SPEECH RECOGNIZERS CARRY THE SURROUNDING SPEECH RATE BUT USE IT LESS AS THEY GROW

Anonymous authors
Paper under double-blind review

ABSTRACT

A probe can decode a variable from a network's activations without the network's behavior depending on it. Separating a variable the network carries from one it uses calls for a manipulation few variables permit: the variable moves while the input the decision is about stays fixed. Human speech perception supplies one. When the speech around a reduced function word such as "or" slows, listeners' reports of it fall from 79 to 33 percent though its audio is unchanged. A recognizer that did the same would drop a correctly transcribed word when only the speech around it slowed, a change that word error counts but cannot attribute to the context. We ran this manipulation on sixteen open checkpoints, twelve Whisper models across five sizes and four CTC recognizers. Whisper tiny through medium lower their belief in the frozen word, the log-probability lead of the transcript with the word over the one without it, as the surrounding speech slows. In the three smaller models the fall is roughly a third to half of the listeners' own, and the word drops out of the transcript. Three further function words and the CTC recognizers show the shift as well. Within Whisper the shift declines with size, persisting at 1.5B in two of three checkpoints, while the context's rate stays decodable from the same frozen frames at every size. Erasing that representation does not selectively remove the shift in Whisper and enlarges it in two of the three primary CTC recognizers. Together, this work separates a speech recognizer's representation from its use, which no benchmark that alters only the target, and no probe, can see. Our approach gives interpretability a stimulus-side test of use, complementary to probes and erasure, that costs one item set and no retraining.

1 INTRODUCTION

What a model's representations contain has become measurable. Probes and benchmarks decode phones, words and prosody out of speech-model activations (Pasad et al., 2021; de Seyssel et al., 2023), and in audio language models speed itself is decodable even where the behavior ignores it (Pang et al., 2026). Whether the behavior uses what a probe decodes is harder to establish (Ravichander et al., 2021), so probes are paired with control tasks and with interventions that remove the variable (Hewitt & Liang, 2019; Elazar et al., 2021; Ravfogel et al., 2021; Lasri et al., 2022; Belinkov, 2022). The design here instead moves the variable in the stimulus. The part of the stimulus the decision is about, the audio of a word and its two neighbors, is held bit-identical while only the speech around it is re-timed. Any change in the model's decision about the word then comes from that context, and matched controls, for the added duration and for the re-timing itself, narrow it to the context's rate, the dependence we call use (Secs. 2 and 3.1).

For a speech recognizer the difference is practical: its layers hold far more than the words, and how much of it reaches the transcript is largely unexplored (Glazer et al., 2026). Whisper's encoder keeps a background sound's identity at every layer, and its word error under that sound rises less than other recognizers'; whether the transcript conditions on the carried variable was inferred, not tested (Gong et al., 2023). Which variables a checkpoint's transcript depends on is the developer's question. A benchmark that alters only the target never creates such a dependence, and aggregate word error cannot attribute one. Probing reports what a layer carries, without a direct connection to what the transcript does (Pouw et al., 2024).

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.

Human speech perception supplies a variable that moves while the input the decision is about stays fixed: listeners hear each stretch of speech relative to the speech around it (Dilley & McAuley, 2008). Slowing the speech surrounding a reduced function word drops reports of that word from 79 to 33 percent while the word itself is acoustically unchanged. Speeding the speech around a region produced without the word raises false reports of it eightfold from 3 percent (Dilley & Pitt, 2010). The effect grows with the amount of slowing (Baese-Berk et al., 2014) and replicates across languages (Dilley et al., 2013). Because the acoustics under the word never change, this distal rate effect (distal because only the speech away from the word changes) shows that perception consults a rate variable in the context. That variable is central to speech perception, from the 4 to 5 Hz amplitude modulation that marks speech to the temporal expectation that moves phoneme boundaries and word reports (Poeppel & Assaneo, 2020; Ding et al., 2017; Chang et al., 2024; Bosker, 2017; Bosker & Ghitza, 2018; Kösem et al., 2018; Giraud & Poeppel, 2012; ten Oever & Martin, 2021).

Speaking rate is the salient case, for a recognizer a long-standing source of error (Siegler & Stern, 1995; Mirghafori et al., 1996; Benzeghiba et al., 2007). Engineering treats it the other way round from perception, measuring robustness to the rate of the target itself (Meyer et al., 2011) and training for invariance through speed perturbation (Ko et al., 2015). Invariance and normalization are opposite stances toward the same variable: discarding the rate of the speech transcribed, or using the rate of the surrounding speech to reinterpret the target. A benchmark that alters only the target cannot tell the two apart, human-model batteries included (Adolfi et al., 2023). If a recognizer normalized as listeners do, a word transcribed at the talker's own rate would drop out when only the speech around it slowed. Whether that is wanted depends on the deployment: a defect for verbatim transcription, a feature for recovering reduced speech. Which stance a checkpoint takes, and whether size or family moves it, only moving the context alone can say.

Context-effect experiments with the target held fixed (Kirby et al., 2026; Pouw et al., 2024; de Heer Kloots & Zuidema, 2024) have begun to reach speech models, but only for context within the neighboring sound or syllable. Measured context integration in base-size encoders runs to about two seconds (Meng et al., 2025), and whether a word decision consults the rate of speech seconds away has not been measured. A model can therefore contain the context's rate as a decodable statistic, use it to re-map a frozen word as listeners do, or neither, and no existing test separates the three.

Here we ran the distal-rate paradigm on open speech recognition models as psychophysics experiments, the model's belief in the frozen word compared with listeners' reports. A linear probe then asks whether the context's rate is present in each model's activations at the frozen frames, and erasing that representation asks whether the behavior runs on it.

Larger models can be less human-like on a documented minority of tasks (Wei et al., 2023; McKenzie et al., 2023; Oh & Schuler, 2023; Kuribayashi et al., 2021; Fel et al., 2022), and whether a human-calibrated behavior changes with size while the variable it runs on stays decodable has not been asked. We therefore ran the paradigm across model sizes and, for generality, across four function words and a second model family, with two metrics so that no size claim rests on either alone (Schaeffer et al., 2023). For evaluation the test needs one item set, no retraining and a checkpoint that passes the measurement validity criteria (Sec. 2.5), as every checkpoint here did.

2 The paradigm and the analysis rules

2.1 Stimuli and manipulations

Speech recordings come from LibriSpeech test-clean (Panayotov et al., 2015), held out from the fine-tuning data of every connectionist temporal classification (CTC) model. Whisper's training corpus is opaque, but memorization cannot produce a context-scale slope, since every condition scores the same two transcripts against re-timed audio of the same item. We selected utterances with "or" between two neighboring words of at least three characters and at least 2.0 s of speech before the proximal window, at most two per speaker: 44 items, the primary item set, the corpus's full yield under the selection rules of Appendix A. The frame "X or Y" mirrors the human materials ("leisure or time" heard as "leisure time"; Dilley & Pitt, 2010).

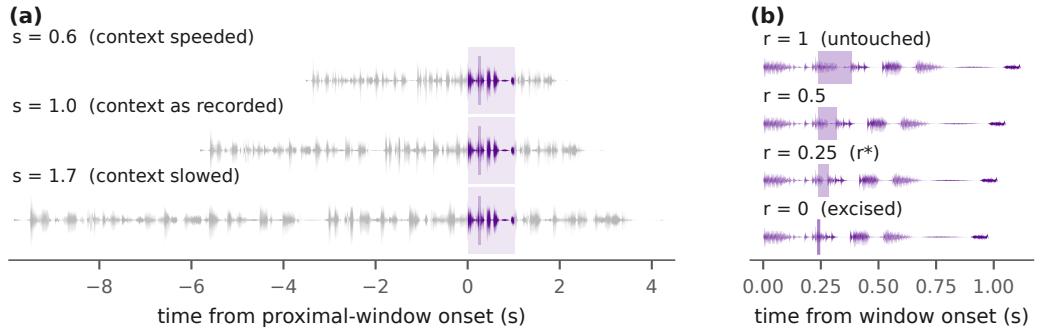


Figure 1: **The frozen word and the re-timed context.** **(a)** One primary item, near the item-set medians, at context scale 0.6, 1.0 and 1.7, aligned at the proximal window (shaded; the darker strip is the reduced "or" at $r = 0.25$): the surrounding speech is time-scaled while the window's samples are bit-identical across rows. **(b)** The same window under the reduction axis, $r = 1$ to 0 (excised).

The **proximal window** runs from the onset of the word before "or" to the offset of the word after it (Fig. 1). Two axes are manipulated by WSOLA time scaling (Verhelst & Roelands, 1993). The **reduction** r time-compresses the "or" span itself to a fraction r of its duration (grid 1, 0.5, 0.35, 0.25, 0, each grid point a **level**; $r = 1$ keeps the whole word, $r = 0.25$ keeps a quarter of it, $r = 0$ cuts it out). At intermediate r it creates an ambiguity between a clear "or" and none, and, with the amount removed set by construction, it gives the belief metric its ground truth (Sec. 2.5). The **context scale** s applies the same time scaling to everything outside the proximal window (grid 0.6, 0.8, 1.0, 1.3, 1.7), so $s = 1.7$ is the slowest context and $s = 0.6$ the fastest. The "or" span and both context chunks pass through WSOLA in every condition that contains them, so the $s = 1$ condition is a matched resynthesis floor, and at fixed r the proximal waveform, the **frozen span**, is bit-identical across all five context scales (Appendix A).

2.2 The metric of the model's belief

The metric is not word error rate but the model's belief in two candidate transcripts, $\Delta = \log P(\text{transcript with the word}) - \log P(\text{transcript without the word})$, where P is the model's own probability of the full transcript given the audio. The candidates are the item's reference transcript with and without "or", so the pair differs by one word, the target, and neither is decoded. For the attention-decoder models, which write the transcript token by token, the candidate's tokens are fed to the decoder in place of its own output (teacher-forced) under one fixed prompt (English, transcribe, no timestamps); $\log P$ sums their log-probabilities and the end-of-transcript marker's. The CTC models have no decoder, so $\log P$ is the negative CTC loss, summed over every frame-level alignment of the transcript to the audio. Δ is positive when the model favors the transcript with the word, and we call it the model's belief in the word. It is measured in the model's own units (nats, natural-log units of probability), so no magnitude is compared across families.

Two further candidates, built the same way and scored against the same audio, serve the controls, each editing the reference transcript at one position outside the proximal window: deleting a word two to three words past the window gives the **unrelated-word control** Δ_{ctrl}, and inserting a second "or" just after that same distant word, with no audio supporting it, gives the **insertion control** Δ_{ins}. They change the text and not the audio, so they separate a percept of the frozen word from a response of the scoring machinery to re-timed audio.

2.3 Controls and item sets

Total duration, window position and the resynthesis artifact dose all change with s, and each gets a matched control. The **duration-matched control** keeps the context at $s = 1$ and appends speech-shaped noise at the file edges, sized so that total duration equals the matching slow-context condition; the noise is not heard as speech, the condition under which listeners show no rate effect (Pitt et al., 2016), and exists on the slow side only. The **duration-clamped twin** pads every intact condition to a fixed total duration per item and reduction level, and the **artifact control** scales the context out to s' and back to 1.0, roughly doubling the artifact dose at zero net change. A 23-item replication set

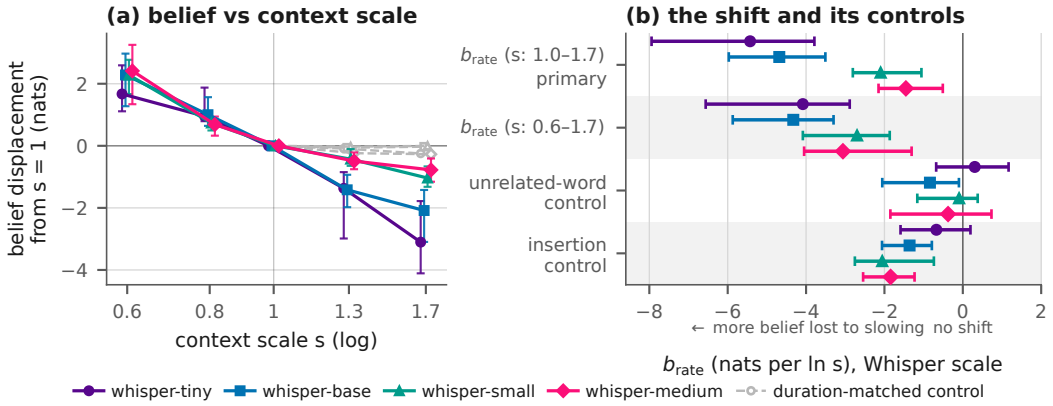


Figure 2: **Whisper's belief in the frozen word follows the rate of the surrounding speech (Sec. 3.1). (a)** Median displacement of Δ from each item's $s = 1$ condition against context scale at $r^* = 0.25$, order-statistic 95 percent intervals; the dashed grey series is the duration-matched control, slow side only. **(b)** Forest of b_{rate} per construction on the Whisper nats scale: the primary form b_{rate} (s: 1.0–1.7), the full-range b_{rate} (s: 0.6–1.7), and the two controls of Sec. 2.2 (both over s: 0.6–1.7); within a row the four models run from tiny to medium (color and marker as in (a)); all filled, since the rows compare constructions, not significance.

from the same corpus family follows the same pipeline. Conditions, their counts per item and their condition codes are in Appendix A, a silence-pad diagnostic in Appendix B.

2.4 Models

Sixteen recognition checkpoints were scored: nine form the main set, and seven further Whisper checkpoints enter as extensions, analyzed separately (Sec. 3.4; Appendix D). Of the nine, four are CTC models, wav2vec2-base-960h, wav2vec2-large-960h, wav2vec2-large-robust-ft-libri-960h and hubert-large-ls960-ft: self-supervised encoders fine-tuned on LibriSpeech, which score a transcript with no decoder or language model (Baevski et al., 2020; Hsu et al., 2021a;b). One of them, wav2vec2-large-960h, is the aligner (it set the word boundaries) and is kept out of primary aggregation (Appendix C); the other three are the primary CTC checkpoints. Five are Whisper models, tiny, base, small, medium and large-v3 (39M to 1.5B parameters): attention encoder-decoders trained by weak supervision on web audio (Radford et al., 2023); the four below 1.5B are the primary Whisper checkpoints. An audio language model, Qwen2-Audio-7B-Instruct (Chu et al., 2024), was scored on the same candidates and is reported only in Appendix B and Tables E1 and E2, as an existence result on a weak transcription base. The Whisper tiny-to-medium and CTC checkpoints document no speed-perturbation augmentation, so their rate behavior was learned rather than engineered; within Whisper only large-v2 documents any augmentation (Radford et al., 2023; Sec. 3.4).

