CLEAN-SPEECH TRANSPARENCY IS NOT DEPLOYMENT TRANSPARENCY: PREDICTING AND REPAIRING THE WORDS NEURAL AUDIO CODECS LOSE ON DEGRADED SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural audio codecs are benchmarked on clean speech, but deployed speech systems encode degraded audio. We find that clean-speech transparency understates the deployed loss several-fold yet predicts it, and that a codec-dependent share of it arises at the codeword lookup. Across ten operating points from six frozen codec families, the word error rate (WER) a codec adds for a recognizer reading its decoded audio is, at 5 dB SNR, 4.5–22$\times$ its clean value, and two recent recognizers replicate it. A power law of the clean loss, fitted on four families and then fixed, predicts held-out codecs, prospective measurements, and six codec families predicted blind (60 measurements: median error 1.32$\times$, 90% within 2$\times$, against 1.47$\times$ and 75% for a constant ratio). Tokenizers built on ASR-trained encoders fall below it. A quantizer-bypassed path through the same decoder attributes 6–96% of the loss, a lower bound, to the lookup, and re-fitting the codebook on noisy speech recovers little. We therefore repair the lookup's input: an identity-initialized causal adapter between the frozen encoder and quantizer, trained by latent regression, halves Mimi's added WER at 5 dB under noise of its training type (21.32 to 11.25 points) with clean WER and 99.8% of clean semantic codes unchanged. Mimi's quantization penalty falls from 17.57 to 6.43 points while its bypassed path does not improve. The adapter removes less on two acoustic codecs and lets a small recognizer trained on clean tokens read degraded speech better than one retrained on degraded tokens. Placed after the lookup or in its place, the same network raises the loss on clean speech and under unseen babble.

1 INTRODUCTION

Neural audio codecs serve as transmission channels whose decoded audio a recognizer transcribes and as tokenizers whose tokens a speech language model reads or emits. They are selected on clean-speech leaderboards (Mousavi et al., 2025; Ye et al., 2025a; Gong et al., 2026), although codecs are known to lose more words on noisy input (Tseng & Harwath, 2025; Della Libera et al., 2025; Behringer et al., 2026), as speech coders did before them (Euler & Zinke, 1994). For frozen checkpoints, three questions remain open: whether clean-speech measurements predict the deployed loss, where in the encoder, quantizer, and decoder the loss arises, and whether a codec whose vocabulary other models depend on can be repaired without changing that vocabulary.

Clean-speech measurements do predict the deployed loss (Figure 1a). At 5 dB SNR, the word error rate a frozen codec adds to its input is 4.5–22$\times$ its clean value at all ten operating points from six families, and two recent recognizers replicate the amplification. A power law of the clean loss, fitted on four families and then fixed, predicts held-out codecs, prospective measurements, and six codec families predicted blind, as accuracy-on-the-line predicts a classifier's accuracy under distribution shift (Miller et al., 2021). Its exponent is below one, so the codecs most transparent on clean speech amplify most. Clean rankings largely survive across rate tiers, but no clean WER budget carries over, and at 5 dB DAC performs as it would on clean speech at $2^{-1.5}$ of its rate.

In quantized codecs a large, codec-dependent share of the loss arises at the lookup. Decoding a quantizer-bypassed path through the same frozen decoder attributes 6–96% of it to the codeword

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

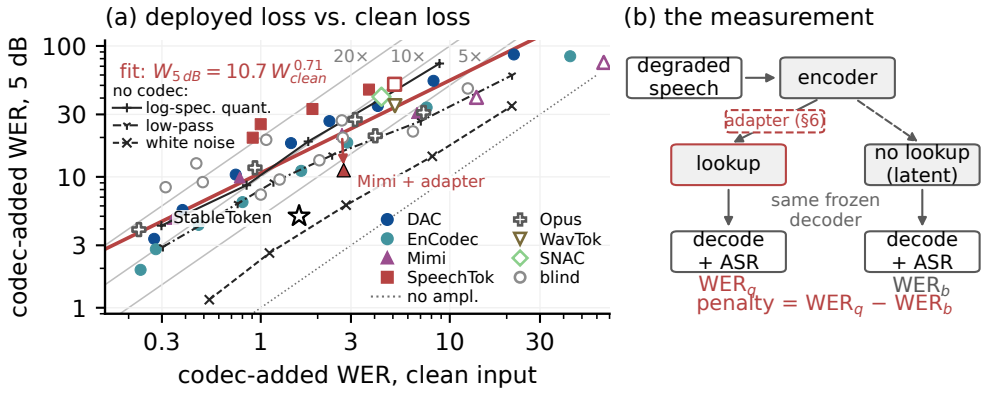


Figure 1: **Clean-speech transparency is not deployment transparency.** (a) Codec-added WER at 5 dB input SNR against clean input for every rate-ladder operating point we measured (2,620 utterances, Whisper-small). Every point sits above the no-amplification diagonal, the ten operating points of Table 1 by 4.5–22×; one power law fitted on the filled points (red line) predicts the hollow ones, held-out codecs (Opus included) and rungs measured after the fit was fixed. Black lines: distortions with no codec, applied to the same inputs; white noise amplifies about 2×, a log-spectral quantizer follows the law. The star is StableToken, a tokenizer trained for noise robustness; grey circles are families predicted blind; the red arrow is Mimi with the pre-lookup adapter of Section 6, which moves it below the line without changing its clean loss. (b) The measurement: the bypassed path decodes the continuous latent through the same frozen decoder, differing from production decoding only in the lookup; the penalty is $\mathrm{WER}_q - \mathrm{WER}_b$. Dashed: where the adapter sits.

lookup. A rate-matched scalar quantizer loses more, recognizers trained on the tokens also incur the penalty, 2.3–3.3× more at 5 dB than on clean speech, and re-fitting the codebook on noisy speech recovers a fifth of one codec's penalty and none of another's. The lookup is not the only amplifier: a continuous tokenizer with no lookup, measured blind, sits above the law.

We repair the lookup's input with an identity-initialized causal adapter that sits between the frozen encoder and the frozen quantizer, maps a degraded pre-lookup latent toward the clean one, and leaves codebooks, decoder, and vocabulary untouched. Trained by latent regression, the loss of latent-space enhancement (Li et al., 2025; Kammoun et al., 2026), it halves Mimi's added WER at 5 and 0 dB with clean WER and 99.8% of clean semantic codes unchanged. The quantization penalty falls by 63% while the bypassed path gains nothing, so the gain is realized at the lookup. On two acoustic codecs it removes less, as much as or more than re-fitting their codebooks. A recognizer trained on clean tokens reads the repaired tokens better than one retrained on degraded tokens reads the original ones, and adapted Mimi loses about what unmodified Mimi loses at twice the bitrate. Our controls show that placement matters more than the loss: the same network after the lookup is worse in every condition, and a learned classifier replacing the lookup loses on clean speech.

Our contributions are the following.

- **A clean-to-deployment law.** Codec-added WER at 5 dB is 4.5–22× its clean value at all ten operating points under Whisper-small, and two recent recognizers replicate it. A power law of the clean loss predicts it for held-out, prospective, and blind codecs and, refitted without the provisioned family, makes 11 of 15 rate decisions right against 9 for a constant ratio (Section 4).
- **An attribution to the lookup.** A quantizer-bypassed path attributes 6–96% of the loss to the codeword lookup (a lower bound). The lookup's displacement is structured rather than isotropic, a rate-matched scalar quantizer built post hoc costs more words, and re-fitting the codebook recovers little (Section 5).
- **A repair that keeps the vocabulary.** Latent regression (Li et al., 2025; Kammoun et al., 2026) as an identity-initialized adapter before the frozen quantizer halves Mimi's added

WER at 5 and 0 dB and removes most of its quantization penalty. A token reader gains without retraining, and in our controls placement matters more than the loss (Section 6).

- **A metric audit.** STOI, WavLM distance, bitrate alone, and log-mel distortion against the codec's input rank codecs under degradation. For PESQ, UTMOS, log-mel distortion against the clean reference, and the token flip rate we detect no ranking (Section 7).

2 Background and Related Work

Residual vector quantization (RVQ) turns speech into the tokens that speech language models read and emit (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023; Zhang et al., 2024; Défossez et al., 2024): each codebook quantizes the residual the previous one left, so the number of codebooks sets the bitrate. Released checkpoints allocate their bits differently (rate ladders in DAC and EnCodec, multi-scale codebooks in SNAC, a semantic first layer in Mimi and SpeechTokenizer, one index per frame in WavTokenizer, semantic supervision in X-Codec (Ye et al., 2025b), a binary spherical codebook in FocalCodec (Della Libera et al., 2025)); each is trained to reconstruct its input, not to be invariant to degradation.

Codec and tokenizer papers rank their models on clean speech (Ye et al., 2025a; Gong et al., 2026); the survey of discrete audio tokens and its benchmarks add downstream tasks, some on noisy input (Mousavi et al., 2025; 2024; Wu et al., 2024). That codecs lose more words on degraded input is known. Tseng and Harwath measure EnCodec, DAC, and SpeechTokenizer under WHAM! noise and reverberation (Tseng & Harwath, 2025), FocalCodec reports resynthesis of noisy LibriSpeech scored against the clean reference (Della Libera et al., 2025), benchmarks add noisy mixtures and token stability under noise (Deng et al., 2025; Vashishth et al., 2024), and listening tests find neural codecs less robust than classical ones at their operating points (Behringer et al., 2026). Speech coding documented the same effect for CELP coders (Euler & Zinke, 1994), with recognition from the bitstream as one remedy (Kim et al., 2002). None of these predicts the deployed loss from clean measurements, expresses it in bitrate, or measures the lookup's causal share of it.

Robust tokenizers change how the codec is trained and therefore its vocabulary: NAST (Messica & Adi, 2024), R-Spin (Chang & Glass, 2024), NDVQ (Niu et al., 2024), progressive quantization perturbation (Zheng et al., 2026), noise-robust variable-bitrate RVQ (Chae et al., 2025; Chae & Lee, 2025), and StableToken (Song et al., 2026). Enhancement in a codec's continuous latent space maps the degraded latent toward the clean one (Li et al., 2025; Kammoun et al., 2026; Li et al., 2026) or fine-tunes the codec (Kammoun et al., 2026; Zhao & Madhu, 2026), where a cross-entropy over the codebook's own distances lost to regression (Kammoun et al., 2026); separate predictors of clean tokens or units keep a frozen consumer (Gat et al., 2023; Poncelet et al., 2024; Lu et al., 2025; Liu et al., 2025b; Ashihara et al., 2026; Godiva et al., 2026). These are scored on perceptual quality or on one consumer, whereas we score the words lost and tokens read by consumers that are never retrained, and compare placements and losses at one budget (Section 6).

Our law reads codecs as accuracy-on-the-line reads classifiers (Miller et al., 2021): the clean loss predicts the deployed one, and a codec's distance below the line is the analogue of effective robustness (Taori et al., 2020). The optimal encoder of a noisy source first estimates the clean one (Wolf & Ziv, 1970); a codec that encodes the noisy observation directly also encodes the noise, so degradation inflates the rate a given loss requires.

3 Measuring the Quantization Penalty

Our measurement isolates one step inside a frozen codec: we decode the same input twice through the same released checkpoint, once as the codec normally runs and once with the codeword lookup replaced by its continuous counterpart, and a recognizer transcribes both.

Frozen codecs and operating points. We evaluate six public codec families exactly as released, retraining nothing: DAC (Kumar et al., 2023), EnCodec (Défossez et al., 2023), Mimi (Défossez et al., 2024), SNAC (Siuzdak et al., 2024), SpeechTokenizer (Zhang et al., 2024), and the single-codebook WavTokenizer (Ji et al., 2025) at 0.9 kbps, with the classical codec Opus (Valin et al., 2012) at 4–16 kbps as a reference (Appendix A). Table 1 fixes ten of these as the operating points

every later section uses. Mimi appears there at 8 codebooks (1.1 kbps) and SpeechTokenizer at 8 layers (4.0 kbps); cross-codec ranges run over these ten points. The lowest rungs of DAC, EnCodec, and Mimi (1.0, 0.75, and 0.14 kbps) exceed 80% WER at 5 dB and are marked as saturated.

The quantizer-bypassed path. In an RVQ codec, layer l projects the residual $\mathbf{r}_l$ with $\mathrm{in_proj}_l$, replaces it by its nearest codeword, projects back with $\mathrm{out_proj}_l$, and passes the remainder on. We keep every step except the codeword replacement. Each layer contributes its continuous projection output, and we decode the sum

$$\hat{\mathbf{z}} \;=\; \textstyle\sum_l \mathrm{out_proj}_l(\mathrm{in_proj}_l(\mathbf{r}_l)) \tag{1}$$

through the same frozen decoder. Re-enabling the lookup inside this path reproduces each codec's official reconstruction exactly, verified at every operating point under deterministic convolutions. The two paths therefore differ in the lookup alone. On clean input the bypassed path adds 0.06–4.10 WER points over the input (Appendix A). The bypass is not a neutral counterfactual, since most of these decoders never saw a continuous latent in training (Mimi, trained with quantization skipped on half its sequences (Défossez et al., 2024), is the exception). At 5 dB DAC 1.5's bypassed path alone scores 41.68 against the input's 8.45. Under Whisper-large-v3 the quantized path beats the bypassed one on clean input at DAC 3.0 and 6.0 kbps (by 0.79 and 0.39 points), where the bypassed path is itself worse than the uncoded input. If decoding a continuous latent only adds errors, the penalty understates the quantizer's cost, and we read every share below as a lower bound. EnCodec, SpeechTokenizer (identity projections), and WavTokenizer have no codebook projections, so their bypassed sum in Eq. 1 is exactly the encoder latent. DAC, SNAC, and Mimi (512→256) keep their projections, so for those three the measurement is a lower bound on the quantizer module's cost.

