FULL-CLIP EVALUATION · CHATTERBOX S3GEN

One full decode.
Sixty paired speech clips.

Original speech and prompt-conditioned reconstruction, measured over complete clips. Explore three spectral metrics and listen to every input/output pair.

20 Cartoon Voices · 20 LibriSpeech Other · 20 Podcast · 24 kHz mono

Full-clip scope only. Windowing and overlap experiments were stopped at the user's request and are excluded from this report. No minimum-window, overlap benefit, or streaming-quality conclusion is claimed.

Full-roundtrip source error

Lower values mean closer spectral agreement with the original input. These are absolute errors—not percentages, human MOS scores, or speech-recognition accuracy. All clips receive equal weight.

MR-STFT error
1.965
Mean absolute source error · 60 clips

Log-mel error
0.6972
Mean absolute source error · 60 clips

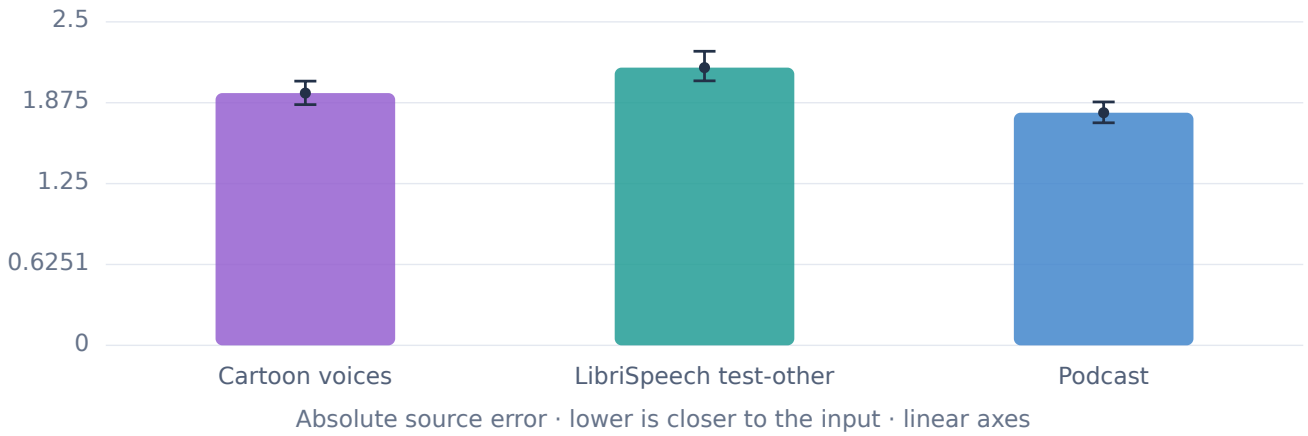
Log1p-mel error
0.01273
Mean absolute source error · 60 clips

Cohort	Metric	Clips	Mean error	95% bootstrap CI
All 60 clips	MR-STFT error	60	1.965	1.901–2.034
All 60 clips	Log-mel error	60	0.6972	0.6711–0.7247
All 60 clips	Log1p-mel error	60	0.01273	0.01098–0.01463
Cartoon voices	MR-STFT error	20	1.95	1.861–2.042
Cartoon voices	Log-mel error	20	0.771	0.7288–0.8174
Cartoon voices	Log1p-mel error	20	0.009803	0.008518–0.01112
LibriSpeech test-other	MR-STFT error	20	2.147	2.045–2.273
LibriSpeech test-other	Log-mel error	20	0.6431	0.6039–0.6888
LibriSpeech test-other	Log1p-mel error	20	0.008566	0.007158–0.01017
Podcast	MR-STFT error	20	1.798	1.72–1.881
Podcast	Log-mel error	20	0.6774	0.6427–0.7097
Podcast	Log1p-mel error	20	0.01982	0.01653–0.02343

Confidence intervals describe uncertainty across the sampled clips, not across speakers, checkpoints, or synthesis seeds. Dataset means are descriptive; the cohorts have different content.

