on the open-source Pipecat framework (Daily, 2024). The stages relevant to this paper are, in order: a transport stage that streams 8 kHz mono audio frames; an optional noise filter; a Silero Voice Activity Detector (Silero Team, 2024); an end-of-turn (“smart-turn”) classifier that runs an ONNX model over sliding 8 s windows at 1 s stride and emits a binary COMPLETE prediction per window; an automatic speech recogniser; and downstream LLM and TTS stages. The full-pipeline replay is run in two configurations: **m01** (smart-turn off, baseline) and **m05** (smart-turn on, the production-deployed configuration).

A.2 Datasets

Production source. A pool of $n=4,999$ live production calls (each more than ≥ 30 s) spanning 6 customer deployments. 8 kHz stereo telephone recordings. Each call ships with a transcript log carrying per-word transcript timestamps and timestamps for each VAD events and agent-speech intervals.

Harper Valley Bank. This dataset comprises caller-side audio from the Wu et al. (2020) open call-centre corpus, featuring 8 kHz mono recordings with a mean clip duration of 60 s. The dataset includes word-level transcripts and timestamps, which we use to establish ground-truth user-speech intervals. To derive binary turn-end decisions, we first process the audio with the Silero VAD model, chunking the audio and its corresponding text into discrete speech segments wherever the VAD confidence exceeds 0.86. We then evaluate each consecutive pair of transcript segments using an LLM to classify whether the two statements represent a single continuous utterance with a mid-turn pause or two entirely distinct conversational turns.

AppTek channel-2 (multi-accent). This dataset comprises caller-side audio from the open call-centre dialogue corpus released by AppTek (2024). The recordings are 16 kHz mono and feature 14 distinct English accents, which we sample using an accent-stratified approach. Because the original data lacks word-level timestamps, we derive ground-truth user-speech intervals via a Connectionist Temporal Classification (CTC) alignment mechanism powered by the Whisper-large-v3 ASR model. To detect conversational turn ends, we employ the identical LLM-based segment classification methodology utilized for the Harper Valley Bank dataset.

TAVC (synthetic). Persona-driven user utterances synthesised across 4 commercial TTS models (ElevenLabs (ElevenLabs, 2025), Hume, Deepgram, OpenAI) covering many voices per model, and composed into one of seven scenario templates:

- **S1 (speech-at-start):** speech in $[0, 2-4]$ s, noise-only tail to 12 s.
- **S2 (noise-only):** no speech, continuous noise for 12 s.
- **S3a/b/c (mid-pause):** speech segment, 1/2/3 s noise-only gap, second speech segment (3–6 s each); total $\sim 12-14$ s.
- **S4 (speech-at-end):** noise-only lead of $\sim 8-10$ s, speech in the final 2–4 s.
- **S5 (continuous-speech):** speech for 9–12 s with noise throughout, no pause.

On top of each composed timeline the pipeline applies a fixed augmentation chain in order: (i) a continuous background noise overlay at 15 dB SNR

with ± 3 dB drift, drawn from a fixed pool of four real-recorded continuous noise files (busy street, TV-noise room, people talking, metro station); (ii) one burst-noise event placed at the geometric centre of a randomly chosen speech segment, drawn from a pool of five burst files (car horn, dog bark, siren, engine idling, ringing phone); (iii) one 150 ms frame drop simulating packet loss; and (iv) a μ -law telephony codec down-sampling to 8 kHz with the standard 300–3400 Hz bandpass. GT speech timelines are exact by construction. $n=6,000$ clips total, balanced across the seven templates and the four TTS models. This method replicates the τ -Voice augmentation protocol (Ray et al., 2026).

VoiceBench (second synthetic). We sample from the TTS-only splits of VoiceBench (HLT-Lab, 2024) (alpacaeval, alpacaeval_full, openbookqa, mmsu, mtbench, ifeval, advbench), all rendered with Google TTS. These splits already carry VoiceBench’s own augmentation chain (far-field low-pass to simulate distant-mic conditions, signal-distortion to simulate poor microphone quality, and reverberation), *none of which simulates a telephony channel*. On top of those clips we apply the same τ -Voice augmentation mechanism as presented in τ -Voice augmentation protocol (Ray et al., 2026).

Misfire-rate definition. For each VAD speech event, let f be the time-fraction overlap with the union of ground truth user-speech intervals. The event is a misfire iff $f < 0.5$. Per-clip misfire rate = #misfires / #total VAD events. Computed on Silero VAD (Silero Team, 2024), WebRTC VAD (Wiseman, 2016), and TEN-VAD (TEN-framework, 2025b).

A.3 Signal-layer and event metrics

Signal layer (audio-only). F0+F1 spectral energy fraction in $[200, 800]$ Hz; spectral flatness; mean F0 over voiced frames and voiced-frame fraction (Praat / parselmouth (Boersma, 2001)); frame-RMS mean over 50 ms windows; mean Silero VAD speech-segment duration.

Sim-real gap effect size. Let $\tilde{x}_d$ denote the median value of a metric on dataset d and let $\mathcal{R}$ be the set of real datasets. The sim-real gap is $g = (\tilde{x}_{\text{syn}} - \text{median}_{d \in \mathcal{R}} \tilde{x}_d) / \text{IQR}_{\mathcal{R}}$, where $\text{IQR}_{\mathcal{R}}$ is the IQR over the union of per-clip values across $\mathcal{R}$ (pooled-real-IQR). $|g| > 1$ means the synthetic median sits more than one pooled-real-IQR from

typical real audio; $|g| > 2$ means the distributions are essentially disjoint.