Degradations, recognizers, and intervals. The degradations are additive noise (WHAM!) (Wichern et al., 2019) at 15, 10, 5, and 0 dB SNR, six-talker babble, and reverberation in simulated rooms (RT60 400 ms). Each utterance's noise draw is keyed to its identity and shared across arms, so every comparison is paired. Whisper-small (Radford et al., 2023), the recognizer FocalCodec's word-error evaluation uses (Della Libera et al., 2025), is the primary recognizer. Both wav2vec2-base-960h (Baevski et al., 2020) and whisper-large-v3 replicate the amplification and the attribution, with one reversal under wav2vec2 (Section 5). Two recent recognizers, the CTC model Parakeet-CTC 1.1B (Rekesh et al., 2023) and the speech language model Voxtral-Mini 3B (Liu et al., 2025a), replicate the amplification at 9 and 10 of the ten points and rank the codecs at 5 dB as Whisper-small does (Spearman 0.99 and 1.00; Table 11). We use the full official LibriSpeech (Panayotov et al., 2015) test-clean, all 2,620 utterances. The random-displacement controls of Section 5 and the metric study of Section 7 use a frozen stratified 500-utterance subset, and we mark those results as such. Decoding is deterministic for five of the six codecs, so we report no seeds; SNAC's decoder injects noise. We attach an utterance-clustered 95% confidence interval (CI) to every result, take penalty intervals from paired differences, and back the metric rankings with permutation tests. WER is the mean of per-utterance WERs on text upper-cased and reduced to letters, apostrophes, and spaces, so Whisper's digits count as errors, a bias that largely cancels in codec-added WER; the uncoded input scores 4.93 on clean speech and 8.45 at 5 dB. At corpus level the amplification spans 5.2–27.5× and the quantizer's share 7–96% (Appendix A.7). Because the decoder is a deterministic function of the token stream, our word counts lower-bound what the tokens retain, and the bound depends on the reader. Readers of the tokens themselves incur the loss too: at Mimi's 4.4 kbps, Kyutai STT 1B (Zeghidour et al., 2025), which reads Mimi tokens directly, matches decoding plus Whisper-small (paired CIs include zero on clean input and at 5 dB). At 1.1 kbps we trained Conformer CTC recognizers that differ only in their input, Mimi's tokens, its continuous pre-lookup latent, or uncoded log-mel (Appendix A.6): the quantizer's penalty for a token reader, tokens minus latent, each reader trained on its own input, grows from 2.56–3.71 points on clean input to 7.36–9.10 at 5 dB over two seeds (2.3–3.3×), and, trained on degraded speech, the tokens' cost over log-mel grows 8.4×.

Definitions. *Codec-added WER* is the WER after encode–decode minus the WER of the same input before it, so it removes the recognizer's own errors on degraded speech. The *amplification factor* is the ratio of codec-added WER on degraded input to codec-added WER on clean input at the same operating point. The *quantization penalty* is $\mathrm{WER}(\text{quantized}) - \mathrm{WER}(\text{bypassed})$ on the same input: the part of codec-added WER that disappears when the quantizer module is bypassed and the same decoder reads the continuous latent. The token flip rate is the fraction of codeword

indices that differ between clean and degraded input; WavLM distance (Chen et al., 2022) and log-mel distortion compare the codec output with the clean signal (definitions in Appendix A).

4 Clean-Speech Transparency Understates Deployment Loss Several-Fold

We find that codec-added WER at 5 dB is 4.5–22× its clean-speech value, at every one of the ten operating points. Table 1 gives the two columns and their ratio, on the full 2,620-utterance test-clean with additive noise (WHAM!). At 5 dB the five points below 2 kbps add 21.32–54.16 WER points to their own input, while on clean speech all ten points add 0.27–8.20, five of them less than 2, and a reader of that column alone would call most of these codecs close to transparent.

The spread of the amplification factor is not noise: the clean loss sets it. Across the 24 rungs of the DAC, EnCodec, Mimi, and SpeechTokenizer rate ladders, the 5 dB loss follows one power law of the clean loss, the clean-to-deployment law $W_{5\,\mathrm{dB}} = 10.7\,[9.4, 13.2] \cdot W_{\mathrm{clean}}^{0.71\,[0.60,0.76]}$ (utterance bootstrap), with a residual factor of 1.5 (Figure 1a). The law predicts codecs it was not fitted on: WavTokenizer within 3%, SNAC within 35%, and the classical codec Opus within 5–15% at 6–12 kbps, so classical codecs found more robust in listening tests (Behringer et al., 2026) are, by the law, codecs with a lower clean loss at their higher rates. It also holds prospectively: on 32 measurements run after its coefficients were fixed (new rungs and SNRs of Mimi and SpeechTokenizer; Opus at 4–12 kbps and 15–0 dB), the loss is within a factor of 1.29 of the prediction at the median and within 2× in 84% of them (Appendix A.8). Predicting from each codec's clean loss before any noisy measurement, it places six families we had never measured within 2× in 90% of 60 measurements (median factor 1.32; a fixed constant ratio: 75%, 1.47; Table 19): X-Codec (v1, LibriSpeech-trained) (Ye et al., 2025b), the Higgs Audio v2 tokenizer (Boson AI, 2025), FocalCodec (Della Libera et al., 2025) and MOSS-Audio-Tokenizer (Gong et al., 2026) at three rates each, the GLM-4-Voice tokenizer (Zeng et al., 2024), and the continuous VibeVoice tokenizer (Peng et al., 2025). Per family the loss runs from 0.51 (GLM-4-Voice) to 1.74× (X-Codec) of the prediction; VibeVoice, which has no lookup, sits 1.49× above the line. The exponent is below one inside every family (families × speakers bootstrap [0.55, 0.78]; Table 17), partly a WER-ceiling effect, so the amplification factor $10.7\,W_{\mathrm{clean}}^{-0.29}$ grows as a codec becomes more transparent on clean speech, which is why the most transparent points amplify most. The same form holds at 15, 10, and 0 dB with exponents 0.90, 0.84, and 0.56 (Appendix A.8). Two families sit off the line in leave-one-family-out prediction: EnCodec at 0.55× and SpeechTokenizer at 2.3× the predicted loss. Every amplification interval in Table 1 has a lower bound of 4.2 or more (4.0 when speakers rather than utterances are resampled).

The amplification is not a generic recognizer effect, and it needs no learned tokens. At DAC 3.0 kbps's clean loss (1.45 points), white noise amplifies 2.3× at 5 dB, a low-pass filter 7.3×, and a uniform quantizer of the log-magnitude spectrum, with no learned part, 10.4×, against the codec's 12.5× (Figure 1a): what amplifies is a lossy transform whose error follows its input, noise included. Recognizer unfamiliarity with codec output explains about half of the loss, not its growth. Fine-tuning Whisper-small on 91 h of degraded training speech decoded by the codec, against a twin fine-tuned on the same waveforms uncoded, roughly halves the loss but keeps its growth, from 1.03 to 9.63 points at DAC 3.0 kbps (9.3×) and from 1.51 to 12.13 at Mimi (8.0×; Appendix A.5). A codec's distance below the line is its robustness at matched clean loss, the analogue of effective robustness (Taori et al., 2020). StableToken (Song et al., 2026), trained for noise robustness and decoded by its released decoder, adds 1.59 points on clean speech and 5.01 at 5 dB, a third of the law's prediction (0.34). Its non-robust counterpart, the GLM-4-Voice tokenizer (Zeng et al., 2024), also a quantized Whisper encoder, already sits at 0.43, so most of that distance comes from the ASR-trained front end. The law's form holds under babble, reverberation (over-predicting every rung), and wav2vec2, on measurements run after condition-specific coefficients were fixed, and on real far-field speech. On AMI meetings (Carletta et al., 2005) recorded by headset and distant microphone, all ten points amplify (1.4–4.8×), and a law fitted on them and fixed predicts 16 further rungs within 1.16× at the median (Appendix A.8). It also reproduces on VoiceBank–DEMAND (Appendix A).

Calibrated with one measurement, the law transfers to new conditions. With b fixed, one measured codec in a new condition (one-anchor calibration) predicts the others to a median factor of 1.28 over eight new conditions, against 1.22 for a full refit and 1.64 for an equally anchored constant ratio, whose worst anchor misses by 13.9× against the law's 3.2× (Table 18). On the held-out

Table 1: Clean-speech codec-added WER understates the 5 dB value by 4.5–22× at all ten operating points, while largely preserving their ordering. Codec-added WER (%, Whisper-small) on the full official LibriSpeech test-clean (2,620 utterances), clean and with additive WHAM! noise at 5 dB SNR. The amplification factor is the 5 dB value divided by the clean value, with an utterance-bootstrap 95% CI (5,000 resamples); the penalty is WER(quantized) − WER(bypassed) at 5 dB, with an utterance-clustered 95% CI. SNAC's decoder is stochastic: over three runs its clean and 5 dB values span 4.26–4.35 and 40.27–42.28. WER is the mean of per-utterance WERs; Table 14 (Appendix A.7) gives the corpus-level version.

Codec (kbps)	Clean	5 dB	Ampl. [95% CI]	Penalty (5 dB)
WavTokenizer 0.9	5.13	34.96	6.8× [6.3, 7.4]	2.10±1.63
SNAC 1.0	4.35	40.91	9.4× [8.5, 10.5]	18.00±1.80
Mimi 1.1	2.69	21.32	7.9× [6.9, 9.3]	17.57±1.95
DAC 1.5	8.20	54.16	6.6× [6.1, 7.2]	20.92±2.46
EnCodec 1.5	7.54	33.88	4.5× [4.2, 4.9]	32.54±1.29
DAC 3.0	1.45	18.11	12.5× [10.6, 15.3]	9.55±0.88
EnCodec 3.0	1.64	11.12	6.8× [5.8, 8.2]	9.77±0.61
SpeechTok. 4.0	0.91	19.92	22.0× [17.6, 28.8]	6.62±1.19
DAC 6.0	0.27	3.36	12.3× [7.8, 28.2]	2.04±0.37
EnCodec 6.0	0.47	4.30	9.2× [6.3, 16.5]	2.95±0.39

test-other corpus, where the amplification is smaller (2.3–9.7×), the law over-predicts by a median factor of 1.53 and one anchor measurement brings it to 1.25 (Table 19), consistent with a corpus-specific constant. Prospectively its error is smaller than a constant ratio's in 22 of 32 measurements (Wilcoxon p=0.012, or 0.045–0.058 clustered by operating point; Table 17). The law also supports rate decisions. Choosing the lowest rate within a 3, 5, or 10-point budget at 5 dB (rate provisioning, where right means a compliant rate, or none when none exists), the in-sample law is right in 12 of 15 cases and the constant ratio in 9. Counting AMI, the laws are right in 26 of 30 (9 of 13 feasible budgets met, 17 of 17 correctly reported infeasible), where a clean leaderboard is right in none. Fitted without the provisioned family, the law is right in 11 of 15.

Finding 1. Codec-added WER at 5 dB is 4.5–22× its clean-speech value at all ten operating points, and one power law of the clean loss predicts it for codecs it was not fitted on, including the classical codec Opus and six families measured blind. Rankings largely survive (Whisper-small, wav2vec2) but budgets do not: at 5 dB DAC performs as on clean speech at $2^{-1.5}$ of its rate.

The law explains why a clean leaderboard mostly picks the right codec and the wrong rate. The map is monotone, so across rate tiers the ordering largely carries over (Spearman $+0.903$, p=0.0003, under both Whisper-small and wav2vec2), and its constant of about 11 means no clean-speech budget does. The ordering is fragile where current tokenizers compete, at 1.5 kbps and below: there the family offsets swap EnCodec and SNAC ($+0.60$ over five points), and under Whisper-large-v3, whose clean loss sits near its floor at most points, it survives only weakly ($+0.552$, p=0.098; $+0.855$ at corpus level). The last column of the table attributes part of the 5 dB loss to the codeword lookup, and Section 5 takes that attribution up.

Degradation also inflates the rate a codec needs. Reading each degraded measurement against its family's clean rate ladder gives its clean-speech equivalent rate. DAC at 5 dB performs like DAC on clean speech at $2^{-1.5}$ of its rate, a rate inflation of 1.47 to 1.56 octaves (rate doublings) at every rate from 3 to 6 kbps. It grows by about one octave per 10 dB (0.58, 0.99, 1.49, and 2.03 octaves at 15, 10, 5, and 0 dB; Figure 2c, Appendix A). EnCodec shows the same median inflation, rising with rate, and the three families not used to define it show their own slopes: at 5 dB Mimi 2.28 and Opus 1.23 octaves, and SpeechTokenizer more than its whole clean ladder below 15 dB. A codec paper's budget at a given rate is therefore a clean-speech budget: holding codec-added WER within 3 points takes DAC from 2.5 kbps on clean speech to its 6 kbps maximum or beyond at 5 dB.

5 THE QUANTIZER'S SHARE, AND A RATE-MATCHED SCALAR QUANTIZER LOSES MORE WORDS

The quantizer accounts for 6–96% of the 5 dB loss (Table 1): 32.54 of a 33.88-point loss at EnCodec 1.5 kbps (96%), and 2.10±1.63 of 34.96 at the single-codebook WavTokenizer (6%). Where the bypassed path is exactly the encoder latent (EnCodec, SpeechTokenizer, WavTokenizer) the share spans 6–96%, and where it keeps rank-limited projections (DAC, Mimi, SNAC) 39–82%. WavTokenizer's 6% is a lower bound, as its bypassed path alone adds 32.86 points. Four results follow; the controls behind them run at the DAC and EnCodec points unless stated.

The added errors look like the recognizer's own: substitutions are 65–75% of the edits the lookup adds at 5 dB, as they are 73% of the recognizer's own errors on the uncoded input (60–79% under wav2vec2-base; Table 4, Appendix A).

Against a random displacement of the latent with the same per-frame norm, decoded by the same decoder and read by the same recognizer, the quantizer's displacement costs 1.7–45× more words, with paired differences excluding zero at all six DAC and EnCodec points ($+1.07$ $[+0.27, +1.87]$ to $+28.68$ $[+26.52, +30.84]$). A random displacement confined to the subspace the decoder reads still costs far fewer words: 0.22, 1.60, and 1.11 WER points at DAC 6, 3, and 1.5 kbps against the quantizer's 1.34, 9.91, and 20.19 (500-utterance subset; paired $+1.12$ $[+0.39, +1.85]$ to $+19.07$ $[+15.34, +22.81]$). The quantizer's displacement is structured, and its word cost is not a property of its norm; at matched WavLM footprint, however, random displacements cost as many words as the lookup (EnCodec 1.5 excepted).

