WHERE CODECS DRAW THE LINE: ALLOCATION PROFILES FOR SPEECH TOKENISERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantising a speech representation is dividing it: a codebook partitions latent space, and every distinction finer than a cell is destroyed. Analyses of speech tokenisers ask what their tokens retain, not where their boundaries fall. We characterise a partition by the reconstruction error each phonetic category incurs, an *allocation profile*, comparable where the partitions themselves are not. It separates two things a single test conflates: added levels may narrow the gaps between categories, or change which ones are finely resolved. Across four widely used codecs and one we train to reverse a single architectural choice, depth does the first, and only quantisation in a learnt low-dimensional projection at every level does the second. And what the first level resolves finely is largely a restatement of the signal: energy, duration and frequency account for 82.9% of one codec's acoustic profile against 16.2% of the distilled stream trained beside it, whose allocation bears no detectable relation to the acoustic one. Depth mostly refines an allocation that architecture has already settled.

1 INTRODUCTION

Neural audio codecs have become a standard front-end of speech language models: a waveform is turned into a sequence of codebook indices, modelled as a vocabulary (Borsos et al., 2023; Défossez et al., 2024). The resulting tokens are widely read as proto-categories, probed for phonetic content, aligned against phone inventories, split into "semantic" and "acoustic" streams on the assumption that the first carries linguistic content (Zhang et al., 2024; Pasad et al., 2021); downstream systems inherit that reading, using codec tokens for unit-based translation or resynthesis, which presuppose tokens that keep phonemic identity and discard other variation (Polyak et al., 2021; Lee et al., 2022). That any of this should work is surprising, as nothing in how a codec is trained refers to language.

A codec is trained to reconstruct its input. An encoder maps the waveform to latent frames, a quantiser replaces each by the nearest of the K codewords of a learnt codebook, and a decoder reconstructs the signal from them. The index a frame receives is the token a downstream model reads,[1] and K (with the depth of the cascade, introduced in §2) is all the resolution the codec has to spend. The codebook is fitted to minimise the expected squared error of the replacement under that budget, so fitting it is fitting a *partition*: everything in a cell is replaced by the same vector and finer distinctions are destroyed. Where the boundaries fall is what a codec decides.

That criterion has a long history in speech, where vector quantisation was introduced to reduce transmission rate (Makhoul et al., 1985). Classical quantisation theory shows that, when codewords are plentiful, they are best placed in proportion to how often each region of the space is visited: resolution follows *mass*, fine where frames are frequent, coarse where they are rare (Zador, 1982; Gersho, 1979). The codecs we study are far from that regime,[2] but it says what a codec is *trying* to do when nothing tells it about language.

[1]Terminology, used consistently throughout: a quantiser assigns each frame an *index*, which points to a *codeword*, a vector in the space being quantised; a *codebook* is the set of codewords one quantiser holds. The region nearer to one codeword than to any other is its *cell*; the cells of a codebook form a *partition* of the space, also called a Voronoi *tessellation*, with as many dimensions as the space, not the two of the figures. What a speech language model reads as a *token* is the index, or the tuple of indices across levels of a cascade; the decoder consumes only codewords. Work on speech tokenisers is almost entirely about indices; this paper is about codewords.

[2]The result is asymptotic in K, requiring cells small enough that the data density is constant within one. The density of codewords per dimension, $K^{1/d}$ in d dimensions, here ranges from 1.01 to 2.38, far from that condition.

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.

The linguistic categories a codec is credited with capturing are not, however, organised by frequency. Under the theory of adaptive dispersion, phonological inventories arrange themselves so that their categories stay mutually discriminable in a perceptual metric, largely regardless of how often they occur (Liljencrants & Lindblom, 1972; Lindblom, 1986; Schwartz et al., 1997): resolution follows *contrast*. Mass and contrast divide the same acoustic space into regions treated as equivalent (Voronoi cells on one side, phonological categories on the other) under objectives that have no reason to agree. Whether a codec's cells end up near the categories speech is organised by motivates this paper, but the question it answers comes first: *what does residual quantisation do to acoustic space?* How does a codec's division of it change as levels are added, and what in its design decides where the boundaries fall? The answer matters beyond description: a distinction finer than a cell cannot be recovered from the tokens, so a contrast a codec leaves unresolved is unresolved for every model built on top of it.

This question cannot be answered by comparing the two divisions directly. In the dimensions codecs use, Voronoi cells can be neither enumerated nor visualised, and computing the volume of one is #P-hard (Dyer & Frieze, 1988). Modern codecs compound this by quantising through a cascade, in spaces whose dimension varies across codecs. We therefore characterise a partition not by the geometry of its cells but by *what it does to phonetic categories*: for each category we record the reconstruction error its frames incur, giving an allocation profile comparable where the cells are not.

To see what in a codec's design decides where the boundaries fall, we study two decisions, one acting on the space in which the partition is drawn and one on the objective that fixes it: whether each quantiser operates directly in the latent space or in a learnt low-dimensional projection of its own, and whether the first level of the cascade is additionally trained to predict a self-supervised speech model. Four widely used codecs span these decisions but differ in everything else, so we add a codec we train ourselves, reversing one decision with the rest held fixed: a comparison between models becomes a counterfactual about a design choice.

Related work. Existing analyses of speech tokenisers do not look at boundaries. Probing, the dominant approach, asks what linguistic information is *present* by training a classifier on frozen features (Belinkov & Glass, 2019): it measures whether a distinction survives somewhere, not where the codec has drawn its boundaries. We ask instead how a codec's finite resolution is *distributed* over the inventory: which categories are carved finely, which coarsely, and whether that changes as levels are added. The two can come apart (a probe can succeed on a category the quantiser resolves poorly, provided the residual is linearly separable), and the partition, unlike probe accuracy, is a property of the codec alone.

Work on the codecs themselves divides along similar lines. Benchmarks evaluate what tokens support downstream (Wu et al., 2024; Mousavi et al., 2026) and report, consistently with what we find, that early codebooks carry most of the phonetic content while later ones add redundancy; they cannot say whether that redundancy refines the existing division or reorganises it, which is the distinction our measure is built to make. Interpretability work aligns individual indices with attributes such as phonetic class or speaker identity (Sadok et al., 2025), and utilisation studies ask how much of a codebook is used (Zheng et al., 2025); both read the index, whereas we read the codeword it points to and its distance to the frame it replaces. Semantic distillation, one of our two design choices, is by now a family rather than a technique (Zhang et al., 2024; Défossez et al., 2024; Ye et al., 2025); we analyse two of its members and make no claim about the rest. Recent surveys map the design space more fully than we can (Guo et al., 2026).

We make three contributions. *Allocation profiles* compare quantisations of acoustic space that cannot be compared cell by cell (§2). Using them, we separate two properties a refinement may or may not preserve: the *gaps* between phonetic categories, and their *ranking*. This matters because depth does the first while architecture does the second (§4–§5); a codec trained under matched conditions attributes that difference to a design choice rather than to the model exhibiting it. We then ask what fixes the allocation at the first level, which the rest of the cascade largely inherits (§6).

2 Framework

What a codec learns, and what fixes its partition. Nearly all current neural audio codecs share one design: a convolutional encoder–decoder trained end to end, with a *residual vector quantiser* (RVQ) in the bottleneck. Two things are learnt at once, and separating them matters. The encoder and decoder learn a *representation*, which properties of the waveform are worth keeping, under

a composite waveform, spectral and adversarial objective. The quantiser learns a *partition* of it, and only the representation is shaped by that objective: the quantiser minimises squared error, its codewords moving towards the centroids of the frames assigned to them, whatever the encoder encodes. Nothing in the objective asks the resulting cells to line up with linguistic categories.

Whether they do is an empirical question, and answering it requires saying what a codec spends where. By the *resolution* devoted to a region we mean the fineness of the partition there: how small the cells are, and hence how little of a frame's detail is lost when it is replaced by a codeword.[3] Two independent design choices distinguish current codecs, and they govern different aspects of how that resolution is spent. *Factorisation* is whether each level holds its own learnt low-dimensional projection and searches a normalised space, rather than quantising in the latent space the encoder produces; *semantic distillation* is whether the first level is additionally trained to predict a self-supervised speech model. A projection shared by a whole cascade is not factorisation in this sense: what the definition turns on is each level having a geometry of its own (§3). Factorisation has a second consequence that will matter: with the lookup in a much lower dimension, the same codebook holds more than twice as many codewords per dimension, so the codecs differ in how densely they populate the space they quantise.

The cascade. A codec is rarely used at one level of precision. A single codebook can only be made finer by enlarging it, and the number of codewords it needs grows exponentially with the precision wanted: matching what L levels of K codewords achieve would take K^L. RVQ instead stacks L quantisers, each with its own codebook of K codewords and each encoding the error the previous ones left (Juang & Gray, 1982):

$$r_1 = z, \qquad k_\ell = \arg\min_{k \le K} \|r_\ell - c_{\ell,k}\|, \qquad r_{\ell+1} = r_\ell - c_{\ell,k_\ell}, \qquad \hat{z}_{1:L} = \sum_{\ell \le L} c_{\ell,k_\ell}, \tag{1}$$

where z is a latent frame, $c_{\ell,k}$ the kth codeword of codebook ℓ, and $\hat{z}_{1:L}$ the reconstruction from the first L levels, which we call the *depth* of the cascade. In a factorised codec each level searches its codebook in a projection of r_ℓ of its own, and c_{ℓ,k_ℓ} in equation 1 stands for the chosen codeword mapped back into the latent (Figure 2). Any prefix of the indices is a valid encoding: the first L levels give a coarser but usable representation. Precision is therefore chosen after training by truncation, which downstream models exploit routinely (Borsos et al., 2023; Défossez et al., 2024).

A cascade hands us directly the successive divisions we want to compare, produced by one model on one dataset. Geometrically, what it does at each level is subdivide. The first level partitions the latent space into K cells; the second splits each of those into at most K again, according to where the residual falls; and so on. A codec of depth L therefore produces not one partition but a nested family of them, each a refinement of the last. Two properties of that construction matter later. The set of reachable reconstructions is the *Minkowski sum* of the L codebooks, so the composite codebook is generated by translation, the same level-ℓ offsets around every level-$(\ell{-}1)$ codeword, and a cascade cannot place its K^L reconstructions as freely as a single codebook of that size. And each level sees only what the previous ones left behind, not the data itself.