A.4 Calibration sweeps

The calibration-misranking figure in section 5.3 (Figure 1) is built from three threshold sweeps on a single paired-design corpus assembled for this section: 4 commercial TTS Vendors: ElevenLabs (eleven_multilingual_v2, 3 voices), Hume (Octave TTS file API, 3 voice descriptions), Deepgram (Aura-2, 3 voices), OpenAI (tts-1-hd, 3 voices). Raw TTS audio is wrapped in the τ -Voice scenario chain documented in section 3: continuous background noise (SNR 15 dB $\pm$ 3 dB drift, drawn from the 4 tau2-bench continuous noise files), one burst-noise event placed inside the speech segment, one 150 ms packet-drop, and a μ -law 8 kHz telephony codec.

VAD threshold sweep (Figure 1(a)). We grid-search Silero $\tau_V \in \{0.10, 0.15, \dots, 0.85\}$ on every clip. Per clip we compute (i) misfire rate (§A.2) and (ii) frame-level recall against the GT speech intervals. The per-group loss is $1 - F_1$ (recall, $1 - \text{misfire}$) aggregated as mean over clips. argmin-loss τ_V^* : Real (Production $\cup$ HVB mean) 0.65; TAVC mean 0.50. The synth-optimum threshold is applied to real-side audio in panel (b).

NF threshold sweep (Figure 1(b)). We hold Silero fixed at the synthetic-optimum threshold, $\tau_{V,\text{synth}}^* = 0.50$ from panel (a), and sweep the noisereduce parameter prop_decrease across $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. For each clip and noise filter (NF) strength, we re-run Silero on the filtered audio and compute the corresponding $1 - F_1$ loss. By anchoring the upstream component to its synthetic optimum, we directly mirror the cascading-misranking failure mode that real-world teams experience when tuning pipelines component-by-component on synthetic data.

To further validate this, we conducted an additional threshold sweep of our internal noise filter from 0.3 to 0.9, again fixing Silero at $\tau_{V,\text{synth}}^*$ and evaluating on the same corpus and loss metric. The results reveal a production argmin of 0.8 (with a loss of 0.447), whereas every tested TTS vendor yields an argmin between 0.3 and 0.4 (with losses ranging from 0.13 to 0.16). This discrepancy demonstrates that synthetic audio understates the absolute production loss by a factor of roughly $3\times$. On the clean, public HVB corpus (median

SNR 31.6 dB), the optimal threshold is 0.4, which is consistent with the findings in §A.6.

Turn analyzer threshold sweep (Figure 1(c)). We sweep Pipecat smart-turn v3 $\tau_S \in \{0.30, 0.35, \dots, 0.90\}$ on the per-window $P(\text{COMPLETE})$ outputs from a single pass of the v3 model (sliding 8 s window, 1 s stride). For each clip we bucket windows by GT: EOT (window end within 0.5 s of the last GT speech end), MIDSPEECH (window end strictly inside a GT speech interval with 0.5 s safety inset), or IGNORE. The per-group loss is $1 - F_1$ of (EOT recall, $1 - \text{MIDSPEECH FP}$) on the pooled window counts. argmin-loss τ_S^* : Real 0.30, TAVC 0.40.

A.5 Cross-NF cross-VAD full matrix

Table 8 reports per-VAD-event misfire rate for every (dataset, NF-or-orig, VAD) cell at $n=30$ clips per dataset. The same directional asymmetry (large NF win on real, small or negative win on synth) reproduces across all three VAD families.

A.6 Noise-filter mechanism on synthetic vs. real audio

This subsection expands the one-sentence note in section 5.1 that the two noise filters behave qualitatively differently on real vs. synthetic audio because they were each designed against a different noise model. We document the Placement-B STT WER numbers (Table 9) and the transcript-level consequence on the text-grounded turn analyzer (Table 10).

The two NF families. (a) *Internal NF.* A learned per-chunk speech-vs-noise classifier deployed as the production NF in our pipeline. It runs at 24 kHz on $\sim 10\text{-}30$ ms chunks, emits a per-chunk noise probability, and mutes or attenuates chunks above a fixed threshold (deployed default 0.7). Training data is real wideband speech with real telephony-class background events: breath bursts, keyboard clicks, transient impulses, multi-talker babble, room reverb, music tails – exactly the multi-source non-stationary noise class Park et al. (2014), Ivry et al. (2019), and Dafna et al. (2015) document as the canonical VAD false-positive triggers. (b) *noisereduce* (Sainburg, 2019–2025). An open-source Python library implementing Boll-style spectral subtraction (Boll, 1979) in stationary=True mode. It estimates a single noise magnitude spectrum from low-energy frames at the start of the clip and subtracts it from every

Dataset	Silero			WebRTC			TEN-VAD		
	orig	internal	nr	orig	internal	nr	orig	internal	nr
Production	0.311	0.287	0.291	0.745	0.503	0.617	0.676	0.525	0.578
HVB	0.029	0.017	0.016	0.629	0.233	0.347	0.263	0.195	0.236
AppTek	0.071	0.017	0.029	0.616	0.230	0.333	0.523	0.194	0.246
VoiceBench*	0.067	0.061	0.091	0.200	0.132	0.114	0.202	0.205	0.133
TAVC*	0.143	0.120	0.157	0.485	0.423	0.498	0.479	0.386	0.373

Table 8: Pooled per-VAD-event misfire rate per (dataset, NF, VAD). *nr* = noisereduce, *internal* = production-internal noise filter. * marks synthetic datasets.