A rate-matched scalar quantizer costs more words than the learned codebook. Because the bypassed path carries no bits, the penalty could be inherent to any quantizer at that rate, or specific to the learned lookup. We test this by replacing the lookup in place with the strongest generic quantizer we could build without a codebook, a post-hoc analogue of the finite scalar quantization that recent codecs train end to end (Mentzer et al., 2024). It is a Karhunen–Loève transform fitted on clean speech, followed by uniform scalar quantization with one step size, bisected on held-out speech at the same SNR until the arm meets its bit budget (Appendix A). Encoder, projections, decoder, and recognizer are unchanged, and the comparison is paired over all 2,620 utterances at 5 dB. We run two budgets, the ladder's nominal rate and the empirical entropy of its index stream (1–14% below nominal at 5 dB, as the released codebooks are not fully used); both favour the scalar arm.

The scalar arm loses more words at all six points under both budgets (Figure 4b and Table 7, Appendix A.2), from $+0.47$ $[+0.05, +0.89]$ at DAC 6.0 kbps to $+15.36$ $[+13.19, +17.52]$ at EnCodec 1.5 kbps, with ten of twelve intervals excluding zero. It loses although the nominal budget hands it up to 17% more bits than the RVQ index stream carries and it overshot its budget at EnCodec 3.0 and 6.0 kbps (Table 7). Its decoder-input distortion is 1.18–1.88× RVQ's, the known advantage of vector quantization (Lookabaugh & Gray, 1989). A dead-zone variant that matches it still loses at the three DAC points, and a second calibration corpus changes none of the 12 signs (Appendix A). It shows which rate-matched quantizer costs fewer words, not that rate alone sets the loss.

Finding 2. At matched bitrate, the strongest generic quantizer we could build post hoc, with the decoder not adapted to it, loses 0.47–15.36 more WER points than the learned codebook at all six DAC and EnCodec points (ten of twelve intervals exclude zero); re-fitting the codebook on noisy speech recovers a fifth of EnCodec's 1.5 kbps penalty under synthetic noise, not on real recordings, and none of DAC's. The lookup's displacement is structured: it costs 1.7–45× more words than an isotropic displacement of the same norm.

Whether the codebook's clean-speech fit matters depends on the codec. Re-fitting DAC's 12 codebooks on degraded training latents at identical bitrate lowers the latent quantization error by 4–7% and leaves the penalty unchanged ($+1.24$ $[-2.07, +4.55]$ at 1.5 kbps). The same re-fit of EnCodec's first eight codebooks recovers 7.04 ± 1.33 of its 32.54-point penalty at 1.5 kbps, and 6.66 ± 5.03 under babble it never saw, but only 1.29 ± 0.59 at 3.0 kbps and nothing on real far-field recordings; SpeechTokenizer's gains nothing significant (Appendix A.2). Noise changes how well the lookup fits only modestly, while the penalty grows 4.4–7.4× at the EnCodec points (Appendix A.2).

Table 2: A pre-lookup adapter on frozen codecs: codec-added WER (Whisper-small, full test-clean). Within each codec, every network arm is the same causal Transformer on the same WHAM!-degraded training speech and differs only in loss or placement (one DAC row: twice the training speech; the linear map is one affine layer; the classifier adds eight output heads). Babble: six talkers at 5 dB, unseen in training. Kept: clean-input codes left unchanged (Mimi's semantic stream, the first stage of DAC and EnCodec). Reader: WER of the recognizer trained on clean Mimi tokens, reading the transmitted tokens at 5 dB, not retrained.

Mimi 1.1 kbps	Clean	15 dB	10 dB	5 dB	0 dB	Babble	Kept (%)	Reader
Frozen codec (Table 1)	2.69	5.47	9.28	21.32	40.44	36.81	100.0	41.62
Linear map, closed form	2.89	5.15	8.15	16.01	30.83	39.10	90.5	34.30
Latent regression (MSE)	2.75	4.03	5.45	11.25	21.62	33.80	99.8	29.96
Code-assignment cross-entropy	2.91	3.84	5.66	10.87	22.01	34.09	91.6	31.06
Regression + cell hinge	2.64	3.80	4.99	11.45	21.40	35.22	99.4	31.36
Same network after the lookup	4.76	7.08	10.97	22.07	41.62	58.62	100.0	41.09
Learned classifier replaces lookup	13.75	15.50	19.60	27.09	39.39	68.55	88.5	33.26
DAC 3.0 kbps								
Frozen codec (Table 1)	1.45	4.31	8.41	18.11	37.42	31.92	100.0	–
Latent regression (MSE)	1.46	3.40	6.39	15.51	32.20	41.03	98.3	–
Latent regression, twice the speech	1.52	2.87	5.57	13.06	27.93	41.06	98.2	–
EnCodec 1.5 kbps								
Frozen codec (Table 1)	7.54	13.52	19.54	33.88	54.02	56.33	100.0	–
Latent regression (MSE)	7.78	11.12	15.67	26.04	42.02	53.67	97.2	–

One attribution reverses: under wav2vec2, SpeechTokenizer's quantized path scores 10.70 ± 0.75 WER points *better* than its bypassed path, perhaps because a self-supervised readout prefers the HuBERT-distilled first codebook (Zhang et al., 2024) to the latent it was distilled from.

6 Repairing the Lookup's Input in a Frozen Codec

Section 5 located a codec-dependent share of the loss at the codeword lookup and found that refitting the codebook on noisy speech recovers little of it; we therefore repair the representation before quantization. A small causal network between the frozen encoder and the frozen quantizer maps the degraded pre-lookup latent toward the clean one, while encoder, codebooks, and decoder stay as released; the vocabulary is unchanged, so models that read the codec's tokens accept the repaired ones as they are. The adapter is a six-layer causal Transformer (19.4M parameters, beside Mimi's 79.3M) whose residual output layer starts at zero: untrained, it leaves the codec bit-identical, and being causal it adds no look-ahead to a streaming codec. We train it on train-clean-360 mixed with WHAM! training noise at 0–20 dB, the audio our degraded token readers saw, with the clean utterance's latent and codes as targets, and test it on the waveforms of Table 1, whose noise recordings it never heard, and on six-talker babble, a degradation it never saw (Appendix A.3).

The arms share network, data, and schedule and differ in one choice each (Table 2). Three losses are compared. Regression of the clean latent (MSE), as in latent-space enhancement (Li et al., 2025; Kammoun et al., 2026), is the method we keep. A code-assignment cross-entropy makes the codebook itself the classifier, with logits the negative distances to the frozen codewords, the clean codes as targets, and residuals teacher-forced through the stages. Regression plus a cell hinge, $\sum_k \max\big(0,\ d(\mathbf{r}_k, \mathbf{e}_{k,t_k}) - \min_{j \neq t_k} d(\mathbf{r}_k, \mathbf{e}_{k,j})\big)$, is zero whenever each residual lies in its clean cell, so the clean latent remains an exact minimizer. Three controls move the network: after the lookup, onto the dequantized codes; in place of it, as a learned classifier; and down to a linear map.

Repairing the lookup's input halves Mimi's loss (Table 2). At 5 dB regression lowers Mimi's codec-added WER from 21.32 to 11.25 and at 0 dB from 40.44 to 21.62. On clean speech it changes nothing measurable ($+0.06 \pm 0.18$ points; 99.8% of its semantic and 93.8% of all clean codes unchanged). Decoded through the bypassed path of Section 3, the adapted latent gains nothing (3.75 to 4.82 points at 5 dB), while the quantization penalty falls from 17.57 to 6.43 (Appendix A.3), so the gain arises at the lookup. On the rate ladder of Section 4, adapted Mimi at 1.1 kbps loses at 5 dB

about what the unmodified codec loses at twice the rate (2.2 kbps: 9.91), and it moves from the law's line (0.99 of its prediction) to 0.51 on the noise type it was trained for. Three training seeds span at most 0.87 points at 5 dB for every loss on Mimi and for regression on DAC and EnCodec (Table 10). The training objective matters less than the placement. The code-assignment cross-entropy reaches 10.87 and regression with the cell hinge 11.45 at 5 dB, but the cross-entropy, whose minimum is not the clean latent, rewrites 8.4% of the test set's clean semantic codes, and without the lookup its output decodes to noise (96.39 points at 5 dB): although its 5 dB first-stage codes match regression's, its latent lies off the decoder's manifold. Placing the same network after the lookup worsens every condition ($+2.07$ points on clean speech, $+0.75$ at 5 dB), and a learned classifier in place of the lookup adds 11.06 points on clean speech, so in our controls the repair had to precede the lookup and the codebook had to stay the classifier. Because the vocabulary is unchanged, token consumers gain without retraining. The recognizer we trained on clean Mimi tokens reads the repaired tokens at 29.96 WER at 5 dB instead of 41.62, better than its twin retrained on degraded tokens reads the original ones (33.84), though both saw the same degraded training audio. Under six-talker babble, which no adapter saw, the gain shrinks to 3.01 ± 2.51 points (36.81 to 33.80), while the linear map and the after-lookup network worsen babble. The recipe transfers with smaller gains to two acoustic codecs whose adapters saw an eighth of the speech. At 5 dB it removes 2.60 ± 1.04 of DAC 3 kbps's 18.11 points, where re-fitting the codebook removed nothing detectable (0.49 ± 1.11), and 7.84 ± 1.22 of EnCodec 1.5 kbps's 33.88, as much as the re-fit (7.04) without changing the vocabulary. On twice the speech the DAC adapter removes 5.05 ± 0.96 points, and on four times 5.71 ± 0.98, so data explains part of the gap, then saturates. On DAC it does not carry over to babble ($+9.11$ points). SepFormer (Subakan et al., 2021), a separation front end also trained on WHAM!, does better in an unpaired run (end-to-end WER at 5 dB 29.15 to 17.25, against the adapter's 29.77 to 19.70; 8.84 against 3.01 points under babble; Appendix A.1), at 233 GFLOP per audio second against the adapter's 0.50. We claim gains over the frozen codec, not enhancement.

Finding 3. An identity-initialized adapter before the frozen lookup, trained by regression, halves Mimi's loss at 5 and 0 dB (21.32 to 11.25; 40.44 to 21.62) with clean WER and 99.8% of clean semantic codes unchanged, removes less on DAC 3 kbps and EnCodec 1.5 kbps (2.60 and 7.84 points; 5.05 on DAC with twice the speech; DAC $+9.11$ under unseen babble), and lets a clean-token recognizer read degraded speech better than one retrained on degraded tokens. After the lookup or replacing it, the same network raises the loss on clean speech and under babble.

7 Which Metrics See the Loss

Across the nine RVQ operating points of Table 1, on the fixed 500-utterance subset, we correlate candidate metrics with codec-added WER (Spearman ρ joint over 10, 5, and 0 dB; permutation tests over 20,000 relabelings; Table 3, Appendix A). Four cheap quantities rank codecs under degradation: STOI (Taal et al., 2011) ($+0.91$, p=0.001), WavLM distance (Chen et al., 2022) ($+0.95$, p<0.001), bitrate alone, known before decoding ($+0.77$, p=0.018), and log-mel distortion against the codec's input ($+0.81$, p=0.007). Against the clean reference, which the noise dominates, log-mel distortion shows no ranking we can detect ($+0.43$, p=0.210), nor do the token flip rate ($+0.17$), PESQ-WB (Rix et al., 2001; ITU-T, 2007) ($+0.62$, p=0.063), or UTMOS (Saeki et al., 2022) ($+0.21$). Nine points detect only $\rho \geq 0.83$ at 80% power. The flip rate saturates at 0.81–0.98 by 5 dB, and no output metric splits the loss between quantizer and decoder.

8 Conclusion and Limitations

Clean-speech transparency is not deployment transparency, but it predicts the deployed loss. Across ten operating points, two recent recognizers, and six codec families measured blind, the word error rate a codec adds at 5 dB is predicted, mostly within $2\times$, by a sublinear power law of its clean value. In quantized codecs a codec-dependent share of this loss arises at the codeword lookup, and an adapter before the frozen quantizer removes much of that share without changing the vocabulary. Evaluations should report added WER at the deployment SNR beside the clean value and read a codec's robustness as its distance below the law. Limits: the adapter is trained on one noise corpus, helps most on Mimi, and worsens unseen babble on DAC; our token readers are small recognizers at one operating point; the law's constant is corpus-specific; and the codebook controls cover acoustic RVQ codecs only.

REPRODUCIBILITY STATEMENT

All numbers in the paper come from per-utterance record files produced by our evaluation scripts, with the codec checkpoints, recognizers, and degradation recipes named in Section 3; noise draws are keyed to utterance identity and SNR, so every arm can be regenerated on the identical waveforms. Appendix A.3 gives the pre-lookup adapter's inputs, architecture, losses, training data, and schedule. We will release with the paper the code, including the adapter's data, training, and evaluation scripts and the scripts that compute every number from the records, the evaluation manifests, the per-utterance records, the adapter checkpoints, the time-stamped log of predictions and decision criteria, and listening examples; third-party model code (FocalCodec, the GLM-4-Voice tokenizer, StableToken) is used as released and not redistributed. Experiments ran on four NVIDIA A100-SXM4-80GB GPUs, mostly as inference with frozen checkpoints; our job log records 462 timed jobs totalling 472 job-hours, jobs often shared a GPU, and runs started outside the job queue are not in the log.

ETHICS STATEMENT

This work evaluates publicly released codecs and recognizers on public speech corpora (LibriSpeech, WHAM!, VoiceBank–DEMAND, and the AMI meeting corpus (Carletta et al., 2005), CC BY 4.0) under their licences; no new human data were collected and no personal data beyond those corpora are used. The findings bear on how speech systems behave on degraded input; we see no direct path to harm.

AI USE STATEMENT

We used large language model assistants solely to aid and polish the writing of this paper: they assisted the authors in editing the text for grammar, phrasing, clarity, and concision. Generative AI tools were not used to produce experimental data, simulate results, conduct research ideation, write codebase implementations, or derive mathematical formulations. The authors reviewed and verified all text, directed all aspects of this research, and take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Takanori Ashihara, Shota Horiguchi, Kohei Matsuura, Tsubasa Ochiai, and Marc Delcroix. Frontend token enhancement for token-based speech recognition. *arXiv preprint arXiv:2602.04217*, 2026.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*, 2020.

