

Categorize Early, Integrate Late: Divergent Processing Strategies in Automatic Speech Recognition

Nathan Roll^{1,2*}, Pranav Bhalerao^{1*}, Martijn Bartelds¹, Arjun Pawar¹, Yuka Tatsumi¹, Tolúlopé Ògúnremí¹, Chen Shani^{1,4}, Calbert Graham³, Meghan Sumner¹, Dan Jurafsky¹

¹Stanford University ²oruk.ai ³University of Cambridge ⁴Tel Aviv University

Abstract

In speech language modeling, two architectures dominate the frontier: the Transformer and the Conformer. However, it remains unknown whether their comparable performance coincides with convergent processing strategies or distinct representational profiles. We introduce *Architectural Fingerprinting*, a probing framework that compares layer-wise representational profiles across architecture labels, and apply it to a suite of 24 pre-trained encoders (39M–3.3B parameters). Our pooled analysis reveals divergent profiles: phoneme-category accessibility peaks at 21% of network depth in Conformers versus 49% in Transformers, while gender peaks at 16% versus 28%. In contrast, Transformers “Integrate Late,” with phoneme, L1-background, and duration information peaking in deeper layers (49–57%). These fingerprints motivate testable hypotheses: Conformers’ front-loaded categorization may benefit low-latency streaming, while Transformers’ deep integration may favor tasks requiring rich context and cross-utterance normalization.

1 Introduction

In text modeling, architectural diversity has largely collapsed onto the Transformer paradigm. Speech is different: two encoder families, Transformers and Conformers, coexist at the frontier, achieving comparable word error rates (WER) on standard benchmarks (Gulati et al., 2020; Radford et al., 2023). This coexistence provides a rare opportunity to compare representational organization across architecture families at similar task performance. We therefore ask: when models from two architecture families solve the same problem equally well, do they exhibit similar layer-wise processing profiles?

This question matters for three reasons. First, interpretability: knowing where phonetic and

speaker information emerges enables targeted debugging and trust calibration. Second, demographic encoding: understanding when gender and L1-background cues become accessible informs fairness audits. Third, deployment: if model groups process information at different depths, this has implications for streaming, latency, and layer-wise pruning.

Prior probing work has established that speech models encode acoustic and linguistic information hierarchically (Chrupała et al., 2017; Belinkov et al., 2019; Pasad et al., 2021; Martin et al., 2023; Pasad et al., 2023), with phonetic information peaking in middle layers and word identity emerging later (Pasad et al., 2021; Bartelds et al., 2022). Early interpretability work on speech representations includes Chrupała et al. (2017), which analyzes layers of a visually grounded speech model to study how linguistic structure emerges across depth. Complementary analyses have revealed speaker-specific and L1-related structure (Ferragne et al., 2019; Graham, 2021). Beyond speech, layer-wise model-comparison work has shown that architectures with similar end-task performance can nevertheless learn systematically different internal representations; for example, Raghu et al. (2021) compare ViTs and CNNs and show clear differences in representation geometry across layers. However, no prior work systematically compares Transformer and Conformer encoders or tests whether their representational hierarchies differ after accounting for scale and model family.

Our main contributions are:

1. We introduce *Architectural Fingerprinting*, a linear-probing framework that quantifies when features become linearly accessible across depth and compares representational hierarchies across model groups. Our probes measure linear accessibility rather than mak-

*Equal contribution.

ing causal claims about mechanism.

2. We present a comprehensive cross-architectural analysis of 24 Transformer and Conformer models, identifying pooled differences across datasets and scales and testing their sensitivity to model family and training regime.
3. In pooled comparisons, we identify two descriptive profiles: the Conformer *Categorize Early* profile, which front-loads phoneme category and gender information, and the Transformer *Integrate Late* profile, which co-locates phoneme, L1-background, and duration cues in deeper layers (Figure 1). These names summarize pooled patterns in our sample.
4. We show that a logistic regression classifier predicts the sampled architecture label from peak position profiles alone (AUC = 0.88), demonstrating profile separation within this model suite.
5. Code and redistributable probing outputs are available at <https://github.com/Nathan-Roll1/categorize-early-asr>.

2 Methods

Using linear probing (§2.3), we analyze 24 Transformer and Conformer models (§2.1) across over 50,000 utterances from seven speech corpora (§2.2). We extract layer-wise representations using standardized depth indexing (§2.4), quantify information accessibility with complementary metrics (§2.5), and apply statistical tests (§2.6) to estimate pooled architecture-label differences and test their sensitivity to model scale and family. This approach characterizes systematic differences in the sampled models’ profiles without treating architecture as isolated from family or training regime.