2.5 The per-item slope and the measurement validity criteria

The per-item slope. Within one item and one condition we fit a straight line of Δ against $\ln s$ by ordinary least squares and keep its slope, written b with the condition as its subscript (b_{intact}, b_{duration}). Its unit is nats per e-fold of context slowing, and its sign is the direction: **a negative slope means the model loses belief as the context slows; the more negative the slope, the more belief lost to the same slowing; a slope of zero means the slowing leaves belief untouched.** The slope is measured at one reduction level r^* per model (r is the fraction of the word's duration kept, Sec. 2.1), the level where the word is most ambiguous to that model, so that a context effect has room to show. With the context untouched ($s = 1$), belief is high with the word whole ($r = 1$) and low with it cut out ($r = 0$); r^* is the level whose $s = 1$ belief lies nearest the midpoint of the two (a model at +6 and -2 nats has midpoint +2, so a level at +2.5 is chosen over one at +5). Only $s = 1$ conditions enter the choice, so the rate manipulation cannot bias it ($r^* = 0.25$ in each of the nine checkpoints; the three extension checkpoints at 0.35 are named in Appendix D; the whole-word level's shift is in Appendix B; the trajectories are Fig. 2a).

The rate-specific shift. The duration-matched control (Sec. 2.3) adds the same seconds as noise with no rate to hear, so its slope is what duration alone does, and the rate-specific belief shift for item i is the within-item difference $b_{\mathrm{rate},i} = b_{\mathrm{intact},i} - b_{\mathrm{duration},i}$. We give the median of $b_{\mathrm{rate},i}$ across items with its 95 percent interval and write it b_{rate}. The control exists on the slow side only, so the primary form takes both slopes over the shared scales, b_{rate} (s: 1.0–1.7). The full-range form b_{rate} (s: 0.6–1.7) lets the intact slope span all five scales and is given beside it (Appendix B).

The equivalent word reduction. A shift in nats is in the model's own units and cannot be compared between models. The word's own reduction axis supplies a unit in the same currency: b_{word} is the slope of Δ against r at $s = 1$, the belief a model gains per unit of word restored (from $r = 0$, cut out, to $r = 1$, whole). $E_i = b_{\mathrm{rate},i}/|b_{\mathrm{word},i}|$ then re-expresses the context shift as the fraction of the word whose removal would cost the same belief as one e-fold of context slowing ($E = -0.5$: slowing the context by e costs as much belief as cutting out half the word); the absolute value keeps the sign of E that of b_{rate}. Its median E is comparable across models (Appendix B).

Measurement validity criteria. Four criteria decide whether a checkpoint's slope may be interpreted, so that a null can only mean the effect is absent. (1) Reduction monotonicity: the median per-item correlation of Δ with r at $s = 1$ must reach +0.6. (2) Dynamic range: Δ must span at least 2 nats on that axis. (3) Finiteness: every score must be a finite number. (4) Resynthesis fidelity: the identity condition's word error, reported for every checkpoint rather than gated, must stay at the same checkpoint's word error on the untouched source audio, which it did in every checkpoint. Sec. 3 reports no quantity for a checkpoint that has not passed all four (values in Appendices B, D and F).

Reporting. We give medians with 95 percent order-statistic confidence intervals across items (two ranked item values chosen for 95 percent coverage; index correction in Appendix B) and claim a direction only where the interval excludes the stated bound, and call the result significant (filled markers in the figures; a result that is not significant is stated at its limits).

3 Experiments

3.1 The use exists

Belief in a bit-identical word falls as the speech around it slows, by more than the added seconds produce: all four whisper tiny-to-medium models move in the listeners' direction (Fig. 2). We call that dependence the model's **use** of the context's rate.

Rate versus added duration. b_{rate} (s: 1.0–1.7) of Sec. 2.5 is negative in all four models with the whole interval below zero, and so is the full-range form over all five scales (Fig. 2b; Table B1). The duration-clamped twin agrees in every model (Table B2), so the shift is not a residue of the pad construction.

Size in word units. In equivalent word reduction (Sec. 2.5) the shift is worth a fifth to three fifths of the word per e-fold, largest in tiny, smallest in medium, every interval below zero (Table B1).

Specificity to the word. The two control quantities of Sec. 2.2, each measured as its own b_{rate}, separate a percept of the frozen word, under which only Δ tracks the context's rate, from a generic response of the scoring machinery, under which all three quantities move together (Fig. 2b; Table B3). Belief in an unrelated word barely moves with the context's rate: its $b_{\mathrm{rate}}(\Delta_{\mathrm{ctrl}})$ is 5 to 28 times smaller than the target's over the same range (s: 0.6–1.7) and spans zero in three of the four models (base just below zero; Tables B1 and B3). The insertion control asks whether the effect is anchored to the frozen span or is a bias toward "or" wherever it is written; the answer splits by size. In tiny and base its $b_{\mathrm{rate}}(\Delta_{\mathrm{ins}})$ over the primary range (s: 1.0–1.7) stays small, about a fifth of the target's, so there the effect is anchored to the span; in small and medium it reaches three fifths to nine tenths of it (Table B3). The shift itself is untouched, since b_{rate} (s: 1.0–1.7) never uses the insertion candidate; what narrows is the specificity claim, which in these two models covers "or" as a word, not the frozen span.

Robustness. The shift survives every artifact control (Appendix B) and the lexical frame (Sec. 3.3).

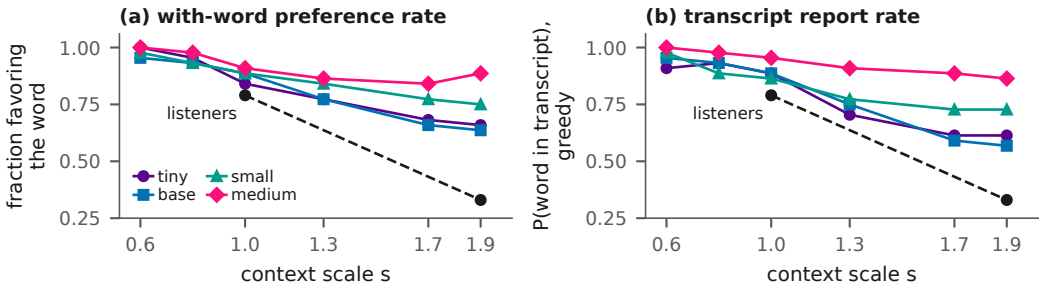


Figure 3: **The human bridge (Sec. 3.2). (a)** The with-word preference rate $P(\Delta > 0)$ against context scale, whisper tiny through medium, extended to the matched scale $s = 1.9$, beside the listeners' anchor (0.79 at the talker's rate falling to 0.33 at expansion 1.9; Dilley & Pitt, 2010). **(b)** The transcript report rate: the fraction of items whose greedy transcript still contains the word, beside the same listeners' anchor (Table E3).

3.2 The use directionally matches listeners

Comparing the effect between models and listeners. We convert belief into the listeners' scale: per context scale, the fraction of items whose Δ favors the with-word transcript at r^*, $P(\Delta > 0)$, the **with-word preference rate** (Fig. 3a; Table E1). The models' baselines land near the listeners' (0.84 to 0.91 against 0.79 at the talker's own rate), and the comparison is made at the human study's expansion factor, $s = 1.9$ (Dilley & Pitt, 2010). Slowing the context to $s = 1.9$ lowers the preference rate in every model, against the listeners' 46-point drop (0.79 to 0.33). The fall is significant in the three smaller models, at 30 to 54 percent of the human effect, and not significant in whisper-medium, whose interval spans zero (Fig. 3a; Tables E1 and E2).

The word's survival in the transcript. Each model transcribed the intact conditions at r^* (greedy decoding, the most probable token at each step, and five-candidate beam search), scored by one fixed presence rule (Appendix B); the fraction of items whose transcript still contains the word is the **transcript report rate**. At $s = 1.9$ it falls from $s = 1$ in every model, up to 14 of 44 items losing the word outright. The drops are 20 to 69 percent of the listeners' 46-point drop, significant in the three smaller models, whisper-small by one point at its lower bound, and not significant in whisper-medium; beam search agrees within one item per model (Fig. 3b; Table E3). Restricted to the items the positional rule matched at both scales, the drops are 15, 23, 17 and 5 points, significant in base and small (Appendix B). The stake is therefore not confined to log-probabilities: a word transcribed at the talker's own rate drops out when the speech around it is slowed.

Three signatures of the listeners' computation. Three signatures ask how much of the listeners' computation the shift shares (Fig. E1; Appendix E). Scaling the window and the context together leaves the word's relative rate unchanged, and listeners' reports return nearly to baseline; most of the window-alone displacement remains, a partial relative-rate response in three of four models (Fig. E1a, b). Re-timing only the speech beyond 2.0 s from the window still moves belief in tiny and base and is not significant in small and medium (Fig. E1c). On a 24-item set built without the word, belief in an unspoken "or" rises as the context speeds in three of four models but never crosses zero (Fig. E1d). The models move toward the word listeners hear; none of them hears it.

3.3 The use is word-general and holds beyond Whisper

Three further function words. The word-generality item set tests three further non-rhotic function words ("and", "but", "that"; 24 items each) and "are" (19 items) as a secondary near-homophone of reduced "or", with no pooling across words. All sixteen word-model pairs (four words by four Whisper models) pass the measurement validity criteria (Sec. 2.5; Appendix F). The shift is significant in twelve of the sixteen pairs (eight of the twelve non-rhotic pairs), in whisper-tiny for every word. The guards hold: the unrelated-word control spans zero in fifteen of sixteen pairs, and the artifact control accounts for at most 14 percent of the shift in every supported pair (Appendix F). The size pattern returns, whisper-medium significant on none of the three non-rhotic words while tiny is strongest on all three. What does not generalize is magnitude: the three non-rhotic words are worth a tenth to three tenths of the word per e-fold, against a fifth to three fifths for "or" (Table F4).

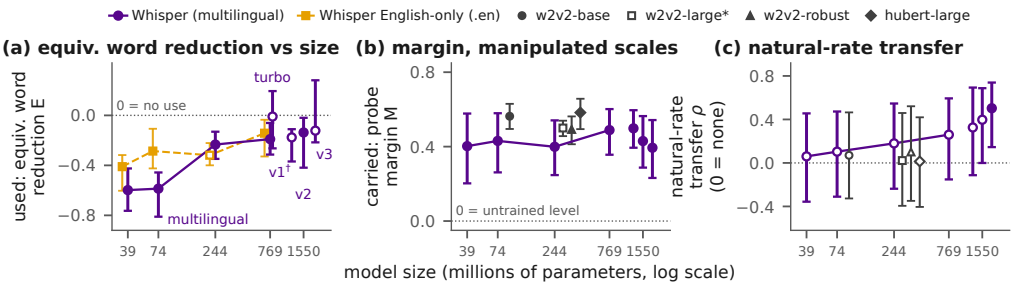


Figure 4: **Representation against use on one size axis (Secs. 3.4 and 3.5). (a)** Equivalent word reduction for twelve Whisper checkpoints: the four primary Whisper checkpoints connected, extensions free-standing, the English-only line dashed; filled = interval below zero with the checkpoint's unrelated-word control at least fivefold smaller than its target (Table D3), open = spans zero or a failed control (large-v3 and its turbo derivative, bounds; small.en and v1, whose unrelated-word controls move with their targets; † = v1; Appendix D); the 1.5B and English-only markers are offset horizontally here and in (b). **(b)** Probe margin M at the frozen frames (trained minus untrained accuracy, chance 0.20); the open marker among the three large CTC checkpoints (about 315M) is the aligner-shared model; wav2vec2-base sits at 95M. **(c)** The natural-rate transfer: the probe's estimate on each item's unmanipulated recording against its natural word rate (Spearman, bootstrap intervals, $n = 23$; open = spans zero).

The CTC family. A CTC model has no decoder. It scores a transcript by adding up the probability of every way the transcript's letters could be lined up with the audio frames, over the whole utterance. Re-timing the context changes the frames outside the frozen span, so it moves the score of every candidate transcript, the two controls of Sec. 2.2 included, since they differ from the reference only outside the span (Appendix G, Fig. G1b). The rule for this family, the agreement rule, therefore asks two estimates to agree before a direction is claimed, the raw shift and the insertion-subtracted shift, the shift with the insertion control subtracted. They do not agree, so the interpretation is withheld as undetermined. These scores cannot separate a percept of the word from the scoring's own response to re-timed audio, a limit of the scores rather than an absence of the effect. The item-level controls correction (Appendix G) then asks the question directly. If the target's shift were the same mechanical response as the controls', the items whose controls move most would be the items whose target moves most. They are not. Across items the controls explain 1 to 7 percent of the variance in the target's shift, and the corrected shift, the shift left after removing what the controls predict, is as large as the raw one in every primary CTC checkpoint (Appendix G; exploratory on this item set). On the word-generality item set that corrected shift is significant in twelve of twelve primary word-model pairs (four words by three primary CTC checkpoints; sixteen of sixteen with the aligner-shared model; Appendix F, Fig. F1b). The CTC family therefore shows a significant, word-general shift; whether it is specific to the frozen span cannot be established for this family, which has no within-span control, a control candidate inside the frozen span (Appendix B).

3.4 The use declines with size

The decline across the four primary sizes. Across the four primary Whisper checkpoints, 39M to 769M parameters, the equivalent word reduction (Sec. 2.5) falls to about a third of its value at the smallest size (Fig. 4a; Table B1). The decline is significant between the two smallest and the two largest sizes (non-overlapping intervals) and concentrated between 74M and 244M. In point value the with-word preference rate and the transcript report rate of Sec. 3.2 each drop less in the two largest models than in the two smallest (their intervals overlap; Fig. 3; Tables E1 to E3). The far context moves belief only in the two smallest (Fig. E1c).

Eight further checkpoints on the same axis. Large-v3 and the seven extension checkpoints of Sec. 2.4 join the axis, analyzed separately, and none overturns the decline (Fig. 4a; Appendix D). At 1.5B the use persists in two of the three checkpoints, and the unaugmented recipe, large (v1), is significant like the augmented large-v2, so the decline is not an augmentation artifact. Large-v3 and its turbo derivative are bounds tied to the v3 encoder, not to size, and the English-only line declines as well. Five sizes and twelve checkpoints in one family are a trend, not a law (Sec. 4).

3.5 THE REPRESENTATION STAYS

The probe and its checks. The decline of Sec. 3.4 leaves two alternative interpretations: the larger models could have lost the context's rate from the frames of the word, or kept it and stopped consulting it. A probe at the frozen frames separates them. At fixed r the proximal frames contain no information about s, so whatever a layer knows about s there was propagated from the context. We call that propagated structure the rate representation: an activation pattern at the frozen frames from which the context's rate can be decoded. A multinomial logistic probe is trained per model and layer to classify the five context scales from the frozen frames (Appendix H). Because even the untrained architecture propagates enough context structure to classify above chance, the probe margin M is the probe's accuracy on the trained checkpoint minus its accuracy on the same architecture untrained, at the selected layer (chosen on a held-out speaker half; Appendix H). In whisper-medium the probe decodes the five scales at accuracy 0.80 against 0.30 for the untrained architecture, a margin M of +0.49 (speaker-averaged paired difference; chance 0.20; Table H1). Three further checks, the input baseline, the position control and the natural-rate transfer, separate a rate representation from what could mimic it; a natural-rate representation is claimed only where all four checks hold.

Decodability across the nine checkpoints. The input baseline, the same probe on the front-end features, sits at chance in every probed checkpoint (at most +0.01 above it), while the context's rate is decodable from trained activations in all eleven probed checkpoints (the nine plus large (v1) and large-v2) with M running +0.39 to +0.58 and every interval above zero (Fig. 4b; Table H1). The position control, the probe cross-applied to duration-matched conditions, collapses rather than transfers in every probed checkpoint, so the probe decodes rate, not position (Table H1).