Dataset	orig WER	internal Δ WER	nr Δ WER
Production	0.136	-2.7pp	-2.0pp
HVB	0.144	+1.4pp	+0.2pp
TAVC*	0.389	+7.6pp	-15.8pp

Table 9: Placement B (NF before STT only): per-clip median WER of Deepgram nova-3-general against ground-truth transcripts. The synthetic baseline WER is high (0.389) because band-passed TTS speech is already hard to transcribe; the internal-NF delta +7.6pp is on top of that. Negative Δ WER means the NF helps STT. **Bold** marks the inverted-sign cells on TAVC* relative to Production — the Placement-B half of the NF-ranking inversion in section 5.1. * marks synthetic datasets.

frame, scaled by a `prop_decrease` parameter (deployed at 1.0). The quasi-stationarity assumption breaks under non-stationary noise (Martin, 2001; Cohen, 2003).

Why each behaves as it does on synthetic. On TAVC each clip carries one HVB noise segment overlaid stationarily under TTS-generated speech. The internal-NF classifier sees no breath bursts, no keyboard transients, no babble onsets, so it acts only weakly (-6.1pp WebRTC misfire on TAVC vs. -24.2pp on production, Table 2). It is also pulled in the opposite direction at Placement B by the TAVC F0 + F1 energy deficit (-3.73 σ , Table 5): under-formant frames look noise-like to the classifier and get attenuated, raising STT WER by +7.6pp (Table 9). noisereduce sees the opposite problem: the synthetic overlay is stationary by construction, so the noise spectrum estimator converges trivially and subtraction is aggressive. Aggressive subtraction of a confidently-estimated stationary noise leaves isolated random spectral peaks behind – the classic “musical noise” artifact (Cappé, 1994), with sharp onsets and quasi-harmonic structure that the downstream Silero/WebRTC VAD reads as additional speech events. This is why noisereduce *regresses* VAD misfire on TAVC (+1.4pp) instead of reducing it.

A.7 Transcript-corruption $\rightarrow$ decision-flip chain

The text-grounded turn-analyzer flip in section 5.2 (33% on TAVC + internal NF versus 3% on Production + internal NF) is the end of a three-step chain: NF $\rightarrow$ STT transcript change $\rightarrow$ $P(c)$ shift $\rightarrow$ binary decision flip. Table 10 walks this chain on nine clips. Per row: the Deepgram nova-3-general transcript, the livekit/turn-detector multilingual $P(c)$, and the binary decision at the no-NF dataset median threshold ($\tau_{\tau\text{-Voice}}^{\text{med}}=0.071$, $\tau_{\text{Production}}^{\text{med}}=0.166$).

A.8 Controlled analyses: null calibration, SNR, gender

UASI null calibration. The $|g|\leq 1$ pass threshold used across all five metrics is calibrated against a real-vs-real null constructed from 200 random within-corpus splits of HVB (seed=17; the 50-clip set is split into two 25-clip halves per iteration). The worst observed null $|g|$ across any (metric $\times$ split) cell is 2.08 (TEN-VAD’s mean-segment- duration tail across the 200 iterations); the 99th-percentile null $|g|$ stays below 0.89 for every Layer-1 metric (M1-M5) and below 1.01 across the routing-layer extension metrics $\times$ three VADs. Zero splits exceed the $|g|>1$ pass threshold on any of the five UASI metrics. The threshold is therefore conservative by construction: a real-vs-real comparison never trips it.

SNR and gender skew. Synthetic and real corpora are not matched on per-clip SNR or speaker gender (Table 11). Production has the lowest SNR (median 14.1 dB); TAVC has a narrower SNR distribution (IQR 21.6-24.6); VoiceBench is mostly clean read speech (median 38.3 dB). Gender is also skewed: TAVC is 93% female, VoiceBench is 73% male, Production is 70% female. Per-clip SNR is computed as $20 \log_{10}(\text{RMS}_{\text{speech}}/\text{RMS}_{\text{non-speech}})$ on the captured signal using the GT speech intervals (§A.2); gender from median voiced F0 with a

165 Hz cutoff.

Matched-SNR comparison. The 15-25 dB SNR bucket contains 10 Production, 8 HVB, and 18 TAVC clips. The Placement-B WER analysis within this bucket (Table 12) gives TAVC’s internal-NF WER penalty at matched SNR as +36.6pp; HVB’s is +0.0pp. The synthetic over-suppression failure is not driven by overall SNR mismatch.

Matched-gender comparison. Within female-estimated clips only (Production 21, HVB 18, TAVC 28), the same analysis shows Production internal-NF −7.7pp (helps), HVB +1.1pp, TAVC +32.3pp (hurts) – a 40pp opposite-direction gap at matched gender (Table 13). The asymmetry is not explained by gender skew.

What remains uncontrolled. Production’s high-SNR clips have ~0% orig WER, leaving no room for the NF to help or hurt in that regime. TAVC has no clips below 15 dB SNR, so the synthetic NF response at production-typical noise levels is not directly measured. The matched-SNR and matched-gender controls rule out SNR mismatch and gender skew as load-bearing explanations for the gap, but do not test the full Cartesian product of SNR, gender, accent, and noise-type.

A.9 Per-band energy distribution

Table 14 reports the per-frequency-band energy distribution that underlies the F0+F1 metric (M1) in Table 5: the median fraction of total spectral energy in each of five bands, computed per clip and aggregated over $n=30$ clips per dataset. All four non-TAVC datasets park 70-74% of energy in the F0+F1 band (200-800 Hz); TAVC parks only 13.5% there and inverts the rest into the 800-1500 Hz presence band. VoiceBench, despite being wide-band read speech rather than telephony, lands inside the F0+F1 distribution – the formant-band departure on TAVC is a TTS-rendering artefact, not a synthetic-vs-real distinction per se.

