Neural/audio codec storage and degradation benchmark

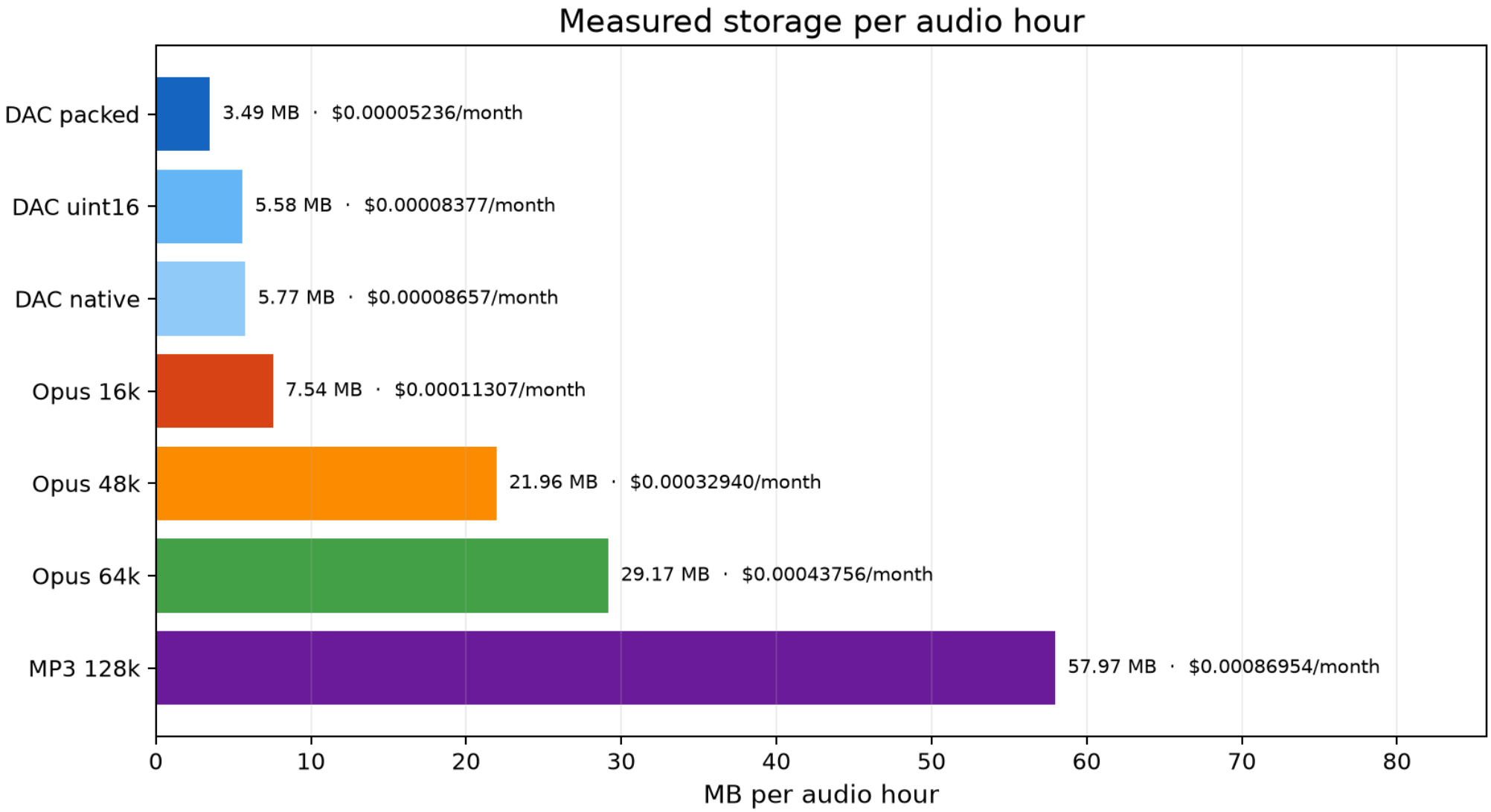
Benchmark date: 2026-08-31. Baseline corpus: 130 survey references (23.26 minutes). New stress cohorts: 20 LibriSpeech `test-other` clips and 20 matched speech-plus-MTG-Jamendo mixtures.

