CONVEX REFORMULATIONS FOR INTERPRETABLE ACOUSTIC-TO-ARTICULATORY INVERSION

Maor Fuks, Daniel BenShushan, Jorgen Bergh
University of California, Berkeley
{maor_fuks, daniel.benshushan, jorbe}@berkeley.edu

ABSTRACT

The Speech Articulatory Coding (SPARC) framework (Cho et al., 2024) achieves state-of-the-art vocal tract articulation through WavLM (Chen et al., 2022) layer-9 features with an ordinary least squares (OLS) head as an acoustic-to-articulatory inversion (AAI) model. This linear probe is accurate but expressively limited. Replacing the probe with a non-convex multi-layer perceptron (MLP) improves both single-speaker and cross-speaker Pearson correlation but forfeits global optimality and reproducibility. We instead replace the head with the convex reformulation of a two-layer ReLU network (Pilanci & Ergen, 2020; Mishkin et al., 2022), training one gated-ReLU model per electromagnetic articulography (EMA) channel via group-ℓ_1-regularized convex optimization. The resulting predictor is globally optimal and expresses each prediction as a sparse sum of linear models on distinct regions of WavLM feature space. Sweeping the regularization path, we recover over 95% of the OLS PCC with a small subset of active gates. Extending the convex probe to all 25 WavLM layers, we find that its lift over OLS is layer-dependent and largest at the deepest layers revealing nonlinear articulatory structure that a linear probe systematically underestimates. Probing the surviving gates against hand-aligned TIMIT phones reveals that the highest-information gates align with classical phonological categories without any phonetic supervision, with vowel and fricative-suppressing gates dominating the rankings across nearly every articulator, while articulator-private gates capture finer articulator-specific specializations invisible to the linear probe. By formulating the probe as a convex problem, we obtain interpretability and global optimality simultaneously and establish a probing framework applicable to any SSL representation, well beyond AAI.

1 INTRODUCTION

Speech is produced by moving the articulators of the vocal tract, so representing speech in articulatory terms gives a low-dimensional description grounded in human physiology (Cho et al., 2024; Browman & Goldstein, 1992). SPARC encodes speech as 12 channels of electromagnetic articulography (EMA) traces plus pitch and loudness, with a separate speaker encoder absorbing speaker variance (Cho et al., 2024). Its articulatory analysis is a two-stage probe. First, a frozen WavLM encoder produces 1024-dimensional layer-9 features. Second, an OLS head, fitted on a single MNGU0 (Richmond et al., 2011) speaker, maps those features to EMA. This design is driven by the hypothesis that individual articulatory spaces are related by an affine map (Cho et al., 2024).

The OLS probe is interpretable only in a weak sense, since each output channel is a fixed linear mapping of the acoustic embedding, which cannot express more complex behavior and mixes SSL-feature dimensions per articulator. This fails to identify which acoustic patterns drive articulators in distinct phonetic contexts. Replacing the head with a non-convex MLP is simple (a 2-layer ReLU MLP raises MNGU0 PCC from 0.876 to 0.890 and HPRC PCC from 0.765 to 0.781), but is susceptible to differences in optimizer initialization yields a nontransparent latent representation. Recent convex reformulations of two-layer ReLU networks (Pilanci & Ergen, 2020; Mishkin et al., 2022) resolve this by replacing the non-convex training objective with a globally optimal group-ℓ_1-sparse program whose surviving gates correspond to distinct activation regions of the feature space.

We bring this idea to AAI. We replace the SPARC linear AAI head with a per-channel gated-ReLU convex reformulation, train it on the same MNGU0 data and the same WavLM-9 features as SPARC, and use the resulting sparse gate structure as a probe of which acoustic classes drive each articulator. Our contributions are:

- We replicate SPARC's AAI setup and train ReLU MLPs as non-convex baselines, raising MNGU0 PCC from 0.876 to 0.890 and HPRC cross-speaker PCC from 0.765 to 0.781.
- We reformulate the AAI head as 12 per-channel gated-ReLU convex programs and characterize the regularization path. $\sim 8\%$ of the sampled gate budget recovers 95% of the per-channel OLS PCC, with the path peak exceeding OLS on all twelve channels.
- We extend the per-channel convex probe to all 25 WavLM layers and characterize the layer-wise PC$-$OLS gap. The gap nearly doubles at WavLM layer 24 ($+0.0246$ vs. $+0.0130$ at layer 9), showing that the convex probe's value as an interpretability instrument grows where linear probes underestimate the encoded articulatory signal.
- We perform a TIMIT phone-aligned analysis of the active gates, showing that a small set of universal gates recovers manner distinctions without phonetic supervision, and articulator-private gates surface specializations invisible to the linear probe.

2 Related Work

Acoustic-to-articulatory inversion. AAI predicts continuous articulator trajectories from speech audio, modeling the mid-sagittal (x, y) positions of the upper lip, lower lip, lower incisor, tongue tip, tongue blade, and dorsum. Early statistical methods (Gaussian mixture models, codebook-based mappings) were superseded by deep recurrent and attention-based architectures that exploit broader temporal context (Wu et al., 2023; Attia & Espy-Wilson, 2024; Siriwardena & Espy-Wilson, 2023), reaching high in-domain Pearson correlation on EMA datasets such as MNGU0 (Richmond et al., 2011) and HPRC (Tiede et al., 2017). Cross-speaker generalization remains the challenge. EMA sensor placement varies with each speaker's anatomy, so models trained on one speaker's coordinates do not directly apply to another's. Cho et al. (2024) and Cho et al. (2023) address this by hypothesizing that all speakers share a common articulatory space up to an affine transformation, which lets a linear probe over self-supervised speech features, trained on a single high-quality speaker (MNGU0), serve as a universal AAI space. The rest of the SPARC pipeline, a HiFi-GAN vocoder (Kong et al., 2020) conditioned on a CREPE pitch tracker (Kim et al., 2018) and a WavLM-based speaker encoder, treats this AAI head as fixed. We replace only that head.

Convex reformulations of neural networks. Pilanci & Ergen (2020) proved that training a two-layer ReLU network with weight decay is equivalent to a finite-dimensional convex program over the discrete set of *gates* (binary activation patterns) induced by the data, with cone constraints tying each linear weight to its gate. Subsequent work extended this to vector outputs, convolutions, and batch normalization (Ergen & Pilanci, 2021). Mishkin et al. (2022) made this framework practical with R-FISTA, an accelerated proximal-gradient solver, and showed that the unconstrained *gated-ReLU* relaxation is tightly equivalent to ReLU training. Ansari et al. (2024) (ConvexECG) applied this approach to reconstructing six-lead ECGs from a single lead, exploiting a key property of the convex formulation, that every ReLU breakpoint in the optimal network is anchored to a specific training point, so any prediction can be traced back to the training samples that produced it.