Allocation profiles. We characterise a codec's partition by what it does to phonetic categories. An inventory of such categories gives a reference division of the same space, annotated independently of any codec, against which the codec's own division can be read. It requires a corpus with phone-level alignment, from which each frame receives a label; each category φ then corresponds to a set of frames F_φ, distributed over whichever cells the codec happens to place there. Frames are labelled by the phone containing their midpoint, so a frame straddling a boundary takes the label of whichever phone owns its centre; we do not exclude boundary frames, and coarticulation is therefore present in every category. Which inventory we use, and at what granularity, from a narrow phonetic transcription to one close to a phonemic inventory, is a question we return to in §3. The *distortion* of φ at depth L is then

$$\varepsilon_\varphi(L) \;=\; \frac{1}{|F_\varphi|} \sum_{z \in F_\varphi} \left\| z - \hat{z}_{1:L} \right\|^2. \tag{2}$$

A category with low distortion is one whose frames are reconstructed accurately: each sits close to the representative standing in for it, which usually means the cells covering that region are small. Note

[3] Resolution here is spectral, not temporal. Frames arrive at a rate fixed by the encoder's stride, which differs across the four codecs (Appendix A); being a property of a codec and not of a level, it is constant along the depth axis we compare on.

that ε_φ assumes neither that a category occupies a single cell nor that a cell holds a single category: it averages over as many cells as the category's frames fall into. Defining it this way makes an otherwise untestable intuition measurable, "the codec resolves this category finely" being a claim about cell sizes that the dimension puts beyond examination. It remains a proxy rather than a measurement of cell volume, and three things can lower a category's distortion: the cells covering it may be small, the category may be compact or quiet to begin with, or its extent may fall inside a single frame. What ε_φ records is the outcome of all three, and how much the category's own properties account for is measured in Appendix G. The vector $\boldsymbol{\varepsilon}(L) = (\varepsilon_{\varphi_1}(L), \ldots, \varepsilon_{\varphi_n}(L))$ over the n categories is the *allocation profile* at depth L; computing it at each depth reads one reference division as the space is refined.

This departs from how quantised representations are usually read. That literature works with the indices: their mutual information with phone labels, their clustering, the token stream as a symbol sequence (Lakhotia et al., 2021; Borsos et al., 2023). We never use the index. What enters equation 2 is the *codeword* and its distance to the frame it replaces, because that is what a codec does: the index only points to a codeword, and the decoder consumes codewords.

What the cascade can redistribute. A category's distortion has no meaning in isolation, only relative to something else. Profiles are useful because they are vectors over a fixed index set, so two can be compared entry by entry, which the partitions cannot. The cascade supplies it: profiles read at successive depths share categories, measurement and model. A profile can also be set against properties of the inventory external to the codec, such as how often each category occurs; that comparison is less direct, requiring a view of what to expect, and we take it up in §6.

The cascade makes a specific prediction. Level 1 sees the full distribution of latent frames and sets the shape of the partition; deeper levels see only residuals, which carry progressively less information about where a frame came from, and by the translation property above they apply what they learn identically around every coarse codeword. Both point the same way: refinements should divide cells everywhere alike, not concentrate codewords where the coarse partition serves some categories poorly. We state this as the hypothesis to test:

> **H_0.** *A codec's allocation profile is fixed by its first level: deeper levels divide every category's distortion ε_φ by the same factor, leaving the profile unchanged up to scale.*

Two vectors over a fixed index set can differ independently in how far apart their entries are and in which are largest. H_0 claims neither changes, and splits into two testable statements:

- **H_0-amplitude**: the *gaps* between categories are preserved, a gap being the ratio of two distortions and hence a difference of logarithms. We measure the spread of the profile, $\sigma_\varphi[\log \boldsymbol{\varepsilon}(L)]$: the standard deviation of $\log \varepsilon_\varphi(L)$ *across the n categories*, which asks how unevenly the codec's error is spread, and which the common factor H_0 posits leaves invariant.
- **H_0-ranking**: the *ranking* of categories is preserved: $\rho(L) = \mathrm{corr}(\log \boldsymbol{\varepsilon}(L), \log \boldsymbol{\varepsilon}(1))$ remains at one. Being a correlation across categories, ρ is blind to any monotone rescaling of the profile and reads only which categories the codec resolves best. We report Spearman unless stated otherwise, since our hypothesis is a statement about ranks.

No single statistic separates the two failures: raising every distortion to a power $\alpha < 1$ shrinks the spread while preserving the ranking exactly, and exchanging two categories' values breaks the ranking while leaving the spread intact. Figure 3 in Appendix C exhibits both on synthetic data, where refinements can be constructed rather than observed. Geometrically, the first failure is *levelling*: the cascade divides cells more finely where the coarse partition served a category badly, evening out how uniformly a codec errs across the inventory. The second is *re-ranking*: it moves resolution from one region of the space to another, changing which phonetic categories are finely resolved. Only the second can move a codec towards or away from the phonetic division of the space; whether it also changes what the tokens support downstream is a question we do not settle here (§7). Our central empirical claim is that the two regimes are realised by different architectures: *depth levels distortion across the inventory in every codec, and only the factorised cascade reorders it.*

3 Experimental setup

Codecs. We study four RVQ codecs in wide use, chosen so that the two design choices of §2 are crossed as far as released models allow: EnCodec (Défossez et al., 2023) adopts neither, SpeechTok-

enizer (Zhang et al., 2024) and Mimi (Défossez et al., 2024) are distilled, and DAC (Kumar et al., 2023) is factorised, each level projecting the latent to 8 dimensions through a learnt map and ℓ_2-normalising encoding and codebook before the lookup. No released codec is both. The two distilled codecs differ in where distillation applies: SpeechTokenizer trains the first level of its cascade, Mimi a separate one-level "semantic" stream beside an acoustic stream with its own projections. We analyse Mimi's acoustic stream throughout, levels counted within it (Appendix A.1). Mimi also projects, from 512 dimensions to 256, but once per stream and without normalising, so it is not factorised in the sense of §2. Lookup dimensions, frame rates, codebook sizes, checkpoints and objectives are in Appendix A.

A codec trained under matched conditions. Released codecs differ in far more than the two design choices we study, so the panel establishes an association at best. We therefore train a toy RVQ codec on LibriSpeech (Panayotov et al., 2015) in two conditions differing only in factorisation: a non-factorised, non-distilled codec against its factorised counterpart, with the same encoder and decoder architecture, codebook size, depth, optimiser, data and objective, ten seeds each. It is not meant to be competitive, but a run is admitted only if it reaches DAC's own reconstruction regime. Architecture, schedule and the admissibility criterion are in Appendix B.

Data. All analyses use TIMIT (Garofolo et al., 1993), read American English, on the full train and test split; no codec here, released or trained by us, has seen it in training. We prefer TIMIT to larger corpora because its phone alignment is manual: a per-category analysis inherits the labels' noise, and forced-alignment errors correlate with phonetic class. We read it under two inventories. The first is the raw 61-label transcription, unfolded, with silence held out, leaving 58 categories: a transcription convention rather than an inventory of contrasts, claimed only to divide the same acoustic space from outside the codec, as a common reference must. The second is the standard 39-label folding (Lee & Hon, 1989), which merges allophonic distinctions and removes the stop closures, leaving 38 speech categories, closer to an inventory of contrasts and with no sparsely populated category. Unless stated otherwise, results are read under the first.

None of our conclusions turns on allophonic distinctions. The folding moves individual values, and one substantially, DAC's levelling in §4; but that comes from removing the closures, not from merging allophones, and no conclusion changes with it. What we report is therefore about phone categories at a granularity coarse enough to be read phonemically, not about how finely a codec resolves allophones.

What we measure. A single extraction pass stores, for every frame and level, the selected index and the frame's phone and speaker labels; $\varepsilon_\varphi(L)$ is recomputed from these as in equation 2, with $\hat{z}_{1:L}$ assembled from the stored indices in the encoder's latent space rather than from a quantiser's internal residual, a different quantity for a factorised codec (Appendix G). Codecs are compared at matched depth $L{=}7$, the depth of Mimi's acoustic stream in its standard configuration, which also rules out DAC's extra levels as the source of any difference; DAC itself is read only up to $L \approx 9$: each of its levels quantises in a 8-dimensional projection of the latent, so its reconstruction spans only a small part of the latent, and by then its cumulative error has stopped falling (Appendix F).

Confidence intervals resample speakers, since frames within a speaker are not independent (Appendix D). We check codebook usage before any analysis, since a partially collapsed codebook would make every subsequent measure hard to read; level-1 usage is in Appendix A. Appendix G records the measurement pitfalls we had to navigate, several of which invalidated intermediate conclusions of our own.

4 Depth levels the tessellation, in every codec

This section tests H_0-amplitude: whether the gaps between categories survive as levels are added. We report $\sigma_\varphi[\log \boldsymbol{\varepsilon}(L)]$ (§2) at every depth to the matched $L{=}7$, with the mean $\mathbb{E}_\varphi[\log \boldsymbol{\varepsilon}(L)]$ as control: a factor common to every category shifts the mean but not σ_φ, a dispersion of logs, so a narrowing of the spread, the levelling of §2, cannot be a by-product of falling error.

Table 1 gives the trajectories, and its first column already separates the codecs: at level 1 EnCodec's error is spread very unevenly over the inventory, DAC's is nearly flat, and SpeechTokenizer and Mimi lie between. Geometrically, EnCodec's first codebook draws cells of markedly different size across the space the categories occupy, where DAC's treats them almost alike: the four do not start from the same tessellation.

Table 1: Allocation profile spread across depth: $\sigma_\varphi(1)$ in absolute terms, then $\sigma_\varphi(L)$ as a ratio to it. The last column, the fall in mean log distortion from $L{=}1$ to $L{=}7$, is a control. CI in Appendix D.

	$\sigma_\varphi(1)$	$\sigma_\varphi(L)\,/\,\sigma_\varphi(1)$					$\Delta\,\mathbb{E}[\log\varepsilon]$
		$L{=}2$	3	4	5	$L{=}7$	
EnCodec	0.4591	0.99	0.97	0.95	0.93	0.89	0.45
SpeechTokenizer	0.0647	1.01	0.96	0.90	0.85	0.77	0.29
Mimi	0.0722	0.75	0.66	0.61	0.57	0.53	-0.08
DAC	0.0211	1.34	1.16	0.95	0.82	0.65	0.24

They nonetheless do the same thing to it. Every codec's spread falls between $L{=}1$ and $L{=}7$, by between -11.4% (EnCodec) and -47.0% (Mimi), with intervals on the paired difference excluding zero in every case (Appendix D): the inventory ends up more evenly served than level 1 left it, so the cells covering a badly served category shrink proportionally more than those covering a well served one. H_0-amplitude therefore fails for every model. Mimi is a useful check on the control: its mean log distortion *rises* monotonically while its spread falls by nearly half, so its levelling cannot be a by-product of error shrinking. The rise itself reflects not the allocation but the projection Mimi quantises through: its cascade converges in the projected space, towards a point that is not the nearest one to its target in the latent (Appendix G).

The intermediate columns show different paths to that change: Mimi does half of it at the first residual level, EnCodec declines gradually, and DAC is not monotone, its spread rising by a third at $L{=}2$ before falling. Where a codec levels, and so where its tessellation stops changing shape, differs across the panel.

The levelling is not an arithmetic consequence of adding codebooks. Subdividing every cell by the same pattern scaled to its size, whatever lies in it, is the refinement H_0 describes; constructed, it moves the spread slightly *upward* while error falls everywhere (Appendix C), and the levelling in Table 1 departs from it in the opposite direction.