2.1 Model Suite

We analyzed a diverse collection of 24 pre-trained spoken language models spanning two dominant speech encoder architectural families: Transformers (N=17) and Conformers (N=7). This imbalance reflects the current landscape of publicly available pretrained models and we address its statistical implications in §2.6. The models ranged

in size from approximately 39M to 3.3B parameters, providing broad coverage across scales and training regimes. A complete list of all models is provided in Appendix A. All models were analyzed in their publicly released, pre-trained state without fine-tuning.

2.2 Dataset and Speech Corpora

For robustness, we probed model representations using seven diverse speech corpora spanning native and non-native English speakers, multiple accent backgrounds, and varied recording conditions: **L2-ARCTIC** (non-native English; Zhao et al., 2018); **CMU ARCTIC** (native English; Kominek and Black, 2004); **Common Voice** (accent-stratified crowdsourced; Ardila et al., 2020); **Speech Accent Archive (SAA)** (elicited passage, 100+ L1s; Weinberger, 2015); **ALLSSTAR** (L2 acquisition research; Bradlow, n.d.); **Cambridge Assessment** (a private, non-redistributable proficiency assessment corpus); and **Speak & Improve (S&I) Corpus 2025** (L2 learner English; Knill et al., 2025).

Our corpus comprises 50,000+ utterances across all datasets. Audio from all corpora was resampled to 16 kHz for model inference and segmented to support phoneme-level analysis; Table 4 reports the source-corpus sampling rates. Results were aggregated across datasets to ensure findings generalize beyond any single corpus.

2.3 Feature Taxonomy and Probing

All analyses in this paper measure *linear accessibility* of information across network depth. Linear probes quantify when a feature can be extracted by a simple readout without additional nonlinear computation. Unlike attention maps or class activation maps, a probe is an auxiliary predictor trained on hidden states. It tests whether a feature is linearly decodable at a given layer; it does not identify which input frames or hidden-state dimensions caused the model’s ASR prediction. This choice is intentional: linear accessibility reflects functional availability to downstream components and is the standard lens through which representational hierarchies are compared across architectures. We do not claim that linear probes exhaustively measure all information present in a representation; rather, our claims concern when information becomes linearly usable as depth increases.

Following the signal-to-symbol hierarchy established in speech perception research (Hickok

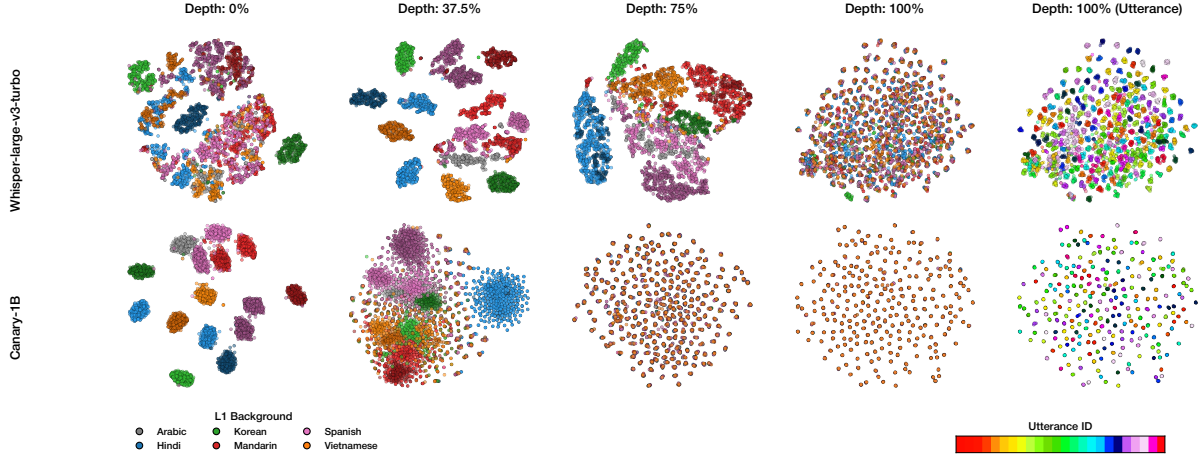


Figure 1: Layer-wise t-SNE representations for Whisper-large-v3-turbo (top) and Canary-1B (bottom) on L2-ARCTIC. The first four columns show depths 0%, 37.5%, 75%, and 100%, colored by L1 background; the fifth repeats the 100% representations colored by utterance identity. Points are plotted in random order to reduce occlusion. In these examples, Whisper shows later, gradual L1 clustering, whereas Canary shows earlier separation.

