WHICH PARTIALS MOVE: A 12-TET PULL ABOVE THE BAND EDGE OF A NEURAL CODEC DECODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural synthesis pulls output tuning toward 12-TET, as recently reported for mel-to-audio vocoders. We ask where inside a neural audio codec that happens. At low bitrate a decoder transmits a tone's lower partials and invents the rest above a *band edge* that rises with bitrate within a codec. The transmitted fundamental stays put through EnCodec, whereas the vocoder we measure moves it by 3.1 cents (a cent is a hundredth of a semitone). The invented partials land most of the way onto the harmonic ladder of the nearest 12-TET pitch, and a pitch estimator fitting across the edge averages the transmitted and invented partials, reporting a displacement no partial underwent: a median of 4.37 cents over twenty-four stimuli. Where the invented partials land follows the decoder's fine-tuning corpus. When only EnCodec's decoder is fine-tuned, so that every arm reads bit-identical tokens, retuning its corpus by 33 cents moves the decoder's grid by 32.7 ± 2.8 cents at seed-level spread, while grid-preserving resamplings of comparable or larger size leave it in place. Classical codecs show no pull. At a 440 Hz reference six of eleven measurable neural conditions show the effect, four of them by more than five cents. Those four hold under every variance treatment and span two codec families; the other two fall below their own detection floors. The practical exposure falls on full-band harmonic fitting, ours more than salience-based tools. After one round trip, 22 of 200 Carnatic and 11 of 60 Turkish makam recordings switch to fitting 12-TET better than their own corpus's measured degrees, while a period tracker is largely spared. The remedy is simple where it applies: fit no partial above the edge. The band-edge account replaced a pre-registered one that failed, and its registered pass on fifteen fresh checkpoints is a consistency check rather than a confirmation, giving $p = 0.070$ against always predicting the middle tier and 0.375 against always predicting the most common tier.

1 INTRODUCTION

Pitch here is in cents, a hundredth of a semitone (c in tables), and 12-TET is the twelve-tone equal-tempered grid of 100-cent steps anchored at A4 = 440 Hz. Modern audio models operate not on waveforms but on discrete tokens produced by a neural audio codec (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023): music generation (Borsos et al., 2023; Agostinelli et al., 2023; Copet et al., 2023), speech synthesis (Wang et al., 2023) and spoken dialogue (Défossez et al., 2024; Yang et al., 2024) all model residual-vector-quantiser codes and hear them through the codec's decoder. The codec fixes the vocabulary of everything built on it, as a subword tokeniser does for text, and tokenisers too were taken as neutral until their quality was shown to vary by language (Rust et al., 2021; Petrov et al., 2023; Ahia et al., 2023). Codec quality is reported as aggregate scores over generic corpora (Wu et al., 2024a;b; Mousavi et al., 2024; Shi et al., 2024), several of which report an F_0 error. A corpus score sees the residual's distribution but not its map: it cannot separate a grid-directed displacement from estimation error, nor say which partials moved.

At low bitrate a learned decoder must regenerate the spectrum the code does not carry, and what it regenerates reflects what its corpus taught it to expect. Figure 1 shows this on one tone. The fundamental stays within 0.2 cents of the input's in the mean, the partial on the edge carries two lines of comparable level, and the partial wholly above it has moved to the grid harmonic. The evidence places the regridding in the decoder, not in the tokens. Codec latents read directly show discretised pitch features (Paek et al., 2025), but we detect no grid in the code assignments (Table

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.

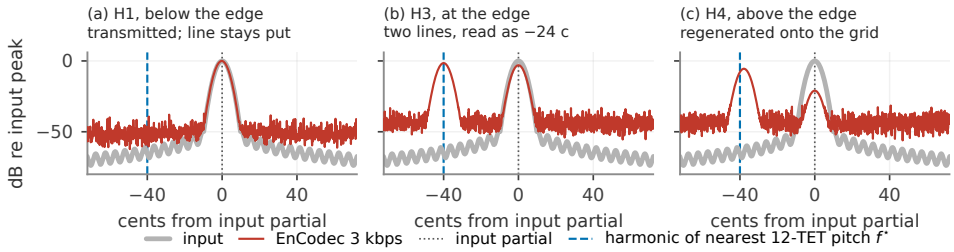

Figure 1: **What moves, and what does not.** A 440 Hz complex detuned $+40$ cents, EnCodec at 3 kbps. Hk is partial k; dashed, $kf^\star$, with $f^\star$ the 12-TET pitch nearest f_0; zero marks the input partial. In (b) the coded trace is shifted up for display; -24 is the two lines' centroid (Table 1).

A10), and removing the quantiser at a matched band edge does not remove the effect (Section 3.2); both bound the code's role. The effect depends on decoder and register: six of eleven conditions show it at a 440 Hz reference (Section 3.1).

We treat the codec as a measurement instrument: a pair of harmonic tones a controlled interval apart is coded, a blind estimator reads the decoded interval, and the residual is studied as the interval sweeps the octave and as the reference pitch is detuned. The detuning is what makes the instrument discriminating, and is the paper's *registration test*: a residual anchored to absolute pitch travels with the reference, advancing its within-semitone phase at slope 1, while one tied to the interval or to the analysis stays at slope 0.

1.1 Contributions

1. **Which partials move.** That a decoder invents rather than transmits above a band edge is the premise of bandwidth extension (Dietz et al., 2002; Nagel & Disch, 2009; Choi et al., 2025); new here is where the invented content lands. In the decoders that show the effect, the fundamental and the partials below the edge stay put, and those above it move most of the way onto the harmonics of the nearest 12-TET pitch. Two measurements summarise this: the *edge*, the lowest partial the decoder regenerates, and the *ladder fraction*, how far the partials at or above it move toward the grid. Section 3.2 reports both.
2. **The ladder follows the corpus.** Fine-tuning only the decoder leaves the tokens bit-identical and moves the residual with the corpus: retuning it by $+33$ cents moves the residual's zero crossing by 32.7 ± 2.8 cents at seed-level spread, a prediction entered in the released pre-registration's amendment log before the arm ran. An extender trained from scratch to fill the band above EnCodec's 1.32 kHz edge places its invented partials 0.68 [0.55, 0.81] of the way to the grid on a grid-peaked corpus, 0.14 [0.00, 0.28] on the same audio flattened, and 0.09 on synthetic harmonics that never carried a grid. EnCodec's corpus is not public, so the causal claim is that changing a decoder's corpus changes its ladder, though only 0.87 cents of its 12.55 is isolable in one already trained.
3. **A measurement hazard, and its fix.** A full-band harmonic fit reads the invented band. On real music the frame-based estimator reads 2.49 cents against the 5.90 the same estimator reads on the 440 Hz tone, and a round trip switches 22 of 200 Carnatic and 11 of 60 Turkish makam recordings to fitting 12-TET better than their own corpus's measured degrees, a confound in Qin & Lerch (2019)'s sense. A period tracker is largely spared, the Carnatic flip rate falling to 3% under pYIN (Mauch & Dixon, 2014); how far salience-based tools inherit it depends on how much weight they place above the edge (Table A33). Fitting only below the edge takes both corpora from 3.2 and 4.4 cents of added degree error to 0.2 each, both intervals covering zero; most of the error it removes is generic codec damage (Section 3.4).

1.2 Related work

Two things about this were already known, and neither locates the effect. A decoder may synthesise the spectrum above a band edge rather than transmit it: spectral band replication in HE-AAC (Dietz et al., 2002), the folding in Opus's CELT layer (Valin et al., 2012), and regeneration now designed on purpose (Choi et al., 2025). Copy-up translates by a frequency offset, misplacing partials against the input's ladder (a mismatch harmonic transposition (Nagel & Disch, 2009) was built to remove), and

carries no frequency prior, so that displacement is uncorrelated with 12-TET. And neural synthesis pulls output tuning toward 12-TET: Berendes et al. (2025) shift recordings to non-standard tunings, pass them through mel-to-audio vocoders, and find the output clustered near A4 = 440 Hz, at up to 18.7 cents, with a listening test finding tuning affects perceived quality; Berendes et al. (2026) then fine-tune BigVGAN-V2 to preserve it. Di Giorgi et al. (2022) report mel inversion destabilising pitch on sustained notes, and Qin & Lerch (2019) that tuning frequency confounds music classifiers. Our downstream claim is the same confound with the codec in the role of the tuning shift, established by the intervention method of Rodríguez-Algarra et al. (2019). The corpus explanation is Berendes et al.'s, who also report the bias shrinking as mel bands and sampling rate rise, an information-loss reading our band edge locates in frequency. Our claim is narrower: we measure which partials move and how far, finding the fundamental unmoved; describe the effect by an edge that rises with bitrate within a codec and a ladder the decoder's corpus shifts; retune a corpus, which *translates* the grid rather than flattening it; and follow the consequence for non-Western tuning. Appendix I sets the boundary out claim by claim.

A full-band estimator forming one pitch across partials, so that displaced upper ones move the reading while the fundamental stays put, is the algorithmic analogue of virtual pitch (Terhardt et al., 1982); new is the spectral condition a codec creates to trigger it. A decoder can also emit tones fixed by its upsampling operator rather than its corpus (Pons et al., 2021; Odena et al., 2016), which registration cannot separate from a learned grid but the retuning of Section 3.3 can.

The question falls between literatures that rarely meet. Codec design has turned toward the tokeniser role (Ye et al., 2025a; Liu et al., 2024; Xin et al., 2024; Yang et al., 2023; Du et al., 2024) while evaluation, even varying one factor (Gichamba & Busogi, 2026; Wang et al., 2026; Altwlkany et al., 2025), scores recorded audio, which averages the residual's structure away. Bandwidth extension (Kuleshov et al., 2017; Liu et al., 2023) and vocoders regenerate spectrum a signal does not carry and inherit their corpus's statistics, and a low-bitrate decoder is both; spectral losses are poor at pitch (Turian & Henry, 2020) and GAN generators carry systematic pitch error (Morrison et al., 2022). Music information retrieval builds corpora on tuning systems 12-TET does not describe (Serra, 2011; Uyar et al., 2014; Bozkurt et al., 2014; Srinivasamurthy et al., 2021; Mehta et al., 2025).

2 Measuring the codec transfer function

2.1 Codecs and the transfer function

A neural audio codec comprises an encoder, a residual vector quantiser and a decoder, at R/D bits per latent dimension for a rate R over D dimensions, all three learned from a training corpus (Appendix A.1). The symbols the results use are below and in full in Appendix B.1; the quantiser account defines its own where they are used.

θ	interval put in, cents	σ	registration slope: 1 absolute, 0 interval
$r_C(\theta)$	interval read back minus θ, cents	f_{e}	band edge, kHz: lowest regenerated partial
δ, δ_0	grid distance of the interval; of the tone	s_k	displacement of partial k, cents
b, $\tilde{b}$	grid bias and its off-grid median, cents	ℓ_k	s_k as a share of $kf_0 \to kf^\star$
d	detuning of the reference, cents	$\bar{\ell}$	mean ℓ_k from f_{e} up; $\bar{\ell}_r$: regenerated
H_k	partial k, in figures and tables	$f^\star$	the 12-TET pitch nearest f_0

Transfer function. Let $x(\theta)$ be a stimulus in which a single property takes the value θ, and let $\mathcal{E}$ be an estimator that recovers that property from a waveform. The *transfer function* of the codec along θ and its residual are (medians over the random phases and gains of repeated trials are reported)

$$T_C(\theta) \;=\; \operatorname*{med}_{\text{trials}} \mathcal{E}\big[C(x(\theta))\big], \qquad r_C(\theta) \;=\; T_C(\theta) - \theta. \tag{1}$$

A corpus-averaged loss scores any two residual maps with the same marginal distribution identically; the transfer function keeps the map.

2.2 The grid, the signed metric and what we report

For the pitch probe θ is a musical interval in cents. Let $g(\theta) = 100 \cdot \mathrm{round}(\theta/100)$ be the nearest 12-TET interval and $\delta(\theta) = \theta - g(\theta) \in [-50, 50]$ the signed distance, round taken half-to-even; at

$|\delta| = 50$ the tone is equidistant from two grid points, so *toward the grid* is a convention there (5% of a sweep). *Grid bias* is the residual's component directed toward the grid,

$$b(\theta) \;=\; \operatorname{sign}\big(g(\theta)-\theta\big)\cdot r_C(\theta) \;=\; -\operatorname{sign}\big(\delta(\theta)\big)\, r_C(\theta), \tag{2}$$

so $b(\theta) > 0$ means the interval was read as moved toward the Western semitone grid. The main text gives every verdict in one bias statistic, the *off-grid bias* $\tilde{b}$: the median of b over trials with $|\delta| \geq 30$, which leaves out the near-grid region where the bias is small by construction. Two companions recur and are not interchangeable with it: the all-trial median, 8.88 cents at 3 kbps, which reads smaller and carries (P2), the quantiser bypass, the rate sweep and repeated coding; and the amplitude A of a sinusoid fitted in δ, 13.72, which reads larger because a sinusoid overshoots a saturating profile and carries the causal arms, whose contrasts are in amplitude. Table A3 repeats this map with its reasons. *Pull* names the phenomenon and, of a condition, the verdict of Section 2.3; it is never itself a statistic. One more recurs: the *attractor*, where the residual crosses zero.

Intervals on the pitch sweeps are trial bootstraps within one sweep. On the headline row, EnCodec at 3 kbps and 440 Hz, that interval is ± 1.5 cents at the Bonferroni $z = 2.84$, a standard error of 0.52, wider than the between-run sd of 0.12 cents that five end-to-end repetitions give; folding the sd in leaves the registration verdict untouched and gives a stricter reading of the pull count of Section 3.1 (Appendix G.2).

2.3 The registration test

A 100-cent-periodic residual is not by itself evidence of a learned grid: any quantiser produces a residual with the period of its own cells, and so may the estimator. A learned grid has its zero crossings at 12-TET *pitches*, a statement about absolute frequency, so detuning the reference by d cents separates two hypotheses: under absolute anchoring the profile translates by d, under interval anchoring it does not. Regressing the fitted sinusoid's phase on d,

$$\varphi(d) \;=\; \varphi_0 + \sigma\,(2\pi/100)\,d, \qquad \sigma = 1 \text{ if anchored to pitch}, \quad \sigma = 0 \text{ if to the interval}, \tag{3}$$

makes the slope σ the test that tells the two apart. It establishes anchoring to absolute pitch, not a temperament. A condition *registers* when its 95% interval lies *inside* 1 ± 0.15, a margin fixed before Table A9 was computed. This is an equivalence test, not a failure to reject. The margin is wide against the precision we reach, so on this data it discriminates on location and not on width (Table A9).

Registration is necessary and not sufficient: any structure fixed in absolute frequency registers, and Appendix D.1 shows a classical codec whose slope passes on a residual too small to carry a phase. We claim a learned grid only where registration meets the *pull* rule, a condition *pulling* when its on-grid phase lies within 45° of 180° and its adjusted off-grid bias interval clears zero; amplitude and octave-constant period are supporting checks, not part of it (Appendices B.3 and C.4). On our data every condition outside the phase window also fails the bias window, so the bias interval decides every verdict.

2.4 What the estimator reads: band edge and ladder fraction

A harmonic-fit estimator combines the first eight partials into one number, so its residual does not say which partials moved; we therefore read the decoded spectrum directly. Let $\hat{f}_k$ be the power-weighted centroid of the lines near the kth input partial kf_0: those within 100 cents of it and within 6 dB of the strongest among them. That is one line except at the edge itself (Figure 1b). Let $f^\star$ be the 12-TET pitch nearest f_0, and $\delta_0 = 1200\log_2(f_0/f^\star)$ the tone's own detuning; the *chromatic pitch* of a partial is the 12-TET pitch nearest $kf^\star$. The two coincide only for k a power of two: at $k = 5$ and 7 the chromatic pitch sits 13.7 and 31.2 cents from $kf^\star$, which is why the two targets are not interchangeable below. The per-partial displacement and its *ladder fraction* are

$$s_k \;=\; 1200\log_2\frac{\hat{f}_k}{kf_0}, \qquad \ell_k \;=\; -\frac{s_k}{\delta_0}, \qquad f_{\mathrm{e}} \;=\; \min\big\{kf_0 : \text{partial } k \text{ regenerated}\big\}, \tag{4}$$

so $\ell_k = 0$ for a partial left on the input's harmonic ladder and $\ell_k = 1$ for one moved onto the grid pitch's ladder (a decoder that does this *regrids*). They are anchors, not bounds: ℓ_k is a ratio, so a partial that overshoots its target takes a value above 1 (Appendix C.5). The *band edge* f_{e} is

the lowest regenerated partial, right-censored at $0.95 f_s/2$ when there is none. An edge reported per condition is a mean over the registers where it resolves, so it need not be a multiple of f_0. $\bar{\ell}$ is read at $|\delta_0| \geq 20$, on the tone's own detuning rather than the interval's, where its denominator is off zero; across 5 to 50 it is flat, a fraction and not a saturating displacement (Appendix C.5).

Regeneration is scored on level *and* placement: partial k counts as transmitted when its decoded line is within 6 dB of the input's, and as regenerated when it has fallen more than 10 dB below that and the strongest line within 100 cents sits elsewhere. The second clause conditions on placement, and so, through Equation 4, does the edge. Dropping it gives the *level-only* reading, which admits attenuated partials that did not move, and where one sits below the edge it lowers both the edge and $\bar{\ell}$ (Appendix G.6). The edge is a minimum, not a boundary: Table A15 shows unmoved partials above it, and $\bar{\ell}$ averages over them rather than hiding them. Taken over the regenerated partials alone, the same mean is the *conditional* ladder fraction $\bar{\ell}_r$. The estimator weights partial k by $w_k \propto k^2 |X_k|^2$, so its residual is a weighted mean of the s_k; weighting the eight partials equally instead gives the *uniform-weight reading* of Appendix C.6.

2.5 Stimuli, estimation and controls

Stimuli and estimation. Each stimulus is a pair of harmonic complex tones in sequence, the first at a reference f_0 and the second at $f_0\, 2^{\theta/1200}$, with $1/n$ partial amplitudes and phases redrawn per repetition; the pitch probe uses eight partials over 500 ms, the spectral sweep twelve over one second. The interval sweeps 0 to 1200 cents in 5-cent steps at twenty draws and the reference at eleven values spanning a semitone in 10-cent steps: $241 \times 11 \times 20 = 53{,}020$ trials, the 100-cent reference folding onto 0, repeating at 220 and 880 Hz. Further sweeps cover four octaves, three spectral envelopes, a population grid of six envelopes by four registers, a pure sinusoid control and, for speech codecs, a source-filter vowel (Appendix B.4).

F_0 is estimated by YIN (de Cheveigné & Kawahara, 2002) seeding a harmonic least-squares refinement over the first eight partials, the coarse stage blind to the synthesis parameters. A learned tracker (Kim et al., 2018; Gfeller et al., 2020) is excluded because, trained on music, it would carry the prior we are measuring; on uncoded stimuli our error stays at or under 5×10^{-4} cents. The independent unit is a reference pitch at one register in one timbre. The off-grid bias and the interval deciding the pull are percentile bootstraps over one condition's trials at the on-grid reference. The resampling unit is the trial, so those intervals cover scatter between trials and cover nothing over detunings, registers or timbres (Appendices C.7, B.6). The only exclusion applied to trials is a tracking gate at 200 cents from nominal. Where we report a *band-limited* reading, it is this same estimator with the partial set truncated to those below that condition's measured band edge, so it reads the transmitted band alone and fits nothing the decoder invented (Appendix B.6). Spectral-sweep tones are analysed over their steady 0.9 s with a Hann window zero-padded to 2^{20} points (0.09-cent bins at 440 Hz and 24 kHz), recording every line's displacement and level around each input partial, so moved, split and added lines are distinguishable.

Codecs. We evaluate EnCodec (Défossez et al., 2023), DAC (Kumar et al., 2023), Mimi (Défossez et al., 2024), SpeechTokenizer (Zhang et al., 2024), SNAC (Siuzdak et al., 2024) and WavTokenizer (Ji et al., 2025b); an operating point is a bitrate or a count Q of residual quantiser stages (Appendix B.2). Twenty-six further checkpoints, their ladder fractions unread while the mechanism was formed, are held out for the pass Section 3.3 reports: twenty from untouched families spanning eighteen systems, and six of the above, the refused SNAC 24k among them (Appendix G.6). A further fifteen, untouched by either earlier set, carry the registered pass (Section 3.3). Figure 2 also carries BigVGAN (Lee et al., 2023), a neural vocoder, and Opus, MP3 and HE-AAC as classical controls; Vocos (Siuzdak, 2024), a second decoder read on another codec's tokens, enters through the held-out pass.

2.6 A quantiser account, as the foil

The hypothesis we pre-specified places the residual in the quantiser: a codebook learned from 12-TET music rounds an off-grid pitch toward an on-grid one. It predicts a sawtooth residual in δ (P1), a positive off-grid bias (P2), a bias falling as $2^{-mR/D}$ in bits per latent dimension, with m set by the dominant error term (P3), and registration to absolute pitch (P4). Appendices A.1 to A.3 state and

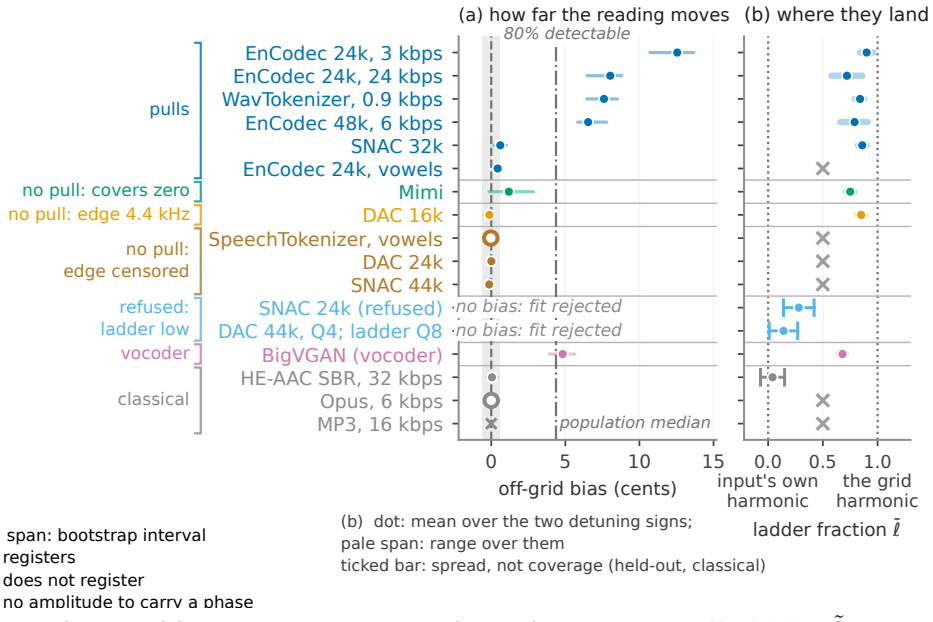

Figure 2: **Which conditions pull, and where their partials land.** (a) Off-grid bias $\tilde{b}$ at 440 Hz, Bonferroni over eleven neural rows; grey band, ± 0.60, the largest of the four null bounds, not Mimi's 2.00; $\times$, no phase. (b) Ladder fraction $\bar{\ell}$, DAC 16k's at 880 Hz, DAC 44k's at Q8; $\times$, edge not resolved.

derive these, and record that EnCodec runs outside the assumed regime, so (P3) is a remark rather than a test.

3 Results

What was fixed in advance. The pre-registration fixed Section 2.6's four predictions, the pull rule, the registration window, the tracking gate and two speech tests on FLEURS (Conneau et al., 2022); of these, registration and (P2) hold, (P1) fails and (P3) is out of regime (Appendix A.5). The band edge and ladder fraction (Section 3.2), the fine-tuning arms (Section 3.3), whose $+33$-cent prediction was nonetheless logged before its arm ran, and the readings on real recordings (Section 3.4) are exploratory and unadjusted, and the registered pass of Section 3.3 is the one test of the band-edge account fixed in advance (Appendices G.1, G.6).

3.1 The residual is anchored to absolute pitch

For EnCodec at 3 kbps the registration slope of Equation 3 is 1.0010 [0.9922, 1.0099], inside the pre-specified window; with no codec the bias is $< 10^{-5}$ cents, and the anchoring holds at 220 and 880 Hz.

Over a population of twenty-four stimuli, six spectral envelopes at four registers, the median off-grid bias is $\tilde{b} = 4.37$ cents [1.73, 9.13], a bootstrap over envelopes and registers read as spread rather than coverage (Table A18, Figure A8). That is the size of the effect for a full-band harmonic fit. Conditions are compared on one tone, a 440 Hz complex reading 12.55 cents through EnCodec at that population's 92nd percentile; across four octaves the bias runs 2.75 to 12.55 (Table A13).

Six of eleven conditions pull. Eleven of the thirteen neural conditions are measurable, the guards of Appendix B.8 refusing DAC 44k and SNAC 24k, and ten register inside the 1 ± 0.15 margin (Schuirmann, 1987); SpeechTokenizer registers only under the estimator's cross-check. At the 440 Hz reference six of the eleven pull under the pre-specified rule: four EnCodec conditions, WavTokenizer and SNAC 32k. Three of those EnCodec conditions and WavTokenizer pull by more than five cents and hold under every variance treatment; SNAC 32k and EnCodec on vowels pull by under a cent, and both sit below their own detection floors, so the count is six on trial intervals, five

Table 1: **Which partials move.** Displacement of each decoded partial, cents, EnCodec 24k at 3 kbps, 440 Hz, $\delta_0 = \pm 40$; means over repetitions; H3, at the edge, splits across two lines and reads at their centroid. Other bitrates are in Table A15.

δ_0	H1	H2	H3	H4	H5	H6	H7	H8	Estimator
-40	-0.0	$+0.0$	$+26.9$	$+40.2$	$+51.9$	$+42.1$	$+35.1$	$+44.6$	$+20.7$
$+40$	-0.2	-0.0	-24.0	-37.9	-27.5	-35.1	-42.8	-37.5	-16.9

once run-to-run variance is folded in, and four under the strictest fold (Appendix G.2). The nulls differ in kind. DAC 16k's the step edge explains, since it pulls at 880 and 1760 Hz where partials clear its edge, though the soft form predicts a 440 Hz pull we do not measure; the three censored nulls have no resolvable edge, so the account makes no prediction for them, as for the censored vowel pull; and Mimi's is a non-detection rather than a bounded null (Table A9, Appendix E.2).

Speech and classical codecs. On speech, neither pre-registered FLEURS test supports its prediction (Appendix H), and on a source-filter vowel EnCodec's bias is under half a cent and SpeechTokenizer's covers zero. A learned grid predicts the classical controls will not pull, and they do not. Opus (Valin et al., 2012) and MP3 (Brandenburg, 1999) stay put, and HE-AAC at 32 kbps, its SBR crossover set to EnCodec's 1.32 kHz edge, leaves those partials at $\bar{\ell} = 0.04$ against EnCodec's 0.90 (Appendix D.2).

3.2 Where a decoder regrids, and where its edge sits

The fundamental stays put and the invented partials move (Table 1). For EnCodec the line at f_0 sits within 0.2 cents of the input on the mean over repetitions, at each of three bitrates, with no companion at the 12-TET frequency above -25 dB, and at these registers every partial below the edge stays within 2 cents. Above it, at 3 kbps, partials land 0.95 and 0.90 of the way to the grid harmonic at 220 and 440 Hz, where the edge resolves, and 0.79 at 880 Hz, where the edge falls between attainable partials (Table A13). A vocoder differs: BigVGAN places its partials 0.68 of the way *and* moves the fundamental by 3.1 cents, so a recording-level estimate (Berendes et al., 2025) cannot separate the two.

The edge rises with bitrate. For EnCodec 24 kHz it moves from the second partial at 1.5 kbps to the fifth at 24 kbps, about 0.9 to 2.2 kHz at a 440 Hz reference (Figure A7b), and the bias falls as it rises, from 12.55 cents at 3 kbps to 8.03 at 24 kbps on the 440 Hz tone. The target is close to, not exactly, the ladder of $f^\star$: the $k = 5$ line lands on its chromatic pitch while $k = 7$ follows $7f^\star$; the resulting $+1.6$-cent mean gap between the two targets is arithmetic, present without any codec, and accounts for part of the fine-tuned arm's 6.9-cent attractor (Section 3.3; Appendix C.5).

Where an edge resolves, a ladder sits above it. Seven of the eleven measurable conditions resolve an edge, and all seven place their invented partials most of the way onto a grid: $\bar{\ell}$ runs 0.72 to 0.90. Two rest on one partial, and DAC 16k is read at 880 Hz (Tables A16, A39). $\bar{\ell}$ says where invented partials land, not whether a condition pulls: Mimi does not pull at 440 Hz yet reads 0.75.

What $\bar{\ell}$ also measures. Averaged over every partial at or above the edge, $\bar{\ell}$ mixes how many partials a decoder regenerates with how far it moves each. Over the twenty-six held-out checkpoints of Section 2.5 it correlates with the regenerated count at $\rho = 0.924$. Eight of them share EnCodec's 1.32 kHz edge and span 0.26 to 0.83 in $\bar{\ell}$ while regenerating two to six partials; over the regenerated partials alone, $\bar{\ell}_r$ spans 0.62 to 0.84 (Appendix G.6). The same mixing is why EnCodec 24k's $\bar{\ell}$ falls from 0.90 at 3 kbps to 0.72 at 24 kbps: above a higher edge a larger share of the partials stay unmoved (Table A16). The edge and the ladder fraction are therefore descriptions, not a clean split between code and decoder. The causal contrasts of Section 3.3 avoid the confound, since every fine-tuned EnCodec arm regenerates every partial above one shared 2.0 kHz edge.

Scoring rule. Scoring regeneration on level alone does not condition on where a partial lands. It reads EnCodec at 3 kbps at 0.78 rather than 0.90, the primary ladder fractions at 0.59 to 0.86 rather than 0.72 to 0.90, and over the held-out rows moves four decoders below the 0.70 boundary of the *ladder* tier defined in Section 3.3 (Appendix G.6). We report the placement-conditioned rule, frozen before the held-out pass, with the level-only readings beside it.

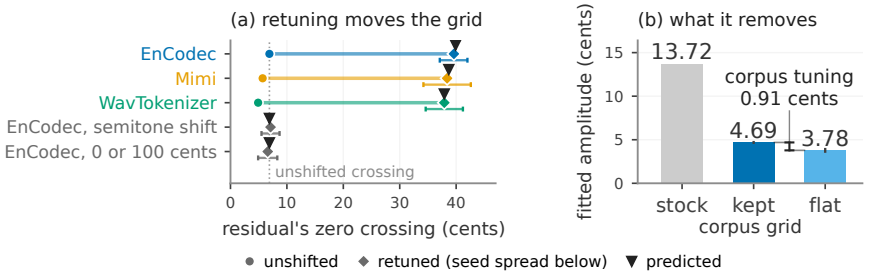

Figure 3: **Retuning the corpus moves the grid, on GTZAN.** (a) Families above, controls below; triangles, predicted crossings; dotted, the unshifted arm's. (b) Amplitude: un-fine-tuned, the magnitude-matched control and the flattened arm; the last two differ by 0.91, 0.87 in $\tilde{b}$; ± 1 sd, tuned arms.

The quantiser account does not explain the effect. Its sawtooth prediction (P1) fails (Appendix A.5), and removing the quantiser at a matched 2.2 kHz edge leaves the 24 kbps condition's all-trial bias within between-run spread, so the effect does not *require* a coarse quantiser; the bypass bounds the account rather than excluding it (Appendix A.1).

3.3 The grid a decoder uses follows its corpus

Every EnCodec corpus arm fine-tunes the one EnCodec decoder for 4000 steps with encoder and quantiser frozen, so all arms read bit-identical tokens and share one 2.0 kHz edge. The arms differ only in the tuning of the GTZAN (Tzanetakis & Cook, 2002) clips: unmodified; resampled by a semitone, or by 0 or 100 cents per clip, a shift matched in magnitude to the flattening and called *magnitude-matched* below, both of which keep the grid; flattened by a uniform offset of 0 to 100 cents; retuned by $+33$ cents; and quantised to 24-TET. Unmodified, GTZAN concentrates near 12-TET, at a within-semitone peak-to-mean of 1.788 against a 1.04 floor (Table A11). Fine-tuning on any of them removes most of the effect, partly because the edge rises from 1.3 to 2.0 kHz (Appendix E.1).

Retuning moves the grid by about the predicted amount. The decoder fine-tuned on unmodified clips has its attractor at 6.9 cents. Retuning the corpus by $+33$ cents moves it to 39.6 [37.1, 42.0] against $6.9 + 33$ predicted, a displacement of 32.7 ± 2.8 cents at seed-level spread (Figure 3a), and its partials follow, at 0.87 of the way to the retuned ladder and 0.11 to 12-TET. On the 24-TET arm, whose corpus peaks at 7.3 times its mean at 12-TET pitches and 3.4 at quarter-tones, a sinusoid of 50-cent period fits the residual with amplitude 3.34 cents at $R^2 = 0.71$, where one of 100-cent period collapses to 0.06, so the 50-cent component dominates the residual (Table A25). Moving the audio without moving its grid does nothing: the two grid-preserving controls keep ladder fractions of 0.88 and 0.86 against 0.89 unmodified. The $+33$ arm reconstructs as well as its control, so a worse decoder cannot explain the shift (Table A24), and the shift transfers to Mimi and WavTokenizer, whose crossings both move about 33 cents on seed-level spread (Appendix E.1).

Flattening removes a third of the ladder. Flattening the corpus, to a peak-to-mean of 1.07 against a 1.04 floor, drops the ladder fraction, over five seeds, to 0.59 against 0.89 unmodified and 0.86 on the magnitude-matched control, the comparator chosen from three once they had run. Over ten independent draws of the offsets, flattening lowers $\tilde{b}$ by 0.87 cents [0.74, 1.00] more than the magnitude-matched control does (Figure 3b). A flattened corpus is also harder to model, but the flattened arm's reconstruction cost, 0.005 on the training objective, has a 95% upper bound at the size of an under-trained arm's deficit that left its ladder in place (Table A24).

A non-Western corpus, and an extender trained from scratch. A decoder fine-tuned on Saraga (Srinivasamurthy et al., 2021) lands 0.81 of the way to that corpus's own histogram of Carnatic scale degrees (svaras) and 0.22 of the way to 12-TET, an arm whose reconstruction deficit is the largest here, so it rests on direction alone (Appendix F.5). A bandwidth extender trained from scratch at EnCodec's 1.32 kHz edge, with no pre-trained corpus behind it, places its partials 0.68 [0.55, 0.81] of the way to the grid on music as recorded, 0.14 [0.00, 0.28] on the same audio flattened and 0.09 on uniformly tuned harmonics, so a generative prior does not default to 12-TET (Appendix D.3).

Held-out and registered passes, as consistency checks. We froze the edge, the ladder fraction and a three-tier rule (*ladder* $\bar{\ell} \geq 0.70$, *part* 0.20 to 0.70, *input's* below 0.20) and scored twenty-six held-out checkpoints, predicted from architecture and bitrate; on the twenty from untouched families the rule is right on 12 against 10 for always predicting *part* (exact McNemar $p = 0.815$). We then registered the rule, that comparator and the sample size before scoring 15 fresh checkpoints, predicting each tier from how many of the probe's eight partials clear its measured edge (6 to 8, *ladder*; 2 to 5, *part*; fewer, *input's*), before reading its ladder fraction: the rule is right on 11 against the comparator's 5 (exact McNemar $p = 0.070$). Always predicting *ladder*, the tier most common in that sample and so knowable only after scoring, is not distinguishable from the rule ($p = 0.375$). Read continuously rather than in tiers, the same rows correlate at $\rho = 0.79$ ($p = 4 \times 10^{-4}$), a post hoc reading that inherits the same count relation; about thirty checkpoints give 85% power at the win rate this pass observed (Appendix G.6).

3.4 Real recordings: size, direction and a remedy

Size. Read frame by frame, EnCodec at 3 kbps gives 2.49 cents [2.05, 2.98] on tuning-flattened GTZAN and 2.63 [2.17, 3.14] on Saraga, both about two fifths of the 5.90 this pipeline reads on the 440 Hz tone. As the account predicts, the bias is 2.68 cents [1.38, 4.30] larger where partials clear the edge on Saraga and 2.74 [0.95, 5.33] on GTZAN, but not under Opus (Appendix F.4). On isolated notes the pitch reading vanishes though the partials still regrid (Appendix F.1; Table A16).

Direction. After one round trip, 22 of 200 Saraga (Carnatic) and 11 of 60 Turkish makam (Uyar et al., 2014) recordings switch to fitting 12-TET at A4 = 440 Hz better than their own corpus's measured degrees placed at each recording's tonic. Against a distortion-matched Opus control the Saraga discordant pairs split 22 to 2 (exact McNemar, $p = 4 \times 10^{-5}$), and an Opus control doing more damage than EnCodec flips 3 of 200 against EnCodec's 22 (Appendix F.4). On Saraga, `librosa.estimate_tuning` (McFee et al., 2015) moves 134 recordings toward 12-TET ($p = 1.7 \times 10^{-6}$), a directional count with no distortion-matched arm.

A downstream task. On 235 recordings, enlarged from sixty, makam identification accuracy falls from 0.81 to 0.75, a paired -0.06 $[-0.11, -0.02]$ (exact McNemar $p = 0.011$, read descriptively: the set was enlarged); on the sixty, the controls lose 0.02 or less (Table A31).

The remedy. Fitting only below the edge takes both corpora from 3.2 and 4.4 cents of added median scale-degree error to 0.2; most of what it removes is generic codec damage, the remainder against matched Opus being 0.4 and 0.7 cents, neither excluding zero (Appendix F.4). pYIN cuts the flip rate to 3%, still adding 0.7 [0.2, 1.2] cents of degree error.

A decoder tuned on Saraga. It nearly removes the Carnatic penalty, 0.3 $[-0.2, 0.8]$ against the 3.2 cents stock coding adds, while making makam, which it never saw, *worse*: 6.6 [5.5, 7.7] against 4.4, flipping 32% against 18%. A pull toward 12-TET would cost both. Its 19 makam flips all move toward Carnatic positions, a directional count with no control arm (Appendix F.5.1). The Carnatic half carries leakage, the makam half out-of-domain degradation, and two thirds of the Carnatic repair is generic fine-tuning (Appendix F.5).

4 Conclusion

In every pulling condition that resolves a band edge, mostly EnCodec operating points, the decoder invents the partials above it near the harmonics of the nearest 12-TET pitch, and a full-band fit over them reports a displacement no partial underwent. Retuning the corpus by 33 cents moves the grid by 32.7. On real recordings it shifts tuning analysis in one direction, reduced by fitting nothing above the edge. We make no audibility claim. Six limits bear on these numbers. The band-edge account was formed after the pre-registered one, and its registered pass on fifteen fresh checkpoints reaches $p = 0.070$, and 0.375 against the most common tier. The comparison tone sits at its population's 92nd percentile, so the effect size we quote is the median, 4.37 cents. Two of the six pull below their own detection floors. The isolable corpus component in a trained decoder is a lower bound. Downstream, the grid-attributable degree error does not exclude zero, and our fit weights the band above the edge more than a salience sum does (Table A33). Table A38 lists the rest.