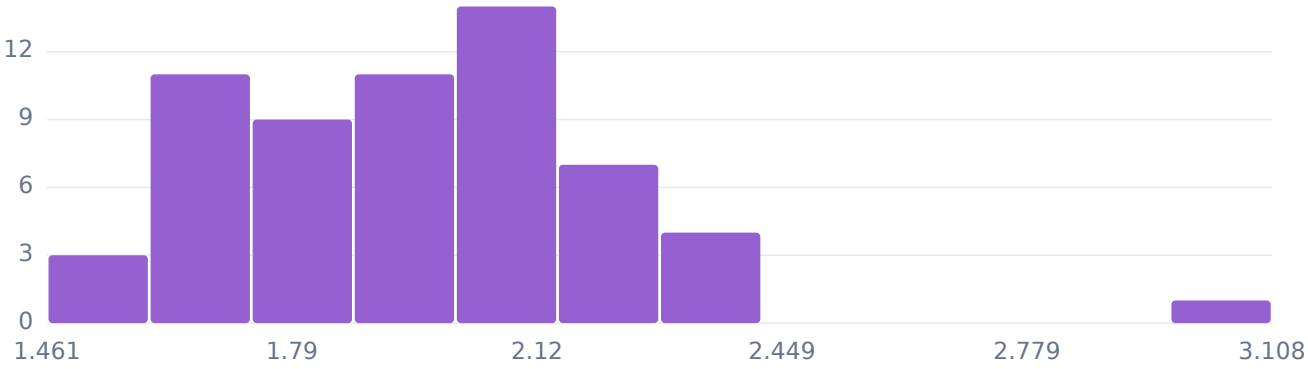
Means and distributions

MR-STFT error · dataset means and 95% CIs



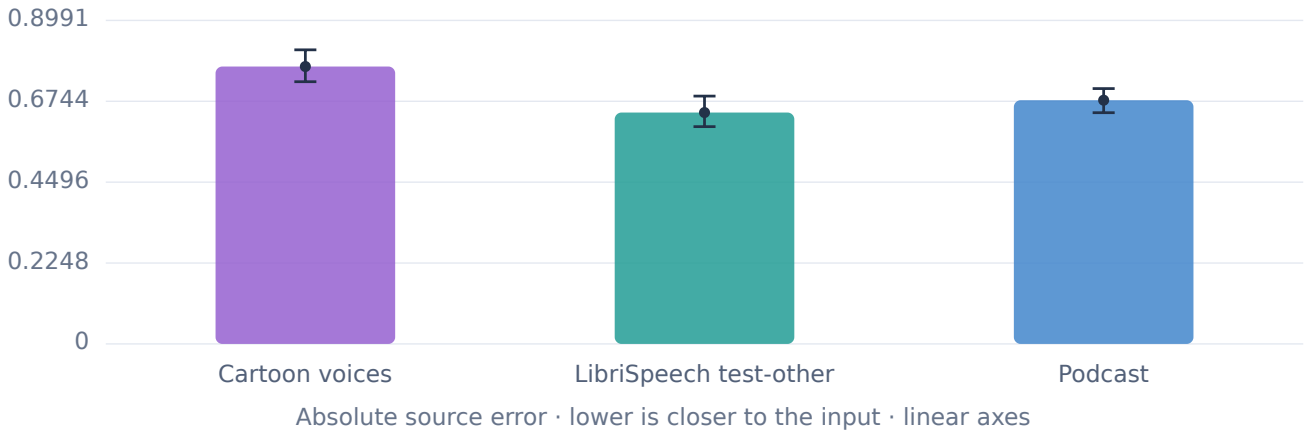
Absolute source error · lower is closer to the input · linear axes