A.10 Multi-TTS extension: signal-layer departure is vendor-agnostic

The two synthetic candidates in the main paper are both rendered through the same TTS family (TAVC via ElevenLabs v3, VoiceBench via studio read speech routed through TTS-style prompt scripts). A natural concern is whether the signal-level departure documented in section 6 is specific to one

TTS vendor. We address this with a multi-vendor extension.

Setup. We generate 27 raw TTS clips covering three prompts (a short turn, a turn with a mid-utterance pause, and a longer multi-sentence turn) across three commercial TTS vendors with three voices each: ElevenLabs eleven-multilingual-v2 (3 voice IDs), Deepgram aura-2 (thalia, orpheus, helena), and Hume Octave 2 (three voice descriptors covering male/female/ casual register). For each raw TTS clip we run the same per-clip signal-panel audit as section 6 against the same Production + HVB + AppTek reference as Table 5.

Result. Table 15 reports the per-vendor and pooled multi-TTS effect sizes. All three vendors fail the same three metrics: voicing fraction, frame-RMS mean, and spectral flatness mean. The pooled multi-TTS departure is essentially indistinguishable from the TAVC departure in the main panel. The signal-layer fingerprint is a property of commercial TTS as a class against real telephony, not a property of any single vendor.

Implication for UASI. The deployment-decision rules in Table 6 return identical safety verdicts on each vendor: a regression-test surface built on any one of the three fails the same NF A/B, VAD-tune, and turn-tune safety checks. Practitioners cannot escape the UASI failure modes by switching vendors within this class.

A.11 End-to-end reproduction in Moshi

Moshi (Défossez et al., 2024) is a single end-to-end voice-to-voice model with no explicit VAD, noise filter, or turn-detector to mis-tune. If the synth-real gap were a property of a particular assembly choice it should not reproduce there. It does. On $n=30$ Production clips Moshi’s median silence-latency after the user finishes is 4.0 s and its overtalk-onset rate is 16.7%; on matched TAVC and VoiceBench clips the silence-latency collapses to 0.0 s and overtalk-onset rises to 100% (Table 16). Moshi treats synthetic user audio as “always ready to be talked over” and real user audio as “probably still talking.” Front-end decoupling is a property of the synthetic-audio / speech-model interface, not of any one assembly choice.

A.12 Paired voice-clone corpus: per-condition results

Construction for section 5.4. From HVB we draw 60 calls balanced 30F/30M by median voiced F0. For each call we clone the caller’s voice from their own audio, re-render each utterance’s exact transcript, and place the render at that utterance’s own start offset, truncating any render that would overrun the next utterance’s onset. Ground-truth intervals for both sides are the native HVB utterance timeline, so the two sides share one timeline by construction and the GT-asymmetry objection does not apply here. The human side is passed through the same μ -law round trip as the synthetic side, so the channel is common to both. Conditions escalate:

- **C0**, timing and codec matched.
- **C1**, + per-utterance gain: each render scaled to the RMS of the human utterance it replaces.
- **C2**, + noise: the clip’s own noise floor (bottom 30th-percentile RMS frames) overlaid across the full timeline at its measured level.

Table 17 reports the median paired delta (synthetic – human) with Wilcoxon signed-rank p over the 60 pairs. Reading down the conditions shows which gaps are confounds and which are not. Spectral flatness collapses from +0.51 to +0.058 once the clip’s own noise is present, and TEN-VAD’s misfire gap falls from –11.2pp to –2.2pp and loses significance. Silero misfire, smart-turn mid-turn false-COMLETE, and WebRTC misfire survive every stage.

A.13 Iterative distortion: signal-layer closure does not imply behavioural closure

We tested whether closing the UASI signal-layer panel (section 7) would also close the downstream front-end behaviour gap measured in section 5. The distortion stages were developed across four progressive experiments, each adding one operation on top of the previous, then evaluated through the same behavioural harness as section 5. All four experiments operate on TAVC paired with HVB as the real reference, with the same random pairing across stages (seed=23).

E1: noise overlay only ($K=5$). For $K=5$ random (τ -Voice, HVB) pairs we extract a low-RMS noise segment from each HVB clip (bottom 30th-percentile RMS frames), tile to the TAVC clip duration, scale to the HVB pooled frame-RMS median (~ 0.0084), and overlay on the original TAVC

voice channel. *Verdict: three Layer-1 cells fail by a wide margin: M1 (F0+F1) $g=-2.75$, M2 (voicing fraction) $g=+9.61$, M3 (frame-RMS mean) $g=+15.96$: the synthetic voice is rendered $\sim 11\times$ louder than real telephony and is voicing-heavy across the clip. Noise overlay alone is necessary but not sufficient.*

E2: + voice RMS normalisation ($K=5$). Same pairs. We scale the voice channel so its speech-region RMS (median of top-30th-percentile frames) matches the HVB pooled-speech RMS median (~ 0.0068), then apply the same noise overlay as E1. *Verdict: M3 closes ($g=+15.96 \rightarrow -1.09$, now within $|g|\leq 1$). M2 voicing remains $g=+9.95$, because RMS normalisation does not change the fraction of audio that is speech.*