ETHICS STATEMENT

This work characterises a decoder-side artefact of neural audio codecs and uses only existing public corpora, with no human subjects and no new data collection. Every corpus and checkpoint is used under its own licence and for research only; we redistribute no corpus audio, and the fine-tuned weights we release carry the licence of the checkpoint they derive from. GTZAN carries documented faults (Sturm, 2013), which Appendix B.4 argues do not bear on the pooled within-semitone histogram we take from it. Two further points bear stating. First, naming specific codecs could be read as disparaging particular released systems; our claim concerns a training-distribution effect whose size differs across codecs, and the paper reports where each codec's band edge sits and what it regenerates above it, not the competence of any research group. Second, we analyse musical traditions that are not our own, on published corpora and descriptions (Uyar et al., 2014; Bozkurt et al., 2014; Srinivasamurthy et al., 2014; 2021). The degree positions we score against are not taken from those traditions' theory tables: they are estimated from each corpus's own recordings, because the tables differ from measured practice by about the size of the effect (Appendix J). For makam a theory-table reading is also reported beside the measured one. The degree error is thus a displacement from measured practice rather than from a tradition's theoretical account. All statements concern measured deviations in cents, and the Western and Carnatic readings differ by 0.13 cents $[-0.5, 0.8]$ (Appendix F.5.3). Whether a displacement of this size inside one band of a complex tone is audible is a question the mistuned-partial and interval-discrimination literatures do not settle, as Appendix E.1 sets out.

REPRODUCIBILITY STATEMENT

The pitch experiments require no dataset: stimuli come from a short deterministic script, so the central result reproduces from the released code and publicly available checkpoints alone. The supplementary material contains the anonymised source code: the stimulus generator, the modules behind every summary statistic in the paper, and a reproduction guide mapping each table and figure to the module that produces it and each script this paper cites by name to the module that implements it. Several of those modules recompute printed values from the paper's own columns, so a number can be checked without running a codec. All 146 raw result files (117 runs, of which nine are a cross-instance replication of the causal experiment, plus 29 files that carry no run sidecar), their sidecars, the scripts' output and the remaining figure scripts will be released with the public code after review, as will the weights of every fine-tuned decoder, EnCodec's, DAC's, Mimi's and WavTokenizer's, and of every bandwidth extender. The WavTokenizer and BigVGAN sweeps are counted separately from those files. Five further results are counted separately: the `librosa` exposure count of Section 3.4, the makam identification rerun at $n = 235$, each of the two arms of the Saraga-tuned decoder in Table A31, and the registered pass of Appendix G.6. Their scripts are in the supplementary material with the rest. Every registration and sweep number uses one exclusion rule, a tracking gate at 200 cents. The isolated-note analysis adds a usability rule, a 60-cent refinement window that removes 3.1% of notes, and the real-music protocol adds the estimator's cross-check, which retains 23 to 40% of frames (Appendix F.2). The gate removes 12 of 53,020 trials of the headline run and under 10% of trials on all but eight reported runs: DAC 16k at 65% and at about two thirds on its 880 and 1760 Hz register runs, Opus at 47% and 18% on the registration sweeps and 40% and 66% on its octave sweep, and the refused SNAC 24k at 93%; Appendices B.7 and D.1 list every run, and the DAC 16k registration slope is 0.98 [0.86, 1.10] with the gate removed. On the sweeps the estimator cross-check is reported only as a robustness variant. Two guards that refuse to fit a registration slope are stated in Appendix B.8, which shows the reported set of conditions is unchanged over a window around their thresholds. The supplementary material, stripped of version-control metadata, includes the pre-registration document with its amendment log. The pitch sweeps are CPU-bound, at about four trials per second per core for EnCodec; fine-tuning, recognition and corpus passes ran on a single-GPU cloud instance and team machines over roughly two weeks.

Each of 117 sweep, corpus, fine-tuning and generation result files carries a sidecar recording the full argument set and the package versions of the run that produced it; the supplementary material defines that record and checks it against the two configurations below. Every checkpoint used in the primary analysis is listed there with its repository and operating point, and the three used at a single point each, the EnCodec 24k rate ladder, EnCodec 32k and the BigVGAN comparison checkpoint, with the operating point at which each is used. The revisions to which the twenty-six held-out and

fifteen registered checkpoints were pinned are released with the code. Of the 108 sidecars outside the replication directory, 104 record one configuration (torch 2.8.0 with CUDA 12.8, transformers 5.16.1) and four are the accumulation-test sweeps, which ran on CPU under transformers 4.57.6 and whose first round trip reproduces that configuration's all-trial bias on the 3 kbps headline sweep, 8.88 cents, to 0.01, though its off-grid bias and amplitude sit outside the five-run spread for a design reason Appendix F.6 gives. The nine replication sidecars all record the standard configuration. The remaining 29 result files are the recognition, phonology and histogram outputs, four retuning summaries, the codebook probe and two spectral summary tables. Their commands are in the reproduction guide in the supplementary material. The spectral sweep, the DAC fine-tuning, the MusicGen decoder swap and their analyses are released on the same terms, with the scripts that produced them (`spectral_sweep.py`, `finetune_dac.py`, `musicgen_swap.py`), which the reproduction guide maps into the supplementary material, and their sidecars; the band edge, the ladder fraction and every number of Tables A16 and A26, and the decoder-swap readings of Appendix F.7, regenerate from their result files.

AI USE STATEMENT

We used a generative AI assistant substantially in this work, and describe that use precisely. It contributed to research ideation, including the framing of the codec as a measurement instrument and the choice to pair synthetic stimuli with corpus validation. It assisted in drafting the text of this paper throughout, in writing analysis and plotting code, and in searching for and summarising related work. It contributed to the stimulus generator of Appendix B.4, which produces the synthetic tones every registration and sweep number is measured on; we checked its output against the intended frequencies and detunings before any of it was used. The propositions and derivations of Appendices A and C.4 are ours, and the assistant drafted the prose that presents them.

We have reviewed all AI-assisted work and take full responsibility for the final content. Every experimental number is produced by code we ran and inspected, from result files we generated; no reported value originates with the assistant. Because this paper's own argument concerns researcher degrees of freedom, we state the boundary explicitly: the eighteen analytic choices of Table A37 were made by us, as was the band-edge account that replaced the failed pre-registered one; the assistant drafted prose describing several of those choices, but every threshold, boundary and inclusion rule was set and applied by us. All 99 bibliography entries carry a provenance comment recording the arXiv identifier, DOI, ISBN or item-level record checked. For the five estimator references behind Table A32 two separate checks were made: the parameter defaults were read off the primary documentation and source, and the bibliographic records were checked against a publisher index. Two of those five, the ATMM proceedings paper and the tutorial webbook, are at venues that issue no DOI, and the comment on each says what was checked instead. The checked entries include every load-bearing one, among them Terhardt et al. (1982), Nagel & Disch (2009), Pons et al. (2021), Berendes et al. (2026) for a whole column of the boundary table in Appendix I, Lee et al. (2023) for the vocoder arm, and Engel et al. (2017) for the NSynth notes behind the isolated-note null, whose author order we take from the ICML proceedings and not from the preprint, which differs. One entry needs a word about its venue. Dietz et al. (2002), cited for spectral band replication, has its title and four authors from an item-level record, the AES-side record obtainable being a convention template with placeholder fields; that item-level record normalises the venue onto the AES journal, and we do not follow it, since the work is Convention Paper 5553 of the 112th Convention and has no journal version. Citations proposed by the assistant that could not be verified were removed rather than retained. The claims, the pre-specified analysis plan and its logged amendments, and the limitations stated in the abstract and at each result are our own, and we can defend each without recourse to the assistant. The novelty of the central claim was checked by a literature survey that did not rely on the assistant's search; the closest prior work, on tuning bias in neural vocoders (Berendes et al., 2025), is discussed in Section 1.2.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Yang Ai, Xiao-Hang Jiang, Ye-Xin Lu, Hui-Peng Du, and Zhen-Hua Ling. APCodec: A neural audio codec with parallel amplitude and phase spectrum encoding and decoding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Kemal Altwlkany, Amar Kuric, and Emanuel Lacic. On the language and gender biases in PSTN, VoIP and neural audio codecs. In *Proceedings of Interspeech*, pp. 1348–1352, 2025.

Hasan Sercan Atlı, Burak Uyar, Sertan Şentürk, Barış Bozkurt, and Xavier Serra. Audio feature extraction for exploring Turkish makam music. In *Proceedings of the 3rd International Conference on Audio Technologies for Music and Media (ATMM 2014)*, pp. 142–153, Ankara, Turkey, 2014.

W. R. Bennett. Spectra of quantized signals. *Bell System Technical Journal*, 27(3):446–472, 1948.

Hans-Ulrich Berendes, Ben Maman, and Meinard Müller. Tuning matters: Analyzing musical tuning bias in neural vocoders. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 166–173, 2025.

Hans-Ulrich Berendes, Ben Maman, and Meinard Müller. Fine-tuning BigVGAN-V2 for robust musical tuning preservation. In *ICASSP*, 2026.

Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R. Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 493–498, Curitiba, Brazil, 2013.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023.

Barış Bozkurt, Ruhi Ayangil, and André Holzapfel. Computational analysis of Turkish makam music: Review of state-of-the-art and challenges. *Journal of New Music Research*, 43(1):3–23, 2014.

Karlheinz Brandenburg. MP3 and AAC explained. In *Proceedings of the AES 17th International Conference: High-Quality Audio Coding*, 1999.

Woongjib Choi, Byeong Hyeon Kim, Hyungseob Lim, Inseon Jang, and Hong-Goo Kang. Neural spectral band generation for audio coding. In *Interspeech 2025*, 2025.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2022.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

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.

Bruno Di Giorgi, Mark Levy, and Richard Sharp. Mel spectrogram inversion with stable pitch. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

Martin Dietz, Lars Liljeryd, Kristofer Kjörling, and Oliver Kunz. Spectral band replication, a novel approach in audio coding. In *Audio Engineering Society 112th Convention*, 2002. Convention Paper 5553.

Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. FunCodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 591–595, 2024.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with WaveNet autoencoders. In *International Conference on Machine Learning (ICML)*, 2017.

Neville H. Fletcher and Thomas D. Rossing. *The Physics of Musical Instruments*. Springer, 2nd edition, 1998.

Allen Gersho. Asymptotically optimal block quantization. *IEEE Transactions on Information Theory*, 25(4):373–380, 1979.

Beat Gfeller, Christian Frank, Dominik Roblek, Matt Sharifi, Marco Tagliasacchi, and Mihajlo Velimirović. SPICE: Self-supervised pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1118–1128, 2020.

Alex Gichamba and Moise Busogi. Probing low frame rate degradation in neural audio codecs. *arXiv preprint arXiv:2606.16969*, 2026.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998.

Yiwei Guo, Zhihan Li, Chenpeng Du, Hankun Wang, Xie Chen, and Kai Yu. LSCodec: Low-bitrate and speaker-decoupled discrete speech codec. In *Interspeech 2025*, 2025.

William M. Hartmann, Stephen McAdams, and Bennett K. Smith. Hearing a mistuned harmonic in an otherwise periodic complex tone. *The Journal of the Acoustical Society of America*, 88(4): 1712–1724, 1990. doi: 10.1121/1.400246.

Bruce Haynes. *A History of Performing Pitch: The Story of "A"*. Scarecrow Press, 2002.

Shengpeng Ji, Minghui Fang, Jialong Zuo, Ziyue Jiang, Dingdong Wang, Hanting Wang, Hai Huang, and Zhou Zhao. Language-Codec: Bridging discrete codec representations and speech language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025a.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. WavTokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling. In *International Conference on Learning Representations (ICLR)*, 2025b.

Harry Julian, Rachel Beeson, Lohith Konathala, Johanna Ulin, and Jiameng Gao. Finite scalar quantization enables redundant and transmission-robust neural audio compression at low bit-rates. *arXiv preprint arXiv:2509.09550*, 2025.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A convolutional representation for pitch estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, 2018.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Volodymyr Kuleshov, S. Zayd Enam, and Stefano Ermon. Audio super-resolution using neural networks. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. BigVGAN: A universal neural vocoder with large-scale training. In *International Conference on Learning Representations (ICLR)*, 2023.

Hanzhao Li, Liumeng Xue, Haohan Guo, Xinfa Zhu, Yuanjun Lv, Lei Xie, Yunlin Chen, Hao Yin, and Zhifei Li. Single-Codec: Single-codebook speech codec towards high-performance speech generation. In *Interspeech*, 2024.

Yunpeng Li, Kehang Han, Brian McWilliams, Zalán Borsos, and Marco Tagliasacchi. SpectroStream: A versatile neural codec for general audio. *arXiv preprint arXiv:2508.05207*, 2025.

Haohe Liu, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D. Plumbley. AudioSR: Versatile audio super-resolution at scale. *arXiv preprint arXiv:2309.07314*, 2023.

Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D. Plumbley. SemantiCodec: An ultra low bitrate semantic audio codec for general sound. *IEEE Journal of Selected Topics in Signal Processing*, 18(8):1448–1461, 2024.

Matthias Mauch and Simon Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 659–663, 2014.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in Python. In Kathryn Huff and James Bergstra (eds.), *Proceedings of the 14th Python in Science Conference*, volume 2015, pp. 18–24, 2015. doi: 10.25080/majora-7b98e3ed-003.

Atharva Mehta, Shivam Chauhan, Amirbek Djanibekov, Atharva Kulkarni, Gus Xia, and Monojit Choudhury. Music for all: Representational bias and cross-cultural adaptability of music generation models. In *Findings of the Association for Computational Linguistics: NAACL*, pp. 4569–4585, 2025.

Brian C. J. Moore, Brian R. Glasberg, and Robert W. Peters. Relative dominance of individual partials in determining the pitch of complex tones. *The Journal of the Acoustical Society of America*, 77(5):1853–1860, 1985. doi: 10.1121/1.391936.

Brian C. J. Moore, Brian R. Glasberg, and Robert W. Peters. Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *The Journal of the Acoustical Society of America*, 80(2):479–483, 1986.

Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. Chunked autoregressive GAN for conditional waveform synthesis. In *International Conference on Learning Representations (ICLR)*, 2022.

Pooneh Mousavi, Jarod Duret, Darius Petermann, Artem Ploujnikov, Luca Della Libera, Anastasia Kuznetsova, Cem Subakan, and Mirco Ravanelli. DASB – discrete audio and speech benchmark. *arXiv preprint arXiv:2406.14294*, 2024.

Frederik Nagel and Sascha Disch. A harmonic bandwidth extension method for audio codecs. In *ICASSP*, pp. 145–148, 2009. doi: 10.1109/ICASSP.2009.4959541.

Thomas Nuttall, Genís Plaja-Roglans, Lara Pearson, Brindha Manickavasakan, Kaustuv Kanti Ganguli, Ajay Srinivasamurthy, and Xavier Serra. *Computational Methods for Supporting Corpus-Based Research on Indian Art Music*. Music Technology Group, Barcelona, Spain, 2022. URL https://mtg.github.io/IAM-tutorial-ismir22.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. doi: 10.23915/distill.00003.

Patrick O'Reilly, Prem Seetharaman, Jiaqi Su, Zeyu Jin, and Bryan Pardo. Code drift: Towards idempotent neural audio codecs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Nathan Paek, Yongyi Zang, Qihui Yang, and Randal Leistikow. Learning interpretable features in audio latent spaces via sparse autoencoders. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

Julian D. Parker, Anton Smirnov, Jordi Pons, C. J. Carr, Zack Zukowski, Zach Evans, and Xubo Liu. Scaling transformers for low-bitrate high-quality speech coding. In *International Conference on Learning Representations (ICLR)*, 2025.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jordi Pons, Santiago Pascual, Giulio Cengarle, and Joan Serrà. Upsampling artifacts in neural audio synthesis. In *ICASSP*, pp. 3005–3009, 2021.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Yi Qin and Alexander Lerch. Tuning frequency dependency in music classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 401–405, 2019. doi: 10.1109/ICASSP.2019.8683340.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, 2023.

Yong Ren, Tao Wang, Jiangyan Yi, Le Xu, Jianhua Tao, Chuyuan Zhang, and Junzuo Zhou. Fewer-token neural speech codec with time-invariant codes. In *ICASSP*, 2024.

Roelof J. Ritsma. Frequencies dominant in the perception of the pitch of complex sounds. *The Journal of the Acoustical Society of America*, 42(1):191–198, 1967. doi: 10.1121/1.1910550.

Francisco Rodríguez-Algarra, Bob L. Sturm, and Simon Dixon. Characterising confounding effects in music classification experiments through interventions. *Transactions of the International Society for Music Information Retrieval*, 2(1):52–70, 2019. doi: 10.5334/tismir.24.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6): 1759–1770, 2012. doi: 10.1109/TASL.2012.2188515.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Xavier Serra. A multicultural approach in music information research. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 151–156, 2011.

Jiatong Shi, Jinchuan Tian, Yihan Wu, Jee-weon Jung, Jia Qi Yip, Yoshiki Masuyama, William Chen, Yuning Wu, Yuxun Tang, Massa Baali, Dareen Alharhi, Dong Zhang, Ruifan Deng, Tejes Srivastava, Haibin Wu, Alexander H. Liu, Bhiksha Raj, Qin Jin, Ruihua Song, and Shinji Watanabe. ESPnet-Codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Hubert Siuzdak, Florian Grötschla, and Luca A. Lanzendörfer. SNAC: Multi-scale neural audio codec. In *NeurIPS 2024 Workshop on Audio Imagination*, 2024.

Joren Six, Olmo Cornelis, and Marc Leman. Tarsos, a modular platform for precise pitch analysis of Western and non-Western music. *Journal of New Music Research*, 42(2):113–129, 2013. doi: 10.1080/09298215.2013.797999.

Ajay Srinivasamurthy, Gopala Krishna Koduri, Sankalp Gulati, Vignesh Ishwar, and Xavier Serra. Corpora for music information research in Indian art music. In *Proceedings of the Joint International Computer Music Conference and Sound and Music Computing Conference (ICMC/SMC)*, 2014.

Ajay Srinivasamurthy, Sankalp Gulati, Rafael Caro Repetto, and Xavier Serra. Saraga: Open datasets for research on Indian art music. *Empirical Musicology Review*, 16(1):85–98, 2021.

Bob L. Sturm. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

Ernst Terhardt, Gerhard Stoll, and Manfred Seewann. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *The Journal of the Acoustical Society of America*, 71(3): 679–688, 1982.

Joseph Turian and Max Henry. I'm sorry for your loss: Spectrally-based audio distances are bad at pitch. In *NeurIPS 2020 Workshop "I Can't Believe It's Not Better!"*, 2020.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

Burak Uyar, Hasan Sercan Atlı, Sertan Şentürk, Barış Bozkurt, and Xavier Serra. A corpus for computational research of Turkish makam music. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology (DLfM)*, pp. 1–7, 2014.

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

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Shih-Heng Wang, Jiatong Shi, Jinchuan Tian, Haibin Wu, and Shinji Watanabe. Do neural codecs generalize? A controlled study across unseen languages and non-speech tasks. *arXiv preprint arXiv:2601.12205*, 2026.

Craig C. Wier, Walt Jesteadt, and David M. Green. Frequency discrimination as a function of frequency and sensation level. *The Journal of the Acoustical Society of America*, 61(1):178–184, 1977.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-Wei Chang, Jiawei Du, Ke-Han Lu, Alexander H. Liu, Ho-Lam Chung, Yuan-Kuei Wu, Dongchao Yang, Songxiang Liu, Yi-Chiao Wu, Xu Tan, James Glass, Shinji Watanabe, and Hung-yi Lee. Codec-SUPERB @ SLT 2024: A lightweight benchmark for neural audio codec models. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024a.

Haibin Wu, Ho-Lam Chung, Yi-Cheng Lin, Yuan-Kuei Wu, Xuanjun Chen, Yu-Chi Pai, Hsiu-Hsuan Wang, Kai-Wei Chang, Alexander H. Liu, and Hung-yi Lee. Codec-SUPERB: An in-depth analysis of sound codec models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10330–10348, 2024b.

Yi-Chiao Wu, Israel D. Gebru, Dejan Marković, and Alexander Richard. AudioDec: An open-source streaming high-fidelity neural audio codec. In *ICASSP*, 2023.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. BigCodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377*, 2024.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Wei Tan, Rongzhi Gu, Shun Lei, Zhiwei Lin, and Zhiyong Wu. MuCodec: Ultra low-bitrate music codec. *arXiv preprint arXiv:2409.13216*, 2024.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. HiFi-Codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Zhou Zhao, Xixin Wu, and Helen Meng. UniAudio: Towards universal audio generation with large language models. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56422–56447, 2024.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25697–25705, 2025a.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025b.

Paul L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, 1982.

Jean Mary Zarate, Caroline R. Ritson, and David Poeppel. Pitch-interval discrimination and musical expertise: Is the semitone a perceptual boundary? *The Journal of the Acoustical Society of America*, 132(2):984–993, 2012.

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

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Youqiang Zheng, Weiping Tu, Yueteng Kang, Jie Chen, Yike Zhang, Li Xiao, Yuhong Yang, and Long Ma. FreeCodec: A disentangled neural speech codec with fewer tokens. In *Interspeech 2025*, 2025.

Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

APPENDIX CONTENTS

A. The pre-registered account, and why it fails. (p. 18) A.1 quantiser notation; A.2 the quantiser account: assumptions and proposition; A.3 the full derivation; A.4 three threats to the density account; A.5 the quantiser account, prediction by prediction; A.6 the (P3) rate test in bits per dimension.

B. Experimental details. (p. 23) B.1 full notation, and the three statistics; B.2 codec checkpoints and operating points; B.3 the registration test in full; B.4 stimulus generator; B.5 estimator implementation and noise floor; B.6 resampling unit and interval construction; B.7 exclusion rules and refused conditions; B.8 the refused conditions in full.

C. The measurement in full. (p. 28) C.1 registration for every condition, the codebook probe and the corpora; C.2 the octave test, the residual's shape and the corpora; C.3 registration at other octaves; C.4 the octave sweep and the edge as a transition; C.5 which partials move; C.6 the uniform-weight reference; C.7 the headline effect over a population of stimuli.

D. Controls. (p. 38) D.1 the classical-codec control; D.2 a classical scheme that regenerates: HE-AAC with SBR; D.3 a learned bandwidth extender, trained three ways.

E. The causal experiment. (p. 40) E.1 the causal experiment in full; E.2 what the DAC causal arm could have detected.

F. Off synthetic tones: real music and downstream. (p. 48) F.1 the effect off synthetic tones; F.2 the pull on real recordings, in full; F.3 the synthetic-to-real ratio, within one instrument; F.4 what a round trip costs a tuning analysis; F.5 tuning analysis through a codec; F.6 accumulation under repeated coding; F.7 downstream generation through a codec-token model.

G. Statistical discipline. (p. 61) G.1 which comparisons were pre-registered; G.2 between-run variance, and which intervals it widens; G.3 what each interval covers; G.4 every decision made after seeing results; G.5 every limit, in one place; G.6 the held-out pass on twenty-six checkpoints, and the registered pass.

H. The speech tests in full. (p. 72) H.1 language groups; H.2 phonological contrasts in full; H.3 downstream recognition in full.

I. What is ours and what is not. (p. 73) The boundary against the closest prior work, claim by claim.

J. Negative results and a replaced validation. (p. 74)

Where to find each claim. The band edge and ladder fraction are Tables A13 and A16; the registration verdicts Table A9; the causal arms Tables A22, A25 and A27; the exclusion accounting Tables A7 and A8; the downstream cost Table A31; the pre-registration boundary Appendix G.1.

A THE PRE-REGISTERED ACCOUNT, AND WHY IT FAILS

A.1 QUANTISER NOTATION

A neural audio codec is a triple $(\mathcal{F}, \mathcal{Q}, \psi)$ with encoder $\mathcal{F} : \mathbb{R}^L \to \mathbb{R}^{N \times D}$ mapping a waveform to N latent frames of dimension D. The quantiser is residual and multi-stage: given codebooks $\mathcal{C}_q = \{e_q^1, \ldots, e_q^K\} \subset \mathbb{R}^D$ for $q = 1, \ldots, S$, it sets $u_0 = z$ and

$$k_q = \arg\min_k \|u_{q-1} - e_q^k\|_2, \qquad u_q = u_{q-1} - e_q^{k_q}, \qquad \hat{z} = \sum_{q=1}^{S} e_q^{k_q}. \tag{5}$$

The decoder returns $C(x) = \psi(\hat{z})$ at a bit rate of $R = S \log_2 K$ bits per frame (Equation 5). Encoder, codebooks and decoder are all learned from a training corpus; the hypothesis of Section 2.6 located the corpus's tuning statistics in the codebooks, and Sections 3.2 and 3.3 locate them elsewhere.

The bypass comparison of Section 3.2, 5.66 against 5.73, is a gap comparable to the 0.12-cent between-run spread Table A35 measures on the 3 kbps off-grid bias, but it is stated on the all-

trial bias rather than on $\tilde{b}$, our declared effect size, and at that edge the ladder fractions of the two conditions disagree (0.84/0.71 against 0.86/0.58). We report it as a bound for both reasons.

High-rate quantisation theory bounds the *centroid* part of such a residual near a cent for a grid-peaked density, though not the cell-offset part, which Appendix A.3 shows is an order larger. The four predictions in full:

(P1) the residual is a sawtooth in δ, or its fundamental sinusoid;
(P2) its off-grid bias is positive, toward the grid;
(P3) that bias falls with exponent -1 or -2 in bits per latent dimension;
(P4) it registers to absolute pitch, which is what would make the grid *learned*.

EnCodec runs at 0.16 to 2.5 bits per dimension, outside the regime the theory assumes, so (P1), (P2) and (P4) are tests and (P3) is a remark (Appendix A.4).

A.2 The quantiser account: assumptions and proposition

We assume **(A1)** a coordinate θ along which the quantiser acts approximately as a scalar quantiser of constant cell width Δ, an idealisation since trained quantisers have density-dependent cell widths, with reconstruction levels $\hat{\theta}(\theta)$; **(A2)** a training density $p(\theta)$ that is positive and four times differentiable, with Δ small relative to its scale of variation (four derivatives, not three, because the remainder in Equation 9 is $O(\Delta^4)$); **(A3)** reconstruction within a cell equal to the conditional mean, that is centroid reconstruction *on a uniform partition*. (A3) is not the Lloyd optimum: a Lloyd-optimal quantiser for non-uniform p has Panter-Dite cell widths $\Delta(\theta) \propto p(\theta)^{-1/3}$ and so contradicts (A1). We take uniform cells because that is what a residual vector quantiser with a fixed codebook approximates. Panter-Dite spacing would give the cell-offset term its own density-registered structure, a variant of this account that our evidence does not separately address.

Proposition 1 (Decomposition of the residual)**.** *Under* (A1)–(A3)*, writing $c(\theta)$ for the centre of the cell containing θ and r_Q for the residual of the idealised quantiser (here θ is the quantiser's own scalar coordinate, not the musical interval of Section 2; the measured r_C is a difference of two absolute-pitch residuals, so the reference term is absorbed by the per-condition intercept),*

$$r_Q(\theta) \;=\; \underbrace{c(\theta)-\theta}_{\text{cell offset, } O(\Delta)} \;+\; \underbrace{\frac{\Delta^2}{12}\,(\log p)'\big(c(\theta)\big)}_{\text{centroid shift, } O(\Delta^2)} + O(\Delta^3). \tag{6}$$

Under (A3) the within-cell mean of the cell-offset term is zero, so a surviving contraction requires levels displaced toward the grid within their cells, which (A1)–(A3) do not supply; we use the contraction only as a descriptive parametrisation that defines λ,

$$r_C(\theta) \;\approx\; -\lambda\,\delta(\theta) \;=\; \lambda\,\big(g(\theta)-\theta\big), \qquad \lambda \in [0,1], \tag{7}$$

a sawtooth of amplitude 50λ cents, with $\lambda \in [0,1]$ an assumption of this parametrisation rather than a property. With R/D bits per dimension the cell width scales as $2^{-R/D}$, so $\log_2 b = -m\,R/D + \text{const}$ with $m = 1$ if a cell-offset term survives and $m = 2$ under the centroid term.

A.3 Full derivation of Proposition 1

Under (A1)–(A3), fix the cell $S_c = [c - \Delta/2,\, c + \Delta/2]$ containing θ. By (A3) the reconstruction level is

$$\hat{\theta}(\theta) \;=\; \hat{\theta}(c) \;=\; \frac{\int_{S_c} u\,p(u)\,du}{\int_{S_c} p(u)\,du}. \tag{8}$$

Substituting $u = c + v$ and expanding p about c to second order by (A2), the terms odd in v vanish over the symmetric cell, and the second-order terms of numerator and denominator cancel when the quotient is expanded, leaving

$$\hat{\theta}(c) - c \;=\; \frac{\Delta^2}{12}\,\frac{p'(c)}{p(c)} + O(\Delta^4) \;=\; \frac{\Delta^2}{12}\,(\log p)'(c) + O(\Delta^4), \tag{9}$$

so the remainder in Proposition 1, stated as $O(\Delta^3)$, is the weaker statement. This is the classical high-rate centroid shift (Gersho, 1979; Zador, 1982; Gray & Neuhoff, 1998); Bennett (1948) gives the noise-power integral from which the high-rate theory descends, not this expression.

A.3.1 WHY THE RESIDUAL IS NOT EQUATION 9.

Equation 9 gives the offset of the reconstruction level from its own cell *centre*. The quantity we measure is the offset from the *input*. Writing $Q(\theta) = \hat{\theta}(c(\theta))$ for the idealised quantiser's reconstruction of θ, and $r_Q(\theta) = Q(\theta) - \theta$ for its residual (the idealised counterpart of the measured r_C), these differ:

$$r_Q(\theta) \;=\; \big[\hat{\theta}(c) - c\big] + \big[c - \theta\big]. \tag{10}$$

The second term is the ordinary quantisation sawtooth, of amplitude up to $\Delta/2$ and therefore $O(\Delta)$, an order larger than Equation 9. Retaining only the centroid shift would discard the dominant term, which is why Proposition 1 carries both, and why the bound of Equation 11 cannot by itself account for an effect of the size we measure.

A.3.2 THE PERIODIC CASE.

With $p \propto \exp(\kappa \cos(2\pi(\theta - g)/100))$ we have $(\log p)'(\theta) = -\kappa(2\pi/100)\sin(2\pi(\theta - g)/100)$, so the centroid term is a sinusoid of period exactly 100 cents, negative immediately above each grid point and positive immediately below the next. Under the dither-averaged picture of the next paragraph the cell-offset term is a sawtooth with the same period, the same zero crossings and the same sign pattern whenever levels register with the grid. Both therefore satisfy (P1)'s period and (P2)'s sign, and only the *waveform* and the *rate exponent* distinguish them, which is why (P1) fits both shapes and (P3) reports a slope rather than merely a decay.

A.3.3 WHY THE CELL-OFFSET TERM IS CONTRACTED RATHER THAN EXACT.

A scalar nearest-level map has a residual that returns to zero at every level, so its period is the level spacing. A period of exactly one semitone would then require exactly one level per semitone, and would force $r_C = g(\theta) - \theta$ with amplitude 50 cents in every codec, which neither varies with rate nor distinguishes one codec from another. Three things separate the measurement from that idealisation: levels register with the grid in density only, so a level near a grid point is not at it; pitch is not an isolable latent coordinate, what we recover being an F_0 estimate from a waveform that depends on every quantised coordinate at once; and onset phase and amplitude jitter enter each cell from a different waveform on every repetition, so a cell's median is an average over assignments rather than a single one. Averaging a grid-registered level density under that dither smooths the staircase into the contraction of Equation 7, $r_C \approx \lambda(g(\theta) - \theta)$, retaining the period, the zero crossings and the sign while replacing the fixed 50-cent amplitude with 50λ. Heuristically, the mean pull across a dithered cell is proportional to its width to leading order, so $\lambda \propto \Delta$ and (P3)'s slope of -1 is unchanged; this is a heuristic rather than a derivation from a stated dither model, and (P3) is inconclusive in any case. Equation 7 gives $b(\theta) = \lambda\,|g(\theta) - \theta|$, so the median bias over off-grid cells recovers $\lambda = \tilde{b}/40$, dimensionless and therefore comparable across codecs and rates in a way a bias in cents is not. For EnCodec at 3 kbps that is $12.55/40 = 0.31$. The recovery presupposes the unsaturated sawtooth, and the measured profile is not one: reading the same λ off the clipped fit of Appendix A.5, whose knee sits at $|\delta| = 22$, gives $12.55/22 = 0.57$. The two disagree by a factor of 1.8, which is a further reading of the same failure (P1) records — the account's single free parameter is not identified, because the measured profile does not have the shape it assumes. We therefore neither fit nor report λ outside this subsection.

A.3.4 MAGNITUDE BOUND.

The peaks of the grid-peaked density p have Gaussian width $w_p = 100/(2\pi\sqrt{\kappa})$, written w_p rather than σ because σ is the registration slope everywhere else in this paper. Writing $\Delta = \alpha w_p$, the centroid amplitude is

$$A_{\text{centroid}} \;=\; \frac{\Delta^2}{12} \cdot \frac{2\pi\kappa}{100} \;=\; \frac{\alpha^2}{12} \cdot \frac{100}{2\pi} \;\approx\; 1.33\,\alpha^2 \text{ cents}, \tag{11}$$

independent of the concentration κ, which cancels. Table A1 tabulates this. The consequence is that any grid bias materially above a few cents is attributable to the cell-offset term, not to the density gradient.

Table A1: Centroid-shift amplitude against cell width, from Equation 11. $\alpha = \Delta/w_p$ is cell width in units of the density peak width; (A2) requires $\alpha \lesssim 1$. The bound is independent of the concentration κ.

$\alpha = \Delta/w_p$	0.5	1.0	2.0	3.0	4.0
A_{centroid} (c)	0.33	1.33	5.31	11.94	21.22
(A2) valid?	yes	marginal	no	no	no

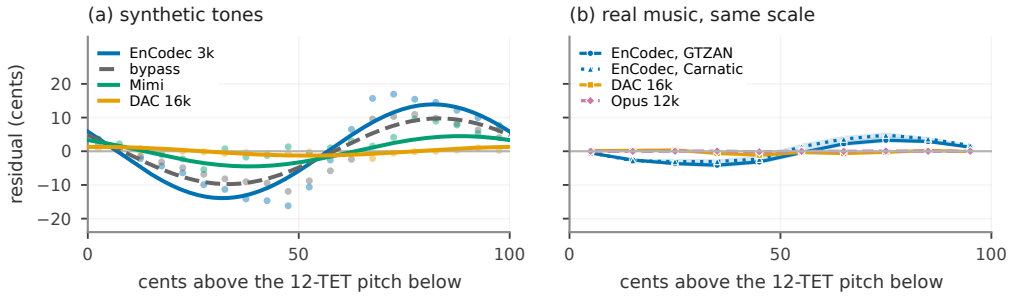

Figure A1: **The residual on synthetic tones and on real music, on one scale.** (a) Binned median residual over a semitone at 440 Hz with its fit. (b) The same on whole clips, bands the clustered interval per bin. Different pipelines: the shared axis compares shape, not size.

A.4 Three threats to the density account

The first is the regime. Proposition 1 is high-rate theory: it assumes cells narrow relative to the variation of p, scalar along θ, with reconstruction at the conditional mean. EnCodec's operating points span 0.16 to 2.5 bits per dimension in a 128-dimensional residual quantiser, and Table A1 shows the expansion already invalid at $\alpha \geq 2$. So the account is being tested outside the regime it was derived in, and neither its failure on (P1) nor its inconclusive (P3) is by itself decisive. What carries the argument against it is empirical and regime-free: removing the quantiser leaves the bias unchanged at a matched band edge (Appendix A.5 sizes it), the codebook shows no grid structure (Table A10), and a quantiser that rounded pitch would move the fundamental, which does not move.

The second is aggregation. The mechanism needs p grid-peaked in absolute F_0 across the training corpus, and aggregation works against that: concert pitch ranges from about 415 to 443 Hz (Haynes, 2002), recordings drift, and much training audio is unpitched or continuously pitched. Pooling smears the peaks and drives κ down. This is a reason not to rest the argument on the centroid term, whose amplitude is independent of κ at fixed $\alpha = \Delta/w_p$ but grows with κ at fixed physical Δ, so its size turns on a parametrisation the measurements do not fix, whereas the coarse-cell term depends only on where levels sit. Under the regeneration account of Section 3.2 the density enters through the decoder's prior over the spectrum it invents rather than through κ, so the second threat weakens: a decoder can lay regenerated partials on the grid ladder however smeared the pooled F_0 density is, as long as the harmonics of its training music sit on that ladder, and the attractor offset of about 7 cents is where that ladder sits.

The third is registration: the effect is evidence of a Western prior only if the residual registers with absolute pitch rather than our interval axis, which is what (P4) tests and why it, not (P1), carries the central claim.

A.5 The quantiser account, prediction by prediction

(P2) holds, with a median bias of 8.88 cents over the sweep. (P1) fails: a clipped model with a free knee at $|\delta| = 22$ [19, 26] cents beats both candidate shapes ($R^2 = 0.83$ against 0.654 and 0.586). It has three free parameters (plateau, knee, offset), as the sinusoid does (amplitude, phase, offset) and one more than the sawtooth (slope, offset), so we compare on information criteria as well: on the twenty 5-cent bins AIC favours it by 14.2 over the sinusoid and 15.8 over the sawtooth, and a held-out split over detunings reproduces the ordering. The profile therefore saturates rather than growing

with δ (Figure A1a). What bounds the account is not the waveform but the bypass. Compared at a matched 2.2 kHz band edge, removing the quantiser leaves the all-trial bias unchanged within its interval, 5.66 [5.28, 6.06] against the 24 kbps condition's 5.73 [5.36, 6.11] at the same 2.2 kHz edge, a difference far inside the interval on either term; the across-edge comparison with the 3 kbps condition's 8.88 is not the right one, since the edge is what the rate sweep moves. The bypass runs the encoder straight into the decoder, which is off the manifold the decoder was trained on, so it is evidence that the effect does not require the quantiser rather than a measurement of the codec at some operating point. (P3) is inconclusive and out of regime at 0.16 to 2.5 bits per dimension, the fitted exponent depending more on which floor is subtracted than on the data (Appendix A.6). What the rate sweep was really measuring is the band edge: the bias falls because the edge rises, and plateaus because the partials above 2.2 kHz are regenerated at every rate. A pure sinusoid shows no all-trial bias (< 0.01 cents), as a tone with nothing above the edge must (Table A28).