How much a codec levels depends on the inventory it is read over. Under the standard 39-label folding (Lee & Hon, 1989), the other three level much as before (Table 4), but DAC's levelling nearly vanishes, from -34.6% to -3.0%, still significant but barely. The folding merges allophones and removes the stop closures, and only the second matters here. At level 1 DAC's closures stand apart from every other category, enough to raise its spread by half ($\sigma_\varphi(1) = 0.0211$ with them, 0.0140 without); by $L{=}7$ they have come into line (0.0138 against 0.0136). DAC's levelling is thus mostly its cascade bringing a handful of near-silent categories into line, where the other three even out their error across the whole inventory. That the closures are the inventory's lowest-energy categories fits what the first-level profile is made of (Appendix G).

Levelling is therefore shared by every architecture in the panel, a residual cascade reconstructing its worst-served categories proportionally better as levels are added whatever its quantisers' design, but not implemented alike: the four start from very different profiles and level along different trajectories. It also closes off an objection: all four level while the ranking survives in three, so a change of ranking cannot be a by-product of levelling.

5 Factorised quantisation changes the best-resolved categories

Are the categories a codec resolves finely at level 1 still the ones it resolves finely at depth? Figure 1(a) answers with $\rho(L)$, the rank correlation with the level-1 profile defined in §2 for exactly this question, and depth does not affect the four alike. Three hold their ranking through the whole cascade, none falling below 0.868: the categories their first codebook reconstructs best are still the best seven levels later, which on the reading of §2 means the refinements of §4 shrink cells without moving resolution between regions. DAC falls to 0.434: its cascade has changed which categories it reconstructs best.

The shapes say more than the endpoints. DAC does not collapse at the first residual level: its ranking survives the first two almost intact, then falls steeply from $L{=}4$ on, with no level at which the decline stops. Reallocation begins part-way down and, once begun, does not stop. The other three are not exactly flat: ρ falls monotonically by 0.046 for SpeechTokenizer and 0.132 for Mimi, a drift rather than strict preservation, and Mimi's is large enough to be worth naming as such. However, their gap

to DAC widens with depth rather than closing, and since all four are read at the same depth, none of it follows from DAC's extra levels.

One reading would be that DAC's deeper levels stop carrying phonetic structure at all, the falling ρ recording a representation going quiet rather than one being reorganised. A frame-level phone probe on $\hat{z}_{1:L}$, asking how much phone identity survives quantisation, does not support it: its accuracy changes little with depth for any of the four, and DAC's rises slightly over the range in which its ρ falls (Appendix E). What changes there is which categories are resolved, not how much.

Is the re-ranking real? Two checks on ρ itself leave the separation where it was. Being a correlation with level 1, ρ would fall with no reallocation at all if the level-1 profile were noisily estimated, a real concern since factorisation degrades level 1. Correcting it for unreliability, with each profile's reliability estimated by recomputing it on two halves of the speakers, leaves DAC far below the other three (0.589 against ≥ 0.940, Pearson since the correction requires it; Appendix I). Nor does changing which categories index the profile: dropping the four sparse categories changes no ordering at any depth, and over the 38 categories of the standard folding DAC reaches 0.363 against 0.766 or more for the other three (Appendix D).

A more substantial objection concerns what ε_φ measures. A category's distortion reflects the size of the cells covering it, but also properties it has before any codec sees it: energy, duration and frequency. To weigh them, we regress $\log \varepsilon_\varphi(1)$ on these three covariates across categories. They account for between 0.347 and 0.971 of it across the four codecs, and they do not cancel between depths, as quantities independent of L might be expected to: the same covariates still account for between 0.298 and 0.933 of the trajectory $\log \varepsilon_\varphi(7) - \log \varepsilon_\varphi(1)$. We therefore recompute ρ on the residuals of these regressions, taken at level 1 and at the matched depth (Appendix G). The control moves the two codecs whose level-1 profiles the covariates explain most, as the decomposition predicts (0.971 for EnCodec and 0.829 for Mimi, against 0.556 for DAC), and leaves the other two almost where they were. EnCodec falls from 0.988 to 0.854. Mimi rises, from 0.868 to 0.952: part of its drift was the covariates' doing rather than reallocation. DAC moves by less than 3.1×10^{-5}, the finest difference a rank correlation over 58 categories can express. The separation survives, the gap between DAC and the lowest of the other three going from 0.434 to 0.420.

The counterfactual. Two questions remain that the four released codecs cannot settle. The codec that re-ranks is their only factorised one, so the association cannot be told apart from anything else particular to DAC. And DAC's level-1 profile is the flattest of the four ($\sigma_\varphi(1) = 0.0211$ against 0.4591 for EnCodec), so its initial ranking is weakly determined: a low ρ may record little order to preserve rather than a cascade overturning one, which no reliability correction can rule out, since a flat profile can be highly reproducible and DAC's is. The two matched conditions of §3 answer both, differing in factorisation alone and starting, on the measure they are read with, from profiles of comparable spread (Appendix B).

The latent-space profile cannot be read on them. A profile is interpretable only while the cascade is still closing the gap between its reconstruction and what it can reach; DAC meets this up to $L \approx 9$ and is bounded there (Appendix F), but the factorised condition of our ablation fails it from the first levels on. (Mimi's latent error rises from the first residual level too, but because its cascade converges in its projected space towards a point that is not the nearest one in the latent; Appendix G.) We therefore read the ablation after decoding, with a per-phone mel distance — a worse measure of the partition, since it sees the decoder too, and used nowhere the latent measure is valid. What licenses it is a rule fixed beforehand: a candidate is admitted only if, on the non-factorised condition, where both measures are valid, its $\rho(L)$ trajectory reproduces the latent one. The mel distance does, following it at 0.918 on average across seeds (a Pearson correlation over the eight depths; Appendix B); a linear probe does not and is excluded. The comparison that carries the conclusion is then internal to the ablation, both conditions read with the same measure, and only its outcome, whether factorisation lowers ρ, is set beside DAC's.

On that measure the effect reproduces. Factorisation lowers ρ at every depth tested (one-sided Mann–Whitney $p \leq 0.0031$, unadjusted), and the gap between conditions narrows with depth without closing (Fig. 1(b)); what grows is the spread between seeds, not a convergence of the two. The ablation also reproduces the pattern visible in DAC, the factorised condition beginning from a markedly worse level 1 and catching up over the cascade on the measure it is read with (Appendix B). Two aspects of the design, a seed count raised after a null result and one run excluded by a pre-registered floor, are disclosed in Appendix B; no conclusion turns on either.

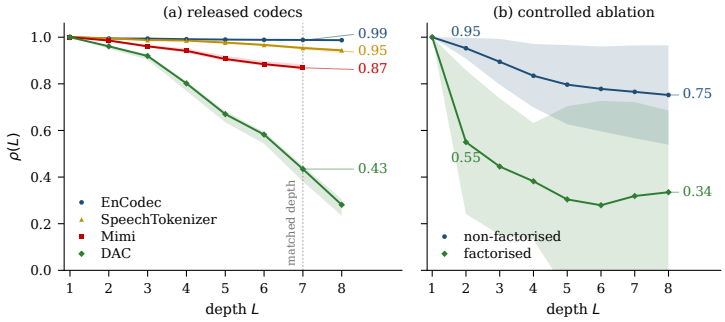


Figure 1: Rank correlation of the allocation profile with its level-1 value against cascade depth, with values at the matched depth annotated. *(a)* Released codecs, Spearman, with 95% speaker-bootstrap intervals; dashed line at the matched depth $L{=}7$. *(b)* The controlled ablation on the qualified measure, mean over seeds with ± 1 standard deviation across them.

H_0-ranking therefore holds for three codecs and fails for the factorised one, and the counterfactual lifts this beyond a single observation. It does not isolate a mechanism: factorisation degrades level 1 and shifts reconstruction work into the cascade, and the re-ranking follows, but whether that is a consequence of the degraded starting point or of the geometry producing it is not separable with the evidence at hand. What the evidence supports is the section title: factorised quantisation, with a learnt low-dimensional projection at every level, changes which categories a codec reconstructs best, where quantising in the latent space, or in one projection shared by the whole cascade, largely preserves the order it starts from.

6 How the distilled and acoustic streams differ at level 1

Both preceding sections concern what depth changes. Since the profile is set at level 1 and depth largely rescales it, the prior question is what fixes it there. Two candidates present themselves. The first is the input: every codec here is trained to reconstruct its waveform, and a category that is loud, long or frequent may incur a distortion of its own whatever the partition does. The second is the objective the first level is trained on, which for two of the four includes predicting a self-supervised speech model.

The first is measurable directly. We regress $\log \varepsilon_\varphi(1)$ across categories on three properties that owe nothing to the codec, a category's mean waveform energy, its mean duration in frames and the log of how often it occurs, none of which passes through the encoder, and read the share of the profile they account for (Appendix G). That share is 0.971 for EnCodec and 0.829 for Mimi: for codecs trained to reconstruct, the level-1 profile largely restates the signal, leaving little room for an objective to settle it, and bounding what this section can claim. Frequency is one of the three, and once within-category dispersion is held constant it pulls the way allocation by acoustic mass predicts, frequent categories being reconstructed better, in three codecs out of four; only two of the three reach significance, and the fourth shows no such relation (Appendix H).

The second candidate is approachable, though not in isolation, because one codec runs two quantisers in parallel. Mimi splits its bottleneck: a one-level semantic stream and a multi-level acoustic stream read the same encoder output, and the decoder receives the sum of what they reconstruct. The semantic stream is additionally trained to predict a self-supervised model, the acoustic one is not, and they share encoder, data and training run (Appendix A.1). Reading the same decomposition on each stream's first level therefore holds those three fixed while varying the objective. It does not vary it alone: each stream projects into and out of the quantised space with weights of its own, and §5 has just shown that the geometry of that projection can move an allocation by itself. This is the closest the paper comes to isolating the objective, and not close enough to license a causal reading.

Table 2 gives the comparison. The share of the level-1 profile accounted for by the three covariates falls from 0.829 in the acoustic stream to 0.162 in the distilled one, and the same contrast holds under the second inventory (0.786 against 0.188): the stream trained with the self-supervised objective depends far less on these signal-level covariates than the one beside it.

Table 2: Mimi's two streams at level 1, read with the decomposition of Appendix G. The residual profile is $\log \varepsilon_\varphi(1)$ with the three covariates partialled out; intervals bootstrap over categories.

	acoustic	distilled
share of $\log \varepsilon_\varphi(1)$ explained by signal properties	0.829	0.162
σ_φ of the profile	0.0722	0.0286
σ_φ of the residual profile	0.0298 [0.0242, 0.0333]	0.0262 [0.0204, 0.0299]

The distilled stream's allocation is also a different one, not a sharpened version of the acoustic stream's. With the covariates removed from both, the two residual profiles correlate at 0.180 ($p = 0.176$) over the 58 categories both resolve: what the distilled stream resolves finely bears no detectable relation to what the acoustic stream does, although the two quantisers see the same encoder output.

