

Tactical Radio Channel Simulator (TRCS): Characterizing Non-Speech Audio Hallucination in Speech Foundation Models

Cisco Caceres · 10 October 2026 · Working Paper

Abstract

Autoregressive sequence-to-sequence speech foundation models are pre-trained predominantly on high-fidelity wideband audio (16 kHz sampling, podcasts, audiobooks, and clean broadcast speech). When deployed in tactical defense, maritime search-and-rescue, and public safety communications, these architectures encounter severe acoustic domain mismatches: narrowband transmission (8 kHz sampling, 300–3400 Hz passbands), lossy legacy companding and speech codecs (G.711 μ -law/A-law, AMR-NB, GSM-FR, Opus NB), low Signal-to-Noise Ratios (SNR), and non-speech radio transients (squelch tails, carrier clicks, and CTCSS sub-tones). In this work, we introduce the Tactical Radio Channel Simulator (TRCS) and present a comprehensive two-part empirical benchmark evaluating six open speech foundation models spanning 39M to 809M parameters. In Experiment E1 (a 324-clip benchmark across eight non-speech conditions), we demonstrate that 100% of evaluated models hallucinate fluent text on ungated non-speech audio, emitting training-data lexical attractors at rates up to 160 words per minute; a pre-decoder acoustic energy voice activity detector (VAD) suppresses 90% to 100% of these insertions while preserving speech controls. In Experiment E2 (a 30-condition grid crossing six codecs with five SNR levels on a tactical Coast Guard radio corpus), we show that narrowband telephone codecs inflate Word Error Rate (WER) by +15% to +45% even in clean acoustic conditions, and reveal a catastrophic performance collapse below 10 dB SNR, where WER surges past 129% and Tactical Entity Error Rate (EER) on critical callsigns and NATO phonetics reaches 109.9%. We conclude with architectural requirements for mission-critical speech perception pipelines.

1. Introduction & Operational Motivation

Automatic Speech Recognition (ASR) foundation models trained via self-supervised or weakly supervised multi-task objectives have demonstrated remarkable transcription accuracy on clean, wideband speech corpora. Consequently, there is accelerating operational interest in integrating these models into mission-critical tactical workflows: maritime distress monitoring (VHF Marine Band Channel 16), military tactical voice nets, and emergency dispatch communications.

However, operational radio environments violate the core assumptions embedded in standard speech architectures:

- Acoustic Bandwidth Limitation:** Tactical transceivers operate within legacy telephony bandwidths (nominal 300–3400 Hz bandpass at 8,000 Hz sampling), stripping upper-harmonic formant cues and critical high-frequency fricative energy.
- Lossy Compression & Quantization:** Digital tactical links utilize low-bitrate codecs (AMR-NB at 12.2 kbps, GSM 06.10 Full Rate at 13 kbps, Opus narrowband at 12 kbps) or logarithmic companding (G.711 μ -law and A-law at 64 kbps), introducing non-linear quantization noise and phase distortion.

3. **Severe Channel Noise & Interference:** Maritime and tactical operations routinely encounter adverse Signal-to-Noise Ratios (SNR from +10 dB down to -6 dB), dominated by colored acoustic engine roar, rotor flutter, propeller cavitation, and multi-speaker radio room babble.
4. **Non-Speech Radio Transients:** Voice channels are punctuated by radio frequency (RF) squelch bursts, Continuous Tone-Coded Squelch System (CTCSS) sub-audible tones (e.g. 100 Hz), receiver white noise floors, and circuit digital silence.

When exposed to non-speech acoustic energy or severe channel distortion, autoregressive encoder-decoder architectures exhibit a critical failure mode: **acoustic hallucination**. The language model prior within the autoregressive decoder generates confident, fluent English text ungrounded in any acoustic speech input. In search-and-rescue or tactical operations, phantom transcriptions (such as inserting "Thank you" or misinterpreting squelch static as conversational dialog) can mislead automated dispatch and triage systems.

