FROM DECODABILITY TO REUSABLE FEATURES: MANDARIN LEXICAL TONE IN SPEECH SSL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech SSL representations can make a linguistic contrast decodable without organizing it into a stable, reusable feature. Mandarin lexical tone makes this distinction testable because T1–T4 form a closed inventory whose realization varies across speakers and syllables. We fit label-blind TopK sparse autoencoders (SAEs) to Mandarin HuBERT and wav2vec 2.0 states from 3,000 AISHELL-3 utterances. Selected features are tested on unseen initial–final types and across independent dictionary fits. Full-state probes, rank-matched decoder-span and PCA controls, and interventions separately assess information availability, inventory recovery and downstream influence. Tone is decodable at every evaluated layer, and many SAE features transfer and recur, but stable sparse localization is uneven and concentrated in T3. With 64 features, sparse SAE activations retain about 80% of the full four-tone probe gain, whereas the span of the same decoder directions retains 99–102%, matching the full-state endpoint and selected PCA directions. The selected SAE geometry supports the complete inventory, but the TopK activation map exposes it incompletely. SAE and PCA directions also influence a frozen downstream tone readout. The gap between decoder-span and TopK recovery supports a feature-absorption account in which tone evidence is distributed across context-conditioned activations rather than assigned to stable category coordinates. The study separates availability, factor-general localization, category coverage and downstream access.

1 INTRODUCTION

Lexical tone is central to word identity in Mandarin: pitch patterns distinguish words even when the segmental syllable is unchanged (Xu, 1997; Moore & Jongman, 1997). The four categories have canonical high-level, rising, low/dipping and falling contours (Figure 1). A speech model can preserve consonants and vowels while losing a lexical contrast if it fails to retain tone. Linear probes show that Mandarin tone is available in self-supervised speech representations, typically most strongly in intermediate layers and even across pretraining languages (de la Fuente & Jurafsky, 2024; Shen et al., 2024). This evidence establishes decodability, but leaves the representational question unresolved: does a tone category align with a reusable feature that generalizes beyond familiar syllables, or does a classifier recover it from distributed or context-conditioned evidence?

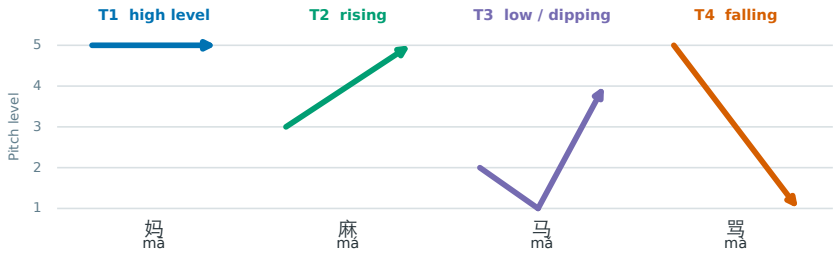


Figure 1: **Canonical citation-form contours of Mandarin lexical tones.** The five horizontal lines show Chao pitch levels from 1 (low) to 5 (high); the targets are T1 55, T2 35, T3 214 and T4 51 (Xu & Wang, 2001). Example characters and tone-marked syllables occur in the analyzed AISHELL-3 subset (Shi et al., 2021). The arrows show canonical targets, not corpus-averaged F0.

Neither standard probes nor candidate-feature searches resolve this distinction. A probe identifies a decision boundary, not the coordinates from which the decision is assembled (Hewitt & Liang,

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.

2019). Sparse autoencoders (SAEs) expose candidate coordinates, but a coordinate that appears selective on discovery data may fail on new segment types or disappear when the dictionary is refit. SAEs trained on the same activations can learn different features across seeds (Paulo & Belrose, 2025); unstable coordinates can occupy reproducible subspaces (Gerasimov et al., 2026); and a learned dictionary need not be a unique or complete set of atomic units (Leask et al., 2025). Existing speech SAEs demonstrate interpretable and steerable features, but mainly through selected concepts rather than complete recovery of a closed linguistic inventory (Parra, 2025; Mariotte et al., 2026; Aparin et al., 2026). Probe accuracy and a compelling feature example are not enough to identify a reusable feature.

The unresolved question needs a setting in which category information is already known to be decodable, while candidate features must generalize across substantial nuisance variation. Mandarin tone provides such a test. T1–T4 form a closed inventory with external labels, so failure to find a sparse coordinate cannot simply be attributed to missing tone information. At the same time, tone realization varies with speaker and syllable composition, allowing a direct test of whether a feature tracks tone rather than vocal range or a familiar tonal syllable. Because the inventory is closed, a compact coordinate set can also be evaluated on whether it recovers all four categories. Mandarin tone connects a speech question about suprasegmental organization with a broader question about what probes, sparse features and interventions establish about a representation.

The resulting evaluation has four linked but distinct parts. We examine Mandarin HuBERT and wav2vec 2.0 using label-blind TopK SAEs. A tone-selective function is called *reusable* when a discovery-selected feature transfers to unseen initial–final types and the same selective function can be rediscovered after independently refitting the SAE. Transfer tests a frozen candidate across segmental contexts; recurrence tests a newly fitted dictionary without requiring the same feature index or direction. The closed inventory supports a separate completeness test of how much of the four-tone system a compact coordinate set retains, while interventions test whether selected directions influence a frozen downstream readout. These analyses distinguish generalization, recurrence, inventory recovery and downstream access rather than treating them as interchangeable evidence for a feature.

This design addresses three research questions. **RQ1.** Do tone-selective SAE features transfer to initial–final types excluded from fitting and selection, and do their selective functions recur after independent refitting? **RQ2.** How completely do sparse SAE activations recover the four-tone inventory, and is any shortfall caused by the selected directions or by the sparse activation map? **RQ3.** Can selected directions influence a downstream tone readout while preserving measured segment margins, and is that access specific to sparse directions?

The three research questions separate claims that are often treated as equivalent. In this study, they lead to different answers. Full-state probes decode tone at every evaluated layer, and tone-selective SAE functions often transfer to unseen initial–final types and recur after refitting. However, only a small subset also retains a consistent selectivity advantage over dense controls. In focused analyses, sparse SAE activations incompletely recover the four-tone readout, while the span of the same decoder directions matches compact PCA and the full-state endpoint. Directions from both bases also influence a frozen downstream tone readout. Tone information is broadly available in the selected geometry, but its allocation to reusable sparse activations is selective rather than guaranteed.

The study makes **three contributions**. *(i)* The paper introduces a factor-generalization protocol that distinguishes transfer of a selected coordinate from functional recurrence after dictionary refitting. *(ii)* A rank-matched decoder-span control localizes incomplete inventory recovery to sparse activation allocation rather than an insufficient set of SAE directions, providing direct evidence for a feature-absorption account in a closed phonological inventory. *(iii)* The paper derives a general evaluation principle for speech SSL: availability, factor-general sparse localization, category coverage and downstream access are distinct claims that require different evidence.

2 Related Work

Interpreting speech SSL representations. wav2vec 2.0 and HuBERT learn contextual speech representations through different self-supervised objectives (Baevski et al., 2020; Hsu et al., 2021). Probes are often used to ask what those representations make predictable, but probe accuracy can also reflect what the classifier learns; control tasks are needed to interpret that accuracy (Hewitt & Liang, 2019). This interpretability question has motivated sparse autoencoders as a more explicit

source of candidate features. Parra identifies HuBERT latents aligned with phonetic events (Parra, 2025), while Mariotte et al. report improved separation of vocal attributes in a singing-technique task (Mariotte et al., 2026). AudioSAE extends this analysis across HuBERT and Whisper, finding recurrent acoustic, semantic, environmental and paralinguistic features and testing selected features through erasure and steering (Aparin et al., 2026). These studies show that sparse speech features can be interpretable and behaviorally relevant. They examine selected concepts or open-ended feature collections, not whether all members of a closed linguistic inventory transfer across nuisance factors and reappear after independent dictionary refits.

Lexical tone in speech representations. Existing tone studies establish that tone information is available in speech SSL states. Layerwise probes place Mandarin suprasegmental information most strongly in intermediate layers (de la Fuente & Jurafsky, 2024), and models pretrained on non-tonal languages also encode lexical tone (Shen et al., 2024). Phone, tone and speaker information occupy largely distinct subspaces across several wav2vec 2.0 models and languages (Gubian et al., 2025). Dense SSL states also retain more lexical-tone information than the discrete units produced by common quantizers (Osakuade & King, 2026). This evidence concerns decodability, subspace geometry and information retained through quantization. It does not identify a factor-general sparse feature. Mandarin makes that distinction demanding. Its four tones have high, rising, low and falling pitch targets (Xu & Wang, 2001), but surface F0 changes with adjacent tones and articulatory constraints (Xu, 1997; Xu & Wang, 2001). Listeners also normalize speaker F0 range (Moore & Jongman, 1997), and low pitch targets, often associated with T3, can co-occur with non-modal voice quality (Kuang, 2017). A coordinate selected without factor holdouts may track vocal range or a familiar tonal syllable rather than tone category. Existing representation studies do not test whether the same tone-selective function transfers across this variation, can be rediscovered after SAE refitting or supports compact recovery of the complete four-tone inventory.

From sparse feature discovery to reusable features. Language-model studies motivate SAEs as a way to replace dense activations with candidate interpretable coordinates (Bricken et al., 2023; Huben et al., 2024). Scaling work improves the reconstruction–sparsity frontier and proposes feature-quality metrics, but reconstruction and sparsity remain proxy objectives (Gao et al., 2025). SAEBench finds that gains on such proxies do not consistently predict disentanglement or practical performance (Karvonen et al., 2025). Reproducibility is also limited: independently initialized SAEs often learn different features (Paulo & Belrose, 2025), although unstable coordinates can occupy reproducible low-rank subspaces (Gerasimov et al., 2026). Stitching and meta-SAE experiments further show that learned dictionaries need not be complete or atomic (Leask et al., 2025). For hierarchical language-model concepts, feature absorption can leave a parent concept without a reliable dedicated latent because its evidence is captured by more specific child features (Chanin et al., 2025). Predictable structure can also remain in SAE reconstruction error (Engels et al., 2024). Finally, gradient-informed SAEs show that a feature's reconstruction role and its downstream influence need not coincide (Olmo et al., 2025). Existing work leaves unresolved whether a feature transfers across nuisance factors, recurs after refitting, recovers a complete inventory and reaches a downstream readout. Mandarin tone supplies a closed, externally labeled inventory for testing these properties together.