Vibrato is the one stimulus manipulation that appears to strengthen the reading rather than weaken it. In Table A12 the coded column is a signed, grid-referred bias and the uncoded column is an unsigned error, so the uncoded column leaves open whether the estimator's own vibrato response carries a grid-directed component that grows with depth. Referred to the grid and read off-grid at $|\delta| \geq 30$, the uncoded arm's signed bias per depth comes to -0.00 [-0.00, 0.00], -0.00 [-0.02, 0.01], 0.01 [-0.03, 0.04], 0.06 [0.01, 0.09] and 0.07 [-0.10, 0.15] cents at depths 0, 5, 10, 20 and 40 — flat, and hundredths of a cent against a coded column that rises by seven over the same depths. The result does not depend on the scoring: the uncoded figure stays within a tenth of a cent of zero under every cut, statistic and reference we tried, while the unsigned column of Table A12 rises from 0.00 to 3.31. So the estimator answers vibrato with a symmetric spread and not with a grid-directed component, and the rise across depth is the codec's. The coded column of Table A12 is scored differently from this one and is not directly comparable in level. Read as it stands, the mechanism is the obvious one: a tone whose fundamental sweeps through the grid point spends more of its duration in the band where the invented partials are displaced, so the estimator integrates a larger share of the effect (Table A12).

A.6 The (P3) rate test in bits per dimension

Table A2 carries the test for EnCodec 24 kHz. Its raw-bias exponent, $[-0.756, 0.108]$, excludes both candidate regimes, and the floor-subtracted exponent is unidentified across the floor's own interval. That, rather than a per-codec comparison, is why (P3) is inconclusive.

Bitrate is the observable (Figure A2), but R/D is the coordinate Proposition 1 speaks in. Only EnCodec 24 kHz exposes a span of R/D wide enough to constrain a slope: its five rates cover R/D from 0.16 to 2.5, a span of 2.34. Mimi was run at one rate, and the rate ladders of DAC and SpeechTokenizer cover too little of the coordinate to contribute.

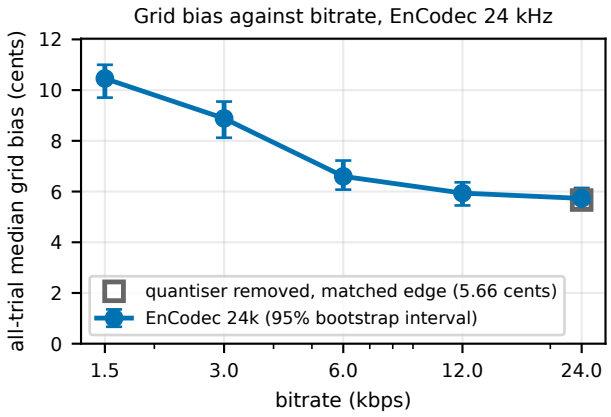

Figure A2: **The bias falls with bitrate, then stops.** All-trial median grid bias for EnCodec 24 kHz, 95% bootstrap intervals. The square is the quantiser removed, plotted only at 24 kbps because that is the point whose 2.2 kHz edge it matches. The plateau is what (P3) does not describe.

Table A2: Test of (P3), and why it is inconclusive. Fit of $\log_2 b$ against bits per latent dimension, EnCodec 24 kHz. The 24 kbps bias sits 0.07 cents above the bypass floor [5.28, 6.06], so the exponent depends on the subtracted value more than on the data.

Fit	**Slope**	**95% CI**	R^2
Raw bias	-0.324	$[-0.756, 0.108]$	0.655
Floor 5.66 subtracted (point estimate)	-2.612	$[-3.802, -1.422]$	0.942

Sweeping the subtracted floor across its own interval moves the fitted exponent through most of the range the two candidate terms occupy: the slope runs -1.43 at a floor of 5.28, -1.62 at 5.40, -2.18 at 5.60, -2.61 at the 5.66 point estimate, -3.75 at 5.72, -4.69 at 5.80 and -6.63 at 6.06, the last two fitted on four and three points because the subtraction leaves the higher rates non-positive.

B Experimental details

B.1 Full notation, and the three statistics

Table A3: **The three statistics we report for one effect, and the verdict built on them.** They are not interchangeable; the effect size is the first row.

Statistic	**At 3 kbps**	**Where it is used**
Off-grid median bias $\tilde{b}$	12.55	**The effect size.** Every *Bias* column and Table A13; the abstract carries the population median instead. Read at $\lvert\delta\rvert \geq 30$, past the near-grid region
All-trial median bias	8.88	(P2), the quantiser bypass, the rate sweep and repeated coding; the last two are read at one or two references rather than the full detuning design
Fitted amplitude A	13.72	Comparison with the periodic account it displaces; the causal arms, whose contrasts are in amplitude
Pull	—	Not a statistic but a binary verdict on a condition, defined in Section 2.3

Table A4 gives the full notation.

Table A4: **Notation.** Left, properties of the measurement; right, of the decoded spectrum. The effect size is $\tilde{b}$ read off-grid; A is reported alongside it for comparison with the account it displaces and is not a second measurement of the same thing.

Symbol	**Meaning**	**Symbol**	**Meaning**
θ	interval put in, cents	f_0, $f^\star$	fundamental; nearest 12-TET pitch
$r_C(\theta)$	interval read back minus θ	δ_0	the tone's own offset, $1200 \log_2(f_0/f^\star)$
d	how far the reference sits off the grid	s_k	how far partial k moved, cents
$\delta(\theta)$	signed distance of θ to the grid	ℓ_k, $\bar{\ell}$	that move as a share of the way to $kf^\star$
b, $\tilde{b}$	residual toward the grid; $\tilde{b}$ its off-grid median	f_e	band edge: lowest regenerated partial
A	fitted sinusoid amplitude	w_k	estimator weight, $\propto k^2\lvert X_k\rvert^2$
σ	registration slope: 1 absolute, 0 interval	b^{unif}	mean per-partial displacement (Eq. 15)
		$\bar{\ell}_\mathrm{r}$	$\bar{\ell}$ over regenerated partials alone ($\bar{\ell}$ itself runs over all at or above f_e)

B.2 Codec checkpoints and operating points

Table A5 lists the checkpoints of the registration sweeps with their operating points; the EnCodec 32k checkpoint of Appendix F.7, the vocoder, and the rate ladder of Appendix C.5 are reported where they are used. Each checkpoint is identified by the Hugging Face repository listed in the table, and the recorded revisions predate the runs.

Table A5: Every checkpoint in the registration sweeps, with its operating point. Frame rate is that of the discrete token stream; Q is the residual stages selected against the maximum the checkpoint exposes.

Codec	Checkpoint	kHz	Frame (Hz)	Q used / max	kbps
EnCodec 24k	`facebook/encodec_24khz`	24	75	4 / 32, 32 / 32	3, 24
EnCodec 48k	`facebook/encodec_48khz`	48	150	4 / 16	6
DAC 16k	`descript/dac_16khz`	16	50	6 / 12	3.0
DAC 24k	`descript/dac_24khz`	24	75	8 / 32	6.0
DAC 44k	`descript/dac_44khz`	44.1	86	4 / 9	3.4
Mimi	`kyutai/mimi`	24	12.5	8 / 32	1.1
SpeechTokenizer	`fnlp/SpeechTokenizer`	16	50	8 / 8	4.0
WavTokenizer[§]	`novateur/WavTokenizer-large-speech-75token`	24	75	1 / 1	0.90
SNAC 24k	`hubertsiuzdak/snac_24khz`	24	12, 23, 47	3 / 3	0.98
SNAC 32k	`hubertsiuzdak/snac_32khz`	32	10, 21, 42, 83	4 / 4	1.9
SNAC 44k	`hubertsiuzdak/snac_44khz`	44.1	14, 29, 57, 115	4 / 4	2.6

[§]The speech-trained release. The held-out pass of Appendix G.6 scores a WavTokenizer at 40 tokens per second, which is the *unify* release, a distinct model we do not pool with this one. The measured condition is speech-trained; its ladder is read against 12-TET all the same, and LibriSpeech's own within-semitone peak-to-mean sits above its flat null (Table A11), so a speech corpus is not grid-free either.

B.3 The registration test in full

Writing $d = 1200 \log_2(f_0/440) \bmod 100$ for the reference's own detuning, and taking the interval residual as the second tone's displacement minus the first's, the two hypotheses are

$$\mathrm{H}_{\text{abs}}:\ r_C(\theta; f_0) = \rho\big((\theta + d) \bmod 100\big) - \rho(d), \qquad \mathrm{H}_{\text{int}}:\ r_C(\theta; f_0) = \rho\big(\theta \bmod 100\big) + \gamma, \tag{12}$$

for some 100-periodic profile ρ and constant γ. Under H_{abs} detuning the reference by d cents translates the profile by d; under H_{int} it does not move. The reference tone's own displacement, $-\rho(d)$, is constant across θ within a condition and is absorbed by the intercept of the fit. Fitting the fundamental of the profile with a free offset, $r_C(\theta; f_0) \approx A(d)\sin\big(2\pi\theta/100 + \varphi(d)\big) + \text{const}$, the two hypotheses of Equation 12 give $\varphi(d) = \varphi_0 + \sigma\,(2\pi/100)\,d$ with $\sigma = 1$ under H_{abs} and $\sigma = 0$ under H_{int} (Equation 3), with $A(d)$ constant in both cases, to the extent that the amplitude does not change over the semitone of absolute pitch a sweep covers. We sweep d across a semitone in 10-cent steps (Appendix B.4), unwrap the fitted phases, and regress $\varphi(d)$ on d by least squares; the relative slope σ, with a 95% t interval on $n - 2 = 8$ degrees of freedom ($n = 10$ detunings, the 100-cent point folded onto 0), is the paper's central statistic. An amplitude that varies with d means neither model holds; the equivalence criterion and the necessity argument are Section 2.3.

B.4 Stimulus generator

Each stimulus is

$$x(\theta) \;=\; y_{f_0} \,\oplus\, \mathbf{0}_{0.1\,\mathrm{s}} \,\oplus\, y_{f_0 2^{\theta/1200}}, \qquad y_f(t) \;=\; a\,w(t) \sum_{n=1}^{N} \frac{1}{n} \sin\big(2\pi n f t + \phi_n\big), \tag{13}$$

with $\oplus$ concatenation in time and $N = 8$ partials for the pitch probe, 12 for the spectral sweep, whose one-second tones are what the 0.9 s analysis window requires.

The generator is deterministic given a seed and requires no dataset. Each trial is specified by an interval θ in cents, a reference frequency f_0, a partial count, and a seed controlling onset phase and amplitude jitter. The eleven detunings below give ten distinct references: the 100-cent value is a semitone and folds onto 0, which is why the sweep is described as eleven values in Section 2.5 and as ten references wherever a per-reference count is meant. The detuning run is:

```
refs = [440 * 2**(d/1200) for d in range(0, 101, 10)]  # 0..100 c
for theta in range(0, 1205, 5):          # interval, cents
  for f1 in refs:                        # reference, Hz
    for rep in range(20):                # repetitions per cell
      f2 = f1 * 2 ** (theta / 1200)
      x  = tone(f1, dur=0.5, partials=8, seed=rep)
```

```
        x += silence(0.1)
        x += tone(f2, dur=0.5, partials=8, seed=rep)
        yield theta, f1, rep, x
```

Partial amplitudes fall as $1/n$ and each tone carries 20 ms raised-cosine ramps. A partial whose frequency would exceed 95% of the codec's Nyquist frequency is dropped from the sum rather than allowed to alias, which matters only at the top of the octave sweep. The onset phase of every partial is drawn uniformly on $[0, 2\pi)$ and the peak gain of each tone uniformly on $[0.55, 0.85]$, both redrawn on every repetition, so no result depends on a fixed phase relationship and the codec never sees an identical waveform across repetitions. Two detuning configurations were run. The headline EnCodec sweep, its 220 and 880 Hz repeats and the classical-codec sweeps of Appendices D.1 and D.2 use the eleven references above at exact 10-cent steps with twenty repetitions, $241 \times 11 \times 20 = 53{,}020$ trials, the 100-cent point folding onto $d = 0$. Every condition run after the change below uses the reduced design instead, which in Figure 2 is thirteen of the seventeen rows: the twelve codec conditions other than the headline EnCodec, plus the BigVGAN arm. Those thirteen use ten references rounded to 0.1 Hz (detunings of 0, 9.8, 20.0, 30.0, 40.1, 50.0, 59.9, 70.2, 80.0 and 90.1 cents, no 100-cent point) with sixteen repetitions, $241 \times 10 \times 16 = 38{,}560$ trials, a compute-saving change made before those sweeps were run; the analysis derives each condition's detuning from the recorded reference frequency, so the rounding enters the regression exactly. The rate sweep and the other controls use two references, 440 Hz and 440 Hz detuned by 50 cents, at five repetitions, $2{,}410$ trials per condition; the octave test uses 110, 220, 440 and 880 Hz.

GTZAN carries documented duplicates, mislabelling and distortion (Sturm, 2013). None of that bears on the quantity we take from it, which is the within-semitone distribution of F_0 over the whole set, and the causal experiment compares that corpus with resampled and flattened copies of itself, so any defect is shared by every arm. It would matter for a genre claim, and we make none. That distribution is taken against a fixed A4 = 440 Hz, and the recordings are not all mastered there: pooling references that differ blurs the peak, so 1.788 is a floor on the corpus's concentration and not an estimate of it. Per-clip tuning estimates put the median clip within 2 cents of 440 Hz, so a shared mastering offset does not account for the attractor near 7 cents.

B.5 Estimator implementation and noise floor

F_0 is estimated in three stages in two domains: YIN in the time domain and harmonic summation in the frequency domain each give a blind coarse estimate, and a harmonic least-squares refinement searches a ± 60-cent window around each expected partial of the coarse estimate, locating the partial by parabolic interpolation of the log-magnitude spectrum and combining the first eight with weights $k^2|X|^2$. The refinement is run once from each coarse estimate; the YIN-seeded value is reported and the disagreement between the two, in cents, is logged per trial as the cross-check, which the robustness variant applies at a 5-cent threshold on the synthetic sweeps and a 20-cent threshold on real recordings. Like the learned trackers of Section 2.5, pYIN (Mauch & Dixon, 2014) is excluded, because it estimates one period from the whole waveform rather than fitting partials, so it cannot read the per-partial structure these sweeps measure, and it returns as the negative control of Appendix F.4; the cost is the octave failures the gate handles. Because the coarse stages work in different domains, an octave error in one is very unlikely to be matched by the other. Table A6 gives the estimator's floor on uncoded stimuli.

Table A6: Estimator noise floor on uncoded stimuli. Every effect in the paper is compared against these values; anything not clearing the floor by more than its confidence interval is reported as null.

Stimulus	**Median |error| (c)**	**95th pct (c)**	**Max (c)**
Harmonic complex, 24 kHz	1×10^{-5}	4×10^{-5}	6×10^{-5}
Harmonic complex, 16 kHz	2×10^{-5}	7×10^{-5}	1×10^{-4}
Harmonic complex, 44.1 kHz	$< 10^{-5}$	3×10^{-5}	6×10^{-5}
Pure sinusoid (control)	4×10^{-5}	4×10^{-4}	5×10^{-4}

B.5.1 THE YIN STAGE ON ITS OWN.

The corpus histograms (Table A11), the MusicGen readings (Appendix F.7) and the phonology contours use the YIN stage alone on 64-ms frames, without the cross-check or the refinement. What separates them from the other real-audio analyses is not polyphony but what is wanted from the frames. These three want a *density* over every frame of arbitrary material, so discarding frames is not free: the refinement fits eight partials of one harmonic series, which a chord does not have, and gating on the cross-check would throw away exactly the frames that make the density. The residual protocol of Appendix F.2 reads GTZAN and Saraga clips and does the opposite, wanting one tracked voice rather than a density: six partials per frame with a hard cross-check gate, retaining 23 to 40% of frames across the runs of Table A30. Two refinements seeded in different domains rarely land within 20 cents unless a single harmonic series dominates, so the gate keeps the monophonic-enough frames and discards the chords — which is what makes the refinement admissible there, at the price of a residual measured on a selected subset, audited there against input position. `test_yin_bias.py` characterises the YIN stage on the sweep's harmonic complexes at 721 fundamentals from 110 to 880 Hz in 5-cent steps: its error is $+0.29$ cents on average at 22.05 kHz and $+0.16$ at 32 kHz (95% range $[-0.2, 1.9]$, maximum 3.0), sharp by the parabolic-interpolation error at short lags, growing with pitch (0.13, 0.17 and 0.58 cents over the three octaves) and flat across the within-semitone position (bin means within 0.15 cents of each other), so on clean tones it neither sculpts the density nor moves its peak by more than a fraction of a cent; the refined estimator reads the same frames to 10^{-2} cents. On MusicGen's polyphonic output the two stages agree on the density's shape (peak-to-mean 4.28 against 4.27) but not its offset: on a 40-clip subset YIN puts the circular-mean position 2.8 cents above the 12-TET pitch and the refined estimator 0.4, against YIN's 3.6 on the full 60, so that offset is uncertain by a few cents and we rest no claim on it. The 7-cent attractor offset has independent support from the fine-tuned arms' zero crossings on synthetic tones (Table A25), read with the refined estimator (Section 3.3). The flat-sampling expectations 1.07 and 1.04 are simulated by `flat_sampling_null.py` (500 multinomial draws over 50 bins at each corpus's count): 1.070 [1.047, 1.099] at LibriSpeech's (Panayotov et al., 2015) 53,678 estimates and 1.037 [1.025, 1.053] at GTZAN's 180,859; the tuning-flattened GTZAN corpus at 1.073 and LibriSpeech at 1.113 both sit just above their nulls' 97.5th percentiles, so neither is exactly flat; the 0.91 cents attributed to the corpus's tuning is therefore a lower bound and the flattened arm's residue may still carry grid structure.

B.6 RESAMPLING UNIT AND INTERVAL CONSTRUCTION

The off-grid bias and the Bonferroni-adjusted interval that decides the pull are both read at a condition's on-grid reference, and both are percentile bootstraps over that condition's trials, with seeds redrawn per condition so stimulus noise is not shared across rows. This is the smallest unit in the design: the interval covers scatter over repetitions of one stimulus at one detuning, and is not a cluster interval over detunings, registers or timbres. The full grid has 132 detuning-by-register-by-timbre cells, which the headline interval does not resample; Appendix C.7 supplies a stimulus-level interval instead, and Table A36 states the unit for every other quantity. The band-limited reading of Section 2.5 is this same estimator restricted to the partials below a condition's band edge.

B.7 EXCLUSION RULES AND REFUSED CONDITIONS

Two rules exist, at different granularities, and only one is primary. The *tracking gate* drops a trial whose estimate is more than 200 cents from nominal. The justification is one of domain: the largest grid effect the signed metric can express is 50 cents, so a trial landing two semitones away is outside the region in which grid bias is defined, whether the estimator failed or the codec moved the pitch that far (SNAC 24k does the latter, with median errors of hundreds of cents). It removes 12 of the headline run's 53,020 trials, none at the 440 Hz reference, but is not zero everywhere: Table A8 gives the rate for every sweep run, and for DAC 16k it is 65%. Ungated its registration slope is 0.98 $[0.86, 1.10]$ against 1.0199 gated, still inside 1 ± 0.15, so the gate does not create the registration; what it buys is precision, the ungated fit being the loosest of any admitted condition. The eighteen conditions we score are the thirteen neural ones, the vocoder, and Opus and MP3 at two bitrates each; the two octave repeats of EnCodec 24k at 3 kbps are excluded from that count, since they repeat a condition already in it, and the table prints twenty rows because they are shown. Those eighteen conditions, as twenty runs, are the set Figure A3 ranges over and Table A8 lists, and eighteen less

the four refused — DAC 44k, SNAC 24k, Opus at 12 kbps and MP3 at 32 kbps — is the fourteen admitted. The HE-AAC and CELT sweeps carry their own gate figures in Table A20 instead. For DAC 16k the gate's rate is flat across the semitone over the whole sweep but not at the on-grid reference, where the bias of Figure 2 is measured (Table A7): there it rises from 48% in the 0 to 10 cent bin to 62.5% in the 40 to 50 cent bin, a gate correlated with grid distance. What it removes are subharmonic returns at residuals near ± 1200 cents, so folding modulo 1200 and keeping every trial is the right check: that gives -0.14 [-0.27, 0.01] against the gated -0.13, the same null with a tighter interval.

Opus is gated at 47%, the highest rate of any admitted condition after DAC 16k, and gets the same check. Folding its residuals modulo 1200 cents and keeping every trial gives an amplitude of 1.61 $\pm$ 0.39 cents and a bias of $+0.02$ [-0.31, 0.38], against 1.57 ± 0.37 and 0.00 [0.00, 0.42] gated, so the classical control's null is not a product of the gate either.

Table A7: Tracking-gate rate for DAC 16k by distance $|\delta|$ from the nearest grid point, over the whole sweep and at the on-grid reference alone; EnCodec at 3 kbps gates 0.0% in every bin.

$\lvert\delta\rvert$ (c)	0–10	10–20	20–30	30–40	40–50
All trials ($n = 38{,}560$)	65.1%	64.4%	65.1%	65.2%	65.7%
On-grid reference ($n = 3{,}856$)	48.0%	49.5%	56.8%	59.4%	62.5%

The *estimator cross-check* additionally drops trials on which the spectral and autocorrelation estimators disagree. It is reported only as a robustness variant because it does not fire uniformly: under an on-grid reference its exclusions concentrate 40 to 50 cents from a grid point, and under a reference detuned by 50 cents they concentrate 0 to 10 cents from one. On the headline run it raises the off-grid-to-on-grid error ratio from 3.27 to 3.48 and the bias from 8.88 to 9.26 cents while leaving the registration slope at 1.0011, so no exclusion rule creates the effect. For the speech-shaped stimuli the cross-check discards most trials, tripping the retention guard of Appendix B.8 on the EnCodec vowel condition, which is the practical reason it cannot be the primary rule.

Table A8: Tracking-gate exclusion rate, twenty runs over the eighteen measured conditions plus two octave repeats (text), with the two guard values of Appendix B.8: the amplitude coefficient of variation across detunings, and retention after the gate.

Run	**Codec**	**Gated**	**Amplitude cv**	**Retention**
detune_encodec3	EnCodec 24k, 3 kbps	0.0%	0.012	1.00
detune_encodec24kbps	EnCodec 24k, 24 kbps	0.0%	0.009	1.00
detune_encodec48	EnCodec 48k, 6 kbps	2.7%	0.027	0.97
detune_mimi	Mimi	0.4%	0.119	1.00
detune_vowel_speechtok	SpeechTokenizer, vowels	0.0%	0.137	1.00
detune_dac16	DAC 16k	65.1%	0.208	0.349
detune_snac32	SNAC 32k	0.0%	0.084	1.00
detune_wavtok†	WavTokenizer, 0.9 kbps	1.2%	0.041	0.99
detune_vowel_encodec	EnCodec 24k, vowels	1.5%	0.052	0.98
detune_dac24	DAC 24k	0.0%	0.104	1.00
detune_snac44	SNAC 44k	2.3%	0.183	0.98
detune_bigvgan†	BigVGAN (vocoder)	1.8%	0.074	0.98
detune_dac44	DAC 44k (refused)	5.5%	0.407	0.95
detune_snac	SNAC 24k (refused)	93.2%	0.572	0.068
detune_encodec3_220	EnCodec 24k at 220 Hz	0.4%	0.017	1.00
detune_encodec3_880	EnCodec 24k at 880 Hz	9.2%	0.096	0.91
detune_opus6	Opus 6 kbps	46.7%	0.233	0.53
detune_opus12	Opus 12 kbps (refused)	17.9%	0.621	0.82
detune_mp316	MP3 16 kbps	0.0%	0.065	1.00
detune_mp332	MP3 32 kbps (refused)	0.0%	0.802	1.00

† The WavTokenizer and BigVGAN sweeps, also reported in Tables A9 and A16.

DAC 16k's null is discussed with its power in Appendix E.2, and Appendix E.1 gives the EnCodec fine-tuning contrast of Section 3.3 on that codec (Table A26).

B.8 THE REFUSED CONDITIONS IN FULL

Four conditions are refused: SNAC 24k and DAC 44k among the neural codecs (Table A9), and Opus at 12 kbps and MP3 at 32 kbps among the classical controls (Table A19). Fitting a phase regression to SNAC at 24 kHz produced a slope of -0.92 at $R^2 = 0.61$, which has the form of a finding and is not one: that codec destroys these stimuli, with median errors of 201 cents on-grid and 551 off-grid, 6.8% of trials surviving, and per-condition sinusoid fits below $R^2 = 0.081$. The phases were noise and the regression a line through noise. A different failure, an amplitude too unstable to carry a phase (cv 0.407), gave DAC 44k an apparent slope of 1.0170 at $R^2 = 0.994$ from amplitudes of 0.06 to 0.18 cents.

Per-condition R^2 does not discriminate these cases, since a genuine small-amplitude effect has low R^2 against trial noise while still giving a precise phase from a thousand trials — SNAC 44k sits at R^2 of 0.03 to 0.06 and regresses at 0.997. Two tests do, and both follow from the prediction: the amplitude must be flat across detuning, because the claim is that detuning moves the phase and nothing else, and a condition retaining almost no trials has no phase to speak of. The thresholds are an amplitude coefficient of variation of at most 0.35 and a retention of at least 25%. Both are recorded in the registration file, and neither decides a verdict: on every run's values, each reported condition has cv at most 0.233 and retention at least 0.349, while the nearest refusals are DAC 44k at cv 0.407 (retention 0.95) and SNAC 24k at retention 0.068 (cv 0.572), so the reported set is the same for any cv limit in $[0.233, 0.407)$ and any retention limit in $(0.068, 0.349]$.

Because DAC 16k's retention of 0.349 is the lowest of any admitted condition, Figure A3 plots how the reported set depends on where the two thresholds are put. The admitted set is the same fourteen conditions at every cv limit from 0.25 to 0.40 crossed with every retention limit from 0.15 to 0.34, and DAC 16k is admitted throughout; only raising the cv limit to 0.45 changes anything, by admitting DAC 44k.

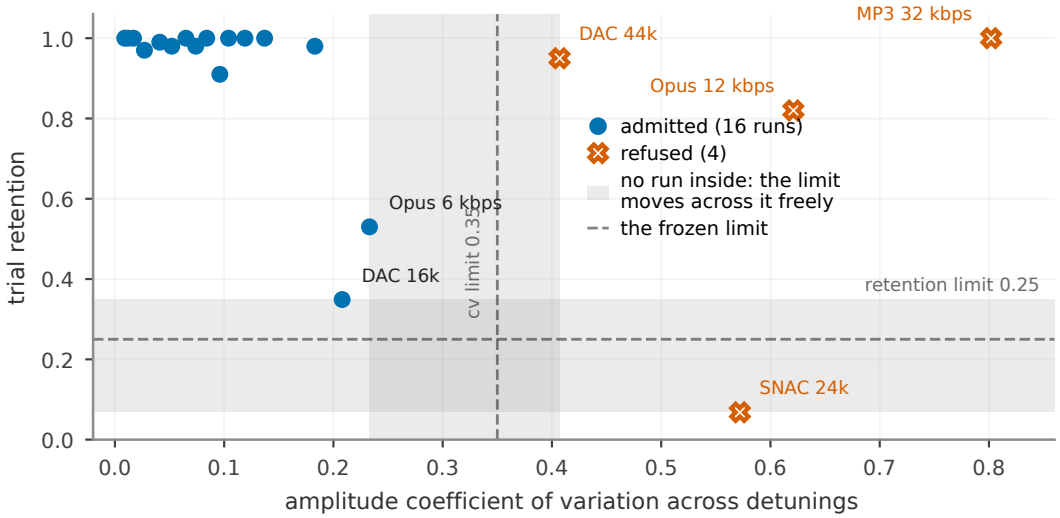

Figure A3: **The two guards, and why their values do not decide the reported set.** Every run of Table A8 — runs, not conditions, so the octave repeats are separate points. Each band spans values with no run inside it, so a limit crossing it changes nothing.

Each admissible range is bounded on one side by Opus and DAC 16k and on the other by a refusal, so DAC 16k is the admitted run nearest the retention limit.

C THE MEASUREMENT IN FULL

C.1 REGISTRATION, THE CODEBOOK PROBE AND THE CORPORA

One caution on reading the table that follows: six of its eleven codec rows clear zero on the printed intervals, which is the paper's count; folding in the between-run component of Appendix G.2 gives five at Satterthwaite degrees of freedom, dropping the vowel condition from [0.19, 0.60]

to $[-0.14, 1.00]$, and four at t_4. Rerunning the registration under the estimator's cross-check leaves every fitted slope interval inside the margin, SpeechTokenizer's included: its cross-check slope is 0.9893 [0.9043, 1.0743] where its primary slope, 0.8559 [0.8240, 0.8878], is the one row that does not register. So the single registration failure holds under the primary rule only: the cross-check, whose retention guard refuses the EnCodec vowel condition, registers all ten conditions it admits. The cross-check column is omitted here for width.

Table A9: Registration for every measured neural condition, the vocoder and two fine-tuned arms included. Slope is against $3.6°$/cent; a dash is a quantity the row lacks. *99.55% over the eleven codec rows, the Bonferroni family ($z = 2.84$); 95% for the vocoder and fine-tuned rows, outside it.

Codec	**Point, stimulus**	**Slope [95% CI]**	R^2	**Phase**	**Bias (c)***
EnCodec 24k	3 kbps, tones	1.0010 [0.9922, 1.0099]	0.99988	$+155°$	$+12.55$ [10.71, 13.67]
EnCodec 24k	24 kbps, tones	0.9972 [0.9922, 1.0021]	0.99996	$+151°$	$+8.03$ [6.45, 8.81]
EnCodec 48k, music	6 kbps, tones	1.0188 [1.0042, 1.0334]	0.99969	$+174°$	$+6.54$ [5.83, 7.80]
WavTokenizer	0.9 kbps, tones	1.0043 [0.9971, 1.0115]	0.99992	$+158°$	$+7.61$ [6.44, 8.52]
BigVGAN (vocoder)	mel input, tones	0.9971 [0.9862, 1.0080]	0.99982	$+167°$	$+4.82$ [3.91, 5.64]
Mimi	Q8, tones	0.9576 [0.9293, 0.9858]	0.99869	$+132°$	$+1.19$ [-0.15, 2.86]
SpeechTokenizer	Q8, vowels	0.8559 [0.8240, 0.8878]	0.99792	$+114°$	-0.03 [-0.21, 0.14]
DAC 16k	Q6, tones	1.0199 [0.9570, 1.0827]	0.99432	$+82°$	-0.13 [-0.55, 0.07]
SNAC 32k	default, tones	0.9638 [0.9259, 1.0018]	0.99767	$+149°$	$+0.61$ [0.18, 1.02]
EnCodec 24k	3 kbps, vowels	1.0109 [0.9897, 1.0321]	0.99934	$+164°$	$+0.43$ [0.19, 0.60]
DAC 24k	Q8, tones	0.9785 [0.9508, 1.0062]	0.99880	$+121°$	$+0.01$ [-0.22, 0.13]
SNAC 44k	default, tones	0.9949 [0.9542, 1.0356]	0.99749	$-125°$	-0.14 [-0.29, 0.07]
EnCodec, FT original clips	3 kbps, tones	0.9991 [0.9914, 1.0068]	0.99991	–	–
EnCodec, FT flattened clips	3 kbps, tones	0.9975 [0.9900, 1.0051]	0.99991	–	–
DAC 44k	Q4, tones	*not measurable*			
SNAC 24k	default, tones	*not measurable*			

Table A10: Codebook probe: positions of code-assignment boundaries within the semitone, pooled over four quantiser levels, tested for circular concentration. The Rayleigh test rules out strong grid alignment, not a weak one; the last three rows say how weak.

Boundaries located	2341
Mean distance to nearest grid point	25.09 cents
Expected under uniform	25.0 cents
Mean resultant length $\bar{r}$	0.008
Rayleigh p	0.86
Rayleigh p on doubled angles	0.83
Chi-square p, ten 10-cent bins	0.65
Per-level Rayleigh p (four levels)	0.16, 0.91, 0.29, 1.00
Detectable $\bar{r}$ at 80% power	0.045
as mean distance to nearest grid point	24.05 cents
Power at the measured $\bar{r}$	7 %

Table A11: Within-semitone F_0 density of the training corpora as peak-to-mean, where 1.0 is flat, against the floor sampling alone gives at that size. The flattened corpus is Section 3.3's manipulation, which leaves timbre, instrumentation and production untouched.

Corpus	F_0 **estimates**	**Peak/mean**	**Flat floor**
GTZAN, Western music	180,859	1.788	1.04
GTZAN, tuning smeared flat	175,552	1.073	1.04
GTZAN, quantised to 24-TET	178,768	7.3 at 0 c, 3.4 at 50 c	1.04
LibriSpeech, speech	53,678	1.113	1.07

On the codebook probe and prior reports of pitch structure. Paek et al. (2025) report discretised pitch features read directly from codec latents, which bears on our Rayleigh null. The two probe different objects: theirs is a learned feature over continuous latents, ours is the circular distribution of code-assignment *boundaries* within a semitone. A latent space can carry pitch information

in a form our test does not probe, and our null bounds the claim that assignment boundaries align *strongly* with 12-TET, leaving a weak alignment open. Neither result excludes the other.

C.2 THE OCTAVE TEST, THE RESIDUAL'S SHAPE AND THE CORPORA

Figures A4 to A6 carry the octave test, the residual's shape and the corpus densities; each caption states what it shows. Two further notes apply. The amplitudes in the titles of Figure A5 are single-condition fits at the 440 Hz reference, so they differ from the mean over detuning: 13.91, from the single-reference pass of Appendix F.6, against Table A3's 13.72 for EnCodec, and 4.50 against 4.97 for Mimi and 1.30 against 1.78 for DAC 16k, the last two read off the same fits as the figure rather than off any table, at $R^2 = 0.07$, which marks them weak. And the waveform comparison of Appendix A.5, sinusoid $R^2 = 0.654$ against sawtooth 0.586, is that same 440 Hz condition.

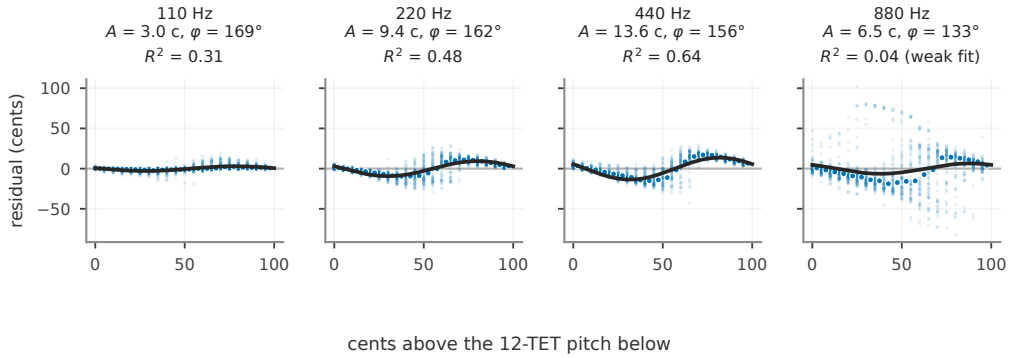

Figure A4: **The period is one semitone, not one frequency interval.** A 100-cent-period sinusoid at four references an octave apart; stride structure would halve its period each octave. Fits are from the octave run, so 440 Hz reads 13.6 against Figure A5's 13.91.

Table A12: **The reading rises with vibrato depth from 10 cents up.** EnCodec at 3 kbps on complexes carrying 5.5 Hz vibrato of the stated depth, on the two-reference sweep of Appendix B.4. Only the trend across depth is read here; the level is not comparable with the headline run.

Vibrato depth (c)	Coded grid bias (c)	Uncoded $\lvert r \rvert$ (c)	Uncoded signed (c)
0	9.06	0.00	−0.00
5	9.03	0.42	−0.00
10	9.86	1.04	0.01
20	12.72	2.42	0.06
40	16.03	3.31	0.07

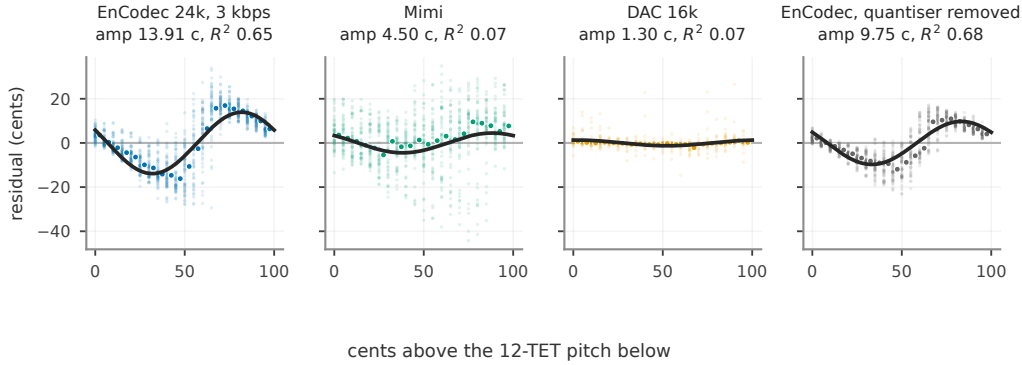

Figure A5: **The residual, one panel per condition.** Folded into one semitone at 440 Hz, the rightmost panel EnCodec with the quantiser removed. Pale points are trials, solid points binned medians, the curve the fit; titles are single-condition fits, not means over detuning.

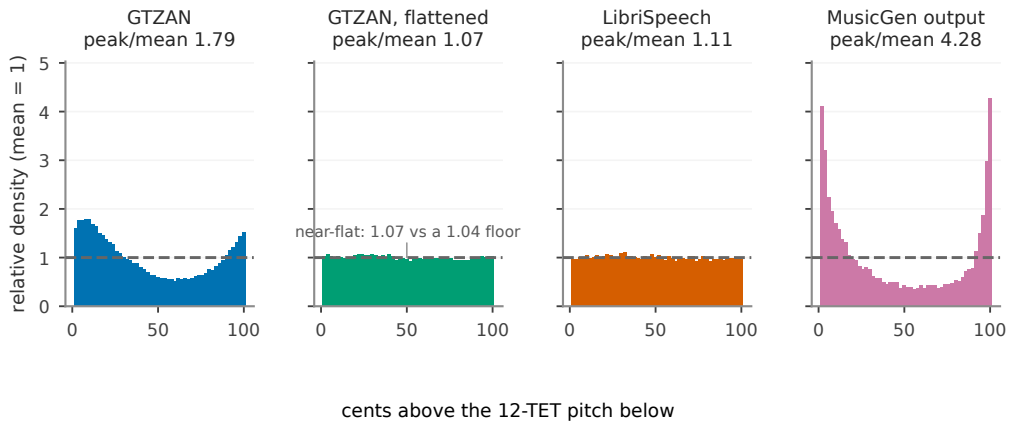

Figure A6: **What the corpora put where, and a generator.** Within-semitone F_0 density of Western music, its tuning-flattened copy, speech and MusicGen's output. Flat-sampling floors: 1.04 for both GTZAN arms, 1.07 speech, 1.12 MusicGen. The flattened arm reads 1.07, near-flat and not flat.