and Poeppel, 2007; Poeppel et al., 2008) and validated in neural probing studies (Belinkov and Glass, 2017; Pasad et al., 2021), we organized features into four levels of abstraction: **Acoustic** (F0, formants, intensity), **Demographic** (gender and **Accent**, operationalized as categorical L1 background), **Phonetic** (phoneme category), and **Temporal** (utterance-level duration). Throughout, *Accent* denotes this L1-background target rather than a direct acoustic measurement of accent. We probe phoneme categories as defined by the CMU pronouncing dictionary (39 categories), which groups allophones of the same phoneme; our probes thus measure categorical sound identity rather than fine-grained allophonic variation. Detailed definitions for each feature are provided in Appendix E. For each model and layer, we trained linear probes to predict these features from hidden states (linear regression for continuous targets; logistic regression for categorical ones). Accordingly, probe curves should be interpreted as measuring *linear accessibility* rather than raw information availability.

2.4 Depth Standardization

To standardize depth across heterogeneous ASR frontends, we extract the sequence of hidden states output by each encoder block. This yields $L + 1$ representations (an initial pre-block representation plus one per encoder block). Throughout, **Layer 0** denotes the *pre-block* encoder representation (i.e., the embedding/projection output immediately before the first encoder block), and Layer L denotes

the final encoder block output. Importantly, this convention does not treat convolutional feature-extractor frontends as additional layers: when present, the feature encoder is upstream of the probed hidden-state stack and is not separately indexed. This ensures comparisons focus on the encoder depth where architectural differences (e.g., per-block convolution vs. attention-only blocks) are expressed. Implementation details for extracting these states across libraries are provided in Appendix A.5.

2.5 Evaluation Metrics

We evaluated probing performance using three complementary metrics that capture where in the network a feature is most linearly accessible, how strongly it is encoded at that point, and how separated two features are in depth.

For each feature f , we define **peak position** π_f as the normalized depth where probe performance is maximized:

$$\pi_f = \frac{\arg \max_l \text{score}_f(l)}{L}$$

where $\text{score}_f(l)$ is the test-set accuracy (for classification) or R^2 (for regression) at layer l , and L is the total number of layers. Peak position ranges from 0 (first layer) to 1 (final layer).

Peak strength s_f is defined as:

$$s_f = \max_l \text{score}_f(l)$$

which measures the maximum probing performance achieved across layers.

To quantify hierarchical separation between feature pairs, we compute **positional deltas**:

$$\Delta_{f_1 \rightarrow f_2} = \pi_{f_2} - \pi_{f_1}$$

Positive deltas indicate that feature f_2 peaks later than f_1 in the network.

As a robustness measure less sensitive to a single maximum, we additionally define a performance-weighted center of mass. Let $z_l = l/L$, let b_f denote the no-skill baseline for feature f , and let $a_f(l) = \max(\text{score}_f(l) - b_f, 0)$. Then

$$c_f = \frac{\sum_{l=0}^L z_l a_f(l)}{\sum_{l=0}^L a_f(l)}.$$

Comparing c_f with the argmax peak π_f indicates whether a broad probing curve and its maximum support the same depth interpretation.

Additional auxiliary metrics, including peak width and layer-wise entropy, are defined in Appendix C.

2.6 Statistical Analysis

Architectural Comparison. To test whether peak positions differ systematically between architectures, we conducted two-sample t-tests comparing Transformer (N=17) and Conformer (N=7) models for each feature type. We report bootstrap 95% confidence intervals (10,000 resamples) for the mean difference in peak position.

Regression Analysis. To estimate architecture-associated differences while controlling for model size, we fit a linear regression model: $\pi_f = \beta_0 + \beta_1 \cdot \text{arch} + \beta_2 \cdot \log(\text{params}) + \epsilon$, where arch is a binary indicator (Conformer = 1, Transformer = 0) and params is the parameter count. We use $\log(\text{params})$ because model size spans nearly two orders of magnitude and scale effects are more plausibly log-linear than linear over this range. We report standardized coefficients (β) and p -values from two-tailed tests. As a family-aware sensitivity analysis, we also fit $\pi_f = \beta_0 + \beta_{\text{arch}} \text{arch} + \beta_{\text{size}} \log(\text{params}) + u_{\text{family}} + \epsilon$, with a random intercept u_{family} for model family.