3 METHOD

3.1 REPRESENTATIONAL CLAIMS AND EVIDENCE STRUCTURE

We distinguish four representational claims that are often conflated: linear availability, reusable sparse localization, compact inventory recovery and causal influence on a frozen readout. Let i index an aligned syllable token with tone $y_i \in \{\mathrm{T1}, \ldots, \mathrm{T4}\}$, initial–final type c_i, speaker p_i and utterance u_i. Encoder m at layer ℓ yields the common input to all analyses, a pooled state $x_i^{m,\ell} \in \mathbb{R}^d$. A full-state readout provides the availability reference. The primary SAE protocol selects a candidate on discovery data and tests whether its tone-selective function transfers to unseen initial–final types and recurs after independent dictionary refitting. Two focused studies then compare sparse activations with rank-matched SAE and PCA subspaces, and test whether selected directions can alter a later tone readout while preserving measured segment margins. Figure 2 summarizes these roles. Because the refit matrix and focused studies use different partitions, Appendix Table 6 lists the layers, partition protocol and endpoint for each analysis.

A class-balanced four-way linear readout tests whether the complete pooled state predicts tone under the same dictionary-segment partitions used for sparse-feature identification. For every model

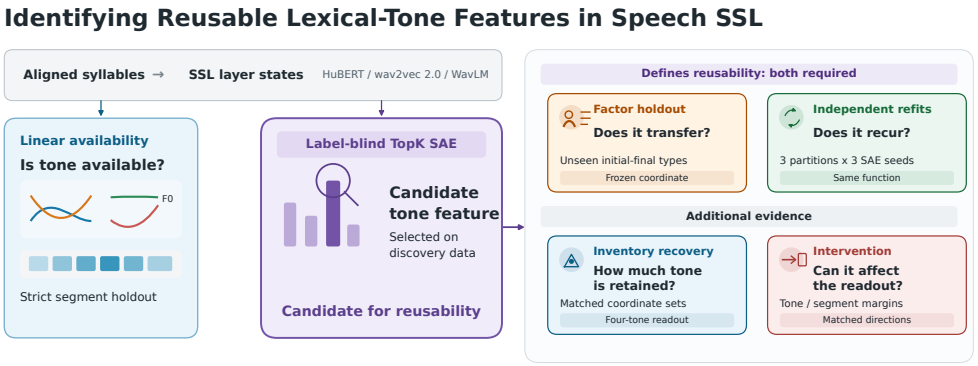


Figure 2: **Experimental framework for identifying reusable lexical-tone features.** Aligned syllables yield SSL states for two parallel starting points: a full-state probe tests linear availability, while a label-blind SAE proposes a candidate tone feature. The right panel tests reusability through transfer to unseen initial–final types and recurrence after independent refits. Rank-matched activation and subspace readouts test inventory recovery; matched interventions test influence on a frozen readout. Refits span layers 4–9 in both encoders; focused analyses use HuBERT L8 and wav2vec 2.0 L4.

and layer, the readout trains on labeled tokens from discovery and validation and is evaluated on the strict test pool. No test token or test initial–final type enters fitting. Three readout seeds are averaged within each partition. Macro recall is the primary endpoint, with macro one-vs-rest AUC as a secondary measure.

The full-state readout provides the availability baseline for the separate question of sparse localization. Appendix D reports the full layerwise matrix.

3.2 Reusable Sparse-Feature Identification

The central protocol first assigns complete initial–final types to four pools: SAE fitting, feature discovery, validation and test (approximately 50/20/15/15%). This assignment occurs before any basis is fitted, so no segment type crosses pools within a partition. Dictionary fitting is label-blind; discovery is the first stage that uses tone labels. Validation calibrates intervention doses, while test labels never select coordinates, fit readouts or set thresholds. Three segment partitions provide repeated corpus splits. A speaker-disjoint protocol is retained as an auxiliary check, while the primary RQ1 endpoint is segment-selection transfer. The segment-selection and utterance-context protocols test the remaining sources of variation separately (Appendices B and C).

Within each partition, states are normalized with statistics from the SAE-fit pool. The SAE computes

$$z = \mathrm{TopK}_{128}(\mathrm{ReLU}(W_{\mathrm{enc}}x + b_{\mathrm{enc}})),$$
$$\hat{x} = W_{\mathrm{dec}}z + b_{\mathrm{dec}}.$$

Each SAE contains 4,096 features and retains at most 128 positive activations per token. It is trained for 30 epochs with three seeds; hyperparameters are selected from label-blind reconstruction and sparsity diagnostics. Three partitions, three seeds, two models and six layers yield 108 independently fitted dictionaries. Appendix A gives the optimizer, normalization and quality-control details.

Tone labels enter only at discovery. For each tone, features are ranked by one-vs-rest AUC within initial–final types that contain both target and non-target tokens. Segment AUCs are averaged with weight n_+n_-, the number of positive–negative pairs; an equal-segment sensitivity removes this weighting. The highest-ranking positive SAE feature is frozen before validation or test evaluation. It *transfers* when its discovery maximum exceeds a within-segment permutation null and its frozen test AUC exceeds 0.5. AUC > 0.55 is reported separately as a descriptive effect-size threshold.

The complete fitting and selection procedure is repeated for three segment partitions and three SAE seeds. A model–layer–tone cell has *recurrent transfer* when all nine runs pass the transfer criterion. Recurrence means that a coordinate with the same tone-selective function can be rediscovered; it does not require the same feature index or decoder direction across dictionaries. All-nine counts serve as a descriptive replication criterion across refits, while effect sizes and segment-clustered intervals quantify uncertainty.

Finally, signed raw dimensions, 256 PCA directions fitted on the SAE pool and 4,096 random unit directions form three dense comparison families. Discovery selects one direction from each family using the same within-segment AUC. Test AUC measures ranking. For purity, each dense direction receives a discovery-calibrated threshold that matches the selected SAE feature's discovery firing rate; the threshold remains frozen on test. A feature has an additional *selectivity advantage* when its test purity exceeds all three discovery-selected dense controls. Selectivity advantage is reported as a separate, stricter SAE-specific diagnostic. Appendix H repeats the comparison at equal test coverage without using test labels to reselect a coordinate.

3.3 Compact Tone-Inventory Recovery

This analysis asks how much of the full-state four-tone readout a small selected coordinate set retains. At HuBERT L8 and wav2vec 2.0 L4, the full-state readout is refitted under a focused protocol with speaker-disjoint basis fitting and initial–final types held out from labeled selection. This protocol differs from the dictionary-segment refits above and evaluates 4,409 test tokens. Its macro recall, R_{full}, is the reference endpoint.

SAE and PCA coordinates are then ranked by discovery AUC and selected in a fixed T1–T4 round robin so that no tone consumes the budget. Budgets are $k \in \{4, 8, 16, 32, 64, 128\}$. A class-balanced readout trains on discovery plus validation coordinates and is evaluated on the shared test tokens. Recovery is reported as retained above-chance macro-recall gain,

$$\frac{R_k - 0.25}{R_{\text{full}} - 0.25}.$$

For each selected SAE set, a rank-matched control orthonormalizes its decoder directions and projects the original normalized SSL state into their span. This preserves the directions proposed by the SAE while removing the encoder amplitudes, ReLU and TopK gate. Comparing sparse activations with this decoder span tests whether incomplete recovery comes from the selected directions or from sparse activation allocation. PCA remains a separate dense information control fitted and selected on the same data split.

A complementary reconstruction control trains readouts on the complete SAE code, reconstruction $\hat{x}$ and residual $x - \hat{x}$; Appendix Table 4 reports these diagnostic results.

3.4 Causal Influence on a Frozen Readout

The focused intervention study uses the same dictionaries and segment-selection protocol as inventory recovery. It tests HuBERT L8 for T1/T3 and wav2vec 2.0 L4 for T1–T4 on 40 held-out tokens per source tone. SAE and PCA directions are unit-normalized and receive the same target-specific coefficient. The $1\times$ coefficient is the selected SAE coordinate's mean positive activation on validation target-tone tokens; doses of $0.5\times$, $1\times$ and $2\times$ are applied without test-label calibration.

Each intervention adds or subtracts the scaled direction at the focused layer and then continues the encoder forward. A separately trained, frozen layer-12 readout supplies the endpoint: the target-tone logit minus the largest non-target logit. SAE ablation removes the selected decoder component, while a value-matched T3 intervention replaces the synthetic coefficient with a validation-derived target value. These operations measure causal influence on the specified frozen tone readout.

Two additional frozen readouts, 24-way for initials and 34-way for finals, measure changes in the gold segment margins on the same tokens. The prespecified segment-equivalence test treats the measured margins as preserved when the speaker-clustered 95% interval for the mean change lies within ± 0.10 SD at the tested dose. Exact resampling procedures and intervention definitions appear in Appendices A and E.

4 Experiments

4.1 Experimental Setup

We use 3,000 AISHELL-3 utterances containing 31,012 aligned syllable intervals from 182 speakers (Shi et al., 2021). The analysis manifest was constructed deterministically by pairing nonempty TextGrid word intervals with the numbered pinyin tokens in the corpus manifest. An utterance was retained only when its audio and TextGrid were available and the word and pinyin counts matched.