The natural-rate transfer. The strictest check asks whether the representation is about speech rate itself: the probe's estimate applied to each item's untouched recording should track its natural word rate. The transfer's point value rises with size and it is significant at large-v3 alone, at +0.50 [+0.15, +0.74] (Fig. 4c; Table H2); every other probed checkpoint's interval spans zero, and at $n = 23$ a spanning interval is inconclusive rather than ruled out. All four checks are therefore met in none of the eight probed checkpoints below 1.5B (the nine without large-v3) and in one of the three at 1.5B: a natural-rate representation passing all four checks exists only at the top of the size range. Across the size axis, then, the use falls while the probe margin does not, and the strictest check is significant only at the top (Fig. 4): size changed what the computation consults, not what the representation contains, a description after the fact rather than a tested prediction.

3.6 ERASING THE REPRESENTATION

A causal step removes the representation the probe decodes and re-measures b_{rate} (s: 1.0–1.7) of Sec. 2.5, with no retraining. The step is a linear erasure at the probe's layer: each frame of the frozen span is projected off the four directions that separate the five context-scale class means, so the context's rate can no longer be read linearly from those frames while the rest of the frame is left as intact as a linear edit allows (Ravfogel et al., 2020; Belrose et al., 2023). The quantity measured is dS, the change in b_{rate} from the clean to the erased pass, positive for attenuation (the direction expected if the decision runs on the representation) and negative for enlargement. Every estimate under this targeted erasure is compared with random erasures of four directions, with disruption-matched random erasures, and with a no-erasure pass (Fig. 5; Appendix I).

Whisper under the erasure. In whisper-tiny the erasure attenuates the shift by about half, with identity word error rising from 0.050 to 0.083, but a disruption-matched random erasure reproduces the attenuation (Fig. 5a; Appendix I). The attenuation is therefore consistent with generic disruption of four directions, the documented failure mode of erasure-based tests (Kumar et al., 2022), and no necessity is claimed. Base, small and medium are not significant at rank 4, and no swept layer yields an attenuation that is both selective and leaves the reduction axis intact (Appendix I).

The CTC family under the erasure. The same erasure makes the shift larger in two of the three primary CTC checkpoints and in the aligner-shared model (Fig. 5b; Appendix I). There every interval is below zero, every random erasure spans zero and transcription is untouched. One account, stated as speculation: if the representation holds the scoring invariant to the context's rate, the opposite stance to normalization, removing it lets that rate act unchecked. Whatever its mechanism, this enlargement is the one intervention effect in the study that holds under a disruption-matched control.

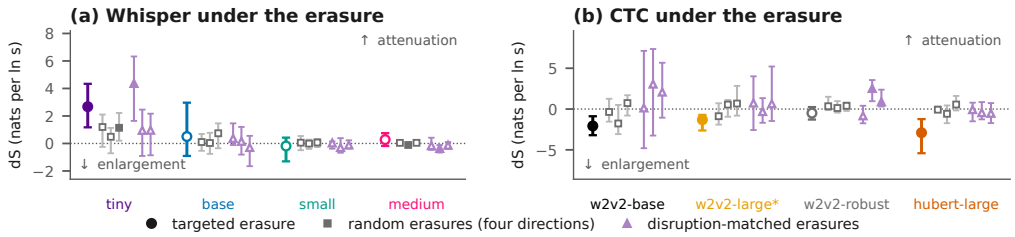


Figure 5: **Erasing the representation (Sec. 3.6).** dS, the paired change of b_{rate} (s: 1.0–1.7) under erasure at each model's probe layer (corrected order-statistic intervals; filled where they exclude zero): the targeted erasure (circles), random erasures (grey squares) and disruption-matched random erasures (triangles). **(a)** Whisper: tiny's attenuation is reproduced by a disruption-matched random. **(b)** CTC: the enlargement holds where every random spans zero (Appendix I).

4 Discussion

We show that the rate of the surrounding speech moves what these models hear of a bit-identical word. The shift is real and partial on the listeners' scale, reproduced by the transcript, three further words and, corrected, the CTC family, and it shrinks with size, persisting at 1.5B in two of three checkpoints. The probe margin M holds across sizes; its natural-rate transfer is significant only where the use is smallest. Erasure returned a bound in Whisper and a selective enlargement in the CTC family. The use exists; it declines with size; the representation stays.

What the approach measures that a probe cannot. The probe alone would have called every probed checkpoint alike; the behavioral experiment, on the same frozen input, separates them by family and by size. In systems neuroscience decodable information is not information the animal uses until a behavioral or causal test says so (Panzeri et al., 2017).

Why use recedes while the representation stays. The account the data allow is reliability weighting (Ernst & Banks, 2002; Kleinschmidt & Jaeger, 2015). A perceiver weights each cue by the quality of the local evidence. The smallest models transcribe this material least accurately, and two checks support the account: at the item level in two of the four primary checkpoints, and under degraded evidence beneath the word in one of three tested (Appendix J). Graded support is not confirmation; a lexical-prior account, larger models weighting text more and audio less, is not excluded.

What the study establishes, and the questions it opens. First, the reduced word here is a time-compressed token of read speech, whereas the human tokens were casually produced and spectrally reduced; naturally reduced conversational tokens and a same-modulation context that is not heard as speech (Pitt et al., 2016; Bosker, 2017) are the next materials. Second, the probe's natural-rate transfer (Sec. 3.5) is significant only at the top of the size range on 23 narrow-range items, and a rate-diverse probe item set is needed below 1.5B. Third, the two experiments establish that the representation is present and that the use declines, on the same frozen input, but not that the decision runs on that representation (Sec. 3.6). Two mechanisms are live for Whisper, its timestamp objective and the production statistics of function-word reduction (Appendix G); an intervention on the decoder's activations (Glazer et al., 2026) is the test.

What generalizes. The test's requirements are not about speech: any domain that can move a variable while a decision-relevant span stays bit-identical, score the decision about that span on a graded metric, match a control to each non-target consequence and probe the frozen positions can ask whether behavior uses a variable, not only whether it is decodable. Whether the dependence is a defect or a feature depends on the deployment (Sec. 1); it should be measured, not averaged into word error. The word is frozen; only the clock around it moves; and what the model reports moves with it.

Reproducibility Statement

Item lists, per-condition scores, summary tables, decode outputs, the eraser subspaces and the figure scripts are released on acceptance. The behavioral, probe and erasure experiments were preregistered with their analysis code before any job ran; the analyses added later either had their rule

fixed before their job ran or re-express committed scores under those frozen rules (the interval and control tables of the appendices). The rest are labeled exploratory where they are reported (the item-level controls correction of Sec. 3.3 and Appendix G, exploratory on the primary item set and fixed before scoring on the word-generality set; the item-level ambiguity check of Appendix J; and the doubled-letter re-scoring of Appendix B). Every extension of the study (the size-axis checkpoints, the $s = 1.9$ scale, the word-generality item set, the erasure and its follow-ups, the probe at 1.5B) first reproduced the quantities it builds on, bit-exactly for stimuli and scores, at 1×10^{-6} or 1×10^{-9} for derived quantities, and at the erasure follow-ups' own gate of 0.02 nats on the paired shift, where agreement was exact as well. An independent implementation reproduced the headline quantities at the same tolerances. The $s = 1.9$ scale serves the listener comparison alone and enters no slope. Intervals are corrected order-statistic 95 percent intervals across items unless stated; values are printed to two decimals, and a bound that would print as zero is given the decimals that show its sign, since significance is read from that sign; the paired preference-rate and transcript drops of Sec. 3.2 are given with paired-proportion intervals (method 10 of Newcombe, 1998) with the counts of items losing and gaining the word. That method replaced a first paired bootstrap, which cannot contain zero when every discordant item runs one way; the replacement moved one call, whisper-medium's transcript drop at $s = 1.9$, to not significant. Scoring and analysis of the eight checkpoints below 1.5B run in about 80 minutes of wall clock once stimuli are built (construction adds 20 minutes), and the whole battery stays under one GPU-hour per added checkpoint including its probe extraction.

Ethics statement

This study uses only public corpora (LibriSpeech) and open model checkpoints; no human subjects were run and no new human data were collected. The human comparison values are published group statistics from the cited literature.

AI use statement

Large language models (a coding and writing assistant) were used under author direction throughout this project. Under the ICLR 2027 policy, the categories that applied are, among those requiring disclosure: feedback on research methodology and experiment design; implementation of methods (stimulus construction, scoring, probe and erasure code); and interpretation of results. Among the recommended-disclosure categories: drafting and editing of the manuscript, figure code, code review, literature search and summarization, reference formatting, and structural suggestions. No synthetic data were generated and no human data were involved. Every analysis rule was fixed by the authors; the headline model-derived numbers were recomputed by an independent implementation against the per-condition scores (released on acceptance) at the tolerances stated in the Reproducibility statement, and every printed number was traced to its results file in a final audit; every citation was checked at the source; and the authors reviewed and take responsibility for all content. The same disclosure is entered in the submission form.

References

F. Adolfi, J. S. Bowers, and D. Poeppel. Successes and critical failures of neural networks in capturing human-like speech recognition. *Neural Networks*, 162:199–211, 2023.

M. Aylett and A. Turk. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56, 2004.

M. M. Baese-Berk, C. C. Heffner, L. C. Dilley, M. A. Pitt, T. H. Morrill, and J. D. McAuley. Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science*, 25(8):1546–1553, 2014.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

A. Bell, J. M. Brenier, M. Gregory, C. Girand, and D. Jurafsky. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60 (1):92–111, 2009.

N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786, 2007.

H. R. Bosker. Accounting for rate-dependent category boundary shifts in speech perception. *Attention, Perception, & Psychophysics*, 79(1):333–343, 2017.

H. R. Bosker and O. Ghitza. Entrained theta oscillations guide perception of subsequent speech: Behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience*, 33(8): 955–967, 2018.

M. Brown, M. K. Tanenhaus, and L. Dilley. Syllable inference as a mechanism for spoken language understanding. *Topics in Cognitive Science*, 13(2):351–398, 2021.

A. Chang, X. Teng, M. F. Assaneo, and D. Poeppel. The human auditory system uses amplitude modulation to distinguish music from speech. *PLOS Biology*, 22(5):e3002631, 2024.

Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou. Qwen2-Audio technical report. arXiv:2407.10759 (preprint), 2024.

M. de Heer Kloots and W. Zuidema. Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in Wav2Vec2.0. In *Proceedings of Interspeech 2024*, pp. 4593–4597, 2024.

M. de Seyssel, M. Lavechin, H. Titeux, A. Thomas, G. Virlet, A. Santos Revilla, G. Wisniewski, B. Ludusan, and E. Dupoux. ProsAudit, a prosodic benchmark for self-supervised speech models. In *Proceedings of Interspeech 2023*, pp. 2963–2967, 2023.

L. C. Dilley and J. D. McAuley. Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, 59(3):294–311, 2008.

L. C. Dilley and M. A. Pitt. Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21(11):1664–1670, 2010.

L. C. Dilley, T. H. Morrill, and E. Banzina. New tests of the distal speech rate effect: Examining cross-linguistic generalization. *Frontiers in Psychology*, 4:Article 1002, 2013.

N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel. Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, 81:181–187, 2017.

Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

T. Fel, I. F. Rodriguez Rodriguez, D. Linsley, and T. Serre. Harmonizing the object recognition strategies of deep neural networks with humans. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

A.-L. Giraud and D. Poeppel. Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, 15(4):511–517, 2012.

N. Glazer, Y. Segal-Feldman, H. Segev, A. Shamsian, A. Buchnick, G. Hetz, E. Fetaya, J. Keshet, and A. Navon. Beyond transcription: Mechanistic interpretability in ASR. In *Proceedings of AAAI-26*, volume 40, pp. 37407–37416, 2026.

Y. Gong, S. Khurana, L. Karlinsky, and J. Glass. Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Proceedings of Interspeech 2023*, pp. 2798–2802, 2023.

J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 2733–2743, 2019.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021a.

W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. In *Proceedings of Interspeech 2021*, pp. 721–725, 2021b.

D. Jurafsky, A. Bell, M. Gregory, and W. D. Raymond. Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee and P. Hopper (eds.), *Frequency and the emergence of linguistic structure*, pp. 229–254. John Benjamins, 2001.

J. Kirby, I. Krehan, and M. Gubian. Perceptual compensation for tonal context in self-supervised speech models. In *Proceedings of Interspeech 2026 (to appear); arXiv:2606.17835*, 2026.

D. F. Kleinschmidt and T. F. Jaeger. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2):148–203, 2015.

T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. Audio augmentation for speech recognition. In *Proceedings of Interspeech 2015*, pp. 3586–3589, 2015.

A. Kösem, H. R. Bosker, A. Takashima, A. Meyer, O. Jensen, and P. Hagoort. Neural entrainment determines the words we hear. *Current Biology*, 28(18):2867–2875.e3, 2018.

A. Kumar, C. Tan, and A. Sharma. Probing classifiers are unreliable for concept removal and detection. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

T. Kuribayashi, Y. Oseki, T. Ito, R. Yoshida, M. Asahara, and K. Inui. Lower perplexity is not always human-like. In *Proceedings of ACL-IJCNLP 2021*, pp. 5203–5217, 2021.

K. Lasri, T. Pimentel, A. Lenci, T. Poibeau, and R. Cotterell. Probing for the usage of grammatical number. In *Proceedings of ACL 2022*, pp. 8818–8831, 2022.

C. McDougall, A. Conmy, C. Rushing, T. McGrath, and N. Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop*, pp. 337–363, 2024.

I. R. McKenzie, A. Lyzhov, M. Pieler, A. Parrish, A. Mueller, A. Prabhu, et al. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*, 2023.

Y. Meng, S. Goldwater, and H. Tang. Effective context in neural speech models. In *Proceedings of Interspeech 2025*, pp. 246–250, 2025.

B. T. Meyer, T. Brand, and B. Kollmeier. Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. *Journal of the Acoustical Society of America*, 129(1): 388–403, 2011.

N. Mirghafori, E. Fosler, and N. Morgan. Towards robustness to fast speech in ASR. In *Proceedings of ICASSP 1996*, pp. 335–338, 1996.

R. G. Newcombe. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17(22):2635–2650, 1998.

B.-D. Oh and W. Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350, 2023.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of ICASSP 2015*, pp. 5206–5210, 2015.

J. Pang, A. Chaubey, and M. Soleymani. Do audio LLMs listen or read? Analyzing and mitigating paralinguistic failures with VoxParadox. In *Proceedings of ICML 2026; arXiv:2605.27772*, 2026.

S. Panzeri, C. D. Harvey, E. Piasini, P. E. Latham, and T. Fellin. Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron*, 93(3):491–507, 2017.

A. Pasad, J.-C. Chou, and K. Livescu. Layer-wise analysis of a self-supervised speech representation model. In *Proceedings of IEEE ASRU 2021*, pp. 914–921, 2021.

M. A. Pitt, C. Szostak, and L. C. Dilley. Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. *Attention, Perception, & Psychophysics*, 78(1):334–345, 2016.