C.3 Registration at other octaves

The headline registration sweep detunes a 440 Hz reference. Repeating it with the reference at 220 and 880 Hz gives relative slopes of 0.992 [0.982, 1.003] with amplitude 9.43 ± 0.16 cents at 220 Hz and 0.968 [0.936, 0.999] with amplitude 5.99 ± 0.58 cents at 880 Hz (ten detuning conditions and a $t(8)$ interval throughout, the 100-cent point folding onto $d = 0$), both at $R^2 > 0.998$. The 880 Hz interval [0.936, 0.999] just excludes 1 but lies inside 1 ± 0.15, so the lock holds at both registers, less precisely at 880. These amplitudes are means over the ten detunings of a registration sweep, where Appendix C.4's 9.38 and 6.46 are single-condition fits from the octave run, which is why the two differ by up to 8%.

C.4 The octave sweep, and the edge as a transition

A period of one semitone is a period in *log* frequency. Convolutional strides, frame rates and filter lengths impose structure in *linear* frequency, whose period in cents halves with every octave. Sweeping four reference pitches an octave apart, a 100-cent-period fit succeeds at three of them, with amplitudes of 2.99, 9.38, 13.64 and 6.46 cents at 110, 220, 440 and 880 Hz and phases of 169, 162, 156 and 133 degrees, a range of 36 degrees. A linear-frequency origin would have made these fits fail (Figure A4). Amplitude varies strongly with octave and peaks near 440 Hz, a pattern taken up below; the 880 Hz fit is the noisiest of the four ($R^2 = 0.04$), so the argument rests on the three lower registers. The 880 Hz reference tone keeps all eight partials, since $8 \times 880 = 7040$ Hz clears 95% of Nyquist at every rate in Table A5. The second tone sits at $880 \times 2^{\theta/1200}$, so partial 8 leaves the band above $\theta \approx 835$ cents and partial 7 above $\theta \approx 1066$, and the generator drops them rather than aliasing (Appendix B.4). The top three tenths of the 880 Hz sweep is therefore read on six or seven partials by an estimator fitting eight with k^2 weights, putting the heaviest weight on partials that are not there. Restricted to $\theta \leq 600$ cents, where all eight survive, the fit's R^2 rises from 0.04 to 0.31 and the amplitude from 6.46 to 9.1 cents. Much of the 880 Hz anomaly is a stimulus artefact; what remains after the restriction is smaller and is taken up below. The lower registers are unaffected: at 440 Hz the second tone's eighth partial reaches 7040 Hz.

Table A13: **The band edge is a property of the codec, not of the octave.** EnCodec 24k at 3 kbps. b^{unif} is Equation 15 and b^{soft} the soft-edge fit. Edge from the twelve-partial stimulus, count from the probe; dash, undefined; brackets, 99.55% trial bootstraps, 110 Hz from the octave run.

Reference	Edge (kHz)	At or above	$\bar{\ell}$	b^{unif} (c)	b^{soft} (c)	Measured (c)
110 Hz	1.32 ($12f_0$)	0 of 8	–	0.0	2.52	2.75 [1.94, 3.55]
220 Hz	1.32 ($6f_0$)	3 of 8	0.95	14.3	9.40	8.66 [7.31, 9.94]
440 Hz	1.32 ($3f_0$)	6 of 8	0.90	27.0	12.50	12.55 [10.71, 13.67]
880 Hz	1.76 ($2f_0$)	7 of 8	0.79	27.7	13.94	5.50 [4.12, 6.85]

Equation 4 makes f_{e} the lowest *regenerated partial*, so per register it can only take values kf_0 and its resolution is the partial spacing — 110 Hz at the bottom register and 880 Hz at the top. Within that resolution the edge is *identically* 1320 Hz at three of the four registers; at 880 Hz the design can bound it only to $(0.88, 1.76]$ kHz and we report the attained value. The 880 Hz register does not resolve an edge, since the attainable edges either side of 1.32 kHz there are 880 and 1760 Hz, and is excluded from the edge columns of Tables A16 and A17, so EnCodec's 3 kbps entry is the 1.3 kHz the three lower registers agree on and not the 1.4 kHz a mean over all four would give, and its $\bar{\ell}$ is the 440 Hz reading.

The table is the sharpest statement of what the account does and does not buy. The edge does not move with the octave — at 110, 220 and 440 Hz it is the same 1320 Hz, reached as the twelfth, sixth and third partial — so it is the moving count of partials above a fixed frequency, not any change in the codec, that drives the reading. But b^{unif} over-predicts the measured bias everywhere it predicts anything — 0.61 of prediction at 220 Hz, 0.46 at 440, 0.20 at 880 — and at 110 Hz it predicts nothing at all while 2.75 cents is measured. Both failures come from the same two idealisations, and correcting them gives the b^{soft} column.

The edge is a transition, and regenerated partials roll off. b^{unif} treats the edge as a step and weights every partial above it equally. Neither holds.

Proposition 2 (Leading-order form of the transition)**.** *Let $T(f) \in (0,1)$ be the fraction of a partial's energy the decoder transmits rather than regenerates, smooth and decreasing, with the soft edge $f_{\mathrm{e}}^{\mathrm{soft}}$ defined by $T(f_{\mathrm{e}}^{\mathrm{soft}}) = \frac{1}{2}$ — a continuous quantity, distinct from Equation 4's discrete f_{e}, which is a multiple of f_0. Write $o = \log_2(f/f_{\mathrm{e}}^{\mathrm{soft}})$, read T as a function of o, and let $L(o) = \log\big(T/(1-T)\big)$. Then, to first order in o,*

$$1 - T \;=\; \frac{1}{1 + e^{-o/w}} + O(o^2), \qquad w \;=\; -\Big(\tfrac{dL}{do}\Big|_{o=0}\Big)^{-1} > 0, \tag{14}$$

so the regenerated share $\rho_k = \bar{\ell}\,\big(1 - T(o_k)\big)$ is logistic in log frequency for any *smooth transmitted fraction.*

Proof. L is smooth wherever $T \in (0,1)$, and $L(0) = 0$ because $T = \frac{1}{2}$ at the edge. Expanding about $o = 0$ gives $L(o) = \dot{L}(0)\,o + O(o^2)$ with $\dot{L} = dL/do$, and $\dot{L}(0) < 0$ since T decreases, so $w > 0$. Inverting the logit, $1 - T = \big(1 + e^{L}\big)^{-1}$, and substituting $L = -o/w$ gives Equation 14. Note that $\dot{L}$ is the derivative in *octaves*: the derivative in frequency differs from it by a factor $f \ln 2$. □

The logistic is the leading-order form near the edge rather than a fitted family, and w is the reciprocal log-odds slope of T there, measured in octaves. It is the model's one *fitted* parameter, but not its only degree of freedom: $\bar{\ell}$ is held and the roll-off ν, defined below, is chosen at the low end of its measured range, so the model carries one fitted parameter and two chosen constants. Two sensitivities follow. At the top of the measured 4.5 to 7 dB attenuation range — 7 dB over the same 0.415 octaves the paragraph below uses, so $\nu = 16.9$ dB per octave — a refitted width holds only two of the four registers within 9%. And the expansion above is local while the fit evaluates it out to $|o| \approx 8w$ at 110 Hz, where none of the fitted 2.5 cents comes from a partial within one width of the edge.

The weighting adds no fitted parameter. The estimator weights partial k by $k^2|X_k|^2$, which is flat across k for the $1/n$ input, so a regenerated partial enters only through its attenuation: returning A_{dB} dB down, it carries weight $10^{-A_{\mathrm{dB}}/10}$ relative to a transmitted one. That is the estimator's definition, not an assumption. Taking A linear in log frequency above the edge and zero below it, $A_{\mathrm{dB}} = \nu \max(o, 0)$ — a transmitted partial is not boosted, only an invented one is attenuated — the slope ν is anchored to a measurement rather than fitted: regenerated partials at 3 kbps return 4.5 dB down, and the first partial above a 1320 Hz edge at a 440 Hz reference sits 0.415 octaves up, so $\nu = 10.8$ dB per octave — anchored at the low end of the 4.5 to 7 dB range Table A14 reports. The predicted reading is then the weight-averaged displacement, $|\delta_0| \sum_k w_k \rho_k / \sum_k w_k$.

We hold $\bar{\ell}$ at 0.90, the 440 Hz value, at every register: Table A13 leaves it undefined at 110 Hz, and using each row's own value moves the 220 Hz prediction outside 9%. Fitting the single remaining parameter on the four registers in log ratio, so that each counts equally, gives $w = 0.45$ octaves

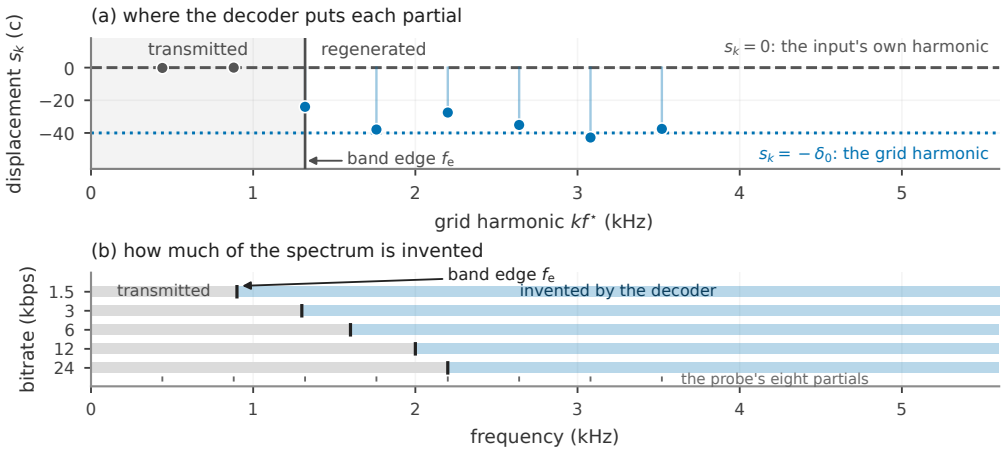

Figure A7: **What the decoder invents, and where it puts it.** (a) The eight partials the estimator fits, 440 Hz at $\delta_0 = +40$ cents, EnCodec at 3 kbps, plotted at the grid harmonics $kf^\star$ with the input partials 40 cents above each. (b) The transmitted/invented split across five bitrates.

(`analyze_softedge.py`). **One fitted width then reproduces, in sample, three of the four registers to within 9%**, 110 Hz among them, and 440 Hz to within a tenth of a cent: 2.52 against 2.75, 9.40 against 8.66, 12.50 against 12.55. The 110 Hz reading needs no second mechanism. It is the same one, acting on the partials the step model places wholly below the edge: at 110 Hz the seventh and eighth partials, at 770 and 880 Hz, sit 1.7 and 1.3 widths below the edge — outside the transition on a one-width reading, but far nearer it than the fundamental's eight — and that is where the 2.75 cents comes from. The prediction is an extrapolation in either case.

The 880 Hz register is over-predicted, by a factor of 2.5. Part of that is the stimulus artefact above. Restricted to $\theta \leq 600$ cents, where all eight survive, the measured off-grid bias rises from 5.50 to 7.8 [5.9, 9.7] and the over-prediction falls from 2.5 to 1.8. It does not fall to 1. Dropping partials from the model changes its prediction by less than 2%, because the roll-off has already given the high partials almost no weight, so the missing partials cannot be the whole story — what they change is the measurement, not the prediction.

What remains at 880 Hz sits in one register: the only one whose partial spacing cannot resolve the edge, bounded only to $(0.88, 1.76]$ kHz, and the only one whose $\bar{\ell}$ falls below 0.85. A model with one fitted parameter and two chosen constants that agrees at three of four registers, and at two of four if the roll-off is taken at the other end of its range, is a better account than the reference it replaces, with one register left open.

C.5 Which partials move

Table A14: **What the estimator reads against what the partials did.** *Uniform-weight* is the mean per-partial displacement over eight partials, *Reading* what the amplitude-weighted fit returns. The ratio is below one because a regenerated partial returns 4.5 to 7 dB down.

Bitrate	Edge	At or above	$\bar{\ell}$	Uniform-weight (c)	Reading (c)	Ratio
$\delta_0 = -40$ *cents*						
1.5 kbps	H2 (880 Hz)	7	1.04	36.5	24.4	0.67
3 kbps	H3 (1320 Hz)	6	1.00	30.1	20.7	0.69
24 kbps	H5 (2200 Hz)	4	0.86	17.3	7.3	0.42
$\delta_0 = +40$ *cents*						
1.5 kbps	H2 (880 Hz)	7	0.81	28.4	18.1	0.64
3 kbps	H3 (1320 Hz)	6	0.85	25.6	16.9	0.66
24 kbps	H5 (2200 Hz)	4	0.58	11.7	5.9	0.50

Table A15 gives the per-partial detail for EnCodec 24 kHz at three bitrates and both detuning signs, all at a 440 Hz reference, the most direct picture of the mechanism; Table A14 is read off it; the sweep over every codec, four registers, twelve partials, the fine-tuned decoders and sixty NSynth notes is summarised in Table A16 and Figure A7. Both are computed from unrounded displace-

ments, so neither reproduces exactly from the printed rows of Table A15. Tables A14 and A16 report $\bar{\ell}$ on different scopes and need not agree: the former is 440 Hz at $\delta_0 = \pm 40$ exactly, where the latter is a single-register reading, 440 Hz except where a row says otherwise, over every $|\delta_0| \geq 20$.

At 24 kbps the edge is H5, but H8 displaces 4.4 cents and, on the sharp side, H5 and H6 displace 1.7 and 5.9, giving $\bar{\ell} = 0.86$ and 0.58. Restricted to partials that moved at all (a 20-cent cut) the same rows give 1.11 and 1.07. So on this reading, which conditions on displacement and is an upper bound (Appendix G.6), regenerated partials go most of the way to the grid harmonic at every bitrate; what falls with bitrate is how many of them there are. We keep the inclusive definition because Equation 15 needs it — the estimator averages over its whole partial set — but the inclusive $\bar{\ell}$ is not "how far a regenerated partial moves". The 1.5 and 3 kbps rows are unaffected.

Three accounts of a residual in a harmonic-fit estimator's reading differ in the high-resolution spectrum of the decoded tone: a moved fundamental gives one displaced line, an added grid component gives a second line beside an unmoved one, and regenerated partials leave the fundamental where it was and move the lines above a band edge. `spectral_check.py` synthesises twelve-partial tones at 440 and 220 Hz, detuned by $\delta \in \{0, \pm 20, \pm 30, \pm 40\}$ cents, codes them through EnCodec 24 kHz at three bitrates, and locates each partial to 0.09-cent bins at 440 Hz under the 4.4 Hz main lobe of the 0.9 s Hann window of Section 2.5. At every bitrate and detuning the decoded line at f_0 sits within 0.2 cents of the input *on the mean over repetitions*, which is what Table A15 prints, with its -3 dB width unchanged, and the level at the 12-TET frequency is 25 to 60 dB below it. Single repetitions scatter wider than that mean, and we give the spread rather than the bound alone: the largest at 440 Hz is 0.8 cents at 1.5 kbps, and over the whole sweep 2.2 against a median of 0.05 (`spectral_sweep.csv`). The claim rests on the mean — that the fundamental carries no displacement locked to the detuning — and not on no repetition ever moving. Above the band edge the line at the input frequency is 10 to 40 dB down and a new line sits near, in the cleanest cases exactly on, the corresponding harmonic of the nearest 12-TET fundamental; the partial at the edge shows two lines of comparable level. Both of Section 2.4's thresholds are referred to the input partial and not to the strongest decoded line, which matters at the edge: the energy there divides between the two lines, so they sit within a few decibels of each other while both fall more than 10 dB below the input.

Table A15: Displacement of each decoded partial, cents, at 440 Hz and $\delta_0 = \pm 40$; means over repetitions, so a partial carrying two lines (H3 at 3 kbps) reads between them. Est. 8 is our estimator on all eight partials, Est. 1–2 on the first two.

Bitrate	δ_0	H1	H2	H3	H4	H5	H6	H7	H8	Est. 8	Est. 1–2
1.5 kbps	-40	-0.0	$+46.5$	$+39.0$	$+35.5$	$+49.2$	$+42.2$	$+39.8$	$+39.9$	$+24.4$	$+4.3$
1.5 kbps	$+40$	-0.2	-29.9	-40.8	-24.6	-34.9	-33.9	-34.3	-28.8	-18.1	-2.0
3 kbps	-40	-0.0	$+0.0$	$+26.9$	$+40.2$	$+51.9$	$+42.1$	$+35.1$	$+44.6$	$+20.7$	-0.0
3 kbps	$+40$	-0.2	-0.0	-24.0	-37.9	-27.5	-35.1	-42.8	-37.5	-16.9	-0.1
24 kbps	-40	$+0.0$	$+0.0$	-0.0	$+0.0$	$+53.2$	$+43.3$	$+37.2$	$+4.4$	$+7.3$	$+0.0$
24 kbps	$+40$	$+0.0$	-0.0	$+0.1$	-0.9	-1.7	-5.9	-42.8	-42.4	-5.9	-0.0

A decoder whose grid sits displaced from $f^\star$ reads the two signs of detuning unequally, so the half-sum of the *Est. 8* column estimates that displacement: $(24.4 - 18.1)/2 = 3.2$ cents at 1.5 kbps and $(20.7 - 16.9)/2 = 1.9$ at 3 kbps, against the 6.9 crossing of the original-clips arm (Table A25) — an arm fine-tuned on GTZAN, where the half-sum is read on stock EnCodec, so the two are not matched and the gap between them is not all displacement. The 24 kbps row gives 0.7 on a residual too small to carry the statistic.

The vocoder is the one condition whose *fundamental* moves: reading BigVGAN's H1 through the same spectral procedure gives 3.1 [2.4, 3.8] cents, against $|s_1| \leq 0.2$ for every EnCodec row of Table A15.

Part of both the half-sum and the 6.9-cent crossing is arithmetic rather than corpus. Equation 4 refers ℓ_k to $kf^\star$, but $\{kf^\star\}$ is not the chromatic grid: $5f^\star$ sits 13.7 cents below the nearest 12-TET pitch and $7f^\star$ 31.2 cents below it, while $k = 3, 4, 6, 8$ sit within 2 cents. Over the six partials above a 3 kbps edge the mean gap is $+6.83$ cents if every partial lands on the chromatic pitch, and $+1.6$ if $k = 7$ follows $7f^\star$ as Table A15 shows; the half-sum statistic above gives 1.9. So part of the 6.9-cent crossing is arithmetic present with no codec in the loop, and the rest arises in the coded path. A partial landing on the 12-TET pitch nearest $5f^\star$ scores $\ell_5 = 1.34$ at $\delta_0 = -40$ and 0.66

at $+40$, which is the direction every row of Table A16 takes and a source of the $\bar{\ell} > 1$ overshoots. We therefore do not read the 6.9-cent crossing as corpus evidence on its own; the retuning arms of Appendix E.1, which move the crossing by a predicted amount, are what carry that claim.

Table A16: **Band edge and ladder fraction** (others: Tables A20, A21). Edge: mean over resolving registers. $\bar{\ell}$ per sign and their mean over $|\delta_0| \geq 20$, at 440 Hz, DAC 16k's at 880 Hz; brackets, BCa over up to twelve clusters, read as spread (Appendix G.3); dash, absent. *Ref.*: grid read against.

			Ladder fraction			
Codec	**Operating point**	**Edge (kHz)**	$\delta_0 < 0$	$\delta_0 > 0$	**Mean**	**Ref.**
EnCodec 24k	1.5 kbps	0.9 kHz	0.94 [0.88, 0.99]	0.81 [0.71, 0.90]	0.88	12-TET
EnCodec 24k	3 kbps	1.3 kHz	0.96 [0.91, 1.00]	0.84 [0.76, 0.92]	0.90	12-TET
EnCodec 24k	6 kbps	1.6 kHz	0.93 [0.86, 0.99]	0.80 [0.69, 0.91]	0.87	12-TET
EnCodec 24k	24 kbps	2.2 kHz	0.86 [0.75, 0.96]	0.58 [0.42, 0.74]	0.72	12-TET
EnCodec 32k	2.2 kbps	1.4 kHz	0.92 [0.86, 0.98]	0.83 [0.74, 0.92]	0.88	12-TET
EnCodec 48k, music	6 kbps	1.6 kHz	0.91 [0.83, 0.99]	0.66 [0.52, 0.80]	0.79	12-TET
EnCodec, quantiser bypass	–	2.2 kHz	0.84 [0.75, 0.93]	0.71 [0.58, 0.84]	0.78	12-TET
EnCodec, fine-tuned on GTZAN	3 kbps	2.0 kHz	0.93 [0.87, 0.99]	0.84 [0.75, 0.93]	0.89	12-TET
EnCodec, fine-tuned on flattened GTZAN	3 kbps	2.0 kHz	0.62 [0.50, 0.74]	0.55 [0.41, 0.69]	0.59	12-TET
EnCodec, fine-tuned on semitone GTZAN	3 kbps	2.0 kHz	0.92 [0.85, 0.99]	0.84 [0.74, 0.94]	0.88	12-TET
EnCodec, fine-tuned on 0/100-cent GTZAN	3 kbps	2.0 kHz	0.90 [0.82, 0.98]	0.82 [0.71, 0.93]	0.86	12-TET
EnCodec, fine-tuned on $+33$-cent GTZAN	3 kbps	2.0 kHz	0.91 [0.84, 0.98]	0.82 [0.72, 0.92]	0.87	own grid
EnCodec, fine-tuned on 24-TET GTZAN	3 kbps	2.0 kHz	0.88 [0.79, 0.96]	0.79 [0.67, 0.90]	0.84	own grid
EnCodec, fine-tuned on Saraga	3 kbps	2.0 kHz	0.85 [0.76, 0.94]	0.76 [0.64, 0.88]	0.81	own grid
DAC 16k	Q6	4.4 kHz	0.89 [0.78, 1.00]	0.81 [0.68, 0.94]	0.85	12-TET
WavTokenizer	0.9 kbps	0.9 kHz	0.88 [0.81, 0.95]	0.79 [0.69, 0.89]	0.84	12-TET
BigVGAN (vocoder)	mel input	2.6 kHz	0.71 [0.60, 0.82]	0.64 [0.51, 0.77]	0.68	12-TET
Mimi	Q8	3.4 kHz	0.79 [0.66, 0.92]	0.71 [0.56, 0.86]	0.75	12-TET
SNAC 32k	default	3.1 kHz	0.90 [0.79, 1.01]	0.82 [0.69, 0.95]	0.86	12-TET
EnCodec 24k, NSynth notes	3 kbps	1.3 kHz	0.83 [0.72, 0.94]	0.74 [0.60, 0.88]	0.79	12-TET

C.5.1 A SATURATING PULL AS AN ALTERNATIVE TO THE LADDER

One alternative to a ladder follows from a clipped model of the residual, and the readings used elsewhere in the paper do not separate it from a ladder. Equation 4 scores a partial against the distance from the tone to the nearest grid pitch, and every tabulated ladder fraction is read at $|\delta_0| \geq 20$ cents. But (P1) failed against a clipped model whose knee sits at 22 [19, 26] cents, so those readings run from just below the knee upward, largely where the residual has saturated. A regenerated partial displaced by a bounded number of cents reproduces the same $\bar{\ell}$ there: at $|\delta_0| = 40$ a cap of about 36 cents and a ladder at 0.90 are the same measurement. One version of that alternative is already excluded. A displacement with no grid in it has a fixed direction, so it would read $\bar{\ell}$ of opposite signs at the two detuning signs, and every row of Table A16 is positive at both — the same argument Appendix E.1 makes against a global transposition. What survives is a pull toward the grid that saturates, which is not a rival to the grid claim but a statement about how far $\bar{\ell}$ may be extrapolated.

Two things weigh against the cap indirectly: the retuning arm of Appendix E.1 moves the whole structure by 33 cents onto a location named in advance, which a cap fixed in the codec does not predict, and the partials land near grid pitches rather than at a constant offset from the input. Neither is decisive on its own, so we read $\bar{\ell}$ as a function of $|\delta_0|$ directly, where a pull saturating at c cents reads $\min(c/|\delta_0|, 1)$. Read at twelve signed detunings from 5 to 50 cents at the 440 Hz reference and 3 kbps, on a run that reproduces Table A15 at the same rate and reference, $\bar{\ell}$ is flat at about 0.87 across the range.

The low end carries the result. Below the clipped model's 22-cent knee they separate: any cap at or above the knee is pinned at $\bar{\ell} = 1.00$ at $|\delta_0| = 5$ and 10, since the partial reaches the grid and can go no further, while a proportional ladder reads its usual value. It reads 0.87. The high end says the same thing with more room: a 22-cent cap requires 0.44 at $|\delta_0| = 50$ against 1.00 at 20, a factor of two, and the measured column does not fall.

We therefore read the ladder as proportional and not as a bounded displacement, on this evidence rather than on the two indirect arguments above, and $\bar{\ell}$ is not a statement confined to $|\delta_0| \geq 20$. The reading is 0.87 against the 0.90 quoted from the primary condition, a difference inside the spread these readings carry.

C.5.2 SENSITIVITY OF THE BAND EDGE TO THE LEVEL THRESHOLDS

Equation 4 calls a partial transmitted when its decoded line is within 6 dB of the input's and regenerated when it is more than 10 dB down with the strongest nearby line elsewhere. Both numbers are choices, and the band edge is the paper's central explanatory variable, so we sweep them. Table A17 gives the EnCodec 24 kHz edge at three bitrates under three threshold pairs, tightened to 3/6 dB and loosened to 10/15 dB.

The placement clause is the one ingredient Section 2.4 names as conditioning on the estimand, so it is swept alongside the two dB numbers rather than held fixed. Dropping the clause at each of the three pairs moves the 3 kbps edge to 0.88 kHz at all three and the headline $\bar{\ell}$ to 0.79, 0.78 and 0.77 from tight to loose. So the clause is worth more than the thresholds are — about 0.12 of ladder fraction against 0.02 across the dB sweep — and the two do not interact. Dropping the clause moves the edge by one partial, within the one-partial tolerance of Table A17. The clause's cost is in $\bar{\ell}$, so a reader who rejects it should read every $\bar{\ell}$ in this paper as up to 0.17 high and every edge as an upper bound, with the tier separation intact (Appendix G.6).

Table A17: Band edge (kHz) for EnCodec 24 kHz under three level-threshold pairs. The edge moves by at most one partial and the ordering across bitrate is never reversed, so no conclusion in Section 3.2 turns on the choice.

Thresholds (transmitted / regenerated)	**1.5 kbps**	**3 kbps**	**24 kbps**
3 / 6 dB (tight)	0.9	1.3	2.0
6 / 10 dB (reported)	0.9	1.3	2.2
10 / 15 dB (loose)	1.3	1.3	2.2

A partial more than 10 dB down with no strong line within 100 cents is neither transmitted nor regenerated by these rules. We count it as *deleted*; it enters neither $\bar{\ell}$ nor the band edge, and deletions are under 4% of partials at every bitrate reported. Where no partial is regenerated at all, the band edge is right-censored at $0.95\, f_s/2$, the highest partial the generator admits. Counted over spectral-sweep tones the rate is 0% for EnCodec at every bitrate and 12% for SNAC 32k, whose edge sits at 3.1 kHz. It is complete for SNAC 44k, DAC 24k and the two vowel conditions, whose partials the level rules classify as neither transmitted nor regenerated at any reference we ran. Those four Table A16 omits for right-censoring, as it omits SNAC 24k and DAC 44k under the guards of Appendix B.8, which Table A40 reads instead; Figure 2 marks six rows *not resolved*: those four plus Opus and MP3, which lie outside the spectral sweep.

C.5.3 MEASURING THE EDGE ON A CODEC WE DO NOT TABULATE

Everything above describes how we measured f_{e} on the conditions of Table A16. A reader whose codec is not in that table needs the same procedure stated in one place, so here it is.

1. Synthesise a harmonic complex at a reference where the codec's plausible edge falls inside the partial set. 440 Hz with eight partials is our default; if the edge lands above every partial, move up an octave, as we do for DAC 16k, whose 4.4 kHz edge exceeds the eighth partial at 440 Hz and forces a reading at 880 Hz.
2. Code and decode it, and take the spectrum on a 0.9 s Hann window at 0.09-cent bins.
3. Around each input partial, classify the strongest line within 100 cents by the rule of Section 2.4, which we do not restate here so that the two cannot drift apart. It scores level and placement jointly, so a full-level partial that has moved counts as transmitted; reading the clauses disjunctively instead would call it regenerated and put the edge lower.
4. The edge is the lowest regenerated partial, right-censored at $0.95 f_{\mathrm{s}}/2$ when no partial qualifies.

A practitioner inherits three things, and they bound what the reading is worth. The first is that the edge is binned: every edge in Table A40 is one of five values, two to six times the 440 Hz reference, and quantities read across checkpoints inherit that: the correlation with regenerated count, the tier rule and the eight checkpoints we describe as sitting at EnCodec's own edge are all statements about a variable binned at 440 Hz, so decoders whose true edges differ by up to a few hundred hertz are

called equal before any of it is computed. Reading each edge again at an incommensurate reference would interleave the bins. The second is resolution: the edge resolves only to the partial spacing, which is f_0, so it is known to $\pm f_0/2$. The third is threshold sensitivity: it moves by at most one partial across the threshold range we swept (Table A17), which is the tolerance the 6 and 10 dB choices carry.

C.6 The uniform-weight reference

The refined estimator weights partial k by $w_k \propto k^2|X_k|^2$ with X_k the decoded amplitude, so for small displacements its residual is the weighted mean of the per-partial displacements. A regenerated partial returns 4.5 to 7 dB down — not the 10 to 40 dB by which its *input-frequency* line has fallen, which is a different line — so those weights are not known before the measurement. We therefore write the reference quantity with the *input-referred* weights $w_k^{\text{in}} \propto k^2|X_k^{\text{in}}|^2$, which for a $1/n$ stimulus are uniform over the estimator's eight partials. With the reference on the grid ($d = 0$, so $\delta_0 = \delta$), that quantity is

$$b^{\text{unif}}(\theta) \;=\; |\delta_0|\;\bar{\ell}\;\frac{\sum_{kf_0 \geq f_e} w_k^{\text{in}}}{\sum_k w_k^{\text{in}}}, \tag{15}$$

where $\bar{\ell}$ is the mean ladder fraction over the partials at or above the band edge, transmitted ones included, and both sums run over the eight partials the estimator fits. This is not a theoretical bound but an identity: under uniform weights the right-hand side is exactly the mean of the measured per-partial displacements, and we use it as the reference against which the estimator's actual reading is compared. Two effects make the measured bias smaller. A regenerated partial comes back attenuated, so the decoded weights favour the transmitted partials the input-referred weights treat equally; and a partial near the edge is a split line, which the least-squares fit reads between its two positions rather than at the stronger one Equation 4 takes. Both push the reading below b^{unif}, and Table A14 reports by how much, at both signs of δ_0.

C.7 The headline effect over a population of stimuli

We sweep six spectral envelopes — the three of the main protocol (Section 2.5) and three more — across the four registers of Table A13 at 3 kbps, twenty-four stimulus cells, and resample both margins of that crossed design (`analyze_stimulus_population.py`; Appendix G.3 gives the three variants and why this is the one we report). The 880 Hz column carries the stimulus artefact of Appendix C.4, but dropping it leaves the median at 4.37 over the remaining eighteen cells, so the population figure does not depend on it.

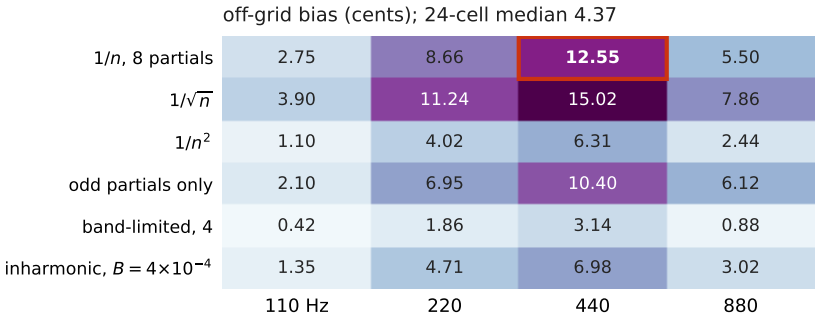

Figure A8: **The headline stimulus against its own population.** Table A18 as a map: rows are spectral envelopes, columns the four reference registers, colour repeats the printed value. The boxed cell is the tone we headline, at the 92nd percentile of the twenty-four.

The population median of Table A18 is **4.37 cents, [1.73, 9.13] over stimuli**. The bracket is a design width over six envelopes by four registers and not a coverage statement, since four register levels cannot calibrate one, and the cells themselves run 0.42 to 15.02. The headline stimulus, at 12.55, sits at the 92nd percentile of this population. It was chosen before any of this was measured, as a conventional harmonic complex at a conventional reference; it is not typical, and trial-level intervals cannot show that.

Table A18: Off-grid median bias in cents by envelope and register, EnCodec at 3 kbps; B is the inharmonicity coefficient. Margins are medians, so the boldface figure is the median of all 24 cells, not of the margins. The $1/n$ row at 440 Hz is the headline stimulus.

Envelope	110 Hz	220	440	880	Median
$1/n$, 8 partials (headline)	2.75	8.66	12.55	5.50	7.08
$1/\sqrt{n}$	3.90	11.24	15.02	7.86	9.55
$1/n^2$	1.10	4.02	6.31	2.44	3.23
Odd partials only	2.10	6.95	10.40	6.12	6.54
Band-limited, 4 partials	0.42	1.86	3.14	0.88	1.37
Inharmonic, $B = 4 \times 10^{-4}$	1.35	4.71	6.98	3.02	3.87
Register median	1.73	5.83	8.69	4.26	5.05
Median over all 24 cells					**4.37**

The ordering the mechanism predicts survives the sweep, which is the part that matters for the account rather than for the effect size. Envelopes with more energy above the edge read higher at every register ($1/\sqrt{n} > 1/n > 1/n^2$), the band-limited stimulus with four partials puts two above the edge at 440 Hz and reads 3.14, and the inharmonic stimulus — whose partials cannot all land on one 12-TET ladder — reads 6.98 against the harmonic 12.55 at the same register. Every one of these is the sign the band-edge account predicts. What it predicts here is the ordering and not the sizes: the soft-edge model of Appendix C.4 is fitted across registers, and nothing in it speaks to the envelope axis this sweep varies.

D CONTROLS

D.1 THE CLASSICAL-CODEC CONTROL

Section 1.2 gives the argument: what these schemes invent is copied from lines already present, so a grid bias has no route in. That is what we measure — a periodic residual of 1.57 cents whose phase does not point at the grid, with an off-grid bias of zero — and it is the control for the word *learned*.

We encode Opus with FFmpeg and libopus 1.4 in VBR mode at complexity 10, application audio, full band, forcing CELT-only mode for the 32 kbps row. Opus and MP3 contain no trained component, so the account of Section 2.6 predicts no pull from either. Opus at 6 kbps does not register (slope 0.929 [0.792, 1.065], Table A19); MP3 at 16 kbps does, on a 0.01-cent residual, and is the classical codec Section 2.3 cites. Both were run through ffmpeg at two bitrates each, on the eleven-reference detuning sweep — the same design as the headline EnCodec condition they are the control for, not the reduced design most of the other learned codecs use (Appendix B.4).

Table A19: Classical codecs through the registration sweep. Opus at 6 kbps is the only *fitted* condition above 0.05 cents; MP3 at 16 kbps fits at 0.01, and the two refused rows print no fit: an amplitude bound for Opus and a zero residual for MP3. Brackets are 95% intervals.

Codec	Amplitude (c)	Slope	95% CI	Grid bias (c)	Gated
Opus 6 kbps	1.57 ± 0.37	0.929	[0.792, 1.065]	0.00 [0.00, 0.42]	47%
Opus 12 kbps	≤ 0.26	*refused: amplitude unstable,* $R^2 \leq 0.002$		0.000	18%
MP3 16 kbps	0.01 ± 0.00	1.001	[0.977, 1.025]	0.000 [0.00, 0.00]	0%
MP3 32 kbps	0.00	*refused: no residual to fit*		0.000	0%

Opus at 6 kbps prints a lower bound equal to its point estimate because most bootstrap draws return exactly zero, not because the interval is one-sided.

The registration slope alone does not separate a learned grid from any structure tied to absolute frequency, which is why (P2), the octave test and the amplitude scale matter. A pull has zero crossings at the grid points and a positive grid bias; the Opus residual peaks at them and has none. Across octaves it does not behave like a pitch effect: its amplitude is 0.21 and 0.13 cents at 110 and 220 Hz, where the tracking gate removes 0% and 1.5% of trials, and 0.86 and 0.91 cents at 440 and 880 Hz, where the gate removes 40% and 66% on this octave sweep, against the 47% Table A8 records for the registration run at 440 Hz. The phase wanders from $-67°$ to $+52°$, and the amplitude scales

with codec failure, not with pitch. On real music (Table A30) Opus at 6 and 12 kbps shows a grid bias of $+0.07$ and $+0.02$ cents against $+2.49$ for EnCodec on the same clips.

D.2 A CLASSICAL SCHEME THAT REGENERATES: HE-AAC WITH SBR

Opus and MP3 control for the word *learned* but not for the word *regenerates*: at these rates they mostly transmit or discard, so a null there does not bear on regeneration. The sharper control is a classical codec with a genuine band edge and genuine synthesis above it and no learned corpus at all. HE-AAC's spectral band replication (Dietz et al., 2002) is exactly that: it copies the lower spectrum upward under a transmitted envelope, so what it puts above the edge comes from the input rather than from a corpus. If the ladder were a generic property of bandwidth extension rather than a learned one, SBR should show it. We encode with libfdk_aac as HE-AAC v1, with the SBR crossover set to 1.32 kHz at 32 kbps, EnCodec's edge at 3 kbps, and to 0.88 kHz at 24 kbps. Both were set through the library rather than an encoder flag, and Table A20's measured edges confirm them.

Table A20: Classical regeneration through the same pipeline, held to Appendix B.8's guards; each scheme has a real edge and so is a measured null. *cv/ret* are its two guard values. Brackets 95%, except ‡: a reference point, bracketed by its range over detuning signs.

Scheme	Point	Edge (kHz)	$\bar{\ell}$	Bias (c)	Slope	Gate	cv/ret
HE-AAC, SBR	32 kbps	1.32	0.04 [−0.07, 0.15]	+0.05 [−0.31, 0.42]	0.981	0.4%	0.07/1.00
HE-AAC, SBR	24 kbps	0.88	0.02 [−0.10, 0.14]	−0.02 [−0.44, 0.39]	1.014	1.1%	0.10/0.99
Opus, CELT	32 kbps	1.76	0.06 [−0.06, 0.18]	+0.08 [−0.28, 0.44]	0.994	2.8%	0.11/0.97
EnCodec 24k‡	3 kbps	1.32	0.90 (0.84 to 0.96)	+12.55	1.001	0.0%	0.01/1.00

It does not (Table A20). HE-AAC at 32 kbps matches EnCodec's 1.32 kHz edge with the same six partials above it; the 24 kbps and CELT rows sit at 0.88 and 1.76 kHz and so regenerate more and less of the spectrum respectively. All three displace those partials without moving them toward the grid, $\bar{\ell}$ within 0.06 of zero — copy-up misplaces against the input's ladder, which $\bar{\ell}$ does not score, and carries no 12-TET prior — against EnCodec's 0.90 at an identical edge.