Numbered pinyin supplies the tone label, while TextGrid boundaries supply timing; tone is not inferred from the TextGrid character. Initial–final types are derived from the numbered pinyin after removing its tone digit. We manually checked a sample of 500 utterances (16.7%) for syllable identity and gross boundary placement. Neutral tones, <unk> tokens and intervals shorter than 50 ms are excluded from labeled evaluation. This leaves 28,886 T1–T4 tokens: 6,576 T1, 7,253 T2, 5,017 T3 and 10,040 T4. All imported intervals remain available for label-free SAE fitting; the filtering flow is reported in Appendix Table 2. Within-segment ranking and class-balanced readouts reduce frequency effects. A tone-balanced control tests the remaining dependence on class frequency.

The two primary encoders are the Tencent Chinese HuBERT and Chinese wav2vec 2.0 base checkpoints. Both have 12 layers and 768-dimensional states and were pretrained on the same WenetSpeech subset (Zhang et al., 2022; Guo & Liu, 2022). Language, pretraining data, depth and hidden size are matched, so the primary comparison mainly changes the pretraining objective. As a focused cross-architecture control, we also run the same layer-8, three-bin and label-blind SAE protocol on the publicly released WavLM Base checkpoint (Chen et al., 2022). This control uses the same 3,000 utterances and three SAE seeds but one layer and one split seed. It is not part of the 108-dictionary primary matrix. Audio is resampled to 16 kHz.

For each aligned syllable, hidden states are averaged within three equal-duration bins and concatenated. The resulting 2,304-dimensional vector preserves coarse temporal order while providing one observation per syllable. Activation normalization is fitted on the SAE-training pool only and then frozen for validation and test. Cross-layer conclusions use the L4–L9 refitting matrix in both encoders; focused inventory and intervention analyses use HuBERT L8 and wav2vec 2.0 L4. Five-bin and voiced-region controls test sensitivity to temporal pooling.

Focused controls test temporal pooling, tone balance, utterance overlap and SAE width. Additional diagnostics evaluate the contribution of measured F0, duration and voicing cues, temporal order and PSOLA contour manipulation to the recurrent T3 feature (Boersma, 2001). Appendices I, J and L give their definitions and scope.

4.2 Main Results

The main results evaluate linear availability, sparse-feature generalization and compact inventory recovery as separated in Figure 2. Full-state macro recall and one-vs-rest AUC measure availability. Frozen-coordinate AUC on held-out initial–final types measures transfer, and all-nine counts summarize recurrence across three segment partitions and three SAE seeds. Retained above-chance macro recall measures compact inventory recovery. Unless noted otherwise, all values are computed on held-out test tokens; Appendix Table 6 maps each result to its layer and partition protocol.

Linear availability. Full-state readouts remain strong at every evaluated layer under strict dictionary-segment holdout: mean macro recall ranges from 0.739–0.882 for HuBERT and 0.848–0.931 for wav2vec 2.0 (Figure 4a). Every sparse-feature analysis starts from a representation with substantial linearly accessible tone information; the full layerwise matrix is in Appendix D.

Sparse-feature transfer and recurrence. Frozen tone-selective SAE coordinates transfer on most runs and recur across most model–layer–tone cells. Across 432 evaluations, 321 exceed test AUC 0.55; 37 of 48 cells remain above chance in all nine refits, and 20 exceed 0.55 in all nine. All four tones exceed 0.55 in every refit at HuBERT L6 (Figure 3a).

An SAE-specific selectivity advantage over dense directions is less recurrent. With discovery-calibrated thresholds, SAE purity exceeds all three dense controls in 259 evaluations, of which 253 also pass the transfer criterion. Seven cells pass both criteria in every refit: five T3 cells, one T2 cell and one T4 cell (Figure 3b).

The advantage remains limited under equal test coverage or a 0.02 purity margin: five cells pass in all nine refits (Table 1). Sensitivity results and clustered intervals are in Appendices G and H.

Compact inventory recovery. Sparse activation values and decoder directions give different answers for the same selected SAE features. At HuBERT L8 and wav2vec 2.0 L4, the full-state probes reach macro recall 0.894 and 0.859. With 64 tone-selected features, sparse SAE activations reach 0.776 and 0.735, retaining 0.816/0.797 of the full above-chance gain. The rank-64 span of the same decoder directions reaches 0.888 and 0.868, retaining 0.990/1.015 of the gain and matching the full-state endpoint. Selected PCA subspaces reach 0.876 and 0.845 (Figure 4b). Every decoder set has effective rank 64 across the three SAE seeds. The comparison places the recovery gap in how the

Table 1: Frozen-coordinate transfer and SAE-specific selectivity advantage under alternative criteria. Columns count passing evaluations out of 432 and model–layer–tone cells passing in all nine refits out of 48. "All nine" means three dictionary-segment partitions × three SAE seeds. The first two rows test transfer; the final three additionally require an advantage over dense controls.

Frozen-feature evaluation	Passing runs / 432	All-nine cells / 48
Transfer, pair-weighted AUC > 0.5	399	37
Transfer, equal-segment AUC > 0.5	421	41
Transfer + advantage, discovery-calibrated coverage	253	7
Transfer + advantage, equal test coverage	239	5
Transfer + advantage > 0.02, discovery-calibrated coverage	231	5

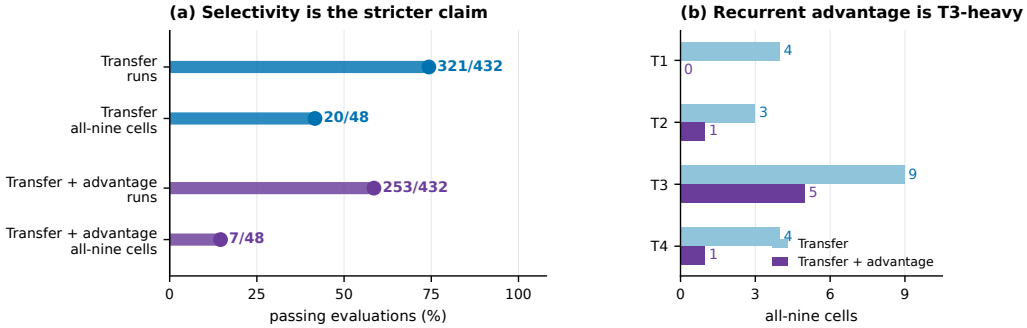


Figure 3: **Transfer is broad; SAE-specific advantage is selective.** (a) Aggregate pass rates for the 432 run-level evaluations and 48 model–layer–tone cells. Transfer uses frozen test AUC > 0.55; the joint criterion additionally requires the discovery null, test AUC > 0.5 and positive SAE-minus-best-dense purity difference. (b) All-nine cells by tone for transfer and the joint criterion. T3 contributes 9 of 20 recurrent transfer cells and 5 of 7 recurrent advantage cells. The complete layer-by-tone matrix and alternative criteria appear in Appendix Table 10 and Table 1.

TopK SAE assigns activation values, not in insufficient tone geometry in the selected directions. Appendix Table 3 reports all budgets.

The complete SAE reconstruction retains 92–93% of the full-state above-chance gain, while the residual remains tone-predictive (macro AUC 0.756/0.758). The decoder-span result places the shortfall in sparse allocation rather than absence of tone information.

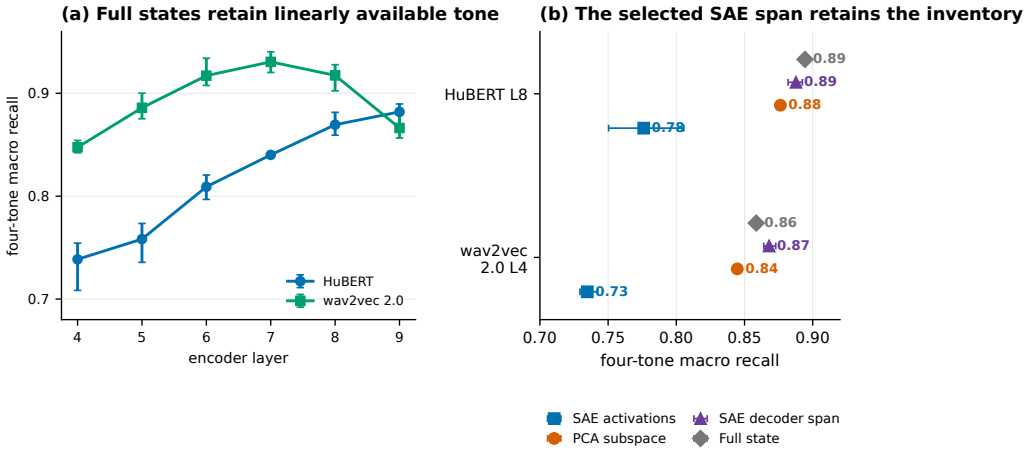


Figure 4: Linear availability and compact inventory recovery. (a) Full-state macro recall across strict layer 4–9 test partitions; points average three partitions and bars show their range. Chance is 0.25. (b) Rank-64 focused-layer results. Sparse SAE activations reach 0.776/0.735, while the linear span of the same selected decoder directions reaches 0.888/0.868 for HuBERT L8/wav2vec 2.0 L4. PCA reaches 0.876/0.845 and full-state recall is 0.894/0.859. Horizontal bars span technical fits.