Lyonel Behringer, Anna Leschanowsky, Anjana Rajasekhar, Emily Kratsch, and Guillaume Fuchs. Assessing the impact of noise and speech enhancement on the intelligibility of speech codecs. In *Proc. Interspeech*, 2026.

Boson AI. Higgs Audio V2 tokenizer. `https://huggingface.co/bosonai/higgs-audio-v2-tokenizer`, 2025.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction (MLMI)*, volume 3869 of *Lecture Notes in Computer Science*, pp. 28–39, 2005.

Yunkee Chae and Kyogu Lee. Towards bitrate-efficient and noise-robust speech coding with variable bitrate RVQ. In *Proc. Interspeech*, 2025.

Yunkee Chae, Woosung Choi, Yuhta Takida, Junghyun Koo, Yukara Ikemiya, Zhi Zhong, Kin Wai Cheuk, Marco A. Martínez-Ramírez, Kyogu Lee, Wei-Hsiang Liao, and Yuki Mitsufuji. Variable bitrate residual vector quantization for audio coding. In *Proc. ICASSP*, 2025.

Heng-Jui Chang and James Glass. R-Spin: Efficient speaker and noise-invariant representation learning with acoustic pieces. In *Proc. NAACL*, 2024.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Topics Signal Process.*, 16(6):1505–1518, 2022.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Luca Della Libera, Francesco Paissan, Cem Subakan, and Mirco Ravanelli. FocalCodec: Low-bitrate speech coding via focal modulation networks. In *Proc. NeurIPS*, 2025.

Ruifan Deng, Yitian Gong, Qinghui Gao, Luozhijie Jin, Qinyuan Cheng, Zhaoye Fei, Shimin Li, and Xipeng Qiu. CodecBench: A comprehensive benchmark for acoustic and semantic evaluation. *arXiv preprint arXiv:2508.20660*, 2025.

Stephen Euler and Joachim Zinke. The influence of speech coding algorithms on automatic speech recognition. In *Proc. ICASSP*, pp. I–621–I–624, 1994.

Itai Gat, Felix Kreuk, Tu Anh Nguyen, Ann Lee, Jade Copet, Gabriel Synnaeve, Emmanuel Dupoux, and Yossi Adi. Augmentation invariant discrete representation for generative spoken language modeling. In *Proc. IWSLT*, 2023.

Bella Godiva, Yeonju Kim, and Yong Man Ro. Noise adaptive streaming audio-visual speech token enhancement for robust full-duplex spoken dialogue models. *arXiv preprint arXiv:2609.08390*, 2026.

Yitian Gong, Kuangwei Chen, Zhaoye Fei, Xiaogui Yang, Ke Chen, Yang Wang, Kexin Huang, Mingshu Chen, et al. MOSS-Audio-Tokenizer: Scaling audio tokenizers for future audio foundation models. *arXiv preprint arXiv:2602.10934*, 2026.

ITU-T. Recommendation P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs, 2007.

Shengpeng Ji et al. WavTokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2025. arXiv:2408.16532.

Sofiene Kammoun, Xavier Alameda-Pineda, and Simon Leglaive. Modeling strategies for speech enhancement in the latent space of a neural audio codec. In *Proc. IEEE ICASSP*, 2026.

Hong Kook Kim, Richard V. Cox, and Richard C. Rose. Performance improvement of a bitstream-based front-end for wireless speech recognition in adverse environments. *IEEE Trans. Speech Audio Process.*, 10(8):591–604, 2002.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Proc. NeurIPS*, 2023.

Haoyang Li, Jia Qi Yip, Tianyu Fan, and Eng Siong Chng. Speech enhancement using continuous embeddings of neural audio codec. In *Proc. IEEE ICASSP*, 2025.

Shaokai Li, Weiping Tu, and Yuhong Yang. Noisy environment adaptation of neural speech codec via focal mask and noise feature separation. In *Proc. Interspeech*, 2026.

Alexander H. Liu et al. Voxtral. *arXiv preprint arXiv:2507.13264*, 2025a.

Fei Liu, Yang Ai, Ye-Xin Lu, Rui-Chen Zheng, Hui-Peng Du, and Zhen-Hua Ling. Universal discrete-domain speech enhancement. *arXiv preprint arXiv:2510.09974*, 2025b.

Tom D. Lookabaugh and Robert M. Gray. High-resolution quantization theory and the vector quantizer advantage. *IEEE Transactions on Information Theory*, 35(5):1020–1033, 1989.

Ye-Xin Lu, Hui-Peng Du, Fei Liu, Yang Ai, and Zhen-Hua Ling. Improving noise robustness of LLM-based zero-shot TTS via discrete acoustic token denoising. In *Proc. Interspeech*, 2025.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *Proc. ICLR*, 2024.

Shoval Messica and Yossi Adi. NAST: Noise aware speech tokenization for speech language models. In *Proc. Interspeech*, 2024.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *Proc. ICML*, 2021.

Pooneh Mousavi et al. DASB – discrete audio and speech benchmark. *arXiv:2406.14294*, 2024.

Pooneh Mousavi et al. Discrete audio tokens: More than a survey! *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=eqNchtvc6v.

Zhikang Niu, Sanyuan Chen, Long Zhou, Ziyang Ma, Xie Chen, and Shujie Liu. NDVQ: Robust neural audio codec with normal distribution-based vector quantization. *arXiv preprint arXiv:2409.12717*, 2024.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *Proc. ICASSP*, 2015.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, and Furu Wei. VibeVoice technical report. *arXiv preprint arXiv:2508.19205*, 2025.

Jakob Poncelet, Yujun Wang, and Hugo Van hamme. Efficient extraction of noise-robust discrete units from self-supervised speech models. In *Proc. IEEE SLT*, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. Fast conformer with linearly scalable attention for efficient speech recognition. In *Proc. IEEE ASRU*, 2023.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (PESQ). In *Proc. ICASSP*, 2001.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Proc. Interspeech*, 2022.

Hubert Siuzdak, Florian Grötschla, and Luca A. Lanzendörfer. SNAC: Multi-scale neural audio codec. *arXiv preprint arXiv:2410.14411*, 2024.

Yuhan Song, Linhao Zhang, Chuhan Wu, Aiwei Liu, Wei Jia, Houfeng Wang, and Xiao Zhou. StableToken: A noise-robust semantic speech tokenizer for resilient SpeechLLMs. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2026. arXiv:2509.22220.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *Proc. ICASSP*, 2021.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(7):2125–2136, 2011.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Proc. NeurIPS*, 2020.

Wei-Cheng Tseng and David Harwath. Probing the robustness properties of neural speech codecs. In *Proc. Interspeech*, 2025.

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In *Proc. SSW*, 2016.

Jean-Marc Valin, Koen Vos, and Timothy B. Terriberry. Definition of the Opus audio codec. IETF RFC 6716, 2012.

Shikhar Vashishth, Harman Singh, Shikhar Bharadwaj, Sriram Ganapathy, Chulayuth Asawaroengchai, Kartik Audhkhasi, Andrew Rosenberg, Ankur Bapna, et al. STAB: Speech tokenizer assessment benchmark. *arXiv preprint arXiv:2409.02384*, 2024.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. WHAM!: Extending speech separation to noisy environments. In *Proc. Interspeech*, 2019.

Jack K. Wolf and Jacob Ziv. Transmission of noisy information to a noisy receiver with minimum distortion. *IEEE Trans. Inf. Theory*, 16(4):406–411, 1970.

Haibin Wu et al. Codec-SUPERB: An in-depth analysis of sound codec models. In *Findings of ACL*, 2024.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, et al. Llasa: Scaling train-time and inference-time compute for Llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025a.

Zhen Ye et al. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proc. AAAI Conf. Artif. Intell.*, 2025b.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 30: 495–507, 2022.

Neil Zeghidour, Eugene Kharitonov, Manu Orsini, Václav Volhejn, Gabriel de Marmiesse, Edouard Grave, Patrick Pérez, Laurent Mazaré, and Alexandre Défossez. Streaming sequence-to-sequence learning with delayed streams modeling. *arXiv preprint arXiv:2509.08753*, 2025.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech language models. In *Proc. ICLR*, 2024.

Haixin Zhao and Nilesh Madhu. Towards robust generative speech enhancement using vector quantisation-based neural audio codec. In *Proc. Interspeech*, 2026.

Rui-Chen Zheng, Yang Ai, Hui-Peng Du, and Li-Rong Dai. Enhancing noise robustness for neural speech codecs through resource-efficient progressive quantization perturbation simulation. In *Proc. ICASSP*, pp. 15392–15396, 2026.

A Additional results

Protocol details. Rate ladders use the released RVQ depths: DAC-16k at 2–12 codebooks (1.0–6.0 kbps), EnCodec-24k at 1–16 (0.75–12 kbps), Mimi at 1–32 of its 32 codebooks (0.14–4.4 kbps), SpeechTokenizer at 1–8 layers (0.5–4.0 kbps), and SNAC at 0.98 kbps, printed as 1.0. On clean input the bypassed path adds 0.06 (DAC 6 kbps), 0.59 (DAC 3 kbps), 0.07 (EnCodec), 0.28 (Mimi), 0.89 (SpeechTokenizer), 2.02 (SNAC), 4.10 (WavTokenizer), and 3.85 (DAC 1.5 kbps, whose three-layer sum carries less than the full latent) WER points over the input. The token flip rate is averaged over layers. The WavLM distance is a layer-8 cosine distance between the codec output and the clean signal; log-mel distortion is the mean absolute log-mel distance of the codec output to the

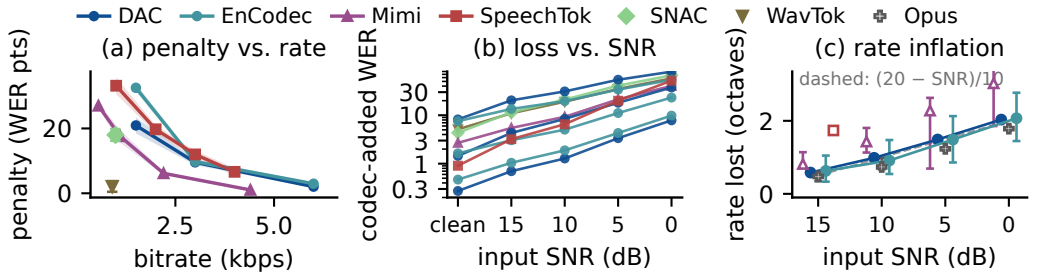


Figure 2: **Degradation inflates the rate a codec needs.** (a) The quantization penalty at 5 dB against bitrate for all six families (utterance-clustered 95% CI bands, full 2,620-utterance test set). (b) Codec-added WER of all ten operating points against input SNR (log scale): every curve rises and the ordering is largely kept. (c) Rate inflation: octaves of bitrate an operating point loses at a given SNR, read against its own family's clean-speech rate ladder (median and range over rates; markers offset for legibility; hollow markers are families whose clean ladder brackets only some operating points, plotted where it does; SpeechTokenizer below 15 dB lies beyond its own ladder and is not drawn). The dashed line is $(20 - \mathrm{SNR})/10$ octaves, one octave per 10 dB.

clean signal (Section 7 also measures it against the codec's input). The SNR is the ratio of whole-utterance mean powers, set at 16 kHz before any codec resamples its input; the noise comes from the WHAM! test split, with the draw keyed to utterance and SNR so that every arm hears the same waveform. The enhancer of Section 6 is the released SepFormer checkpoint trained for enhancement on WHAM! at 16 kHz, so it faces noise from its own training distribution, which favours it.

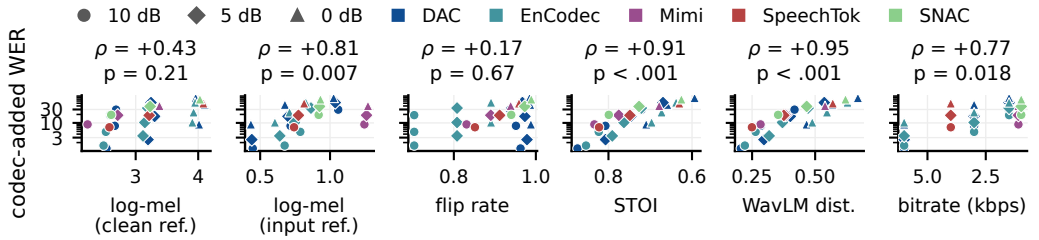


Figure 3: **Four cheap metrics rank codecs under degradation; log-mel distortion against the clean reference and token flip rate do not.** One candidate metric per panel against codec-added WER at the nine RVQ operating points of Table 1 at 10, 5, and 0 dB (27 point–SNR pairs, 500-utterance subset; colour = family, marker = SNR); titles give the joint Spearman ρ and its permutation p (Table 3, Appendix A); STOI and bitrate axes are reversed so that more damage reads left to right.

Are the ranking and non-ranking metrics different? The asymmetry is borderline at nine points: a paired permutation test on the difference of joint ρ gives p=0.04–0.07 for STOI and WavLM distance against log-mel distortion (vs. clean) or token flip rate and p=0.11–0.27 for bitrate and input-referenced log-mel.

Replication. We reproduce the effect on a second corpus, under two further recognizers, and under two further degradation types. On the noisy recordings of VoiceBank–DEMAND (Valentini-Botinhao et al., 2016), codec-added WER ranks the nine RVQ operating points as Table 1 does (Spearman +0.98). The quantization penalty there is 0.45–0.89 of its 5 dB WHAM! value at eight of the nine points (Table 5). It is indistinguishable from zero only for SNAC (0.96 ± 3.86). Its nominally clean recordings already incur a penalty, and scoring them with wav2vec2 reproduces it (4.26 ± 0.96 for DAC 3.0; 9.30 ± 1.27 for Mimi). Under two further recognizers on the full test-clean at 5 dB, the penalty is positive with a CI excluding zero at all nine RVQ operating points

Table 3: STOI, WavLM distance, bitrate, and log-mel distortion against the codec's input rank codecs under degradation; for PESQ-WB, UTMOS, log-mel distortion against the clean reference, and token flip rate we detect no ranking, at a sample size that can only detect $|\rho| \geq 0.83$. Spearman ρ against codec-added WER, joint over the nine RVQ operating points of Table 1 (WavTokenizer is not in the metric subset) at 10, 5, and 0 dB, 500-utterance subset; p from 20,000 sampled codec relabelings. Power is for the observed ρ at two-sided α=0.05.

