# A STUDY ON THE GEOMETRY OF ACCENT IN SPEECH FOUNDATION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Speech foundation models are known to encode phonetic, speaker, and higher-level linguistic information, as measured by decoding probes. Within the representation space, how these factors are geometrically organized is comparatively underexplored. We study this organization through accent, an attribute that self-supervised learning (SSL) models are not trained to represent. We ask how accent is arranged in representation space, how that arrangement changes across network depth, and whether it is shared across independently trained models, using three corpora (L2-ARCTIC, VCTK, GLOBE) and three encoders (WavLM, HuBERT, Whisper). We report three properties. First, accent variation forms a continuous, low-dimensional structure rather than isolated clusters, with softer cluster boundaries than speaker identity, populated interpolation paths, a connected accent graph, and a low-rank set of accent displacement directions (effective rank $\approx$5 on the parallel corpora). Removing the top five accent directions reduces a speaker-disjoint accent probe to near chance while leaving speaker identity largely intact, and ablating random subspaces of equal rank does neither. Second, because accent is deterministic given speaker, a probe whose training and test folds share speakers can classify accent by memorizing speaker identity. Under speaker-disjoint evaluation, where no speaker appears in both training and test folds, accent decodability is weakest in early layers and strengthens to a mid-to-late peak, whereas speaker identity peaks in the earliest layers. The apparent early accent structure under speaker-shared evaluation is largely speaker leakage, which shrinks with depth. Third, accent geometry agrees across the two SSL encoders ($\rho = 0.85$–$0.97$, permutation $p < 0.005$). On GLOBE this agreement exceeds both the untrained-encoder and acoustic (MFCC) baselines and reaches the split-half noise ceiling, while on the smaller corpora the baselines approach the SSL–SSL correlation, so our learned-convergence claim is based on GLOBE. One initial expectation fails, where accent centroids cannot be expressed as convex combinations of other accents, so continuity does not imply compositionality. The speaker-first depth ordering we observe matches prior layer-wise analyses of these encoders. We release and report all code and results.

## 1 INTRODUCTION

Speech foundation models pretrained with SSL have become the default front end for recognition, verification, and paralinguistics (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Yang et al., 2021). Layer-wise probing has characterized what these models encode. Phonetic, speaker, and prosodic information are linearly decodable to varying degrees at varying depths. Less is known about how that information is geometrically organized, that is, what shape the encoded factors take in representation space, how that shape changes across depth, and whether different models encode the same one. Geometry is not equivalent to decodable content as two layers can encode identical accent information while organizing it differently, so probe accuracy is by construction blind to organization.

Geometry provides information not captured solely by accent decodability. Accent identity representation depends on whether accent varies continuously or forms discrete clusters. The rank of the variation indicates how many directions are involved, while the layer where accent becomes separable from speaker identity can guide the choice of representation for downstream systems. Finally, cross-encoder similarity indicates whether these properties are model-specific or shared across encoders. Thus, we study the geometry directly and use accent as a probe due to its social



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.

relevance (Koenecke et al., 2020). Accent is not an explicit pretraining target, so any associated structure emerges without direct supervision from accent labels. Linguistically, accent is expected to vary continuously rather than form discrete clusters, giving us a testable prediction about its geometry.

We hypothesize that accent variation forms a continuous, low-dimensional manifold whose organization reflects phonological relationships, with mixed or transitional speakers at intermediate positions. We further hypothesize this geometry forms progressively across the network and is shared across independently trained encoders. To study these claims, we freeze all encoders for analysis. The only fitted components are a logistic-regression probe used to select an analysis layer and an unsupervised PCA projection defining the analysis subspace, neither of which generates accent representations. Each property is tested against a corresponding null hypothesis and we report all findings. Because accent is usually constant within each speaker, standard cross-validation can capture speaker identity instead of accent, so we use speaker-disjoint cross-validation for accent decoding, quantify residual speaker effects, and validate the geometry against untrained-encoder and acoustic baselines.

**Contributions.** We establish three geometric properties of accent in SSL representations and report one divergence from our predictions.

- **Accent forms a continuous, low-rank manifold.** Accent clusters are soft, interpolation paths between them stay populated, the nearest-neighbor graph is fully connected, and intrinsic dimension is below the ambient width. The directions separating accents span an effective rank of $\approx 5$ on the parallel corpora, and this rank is functional as ablating the top five accent directions collapses a speaker-disjoint accent probe while preserving speaker identity, unlike ablating random subspaces of equal rank (Section 4).
- **A speaker-invariant accent code emerges only with depth.** Under speaker-disjoint probing, a speaker-invariant accent representation is weakest early and strengthens to a mid-to-late peak, while speaker identity peaks earliest, an ordering consistent with prior layer-wise analyses of these encoders (Figure 1, Section 5). Apparent early accent structure under speaker-shared evaluation is largely speaker leakage, which decreases with depth.
- **Accent geometry is shared across independently trained models.** Accent geometry agrees across WavLM and HuBERT ($\rho = 0.85$–$0.97$, permutation $p < 0.005$). On GLOBE it exceeds both baselines and reaches the split-half noise ceiling. On the small corpora the baselines are close to the SSL–SSL value, so they cannot separate learned convergence from shared acoustics, and we base the learned-convergence claim on GLOBE (Section 6).
- **One divergence.** The manifold is continuous without being a convex simplex of accents, so continuity does not imply compositionality.

## 2 Related work

**Geometry of learned representations.** Intrinsic-dimension estimators such as TwoNN (Facco et al., 2017) show that deep representations concentrate on low-dimensional manifolds (Ansuini et al., 2019). RSA (Kriegeskorte et al., 2008), Procrustes alignment (Gower, 1975), and CCA-family methods (Hotelling, 1936; Raghu et al., 2017; Kornblith et al., 2019) compare representations across models and layers, and a recent line of work argues that independently trained models converge toward a shared representational geometry (Huh et al., 2024). Most evidence for such convergence comes from vision and language, and to our knowledge, whether it holds for an unsupervised sociophonetic attribute in speech like accent is open. We treat accent as a test case, and add a non-negative decomposition of accent centroids (Lawson & Hanson, 1995) to probe compositional structure.

**SSL speech encoder representations.** Layer-wise analyses of HuBERT (Hsu et al., 2021), wav2vec2.0 (Baevski et al., 2020), and WavLM (Chen et al., 2022) show that phonetic information is strongest in intermediate layers, that speaker and prosodic information are distributed differently across depth, and that these encoders capture articulatory and phonemic structure without supervision (Pasad et al., 2021; 2023). Speaker information lies in the lower layers of WavLM with content above it (Chen et al., 2022; Ashihara et al., 2024). The hierarchy inverts under speaker supervision (Ashihara et al., 2024), so the ordering follows from the pretraining objective rather than from speech encoders in general. These works study decodability across depth. We instead examine the geometric organization of accent, its evolution across layers, and consistency across models.

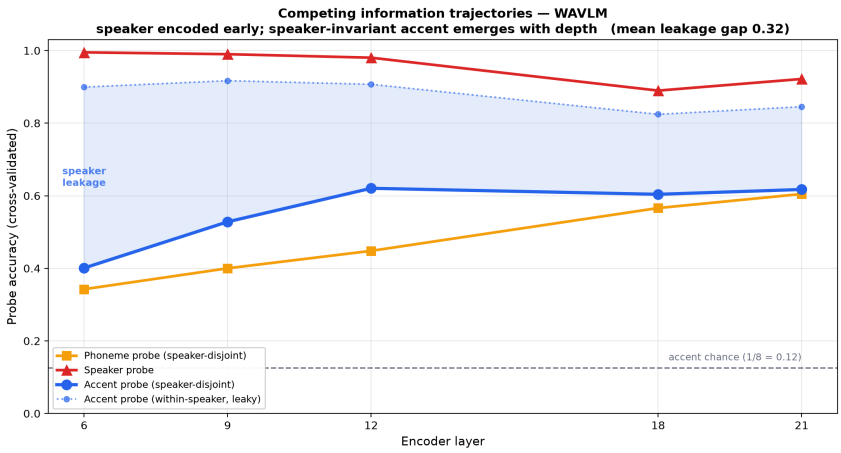



Figure 1: **Accent geometry disentangles from speaker across depth.** Competing information trajectories (WavLM, VCTK). The speaker probe (red) peaks earliest and the phoneme probe (orange) peaks latest. The speaker-disjoint accent probe (solid blue) is weakest in early layers and strengthens to a mid-to-late plateau. The dotted blue curve is the same probe under speaker-shared cross-validation, and the shaded band between the two is speaker leakage (Section 5).