4.3 Intervention Results

Selected SAE and PCA directions change a frozen downstream tone readout while leaving measured initial and final margins within the prespecified equivalence bounds. Both sparse and dense directions alter the tone readout under matched perturbation norms. At HuBERT L8, ablating the selected

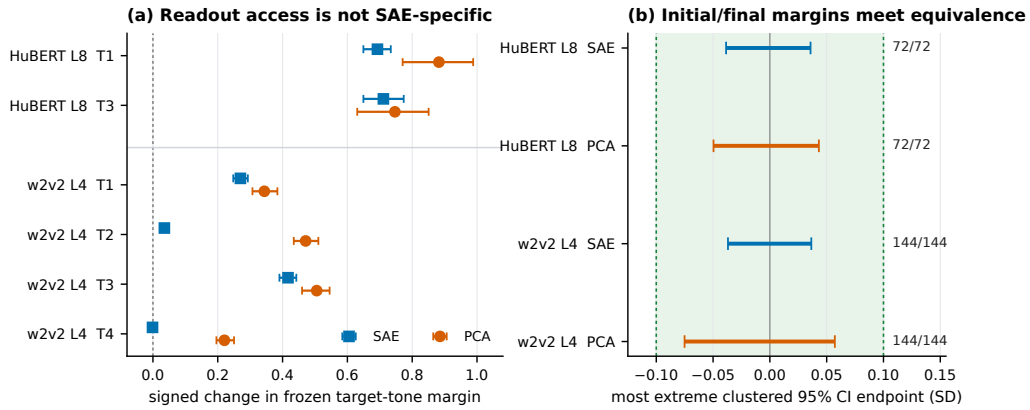


Figure 5: Forward intervention at the validation-calibrated $1\times$ dose, defined as mean positive target-feature activation on validation tokens and applied to unit-normalized SAE and PCA directions. (a) Mean sign-aligned change in the frozen target-tone margin; bars show speaker-clustered 95% bootstrap intervals. (b) Most extreme clustered interval endpoints across initial and final margin tests; green lines mark the ± 0.10 SD equivalence bounds and labels show passed/total cells.

T1 and T3 SAE directions lowers their target-tone margins, while steering along the same directions raises the margins. All initial and final equivalence tests pass for both SAE and PCA steering.

Norm-matched PCA has the larger mean steering effect for every tested tone, especially wav2vec 2.0 T2 and T4, whose SAE effects are near zero (Figure 5). All wav2vec 2.0 segment-equivalence tests also pass. Exact effects and speaker-clustered intervals appear in Appendix E.

The value-matched T3 check also preserves the segment margins; its source-wise effects and sign agreement are reported in Appendix E.

4.4 Robustness and Cross-architecture Control

Pooling, tone-balance, width and acoustic controls preserve the recurrent T3 pattern; the width sweep is reported in Appendix J, and the acoustic diagnostics in Appendix I. The focused WavLM Base layer-8 run preserves broad transfer but changes the allocation across tones: 7 of 12 seed–tone evaluations exceed 0.55, while 4 of 12 show a positive SAE advantage over dense controls. The WavLM result limits the interpretation to a controlled Mandarin case and shows why model and training conditions matter. Exact values appear in Appendix K.

5 Discussion

Across the three research questions, Mandarin tone shows reusable sparse handles embedded in distributed structure. SAE functions transfer and recur, but sparse-specific selectivity is limited to fewer model–layer–tone combinations. The selected SAE decoder span recovers the inventory even when its sparse activations do not, while both SAE and PCA directions influence the frozen readout. Availability, localization, sparse allocation, completeness and downstream access do not coincide.

Reusable sparse functions coexist with distributed tone structure. Transfer and recurrence extend prior tone-decoding results (de la Fuente & Jurafsky, 2024; Shen et al., 2024) from full representations to individual coordinates under factor holdouts. Dense controls show that transfer is broader than sparse-specific advantage. The WavLM control also shows that allocation varies by model: T3 remains strongest, while T1/T4 vary and T2 stays near chance. Sparse functions can be reusable without forming a complete tone inventory, and localization must be tested for each model and training condition (Paulo & Belrose, 2025; Gerasimov et al., 2026; Leask et al., 2025).

The recurrent T3 feature integrates ordered acoustic cues. T3 contributes five of the seven model–layer–tone cells with an all-nine selectivity advantage. Its low pitch target and voice-quality cues may help explain why this ordered multi-cue function remains sparse and recoverable across segmental contexts; the acoustic and width controls are reported in Appendix I.

Incomplete sparse recovery supports a feature-absorption account. The selected decoder directions recover nearly the complete tone inventory when treated as a linear span, but their TopK activations do not. The decoder-span result indicates that the relevant geometry is present; the loss

occurs when tone evidence is assigned to sparse activation coordinates. The complete reconstruction and tone-predictive residual further show that evidence extends beyond the strongest selected features. This allocation is consistent with feature absorption: context-conditioned features divide evidence that a category-level analysis expects to find in stable individual coordinates, extending an account developed for hierarchical language-model concepts to a flat phonological inventory (Chanin et al., 2025). Encoder–decoder mismatch and TopK competition are concrete routes by which this allocation can arise.

Sparse and dense directions share downstream access. Selected SAE and matched PCA directions both affect the frozen tone endpoint while preserving segment margins, showing that downstream access is not specific to sparse coordinates and complementing prior speech SAE steering work (Parra, 2025; Mariotte et al., 2026; Aparin et al., 2026).

From Mandarin tone to speech representation analysis. The Mandarin case suggests a practical change in how speech SSL features should be retrieved. A probe shows that tone information can be decoded, but it does not show that the information is localized in a feature that survives changes in speaker or segmental context. The decoder-span result adds a further distinction. The selected SAE directions can span the tone inventory even when their TopK activations fail to place that information in stable coordinates. Incomplete sparse recovery can reflect activation allocation rather than an absence of tone information. For claims about phones, accent, voice quality, emotion or other speech attributes, the evidence should match the claim. Factor-held-out transfer and independent refits test reuse. Rank-matched dense and decoder-span controls test sparse specificity and coverage. Interventions test downstream access. Mandarin tone provides a controlled demonstration that decodability, localization, completeness and access can diverge. A retrieved feature should be interpreted according to the evidence it passes, not only its probe score or one successful example.

6 Conclusion

Mandarin lexical tone is readily decodable in the primary speech SSL models, but T1–T4 do not consistently form reusable sparse activations. Tone-selective SAE functions transfer and recur, while the same decoder directions recover the inventory only when read as a linear span; the resulting allocation pattern supports a feature-absorption account. SAE and PCA directions also influence the same frozen downstream readout. A focused WavLM control shows that this allocation changes with model conditions, keeping the empirical claim at the level of a controlled Mandarin case. The methodological lesson is that probes, factor-held-out feature tests, decoder-span controls and interventions provide distinct evidence for a learned feature.

7 Limitations

The empirical scope comprises two matched Mandarin checkpoints, a focused WavLM Base layer-8 control, 3,000 AISHELL-3 utterances and TopK SAEs fitted to pooled syllable representations. The WavLM run broadens the architecture comparison but uses one layer and one split seed, so it does not establish architecture-independent recurrence. We manually checked a sample of 500 utterances for syllable identity and gross boundary placement. Five-bin and voiced-pooling controls test dependence on temporal aggregation. The labels describe observed lexical-tone tokens without a separate annotation of underlying tone. Interventions target frozen tone and segment readouts, so they establish model-level access rather than perceptual tone conversion. Replication across additional architectures, corpora and speech factors is needed to determine how broadly the observed sparse–distributed organization generalizes.

Reproducibility statement

The main text specifies the data filters, representation construction, data partitions, SAE architecture, optimization settings, feature-selection rule, dense baselines, evaluation criteria and intervention protocol. Appendix A provides implementation details; Appendix C reports complete split and overlap audits; Appendix D reports the full-state control; and Appendices E–J provide the intervention intervals, layer matrix and sensitivity analyses. The anonymized code, configuration files and derived evaluation tables are included with the supplementary submission.

AI USE STATEMENT

We used LLMs to assist with grammar, clarity, manuscript organization, code drafting and figure revision. All experiments and reported results were executed and verified by the authors. The authors reviewed and edited all LLM-assisted material and take full responsibility for the scientific claims and final content of this paper.

ETHICS STATEMENT

This study analyzes an existing speech corpus and does not collect new human-subject data, attempt speaker identification or infer speaker demographics. The claims concern model representations and should not be interpreted as assessments of individual speakers or linguistic competence. Dataset use follows the corpus release terms.

REFERENCES

Georgii Aparin, Tasnima Sadekova, Alexey Rukhovich, Assel Yermekova, Laida Kushnareva, Vadim Popov, Kristian Kuznetsov, and Irina Piontkovskaya. Audiosae: Towards understanding of audio-processing models with sparse autoencoders. In *Proceedings of EACL*, pp. 3221–3254, 2026. doi: 10.18653/v1/2026.eacl-long.149.

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

Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2755. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/764ff7477b8e24dbe01531f6791e8bdf-Abstract-Conference.html.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.

Antón de la Fuente and Dan Jurafsky. A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models. In *Proceedings of Interspeech*, pp. 1290–1294, 2024. doi: 10.21437/Interspeech.2024-2341.

Joshua Engels, Logan Riggs, and Max Tegmark. Decomposing the dark matter of sparse autoencoders, 2024. URL https://arxiv.org/abs/2410.14670.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *Proceedings of ICLR*, 2025.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders, 2026. URL https://arxiv.org/abs/2606.12138.

Michele Gubian, Ioana Krehan, Oli Liu, James Kirby, and Sharon Goldwater. Analyzing the relationships between pretraining language, phonetic, tonal, and speaker information in self-supervised speech models, 2025. URL https://arxiv.org/abs/2506.10855.

Pengcheng Guo and Shixing Liu. chinese_speech_pretrain: Chinese speech pretrained models. GitHub repository, 2022. URL https://github.com/TencentGameMate/chinese_speech_pretrain.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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, 2021. doi: 10.1109/TASLP.2021.3122291.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of ICLR*, 2024.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of ICML*, volume 267, pp. 29223–29264, 2025.

Jianjing Kuang. Covariation between voice quality and pitch: Revisiting the case of mandarin creaky voice. *The Journal of the Acoustical Society of America*, 142(3):1693–1706, 2017. doi: 10.1121/1.5003649.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *Proceedings of ICLR*, 2025.