Two limits on this. The distilled stream starts from a much flatter profile, which invites the reading that its low share reflects having little left to explain rather than having been decoupled. Two things tell against it: the share is scale-invariant, so a uniformly flattened profile would keep the one it had, and the explained part shrinks by 5.7 between the two streams where the profile as a whole shrinks by 2.5 — faster than flattening alone would detach it. Nor is what remains noise: the distilled stream's residual profile is as reproducible across two halves of the speakers as its raw profile (0.9937 against 0.9937, split-half), and the acoustic stream's nearly so (0.9864). The reading is bounded rather than excluded, and no second matched pair of streams exists among the four released codecs to settle it against. The second limit: the effect is not a gradation across those four. Ordering them by how much their level-1 indices tell one about phone identity — the phone-normalised mutual information or PNMI, the standard index-based measure of what distillation is claimed to improve, normalised by label entropy, plug-in, with a bias-corrected estimator leaving the ordering unchanged — does not order the dispersion of their residual profiles, so nothing generalises from the within-Mimi contrast to the rest.

7 Conclusion

Across four widely used codecs, adding levels narrows the differences in how well phonetic categories are reconstructed, universally enough to tell us nothing about architecture, but it leaves the ranking almost exactly where level 1 put it, in three codecs out of four. The exception is the factorised codec, and a matched-condition codec we train ourselves, varying that choice alone, reproduces the effect. What level 1 puts there is, for a codec trained to reconstruct, largely a restatement of the signal: energy, duration and frequency account for most of it. Across two streams of one codec, sharing encoder, data and run, the one trained to predict a self-supervised speech model is far less explained by those properties and bears no detectable relation to the allocation beside it; the streams also differ in their projections, so the objective is not isolated.

Three limits bound this. Our reference is the TIMIT phone inventory and our corpus read laboratory speech, so whether these patterns survive spontaneous speech or other languages is untested; under a folding closer to an inventory of contrasts the re-ranking stays where it was but DAC's levelling shrinks by an order of magnitude, so the two diagnostics do not depend on the inventory alike. The counterfactual isolates one design choice, but that choice is a compound — factorisation here means projecting to a low-dimensional space *and* normalising within it — and no panel codec is both factorised and distilled, so the two choices are crossed only three ways out of four. Most importantly, our measure is centred on categories where §1's motivation is on contrasts: an allocation profile records how accurately each category is reconstructed one at a time, whereas adaptive dispersion is a claim about how discriminable categories are from one another — two categories can both be reconstructed well without the boundary between them being finely drawn. Re-ranking therefore establishes that a codec reconstructs a different set of categories best, not that a different set of contrasts is resolved better; whether it propagates that far, to phone recognition, minimal-pair discrimination or spoken language modelling, is the natural next question, not shown here.

Practically, a tokeniser's phonetic organisation is largely settled before its cascade ends, and adding levels does not revise it: extra codebooks buy a narrower gap between categories, not a different order among them, so the levers are level 1 and the quantiser's geometry, not the codewords spent after it. A probe cannot see this, asking whether a distinction survives rather than whether the codec has quietly moved resolution from one part of the inventory to another.

THE USE OF LARGE LANGUAGE MODELS

Generative AI tools were used in this work for writing assistance only: drafting and revising prose whose content had been fixed by the authors beforehand. They were not used for research ideation or hypothesis formulation, for producing the extraction, analysis or figure code, or for the literature search; these are the authors' own work. Every AI-assisted passage was read and revised by the authors. No claim, number or citation originates from a generative model: all reported values are produced programmatically from result files, with the exceptions noted in the reproducibility statement, and every reference was checked against its source. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work analyses publicly released pretrained speech codecs on TIMIT (Garofolo et al., 1993), a corpus distributed under licence by the Linguistic Data Consortium and long established as a benchmark in the community, and LibriSpeech (Panayotov et al., 2015), released under CC BY 4.0, for the codec we train ourselves. We collected no data and conducted no experiments involving human subjects. Speaker labels enter our analyses only as a grouping variable (as the resampling unit for the bootstrap, and to draw the subsample the probes are fitted on) and we make no attempt to identify speakers or to recover speaker attributes; the speaker probe we mention is a control on our own measurement, read under a split that explicitly does not license claims about recoverable speaker information (Appendix E).

One concern is substantive rather than procedural. If a codec spends its resolution unevenly across phonetic categories, and if that allocation is settled at level 1 and left largely unrevised by depth (§4, §5), then systems built on these tokenisers inherit whatever the allocation favours, without any stage at which it is reconsidered. TIMIT is read American English, and our measurements therefore say nothing about how these codecs treat the languages, accents and speaking styles they were not measured on: the method transfers, the numbers do not. We would not want the ranking we report on one corpus read as a general statement about which distinctions a codec preserves.

REPRODUCIBILITY STATEMENT

The setup is given in §3 and detailed in the appendices: the four codecs, with the configuration values read from the loaded checkpoints rather than from the corresponding publications (Appendix A); the convention by which Mimi's two streams are indexed (Appendix A.1); the architecture, training objective, admissibility criterion and seed count of the codec we train ourselves, together with a disclosure of how that count was set (Appendix B); the split, the two inventories and the labelling rule (Appendix D); the depth each codec is scoped to and why (Appendix F); the probe protocol, including the split that bounds how its accuracies may be read (Appendix E); the construction of the synthetic figure (Appendix C); the reliability control on the allocation profile (Appendix I); and the measurement relating level-1 distortion to category frequency (Appendix H). Appendix G records the measurement choices that a reimplementation would otherwise have to rediscover.

Every measured value in this paper, in the text, tables and captions alike, is generated programmatically from the stored result files rather than transcribed by hand, so the prose cannot drift from the results it reports; the few exceptions are fixed protocol constants and a handful of figures from exploratory runs whose scripts write no result file, all of them in Appendix G. The extraction, analysis and figure code, together with the training code for the matched-condition codec and the scripts that regenerate every reported value, are included in the anonymous supplementary material.

REFERENCES

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019. doi: 10.1162/tacl_a_00254.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Au-

dioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. arXiv:2210.13438.

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. doi: 10.48550/arXiv.2410.00037.

Martin E. Dyer and Alan M. Frieze. On the complexity of computing the volume of a polyhedron. *SIAM Journal on Computing*, 17(5):967–974, 1988. doi: 10.1137/0217060.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren. DARPA TIMIT acoustic-phonetic continuous speech corpus. Technical Report NISTIR 4930, National Institute of Standards and Technology, 1993.

Allen Gersho. Asymptotically optimal block quantization. *IEEE Transactions on Information Theory*, 25(4):373–380, 1979. doi: 10.1109/TIT.1979.1056067.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. Recent advances in discrete speech tokens: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4184–4204, 2026. doi: 10.1109/TPAMI.2025.3643619.

Biing-Hwang Juang and Augustine H. Gray. Multiple stage vector quantization for speech coding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 7, pp. 597–600, 1982. doi: 10.1109/ICASSP.1982.1171604.

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)*, volume 36, pp. 27980–27993, 2023. doi: 10.52202/075280-1214. arXiv:2306.06546.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. doi: 10.1162/tacl_a_00430.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR – half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019. doi: 10.1109/ICASSP.2019.8683855.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 860–872, 2022. doi: 10.18653/v1/2022.naacl-main.63.

Kai-Fu Lee and Hsiao-Wuen Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, 1989. doi: 10.1109/29.46546.

Johan Liljencrants and Björn Lindblom. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48(4):839–862, 1972. doi: 10.2307/411991.

Björn Lindblom. Phonetic universals in vowel systems. In John J. Ohala and Jeri J. Jaeger (eds.), *Experimental Phonology*, pp. 13–44. Academic Press, Orlando, FL, 1986.

John Makhoul, Salim Roucos, and Herbert Gish. Vector quantization in speech coding. *Proceedings of the IEEE*, 73(11):1551–1588, 1985. doi: 10.1109/PROC.1985.13340.

Pooneh Mousavi, Jarod Duret, Darius Petermann, Artem Ploujnikov, Luca Della Libera, Anastasia Kuznetsova, Cem Subakan, and Mirco Ravanelli. DASB - discrete audio and speech benchmark. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=vGWrp0NjaE.

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. doi: 10.1109/ICASSP.2015.7178964.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921, 2021. doi: 10.1109/ASRU51503.2021.9688093.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. In *Interspeech*, pp. 3615–3619, 2021. doi: 10.21437/Interspeech.2021-475.

Samir Sadok, Julien Hauret, and Éric Bavu. Bringing Interpretability to Neural Audio Codecs. In *Interspeech 2025*, pp. 5023–5027, 2025. doi: 10.21437/Interspeech.2025-115.

Jean-Luc Schwartz, Louis-Jean Boë, Nathalie Vallée, and Christian Abry. The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25(3):255–286, 1997. doi: 10.1006/jpho.1997.0043.

Haibin Wu, Ho-Lam Chung, Yi-Cheng Lin, Yuan-Kuei Wu, Xuanjun Chen, Yu-Chi Pai, Hsiu-Hsuan Wang, Kai-Wei Chang, Alex 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, 2024. doi: 10.18653/v1/2024.findings-acl.616.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, et al. 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, 2025. doi: 10.1609/aaai.v39i24.34761.

Paul L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, 1982. doi: 10.1109/TIT.1982.1056490.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2308.16692.

Rui-Chen Zheng, Hui-Peng Du, Xiao-Hang Jiang, Yang Ai, and Zhen-Hua Ling. ERVQ: Enhanced residual vector quantization with intra-and-inter-codebook optimization for neural audio codecs. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2539–2550, 2025. doi: 10.1109/TASLPRO.2025.3579310.

A THE FOUR CODECS

Table 3 gathers the properties of the four codecs that the analysis depends on: the frame rate, which fixes how much signal a frame covers (§2); the dimension of the latent and of the space in which the lookup happens, which says whether and how a codec projects (§3) and where a reconstruction floor can arise (Appendix G); the codebook size and number of levels, which set the resolution a codec has to spend (§1); the number of codewords per dimension, which places the codecs relative to the high-resolution regime (§1); the codewords actually used at level 1 (§3); and whether and how each codec is distilled (§6). Every value was read by loading the checkpoint and querying its attributes and weights, except the row marked PUB, which frozen weights cannot reveal and which comes from the codec's paper.

None of the four codecs' papers lists TIMIT among its training data, and the codec we train ourselves uses LibriSpeech (Appendix B). TIMIT is therefore held-out data for every model in this paper, which is why our profiles are read on its full train and test split rather than on the test split alone.

Table 3: Properties of the four codecs, read from the loaded checkpoints except where marked PUB. Mimi's figures are for its acoustic stream (below); K_{used} is measured on full TIMIT with silence excluded; §3 reads DAC only to $L \approx 9$ (Appendix F).