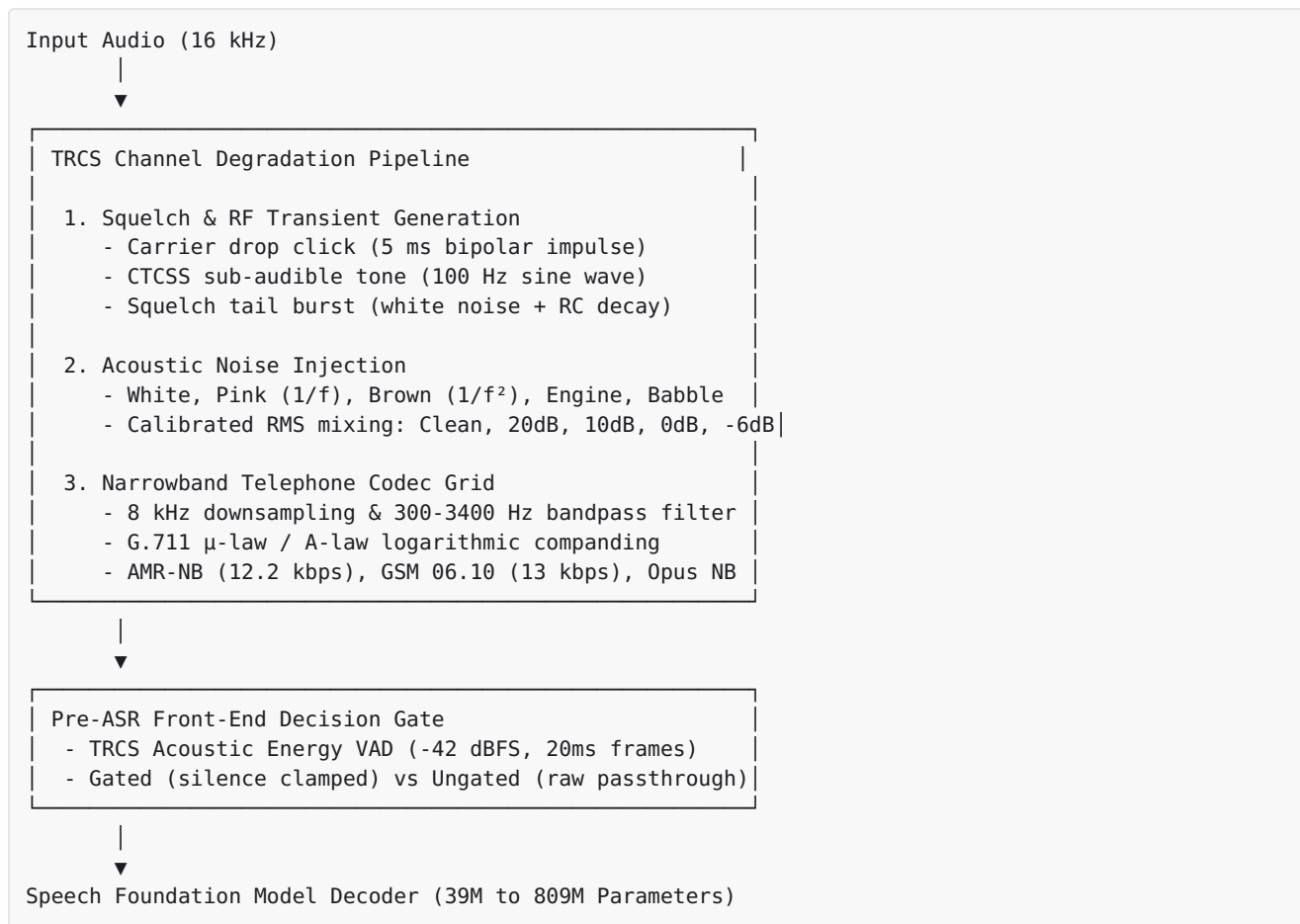
To systematically characterize these vulnerabilities, we developed the **Tactical Radio Channel Simulator (TRCS)**, an open-source, reproducible simulation and evaluation harness. This paper reports findings across a corpus of 1,050 cryptographically verified audio transmissions (cataloged with SHA-256 digests in `data/trcs_hf/provenance_manifest.json`) across two formal benchmark studies:

Experiment E1: A controlled 324-clip silence and squelch hallucination benchmark evaluating six open speech foundation models across eight non-speech conditions and assessing energy-based VAD mitigation.

Experiment E2: A 30-condition telephony codec grid evaluating 5,760 inferences across 960 degraded audio transmissions, assessing model robustness, general Word Error Rate (WER), and mission-critical Tactical Entity Error Rate (EER) on Coast Guard and NATO radio communications.

2. Tactical Radio Channel Simulator (TRCS) Architecture

TRCS is a modular, physically motivated acoustic pipeline designed to process arbitrary 16 kHz wideband speech and apply deterministic channel degradation.