Théo Mariotte, Martin Lebourdais, Antonio Almudévar, Marie Tahon, Alfonso Ortega, and Nicolas Dugué. Sparse autoencoders make audio foundation models more explainable. In *Proceedings of ICASSP*, 2026. doi: 10.1109/ICASSP55912.2026.11463660.

Corinne B. Moore and Allard Jongman. Speaker normalization in the perception of mandarin chinese tones. *The Journal of the Acoustical Society of America*, 102(3):1864–1877, 1997. doi: 10.1121/1.420092.

Jeffrey Olmo, Jared Wilson, Max Forsey, Bryce Hepner, Thomas Vincent Howe, and David Wingate. Features that make a difference: Leveraging gradients for improved dictionary learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7624–7634, 2025. doi: 10.18653/v1/2025.findings-naacl.423.

Open Speech and Language Resources. Aishell-3 speech corpus, slr93. OpenSLR resource page, 2020. URL https://www.openslr.org/93/.

Opeyemi Osakuade and Simon King. Lexical tone is hard to quantize: Probing discrete speech units in mandarin and yorùbá, 2026. URL https://arxiv.org/abs/2604.07467.

Iñigo Parra. Interpretable sparse features for probing self-supervised speech models. In *Proceedings of the IJCNLP-AACL Student Research Workshop*, pp. 1–9, 2025. doi: 10.18653/v1/2025.ijcnlp-srw.1.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features, 2025. URL https://arxiv.org/abs/2501.16615.

Gaofei Shen, Michaela Watkins, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. Encoding of lexical tone in self-supervised models of spoken language. In *Proceedings of NAACL*, pp. 4250–4261, 2024. doi: 10.18653/v1/2024.naacl-long.239.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus. In *Proceedings of Interspeech*, pp. 2756–2760, 2021. doi: 10.21437/Interspeech.2021-755.

Yi Xu. Contextual tonal variations in mandarin. *Journal of Phonetics*, 25(1):61–83, 1997. doi: 10.1006/jpho.1996.0034.

Yi Xu and Q. Emily Wang. Pitch targets and their realization: Evidence from mandarin chinese. *Speech Communication*, 33(4):319–337, 2001. doi: 10.1016/S0167-6393(00)00063-7.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *Proceedings of ICASSP*, pp. 6182–6186, 2022. doi: 10.1109/ICASSP43922.2022.9746682.

This appendix provides implementation details, factor-specific transfer protocols, the strict full-state probe matrix, the complete 108-dictionary matrix, intervention intervals, the expanded layer screen and sensitivity analyses. The main paper contains the evidence required for its central claims.

A Implementation and Evaluation Details

Corpus preparation The 3,000-utterance subset contains 31,012 aligned syllable intervals from 182 speakers. Numbered pinyin in the AISHELL-3 manifest supplies the tone label; we do not infer tone from the TextGrid character. Neutral-tone tokens remain available during label-free SAE fitting but are excluded from T1–T4 evaluation. Labeled evaluation also excludes <unk> intervals and intervals shorter than 50 ms, leaving 28,886 labeled tokens (T1: 6,576; T2: 7,253; T3: 5,017; T4: 10,040). All four tones therefore have substantial support, although the retained natural distribution is not exactly balanced. Within-segment ranking, matched-coverage purity, class-balanced readouts, and the size-matched balance sensitivity reduce sensitivity to category frequency. The release reports more than 98% word and tone accuracy for its text and pinyin labels (Open Speech and Language Resources, 2020). TextGrid boundaries were generated by forced alignment. We manually checked a sample of 500 utterances (16.7%) for syllable identity and gross boundary placement; this is a quality-control sample rather than independent frame-level phonetic adjudication.

The import is deterministic. It reads the nonempty word intervals from each TextGrid, pairs them in order with the numbered pinyin tokens in the AISHELL-3 human manifest, and accepts an utterance only when its audio and TextGrid are present and the two token counts match. The first 3,000 successfully imported utterances form the analysis subset. Tone labels come from the pinyin digit; the TextGrid character is used only as the aligned syllable identity. Initial and final strings are obtained by removing the tone digit and matching the onset against a fixed Mandarin-initial list. No audio-derived tone relabeling is performed. Table 2 gives the interval-level filtering flow. The exclusion counts overlap, so the union row, rather than the sum of the preceding rows, determines the retained evaluation count.

Table 2: Preprocessing and labeled-evaluation flow. Counts are interval-level; the exclusion rows overlap. All imported intervals remain in the label-free SAE-fitting pool, whereas only the final row enters T1–T4 labeled evaluation.

Stage	Utterances	Intervals
Imported analysis subset	3,000	31,012
Neutral tone (T5)	–	1,856
<unk> syllable	–	73
Duration < 50 ms	–	234
Union of evaluation exclusions	–	2,126
Retained T1–T4 evaluation tokens	–	28,886

Representations and SAE fitting Audio is resampled to 16 kHz. Each aligned syllable is split into three equal-duration intervals; hidden states are averaged within each interval and concatenated into a 2,304-dimensional vector. Normalization uses a dimension-wise mean and one global root-mean-square scale estimated from the SAE-fit pool. Each TopK SAE has 4,096 features and at most 128 positive activations. Decoder columns are renormalized after every update. Training uses mean squared reconstruction loss, AdamW, learning rate 10^{-3}, batch size 512, 30 epochs, and seeds 17, 23, and 29. Reconstruction MSE, explained variance, mean L0, activation density, and dead-feature fraction are recorded on validation. A label-blind $k \in \{64, 128, 256\}$ comparison selected $k = 128$; cells above the 20% dead-feature gate are retained only as quality-flagged diagnostics.

Selection, baselines, and inference For every tone, within-segment one-vs-rest AUC is computed only for initial–final types containing positive and negative examples; segment AUCs are averaged with weight $n_+ n_-$. TopK features retain their positive orientation. Dense raw, PCA, and random directions may use either sign. Matched-coverage purity thresholds a dense direction at the discovery quantile corresponding to the selected SAE feature's firing rate, freezes the threshold, and measures target-tone purity on test. Coverage is therefore matched on discovery, not necessarily on test. For each tone, the dense comparison is the maximum test purity among the three discovery-selected raw, PCA and random coordinates. This maximum is an evaluation envelope; it does not select a new coordinate for downstream analysis. Searching the full dictionary is calibrated with within-segment max-statistic permutations: 999 for the layer screen and 99 for each dictionary-refitting

Table 3: Retained above-chance four-tone gain across coordinate budgets. Act. uses selected SAE activations; span projects the original SSL state into the span of the same selected decoder directions. PCA uses discovery-selected principal-component coordinates. Values average three readout seeds and, for SAE conditions, three dictionary seeds. The full-state reference equals 1.

	HuBERT L8			wav2vec 2.0 L4		
k	Act.	Span	PCA	Act.	Span	PCA
4	0.389	0.609	0.354	0.395	0.623	0.634
8	0.488	0.716	0.454	0.513	0.792	0.706
16	0.562	0.836	0.654	0.620	0.880	0.790
32	0.675	0.925	0.824	0.697	0.959	0.845
64	0.816	0.990	0.972	0.797	1.015	0.977
128	0.887	1.014	0.972	0.891	1.042	1.016

run. The latter is a calibration check rather than a source of fine-grained p-values. Test uncertainty uses speaker- or segment-clustered bootstrap intervals as appropriate. SAE and readout seeds are technical replicates.

Acoustic diagnostics The cue analysis retains 0.10–0.80 s syllables with at least four voiced frames and 40% voiced coverage. Predictors are speaker-normalized F0 level, F0 at the start, middle, and end, slope, curvature, F0 range, duration, and voiced-frame fraction. A ridge model trained on discovery and validation predicts the frozen feature preactivation; raw, predicted, and residual tone AUC and held-out R^2 are evaluated on segment-test tokens. Temporal reversal reverses the three pooled bins without reselecting the feature. The Praat PSOLA analysis resynthesizes 80 held-out tokens per tone toward a target contour and uses a sham resynthesis control. The paired double difference tests acoustic responsiveness, not model-level causal control.

Forward intervention A four-way tone readout is trained on layer-12 discovery and validation states. The endpoint is the target logit minus the largest non-target logit. Two additional frozen readouts predict 24 initials and 34 finals. Studies use 40 segment-test tokens per source tone. Steering adds or subtracts a unit-normalized direction at 0.5, 1, and 2 times its validation-derived coefficient; dense and SAE directions receive matched L_2 perturbations. SAE ablation subtracts the selected coefficient times its decoder direction and is interpreted only within method. Segment preservation requires the speaker-clustered 95% interval for each gold-class margin change to remain inside $[-0.10, 0.10]$ SD. Paired sign-flip tests, clustered bootstrap intervals, and the paired 2x-minus-0.5x contrast assess direction and dose consistency. In set-to-real interchange, the synthetic coefficient is replaced by the difference between the source activation and the held-out validation mean for the target tone. The resulting hidden state remains a hybrid.

Full-state probe and equal-budget recovery The strict information control fits a class-balanced four-way linear readout to the complete 2,304-dimensional pooled state. Training uses labeled T1–T4 tokens from discovery and validation; evaluation uses the dictionary-segment test pool. The readout is trained for 80 epochs with seeds 17, 23 and 29. Reported cell values first average these technical seeds within each partition and then summarize the three overlapping partitions. Frozen test logits and row ids are saved for every run.

Discovery AUC ranks coordinates for budgets $k \in \{4, 8, 16, 32, 64, 128\}$. A fixed T1–T4 round robin selects unused coordinates so that one tone cannot consume the budget. PCA uses a frozen 256-component candidate set and the SAE uses its 4,096-coordinate dictionary. Class-balanced readouts train on discovery plus validation and evaluate on 4,409 test tokens using three SAE seeds and three readout seeds. Retained gain is $(R_k - 0.25)/(R_{\text{full}} - 0.25)$.

