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

Cisco Caceres · 11 October 2026 Pacific · Working Paper / Corrected Diagnostic Record

Abstract

E1 retains 27 authored synthetic input clips across 6 models and two decoding regimes (324 output evaluations). Overall per-model ungated rates are 24–55 words per minute; 160 is a condition-cell peak. The gate reduces overall WPM by 19.5–25.0%, leaving hallucinations on 108/144 non-speech evaluations. E2 retains 5 authored synthesized utterances across 30 declared conditions and 6 models (900 model-by-condition records). Mean WER across recorded models and paths is 129.4% at 0 dB. These are repeated views of a small synthetic input set, not independent samples or authentic Coast Guard recordings. No population confidence interval, genuine-voice preservation, codec-compliance guarantee, operational-radio accuracy or proprietary AMBIE model accuracy is established.

Correction · 11 October 2026 Pacific: Earlier 90–100% gating suppression, generic 15–45% codec penalties, authentic Coast Guard corpus, independent-clip denominators, confidence and voice-preservation claims are withdrawn. The tables below are bounded retained observations; the released 1,050-file fixture package has a different denominator.

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:

1. **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.
2. **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.

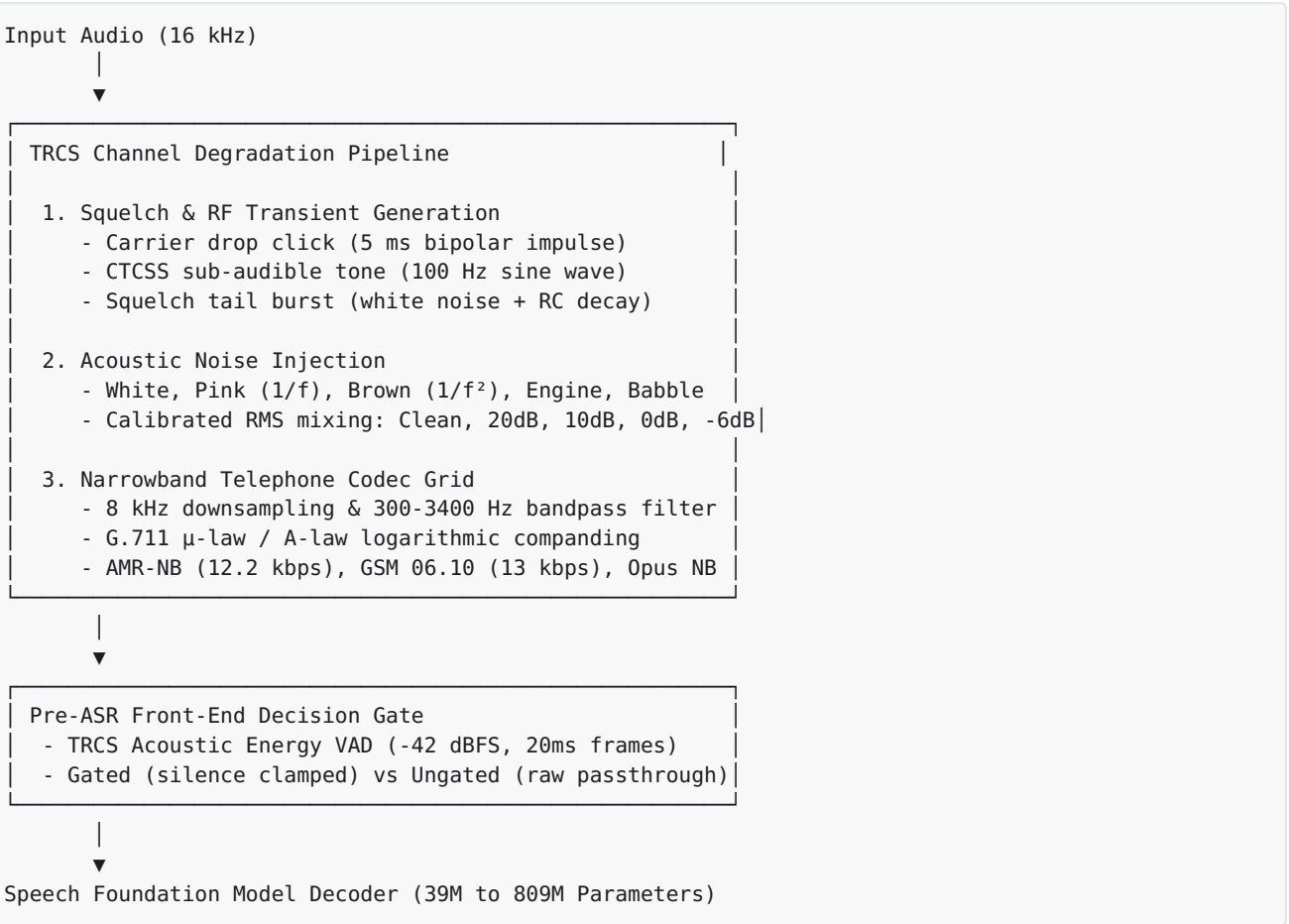
To systematically characterize these vulnerabilities, we developed the **Tactical Radio Channel Simulator (TRCS)**, an open-source, reproducible simulation and evaluation harness. The released TRCS package contains 1,050 authored audio fixtures with per-file SHA-256 identities. That released corpus is separate from these two smaller recorded evaluations:

E1: 27 synthetic inputs (eight non-speech conditions plus a harmonic control, three seeds), six models and two regimes: 324 output evaluations, including 144 non-speech evaluations per regime.

E2: Five synthesized utterances across six declared channel paths and five noise levels: 150 input variants and 900 retained model outputs. They are not authentic Coast Guard recordings or 900 independent utterances.

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 (τ = 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):

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

Decision threshold: $\theta_{vad} = -42.0$ 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 envelope; it does not establish preservation of genuine recorded 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 ungated regime retains hallucinations on all 144 non-speech evaluations; the gated regime retains them on 108/144. The gate clears digital and ambient silence in these fixtures, while all six remaining non-speech conditions retain positive hallucinations. Per-model overall WPM suppression is 19.5–25.0%.

The retained outputs contain short recurring words and phrases. This does not isolate a training-prior mechanism, establish a statistical architecture difference or measure confidence: the ledger contains no confidence score supporting a higher-confidence squelch claim. The harmonic control is not genuine speech, and preservation of real voice remains untested.

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

Experiment E2 is a small synthetic-input channel diagnostic. Its maritime vocabulary motivates future work, but the current inputs are not operational recordings. Some declared AMR/GSM paths can use emulation fallback; the ledger does not establish which implementation path ran.

4.1 EVALUATION CORPUS & TACTICAL ENTITIES

The five authored utterances were synthesized with espeak-ng using one voice. Their text contains maritime-style entities such as:

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 the declared extraction and matching rules.

4.3 OVERALL MODEL PERFORMANCE ACROSS THE 30-CONDITION GRID

Model Architecture	Parameters	Unencoded/no-noise WER	Unencoded/no-noise 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

These descriptive path means average across models and noise conditions; they do not isolate a codec mechanism or prove codec compliance. Unencoded/no-noise WER in the model table is a different scope from the no-added-noise mean across all paths in the SNR table:

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**

In these retained fixtures, mean WER increases by about 49 percentage points from 10 dB to 0 dB. Entity F1 is 0.175 at 0 dB and 0.012 at −6 dB. These are descriptive repeated measurements, not independent population estimates or evidence that every entity extraction fails.

4.6 TACTICAL ENTITY ERROR ANALYSIS

The records motivate closer examination of alphanumeric substitutions, phonetic confusions and channel-number errors. No independently coded failure taxonomy, confidence analysis or causal mechanism is established here. A follow-up should prespecify critical-field scoring and retain source-disjoint real recordings.

5. Architectural Recommendations & Systems Design

These diagnostics motivate prospective engineering tests, not definitive deployment requirements:

1. Compare energy gating with stronger speech/noise discrimination and abstention. The present gate does not clear most non-speech evaluations or establish real-voice preservation or compute savings.
2. Test grammar-aware critical-field checks under a separately frozen protocol. The VEC-SCR study is unrun; it supplies no demonstrated speech-verification benefit.
3. Qualify exact codec implementations and evaluate licensed, speaker-diverse real-domain inputs before adaptation or operational claims. The five synthesized utterances cannot establish a general codec penalty.

6. Artifacts & Reproducibility

The fixture package is published. Local harness and ledger paths below identify retained author artifacts; they do not establish public availability of every script, exact model identity, outside reproduction or independent review:

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`