Architectural Fingerprinting. To quantify the distinctiveness of the observed fingerprints, we trained a logistic regression classifier to predict the sampled architecture label (Transformer vs. Conformer) from the 5-dimensional feature vector $[\pi_{\text{acoustic}}, \pi_{\text{gender}}, \pi_{\text{accent}}, \pi_{\text{phoneme}}, \pi_{\text{duration}}]$. We used leave-one-out cross-validation and report the area under the ROC curve (AUC) as our primary metric.

3 Results

In pooled comparisons, Transformers and Conformers show systematically different representational profiles.

3.1 Divergent Processing Hierarchies

The pooled Transformer group exhibits an integrated processing profile (Acoustic $\rightarrow$ Gender $\rightarrow$ {Phoneme $\approx$ Duration $\approx$ Accent}), where high-level linguistic and sociolinguistic information co-locate in layers at 49–57% of total depth.

In contrast, the pooled Conformer group exhibits a segregated hierarchy. Using mean peak depth, where smaller values indicate earlier layers, Conformers follow Gender ($\pi = 0.16$) $\rightarrow$ Phoneme ($\pi = 0.21$) $\rightarrow$ Acoustic ($\pi = 0.23$) $\rightarrow$ Accent ($\pi = 0.35$) $\rightarrow$ Duration ($\pi = 0.70$). Thus, categorical gender and phoneme information is front-loaded, while temporal integration is deferred to the final layers. These pooled summaries mask family heterogeneity. In our sample, 71% of Transformers (12/17) show co-located deeper peaks for Phoneme, Duration, and Accent (all $> 35\%$ of normalized depth), while 86% of Conformers (6/7) show earlier Gender and Phoneme peaks (both $< 50\%$ depth) together with deeper Duration peaks ($> 50\%$ depth). The pooled Transformer profile is concentrated largely among Whisper checkpoints: Whisper models show later Gender and L1-background peaks than non-Whisper Transformers, while none of the five Conformer–non-Whisper Transformer contrasts is significant (§3.4). These thresholds are descriptive summaries of the observed distributions in our model set, not theoretical cutoffs. Figure 2 summarizes these peak positions across all features, and Table 1 shows individual variation around these patterns.

3.2 Positional Dynamics of Individual Features

We examine architecture-associated differences in the linear accessibility of specific features. These feature categories are not independent: speech cues often covary across acoustic, demographic, phonemic, and prosodic dimensions. This covariance motivated our focus on depth-wise separability, since the goal is not to assume isolated signals, but to test whether model groups differ in when particular information types become linearly accessible. Despite this overlap, we still observe

Model	Acoustic	Gender	Accent	Phoneme	Duration
<i>Conformer Models</i>					
canary-1b	0.153	0.139	0.347	0.215	0.611
canary-1b-flash	0.146	0.089	0.260	0.130	0.849
canary-qwen-2.5b	0.151	0.099	0.255	0.312	0.464
granite-speech-3.3-2b	0.466	0.323	0.375	0.229	0.812
parakeet-tdt-0.6b-v2	0.194	0.167	0.340	0.340	0.611
speechbrain-loq	0.206	0.083	0.537	0.157	0.759
w2v2-conformer	0.192	0.194	0.319	0.056	0.764
<i>Transformer Models</i>					
Phi-4-multimodal	0.332	0.281	0.266	0.896	0.109
hubert-large	0.140	0.181	0.278	0.000	0.681
hubert-xlarge	0.091	0.073	0.271	0.156	0.538
wav2vec2-large	0.137	0.090	0.278	0.014	0.424
wavlm-large	0.186	0.118	0.618	0.500	0.667
whisper-base	0.306	0.333	0.722	0.472	0.472
whisper-base.en	0.258	0.278	0.667	0.778	0.472
whisper-large	0.217	0.297	0.698	0.245	0.552
whisper-large-v2	0.202	0.328	0.667	0.797	0.990
whisper-large-v3	0.202	0.406	0.708	0.604	1.000
whisper-large-v3-turbo	0.208	0.396	0.651	0.495	0.380
whisper-medium	0.172	0.292	0.618	0.417	0.465
whisper-medium.en	0.238	0.403	0.750	0.611	0.493
whisper-small	0.225	0.306	0.639	0.542	0.403
whisper-small.en	0.214	0.306	0.653	0.625	0.486
whisper-tiny	0.208	0.375	0.667	0.667	0.458
whisper-tiny.en	0.262	0.333	0.583	0.542	0.375

Table 1: Mean peak layer position for each model. Values represent the fraction of model depth (0 = first layer, 1 = final layer) where probing accuracy/ R^2 is maximized. At the pooled-group level, Conformers show earlier Gender and Phoneme peaks and later Duration peaks than Transformers.

clear differences in relative positioning; for example, in Conformers, Gender peaks earlier (mean $\pi = 0.16$) than the aggregated Acoustic category (mean $\pi = 0.23$).