E3: + pause injection ($K=5$). Same pairs. We chunk the voice channel into 2-second windows, randomly retain 30% as speech and mute the remaining 70%, with 50 ms linear fades at chunk boundaries. HVB-derived noise is overlaid across the full duration so muted regions carry the real noise floor. *Verdict: three of five UASI cells close (M2 voicing $g=+0.65$, M3 frame-RMS $g=-1.09$, M5 segment duration $g=+1.40$). M1 (F0+F1) remains the load-bearing failure at $g=-2.61$.*

E4: same recipe at $K=50$. Identical algorithm to E3, scaled to 50 random pairs (the $K=5$ E3 set as the prefix). *Verdict: three-of-five closure confirmed at scale. M2 voicing $g=+0.50$, M3 frame-RMS $g=-1.40$, M5 segment duration $g=-1.13$; routing-layer extension metrics also pass (speech-starts-per-minute $g=-0.51$, speech-ratio $g=-0.79$ across all three VADs). The only remaining failure is M1 (F0+F1) at $g=-2.25$. The TAVC voice’s spectral envelope is not addressed by RMS normalisation, pause injection, or noise overlay; closing M1 would require a telephony codec pass (μ -law band-limit + quantisation) or a spectral-envelope warp on the voice channel.*

Behavioural evaluation through the §5 harness. When the E4 distorted output is evaluated through the same Silero / WebRTC VAD + noisereduce + Deepgram STT + smart-turn harness used in section 5, downstream behaviour remains closer to raw TTS than to production despite the three-of-five panel closure: Silero misfire 0% vs. production 30.6%; WebRTC misfire 3.9% vs. production 61.7%. Adding transient impulse events on top of

E4 (keyboard-click-like impulses, breath-like low-frequency bursts, babble snippets in non-speech regions) overshoots in the opposite direction (60%+ Silero misfire) without converging on production’s distribution shape on any sweep.

Conclusion. The E1→E4 progression closes three of five UASI cells (frame-RMS via E2, voicing fraction and segment duration via E3, holding at $K=50$ via E4) measurably and at scale. Neither E4 nor any tested transient-injection variant closes M1 (formant-band envelope) or reproduces the downstream VAD behaviour. The signal-layer panel is necessary but not sufficient: additive overlay of independently-recorded noise cannot reproduce the shared acoustic channel real speech and real noise traverse together, consistent with the RIR-augmentation literature (Ko et al., 2017; Ravanelli et al., 2015) and with the speech-enhancement artefact studies (Iwamoto et al., 2022) that locate downstream ASR degradation in artefact insertion rather than residual noise.

B Worked CI example: integrating UASI into a production voice-agent release pipeline

We illustrate the three deployment recommendations from section 8 on a concrete pre-merge CI workflow. Suppose a team ships a customer-support voice agent and currently uses a single synthetic regression suite (TAVC) as a nightly gate for every front-end component: VAD threshold, noise-filter strength, smart-turn threshold, and agent-prompt regression. The observed grievance is the misranking that motivates this paper: an NF threshold A/B that wins on TAVC reproducibly degrades production NPS on the same release. The workflow below restructures the team’s release pipeline so that each component A/B is routed to the evaluation surface its UASI report card declares it can trust.

Step 1 – Audit each candidate suite against UASI before adoption. For every regression suite already in use (or proposed), the team runs the five-metric ASFS panel (§A.3) on the suite *and* on a held-out production trace (50–100 calls sampled across deployment cohorts, with Deepgram word-level timestamps as GT). For the running example, the report card (Table 7) shows TAVC failing M1 (formant-band energy: $g = -3.73$, a $2\times$ disjoint shift) and M4 (spectral flatness: $g = -1.12$) while

passing M5 (segment duration: $g = +0.41$). Per the decision-conditional routing table (Table 6), M1+M4 failure *structurally* disqualifies TAVC for any audio-grounded component decision: VAD threshold, NF strength, audio-grounded smart-turn threshold, and any ASR routing that depends on the front-end audio path. TAVC remains valid whenever the only input into the gated decision is text.

Step 2 – Route each release decision by its UASI verdict. The team rewires CI so each gated decision runs on the surface its UASI verdict says it can trust:

- **TAVC nightly regression** – text-grounded decisions only: LLM tool-selection, agent-persona regression, knowledge-retrieval recall, intent classification. TAVC’s $\sim 6,000$ clips give wide scenario coverage at near-zero cost per run.
- **Production-trace holdout** – every audio-grounded decision: VAD threshold, NF strength (internal-NF `speech_threshold` vs `noisereduce_prop_decrease`), smart-turn threshold, ASR routing. The holdout is sampled per release, refreshed monthly, and labelled with Deepgram word-level GT. 50–100 clips is sufficient at the effect sizes section 5.3 reports.
- **TAVC + holdout joint replay** – end-to-end full-pipeline regression on the union, run weekly as a sanity check on whether the front-end and decision-layer gates remain composable. Not a per-PR gate.

Step 3 – Add UASI as a pre-merge gate on every new candidate suite. The release pipeline runs UASI on every newly proposed suite (a new TTS model, a new noise overlay preset, a new scenario template, a new τ -Voice variant) before the suite is allowed into the CI matrix. The gate accepts a suite iff $|g| \leq 1$ on every UASI metric required for the deployment decision the suite’s caller declares it intends to support. UASI compute is small: under 5 min on a single CPU per candidate suite of 30–100 clips, dominated by Praat F0 extraction. The gate runs as a 5-minute job on every PR that adds or modifies a regression-suite manifest.