Metric	Domain	Joint ρ	p	Power at observed ρ
WavLM distance (layer 8)	representation	$+0.95$	< 0.001	—
STOI	intelligibility	$+0.91$	0.001	—
Bitrate	none (known a priori)	$+0.77$	0.018	—
PESQ-WB	perceptual quality	$+0.62$	0.063	0.41
UTMOS	naturalness (MOS)	$+0.21$	0.193	0.08
Log-mel distortion (vs. clean)	spectral	$+0.43$	0.210	0.19
Log-mel distortion (vs. input)	spectral	$+0.81$	0.007	—
Token flip rate	token	$+0.17$	0.668	0.07

for Whisper-large-v3 and at eight of them for wav2vec2-base; the ninth, SpeechTokenizer, reverses (Section 5). Whisper-large-v3 gives 0.43–1.08 of Whisper-small's values (1.04–23.14 points), and both its rate ladders are monotone. Wav2vec2-base, whose own 5 dB WER is 20.6, gives 3.55–32.32 points. WavTokenizer, added after these runs, gives a 5 dB penalty of 3.66 ± 0.65 under Whisper-large-v3 (codec-added 26.90) and 1.14 ± 0.45 under wav2vec2-base (codec-added 30.34). Two 200-utterance checks with Whisper-small, WHAM! cross-validation noise on test-clean and test noise on dev-clean speakers, reproduce the DAC ladder's ordering at 10 and 5 dB but not at 0 dB. Resampling the 40 test speakers instead of utterances widens the penalty intervals of Table 1 by at most 2.2× and lowers the smallest amplification bound from 4.2 to 4.0; no interval changes sign. The Opus reference uses libopus 1.3.1 in VOIP mode at constant bitrate with 20 ms frames on the same 16 kHz inputs; libopus chooses its internal bandwidth, narrowband at the lowest rates. Six-talker babble makes DAC add 4.68 ± 1.53 points at 6 kbps and 31.92 ± 2.30 at 3 kbps, and reverberation at RT60 400 ms costs 2.88–53.50 penalty points on the full test-clean at all nine RVQ points (DAC 6.0/3.0/1.5: 2.88, 11.22, 17.79; EnCodec 6.0/3.0/1.5: 6.34, 18.58, 53.50; Mimi 30.64; SpeechTokenizer 16.77; SNAC 19.10), every CI excluding zero.

Substitution distance at matched WER. Normalized character distance between a substituted reference word and its replacement rises steeply with the WER of the hypothesis it comes from, for every error source: the bypassed path's own substitutions have median distance 0.50 below 10% WER and 0.86 above 100%. We therefore compare the lookup-added substitutions with the bypassed path's own at matched hypothesis WER (direct standardisation over WER bins of 10 points up to 60%, then 60–80, 80–100, and above 100%, utterance bootstrap over the nine RVQ points). The lookup's are farther by $+0.065$ $[+0.058, +0.073]$ in mean distance, 30% of the unmatched contrast with the recognizer's own substitutions ($+0.219$), and by $+0.034$ $[+0.027, +0.043]$ against bypass-added substitutions built by the same set difference; the remaining gap to the uncoded input's substitutions ($+0.155$ $[+0.133, +0.177]$) is mostly shared by the bypassed path ($+0.090$ $[+0.077, +0.104]$). Distant substitutions thus mark codec-path errors in general, not the lookup in particular.

A.1 An enhancer in front of the frozen codec

An enhancer in front of the codec, facing noise from its own training distribution, recovers most of the loss at low rates and nets nothing at 6 kbps: at 5 dB a SepFormer enhancer (Subakan et al., 2021) costs 1.76 WER points on the input and cuts codec-added WER by 58.1–76.3%, netting $+11.51$ to $+13.17$ points at DAC 3 kbps, Mimi, and SpeechTokenizer, while at DAC 6 kbps its own 1.76 points offset the 1.93 it removes. Under six-talker babble, outside its training noise, the enhancer raises the uncoded input's WER from 18.88 to 28.77, yet still cuts codec-added WER by 58.7% at DAC 3 kbps and 49.7% at Mimi, netting $+9.01$ $[+6.45, +11.58]$ and $+8.84$ $[+5.84, +11.83]$ points.

Table 4: Substitutions are 65–75% of the edits the quantizer adds at every RVQ point, a share like that of the recognizer's own errors on the uncoded 5 dB input (73%). Quantization penalty clean→5 dB (utterance-clustered 95% CIs) and substitution/deletion/insertion shares of the lookup-added edits at 5 dB, nine RVQ operating points of Table 1, full test-clean (2,620 utterances, Whisper-small); substitution-share CI lower bounds ≥60%, half-widths ≤7 points. The shares count edits over the whole test set (corpus level). For the deterministic codecs the per-utterance results are those of Table 1, so the corpus-level penalty totals (Table 14) differ from the first column only in aggregation; SNAC's shares come from a second run of its stochastic decoder.

Codec (kbps)	Penalty clean→5 dB	Sub./Del./Ins. (%)	Sub./Del./Ins., wav2vec2 (%)
SNAC 1.0	2.33±0.39→18.00±1.80	65/11/23	—
Mimi 1.1	2.41±0.33→17.57±1.95	71/14/15	—
DAC 1.5	4.35±0.45→20.92±2.46	68/10/22	60/40/ 1
EnCodec 1.5	7.48±0.52→32.54±1.29	73/12/14	75/22/ 3
DAC 3.0	0.86±0.22→9.55±0.88	69/12/18	73/24/ 3
EnCodec 3.0	1.57±0.28→9.77±0.61	75/11/14	76/21/ 3
SpeechTok. 4.0	0.01±0.26→6.62±1.19	75/10/14	—
DAC 6.0	0.21±0.17→2.04±0.37	75/16/ 9	79/18/ 3
EnCodec 6.0	0.40±0.20→2.95±0.39	74/15/12	75/22/ 3

Table 5: A second corpus, clean VCTK speech mixed with recorded DEMAND noise at 2.5–17.5 dB, reproduces the ordering of the operating points and the quantization penalty; at its milder SNRs the amplification is 2.2–11.7×. VoiceBank–DEMAND, all 824 test utterances; references are whisper-large-v3 transcripts of the clean signals, so these values are disagreements with that transcriber rather than with human labels. Four of the nine operating points are listed; the text reports the rest.

Codec (kbps)	Clean	Noisy	Pen. cl.	Pen. noisy	Ampl.
DAC 6.0	+0.19	+2.27	0.00	1.48±0.76	11.7×
DAC 3.0	+3.57	+11.77	2.01	6.96±1.46	3.3×
EnCodec 3.0	+3.75	+8.26	3.96	7.17±1.42	2.2×
Mimi 1.1	+6.32	+17.88	6.17	15.62±1.87	2.8×

A.2 Codebook re-fit and rate-matched scalar quantization

Codebook re-fit: design and results. Encoder, decoder, and codebook size and dimension are unchanged, so the bitrate is identical; only the stored vectors are re-fitted, on the latents of 2,000 train-clean-360 utterances mixed with WHAM! training noise at clean, 15, 10, 5, and 0 dB, by Lloyd iterations warm-started from the released codebooks, layer by layer on the residual left by the re-fitted layers above, under each codec's own assignment rule. Our quantization path reproduces each codec's official codes bit-exactly. The re-fit lowers the quantization error by 4–7% for DAC (on held-out utterances) and, on its training latents, by 3–9% for EnCodec (first eight codebooks) and 1–19% for SpeechTokenizer. With the released codebooks the evaluation reproduces Tables 1 and 4 exactly for EnCodec and SpeechTokenizer and within 0.6 WER points for DAC. Table 6 gives the paired differences: every DAC interval contains zero; EnCodec gains at 1.5 and 3.0 kbps and keeps most of its 1.5 kbps gain under babble; SpeechTokenizer gains nothing significant at 5 dB and loses 1.42 clean points at 1 kbps.

How well the lookup fits under noise. Noise changes how well the lookup fits only modestly: on 500 utterances, matching the clean-speech latent error at 5 dB takes EnCodec one or two more codebooks (3→4, 6→8), and the first-layer gap between the nearest and the second codeword widens rather than shrinks, while the penalty grows 4.4–7.4× at the EnCodec points. The first-layer indices that change under noise belong to frames that sat near a boundary on clean speech (96% of the lowest-margin fifth of EnCodec frames change, 54% of the highest).

Rate-matched scalar quantizer: design. Replacing the codebook lookup in place with the strongest generic quantizer we could build without a learned codebook, a rate-matched scalar quantizer. A Karhunen–Loève transform estimated on clean speech rotates each layer's projected residual

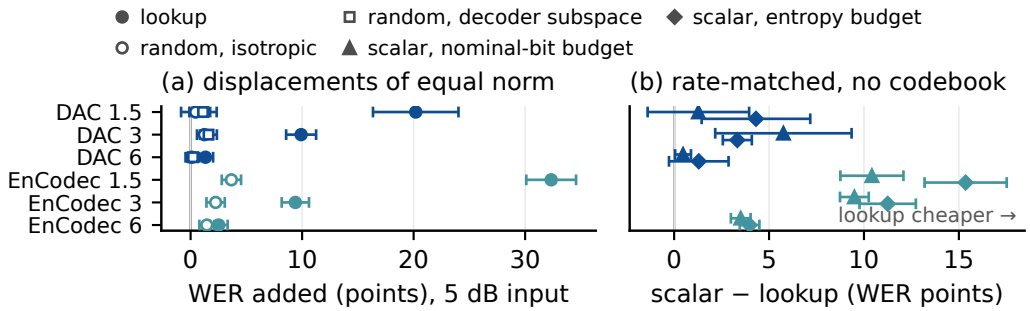


Figure 4: **The penalty is structured, and a rate-matched scalar quantizer loses more words.** (a) At 5 dB a random displacement of the latent with the quantizer's per-frame norm, isotropic or confined to the decoder's subspace, adds almost no words; the quantizer's adds up to 32 WER points (500-utterance subset, paired 95% CIs), 65–75% of its added edits being substitutions (Table 4, Appendix A). (b) Replacing the lookup in place by a rate-matched scalar quantizer, under either budget (the ladder's nominal kbps or the entropy of its index stream), costs more words at every DAC and EnCodec point (full test-clean, paired; Table 7).

Table 6: Re-fitting the codebook on degraded speech at identical bitrate helps EnCodec at low rate under synthetic noise, including a noise type it never saw, but not on real far-field recordings, and helps neither DAC nor SpeechTokenizer. Paired difference in WER points, re-fitted minus released codebook, full test-clean (utterance 95% CI); babble is six-talker babble, absent from the re-fit data; AMI far-field is the distant-microphone recording of Table 16's 2,000 meeting segments.

Codec (kbps)	Clean	5 dB WHAM!	5 dB babble	AMI far-field
DAC 1.5	-0.11 ± 0.47	$+1.24 \pm 3.31$	—	—
DAC 3.0	$+0.07 \pm 0.25$	-0.49 ± 1.11	—	—
EnCodec 1.5	-1.11 ± 1.55	-7.04 ± 1.33	-6.66 ± 5.03	$+0.68 \pm 8.75$
EnCodec 3.0	-0.41 ± 0.26	-1.29 ± 0.59	-0.49 ± 1.15	$+0.60 \pm 5.19$
EnCodec 6.0	$+0.06 \pm 0.21$	-0.40 ± 0.44	—	—
SpeechTok. 1.0	$+1.42 \pm 0.38$	-2.37 ± 2.86	—	—
SpeechTok. 2.0	$+0.26 \pm 0.30$	-1.96 ± 1.97	—	—
SpeechTok. 4.0	$+0.11 \pm 0.22$	-0.52 ± 1.16	—	—

onto its principal axes, and a uniform scalar quantizer with one step size shared across all dimensions quantizes every coordinate. The step size is bisected on held-out speech at the same SNR until the arm meets its bit budget. Encoder, projections, decoder, and recognizer are unchanged, and the comparison is paired over the full 2,620 utterances at 5 dB. We run two budgets. The nominal-bit budget is the kbps the RVQ ladder advertises. The entropy budget is the empirical entropy of the RVQ index stream, which lies 1–14% below the nominal rate at 5 dB because the released codebooks are not fully used. The scalar arm is charged its ideal entropy-coded rate and RVQ its raw index rate; both conventions favour the scalar arm. At each layer the quantization is carried into the residual of the next layer exactly as the codebook's would be; with the step size at zero the arm reproduces the bypassed path, and with the codebook restored it reproduces the codec's own output bit-exactly.

Rate-matched scalar quantization at 10 and 0 dB. On DAC, with the step size calibrated at the matching SNR, the scalar arm again loses more words than the codebook: at 10 dB the paired differences are $+3.55$ $[+1.40, +5.70]$, $+1.49$ $[+0.95, +2.02]$, and $+0.19$ $[-0.07, +0.46]$ at 1.5, 3.0, and 6.0 kbps under the nominal budget ($+4.60$, $+2.34$, $+0.20$ under the entropy budget); at 0 dB they are $+5.51$ $[+2.09, +8.94]$ and $+1.65$ $[+0.45, +2.86]$ at 3.0 and 6.0 kbps ($+4.26$, $+1.96$), while at 1.5 kbps both arms exceed 80% WER and the cell is excluded. The sign is the same at all ten reported cells; the deficit grows with degradation at 6.0 kbps and not monotonically at 3.0 kbps.