MR-STFT error · all 60 per-clip errors



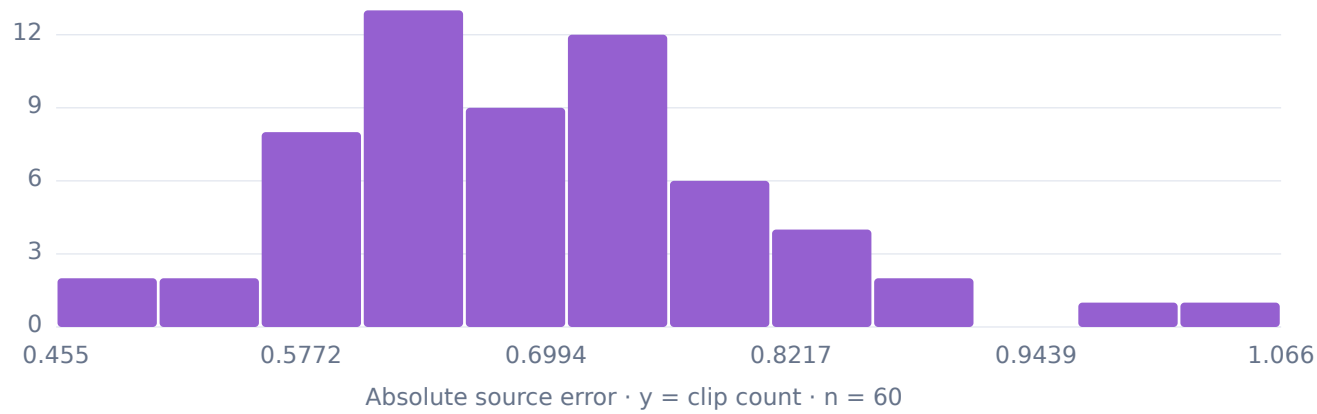
Absolute source error · y = clip count · n = 60