For the decoder-span control, let $D_S \in \mathbb{R}^{2304\times k}$ contain the decoder directions of the discovery-selected SAE coordinates. A thin SVD $D_S = U\Sigma V^\top$ gives an orthonormal basis U_r for their numerical span, and the readout receives xU_r rather than the SAE activation values. The selected feature ids, data split, readout training and test tokens remain unchanged. This control preserves the geometry proposed by the SAE while removing encoder amplitudes, ReLU and TopK gating. All selected decoder sets have effective rank $r = k$ in all three SAE seeds.

The selected decoder span reaches 90% retained gain at $k = 32$ in both encoders, whereas selected SAE activations remain below 90% through $k = 128$. Values slightly above 1 indicate that the lower-dimensional readout exceeds the separately optimized full-state readout on the same test set; they are not interpreted as additional tone information.

Table 4: Reconstruction control on the focused segment-held-out split. Each value averages three SAE seeds and three readout seeds. The original state is the full pooled SSL state; the code uses all 4,096 SAE activations; the reconstruction is $\hat{x}$; and the residual is $x - \hat{x}$. These probes are diagnostic and do not participate in feature selection.

Model/layer	View	Macro recall	Macro AUC	AUC T3	AUC T4
HuBERT L8	Original	0.894	0.972	0.959	0.980
	SAE code	0.798	0.927	0.892	0.942
	Reconstruction	0.848	0.954	0.926	0.968
	Residual	0.514	0.756	0.770	0.733
wav2vec 2.0 L4	Original	0.859	0.960	0.942	0.966
	SAE code	0.781	0.922	0.889	0.928
	Reconstruction	0.808	0.939	0.919	0.948
	Residual	0.515	0.758	0.758	0.740

Table 5: The factor-specific transfer protocols. SAE fitting is label-free; discovery is the first stage that uses tone labels.

Analysis	Partition rule	May still recur	Supported interpretation
Speaker transfer	SAE fit, discovery, validation, and test use disjoint speakers	Segment and sentence context	Coordinate ranking transfers across vocal ranges
Segment-selection transfer	The speaker-disjoint SAE pool is fixed; discovery, validation, and test use disjoint initial–final types	Test segments in label-free SAE fitting; utterance context	A coordinate selected in one dictionary transfers to new labeled segment types
Dictionary-segment transfer	Initial–final types are disjoint across SAE fit, discovery, validation, and test; every basis is refitted	Speaker and utterance context	A selective coordinate is recovered when dictionary fitting never sees test segment types
Utterance-context control	The speaker-disjoint SAE pool is fixed; complete evaluation utterances are assigned to discovery, validation, or test	Segment types	Ranking is not explained by syllables from the same sentence crossing labeled partitions

B Factor-Specific Transfer Protocols

All protocols reserve approximately 50% of tokens for label-free dictionary or basis fitting, 20% for feature discovery, 15% for validation, and 15% for test. Table 5 states what each partition changes and what may still recur. The protocols answer complementary questions rather than form one scalar strictness ordering. Feature selection is repeated within each protocol, so passing several protocols does not establish joint invariance of one frozen coordinate. The dictionary-segment partitions have disjoint pools within each run but overlapping SAE-fit support across partition seeds. Their fit sets contain 191, 191 and 190 segment types; pairwise overlaps are 107, 103 and 105 types. Seeds are repeated fits on shared data, not new corpora.

C Transfer-Test Audit

The dictionary-segment protocol assigns complete initial–final types to SAE fit, discovery, validation, or test before any basis is fitted. Table 7 reports the test partitions. Activation tables for both encoders and all central layers have identical row assignments. Each partition contains all four evaluated tones; test counts are approximately 1,019 T1, 1,095 T2, 755 T3, and 1,518 T4 tokens in every split.

For each split seed, the activation normalizer, PCA basis, random directions, and three TopK SAEs are fitted again on that split's unlabeled SAE pool. SHA-256 checks confirm 108 distinct SAE checkpoints across two models and layers 4–9. Feature selection uses discovery labels only. The selected feature id and orientation are identical in the saved validation and test evaluations; all SAE orientations are positive because TopK codes are nonnegative. Dense directions alone may use either sign.

We describe a selected feature as transferring in a run when (i) its discovery maximum exceeds the 95th percentile of the within-segment max-statistic permutation null and (ii) its frozen test segment-conditioned AUC exceeds 0.5. It has a selectivity advantage when (iii) its test purity also exceeds

Table 6: Which analysis supports which claim. The inventory and intervention case studies do not use the strict dictionary-segment refits. Their test segments can have occurred during label-free dictionary fitting.

Evidence	Layers	Dictionary and evaluation split	Endpoint
Full-state availability	Both models, L4–L9	Linear probe fitted on discovery + validation in each dictionary-segment partition	Four-tone macro recall and one-vs-rest AUC
Transfer and recurrence	Both models, L4–L9	Refit within each of three dictionary-segment partitions; three SAE seeds	Frozen single-coordinate AUC and purity
Inventory recovery	HuBERT L8; wav2vec 2.0 L4	Speaker-disjoint fitting; segment-selection holdout	Four-tone macro recall from selected coordinate sets
Forward intervention	HuBERT L8; wav2vec 2.0 L4	Same focused dictionaries and segment-selection protocol as inventory recovery	Frozen layer-12 tone, initial and final margins
Speaker and utterance controls	Focused layers	Protocol-specific selection; dictionary fitting as specified in Table 5	Single-coordinate transfer under the named factor

Table 7: Dictionary-segment test partitions. Token totals include neutral tones; labeled evaluation uses T1–T4 only. Segment overlap with SAE fit, discovery, and validation is zero. Speakers and utterances may recur across partitions; separate controls address those factors.

Split seed	Tokens	Segments	Speakers	Utterances
20260714	4,637	58	162	2,260
20260715	4,641	59	160	2,272
20260716	4,637	58	163	2,195

Table 8: Run-level dictionary-segment audit. "Purity > 0" means a positive SAE-minus-best-dense purity difference with discovery-calibrated coverage.

Tone	Transfer + advantage	AUC > 0.55	Purity > 0	Minimum AUC
T1	7/18	7/18	9/18	0.435
T2	3/18	2/18	3/18	0.498
T3	**18/18**	**18/18**	**18/18**	**0.657**
T4	9/18	12/18	9/18	0.505

each discovery-selected raw, PCA and random coordinate with discovery-calibrated thresholds. The complete matrix appears in Table 10; Figure 3 summarizes the aggregate rates and tone distribution. Table 8 reports strict refits at the same layer locations as the inventory and intervention case studies, but with different dictionaries and partitions. Its 18 runs are technical replications, not independent speech samples.

For a concrete uncertainty check, we cluster the same frozen T3 evaluation by test segment type. The HuBERT-L8 and wav2vec-2.0-L4 examples used in the main text are shown below; these intervals quantify variation over segment types, not uncertainty over independent corpora.

Table 10 reports every central-layer cell. Each entry is "A/S": the number of runs with test AUC above 0.55, followed by the number with transfer and a selectivity advantage. The second count can exceed the first because its transfer boundary is chance (0.5), while 0.55 is a separate descriptive effect-size threshold.

In the full matrix, all 432 discovery maxima exceed their null, 399 frozen coordinates remain above chance, 321 exceed AUC 0.55, and 259 have positive SAE-minus-best-dense purity. Transfer remains above chance in all nine runs for 37 of 48 model–layer–tone cells; 20 cells exceed AUC 0.55 in all nine. The transfer-plus-advantage criterion holds in 253 runs, but only seven cells pass all three checks in 9/9 runs. T3 contributes HuBERT L7–L9 and wav2vec 2.0 L4/L9; T2 contributes HuBERT L5, and T4 contributes wav2vec 2.0 L8. The first result establishes repeated recovery of a tone-selective function, while the second shows a repeated advantage over the discovery-calibrated dense controls. Neither implies a shared feature id or decoder direction. Decoder cosine and activation-correlation matching in the main analysis applies to the three original dictionaries per focused cell, not to the strict refits.

Table 9: Segment-clustered bootstrap intervals for SAE seed 23 and dictionary-segment split 20260714 (2,000 resamples). Evaluation excludes neutral tones, retaining 4,387 T1–T4 tokens and 30 eligible T3 comparison segments.

Case	AUC	2.5%	97.5%
HuBERT L8, T3	0.860	0.851	0.903
wav2vec 2.0 L4, T3	0.675	0.600	0.693

Table 10: Complete dictionary-segment matrix. A is the number of test AUCs above 0.55; S is the number passing the discovery max-statistic, test AUC above 0.5 and positive SAE-minus-best-dense purity at discovery-calibrated coverage. Every cell contains nine technical runs.

Model	Layer	T1 A/S	T2 A/S	T3 A/S	T4 A/S
HuBERT	4	7/5	9/8	6/2	7/3
HuBERT	5	6/6	**9/9**	8/0	8/5
HuBERT	6	9/5	9/7	9/6	9/6
HuBERT	7	9/7	6/8	**9/9**	9/6
HuBERT	8	5/5	0/1	**9/9**	5/5
HuBERT	9	1/2	1/1	**9/9**	6/3
wav2vec 2.0	4	2/2	2/2	**9/9**	7/4
wav2vec 2.0	5	1/1	1/1	9/7	4/1
wav2vec 2.0	6	3/4	3/3	8/8	8/4
wav2vec 2.0	7	9/6	6/6	9/8	8/7
wav2vec 2.0	8	7/2	7/7	9/6	**9/9**
wav2vec 2.0	9	9/6	8/7	**9/9**	9/7

Table 11: Strict full-state linear probes. Macro recall and macro one-vs-rest AUC are averaged across three dictionary-segment partitions after averaging three readout seeds. Brackets give the partition range, not a confidence interval. Four-way chance macro recall is 0.25.