Table 7: At matched bitrate the rate-matched scalar quantizer loses more words than the learned codebook at all six operating points, under both budgets. WER (%, Whisper-small) at 5 dB on the full test-clean (2,620 utterances); SQ−RVQ is the paired difference with an utterance-clustered 95% CI; the distortion ratio is the scalar arm's decoder-input error relative to RVQ's. Rates are the scalar arm's measured entropy-coded rates; at EnCodec 3.0 and 6.0 neither arm stayed within its budget, the nominal-budget arm exceeding it by 13% and 50%, and both still lose. Calibration on dev-clean; the first-round calibration is in Appendix A.2.

		nominal-bit budget				entropy budget			
Codec (kbps)	RVQ WER	rate	SQ WER	SQ−RVQ [95% CI]	dist.	rate	SQ WER	SQ−RVQ [95% CI]	dist.
DAC 1.5	62.61	1.48	63.88	+1.28 [-1.40, +3.95]	1.18	1.45	66.92	+4.31 [+1.45, +7.17]	1.20
DAC 3.0	26.56	3.00	32.31	+5.76 [+2.16, +9.35]	1.40	2.97	29.89	+3.33 [+2.56, +4.09]	1.41
DAC 6.0	11.81	6.09	12.28	+0.47 [+0.05, +0.89]	1.57	6.07	13.11	+1.30 [-0.27, +2.87]	1.58
EnCodec 1.5	42.33	1.40	52.75	+10.42 [+8.76, +12.08]	1.23	1.27	57.69	+15.36 [+13.19, +17.52]	1.27
EnCodec 3.0	19.57	3.38	29.06	+9.49 [+8.74, +10.25]	1.46	3.22	30.82	+11.25 [+9.77, +12.73]	1.50
EnCodec 6.0	12.75	9.03	16.26	+3.51 [+2.99, +4.03]	1.84	8.85	16.72	+3.97 [+3.46, +4.49]	1.88

Rate-matched scalar quantization with a dead zone. Under an entropy constraint a step size shared across the KLT axes is the optimal high-rate allocation, so no reverse water-filling is needed; at low rate the standard improvement is a dead zone around zero. Table 8 repeats the DAC comparison with the dead-zone width chosen per layer, from $\{0, 0.15, 0.30, 0.45\}$ steps, by the lowest calibration distortion at the budget. The dead zone lowers the scalar arm's WER at 3.0 and 6.0 kbps and raises it at 1.5 kbps, brings its distortion ratio to 0.99–1.04, and the codebook is still cheaper at all three points ($+3.66$ $[+0.34, +6.99]$, $+1.77$ $[+0.83, +2.72]$, $+0.37$ $[-0.05, +0.79]$).

Table 8: Dead-zone entropy-constrained scalar quantizer against the released codebook lookup on DAC, 5 dB, nominal-bit budget, full test-clean, paired per utterance. Columns as in Table 7.

Codec (kbps)	RVQ WER	rate	SQ WER	SQ−RVQ [95% CI]	dist.
DAC 1.5	62.61	1.44	66.27	+3.66 [+0.34, +6.99]	0.99
DAC 3.0	26.56	2.91	28.33	+1.77 [+0.83, +2.72]	1.01
DAC 6.0	11.81	5.91	12.18	+0.37 [-0.05, +0.79]	1.04

Rate-matched scalar quantization on SpeechTokenizer. SpeechTokenizer quantizes a 1024-dimensional residual with 1024-entry codebooks, 0.01 bit per dimension, and there the scalar arm collapses: 92.32 and 95.91% WER at 5 dB under the nominal and entropy budgets against 28.37 for the codebook. At that rate per dimension no scalar quantizer is usable, so the comparison is uninformative about the codebook and the attribution of Section 5 stays scoped to the DAC and EnCodec points.

Rate-matched scalar quantization, first calibration round. Table 9 repeats the comparison of Section 5 with the scalar arm's step size calibrated on 300 train-clean-360 utterances instead of 500 dev-clean utterances. Every sign is the same as in the main table; magnitudes move by up to 5 WER points.

A.3 The pre-lookup adapter

Where it sits. For Mimi the adapter reads and writes $\mathbf{u} = [\mathrm{in_proj}_{\mathrm{sem}}(\mathbf{z}_e);\ \mathrm{in_proj}_{\mathrm{ac}}(\mathbf{z}_e)]$, the 512-dimensional (256 + 256) vector Mimi's semantic and acoustic RVQ groups quantize at 12.5 Hz; for DAC 16 kHz at 3 kbps it reads and writes the 1024-dimensional encoder output at 50 Hz, which DAC's six stages project to 8 dimensions, L2-normalize, and look up; for EnCodec 24 kHz at 1.5 kbps, the 128-dimensional encoder output at 75 Hz, which its two stages look up by Euclidean distance. Our re-implementation of each lookup reproduces the released encoder's codes bit for bit on the evaluation waveforms, so with the adapter's output layer at zero the codec is unchanged: the untrained arm reproduces Table 1's rows for all three codecs utterance by utterance. The transmitted codes go through the released decoder, and the recognizer is Whisper-small as in Table 1.

Table 9: Rate-matched scalar quantization against the released codebook lookup, first calibration round (step size from train-clean-360). Columns as in the main text; 5 dB input, full test-clean, paired per utterance.

		budget = nominal bits				budget = code entropy			
Codec (kbps)	RVQ WER	rate	SQ WER	SQ−RVQ [95% CI]	dist.	rate	SQ WER	SQ−RVQ [95% CI]	dist.
DAC 1.5	62.61	1.49	67.18	+4.57 [+0.98, +8.16]	1.17	1.46	70.81	+8.21 [+2.29, +14.12]	1.19
DAC 3.0	26.56	3.02	29.41	+2.85 [+1.87, +3.84]	1.39	2.99	30.26	+3.70 [+1.67, +5.74]	1.41
DAC 6.0	11.81	6.12	12.38	+0.57 [+0.13, +1.01]	1.56	6.10	12.41	+0.59 [+0.19, +0.99]	1.58
EnCodec 1.5	42.33	1.56	47.60	+5.27 [+3.66, +6.88]	1.18	1.40	53.35	+11.02 [+9.04, +12.99]	1.22
EnCodec 3.0	19.57	3.62	26.18	+6.62 [+5.91, +7.32]	1.39	3.45	28.07	+8.50 [+7.42, +9.59]	1.43
EnCodec 6.0	12.75	9.50	16.29	+3.55 [+2.89, +4.21]	1.76	9.31	16.20	+3.46 [+2.96, +3.96]	1.80

Network and training. Input LayerNorm and projection to width 512, a causal depthwise convolution of width 3, six pre-norm causal Transformer layers (8 heads, feed-forward 2048, dropout 0.1), a LayerNorm, and a zero-initialized projection back to the input width, added to the input: 19.4M parameters for Mimi, 20.0M for DAC. AdamW (learning rate $3 \cdot 10^{-4}$, weight decay 0.01), 1,000 warm-up steps, cosine decay, 20,000 steps, gradient clipping at 1, bf16. Batches hold 64 crops of at most 250 frames for Mimi and 32 of at most 500 for DAC and EnCodec, the same number of frames per step. Mimi's arms train on all of train-clean-360 (103,014 utterances after holding out 1,000), DAC's and EnCodec's on 13,002 of its utterances, every fourth of the first 52,008 (12,002 after the hold-out), one DAC run on these plus the next 13,002 at the same stride (25,004 after the hold-out), one on 52,008 utterances (51,008 after the hold-out), and one EnCodec run on 26,004 (25,004 after the hold-out); inputs are the codec's pre-lookup vectors of the training-noise mixtures of Appendix A.6 (WHAM! training recordings, 0–20 dB, 20% left clean), targets the clean utterance's vectors and codes. Nothing is selected on the evaluation set: every arm is scored after its last step.

Losses. With t_k the clean code at stage k and $\mathbf{r}_k$ the adapter output's residual at stage k computed with the clean codes of the earlier stages (teacher forcing; Mimi's semantic stage reads the first 256 dimensions, its acoustic stages the rest), $d_{k,j}$ is the squared distance from $\mathbf{r}_k$ (DAC: its normalized 8-dimensional projection) to codeword j in the lookup's own space, and τ_k is a quarter of the median squared distance from a codeword to its nearest neighbour. The code-assignment loss is the mean over stages of the cross-entropy of $\mathrm{softmax}(-d_{k,\cdot}/\tau_k)$ against t_k. Regression is the squared error to the clean vector divided by the per-dimension variance of clean vectors. The cell hinge adds $\frac{1}{K}\sum_k \max(0,\ d_{k,t_k} - \min_{j \neq t_k} d_{k,j})/\tau_k$ with weight 1, fixed before training. The learned-classifier control replaces the output projection by eight linear heads over the 2,048 codes (27.8M parameters) trained with cross-entropy; the after-lookup control applies the network to the dequantized degraded codes and re-quantizes its output; the linear control is a ridge regression from degraded to clean vectors, $\lambda = 10^{-3}$ times the mean diagonal of the Gram matrix, fitted on all training frames.

Two paths. Decoding the adapter's output through the bypassed path of Eq. 1 separates a repair of the lookup from generic denoising. At 5 dB the frozen codec's bypassed path adds 3.75 points and its lookup 17.57 more; with the regression adapter the bypassed path adds 4.82 and the lookup only 6.43, so the whole gain is the lookup's (Table 10), and two more seeds agree (penalty 5.65 and 5.35, bypassed path 5.00 and 5.03). With clean waveforms' codes as an oracle, the quantized path would score -0.83 at 5 dB (7.62 against the input's 8.45), so regression recovers 45% of the gain available at the lookup's input. On DAC 3 kbps the adapter cuts the lookup's penalty about as much (9.55 to 3.71 points; 2.50 on twice and 2.58 on four times the speech) but worsens the latent its decoder reads (the bypassed path's loss rises from 8.56 to 11.80; 10.56 and 9.82), which is why its net gain is smaller; on EnCodec 1.5 kbps the penalty falls from 32.54 to 19.90 (20.04 for a second seed) while the bypassed path rises from 1.35 to 6.14 (6.81); on twice the speech, 19.52 and 5.54. In all three codecs the net gain is the penalty the adapter removes minus what the continuous path loses: trained by squared error, its output approximates a conditional mean of the clean latent, which the lookup snaps to the clean code more often (Table 10) but which the decoder, reading it unquantized, does not decode better.

Unseen degradations. Trained on WHAM! noise only, the adapter helps a little under six-talker babble on Mimi (-3.01 ± 2.51, -2.90 ± 2.60, and -1.81 ± 3.11 points over three seeds) and not detectably under reverberation (RT60 400 ms: the frozen codec adds 38.67 points, regression 38.18, -0.49 ± 1.48); on DAC it makes babble worse for every seed ($+9.11 \pm 4.42$, $+7.18 \pm 3.62$, $+6.99 \pm 3.15$; $+9.13 \pm 3.93$ on twice and $+4.92 \pm 3.03$ and $+4.17 \pm 2.61$ on four times the speech), and on EnCodec the babble change is within noise for every seed (-2.66 ± 5.25, $+1.50 \pm 5.62$, -1.69 ± 5.42; -1.13 ± 5.81 on twice the speech).

Why the cross-entropy rewrites clean codes. Its minimum is not the clean latent: a point inside its correct cell still lowers the loss by moving deeper, so clean inputs are moved too, and through the teacher-forced residuals the later acoustic stages drift. On held-out training utterances the cross-entropy arm keeps 91.8% of clean semantic codes and 13.5% of the eighth stage's, regression 99.8% and 88.9%. The hinge vanishes at the clean latent by construction.

Table 10: The pre-lookup adapter, three training seeds of each loss on Mimi 1.1 kbps and of regression on DAC 3.0 kbps and EnCodec 1.5 kbps; DAC regression runs on twice and four times the training speech (25,004 and 51,008 utterances instead of 12,002) and an EnCodec run on twice (25,004). Codec-added WER (Whisper-small) on clean input, at 5 and 0 dB (WHAM! test noise), and under unseen six-talker babble at 5 dB; the share of transmitted codes at 5 dB equal to the codes the frozen codec assigns to the clean waveform, for the semantic stream and the first acoustic stage (DAC and EnCodec: first and second stage); the WER at 5 dB of the token reader trained on degraded tokens (33.84 on the original tokens), not retrained; and the two-path measurement of Section 3 applied to the adapter's output at 5 dB: codec-added WER of the bypassed path (the adapted latent decoded without the lookup) and the quantization penalty, quantized minus bypassed.

Arm	Seed	Clean	5 dB	0 dB	Babble	Sem. (%)	Ac. 1 (%)	Reader$_N$	Bypass	Penalty
Frozen codec	–	2.69	21.32	40.44	36.81	42.1	15.3	33.84	3.75	17.57
Latent regression (MSE)	0	2.75	11.25	21.62	33.80	55.8	44.3	29.98	4.82	6.43
Latent regression (MSE)	1	2.60	10.66	21.19	33.91	55.8	44.2	–	5.00	5.65
Latent regression (MSE)	2	2.81	10.38	20.89	35.00	55.8	44.2	–	5.03	5.35
Code-assignment cross-entropy	0	2.91	10.87	22.01	34.09	56.2	44.4	31.01	96.39	-85.52
Code-assignment cross-entropy	1	2.78	10.51	22.09	33.06	56.2	44.4	–	–	–
Code-assignment cross-entropy	2	2.88	11.02	20.66	35.21	56.3	44.3	–	–	–
Regression + cell hinge	0	2.64	11.45	21.40	35.22	54.7	44.7	31.11	4.15	7.30
Regression + cell hinge	1	2.67	11.11	23.17	32.13	54.6	44.6	–	–	–
Regression + cell hinge	2	2.55	11.21	22.27	31.35	54.6	44.7	–	–	–
Linear map	–	2.89	16.01	30.83	39.10	51.9	22.8	32.69	5.85	10.16
After the lookup	–	4.76	22.07	41.62	58.62	49.8	34.2	40.18	–	–
Learned classifier	–	13.75	27.09	39.39	68.55	56.4	45.2	33.08	–	–
DAC 3.0 kbps										
Frozen codec	–	1.45	18.11	37.42	31.92	6.1	2.1	–	8.56	9.55
Latent regression (MSE)	0	1.46	15.51	32.20	41.03	15.9	4.1	–	11.80	3.71
Latent regression (MSE)	1	1.46	16.07	33.49	39.11	15.9	4.1	–	–	–
Latent regression (MSE)	2	1.42	15.79	32.72	38.92	15.8	4.1	–	–	–
Latent regression, twice the speech	0	1.52	13.06	27.93	41.06	16.4	4.3	–	10.56	2.50
Latent regression, four times the speech	0	1.51	12.40	27.34	36.84	16.8	4.4	–	9.82	2.58
Latent regression, four times the speech	1	1.46	12.14	24.98	36.09	16.9	4.5	–	–	–
Code-assignment cross-entropy	0	1.33	16.80	36.03	41.72	15.4	4.7	–	–	–
Regression + cell hinge	0	1.47	16.35	34.96	42.25	15.2	4.6	–	–	–
EnCodec 1.5 kbps										
Frozen codec	–	7.54	33.88	54.02	56.33	19.5	12.2	–	1.35	32.54
Latent regression (MSE)	0	7.78	26.04	42.02	53.67	38.1	23.8	–	6.14	19.90
Latent regression (MSE)	1	7.74	26.85	41.90	57.83	38.1	23.7	–	6.81	20.04
Latent regression (MSE)	2	8.15	26.39	41.81	54.64	38.2	23.8	–	–	–
Latent regression, twice the speech	0	8.39	25.06	40.82	55.20	38.4	23.9	–	5.54	19.52
Latent regression, twice the speech	1	8.54	25.06	41.22	53.99	38.4	23.9	–	–	–
Code-assignment cross-entropy	0	7.90	28.95	46.26	59.68	38.3	23.5	–	–	–

A.4 Two recent recognizers

Parakeet-CTC 1.1B (FastConformer, CTC, no language model) and Voxtral-Mini 3B (a speech language model in its transcription mode, language fixed to English, greedy decoding), both through their transformers implementations, transcribe the decoded waveforms of Table 1; text normalization and the codec-added definition are unchanged, and batched decoding matched one-utterance decoding on every smoke-test utterance. Table 11 gives the results.