Log-mel error · dataset means and 95% CIs

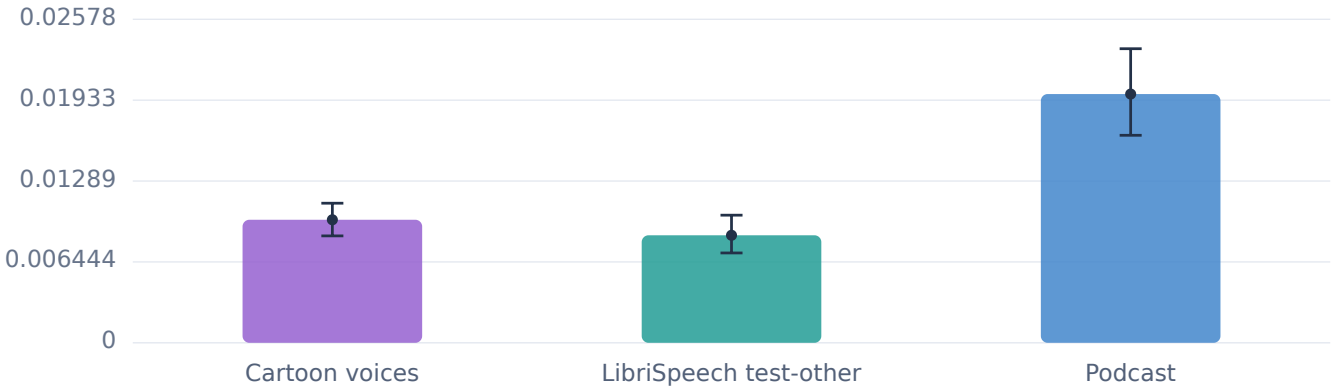


Absolute source error · lower is closer to the input · linear axes

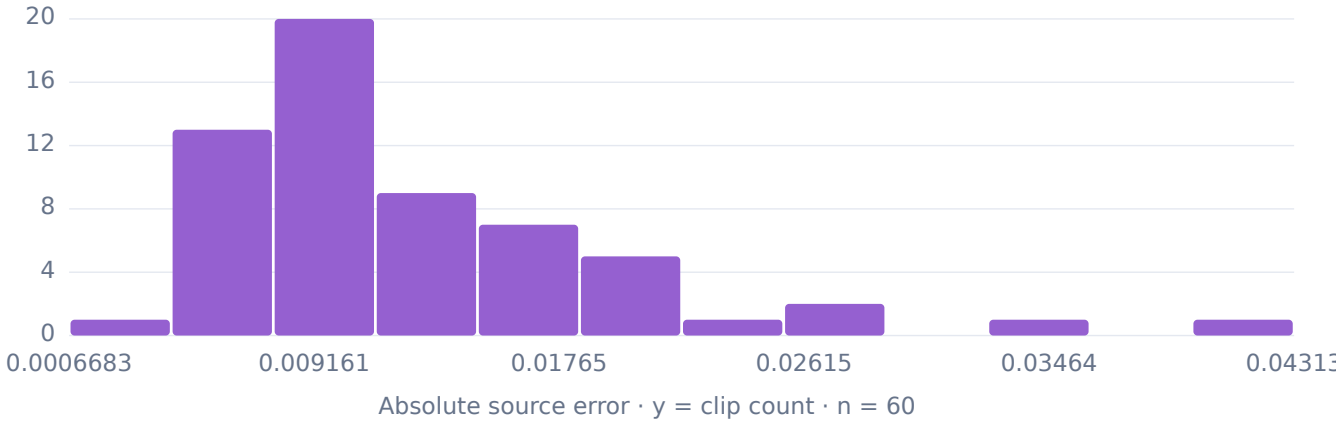
Log-mel error · all 60 per-clip errors



Log1p-mel error · dataset means and 95% CIs



Log1p-mel error · all 60 per-clip errors



Full-source error alongside earlier codecs

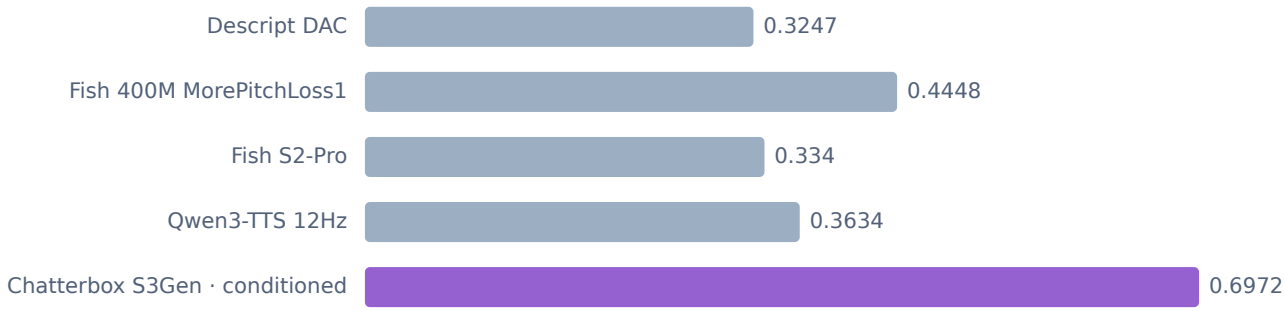
The prior four codecs' published full-decode mean errors are preserved on the same 60 source clips. This is a descriptive comparison, not bitrate matching: S3Gen has a semantic objective and receives additional same-source conditioning. No new codec inference or windowed measurements are included.

MR-STFT error · full-source error, all 60 clips



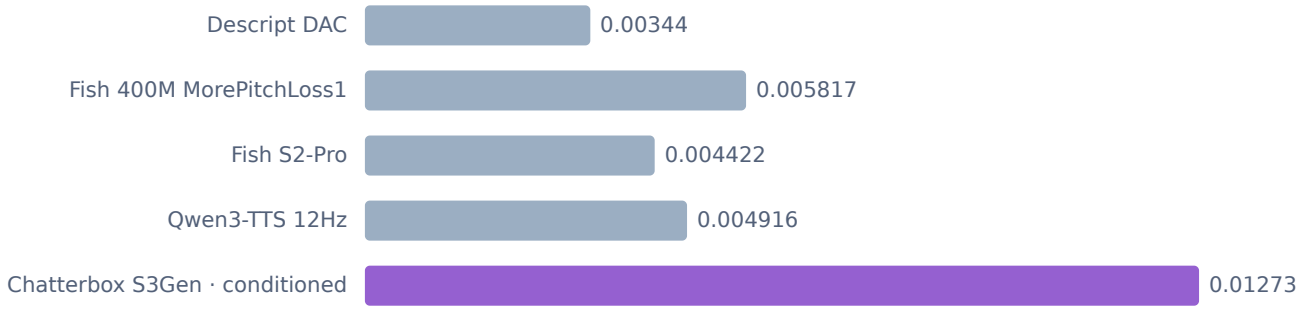
Mean absolute source error · lower is closer · not bitrate-matched

Log-mel error · full-source error, all 60 clips



Mean absolute source error · lower is closer · not bitrate-matched

Log1p-mel error · full-source error, all 60 clips



Mean absolute source error · lower is closer · not bitrate-matched

Bars show means only. Earlier baseline summaries did not provide matching confidence intervals; none are fabricated.