Executive result

Packed all-codebook DAC remains the smallest representation at about 7.8 kbps. On the new stress tests it substantially outperforms Opus 16 kbps in SI-SDR and spectral distance while using roughly half the storage. Opus 48/64 kbps and MP3 128 kbps preserve more waveform detail, at 6–17× packed-DAC storage. No single objective score is treated as ground truth: intelligibility, perceptual prediction, waveform fidelity, and spectral fidelity are reported separately.

Storage and original survey result

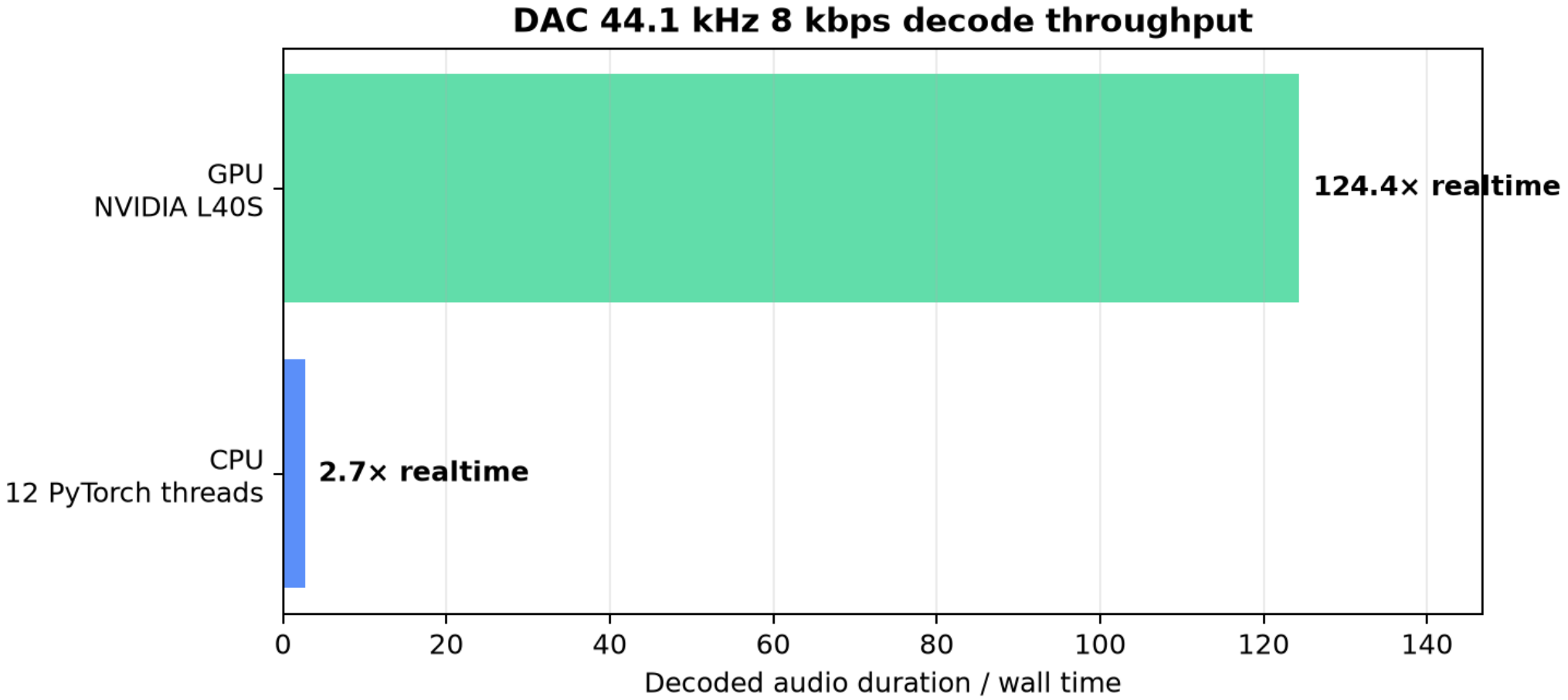
The prior 130-clip survey benchmark remains unchanged: 3.49 MB/hour for packed DAC, 7.54/21.96/29.17 MB/hour for Opus 16/48/64 kbps, and 57.97 MB/hour for MP3 128 kbps. At R2 Standard's $0.015/GB-month marginal rate, those are approximately $0.000052/$0.000113/$0.000329/$0.000438/$0.000870 per stored audio-hour-month. SQUIM mean MOS losses were 0.011/0.207/0.059/0.007/0.004 respectively.