This control has a specific scope. What goes to zero is the grid-directed component alone: the translation carries no 12-TET prior, so $\bar{\ell} \approx 0$ is entailed by the algorithm rather than discovered by the measurement, and it bounds the grid-directed part rather than the displacement. The arm is therefore a check that our pipeline reads a known-zero ladder as zero at a matched edge, and a demonstration that regeneration alone does not imply a grid — not an exclusion of the more interesting alternative, that any *generative* model filling an empty band must place partials somewhere and 12-TET is where a plausible prior lands. SBR instantiates no such prior, so it cannot test that. The discriminating experiment is a learned bandwidth extender at a matched edge, trained on Western, tuning-flat and synthetic audio, and Appendix D.3 runs it.

D.3 A LEARNED BANDWIDTH EXTENDER, TRAINED THREE WAYS

We train a bandwidth extender — convolutional, 4.1M parameters, mel-conditioned, HiFi-GAN discriminators (Kong et al., 2020) — to reconstruct everything above 1.32 kHz from the band below it. That edge is EnCodec's at 3 kbps, so the model faces the same six partials of a 440 Hz complex that EnCodec regenerates. Architecture, schedule, optimiser and step count are identical across arms; only the training corpus differs. Five seeds per arm (`analyze_bwe.py`).

Table A21: A learned extender at EnCodec's 1.32 kHz edge; only the corpus differs across arms. Arm intervals combine the five-seed spread with each arm's own within-arm estimation half-width. ‡A reference point, not an arm: its bracket is the range over detuning signs.

Trained on	$\bar{\ell}$	Off-grid bias (c)
Western music (GTZAN, as recorded)	0.68 [0.55, 0.81]	+8.10 [6.48, 9.72]
The same audio, tuning flattened	0.14 [0.00, 0.28]	+1.12 [−0.44, 2.68]
Synthetic harmonics, uniform tuning	0.09 [−0.05, 0.22]	+0.60 [−0.93, 2.13]
EnCodec 24k at 3 kbps, for reference‡	0.90 (0.84 to 0.96)	+12.55

Three things follow from Table A21, and they are worth separating.

A generative prior filling an empty band does not land on 12-TET by default. The flattened arm fills exactly the same band above exactly the same edge with exactly the same architecture, and leaves what it invents on the input's own ladder, $\bar{\ell} = 0.14$ [0.00, 0.28]. This is the alternative the SBR control could not test, and it is now bounded for one architecture: whatever a grid-free prior contributes is under 0.28 in ladder fraction and under 2.7 cents of bias. That bias bound's size depends on what it is read against, and we give both readings: against the headline stimulus's 12.55 cents the bound is a fifth of the effect, but against the 4.37 cents of Appendix C.7, the same instrument and statistic on a population rather than one tone, it is three fifths. The figures to compare are 0.14 and 1.12 cents, bounded by 0.28 and 2.7, against the grid corpus's 0.68 and 8.10. A bound is not an exclusion, but it is small enough that the grid is not plausibly a consequence of having to put invented partials somewhere.

The corpus is what carries the grid. The contrast between the two music arms is 0.54 [0.35, 0.73] in ladder fraction, from a manipulation intended to change nothing but the tuning of the training audio. This is the same conclusion as the fine-tuning experiment of Section 3.3, reached without a pre-trained decoder anywhere in the design, whereas every arm of Section 3.3 starts from a codec pre-trained on a corpus outside the experiment.

The synthetic arm bears on the residue of the flattened fine-tuning arm. That arm still placed partials 0.59 of the way onto the grid: flattening the corpus removed a third of the ladder rather than all of it. An extender trained on synthetic harmonics at uniform tuning, which never saw 12-TET in any form, reaches 0.09 [-0.05, 0.22]; the flattened-music arm, whose clips keep their internal 12-TET intervals, reaches 0.14. This is consistent with the residue being pre-training that fine-tuning on a grid-free corpus does not undo, as a competing grid does on the $+33$-cent and 24-TET arms (Section 3.3), but the extender is a weaker and differently scaled instance — 0.68 against EnCodec's 0.90 on the grid corpus, consistent with a smaller model on a smaller corpus — so the comparison does not isolate pre-training.

The extender experiment has three limits. The first concerns resampling. Flattening resamples each clip, so it changes tempo, formant scaling and spectral detail as well as tuning, and the phrase "nothing but the tuning" is demonstrated for the codec arms and asserted for these. Appendix E.1 closes that gap on the codec with grid-preserving resamplings, and the codec's own controls move the ladder fraction by at most 0.03, small against this arm's 0.54 contrast. The second is that this is one architecture at one edge, and a bandwidth extender is not a codec: it is handed a clean low band rather than a quantised one, and it never has to allocate bits. The third, the "grid removed versus harder to model" confound, which Appendix E.1 bounds for the fine-tuning arms and the retuning arm excludes outright, is addressed here by one observation: synthetic harmonics are the easiest of the three corpora to model and read the *lowest* ladder fraction, 0.09, so difficulty and ladder fraction are not monotone across the arms.

E THE CAUSAL EXPERIMENT

E.1 THE CAUSAL EXPERIMENT IN FULL

Protocol. Every arm fine-tunes the decoder alone for 4000 steps. The loss is a multi-resolution STFT loss — spectral convergence plus log-magnitude ℓ_1 at 512, 1024 and 2048 points — plus ten times the waveform ℓ_1, under AdamW at a learning rate of 10^{-5}, in batches of sixteen 0.5-second segments. Code assignment is an argmin, so it is unchanged by decoder training: we verified that the encoder and quantiser of every reported arm are bit-identical to stock, which is what makes the comparison a decoder experiment rather than a whole-codec one (Table A27).

Arms. The flattened arm shifts each clip upward by an offset uniform on $[0, 100)$ cents, which destroys the corpus's grid while preserving its spectral and rhythmic statistics. Three controls keep a grid: the original clips, the same clips shifted by exactly one semitone (which controls for the resampling operation, since it changes tempo and spectrum but leaves the density as peaked), and clips shifted by 0 or exactly 100 cents with equal probability (matched to the flattened arm in mean resampling magnitude). The retuning arms replace flattening with a fixed $+33$-cent shift, which keeps the density peaked but moves its grid off 12-TET, and with quantisation to 24-TET, each short

frame of the mix pitch-shifted so that its dominant F_0 lands on its nearest 50-cent point. Since 71% of GTZAN's pitch lies within 25 cents of a 12-TET note, the quantised corpus snaps those pitches to 12-TET points and the remaining 29% to quarter-tones: its density peaks at 7.3 times its mean at the 12-TET pitch and 3.4 at the quarter-tone, against 1.788 for GTZAN as recorded (Table A11).

What the intervals cover. Table A22's five seeds vary the fine-tuning initialisation and data order within *one* draw of the flattening offsets, so those intervals cover optimisation noise and not the draw. The corpus-level claim needs independent draws on both sides, which is why the reported contrast is the ten-draw one. Mimi's and WavTokenizer's arms carry seed-level spread only, so they establish that the direction replicates and not the size of the corpus component.

Tables A22 (four arms, five seeds), A23 (the ten draws), A24 (reconstruction), A25 (the retuning arms), A26 (DAC), A27 (which half carries the change) and A28 (the controls that bound the measurement) follow.

Table A22: **Causal experiment: four arms, five seeds.** Fitted amplitude in cents, registration slope in parentheses. The seeds vary initialisation within *one* draw of the flattening offsets, so these intervals cover optimisation noise; the contrast in the text is the ten-draw one.

Fine-tuned on	Peak/mean	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4	Mean $\pm$ sd
Original clips (grid)	1.788	4.41 (0.999)	4.50 (0.999)	4.45 (0.998)	4.59 (0.999)	4.61 (0.998)	4.51 ± 0.09
Shifted by one semitone (grid)	1.788	4.65 (0.996)	4.68 (0.996)	4.60 (0.996)	4.59 (0.996)	4.60 (0.997)	4.62 ± 0.04
Shifted by 0 or 100 cents (grid)	1.788	4.74 (0.997)	4.65 (0.997)	4.65 (0.997)	4.68 (0.996)	4.75 (0.995)	4.69 ± 0.05
Random offsets (flat)	1.073	3.91 (0.997)	3.68 (0.996)	3.64 (0.997)	3.80 (0.998)	3.71 (0.998)	3.75 ± 0.11

Table A23: **The corpus-level contrast, draw by draw.** Ten independent draws of the flattening offsets, each fine-tuned against a grid-preserving control on the same draw. Brackets are a 95% $t(9)$ interval on the paired difference.

Draw	1	2	3	4	5	6	7	8	9	10	Mean
Grid control (c)	4.72	4.65	4.71	4.68	4.74	4.63	4.70	4.66	4.73	4.67	4.69
Flattened (c)	3.55	3.95	3.72	3.86	3.63	3.88	3.76	3.62	3.85	3.97	3.78
Difference (c)	1.17	0.70	0.99	0.82	1.11	0.75	0.94	1.04	0.88	0.70	**0.91** [0.79, 1.03]

What the arms cost in reconstruction, and what an aggregate score sees. A corpus manipulation that changed how well the decoder reconstructs would offer a rival reading of the whole causal result: the flattened arm might place its partials differently because it is simply a worse decoder. Table A24 bounds that reading, and one arm in it excludes the version that matters. Every arm is scored on the same 200 held-out GTZAN clips, none of them in any fine-tuning set.

The near-equality comes first and settles least. Against the original-clips arm the flattened arm is 0.005 worse on the objective, 0.9 standard errors of the difference of means (Welch $t = 0.87$, $p = 0.41$), while its ladder fraction is 0.30 lower; against the 0/100-cent comparator of Table A37 the two figures are 0.002 and 0.27. We do not read a rank correlation across the six arms: five of them sit within 0.05 of each other on $\bar{\ell}$, so at $n = 6$ that statistic has no power in either direction.

Two arms do the work instead. The first is a bound rather than a null. Eight EnCodec decoders were fine-tuned in all, and the paper counts seven corpus arms: the six GTZAN variants and Saraga. The eighth is a training-duration control, and fine-tuning on the grid-preserving corpus but stopping at 1200 steps costs 0.017 on the objective against the fully trained control (Welch $t = 2.79$, $p = 0.026$ unadjusted, and selected by inspecting the table, so it should be read as descriptive) and leaves the ladder fraction at 0.88 against that control's 0.89. The flattened arm's own deficit is 0.005 with a standard error of 0.006, so its 95% upper bound is 0.017, the same as the deficit we have just shown does not move the ladder, so the flattening arm's reconstruction cost is at most about the size of a deficit shown harmless.

That bound covers the flattening arm and not the Saraga one, whose deficit is 0.022 against the fully trained control of Appendix E.1, past both the 0.017 shown harmless and the flattened arm's upper bound. Saraga is out of domain, so some deficit is expected; what it means is that the Saraga arm's 0.81 against its own svara histogram is not protected by this argument. It rests instead on direction,

as the retuning arm does: a decoder that merely reconstructed worse would not move its ladder onto the scale degrees of the corpus it was fine-tuned on while reading 0.22 against 12-TET.

The second is the retuning arm, and it is the stronger of the two. In Table A24 the $+33$-cent arm reconstructs at least as well as its control — 1.041 on the objective against 1.046 and 1.16 above the edge against 1.18, differences inside the seed spread either way — and its partials sit at 0.87 against its own retuned grid and 0.11 against 12-TET, the latter at both detuning signs. A reconstruction deficit cannot displace a ladder by 33 cents onto a location named before the arm was run, in a decoder that reconstructs as well as the arm it is compared with. For the translation manipulation the rival is not bounded but excluded.

For the flattening manipulation it is bounded and not excluded, and the bound has a shape worth stating: flattening a corpus gives it more distinct pitch positions to fit, so "grid removed" and "harder to model" covary by construction. The under-trained arm cannot speak to that, because its degradation route is under-training from a grid-bearing initialisation and both ends of that trajectory read high, 0.90 un-fine-tuned and 0.89 trained. Settling it needs a degradation that does not itself supply a grid, trained to convergence.

We report the objective above the band edge as well as full-band because the full-band figures cannot settle this on their own, and it would be circular to claim they do. Section 2.1 argues that a corpus-averaged spectral loss sees the residual's distribution and not the map; a loss with that property cannot certify placement in the invented band, and its near-equality across arms is entailed rather than informative. What Table A24 does measure is that argument: the quantity we report moves by a third where the objective the decoder is trained on moves by under one standard error. One column of Table A24 needs care: its $\bar{\ell}$ is read against each arm's own grid, so the retuned and 24-TET rows are not 12-TET readings. The $+33$ arm reads 0.87 against the grid it was trained on (Table A16) and 0.11 against 12-TET (Figure A9).

Table A24: **Reconstruction quality of the fine-tuning arms.** EnCodec 24k at 3 kbps on 200 held-out GTZAN clips. LSD is above the edge; LSD and Mel lower-better, ViSQOL MOS-LQO. Five seeds, own-grid $\bar{\ell}$ too but for the top and 1200-step rows. *Edge 1.3 kHz. †Out of domain.

Decoder fine-tuned on	mr-STFT ↓	$\text{LSD}_{>fe}$ ↓	Mel ↓	ViSQOL ↑	$\bar{\ell}$
— (not fine-tuned)*	1.121	1.42	0.941	3.71	0.90
GTZAN, grid kept	1.046 ± 0.008	1.18	0.884	3.77	0.89 ± 0.02
GTZAN, tuning flattened	1.051 ± 0.010	1.21	0.887	3.76	0.59 ± 0.02
GTZAN, shifted $+33$ cents	1.041 ± 0.008	1.16	0.881	3.78	0.87 ± 0.01
GTZAN, quantised to 24-TET	1.044 ± 0.009	1.19	0.883	3.77	0.84 ± 0.01
GTZAN, shifted one semitone	1.052 ± 0.008	1.22	0.888	3.76	0.88 ± 0.01
GTZAN, shifted 0 or 100 cents	1.049 ± 0.009	1.20	0.886	3.76	0.86 ± 0.01
GTZAN, grid kept, 1200 steps	1.063 ± 0.011	1.29	0.897	3.73	0.88
Saraga†	1.068 ± 0.011	1.31	0.903	3.72	0.81 ± 0.01

Some of the drop any corpus produces is the edge moving. Fine-tuning on any corpus takes the amplitude from 13.72 (Table A27) to 4.51–4.69 on the three GTZAN grid controls (Table A22; 4.12 to 4.69 over the EnCodec arms of Table A25 other than the flattened and 24-TET ones), a ratio of 0.33, and raises the edge from 1.3 to 2.0 kHz, which at a 440 Hz reference takes the partials at or above it from six of eight to four while leaving $\bar{\ell}$ at 0.89. Equation 15 then predicts a ratio of $(0.89 \times 4/8)/(0.90 \times 6/8) = 0.66$, not 0.33. Read off that equation the edge shift would account for about half the drop; the input-referred correction below moves the same reading to four fifths, which is why Section 3.3 puts no share on it at all. It does not touch the flattened-versus-grid contrast, which is read at a matched 2.0 kHz edge.

The identity does not carry onto the amplitude scale, in either direction. Both arms sit at the 2.0 kHz edge with four of eight partials above, so the partial-count term cancels and Equation 15 predicts an amplitude ratio equal to the ladder ratio. Read on the matched comparator throughout, whose $\bar{\ell}$ is 0.86 against the flattened arm's 0.59, that ratio is $0.59/0.86 = 0.69$: a flattened amplitude of 3.22 and a contrast of 1.47 cents. The measured contrast is 0.91 [0.79, 1.03], whose interval excludes that. A saturation artefact does not rescue the gap, on the argument that a sinusoid fitted to a clipped response overshoots more where the response is larger: that argument fails twice over. It has the sign backwards — an overshoot growing with the response puts the amplitude ratio *below*

0.69, where the measured ratio $3.78/4.69 = 0.81$ is above it, 3.78 being the ten-draw mean of Table A23 where 3.75 is the five-seed mean of the same arm — and it has no room to act, because for a vertically scaled clipped profile the overshoot depends on the knee's position rather than the plateau's height, so it cancels from any ratio. The stock condition's $13.72/12.55 = 1.09$ is that overshoot at the fitted 22-cent knee.

The identity meets a measured contrast twice in this appendix and misses both times, in opposite directions. On the fine-tuning drop it *under*-predicts: a predicted ratio of 0.66 is a predicted drop of 34% against a measured 67%. On the flattening contrast it *over*-predicts: the ladder ratio of 0.69 gives 1.47 cents against a measured 0.91.

One of those two misses is smaller than it looks, and the correction is a quantity we already report. Equation 15 is written on input-referred weights, so a ratio of *measured* amplitudes equals a ratio of b^{unif} only where the estimator's attenuation cancels. It cancels on the flattening contrast, whose arms both sit at 2.0 kHz with four of eight partials above. It does not cancel on the fine-tuning drop, which compares six of eight above at 1.3 kHz against four of eight at 2.0 kHz, and Table A14 gives that factor as 0.66 to 0.69 at six partials and 0.42 to 0.50 at four. Folding it in moves the prediction from 0.66 to about 0.45 against the measured 0.33 — a quarter off, not a factor of two.

What survives is that the identity still misses in both directions after the input-referred correction, and that a knee refitted per arm remains an open route: the clipped profile's knee is fitted for the stock condition alone. We therefore claim Equation 15 for the ordering it gets right and not for magnitudes on the amplitude scale.

The effect size, per arm. Whether the amplitude contrast carries onto the bias scale is settled by reading the off-grid median bias on each fine-tuned decoder rather than inferring it from the fitted amplitude. Amplitudes below are the ten-draw means of Figure 3b for the two arms it contrasts, so that contrast is the one Section 3.3 quotes, and the five-seed mean of Table A25 for the $+33$ arm, which has no ten-draw run; that table prints five-seed means for the other arms too, and the flat arm's $\tilde{b}$ of 3.75 happens to equal that table's five-seed *amplitude* for it, a coincidence of two decimals and not the same quantity; the magnitude-matched arm's $\tilde{b}$ of 4.62 collides the same way with the semitone arm's five-seed amplitude, and is likewise a different quantity on a different arm. That table also gives the flat arm no crossing at all: its residual does not pass zero over the range, so the arm has no located attractor. An off-grid median survives that, being read at $|\delta| \geq 30$ as a displacement magnitude and not as a crossing, and the absence is the point — it is why the flat arm's 3.75 is subtracted from the grid arm's 4.62 below rather than credited to a grid of its own:

Arm	**Amplitude A (c)**	$\tilde{b}$ **(c)**	$\tilde{b}/A$
Stock, no fine-tuning	13.72	12.55	0.915
Shifted by 0 or 100 cents	4.69	4.62 [4.03, 5.21]	0.985
Shifted by $+33$ cents	4.55	4.49 [3.91, 5.07]	0.987
Random offsets (flat)	3.78	3.75 [3.20, 4.30]	0.992

The ratio $\tilde{b}/A$ is *not* constant — 0.915 at stock against 0.985 to 0.992 on the arms — which is the saturation noted above, the knee having been fitted to the stock condition alone: the overshoot depends on the knee's position, so a ratio that changes means the arms' clipped profiles are not vertical scalings of the stock one. So an amplitude contrast does not carry onto the bias scale *between* stock and an arm, and we do not carry it there. But the corpus contrast is between two arms that sit within a cent of each other in amplitude, where the ratio is flat to within 0.007, and it survives the transfer: on the bias scale the isolable corpus component is $4.62 - 3.75 = 0.87$ cents [0.74, 1.00] over the same ten draws, on the two arms Figure 3b contrasts at $4.69 - 3.78$, against 0.91 [0.79, 1.03] in amplitude, 6.9% of the stock bias against 6.6% of the stock amplitude. Between two arms the transfer leaves the number unchanged.

One asymmetry belongs here rather than in the main text. At the matched 2.2 kHz edge the quantiser bypass and EnCodec at 24 kbps agree on the bias, 5.66 against 5.73, but not on the ladder fraction: 0.84/0.71 against 0.86/0.58 (Table A16), a gap of 0.13 on the sharp side. The bias equivalence holds, and the mechanism variable supports it less closely; twelve clusters do not resolve whether that gap is real.

[§]Mimi's stock $\bar{\ell}$ is 0.75 (Table A16) and its fine-tuned arms read 0.31. Fine-tuning raises every decoder's edge, and Mimi's is 3.4 kHz to begin with, so after fine-tuning few of the probe's eight partials clear it: the 0.31 is a mean over one or two partials and is not comparable with the stock 0.75. The transfer claim rests on the crossing, which moves as predicted, not on that $\bar{\ell}$.

Every row of Table A25 is a 100-cent-period fit except the 24-TET arm, whose amplitude, slope and crossing are read at the 50-cent period its corpus imposes; those three entries are therefore not comparable down their columns with the rest.

The ± 2.8 quoted on the displacement in Section 3 is a nominal half-width on the *difference* of two crossings, combined in quadrature from the $+33$ arm's [37.1, 42.0] and the unshifted arm's [5.4, 8.3]; both are five-seed percentile bootstraps, which Table A36 reads as spread rather than coverage, so it is a seed-level spread and neither a 95% nor a corpus-level interval.

Is the flattening contrast placement, or count? The inclusive $\bar{\ell}$ is the conditional fraction scaled by the regenerated fraction, up to a nuisance bounded by 0.11 (Appendix G.6). The ladder fractions of the fine-tuned arms are means over the same five seeds as Table A24, where the matched-comparator contrast reads 0.27 ± 0.01 across seeds (paired $t(4)$ interval [0.26, 0.29]) and the grid-kept contrast 0.30 [0.29, 0.31]; the amplitude contrast of Table A23 has ten independent draws per arm. What follows tests the 0.86-to-0.59 drop on the matched comparator, 0.89 against the grid-kept arm, against a confound.

Fine-tuning raises the edge, so a drop from 0.89 to 0.59 could in principle be the flattened decoder regenerating *fewer* partials rather than placing them differently. It is not. At the 2.0 kHz edge the seven EnCodec arms share, four of the probe's eight partials lie above, and all four are classified regenerated in each: $r = 4$ of $a = 4$ throughout, so $\bar{\ell}_{\mathrm{r}} = \bar{\ell}$ identically and the count term cannot carry any of the contrast. Mimi's and WavTokenizer's arms sit at their own edges and are not part of this comparison. The same holds for the extender at its 1.32 kHz edge, where six of eight lie above and all six are regenerated in each of the three arms. The contrast is where the partials went, not how many of them there were — which is the one thing the identity of Appendix G.6 would otherwise have left open, and the reason we report the counts here rather than only the fractions.

The grid-preserving controls. A $+33$-cent shift moves the corpus grid; a shift of exactly one semitone, or of 0 or 100 cents with equal probability, resamples the corpus by a comparable or larger amount and maps 12-TET onto itself. If the crossing followed the *resampling* rather than the *grid*, those arms would move it too. They do not: they read 7.1 [5.5, 8.7] and 6.6 [4.9, 8.3] against the unshifted arm's 6.9 [5.4, 8.3], displacements of $+0.2$ and -0.3 cents where the $+33$ arm displaces 32.7. Their own-grid and 12-TET ladder fractions coincide, at 0.88 and 0.86, because a semitone shift leaves the grid where it was. This is the specificity control for the retuning result; it was run for the amplitude comparison of Table A22, and its crossing was read afterwards.

The transposition alternative. A decoder fine-tuned on audio resampled uniformly upward by 33 cents could move its crossing by 33 for a reason that has nothing to do with a grid: it might simply have learned to transpose everything it emits by 33 cents. That account predicts the headline number exactly, so it has to be excluded rather than assumed away, and two measurements do it. First, a global transposition would carry the fundamental with it. It does not move: s_1 reads $+0.2$ $[-0.5, +0.9]$ cents on the $+33$ EnCodec arm, $+0.3$ $[-0.6, +1.2]$ on Mimi's and -0.1 $[-0.9, +0.7]$ on WavTokenizer's, against the $+33$ a transposition requires and within noise of the stock decoder's -0.0. The arms move their upper partials and leave the transmitted band where it was, which is the same dissociation the stock codec shows. Second, a transposition is a signed constant, so at $\delta_0 = \pm 40$ it would push the ladder fraction to opposite signs; the $+33$ arm instead reads 0.91 and 0.82, the same sign at both detunings (Table A16). A constant offset cannot produce that and an attractor can.

Table A27's second and third rows need an encoder that is *not* bit-identical, so they are drawn from a separate control in which the whole model was fine-tuned on each corpus; only that control's encoder is swapped in, and the quantiser and decoder beside it are stock. Without those rows the table would compare each arm against itself. Its *grid* rows are the original-clips arm, so its last two amplitudes are Table A22's five-seed means for original clips and for random offsets. They are not the pair Figure 3b contrasts, which reads the magnitude-matched control at 4.69 in place of the

Table A25: **Retuning the corpus moves the attractor.** *Crossing*: where the residual passes zero; dash, absent; brackets, five-seed spreads, not coverage (Appendix G.3). Each +33 arm was predicted at its unshifted arm's crossing plus 33. The 24-TET row is a 50-cent fit; ladder fractions, Figure A9.

Fine-tuned on	Amplitude (c)	Slope	Zero crossing (c)
Original clips (grid)	4.51 ± 0.09	0.999	6.9 [5.4, 8.3]
Shifted by one semitone	4.62 ± 0.04	0.996	7.1 [5.5, 8.7]
Shifted by 0 or 100 cents	4.69 ± 0.05	0.996	6.6 [4.9, 8.3]
Shifted by +33 cents	4.55 ± 0.10	0.9962	39.6 [37.1, 42.0]
Quantised to 24-TET	3.34 ± 0.12	1.0074	7.4 [5.1, 9.6]
Random offsets (flat)	3.75 ± 0.11	0.997	–
Mimi, original clips[§]	2.02 ± 0.14	0.961	5.7 [3.1, 8.2]
Mimi, shifted by +33 cents	1.98 ± 0.16	0.958	38.4 [34.2, 42.6]
Mimi, random offsets (flat)	1.74 ± 0.15	0.957	–
WavTokenizer, original clips	3.41 ± 0.12	1.002	4.9 [3.2, 6.6]
WavTokenizer, shifted by +33 cents	3.38 ± 0.13	0.999	37.9 [34.6, 41.2]
WavTokenizer, random offsets (flat)	2.78 ± 0.14	1.001	–
EnCodec, fine-tuned on Saraga	4.12 ± 0.14	1.001	15.4 [12.1, 18.7]

The 24-TET row's 50-cent fit reads $R^2 = 0.71$; a 100-cent one collapses to 0.06.

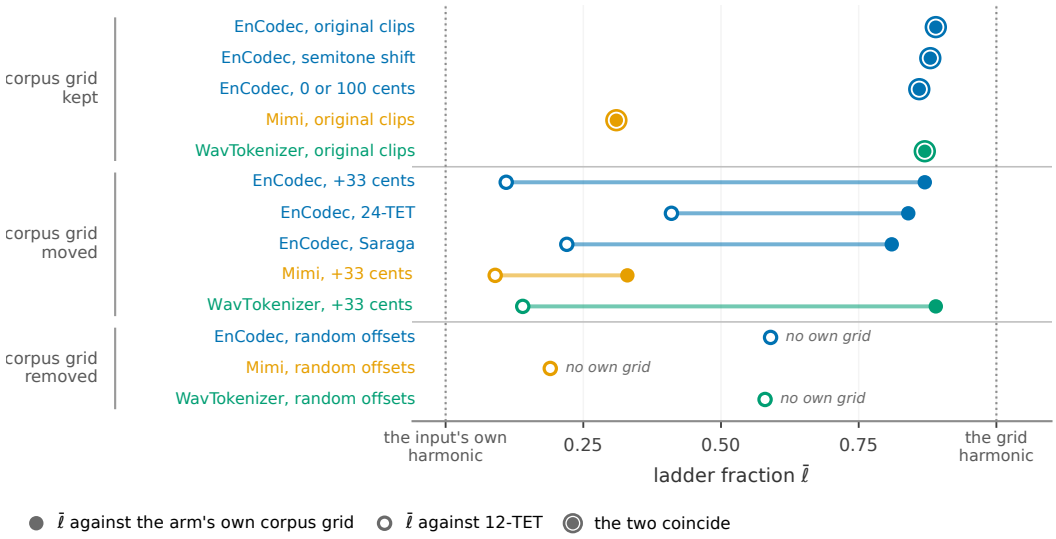

Figure A9: **The segment is the corpus's tuning.** Each arm of Table A25 read against its own corpus grid and against 12-TET. The two coincide exactly where the corpus grid did not move and separate where it did; the flat arms, having no own grid, are drawn once.

Table A26: Fitted amplitude, cents. DAC 16k, two arms, five seeds, decoder fine-tuned with encoder and quantiser bit-identical. The arms differ by 0.08 cents against a matched prediction of 0.12 and a detectable 0.34, so the arm is uninformative, not null. A dash is a quantity the row lacks.

Fine-tuned on	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4	Mean ± sd
Stock DAC 16k (no fine-tuning)	–	–	–	–	–	1.8
Original clips (grid)	1.7	1.7	1.8	1.7	1.7	1.72 ± 0.04
Random offsets (flat)	1.6	1.7	1.6	1.6	1.7	1.64 ± 0.05

original clips' 4.51. That change of comparator adds 0.18 cents, and moving from five seeds to ten draws takes off 0.03, which together turn the 0.76 gap here into the 0.91 there.

Table A27: Which half carries the change. Encoder from one checkpoint, quantiser and decoder from another. Rows sharing a quantiser and decoder differ by at most 0.01 in amplitude, within Table A35's between-run variation; rows sharing only an encoder do not. Last two rows: Table A9's arms.

Encoder	Quantiser and decoder	Amplitude (c)	Slope
stock	stock	13.72 ± 0.16	1.0010
fine-tuned, grid	stock	13.71 ± 0.17	1.0013
fine-tuned, flat	stock	13.71 ± 0.17	1.0013
stock	fine-tuned, grid	4.51 ± 0.09	0.9991
stock	fine-tuned, flat	3.75 ± 0.11	0.9975

Table A28: Controls. Each isolates one way the measurement could produce the effect on its own; the bypass is matched at one edge only, so it bounds the codebook account rather than excluding it.

Control	What it excludes	All-trial bias (c)
Identity codec	stimuli, estimator, analysis	$< 10^{-5}$
Pure sinusoid	pitch without harmonic structure	< 0.01
Quantiser removed	the codebook account (a bound, not an exclusion)	5.66
Uncoded noise floor	estimator resolution	$< 5 \times 10^{-4}$

What these arms do not attribute. They localise the ladder to the decoder, not to a design choice. EnCodec and DAC differ in corpus, activations, discriminator family, loss weighting and sample rate (Kumar et al., 2023; Ziyin et al., 2020), and nothing here says which of those matters. Whether any of it is audible is a separate question, on which the literature below bears only indirectly. Mistuned-partial thresholds (Moore et al., 1986) are stated for a single mistuned harmonic, where the partials that move here move together; and interval discrimination near the semitone (Zarate et al., 2012) is measured on the interval, not on a displacement inside one band of a complex. The pure-tone limen of about 5 cents near 500 Hz (Wier et al., 1977) is further still from the case at hand, being a threshold on an isolated sinusoid. The literature that speaks most directly to the mechanism is the pitch-dominance work, beginning with Ritsma (1967): Moore et al. (1985) measure how much a shifted individual partial moves the pitch of the whole complex, and Hartmann et al. (1990) how a mistuned harmonic is heard within an otherwise periodic tone. Both concern a *single* mistuned partial, where six of ours move together, so they bound the perceptual analogue without settling it. Berendes et al. (2025) did run a listening test, on whole recordings whose *global* tuning differs, and found tuning affects perceived vocoder quality; that is a different manipulation from a displacement inside one band of one tone, so we do not read it as evidence either way. Perceptual consequences are outside this paper's scope.

E.2 What the DAC causal arm could have detected

Computed on the seed-level standard deviation this arm looks sensitive, but that is not the component we argue applies to EnCodec. Read on independent flattening draws, as for EnCodec, it is uninformative; the numbers are below.

Two of the *pull* nulls, DAC 16k and SpeechTokenizer, call for a different calculation. DAC 16k registers and SpeechTokenizer does not; for both, what the pull rule finds null is the bias, which is what the power calculation below concerns. Taking the standard error from each condition's printed interval in Table A9 and testing against the Bonferroni-adjusted $\alpha = 0.05/11 = 0.0045$ the pull rule uses, the minimum detectable off-grid bias at 80% power is $(z_{0.99773} + z_{0.80})\,\mathrm{SE} = 3.68\,\mathrm{SE}$. DAC 16k's 99.55% interval $[-0.55, 0.07]$ gives a trial-level $\mathrm{SE} = 0.109$ and SpeechTokenizer's $[-0.21, 0.14]$ gives 0.062. Those cover one stimulus at one detuning, so we add the 0.122-cent between-run component of Appendix G.2 in quadrature, giving $\mathrm{SE} = 0.164$ and 0.137 and detectable biases of **0.60** and **0.50** cents. The account makes a prediction for only one of the two. SpeechTokenizer's edge is right-censored rather than measured high, so the account makes no prediction there, as Section 3.1 reports. DAC 16k is the case the step and soft forms of the edge (Ap-

pendix C.4) disagree about. Its edge is 4.4 kHz, so at a 440 Hz reference the eighth partial reaches 3.52 kHz and *no* partial the estimator fits lies above the edge: Equation 15 gives $b^{\mathrm{unif}} = 0$ however high the ladder fraction is, and DAC's is high (0.85). Had DAC regenerated from 1.5 kHz, five of the eight partials would have been above it and the same equation would imply a bias of about 21 cents before the estimator's 0.42 to 0.69 attenuation, so roughly 9 after it — far above either detectable figure, and excluded by both nulls. Appendix E.1 disclaims Equation 15 for magnitudes, but the exclusion does not need 9 to be right: the error runs one way and is bounded, over-predicting by 1.6 to 5 times at every register of Table A13 where it predicts anything (Table A13), and 9 over the worst of those is still 1.8. What it needs is the sign, which is measured. That reading uses the step form of the edge, and it is the step form Appendix C.4 shows failing: at 110 Hz EnCodec also has no partial at or above its edge, so the step form predicts nothing there while 2.75 cents is measured. The soft-edge form fitted there puts DAC's eighth partial 0.32 octaves below a 4.4 kHz edge and predicts about 3.9 cents at $\bar{\ell} = 0.85$, or 1.6 after attenuation — above the 0.60 this arm could detect, and not what we measure. The two forms disagree, and the register sweep below adjudicates them: the step form predicts the 440 Hz null and the 880 Hz pull but not the size of the 1760 Hz pull, since the same share of partials clears the edge there as at 880 Hz; the soft form's 1.6 cents at 440 Hz is not measured, though it still describes EnCodec's 2.75 cents at 110 Hz. The soft form was fitted on one codec's registers, the likeliest reason it transfers less well here.

The DAC 16k fine-tuning arms give 1.72 ± 0.04 and 1.64 ± 0.05 cents over five seeds. Computed on the between-draw variance of EnCodec's ten-draw run (Table A23), applied here to DAC, the arm detects 0.34 cents at 80% power against a proportionally matched expected effect of 0.12, so it is uninformative: it could not have found an effect of the size the account predicts for DAC. DAC's stock amplitude is 1.78 cents, comparable to Opus's 1.57, so the dynamic range available to the manipulation is small. Nothing rests on this null: the spectral evidence carries the contrast, and the edge reading of the null is the one the register sweep below settles.

Does DAC pull where its edge is reached? The edge reading of DAC's null says the estimator never fits above 4.4 kHz at 440 Hz, which would make the null a scope statement rather than a counterexample. The test is a register where partials do clear the edge, and the 440 Hz pair sweep looks as though it already contains one. The first tone sits at the reference, and eight partials of 440 Hz reach 3.52 kHz, so it never clears; the second sits at $440 \cdot 2^{\theta/1200}$, so θ alone fixes how many of *its* partials clear 4.4 kHz — none below 386 cents, one from 386, two from 617 and three from 884 — with the first tone held fixed throughout as its own control. Scored as Section 3.1 scores the bias, on the off-grid trials at the on-grid reference and under the octave gate alone, DAC reads $+0.67$ [-0.11, 1.17] with no partial above its edge, -0.51 [-0.77, 0.00] with one, $+0.03$ [-0.19, 0.52] with two and -0.55 [-1.42, -0.27] with three, which reads as no pull anywhere. That bucketing does not isolate the quantity a register sweep measures: in a pair sweep the residual is an *interval* error to which both tones contribute, while only the second tone's partials ever clear 4.4 kHz, and the informative bucket rested on 58 trials. Run directly at the registers where the edge is reached, DAC pulls.

Four of eight clear at 880 Hz, which through Equation 15 at $\bar{\ell} = 0.85$ and the off-grid median $|\delta_0|$, scaled by the 0.42 and 0.45 at which EnCodec's six- and seven-partial buckets of its own pair sweep read against that equation, predicts 7.1 to 7.7 cents. That scale is not the ratio of Table A14, which compares one tone's reading with its uniform-weight mean. At 1760 Hz the generator keeps the four partials below its 7.6 kHz cutoff for a 16 kHz codec, and two of them clear the edge. The 880 Hz prediction was fixed before the arm ran, and the arm returned 7.0 to 7.6 there and 10.9 to 11.3 at 1760 Hz, registering inside the pre-specified 1 ± 0.15 window at both.

That settles what the 440 Hz null was. It was not a decoder that fails to regrid — DAC's ladder is 0.85, among the highest in the paper — but a decoder whose invented band sits above everything the estimator fits at that register. Raise the register until partials clear the edge and the pull appears, at 880 Hz at the size the edge share predicts. This is a within-condition test of the account: one decoder, one bitrate, one ladder, and only the register varying, so no comparison across systems carries any of it.

E.2.1 THE SAME FLOOR FOR EVERY CONDITION

Table A29 extends that floor from the two conditions above to all eleven, so that a bounded null can be told from a non-detection directly. The table rests on one assumption: that the 0.122-cent between-run component, measured on EnCodec 24k at 3 kbps (Table A35), is applied to every row. Folding it in widens the nulls' detectable floors, and understates them where the true between-run spread is larger. Mimi's non-detection is not, however, a wide-interval problem: Mimi's trial interval is within two per cent of EnCodec's at 3 kbps (half-widths 1.51 and 1.48), and its floor of 2.00 cents sits beside EnCodec's own 1.97, so what makes Mimi a non-detection is the size of its effect, $+1.19$ cents, and not the width of its interval. Five end-to-end repetitions each of SNAC 32k and the vowel condition would settle both marginal pull verdicts.

Table A29: Smallest off-grid bias each condition could detect at 80% power, by the method of this appendix: $3.68\sqrt{\mathrm{SE}^2 + 0.122^2}$ with SE from the printed interval of Table A9. The two rows the text derives above are recovered exactly.