Operating envelope. The production-trace holdout is not a static artefact: its acoustic distribution drifts as deployment surface evolves (new contact centres, microphone upgrades, codec migra-

tions, customer-mix changes). The team refreshes the holdout monthly with 50 new calls sampled across cohorts, re-extracts Deepgram timestamps, and re-runs UASI on the refreshed pooled-real reference. A material shift in any pooled-real metric (e.g. formant-band energy moving by > 0.5 IQR after a microphone upgrade) triggers a re-baseline: the new pooled-real becomes the reference distribution, every previously-accepted candidate suite is re-audited, and any suite that now fails a required metric is demoted out of its previous routing slot. The trigger keeps the gate honest under deployment drift without requiring every release to refresh the reference from scratch.

Cost summary. The full workflow costs:

- Per-PR: 5 min UASI on any modified suite + the existing TAVC nightly run (text-grounded only, several hundred clips through the agent loop) + the production-trace holdout job (front-end components only, 50–100 clips, run only on PRs that touch a front-end component).
- Per-month: 50-call holdout refresh (~ 2 h of an annotator’s time using Deepgram word-level timestamps as first pass).
- Per-trigger: re-audit and routing-table re-derivation whenever the pooled-real reference shifts.

The first-time setup – standing up the production-trace holdout, wiring UASI as a pre-merge job, and re-routing each gated decision to its UASI-approved surface – is a one-week engineering project. The recurring cost lives inside the existing CI compute envelope.

C Open Questions and Discussion

The synthetic-real front-end gap is a paradigm-level property of the signal-to-model interface. While real datasets align acoustically ($|g| \leq 1$), this divergence persists in end-to-end V2V architectures like Moshi (Défossez et al., 2024) that omit explicit VAD or turn gates. At matched non-silence wall-clock rates (9–17%), Moshi places non-silence frames inside ground-truth user-speech boundaries 83–100% of the time on synthetic data, but only 17% on real telephony.

Configuration-level misalignment. Unlike text-level metrics (WER, intent F1), we expose a configuration gap: optimal VAD, NF, and turn-detection

thresholds minimised on synthetic data diverge sharply from production-optimal parameters (section 5.3). Simulators actively misdirect optimisation toward broken production configurations. The signal-layer properties in section 6 ground this behaviour, and UASI (section 7) formalises it as a pre-adoption screen.

Prosodic vs. noise drivers. The divergence is not a simple noise-floor artefact. Text-grounded turn-detectors process text alone yet exhibit severe synthetic distribution shifts and an $11 \times$ NF-driven decision-flip surge (section 5.2), confirming the driver is prosodic/lexical-tail distortion rather than acoustic noise. Four of the five section 6 metrics track speech-production and channel constraints that persist even under perfectly matched noise overlays.

Limits of proof. The structural gap is verified across six independent measurement axes, three threshold sweeps, and the V2V reproduction. Lacking a live agent-in-the-loop A/B, the signal-to-failure chain leans on the production silence anchor ($n=4,999$, section 5.1), Moshi trials ($n=30$ per source), and the noise-filter matrix and WER analyses ($n=30$ per dataset).

Backend commonality. Evaluated suites span current commercial (TAVC) and academic (VoiceBench) baselines. Shared upstream TTS backends (ElevenLabs, Deepgram, Hume) across nine voices fail the identical three UASI metrics (§A.10), establishing a vendor-agnostic architectural fingerprint. Simulators must fall inside the measured real-vs-real null on the UASI dimensions required by the intended deployment decision (Table 6) to yield valid production targets.

Signal match is necessary, not sufficient. The UASI panel captures structural pre-conditions but does not guarantee behavioural equivalence. Our iterative distortion experiments (§A.13) closed three of five panel cells yet did not close downstream front-end behaviour. Closing the remaining gap likely requires joint speech-noise processing or learned source-to-target translation (Ko et al., 2017; Iwamoto et al., 2022), both out of scope.

Where the gap injects itself. The Moshi reproduction (§A.11) shows the divergence persists in an end-to-end model with no explicit VAD, NF, or turn-detector. This locates the failure at the synthetic-signal-meets-speech-model interface

rather than at any specific assembly choice. Our evidence is sufficient to claim the gap is structural and signal-layer; attributing it to a single mechanism (acoustic noise, prosody, or channel) remains open.