Table 11: Amplification is not a property of older recognizers. The ten operating points of Table 1, same waveforms, read by two recent recognizers: Parakeet-CTC 1.1B (FastConformer, CTC, no language model) and Voxtral-Mini 3B (a speech language model, transcription mode, greedy). Codec-added WER on clean input and at 5 dB, and their ratio with a 95% utterance-bootstrap interval. Points whose ratio has a lower bound above 1.5: 9 and 10 of 10 (DAC 6.0 adds no clean-speech error under Parakeet). Spearman correlation of the 5 dB column with Whisper-small's: 0.99 and 1.00; median ratio of the observed 5 dB loss to the frozen Whisper-small law: 1.35 and 1.29.

	Parakeet-CTC 1.1B			Voxtral-Mini 3B		
Codec (kbps)	Clean	5 dB	Ratio	Clean	5 dB	Ratio
WavTokenizer 0.9	1.70	22.67	13.3× [11.6, 15.3]	2.99	29.97	10.0× [9.0, 11.3]
SNAC 1.0	1.73	24.98	14.4× [12.4, 16.9]	2.38	33.86	14.2× [12.5, 16.3]
Mimi 1.1	0.97	10.83	11.1× [9.1, 14.2]	1.50	16.45	11.0× [9.2, 13.3]
DAC 1.5	3.01	35.69	11.9× [10.5, 13.5]	4.25	44.69	10.5× [9.5, 11.7]
EnCodec 1.5	3.34	20.55	6.1× [5.5, 6.9]	3.91	26.89	6.9× [6.3, 7.6]
DAC 3.0	0.58	9.83	17.1× [12.4, 28.1]	0.87	12.83	14.8× [11.7, 19.7]
EnCodec 3.0	0.72	5.89	8.2× [6.3, 11.1]	0.87	7.47	8.6× [6.8, 11.5]
SpeechTok. 4.0	0.26	11.78	45.9× [28.8, 100.7]	0.47	15.87	34.1× [24.3, 55.3]
DAC 6.0	-0.02	1.56	–	0.13	2.03	16.0× [7.3, 98.0]
EnCodec 6.0	0.29	2.39	8.2× [5.1, 18.7]	0.26	2.83	10.9× [6.3, 32.6]

A.5 Adapting the recognizer to decoded audio

Every amplification in Table 1 is measured with a released recognizer that never heard codec output. To test whether the amplification is that mismatch, we fine-tuned Whisper-small three times on the same 26,003 train-clean-360 utterances (about 91 h), each degraded once by a training-noise rule disjoint from the evaluation noise (clean with probability 0.2, otherwise WHAM! training-split noise at an SNR drawn uniformly from 0 to 20 dB): on the uncoded waveforms, on the same waveforms decoded by DAC at 3.0 kbps, and decoded by Mimi at 1.1 kbps. The three runs share the steps (3,000 at batch 32), the learning rate (10^{-5} with 300 warm-up steps and linear decay), the seed, and the decoding of Table 1. Each path is read by the recognizer adapted to it: codec-added WER after adaptation is the WER of a codec's adapted recognizer on its decoded evaluation audio minus that of the uncoded twin on the uncoded audio, on the waveforms of Table 1. We registered in advance that an adapted amplification of at least 4, with a bootstrap lower bound of at least 2, would mean that adaptation does not remove the amplification, and one of at most 2 that it is mostly recognizer mismatch. Table 12 gives the result: adaptation roughly halves codec-added WER at every SNR and leaves the amplification at 9.3× for DAC and 8.0× for Mimi. The adapted uncoded twin itself improves from 4.93 to 4.12 on clean input and from 8.45 to 6.91 at 5 dB.

Table 12: Adapting the recognizer to decoded audio lowers codec-added WER but keeps its growth under noise. Codec-added WER (%) on the full test-clean with the WHAM! mixtures of Table 1. Released: Whisper-small as in Table 1, recomputed by this pipeline. Adapted: Whisper-small fine-tuned on 91 h of train-clean-360 degraded by the same training-noise rule and decoded by the codec, minus a twin fine-tuned on the same waveforms uncoded (same utterances, steps, and seed); each path is read by the recognizer adapted to it. Amplification: 5 dB over clean, utterance bootstrap 95% interval.

Codec, recognizer	Clean	15 dB	10 dB	5 dB	0 dB	Ampl. [95% CI]
DAC 3.0, released	1.45	4.31	8.41	18.11	37.42	12.5× [10.5, 15.4]
adapted, seed 0	1.03	2.44	4.64	9.63	23.64	9.3× [7.5, 12.1]
adapted, seed 1	1.06	2.50	4.45	9.48	23.86	9.0× [7.4, 11.3]
Mimi 1.1, released	2.69	5.47	9.28	21.32	40.44	7.9× [6.9, 9.3]
adapted, seed 0	1.51	3.02	5.51	12.13	28.80	8.0× [6.8, 9.7]
adapted, seed 1	1.61	3.05	5.39	12.17	29.02	7.5× [6.5, 9.0]

A.6 RECOGNIZERS TRAINED ON THE TOKENS

Every codec-added WER in Table 1 is measured on decoded audio. To ask whether a reader of the tokens themselves incurs the amplification, we trained character CTC recognizers that differ only in their input: Mimi's first eight codebooks (1.1 kbps, the Mimi point of Table 1), looked up in Mimi's own frozen codebooks and summed, which is what the codec's decoder receives; Mimi's continuous latent before the lookup, which the bypassed path of Eq. 1 decodes; and the 80-dimensional log-mel spectrogram of the uncoded audio. All share a Conformer encoder (12 layers, width 256, four heads, 50 frames per second), the batches and their order, the optimizer (AdamW, peak learning rate 10^{-3}, 4,000 warm-up steps, cosine decay over 45,000 steps), time masking (four masks of up to 5% of each utterance), greedy decoding, and the text normalizer; token and latent front ends map 512 dimensions to 256 and learn a four-fold upsampling from 12.5 frames per second, and the log-mel front end is two convolutions. Training uses all 104,014 utterances (364.7 h) of LibriSpeech train-clean-360, either clean or each degraded once by a training-noise rule disjoint from the evaluation noise (clean with probability 0.2, otherwise WHAM! training-split noise at an SNR drawn uniformly from 0 to 20 dB); tokens, latent, and log-mel of a degraded utterance come from the identical waveform, and the tokens of utterances the rule left clean are bit-identical to those of clean training. Evaluation uses the 2,620 test-clean utterances and the WHAM! mixtures of Table 1.

The log-mel twins reach 12.10 and 13.44% WER on clean input, and we count a difference as amplified when its 5 dB value is at least twice its clean value with an interval excluding zero. Table 13 gives the result. The quantizer's penalty for token readers, tokens minus latent, amplifies in both training conditions (3.3× and 2.3×, bootstrap intervals [2.83, 3.94] and [2.11, 2.62]), less than the decoded path's 7.3×. Trained on degraded speech, the token reader's cost over its log-mel twin amplifies 8.4× [6.36, 12.60]. Trained on clean speech, the token reader beats its log-mel twin at 5 dB because Mimi's encoder is far more robust than a log-mel front end that never heard noise (latent minus log-mel -14.25). Second seeds agree (Table 13); one degraded-condition log-mel run diverged and was rerun with a new seed. At Mimi's 4.4 kbps, Kyutai STT 2.6B scores 2.49 and 5.17 on clean input and at 5 dB, as Whisper-large-v3 does on the uncoded 5 dB input (5.10).

Table 13: Readers of Mimi's 1.1 kbps tokens incur an amplified quantization penalty. Paired per-utterance differences in WER points on the full test-clean with the WHAM! mixtures of Table 1; values separated by a slash are separate seeds (the rerun degraded log-mel twin reaches 15.24% clean WER). Conformer CTC recognizers (12 layers, width 256) trained on train-clean-360, clean or degraded by the training-noise rule, that differ only in their input: Mimi's tokens (dequantized by Mimi's frozen codebooks), Mimi's continuous pre-lookup latent, or uncoded log-mel. Tokens minus latent is the token-domain counterpart of the penalty in Table 1; latent minus log-mel is what Mimi's encoder, before the lookup, costs or saves. Ratio: 5 dB over clean where both are positive.

Comparison	Training	Clean	5 dB	0 dB	Ratio
Decoded audio, Whisper-small: codec-added	–	2.69	21.32	40.44	7.9×
Decoded audio, Whisper-small: penalty ($q-b$)	–	2.41	17.57	31.46	7.3×
Tokens − latent (quantizer penalty)	clean	+2.56 / +2.81	+8.46 / +9.10	+10.66 / +11.12	3.3× / 3.2×
Tokens − log-mel	clean	+0.75 / +1.60	-5.79 / -6.06	-3.88 / -3.29	– / –
Latent − log-mel (encoder)	clean	-1.81 / -1.21	-14.25 / -15.16	-14.54 / -14.42	–
Tokens − latent (quantizer penalty)	degraded	+3.71 / +3.04	+8.69 / +7.36	+11.36 / +10.75	2.3× / 2.4×
Tokens − log-mel	degraded	+1.30 / -0.13	+10.92 / +8.66	+19.27 / +17.74	8.4× / –
Latent − log-mel (encoder)	degraded	-2.41 / -3.17	+2.23 / +1.29	+7.91 / +6.99	–

A.7 CORPUS-LEVEL WER

Codec benchmarks report corpus-level WER, total edits over total reference words; Table 14 repeats Table 1 that way. Each utterance's edit count is its stored WER times its reference length; recomputing the edit distance from stored hypotheses reproduces these counts exactly for the deterministic codecs. Amplification spans 5.2–27.5× (4.5–22× in Table 1) and the quantizer's share 7–96%; the Spearman correlation between the clean and 5 dB columns stays $+0.903$, and the law refitted on the same 24 rungs is $W_{5\,\mathrm{dB}} = 12.0\,[10.8, 14.6] \cdot W_{\mathrm{clean}}^{0.69\,[0.55,0.74]}$. The normalizer keeps only letters, apostrophes and spaces. Applying Whisper's English normalizer on top of it lowers the uncoded

clean WER from 4.04 to 3.59 and changes codec-added WER by at most 0.89 points (amplification 5.3–27.0× over the nine RVQ points).

Table 14: Table 1 at corpus level: codec-added WER as total edits over total reference words (52,576 words) instead of the mean of per-utterance WERs. Same runs, normalizer and utterances; each utterance's edits are its stored WER times its reference length. Amplification: utterance bootstrap (5,000 resamples, each recomputing the corpus ratio) and speaker-clustered bootstrap (40 speakers). Penalty at 5 dB: 95% half-width, utterance-level (linearized) and speaker-clustered. Share: penalty over codec-added WER at 5 dB. Uncoded input: 4.04 on clean speech, 6.89 at 5 dB (per-utterance means 4.93 and 8.45). DAC 6.0's clean value sits at the recognizer's floor, so its ratio is poorly determined. SNAC's decoder is stochastic: over three runs its clean and 5 dB values span 3.26–3.34 and 37.57–39.17.

Codec (kbps)	Clean	5 dB	Ampl. [95% CI]	Speaker CI	Penalty (5 dB)	Speaker	Share
WavTokenizer 0.9	4.19	32.65	7.8× [7.3, 8.4]	[6.9, 9.0]	2.43±1.17	±1.08	7%
SNAC 1.0	3.34	37.76	11.3× [10.4, 12.4]	[9.8, 13.3]	17.01±1.26	±1.60	45%
Mimi 1.1	2.01	18.84	9.4× [8.4, 10.6]	[8.1, 11.2]	15.59±0.92	±1.76	83%
DAC 1.5	6.74	50.80	7.5× [7.0, 8.2]	[6.6, 8.8]	19.70±1.98	±2.04	39%
EnCodec 1.5	5.88	30.86	5.2× [4.9, 5.6]	[4.7, 5.9]	29.61±1.15	±2.69	96%
DAC 3.0	1.15	16.16	14.1× [11.1, 18.1]	[10.9, 19.2]	8.84±0.83	±1.05	55%
EnCodec 3.0	1.21	9.65	8.0× [6.9, 9.6]	[6.6, 10.6]	8.40±0.44	±1.08	87%
SpeechTok. 4.0	0.67	18.37	27.5× [22.5, 36.2]	[21.4, 37.6]	6.51±0.71	±1.28	35%
DAC 6.0	0.15	2.97	20.0× [10.3, 97.6]	[10.2, 95.1]	1.77±0.24	±0.24	60%
EnCodec 6.0	0.28	3.83	13.8× [9.3, 29.6]	[9.0, 29.2]	2.58±0.28	±0.36	67%

A.8 THE CLEAN-TO-DEPLOYMENT LAW

For each SNR we fit $\log W_{\mathrm{SNR}} = \log A + b \log W_{\mathrm{clean}}$ by least squares over every rate-ladder rung of DAC, EnCodec, Mimi, and SpeechTokenizer whose clean codec-added WER exceeds 0.15 (Whisper-small, full test-clean). The coefficients below were frozen on 2026-09-23 before the cells of the prospective test were run. Leave-one-family-out entries give the median ratio of observed to predicted loss for a family when the law is refitted without it. On the ten Table 1 points alone,

Table 15: Clean-to-deployment law per SNR (Whisper-small). Residual is the standard deviation of the log residual; LOFO is the median observed/predicted ratio for the held-out family.