D. Poeppel and M. F. Assaneo. Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, 21(6):322–334, 2020.

C. Pouw, M. de Heer Kloots, A. Alishahi, and W. Zuidema. Perception of phonological assimilation by neural speech recognition models. *Computational Linguistics*, 50(4):1557–1585, 2024.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML 2023, PMLR 202*, pp. 28492–28518, 2023.

S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL 2020*, pp. 7237–7256, 2020.

S. Ravfogel, G. Prasad, T. Linzen, and Y. Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of CoNLL 2021*, pp. 194–209, 2021.

A. Ravichander, Y. Belinkov, and E. Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of EACL 2021*, pp. 3363–3377, 2021.

R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, 2023.

M. A. Siegler and R. M. Stern. On the effects of speech rate in large vocabulary speech recognition systems. In *Proceedings of ICASSP 1995*, pp. 612–615, 1995.

S. ten Oever and A. E. Martin. An oscillating computational model can track pseudo-rhythmic speech by using linguistic predictions. *eLife*, 10:e68066, 2021.

W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *Proceedings of ICASSP 1993*, pp. 554–557, 1993.

K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *Proceedings of ICLR 2023*, 2023.

J. Wei, N. Kim, Y. Tay, and Q. V. Le. Inverse scaling can become U-shaped. In *Proceedings of EMNLP 2023*, pp. 15580–15591, 2023.

A STIMULUS CONSTRUCTION AND CONDITION CODES

This appendix gives the construction detail behind Secs. 2.1 and 2.3 and the condition codes under which the released files are labeled (Table A1).

Table A1: Conditions per item, the code under which the released files label each, and what each isolates.

Condition	Code	Conditions per item	Isolates
intact	A	5 r × 5 s = 25	the distal metric and the reduction-axis ground truth
duration-matched control	B1a	3 r × s' in {1.3, 1.7} = 6	duration and window position without a rate percept
duration-clamped twin	B1b	3 r × 5 s = 15	the rate axis at fixed total duration
artifact control	B2	2 r × 4 s' = 8	artifact dose without net rate or duration change
silence-pad check	B3	1 r × 2 s' = 2	pad-content pathology (diagnostic only)
together-scaled / window-alone	P3T / P3O	2 factors × 2 r, twice = 8	relative versus absolute rate
far-context / near-context (subset)	FAR / NEAR	2 × 2 s × 1 r = 4	proximity of the effective context
word-absent item set (label p2 in the released item list)	A, B1a, B1b, B2	5 + 2 + 5 + 4 = 16 (24 items × 16 = 384)	the appearing-word direction, with controls (Sec. 3.2)
word-generality item sets (`wg_and`, `wg_but`, `wg_that`, `wg_are`)	A, B1a, B1b, B2	25 + 6 + 15 + 8 = 54	lexical generality of the shift (Sec. 3.3; Appendix F)
untouched source audio	ORIG	1	the natural-rate transfer input (Sec. 3.5) and source-audio word error rate (Appendix B)

Item selection. Eligible utterances run 6 to 16 s, contain "or" between two neighboring words of at least three characters, keep at least 2.0 s of speech before and 1.0 s after the proximal window, and contribute at most two items per speaker. Boundaries come from the aligner's (wav2vec2-large-960h) Viterbi emission-peak convention, so window edges approximate the word onset and offset. Six candidates were dropped (five failing the context floor, one with an "or" span of 100 ms), with no alignment failures. A planned manual screen of the items' text plausibility was waived, and no item was struck. The primary item set is the full yield of test-clean (2,620 utterances, 40 speakers, 5.4 h) under these rules, 44 items from 25 speakers; no eligible utterance was sampled out except by the per-speaker cap. The replication set is the 23-item pilot set from train-clean-100, the same corpus family, treated as replication only because the family-level decision rules were set on it.

Resynthesis. The "or" span (10 ms WSOLA analysis frames) and both context chunks (40 ms frames) pass through WSOLA in every condition that contains them. The flanking words inside the proximal window are resynthesized only in the together-scaled and window-alone conditions and are bit-identical constants everywhere else. The duration-matched control's noise pad is a phase-randomized, RMS-matched surrogate of the item's own context, and pads are multiples of 320 samples, which preserves the proximal bit-identity.

Why the grid stops at 1.7. The 6 to 16 s eligibility band and the 1.7 cap were set together so that the slowest condition of the longest eligible utterance stays inside Whisper's fixed 30 s input window. At the cap, 16 s scales to 27.2 s; the human study's expansion of 1.9 would reach 30.4 s and truncate exactly the manipulated far context. The grid is symmetric on the log scale the slopes are fit on (1.7 is close to 1/0.6), and the realized item set's longest utterance is 15.5 s, the headroom that later admitted the $s = 1.9$ comparison of Sec. 3.2.

B THE BEHAVIORAL EXPERIMENT PER MODEL AND ITS CHECKS

This appendix gives the per-model values of the behavioral experiment (Tables B1 to B3), the measurement validity values of Sec. 2.5, the equivalent word reduction in display form, and the checks that Secs. 2.5 and 3.1 to 3.3 cite without showing, each under the paragraph it serves.

Measurement validity criteria (Sec. 2.5), per family. Reduction monotonicity: the median per-item correlation of Δ with r at $s = 1$ is +0.90 to +1.00. Dynamic range: Δ spans 5.8 to 9.6 nats on the reduction axis in the five Whisper checkpoints of the nine (7.2 to 9.6 in tiny through medium) and 44.4 to 53.1 in the CTC family. Finiteness: every score is finite. Resynthesis fidelity: identity-condition word error is 0.000 to 0.050 in the Whisper family and 0.120 in the three primary CTC checkpoints (0.125 in the aligner-shared model, Appendix C). The CTC value is a floor set by our greedy CTC decoding and not a resynthesis cost: the word error on the untouched source audio of the same items is also 0.120, and 0.098 to 0.140 across the replication and word-absent item sets. The decoding merges repeated letters twice, so a doubled letter is lost ("yellow" is transcribed "yelow") and every reference word with one counts as an error. Such words are about a tenth of each transcript (median share 0.096 on the primary item set). In a re-scoring added after the primary results (exploratory), doubled letters are merged on both sides before scoring. The source-audio word error of the four CTC checkpoints then has medians of 0.000 to 0.012 on the primary item set (means 0.023 to 0.030), and the Whisper values there do not change. The floor is set by the reference text, so it is the same in every condition of an item to within one word. It therefore cancels in the two comparisons that use it, identity against source audio here and clean against erased in Appendix I. No belief score depends on decoded text. Qwen2-Audio sits on a weaker base at 0.31 and is reported in this appendix only. All nine checkpoints pass; the extension checkpoints' values are in Appendix D and the word-generality item set's in Appendix F.

The equivalent word reduction in display form (Sec. 2.5). The definition of Sec. 2.5 in one line:

$$E_i \;=\; \frac{b_{\mathrm{rate},i}}{|b_{\mathrm{word},i}|}.$$

The absolute value in the denominator keeps the sign of E that of b_{rate}. A model that loses 6 nats at full excision and has $b_{\mathrm{rate},i} = -3$ has $E_i = -0.5$, half the word per e-fold of context slowing. One bound comes with the unit: the context scales reach 1.7, so a value per e-fold is a projection about twice past the widest slowing built, and it describes the fitted slope rather than a condition we ran.

Why the slope is measured at the most ambiguous level (Sec. 2.5). At the whole-word level ($r = 1$) the raw slope of Δ on $\ln s$ (no control subtracted) is -0.87 to -0.32 across the four primary Whisper checkpoints, four to fourteen times smaller than the same slope at the most ambiguous level: a context effect needs room to show, and the whole word leaves it none.

Robustness of the shift (Sec. 3.1). Replacing the noise pad of the duration-matched control with silence changes the control's slope by -0.09 to +0.08 in the Whisper models (by +3.54 in wav2vec2-base, consistent with that family's mechanical sensitivity), so the result does not depend on the pad's content. The artifact control's slope is 0.0 to 3.4 percent of b_{rate} (s: 1.0–1.7) in whisper-tiny and whisper-base and 12 and 23 percent in whisper-small and whisper-medium, where the shift is smallest, so it does not depend on the artifact dose either. On the 23-item replication set the shift (full-range form) is negative with every interval below zero in all four Whisper models (-3.21 to -4.70). A speaker-clustered bootstrap reproduces the item-level interval of the full-range form in every model (whisper-base [-5.57, -3.33] against [-5.87, -3.30]), and correcting the order-statistic index of the intervals, which had given 92 to 96 percent coverage, changes 0 of 24 calls (three quantities by the eight checkpoints below 1.5B).

Table B1: The behavioral experiment per model (median [95 percent order-statistic CI, corrected] over the 44 items; nats per $\ln s$). A negative slope is the listeners' direction, and the further below zero, the more belief that model loses to the same context slowing (Sec. 2.5); no magnitude is compared across families. The full-range column takes the intact slope over all five scales, with its own interval. The controls are in Table B3, the duration-clamped twin and the together-scaled ratio in Table B2; the extension checkpoints are in Appendix D, the aligner-shared model in Appendix C, the word-generality item set in Appendix F.

Model	b_{rate} (s: 1.0–1.7), primary	b_{rate} (s: 0.6–1.7)	Equivalent word reduction (s: 1.0–1.7)	Result
whisper-tiny	−5.43 [−7.94, −3.79]	−4.08 [−6.56, −2.89]	−0.60 [−0.76, −0.42]	listeners' direction; qualifies as rate normalization (Appendix E)
whisper-base	−4.68 [−5.97, −3.51]	−4.33 [−5.87, −3.30]	−0.59 [−0.81, −0.46]	listeners' direction; qualifies as rate normalization (Appendix E)
whisper-small	−2.10 [−2.81, −1.06]	−2.70 [−4.08, −1.87]	−0.23 [−0.35, −0.13]	listeners' direction; qualifies as rate normalization (Appendix E)
whisper-medium	−1.45 [−2.15, −0.51]	−3.06 [−4.05, −1.31]	−0.19 [−0.31, −0.06]	listeners' direction; does not qualify as rate normalization (Table B2)
wav2vec2-base	−13.77 [−24.53, −8.86]	−11.32 [−19.22, −8.87]	−0.37 [−0.58, −0.16]	raw result undetermined (mechanics-dominated); corrected shift significant, specificity unestablished (Sec. 3.3; Appendix G)
wav2vec2-robust	−15.69 [−24.66, −11.01]	−18.09 [−24.14, −14.08]	−0.42 [−0.61, −0.35]	raw result undetermined (mechanics-dominated); corrected shift significant, specificity unestablished (Sec. 3.3; Appendix G)
hubert-large	−11.90 [−16.95, −8.05]	−10.16 [−12.25, −5.56]	−0.30 [−0.42, −0.18]	raw result undetermined (mechanics-dominated); corrected shift significant, specificity unestablished (Sec. 3.3; Appendix G)
Qwen2-Audio-7B	−3.32 [−5.00, −1.00]	−6.45 [−9.28, −2.90]	−0.32 [−0.70, −0.07]	listeners' direction, word-specific by its unrelated-word control; existence result only: weak transcription base (identity word error 0.31), no probe, signature or erasure experiment

Table B2: The duration-clamped twin, measured as b_{rate} over the same range (Sec. 2.3), and the together-scaled ratio of Appendix E, which decides whether a model's result is called rate normalization (its interval's upper bound below 1); Qwen2-Audio's values are computed but not reported (existence result; Table B1).

Model	Duration-clamped twin (s: 1.0–1.7)	Together-scaled ratio
whisper-tiny	−4.94	0.71 [0.52, 0.98]
whisper-base	−5.21	0.59 [0.43, 0.70]
whisper-small	−2.46	0.81 [0.67, 0.97]
whisper-medium	−1.94	0.74 [0.59, 1.06]
wav2vec2-base	−9.21	0.58 [0.51, 0.80]
wav2vec2-robust	−17.43	0.55 [0.47, 0.68]
hubert-large	−11.37	0.70 [0.65, 0.82]
Qwen2-Audio-7B	not reported	not reported

Table B3: The two control quantities of Sec. 2.2 per model, each measured as its own b_{rate} (median [95 percent order-statistic CI, corrected]); the insertion control on both ranges so that its comparison with the target in Sec. 3.1 can be made on the primary form's range.

Model	Insertion-alone $b_{\text{rate}}(\Delta_{\text{ins}})$ (s: 0.6–1.7)	Insertion-alone $b_{\text{rate}}(\Delta_{\text{ins}})$ (s: 1.0–1.7)	Unrelated-word $b_{\text{rate}}(\Delta_{\text{ctrl}})$ (s: 0.6–1.7)
whisper-tiny	−0.67 [−1.59, +0.19]	−1.02 [−1.66, −0.45]	+0.30 [−0.68, +1.17]
whisper-base	−1.36 [−2.06, −0.79]	−1.00 [−1.70, −0.30]	−0.84 [−2.06, −0.10]
whisper-small	−2.06 [−2.75, −0.74]	−1.35 [−1.96, −0.60]	−0.10 [−1.16, +0.38]
whisper-medium	−1.84 [−2.54, −1.23]	−1.32 [−1.82, −0.88]	−0.38 [−1.85, +0.73]
wav2vec2-base	−29.43 [−38.19, −19.40]	−10.59 [−18.82, −8.34]	+36.75 [+27.98, +48.88]
wav2vec2-robust	−31.06 [−43.24, −20.29]	−18.87 [−24.88, −12.90]	+12.20 [+10.79, +16.03]
hubert-large	−33.56 [−43.73, −16.67]	−13.98 [−23.09, −8.98]	+20.08 [+17.28, +25.06]
Qwen2-Audio-7B	−1.81 [−2.88, −0.89]	−1.45 [−2.73, −0.80]	−0.71 [−2.75, +0.0002]

Decode detail (Sec. 3.2). Decoded were the primary item set's intact conditions at $r = 0.25$ and $r = 0$, the word-absent item set's intact conditions and the untouched source audio. The $s = 1.9$ scale was decoded at $r = 0.25$ in the four primary Whisper checkpoints only, and the CTC checkpoints were decoded greedily only. The presence rule counts a transcript as containing the word when "or" occurs between the two words that neighbor it in the reference. When those neighbors are not both found, it falls back to counting "or" tokens against the without-word transcript. The fallback fires for 14 to 40 percent of Whisper hypotheses (greedy and beam pooled), most often at the slowest scale, and for 53 to 66 percent of CTC hypotheses. The CTC rate is inflated by the decoding floor above, since a neighbor with a doubled letter never matches. With doubled letters merged before matching it is 24 to 38 percent, and no presence call changes in either family. Restricted to items the positional rule matched at both scales, the paired transcript drops at $s = 1.9$ are 15, 23, 17 and 5 points in tiny, base, small and medium (base and small exclude zero, tiny and medium do not). This is a conservative bound, since the restriction keeps the items whose neighbors survived the re-timing.

The CTC scores (Sec. 3.3; Appendices E and G). On the word-absent item set the CTC slopes are negative and typically three to five times Whisper's, and on the primary item set the family's transcript report rate drops 18 to 39 points from $s = 1.0$ to 1.7. Both estimates sit under the same limit as every CTC score (Appendix G). The candidate that could have given the family a within-span control, a second "or" inserted at the frozen site, is degenerate. The doubled token costs a median 36 to 40 nats at $s = 1$ (21 to 61 across the hand-checked items), so it never sits near the decision boundary, and its slope is kept in the released tables only.