Condition	Bias (c)	Floor (c)	Counted
EnCodec 24k, 3 kbps	$+12.55$	1.97	pull
EnCodec 24k, 24 kbps	$+8.03$	1.59	pull
EnCodec 48k, 6 kbps	$+6.54$	1.35	pull
WavTokenizer	$+7.61$	1.42	pull
SNAC 32k	$+0.61$	0.71	pull
EnCodec 24k, vowels	$+0.43$	0.52	pull
Mimi	$+1.19$	**2.00**	null
SpeechTokenizer	-0.03	0.50	null
DAC 16k	-0.13	0.60	null
DAC 24k	$+0.01$	0.50	null
SNAC 44k	-0.14	0.51	null

Two things follow. First, **Mimi's null is not bounded**. Its floor of 2.00 cents is above every null point estimate in the table and above the two smallest effects we count as pulls, SNAC 32k's $+0.61$ and the vowel condition's $+0.43$: the arm could not have detected an effect three times the size of either. The structural reason is that almost nothing of the probe clears Mimi's 3.4 kHz edge (Table A16): at 440 Hz only the eighth partial does. The eleven of "six of eleven" are not power-homogeneous. Second, **SNAC 32k sits below its own floor**: we count it as a pull at $+0.61$ where the arm's 80%-power floor is 0.71, and the vowel condition likewise at $+0.43$ against 0.52. Both are marginal, consistent with the between-run fold removing one of them, and the floor comparison states it more sharply. The four remaining nulls are bounded, at 0.50 to 0.60 cents, about the size of the two marginal pulls.

F OFF SYNTHETIC TONES: REAL MUSIC AND DOWNSTREAM

F.1 THE EFFECT OFF SYNTHETIC TONES

The two EnCodec whole-recording readings of Table A30 sit at a phase within 11° of the tones' (Table A9), which is what makes them the same effect rather than two coincidences.

Note by note it disappears, and that null is the mechanism switched off rather than a failure to find it: a recorded note's upper partials are weak and often inharmonic (Fletcher & Rossing, 1998), so there is less above the edge to act on. That is testable, and tested — sorting 884 NSynth notes by above-edge energy, the top decile reads $+2.9$ cents and the bottom -0.1.

This also reconciles two numbers that would otherwise appear to conflict. Table A16 reads $\bar{\ell} = 0.79$ on sixty of these notes: their partials above the edge do land most of the way onto the grid, and Equation 15 would put several cents of bias on that. Table A30 reads $+0.007$ [-0.232, $+0.260$]. Both readings hold, and Equation 15 already contains the reason. Its weight ratio $\sum_{kf_0 \geq f_e} w_k^{\text{in}} / \sum_k w_k^{\text{in}}$ is 6/8 for the $1/n$ probe at 440 Hz: its input-referred weights are uniform across the eight partials by construction (Appendix C.6), and six of them sit at or above the edge. On a recorded note it is far smaller: the energy sits in the low partials, so the partials above the edge carry little of the weight the

estimator forms its reading from. The identity therefore predicts a small bias at $\bar{\ell} = 0.79$, which is what is measured. So *where* a decoder puts what it invents and *how far a reading moves* are different questions on real audio, and the identity is what relates them rather than what fails.

Two further checks. Repeated coding does not compound the bias, where O'Reilly et al. (2025) report drift under it, and MusicGen (Copet et al., 2023) sharpens the grid rather than inheriting it (Appendices F.6, F.7), though its continuation residual separates from only one of its three null controls.

F.2 THE PULL ON REAL RECORDINGS, IN FULL

Two protocols test whether the pull operates on real audio. The first takes isolated instrument notes from NSynth (Engel et al., 2017), resamples each to sit a controlled distance from the nearest 12-TET pitch, and measures where the codec returns it; displacement is reported as notes between grid points minus notes near them, so a pull gives a positive value. The second takes whole polyphonic clips of the tuning-flattened GTZAN corpus of Section 3.3, codes each, tracks F_0 frame by frame before and after coding with the same blind estimator, and studies the per-frame residual $1200 \log_2(f_{\text{out}}/f_{\text{in}})$ as a function of where the input pitch falls within the semitone, exactly as the sweeps do for θ. Each clip is resampled once, with the same polyphase resampler for every codec arm, from GTZAN's 22.05 kHz (or Saraga's 44.1 kHz) to the codec's rate, and both sides of the residual are read at that rate, so the resampler cancels; Opus's zero bias on the same chain confirms it adds nothing. Frames are kept when the cross-check agrees within 20 cents on both input and output and the residual is within the tracking gate; the flattened corpus puts input positions uniformly within the semitone (mean 50.3 cents), so the off-grid band is well populated. Retention is conditioned on the coded output, so a selection depending on within-semitone position could sculpt a periodic residual on its own; the exclusion rate per 10-cent bin of input position is flat, 68 to 73% for EnCodec on GTZAN and within four points of its mean on every other run. Moving the agreement threshold from 20 to 10 or 40 cents, or applying it to the input only, moves the EnCodec bias between 2.46 and 2.69 cents on GTZAN and between 2.14 and 2.75 on Carnatic, with retention between 27 and 46%. Intervals on the off-grid median are percentile bootstraps over clips, not frames, since frames within a clip are strongly autocorrelated; the frame-level intervals are narrower (for EnCodec on GTZAN, $[2.21, 2.81]$ against $[2.05, 2.98]$ over clips) and are not used. The fitted amplitudes of 3.05 cents on both corpora agree to two decimals by coincidence (3.049 and 3.051 cents).

Table A30: Real recordings. Top: isolated NSynth notes, with a synthetic positive control. Middle: GTZAN clips, tuning flattened. Bottom: 200 Carnatic recordings, whose pitches are not on the Western grid. Grid bias is the off-grid median over clips; a dash is a fit not attempted.

Protocol	Codec	Usable / total	Off-grid bias (c)	Amplitude (c)	Phase, R^2
Isolated notes	EnCodec 3 kbps	884 / 3455	$+0.007$ $[-0.232, +0.260]$	–	–
Isolated notes	DAC 16k	932 / 3443	$+0.011$ $[-0.018, +0.036]$	–	–
Isolated notes	synthetic control	355 / 2000	$+7.07$ $[+4.71, +10.25]$	–	–
Whole clips	EnCodec 3 kbps	18279 / 61714	$+2.493$ $[+2.053, +2.980]$	3.05	$+163°$, 0.015
Whole clips	DAC 16k	14260 / 61680	$+0.173$ $[-0.163, +0.527]$	0.66	$+92°$, 0.001
Whole clips	Opus 6 kbps	14907 / 61686	$+0.068$ $[-0.191, +0.361]$	0.32	$-129°$, 0.000
Whole clips	Opus 12 kbps	20819 / 61703	$+0.018$ $[-0.056, +0.093]$	0.07	$+106°$, 0.000
Carnatic clips	EnCodec 3 kbps	19082 / 53815	$+2.625$ $[+2.173, +3.140]$	3.05	$+166°$, 0.009
Carnatic clips	DAC 16k	13776 / 52219	-0.376 $[-0.663, -0.109]$	0.53	$+70°$, 0.001
Carnatic clips	Opus 12 kbps	21337 / 52854	-0.012 $[-0.066, +0.033]$	0.05	$+60°$, 0.000

The two protocols disagree, and the disagreement is informative. Isolated notes carry attack transients, decay and the estimator's difficulty with inharmonic instruments, and only a quarter of them survive: 884 of 3455, of which the 60-cent usability rule of Appendix F.5.3 accounts for 3.1 points and the estimator's cross-check for the rest. The null is measured on that retained quarter and we report it as such; the paired whole-clip measurement has an order of magnitude more usable frames and finds the pull for EnCodec with the sign pattern and phase of the synthetic sweeps, negative just above each grid point and positive just below the next. DAC 16k gives $+0.17$ cents on GTZAN and -0.38 on Carnatic; the second interval excludes zero on the side away from the grid. Both fits have amplitudes near half a cent at phases $70°$ to $96°$ from the pull's, so neither has the shape of a pull or of its reverse, and we read both as the floor of a measurement whose per-frame scatter is several cents; the Carnatic sign is reported as measured. Opus, which has no learned component, is at zero.

The binned medians for EnCodec run $-0.5, -2.6, -3.7, -4.1, -3.2$ cents over the first five 10-cent bins above a grid point and $-0.4, +2.2, +3.3, +2.9, +1.2$ over the five below the next.

F.3 The synthetic-to-real ratio, within one instrument

Comparing 2.49 cents on real music against 12.55 on synthetic tones would make the effect a fifth of its synthetic size. That comparison is not licensed: the real-music reading comes from the frame-by-frame pipeline of Appendix F.2, a six-partial fit on short frames behind the cross-check gate, the synthetic reading from the eight-partial fit over a whole steady tone, and the two instruments do not agree even on clean tones. Quoting their ratio confuses an attenuation with an instrument change.

We therefore ran the frame-based pipeline over the synthetic sweep, which costs nothing but CPU. It reads $+5.90$ cents [5.41, 6.38] off-grid where the refined estimator reads 12.55 on the same audio. The gap is expected and is the paper's own point restated: a six-partial fit on short frames puts fewer partials above the edge, and less weight on them, than an eight-partial fit over a steady tone, so it weights the regenerated band less. Table A15's two-partial column, reading zero, is the limiting case of the same effect.

Against that within-instrument baseline, real music retains $2.49/5.90 = 0.42$ of the synthetic effect on GTZAN and 0.44 on Saraga, nearer two fifths than a fifth.

F.4 What a round trip costs a tuning analysis

A 2.5-cent bias matters where cents are the measurement. Tuning analysis locates a scale's degrees to a few cents, and the traditions whose degrees sit away from 12-TET are the ones a pull toward it would cost most. The two corpora read here, Saraga and makam, sit closer to 12-TET than that framing suggests — Appendix F.5 gives the distances, and much of both octaves lies within a few cents of the grid — so they are corpora where the distinction under test is a few cents wide rather than traditions maximally far from 12-TET. We take each degree's reference from the corpus's own measured F_0 histogram rather than from theory tables, and read 200 Saraga and 60 Turkish makam recordings (Uyar et al., 2014) within-recording, before and after a 3 kbps round trip.

One provenance point concerns the source format. Saraga ships as 128 kbps MP3 at 44.1 kHz, so its "uncoded" arm is uncoded only with respect to the neural round trip we apply. This is the condition an analyst actually works in, which is why we leave it as it is, and MP3 is one of the classical controls that does not pull (Section 3.1); its band edge also sits above the partials our fit reads, so it cannot supply the ladder we attribute to the decoder. The contrast is within-recording, so any grid the MP3 stage left is present on both sides and differences out. It does mean the Saraga arm cannot speak to a chain with no lossy stage at all, and the makam corpus, which we read the same way, carries the same caveat.

Most of the damage is generic, and which control is used decides whether that null holds, so we give both. Against a distortion-matched Opus the grid-attributable part of the degree error is 0.4 cents $[-0.2, 1.0]$ on Saraga and 0.7 $[-0.3, 1.7]$ on makam, neither excluding zero; against DAC 16k, which regenerates above 4.4 kHz and regrids what it regenerates — Table A16 puts its $\bar{\ell}$ at 0.85, and its 4.4 kHz edge is the candidate reason it is protected, which the register sweep of Appendix E.2 supports — it is 0.9 [0.2, 1.6] and 1.5 [0.5, 2.5]. On magnitude alone the grid adds little.

Distortion-matched Opus is the bitrate whose *coded* median absolute degree error comes closest to EnCodec's from below — 4.6 kbps on Saraga, where it reaches 6.9 cents against EnCodec's 7.3, and 5.2 kbps on makam, 8.9 against 9.6. The match is to within 0.4 and 0.7 cents. Those are the same two numbers as the grid-attributable degree error above, and necessarily so: with a common uncoded arm, the grid-attributable part just is the gap between EnCodec's coded error and the control's, which is what the matching procedure minimises. It therefore measures how finely the Opus bitrate grid could be matched and not a property of the codec, and under the bracketing arm below it changes sign. No magnitude claim rests on it, and Section 3.4 rests its argument on the flip contrast instead. The match is also from below: EnCodec does that much more damage than the control, so part of its flip excess could be generic. The bracketing arm tests this: Opus at 4.0 kbps reaches a coded median absolute degree error of 7.8 cents on Saraga against EnCodec's 7.3, and 4.2 kbps reaches 10.2 on makam against 9.6 — overshooting EnCodec's damage by 0.5 and 0.6 cents rather than falling short of it by 0.4 and 0.7. It flips 3 of 200 Saraga and 3 of 60 makam, against EnCodec's 22 and 11, with

discordant pairs splitting 21 to 2 (exact McNemar, $p = 7 \times 10^{-5}$) and 9 to 1 ($p = 0.021$). The contrast is now bracketed: on Saraga, a control doing less damage than EnCodec flips 1 per cent of recordings, a control doing more flips 1.5 per cent, and EnCodec flips 11. Flip rate does not track damage magnitude on either side, which a generic account of the flips would require.

What the grid changes is direction, and that is where the controls do the work. Fitting each recording to its own system against 12-TET by likelihood, 33 recordings change which fits better and every one moves toward 12-TET: 22 of 200 Saraga flip under EnCodec against 2 under the matched Opus, whose discordant pairs split 22 to 2 (exact McNemar, $p = 4 \times 10^{-5}$); the makam set flips 11 of 60, whose discordant pairs split 10 to 1 at $p = 0.012$; those 22 and 11 are the 33 of Table A31's caption, which counts the two primary rows and not the band-limited or pYIN rows below them; and the nineteen control flips across both corpora (the control arms' own, not the nineteen makam flips of the Saraga-tuned decoder below) split ten toward 12-TET and nine away, which is what makes the direction of the EnCodec flips a finding rather than a property of the metric. The nineteen are, on Saraga, 4 band-limited, 0 at Opus 12 kbps, 2 at distortion-matched Opus, 4 under DAC 16k and 2 under the Saraga-tuned decoder, and on makam, 2 band-limited, 0 at Opus 12 kbps, 2 at distortion-matched Opus and 3 under DAC 16k; pYIN's 6 are left out, since that row reads EnCodec-coded audio with another estimator rather than a control coding. The Saraga-tuned decoder sits on both sides of this accounting, and the asymmetry is deliberate: on the corpus it was tuned to it repairs rather than imposes, so a flip there counts against the account and is admissible among the controls, while its makam flips are the positive result of Appendix F.5.1 and are never pooled with them. The pooled control is broad: the arms summed here change the fitted partial set and the decoder as well as the distortion, so ten-to-nine is evidence the metric is two-sided and not specifically that a distortion-matched codec flips both ways; that narrower claim would rest on the matched and bracketing Opus arms alone. At task level, on the sixty recordings first scored, makam identification from the degree positions falls from 0.80 to 0.73, a raw paired change of -0.07 [-0.16, 0.02]: six of sixty recordings losing their class and two gaining one, an exact McNemar p of 0.29 on that 6-to-2 split. By the non-detection standard of Appendix E.2.1: with eight discordant recordings only 8-to-0 reaches $p < 0.05$ (7-to-1 gives 0.070), so the smallest drop sixty can detect is $8/60 = \mathbf{0.133}$ against the $4/60 = 0.067$ measured. At that size the task is a non-detection and not a null, which is why we enlarged it.

Rerun over 235 recordings, the change is -0.06 [-0.11, -0.02] at an exact McNemar p of 0.011. That p is computed as though 235 had been fixed in advance and it was not: we enlarged after the sixty-recording arm returned 0.29, which is the same stopping rule that makes us decline to quote a permutation p for the checkpoint arm of Appendix G.6. It should be read as a descriptive statistic on the enlarged set, not as a test. The enlargement grew the reference pool with the test set: over twenty classes sixty recordings give about three exemplars per class and 235 about twelve. The enlarged arm is therefore a better-supported classifier than the controls that stayed at sixty, and the two rows of Table A31 are not one instrument. What survives that is the paired change itself, which is within-recording — the same 235 recordings clean and coded, through one classifier — so the drop is measured against itself and not against the smaller arms. What does not survive is the comparison across rows: the band-limited and Opus-matched figures sit at sixty with three exemplars per class, and the difference between them and this arm mixes the remedy with the reference pool. The sixty pilot recordings are among the 235, so this arm extends the pilot rather than replicating it. The split is 23 recordings losing their class against 8 gaining one, which is the -0.06 at $15/235$ and the exact p above. The enlargement was sized from the pilot's discordance rate of 8 in 60, which projects about 31 discordant pairs at 235 and a detectable drop near 0.055; that is the design calculation, not a reported count. The enlarged-arm change is the reported result; the sixty-recording figures above are retained because the flip contrast and the degree errors elsewhere in this appendix are scored on that set. The controls move the smaller set by one recording, so its grid-attributable remainder is -0.05 [-0.11, 0.01] (the difference of unrounded paired differences, $-4/60$ against $-1/60$), which does not clear zero and is computed at the sixty-recording size. The paired change reads 0.00 [-0.05, 0.05] under a band-limited estimator and under a corpus-matched decoder, but both of those rows sit at sixty recordings, where by the calculation above the task cannot detect a drop below 0.133. Neither is evidence that the remedy restores identification; both are non-detections at that size, and only the first is a fix we recommend on other grounds.

The band-limited fix has a register condition, which the abstract carries as a clause and this paragraph makes precise. It is demonstrated at 440 Hz, where six of the probe's eight partials sit above the

edge, and on the real recordings above. At 110 Hz its premise fails: Table A13 has no partial at or above the edge in that register and 2.75 cents is still measured, which Appendix C.4's soft-edge form attributes to the seventh and eighth partials sitting just *below* it, which a limit at f_{e} keeps. The remedy is for the registers where the edge falls inside the fitted partials, which are the registers where the effect is large; at the bottom of the range the effect is small.

Degree error is scored against the corpus's own aggregate degree positions, and the leave-one-recording-out reference removes recording-level leakage and not corpus-level leakage, because the quantity scored is a corpus-level constant. A decoder fine-tuned on that corpus therefore has its invented partials pulled toward the positions the metric scores against, which is the same circularity we refuse for the pooled-in reference (Appendix F.5). Table A25 shows the attractor is not gone but re-aimed: that arm's crossing sits at 15.4 cents with $\bar{\ell} = 0.81$ against Saraga's own grid. Every reference here is built from uncoded audio, and the coded arm is scored against it. Leave-one-recording-out removes a recording from its own reference; it does not remove the mismatch between a coded estimate and an uncoded target. An analyst working entirely from codec-distributed audio would estimate both the positions and the reference from the same coded corpus, and a shift shared by every recording would largely cancel in that comparison. So some part of what we report as a cost may be the cost of mixing coded estimates with uncoded references rather than of the grid. That is the realistic pipeline for someone auditing an archive, and not for someone whose whole corpus arrived coded. Rerunning degree error, flips and identification with references re-estimated from coded audio would separate the two.

The register prediction, tested where the cost is claimed. The exposure figures above are also a prediction. If the pull is carried by partials regenerated above the edge, then recordings whose register puts partials there should carry more of it, and recordings below should carry less. Splitting the same retained frames at the 220 Hz where the sixth partial clears the 1.32 kHz edge, and reading the off-grid grid bias with the conventions of Appendix F.2 throughout, the 195 recordings the cross-check retains divide 79 above and 116 below with bias 4.62 [3.38, 6.25] against 1.95 [1.51, 2.40] cents, a difference of 2.68 [1.38, 4.30] that clears zero on a bootstrap over clips. It is monotone across register: 1.76, 1.80, 2.46, 3.78 and 8.34 cents over clips whose median f_0 falls below 160 Hz, in 160–200, 200–240, 240–300 and above 300 Hz. It is stable from a 180 Hz threshold to a 280 Hz one, where the difference runs 1.35 to 5.72 and clears zero throughout. And it is a property of the pulling codec, not of high-voiced Carnatic recordings: the same split gives -0.11 [-0.24, -0.02] under a distortion-matched Opus and -0.19 [-0.63, 0.43] under DAC 16k, neither of which is the gradient and the second of which covers zero. Read on the whole corpus this pipeline returns 2.63 [2.17, 3.14], which is the arm's published reading and the check that the split is computed the same way. It also replicates on the other corpus: tuning-flattened GTZAN through the same codec at the same rate, so at the same threshold, gives 4.88 against 2.15 cents, a difference of 2.74 [0.95, 5.33], against -0.10 [-0.25, 0.09] for Opus at 12 kbps and -0.66 [-1.21, 0.09] at 6 kbps, both covering zero. Two corpora from different traditions give differences 0.06 cents apart at a threshold fixed by the edge and the partial count rather than chosen. GTZAN's upper bins are thin — 152 of its clips sit below 160 Hz — so the split is supported there but the bin-by-bin monotonicity is claimed for Saraga only. The test establishes that the effect on this corpus is concentrated where the mechanism says it must be (`analyze_register_split.py`).

Two links in the chain from mechanism to cost are shown only indirectly. The per-partial reading is made on synthetic tones and on isolated NSynth notes, and on no Saraga or makam recording; on those recordings the register test above shows the effect concentrating where the mechanism predicts. And the remedy is demonstrated with our own estimator truncated at the edge; the external route, low-passing the decoded audio below the edge, applies the same principle. Low-passing the 200 Saraga files at the edge and rerunning `estimate_tuning`, whose uncorrected count Section 3.4 reports, would test the external route directly.

Band-limiting declines to read the invented band at all, and it costs nothing. That is also why it removes far more than the grid-attributable figure above: it drops every partial the decoder invented, and with them the generic displacement in that band as well as any grid-directed part, so 3.0 and 4.2 cents removed against a grid-attributable 0.4 and 0.7 is what the mechanism predicts rather than a tension in it. The remedy is worth more than the grid claim is, and the two quantities answer different questions. Three points about it, since the remedy is what a practitioner acts on. First, its detectability floor: at 80% power an arm resolves about 2.80 standard errors, so the band-limited

degree-error readings, 0.2 [−0.3, 0.7] on Saraga and 0.2 [−0.5, 0.9] on makam, are blind to a residual below 0.71 and 1.00 cents respectively. They are bounded nulls up to about a cent and not demonstrations that nothing is left — and a cent is larger than the 0.7 that pYIN leaves. Second, the register condition above applies. Third, raising the edge should help too, but the decoders we compare happen to differ in edge rather than having been made to, so that one is a correlation and not an intervention (Table A31). That comparison also rests on a net three recordings of sixty, in an exploratory family of three carrying no adjustment, so we rest the section on the Saraga flip contrast rather than on it.

F.5 Tuning analysis through a codec

Scale degrees are estimated as the modes of each recording's within-octave F_0 histogram, folded in absolute pitch so that 12-TET sits at A4 = 440 Hz for every recording, as the codec's grid does, and matched by nearest neighbour to the corpus's aggregate measured positions, which are measured relative to each tonic and placed at the recording's own absolute tonic before matching and scored as the median absolute deviation from those. We call it the measured-histogram reference throughout, since a reference and an annotation are not the same thing here: neither corpus ships svara or perde *positions*. What they ship are labels, the rāga or makam of each recording and its tonic, and the positions are estimated from the audio, which is why the reference is built leave-one-recording-out below; the tuning-system fit is a likelihood ratio over the same histogram between 12-TET and the corpus's own degree positions, measured rather than theoretical. That distinction matters: the reported flip counts are scored against the measured-histogram reference, and the theory tables are not the comparator. For makam the Arel-Ezgi-Uzdilek positions of Appendix J appear as the secondary comparison of Table A31, for the reason that appendix gives; for Carnatic the 22-śruti comparison quoted below is a distance illustration and not a reference the flips were scored against. A flip therefore says a recording's histogram has moved nearer 12-TET than to its own corpus's measured positions, which is a narrower statement than a change of tuning system.

How far a tradition's degrees sit from 12-TET decides how much work a flip metric can do, so we state it for the theory tables, which are the positions we can quote exactly. Across the 22 śruti the gap to the nearest 12-TET pitch runs 0 to 19.6 cents with a median of 9.8, but five of the twenty-two lie within 4 cents: $1/1$, $9/8$, $4/3$, $3/2$ and $16/9$ — the tonic, the Pythagorean whole tone, the fourth, the fifth and the Pythagorean minor seventh, so structurally central rather than incidental. Taking Rast in Arel-Ezgi-Uzdilek commas, six of its eight degrees sit within 6 cents of a 12-TET pitch and only segah and evc are materially off, at 15.1 and 13.2. Rast is the makam closest to the diatonic, so that is a lower bound on the distance rather than a characterisation of the corpus: the perdes that distinguish the hicaz, uşşak and saba families, among the twenty makams present, sit further out. At the near-coincident degrees a round trip's few cents is the size of the distinction being tested and a flip is close to arithmetic, so the counts are a directional statistic over a mixture of well-separated and near-coincident degrees rather than a per-degree one. The Uncoded and Coded columns of Table A31 are computed before and after one round trip, so the comparison is within-recording with a common estimator. The uncoded column is a self-consistency figure and not an accuracy — each recording's deviation from the corpus's own pooled positions, so how tightly that corpus agrees with itself under this estimator, not how close either sits to what a musician intended — and should not be read as the precision of tuning analysis in general. The paired difference is what the round trip costs. The degree reference is built leave-one-recording-out, since pooling a recording into the histogram it is then scored against would benefit the uncoded arm where the coded arm does not; the pooled-in reference instead moves the Saraga paired difference from 3.2 to 3.6 and the makam from 4.4 to 4.9, so the circular version overstates by about a tenth.

The identification task is specified as follows, since the accuracies above cannot be read without it. The label set is twenty makams, so chance is 0.05 and the uncoded 0.80 sits well above it. The classifier is a nearest neighbour over the recording's degree histogram against per-makam reference histograms, built leave-one-recording-out as the degree reference is and for the same reason. There is no training beyond forming those references. This is deliberately a scalar-component instrument. Makam identity is not fixed by scale alone — *seyir*, the melodic progression, is constitutive — so a histogram classifier isolates the part of makam identity that tuning can damage and ignores the rest. It is sensitive to a tuning perturbation by construction, which is what makes it a useful probe here and what stops it from being a claim about makam recognition in general.

Table A31: **What the artefact costs a tuning analysis.** Median absolute scale-degree error before and after one round trip, within-recording. All 33 EnCodec flips against the measured-histogram reference are toward 12-TET. Brackets are 95% intervals, Clopper–Pearson for flip rates.

Corpus	Codec, estimator	Uncoded	Coded	Paired diff.	Flips
Degree-error blocks are in cents; the identification block below is accuracy, over 60 recordings except where a row says otherwise.					
Saraga, Carnatic	EnCodec 3 kbps	4.1	7.3	3.2 [2.6, 3.8]	11% [7, 16]
Saraga, Carnatic	band-limited	4.1	4.3	0.2 [−0.3, 0.7]	2% [1, 5]
Saraga, Carnatic	Opus 12 kbps	4.1	4.2	0.1 [−0.4, 0.6]	0% [0, 2]
Saraga, Carnatic	Opus, distortion-matched	4.1	6.9	2.8 [2.2, 3.4]	1% [0, 4]
Saraga, Carnatic	DAC 16k	4.1	6.4	2.3 [1.7, 2.9]	2% [1, 5]
Saraga, Carnatic	pYIN, EnCodec 3 kbps	4.4	5.1	0.7 [0.2, 1.2]	3% [1, 6]
Saraga, Carnatic	decoder tuned on Saraga	4.1	4.4	0.3 [−0.2, 0.8]	1% [0, 4]
Turkish makam	EnCodec 3 kbps	5.2	9.6	4.4 [3.5, 5.4]	18% [10, 30]
Turkish makam	band-limited	5.2	5.4	0.2 [−0.5, 0.9]	3% [0, 12]
Turkish makam	Opus 12 kbps	5.2	5.3	0.1 [−0.6, 0.8]	0% [0, 6]
Turkish makam	Opus, distortion-matched	5.2	8.9	3.7 [2.8, 4.6]	3% [0, 12]
Turkish makam	DAC 16k	5.2	8.1	2.9 [2.1, 3.8]	5% [1, 14]
Turkish makam	decoder tuned on Saraga	5.2	11.8	6.6 [5.5, 7.7]	32% [20, 45]
Flip rate against a log-likelihood margin, EnCodec 3 kbps, both corpora pooled					
margin ≥ 0	(any crossing)				13%
margin ≥ 1					12%
margin ≥ 3					9%
Makam identification from the degree positions; accuracy, not cents. Twenty classes, chance 0.05.					
Turkish makam	EnCodec 3 kbps, $n = 60$	0.80	0.73	−0.07 [−0.16, 0.02]	
Turkish makam	rerun at $n = 235$	0.81	0.75	−0.06 [−0.11, −0.02]	
Turkish makam	band-limited	0.80	0.80	0.00 [−0.05, 0.05]	
Turkish makam	Opus, distortion-matched	0.80	0.78	−0.02 [−0.05, 0.02]	
Turkish makam	DAC 16k	0.80	0.78	−0.02 [−0.05, 0.02]	
Turkish makam	decoder tuned on Saraga	0.80	0.80	0.00 [−0.05, 0.05]	
Reference degree positions, makam, EnCodec 3 kbps (first row repeats the block above; these re-score the same 60 recordings, so their flips are not additional to the 33. The two references do not flip the same recordings: 22% of sixty is thirteen against the primary reference's eleven, so at least two lie outside it; the AEU flips are counted without a direction score)					
measured histogram	(as above)	5.2	9.6	4.4 [3.5, 5.4]	18%
Arel-Ezgi-Uzdilek tables	(secondary)	8.7	12.4	3.7 [2.6, 4.9]	22%

What the downstream instrument cannot see. Every magnitude in this appendix is an unsigned median absolute deviation and every direction is a whole-recording binary, which one system fits better. Neither is the mechanism's distinguishing prediction. The mechanism says each degree moves toward *its own* nearest 12-TET pitch, so the signs should differ across degrees within a single recording; a coherent offset of a couple of cents, applied to a corpus whose degrees sit flat of 12-TET on average, reproduces the raised absolute error and the unanimous flips just as well. The offset account can be tested without assigning degrees to frames. Write d_f for a frame's signed distance from the nearest 12-TET pitch. A constant offset reads the same residual at every d_f; a grid-directed displacement reads one that runs opposite to d_f and crosses zero at the grid point. On the 199 Saraga recordings carrying a frame inside the octave gate — four fewer also survive the cross-check, and that smaller set is what the exposure reading below uses — resampled over recordings, the residual reads $+1.11$ [0.26, 2.06], $+3.13$ [2.41, 3.62], $+0.39$ [0.17, 0.62], -2.37 [-3.05, -1.60] and -2.21 [-3.09, -1.64] cents across the five bins of d_f from -50 to $+50$: a sawtooth crossing zero at the grid point, not a constant. Taking the two off-grid halves, $d_f < -20$ gives $+2.08$ [1.78, 2.54] against $d_f > +20$'s -2.53 [-3.12, -1.86], where an offset requires them equal and the grid account requires them opposite. Two controls read the same two halves. Opus at 12 kbps gives 0.01 [-0.06, 0.07] and 0.00 [-0.08, 0.08], but it barely alters the signal, so its null is close to entailed. DAC 16k is the informative one — it damages heavily and regrids what it regenerates, $\bar{\ell} = 0.85$ — and it reads -0.23 [-0.68, 0.13] and -0.29 [-0.63, 0.18]: flat, and the same sign on both halves, which is what a small coherent offset looks like and what the artefact of reading d_f off a corpus whose own density is peaked would also look like. Neither control shows the sawtooth, so it is not the estimator's and not the corpus density's. The coherent-offset account is therefore excluded on these recordings. The per-degree sign pattern within a recording is a separate, finer quantity, which would say *which* degrees move rather than that the movement is grid-directed.

How far the edge reaches into this corpus, and what that predicts. The remedy truncates the fitted set at the measured edge, so on a recording low enough that no fitted partial reaches it the band limit removes nothing. The frame file measures how often that happens. The estimator refines

over six partials on this arm (Section 2.5's probe uses eight; the real-music fit uses six), so partial k clears stock EnCodec's 1.32 kHz edge when $kf_0 \geq 1320$. Across the retained Saraga frames the median f_0 is 209 Hz and the mean count above the edge is 0.95 of six, against the six of eight that the 440 Hz probe of Table A13 carries; 53% of retained frames have none at all, and 116 of the 195 recordings carrying cross-check-retained frames have a median f_0 below the 220 Hz at which even the sixth partial clears. Over all frames before the cross-check the median is 184 Hz and the count 1.90, that count raised by octave-error frames read far above their true f_0; the retained set's median skews high because the cross-check keeps the frames it tracks best. So this corpus sits at roughly a fifth of the probe's exposure, and on over half its frames the remedy is inert.

That is a limit on the remedy and, at the same time, a prediction about the cost: if the displacement is the invented band being fitted, it must grow with how much of the fitted set lies above the edge. Scoring the grid-directed component of Section 3.4 by that count gives 1.34 [0.88, 1.80] cents on frames with no partial above the edge, 4.40 [3.41, 5.03] on those with one or two, and 12.04 [7.60, 17.32] on those with three or more, resampled over recordings. The same two controls read the same frames of the same recordings, so each bucket holds the identical register composition in all three: Opus at 12 kbps gives 0.02 [-0.14, 0.13], -0.00 [-0.09, 0.08] and -0.03 [-0.09, 0.03], and DAC 16k gives -0.51 [-0.90, 0.08], -0.35 [-0.90, 0.02] and -0.28 [-0.62, 0.14]. Neither approaches that gradient: across the three buckets Opus spans 0.05 and DAC 16k 0.23 against EnCodec's 10.70, and every control interval overlaps its neighbours. Register as such and the estimator as such therefore do not produce the gradient.

Two things follow. The first is that the 53% of frames the remedy cannot touch are also the frames carrying least of the cost, which is why the band-limited row of Table A31 reaches 0.2 cents despite the remedy being inert on over half the material; they are not costless, at 1.34 cents with an interval clear of zero, which is the frame-level counterpart of the 2.75 cents Table A13 measures below the edge at 110 Hz. The second is that DAC 16k's flat reading above is explained by its 4.4 kHz edge: the real-music fit takes six partials, so that edge is reached only above $f_0 = 733$ Hz, against the retained Saraga median of 209 Hz. Over its own retained set the arm reads -0.38 [-0.66, -0.11] (Table A30). A codec whose invented band lies above everything the estimator fits should displace nothing toward the grid, and it does not: its small reading runs away from it.

The sharper test, a rise following each codec's *own* edge, needs more than one codec whose edge falls inside the range the corpus reaches; of the three codecs run on this corpus, only EnCodec's does.

F.5.1 The two fine-tuned rows read as a pair

Section 3.4 rests its generalisation claim on the two *decoder tuned on Saraga* rows of Table A31 together rather than on the makam row alone, and the reason is that one row cannot identify a grid. A rise in makam degree error says the decoder imposed something the makam histogram does not match; it is equally consistent with 12-TET, with Carnatic positions, and with no grid at all. The pair is more specific. The same decoder, at the same bitrate and under the same six-partial estimator, takes the Carnatic paired difference from 3.2 to 0.3 [$-0.2, 0.8$] and the makam one from 4.4 to 6.6 [$5.5, 7.7$]. A decoder pulling its output toward 12-TET would raise both, since neither corpus is in 12-TET; only a decoder carrying a corpus-specific grid can repair one while damaging the other.

One confound belongs on the Carnatic half specifically. Appendix E.1 finds that fine-tuning on *any* corpus takes the amplitude from 13.72 to about 4.6 and raises the edge from 1.3 to 2.0 kHz, which moves partials below the edge and reduces the reading with no corpus-specific grid involved. We separate the two by scoring a decoder fine-tuned on a different corpus on Saraga. Against a stock arm rerun alongside at 3.3 cents, the GTZAN-fine-tuned decoder takes the Carnatic paired difference to 1.3 cents, five seeds spanning 1.1 to 1.5 with the interval clear of zero, and flips seven or eight of 200 in every seed against stock's 22. The Saraga-tuned decoder, rerun beside it, reads 0.4. Every figure in this decomposition is from that rerun, so it uses 3.3 and 0.4 where Table A31 carries the original run's 3.2 and 0.3; the repair is 2.9 cents either way. Of it, about two thirds is the generic effect and about one third, 1.3 to 0.4, is Saraga's own grid; the flips divide in nearly the same proportion.

That third is a grid-*position* effect rather than a further exposure effect, and the pair is unusually clean on this point: both fine-tuned arms sit at the same 2.0 kHz edge and carry comparable ladders, read as means over the two detuning signs: 0.89 for the GTZAN arm against Table A16's 0.81 for

the Saraga one. Edge and ladder strength are held fixed across the comparison, so what separates them is which grid the ladder points at, which is the quantity the claim is about. The corpus-specific component is nonetheless the minority of the repair, and at 0.9 cents it sits near the floor below which our band-limited reading is blind, so we state it as a measured share and not as a large effect.

We record how this number arose. The value was predicted before the arm was scored, from the edge share alone: raising the edge from 1.32 to 2.0 kHz takes the mean count of fitted partials above it on the retained Saraga frames from 0.95 of six to 0.20, a factor of 0.21, which applied to the stock reading gives 0.7 cents; re-weighting the measured buckets by the counts a 2.0 kHz edge produces gives 1.6. The arm returned 1.1 to 1.5, between the two routes. The prediction was recorded before the run, at eight partials rather than the six this arm uses, where the first route gives 1.1 cents rather than 0.7. A single arm is not a registered pass; what it shows is that the edge-share account predicted a measurement it had not seen. The makam half is unaffected by this confound, because the generic effect reduces the reading, and the makam arm raises it, so whatever the makam decoder is doing runs against the generic direction rather than with it.

The two halves of the pair are not equally clean, and the makam half is the load-bearing one. Appendix F.4 states why: the leave-one-recording-out reference removes recording-level leakage and not corpus-level leakage, since the quantity scored is a corpus-level constant. The 200 Carnatic recordings being held out of the fine-tuning corpus therefore answers a smaller objection than the one that applies — it rules out memorising these recordings, not being trained toward the positions the metric scores against. Read strictly, the Carnatic repair is the fine-tuning seen through the metric rather than an independent confirmation of it.

The makam arm carries no such leakage. That decoder never saw the makam corpus, and the reference it is scored against is a constant of a corpus it was not trained on. But clean of leakage is not clean: Appendix E.1 records that this arm carries the largest reconstruction deficit of any in the paper, 0.022 against the same control, past both the 0.017 shown harmless and the flattened arm's bound, and it has no distortion-matched control. A decoder simply worse out of domain predicts a raised makam degree error with no grid involved. What the pair licenses is therefore this: a decoder trained on one tradition damages the other, which a pull toward 12-TET does not predict, and the direction of the Carnatic arm is consistent with the grid being Saraga's without independently establishing it.

The identification block of the downstream table bears on this reading. Across the six makam arms, identification accuracy tracks neither degree error nor flip rate. This decoder displaces degrees hardest of any arm, 6.6 against EnCodec's 4.4, and flips most, 32% against 18%, while costing identification exactly nothing — 0.00 $[-0.05, 0.05]$, the same as the band-limited fix. A plausible reason is that a corpus-specific grid displaces the degrees coherently, preserving the relative pattern a nearest-neighbour classifier reads, where an incoherent displacement would not; the present design does not separate that from the classifier being insensitive at this size. Either way the consequence is the one to state: no arm in this paper links the measured displacement to the identification drop. The task cost is a round-trip cost, and the attribution to the grid rests on the degree-error and flip contrasts rather than on the task.