DAC decoder size and throughput

The DAC synthesis decoder contains **54.104M parameters**. Codebook lookup and quantizer projections add 0.240M, making the complete decode path 54.344M parameters. The full encoder–quantizer–decoder codec contains 76.652M parameters.

Device	Median throughput	Observed range	Repeats
CPU, 12 PyTorch threads	2.7× realtime	2.6–2.7×	5
NVIDIA L40S GPU	124.4× realtime	111.0–139.6×	20



This is warm-cache, batch-size-one, application-level decoding of an 8.29-second all-nine-codebook in-memory DAC object. Timing includes codebook reconstruction, the neural synthesis network, GPU-to-CPU output materialization, loudness normalization, and output-length restoration. It excludes disk I/O, model loading, and token encoding. CUDA is synchronized around each measurement. Results characterize this machine and software build, not universal codec throughput.

New 20-clip stress cohorts

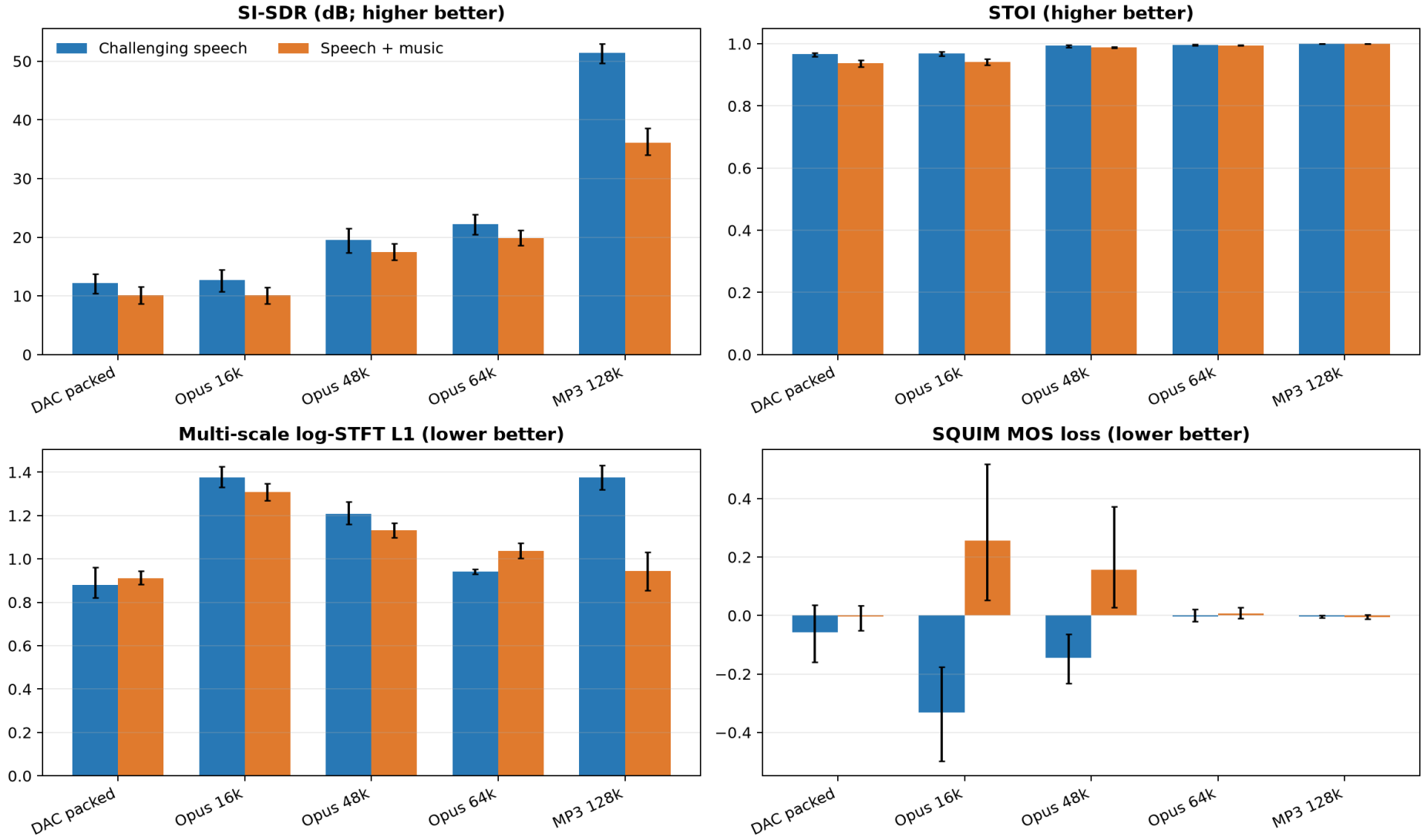
“Challenging speech” is the official LibriSpeech `test-other` set, sampled deterministically across 20 speakers. It is harder read speech, not a controlled additive-noise dataset, so this report does not