C THE ALIGNER-SHARED MODEL

This appendix reports the aligner-shared model of Sec. 2.4 beside its family. wav2vec2-large-960h supplies the forced-alignment boundaries for every item, so it is kept out of primary aggregation, for its role rather than its behavior: on every quantity its interval overlaps its family's, although in point value it is the family's largest shift. Its values run in its family's direction throughout: b_{rate} (s: 1.0–1.7) −26.41 [−34.52, −14.03]; the insertion-subtracted shift +12.21 [+6.08, +20.96]; the unrelated-word control +21.51 [+14.47, +34.66], opposite in sign to its target shift as in its family peers. The segment-restricted score of Appendix G leaves the target shift unchanged and moves the insertion-subtracted shift to +13.01 [+5.74, +20.52]. In the item-level controls correction of Appendix G it also matches its family: rank correlations +0.15 and +0.03 between target and controls, 2 percent of item variance explained, corrected shift −24.23 [−31.56, −13.29].

D THE EXTENSION CHECKPOINTS

The seven extension checkpoints of Sec. 2.4 were scored on the full battery and are reported in Sec. 3.4 beside the four primary Whisper checkpoints; this appendix gives their values (Tables D1 to D3) and the reasoning behind Sec. 3.4's second paragraph.

Their measurement validity (Sec. 2.5). Identity-condition word error is 0.000 to 0.036 across the seven, every checkpoint passes the four criteria, and r^* is 0.25 in four of the seven and 0.35 in whisper-medium.en, whisper-large-v3-turbo and whisper-large-v2. Whisper-large-v3's row is repeated for comparison, and the English-only line was not run at $s = 1.9$. Values are medians with 95 percent order-statistic intervals, corrected, over the 44 items; the listeners' direction is negative, so a row further below zero is a checkpoint losing more belief to the slowed context.

Table D1: The seven extension checkpoints on the full battery (Sec. 3.4), whisper-large-v3 repeated from the nine.

Checkpoint	Params	b_{rate} (s: 1.0–1.7)	b_{rate} (s: 0.6–1.7)	Equivalent word reduction
whisper-large (v1)	1550M	−1.62 [−2.35, −0.69]	−3.39 [−4.16, −2.21]	−0.18 [−0.37, −0.11]
whisper-large-v2	1550M	−1.16 [−2.49, −0.14]	−2.85 [−3.87, −1.45]	−0.14 [−0.42, −0.02]
whisper-large-v3 (one of the nine)	1550M	−0.67 [−2.03, +0.99]	+0.18 [−1.08, +2.55]	−0.12 [−0.22, +0.28]
whisper-large-v3-turbo	809M	−0.14 [−1.96, +1.42]	−1.10 [−2.27, +2.16]	−0.01 [−0.26, +0.19]
whisper-tiny.en	39M	−3.80 [−4.65, −2.23]	−3.82 [−5.57, −2.58]	−0.41 [−0.60, −0.32]
whisper-base.en	74M	−1.93 [−3.24, −0.71]	−3.28 [−4.19, −2.27]	−0.29 [−0.42, −0.11]
whisper-small.en	244M	−1.89 [−2.81, −1.29]	−3.23 [−4.72, −2.29]	−0.32 [−0.40, −0.22]
whisper-medium.en	769M	−1.33 [−2.67, −0.22]	−3.96 [−4.50, −2.04]	−0.14 [−0.33, −0.04]

What the rows say for Sec. 3.4. At 1.5B the shift is significant in whisper-large (v1) and large-v2 and not in large-v3; in equivalent word reduction all three sit below every primary size, though not in nats for v1. Whisper-large (v1) documents no augmentation and is significant like the augmented large-v2, so the decline is not an augmentation artifact. Large-v3 is a bound rather than a measurement: its shift is not significant while its unrelated-word control is (Table D3). Large-v3-turbo uses large-v3's encoder and is not significant either, while both 1.5B recipes beside it are, so the null is tied to the v3 encoder rather than to size. The English-only line, outside the multilingual training mix, declines in nats from tiny.en to medium.en with all four checkpoints significant, and in equivalent word reduction it ends lower without being monotone. Their with-word preference rate drops are small as well: 2 to 7 points at $s = 1.9$ for the 1.5B checkpoints and 7 to 23 points at $s = 1.7$ for the English-only line (Table D2).

Table D2: The with-word preference rate (Sec. 3.2) of the same checkpoints as point values: the fall in $P(\Delta > 0)$ from $s = 1.0$ to the named scale, in points, with the fraction of the human 46 beside it; the matched scale $s = 1.9$ first.

Checkpoint	Drop 1.0 → 1.9 (pts, point value)	Fraction at 1.9	Drop 1.0 → 1.7 (pts, point value)	Fraction at 1.7
whisper-large (v1)	2	0.05	2	0.05
whisper-large-v2	2	0.05	2	0.05
whisper-large-v3 (one of the nine)	7	0.15	2	0.05
whisper-large-v3-turbo	5	0.10	2	0.05
whisper-tiny.en	not run	not run	23	0.49
whisper-base.en	not run	not run	11	0.25
whisper-small.en	not run	not run	11	0.25
whisper-medium.en	not run	not run	7	0.15

Large-v3's clamped slope. Whisper-large-v3's duration-clamped slope over s: 0.6–1.7 is -3.08 [-4.17, -2.26]. It is significant, but it confounds the amount of pad with the context scale, and under the matched-support construction of Table B2 it spans zero (-0.66 [-1.12, +1.46]), so it is not counted as a shift.

Their controls (Table D3). Four of the eight rows keep the specificity margin of the four primary Whisper checkpoints. In whisper-tiny.en, base.en, medium.en and large-v2 the target's b_{rate} (s: 0.6–1.7) runs 5.5 to 12.1 times the unrelated-word control, inside the four primary Whisper checkpoints' band of 5 to 28, and three of those four control intervals span zero. Whisper-small.en and whisper-large (v1) do not: their unrelated-word controls are significant in the target's own direction at roughly two fifths of its magnitude (ratios 2.5 and 2.4). For these two a shift specific to the frozen word is not separable from a response to the re-timed context at this n. In whisper-large-v3 and large-v3-turbo the control is significant and positive while the target is not, so their near-null results are undetermined rather than an absence. Sec. 3.1's specificity result rests on Table B3 and is untouched.

Table D3: Control slopes per extension checkpoint, whisper-large-v3 repeated, over s: 0.6–1.7 (median [95 percent order-statistic CI, corrected]; nats per $\ln s$); the ratio is the target's b_{rate} (s: 0.6–1.7) over the unrelated-word control, in absolute value, as in Sec. 3.1.

Checkpoint	Insertion-alone $b_{\text{rate}}(\Delta_{\text{ins}})$	Unrelated-word $b_{\text{rate}}(\Delta_{\text{ctrl}})$	$\lvert b_{\text{rate}} \rvert$ ratio
whisper-large (v1)	-2.07 [-3.07, -1.54]	-1.41 [-1.72, -0.55]	2.4
whisper-large-v2	-1.84 [-2.79, -0.52]	-0.47 [-1.02, -0.07]	6.0
whisper-large-v3 (one of the nine)	+0.13 [-0.88, +1.38]	+2.16 [+0.86, +2.88]	0.1
whisper-large-v3-turbo	-1.02 [-2.73, +0.47]	+1.17 [+0.37, +2.20]	0.9
whisper-tiny.en	-0.70 [-1.79, +0.05]	+0.33 [-0.40, +1.09]	11.6
whisper-base.en	-0.49 [-1.40, +0.25]	-0.60 [-1.00, +0.32]	5.5
whisper-small.en	-2.55 [-3.49, -1.84]	-1.30 [-1.93, -0.37]	2.5
whisper-medium.en	-1.95 [-3.49, -0.60]	-0.33 [-1.44, +0.01]	12.1

E THE LISTENER-CORRESPONDENCE BATTERY

Three signatures ask how much of the listeners' computation the shift shares (Sec. 3.2). This appendix gives each signature's construction and its values per model, then the with-word preference rate tables (E1 and E2) and the transcript report rate table (E3); the four signature panels are in Fig. E1.

Relative rate or absolute duration (Fig. E1a, b). Listeners judge the word against the rate of the speech around it, not by its duration in seconds. Scaling the proximal window (the word and its two neighbors) and the context by the same factor (0.6 or 1.7) separates the two accounts. The word's absolute duration changes while its duration relative to the context does not. A system

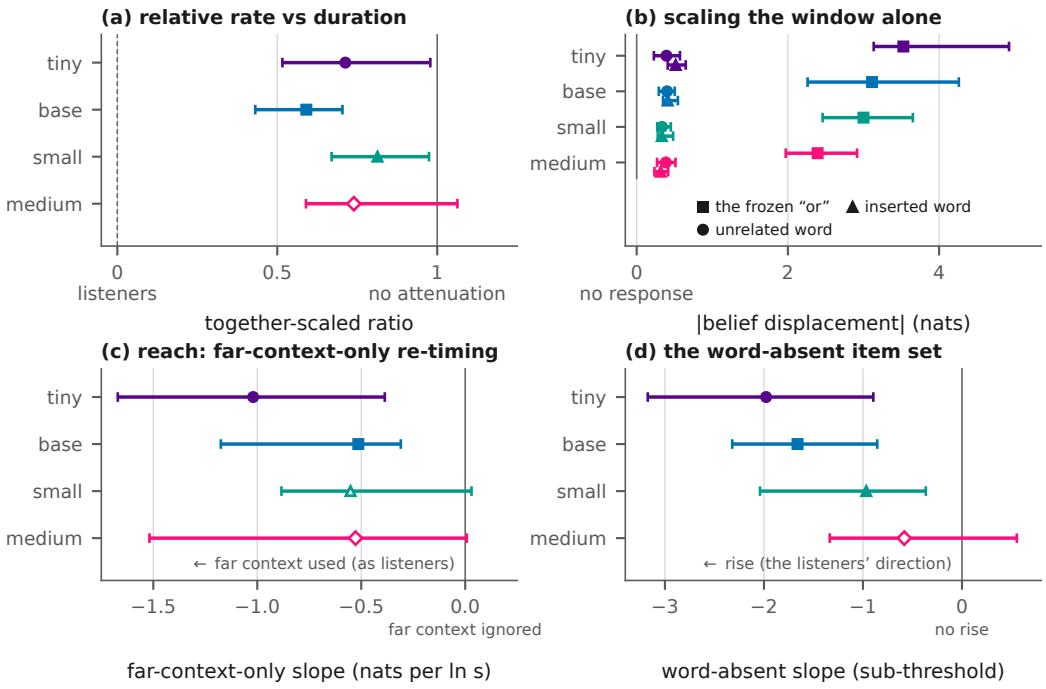


Figure E1: **Three signatures of the listeners' computation (Sec. 3.2). (a)** The together-scaled ratio per model: 1 is no attenuation (the absolute-duration account), 0 marks the listeners' near-complete restoration. Filled where the interval's upper bound sits below 1, which whisper-medium's interval does not meet. **(b)** The same contrast in the model's own units: the belief displacement produced by scaling the proximal window alone, for the frozen word (squares) and the two control candidates, the unrelated word (circles) and the inserted word (triangles). Only the frozen word shows a displacement for the context to cancel, which is why the ratio in (a) is measured on the target alone. **(c)** The belief slope when only speech beyond 2.0 s from the proximal window is re-timed (0 = the far context ignored). **(d)** The belief slope on the word-absent item set, sub-threshold throughout ($\Delta > 0$ in 0 of 384 conditions, where listeners' reports rise eightfold from 3 percent). Corrected order-statistic 95 percent intervals throughout; filled where the interval excludes the panel's reference line.

tracking absolute duration then responds as it did when the window alone was scaled, and a system tracking relative rate returns to baseline, the near-complete restoration listeners show (Dilley & Pitt, 2010). The together-scaled ratio divides the belief displacement when window and context are scaled together by the displacement when the window alone is scaled: 1 for absolute duration, near 0 for the listeners' relative-rate account. Both displacements are measured at $r = 0.35$, the level nearest r^* at which these conditions were built. We call a model's result rate normalization only where the ratio's interval has its upper bound below 1. The ratio is 0.59 to 0.81 in the four Whisper models, three of the four qualifying (Table B2; Fig. E1a): the attenuation is partial everywhere. A two-point ratio understates relative coding. A joint fit of Δ on $\ln r$ and $\ln s$ over the intact grid at $r > 0$ gives a context-to-reduction coefficient ratio of -1.07 [-1.33, -0.83] in tiny and -0.97 [-1.21, -0.71] in base (-0.64 and -0.74 in small and medium), where -1 is pure relative coding. On the full grid, then, the two smallest models weight the context's rate nearly as strongly as the word's own duration. The signature is measured on the frozen word alone (Fig. E1b). Scaling the window alone displaces belief in the frozen word by 2.39 to 3.53 nats but belief in the two control candidates by only 0.32 to 0.52, so no control shows a comparable displacement for the context to cancel. The controls' own ratio (2.1 to 3.7) is a quotient on a near-floor denominator, not an amplified response.

The far context (Fig. E1c). In listeners the effect is distal: the re-timed speech sat several phonemes beyond the word (Dilley & Pitt, 2010), and listeners track context rate over seconds (Baese-Berk et al., 2014). In the far-context condition (items with at least 4.0 s of speech before the window, $n = 25$) only the speech beyond 2.0 s from the proximal window is scaled, at two context scales (0.6 and 1.7) and one reduction level ($r = 0.35$), against a near-context condition. It is a two-point contrast with no duration-matched subtraction of its own, so we call the effect distal only where

the far context alone moves belief. Re-timing only the far speech still moves belief in tiny (−1.02 [−1.67, −0.39]) and base (−0.51 [−1.18, −0.31]), while in small and medium the interval spans zero. The pad slope at the same level spans zero in every Whisper checkpoint (medians −0.19 to +0.42), so duration does not produce the effect. The effect is distal in the two smallest models, and we claim no more.

The appearing word (Fig. E1d). In listeners, speeding the context around a region produced without the word raises false reports of it. The word-absent item set ($n = 24$ pairs of adjacent words of at least four letters, function words excluded by a stoplist, with "or" scored at their boundary; context scale the only manipulated axis, with the duration and artifact controls of the primary item set) tests that direction on its own items. The insertion was not screened for text plausibility, and many of the insertions are ungrammatical. Belief in an unspoken "or" rises as the context speeds in three of four models (slopes −1.98 [−3.17, −0.90], −1.66 [−2.32, −0.86] and −0.97 [−2.04, −0.37]; medium not significant, [−1.34, +0.55]). It stays sub-threshold everywhere: Δ crosses zero in 0 of 384 conditions, and the movement is a small fraction of the roughly 15-nat gap to the with-word transcript. In listeners the word is heard; in these models belief moves toward it and never arrives. With the word cut out entirely ($r = 0$) the models do prefer the with-word transcript on 48 to 66 percent of items at $s = 1$, falling to 25 to 52 percent at $s = 1.7$, a preference-level effect on a splice that we do not interpret.