**Accent and dialect in speech models.** Accent conversion (Zhao et al., 2019), accent-robust recognition (Koenecke et al., 2020), and dialect identification typically treat accent as a classification or transfer target. Recent work instead identifies accent-related subspaces in hidden representations and intervenes on them. Sun et al. (2026) estimate layer-wise accent directions in the audio encoder of a large audio language model and add them at inference for recognition adaptation, finding that accent information is strongest in the middle encoder layers. Parekh (2026) extract an accent-discriminative subspace from a fine-tuned ASR model and use it to constrain adversarial perturbations, finding that attenuating the subspace increases word error rate. Accent directions have also been used for accent-neutral synthesis (Yang & Hansen, 2026). Our analysis focuses on the geometry of accent representations, including their continuity, rank, formation across layers, and agreement across encoders, using native (VCTK), non-native (L2-ARCTIC), and crowd-sourced (GLOBE) speech.

# 3 METHOD

## 3.1 DATA, ENCODERS, AND THE ANALYSIS SUBSPACE

**Corpora.** We use three corpora. **L2-ARCTIC** (Zhao et al., 2018) comprises six non-native L2 backgrounds (Arabic, Hindi, Korean, Mandarin, Spanish, Vietnamese), four speakers each, in studio-recorded parallel speech. **VCTK** (Yamagishi et al., 2019) comprises eight native English dialects (American, Australian, Canadian, English/Southern British, Indian, Irish, Scottish, Welsh), studio-recorded and parallel. **GLOBE** (Wang et al., 2024), derived from Common Voice (Ardila et al., 2020), provides 40 self-reported accent labels from crowd-sourced, non-parallel speech. The studio corpora control recording conditions and content but offer few accents and few speakers per accent, whereas GLOBE provides a large accent pool at the cost of uncontrolled channel conditions and non-parallel content. Because GLOBE also mixes native and non-native varieties, we analyze it separately and use it as a replication dataset instead of pooling it with the controlled corpora.

**Encoders.** We extract frozen representations from WavLM-large (Chen et al., 2022) (primary, strong accent-task performance), HuBERT-large (Hsu et al., 2021) (secondary, phonetic SSL objective), and Whisper-large (Radford et al., 2023) (a negative control whose transcription objective should suppress accent-specific detail). No encoder parameter is updated. For each utterance we obtain a per-layer, mean-pooled representation, a tensor of shape $(N, L, H)$ with ambient width $H = 1024$.

**Analysis subspace.** For each (corpus, encoder) pair we fix a single analysis layer before computing any geometry, selecting the layer that maximizes the cross-validated accuracy of a logistic-regression

accent probe, and project its representations onto a 64-dimensional PCA subspace. The probe provides only a scalar layer index and the PCA is unsupervised, so the geometry we measure is a property of the frozen encoder, not a fitted accent model.

### 3.2 GEOMETRIC CRITERIA

We test the hypothesis that accent forms a continuous, low-rank manifold that is shared across encoders and emerges with depth, through a series of falsifiable tests organized into three stages.

**Stage I (structure).** Intrinsic dimension $d^\star$ is estimated with TwoNN (Facco et al., 2017) and compared to number of accent classes $K$ and to a matched Gaussian random baseline $d^\star_{\text{rand}}$. We also report the PCA dimension at $90\%$ variance. Cluster softness compares the silhouette (Rousseeuw, 1987) of accent labels to that of speaker labels (softness $> 0.5$ means accent clusters are softer than speaker clusters). Interpolation density measures minimum $k$NN density along linear (lerp) and spherical (slerp) paths between accent centroids, reported as the fraction of paths that drop below a density floor. Graph connectivity builds a $k$NN graph ($k$=15) and reports the fraction of connected accent-centroid pairs and mean path length. Transitional-speaker placement tests whether contact or mixed-variety speakers have higher neighborhood accent entropy and positive betweenness than monolithic speakers. For direction structure, we form the matrix $V$ of pairwise accent-centroid difference vectors, report its effective rank and compression ratio against the baseline, correlate each principal direction with acoustic features (F1/F2, rhoticity, VOT, pitch, speech rate) via Spearman $\rho$, and test whether phonologically similar contrasts reuse directions (within- vs. between-group cosine similarity). To test whether the low effective rank is functional, we ablate the top-$r$ accent directions and track a speaker-disjoint accent probe as a function of $r$. We estimate the directions inside each training fold, and not on the held-out speakers. We compare against two controls, random subspaces of equal rank and a speaker probe on the same ablated space. Convex decomposability solves a leave-one-out non-negative least-squares (NNLS) reconstruction of each accent from the remaining centroids, comparing reconstruction quality $\bar{\delta}$ to a label-permutation null (Wilcoxon signed-rank) and reporting the sparsity of the recovered weights.

**Stage II (formation).** We recompute the Stage I geometry per layer and track accent, speaker, and phoneme decodability to locate the depth at which each peaks. Because accent is deterministic given speaker, the accent probe uses speaker-disjoint cross-validation (folds grouped by speaker), so above-chance accuracy reflects accent structure that generalizes to unseen speakers. The per-layer leakage gap is the difference from the same probe under speaker-shared cross-validation, which isolates the apparent accent structure attributable to speaker identity. We also track per-layer global geometry ($d^\star$, continuity, connectivity) and between-layer dynamics (Procrustes distance, neighborhood Jaccard, RSA), and identify plateaus by rate-of-change falling below a data-adaptive slope.

**Stage III (universality).** We align accent centroids across encoders in a shared PCA frame and report RSA and Procrustes distance. Significance for each RSA correlation is assessed with a Mantel permutation test (permuting accent labels of one Representational Dissimilarity Matrix (RDM)) (Kriegeskorte et al., 2008), with bootstrap confidence intervals over accents, and each encoder's split-half RDM reliability gives a per-corpus noise ceiling. The ceiling is estimated from half-data RDMs, so it is lower than the reliability of the full-data RDMs. Thus, a full-data cross-encoder correlation may approach or exceed it. To distinguish learned convergence from convergence inherited from shared inputs or architecture, we compare SSL–SSL agreement against two baselines, an acoustic RDM built from MFCC accent centroids and an RDM from an untrained (randomly initialized) encoder run through the same pipeline. We test whether encoders discover the same phonological axes by correlating each encoder's best per-feature direction across encoder pairs (Appendix B).

## 4 STAGE I: ACCENT VARIATION FORMS CONTINUOUS, LOW-RANK MANIFOLD

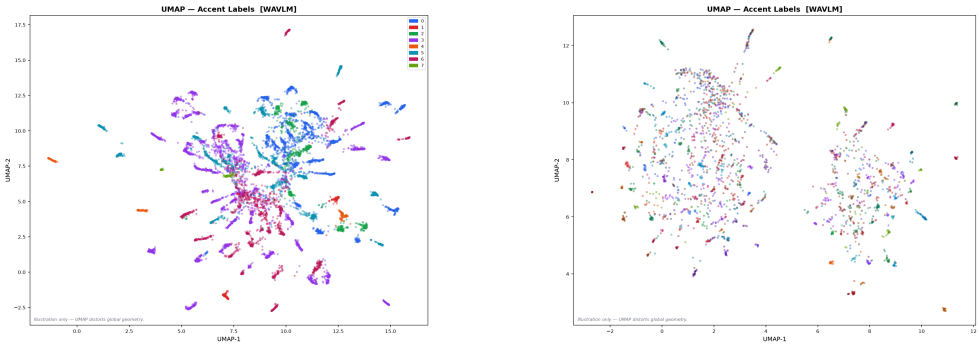
### 4.1 CONTINUITY

Continuity is not directly observable from finite samples, so we test for the signatures of a discrete organization. This includes hard cluster separation, depopulated interpolation paths, and a fragmented accent graph, which were not observed. Across all corpora and encoders, accent clusters are softer than

speaker clusters (WavLM softness 0.57–0.72 in Table 1, 0.57–0.93 across all encoders, Appendix A), so accent identity does not partition the space like how speaker identity partially does. Both linear and geodesic interpolation paths between accent centroids remain populated (drop fraction 0 in all configurations), and the $k$NN accent graph forms a single connected component with all centroid pairs reachable. We thus find no evidence of the discreteness that would falsify continuity. Figure 2 shows the resulting overlapping structure.