Probing speech SSL models. Linear and shallow probes on speech SSL representations are standard tools for understanding what those representations encode (Pasad et al., 2021; Choi et al., 2022). Cho et al. (2023; 2024) showed that a single linear layer on WavLM-9 already recovers EMA with high fidelity.

3 Baseline Setup

Data. We use MNGU0 (Richmond et al., 2011) for single-speaker training and 5-fold cross-validation, and HPRC (Tiede et al., 2017) (7 speakers with F03 dropped per the SPARC setup) for cross-speaker evaluation. Following SPARC, we use only the mid-sagittal x and y coordinates

Table 1: Baseline accuracy results. PCC is averaged over EMA channels and folds (MNGU0) and speakers (HPRC).

Model	MNGU0 PCC	HPRC PCC
OLS	0.876	0.765
2-Layer MLP (no context)	0.890	0.781
3-Layer MLP (with context)	0.908	0.780

of six standard EMA sensors: upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue blade (TB), and tongue dorsum (TD), yielding 12 output channels.

Preprocessing. We replicate the SPARC pipeline. Audio is resampled to 16 kHz, zero-meaned and unit-scaled. EMA is downsampled to 50 Hz to match the WavLM-Large frame rate, and each channel is z-scored within each utterance. We extract WavLM-Large layer-9 features (1024 dimensions, 50 Hz), and apply a 5th-order Butterworth low-pass filter at 10 Hz.

Cross-speaker alignment. Models are trained on MNGU0 only. For HPRC cross-speaker evaluation we use the lasso alignment of Cho et al. (2024) to map each HPRC speaker's EMA to the MNGU0 template space.

Linear OLS baseline. The SPARC AAI head fits a single linear map $W \in \mathbb{R}^{1024\times 12}$ for all channels by ordinary least squares. This baseline obtains 0.876 average PCC on our MNGU0 5-fold split. On HPRC it obtains 0.765 after alignment.

Non-convex MLP baselines. A 2-layer ReLU MLP without temporal context, trained on the same WavLM-9 features (see Appendix A), obtains a MNGU0 5-fold PCC of 0.890 and an HPRC cross-speaker PCC of 0.781. A 3-layer MLP with temporal context performs similarly in-domain but slightly worse cross-speaker, suggesting that temporal context introduces a small speaker-specific overfit that does not transfer (Section 7).

4 METHOD

We replace the AAI linear head with a convex reformulation of a two-layer ReLU network, applied independently to each of the 12 EMA channels. We use the gated-ReLU formulation of Mishkin et al. (2022), which drops the polyhedral cone constraints of the full convex reformulation for an unconstrained group-ℓ_1 program, exactly equivalent to two-layer ReLU training when $\lambda = 0$.

4.1 PER-CHANNEL CONVEX FORMULATION

Let $X \in \mathbb{R}^{n\times d}$ stack n training frames of $d = 1024$-dimensional preprocessed WavLM-9 features (Section 3), and let $y \in \mathbb{R}^n$ be a z-scored EMA channel. The non-convex two-layer ReLU training objective solves

$$\min_{W^{(1)},w^{(2)}} \tfrac{1}{2}\big\| \textstyle\sum_{i=1}^m (XW_i^{(1)})_+ w_i^{(2)} - y\big\|_2^2 + \tfrac{\lambda}{2} \textstyle\sum_{i=1}^m \big(\|W_i^{(1)}\|_2^2 + (w_i^{(2)})^2\big), \tag{1}$$

where $(\cdot)_+ = \max(\cdot, 0)$ is ReLU, $W_i^{(1)} \in \mathbb{R}^d$ is the i-th hidden weight, and $w_i^{(2)} \in \mathbb{R}$ its output coefficient. Pilanci & Ergen (2020) showed that this problem is equivalent to a convex program over the *hyperplane arrangement* of X, the finite set of diagonal indicator matrices

$$\mathcal{D}_X = \big\{\mathrm{diag}(\mathbf{1}[Xu \geq 0]) : u \in \mathbb{R}^d\big\}. \tag{2}$$

We sample a subset $\tilde{\mathcal{D}} \subset \mathcal{D}_X$ of size P, calling each surviving D_i a *gate*. The gated-ReLU convex program replaces the ReLU with a fixed gate and removes the cone constraints,

$$\min_{\{v_i,w_i\}_{i=1}^P} \tfrac{1}{2}\Big\| \textstyle\sum_{i=1}^P D_i X(v_i - w_i) - y\Big\|_2^2 + \lambda \textstyle\sum_{i=1}^P \big(\|v_i\|_2 + \|w_i\|_2\big), \tag{3}$$

with $v_i, w_i \in \mathbb{R}^d$. The group-$\ell_1$ penalty induces gate sparsity in the active gates. The prediction at test input $x \in \mathbb{R}^d$ is

$$\hat{y}(x) = \sum_{i \in \mathcal{A}} \mathbf{1}[u_i^\top x \geq 0] \cdot x^\top (v_i^\star - w_i^\star), \tag{4}$$

where the active set $\mathcal{A} = \{i : \|v_i^\star - w_i^\star\|_2 > \varepsilon\}$ contains the gates with non-trivial weights. We solve Eq. 3 once per EMA channel, giving 12 independent convex models (see Appendix A for implementation).

4.2 Phonetic gate analysis

We project the TIMIT (Garofolo et al., 1993) corpus through WavLM layer-9 (Section 3), normalized by the per-column norms from MNGU0 training, and let $A_{i,t} \in \{0, 1\}$ indicate whether gate $i \in \mathcal{A}_k$ fires on frame t. Each frame's hand-aligned phone label is classified into one of $\mathcal{C} = \{\text{vowel}, \text{stop}, \text{fricative}, \text{affricate}, \text{nasal}, \text{glide}, \text{silence}\}$ (Appendix B), giving the class variable $C \in \mathcal{C}$. Let $p(c) = \Pr(C = c)$ be the corpus prior, $\pi_i = \Pr(A_i = 1)$ the gate firing rate, and $p(c \mid A_i = a)$ the conditional class distribution. For each gate we report mutual information

$$I(A_i; C) = \pi_i \, D_{\mathrm{KL}}(p(\cdot \mid A_i = 1) \,\|\, p) + (1 - \pi_i) \, D_{\mathrm{KL}}(p(\cdot \mid A_i = 0) \,\|\, p)$$

in bits (using $\log_2$ KL), the normalized mutual information

$$\mathrm{NMI}(A_i; C) = I(A_i; C)/H(C) \in [0, 1],$$

the per-class enrichment

$$\mathrm{enr}(i, c) = p(c \mid A_i = 1) - p(c),$$

and a z-score

$$z(i, c) = (p(c \mid A_i = 1) - p(c))/\sqrt{p(c)(1 - p(c))/n_i},$$

testing the on-conditional against the null $p(c)$. Gates shared by multiple articulators are ranked globally by NMI averaged across sharing channels; per-articulator profiles use each channel's own MI ranking. We also record *gate sharing*, $\big|\{j : i \in \mathcal{A}_j\}\big|$, and optionally filter out gates whose sharing degree exceeds a threshold k (Appendix B).