mislabel it as artificially noisy. The music cohort pairs those same clips with 20 unique-artist tracks from MTG-Jamendo's artist-disjoint mood/theme test split, using adaptation-compatible CC BY/CC0 licensing. Music is mixed at +5 dB active-speech RMS SNR; one common peak gain preserves the SNR.

LibriSpeech test-other — challenging speech

Codec	kbps	SI-SDR dB	STOI	log-STFT L1	SQUIM MOS loss
DAC packed	7.76	12.20	0.966	0.881	-0.057
Opus 16k	16.86	12.66	0.969	1.375	-0.332
Opus 48k	48.94	19.55	0.994	1.208	-0.144
Opus 64k	64.99	22.24	0.996	0.941	-0.003
MP3 128k	129.38	51.39	1.000	1.375	-0.004

Speech plus MTG-Jamendo music

Codec	kbps	SI-SDR dB	STOI	log-STFT L1	SQUIM MOS loss
DAC packed	7.76	10.13	0.938	0.911	-0.003
Opus 16k	16.86	10.09	0.942	1.307	0.256
Opus 48k	48.94	17.50	0.989	1.131	0.156
Opus 64k	64.99	19.87	0.995	1.037	0.007
MP3 128k	129.38	36.10	1.000	0.943	-0.005


Codec degradation on 20 clips per stress cohort (mean ± 95% paired bootstrap CI)
SI-SDR (dB; higher better)
Challenging speech
Speech + music
STOI (higher better)
Multi-scale log-STFT L1 (lower better)
SQUIM MOS loss (lower better)
DAC packed
Opus 16k
Opus 48k
Opus 64k
MP3 128k