Table 1: Stage I summary, WavLM (analysis layer in parentheses). Intrinsic dimension $d^\star$ is reported against number of accent classes $K$ and matched baseline $d^\star_{\text{rand}}$. Full tables are in Appendix A.

| Metric | L2-ARCTIC (L6) | VCTK (L9) | GLOBE (L6) |
|---|---|---|---|
| $K$ (accent classes) | 6 | 8 | 40 |
| TwoNN $d^\star$ / $d^\star_{\text{rand}}$ | 21.51 / 31.68 | 11.55 / 32.83 | 6.74 / 32.91 |
| ratio | 0.68 | 0.35 | 0.20 |
| PCA dims @90% var. (of 1024) | 40 | 45 | 48 |
| Softness (accent vs. speaker) | 0.71 | 0.57 | 0.72 |
| Interpolation drop fraction (lerp) | 0.00 | 0.00 | 0.00 |
| Graph connectivity (pairs) | 1.00 | 1.00 | 1.00 |
| Direction-matrix eff. rank / random | 5 / 12.0 | 5 / 19.95 | 21 / 54.0 |
| as fraction of $K(K-1)/2$ pairs | 0.333 | 0.179 | 0.027 |




(a) VCTK, native dialects.

(b) GLOBE, 40 crowd-sourced accents.

Figure 2: The accent manifold (WavLM, UMAP (McInnes et al., 2020), illustration only. UMAP distorts global geometry). Accent labels form overlapping, continuous structure not disjoint clusters, consistent with the measurements of Table 1. HuBERT/Whisper projections are in Appendix D.1.

### 4.2 Low dimension and functional low rank

Both properties hold across all corpora and encoders (Table 1). The intrinsic dimension remains below the ambient width throughout ($d^\star \approx 6$–$22$ vs. $H = 1024$). Relative to a matched Gaussian baseline, it ranges from 0.68 on L2-ARCTIC to 0.20 on GLOBE, indicating that accent variation occupies a small fraction of the available dimensions. The accent-difference directions show a similar pattern. Their effective rank is low relative to the random baseline in all three corpora, using only 0.33, 0.18, and 0.03 of the $K(K-1)/2$ available pairwise directions. This suggests that accent variation is captured by a small set of shared directions rather than a separate direction for each accent pair. The narrower "$d^\star < K$" criterion is informative only at large $K$, holding on GLOBE ($d^\star = 6.7 \ll 40$) but not at $K$=6–8, so we do not rely on that criterion (Appendix C.4).

Effective rank is a spectral statistic, so we test whether these directions carry accent information. On the parallel corpora, a speaker-disjoint accent probe moves to near chance by $r \approx 5$ (VCTK WavLM $0.49 \to 0.30$, chance 0.33. L2-ARCTIC WavLM $0.44 \to 0.17$, chance 0.167) when ablating top-$r$ accent directions, while ablating random subspaces of the same rank leaves the probe near its original value (VCTK WavLM stays $\approx 0.49$) and the speaker probe is unaffected ($\approx 1.0$). The main result is the contrast between accent and speaker. Removing the five-dimensional accent subspace removes

linear accent decodability while leaving speaker identity intact. Removing a random subspace of the same rank does not affect either. This double dissociation suggests that the low-rank structure is the subspace in which accent information is represented. On GLOBE the speaker-disjoint accent signal is weak (probe $\approx 0.25$ at $r{=}0$ against an empirical chance rate of $0.07$), and the ablation and random curves nearly coincide. The GLOBE ablation uses the 31 accents with enough speaker coverage for speaker-disjoint cross-validation, and we base this result on the parallel corpora.

### 4.3 Phonological structure and transitional speakers

Individual directions reach high correlations with acoustic features on the parallel corpora, with $|\rho|$ up to $0.98$ for F1/F2 and rhoticity. Each correlation is computed over $K{=}6$–$8$ centroids, where values this large are common under the null (Appendix C.6), so they are not interpreted. On GLOBE, with $K{=}40$, the largest is $0.74$. Direction reuse is measurable at all three accent counts, and phonologically similar contrasts reuse directions more than dissimilar ones, comparing within-group against between-group similarity ($p = 0.036$ for WavLM on L2-ARCTIC and $p < 10^{-3}$ for both SSL encoders on GLOBE). Thus, the phonological claim is based on reuse structure instead of any individual axis.

On GLOBE, speakers whose self-reported labels denote contact or mixed varieties occupy higher-entropy neighborhoods than speakers with monolithic labels (entropy $1.77$ vs. $1.57$, $p < 10^{-3}$ for WavLM, similar for HuBERT). This is consistent with intermediate positioning on the manifold, and all nine cells are shown in Appendix D.2. The complementary betweenness test does not return a positive signal. On L2-ARCTIC and VCTK the centroid-betweenness statistic is uniformly negative given the small number of accents and speakers, which we treat as a limit on what these corpora can resolve, not as evidence against placement.

### 4.4 Convex decomposability is not supported

We predicted that each accent centroid would be a convex combination of the others, placing contact varieties in the simplex spanned by their contributing accents. Our results do not support this. Leave-one-out NNLS reconstruction quality $\bar{\delta}$ never separates from its permutation null. Wilcoxon $p$ values are $0.34/0.67/0.59$ on L2-ARCTIC, $0.58/0.68/0.44$ on VCTK, and $0.13/0.07/0.50$ on GLOBE, none below $0.05$. The recovered weights are denser than a compositional code predicts. On GLOBE, each accent is reconstructed using an average of $\approx 28$ basis accents. Part of this reflects limitations of the NNLS formulation instead of absence of geometric structure, as the leave-one-out simplex is degenerate at small $K$ and underdetermined at large $K$ (Appendix C.4). It cannot explain the GLOBE null on its own, where $K$ is large enough to avoid the small-$K$ degeneracy. Thus, it suggests the accent manifold is continuous and low-rank but not convexly decomposable into a set of basis accents, so continuity and compositionality are distinct properties and only continuity is supported.

## 5 Stage II: Accent geometry disentangles speaker across depth

Stage I shows the accent manifold is continuous and low-rank. This section asks how it forms. Under speaker-shared cross-validation accent structure appears present early, but this does not hold under speaker-disjoint analysis. Because accent is deterministic given speaker, a probe trained and tested on overlapping speakers can classify accent by memorizing speaker identity, which is most decodable in the early layers where the apparent accent signal lay. Under speaker-disjoint folds that early signal largely disappears and a speaker-invariant accent representation emerges with depth (Section 5.1).

### 5.1 Formation order across depth

Under speaker-disjoint probing, the accent trajectory reverses relative to the speaker-shared setting. Speaker-invariant accent accuracy is lowest in the early layers and peaks mid-to-late in all nine corpus–encoder pairs, never peaking at the first sampled layer (Table 2, Figure 1, with remaining cells in Appendix D.3). On VCTK WavLM, speaker-disjoint accent accuracy increases from $0.40$ at layer 6 to a plateau of $\approx 0.60$–$0.62$ at layers 12–21, whereas the speaker-shared probe stays high early ($0.90$–$0.92$ at layers 6–12) and declines to $0.82$–$0.85$ by layers 18–21. The gap between the two curves is speaker leakage. It is largest on the small corpora (mean $0.31$–$0.46$), where each accent has three to four speakers, and smallest on GLOBE (mean $0.17$–$0.21$), where more speakers per accent

Table 2: Stage II formation order under speaker-disjoint probing (WavLM/HuBERT/Whisper), reported by peak layer. The final row is the mean speaker-leakage gap (speaker-shared minus speaker-disjoint accent accuracy, averaged over layers). †L2-ARCTIC phoneme figures index alignment to native phoneme targets and understate phoneme information for L2 speech (Section 7). ‡The speaker probe attains its maximum at the shallowest sampled layer and decreases monotonically, so these entries bound the peak location.

| Peak layer | L2-ARCTIC | VCTK | GLOBE |
|---|---|---|---|
| Speaker‡ | ≤6 / ≤6 / ≤8 | ≤6 / ≤6 / ≤8 | ≤6 / ≤6 / ≤8 |
| Accent (speaker-disjoint) | 18 / 18 / 28 | 12 / 12 / 24 | 12 / 12 / 24 |
| Phoneme† | 18 / 21 / 28 | 21 / 21 / 31 | 21 / 21 / 28 |
| Mean leakage gap | 0.39 / 0.46 / 0.31 | 0.32 / 0.31 / 0.32 | 0.20 / 0.17 / 0.21 |

reduce the opportunity to exploit speaker-specific information. This pattern is consistent with greater speaker leakage when fewer speakers represent each accent.