5 Experiments

Baselines. (i) SPARC's OLS linear head; (ii) a 2-layer ReLU MLP; (iii) a 3-layer ReLU MLP. Both MLPs are trained with Adam and cosine learning-rate annealing (see Appendix A).

Convex models. Per-channel gated-ReLU networks with $P = 1000$ sampled gates and regularization λ swept over a logarithmic grid (see Appendix A). The phonetic analysis in Section 6 is performed at the per-channel *knee* (Section 5.1), the operating point that trades a small accuracy concession for the sparsity required for mechanistic interpretation.

Baseline results. Table 1 summarizes the baseline Pearson correlation results on single-speaker (MNGU0) and cross-speaker (HPRC) evaluation.

The 3-layer MLP improves MNGU0 by $+0.018$ PCC over the 2-layer no-context MLP but the two converge on HPRC, suggesting that the added depth and context does not transfer across speakers. This is consistent with the affine cross-speaker hypothesis (Cho et al., 2024), that per-frame heads learn the speaker-invariance of the SSL feature space, while context-dependent heads partially break it.

5.1 Regularization path: minimum capacity to match the linear probe

To quantify how much nonlinear capacity the convex probe needs to match the OLS linear probe, we sweep λ over $T_{\text{train}} = 100{,}000$ MNGU0 frames per channel (Appendix A). For each λ we record the number of active gates and the held-out test PCC.

Table 2: Per-channel regularization-path summary. **Active gates at 95% OLS** is the smallest count reaching 95% of OLS PCC. **Peak PCC** and **Num active** are the path maximum and its active-gate count.

Channel	OLS PCC	Num active at 95% OLS	Peak PCC	Num active at peak
UL_x	0.8291	136	0.8534	855
UL_y	0.8882	85	0.9071	920
LL_x	0.8156	101	0.8419	890
LL_y	0.9147	83	0.9313	986
LI_x	0.8465	83	0.8673	841
LI_y	0.9080	77	0.9229	978
TT_x	0.8416	114	0.8627	967
TT_y	0.9067	57	0.9201	935
TB_x	0.8446	92	0.8680	929
TB_y	0.8613	73	0.8809	957
TD_x	0.8775	77	0.8952	929
TD_y	0.9080	52	0.9218	857
Mean	**0.8701**	**85.8**	**0.8894**	**920.3**

Path geometry. At large λ the group-ℓ_1 penalty zeroes every gate and the predictor mapping becomes a constant. As λ shrinks, gates enter the active set monotonically and the count saturates near 90% of the $P = 1000$ sampled gates. PCC rises sharply through the early gate additions, peaks at a moderate sparsity level, and then declines very slightly as the unconstrained formulation introduces mild overfitting (full per-channel shown in Figure 3). The peak PCC exceeds the per-channel OLS PCC for all twelve EMA channels by an average of $+0.019$, confirming that the gated-ReLU reformulation is more expressive than the linear probe on WavLM-9 features.

Per-articulator knee. The model achieving the peak PCC is dense and not very interpretable. We instead select the per-channel *knee*, which we define to be the smallest active-gate count at which the convex probe reaches 95% of the per-channel OLS PCC. Table 2 reports the knee for each EMA channel. The mean knee is 85.8 active gates, approximately 8.6% of the $P = 1000$ sampled gates, with a range of 52 (TD_y) to 136 (UL_x). Knees are tighter for articulators whose articulatory information is concentrated in a small number of regions of the WavLM-9 space, and wider for articulators where the linear probe is already near its ceiling.

Implication for capacity selection. The knee is the natural operating point for interpretation, as we define it as the smallest active set whose predictive capacity already matches the linear probe. The phonetic analysis in Section 6 aggregates the per-channel knee active sets, totaling $1{,}030$ gate–channel pairs across the 12 channels, decreasing to 376 unique gates after deduplication.

5.2 Layer-resolved convex probe

The SPARC pipeline fixes the AAI head's input layer to WavLM-9. To test whether the convex probe's expressiveness depends on this choice, we apply the same per-channel formulation (Section 4.1) to each of the 25 WavLM-Large layers, fixing $P = 1000$ and a low regularization $\lambda_{\text{unnorm}} = 0.01/12$ (the per-channel equivalent of $\lambda_{\text{unnorm}} = 0.01$ in the multi-output formulation; see Section 5.3 for the SCNN λ normalization). All 1000 sampled gates remain active at this λ, so the comparison isolates the ceiling expressiveness of the convex reformulation rather than its sparsity behavior.

Layer-wise PCC. The convex probe beats OLS at every layer except the embedding layer (L0), where it falls 0.006 PCC below OLS: the gated-ReLU formulation needs nonlinear acoustic structure that the WavLM transformer adds but is absent from raw mel-conv features (Table 5, Figure 1). From L1 onward the PC$-$OLS gap is positive and grows from $+0.004$ (L1) to $+0.025$ at WavLM-24. In absolute terms, layer 24 nearly ties layer 9 under the convex probe (0.8827 vs. 0.8840) despite trailing it by 0.013 under OLS, indicating that articulatory information at deeper WavLM layers is preferentially encoded in nonlinear directions that the linear head cannot recover.

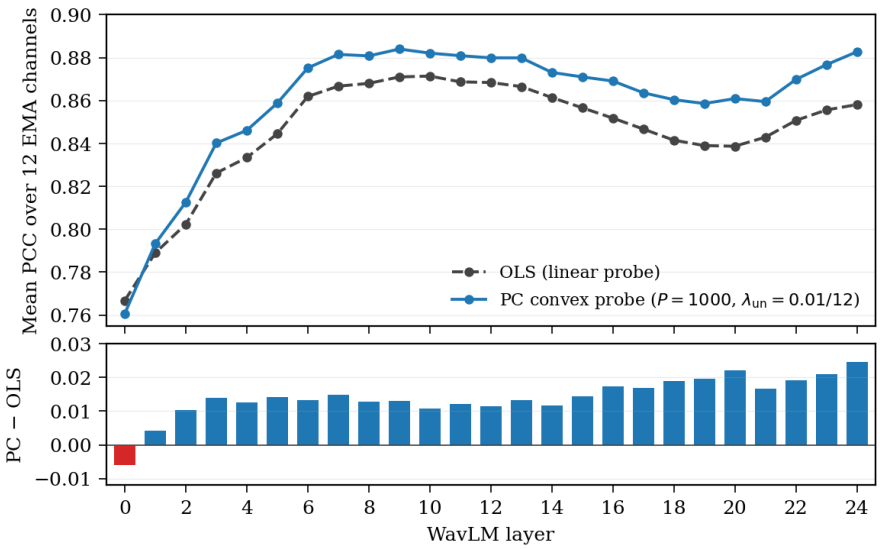

Figure 1: Layer-resolved convex probe (fold 0, $P = 1000$, $\lambda_{\text{unnorm}} = 0.01/12$). **Top:** mean PCC over the 12 EMA channels for the OLS linear probe (dashed) and the per-channel convex probe (solid), as a function of WavLM layer. The dotted vertical marks the SPARC default (L9). **Bottom:** per-layer PC−OLS gap. The convex probe beats OLS at every layer except the embedding layer (L0, red); the gap is largest at WavLM-24 ($+0.0246$), where the convex-probe accuracy nearly ties the L9 peak despite OLS underperforming L9 by 0.013.