What metrics are commonly used, and what was selected

Metric	What it measures	Typical role here	Included
ViSQOL MOS-LQO	Full-reference spectro-temporal perceptual similarity	Widely used headline codec metric for speech/audio	Research comparison; not scored
SI-SDR	Scale-invariant waveform reconstruction fidelity	Common neural codec/separation diagnostic	Yes
Multi-scale log-STFT	Spectral-envelope and fine-structure mismatch at several resolutions	Used in DAC and neural audio-codec work	Yes
STOI	Short-time speech intelligibility	Speech-bearing cohorts	Yes
SQUIM Subjective	Learned non-intrusive MOS estimate	Secondary perceptual diagnostic	Yes
POLQA (P.863)	Standardized speech quality	Telecom-grade validation	Not included: licensed/proprietary
PESQ (P.862)	Legacy standardized speech quality	Frequent historical benchmark	Not included: ITU deleted P.862 in 2024 in favor of P.863
PEAQ / PEMO-Q	Perceptual audio quality, especially music	Music codec evaluation	Not included: reproducible validated tooling/licensing constraints
MUSHRA listening test	Human subjective quality against hidden reference/anchor	Gold standard for final codec decisions	Recommended follow-up

ViSQOL was not numerically included because the official project supplies a source/Bazel build rather than a safe current wheel for this Python 3.12 system; the similarly named PyPI package is explicitly reported upstream as unrelated/malicious. Substituting an unofficial implementation would weaken reproducibility. The report therefore uses four independently interpretable metrics and documents ViSQOL as a recommended follow-up.

Interpretation

- SI-SDR is strict: neural codecs can sound good while receiving lower scores because harmless phase/timing changes count as waveform error.
- STOI is speech-specific and should not be interpreted as music quality; here it tests preservation of the intelligible speech component.
- Multi-scale log-STFT L1 is an uncalibrated distance: compare codecs within the same cohort only.
- SQUIM is a speech model and can be unstable on music mixtures. Its paired reference-minus-codec delta is a secondary diagnostic, not a music MOS.
- ViSQOL audio mode is itself reported as calibrated only down to roughly 24 kbps, so 8 kbps DAC and 16 kbps Opus require caution even when it is available.

Reproducibility and provenance

LibriSpeech `test-other` was downloaded from OpenSLR and verified against MD5 `fb5a50374b501bb3bac4815ee91d3135`. Selection uses a stable SHA-256 ordering, 4–12 second eligibility, and one utterance per speaker. Jamendo selection pins repository commit `cafd8e20c265ed84f1e61f1c875327971f43a62f`, split 0 test data, unique artists, and adaptation-compatible licenses. Exact source hashes, attribution, offsets, gains, SNR, and mixture hashes are in `stress_sets/provenance.json`.

All codecs use mono 44.1 kHz evaluation references; Opus is encoded at 48 kHz CBR and decoded back to 44.1 kHz. DAC uses Descript's 44.1 kHz 8 kbps model with all nine codebooks on an NVIDIA L40S GPU. Confidence intervals are deterministic 10,000-resample clip-level bootstraps.

Sources

- Descript Audio Codec paper (NeurIPS 2023): multi-scale spectral loss, SI-SDR, ViSQOL, bitrate, and listening evaluation.
- Hines et al., ViSQOL v3; Google ViSQOL usage guidelines.
- Le Roux et al. (2019), SI-SDR.
- Taal et al. (2011), STOI.
- TorchAudio SQUIM tutorial and Kumar et al. (2023).
- Panayotov et al. (2015), LibriSpeech; OpenSLR release/checksum.
- Bogdanov et al. (2019), MTG-Jamendo dataset.
- ITU-T P.862 status and P.863/POLQA recommendation.

Artifacts

- `stress_sets/provenance.json`: exact 40-clip cohort provenance and mixing recipe
- `stress_metric_measurements.csv`: 280 raw codec/cohort measurements
- `stress_metric_aggregate.csv`: cohort means and bootstrap intervals
- `dac_decode_throughput.json`: raw per-repeat CPU/GPU timings, environment, and parameter counts
- `benchmark_stress_metrics.py`, `build_stress_sets.py`, `build_expanded_report.py`: reproducible pipelines
- Existing `measurements.csv` and `mos_measurements.csv`: original 130-clip storage and SQUIM results