Speaker identity is maximal at the shallowest layer we sample (layers 6–8) and decreases with depth in all nine cells. Since the maximum may lie below the sampled range, the trajectory is more of a monotone decay instead of a peak. The phoneme probe rises monotonically to a late peak (layers 18–31). The early emergence of speaker information is consistent with past layer-wise analyses of these encoders (Chen et al., 2022; Ashihara et al., 2024), providing validation on our probing setup. For accent, Sun et al. (2026) sweep injection layers across a 32-layer audio encoder on VCTK and L2-ARCTIC and locate the effective window at layers 15 to 19. The model family and the measurement both differ from ours, so the comparison is loose, but the mid-depth location agrees with the accent peaks in Table 2. Under speaker-disjoint probing, speaker information develops first, followed by a speaker-invariant accent representation in the middle-to-late layers, and finally phonetic detail, rather than the early-accent interpretation suggested by the speaker-shared analysis.

### 5.2 Erasure and plateau controls

Within-accent linear concept erasure (Belrose et al., 2023) removes speaker identity while preserving accent centroids by construction, so the centroid-based Stage I geometry and Stage III RSA are unaffected and the low-rank and convergence results do not depend on within-accent speaker variation. On L2-ARCTIC continuity is also preserved directly, with interpolation paths still populated and the accent graph still connected after erasure. On the larger corpora the speaker count exceeds the subspace dimension, so full erasure over-collapses the point cloud and the point-cloud metrics become uninformative (Appendix C.1). A data-driven plateau detector (rate of change of Procrustes distance, neighborhood Jaccard, or RSA falling below a small slope) does not identify any plateau for any metric, corpus, or encoder (Appendix C.7). Therefore, local neighborhoods continue to reorganize with depth while the global organization reported in Stage I is preserved.

## 6 Stage III: Cross-model agreement in accent geometry

Stages I and II show that accent geometry is continuous and emerges with depth. This section asks if the geometry is model-specific or shared across independently trained encoders, providing a socio-phonetic test of the convergence hypothesis (Huh et al., 2024). If independently trained models converge, two SSL encoders with different objectives and data pipelines should arrange the same accents in the same relative configuration despite not being trained on accent. Agreement does not imply learned convergence, since it could follow from shared acoustic input or architectural priors, which is why we report the MFCC and untrained-encoder baselines alongside it.

### 6.1 Convergence exceeds baselines on the large corpus

Accent geometry agrees across WavLM and HuBERT, with $\rho = 0.85$ (L2-ARCTIC), $0.93$ (VCTK), and $0.97$ (GLOBE), all Mantel $p < 0.005$ and, on GLOBE, a small bootstrap interval ($[0.95, 0.97]$) (Table 3, Figure 3). The baselines separate learned convergence from inherited agreement differently

Table 3: Stage III cross-model alignment with baselines. WavLM–HuBERT representational similarity (all Mantel $p < 0.005$) is compared against an untrained-encoder RDM, an acoustic (MFCC) RDM, and the split-half noise ceiling. The margin row is SSL–SSL agreement minus the higher of the two baselines. Baseline entries are the strongest SSL value per corpus.

| | L2-ARCTIC ($K$=6) | VCTK ($K$=8) | GLOBE ($K$=40) |
|---|---|---|---|
| RSA $\rho$ WavLM–HuBERT (SSL–SSL) | 0.850 | 0.930 | 0.966 |
| RSA $\rho$ WavLM–Whisper | 0.675 | 0.707 | 0.567 |
| RSA $\rho$ HuBERT–Whisper | 0.839 | 0.712 | 0.566 |
| Untrained-encoder baseline (best SSL) | 0.62 | 0.69 | 0.48 |
| MFCC acoustic baseline (best SSL) | 0.82 | 0.85 | 0.83 |
| Split-half noise ceiling (SSL) | 0.99 | 0.99 | 0.89 |
| SSL–SSL margin over higher baseline | 0.03 | 0.08 | 0.14 |

across corpora. On GLOBE, agreement is higher than both baselines (SSL–SSL 0.97 against untrained 0.48 and MFCC 0.83) and exceeds the split-half noise ceiling (0.89), with the bootstrap interval lying above it. Since the noise ceiling is a conservative reference (explained in Section 3.2), this indicates agreement at the limit of the reliability we can estimate. Independently trained SSL encoders therefore share an accent geometry on the largest and least-controlled corpus, and we treat GLOBE as the primary evidence for representational convergence.

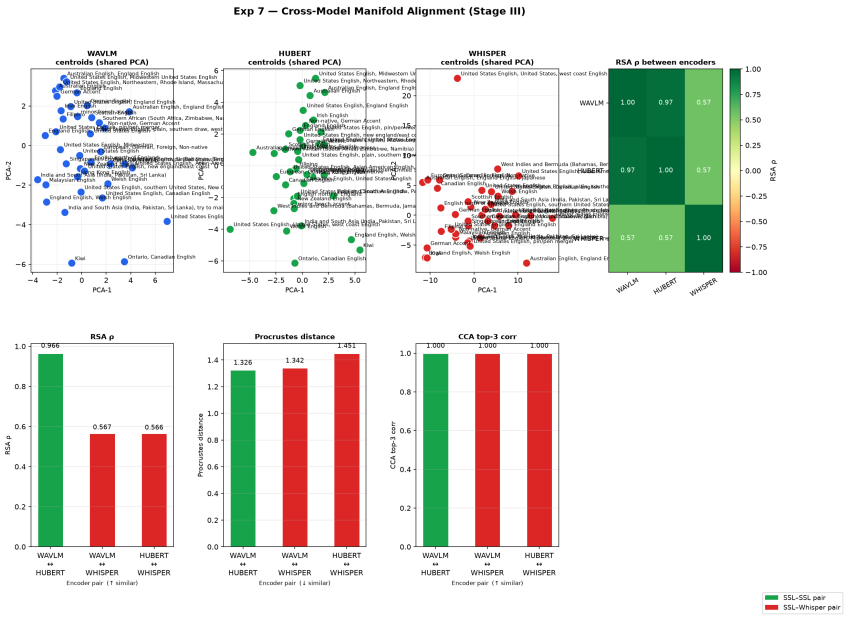



Figure 3: **Cross-model manifold alignment on GLOBE, the corpus carrying the convergence claim.** WavLM and HuBERT centroid configurations align closely (RSA $\rho = 0.97$, Mantel $p \leq 10^{-4}$), the Whisper control much less closely ($\rho = 0.57$ against both). Parallel corpora are in Appendix D.4. CCA (bottom right) is degenerate at this centroid count and has no claim (Section 7).

### 6.2 The acoustic baseline on the controlled corpora

On the controlled corpora, high SSL–SSL agreement is mainly explained by the acoustic baseline. On VCTK, SSL–SSL is 0.93 versus an MFCC baseline of 0.85 (margin 0.08), while on L2-ARCTIC the gap is smaller (SSL–SSL 0.85 vs. HuBERT–MFCC 0.82). With only $K$=6–8 accents, the RDM contains just 15–28 independent entries, allowing an acoustic representation to recover most of the geometry. Thus, the learned-convergence claim only applies to GLOBE, where the transcription-trained control behaves as expected (Whisper–SSL 0.57, below SSL–SSL 0.97 and below its own

noise ceiling), consistent with an accent-normalizing objective. L2-ARCTIC is an exception, where Whisper nearly matches the SSL–SSL agreement (HuBERT–Whisper $0.84$ vs. $0.85$) and achieves the highest speaker-disjoint accent accuracy ($0.88$, compared with $0.68$ for WavLM and $0.60$ for HuBERT). We attribute this to small-$K$ effects (Appendices C.4 and C.6), with similar per-feature inversions on L2-ARCTIC.