Effective rank. The effective rank of the per-channel weight tensor $W \in \mathbb{R}^{12\times P\times d}$ is the number of singular directions accounting for 90% of $\sum \|W_{c,p,:}\|^2$. We find this is not constant across layers. It rises from 109 at L0 to a peak of 467 at L19, then collapses back to ≈ 330 at L22–L24. The layers with the largest PC−OLS gap (L20 onward, gap $\geq +0.020$) are lower-dimensional than the mid-network layers, suggesting that deeper articulatory codes are more compactly represented even as their fraction of nonlinear-only signal grows.

Firing geometry. The distribution of per-pattern firing rates on held-out frames differs sharply across layers (Figure 2). At layer 9 the 10th/90th-percentile firing rates are $(0.13, 0.85)$, with a heavy mix of selective and broadcast gates. At layer 24 they collapse to $(0.38, 0.62)$, with every gate firing roughly half the time. L24's firing-rate distribution is qualitatively distinct from every other WavLM layer (next-tightest is L22 at $(0.22, 0.77)$). The convex probe at layer 24 thus partitions WavLM space into more uniform halves, despite reaching the same predictive accuracy as the more heterogeneous layer-9 partition.

5.3 Per-channel vs. vector-output convex probe at equal capacity

The per-channel formulation (Section 4.1) solves 12 independent convex programs with no gate-sharing across articulators. A natural alternative is the vector-output (multi-output) formulation, which solves a single program with $c = 12$ output dimensions and shares both gates and weights across articulators.

Calibrating the comparison. A subtlety of the SCNN solver (Mishkin et al., 2022) is that the user-supplied regularization weight λ_{unnorm} is internally normalized by the number of training frames n and the number of outputs c, yielding the effective penalty $\lambda_{\text{scnn}} = \lambda_{\text{unnorm}}/(n \cdot c)$ on the loss. The per-channel formulation has $c = 1$ and the multi-output formulation has $c = 12$, so passing the same λ_{unnorm} to both solvers would compare them at effective penalties differing by a factor of 12. To match the two at the same λ_{scnn} we set $\lambda_{\text{unnorm,pc}} = \lambda_{\text{unnorm,mo}}/12$, so any PC−MO accuracy gap reflects gate sharing rather than a difference in regularization strength. We use this convention throughout this section and in Section 5.2.

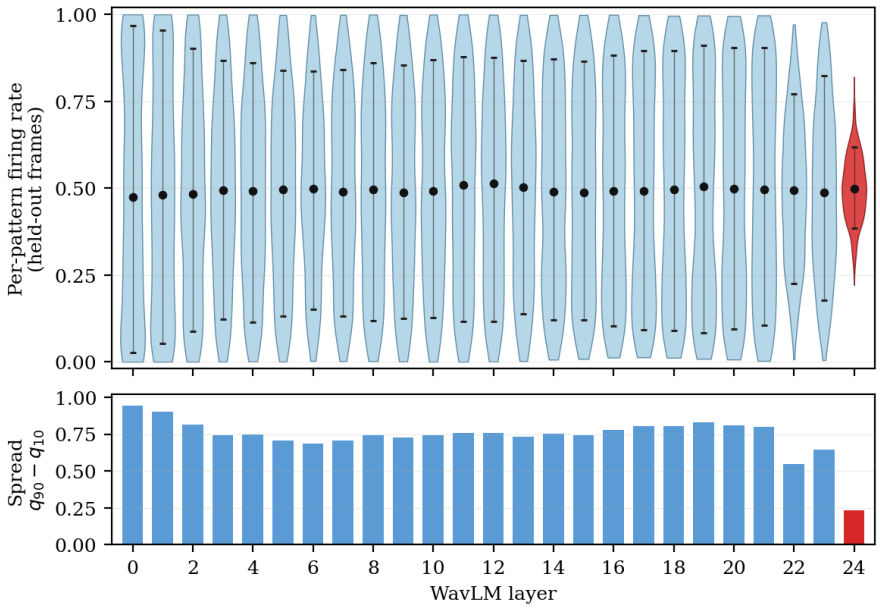

Figure 2: Distribution of per-pattern firing rates on held-out frames, across the 25 WavLM layers (fold 0, $P = 1000$). **Top:** violin per layer over the 1000 PC gates' firing rates; bars mark q_{10}/q_{90}, dots mark medians. **Bottom:** firing-rate spread $q_{90} - q_{10}$ per layer, with the L0–L23 mean as a dashed reference. Layers L0–L21 hold a wide spread of selective and broadcast gates (spread 0.69–0.94). At L24 the distribution collapses to a narrow band around 0.5 (spread 0.23, less than a third of the mean), indicating that the convex probe at the deepest WavLM layer partitions the feature space into approximately uniform halves while still attaining the same predictive accuracy as the more heterogeneous L9 partition (Figure 1).

At WavLM-9, $P = 500$, $\lambda_{\text{unnorm,pc}} = 0.01/12$ (equivalently $\lambda_{\text{unnorm,mo}} = 0.01$), the per-channel formulation reaches PCC 0.8846 while the vector-output formulation reaches PCC 0.8722 (Table 6). Sharing gates across articulators thus costs 0.0124 PCC at this operating point, collapsing nearly all of the convex probe's lift over OLS (0.8706) and leaving the vector-output formulation barely above the linear baseline. The gap suggests that the 12 EMA channels prefer distinct directional specializations within each gate, even when the firing geometry of the gate set itself can be shared; the x articulators in particular benefit disproportionately from per-channel directions (mean PC−MO gap $+0.0144$ on x vs. $+0.0105$ on y).MO accuracy is also regularization-insensitive across two orders of magnitude in λ_{unnorm}: we obtain 0.8723 at $\lambda_{\text{scnn}} = 4.6 \times 10^{-7}$ ($\lambda_{\text{unnorm,mo}} \approx 1.0$) and 0.8722 at 4.6×10^{-9} ($\lambda_{\text{unnorm,mo}} = 0.01$), indicating that the gate-sharing constraint, not the ℓ_1 penalty, is the binding constraint on the multi-output formulation. We leave a layer-resolved version of this comparison for future work.