	EnCodec	SpeechTokenizer	Mimi	DAC
Checkpoint	`encodec_24khz-d7cc33bc.th`	`fnlp/SpeechTokenizer`	`kyutai/mimi 89091b3e`	`weights_24khz_8kbps_0.0.4.pth`
Sample rate (Hz)	24,000	16,000	24,000	24,000
Frame rate (Hz)	75	50	12.5	75
Raw latent dim	128	1,024	512	1,024
Lookup dim d	128	1,024	256	8
Codebook K	1,024	1,024	2,048	1,024
Levels available	32	8	31	32
Levels used	8	8	7	32
Factorised	no	no	no	yes, $\to$ 8-d, ℓ_2
Distillation PUB	none	HuBERT, level 1	WavLM, sem. stream	none
kbps at L=7	5.3	3.5	1.0	5.3
$K^{1/d}$	1.06	1.01	1.03	2.38
K_{used} at L=1	372	929	874	665

Two entries need care, because reading them at face value would misplace a codec in the analysis. EnCodec's `model.quantizer.n_q` is 32, the checkpoint's native capacity, but at its nominal bandwidth of 6.0 kbps encoding uses 8 levels, through `get_num_quantizers_for_bandwidth`: taking the first number as the length of its cascade would put four times as many levels into the analysis as the codec uses. DAC's codebook holds 1,024 codewords, but over 1,243,625 level-1 TIMIT frames the codeword at index 1,023 is never selected and the one at index 1,022 only once, so a codebook size inferred from the indices observed would come out at 1,023. The difference is one unused codeword, not an off-by-one. Finally, the 8 kbps in the name of DAC's checkpoint is an operating point of its quantiser dropout, about 10.7 levels, not its capacity: all 32 levels run at 24.0 kbps.

One field is **undetermined**: the WavLM layer Mimi distils from. `MimiConfig` does not store it, no `distill`, `wavlm` or `teacher` field appearing in `model.config.to_dict()`, and we have not resolved it against the Moshi paper. We leave it open rather than guess. Nothing in our analysis depends on it: §6 compares Mimi's two streams as they are, whatever layer the semantic one was trained against.

A.1 MIMI'S TWO STREAMS

Mimi's quantiser is not one cascade but two, run in parallel on the same 512-dimensional pre-quantisation tensor: `semantic_residual_vector_quantizer`, with exactly one level, and `acoustic_residual_vector_quantizer`, whose standard configuration uses 7 of its 31 levels. Each has its own `input_proj` and `output_proj`, verified to be distinct objects carrying numerically different weights rather than one module referenced twice, and the decoder receives the sum of the two streams' reconstructions. The acoustic cascade is therefore trained to supply its share of a sum: its target is the latent minus what the semantic stream reconstructs, not the latent itself, and every Mimi number in this paper is measured against that target (Appendix G).

Keeping the streams apart matters because mixing them in one cumulative cascade would introduce a change of projection part way through it, the confound that factorisation introduces in DAC. It does not occur in any number we report: each analysis loops over streams and computes the available depths within each, and the reconstruction routine fixes which quantiser to use from its `stream` argument before any level is decoded, so it cannot apply one stream's `output_proj` to the other's codewords.

How L is counted. Whenever this paper writes "Mimi at L", it means the acoustic stream, with L counted from one inside that stream. The semantic level is neither level zero of that cascade nor

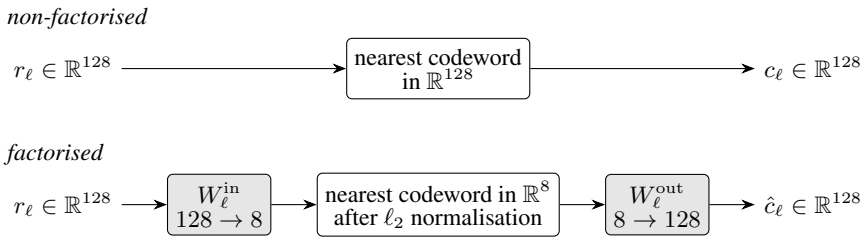


Figure 2: One level of the matched-condition codec's quantiser in its two conditions. Everything else is shared: encoder and decoder, codebook size (K=1024), depth (L=8), EMA codebook updates, objective and data. The factorised condition adds only the shaded per-level projections, and searches its codebook after ℓ_2-normalising both the projected residual and the codewords, as DAC does. In both, each level subtracts its output from the residual it received.

excluded from it: it belongs to a separate quantiser that the reported cascades never traverse. Mimi's L=1 is thus its first acoustic codebook, directly comparable to the other three codecs' first codebook, and the semantic level appears only where it is named, as in the within-model comparison of §6. The two streams also differ in how much of their codebook they use (K_{used} 1,707 for the semantic stream against 874 for the acoustic one), a property of their objectives and projections rather than of the encoder they share.

B THE MATCHED-CONDITION CODEC

The four released codecs differ in far more than factorisation, so the association found in §5 between factorisation and re-ranking could reflect anything else particular to DAC. This appendix describes the codec we train to separate the two: one architecture in two conditions that differ in factorisation alone (Figure 2), how it is trained, which runs are admitted and why, how the measure it is read with was qualified, and how its p values should be read. Each of these bears on whether the comparison is a counterfactual at all: a second difference between the conditions, an admission rule set after the fact or a measure chosen for the answer it gives would each undo it.

Architecture. A convolutional encoder of EnCodec's shape, dilated residual units followed by strided downsampling (base width 32, strides $(2, 4, 4, 5)$, hence a hop of 160 samples at 16 kHz), maps the waveform to a 128-dimensional latent; a mirrored decoder reconstructs it, and an 8-level residual quantiser with 1,024 codewords per level and EMA codebook updates sits between them. The two conditions share all of this. They differ in one mechanism: the factorised condition inserts a learnt per-level projection into 8 dimensions before the lookup and back after it, with the lookup performed in ℓ_2-normalised space, as in DAC. Encoder, decoder, codebook size, depth, decay and dead-code reset are shared code paths, so the contrast is confined to factorisation, itself a compound of projection and normalisation that this design does not separate (§7).

Training. Both conditions are trained on LibriSpeech (Panayotov et al., 2015), with a validation set of 500 utterances held out from the training split at a fixed seed; TIMIT is never seen, so the ablation, like the released codecs, is read on held-out data. Both use an adversarial objective, a multi-scale waveform discriminator with feature matching on top of a multi-scale mel L_1 loss and the commitment term, since a reconstruction-only objective left the codec unable to improve scale-invariant signal-to-distortion ratio (SI-SDR; Le Roux et al., 2019) across the cascade at all. Generator and discriminator use Adam at 3×10^{-4}, with no separate tuning budget. Training stops on a plateau of held-out mel L_1 (patience ten epochs, minimum relative improvement one per cent) rather than at a fixed step count; runs converged between 18 and 59 epochs, none reaching the 200,000-step safety ceiling. The model is not meant to be competitive: it exists to vary one factor while holding the rest constant.

Admissibility. A run is admissible only if its reconstruction actually improves over the cascade, which we require in three ways: SI-SDR at L=8 within DAC's own regime (≥ -13 dB), a significant mel-distance gain from L=1 to L=8, and an SI-SDR gain that is both significant and at least 3 dB. A run that fails to use its cascade cannot inform a question about how the cascade reallocates: it would contribute a flat profile that is an artefact of undertraining, not a measurement.

Of twenty runs, nineteen are admissible. The single exclusion is the non-factorised seed 0, whose SI-SDR gain is 2.95 dB against the pre-registered threshold of 3 dB, a miss narrow enough that we report the analysis both ways. On the admissible set (9 against 10 runs) the one-sided Mann–Whitney p is 0.00047 at L=2, 0.00047 at L=4, 0.0019 at L=6 and 0.0031 at L=8; on all twenty runs, 0.00029, 0.00038, 0.0018 and 0.0023. No conclusion turns on the exclusion.

On the measure they are read with, the two conditions start from profiles of comparable spread: $\sigma_\varphi(1)$ averages 0.156 for the non-factorised condition and 0.160 for the factorised one (Mann–Whitney $p = 0.903$). The factorised condition's level-1 reconstruction is nonetheless much worse, its latent error 2.81 times that of the non-factorised one ($p = 0.00028$): factorisation degrades how well level 1 reconstructs, not how unevenly.

Disclosure: the seed count was increased after a null result. The first analysis used five seeds per condition and found the effect non-significant at L=8; we then extended to ten. This is optional stopping, and we report it rather than presenting the ten-seed analysis as the design. At five seeds the p values were 0.0754 at L=8 and 0.0040 at L=2, against the ten-seed values given above. What changed is not the separation between conditions but its precision: the spread between seeds grows with depth, and the five-seed sample could not resolve it.

Instrument qualification. The latent-space profile cannot be read on the factorised condition, which fails the criterion of §5 from the first levels on, so rather than choosing a substitute after seeing the result we fixed a qualification in advance: a candidate measure is admitted only if it reproduces the latent-space $\rho(L)$ trajectory on the *non-factorised* condition, where the latent profile is valid. A per-phone mel distance qualifies: its $\rho(L)$ trajectory correlates with the latent-space one at 0.918 on average over ten seeds (range 0.835–0.986). The correlation here is Pearson, taken between two trajectories of Spearman ρ: the question is whether the two instruments follow the same trajectory and not only the same order, and a rank correlation over eight depths has too few ranks to tell them apart. A linear probe does not qualify: 0.766 on average over the five seeds it was run on (range 0.573–0.961), with a qualitatively different trajectory, falling sharply from L=1 to L=2 and then flat where the reference declines steadily. Its outputs are kept in the released results but excluded from interpretation.

On reading the p values. The tests of the effect on ρ are one-sided, factorisation being predicted to lower it; the two comparisons of starting points above are two-sided. With so few runs per condition the exact test has a floor: 92,378 arrangements give a smallest attainable p of 1.1×10^{-5}, reached only if every run of one condition ranks below every run of the other. No depth reaches it, so at every depth at least one pair of runs is out of order: the two conditions are well separated, not perfectly. Where a p value sits against its floor says this directly, which is why we report it rather than a comparison with a significance threshold. The values are unadjusted for the extension from five to ten seeds disclosed above, and we read them as descriptive.

C Levelling and re-ranking are independent: a constructed example

Section 2 defines two ways a refinement may fail to preserve an allocation profile: *levelling*, which narrows the gaps between categories, and *re-ranking*, which changes which categories are resolved best. The claim that they are independent, and that no single statistic separates them, is what licenses reporting σ_φ and ρ side by side throughout the paper. This appendix establishes it by construction rather than by argument: we build two refinements of the same level-1 partition, one that levels without re-ranking and one that re-ranks without levelling, and measure both with the statistics of §2. Figure 3 shows them. Everything here is synthetic, two-dimensional so that cells can be drawn, and deterministic from seed 7.