## 7 Discussion, limitations, and threats to validity

**Implications for speech representation learning.** Accent varies continuously in these encoders, suggesting that discrete accent labels do not fully capture how the representations are organized. This continuity does not imply compositionality as accents are likely not mixtures of neighboring accents. The variation is also low-dimensional, with about five directions accounting for accent information on the parallel corpora. Depth is also important as accent becomes separable from speaker identity in the middle/later layers, while earlier layers retain more speaker information. These patterns hold across three datasets and two independently trained encoders. On GLOBE, the two encoders show similar accent geometry beyond the acoustic and untrained-encoder baselines. These properties are not captured by probe accuracy alone and require speaker-disjoint evaluation and acoustic controls.

**Threats to validity.** To address the main speaker–accent confound, where accent is constant in a speaker and speaker-shared evaluation may reflect speaker identity, we use speaker-disjoint probing throughout Stage II. Because L2-ARCTIC and VCTK are parallel, lexical content is held fixed across speakers there. GLOBE is non-parallel, so we rely on cross-corpus consistency to prevent a content–accent confound. At $K{=}6$ (L2-ARCTIC) several statistics are poorly conditioned, including the "$d^\star < K$" criterion, the NNLS simplex analysis, centroid betweenness, and CCA (which we exclude), and we rely on GLOBE ($K{=}40$) where these are more informative. For Stage III the risk is that RSA agreement can be inherited from shared acoustics. On the small corpora the MFCC baseline is close to the SSL–SSL value, so we base the learned-convergence claim on GLOBE. The L2-ARCTIC phoneme probe measures alignment to native phoneme targets and likely underestimates phoneme content in L2 speech, so the "refined late" claim is based on the monotone-with-depth trajectory, corroborated on native VCTK (Appendix C.5). Appendix C has further details, including two corrected reporting verdicts that do not affect our conclusions. Our ablation shows that the accent subspace can be removed while preserving a speaker probe. We do not establish whether this impacts recognition, and is not measured here. Parekh (2026) report in a preprint that attenuating an accent subspace in a fine-tuned ASR model increases word error rate, but their setting differs from ours in encoder, layer, and objective. Their subspace is estimated at an early layer selected by probe accuracy, where Section 5 finds substantial speaker leakage in accent decoding. It may therefore also contain speaker variance. We have not tested whether this affects their result and leave the question open.

**Scope.** All representations are mean-pooled per utterance in a selected layer per encoder for Stage I/III. We do not model within-utterance dynamics and layer selection by probe accuracy could interact with geometry, mitigated by fixing the selection criteria, but a layer-agnostic analysis is future work.

## 8 Conclusion

We used accent as a probe to study how one attribute is organized in frozen speech encoders under speaker-disjoint evaluation. Accent variation forms a continuous, low-dimensional representation, with about five shared directions carrying accent information on the parallel corpora, and removing those directions moves a speaker-disjoint accent probe to near chance while leaving speaker identity intact. Across layers, the speaker-invariant accent representation emerges gradually with depth, while the early accent signal under speaker-shared probing is explained by speaker leakage. On GLOBE, WavLM and HuBERT show closely aligned accent geometry that exceeds both acoustic and untrained-encoder baselines, though on the smaller parallel corpora, we cannot distinguish learned convergence from shared acoustics. One expectation did not hold, where accent manifold is continuous but not a convex simplex of accents. While previous work has shown that accent can be decoded from speech representations, our analyses ask how it is organized, when a speaker-invariant form emerges, and whether that organization is shared across independently trained speech foundation models.

### AI USE STATEMENT

In this work, we used generative AI tools for editing and proofreading manuscript text, and for assisting with code review and debugging, under author direction. We have not used generative AI tools for generating research ideas, designing experiments, or producing any experimental results or claims independent of author verification, and literature discovery/summarization and automated plagiarism or novelty checking are not applicable to this work. We have reviewed all AI-assisted work. All experiments were designed, executed, and their results verified by the authors prior to any AI involvement. AI-assisted edits to code were tested against the underlying JSON result files to confirm numerical consistency, and all manuscript text produced or modified with AI assistance was read and edited by the authors for accuracy and correctness of scientific claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

Accent is a socially meaningful, demographic-adjacent attribute. All three corpora are public research datasets collected with speaker consent. GLOBE accent labels are self-reported and are analyzed as given, without collapsing them into externally imposed categories. We study aggregate geometric structure only and make no attempt to identify individual speakers. We note that a clearer understanding of how accent is encoded could inform both bias mitigation and accent-based profiling. We believe transparent measurement of what frozen public models already encode weighs in favor of publication.

### REPRODUCIBILITY STATEMENT

All encoders are loaded from public checkpoints with no weight modification (except the untrained-encoder baseline, which re-initializes weights and is marked as such). The only fitted components are a logistic-regression probe and a PCA projection. Accent probes in Stage II and the rank ablation use speaker-disjoint cross-validation (folds grouped by speaker). Random seeds are fixed (seed $= 42$), embeddings are cached so that all stages share identical representations, and the layer-selection criterion is documented and fixed before any geometry analysis. Per-encoder, per-corpus numerical tables and analysis code, including the speaker-disjoint probe trajectories, the rank-ablation sweeps, the layerwise speaker-erasure control, and the Stage III baseline and significance results, are released with the paper. All encoders are used under their public licenses (WavLM-large and Whisper-large-v3 under MIT, HuBERT-large under Apache-2.0). L2-ARCTIC is used under its research-use license agreement, VCTK under CC BY 4.0, and GLOBE_V2 under CC0-1.0 (derived from Mozilla Common Voice, CC0-1.0). Our released code is distributed under the MIT license.

## REFERENCES

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 2020.

Takanori Ashihara, Marc Delcroix, Takafumi Moriya, Kohei Matsuura, Taichi Asami, and Yusuke Ijima. What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10166–10170. IEEE, 2024.

Alexei Baevski, Henry 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.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017.

John C. Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

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.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *The Forty-First International Conference on Machine Learning*, 2024.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *The Thirty-Sixth International Conference on Machine Learning*, pp. 3519–3529. PMlR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, Volume 2 - 2008, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008.

Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*. SIAM, 1995.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL `https://arxiv.org/abs/1802.03426`.

Swapnil Parekh. Aces: Accent subspaces for coupling, explanations, and stress-testing in automatic speech recognition, 2026. URL `https://arxiv.org/abs/2603.03359`.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *The Fortieth International Conference on Machine Learning*, 2023.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Jinuo Sun, Yang Xiao, Sung Kyun Chung, Qiuchi Hu, Gongping Huang, Eun-Jung Holden, and Ting Dang. Activation steering for accent adaptation in large audio language models. In *Proc. Interspeech*, 2026.

Wenbin Wang, Yang Song, and Sanjay Jha. GLOBE: A high-quality English corpus with global accents for zero-shot speaker adaptive text-to-speech. *arXiv preprint arXiv:2406.14875*, 2024.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). University of Edinburgh, The Centre for Speech Technology Research, 2019.

Mu Yang and John H. L. Hansen. Activation steering for accent-neutralized zero-shot text-to-speech, 2026. URL `https://arxiv.org/abs/2603.05977`.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal performance benchmark. In *Proc. Interspeech*, 2021.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-ARCTIC: A non-native English speech corpus. In *Proc. Interspeech*, 2018.

Guanlong Zhao, Shaojin Ding, and Ricardo Gutierrez-Osuna. Foreign accent conversion by synthesizing speech from phonetic posteriorgrams. In *Proc. Interspeech*, 2019.

# A Full Stage I tables

Per-encoder Stage I metrics for all three corpora (TwoNN $d^\star$, softness, interpolation, connectivity, effective rank, compression ratio, phonetic-alignment and direction-consistency statistics, and the full NNLS decomposition tables) are provided in the released `stage1_report_*.md` files and `results_stage1_*.json`. The cross-encoder summaries are softness $0.57$–$0.93$ (all $> 0.5$), connectivity $= 1.0$ throughout, compression ratio $\leq 0.33$ throughout, and NNLS Wilcoxon $p > 0.05$ in all nine (corpus $\times$ encoder) cells.

# B Full Stage III tables

Per-feature best-alignment and cross-encoder direction-consistency matrices for the shared-directions analysis are in `results_stage3_*.json`. Cross-encoder direction consistency is high for some axes (e.g. rhoticity for WavLM–HuBERT on VCTK, $\rho = 0.83$, and F1 for HuBERT–Whisper on L2-ARCTIC, $\rho = 0.94$) but low or sign-inconsistent for others, so we do not claim a fully shared phonological basis. The cross-model result we rely on is the RSA convergence in Table 3.