Every clip, with its matching audio

Open any clip ID to compare its exact original and full reconstruction. The table and player use the same measured error record.

Dataset	Clip / listen	Duration	MR-STFT error	Log-mel error	Log1p-mel error
Podcast	04c87851-859c-4a3d-be14-b501dadb468b_seg00187_0001366305_0001375009	8.70 s	2.105	0.7455	0.0147
Podcast	0f1e09b6-2974-4bad-99e5-df0f98bab544_seg00121_0001086971_0001092790	5.82 s	1.575	0.6345	0.01092

Dataset	Clip / listen	Duration	MR-STFT error	Log-mel error	Log1p-mel error
LibriSpeech test-other	3080-5032-0004	5.92 s	2.145	0.6365	0.008766
LibriSpeech test-other	3080-5040-0007	20.38 s	3.033	0.8422	0.01799
LibriSpeech test-other	4198-12259-0014	2.47 s	2.078	0.5664	0.007399
LibriSpeech test-other	4198-61336-0000	13.77 s	2.388	0.8914	0.009668
LibriSpeech test-other	4198-61336-0027	3.53 s	2.101	0.7869	0.006541
Podcast	4777eab3-7251-48c3-8190-846084fe32ec_seg00113_0001219121_0001234983	15.86 s	2.085	0.7688	0.03392
LibriSpeech test-other	5764-299665-0007	17.01 s	2.266	0.6742	0.01032
LibriSpeech test-other	5764-299665-0035	5.13 s	2.069	0.567	0.004986
LibriSpeech test-other	5764-299665-0043	4.63 s	1.928	0.5287	0.002599
LibriSpeech test-other	5764-299665-0071	8.46 s	2.28	0.6379	0.007852
LibriSpeech test-other	5764-299665-0085	4.83 s	2.078	0.5862	0.006141
LibriSpeech test-other	5764-299665-0094	3.80 s	2.17	0.5944	0.005236
Podcast	5c820ab0-e0d2-4a5c-a32d-3725192328c0_seg00025_0000231664_0000237125	5.46 s	1.98	0.7465	0.02769
Podcast	653893f4-b4c4-4b8b-958f-ee1d8e92de3d_seg00303_0001811032_0001817108	6.08 s	1.714	0.6725	0.02186
LibriSpeech test-other	7018-75788-0009	3.40 s	1.992	0.5891	0.00637
LibriSpeech test-other	7018-75788-0012	7.96 s	2.098	0.6567	0.009034
LibriSpeech test-other	7018-75788-0018	20.84 s	2.122	0.6511	0.01202
LibriSpeech test-other	7018-75789-0031	16.13 s	2.375	0.6684	0.0142
LibriSpeech test-other	7902-96591-0004	3.60 s	1.94	0.5786	0.007605
LibriSpeech test-other	7902-96591-0008	14.73 s	2.017	0.6841	0.008818
LibriSpeech test-other	7902-96592-0016	2.08 s	2.262	0.6435	0.01214
LibriSpeech test-other	7902-96594-0001	3.04 s	1.717	0.5031	0.005901
LibriSpeech test-other	7902-96595-0010	4.95 s	1.876	0.5755	0.007748
Podcast	826dc812-843b-4fe8-b6f6-4e18d7e4ca67_seg00357_0002033234_0002038319	5.09 s	1.862	0.6612	0.007429
Podcast	8603eaa6-409f-4a10-9f53-f245cea9ded1_seg00160_0001812739_0001829522	16.78 s	1.642	0.7368	0.02065
Podcast	94cbf45c-d51c-4773-9059-8edc5d17f7cb_seg00025_0000325964_0000339070	13.11 s	1.703	0.6309	0.0147
Podcast	9593757b-553f-4353-b1e3-539407999da4_seg00245_0002721937_0002726749	4.81 s	1.924	0.6933	0.01629
Cartoon voices	Apple-Bloom-outtakes_s7e13-4170.998808-4175.178794--ANXIOUS	3.98 s	1.994	0.8047	0.006029
Cartoon voices	Applejack-fim_s8e9-497.661734-502.234059--NEUTRAL	4.37 s	1.737	0.7171	0.006564
Cartoon voices	Cheese-Sandwich-fim_s9e14-653.486000-657.635809--SAD	3.84 s	1.709	0.6343	0.009135
Cartoon voices	Countess-Coloratura-fim_s5e24-727.269642-731.487736--HAPPY	3.74 s	1.882	0.7439	0.008827
Cartoon voices	Fluttershy-eqg_original-1259.653566-1265.456359--NEUTRAL	5.27 s	1.735	0.6567	0.004316
Cartoon voices	Gallus-fim_s8e21-1046.852208-1051.506746--NEUTRAL	4.56 s	1.935	0.7342	0.01134
Cartoon voices	Granny-Smith-fim_s2e15-534.803485-539.480893--ANNOYED	4.42 s	2.285	1.038	0.01053
Cartoon voices	Pinkie-Pie-fim_s5e2-489.168089-493.85984--SAD	4.20 s	2.047	0.8282	0.009678
Cartoon voices	Rarity-fim_s5e14-190.023515-195.042426--NEUTRAL	4.93 s	2.097	0.866	0.01061
Cartoon voices	Sonata-Dusk-eqg_rainbow-rocks-3430.634575-3435.144400--NEUTRAL	4.34 s	2.139	0.8437	0.01157