Model	Layer	Macro recall [range]	Macro AUC
HuBERT	4	0.739 [0.708, 0.754]	0.906
HuBERT	5	0.758 [0.736, 0.774]	0.912
HuBERT	6	0.809 [0.797, 0.821]	0.937
HuBERT	7	0.840 [0.840, 0.841]	0.949
HuBERT	8	0.869 [0.859, 0.881]	0.959
HuBERT	9	0.882 [0.869, 0.890]	0.966
wav2vec 2.0	4	0.848 [0.842, 0.854]	0.951
wav2vec 2.0	5	0.886 [0.875, 0.900]	0.967
wav2vec 2.0	6	0.917 [0.908, 0.934]	0.977
wav2vec 2.0	7	0.931 [0.920, 0.940]	0.980
wav2vec 2.0	8	0.917 [0.902, 0.928]	0.977
wav2vec 2.0	9	0.866 [0.856, 0.881]	0.956

D Strict Full-State Probe Control

The full-state control uses the same dictionary-segment partitions as the strict SAE matrix. Probe fitting uses 10,225–10,239 labeled discovery and validation tokens; evaluation uses 4,386–4,389 labeled test tokens. Within a partition, the 56–58 test initial–final types do not occur in the probe training pools or the label-free SAE-fit pool. Table 11 reports partition-level means after averaging three technical readout seeds. Every model–layer cell remains well above four-way chance. The saved logits reproduce the run-level metrics to numerical precision.

These probes answer whether tone is linearly available under the same segment holdout as the SAE matrix. They do not identify a coordinate or imply that the probe weights correspond to the model's internal feature basis. The focused full-state values in Figure 4b differ because that experiment uses the segment-selection protocol and its 4,409-token test set.

E Intervention Intervals and Equivalence Tests

Table 12 reports the uncertainty omitted from a means-only intervention summary. Positive and negative 1x steering are sign-aligned before averaging. Intervals resample held-out speakers while retaining SAE seeds as technical replications.

Table 12: Mean sign-aligned change in the frozen target-tone margin under norm-matched 1x steering, with speaker-clustered 95% bootstrap intervals. Each target aggregates 120 unique held-out source tokens; positive/negative directions and three SAE seeds are repeated measurements, not independent speech samples.

Model	Target	SAE effect [95% CI]	PCA effect [95% CI]
HuBERT L8	T1	0.693 [0.649, 0.734]	0.882 [0.770, 0.988]
HuBERT L8	T3	0.711 [0.649, 0.774]	0.747 [0.630, 0.851]
wav2vec 2.0 L4	T1	0.270 [0.248, 0.293]	0.344 [0.307, 0.384]
wav2vec 2.0 L4	T2	0.035 [0.026, 0.045]	0.471 [0.435, 0.510]
wav2vec 2.0 L4	T3	0.417 [0.390, 0.442]	0.505 [0.460, 0.545]
wav2vec 2.0 L4	T4	-0.001 [-0.006, 0.004]	0.221 [0.196, 0.250]

Table 13: Eight-layer segment-selection screen. Layers 5, 7, and 9 provide criterion-frozen interpolation within the central-depth analysis. wav2vec 2.0 layers 10 and 12 cross the predeclared 20% dead-feature gate; their AUCs are included for completeness but do not support representational-absence claims.

Model	Layer	T1	T2	T3	T4	Explained variance	Dead features
HuBERT	4	0.592	0.717	0.611	0.625	0.798	5.7%
HuBERT	5	0.636	0.695	0.656	0.670	0.808	6.1%
HuBERT	6	0.694	0.766	0.741	0.626	0.827	6.6%
HuBERT	7	0.784	0.739	0.806	0.631	0.837	6.2%
HuBERT	8	0.671	0.508	0.797	0.506	0.820	2.0%
HuBERT	9	0.701	0.503	0.764	0.520	0.802	1.2%
HuBERT	10	0.640	0.499	0.713	0.562	0.821	1.8%
HuBERT	12	0.643	0.603	0.635	0.567	0.921	7.6%
wav2vec 2.0	4	0.669	0.520	0.747	0.521	0.773	1.6%
wav2vec 2.0	5	0.636	0.504	0.676	0.508	0.761	0.8%
wav2vec 2.0	6	0.640	0.508	0.731	0.569	0.768	1.4%
wav2vec 2.0	7	0.670	0.565	0.750	0.652	0.790	3.7%
wav2vec 2.0	8	0.640	0.593	0.672	0.624	0.802	6.3%
wav2vec 2.0	9	0.607	0.558	0.657	0.683	0.831	11.5%
wav2vec 2.0	10	0.575	0.563	0.701	0.628	0.923	44.3%
wav2vec 2.0	12	0.570	0.581	0.674	0.560	0.973	61.2%

The prespecified segment-equivalence result is complete rather than based on mean margin changes. For HuBERT, every initial/final test cell passes for SAE (72/72) and PCA (72/72); the corresponding wav2vec 2.0 counts are 144/144 for each method. Across these cells, the most extreme speaker-clustered interval endpoints are $[-0.039, 0.036]$ (HuBERT SAE), $[-0.050, 0.043]$ (HuBERT PCA), $[-0.037, 0.036]$ (wav2vec 2.0 SAE), and $[-0.075, 0.057]$ (wav2vec 2.0 PCA), all inside the predeclared $[-0.10, 0.10]$ SD band. HuBERT SAE ablation reduces the T1 and T3 target margins by 0.538 [0.397, 0.676] and 0.739 [0.552, 0.899], respectively.

For the HuBERT T3 set-to-real control, target-margin changes from T1, T2, and T4 sources are 0.675 [0.586, 0.751], 0.483 [0.417, 0.547], and 0.287 [0.235, 0.342], respectively. These are speaker-clustered intervals. The corresponding initial/final intervals all lie inside $[-0.008, 0.018]$ SD. Agreement in sign with synthetic 1x steering is 98–100% across the three source tones.

F Eight-Layer Localization Screen

Table 13 reports the discovery-selected SAE coordinate at every tested layer. Entries are mean segment-conditioned test AUC over three SAE seeds. This screen uses one-vs-rest selection uniformly for all tones. It is descriptive layer coverage, whereas the dictionary-refitting analysis above is the targeted transfer stress test.

The expanded screen changes the interpretation in a useful way. HuBERT T2 is strong at layers 4–7, while L7 supports high T1/T2/T3 AUC and L8–L9 favor T3. In wav2vec 2.0, T3 is strongest through L7, whereas T4 is strongest at L8–L9. The paper therefore does not claim that T1, T2, or T4 is unavailable to sparse localization. Instead, category localization is reorganized across depth, and the strict matrix distinguishes broad transfer from the narrower case in which an SAE coordinate is repeatedly more selective than matched dense alternatives.

Table 14: Sensitivity of the cell-level transfer-plus-advantage count to the required number of passing runs.

Required runs	All	T1	T2	T3	T4	Cells
$\geq 7/9$	18/48	1	6	8	3	48
$\geq 8/9$	11/48	0	3	7	1	48
$\geq 9/9$	7/48	0	1	5	1	48

Table 15: Minimum SAE-minus-best-dense purity difference across nine runs. Original thresholds are calibrated on discovery; equal-coverage thresholds use unlabeled test scores. Negative minima identify failures of the all-nine rule, not absence of a tone-selective SAE feature.

Originally all-nine cell	Min. original Δ	Min. equal-coverage Δ	Equal-coverage passes
HuBERT L5–T2	0.109	0.114	9/9
HuBERT L7–T3	0.073	0.048	9/9
HuBERT L8–T3	0.118	0.156	9/9
HuBERT L9–T3	0.067	0.062	9/9
wav2vec 2.0 L4–T3	0.018	-0.014	8/9
wav2vec 2.0 L8–T4	0.067	0.092	9/9
wav2vec 2.0 L9–T3	0.006	-0.019	8/9

G Sensitivity of the Recurrence Criterion

The all-nine transfer-plus-advantage rule is a descriptive reproducibility criterion. Table 14 varies the required number of passing runs while keeping discovery-calibrated coverage. T3 contributes the largest number of cells at each threshold. It is the only tone spanning both encoders at $\geq 8/9$ and 9/9; T2 also spans both at $\geq 7/9$. This statement is conditional on the original coverage rule, not a threshold-independent property of T3. Because seeds share observations and partition fit sets overlap, we do not interpret these counts through an independent-trial binomial posterior.

With a minimum positive purity difference of 0.01, 0.02 or 0.05, the number of all-nine cells is six, five or five, respectively (242, 231 and 208 passing runs). At 0.02, wav2vec 2.0 L4–T3 and L9–T3 no longer pass all nine runs. Continuous effects and conditional uncertainty are reported next rather than converting the nine dependent outcomes into a pass-probability estimate.

H Coverage, Weighting and Effect-Size Audit

This post hoc audit replays all 108 checkpoints and frozen dense selections. Recovered test purity agrees with the original tables within 2×10^{-16}; pair-weighted AUC agrees within 4×10^{-8}. No feature id, direction sign or readout is selected again. All sensitivities use only T1–T4 tokens.

Equal test coverage For each frozen SAE coordinate, let m be the number of test tokens with positive activation. Each frozen dense direction is evaluated on its top m test scores. Boundary ties share the remaining activation weight equally, without consulting labels. SAE activation remains unchanged. Purity is the target-tone count divided by the same effective activation count m. This is a label-blind test-distribution calibration diagnostic, not a replacement for the discovery-calibrated deployment threshold. It produces 239 transfer-plus-advantage runs and five all-nine cells. The five are HuBERT L5–T2, L7–T3, L8–T3, L9–T3 and wav2vec 2.0 L8–T4.

Equal segment weighting For each tone, we retain test initial–final types with at least one positive and one negative example, then average their AUCs equally. There are 32–38 eligible types for T1, 34–42 for T2, 30–35 for T3 and 37–39 for T4 across partitions. Equal weighting yields 421 above-chance runs and 41 all-nine cells; at AUC > 0.55 the corresponding counts are 314 and 19. Applying both equal segment weighting and equal test coverage gives 244 transfer-plus-advantage runs and the same five all-nine cells. This tests evaluation weighting, not an alternative feature-selection objective.