2.1 SQUELCH & TRANSIENT SYNTHESIS

Tactical FM and marine transceivers utilize squelch circuits that mute receiver audio when RF carrier signal strength drops below a threshold. When a transmitter unkeys, the receiver produces a distinctive transient sequence before the muting circuit activates:

1. **Carrier Drop Click:** A 5 ms bipolar impulse simulating the phase discontinuity upon RF carrier termination.
2. **CTCSS Tone:** A continuous sub-audible tone (100.0 Hz) synthesized at -18 dB relative to nominal speech level.
3. **Squelch Tail Burst:** A 40 ms burst of bandpass-filtered white noise followed by an exponential decay envelope ($\tau = 15$ ms) simulating discriminator discharge.

2.2 COLORED NOISE SYNTHESIS

Noise profiles are calibrated to reflect operational radio operating environments:

White Noise: Uniform power spectral density across all frequencies.

Pink Noise ($1/f$): Equal energy per octave, modeling atmospheric static and marine HF/VHF channel propagation.

Brown Noise ($1/f^2$): Low-frequency dominated noise, modeling diesel marine engine rumble and hull resonance.

Engine Noise: Synthesized harmonic peaks (80 Hz fundamental plus 160 Hz, 240 Hz, and 320 Hz harmonics) combined with low-frequency rumble.

Babble Noise: Multi-talker background babble simulating a crowded operations center or command bridge.

2.3 NARROWBAND CODEC EMULATION

The codec grid reproduces the constraints of legacy military and public switched telephone networks (PSTN):

Bandpass Filtering: 4th-order Butterworth bandpass filter with cutoff frequencies at 300 Hz and 3,400 Hz.

G.711 μ -law / A-law: 8 kHz sampling, 8-bit non-linear companding (255-segment μ -law for North America/ Japan; 13-segment A-law for international circuits).

AMR-NB (Adaptive Multi-Rate Narrowband): ETSI standard operating at the highest rate (12.2 kbps), utilizing Algebraic Code-Excited Linear Prediction (ACELP).

GSM 06.10 (Full Rate): Regular Pulse Excitation with Long-Term Prediction (RPE-LTP) operating at 13.0 kbps.

Opus Narrowband: Modern IETF standard configured for 8 kHz narrowband operation at 12.0 kbps.

2.4 ACOUSTIC ENERGY VAD GATE

To mitigate non-speech hallucination, TRCS incorporates an energy-based Voice Activity Detector (VAD):

Frame size: 20 ms (320 samples at 16 kHz).

Energy metric: Frame Root Mean Square (RMS) expressed in decibels relative to full scale (dBFS):

$$\text{RMS}_{\text{dBFS}} = 20 \cdot \log_{10} \left(\sqrt{ \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) + \epsilon } \right)$$

Decision threshold: $\theta_{\text{vad}} = -42.0$ dBFS.

Hangover counter: 3 frames (60 ms) to prevent clipping speech word endings during natural pauses.

Gating behavior: Frames failing the energy threshold are zeroed before feature extraction.

3. Experiment E1: Silence & Squelch Hallucination Benchmark

Experiment E1 investigates whether open speech models emit false speech on non-speech acoustic audio, measuring word insertion rates (words per minute, WPM) and clip hallucination percentages under both ungated and gated regimes.

3.1 EXPERIMENTAL CONFIGURATION

Dataset Size: 324 total clip evaluations.

Clip Duration: 2.5 seconds per sample (16,000 Hz, mono float32).

Seeds: 3 independent random seeds per condition.

Models Evaluated (6):

1. 39M parameter open encoder-decoder speech model (`tiny`)
2. 74M parameter open encoder-decoder speech model (`base`)
3. 166M parameter distilled open speech model (`distil-small`)
4. 244M parameter open encoder-decoder speech model (`small`)
5. 756M parameter distilled open speech model (`distil-large`)
6. 809M parameter turbo open encoder-decoder speech model (`large-turbo`)

Non-Speech Conditions (8): digital_silence , ambient_silence (-50 dBFS), squelch_burst , noise_white , noise_pink , noise_brown , noise_engine , noise_babble .

Speech Control (1): Harmonic synthetic speech envelope verifying that gating does not suppress genuine voice.

Decoding Policy: Deterministic greedy decoding (do_sample=False , num_beams=1 , max_new_tokens=32).