6 Analysis

Gate budget per articulator. The active sets across the 12 channels contain 376 unique gates ($\sim$86 per channel), with 132 (35%) private to a single channel and the most universal gate appearing in 11. The per-channel top gates and their phonetic profiles are shown in Figure 5.

A small universal gate set dominates the NMI rankings. A small set of universal gates accounts for the top of nearly every per-articulator NMI ranking (Figure 4). Gate 1, a vowel detector with sharing degree 9 across the 12 EMA channels, attains the greatest NMI in the active set (NMI $=$ 0.148, $z_{\text{vowel}} = +66.2$) and heads the per-channel top-1 ranking on 9 of the 12 channels. Gate 6 is the most informative non-vowel-aligned universal gate (NMI $= 0.041$, shared by 7 channels), with strongly negative enrichment on fricatives ($z_{\text{fricative}} = -32.7$). Gate 7 is a private nasal detector for TD_y (NMI $= 0.049$, $z_{\text{nasal}} = +25.4$). Without phonetic supervision, the most informative gates

of the convex probe thus recover the classical distinctions of articulatory phonology (Browman & Goldstein, 1992), namely vowel detection, stop closure, nasal coupling, and silence.

Articulator-specific specialization. Filtering out shared gates reveals per-channel gate specializations invisible to the linear probe (Table 4). For example, on tongue-tip x (TT_x), gate 601 is fricative-selective ($z = +32.1$), isolating a tongue-tip structure specific to fricatives. On tongue-dorsum y (TD_y), gate 7 is the highest-NMI nasal detector in the active set. Additional private specializations are reported in Appendix B.

Private-gate phonetic structure. Restricting the analysis to articulator-private gates (sharing degree $= 1$) yields 132 gates whose phonetic profiles differ from the universal-gate distribution (Figure 6, Figure 7). Silence becomes the single most common dominant class (33% of private gates, against a corpus prior of 13.5%), with most silence gates suppressing rather than enriching silence. The next most common classes are vowel (23%), fricative (17%), and stop (14%), each of which shows a mix of enrichment and suppression depending on the articulator. Maximum private-gate NMI is 0.049 (gate 7, TD_y, nasal-selective), about a third of the universal-gate maximum (0.148 for gate 1), confirming that private gates encode finer distinctions than the universal vowel/stop axis. The 10 highest-NMI private gates, ranked globally (Figure 6), span six of the seven phonetic categories.

Sample-anchored interpretability. Following the geometric interpretation of Ansari et al. (2024); Pilanci (2023), each gate's breakpoint hyperplane $\{x : u_i^\top x = 0\}$ is supported by a specific subset of training frames. For any test prediction we can re-trace which gates fired and which MNGU0 training frames anchor them. Whether the anchoring frames carry genuinely phonetic structure or speaker-specific signal is a question this framework makes tractable but that we leave to future analysis.

Layer-wise comparison of probe codes. The phonetic gate analysis above is at WavLM-9, but the layer-resolved sweep (Section 5.2) shows that WavLM-24 reaches comparable predictive accuracy through a structurally distinct convex probe. On held-out frames, the per-pattern firing-vector overlap between the WavLM-9 gate bank and the WavLM-24 gate bank has mean cosine 0.64 (L9$\rightarrow$L24) and 0.71 (L24$\rightarrow$L9): the two probe banks are partially aligned but not redundant. Layer 24 thus offers an independent vantage point for the same articulator-level decoding problem; whether its universal gates carry the same phonological structure as the WavLM-9 universal vowel/stop/nasal detectors is a direct question for future phoneme-level analysis.

Limitations of the analysis. Gate sampling is random, so the gate set we analyze is one of many possible. The dominating universal gates and the shared/private split are robust across seeds, but specific gate identities are not. NMI may be biased by phone-class imbalance, and we report z-scores to make this concrete.

7 Limitations and Broader Impact

Accuracy gap to non-convex MLP. Our convex probe exceeds the per-channel OLS PCC on all twelve EMA channels by an average of $+0.019$ PCC, but does not yet reach the accuracy of the non-convex MLPs. Closing this gap will likely require larger gate samples, deeper convex architectures, or vector-output convex programs that share gates across articulators. We provide a single-layer comparison of per-channel and vector-output formulations at $P = 500$, $\lambda_{\text{unnorm,pc}} = \lambda_{\text{unnorm,mo}}/12$ in Section 5.3 (vector-output trails per-channel by 0.0124 PCC at WavLM-9, collapsing nearly all of the convex probe's lift over OLS).

Context windows. A 2-layer MLP with no temporal context generalizes slightly better cross-speaker than a 3-layer MLP with context, suggesting that context windows can introduce speaker-specific overfit. We did not investigate this systematically, and this is left for future work.

Computational cost. R-FISTA on $P = 1000$ gates with $n = 100{,}000$ training frames per channel, sweeping the full 25-point λ grid, completes in roughly twenty minutes per channel on a single GPU.

Broader impact. Articulatory representations have applications in speech-based diagnostics, brain-computer interfaces (Anumanchipalli et al., 2019), and pronunciation training. Replacing an opaque non-convex probe with a globally optimal, sample-anchored one improves auditability of these downstream systems. We do not foresee specific dual-use risk beyond those already present in SPARC.

8 Conclusion

We have shown that the AAI head of SPARC can be replaced with a convex two-layer gated-ReLU model that exceeds the OLS linear probe on every articulator while remaining globally optimal, group-ℓ_1-sparse, and structurally interpretable. A small set of universal gates (vowel, fricative, and silence) dominates the per-articulator information rankings, recovering manner-of-articulation distinctions without phonetic supervision, while articulator-private gates, including the highest-NMI nasal detector (gate 7, unique to tongue-dorsum y), capture finer specializations that a linear probe cannot represent. We see this as a step toward an articulatory-coding pipeline whose accuracy is competitive with non-convex baselines and whose internal structure is fully transparent.

References

Rayan Ansari, John Cao, Sabyasachi Bandyopadhyay, Sanjiv M. Narayan, Albert J. Rogers, and Mert Pilanci. ConvexECG: Lightweight and explainable neural networks for personalized, continuous cardiac monitoring. In *IEEE Conference Proceedings*, 2024.

Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

Ahmed Attia and Carol Espy-Wilson. Improving acoustic-to-articulatory inversion with self-supervised learning. In *Interspeech*, 2024.

Catherine P. Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49 (3-4):155–180, 1992.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 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.

Cheol Jun Cho, Abdelrahman Mohamadzadeh, and Gopala K. Anumanchipalli. Self-supervised models of speech infer universal articulatory kinematics. In *Interspeech*, 2023.

Cheol Jun Cho, Peter Wu, Tejas S. Prabhune, Dhruv Agarwal, and Gopala K. Anumanchipalli. Coding speech through vocal tract kinematics. *arXiv preprint arXiv:2406.12998*, 2024.