The data and the level-1 partition. The configuration has 8 two-dimensional Gaussian categories, four of them frequent and concentrated and four rare and diffuse, so that mass and spread vary together across the inventory as they do in a phone inventory. The level-1 codebook is obtained by drawing 14 initial centroids uniformly at random from a subset of the pooled samples and running twenty-five Lloyd iterations over all of them, with no k-means++ initialisation and a fixed iteration count rather than a convergence test; each category's distortion $\varepsilon_\varphi(1)$ is then the mean squared distance from

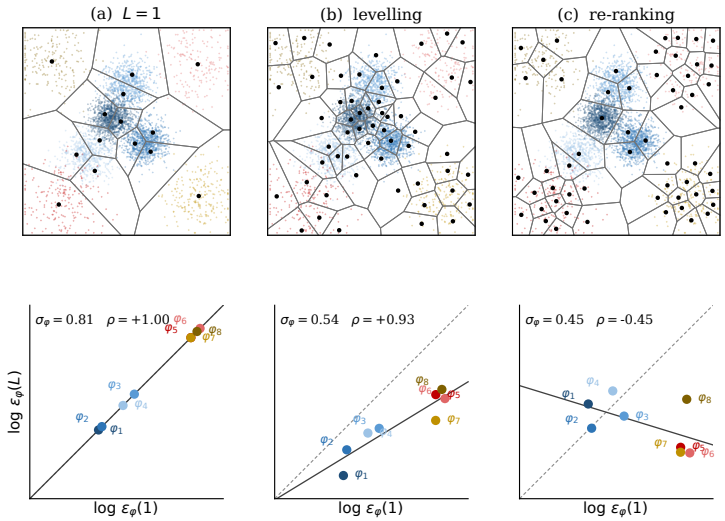


Figure 3: Levelling and re-ranking on synthetic two-dimensional data. *Top:* eight categories, four concentrated and four diffuse, under a fitted level-1 codebook (a), a levelling refinement (b) and a re-ranking one (c); cell boundaries are the Voronoi diagram of the codewords. *Bottom:* the corresponding allocation profiles, $\log \varepsilon_\varphi(L)$ against $\log \varepsilon_\varphi(1)$, one point per category; the dashed line is $y = x$, the solid line the least-squares fit, drawn where the level-1 profile varies. Points lying on a line flatter than $y = x$ indicate levelling; points out of order along the x axis indicate re-ranking.

its samples to their nearest centroid. Fitting the codebook is what makes the exercise meaningful: a uniform lattice would give every category nearly the same distortion, so the spread would start near zero and neither preserving nor reducing it would mean anything. Here it starts at 0.807. Both refinements take the codebook to 56 codewords, so they spend nearly the same budget and differ in where they spend it. Two conventions differ here from the body, for legibility of the figure rather than for any reason of substance: ε_φ is a squared distance and its dispersion across categories is taken with no correction for degrees of freedom, so the σ_φ values below are not on the scale of Table 1. The rank correlations are unaffected, being invariant under any monotone transformation.

The levelling refinement (b). The new codewords are drawn towards the regions the level-1 codebook resolves worst. We sample candidate positions from the data with replacement, with probability proportional to a power of the distance from each sample to its nearest level-1 codeword, and run twenty-five Lloyd iterations over those candidates. The exponent, 1.6, is an illustration parameter: we chose it to compress the spread between categories without reordering them, which is the regime this panel is built to exhibit. It gives $\sigma_\varphi[\log \varepsilon] = 0.538$, down from 0.807, with rank correlation 0.929 against the level-1 profile: the categories the codebook resolved best are still the ones it resolves best, and in the lower panel the points lie on a flatter line without changing order.

The re-ranking refinement (c). Here the budget is reallocated *between* categories rather than across the space as a whole. For each category we count the distinct level-1 codewords nearest to at least one of its samples, raise the inverse of that count to a power, normalise, and round to obtain a new budget, with at least one codeword per category; each category's share is then placed by twenty Lloyd iterations inside that category. The exponent, 3.4, is again an illustration parameter, chosen to make the reallocation sharp: at an exponent of one the budgets would be nearly flat and this panel would resemble the previous one. Rounding leaves this refinement one codeword short of the other, which we note rather than correct, the figure being demonstrative. The rank correlation with the level-1 profile is -0.452, at $\sigma_\varphi[\log \varepsilon] = 0.453$: the points of the lower panel are scattered across the diagonal rather than shifted along it. With 8 categories that correlation is not distinguishable from zero on its own; what the figure establishes is the contrast with panel (b), where the same statistic sits at 0.929.

The two refinements therefore reach almost the same spread (0.538 against 0.453) and almost the same mean distortion (0.280 against 0.271, from 0.860 at level 1) while doing quite different things

to the ordering. That is the figure's whole content, and the reason σ_φ and ρ are reported together: a measure of how much distortion varies across categories cannot see which categories those are.

Uniform refinement, the null of §4. Neither construction is what a refinement that ignores what lies in each cell would do. Subdividing every cell by the same pattern, scaled to that cell's own size, divides each cell's internal distortion by a common factor, and a common factor is an additive shift in the log domain, which a spread across categories cannot see: this is H_0-amplitude realised geometrically, and the levelling of §4 is a departure from it. Measured on a separate configuration built for this control, with 8 categories, $4{,}000$ samples each, 256 codewords and each cell subdivided 3^2 times, $\sigma_\varphi[\log \varepsilon]$ goes from 0.707 to 0.752, a change of 6.4% and in the direction opposite to levelling, while $\mathbb{E}_\varphi[\log \varepsilon]$ falls by 1.79 nats and the per-category distortion ratios stay within 6.6% of one another. The residual change is the finite-sample departure from the ideal common factor, and it is small and of the wrong sign, which is what makes every levelling measured in §4 a finding rather than an arithmetic consequence of adding codebooks.

D DATA, INVENTORIES AND INTERVALS

Corpus and split. Every measurement in this paper is taken on the whole of TIMIT, its training and test splits together: $4{,}620$ utterances from 462 speakers ($14{,}170$ s) and $1{,}680$ from 168 ($5{,}174$ s), rather than the 24-speaker core test set often used in ASR work. Nothing here is trained on TIMIT, so no split of it is held out for us in the usual sense (Appendix A); reading both simply gives the thinner categories as many frames as possible. The probes of Appendix E are the one exception, being fitted rather than measured, and they are cross-validated within a fixed subsample of speakers.

Inventories. Labels come from the raw 61-phone TIMIT inventory, read under the two inventories defined in §3. The first applies *no folding*: stop closures are not merged with their releases and syllabic variants are not merged with their non-syllabic counterparts, and removing the 3 silence labels (`h#`, `pau`, `epi`) leaves 58 categories. It is a transcription convention, and a narrow one. The second is the standard 39-label folding (Lee & Hon, 1989), which sends the stop closures to silence along with `q` and merges allophonic distinctions, leaving 38 speech categories. We report both because the first is finer than the contrasts of the language and the second is close to them: a result that holds under both does not depend on where between the two a reader would place the reference division, and where the two disagree, as they do for DAC's levelling (§4), the disagreement is itself informative. Unless we say otherwise, a number in this paper is read under the first.

Frames and categories. Frames are assigned to phones by *midpoint*: a frame takes the label of the TIMIT interval containing its centre, and frames whose centre falls outside any interval are dropped. There is no overlap-majority rule and no exclusion of boundary frames, so a frame straddling two phones is labelled like any other, by whichever interval owns its midpoint. A category is excluded from an aggregate when it has fewer than 5 frames in the partition that aggregate is computed on, so that a per-category mean is never taken on a handful of frames. The threshold is a guard rather than a filter: it removes no category from any analysis reported here, on the full corpus, in the $1{,}000$ resamples, on the speaker subsample the probes are fitted on, or on the frame subsample. The thinnest category of the raw inventory, `eng`, holds 0.021% of the corpus, and the frames it contributes outnumber the threshold by at least six to one in every codec, frame rates differing (Appendix A).

Statistical protocol. Confidence intervals throughout the paper come from a bootstrap over *speakers*, with $1{,}000$ resamples: frames within a speaker are not independent, and the speaker is the unit our statistics range over (Appendix G). The intervals are percentile intervals and are not symmetric about the estimate, so we give both ends rather than a margin. Silence is excluded from every aggregate unless we say otherwise. Since ε_φ is a per-category mean, every category weighs the same in σ_φ and in ρ whatever its frame count. Statistics computed across categories are reported under both inventories, and under the first a second time without its four sparsely populated categories.

D.1 INTERVALS BEHIND §4

Table 4 gives the speaker-bootstrap intervals underlying §4, under both inventories. What makes the levelling significant is the last column, the interval on the *difference* with both ends read off the same resample; the two marginal intervals are given beside it as description of each end, not as the test.

Under the first inventory they happen to be disjoint as well, but that is the conservative reading and it is not equivalent: under the second they are not, for DAC, while the paired interval still excludes zero (Appendix G). The intervals are narrow in absolute terms and the four codecs sit on very different scales, so overlap *across* codecs is not meaningful: σ_φ is a property of one codec's profile, and only its change with depth is comparable between them.

Table 4: Dispersion at the two ends of the matched range, and its relative change, with 95% speaker-bootstrap intervals over $1{,}000$ resamples, under both inventories. The interval on $\Delta\sigma_\varphi$ is *paired*: both ends are read off the same resample. It is the test; the two marginal intervals describe the ends.

	$\sigma_\varphi(1)$	$\sigma_\varphi(7)$	$\Delta\sigma_\varphi$ (%)
Raw inventory, 58 categories			
EnCodec	0.4591 [0.4564, 0.4623]	0.4067 [0.4042, 0.4094]	-11.4 [-11.6, -11.3]
SpeechTokenizer	0.0647 [0.0641, 0.0655]	0.0501 [0.0495, 0.0507]	-22.7 [-23.2, -22.1]
Mimi	0.0722 [0.0711, 0.0733]	0.0383 [0.0376, 0.0390]	-47.0 [-47.6, -46.3]
DAC	0.0211 [0.0208, 0.0215]	0.0138 [0.0135, 0.0141]	-34.6 [-36.1, -32.9]
Standard folding, 38 categories			
EnCodec	0.3016 [0.2981, 0.3053]	0.2686 [0.2656, 0.2717]	-10.9 [-11.2, -10.7]
SpeechTokenizer	0.0726 [0.0718, 0.0734]	0.0556 [0.0550, 0.0563]	-23.3 [-23.9, -22.8]
Mimi	0.0577 [0.0568, 0.0588]	0.0283 [0.0277, 0.0289]	-51.1 [-51.8, -50.2]
DAC	0.0140 [0.0137, 0.0144]	0.0136 [0.0133, 0.0139]	-3.0 [-5.8, -0.3]

D.2 The 54-category variant

Four categories of the raw inventory are thinly populated, for linguistic rather than accidental reasons. Three record connected-speech processes: a syllabic velar nasal left by vowel elision (`eng`, [ŋ̩], as in *Washington*), a devoiced schwa (`ax-h`, [ə̥], *suspect*) and a nasalised alveolar flap (`nx`, [ɾ̃], *winner*). The fourth, the voiced postalveolar fricative (`zh`, [ʒ], *azure*), is a phoneme of restricted distribution in English, rare outside intervocalic position. Any result driven by four such categories would be fragile, so wherever a claim could depend on them we recompute it on the remaining 54.