Dataset	Clip / listen	Duration	MR-STFT error	Log-mel error	Log1p-mel error
Cartoon voices	Starlight-fim_s9e11-388.371879-393.302612--ANXIOUS	4.78 s	2.187	0.9663	0.01672
Cartoon voices	Starlight-fim_s9e20-1139.339000-1145.200306--NEUTRAL	5.58 s	1.953	0.74	0.0111
Cartoon voices	Steve-fim_s5e9-866.113039-871.879677--NEUTRAL	5.72 s	1.809	0.6587	0.00991
Cartoon voices	Timber-Spruce-egg_legend-of-everfree-1187.460833-1191.773325--NEUTRAL	4.05 s	1.615	0.6572	0.004951
Cartoon voices	Trixie-fim_s7e17-469.870933-474.054965--NEUTRAL	4.03 s	1.83	0.739	0.009147
Cartoon voices	Trixie-fim_s7e2-1274.942270-1279.107167--ANXIOUS	4.11 s	1.762	0.7573	0.009911
Cartoon voices	Twilight-Sparkle-egg_original-2977.865963-2983.116455--NEUTRAL	5.03 s	1.916	0.7611	0.00907
Cartoon voices	Twilight-Sparkle-fim_s7e1-81.575351-87.356653--NEUTRAL	5.55 s	2.24	0.7735	0.01196
Cartoon voices	Twilight-Sparkle-fim_s9e20-1085.752000-1090.464169--NEUTRAL	4.48 s	1.755	0.6671	0.009198
Cartoon voices	Yona-fim_s8e15-653.735471-658.372311--HAPPY	4.35 s	2.371	0.8325	0.01551
Podcast	a2afed6b-2524-4c0c-b38c-a15f7628c0c1_seg00177_0002197031_0002206186	9.16 s	2.043	0.7588	0.0412
Podcast	b86350d9-d8f9-4929-8624-2c3b8ef1776f_seg00155_0001090009_0001106689	16.68 s	1.66	0.642	0.01991
Podcast	cca6bedd-8fef-4712-9f47-5bd15d5955dd_seg00118_0001023234_0001040230	17.00 s	1.586	0.6662	0.02426
Podcast	cf050539-0950-42f6-a81c-cb5f0c8f2315_seg00080_0001132491_0001152474	19.98 s	1.732	0.7448	0.0209
Podcast	cfa57dd8-929a-4180-89b6-dea7060fc5bd_seg00034_0000229172_0000235913	6.74 s	1.689	0.5951	0.009893
Podcast	d98de9e2-a658-4fbb-af5d-9e9244ec29ad_seg00085_0000536937_0000547500	10.56 s	1.639	0.6412	0.02643
Podcast	e5c50436-4c39-4121-9e8d-338332db4b01_seg00022_0000317500_0000335290	17.79 s	1.821	0.6607	0.01708
Podcast	e72f590b-fb46-460a-ab6d-37b61b8f9ae1_seg00165_0001808575_0001827423	18.85 s	1.811	0.7211	0.01768
Podcast	e83d0e91-4673-47c7-b083-641fdf00dd65_seg00026_0000377039_0000392611	15.57 s	2.158	0.7927	0.02108
Podcast	ea616a91-8a10-4b6d-88ad-bd6106924304_seg00307_0001820742_0001824275	3.53 s	1.536	0.4828	0.01231
Podcast	f2621074-e35d-4b3a-bbf3-a541fc75905d_seg00059_0000424275_0000427705	3.43 s	1.698	0.5519	0.01751

What this model actually does

01 / SEMANTIC TOKENS

25 semantic IDs per second

The S3 tokenizer emits one 6,561-way FSQ ID every 40 ms. These are semantic tokens, not independent acoustic RVQ codebooks.