Acoustic features are front-loaded in both architectures (Transformer mean $\pi = 0.22$; Conformer mean $\pi = 0.23$), as expected for low-level signals.

Gender peaks very early in Conformers (mean $\pi = 0.16$), *before* the aggregated Acoustic category ($\pi = 0.23$), while Transformers distribute gender encoding more broadly (mean $\pi = 0.28$). Note that our use of binary gender labels reflects dataset annotations and may not capture the full spectrum of gender identity; similar caveats apply to coarse L1-background categories.

Phoneme category shows the largest divergence ($\Delta\pi = -0.29$, $p < 0.01$). Conformers peak early (mean $\pi = 0.21$) while Transformers peak late (mean $\pi = 0.49$).

Accent peaks earlier in Conformers (mean $\pi = 0.35$) than Transformers (mean $\pi = 0.57$).

Duration, requiring the longest-range integration, represents the final stage of processing for both architectures (Transformer $\pi = 0.53$; Con-

former $\pi = 0.70$).

Figure 3 visualizes these dynamics as smoothed probing trajectories across normalized depth. The Conformer curves show sharp early rises for Gender and Phoneme followed by plateau, while Transformer curves rise more gradually and peak in mid-to-deep layers. These trajectories underlie the summary statistics in Figure 2 and Table 1.

3.3 Architectural Fingerprints

The systematic differences we observe allow us to classify the sampled architecture label from its representational profile alone with high accuracy (AUC = 0.88). The most discriminative classifier weights are Accent ($w_{\text{accent}} = -0.85$), Gender ($w_{\text{gender}} = -0.83$), and Duration ($w_{\text{duration}} = +0.81$; Appendix B.3). In a size-adjusted regression, Accent has a significant architecture-label coefficient ($\beta_{\text{arch}} = -0.21$, $p < 0.05$), but no architecture coefficient is significant in the family-aware analysis (Appendix B.4). Table 2 reports the complete statistical comparison, showing large effect sizes (Cohen’s $d > 0.8$) for all features except Acoustic. With Bonferroni correction ($\alpha = 0.01$), Accent and Phoneme differences remain signifi-

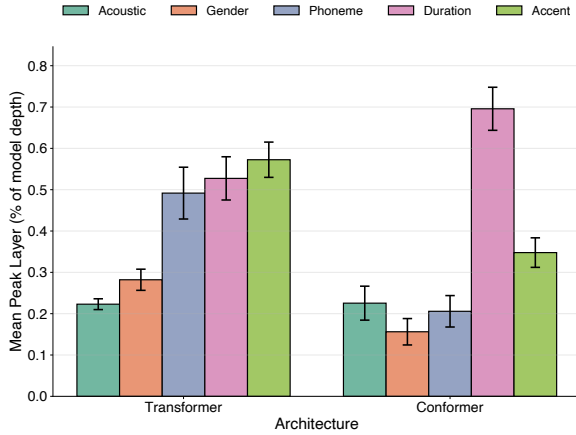


Figure 2: Conformers front-load Gender and Phoneme while deferring Duration, whereas Transformers place the high-level features in a narrower mid-to-deep band. Bars show mean peak layer position (percentage of model depth) for five feature categories across Transformer (N=17) and Conformer (N=7) architectures; error bars indicate standard error of the mean. Temporal ordering is indicated by bar height, not left-to-right bar position.

Feature	$\Delta\pi$	t	df	p	Cohen's d
Acoustic	+0.003	-0.08	22	0.940	0.03
Gender	-0.126	2.80	22	0.011	-1.26
Accent	-0.225	3.17	22	0.004	-1.42
Phoneme	-0.286	2.81	22	0.010	-1.26
Duration	+0.168	-1.90	22	0.071	0.85

Table 2: T-tests comparing peak positions between architectures. $\Delta\pi$ = Conformer - Transformer (negative values indicate Conformers peak earlier). Cohen's $d > 0.8$ indicates large effect size. With Bonferroni correction for 5 comparisons ($\alpha = 0.01$), Accent and Phoneme survive correction.

cant. Full statistical details including bootstrap confidence intervals are in Appendix B.