The positive half on makam is measured directly. The 32% figure is a flip rate, and scoring those flips for direction — as the stock-EnCodec flips were scored toward 12-TET — puts all nineteen toward Carnatic degree positions rather than merely away from makam ones (two-sided exact binomial against an even split, $p = 4 \times 10^{-6}$). This is the half that does not depend on the Carnatic arm. The recordings are makam ones the decoder never saw, the reference they move toward is a third corpus's grid rather than the one they are normally scored against, and neither the leakage of the Carnatic arm nor the 12-TET account predicts it: a decoder pulling toward 12-TET would move makam histograms toward 12-TET, not toward Saraga.

Four points bound this reading. The first is whether the two candidate directions are separable at all on these recordings. Carnatic degree positions themselves sit close to 12-TET over much of the octave, so a displacement of a few cents may move a histogram toward both at once, and each flip is scored toward Carnatic without a recording-by-recording check that Carnatic and 12-TET lie in different directions from it. Scoring only the flips where the two diverge would settle that. The count therefore establishes that the flips are not random, and is weaker evidence against 12-TET than its unanimity suggests. Second, the nineteen are not a random nineteen: they are selected as the

recordings on which 12-TET became the better fit, so the subset is enriched for motion toward 12-TET before direction is scored at all. Given how close Carnatic sits to 12-TET, some of the unanimity is a property of that selection rule rather than of where the partials went. Scoring direction across all sixty recordings rather than the flippers would remove it. Third, this count has no control arm for flip direction, so how many flips a generically degraded decoder would send the same way is open — the same gap the `librosa` count carries, and we report this count on the same terms. Fourth, nineteen flips on sixty recordings is a small base; the count is unanimous, which is what makes it informative, but a single recording scored the other way would move the p by an order of magnitude.

Applying the remedy in a pipeline you did not write. The band limit above is demonstrated inside our own six-partial fit, which is not what either corpus is usually analysed with, and one of the tools that is, `tomato`, does not expose its harmonic count at all (Table A32). A reader could reasonably conclude the fix is unavailable to them. It is not, because it need not be applied inside the estimator. Low-pass the decoded audio below the codec's measured band edge before analysis and any estimator downstream inherits the remedy, whatever it exposes; the edge is measured by the procedure of Appendix C.5.3, which needs only the codec and a tone. Where the estimator does expose its reach, setting it directly is equivalent and cheaper: Essentia's `PredominantPitchMelodia` takes both a harmonic count and a maximum frequency. The one thing that does not work is applying a fixed limit across a corpus, because the edge sits at a fixed frequency while these traditions' tonics do not: at a low enough tonic no fitted partial reaches the edge, the limit removes nothing, and Appendix F.4 shows the residual that survives.

F.5.2 A STANDARD ESTIMATOR ON THE SAME CORPORA

The within-octave pitch histogram we use has a direct ancestor: Tarsos (Six et al., 2013) is the standard platform for exactly this analysis on non-Western material, and our svara and makam histograms are that method applied to coded audio rather than a new one. Tarsos sits on the protected side: its five pitch estimators are all time-domain period trackers and none fits a partial above the first. That is a statement about what they fit, not a guarantee of immunity — pYIN, the one member of that family we ran, still moves by 0.7 $[0.2, 1.2]$ cents under coding. The exposure divides by estimator family rather than by field, and Table A32 sets the defaults side by side. Period-based trackers, Tarsos's five and pYIN, form their estimate from the signal's period rather than from measured partial positions. That attenuates the exposure rather than removing it — the difference function is computed on the full-band waveform, so displaced upper partials can still move the estimated period, and our own pYIN arm, run through librosa's `pyin` at a pitch resolution of one cent or finer so that its output sits on no 10-cent lattice, moves 0.7 cents under coding. The division in this table is a gradient, and Tarsos is listed by its configuration. Salience- and peak-based tools read far above the fundamental: Essentia's salience function (Bogdanov et al., 2013) sums 20 partials by default, and `librosa.estimate_tuning` (McFee et al., 2015) estimates no fundamental at all, folding every local spectral maximum between 150 and 4 000 Hz into its histogram. Our eight-partial fit belongs to that second family and not to the first, which is the comparison that matters here, because the two corpora of Section 3.4 are analysed with the second family, as set out below.

We repeat the Saraga arm with pYIN (Mauch & Dixon, 2014) at its defaults, which fits no partials above the first. Uncoded error is 4.4 cents and coded 5.1, a paired 0.7 [0.2, 1.2], against 4.1 and 7.3 for the six-partial fit that arm uses, whose own paired difference is 3.2 cents. A tool that does not read the invented band is largely protected, which is the same conclusion the band-limited arm reaches by a different route, and it means the hazard is specific to full-band harmonic estimation rather than to tuning analysis as such.

Three points bound how far Table A32 extends. First, the hazard is a property of the tool and the codec together, not of the tool alone: it takes combined partials above *that codec's* edge, and Table A16 spans edges from 0.9 to 4.4 kHz. We resist reading any row as protected on this basis. A salience tool's f_0 ceiling is not a ceiling on what it consults — `tomato` forms candidates only to 1.76 kHz but sums twenty harmonics of each, so it reaches far above any edge here — and even a genuine ceiling below the edge would attenuate rather than protect.

Second, the rows measured on coded audio are our own fit, `librosa.estimate_tuning` and pYIN; every other row is read off documentation and source, so the table records configurations and

Table A32: Partials each estimator combines, at its own defaults. The middle column bounds f_0 *candidates*, not the spectrum read: summing H harmonics consults up to Hf_0. The four starred rows are three estimators, ours at two partial counts, run on coded audio; the rest come from documentation.

Estimator	Partials combined	f_0 candidates	Spectral reach
Salience- and peak-based			
`librosa.estimate_tuning`⋆	every peak ≥ 0.1 of frame max	none formed	150–4 000 Hz
Essentia `PitchSalienceFunction`	20 (decay 0.8)	from 55 Hz	$20f_0$
Melodia / `PredominantPitchMelodia`	20 (decay 0.8)	80–20 000 Hz	$20f_0$
`compIAM` (Carnatic)	20	80–20 000 Hz	$20f_0$
`tomato` (makam)	20, not user-exposed	55–1 760 Hz	$20f_0$
This paper, synthetic tones⋆	**8**	—	$8f_0$
This paper, Saraga and makam⋆	**6**	—	$6f_0$
Learned, time-domain input			
CREPE	learned; filters peak 600–1 500 Hz	outputs 31–2 006 Hz	raw 16 kHz frames
Period-based			
Tarsos (YIN, FastYIN, MPM, AMDF, wavelet)	1, the period	—	full band, time domain
pYIN⋆	1, the period	—	full band, time domain

not measured displacements. What it supports is that the configuration we tested is an ordinary one, not that any named tool has been shown to move.

Third, partial count is not the quantity the mechanism turns on, and on the quantity that is, the ordering inverts. The reading is a weighted mean, so what matters is the share of estimator weight above the edge. Our weights are $w_k \propto k^2|X_k|^2$; a salience function's contribution from harmonic h is $0.8^{h-1}|X_h|$, which puts $0.8^{19} \approx 0.014$ of its decay on a twentieth harmonic. Both must be read on the same stimulus envelope or the comparison is not like for like. Table A33 does that at a 440 Hz reference and a 1.32 kHz edge.

Table A33: Share of estimator weight above the edge, 440 Hz reference, 1.32 kHz edge, each estimator read on the same envelope. The six-partial column is the fit that produced Table A31.

Envelope	**Ours, 8**	**Ours, 6**	**20, decay 0.8**
$1/n$	0.75	0.67	0.30
$1/\sqrt{n}$	0.92	0.86	0.46
$1/n^2$	0.18	0.16	0.11

On every envelope our fit puts more weight above the edge than a twenty-partial sum does, at either partial count. So a tool combining more partials is *not* thereby more exposed, and on these envelopes it is less exposed than we are. Two caveats keep this indicative: a salience function uses those weights to select a candidate on a binned grid rather than to average measured partial positions, so the transfer from a displaced partial to the reported pitch is not the same function in the two cases; and the share depends on the envelope and on f_0, both of which a corpus varies. We therefore claim only that these tools combine partials well above the fundamental and that the corpora here are analysed with them.

`librosa.estimate_tuning` is the one row of that family we ran on coded audio. It returns a single tuning offset per recording rather than a degree histogram, so it is scored differently from the rest of this appendix: for each of the 200 Saraga recordings we take the offset before and after the round trip and ask whether it moved toward 12-TET. 134 of 200 did, against the 100 expected if coding moved them at random, a two-sided binomial p of 1.7×10^{-6}. One competing explanation, dilution, does not apply here. `estimate_tuning` takes the mode of a histogram of folded residuals rather than their mean, so attenuating scattered peaks does not dilute the estimate toward the centre of its range the way it would an average. That leaves the plainer point: these 200 recordings

have no distortion-matched arm, so this count is a direction the practice exhibits under coding and not evidence that the direction is grid-driven. Because a directional binomial reaches significance at any effect size, the count establishes that coding moves this tool's offsets toward 12-TET systematically, and not that it moves them far. How far is a question we answer only for our own six-partial fit, where the paired cost is 3.2 cents on the same corpus, and a reader should not carry that figure across to `librosa`, whose reach and weighting differ (Table A32).

The Carnatic and makam rows are not incidental. Saraga's released F0 is Melodia (Salamon & Gómez, 2012), and MTG's own `compIAM` (Nuttall et al., 2022) wraps it at 20 partials; makam analysis runs through `tomato` (Atlı et al., 2014), which calls Essentia's salience function without exposing the harmonic count, so a makam tuning study runs at 20 partials whether or not its author knows it. Against WavTokenizer's 0.9 kHz edge, 45% of the log-frequency span `librosa` draws peaks from lies above the edge. That is a statement about the span and not about the reading: peaks are counted rather than magnitude-weighted, magnitude entering as a threshold and not as a weight, so the share of the *estimate* that comes from invented spectrum is a different number, read here for our own fit and, directionally and without magnitude, for `librosa` as reported above.

Under pYIN the flip contrast follows the degree error: the Saraga recordings that change which system fits them better fall from 22 of 200 to 6, against 3 for the distortion-matched Opus control read the same way, 2 under our own fit: a discordant split of 5 to 2, which does not clear zero (exact McNemar, $p = 0.45$). That arm carries the same caveat we attach to the pull nulls: with seven discordant recordings only a 7-to-0 sweep reaches $p < 0.05$, so it could not have separated the two estimators except unanimously, and the null is correspondingly weak evidence. We read the pYIN null as the negative control the mechanism predicts rather than as a failure to replicate: pYIN sits with Tarsos in the one group of Table A32 that estimates a period rather than fitting partials, and that is the minority configuration among the tools these two corpora are actually analysed with.

F.5.3 Equivalence between the traditions, and the note-level unit

The GTZAN and Saraga readings differ by 0.13 cents with a bootstrap interval of $[-0.5, 0.8]$ over clips. We treat this as an equivalence question rather than a null one and declare a margin of one cent: a fifth of the frequency difference limen, and about two fifths of the real-music readings being compared, which is the denominator Appendix F.3 requires rather than the synthetic effect. Two one-sided tests at 5% both reject ($p = 0.002$ and $p < 10^{-3}$, from the bootstrap draws), so the two traditions are equivalent within one cent. The margin was chosen after the difference was seen and is reported as such; the full curve is monotone, equivalence holding for every margin above 0.62 cents and failing below it. They are not equivalent within a quarter of a cent: the same procedure at a 0.25-cent margin fails. The comparison is also not like for like, since GTZAN is tuning-flattened and Saraga is not; flattening Saraga would remove the very property under test, and leaving GTZAN unflattened would let its own 12-TET grid, rather than the codec's, produce the reading.

The isolated-note interval is a cluster bootstrap over the 108 instruments the notes come from, not over readings: notes from one instrument share a partial structure and are not independent, and resampling readings would narrow the interval on the side of the null. The usability rule discards a reading whose shift leaves the 60-cent refinement window, which censors in the direction of the null it supports. It removes 3.1% of otherwise usable notes. Widening the window to 120 cents keeps them and moves the reading from $+0.007$ [-0.23, $+0.26$] to $+0.09$ [-0.28, $+0.44$], still a null, and leaves the above-edge energy correlation at $r_s = 0.51$.

F.6 Accumulation under repeated coding

Table A34 gives the headline sweep passed through EnCodec at 3 kbps one, two, four and eight times in succession, each pass decoding and re-encoding the previous output. This test was not pre-specified; it was added to answer whether a generate-and-decode loop over codec tokens would compound the pull. It does not: the fitted amplitude falls by 6% over eight passes and the phase moves by 3°, while the all-trial bias falls from 8.88 to 7.26 cents as the codec settles on its own output; nothing grows. The single-pass row is this CPU pass's first round trip, whose all-trial bias of 8.88 matches the headline sweep's to the printed digits; the accumulation rows are read against it. Its off-grid bias and amplitude, 12.35 and 13.91, sit outside the five-run spread of Table A35 because that CPU pass used a single on-grid reference and five repetitions, where the headline design uses

eleven references and twenty draws: 1205 trials against 53,020. The ± 0.16 quoted on the amplitude is a spread across the ten detunings of one sweep, not an uncertainty on the reported quantity, so the headline figures should be read to about a tenth of a cent and not to three significant figures.

Table A34: Repeated coding. EnCodec at 3 kbps applied k times to the headline sweep (on-grid reference, 1205 trials per row, no trial gated). Grid bias is the all-trial median; amplitude and phase are the sinusoid fitted against the original input. Brackets are 95%.

Round trips k	All-trial bias (c)	Off-grid bias (c)	Amplitude (c)	Phase
1	8.88 [8.12, 9.54]	12.35	13.91	$+155^\circ$
2	8.20 [7.56, 8.83]	11.51	13.49	$+155^\circ$
4	7.39 [6.90, 8.08]	11.44	12.98	$+157^\circ$
8	7.26 [6.82, 7.86]	11.60	13.06	$+158^\circ$

F.7 Downstream generation through a codec-token model

MusicGen decodes with the 32 kHz EnCodec checkpoint rather than the 24 kHz one the rest of the paper measures. We therefore ran the band-edge and ladder-fraction sweep on that checkpoint too, so this section rests on a characterised decoder: its edge is 1.4 kHz at its single 2.2 kbps operating point and its ladder fraction 0.92 and 0.83 at the two signs of the detuning, within the spread of the 24 kHz rows (Table A16). That row and the held-out row of Table A40 are the same checkpoint read two ways: the first is the cross-register mean of Section 2.4, the second the binned 440 Hz reading the held-out arm uses throughout (Appendix C.5.3), which is why their edges and ladder fractions differ.

MusicGen-small (Copet et al., 2023) generates EnCodec tokens at 32 kHz and decodes them with EnCodec's decoder. Two measurements, neither pre-specified, read its output with the YIN stage of the estimator of Appendix B.5 on 64-ms frames, the instrument of the corpus histograms. First, 60 text-prompted 10-second generations (forty instrument and genre prompts): their pooled within-semitone F_0 density has peak-to-mean 4.28 against 1.788 for GTZAN, over a flat-sampling floor of 1.12 of its own at this sample size (GTZAN's, at its larger count, is 1.04), with the circular-mean position 3.6 cents above the 12-TET pitch (Figure A6). Second, 80 audio-prompted continuations: the middle five seconds of a tuning-flattened GTZAN clip is given as the prompt and ten seconds are generated. The prompt's tuning offset is read by the same estimator (it is the clip's own offset plus the applied shift), and the continuation's offset relative to it is folded into $(-50, 50]$. Overall the continuation stays with the prompt (median displacement $+0.2$ cents, interquartile range -6.3 to $+9.7$; Spearman correlation of continuation with prompt offset 0.60 over the 80 clips), and the grid-directed component of the residual is positive but does not clear two of the three null models below. For the 49 prompts at least 20 cents off the grid the continuation moves toward the grid by a median 7.5 cents, bootstrap interval [4.3, 11.4], or 0.23 of the median prompt offset of 33 cents, and by $+2.4$ [0.7, 6.0] cents for the 31 prompts within 20 cents of it. A linear slope of continuation offset on prompt offset is not a pull statistic here, since a pull toward the two grid points makes an S-shaped map, and is not used.

The bootstrap interval measures sampling noise, not grid attraction, so we test the statistic against three null models (`analyze_musicgen.py`, 4000 draws each). A *shuffle* null pairs each prompt with another clip's continuation: continuations that ignore the prompt and revert to MusicGen's grid-peaked output give a mean of 13.3 cents, range [0.4, 23.5], above the observed value (83% of draws exceed it). A *tracking* null keeps each continuation's distance from its prompt but randomises the direction: mean 0, range $[-4.1, 4.1]$, which the observed value exceeds in all but 0.05% of draws. A uniform-tuning null spans $[-13.6, 13.7]$ and does not exclude the observed value at $n = 49$. Selecting prompts on the applied shift alone, which removes regression to the mean from estimation error, gives $+6.9$ [2.2, 10.3] against that selection's own shuffle null of 12.0. The continuations therefore track the prompt (median 9.7 cents from it, 25.2 from the grid, 29% ending closer to the grid than to the prompt) with a grid-directed residual, but the statistic cannot tell a per-clip pull from partial reversion to the model's own distribution, and we do not report it as a displacement magnitude or compare it with the codec's corpus pull.

Swapping the decoder under 40 of those token sequences moves the peak/mean from 4.28 with the stock decoder, the same figure on these 40 as on the full 60, to 4.3 grid-tuned and 3.7 flattened, a

paired per-clip drop of 0.5 [0.3, 0.7] against -0.02 [-0.2, 0.2] for the grid-tuned arm. The drop is stock minus arm, so the grid-tuned arm's negative entry is a slight rise; and the flattened arm's paired median is not the difference of the pooled figures above it, 0.58.

G STATISTICAL DISCIPLINE

G.1 WHICH COMPARISONS WERE PRE-REGISTERED

The comparisons the pre-registration fixed (Section 3) are the confirmatory ones, and the Bonferroni adjustment of Figure 2 is taken over the eleven measurable codec conditions they apply to. The vocoder row sits outside that family and carries an unadjusted 95% interval, as the fine-tuned rows do. The registered pass of Appendix G.6 is confirmatory on its own terms, registered after this pre-registration and against one comparator fixed with it, and carries no further adjustment. Everything else in this paper is exploratory and is reported without adjustment: the band edge, the ladder fraction, every fine-tuning arm, the octave and vibrato sweeps, repeated coding, the MusicGen analyses, the decoder swap, the equivalence tests of Appendix F.5.3, the per-language recognition intervals, and the two fitting guards, which were written after the SNAC 24k slope was seen. The downstream tuning analysis of Section 3.4 is exploratory too, and its three measures — degree error, flips and makam identification — were computed together and carry no adjustment across that family of three; the flip contrast clears zero where the degree-error magnitude does not against the matched Opus control (it does against DAC 16k, Appendix F.4). Adjusting them does not change which ones hold. The magnitude covers zero at 95%, so it orders last, and a Holm step-down over the three leaves the flip contrast at $3 \times 4 \times 10^{-5} = 1.2 \times 10^{-4}$ and the identification drop, whose p treats the enlarged set as fixed, at $2 \times 0.011 = 0.022$, both still below 0.05. We report it because the question is the obvious one to ask of three measures computed together, not because the adjustment is what licenses them: the identification p carries the stopping-rule caveat of Appendix F.4 whatever the multiplicity correction does, and that caveat is the binding one. We report exploratory intervals as descriptions rather than tests.

G.2 BETWEEN-RUN VARIANCE, AND WHICH INTERVALS IT WIDENS

The intervals on the pitch sweeps are within-sweep intervals; Table A36 gives the unit for every other quantity, and none of them covers variation between runs. That component cannot be inferred from the single-reference repetition of Appendix F.6, whose design differs, so we measured it directly: we ran the headline sweep end to end five times in the headline design, redrawing stimulus phases and gains and re-coding from the checkpoint each time. In Table A35, *Combined* is $\sqrt{\mathit{Within}^2 + (2\,\mathit{Sd})^2}$, the 2 being a nominal display multiplier; the pull intervals below fold the same component at their own Bonferroni multiplier, and a t interval at five runs would carry $2.78\,\mathit{Sd}$ and widen the amplitude to 0.27.

Table A35: Five end-to-end repetitions of the headline sweep, EnCodec 24k at 3 kbps, 440 Hz. *Combined* adds the between-run spread in quadrature as a nominal two-sd band. The bias interval absorbs it; the amplitude interval does not.

Quantity	Five runs	Mean	Sd	Within	Combined
Off-grid bias (c)	12.55, 12.41, 12.71, 12.44, 12.60	12.54	0.122	1.48	1.50
Amplitude (c)	13.72, 13.83, 13.64, 13.80, 13.69	13.74	0.078	0.16	0.22
Slope	1.0010, 0.9962, 1.0034, 1.0021, 0.9988	1.0003	0.0028	0.0089	0.0105

The **registration verdict survives**. Widened to 1.0003 ± 0.0105, the slope interval is $[0.990, 1.011]$, still far inside the 1 ± 0.15 equivalence margin. The margin is 14 times the combined half-width, so no plausible run-to-run component threatens it.

The **bias intervals move where it matters**. On the headline row the trial bootstrap at the on-grid reference is already wide, so combining is nearly free: 1.48 becomes 1.50, and the headline 12.55 is stable across runs to 0.12 cents. That is a statement about one row. The marginal pull verdicts have far tighter trial intervals, so the same 0.122 cents is a much larger share of them. Combined in quadrature at the same Bonferroni multiplier $z = 2.84$, and taking one standard error from the

mean of the percentile interval's two half-widths rather than folding each side separately, SNAC 32k's $+0.61$ [0.18, 1.02] becomes [0.07, 1.15] and EnCodec's vowel condition's $+0.43$ [0.19, 0.60] becomes [0.03, 0.83]: both still clear zero, by hundredths of a cent. And a component estimated from five runs carries four degrees of freedom, not infinitely many.

Applying $t_4 = 5.75$ to the combined half-width would overcorrect, though. The combination pools a trial bootstrap on $n = 1728$ off-grid trials of the reduced design with a five-run component, so its effective degrees of freedom are Satterthwaite's rather than the smaller component's: $\nu = 24.3$ for SNAC 32k and $\nu = 7.3$ for the vowel condition, at $t = 3.13$ and $t = 4.04$ in place of 5.75. SNAC 32k then reads $[+0.01, +1.21]$ and the vowel condition $[-0.14, +1.00]$.

So six of the eleven clear zero on trial bootstraps, the pre-specified count Section 3.1 leads with, and five survive run-to-run variance at Satterthwaite degrees of freedom. The four exceeding five cents survive every reading; SNAC 32k clears zero by a hundredth of a cent under the fold, a verdict to read as marginal, and the vowel condition does not. Read at t_4 the count would be four; the Satterthwaite reading keeps SNAC 32k and with it a third codec family, after EnCodec and WavTokenizer. The fold is the stricter reading rather than the primary one: the five runs redraw the stimulus phases and gains that the trial bootstrap already resamples, and the headline row's 0.52 standard error already exceeds their 0.122 sd (Section 2.2), so adding the two in quadrature can count the same variation twice. The component is applied both to the positive intervals, where it makes a pull harder to claim, and to the nulls' detectable floors in Appendix E.2, where it makes a null harder to claim.

The **amplitude interval widens**. The ± 0.16 printed throughout is a spread across the ten detunings of one sweep and does not include run-to-run variation; combined it is ± 0.22. This matters most for the causal arms of Section 3.3, which are conducted in amplitude with seed-level intervals of ± 0.09 to ± 0.15. Those intervals cover optimisation noise within one draw and not run-to-run variation, so a combined half-width on each arm is nearer 0.2. The 0.91-cent contrast survives that — it is more than four times the combined component — but the five-seed spreads should not be read as the uncertainty on an arm.

G.3 What each interval covers

Two readings in Appendix C.5 (Table A15) are larger than panel (a) of Figure 2, and the reason is structural rather than a difference of stimulus. It is not the twelve partials of the spectral sweep against the pitch probe's eight: Table A15's reported statistic is *Est. 8*, the same estimator on the same eight partials.

What differs is that the pitch probe measures a *two-tone interval* while the spectral sweep reads a single tone. Write $\rho(\cdot)$ for the per-tone displacement the spectral sweep measures. The probe's residual is the difference of two such displacements,

$$r_C(\theta; d) \;=\; \rho\big((\theta + d) \bmod 100\big) \;-\; \rho(d), \tag{16}$$

so the reference tone's own displacement subtracts. Three things follow. The subtraction removes any component common to both tones. The probe's bias is then read at the on-grid reference, where the spectral sweep reads one detuning. And it is medianed over the off-grid band $|\delta| \geq 30$, where the spectral sweep reads exactly $|\delta_0| = 40$, the largest displacement available.

Evaluating Equation 16 on the measured ρ and taking that median gives 12.4 cents against the probe's measured 12.55 — so the two instruments do agree once the interval structure is accounted for, and the 20.7 is a single-tone, single-detuning reading of the same effect. A further point holds independently of all this: the reading is a property of the estimator's partial set as much as of the codec, since at 3 kbps and $\delta_0 = -40$ the eight-partial fit reads $+20.7$ cents where the same estimator restricted to the first two partials reads zero.

† A percentile bootstrap over five seeds cannot place a 2.5th percentile: its tails are pinned inside the observed range of five points, so those brackets are spread rather than coverage, on the same footing as the twelve-cluster ladder intervals above. ‡ GTZAN 199 and Saraga 195. One GTZAN clip of the 200 is dropped: its estimator cross-check fails on the uncoded audio, so it has no input reading to difference against. Saraga's 195 are the recordings the cross-check retains.

Five of these deserve a caution. The first is the sharpest: the headline sweep is *one* register in *one* timbre read at *one* reference, so its interval resamples trials of a single stimulus. That interval

Table A36: What each interval covers. The resampling unit differs by statistic because the designs differ; the point of the table is that the differences are visible in one place.

Quantity	Unit resampled	n	Level	Method
Off-grid bias, tones	trial, on-grid reference	2160 (1728 reduced design)	99.55%	percentile, adjusted
One interval, printed once: the off-grid bias and the pull verdict share it.				
Registration slope	per-condition phases	10	95%	OLS, $t(8)$
Band edge, ladder fraction	register × timbre	12	95%	BCa; under-covers
registered pass	—	—	—	pooled half-width
Zero crossing	fine-tuning seed	5	95%	percentile bootstrap; spread, not coverage[†]
Real-music bias	clip	199; 195[‡]	95%	percentile bootstrap
Isolated notes	instrument	108	95%	cluster bootstrap
Causal arms	fine-tuning seed	5	95%	Welch, exact permutation
Causal arms, ten-draw	flattening draw	10	95%	paired $t(9)$ on draw differences
Downstream flips	recording	200, 60	95%	exact McNemar; Clopper–Pearson
Makam identification	recording	235	95%	exact McNemar
Off-grid bias, population	envelope × register	6, 4	—	two-way bootstrap, both margins; design width
Extender arms	seed, plus half-width	5	95%	combined (Appendix G.2)

describes variation within one condition, not across stimuli, and the headline bias therefore carries no interval over timbres or registers, which is what Appendix C.7 supplies. Table A13 gives the effect at four octaves, but as four point estimates rather than as a distribution the bias is resampled over. The twelve-unit band-edge and ladder-fraction bootstraps under-cover at that cluster count, which is why those intervals are offered as spread rather than coverage. The stimulus-population interval belongs on this list too, and for the same reason. Its design is crossed, six envelopes by four registers, and resampling the twenty-four cells exchangeably ignores both margins: the register factor dominates, its medians running 1.73, 5.83, 8.69 and 4.26, so the effective replication there is four. Resampling each margin separately gives $[2.10, 6.98]$ over registers and $[2.50, 7.41]$ over envelopes, and resampling both gives $[1.73, 9.13]$, which is 76% wider than the $[2.75, 6.95]$ the exchangeable resample returns. We report the two-way interval because it is the one that answers "another envelope grid at another set of registers", which is what a population claim asserts. The six envelopes are a grid we chose rather than a sample from a population a reader would have in mind. The registration interval is over the ten detuning conditions of one sweep, so it describes variation across a manipulated variable and not across codecs or stimuli. And the five-seed causal interval covers optimisation noise within one draw of the flattening offsets; the corpus-level claim rests on the ten-draw row beneath it.

G.4 Every decision made after seeing results

Table A37: Analytic choices made after results were seen. Each is also stated in the section that uses it; the point of the table is that there are eighteen, and that a reader can weigh them together.

Decision	Occasioned by	What turns on it
The mechanism, and the thresholds it is read on		
The band edge and ladder fraction themselves	(P1) failing against the clipped model	The paper's mechanism, and Sections 3.2 onward
The tier boundaries at $\bar{\ell} = 0.20$ and 0.70	The exploratory phase having assigned tiers by eye	How many tier *predictions* hold: 20 of 26 at a 0.60 boundary, 17 at our frozen 0.70, 15 at 0.75. The count of checkpoints *assigned* below the boundary moves the other way (Appendix G.6)
Amplitude stability guard, $\mathrm{cv} \leq 0.35$	The SNAC 24k slope	Refuses 4 of the 18 measured (2 of the 13 neural). Stable over $\mathrm{cv} \in [0.233, 0.407)$
Retention guard, ≥ 0.25	The same	DAC 16k is admitted at 0.349; nearest refusal 0.068
Comparators and controls		
Taking the 0-or-100-cent arm as the causal comparator	Three grid controls being available once the arms were run	0.18 cents of the 0.94-cent five-seed contrast; over five seeds the three comparators give 0.94, 0.87 and 0.76
Matching the Opus control's bitrate on distortion rather than rate	EnCodec's damage at 3 kbps being read after the fact	The size of the downstream flip contrast, which the match affects; Table A31 gives the unmatched Opus 12 kbps row
Considering a 155° phase anchor, EnCodec's own, in place of the nominal 180°	Mimi's $+132^\circ$ falling outside the 45° window	Nothing: we kept the 180° anchor of Section 2.3, and Mimi fails on bias under either

Table A37, continued

Decision	Occasioned by	What turns on it
What was measured, and over how much		
The stimulus population of Appendix C.7	The observation that the headline had no unit covering stimuli	Nothing already reported; it adds the population figure and shows the headline stimulus is at that population's 92nd percentile
The extender arms of Appendix D.3	The SBR control being unable to test a generative prior	The corpus claim, which rests on the extender arms as well as on pre-trained decoders
Enlarging the makam identification set from sixty recordings to 235	The sixty-recording version returning $p = 0.29$ against a detectable drop of 0.133, i.e. a non-detection	The verdict on that task: it does not clear zero at sixty and does at 235. The enlargement was to reach a size that could resolve the measured effect, and the point estimate moves only from -0.07 at sixty to -0.06 at 235
Enlarging the held-out pass from ten new families to twenty	The ten-family version returning five of ten, which does not clear chance	Its permutation p, which assumes a size fixed in advance and so is not quoted. Reported in full in Appendix G.6
Restricting the 880 Hz sweep to $\theta \leq 600$ cents	The second tone losing partials over the top of that sweep	The 880 Hz over-prediction, 2.5 unrestricted against 1.8 restricted; it does not reach 1 either way
Regrouping the FLEURS languages to match each language's phonology	A phonological review of the grouping, which had placed Amharic in two groups and assigned pharyngealisation to three languages that do not share it	The language count, 20 in four groups against 24 before regrouping, and the grouping of a *pre-registered* test. The regrouping was made on linguistic grounds, independently of the F_0 results, and changes a confirmatory design (Appendix H)
How the results are counted and reported		
Folding the between-run component into the positive bias intervals, at Satterthwaite degrees of freedom rather than z	Applying it to the positive intervals as well as the nulls' detectable floors	A stricter reading of the pull count: five of eleven with it, against the pre-specified six on trial bootstraps. We lead with six and print both
Equivalence margin of Appendix F.5.3	The size of the observed difference	The equivalence claim between the two traditions
Reading the pull count as three families	Each family being counted once its condition clears zero	What the three are made of: EnCodec; WavTokenizer; and SNAC 32k, which sits below its own 0.71-cent power floor, clears zero by a hundredth of a cent once the between-run component is folded in, and drops out at t_4
The soft edge's two constants	$\bar{\ell}$ held at the 440 Hz value and the roll-off anchored at the low end of the measured range	Both were set in the direction that keeps Appendix C.4's fit inside its 9% tolerance; the alternatives are stated there
Mimi's separate tier	Its ladder resolving with a bias below threshold	Presentation of Figure 2 only; no interval changes

G.5 THE LIMITS, COLLECTED

The conclusion carries the limits that bear on how the headline numbers should be read. This subsection collects the limits the sections raise in Table A38, so that a reader need not assemble them one section at a time.

Of the decisions made after results were seen, the band-edge and ladder-fraction row of Table A37 carries the most weight: the mechanism this paper reports was arrived at after its pre-specified account failed, and its definitions were not committed to in advance, although individual predictions derived from it were logged before their arms ran. The registration test, the pull rule and their window were pre-specified, which is why they carry an adjustment and the rest do not.

G.6 THE HELD-OUT PASS ON TWENTY-SIX CHECKPOINTS

This subsection reports a held-out test of the frozen mechanism, read as a consistency check for the reasons set out at the end. We froze the band edge, the ladder fraction and a tier rule as the exploratory phase left them, recorded a tier prediction for twenty-six checkpoints whose ladder fraction none of us had read, and then ran them. Twenty of the twenty-six are families the exploratory phase never touched; of the remaining six, three are new operating points of codecs it measured, two are further checkpoints of families it measured, WavTokenizer at 40 tokens per second and EnCodec 32k, whose Table A16 row was measured after the freeze, so its ladder fraction was unread when the mechanism was formed, and one — SNAC 24k at its default point — is a primary condition

Table A38: The stated limits. The conclusion names the opening block; the rest are raised where they arise and collected here.

Limit	Where it is established
Carried in the conclusion	
The band-edge account follows the pre-registered one; its registered pass has fifteen checkpoints and does not separate from the most common tier	§3.3, Appendix G.6
The headline stimulus sits at its population's 92nd percentile	Appendix C.7
Two of the six pulling conditions pull by under a cent	§3.1, Appendix G.2
The corpus component isolable on an already-trained decoder is a lower bound	Appendix E.1
The grid-attributable downstream degree error does not exclude zero	Appendix F.4
The estimator we use puts more weight above the edge than a twenty-partial salience sum	Table A33
Collected here	
The cost is to measurement and to one task, not to listening	Ethics statement
SNAC 32k and the vowel condition are counted as pulls below their own 0.71- and 0.52-cent floors, which include the between-run component their verdict intervals omit	Appendix E.2.1
Mimi's null is a non-detection, not a bounded null	Appendix E.2.1
The between-run component is measured on one condition and applied to all eleven	Appendix E.2.1
Corpus-level causal evidence is GTZAN plus one Saraga arm; the pretraining corpus is never manipulated	Appendix E.1
The retuning arm varies one corpus at one shift, so its spread is over seeds	Appendix E.1
The cross-family swap rests on one reversing pair	Appendix G.6
Three estimators were run on coded audio; the rest are read off documentation	Table A32
The band limit is not demonstrated below the edge, where 2.75 cents is still measured, and it is inert on the 53% of retained Saraga frames that carry no fitted partial above the edge	Appendices F.4, F.5
The real-music reading uses the 23–40% of frames the cross-check retains	Appendix B.5
The downstream flip is exploratory and holds only for full-band fitting	Appendix F.5
The extender arms have no grid-preserving resampling control or reconstruction score	Appendix D.3
The per-partial reading is made on tones and isolated notes, on no Saraga or makam recording, though the frame-level residual there is grid-directed	Appendix F.5
Degree references are estimated from uncoded audio while the arm is coded	Appendix F.4
The remedy is demonstrated with our own estimator truncated at the edge	Appendix F.4
The ladder's correlation with regenerated count depends on which identity is read, 0.924 against 0.696	Appendix G.6
Placement-blind scoring puts 20 of 26 below the boundary rather than 16	Appendix G.6
Four of the six pulling conditions are EnCodec operating points, and at a matched edge held-out decoders differ in how much they regenerate: the mechanism is a property of particular decoders	Appendix G.6
The pre-registration and its amendment log ship in the supplement and carry no external timestamp, so the freeze dates, and what had been read when each prediction was logged, are self-certified	Reproducibility statement
SoundStream, Single-Codec and SpectroStream have no public release, so three held-out rows cannot be re-scored independently	Appendix G.6
The 100-cent radius within which a decoded line is associated with an input partial is not swept, although the level thresholds of the same rule are	Section 2.4, Appendix C.5.2

our guards refuse for the pitch probe, so it is neither a new operating point nor strictly held out (Appendix B.8 states the refusal).

Separating how many partials move from how far they go. The inclusive $\bar{\ell}$ averages over every partial at or above the edge, transmitted ones included, so it mixes two quantities: how many partials the decoder regenerated, and how far it put each one. Over the twenty-six rows the Spearman correlation between partials at or above the edge and the inclusive $\bar{\ell}$ is $\rho = 0.635$ — the quantity we assign to the decoder is substantially predicted by the quantity we assign to the code. Against the *regenerated* count, which is the one that enters the identity below, it is stronger still: $\rho = 0.924$. We report the larger figure because it is the stricter of the two.

We therefore also report a *conditional* ladder fraction $\bar{\ell}_{\mathrm{r}}$, the mean of ℓ_k over the partials the level rules classify as regenerated, with that count printed beside it. It carries no interval: at the one to three regenerated partials many rows have, a resample of it would not be meaningful. The $\bar{\ell}$ column beside it is the inclusive metric used everywhere else, including the tier rule, so the two should not be read down the same column. Transmitted partials above the edge are not scored zero: they move a little, between 0.04 and 0.11 of the way. The two metrics nonetheless carry largely the same information. Writing a for the partials at or above the edge and r for those classified regenerated, $\bar{\ell} = (r/a)\,\bar{\ell}_{\mathrm{r}} + (1 - r/a)\,t$; solving for t on the seventeen rows of Table A40 where $r < a$ — the other nine, those with $r = a$, regenerate everything above their edge and leave t undefined — gives $t \in [0.040, 0.110]$, the same interval we print for transmitted motion. So the inclusive metric is the conditional one scaled by the regenerated fraction, up to a nuisance bounded by 0.11 — which is why $\rho = 0.924$ against the count. The consequence is that $\bar{\ell}$, and therefore the tier rule and the 16-of-26 count that runs on it, measures *how much* a decoder regenerates as much as *where* it puts it. The first of those cannot be cleanly assigned to the decoder: the edge is set largely by the code, but decoder-only fine-tuning moves both the edge and the count with the tokens unchanged, so the count sits on neither side of the partition. One comparison the confound does not reach is the flattening contrast of Section 3.3: its seven arms share a 2.0 kHz edge at which every partial above is regenerated in each, $r = a = 4$ throughout, so $\bar{\ell}_{\mathrm{r}} = \bar{\ell}$ identically and no count term is present to carry any of it (Appendix E.1).

Against the conditional metric the at-or-above correlation falls to $\rho = 0.504$, and the regenerated-count correlation from 0.924 to 0.696. Only the second is the arithmetic consequence of removing a count factor, and we print both because quoting 0.504 beside 0.924 invites the reader to treat them as one comparison. The 0.924-to-0.696 drop is partly arithmetic — removing a count factor from a quantity that contains one must reduce its correlation with the count — and the at-or-above $\rho = 0.504$ at $n = 26$ is $p = 0.004$ one-sided, 0.009 two-sided, so a real residual association remains. The conditional metric reduces the confound without removing it.