02 / CONDITIONAL FLOW

Ten flow evaluations

The original S3Gen checkpoint uses token-to-mel generation with global attention and ten Euler steps. It also receives a fixed reference prompt and speaker embedding.

03 / WAVEFORM

HiFT at 24 kHz

A neural vocoder converts the generated mel features into the reconstructed waveform. This evaluates original S3Gen, not T3, Turbo, or another Chatterbox checkpoint.

Stage	Parameters	Exact count	Non-parameter buffer elements
Semantic tokenizer encoder	123.715 M	123,714,568	26,128
Speaker conditioner	6.849 M	6,848,544	82,298
Token-to-mel generator	112.549 M	112,549,360	0
HiFT vocoder	20.821 M	20,821,295	0
Total installed model	263.934 M	263,933,767	109,386
Decoder only (flow + vocoder; subtotal)	133.371 M	133,370,655	0

> **Conditioning is additional side information.** Each clip uses prompt tokens, 50 Hz mel features, and a 192-dimensional speaker embedding extracted once from its first up to three seconds (observed range 2.04–3.00 s). This same-source reference is available offline. The experiment is conditional semantic resynthesis, not a bitrate-matched ranking against unconditional audio codecs.

Architecture evidence and limits

The pinned implementation’s token encoder and flow network use global attention. Future context can change earlier representations, and the flow state changes at each Euler step; a naive persistent transformer KV cache is therefore not equivalent to full decoding.

Earlier three-anchor perturbation probes changed future context and observed changes in earlier waveforms; they also observed changing attention keys across flow steps. These support the architectural explanation only. They do not establish population-level chunking quality, a minimum usable window, or an online streaming guarantee.

The existing full-control audit found zero above-threshold harmonic-excitation frames on these 60 full clips. Internal predictor outputs are not measurements of speaker pitch, and the predictor is not mathematically bounded. A three-anchor phase intervention produced no excitation-source difference under those conditions; this is a checkpoint-and-sample-qualified negative result, not a universal statement.

Primary sources: S3Gen composition, token-to-mel flow, Euler integration, attention estimator, and HiFT vocoder. Architecture probes and the exact protocol are downloadable below.

Full-decoder throughput

13.8567× realtime: 34.5676 seconds of cached full-decoder time for 478.9917 seconds of audio. Measured on the shared NVIDIA L40S at batch size 1, FP32, TF32 off.

This includes full decoding and return-details overhead but excludes model loading, semantic encoding, source and prompt preparation, metrics, and file export. It uses already prepared conditioning. It is neither end-to-end throughput nor streaming latency, and no CPU result is inferred.

Measurement protocol

The original listening files are the exact analysis references: channels averaged to mono and resampled once to 24 kHz. They are not bit-identical to original-rate corpus files. Paired waveforms use source-length cropping, no independent gain normalization, and no alignment search. The native startup fade is applied once per complete reconstruction. Random inputs and conditioning are fixed per clip. Listening exports are float32 WAV, with no extra lossy transcoding.

MR-STFT
Spectral convergence plus log-magnitude L1 at FFT sizes 512, 1024, and 2048.

Log-mel

L1 of $\ln(M + 10^{-5})$, using 80 magnitude-mel bands.

Log1p-mel
L1 of $\ln(1 + M)$, using the same mel transform.

These 60 speech clips do not establish music performance, multilingual ASR preservation, perceptual equivalence, or production streaming latency. Float WAV values can exceed ±1; device playback may clip without changing measured file bytes.

▸ **Exact full-only report protocol**

Download the evidence

Measured report JSON · Aggregate CSV · Per-clip CSV · Independent validation · All 60 audio pairs and hashes · Architecture probes · PDF report

---