Kwanghee Choi et al. Self-supervised models of speech: A survey. In *Interspeech*, 2022.

Tolga Ergen and Mert Pilanci. Convex geometry and duality of over-parameterized neural networks. In *Journal of Machine Learning Research*, 2021.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. TIMIT acoustic-phonetic continuous speech corpus. Technical report, Linguistic Data Consortium, 1993.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A convolutional representation for pitch estimation. In *ICASSP*, 2018.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*, 2020.

Aaron Mishkin, Arda Sahiner, and Mert Pilanci. Fast convex optimization for two-layer ReLU networks: Equivalent model classes and cone decompositions. In *ICML*, 2022.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *ASRU*, 2021.

Mert Pilanci. From complexity to clarity: Analytical expressions of deep neural network weights via Clifford's geometric algebra and convexity. *arXiv preprint arXiv:2309.16512*, 2023.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *ICML*, 2020.

Korin Richmond, Phil Hoole, and Simon King. Announcing the MNGU0 articulatory corpus. In *Interspeech*, 2011.

Patricia Scanlon, Daniel P. W. Ellis, and Richard B. Reilly. Using broad phonetic group experts for improved speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):803–812, 2007.

Yashish M. Siriwardena and Carol Espy-Wilson. Speaker-independent speech inversion for recovery of velopharyngeal function. In *Interspeech*, 2023.

Mark Tiede, Carol Espy-Wilson, Dolly Goldenberg, Vikramjit Mitra, Hosung Nam, and Ganesh Sivaraman. Quantifying kinematic aspects of reduction in a contrasting rate production task. *Journal of the Acoustical Society of America*, 2017.

Peter Wu, Li-Wei Chen, Cheol Jun Cho, Shinji Watanabe, Louis Goldstein, Alan W. Black, and Gopala K. Anumanchipalli. Speaker-independent acoustic-to-articulatory speech inversion. In *ICASSP*, 2023.

AUTHOR CONTRIBUTIONS

Maor Fuks implemented all preprocessing, baseline training, and evaluation code for MNGU0 and HPRC. Implemented and ran the convex probe regularization-path experiments, and designed and implemented the TIMIT phonetic gate analysis. Wrote final paper.

Daniel BenShushan originated the project's central idea of applying Pilanci-style convex reformulations to acoustic-to-articulatory inversion, which then grew into the convex "Pilanci probes" framework that anchors this paper. Extended the convex probe from the SPARC default WavLM layer to a layer-resolved sweep across all 25 WavLM layers, ran the per-channel vs. vector-output comparison, and contributed the layer-wise interpretability analysis of the convex probe's geometry. Wrote final paper.

Jorgen Bergh contributed to the initial research around the topic and helped explore the problem statement. Also ran experiments to test the feasibility of building a fully end-to-end convex model for interpretability independent of WavLM, but this direction was dropped after expert consultation with Cheol Jun Cho from the Berkeley Speech Group, as the team decided the final direction was more promising.

A IMPLEMENTATION DETAILS

MLP training. The reported 2-layer MLP has architecture $1024 \to 512 \to 12$ with no temporal context. The 3-layer MLP has architecture $1024 \cdot K \to 2048 \to 1024 \to 12$ with a context window of $K = 4$ frames centered on the target. Both are trained with Adam at initial learning rate 5×10^{-4}, cosine annealing over 200 epochs, and weight decay 10^{-3}.

Convex solver and hyperparameter sweeps. We solve the gated-ReLU convex program (Eq. 3) using the R-FISTA accelerated proximal-gradient solver from the SCNN package (Mishkin et al., 2022). Reported numbers correspond to the best configuration on a held-out fold. For the convex probe we swept λ on a 25-point logarithmic grid in $[10^{-4}, 10^{0}]$ and selected the operating point at the per-channel knee (Section 5.1). We use $\varepsilon = 10^{-4}$ for the active gate tolerance.

Gate sampling. We sample $\lfloor P/6 \rfloor$ raw top principal components of the centered training features, $\lceil P/6 \rceil$ random Gaussian combinations of those components, and $\lceil 2P/3 \rceil$ standard normal directions $u \sim \mathcal{N}(0, I_{1024})$ for the gates, shared across the 12 EMA channels. Duplicates and the all-zero pattern are removed, leaving $P = 1000$ unique gates.

Note on the regularization-path training subset. The regularization-path experiment (Section 5.1, Table 2) uses a fixed $T_{\text{train}} = 100{,}000$ MNGU0 subset across all 12 channels and 25 λ values, keeping knees directly comparable. The OLS reference in Table 2 (0.8701) is computed on this same subset and is therefore lower than the full-fold OLS PCC (0.876) in Table 1. The two numbers measure the same model on different data slices. A full-fold path sweep is left for future work.

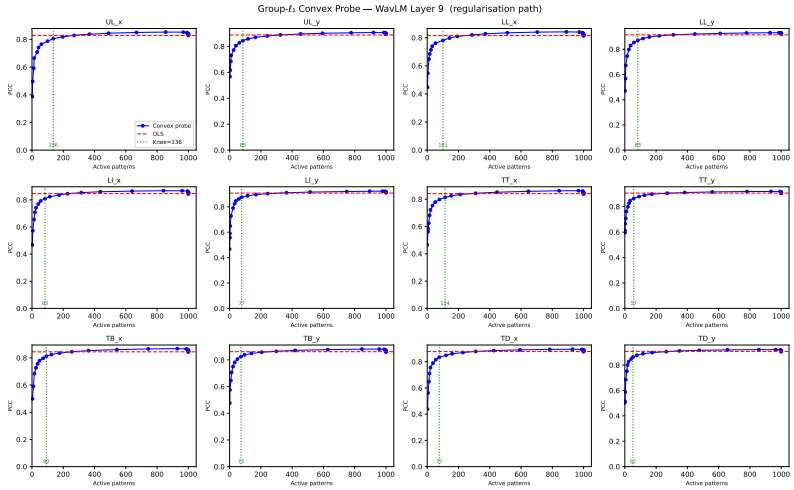

Figure 3: Regularization paths for all 12 EMA channels at $P = 1000$, $T_{\text{train}} = 100{,}000$. Each panel shows PCC against the number of active gates as λ decreases. The dashed red line is the per-channel OLS baseline. The dotted vertical line is the knee.

B Phonetic Analysis

TIMIT phone-class mapping. TIMIT (Garofolo et al., 1993) labels each frame with one of 61 ARPABET phone symbols. Following prior work collapsing the TIMIT corpus into articulatory classes (Scanlon et al., 2007), we collapse these into seven broad categories: *vowel* (`aa, ae, ah, ao, aw, ax, ax-h, axr, ay, eh, el, em, en, er, ey, ih, ix, iy, ow, oy, uh, uw, ux`), *stop* (`b, d, g, p, t, k, dx, q, bcl, dcl, gcl, pcl, tcl, kcl`), *fricative* (`f, v, th, dh, s, z, sh, zh, hh, hv`), *affricate* (`ch, jh`), *nasal* (`m, n, ng, nx`), *glide* (`l, r, w, y`), and *silence* (`h#, pau, epi`).