3.4 Architecture, Model Family, and Training Regime

The pooled results differ across Transformer and Conformer groups, but model family and training regime are entangled with architecture, as illustrated by the later Gender and Accent peaks in Whisper than in SSL Transformers such as HuBERT and WavLM. We therefore test the pooled contrast after excluding Whisper and with family random intercepts.

We additionally separated the Whisper family from the architecture label. Compared with the five non-Whisper Transformers, the seven Conformers differ by +0.038 in Acoustic, +0.008 in

Gender, +0.005 in Accent, -0.108 in Phoneme, and +0.212 in Duration peak position; none of these contrasts is significant ($p \geq 0.120$). Family-random-intercept models likewise yield no architecture coefficient significant at $p < 0.05$ (Appendix B.4). Because these controls are weakly identified by the available model families, the pooled differences should be interpreted as associations within this model suite rather than isolated architecture effects.

As an approximately scale-matched comparison, Canary-1B (1B) peaks earlier than Whisper-large-v3 (1.55B) by 0.27 normalized depth for Gender, 0.36 for Accent, and 0.39 for Phoneme. This pair follows the pooled fingerprint, but training data and model lineage still differ, so it is descriptive rather than a controlled architecture experiment.

3.5 Robustness and Controls

To ensure our findings are not driven by outliers or specific training factors, we conducted extensive robustness checks (detailed in Appendix D). A leave-one-out sensitivity analysis shows that the pooled contrasts become *stronger* when excluding the outlier model *granite-speech-3.3-2b* (e.g., Gender $\Delta\pi$ changes from -0.13 to -0.15, $p = 0.003$). Additionally, a paired comparison of multilingual vs. English-only Whisper models reveals no significant difference in peak positions for any feature (though Phoneme shows a non-significant trend, $\Delta\pi = -0.12$, $p = 0.30$), indicating that multilingual data diversity does not explain the delayed processing in Whisper. Performance-weighted center of mass, gold-transcript-only phoneme probing, and leave-Cambridge-out analyses retain the principal pooled directions. A depth-matched truncation experiment, however, does not support the predicted Conformer early-exit advantage (Appendix D).

4 Discussion

In the pooled comparison, Conformers place Gender, Accent, and Phoneme information earlier and Duration later than Transformers. The Transformer average is shaped strongly by Whisper, whose Gender and L1-background peaks occur later than those of non-Whisper Transformers. In family-aware models, no architecture coefficient is significant. We therefore use “Categorize Early” and “Integrate Late” as descriptive names for