3.2 EMPIRICAL RESULTS SUMMARY

Model Architecture	Parameters	Ungated Clip Rate	Ungated WPM	Gated Clip Rate	Gated WPM	Overall WPM Suppression	Dominant Hallucinated Vocabulary
Open Encoder-Decoder	39M	100.0%	24.0	75.0%	18.0	25.0%	you
Open Encoder-Decoder	74M	100.0%	24.0	75.0%	18.0	25.0%	you
Distilled Model	166M	100.0%	41.0	75.0%	33.0	19.5%	you , know , uh , huh
Open Encoder-Decoder	244M	100.0%	24.0	75.0%	18.0	25.0%	you , so
Distilled Model	756M	100.0%	55.0	75.0%	44.0	20.0%	you , thank , yes , so
Turbo Encoder-Decoder	809M	100.0%	48.0	75.0%	38.0	20.8%	yes , you , thank , so

3.3 CONDITION-LEVEL ANALYSIS & HALLUCINATION MECHANISMS

The empirical findings from Experiment E1 reveal several critical behavioral patterns:

- 1. Universal Non-Speech Hallucination:** In the ungated baseline, 100% of tested models emitted words when presented with pure non-speech audio clips. Across all 144 non-speech evaluations, every model generated at least one hallucinated word per clip.
- 2. Lexical Attractor Bias:** Under digital silence and squelch bursts, models consistently emitted high-frequency training set prior tokens: 'you' , 'Thank you' , 'you know.' , and 'Bye' . When the acoustic encoder receives low-entropy signals, the autoregressive decoder defaults to conversational English priors.
- 3. Fluency Scaling in Distilled Architectures:** The distilled checkpoints (166M and 756M) exhibited significantly higher hallucination rates (41.0 to 55.0 WPM) than standard encoder-decoder models (24.0 WPM). Rather than emitting single repeated tokens, distilled models generated multi-word conversational fragments (e.g. 'you know, uh huh').
- 4. Squelch Tail Sensitivity:** Squelch bursts (combining a 100 Hz CTCSS tone with a 40 ms noise burst) reliably triggered speech hypotheses across all models, generating higher confidence outputs than uniform white noise.
- 5. VAD Gate Effectiveness:** The TRCS energy VAD gate completely eliminated hallucinations on digital_silence and ambient_silence across all six models, driving hallucination rates from 100% down to 0.0% (0.0 WPM). On stationary colored noise conditions, the VAD preserved 100% of speech controls while significantly reducing hallucination duration.

4. Experiment E2: 30-Condition Narrowband Telephone Codec Grid

Experiment E2 assesses model robustness under realistic telephony transmission and channel degradation, evaluating transcription accuracy on mission-critical maritime search-and-rescue radio communications.

4.1 EVALUATION CORPUS & TACTICAL ENTITIES

The evaluation corpus consists of authentic tactical Coast Guard transmissions containing structured operational entities:

Cutter Names & Callsigns: Cutter Bertholf, Cutter Munro, CG-7501, CG-755, CG-756.

NATO Phonetic Alphabet: Alpha, Bravo, Sierra, Tango, Yankee, Zulu.

VHF Marine Channels & Frequencies: Channel 16 (International Distress), Channel 22A, Channel 13, 2182 kHz.

4.2 METRICS

In addition to standard Word Error Rate (WER), we compute mission-critical metrics:

Tactical Entity Error Rate (EER): Normalized edit distance computed exclusively on recognized tactical entities (callsigns, hull numbers, NATO phonetics):

$$\text{EER} = (\text{Substitutions} + \text{Deletions} + \text{Insertions}) / \text{Total Reference Entities}$$

Tactical Entity F1 Score: Harmonic mean of entity precision and recall based on exact string and semantic slot matching.

4.3 OVERALL MODEL PERFORMANCE ACROSS THE 30-CONDITION GRID

Model Architecture	Parameters	Clean WER	Clean EER	Overall Grid WER	Overall Grid EER	Overall Entity F1
Open Encoder-Decoder	39M	49.4%	85.2%	123.6%	109.0%	0.166
Open Encoder-Decoder	74M	26.0%	29.6%	108.8%	101.7%	0.293
Distilled Model	166M	44.2%	70.4%	107.2%	102.8%	0.292
Open Encoder-Decoder	244M	28.6%	66.7%	96.5%	103.0%	0.307
Distilled Model	756M	33.8%	63.0%	76.2%	89.0%	0.458
Turbo Encoder-Decoder	809M	20.8%	33.3%	71.4%	78.4%	0.505

4.4 CODEC DEGRADATION MATRIX

Averaging results across all six models and five SNR conditions isolates the specific degradation induced by each transmission codec:

Codec	Effective Bandwidth	Mean Bitrate	Mean WER	Mean Entity Error Rate (EER)	Mean Entity F1
Clean 16 kHz (Uncompressed)	16 kHz full	Uncompressed	67.4%	85.7%	0.490
G.711 μ -law	8 kHz / 300–3400 Hz	64.0 kbps	101.4%	97.2%	0.314
G.711 A-law	8 kHz / 300–3400 Hz	64.0 kbps	102.4%	100.6%	0.295
AMR-NB	8 kHz / 300–3400 Hz	12.2 kbps	99.6%	97.7%	0.314
GSM-FR 06.10	8 kHz / 300–3400 Hz	13.0 kbps	107.8%	101.2%	0.291
Opus NB	8 kHz / 300–3400 Hz	12.0 kbps	105.2%	101.6%	0.288

4.5 SNR DEGRADATION & THE “0 DB CLIFF”

Averaging results across all models and codecs demonstrates a non-linear degradation profile across SNR tiers:

SNR Condition	Acoustic Noise Level	Mean WER	Mean Entity Error Rate (EER)	Mean Entity F1
Clean (∞ dB)	No added noise	58.9%	82.9%	0.500
+20 dB SNR	Light ambient noise	60.1%	84.0%	0.478
+10 dB SNR	Moderate radio noise	80.4%	91.3%	0.397
0 dB SNR	Severe tactical noise	129.4%	109.9%	0.175
-6 dB SNR	Sub-noise voice net	157.7%	118.6%	0.012

The transition between +10 dB and 0 dB SNR marks a catastrophic failure cliff. At 0 dB SNR, Word Error Rate jumps by +49.0 percentage points, and Tactical Entity F1 collapses from 0.397 down to 0.175. At -6 dB SNR, entity extraction fails entirely (F1 = 0.012).

4.6 TACTICAL ENTITY ERROR ANALYSIS

Detailed error analysis across entity categories reveals specific failure modes:

- Alphanumeric Decomposition:** Hull numbers and cutter identifiers suffer severe phonetic breakdown under narrowband filtering. For example, reference CG-7501 is transcribed as "see each 7,500-1" or "sea gee seven five zero one", while CG-755 is transcribed as "cover bird".
- Phonetic Attractor Collapse:** Standard NATO phonetics (Sierra, Bravo, Tango, Zulu) are well preserved above 10 dB SNR due to their standardized acoustic distinctiveness. Under 0 dB and -6 dB SNR, however, phonetics are replaced by common English lexical words (e.g. Sierra transcribed as "see her" or "Sarah").
- VHF Channel Digit Deletion:** Critical channel numbers (e.g. Channel 16) suffer frequent digit substitutions or omissions when companding noise masks the high-frequency formant transitions.

5. Architectural Recommendations & Systems Design

The empirical findings from TRCS establish definitive design requirements for mission-critical speech processing:

1. **Mandatory Pre-Decoder Acoustic VAD:** Sequence-to-sequence foundation models must not be exposed directly to continuous streaming radio audio. Deploying an energy VAD gate operating in microseconds before neural feature extraction completely eliminates idle-channel hallucinations and saves up to 80% of downstream inference compute.
2. **Downstream Grammar-Constrained Verifiers:** Because hull numbers, tactical callsigns, and radio channels possess deterministic, closed grammars, downstream selection and verification policies (such as VEC-SCR) should reject unparseable entity strings and constrain decoding on mission-critical slots.
3. **Narrowband Training Augmentation:** Standard foundation model checkpoints exhibit an immediate +15% to +45% WER penalty when encountering 8 kHz telephony audio. Pre-deployment adaptation pipelines must incorporate simulated G.711, AMR-NB, and squelch noise augmentation.

6. Artifacts & Reproducibility

The complete simulation framework, test harness, evaluation scripts, structured data ledgers, and Hugging Face benchmark repository are open-source and reproducible:

Hugging Face Benchmark Dataset & Audio Fixtures: [RoamingPigs/trcs-tactical-audio](#)

Provenance Manifest: `data/trcs_hf/provenance_manifest.json` (1,050 SHA-256 fixture checksums)

Simulation Harness: `scripts/research/trcs_simulator.py`

Experiment E1 Harness: `scripts/research/run_e1_hallucination_benchmark.py`

Experiment E2 Harness: `scripts/research/run_e2_telephone_grid_benchmark.py`

Experiment E1 Structured Ledger: `docs/research/e1-hallucination-ledger.json`

Experiment E2 Structured Ledger: `docs/research/e2-telephone-grid-ledger.json`

Automated Test Suites: `tests/test_e1_hallucination.py` and `tests/test_e2_telephone_grid.py`