What the conditional metric does settle is the cross-family comparison. The cross-family swap described below is set aside for a different reason — it cannot separate a decoder property from DAC 44k being unusual — and Section 3.2 rests the attribution on the fine-tuning arms instead; what the swap does show is narrower, and survives that objection. Vocos and DAC 44k's own decoder both regenerate three of the four partials above their shared 2.20 kHz edge, and place them at $\bar{\ell}_{\mathrm{r}} = 0.89$ and 0.17. That gap is where the partials went, not how many there were, which is the one thing the count confound could have explained away.

A caution on what sets the edge at the highest rate. At 24 kbps the edge is H_5, and Table A15 reads $s_5 = -1.7$ cents at $\delta_0 = +40$ against $+53.2$ at -40. At the first of those the placement clause — the strongest line within 100 cents sitting away from kf_0 — is satisfied only nominally, by less than the 2 cents within which Section 3.2 calls a partial below the edge unmoved, so the edge at that rate and sign rests on the level clause alone. This is also the row carrying the largest per-sign gap in the paper, 0.86 against 0.58, which average to 0.72. We report the mean over signs throughout; at the highest rate it averages two quite different readings.

A further caution, on what the column can discriminate. Across the seven primary conditions $\bar{\ell}$ runs 0.72 to 0.90 while their off-grid bias runs from -0.13 to 12.55, two of the seven being counted null. So on the set the mechanism is built from, the decoder-assigned quantity is close to constant, and what separates a pulling condition from a null one is the edge and the weighting rather than the ladder fraction. That is consistent with the account, since the ladder says where invented partials go and not how much of the spectrum is invented, but it means $\bar{\ell}$ carries little cross-condition information

there, and the correlation with regenerated count reported above is computed on the held-out rows and not on these.

One reading accounts for part of the per-sign gap. Every row of Table A16 carrying both signs reads higher at $\delta_0 < 0$ than at $\delta_0 > 0$, and on fourteen of the twenty the gap is 0.07 to 0.09 — what a constant additive offset in the measured shift would produce, since an offset of c cents gives a per-sign gap of exactly $c/20$ at $|\delta_0| = 40$, so $c \approx 1.6$ reproduces it. Section 3.2 already reports that same $+1.6$ cents as the arithmetic mean gap between the two ladder targets, measured with no codec involved.

That accounts for the modal gap and not for the column. Four rows read 0.12 to 0.13 and two read 0.25 and 0.28, which would need c from 2.4 to 5.6, and a constant offset is constant. We do not read the excess as a second effect either: the two largest gaps sit on rows where the intervals nearly meet, 0.86 [0.75, 0.96] against 0.58 [0.42, 0.74] and 0.91 [0.83, 0.99] against 0.66 [0.52, 0.80]. A 1.6-cent offset is sufficient for the gap most rows show. The gap also appears on arms read against their own grids and on the flattened arm, which has no grid to displace toward, so it is not read as a learned displacement; Appendix C.5's half-sum reading treats the same asymmetry.

Two cautions on reading the column. Nine of the twenty-six rows regenerate one or two partials, so their $\bar{\ell}_{\mathrm{r}}$ is a mean over one or two numbers and is not comparable with a six-partial reading — the same caution Appendix E.1 applies to fine-tuned Mimi. And *regenerated* is defined partly by displacement (the strongest line within 100 cents sits away from kf_0), so conditioning on it selects for partials that moved: the conditional metric is an upper bound on how far a regenerated partial travels, not an unbiased estimate of it. That is a bias and not a floor, so a low conditional reading is evidence and not an impossibility — the clause asks only that the strongest line sit away from kf_0, not far away, so a partial moving a few cents against a δ_0 of forty still counts, which is what DAC 44k's own decoder and SpeechTokenizer do at 0.17 in Table A40 and what the cross-family swap's 0.17 is. Nor does it follow that the low checkpoints merely regenerate less; that reading would be an artefact of the selection rule.

Read with those cautions, the conditional metric narrows the held-out picture no further than the inclusive one does. The two metrics disagree here, and the disagreement can be settled because the two selection rules are not equally entangled with the quantity they score. The conditional rule admits a partial when the strongest line within 100 cents sits *away from* kf_0, a condition on displacement, which is what the ladder fraction measures: it selects on the estimand. The level-only rule admits a partial when its line has fallen more than 10 dB, wherever the strongest nearby line sits; it selects on attenuation and is blind to placement. Where the two disagree, the level-only rule agrees with the inclusive metric: all twelve checkpoints reading $\bar{\ell} \leq 0.44$ still read ≤ 0.44 under it, while on the conditional metric nine of the twelve instead place what they regenerate high, which is what conditioning on displacement predicts — the partials it discards are the ones that did not move, and those are the ones sitting near the input's own harmonic. This is an argument about which selection is less circular, not a demonstration that the level rule is unbiased: attenuation and regeneration are themselves correlated, so a 10 dB rule is only *less* entangled with placement — a better instrument than a displacement rule for deciding whether a low reading is real, and not a clean one.

That settles one direction. Low rows staying low leaves the mechanism's reading unchanged; the rows the mechanism runs on are the high ones of Table A16 — 0.90 for EnCodec 24k at 3 kbps, 0.89 for its GTZAN fine-tune, 0.85 for DAC 16k — so every primary row whose edge resolves, the fine-tuning base arm and the fourteen held-out rows above 0.44 are re-scored under the level-only rule, which lowers most of them. Table A39 is the result on the seven primary rows whose edge resolves, plus the fine-tuning base arm.

Six of the eight rows fall, by 0.11 to 0.17, and the two that do not are the two whose edge is high enough that everything below it is transmitted cleanly: with $a = 1$ there is no attenuated-but-unmoved partial for the level rule to admit. So the bias is real, it is the direction Section 2.4 predicts, and at up to 0.17 it exceeds the 0.12 of ladder fraction Appendix C.5 quotes as the clause's cost.

The fourteen held-out rows above 0.44 run the same way. Nine fall by 0.10 to 0.17 and five do not move, having no attenuated-but-unmoved partial below their edge for the level rule to admit. Four of the nine cross the frozen 0.70 boundary, so the count of checkpoints below it reads 20 of 26 placement-blind against 16 of 26 as scored. Under the level-only rule *fewer* held-out decoders reach the ladder tier. We report the frozen count, and this beside it.

Table A39: **The same rows scored without the placement clause.** a is the count at or above the edge, t the ladder fraction of the partial the lower edge admits; blind reading $(a\bar{\ell}+t)/(a{+}1)$; dash, edge unmoved. Edge is Table A16's cross-register mean; counts are at 440 Hz.

Condition	Edge (kHz)	Blind edge	$a \to a'$	t	$\bar{\ell} \to \bar{\ell}$ blind
EnCodec 24k, 3 kbps	1.3	0.88	$6 \to 7$	0.05	$0.90 \to 0.78$
EnCodec 24k, 24 kbps	2.2	1.76	$4 \to 5$	0.05	$0.72 \to 0.59$
EnCodec 48k, 6 kbps	1.6	1.32	$5 \to 6$	0.06	$0.79 \to 0.67$
WavTokenizer	0.9	0.88	$6 \to 7$	0.05	$0.84 \to 0.73$
DAC 16k[†]	4.4	3.5	$4 \to 5$	0.05	$0.85 \to 0.69$
Mimi	3.4	–	$1 \to 1$	–	$0.75 \to 0.75$
SNAC 32k	3.1	–	$1 \to 1$	–	$0.86 \to 0.86$
EnCodec, GTZAN fine-tune	2.0	1.76	$4 \to 5$	0.06	$0.89 \to 0.72$

[†]Counted at 880 Hz, the lowest register at which a probe partial clears its 4.4 kHz edge; the other rows are counted at 440 Hz.

What survives is the separation the mechanism actually runs on. The twelve low checkpoints stay at or below 0.44 under the same rule, and the held-out column above them is continuous from 0.55 up, so the blind rule compresses the column without opening or closing a gap. We keep the placement-conditioned $\bar{\ell}$ as the primary metric because it is the one the freeze before this pass named, not because it was pre-registered (Appendix G.4). A reader who prefers the placement-blind rule should read the headline as nearer four fifths of the way onto the grid than nine tenths, and every ladder comparison in this paper as compressed by about that much. The other fine-tuned arms, the bypass, the extender and the remaining EnCodec rates are read on the placement-conditioned rule, as is the 0.86-to-0.59 contrast of Section 3.3; every partial above those arms' shared 2.0 kHz edge is classified regenerated in every arm (Appendix E.1), so a level-only re-score would add a similar admitted-partial term to each and compress the contrast rather than remove it. Appendix C.5.2 sweeps the placement clause alongside the two dB thresholds rather than holding fixed the one ingredient we name as circular.

The held-out generalisation is therefore narrow on the inclusive metric and broader on the conditional one. Twelve of twenty-six read an inclusive $\bar{\ell} \leq 0.44$, and that survives scoring regeneration on level alone; on the conditional metric nine of those twelve instead place their regenerated partials between 0.49 and 0.81, and only SemantiCodec (0.19), DAC 44k (0.17) and SpeechTokenizer (0.17) stay low on both. Where the question is how far the partials a decoder does regenerate have moved, the conditional reading is the better instrument, and we use it as one, the cross-family swap being read on it precisely because both decoders are scored on the partials each regenerated. What separates the twelve is not the edge. Eight held-out checkpoints sit at EnCodec's own 1.32 kHz edge with six of eight partials above it, and their inclusive $\bar{\ell}$ still spans 0.26 to 0.83; what varies across them is the regenerated count, two to six, a third quantity belonging cleanly to neither side of our partition and one that decoder-only fine-tuning moves. Conditional on regenerating, those eight agree closely: $\bar{\ell}_{\mathrm{r}}$ from 0.62 to 0.84. At a matched edge these decoders differ in how much they regenerate and agree about where they put what they do. The mechanism is a property of particular decoders rather than of neural codecs at large.

What was frozen. The band edge is Equation 4 with the 6 and 10 dB level thresholds unchanged; the ladder fraction is the mean of ℓ_k over partials at or above it, read at $|\delta_0| \geq 20$. The tier rule, written for this pass since the exploratory phase assigned tiers by eye, is $\bar{\ell} \geq 0.70$ for *on the ladder*, $0.20 \leq \bar{\ell} < 0.70$ for *part of the way*, and $\bar{\ell} < 0.20$ for *on the input's ladder*. Predictions came from published architecture and bitrate alone, on the reasoning that a codec regenerating more of its spectrum lands further onto the grid, and were recorded before any checkpoint was coded.

That reasoning is the one Section 3.2 replaces, and this arm is where it fails. It ties the *ladder* to bitrate, where the band-edge account ties only the *edge* to it; and the rule's misses run one way, over-predicting how far a decoder goes on seven of its nine misses. Read through the edge, a low measured $\bar{\ell}$ on this pass means the rule mistook an edge for a ladder. It predicts a high $\bar{\ell}$ wherever the bitrate is low, but the checkpoints it misses sit at comparable band edges, with four to seven of eight partials at or above them against the hits' three to seven. They do not, however, regenerate as many: on the regenerated count the misses run two to five (median three) against the hits' one to seven (median five), so the rule's misses are decoders that invent less, not decoders that invent as much and place it lower; the at-or-above column, read as though it were the regenerated one,

would say the opposite. That is the finding of this pass, and it is why the edge is the first quantity to measure and not the only one. We have kept the rule and its predictions exactly as frozen rather than re-scoring them under the later account, which would not be a held-out test of anything.

The twenty new-family rows are BigCodec (Xin et al., 2024), NeuCodec (Julian et al., 2025), X-Codec 2.0 (Ye et al., 2025b), Vocos (Siuzdak, 2024) on two token sets, TAAE and its released Stable Codec (Parker et al., 2025), HiFi-Codec (Yang et al., 2023), FunCodec (Du et al., 2024), SemantiCodec (Liu et al., 2024), SoundStream (Zeghidour et al., 2022), Single-Codec (Li et al., 2024), AudioDec (Wu et al., 2023), APCodec (Ai et al., 2024), SpectroStream (Li et al., 2025), TiCodec (Ren et al., 2024), MuCodec (Xu et al., 2024), Language-Codec (Ji et al., 2025a), FreeCodec (Zheng et al., 2025) and LSCodec (Guo et al., 2025). SoundStream, Single-Codec and SpectroStream have no public release, and their weights were provided by the authors on request; the Vocos that decodes DAC tokens is a community checkpoint trained on DAC 44k tokens. Each is identified with its revision in the shared code. The six further checkpoints come from codec families already in Table A5.

Table A40: Held-out pass, twenty-six checkpoints. Tiers by inclusive $\bar{\ell}$: *ladder* ≥ 0.70, *part* 0.20 to 0.70, *input's* < 0.20; ✓ marks prediction meeting assignment. Brackets, twelve-cluster BCa read as spread (Appendix G.3); **Regen.** counts regenerated partials, over which alone $\bar{\ell}_{\mathrm{r}}$ is taken.

Checkpoint	Point	Edge (kHz)	At or above	$\bar{\ell}$	Predicted	Assigned	Regen.	$\bar{\ell}_{\mathrm{r}}$
Families never measured								
BigCodec	1.04 kbps	0.88	7 of 8	0.88 [0.74, 1.01]	ladder	ladder ✓	7	0.88
NeuCodec	0.8 kbps	0.88	7 of 8	0.86 [0.72, 0.99]	ladder	ladder ✓	7	0.86
X-Codec 2.0	1 kbps	1.32	6 of 8	0.83 [0.70, 0.95]	ladder	ladder ✓	6	0.83
Vocos	EnCodec tokens	1.32	6 of 8	0.71 [0.55, 0.86]	ladder	ladder ✓	5	0.84
Vocos	DAC tokens	2.20	4 of 8	0.68 [0.52, 0.83]	ladder	part	3	0.89
TAAE	research ckpt	1.32	6 of 8	0.62 [0.47, 0.77]	ladder	part	5	0.73
HiFi-Codec	default	1.76	5 of 8	0.34 [0.20, 0.48]	ladder	part	3	0.53
Stable Codec	released ckpt	2.20	4 of 8	0.31 [0.17, 0.45]	input's	part	2	0.54
FunCodec 16k	Q8	1.76	5 of 8	0.24 [0.10, 0.39]	ladder	part	2	0.49
SemantiCodec	default	2.64	3 of 8	0.09 [−0.04, 0.22]	input's	input's ✓	1	0.19
SoundStream	3 kbps	0.88	7 of 8	0.84 [0.71, 0.97]	ladder	ladder ✓	7	0.84
Single-Codec	0.3 kbps	0.88	7 of 8	0.81 [0.68, 0.94]	ladder	ladder ✓	7	0.81
AudioDec	6.4 kbps	1.32	6 of 8	0.76 [0.62, 0.90]	ladder	ladder ✓	6	0.76
APCodec	6 kbps	1.76	5 of 8	0.72 [0.57, 0.87]	ladder	ladder ✓	5	0.72
SpectroStream	4 kbps	2.20	4 of 8	0.63 [0.48, 0.78]	ladder	part	3	0.81
TiCodec	1 kbps	0.88	7 of 8	0.55 [0.41, 0.69]	ladder	part	5	0.75
MuCodec	0.35 kbps	2.20	4 of 8	0.44 [0.29, 0.59]	input's	part	2	0.81
Language-Codec	2 kbps	1.32	6 of 8	0.38 [0.24, 0.52]	part	part ✓	3	0.71
FreeCodec	0.45 kbps	1.32	6 of 8	0.29 [0.15, 0.43]	part	part ✓	2	0.77
LSCodec	0.25 kbps	1.76	5 of 8	0.17 [0.03, 0.31]	input's	input's ✓	1	0.57
Further checkpoints of codec families already measured								
WavTokenizer	40 tok/s	0.88	7 of 8	0.81 [0.68, 0.94]	ladder	ladder ✓	7	0.81
EnCodec 32k	2.2 kbps	1.76	5 of 8	0.79 [0.66, 0.92]	ladder	ladder ✓	5	0.79
SNAC 24k	default	1.32	6 of 8	0.28 [0.14, 0.42]	ladder	part	2	0.62
Mimi	Q4	1.32	6 of 8	0.26 [0.12, 0.40]	part	part ✓	2	0.68
DAC 44k	Q8	2.20	4 of 8	0.14 [0.01, 0.27]	input's	input's ✓	3	0.17
SpeechTokenizer	Q4	2.20	4 of 8	0.11 [−0.02, 0.24]	input's	input's ✓	1	0.17

What happened, and where the predictive content actually is. Seventeen of twenty-six predictions hold (Table A40), an accuracy of 0.65 with an exact interval of [0.44, 0.83]. The split matters more than the total.

First, the split. On the six remaining checkpoints — three new *operating points* of codecs already measured, two further checkpoints of measured families, plus the refused SNAC 24k — five of six hold. But predicting Mimi at Q4 after measuring Mimi at Q8, or DAC 44k at Q8 after Q4, is closer to a lookup than to a prediction. Those six carry the least information and we do not rest anything on them. On the twenty checkpoints from families the exploratory phase never touched, twelve hold, an accuracy of 0.60 [0.36, 0.81]. We read both Clopper–Pearson intervals as descriptions of the rows scored.

One row needs a word. SNAC 24k at its default point is the condition Appendix B.8 refuses, so it is neither a new family nor a new operating point; we include it because the refusal is of the *pitch tracker* on the two-tone task, while the ladder fraction is read off the decoded spectrum and does not use the tracker. The two readings come from different stimuli, the two-tone probe and the single steady spectral-sweep tone, so each is read on its own task. It is scored as it fell, a miss.

The twenty new-family rows do not span twenty independent systems. Vocos appears twice, once on each codec's tokens, so those two rows share an architecture but not one set of weights; and Stable Codec is TAAE's released checkpoint, so those two share an architecture at different operating points, which is why their edges and frozen predictions differ. Eighteen distinct systems are represented. The permutation null below treats all twenty rows as exchangeable, which they are not; that is a further reason its p is not quoted.

Second, the null. Treating the three tiers as equiprobable would understate the baseline: our predictions were seventeen *ladder*, three *part* and six *input's*, where the assignments came out ten, twelve and four, so the majority-class predictor, always *part*, already scores twelve of twenty-six, and always *ladder* ten. A margin-fixed permutation null (`confirmatory_null.py`, 2×10^5 draws) expects 8.85 hits of twenty-six and 7.00 on the twenty new checkpoints, against the seventeen and twelve observed. We do not quote its p-values. The arm was enlarged mid-course after the first ten families returned five hits, and no permutation p computed as though the size were fixed in advance is valid under that stopping rule.

That null is not the decisive comparison either. It holds both margins fixed, which excludes constant predictors from its reference set, and the constant predictor is the comparison a reader will reach for. The twenty new-family rows are assigned 8 *ladder*, 10 *part* and 2 *input's*, so answering *part* every time scores 10 of 20 against the frozen rule's 12. The two disagree on 18 rows, 10 in the rule's favour and 8 against, an exact McNemar p of 0.815. The rule is therefore not distinguishable from the majority class on this arm, and we report the pass as a consistency check rather than as confirmation.

We then ran a registered pass. Before reading any $\bar{\ell}$ we froze the edge definition, the ladder fraction, the tier rule, both boundaries, the comparator and the sample size, registered them, and scored 15 public checkpoints untouched by either earlier arm. Unlike the held-out arm, whose predictions came from architecture and bitrate, each registered prediction came from the checkpoint's band edge, measured from its spectrum before its ladder fraction was read and taken as the mean over the registers where it resolves, so those edges are not confined to multiples of the reference as the held-out arm's are: the number of the probe's eight partials at or above that edge set the predicted tier, six to eight predicting *ladder*, two to five *part*, and zero or one *input's*. No checkpoint drawn had fewer than two partials above its edge, so the *input's* tier was never predicted and the two rows scored into it miss by construction. The prediction is a partial count read off an edge that carries the placement clause of Section 2.4, so this pass tests the count relation described above as much as placement. The frozen rule holds 11 of 15 against the constant predictor's 5, the two disagreeing on eight rows, seven in the rule's favour and one against: an exact McNemar p of 0.070. That is a long way from this arm's 0.815, and short of 0.05 at this size: with eight discordant rows only an 8-to-0 sweep reaches $p < 0.05$. Enumerating the exact McNemar distribution over discordant rows, nine are the fewest that can reject with one row against and twelve with two; a rule winning discordant rows at the rate this one did reaches 85% power at about thirty checkpoints, drawn from outside the forty-one already scored. Table A41 gives the contingency the test runs on, so that both counts and the p can be checked. No row's $\bar{\ell}$ falls between 0.60 and 0.70, so the pass reads 11 against 5 at both lower boundaries by construction rather than by robustness; the informative endpoint is 0.75, where it weakens to 11 against seven at $p = 0.289$, the two rows whose $\bar{\ell}$ lies between 0.70 and 0.75 moving into *part*. Read continuously instead of in tiers, the same rows give a rank correlation of 0.79 between partials above the edge and $\bar{\ell}$ ($t(13) = 4.7$, $p = 4 \times 10^{-4}$); that is a post hoc reading of the registered data, not the registered test, and it inherits the count relation the tier test does. Four of the fifteen intervals cross a boundary at the half-width Table A42 prints, which is pooled rather than per-row, so that count rests on it. The 15 are 15 distinct systems, no family appearing twice, assigned 8 *ladder*, 5 *part* and 2 *input's*; the constant predictor is the same always-*part* rule the previous arm used, so it is right on exactly those five rows, and its lower rate here than there — 5 of 15 against 10 of 20 — is the tier distribution differing rather than a change of comparator. Against that split the frozen rule takes four of the five *part* rows and seven of the other ten, which is the 11 and the seven-to-one discordance the table reports. Against always *ladder*, the majority tier in this sample, the discordant split is 4 to 1, since the rule got neither *input's* row right and so all seven of its other hits are *ladder* rows, for an exact McNemar $p = 0.375$.

Table A42 gives those 15 rows individually, so that each prediction can be read against the tier its $\bar{\ell}$ placed it in. The 15 are `Plachta/FAcodec`, `Zuellni/BiCodec` (a community copy of Spark-TTS's codec), `zhai-lw/SQCodec`, `amphion/dualcodec`,

Table A41: **The registered pass, rule against constant predictor.** The 15 checkpoints of the registered arm, cross-classified by whether each predictor assigned the tier the checkpoint was scored into. The exact McNemar test reads the off-diagonal alone: 7 against 1, two-sided $p = 0.070$.

	Constant right	**Constant wrong**	**Total**
Frozen rule right	4	7	11
Frozen rule wrong	1	3	4
Total	5	10	15

Table A42: The registered pass, row by row. The edge is measured; the partial count and predicted tier follow from it under the registered rule, the assigned tier from $\bar{\ell}$. **Regen.** counts the regenerated partials $\bar{\ell}_{\mathrm{r}}$ is taken over. Brackets are a pooled half-width of 0.13, not per-row resamples.

Checkpoint	**Edge (kHz)**	**At or above**	$\bar{\ell}$	**Predicted**	**Assigned**	**Regen.**	$\bar{\ell}_{\mathrm{r}}$
FAcodec	1.01	6 of 8	0.85 [0.72, 0.98]	ladder	ladder ✓	5	1.00
BiCodec	0.90	6 of 8	0.89 [0.76, 1.02]	ladder	ladder ✓	6	0.89
SQCodec	1.12	6 of 8	0.75 [0.62, 0.88]	ladder	ladder ✓	5	0.88
DualCodec	1.10	6 of 8	0.74 [0.61, 0.87]	ladder	ladder ✓	5	0.88
RepCodec	1.13	6 of 8	0.83 [0.70, 0.96]	ladder	ladder ✓	5	0.98
LLM-Codec	0.96	6 of 8	0.84 [0.71, 0.97]	ladder	ladder ✓	5	1.00
NanoCodec 22k	0.99	6 of 8	0.89 [0.76, 1.02]	ladder	ladder ✓	6	0.89
FlowDec	2.13	4 of 8	0.72 [0.59, 0.85]	part	ladder	3	0.94
PAST	2.33	3 of 8	0.45 [0.32, 0.58]	part	part ✓	2	0.64
MagiCodec	2.44	3 of 8	0.55 [0.42, 0.68]	part	part ✓	2	0.79
UniCodec	1.95	4 of 8	0.34 [0.21, 0.47]	part	part ✓	2	0.61
FocalCodec	2.22	3 of 8	0.34 [0.21, 0.47]	part	part ✓	1	0.87
XY-Tokenizer	1.21	6 of 8	0.36 [0.23, 0.49]	ladder	part	2	0.93
HILCodec	2.86	2 of 8	0.11 [−0.02, 0.24]	part	input's	1	0.15
U-Codec	2.55	3 of 8	0.07 [−0.06, 0.20]	part	input's	1	0.07

`Hecheng0625/RepCodec`, `voidful/llm-codec-librispeech` (itself a fine-tune within that family), `nvidia/nemo-nano-codec-22khz-0.6kbps-12.5fps`, `facebookresearch/FlowDec`, `slprl/PAST`, `Ereboas/MagiCodec_16k_50hz`, `Jiang-Yidi/UniCodec`, `lucadellalib/focalcodec_50hz`, `MCplayer/XY_Tokenizer`, `aask1357/hilcodec` and `shaunxsyang/U-Codec`, each with its revision in the shared code. Three of them read $\bar{\ell}_{\mathrm{r}}$ at 0.98 or above, so what those decoders regenerate sits on the grid harmonic rather than part way to it, and thirteen of the fifteen read 0.61 or higher.

Third, the intervals, on the held-out rows. Eighteen of the twenty-six $\bar{\ell}$ intervals cross one of the two tier boundaries and two more, X-Codec 2.0 and HiFi-Codec, touch one at an endpoint; only BigCodec, NeuCodec, SoundStream, TiCodec, MuCodec and Language-Codec are boundary-safe. The tier assignments are therefore sensitive to the boundaries in a way the guard thresholds of Appendix B.8 are not, and the boundaries themselves were chosen knowing the exploratory $\bar{\ell}$ values. Two different counts are read off this table: the number of checkpoints *not assigned* the *ladder* tier is 13, 15, 16 and 18 at ladder boundaries 0.60, 0.65, 0.70 and 0.75, while the count of predictions holding moves the other way, 20 to 15. The first count at the frozen boundary is the 16-of-26 figure, and it is a different quantity from the number of checkpoints whose *predicted* tier is the one assigned; both are reported here. Seven of the 16 place what they do regenerate at 0.71 or higher. Sweeping them: at a ladder boundary of 0.60 the predictions holding number 20 of 26; at 0.65 it is 18; at our pre-frozen 0.70 it is 17; at 0.75 it is 15. That count is not the placement-blind count of checkpoints *assigned* below the boundary, which is also 20 of 26. We report the pre-frozen value rather than the best of the four.

The two held-out misses that under-predict, Stable Codec and MuCodec, were predicted on the input's own ladder and landed part way; the registered arm's one under-predicting miss is FlowDec.

Two rows carry more than their share, and they carry unequal weight.

The **two decoders on one token set** comparison is the strong one. On DAC 44k's tokens Vocos lays its partials at $\bar{\ell} = 0.68$ [0.52, 0.83] where DAC 44k's own decoder reads 0.14 [0.01, 0.27]: the same code, the same 2.20 kHz edge and the same four partials above it, a gap of 0.54. The intervals do not

overlap, but Appendix G.3 reports them as spread rather than coverage, so we read the gap and not the non-overlap: whatever fixes the ladder is downstream of the tokens. On EnCodec's tokens the same swap goes the other way and is smaller, 0.71 [0.55, 0.86] against EnCodec's own 0.90, whose lower per-sign reading is 0.84 — the two decoders differ on both token sets, in opposite directions.

The **one decoder on two token sets** comparison is not resolved. Vocos reads 0.71 and 0.68 on the two, which differ by far less than either half-width, so the data are consistent with a constant ladder and equally consistent with a drift of up to about 0.3; and since the two rows are different weights, a constant reading would not have isolated one decoder in any case. Edge inheritance (1.32 and 2.20 kHz) is close to entailed, since a decoder cannot regenerate below where its code stops carrying. Of the two swap comparisons only the first bears on the partition, and Section 3.2 rests the partition on the fine-tuning arms rather than on it. The DAC row is scored a miss, 0.68 falling under the 0.70 boundary.

What this pass establishes. The pass also shows, on two checkpoints we did not build, that the ladder differs at fixed tokens; whether that is a property of decoders in general or of DAC 44k's decoder in particular is not settled by one pair, which Table A38 carries as a stated limit. The corpus claim rests on Section 3.3, since none of the twenty-six was fine-tuned.

H THE SPEECH TESTS IN FULL

Speech is quiet ground but not silent: the bias on speech-shaped stimuli is 0.43 cents [0.19, 0.60] at $+164°$, which clears zero on trial bootstraps, covers it once the between-run component is added at Satterthwaite degrees of freedom (Appendix G.2), and reads thirty times smaller than the tone condition, a voice's above-edge energy being spread over dense low-amplitude harmonics rather than eight strong partials.

Of two pre-specified tests on FLEURS (Conneau et al., 2022), one refutes its prediction and the other does not resolve its prediction. Over twenty languages the tone group's F_0-contour error ratio to the control is 1.048 [0.86, 1.25] under EnCodec, excluding the predicted 1.6; with six tone languages the test detects a ratio of 1.31 at 80% power, so the exclusion is real but rests on six units. We draw no conclusion from the recognition test, and read the F_0 test as evidence against its own prediction.

H.1 LANGUAGE GROUPS

FLEURS languages are assigned to the contrast each carries: lexical tone (Vietnamese, Thai, Cantonese, Mandarin, Yoruba, Igbo), ejective release (Amharic, Georgian, Oromo), click consonants (Zulu, Xhosa), and a non-tonal high-resource control (English, French, German, Spanish, Italian, Portuguese, Polish, Dutch, Czech). Arabic is measured and reported individually. Hebrew, Amharic and Maltese are not grouped with it under pharyngealisation: Modern Israeli Hebrew has largely lost the pharyngeals for most speakers, Maltese ġ is generally silent or realised as vowel lengthening, and Amharic lost the Ge'ez pharyngeals and carries ejectives instead. The design has 20 languages in four groups, 21 with Arabic, and no language in more than one group.

H.2 PHONOLOGICAL CONTRASTS IN FULL

Table A43 gives the F_0-contour test in full. The ratio reported is of the group's median error to the control group's, under each codec, with intervals bootstrapped over languages. The tone group's ratio under EnCodec, 1.048 [0.86, 1.25], is the number quoted at the head of this appendix; the ejective and click groups have three and two languages, too few to bootstrap; their ranges are reported for completeness and should not be read as tests. The tone row is an F_0-contour ratio against a predicted 1.6; the ejective and click rows are log-spectral distance (LSD) ratios against a predicted 1.4, a different registered prediction. No row reaches its own threshold.

H.3 DOWNSTREAM RECOGNITION IN FULL

This test does not resolve its prediction, and it is reported because it was pre-registered. The relative word-error increase under coding is 32.7% [11, 45] for the tone group against 64.8% [29, 94] for

Table A43: Error ratios of group median to control median for three codecs. *Range over per-language ratios, not an interval; those rows are not tests. Brackets are 95% intervals.

		EnCodec 3 kbps		Mimi		DAC 16k	
Contrast	**Metric**	**ratio**	**CI / range**	**ratio**	**CI / range**	**ratio**	**CI / range**
Lexical tone (6)	F_0 (c)	1.048	[0.86, 1.25]	0.921	[0.73, 1.20]	1.041	[0.82, 1.24]
Ejective (3)*	LSD (dB)	1.080	0.96 to 1.12	1.044	0.96 to 1.11	1.024	1.00 to 1.03
Click (2)*	LSD (dB)	1.077	1.07 to 1.08	1.120	0.95 to 1.29	0.972	0.96 to 0.98

the control, so the point estimate falls on the wrong side of the predicted $1.5\times$ while the intervals overlap. The sign depends on a normalisation the pre-registration did not fix: in absolute terms the tone group rises $+0.090$ against the control's $+0.059$, a ratio of 1.53 and the right side of the prediction. The tone group's baseline error is 3.7 times the control's, which is what makes the two disagree.

Group sizes differ between recognisers because a language enters a group only where the recogniser supports it. Under MMS (Pratap et al., 2024), which covers 18 of the 21 languages, the groups are tone (4), ejective (2), click (2) and control (9), with Arabic the eighteenth and reported on its own rather than in a group. Table A44 gives the group summaries under three normalisations; figures are the median over each language's own relative increase, so they do not equal the ratio of the pooled columns, and intervals are bootstraps over languages, except on the two-language ejective and click rows, where the bracket is the range over the two languages. Whisper (Radford et al., 2023) was run as well and is reported with the accompanying results rather than in the table: it cannot transcribe seven languages, four of them contrast-bearing, and three of its controls have baselines near zero, so a relative increase over them is noise over noise. Its failure modes on low-resource audio are documented (Koenecke et al., 2024), so no contrast here rests on it. Per-language values, the level normalisation and the recogniser-failure exclusions are in the accompanying result tables.

Headroom normalises the increase by the error the recogniser had left to make, $1-$baseline, averaged over recordings rather than formed from the pooled margins printed in the table; the two agree only where a group's recordings share a baseline, which is why five of the eight rows recompute from the printed columns to within 0.002 and three — the two tone rows and the EnCodec control — do not.

Table A44: Downstream recognition by contrast group, under three normalisations of the error increase. Non-control baselines are 2.7 to 3.7 times the control's, and the sign of the tone-versus-control difference depends on the normalisation; *Headroom* and the brackets are defined above.

Codec, recogniser	**Group**	**Baseline**	**Absolute**	**Headroom**	**Relative**
EnCodec, MMS	Control (9)	0.091	$+0.059$ [0.015, 0.101]	0.067 [0.016, 0.114]	64.8% [29, 94]
	Tone (4)	0.340	$+0.090$ [0.041, 0.126]	0.150 [0.047, 0.200]	32.7% [11, 45]
	Ejective (2)	0.244	$+0.114$ [0.062, 0.166]	0.150 [0.082, 0.217]	47.1% [25, 70]
	Click (2)	0.315	$+0.185$ [0.167, 0.203]	0.269 [0.250, 0.289]	59.2% [50, 68]
Mimi, MMS	Control (9)	0.091	$+0.129$ [0.079, 0.213]	0.142 [0.083, 0.240]	154.0% [111, 188]
	Tone (4)	0.340	$+0.134$ [0.108, 0.276]	0.238 [0.136, 0.383]	67.8% [20, 103]
	Ejective (2)	0.244	$+0.253$ [0.205, 0.301]	0.334 [0.273, 0.395]	104.2% [82, 127]
	Click (2)	0.315	$+0.308$ [0.289, 0.327]	0.451 [0.411, 0.490]	97.6% [97, 98]

I WHAT IS OURS AND WHAT IS NOT

Our nearest neighbour is the work of Berendes and colleagues, and because the boundary is narrow we set it out claim by claim in Table A45 rather than in prose. The table bounds against that work only. It is not a full accounting against every neighbour: the copy-up literature (Appendix D.2), Terhardt's virtual pitch, the upsampling-artefact account (Pons et al., 2021; Odena et al., 2016), Paek et al. (2025)'s report of pitch structure in codec latents (Appendix C), Di Giorgi et al. (2022)'s report of mel inversion destabilising pitch on sustained notes, and the tuning-confound work of Qin & Lerch (2019) and Rodríguez-Algarra et al. (2019) bound different claims, and Section 1 separates them in prose. A ✓ in the *Ours* column of the lower block therefore means new relative to Berendes, not new absolutely — the last row in particular states a consequence whose underlying confound Qin & Lerch (2019) established and whose intervention method is Rodríguez-Algarra et al. (2019)'s.

Table A45: The boundary against the closest prior work. ✓ established there; ∼ raised but not measured, or partly; — absent. 6/11 is the pull count at a 440 Hz reference. pYIN, an off-the-shelf tracker, largely avoids the effect (Section 3.4).

Claim	Berendes 2025	Berendes 2026	Ours
Neural synthesis pulls output tuning toward 12-TET	✓	✓	✓
Magnitude on real recordings	✓ 18.7 c	✓	✓ 2.5 c
(different systems and estimators; not comparable)			
The residual is anchored to absolute pitch	✓ observed	—	✓ slope test
The fundamental moves with it	✓ entangled	—	vocoder only
It does not happen in every system	∼ "some"	—	✓ 6/11 at 440 Hz
The corpus is the source	∼ posed	✓ mitigated	✓ translated
Capacity governs its size	✓ mel bands	—	✓ a band edge
A perceptual test	✓	—	—
A mitigation	—	✓	✓ band-limited
Which partials move, and how far	—	—	✓
The fundamental measured and unmoved	—	—	✓ EnCodec
Edge and ladder separated	—	—	∼ partly
The grid *translated* by a predicted amount	—	—	✓
Shown in RVQ token codecs as well as vocoders	—	—	✓
(and not to require the quantiser; bounded)			
Consequence for non-Western tuning analysis	—	—	✓

J NEGATIVE RESULTS AND A REPLACED VALIDATION

What did not work. Training a codec from scratch on the stimuli. A small residual quantiser trained on these tone pairs either collapses to silence or reproduces only the fundamental, with h_2/h_1 falling from 0.500 to 0.003. Since a pure sinusoid shows no grid effect through any codec we tested, as a signal with nothing above the band edge must, a model that outputs near-sinusoids cannot exhibit the phenomenon whatever it was trained on.

Fine-tuning on synthetic tones. It cuts reconstruction error eighteenfold and grid bias fortyfold regardless of the training pitch density, including on held-out vowel stimuli the codec never saw.

Both leave no band above the edge for a ladder to act on: the from-scratch model outputs near-sinusoids, and fine-tuning makes the codec transmit the tones' few partials and regenerate nothing. Training on broad real audio, as in Section 3.3, is what made the causal experiment work.

Randomising the codebook. Replacing entries with random vectors of matched scale destroys the representation, giving 785-cent errors with no pitch left to measure, too destructive to isolate anything.

An untrained network. A randomly initialised convolutional autoencoder outputs noise, 405-cent errors. The classical controls of Appendix D.1 answer that question without this problem.

Speech codecs on musical tones. SpeechTokenizer fed harmonic complexes gives roughly 100-cent errors, the two estimators disagreeing on 97% of trials. A 16 kHz speech codec treats an isolated harmonic complex as far out of distribution; the source-filter vowel of Section 2.5 cuts its error to 0.65 cents.

Makam validation. The first design read scale-degree offsets from Arel-Ezgi-Uzdilek comma tables (Bozkurt et al., 2014). In that theory Segah and Evc are the third and seventh degrees of Rast, at 17 and 48 Holdrian commas of $1200/53 \approx 22.6$ cents, one comma below the Pythagorean major third and major seventh of 407.8 and 1109.8 cents. That places them 15.1 and 13.2 cents below the nearest 12-TET pitch, as Appendix F.5 reports. Measured practice differs from the tables by several cents, the size of the effect, so the theory positions would have contributed an annotation error as large as the quantity measured. The reported analysis (Section 3.4) therefore takes each degree's reference from the corpus's own measured histogram, pooled over that corpus's recordings in that makam, keeping the theory tables as the secondary comparison of Table A31.