Clustered uncertainty Within each model–layer–partition–tone condition, 2,000 bootstrap draws resample test segment types, retaining all their tokens. The same draw is used for all SAE seeds and baseline families. Discovery-calibrated firing thresholds remain fixed; the maximum dense purity is recomputed within every draw, and the three SAE-seed differences are averaged. These are pointwise 95% percentile intervals conditional on fitted coordinates and one partition, not intervals

Table 16: Mean purity advantage across three technical SAE seeds, with segment-clustered 95% intervals under discovery-calibrated thresholds. The seven cells were highlighted by the original all-nine rule; these conditional intervals are descriptive follow-up, not independent confirmation. All 144 condition-level intervals are in the reproducibility tables.

Cell	Partition 20260714	Partition 20260715	Partition 20260716
HuBERT L5–T2	0.182 [0.050, 0.267]	0.139 [−0.005, 0.230]	0.125 [0.053, 0.188]
HuBERT L7–T3	0.232 [0.077, 0.308]	0.104 [−0.042, 0.234]	0.222 [0.120, 0.283]
HuBERT L8–T3	0.417 [0.234, 0.461]	0.437 [0.250, 0.491]	0.240 [0.137, 0.349]
HuBERT L9–T3	0.263 [0.149, 0.372]	0.266 [0.144, 0.398]	0.147 [0.084, 0.211]
wav2vec 2.0 L4–T3	0.121 [0.035, 0.248]	0.273 [0.181, 0.361]	0.324 [0.165, 0.434]
wav2vec 2.0 L8–T4	0.178 [0.078, 0.240]	0.188 [0.011, 0.309]	0.146 [0.081, 0.225]
wav2vec 2.0 L9–T3	0.223 [0.029, 0.271]	0.080 [0.006, 0.129]	0.090 [0.017, 0.163]

Table 17: Frozen-feature T3 context sensitivity over nine focused runs. The 77 flagged T3–T3 tokens are removed from the manifest before evaluation; 10.3 is the mean number removed per focused run.

Model–layer	Mean removed/run	AUC all	AUC excl.
HuBERT L8	10.3	0.782	0.786
wav2vec 2.0 L4	10.3	0.735	0.738

over independent corpora or simultaneous tests of all cells. We do not pool overlapping partitions as independent observations. Of 144 conditions, some have empty active sets in rare resamples; valid draws range from 1,984 to 2,000 and counts are saved. Every cell in Table 16 has all 2,000 valid draws.

Tone-context sensitivity The primary manifest uses the numbered pinyin supplied by the AISHELL-3 annotation rather than an inferred label from the TextGrid character. It does not provide an independent underlying-versus-surface annotation, so we do not claim to identify all T3 sandhi or half-T3 tokens. As a conservative check, we flagged 77 T3 tokens immediately followed by T3 and removed them only from the frozen focused test evaluation. The mean T3 segment-conditioned AUC rose from 0.782 to 0.786 at HuBERT L8 and from 0.735 to 0.738 at wav2vec 2.0 L4. Mean changes for T1, T2, and T4 were below 0.001 in magnitude. This is a context sensitivity analysis, not a reannotation of the corpus.

I Focused Coordinate Controls

Inventory-recovery uncertainty At 64 coordinates, the paired PCA-minus-SAE macro-recall difference is 0.100 [0.082, 0.115] for HuBERT L8 and 0.110 [0.088, 0.135] for wav2vec 2.0 L4. These percentile intervals use 2,000 paired segment-clustered resamples of the 4,409 segment-selection test tokens, averaging technical fit pairs. They concern the focused dictionaries, not the strict refit matrix.

Contrastive and acoustic controls The weak HuBERT-L8 T2/T4 result is not specific to one-vs-rest scoring. Selecting a T2-versus-T4 contrast raises pairwise AUC under speaker transfer from 0.648/0.645 to 0.697/0.695 for T2/T4, but the same contrast reaches only 0.511/0.503 when initial–final types are held out. The contrast therefore separates the tones on familiar segment support without producing a segment-general feature at this layer.

At the focused layers, independently selected T3 features pass the original discovery-calibrated transfer-plus-advantage criterion in all 18 strict runs; equal test coverage gives 17/18. Mean strict AUC is 0.782 at HuBERT L8 and 0.735 at wav2vec 2.0 L4; the utterance-disjoint control gives 0.789 and 0.734. Across independently fitted focused dictionaries, mean decoder cosine is 0.62/0.68 and activation correlation is 0.72/0.76 for HuBERT/wav2vec 2.0. These comparisons support recurrence of a tone-selective function rather than reuse of one feature index.

Measured acoustic cues predict T3 preactivation with held-out $R^2 = 0.245$ at HuBERT L8 and 0.248 at wav2vec 2.0 L4. Residual tone AUC remains 0.705 and 0.683. Reversing the three temporal bins lowers preactivation AUC from 0.882 to 0.727 and from 0.826 to 0.720. PSOLA manipulation moves selected T1/T3 features in the predicted direction across all three seeds but does not consistently induce the target feature. The selected features therefore respond to ordered acoustic evidence, but these tests do not identify a unique acoustic mechanism.

In the tone-balanced HuBERT-L8 control, segment-held-out purity remains positive for T1 (+0.163) and T3 (+0.133), improves to +0.075 for T4, and remains negative for T2 (-0.107) in all three seeds. Balancing affects feature allocation without accounting for the full tone-specific pattern.

J Dictionary-Width Sensitivity

To test whether the selective result is a trivial consequence of fitting a 4,096-feature dictionary to a modest unlabeled pool, we refit HuBERT L8 SAEs on the same strict dictionary-segment training partition with 1,024, 4,096, and 8,192 features. TopK remains 128, all other optimization settings and the three SAE seeds are unchanged, and the T3 coordinate is selected on discovery and evaluated once on the frozen strict test partition. This is a focused scaling check rather than a new model comparison. Table 18 reports transfer AUC together with reconstruction and dead-feature diagnostics.

Table 18: Focused HuBERT-L8 T3 dictionary-width sensitivity. Test AUC is segment-conditioned and averaged over three seeds. Explained variance and dead features are computed on the unlabeled SAE fit pool; they are reported as reconstruction diagnostics rather than as test estimates.

Width	AUC	EV (fit)	Dead	n
1,024	0.811	0.879	0.0%	3
4,096	0.789	0.928	0.3%	3
8,192	0.710	0.939	1.1%	3

T3 transfer remains above chance for all nine width–seed runs (mean AUCs 0.811, 0.789, and 0.710), while explained variance increases with width and dead-feature fraction remains below 1.2%. Thus the existence of a T3 transfer signal is not specific to the 4,096-feature dictionary, but its localization strength is not monotonic in reconstruction quality. This result supports the paper's narrower claim: dictionary width changes the allocation of sparse handles, and reconstruction quality alone cannot be used as a proxy for category-level recurrence.

K Focused WavLM Cross-Architecture Control

We repeated the segment-held-out analysis at layer 8 of the WavLM Base checkpoint using the same 3,000-utterance manifest, three-bin pooling, 4,096-feature TopK SAEs with $K = 128$, and SAE seeds 17, 23 and 29. The run uses one split seed and is intended as a focused architecture control, not as an additional layer-by-layer recurrence matrix. Table 19 reports the frozen test AUC after segment conditioning and the number of positive SAE-minus-best-dense purity differences for each seed.

Table 19: Focused WavLM Base layer-8 control. Entries are segment-conditioned test AUCs for the tone-selective coordinate selected on each SAE seed. The final column counts tones for which the SAE purity exceeds all three dense controls on the same held-out test split.

SAE seed	T1	T2	T3	T4	Mean	SAE > dense
17	0.562	0.502	0.555	0.641	0.565	2/4
23	0.562	0.503	0.596	0.572	0.558	2/4
29	0.546	0.512	0.668	0.513	0.560	0/4
Mean	0.557	0.506	0.607	0.575	0.561	4/12

The WavLM SAEs have mean explained variance 0.766, dead-feature fractions of 0.17–0.51%, and mean $L_0 = 128$. Thus the weaker and less stable selectivity pattern is not explained by a failed or unusually dense SAE fit. It instead qualifies the generality of the main result: the recurrent sparse allocation observed in the two Mandarin-matched checkpoints should be read as a controlled case study, with model and pretraining conditions treated as part of the object of analysis.

L Scope and Additional Caveats

The paper's claims are intentionally scoped to three-bin pooled syllable representations and label-blind TopK SAEs. Frame-level dictionaries may expose different temporal features, and alternative SAE objectives may allocate coordinates differently.

A natural concern is that the three-bin resolution itself disfavors the dynamic rising and falling contours of T2 and T4 and thereby manufactures the focused pattern. We therefore refit SAEs at

HuBERT layer 8 and wav2vec 2.0 layer 4 using five equal-duration whole-syllable bins ($x \in \mathbb{R}^{3840}$) and three bins over the first-to-last voiced frame. With three-bin, five-bin, and voiced-span pooling, T3 AUC is 0.797/0.777/0.764 for HuBERT and 0.747/0.719/0.733 for wav2vec 2.0. Across the same six settings, T2 and T4 AUCs remain between 0.503 and 0.521. Thus the focused result is not specific to one whole-syllable bin count or to gross unvoiced-onset inclusion. These controls do not establish invariance to frame-level dictionaries or nucleus-normalized representations. AISHELL-3 provides professionally reviewed text and pinyin labels. The TextGrid boundaries used here were generated by forced alignment; 500 of the 3,000 utterances were manually inspected for syllable identity and gross placement. This sample and the short-interval sensitivity analysis are quality controls rather than independent frame-level adjudication. The corpus is read, studio-style speech, so the present evidence does not establish the same recurrence pattern in conversational or prosodically diverse speech.