Variant	Transcript (Deepgram nova-3-general)	WER	P(c)	Decision
TAVC S3c__057: GT = “Are you the technician that’ll be coming, or is this just a call from the company?”				
orig	“Are you the technician that will be coming? Or is this just the call from the company?”	0.19	0.683	COMPLETE
+internal NF	“Are you on the technician? That will be Oh, the job that was not the company.”	0.69	0.039	INCOMPLETE (flip)
+nr	“Are you the techies and battles becoming? Or is it just the call from the company?”	0.38	0.556	COMPLETE (no flip)
TAVC S5__022: GT = “Yes, Scott. I was calling up from Mobile, Alabama, and I was trying to get a passport. . .”				
orig	“Yes, Scott. I was calling up from Mobile, Alabama, and I was trying to get a passport. . .”	0.00	—	INCOMPLETE
+internal NF	“Scott. Calling up from Alabama, and I was trying out of New Orleans. Take a long. . .”	0.49	—	(flip)
+nr	“Yes, Scott. I was calling up from Mobile, Alabama, and I was trying to get a passport. . .”	0.02	—	no flip
TAVC S3b__017: GT = “Yeah. There’s a center gray male. . . she has to go here. She has to go down that lane. . .”				
orig	“Yeah. They just sent you a mail, but she has to go here. She has to go down that lane. . .”	0.49	0.125	COMPLETE
+internal NF	“Okay. No. But she has to go here. She has to go down lane. And then she. . .”	0.63	0.000	INCOMPLETE (flip)
+nr	“Yeah. They just sent you a mail, but she has to go here. She has to go down that lane. . .”	0.49	0.125	COMPLETE (no flip)
TAVC S4__084: GT = “Can you please notate that my swipe out time for today’s shift is correct? The event ended early.”				
orig	“Way. The white van that we know of did have a passenger. . .”	1.50	0.003	INCOMPLETE
+internal NF	“Take them out time.”	0.89	0.101	COMPLETE (flip)
+nr	“Can you please notate that my swipe out time”	0.50	0.025	INCOMPLETE (no flip)
TAVC S4__082: GT = “One zero one zero nine eight eight eight one two nine nine.”				
orig	“One zero one zero nine eight eight eight.”	0.33	0.032	INCOMPLETE
+internal NF	“One zero nine eight eight.”	0.58	0.161	COMPLETE (flip)
+nr	“One zero one zero nine eight eight eight.”	0.33	0.032	INCOMPLETE (no flip)
TAVC S5__019: GT = “It’s four seven three seven Hunter h-u-n-t-e-r, Ridge r-i-d-g-e, Drive. . . Florida”				
orig	“It’s four seven three seven Hunter, h-u-n-t-e-r, Ridge, r-i-d-g-e, Drive”	0.21	—	COMPLETE
+internal NF	“It’s four seven three seven Ridge d-g-e drop”	0.62	—	(flip)
+nr	“It’s four seven three seven Hunter, h-u-n-t-r, Ridge, r-i-d-g-e, Drive.”	0.25	—	no flip
TAVC S3b__049: GT = “I need to make sure that somebody help my brother to the car from the facility he’s at because he’s blind.”				
orig	“I need to make sure that somebody helped my brother too. The car from the facility he’s at because he’s blind.”	0.10	—	INCOMPLETE
+internal NF	“I need to make help my brother too. The call from that because he’s”	0.48	—	(flip)
+nr	“I need to make sure that somebody helped my brother too. The car from the facility he’s at because he’s blind.”	0.10	—	no flip
TAVC S4__018: GT = “four three four three nine nine.”				
orig	“Four three four three nine”	0.17	—	INCOMPLETE
+internal NF	“O three nine nine.”	0.50	—	(flip)
+nr	“Four three four three nine”	0.17	—	no flip
Production 1b2105cc: GT = “questions about passports new passport”				
orig	“Questions about passports. New passports.”	0.40	0.263	COMPLETE
+internal NF	“Questions about passports. New passport.”	0.20	0.166	INCOMPLETE (marginal flip)
+nr	“Questions about passports. New passport.”	0.20	0.163	INCOMPLETE (flip)

Table 10: The $NF \rightarrow STT \rightarrow P(c) \rightarrow \text{decision}$ chain that produces the 33% vs. 3% flip-rate gap reported in section 5.2. Each block of three rows is one clip: the original (no-NF) Deepgram transcript, the transcript after the internal NF, and the transcript after noisereduce. The eight TAVC blocks all show the internal NF collapsing the transcript to a syntactically different sentence or a short fragment; noisereduce preserves the transcript near-verbatim on the same clips and does not flip the turn-detector decision. The final block (Production) is the only internal-NF flip in our Production sample: the NF removes a stray plural “passports” $\rightarrow$ “passport,” moving P from 0.263 to 0.166 (at the threshold) – a binary flip with no semantic content change, contrasting sharply with the synthetic rows above where the NF swaps the sentence entirely. Dashes in the $P(c)$ column indicate clips outside the LiveKit per-clip $P(\text{COMPLETE})$ measurement subset.

Dataset	p25 (dB)	median (dB)	p75 (dB)	F0 med (Hz)	M/F
Production	10.4	14.1	18.2	183	9/21
HVB	21.7	31.6	37.3	193	12/18
AppTek	15.7	21.8	33.5	195	10/20
VoiceBench*	25.9	38.3	57.2	156	22/8
TAVC*	21.6	22.3	24.6	211	2/28

Table 11: **SNR and gender distribution per dataset, $n=30$ clips each.** Per-clip signal-to-noise ratio (dB) is computed as $20 \log_{10}(\text{RMS}_{\text{speech}}/\text{RMS}_{\text{non-speech}})$ on the captured signal, with the speech / non-speech partition taken from the GT timeline (§A.2). *p25*, *median*, *p75* are the per-clip SNR quartiles across the $n=30$ sample; *F0 med* is the median voiced-frame fundamental frequency in Hz from Praat / parselmouth; *M/F* is the estimated speaker-gender count (male / female) using a 165 Hz voiced-F0 cutoff. This table establishes the per-corpus SNR and gender baselines for the two matched-bucket re-analyses in Tables 12 and 13; TAVC’s narrow SNR window (21.6–24.6 IQR) and female skew (28/30) motivate the matched-bucket controls. **Bold** marks the highest median SNR per dataset group (Production for real, VoiceBench* for synthetic) and the headline TAVC* median; * marks synthetic datasets.

SNR $\in [15, 25]$ dB bucket, median across clips			
Dataset	n	orig WER	internal Δ WER
Production	10	0.000	+0.0pp
HVB	8	0.173	+0.0pp
TAVC*	18	0.194	+36.6pp

Table 12: **Placement-B WER within the matched 15-25 dB SNR bucket: the synth-real internal-NF asymmetry survives the SNR control.** For each dataset we re-aggregate the Placement-B noise-filter analysis (NF applied before STT only; Table 9) on the subset of clips whose per-clip SNR falls inside the matched $[15, 25]$ dB window (§11). n is the per-dataset clip count surviving the SNR filter; *orig WER* is the per-clip median word error rate of Deepgram nova-3-general against the GT transcript with no NF applied; *internal Δ WER* is the change after the production internal NF is inserted before STT, sign-convention +pp = WER increase. TAVC retains the **+36.6pp** internal-NF WER penalty at matched SNR; HVB at the same SNR window is neutral (+0.0pp). The synth-vs-real internal-NF over-suppression failure is therefore not driven by overall SNR mismatch between corpora. **Bold** marks the TAVC* internal-NF penalty (+36.6pp) that survives the SNR control; * marks synthetic datasets.