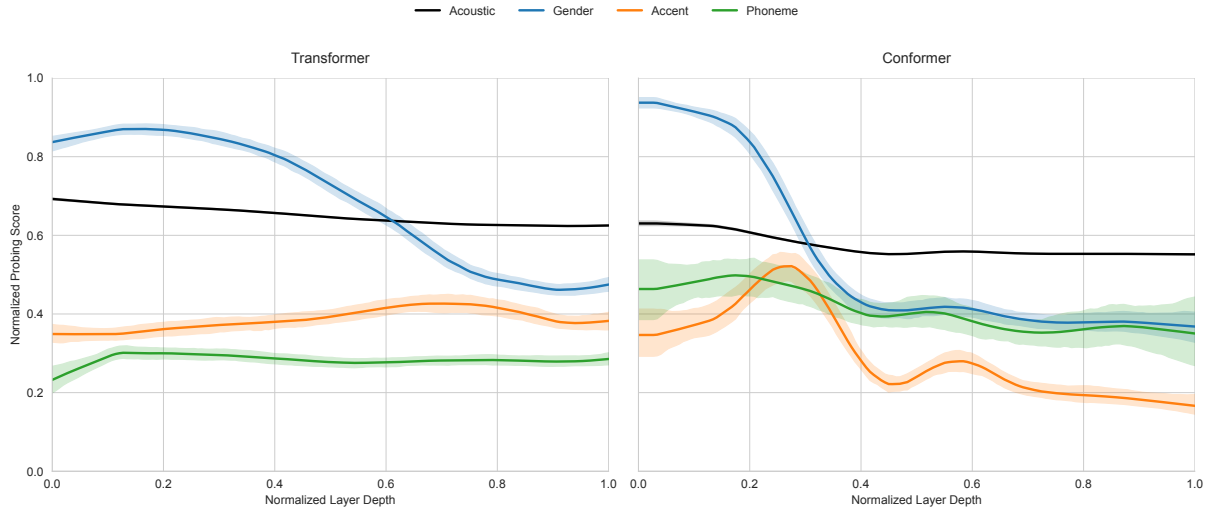


Figure 3: Smoothed layer-wise probing trajectories across normalized depth for representative feature groups (Acoustic, Gender, Accent, Phoneme), shown separately for Transformers (left) and Conformers (right). Curves are LOWESS fits with 95% bootstrapped confidence intervals. To emphasize *where* features become accessible (rather than absolute probe difficulty), probing scores within each feature group are min-max normalized before smoothing. Duration is omitted from this visualization for clarity; its peak positions are reported in Figure 2 and the main text.

pooled profiles in our sample, not as claims that architecture alone determines them (Appendices B.4 and D).

Prior probing work has established that self-supervised speech models encode features hierarchically (Pasad et al., 2021, 2023). Early work demonstrated that neural models of speech develop linguistically meaningful representations across layers, providing an initial interpretability perspective on spoken representations. However, these studies primarily analyzed models individually without systematic architectural comparison. Our pooled comparison shows that similar task performance can coincide with different processing profiles across sampled models.

Work in other domains has shown that architectures with similar performance can nevertheless differ substantially in their internal representations. For example, Raghu et al. (2021) compare Vision Transformers and CNNs and find clear differences in layer-wise representation structure. Our pooled results extend this comparative perspective to speech: comparable ASR performance coincides with distinct representational timelines.

Taken together, our findings demonstrate that benchmark parity does not imply representational convergence—a result with implications for interpretability, transfer learning, and model selection.

Why Do These Divergent Strategies Emerge?

We consider how architectural differences could contribute to the pooled patterns, while recognizing that architecture, model family, and training objective are entangled.

Pooled Conformer “Categorize Early” Profile. Depthwise separable convolutions could contribute a bias toward local pattern detection. This is consistent with the rapid linear accessibility of categorical features (gender from coarse spectral patterns, phonemes from local acoustic templates) we observe in early layers. The pooled 54% depth separation between gender and duration suggests later layers perform higher-order integration.

Pooled Transformer “Integrate Late” Profile. Pure self-attention could contribute to the gradual, distributed profile observed in the pooled Transformer group. The co-location of Accent, Phoneme, and Duration in the latter half of the network (all peaking between 49–57% depth) suggests these models resolve multiple feature types concurrently rather than through staged processing. This pattern is most pronounced in Whisper.

Architecture, Family, and Objective Are Entangled. These profile names describe pooled patterns; they are not universal or causal properties of the architectures. A factorial comparison holding data, objective, and lineage constant is needed.

Implications for Model Design

Our findings motivate three testable deployment hypotheses:

- **Streaming and low-latency:** A depth-matched Wav2Vec2/Conformer truncation test does not support the prediction that early Conformer probe peaks yield smaller WER degradation. Both unmodified checkpoints fail below full depth (Appendix D); trained early-exit systems require direct evaluation.
- **Speaker variability:** Transformers co-locate Gender and Accent with Phoneme categories in deep layers, suggesting tighter integration of speaker normalization. We predict Transformer-based systems will show smaller performance gaps across speaker demographics when using full-depth representations. Recent work by Roll and Graham (2025) finds that relative L2 performance gaps can widen as L1 WER improves, with this pattern especially notable among the Conformer-based systems evaluated.
- **Layer-wise distillation:** The staged Conformer hierarchy (Gender $\rightarrow$ Phoneme $\rightarrow$ Acoustic $\rightarrow$ Accent $\rightarrow$ Duration) suggests natural cut-points for layer-wise knowledge distillation; the integrated Transformer profile may require different distillation strategies.

Future Directions. Controlled comparisons should test models that combine early convolutional processing with deep integration while holding training data, objective, and lineage constant.

5 Conclusion

In our pooled model suite, state-of-the-art spoken language models do not converge on a single representational profile. Conformers exhibit a *segregated* “categorize early” hierarchy, where Gender and Phoneme categories become linearly accessible in shallow layers before higher-order integration, whereas Transformers—especially Whisper—exhibit an *integrated, distributed* “integrate late” hierarchy, which co-locates Phoneme, Accent, and Duration cues in deeper layers.