Table E1: The with-word preference rate $P(\Delta > 0)$, the fraction of items whose belief favors the with-word transcript at r^*, for the attention-decoder checkpoints (44 items); the human anchor is Dilley and Pitt's 46-point drop (0.79 to 0.33) at expansion 1.9. The four primary Whisper checkpoints are given with a 95 percent interval for a paired difference of proportions (method 10 of Newcombe, 1998) and the counts of items losing and gaining the word, of 44. The other rows are point values (a 2-point drop is one item). r^* is 0.35 for large-v2 and large-v3-turbo and 0.25 otherwise (Appendix D); the English-only line is in Appendix D; Qwen2-Audio was not run at $s = 1.9$.

Model	$P(\Delta > 0)$ at s = 1	Drop 1.0 → 1.9 (points)	Fraction at 1.9	Items losing / gaining at 1.9
whisper-tiny	0.84	18 [+6, +31]	0.40 [0.13, 0.67]	8 / 0
whisper-base	0.89	25 [+10, +40]	0.54 [0.21, 0.86]	12 / 1
whisper-small	0.89	14 [+2, +26]	0.30 [0.05, 0.56]	6 / 0
whisper-medium	0.91	2 [−10, +15]	0.05 [−0.21, 0.32]	3 / 2
whisper-large-v3	0.95	7	0.15	not computed
whisper-large (v1)	0.98	2	0.05	not computed
whisper-large-v2	0.98	2	0.05	not computed
whisper-large-v3-turbo	0.98	5	0.10	not computed
Qwen2-Audio-7B	0.93	not run	not run	not computed

Table E2: The same drop at $s = 1.7$, point values with the fraction of the human 46 beside them; the four primary Whisper checkpoints are given with the Newcombe interval as in Table E1, r^* as in Table E1, and a 2-point drop is one item.

Model	Drop 1.0 → 1.7 (points)	Fraction at 1.7
whisper-tiny	16 [+4, +28]	0.35 [0.09, 0.61]
whisper-base	23 [+8, +37]	0.49 [0.17, 0.81]
whisper-small	11 [+1, +23]	0.25 [0.02, 0.50]
whisper-medium	7 [−5, +19]	0.15 [−0.10, 0.42]
whisper-large-v3	2	0.05
whisper-large (v1)	2	0.05
whisper-large-v2	2	0.05
whisper-large-v3-turbo	2	0.05
Qwen2-Audio-7B	2	0.05

Table E3: The transcript report rate at $s = 1.9$: greedy decodes under the presence rule of Appendix B, the drop from $s = 1$ in points, its fraction of the listeners' 46-point drop, and the counts of items losing and gaining the word (Sec. 3.2); beam search agrees within one item per model; large-v3 and the extension checkpoints were not decoded.

Model	Transcript drop at 1.9 (points)	Fraction of the human 46	Items losing / gaining the word at 1.9
whisper-tiny	27 [+13, +41]	0.59 [0.28, 0.89]	12 / 0
whisper-base	32 [+17, +46]	0.69 [0.36, 0.99]	14 / 0
whisper-small	14 [+1, +27]	0.30 [0.01, 0.58]	7 / 1
whisper-medium	9 [−1, +21]	0.20 [−0.02, 0.46]	4 / 0
whisper-large-v3	not decoded	not decoded	not decoded
whisper-large (v1)	not decoded	not decoded	not decoded
whisper-large-v2	not decoded	not decoded	not decoded
whisper-large-v3-turbo	not decoded	not decoded	not decoded
Qwen2-Audio-7B	not decoded	not decoded	not decoded

F THE WORD-GENERALITY ITEM SET PER WORD

The word-generality item set of Sec. 3.3 tests three further non-rhotic function words and the near-homophone "are"; this appendix gives the per-word values behind Fig. F1 (Tables F1 to F4).

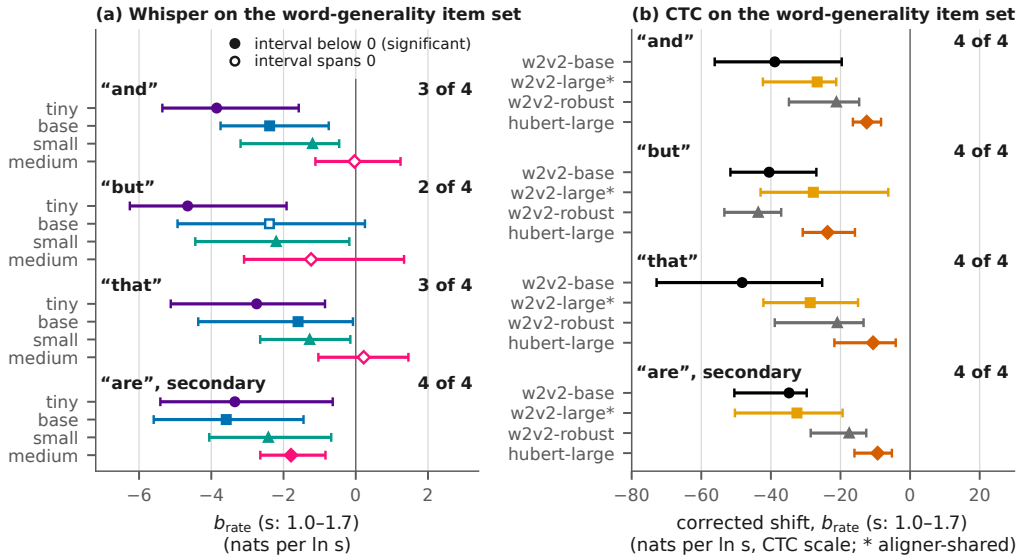


Figure F1: **The word-generality item sets (Sec. 3.3). (a)** b_{rate} (s: 1.0–1.7) per word and model on the word-generality item sets, whisper tiny to medium within each word. Filled where the interval excludes zero; "are" flagged as the secondary rhotic near-homophone. **(b)** The CTC family on the same item sets after the item-level controls correction (Sec. 3.3), per word and checkpoint; every word-model pair is significant; the asterisk marks the aligner-shared model. Per-word values in the tables of this appendix.

Every word-model pair of Table F1 passes the measurement validity criteria (Sec. 2.5), with $r^* = 0.25$ in all sixteen. "Are" is the secondary rhotic near-homophone of reduced "or" and adds no lexical-generality evidence.

The controls per word (Table F2). The unrelated-word candidate does not show the shift. The unrelated-word control spans zero in fifteen of the sixteen pairs, the exception "but" in whisper-tiny (+0.69 [+0.04, +2.88]), so the per-word shifts are as specific to the frozen word as the primary item

Table F1: b_{rate} (s: 1.0–1.7) per word and model (median [95 percent order-statistic CI, corrected]; nats per $\ln s$); the listeners' direction is negative, and a more negative entry is a model losing more belief in that word as its context slows (Sec. 2.5).

Word	and	but	that	are (secondary)
n	24	24	24	19
Span (ms)	185	185	215	145
whisper-tiny	−3.85 [−5.36, −1.58]	−4.66 [−6.26, −1.92]	−2.74 [−5.12, −0.85]	−3.35 [−5.41, −0.64]
whisper-base	−2.38 [−3.74, −0.75]	−2.39 [−4.93, +0.25]	−1.60 [−4.36, −0.08]	−3.59 [−5.59, −1.45]
whisper-small	−1.20 [−3.19, −0.46]	−2.20 [−4.44, −0.18]	−1.28 [−2.65, −0.15]	−2.42 [−4.06, −0.68]
whisper-medium	−0.03 [−1.12, +1.24]	−1.24 [−3.09, +1.34]	+0.22 [−1.04, +1.46]	−1.79 [−2.64, −0.84]
Supported	3 of 4	2 of 4	3 of 4	4 of 4

set's. The insertion control repeats the primary item set's size pattern, small in whisper-tiny (−0.04 to −0.59 across the four words) and larger in every pair of the three larger models (−1.23 to −2.90, eleven of the twelve intervals below zero). The finding of Sec. 3.1, that the larger models' relative belief in a second "or" also moves with the context, therefore extends to these words and to whisper-base. The remaining guards behave as on the primary item set: the duration-clamped twin agrees in ten of the twelve supported pairs, and the artifact control's share of the shift stays at or below 14 percent in every supported pair (per-pair values in the released score tables).

Table F2: Control slopes per word and model over s: 0.6–1.7 (median [95 percent order-statistic CI, corrected]; nats per $\ln s$): Δ_{ctrl} is the unrelated-word control $b_{\mathrm{rate}}(\Delta_{\mathrm{ctrl}})$ and Δ_{ins} the insertion control $b_{\mathrm{rate}}(\Delta_{\mathrm{ins}})$ of Sec. 2.2.

Word, control	whisper-tiny	whisper-base	whisper-small	whisper-medium
and, Δ_{ctrl}	+0.79 [−0.28, +1.39]	+0.07 [−0.88, +0.50]	+0.94 [−0.85, +2.14]	+0.35 [−1.99, +1.44]
and, Δ_{ins}	−0.59 [−2.97, +0.93]	−1.92 [−3.52, −0.62]	−2.31 [−3.52, +0.77]	−2.47 [−4.73, −0.21]
but, Δ_{ctrl}	+0.69 [+0.04, +2.88]	−0.30 [−1.65, +0.94]	−0.39 [−1.70, +1.34]	−0.76 [−2.83, +0.68]
but, Δ_{ins}	−0.04 [−2.84, +0.84]	−2.52 [−5.03, −1.84]	−2.38 [−3.26, −1.22]	−2.90 [−4.32, −0.84]
that, Δ_{ctrl}	+0.64 [−1.00, +1.28]	+0.02 [−0.69, +0.90]	+0.13 [−0.87, +1.02]	+0.21 [−0.82, +1.47]
that, Δ_{ins}	−0.41 [−1.69, −0.02]	−1.23 [−2.40, −0.17]	−1.95 [−2.96, −1.23]	−2.46 [−3.44, −1.95]
are (secondary), Δ_{ctrl}	+0.78 [−0.65, +2.00]	−0.89 [−1.83, +0.60]	−0.76 [−2.74, +0.81]	−0.37 [−2.35, +0.59]
are (secondary), Δ_{ins}	−0.41 [−2.51, +0.49]	−1.91 [−3.77, −1.12]	−1.67 [−5.06, −1.06]	−2.24 [−6.07, −1.30]

The CTC family (Sec. 3.3; Table F3). All sixteen word-model pairs are significant under the item-level controls correction of Appendix G. Identity-condition word error is 0.100 to 0.107, scored with the greedy decoding of Appendix B and so on the same kind of floor (not re-scored on this item set). The share of the target's item variance the controls explain (R^2) is 0.00 to 0.14 in thirteen of sixteen pairs. The three pairs where they covary, "but" in wav2vec2-robust (R^2 0.64) and hubert-large (0.59) and the secondary "are" in wav2vec2-base (0.41), are significant after the correction as well. The same correction over s: 0.6–1.7 is significant in fourteen of sixteen; the two exceptions, "that" in the aligner-shared model (upper bound +0.38) and "and" in hubert-large (+1.69), span zero.

Table F3: The corrected shift for the CTC family: the target's b_{rate} (s: 1.0–1.7) with the two controls' fitted contribution removed across the word's items (median [95 percent order-statistic CI, corrected]), significant where the interval is below zero.

Word	wav2vec2-base	wav2vec2-large (aligner)	wav2vec2-robust	hubert-large
and	−38.9 [−56.1, −19.6]	−26.7 [−42.3, −21.2]	−21.2 [−34.8, −14.6]	−12.5 [−16.4, −8.3]
but	−40.5 [−51.6, −26.9]	−27.8 [−43.0, −6.3]	−43.6 [−53.4, −37.0]	−23.7 [−30.9, −15.8]
that	−48.3 [−72.8, −25.3]	−28.7 [−42.2, −15.0]	−20.9 [−38.9, −13.4]	−10.6 [−21.8, −4.1]
are (secondary)	−34.8 [−50.5, −29.7]	−32.5 [−50.3, −19.4]	−17.5 [−28.6, −12.6]	−9.3 [−16.0, −5.2]

Table F4: Equivalent word reduction E per word and model (Sec. 2.5; median [95 percent order-statistic CI, corrected]): the fraction of the word whose removal would cost the model the same belief as one e-fold of context slowing. Sec. 3.3's "a tenth to three tenths" is the range of the non-rhotic pairs whose b_{rate} interval lies below zero (Table F1); "are" is the secondary rhotic near-homophone.

Word	whisper-tiny	whisper-base	whisper-small	whisper-medium
and	−0.25 [−0.47, −0.15]	−0.22 [−0.31, −0.07]	−0.13 [−0.27, −0.02]	0.00 [−0.15, +0.17]
but	−0.30 [−0.56, −0.12]	−0.18 [−0.34, +0.02]	−0.14 [−0.32, −0.02]	−0.10 [−0.23, +0.08]
that	−0.19 [−0.44, −0.09]	−0.15 [−0.38, −0.01]	−0.11 [−0.22, −0.01]	+0.02 [−0.16, +0.15]
are (secondary)	−0.61 [−0.93, −0.10]	−0.81 [−1.03, −0.26]	−0.35 [−0.72, −0.11]	−0.37 [−0.66, −0.14]

G THE MEASURE ON CTC SCORES

A CTC score responds to a re-timed context in its own right, so applying the measure to this family took three additions: the agreement rule, a score restriction that locates the response, and the item-level controls correction that Sec. 3.3 reports. This appendix gives each with its values, the signatures of Appendix E under the same correction, and two accounts of why the families differ.

The agreement rule. A CTC score sums the probability of every alignment of the transcript's letters to the frames over the whole utterance. Re-timing the context therefore moves the score of every candidate, including the two controls, which differ from the reference only outside the frozen span. On the pilot set this showed as an insertion control that moved with s, and the family's rule comes from there. A direction is claimed only when the raw shift and the insertion-subtracted shift (Δ with Δ_{ins} subtracted, the solid and dotted series of Fig. G1a) both have intervals below zero; where they disagree the family is mechanics-dominated and its direction undetermined. Whisper's insertion control spanned zero on the pilot set, so that family is assessed on the raw shift; since that premise holds only in part on the primary item set (Sec. 3.1), both control shifts are printed for every model (Table B3).

What the controls show (Fig. G1). The family's shift is large in its own units, medians −10.2 to −18.1 over s: 0.6–1.7 and −11.9 to −15.7 over s: 1.0–1.7 with every interval below zero. The question is how much of it is a percept of the target. Under the agreement rule none of it can be claimed: the insertion control responds more strongly than the target itself over s: 0.6–1.7 (insertion-subtracted shifts +8.0 to +21.9, intervals above zero; the insertion row sits beyond the target's row in Fig. G1b), and the unrelated-word control responds in the opposite direction to the target (+12.2 to +36.7). The second observation rules out the simplest mechanical account, a uniform shift of every word's score with the context, since a uniform response would move both deletion candidates the same way. What remains in the raw result is a score in which each candidate's own alignment to the re-timed frames dominates any percept-sized signal.