Nothing turns on them. The levelling of §4 keeps its sign and its ordering, and no codec's relative change to $L{=}7$ moves by as much as a point (-11.4%, -23.5%, -46.5% and -35.3% against the values in Table 4). For the re-ranking of §5, ρ at the matched depth becomes 0.989 for EnCodec, 0.962 for SpeechTokenizer, 0.852 for Mimi and 0.447 for DAC, against 0.988, 0.954, 0.868 and 0.434 on the full set: only Mimi falls, the ordering is unchanged at every depth, and the separation between DAC and the lowest of the other three narrows slightly without coming near to closing. Both conditions are computed in a single pass, on the same frames, the same codes and the same bootstrap resamples, so the two are comparable pairwise rather than being two independent runs.

The question does not arise under the standard folding, which absorbs all four labels into larger categories: its thinnest category holds 0.369% of the corpus, against 0.021% for the thinnest of the raw inventory.

E Probes do not register the re-ranking

Section 5 uses frame-level probes to test a reading of DAC's falling ρ that the allocation profile cannot rule out on its own. If DAC's deeper levels simply stopped carrying phonetic structure, ρ would fall without any reallocation, recording a representation going quiet rather than one being reorganised. Distortion does not say how much phone identity survives quantisation; a probe does. A speaker probe, fitted on the same features and folds, serves as a control on the phone probe. Neither is read as a measure of allocation, which a probe cannot see (§1).

What the probes see. Both probes read the cumulative reconstruction $\hat{z}_{1:L}$ in the encoder's latent space, the space in which ε_φ is measured, rather than the indices or the space in which a quantiser performs its lookup, so that probe and profile describe the same object at each depth. Each instance is a single frame, with no pooling over an utterance. Probes are fitted on 20 speakers, half of them women and spread across TIMIT's eight dialect regions, the same for every codec and depth, so that the values along a row of Table 5 are computed on identical data.

The classifier. Each probe is a multinomial logistic regression (Pedregosa et al., 2011) with an ℓ_2 penalty ($C = 1.0$, `lbfgs`, at most 1,000 iterations), fitted on features standardised with statistics from the training folds alone. C is left at its default and never tuned, so accuracies are a fixed-capacity reference point rather than the best a probe could reach. Since they are read only along depth within one codec, a capacity that stays fixed matters more here than one that is optimal.

Why the folds are utterance-disjoint. Accuracies are averaged over 5-fold cross-validation grouped by utterance, so the same speakers appear on both sides of every fold in different utterances. A speaker-disjoint split would rule out the speaker probe altogether, since a classifier cannot name speakers it has never seen, and would leave the phone probe trained on a handful of voices. The price is that the phone probe can exploit speaker-specific regularities it has already met, so its accuracy is an upper bound on what the representation offers about phone identity independently of speaker, and the speaker probe says nothing about whether speaker information is recoverable across speakers. Neither matters for the use made of them here: both are read only as trajectories along depth, never as absolute measures.

E.1 Results

Table 5: Probe accuracy against cascade depth, to be read along a row; mean over 5 folds, fold standard deviation at most 0.024.

	L=1	L=2	L=3	L=4	L=5	L=6	L=7	L=8
Phone probe, 58 classes								
EnCodec	0.209	0.239	0.253	0.263	0.269	0.273	0.275	0.278
SpeechTokenizer	0.621	0.607	0.607	0.618	0.621	0.624	0.630	0.635
Mimi	0.210	0.207	0.225	0.220	0.224	0.224	0.225	—
DAC	0.193	0.220	0.230	0.238	0.241	0.242	0.244	0.243
Speaker probe, 20 classes								
EnCodec	0.071	0.103	0.130	0.147	0.158	0.170	0.177	0.186
SpeechTokenizer	0.054	0.102	0.149	0.189	0.223	0.249	0.278	0.299
Mimi	0.103	0.161	0.183	0.208	0.221	0.238	0.252	—
DAC	0.051	0.083	0.097	0.111	0.117	0.121	0.125	0.130

Table 5 gives accuracy at every depth. It is to be read along a row: the four codecs differ in frame rate and in the space the probe is given, so a difference between rows says nothing about what a cascade contributes. The speaker probe is the 20-way one; the 630-way variant, whose error is flat at every depth, is discussed in Appendix G. Mimi's acoustic stream stops at L=7 in its standard configuration.

Phone-probe accuracy changes little with depth for any codec, while speaker-probe accuracy rises steadily in all four: between L=5 and L=7 the phone probe gains 0.003 for DAC and between 0.002 and 0.009 for the others, on values that differ far more between codecs than across depth. SpeechTokenizer sits well above the rest on the phone probe at every depth, as expected of a first level trained to predict HuBERT (§3). DAC is lowest on the speaker probe at every depth; on the phone probe it shares the bottom with Mimi, the two exchanging places once, by margins of the order of the gains just quoted, and we read nothing into the exchange.

This answers the question the probes were asked. Between L=3 and L=7, where DAC's ρ falls from 0.920 to 0.434, its phone probe rises from 0.230 to 0.244, so its deeper levels are not going quiet: what changes is which categories the cascade reconstructs best, not how much phone identity survives. The same table illustrates, on released codecs, the contrast §1 draws in principle. A phone probe that registers little across depth sits beside an allocation profile that is reordered, because the two answer different questions: whether a distinction survives, and where the codec has drawn its boundaries.

The probes also bear on §6, by a route independent of the allocation profile. Read at its single level, Mimi's distilled stream gives the phone probe 2.4 times the accuracy of the acoustic stream's first level (0.514 against 0.210) and the speaker probe half of it (0.051 against 0.103), at the same rate. The two streams differ in their projections as well as their objective, so this is no more a clean isolation of distillation than the comparison of §6, but it points the same way.

F DEPTH SCOPE FOR DAC

DAC results stop at $L \leq 9$. Its cumulative latent-space reconstruction error reaches a minimum at $L{=}9$ and then rises monotonically to $L{=}32$, which needed ruling out as an assembly error before it could be interpreted.

It is not one. The same error computed from DAC's own forward pass, with `n_quantizers` truncated, rather than from our reconstruction, agrees with ours to floating-point precision at every level and rises after $L{=}9$ in the same way. The behaviour belongs to the checkpoint, not to our code.

Two things constrain it. Each of DAC's levels projects the encoder's $1{,}024$-dimensional latent into 8 dimensions before its lookup, so after L levels the reconstruction spans at most $8L$ of those dimensions, a small part of the latent at the depths we read. And beyond $L \approx 9$ the error stops falling even as that span grows, while the decoded audio keeps improving: between $k{=}9$ and $k{=}32$ the mel L_1 distance falls from 0.2913 to 0.1455 across all ten utterances tested, while SI-SDR stays essentially flat. The deeper levels are still doing something for the decoder, just not in the direction our measure reads. We bound DAC's depth there rather than report quantities taken past the point where its cumulative latent error stops falling.

Mimi's error rises from the first level for a related reason, a projection shared by its whole acoustic cascade (Appendix G).

G MEASUREMENT PITFALLS IN DETAIL

Several quantities this paper relies on look interchangeable with quantities that are easier to compute, and are not. Each item below invalidated a conclusion we had drawn, and each will be met by anyone comparing quantised speech representations across architectures.

A quantiser's internal residual is not comparable across codecs. Every RVQ implementation exposes the residual it computes at each level, and reading it is the obvious way to obtain $\varepsilon_\varphi(L)$. It is wrong for a factorised codec: the residual there lives in a per-level projected and renormalised space, so it measures a local, bounded quantisation error rather than the cumulative error in the encoder's latent space that the same variable holds for the other three codecs. The two are not merely on different scales — the normalisation makes geometric decrease impossible — and an early version of our central result was computed on this mixture. We therefore recompute every distortion from the stored codes, decoding back into the encoder's latent space, which is the only space the four codecs share.

A cascade quantising in a subspace cannot be measured in the whole space. Two of the four codecs project the encoder's latent before the lookup, both into a strict subspace: DAC into 8 dimensions at every level, each with its own projection, Mimi into 256 once for its whole acoustic cascade. Mimi's case is the clean one. Its reconstruction is confined to the image of a single `output_proj`, so the component of the target outside that image is a floor no depth can reach; and because `output_proj` is not the inverse of `input_proj`, a cascade converging in the projected space converges towards a point that is not the nearest one to the target in the latent. Measured in the latent, its error therefore rises from the first residual level, while in the 256-dimensional space where its lookup happens its residual falls monotonically, from 4.330 at the input to 0.683 at the matched depth. DAC's case is looser: each new level adds a projection of its own, so the span of its reconstruction grows with depth, and its internal residual, renormalised at every level, is a per-level quantisation error rather than a cumulative one (the first item of this appendix), so its convergence is legible only as the plateau of its cumulative error (Appendix F). What this bounds is how far the *level* of ε_φ can be read as a property of the partition. Whether it also shapes Mimi's profile we measured directly: the unreachable component varies far less across frames than ε_φ does, and its mean by category is only weakly related to the level-1 profile (rank correlation -0.256), negatively, so it flattens the profile a little rather than making it. With it removed, the levelling of §4 is essentially unchanged and ρ at the matched depth rises from 0.868 to 0.893: if anything, the floor slightly overstates Mimi's drift.

Mimi compounds this with a second error we made first. Its acoustic cascade contributes one term of a sum, the other being its semantic stream, so it is not trained to reconstruct the encoder's latent but the part of it the semantic stream leaves. Measuring it against the whole latent evaluates a cascade

against a target it never aims at. Every Mimi number in this paper is computed against the corrected target. The correction changes Mimi's profiles themselves, halving its level-1 dispersion and moving ρ outside its interval, but leaves the comparisons we draw from them within their intervals: the relative levelling $\Delta\sigma_\varphi$, the covariate decomposition and the within-model comparison of §6.

A latent proxy stops tracking what the model optimises. Beyond $L \approx 9$, DAC's cumulative latent reconstruction error rises while its decoded audio keeps improving spectrally (Appendix F). The behaviour is the model's own, not an artefact of our reconstruction, since the same error computed from DAC's own forward pass rises identically; but it means the quantity every measure in this paper reads has left the regime where it stands for anything. We bound DAC's depth rather than report numbers the proxy no longer supports. The general point is that a proxy validated at one operating point need not remain valid at another, and depth is exactly the axis along which one leaves it.