Female-estimated clips, Placement-B WER			
Dataset	n	orig WER	internal Δ WER
Production	21	0.077	−7.7pp
HVB	18	0.103	+1.1pp
TAVC*	28	0.177	+32.3pp

Table 13: **Placement-B WER within female-estimated clips only: the synth-real internal-NF direction reverses, ruling out gender skew as the driver.** For each dataset we restrict to clips whose median voiced F0 sits above the 165 Hz cutoff used in Table 11 (female-estimated subset), then re-run the Placement-B analysis. n is the per-dataset clip count surviving the gender filter; *orig WER* is the per-clip median Deepgram nova-3-general WER against the GT transcript with no NF; *internal Δ WER* is the WER change after the production internal NF is inserted before STT (sign-convention +pp = WER increase). Within female-only clips, the internal NF *helps* Production by **−7.7pp**, is neutral on HVB (+1.1pp), and *hurts* TAVC by **+32.3pp** — a ~ 40 pp opposite-direction gap at matched gender. The synth-vs-real asymmetry is therefore not reducible to the female skew of TAVC* (93%, Table 11). **Bold** marks the inverted-direction internal-NF effects on Production (−7.7pp helps) vs. TAVC* (+32.3pp hurts); * marks synthetic datasets.

Dataset	< 200 Hz	200-800 Hz	800-1500 Hz	1500-3000 Hz	3-4 kHz
Production	4.9	70.5	13.7	4.9	0.3
HVB	4.0	73.4	13.0	5.7	0.6
AppTek	2.8	74.1	10.0	4.1	0.3
VoiceBench*	9.2	72.7	11.9	4.2	0.2
TAVC*	0.0	13.5	57.2	20.6	0.1

Table 14: Per-band energy distribution (% of total spectral energy), median over $n=30$ clips per dataset. **Bold** marks the median F0+F1 band concentration per row (the dominant band); * marks synthetic datasets.

Group	M1 F0+F1	M2 voicing	M3 RMS	M4 flat.	M5 seg.dur.
multi-TTS* (all 27)	-0.64	+ 2.86	+ 3.20	- 1.21	+0.07
ElevenLabs* ($n=9$)	- 1.31	+ 3.45	+ 2.90	- 1.23	+0.51
Deepgram* ($n=9$)	-0.34	+ 2.46	+ 3.20	- 1.15	-0.17
Hume* ($n=9$)	-0.64	+ 3.02	+ 3.50	- 1.22	+0.07
TAVC* (section 6)	- 3.73	+ 1.69	+ 3.19	- 1.12	+0.41
VoiceBench* (section 6)	-0.03	+ 3.56	+ 3.92	- 1.15	+ 1.77

Table 15: Multi-TTS extension: pooled-real-IQR effect size g for 27 raw TTS clips (3 prompts $\times$ 3 voices $\times$ 3 vendors). **Bold** marks $|g|>1$ (UASI failure); * marks synthetic datasets (every row is synthetic; the real reference is the pooled-real R from section 6). All three vendors fail the same three signal metrics; the pattern matches the TAVC* main-text row.

Source	silence (s)	overtalk
Production	4.0	16.7%
TAVC*	0.0	100%
VoiceBench*	0.0	100%

Table 16: Moshi V2V (Défossez et al., 2024) on $n=30$ clips per source. *silence* = median time after the GT user-speech end before Moshi emits non-silence; *overtalk* = fraction of GT user-speech segments during which Moshi already emits non-silence. * marks synthetic datasets.

Measure (synth – human)	C0	C1	C2
<i>Signal panel</i>			
F0+F1 energy (M1)	+0.003	-0.006	-0.006
Voicing fraction (M2)	-0.009	-0.009	-0.009
Frame RMS (M3)	-0.000	-0.001	-0.001
Spectral flatness (M4)	+0.507	+0.507	+ 0.058
Segment duration (M5)	-0.102	-0.075	-0.080
<i>Front-end behaviour</i> (pp)			
Silero misfire $\tau=0.5$	+4.2	+4.5	+ 6.4
Silero misfire $\tau=0.7$	+9.2	+7.3	+ 7.3
WebRTC misfire	-33.9	-32.5	- 32.7
TEN-VAD misfire	-11.2	-5.5	-2.2
smart-turn mid-turn FP	0.0	0.0	+ 15.5
smart-turn EOT recall	0.0	0.0	0.0

Table 17: **Paired voice-clone experiment, per condition.** Median paired delta over $n=60$ (call, vendor) pairs, synthetic minus human; signal metrics in native units, behavioural metrics in percentage points. **Bold** marks the C2 cells discussed in section 5.4. Wilcoxon signed-rank p at C2: 0.17 (M1), < 0.001 (M2, M3, M4), 0.078 (M5), 0.001 (Silero at both thresholds), < 0.001 (WebRTC), 0.11 (TEN-VAD), 0.001 (mid-turn FP), 0.44 (EOT recall). TEN-VAD and EOT recall are the two measures whose gaps do not survive full matching.