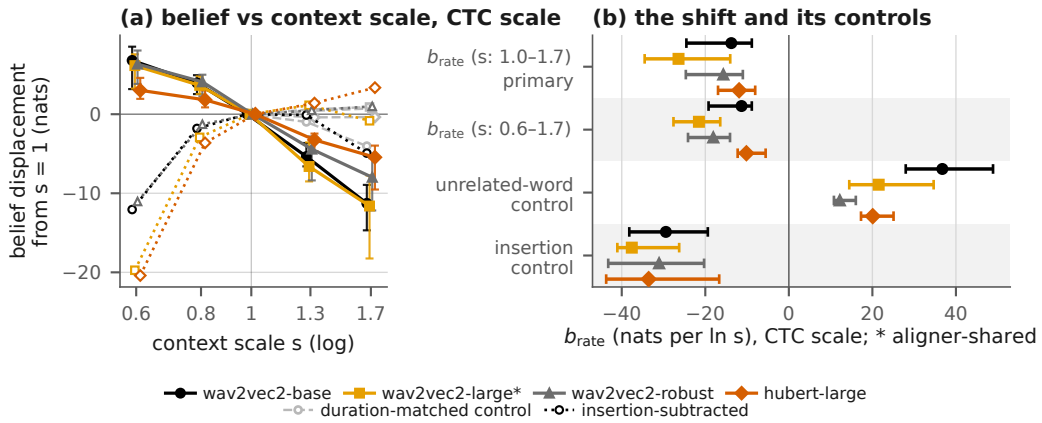


Figure G1: **The CTC family's trajectories and constructions.** **(a)** Median displacement of Δ against context scale on the CTC scale, with the duration-matched control (dashed grey, slow side only) and the insertion-subtracted belief, Δ minus Δ_{ins} (dotted, open). **(b)** Forest as Fig. 2b on the CTC scale: the unrelated-word control runs opposite in sign to the target and the insertion control moves further than the target, which is what leaves the family's raw result undetermined (Sec. 3.3; the item-level controls correction is in this appendix).

The segment-restricted score. The objection that the mechanics live in the context has a direct test: hold the alignment outside the frozen span at its best (Viterbi) path and sum only the span's alignments, the CTC analog of a local lexical decision. The restriction leaves the target quantity unchanged, since the two target candidates differ only inside the span (median change below 0.0001 nats; above 0.1 nat on at most two items per model), and it pins down the two controls, whose candidates differ in the re-timed context. Under the restricted score the target shift stays negative (-10.2 to -18.1 in the three primary CTC checkpoints, -21.5 in the aligner-shared model), the insertion-subtracted shift stays positive (+7.3 to +22.6), and the two estimates still disagree, so the raw result is unchanged. A second "or" inserted inside the span, which would have given the family a within-span control, is degenerate (Appendix B).

The item-level controls correction (Sec. 3.3). The agreement rule treats the controls' size as evidence about the target: if the controls move this much for mechanical reasons, the target's shift may be the same mechanics. The segment restriction locates those mechanics outside the span, where the controls differ from the reference and the target does not. That leaves an item-level question: if one process drove both, the items where the controls respond most would be the items where the target responds most. Across the 44 items we regress the target's b_{rate} on the two controls' b_{rate} and take three quantities: the rank correlations, the share of the target's item variance the controls explain, and the corrected shift, the median of the target's b_{rate} with the controls' fitted contribution removed, with its order-statistic interval. The analysis was added after the primary results and is exploratory on this item set; on the word-generality item set it was fixed before scoring. The controls hold almost no information about the target. On the primary form (s: 1.0–1.7) they explain 1 to 7 percent of its item variance in the three primary CTC checkpoints, with rank correlations of -0.13 to +0.10 (+0.15 with the aligner-shared model). The corrected shift is as large as the raw one or larger, -14.63 [-25.30, -8.28] in wav2vec2-base, -20.34 [-27.42, -14.49] in wav2vec2-robust and -13.71 [-18.15, -7.81] in hubert-large, every interval below zero (the aligner-shared model in Appendix C), and the correction gives the same answer over s: 0.6–1.7 and under the segment-restricted score. The correction behaves as it should where the answer is known. In whisper-tiny and whisper-base, whose controls are small, it returns the target essentially unchanged (residuals -6.50 and -4.13 against -5.43 and -4.68). The one model it flags, whisper-medium, is one of the two Sec. 3.1 flagged for an insertion control that accounts for 60 percent of its full-range shift its partialled shift, the regression intercept with its bootstrap interval, spans zero on the primary range (-0.41 [-1.52, +0.85]) while its corrected shift stays below zero (-0.80 [-1.27, -0.07]). What the test rules out is a shared process that varies across items; an offset common to every item is not testable this way, and the family has no within-span control that could close that gap.

The signatures under the same correction. The three signatures of Appendix E were computed for this family with the primary battery and first reported descriptively, since the raw result is undetermined. Under the same correction, fit per model across items, the word-absent slope is significant in three of four checkpoints (aligner -7.30 [-13.96, -5.43], robust -2.43 [-3.88, -1.73], hubert -6.06 [-9.24, -4.89]; wav2vec2-base spans zero), sub-threshold as everywhere ($\Delta > 0$ in 0 of 384 conditions), and the far-context slope is significant in wav2vec2-base alone (-3.34 [-5.83, -0.81]). The together-scaled ratio is not corrected (the correction is defined for slopes regressed across items, not ratios); the target's ratio runs 0.55 to 0.70, overlapping the Whisper range of 0.59 to 0.81, and the controls' ratio of 4.4 to 18 sits, as in Whisper, on a near-floor denominator (scaling the window alone displaces the controls by 1.0 to 3.6 nats against the target's 16.9 to 22.0).

Two accounts of the family split, not separated here. One is Whisper's timestamp-token objective, a text-to-time prior perfectly confounded with the family split. The other is the production statistics of function-word reduction: a web-scale model of $P(\text{text}, \text{audio})$ learns that talkers shorten predictable function words (Aylett & Turk, 2004; Bell et al., 2009; Jurafsky et al., 2001). The account needs the further regularity that such reduction accompanies fast surrounding speech. The second account is not deflationary, since a system that learned the production regularity listeners may themselves exploit is the syllable-inference story told from the corpus side (Brown et al., 2021). Both predict that the effect should track the target's reducibility, which the four-word set is too small to test.

The with-word preference rate in this family (Sec. 3.2). Its drop from $s = 1.0$ to 1.7 is 36, 30, 23 and 16 points in wav2vec2-base, the aligner-shared model, wav2vec2-robust and hubert-large (0.79, 0.64, 0.49 and 0.35 of the human 46), reported under the same limit as every CTC score; the transcript drops are in Appendix B.

H THE PROBE EXPERIMENT

This appendix gives the probe of Sec. 3.5, its three further checks and the values per checkpoint (Tables H1 and H2).

Construction. Activations are averaged over the proximal frames, excluding 100 ms at each edge. A multinomial logistic probe is trained per model and layer to classify the five context scales, with speaker-grouped cross-validation, and layer selection and evaluation use disjoint speaker-grouped halves. The untrained reference is the same architecture with randomly initialized weights (three seeds), and the probe margin M of Sec. 3.5 subtracts its accuracy at the selected layer. The context's word rate is words per second over the context, from the aligner's boundaries; the natural-rate transfer uses each item's unmanipulated source audio ($n = 23$, the unmanipulated recordings of the held-out evaluation half), and its partial correlation removes the proximal window's own local word rate. Large-v3-turbo is not probed because its encoder is large-v3's.

What each check rules out. The input baseline, the same probe on the front-end output, must sit at chance, or the stimuli themselves contain the class information. The margin must exceed the untrained architecture's accuracy, or what the probe decodes is architecture rather than learning. The position control cross-applies the probe to the duration-matched conditions, where a position probe transfers and a rate probe collapses, and it must collapse, or the representation is position. The natural-rate transfer must hold, or the representation is about our manipulation rather than about rate. The first three checks are met in every probed checkpoint (Table H1, with the position-transfer index and the raw accuracy on the clamped conditions); the fourth is met at large-v3 alone (Table H2), interpreted in Sec. 3.5 as a rising gradient.

Table H1: The probe experiment per model (evaluation half; 1.5B = the three 1.5B Whisper checkpoints, analyzed separately: large-v3 of the nine, large (v1) and large-v2 of the extensions). Trained and untrained accuracy are the probe's five-way accuracy at the selected layer on the trained checkpoint and on the same architecture with random weights; the probe margin M is their speaker-averaged paired difference (chance 0.20); position transfer is a chance-normalized index (1 = full transfer, 0 = chance, negative = below chance); clamped accuracy is raw accuracy on the duration-clamped conditions (chance 0.20).

Model	Trained acc.	Untrained acc.	Probe margin M	Input baseline	Position transfer	Clamped accuracy
whisper-tiny	0.65	0.25	+0.40 [+0.20, +0.58]	0.00	−0.17	0.59
whisper-base	0.70	0.27	+0.43 [+0.26, +0.58]	0.00	−0.10	0.57
whisper-small	0.70	0.28	+0.40 [+0.25, +0.54]	0.00	−0.12	0.63
whisper-medium	0.80	0.30	+0.49 [+0.36, +0.60]	0.00	−0.22	0.66
wav2vec2-base	0.82	0.27	+0.56 [+0.49, +0.63]	+0.01	−0.25	0.70
wav2vec2-robust	0.80	0.30	+0.49 [+0.41, +0.56]	0.00	−0.19	0.76
hubert-large	0.89	0.30	+0.58 [+0.49, +0.66]	+0.01	−0.21	0.86
wav2vec2-large (aligner)	0.80	0.30	+0.50 [+0.46, +0.54]	+0.01	−0.20	0.65
large-v3 (1.5B)	0.73	0.33	+0.39 [+0.23, +0.54]	+0.01	+0.03	0.57
large (v1, 1.5B)	0.81	0.31	+0.50 [+0.39, +0.60]	0.00	−0.22	0.63
large-v2 (1.5B)	0.75	0.31	+0.43 [+0.29, +0.56]	0.00	−0.13	0.68

Table H2: The natural-rate transfer per model (Sec. 3.5; 1.5B as in Table H1): the Spearman correlation between the probe's estimate on each item's unmanipulated recording and its natural word rate, and the partial correlation with the local rate removed (bootstrap intervals, $n = 23$; the partial correlation of the three 1.5B rows is a point value; the ridge penalty is chosen by inner cross-validation, 100 for every checkpoint but whisper-large, 10). Every interval among the eight probed checkpoints below 1.5B (the nine without large-v3) spans zero; whisper-large-v3's is the one that is significant.

Model	Natural-rate ρ	Partial ρ
whisper-tiny	+0.06 [−0.36, +0.46]	+0.09 [−0.35, +0.57]
whisper-base	+0.10 [−0.31, +0.47]	+0.13 [−0.31, +0.58]
whisper-small	+0.18 [−0.24, +0.55]	+0.24 [−0.22, +0.72]
whisper-medium	+0.26 [−0.15, +0.59]	+0.33 [−0.16, +0.75]
wav2vec2-base	+0.07 [−0.33, +0.46]	+0.06 [−0.35, +0.45]
wav2vec2-robust	+0.10 [−0.35, +0.52]	+0.10 [−0.38, +0.53]
hubert-large	+0.01 [−0.40, +0.42]	+0.03 [−0.42, +0.49]
wav2vec2-large (aligner)	+0.02 [−0.39, +0.46]	+0.05 [−0.43, +0.52]
large-v3 (1.5B)	+0.50 [+0.15, +0.74]	+0.56
large (v1, 1.5B)	+0.33 [−0.11, +0.69]	+0.36
large-v2 (1.5B)	+0.40 [−0.0005, +0.69]	+0.45

I THE ERASURE PER MODEL AND ERASURE TYPE

This appendix gives the erasure of Sec. 3.6, what its rank means, its values per model and erasure type (Tables I1 and I2), and the layer sweep and disruption-matched controls that condition them.

The erasure and its rank. At the probe's layer each frame of the frozen span is a vector with as many entries as the layer has channels (384 in whisper-tiny, 1024 in whisper-medium and the large CTC models). The five context scales have five mean vectors there, and the rate is decodable because those means sit apart along some directions. The erasure is fit on the intact conditions at the three intermediate reduction levels ($r = 0.25$, 0.35 and 0.5), over both speaker halves. These are the conditions the probe is trained on: 44 items by three levels by five context scales, 660 span-averaged vectors. The erasure is the set of directions along which the five class means differ, computed after whitening (each direction's variance equalized, with shrinkage); each frame is projected off those directions and mapped back, which leaves every other direction as it was (Ravfogel et al., 2020; Belrose et al., 2023). The rank of the erasure is the number of directions removed. Five means, once centered, span at most four directions. Rank 4 therefore makes the five class means identical and is the smallest erasure that leaves the context's rate not decodable by a linear probe; there is no rank 5 for a five-class variable. Ranks 1 to 3 remove the largest one to three of the four and are partial lesions. As a manipulation check, a linear classifier is refit on the 660 fit vectors after each erasure and scored on those same vectors. It is an in-sample check and not the cross-validated probe of Appendix H: with no erasure it scores 1.00 in each of the eight checkpoints, so its values are not comparable with Table H1. It stays at 1.00 after rank 1, 0.97 to 1.00 after rank 2 and 0.56 to 0.96 after rank 3 across the eight probed checkpoints below 1.5B, and falls to 0.18 to 0.21 after rank 4 (chance 0.20; the rank-4 values are the erased-probe rows of Tables I1 and I2). Only the rank-4 value carries an interpretation: with the class means equal, no linear classifier reaches a lower cross-entropy loss than a constant prediction on the fit vectors, so its accuracy sits at chance. Rank 4 is therefore the primary erasure, and ranks 1 to 3 are analyzed as a dose series (Fig. I1). Two kinds of control remove the same number of directions and differ from the targeted erasure only in which directions go: three random erasures of four directions, and the disruption-matched random erasures described below. A whole-utterance erasure applies the targeted erasure to every frame rather than to the span alone.

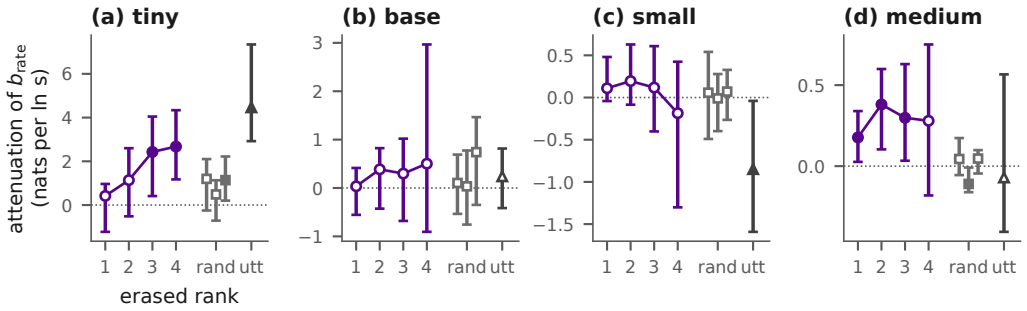


Figure I1: **The Whisper erasure, full dose series.** Per model, dS, the paired change of b_{rate} (s: 1.0–1.7) under the targeted erasure at ranks 1 to 4 (connected; the rank is the number of directions removed, defined in this appendix), under the three random erasures of four directions (rand) and under the whole-utterance erasure (utt), with corrected order-statistic intervals; filled where the interval excludes zero. Of the rank-4 erasures only whisper-tiny's is significant, and one of its random erasures is significant beside it (Table I1).