Geometric quantities need a null model in high dimension. We initially took the near-orthogonality of successive projection subspaces in a factorised codec (85.8 degrees between consecutive levels) as evidence of a rotation mechanism. Two random 8-dimensional subspaces of $\mathbb{R}^{1024}$ are separated by 85.7 ± 0.4 degrees: the observation is indistinguishable from concentration of measure, and the mechanism was withdrawn. High-dimensional geometry supplies many quantities whose typical value is far from intuition, and none of them can be read without the corresponding null.

Counting distinct codes is a rarefaction statistic. The number of distinct codes a category activates, divided by its frame count, correlates at between -0.890 and -0.857 with category frequency in all four codecs, an apparently strong result. A permutation null gives between -1.000 and -0.998: the correlation is what one obtains when the number of observed types grows sublinearly with sample size, which it always does, and at matched sample size the sign reverses. Any statistic based on counting types needs its rarefaction null before it means anything.

A probe with too many classes reports nothing. Our speaker control was a 630-way probe whose error sat at 0.99 at every depth in every codec, which we first read as evidence that no speaker information was being added. On a fixed 20-speaker subset the same probe declines steadily with depth. A flat curve at a value close to chance is a statement about the task, not about the representation.

A raw distance ranks categories by energy. The margin between a frame and its runner-up codeword is a natural measure of how precariously a category sits with respect to a cell boundary. Taken raw it is a distance, so it is mechanically larger where the residual is larger, and averaging it by category ranks categories by energy: low-energy stop closures come out as the most precarious. Normalising per frame before ranking gives a different ordering, anti-correlated with the first. Neither is wrong, but they answer different questions, and the choice must be made before the categories are read off. We met this on the margin first and set the margin aside; it turned out to hold of ε_φ itself, which is a distance too, and that is the next item.

ε_φ is not purely a property of the partition. It is read as one, the resolution a codec spends on a category, but it is a distance, and a category's loudness, duration, frequency or variability can move it whatever the partition does. Regressing $\log \varepsilon_\varphi(1)$ on three properties of the category that owe nothing to the codec (waveform energy, phone duration and $\log n_\varphi$) accounts for 0.971 of it for EnCodec, 0.829 for Mimi, 0.556 for DAC and 0.347 for SpeechTokenizer; adding the within-category dispersion of the encoder latent gains a further 0.360 for DAC and at most 0.100 for the others. Reading a single profile as a statement about cell sizes therefore overstates what the measurement supports, and §6 is the section most exposed to this.

It is tempting to conclude that comparisons *across depth* are immune, since these properties do not change with L and would enter $\log \varepsilon(1)$ and $\log \varepsilon(L)$ as one additive term that cancels in any comparison of the two. That argument fails on the data: the same covariates account for between 0.298 and 0.933 of the per-category trajectory $\log \varepsilon_\varphi(7) - \log \varepsilon_\varphi(1)$, so they do not enter as a constant and do not cancel. What protects §5 is the residual control it reports rather than that cancellation: with the three covariates partialled out, DAC sits at 0.434 against 0.854 to 0.952 for the other three.

The control behaves as the decomposition says it should, which is worth more than the survival of the separation on its own. EnCodec and Mimi, whose level-1 profiles the covariates account for most, are the two it moves, in opposite directions, shifting EnCodec's ρ by -0.134 and Mimi's by $+0.084$:

part of Mimi's drift across depth came from the covariates rather than from reallocation. DAC, whose profile they account for barely more than half of, moves by less than 3.1×10^{-5}, the finest difference a rank correlation over 58 categories can express, since ρ is a function of an integer sum of squared rank differences. That is not an invariance: all but two of DAC's 58 categories change rank under the control, and the sum happens to land on the same integer. None of this disposes of the confound: it bounds what a ε_φ profile licenses, and every reading of one in this paper inherits the bound.

Bootstrap over the unit the statistic ranges over. Confidence intervals throughout this paper resample speakers, since frames within a speaker are not independent. For a regression across categories this is the wrong unit: every resample refits the line on the same 58 points, so the dominant source of uncertainty never enters, and the intervals came out roughly sixteen times too narrow. The unit to resample is the one the statistic ranges over — speakers for a frame-level quantity, categories for a fit across categories, and both when the statistic depends on both.

Test the difference, not the two ends. Whether a quantity changes between two conditions is a question about the difference, and the interval that answers it is the one taken on the difference with both ends read off the same resample. Asking instead whether the two marginal intervals are disjoint is a different and more conservative test: the two ends move together under a resample, so their intervals can overlap while the paired difference is plainly non-zero. We used the marginal form at first, and on TIMIT's 58 labels it is harmless — the two tests agree for all four codecs, so nothing we reported on it was wrong. They part company on the 38-category inventory, and exactly where it matters: DAC's $\sigma_\varphi(1) = 0.0140$ [0.0137, 0.0144] and $\sigma_\varphi(7) = 0.0136$ [0.0133, 0.0139] have overlapping intervals, while the paired change, -3.0% [-5.8, -0.3], excludes zero. Reporting the marginal test would have meant reporting a verdict that flips with the inventory where the right one does not. Both are now given in Table 4, the paired one as the test and the marginals as description.

H LEVEL-1 DISTORTION AND CATEGORY FREQUENCY

Section 6 treats category frequency as one of the properties of the input that shape level-1 allocation, but not as a principle that organises it. This appendix gives the measurement behind that.

If a codec allocated resolution by acoustic mass, frequent categories would be reconstructed somewhat better than rare ones. Under the high-resolution theory of §1, codewords placed in proportion to a power of the data density leave a local error that falls as the density rises, though ever more weakly as the dimension of the quantised space grows (Zador, 1982; Gersho, 1979). The prediction is a negative relation between frequency and distortion, faint in high dimension, and the codecs are far from the regime in which it holds exactly. We therefore regress $\log \varepsilon_\varphi(1)$ on $\log n_\varphi$ across the 58 categories and, separately, take the rank correlation between them.

The least-squares slopes are small for all four codecs. A bootstrap over categories, the unit the fit ranges over (Appendix G), excludes zero only for DAC, whose slope is negative but at the limit, the analytic test not reaching significance. The prediction of the high-resolution theory for such a slope is $-2/(d+2)$, d being the dimension of the lookup space: near zero for the three codecs that search in 128 dimensions or more, with which they are consistent, but -0.2 for DAC's 8, which DAC falls short of by more than an order of magnitude. In EnCodec the six stop closures sit well below the fitted line; in the other three they do not stand out. The rank correlations are better behaved and are what §6 refers to.

Controlling for within-category dispersion. Frequent categories are also more variable, which would inflate their distortion for reasons unrelated to allocation, so the relation has to be read with that dispersion held constant. The control must be measured in the space ε_φ itself lives in, the encoder's latent, and we take the mean per-dimension variance of a category's frames there, the ambient dimension differing across codecs (Appendix A) so that a raw trace would not be comparable between them. This applies to DAC as it does to the rest: DAC's ℓ_2 normalisation constrains the *lookup* space, in which every encoding has unit norm, and not the encoder latent, where its categories have per-dimension variances ranging from 5.00 to 8.66.

What the panel shows. Read raw, only EnCodec shows a significant association, and it is positive: its frequent categories are reconstructed somewhat *worse* than its rare ones, the opposite of what allocation by mass predicts. Held against dispersion, that relation dissolves ($+0.135$, $p = 0.311$):

Table 6: Rank correlation between category frequency and level-1 distortion, raw and controlling for within-category dispersion in the encoder's latent space, across the 58 categories. A negative value means frequent categories are reconstructed better.

	raw		controlled	
	ρ	p	ρ	p
EnCodec	$+0.328$	0.012	$+0.135$	0.311
SpeechTokenizer	$+0.032$	0.811	-0.338	0.0095
Mimi	-0.229	0.084	-0.228	0.085
DAC	-0.182	0.171	-0.434	0.00066

what looked like a frequency effect was dispersion. The other three turn negative, frequent categories being reconstructed better, significantly for DAC (-0.434, $p = 0.00066$) and SpeechTokenizer (-0.338, $p = 0.0095$), short of significance for Mimi (-0.228, $p = 0.085$). For SpeechTokenizer, which shows no marginal relation at all, this is a suppression effect: frequency and dispersion pull in opposite directions and cancel until one is held constant.

Once dispersion is controlled, then, the sign is the one allocation by mass predicts in three codecs out of four. The magnitude is another matter: the relation is absent in EnCodec, marginal in Mimi, a moderate rank correlation at most where it is present, and in DAC, where the theory predicts the largest effect, the slope falls far short of it. Frequency is therefore one of the input properties that shape level-1 allocation, which is why it enters the decomposition of Appendix G, not a principle that organises it. It is also a caution about ε_φ itself: a substantial part of a category's level-1 distortion is accounted for by properties of the category, before any question of where the codec has placed its cells.

I Reliability of the allocation profile

Section 5 reads $\rho(L)$ as evidence that a cascade has reordered its allocation. Since ρ is a correlation with $\varepsilon(1)$, a noisily estimated level-1 profile depresses it whether or not anything has been reordered. This appendix gives the control.

Estimating reliability. We split the speakers of the corpus at random into two halves, recompute the allocation profile on each, and correlate the two. That correlation estimates the reproducibility of a profile computed on *half* the data, and is therefore pessimistic; the Spearman–Brown formula corrects it to the full sample, giving $r_{\ell\ell}$ for the profile at level ℓ. The disattenuated correlation is then $\rho/\sqrt{r_{11}r_{LL}}$, which is defined for Pearson correlations only, so this control and this control alone reports Pearson rather than Spearman; the two orderings agree throughout.

Table 7: Split-half reliability of the allocation profile and its effect on the level-1 correlation at the matched depth. r_{11} and r_{77} are Spearman–Brown-corrected split-half reliabilities; ρ_{Pearson} is the raw correlation and $\rho/\sqrt{r_{11}r_{77}}$ the same correlation corrected for unreliability.

	r_{11}	r_{77}	$\rho_{\text{Pearson}}(7)$	$\rho/\sqrt{r_{11}r_{77}}$
EnCodec	0.998	0.998	0.997	0.999
SpeechTokenizer	0.995	0.992	0.962	0.968
Mimi	0.995	0.992	0.934	0.940
DAC	0.961	0.953	0.564	0.589

What the control shows. DAC's level-1 profile is the least reproducible of the panel, which is what the objection predicts, but the gap is small: its reliability sits at 0.961 against 0.995–0.998 for the other three, a difference of a few per cent rather than the collapse that would be needed to manufacture a fall from unity to 0.434. Correcting for it moves DAC's correlation at the matched depth from 0.564 to 0.589, still far below the corrected values of the other three.

One further observation tells against the objection. If low reliability followed from an uninformative first level, the panel would show the two moving together. It does not: EnCodec has nearly the lowest level-1 PNMI (0.255, against 0.222 for DAC) and the highest reliability of the four. What a profile records and how reproducibly it records it are separate properties.

Reliability does not address the second objection of §5, which concerns how much ordering a flat profile contains rather than how precisely it is measured: a profile can be both nearly flat and highly reproducible, and DAC's is.