Top-NMI gates per articulator. Table 3 reports the top-1 gate per EMA channel ranked by NMI with phonetic class, together with the most-enriched and most-suppressed phonetic class for each.

Gate sharing distribution. Of the 376 unique gates in the union of active sets, the sharing-degree histogram is: 132 private (degree 1), 72 at degree 2, 78 at degree 3, 36 at degree 4, 20 at degree 5, 14 at degree 6, 11 at degree 7, 6 at degree 8, 4 at degree 9, 2 at degree 10, and 1 gate shared across 11 articulators. The fat tail of high-degree gates is consistent with a small set of universal phonetic detectors driving most of the inversion signal, while the substantial private fraction (35%) suggests the $P = 1000$ gate budget is used to discover articulator-specific structure.

Table 3: Top-1 highest-NMI gate per EMA channel from the active gate set ($P = 1000$, $T_{\text{train}} = 100{,}000$). **Top Class** has the largest-magnitude z-score. Over-represented ($+$) or under-represented ($-$) relative to the corpus prior.

Channel	Gate	NMI	Top class
UL_x	522	0.045	silence ($+35.0$)
UL_y	6	0.041	fricative (-32.7)
LL_x	1	0.148	vowel ($+66.2$)
LL_y	1	0.148	vowel ($+66.2$)
LI_x	1	0.148	vowel ($+66.2$)
LI_y	1	0.148	vowel ($+66.2$)
TT_x	1	0.148	vowel ($+66.2$)
TT_y	1	0.148	vowel ($+66.2$)
TB_x	1	0.148	vowel ($+66.2$)
TB_y	6	0.041	fricative (-32.7)
TD_x	1	0.148	vowel ($+66.2$)
TD_y	1	0.148	vowel ($+66.2$)

Table 4: Top-1 highest-NMI gate per EMA channel restricted to private gates (sharing degree $= 1$), comparing to Table 3.

Channel	Num private	Top private gate	Top class
UL_x	25	808	stop ($+20.2$)
UL_y	9	865	vowel ($+20.2$)
LL_x	2	999	fricative (-19.5)
LL_y	6	410	vowel ($+20.9$)
LI_x	6	916	silence (-22.1)
LI_y	15	428	silence (-16.7)
TT_x	26	601	fricative ($+32.1$)
TT_y	6	559	glide ($+20.4$)
TB_x	13	526	vowel ($+21.4$)
TB_y	14	853	silence ($+21.2$)
TD_x	3	594	fricative (-19.3)
TD_y	7	7	nasal ($+25.4$)

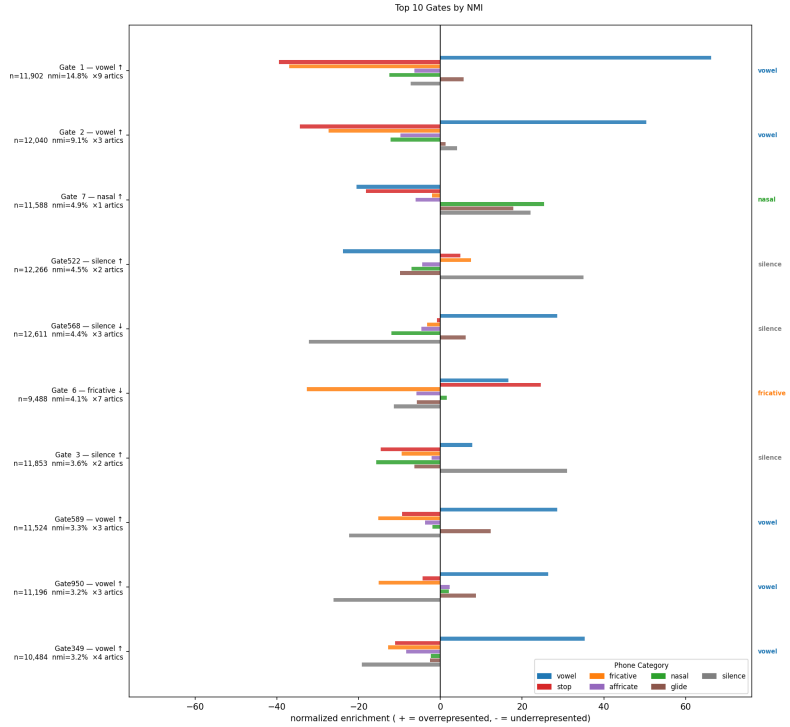

Figure 4: Top 10 gates ranked by NMI with phonetic class. Each row shows one gate's z-score values across the phone categories. Positive bars indicate over-representation, negative bars under-representation. The right margin labels the largest-magnitude category.

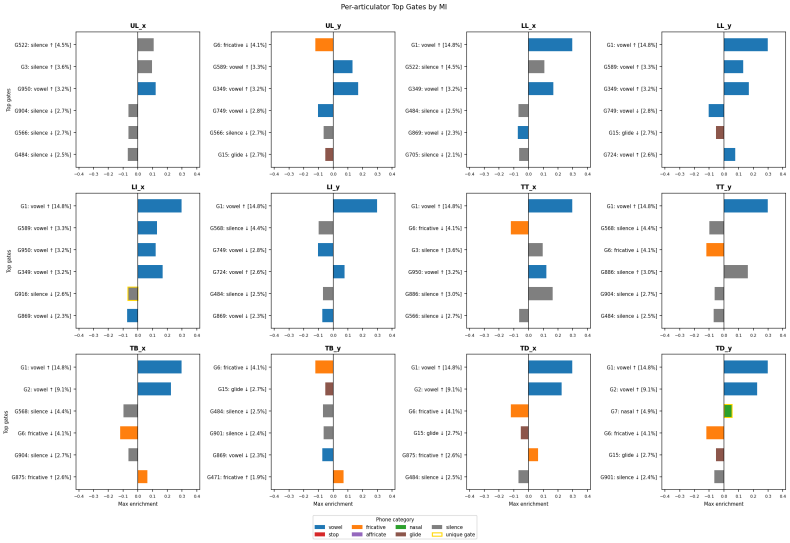

Figure 5: Per-articulator top gates ranked by per-channel mutual information. Each subplot shows up to six gates for a single EMA channel. Bar height is the maximum class enrichment over phonetic categories, color-coded by category. Yellow borders mark articulator-private gates (sharing degree $= 1$). Per-channel NMI is shown in brackets. The corresponding view restricted to private gates only is in Figure 7.