# C Diagnostics, degeneracies, and corrected verdicts

Several initially planned analyses proved statistically or geometrically inappropriate once evaluated on small accent inventories, and two reporting verdicts in our pipeline were mis-specified. We document them here with the rationale for excluding each from the main scientific claims, so that the main text can report the outcome while the experiments remain on record.

## C.1 Within-accent speaker erasure

We document two Stage II diagnostics. First, the within-accent speaker erasure of Section 5 is well-conditioned only when the number of speakers is below the PCA dimension. On VCTK and GLOBE the speaker count exceeds the $64$-dimensional subspace, so the within-accent speaker subspace spans the space and full erasure collapses each accent onto its centroid, driving accent silhouette to a degenerate $1.0$, interpolation-drop to $1.0$, and connectivity to $0$ at every layer. These

values are arithmetic artifacts of an over-determined erasure and they are not interpreted. The erasure is informative on L2-ARCTIC (where speakers are fewer than dimensions), and the centroid-preservation argument, which is exact, covers the low-rank and RSA claims on all corpora. Second, an earlier version of this work used a per-accent neighborhood-consistency figure as its central Stage II result. That metric shares the speaker confound. With only three to four speakers per accent on the parallel corpora, a point's $k$ nearest neighbors are dominated by same-speaker and therefore same-accent utterances, so the high early consistency it showed is largely a same-speaker-neighborhood effect. Consistent with this, the same metric is low throughout on GLOBE ($0.13$–$0.29$), where many speakers per accent prevent same-speaker neighbors from dominating. We therefore dropped that figure in favor of the speaker-disjoint probe trajectory (Figure 1).

### C.2 Stage III baselines and significance

All Stage III RSA correlations are tested with a Mantel permutation test ($10{,}000$ permutations of accent labels) and reported with bootstrap confidence intervals over accents. Every SSL–SSL correlation has $p < 0.005$. The acoustic baseline is an RDM of per-accent MFCC (mean and standard deviation) centroids in a standardized feature space. The untrained baseline re-initializes the encoder weights and runs the identical extraction and subspace pipeline, isolating architecture and front end from learned structure. Per-encoder split-half RDM reliability gives the noise ceiling. The full baseline table (all encoder pairs against both the MFCC and untrained-encoder baselines, together with the per-encoder split-half noise ceilings, per corpus) is released in `stage3_baselines_*.json`. Across corpora, SSL–SSL agreement exceeds both baselines and, on GLOBE, exceeds the split-half noise ceiling (which, being estimated from half-data RDMs, understates full-data reliability), while on the small controlled corpora the MFCC baseline is close to the SSL–SSL value.

### C.3 CCA exclusion

A naive CCA alignment of the accent centroids returns a top-3 canonical correlation of $1.0$ for every encoder pair on every corpus, which suggests a degeneracy of the setup. We ran CCA on the per-accent centroid matrices ($K \leq 40$ points) in the $64$-dimensional PCA subspace, without regularization, so the number of points is at or below the subspace dimension and three canonical variates fit any two configurations almost exactly. At $K$=6, six points in $64$ dimensions can be aligned under little constraint. We therefore omit CCA from Table 3 and the cross-model conclusions is based on RSA and Procrustes, which use the full pairwise dissimilarity structure. Regularized or dimensionality-matched CCA on many-sample (per-utterance) representations would be more appropriate for future work.

### C.4 Small-$K$ degeneracy

At $K = 6$ (L2-ARCTIC) several statistics are unreliable, namely the "$d^\star < K$" rule (which cannot hold when $K$ is below the number of continuous degrees of freedom), the leave-one-out NNLS simplex (which collapses to uniform mixtures), the centroid-betweenness test (uniformly negative), the per-axis acoustic-alignment correlations (whose rank-permutation null is too coarse to test and whose large coefficients are the expected behavior of the estimator at $K$=6, Appendix C.6), and CCA (Appendix C.3). We flag each in place and rely on GLOBE ($K = 40$), where these tests are more informative.

### C.5 The L2-ARCTIC phoneme probe measures native-target alignment

The phoneme-probe accuracies on L2-ARCTIC ($0.17$–$0.30$) are below native VCTK ($0.37$–$0.62$) and only slightly above chance. This gap is an artifact of the labeling pipeline. Targets are forced-aligned to native English phoneme categories, which systematically mismatch L2 segmental realizations, so the probe measures agreement-with-native-targets rather than phoneme content. We (i) mark these figures in Table 2 and do not compare their levels against VCTK, and (ii) base the "refined late" claim on the monotone-with-depth trajectory, which is direction-only and holds on native VCTK where the mismatch does not arise. An L2-aware aligner permitting L1-conditioned phone substitutions would be needed to read L2 phoneme levels. We do not attempt this and draw no level-based conclusion from that column.

### C.6 PER-FEATURE ACOUSTIC ALIGNMENT AT SMALL $K$

We correlate each of the top $8$ direction-matrix axes with $8$ accent-level acoustic features (F1, F2, F1/F2 ratio, rhoticity, VOT, pitch mean, pitch range, speech rate), giving $64$ Spearman coefficients per (corpus, encoder) cell. The unit of analysis is the accent centroid, so each coefficient is computed over $K$ points, 6 on L2-ARCTIC, $8$ on VCTK, and $40$ on GLOBE. We report these coefficients as effect sizes instead of testing them per cell. At $K$=6 the exact rank-permutation null has $6! = 720$ points, so the smallest attainable $p$-value is $2.8 \times 10^{-3}$, reached only by a direction that orders all six accents perfectly, and this does not hold under correction over $64$ comparisons. Large coefficients are also frequent at this sample size. Under the exact null, $\Pr(|\rho| \geq 0.5)$ is $0.297$ at $K$=6, $0.216$ at $K$=8 and $1.0 \times 10^{-3}$ at $K$=40, so of $64$ cells, we would expect $19.0$, $13.8$ and $0.1$ respectively to clear that threshold with no structure present. Table 4 gives the observed counts. The same sample-size

Table 4: Per-feature alignment reported descriptively. Counts are cells with $|\rho| \geq 0.5$ out of the $64$ (direction $\times$ feature) pairs, alongside the count expected under the exact rank-permutation null at the corresponding $K$. Observed counts match the chance expectation on the parallel corpora and exceed it only on GLOBE at $K$=40.

| Corpus | $K$ | Max $\|\rho\|$ (WavLM/HuBERT/Whisper) | Cells $\|\rho\| \geq 0.5$ | Expected by chance |
|---|---|---|---|---|
| L2-ARCTIC | 6 | 0.94 / 0.94 / 0.89 | 20 / 20 / 19 | 19.0 |
| VCTK | 8 | 0.98 / 0.98 / 0.88 | 19 / 16 / 14 | 13.8 |
| GLOBE | 40 | 0.49 / 0.74 / 0.63 | 0 / 3 / 2 | 0.1 |

effect accounts for the encoder ordering referenced in Section 6. Counting the directions whose best-matching feature reaches $|\rho| \geq 0.5$, L2-ARCTIC gives $6/8$ for WavLM, $6/8$ for HuBERT and $8/8$ for Whisper, placing the transcription-trained control above both SSL encoders. Treating the eight features as independent, which is an approximation since they are correlated, the chance probability that a direction's best feature clears $0.5$ is $1 - (1 - 0.297)^8 \approx 0.94$ at $K$=6, or about $7.5$ of $8$ directions. The observed counts of $6$ to $8$ fall within this range, so the ordering across encoders is not informative about the control. On GLOBE the corresponding chance figure is about $0.07$ of $8$ directions, and the counts are $0/8$, $2/8$ and $1/8$, with the control below both SSL encoders. The phonological claim in Section 4 therefore is based on the direction-reuse test, which compares within-group against between-group cosine similarity over accent pairs rather than over accents and is not subject to this limitation.

### C.7 CORRECTED REPORTING VERDICTS

We had two verdict strings in our original pipeline that were logically inverted or mis-thresholded. The first one which is an interpolation-drop verdict, printed "gaps present" for the ideal drop fraction of $0.0$. The second, a redundant hard-cutoff stabilization detector (Section 5), had an overlap threshold that was never reached and so silently returned the last sampled layer. This does not affect our claims as the interpolation measurement was always correct, and the "no stabilization" result comes from an independent, data-driven plateau detector that finds no plateau in any metric for all nine (corpus $\times$ encoder) cells, with the between-layer change rate remaining above any plausible plateau throughout. Both verdicts are corrected in the reported results.