SNR	A	b	resid. (log)	points	LOFO (median ratio)
15 dB	2.44	0.90	0.20	18	DAC 1.34, EnCodec 0.74, Mimi 0.91, SpeechTok. 1.47
10 dB	4.25	0.84	0.29	18	DAC 1.50, EnCodec 0.65, Mimi 0.95, SpeechTok. 1.73
5 dB	10.72	0.71	0.42	24	DAC 1.24, EnCodec 0.55, Mimi 0.96, SpeechTok. 2.28
0 dB	20.50	0.56	0.40	18	DAC 1.42, EnCodec 0.59, Mimi 1.14, SpeechTok. 2.86

the 5 dB law's constants depend on the recognizer: Whisper-small A=10.6, b=0.75 (residual factor 1.42, Spearman +0.90); Whisper-large-v3 A=15.4, b=0.34 (residual factor 2.86, Spearman +0.55); wav2vec2-base A=11.4, b=0.59 (residual factor 1.28, Spearman +0.90). Whisper-large-v3's clean codec-added WER sits near its floor at most operating points, which loosens both the fit and the ranking; Table 16 adds five low-rate rungs. In a prospective test, the coefficients were frozen before 32 further cells were run (new rungs and SNRs of Mimi and SpeechTokenizer, and Opus at 4–12 kbps and 15–0 dB); the observed loss is within a factor 1.29 of the prediction at the median and within a factor 2 in 84% of cells. The 5 cells outside a factor of 2 (observed/predicted): SpeechTokenizer 1 kbps at 0 dB (2.11), SpeechTokenizer 2 kbps at 0 dB (2.46), SpeechTokenizer 3 kbps at 10 dB (2.09), SpeechTokenizer 3 kbps at 0 dB (2.88), Mimi 0.14 kbps at 5 dB (0.36). The three references use the Table 1 waveforms and recognizer and add no codec: white Gaussian noise at a fixed SNR relative to the input's own power; an eighth-order Butterworth low-pass applied forward and backward; and uniform quantization of the log-magnitude STFT (512-point Hann window, hop 128) with the phase kept. Figure 1a draws them; the log-spectral quantizer's error follows its input, white noise does not.

Table 16: The law's form carries across degradations, a second recognizer, and real far-field recordings; its constants change with the condition, and a recognizer whose clean loss sits at the floor cannot use it. Tests marked frozen score cells run after the coefficients were fixed.

Condition	Test	Result
Six-talker babble	EnCodec ladder (8 rungs) against the DAC-ladder law	exponent 0.88/0.69/0.50 against DAC's 0.92/0.71/0.43 at 15/5/0 dB; EnCodec sits at 0.67/0.69/0.56× DAC's line, as under noise
Babble, 5 dB	Mimi and SpeechTokenizer rungs (4+4) against the DAC-babble law	median ratio 0.92 (Mimi) and 1.77 (SpeechTokenizer)
Reverberation, RT60 400 ms	10 new rungs, frozen 9-point law	within 1.23× at the median, 90% within 2×; every rung below the law
wav2vec2-base, 5 dB	10 new rungs, frozen 10-point law	within 1.19× at the median, 100% within 2×
Whisper-large-v3, 5 dB	Table 1 points plus five low-rate rungs; the 7 with clean loss >1	b=0.72 but residual 2.89×: the clean loss sits at the floor
AMI, headset → distant microphone (real)	ten Table 1 points, 2,000 paired segments	amplification 1.4–4.8×; A=7.04, b=0.61, residual 1.24×, Spearman +0.84
AMI, prospective	16 new rungs, frozen AMI law	within 1.16× at the median, 100% within 2×
AMI, Opus 4–16 kbps	6 rates, frozen AMI law	within 1.45× at the median, 83% within 2×
StableToken (noise-robust tokenizer), 5 dB	released tokenizer and decoder, Table 1 waveforms	clean 1.59, 5 dB 5.01 (3.1×); 0.34 of the law; token flip rate 0.41
White noise, no codec	same inputs, SNR 30–0 dB re. input	amplification 1.6–2.4×, 0.17–0.37 of the law; 2.3× at DAC 3.0's clean loss
Low-pass filter, no codec	same inputs, cutoff 4–1 kHz	amplification 2.8–9.6×, 0.63–0.76 of the law; 7.3× at DAC 3.0's clean loss
Log-spectral quantizer, no codec	same inputs, step 1.5–4 nepers	amplification 8.3–14.2×, 0.91–1.47 of the law; 10.4× at DAC 3.0's clean loss

Table 17: Statistics of the 5 dB clean-to-deployment law. Paired tests compare $|\log(\text{obs}/\text{pred})|$ of the fixed law with the frozen nulls on the 32 prospective cells. Family-level intervals resample the four fitted families (and, in the two-level bootstrap, the 40 test-clean speakers). The prediction interval is for a codec from a new family, from a random-intercept model whose between-family variance is 87% of the total.

Quantity	Value
Law vs. constant ratio (32 cells)	smaller error in 22; Wilcoxon p=0.012, sign test p=0.050, rank-biserial r=0.50; median factor 1.29 vs. 1.48; within 2×: 84% vs. 78% (McNemar p=0.69); clustered by operating point p=0.045–0.058 (sign-flip, Wilcoxon on 14 rung means)
Law vs. bitrate only (32 cells)	smaller error in 25; Wilcoxon $p=3.9\times10^{-5}$; median factor 1.29 vs. 2.43
b, frozen fit (24 rungs)	0.707; leave-one-family-out 0.69–0.72; family bootstrap [0.62, 0.73]; families × speakers [0.55, 0.78]; random intercept 0.70 [0.65, 0.76]
b without the two saturated rungs	0.80; family bootstrap [0.62, 0.88]; jackknife (3 df) [0.57, 1.02]
Within-family slopes, 5 dB	DAC 0.73 [0.60, 0.85]; EnCodec 0.71 [0.62, 0.80]; Mimi 0.51 [0.39, 0.63]; SpeechTok. 0.51 [0.33, 0.68]; Opus 0.59 [0.34, 0.84]
Within-family slopes, all SNRs	20 of 20 below one; 17 of 20 with a 95% CI below one
Within-family b away from the ceiling (deployed WER $\leq$30)	0.90 [0.86, 0.94], 0.84 [0.78, 0.90], 0.77 [0.68, 0.87], 0.74 [0.58, 0.90] at 15/10/5/0 dB (all cells: 0.89/0.82/0.64/0.55); babble 5 dB 0.94 [0.70, 1.18], AMI 0.74 [0.38, 1.10]
Measurement error in W_{clean}	reliability of $\log W_{\text{clean}}$ 0.98; corrected b 0.72 (pooled), 0.71 (within family)
95% prediction interval, new family	factor 2.9 (normal) to 5.5 (t, 3 df); one-SD residual factor 1.52; held-out and prospective 5 dB points inside (normal): 10 of 11 (StableToken at 0.34 is the exception)

Table 18: The clean-to-deployment law as a calibrated tool. **Top:** one-anchor calibration. The exponent is frozen at the WHAM! value for the matching SNR (5 dB where the condition has none) and A is set from one measured codec; each operating point in turn is the anchor and the others are predicted from their clean-side loss. Entries are the median over anchors of the median factor $\exp|\log(\mathrm{obs}/\mathrm{pred})|$, with the worst anchor in parentheses. Constant ratio: $W = kW_{\mathrm{clean}}$ with k from the same anchor. Refit: A and b fitted on all points of the condition (in sample). Same family: the anchor is a rung of the family being predicted. $^{\dagger}b$ was fitted on 18 of these points. ‡Clean loss at the recognizer's floor; outside the law's domain. **Bottom:** rate provisioning with frozen coefficients: the lowest rate of each family (DAC, EnCodec, Mimi, SpeechTokenizer, Opus) whose deployed codec-added WER is at most 3, 5 or 10 points; 15 decisions per deployment. A decision is right if it picks a rung that meets the budget when one exists, or reports that none does when none exists. Fitted without the provisioned family, the law is right in 11 of 15 at 5 dB; the AMI law shares eight rungs with the provisioned ladders.

		One anchor (any codec)			Same-family anchor	
Condition	n	Law	Const. ratio	Refit	Law	Const. ratio
Babble 15 dB	16	1.25 (1.58)	1.31 (2.18)	1.26	1.12	1.15
Babble 10 dB	8	1.19 (1.34)	1.09 (2.32)	1.14	1.19	1.09
Babble 5 dB	24	1.53 (2.50)	1.74 (7.31)	1.31	1.27	1.38
Babble 0 dB	16	1.51 (2.41)	1.75 (13.88)	1.45	1.29	1.72
Reverberation 400 ms	19	1.16 (2.97)	1.54 (6.87)	1.10	1.10	1.59
wav2vec2, 5 dB	20	1.27 (2.21)	1.79 (6.30)	1.17	1.20	1.66
AMI, headset→distant	32	1.28 (3.18)	1.53 (2.62)	1.19	1.19	1.39
AMI, LibriSpeech→distant	31	1.70 (3.12)	2.29 (7.84)	1.39	1.65	2.45
WHAM! 15 dB†	29	1.20 (1.50)	1.21 (2.05)	1.16	1.09	1.20
WHAM! 10 dB†	29	1.29 (1.89)	1.37 (3.07)	1.29	1.13	1.30
WHAM! 0 dB†	29	1.46 (2.48)	1.74 (9.58)	1.26	1.17	1.77
large-v3, 5 dB‡	12	2.22 (13.71)	3.49 (13.53)	1.52	2.62	4.81

Rate provisioning (15 decisions each): right decisions (violations); one anchor: mean

Deployment	Clean board	Fixed law	Frozen const.	One anchor: law / const.
WHAM! 5 dB	0 (15)	12 (1)	9 (6)	–
AMI, distant	0 (15)	14 (0)	–	88% / 74%

Table 19: Blind test of the fixed laws on codec families we had not measured, and the held-out corpus test-other. **Top:** X-Codec (Ye et al., 2025b), the Higgs Audio v2 tokenizer (Boson AI, 2025), FocalCodec (Della Libera et al., 2025) at 50, 25, and 12.5 Hz, and the continuous VibeVoice acoustic tokenizer, which has no lookup, on the Table 1 waveforms, Whisper-small; predictions for 15, 10, 5, and 0 dB were registered from each point's measured clean loss before any noisy cell was run. Entries: codec-added WER, with observed over predicted in parentheses. X-Codec and Higgs (the first test, 28 cells): median factor 1.32 with 86% within 2×; constant ratio 1.48 and 68%. All 60 blind cells: 1.32 and 90% (constant ratio 1.47 and 75%); per-family median observed over predicted: FocalCodec 0.74, GLM-4-Voice 0.51, Higgs v2 0.86, MOSS 0.91, VibeVoice 1.49, X-Codec 1.74. **Bottom:** the ten operating points of Table 1 on LibriSpeech test-other (2,939 utterances), clean and 5 dB; the fixed 5 dB law is applied to each point's test-other clean loss. Median factor 1.53 with 90% within 2×; one anchor, b fixed: 1.25 (constant ratio 1.28).

Codec (kbps)	Clean	15 dB	10 dB	5 dB	0 dB
Higgs v2 0.5	12.45	22.06 (0.93)	31.66 (0.90)	47.36 (0.74)	65.70 (0.78)
Higgs v2 1	2.72	5.69 (0.95)	9.59 (0.97)	19.93 (0.92)	37.16 (1.03)
Higgs v2 2	1.00	1.90 (0.78)	3.32 (0.78)	7.32 (0.68)	16.81 (0.82)
X-Codec 0.5	2.69	5.09 (0.85)	10.28 (1.06)	27.00 (1.25)	66.38 (1.86)
X-Codec 1	1.07	2.82 (1.09)	6.44 (1.43)	19.20 (1.71)	53.68 (2.52)
X-Codec 2	0.46	1.66 (1.38)	3.92 (1.78)	12.68 (2.06)	34.99 (2.65)
X-Codec 4	0.31	1.25 (1.46)	2.89 (1.80)	8.34 (1.77)	24.17 (2.26)
FocalCodec 0.65	1.26	1.77 (0.59)	3.50 (0.68)	9.55 (0.76)	24.46 (1.05)
FocalCodec 0.33	2.04	3.50 (0.75)	5.87 (0.76)	13.39 (0.75)	32.46 (1.06)
FocalCodec 0.16	6.38	8.59 (0.66)	11.91 (0.59)	22.23 (0.56)	42.21 (0.73)
GLM-4-Voice 0.175	2.32	3.70 (0.71)	4.82 (0.56)	8.32 (0.43)	15.12 (0.46)
MOSS 0.5	3.39	8.46 (1.15)	13.21 (1.12)	22.91 (0.90)	35.93 (0.88)
MOSS 1	0.78	2.11 (1.08)	3.91 (1.14)	7.81 (0.87)	16.49 (0.93)
MOSS 2	0.23	0.60 (0.92)	1.06 (0.85)	3.32 (0.87)	7.39 (0.82)
VibeVoice (cont.)	0.49	1.86 (1.44)	3.63 (1.54)	9.12 (1.40)	22.22 (1.61)
test-other: clean, 5 dB (observed over predicted), amplification (Table 1 in brackets)					
WavTokenizer 0.9	15.53		51.62 (0.69)		3.3× [6.8×]
SNAC 1.0	15.61		57.23 (0.76)		3.7× [9.4×]
Mimi 1.1	9.93		34.99 (0.64)		3.5× [7.9×]
DAC 1.5	22.65		65.02 (0.67)		2.9× [6.6×]
EnCodec 1.5	21.58		49.42 (0.53)		2.3× [4.5×]
DAC 3.0	6.08		31.85 (0.83)		5.2× [12.5×]
EnCodec 3.0	6.43		21.05 (0.53)		3.3× [6.8×]
SpeechTok. 4.0	3.75		36.29 (1.33)		9.7× [22.0×]
DAC 6.0	1.28		8.07 (0.63)		6.3× [12.3×]
EnCodec 6.0	2.24		9.34 (0.49)		4.2× [9.2×]