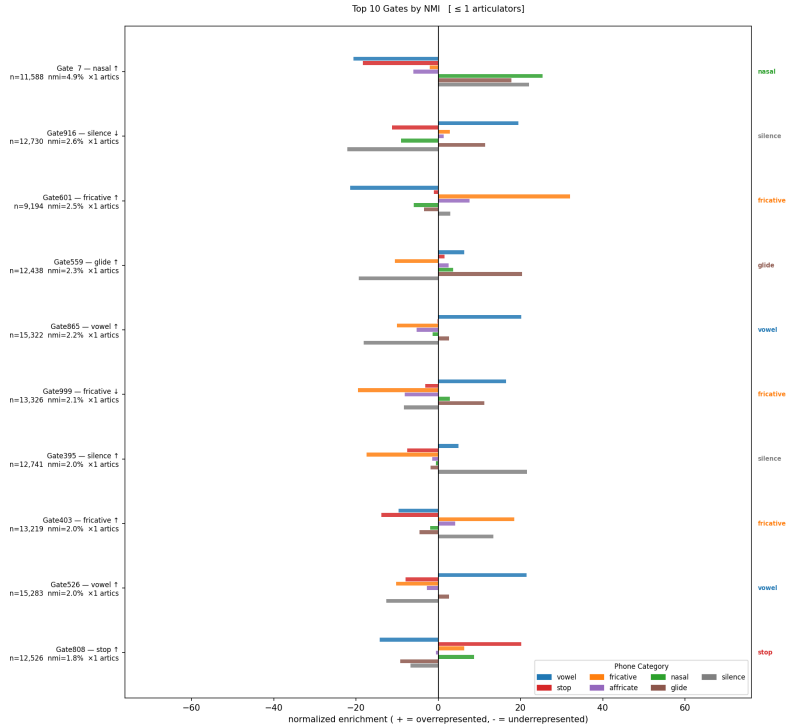

Figure 6: Top 10 private gates (sharing degree $= 1$) ranked by NMI. Each row shows one gate's z-score values across the phone categories. Compared to the unfiltered ranking (Figure 4), the private set is more diverse, spanning fricative, glide, silence, and vowel specializations in the top 10.

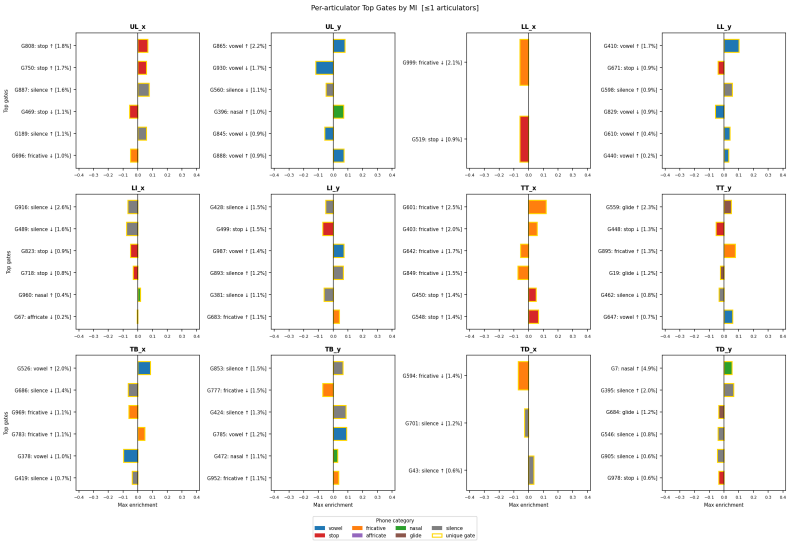

Figure 7: Per-articulator top private gates (sharing degree $= 1$), ranked by per-channel mutual information. Each subplot shows up to six gates for a single EMA channel. Bar height is the maximum class enrichment over phonetic categories, color-coded by category. Per-channel NMI is shown in brackets.

C LAYER-SWEEP CONVEX PROBE

Full Per-Channel Table. Full per-layer numbers underlying Section 5.2. **eff-50/eff-90** are the number of singular directions of $W \in \mathbb{R}^{12\times P\times d}$ accounting for $50\%/90\%$ of its total energy. q_{10}/q_{90} are quantiles of the per-pattern firing rate on held-out frames.

Table 5: Layer-resolved convex probe, full table (fold 0, $P = 1000$, $\lambda_{\text{unnorm}} = 0.01/12$).

Layer	OLS	PC	PC−OLS	eff-50	eff-90	q_{10} / q_{90}
0	0.7668	0.7607	−0.0061	27	109	0.026 / 0.968
1	0.7890	0.7933	+0.0043	38	179	0.052 / 0.954
2	0.8023	0.8126	+0.0103	43	212	0.088 / 0.902
3	0.8262	0.8402	+0.0140	48	245	0.123 / 0.867
4	0.8334	0.8461	+0.0127	50	273	0.114 / 0.860
5	0.8445	0.8588	+0.0143	53	294	0.131 / 0.838
6	0.8619	0.8752	+0.0133	55	313	0.151 / 0.836
7	0.8667	0.8815	+0.0148	60	339	0.131 / 0.840
8	0.8680	0.8808	+0.0128	59	342	0.119 / 0.861
9	0.8710	0.8840	+0.0130	56	334	0.125 / 0.854
10	0.8714	0.8821	+0.0107	60	351	0.128 / 0.870
11	0.8687	0.8809	+0.0122	58	351	0.117 / 0.877
12	0.8684	0.8799	+0.0115	59	351	0.117 / 0.876
13	0.8665	0.8799	+0.0134	60	356	0.138 / 0.867
14	0.8614	0.8731	+0.0117	62	367	0.121 / 0.872
15	0.8566	0.8710	+0.0144	67	400	0.119 / 0.864
16	0.8518	0.8691	+0.0173	71	418	0.102 / 0.882
17	0.8467	0.8636	+0.0169	77	443	0.091 / 0.895
18	0.8415	0.8604	+0.0189	80	461	0.090 / 0.896
19	0.8390	0.8586	+0.0196	80	467	0.083 / 0.911
20	0.8387	0.8609	+0.0222	73	456	0.094 / 0.904
21	0.8429	0.8595	+0.0166	67	435	0.104 / 0.904
22	0.8507	0.8698	+0.0191	43	332	0.225 / 0.772
23	0.8556	0.8767	+0.0211	48	337	0.177 / 0.824
24	0.8581	0.8827	**+0.0246**	45	330	**0.384 / 0.617**

Table 6: Per-channel (PC) vs. vector-output (MO) convex probe at WavLM-9, $P = 500$, fold 0, $\lambda_{\text{unnorm,pc}} = \lambda_{\text{unnorm,mo}}/12$.

Formulation	Mean PCC
OLS	0.8706
PC, $P = 500$	**0.8846**
MO, $P = 500$	0.8722