## D ADDITIONAL FIGURES

Every figure in the main text uses WavLM. This appendix reports the remaining corpus–encoder cells so that the consistency claims in Sections 5 and 6 can be checked directly.

### D.1 STAGE I: ACCENT PROJECTIONS FOR HUBERT AND THE WHISPER CONTROL

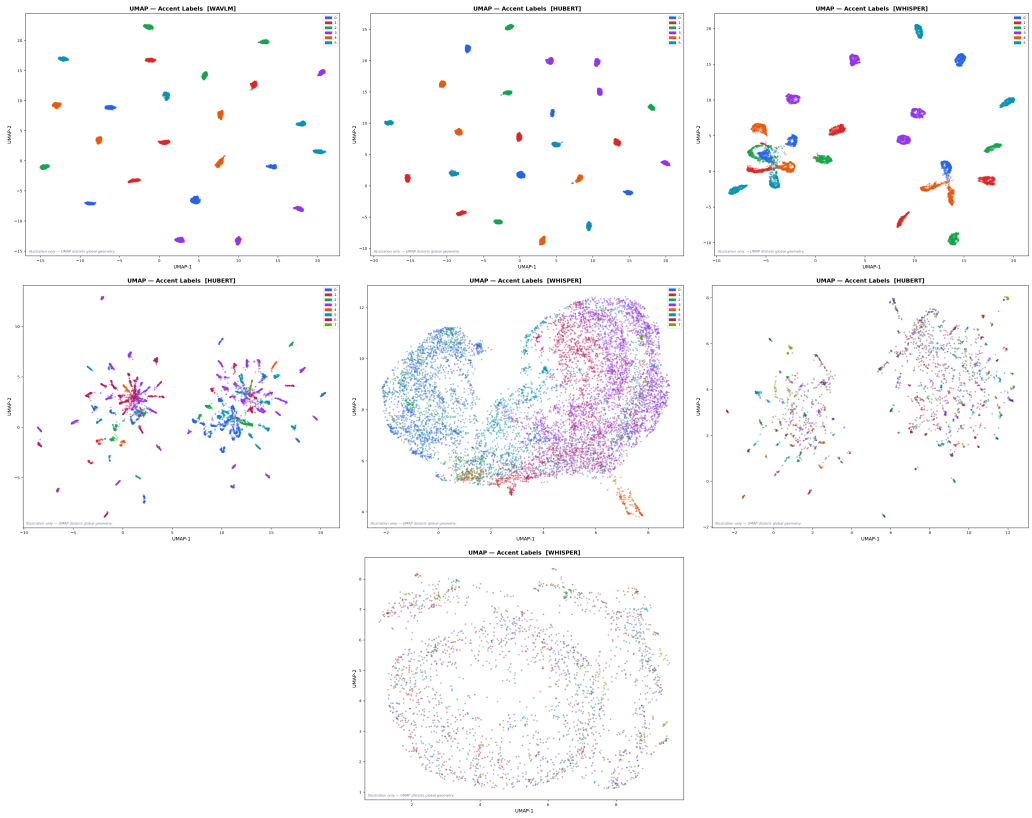



Figure 4: **Accent projections for the remaining corpus–encoder cells.** UMAP projections of accent-labelled utterances for the cells not shown in Figure 2. Rows group L2-ARCTIC (WavLM, HuBERT, Whisper), then VCTK (HuBERT, Whisper) and GLOBE (HuBERT, Whisper). These projections are for illustration only. The quantitative counterparts are the cluster-softness and connectivity measures in Table 1 and Appendix A.

### D.2 Stage I: Transitional-speaker placement

These plots supplement Section 4 and show transitional-speaker placement on the three corpora. On WavLM GLOBE, contact/mixed-variety speakers occupy higher accent-entropy neighborhoods than monolithic-label speakers ($p < 10^{-3}$). The betweenness panel is uninformative on small corpora and is not interpreted. The top three panels in Figure 5 are on GLOBE, the subsequent panels with no data belong to L2-ARCTIC, and the final three panels belong to VCTK.

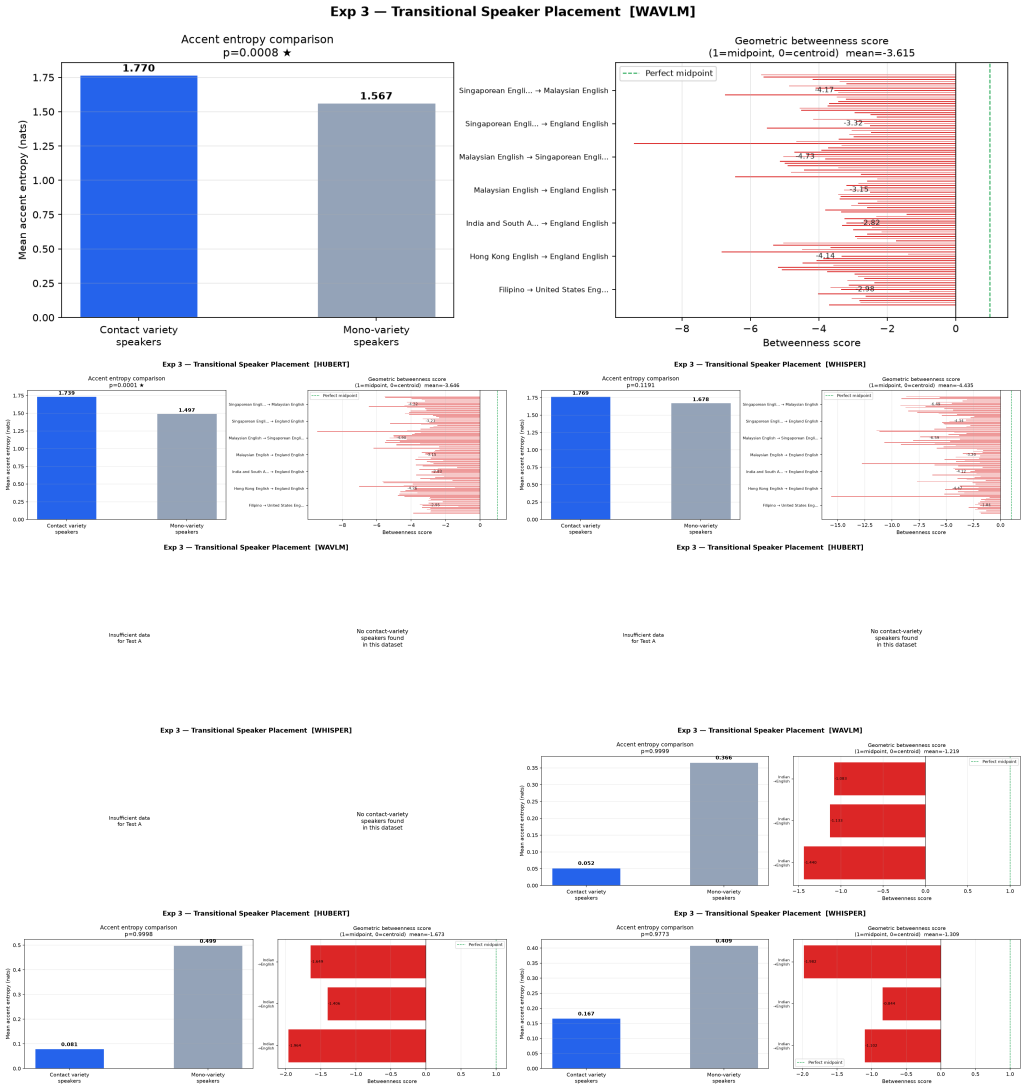



Figure 5: **Transitional-speaker placement across the three corpora.** Neighborhood accent entropy and betweenness for all nine corpus–encoder cells. The L2-ARCTIC and VCTK panels illustrate the resolution limit discussed in Section 4, where too few speakers per accent leave the contact/monolithic comparison weak.

### D.3 Stage II: Information trajectories for the remaining eight cells

These plots supplement Section 5. The top two panels in Figure 6 are for WavLM (L2-ARCTIC and GLOBE). The second row is for HuBERT (L2-ARCTIC and VCTK). The next two panels are HuBERT GLOBE and Whisper L2-ARCTIC. The final two panels are for Whisper (VCTK and GLOBE).