Whisper (Table I1). dS is the per-item paired change of b_{rate} (s: 1.0–1.7), erased minus clean (median [95 percent order-statistic CI, corrected]); attenuation is positive. Only whisper-tiny's rank-4 erasure is significant, with the reduction axis intact and identity word error at 0.083, and its attenuation grows with rank as the manipulation check's accuracy falls. One of its three random erasures is significant as well, however, so the attenuation is not selective. In base, small and medium the rank-4 erasure spans zero. The random erasures disturb the frames far less than the targeted one (0.006 to 0.030 of its disruption in the Whisper models; about a forty-fourth for the random that

is significant in tiny), which is why the disruption-matched controls below were built. The whole-utterance erasure is a coarse anchor with a limit. Erasing the representation over the whole utterance approaches removing the stimulus property itself, so attenuation there bounds the class of usable representations and does not localize use. In whisper-tiny it also breaks the measurement (identity word error to 0.905), so its attenuation is a bound rather than a dose level. In whisper-base it degrades transcription past 0.05 identity word error, the erasure's transcription margin, with its shift not significant. In whisper-small it is significant in the enlargement direction, the direction of the CTC family's effect, noted without interpretation.

Table I1: The erasure per Whisper model at the probe's layer (hs = hidden-state index): clean = the no-erasure pass; dS = the paired change of b_{rate} under each erasure; db_{word} = the paired change of the reduction-axis slope; WER = identity-condition word error; disruption-matched 1 to 3 = the three matched random erasures of this appendix's second follow-up; the erased-probe accuracy is the in-sample manipulation check on the fit vectors (chance 0.20; 1.00 with no erasure).

Erasure	whisper-tiny (hs 2)	whisper-base (hs 3)	whisper-small (hs 7)	whisper-medium (hs 12)
b_{rate}, clean	−5.43 [−7.94, −3.79]	−4.68 [−5.97, −3.51]	−2.10 [−2.81, −1.06]	−1.45 [−2.15, −0.51]
dS, rank 1	+0.42 [−1.23, +0.96]	+0.03 [−0.56, +0.41]	+0.11 [−0.04, +0.48]	+0.18 [+0.03, +0.34]
rank 2	+1.14 [−0.52, +2.60]	+0.39 [−0.43, +0.82]	+0.19 [−0.09, +0.63]	+0.38 [+0.10, +0.60]
rank 3	+2.43 [+0.41, +4.05]	+0.30 [−0.68, +1.02]	+0.12 [−0.40, +0.61]	+0.30 [+0.03, +0.63]
rank 4 (primary)	+2.68 [+1.17, +4.34]	+0.50 [−0.91, +2.97]	−0.19 [−1.30, +0.42]	+0.28 [−0.18, +0.75]
random 1	+1.20 [−0.25, +2.10]	+0.10 [−0.53, +0.69]	+0.06 [−0.49, +0.54]	+0.04 [−0.06, +0.17]
random 2	+0.49 [−0.71, +1.13]	+0.03 [−0.76, +0.77]	−0.01 [−0.40, +0.28]	−0.11 [−0.16, −0.01]
random 3	+1.14 [+0.20, +2.22]	+0.74 [−0.35, +1.46]	+0.07 [−0.26, +0.33]	+0.05 [−0.05, +0.10]
whole utterance	+4.47 [+2.92, +7.34]	+0.24 [−0.41, +0.82]	−0.85 [−1.59, −0.04]	−0.07 [−0.41, +0.57]
disruption-matched 1	+4.37 [+1.64, +6.33]	+0.37 [−0.12, +1.46]	+0.04 [−0.38, +0.14]	−0.16 [−0.47, +0.42]
disruption-matched 2	+0.95 [−0.91, +2.46]	+0.16 [−0.81, +1.20]	−0.30 [−0.68, +0.39]	−0.36 [−0.66, −0.11]
disruption-matched 3	+0.96 [−0.86, +2.17]	−0.29 [−1.65, +0.55]	−0.11 [−0.36, +0.23]	−0.13 [−0.40, +0.10]
db_{word}, rank 4	−1.15 [−2.86, +0.33]	−0.71 [−1.57, −0.002]	−0.54 [−1.17, +0.29]	−0.27 [−0.43, −0.03]
WER, clean to rank 4	0.050 to 0.083	0.036 to 0.037	0.000 to 0.000	0.000 to 0.000
Erased probe acc.	0.19	0.18	0.21	0.19

The layer sweep. A follow-up swept the same erasure across an encoder-layer grid per checkpoint, added an erasure that removes every swept layer's directions at once, and tried to draw disruption-matched controls from random subspaces (sixty candidates per layer, the three closest kept). Random subspaces cannot match the disruption (the kept randoms sit at 0.008 to 0.090 of the targeted one; four random directions out of several hundred capture almost none of the variance), so they match the targeted erasure in rank only, not in disruption. In Whisper no swept layer and no all-layer erasure yields an attenuation that is both selective and leaves the reduction axis intact, so two alternatives, redundancy across layers or a decoder-side representation, stay open. The erasures that attenuate with quiet randoms also move the reduction axis, and the all-layer erasure moves it in all four models while degrading transcription past the word-error guard in three. In the CTC family the enlargement appears, with every kept random spanning zero, at the probe layers of wav2vec2-base, the aligner-shared model and hubert-large. It also appears at three of hubert-large's neighbors under the corrected form as well (four on dS alone, Fig. I2) and at its all-layer erasure, and at four of thirteen layers in wav2vec2-robust, though not at that model's own probe layer. The all-layer

enlargement does not survive the disruption-matched control below, so the claim is limited to the probe's layer; per-layer values are in the released summary tables.

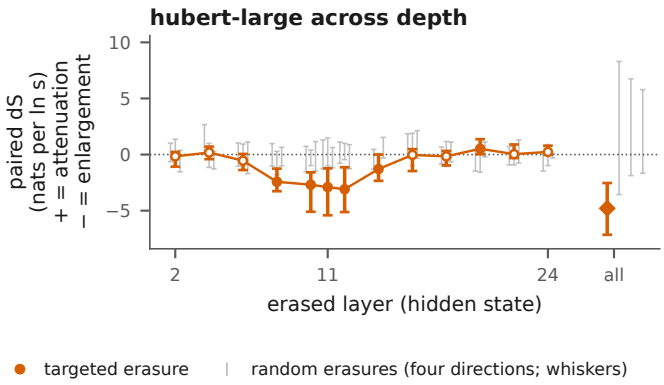


Figure I2: **The layer sweep in hubert-large.** dS against encoder depth: the enlargement sits at mid-depth (filled where the interval excludes zero; grey whiskers are the per-layer random erasures) and appears when every swept layer is erased at once (the diamond, all). Under disruption-matched controls the all-layer erasure is not selective: matched erasures move the shift as far in the opposite direction, which limits the claim to the probe's layer (this appendix).

Disruption-matched controls. A second follow-up built the control the sweep could not sample, at the probe's layer of all eight probed checkpoints below 1.5B (the nine without large-v3). Each is an erasure of four random directions drawn within the leading directions of the frames' own variance (their principal components, after the same whitening), three per model, with the number of components chosen so that the disruption of the fit set matches the targeted erasure's (realized ratios 0.976 to 1.030; 32 to 96 components at the probe layers, up to 192 in hubert-large's all-layer erasures). At the probe's layer every such erasure leaves the in-sample manipulation check at 1.00, its no-erasure value: the erasures match the damage without equalizing the class means. Their four directions were not made orthogonal to the rate directions. Under these controls whisper-tiny's attenuation is reproduced by a matched erasure (+4.37 [+1.64, +6.33] against the targeted +2.68), the result of Sec. 3.6. Base, small, medium and wav2vec2-robust have no targeted effect to separate; two matched erasures in wav2vec2-robust and one in whisper-medium are nonetheless significant (Fig. 5), consistent with generic disruption at that layer. The CTC enlargement holds in all three significant checkpoints, every matched interval spanning zero with word error unchanged, while hubert-large's all-layer enlargement does not: the matched all-layer erasures move the shift as far in the opposite direction (+7.2 to +9.1, attenuation), with collapsed reduction axes.

The CTC family (Table I2). A component whose removal increases rather than removes a behavior has precedent in the interpretability literature (McDougall et al., 2024; Wang et al., 2023). The family split is what the two architectures would predict. The erasure alters the span's frames only, and Whisper's decoder can draw on any encoder frame when it emits a word, so the propagated copy of the rate need not be the copy the decision consumes; a CTC model's probability for the span depends on the span's own frames, the frames the erasure altered. The enlargement is selective in wav2vec2-base, the aligner-shared model and hubert-large. dS is below zero at rank 4 with the corrected form ($d\mathrm{Resid}$) matching, every random erasure spans zero, identity-condition word error is unchanged in every erasure (0.120 to 0.125, the decoding floor of Appendix B), and the reduction-axis slope is unchanged in all three. It is not significant in wav2vec2-robust, whose reduction-axis slope moves under the erasure (-1.29 [-1.62, -1.14]) while its rate shift spans zero. No necessity claim follows from an enlargement; it is stated at its intervals.

Table I2: The same construction at the CTC probe layers (clean, dS, db_{word}, WER and the disruption-matched rows as in Table I1); $d\mathrm{Resid}$ is the paired change of the corrected shift of Appendix G, its control coefficients fit on the no-erasure pass.

Erasure	wav2vec2-base (hs 3)	wav2vec2-large (hs 6, aligner)	wav2vec2-robust (hs 17)	hubert-large (hs 11)
b_{rate}, clean	−13.77 [−24.53, −8.86]	−26.41 [−34.52, −14.03]	−15.69 [−24.66, −11.01]	−11.90 [−16.95, −8.05]
dS, rank 1	−0.72 [−1.79, +0.01]	−0.93 [−2.56, +0.50]	−0.57 [−1.22, −0.26]	−2.20 [−4.18, −1.44]
rank 2	−1.03 [−2.68, −0.30]	−1.89 [−3.05, +0.10]	−0.01 [−0.44, +1.27]	−2.15 [−4.25, −0.61]
rank 3	−1.33 [−2.89, −0.62]	−1.43 [−3.60, −0.18]	−0.41 [−1.08, +0.17]	−2.29 [−5.03, −0.80]
rank 4 (primary)	−2.07 [−3.21, −0.88]	−1.31 [−2.63, −0.72]	−0.49 [−1.29, +0.24]	−2.90 [−5.41, −1.23]
$d\mathrm{Resid}$, rank 4	−2.10 [−3.16, −0.90]	−1.42 [−2.57, −0.71]	−0.48 [−1.31, +0.26]	−2.99 [−5.46, −1.41]
random 1	−0.36 [−1.54, +1.26]	−0.88 [−1.91, +0.71]	+0.34 [−0.13, +1.50]	−0.08 [−0.28, +0.30]
random 2	−1.77 [−3.07, +0.51]	+0.55 [−0.94, +1.12]	+0.12 [−0.44, +0.97]	−0.58 [−1.73, +0.44]
random 3	+0.73 [−0.80, +1.69]	+0.67 [−0.81, +2.84]	+0.39 [−0.23, +0.79]	+0.56 [−0.17, +1.61]
whole utterance	−3.74 [−5.39, −1.76]	−1.48 [−3.44, −0.01]	−0.65 [−1.36, +0.41]	−2.58 [−4.73, −0.96]
disruption-matched 1	+0.13 [−4.80, +7.13]	+0.74 [−2.57, +4.01]	−0.83 [−1.75, +0.40]	−0.07 [−1.50, +0.78]
disruption-matched 2	+3.06 [−3.25, +7.36]	−0.31 [−1.68, +1.34]	+2.52 [+0.96, +3.55]	−0.44 [−1.26, +0.85]
disruption-matched 3	+2.08 [−1.06, +5.67]	+0.64 [−1.47, +5.21]	+0.88 [+0.31, +2.37]	−0.52 [−1.73, +0.73]
db_{word}, rank 4	−0.27 [−1.02, +0.02]	−0.15 [−0.44, +0.07]	−1.29 [−1.62, −1.14]	+0.03 [−0.51, +0.70]
WER, clean to rank 4	0.120 to 0.120	0.125 to 0.125	0.120 to 0.120	0.120 to 0.120
Erased probe acc.	0.21	0.21	0.20	0.20

J THE RELIABILITY-WEIGHTING CHECKS

Two checks test the reliability-weighting account of Sec. 4: if a model weights the context's rate by the quality of the evidence under the word, the use should be largest where that evidence is poorest. The smallest models transcribe this material least accurately (identity word error 0.050 and 0.036 in whisper-tiny and whisper-base against 0.000 in small, medium and large-v3).

Item-level ambiguity (exploratory). Within a checkpoint the account predicts that the items whose reduced word sits nearest the decision boundary shift most with the context. Ambiguity is how close the model's belief sits to zero with the context unchanged, $-|\Delta|$ at $s = 1$ and r^*. The shift is the item's b_{rate} (s: 1.0–1.7) with its sign flipped so that a larger shift is positive. The statistic is the Spearman correlation across the 44 items with a 10,000-draw item bootstrap, predicted positive for the four primary Whisper checkpoints and descriptive elsewhere. The prediction holds in whisper-base (+0.32 [+0.01, +0.58]) and whisper-small (+0.32 [+0.005, +0.60]) and is not significant in whisper-tiny (+0.15 [−0.15, +0.43]) and whisper-medium (+0.02 [−0.32, +0.37]). A secondary form with both variables residualized on the realized span length agrees in base and just spans zero in small, and the CTC checkpoints (+0.03 to +0.26) and the 1.5B checkpoints are not significant. Two of four is graded item-level support, and the check makes no cross-model claim.

Degrading the evidence under the word. The account's own manipulation degrades the evidence under the word and never the context. Speech-shaped noise, a phase-randomized surrogate of the proximal span's own content, is added over the proximal span only, at 10, 0 and −5 dB relative to the span. The noisy span is bit-identical across context scales at a fixed reduction, so the attribution to the surrounding rate is untouched. The prediction, a larger shift under noise (b_{rate} under noise

minus b_{rate} clean below zero), is tested in whisper-medium, large (v1) and large-v2. Tiny, base and small are dose anchors without a prediction, and large-v3 is excluded for want of a clean baseline, its shift being a bound (Appendix D). In whisper-medium the prediction holds at 0 dB (-1.80 [-4.10, -0.55] nats per $\ln s$, added to a clean -1.45): degrading the evidence under the word more than doubles the shift, and at -5 dB the point sits at -1.96 with the interval reaching +0.28. Whisper-small, an anchor, grows as well (-2.97 [-3.99, -0.16] at 0 dB, -1.94 [-2.63, -0.92] at -5 dB). Large (v1) orders with dose in the predicted direction without reaching significance (-0.01, -0.36 and -1.07 at 10, 0 and -5 dB), and large-v2 is neither significant nor ordered (-0.43, +0.57, -0.52). Tiny and base reverse under the heaviest noise (+4.66 [+2.17, +6.74] and +2.95 [+2.41, +4.51] at -5 dB), exactly where their validity criteria weaken (reduction monotonicity 0.60 and 0.70 against the clean +1.00; identity word error up by 0.111 and 0.065). At the top of the use range the noise therefore begins to break the measurement rather than re-weight the cue, and the reversal is not taken as evidence against the account. One significant checkpoint of three is graded support, not a confirmation. Growth where the clean use is small and reversal with degradation where it is already largest is what an evidence-weighting policy with limited headroom would produce, an account after the fact; a finer dose series in the largest checkpoints is the next step.