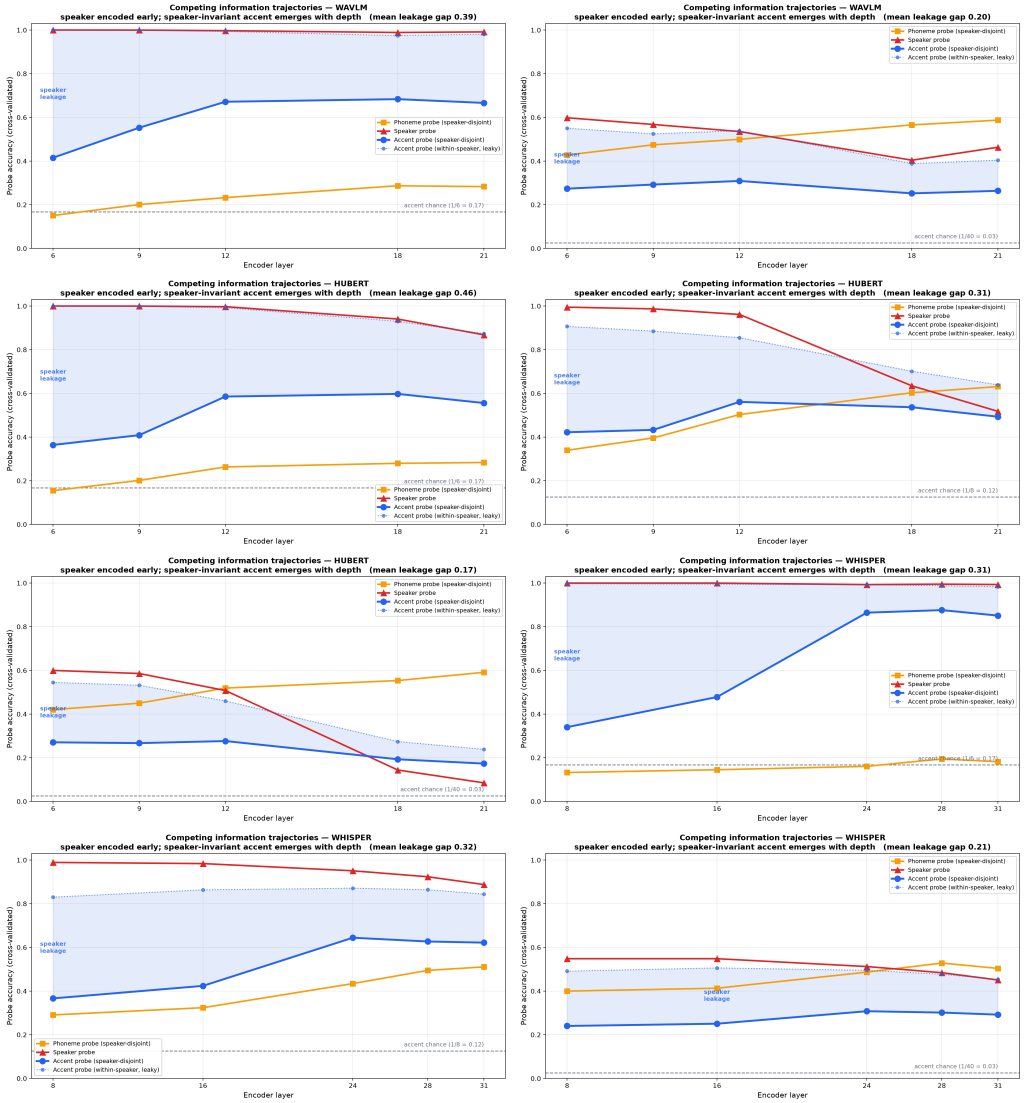



Figure 6: **Information trajectories for the remaining eight corpus–encoder cells.** The speaker-disjoint accent probe strengthens with depth and the leakage band narrows with depth in every cell. The eight pairs are those not shown in Figure 1 (WavLM/VCTK). Reading order is WavLM (L2-ARCTIC, GLOBE), HuBERT (L2-ARCTIC, VCTK, GLOBE), Whisper (L2-ARCTIC, VCTK, GLOBE). In each panel the speaker probe (red) is maximal at the shallowest sampled layer and decays, the phoneme probe (orange) rises to a late peak, and the solid blue speaker-disjoint accent probe is weakest early on. The shaded band between the solid and dotted blue curves is speaker leakage, and it is wider on the small corpora (L2-ARCTIC, VCTK) than on GLOBE, matching the leakage gaps in Table 2.

### D.4 Stage III: Cross-model alignment on the parallel corpora

These plots show cross-model alignment for L2-ARCTIC and VCTK and supplement the GLOBE plot in Section 6.

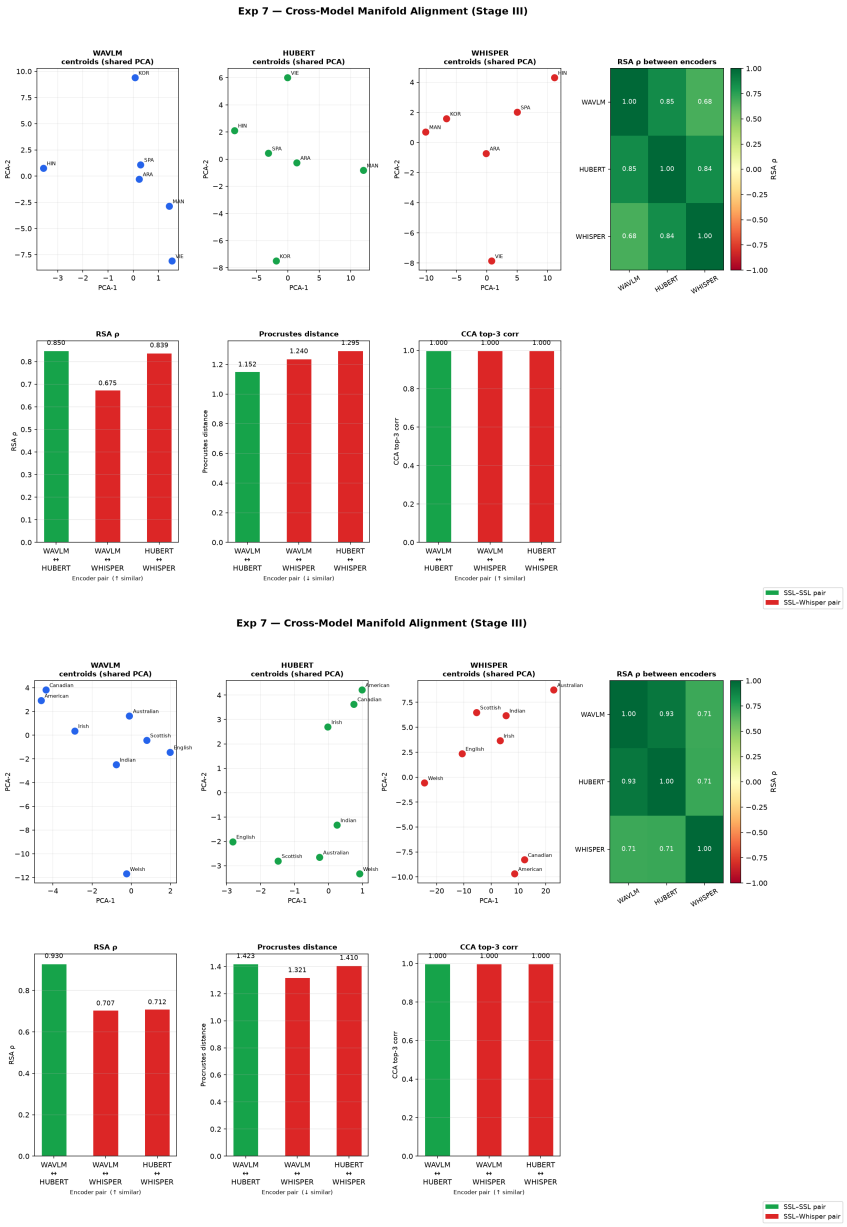



Figure 7: **Cross-model alignment on the two parallel corpora.** L2-ARCTIC (top, SSL–SSL RSA $\rho = 0.85$) and VCTK (bottom, $\rho = 0.93$). Agreement is high on both, along with the untrained-encoder and MFCC baselines (Table 3). We do not treat these corpora as evidence for learned convergence and base that claim on GLOBE (Figure 3, Section 6). The CCA sub-panels are degenerate at this centroid count and has no claim.