These differences are quantifiable: Phoneme-category accessibility peaks at mean normalized

depth 0.21 in Conformers and 0.49 in Transformers ($\Delta\pi = -0.286$), the largest pooled group difference we observed and one that survives Bonferroni correction. In contrast, Duration exhibits a trend toward later processing, peaking 17 percentage points deeper ($p = 0.071$). We show that a simple classifier can distinguish the sampled architecture labels from representational profiles alone (AUC = 0.88, 95% CI [0.73, 1.00]), demonstrating profile separation in this model suite (though the wide confidence interval reflects our moderate sample size).

Our findings show that comparable speech-recognition performance can coincide with different representational profiles. The pooled timelines may reflect architecture, model lineage, or training objective; the present sample cannot separate their contributions. These pooled patterns motivate hypotheses for model selection and more intentional model design rather than established architecture-specific prescriptions. The truncation experiment likewise does not support an early-exit advantage. Architectural Fingerprinting should therefore be interpreted as a descriptive diagnostic rather than an architecture-only causal analysis.

Limitations

Linearity vs. availability. Linear probes measure *linear accessibility* and cannot distinguish late *integration* from late *linearization*. Our claims are thus restricted to when information becomes linearly usable, not when it first emerges in any form. Testing nonlinear accessibility would refine the interpretation of late accessibility without altering the pooled descriptive differences reported here.

Error pattern analysis. Our analysis characterizes *when* features emerge but not *how* they are represented. We cannot distinguish, for example, whether early Conformer phoneme encoding reflects local template matching or a different mechanism. Analyzing confusion matrices of phoneme probes could address this limitation, but such mechanistic validation is beyond our current scope.

Feature taxonomy scope. Our five-feature taxonomy captures the signal-to-phoneme transformation central to ASR but does not address lexical, syntactic, or discourse-level organization. Extending this to word-level semantics would provide a more complete picture of speech understanding.

Native vs. non-native speech. Our corpus

includes both native speakers (CMU ARCTIC) and non-native speakers (L2-ARCTIC and others), but we do not systematically test for interactions between architecture label and speaker nativeness. This limits our ability to characterize how architecture-associated profiles relate to robustness under pronunciation variation, despite evidence that human listeners show processing delays for non-native accents (Munro and Derwing, 1995).

Model diversity and sample imbalance. Our model suite is unbalanced across architectures, with fewer Conformers (N=7) than Transformers (N=17), largely due to the present distribution of publicly available pretrained ASR models. This limits the strength of claims about within-architecture variability, especially for Conformers, and may reduce sensitivity to less common representational profiles. We therefore interpret our conclusions as pooled group tendencies rather than exhaustive coverage of each family, and we include robustness checks in Appendix D showing that the main Transformer–Conformer differences remain stable under leave-one-out analyses. Our suite also does not include transducers (Graves, 2012) or state-space models, limiting generalization to the full ASR landscape.

Correlational nature. Our analysis is correlational: architecture label predicts hierarchy in the pooled analysis but does not establish causation. Controlled experiments (e.g., adding convolutions to Transformers) would provide stronger evidence but require training models from scratch, a substantial undertaking given the scale of modern ASR systems.

Robustness and deployment scope. Center-of-mass, gold-transcript-only, and leave-Cambridge-out analyses retain the principal pooled directions, but reuse the same confounded model suite. The layer-truncation experiment tests one depth-matched pair and does not support the proposed Conformer early-exit advantage; it does not evaluate trained early-exit or streaming systems generally (Appendix D).

Ethical considerations. Our use of binary gender labels and coarse L1-background categories reflects dataset annotations but does not capture the full spectrum of gender identity or accent diversity. Probing can expose sensitive attributes from representations, raising privacy concerns for deployment. We encourage practitioners to conduct fairness audits before using fingerprinting for model

selection.

Acknowledgments

We are grateful to the Stanford NLP Group for thoughtful discussions, generous feedback, and the collaborative intellectual community that supported this work. We also thank the anonymous reviewers for their constructive comments.

References

- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common voice: A massively-multilingual speech corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. [Neural representations for modeling variation in speech](#). *Journal of Phonetics*, 92:101137.
- Yonatan Belinkov, Ahmed Ali, and James Glass. 2019. [Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition](#). In *Interspeech 2019*, pages 81–85.
- Yonatan Belinkov and James Glass. 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems. In *Advances in Neural Information Processing Systems*, volume 30, pages 2441–2451.
- Paul Boersma and David Weenink. 2024. *Praat: doing phonetics by computer [Computer program]*. Version 6.4, retrieved from <https://www.praat.org/>.
- Ann R. Bradlow. n.d. ALLSTAR: Archive of L1 and L2 scripted and spontaneous transcripts and recordings. Retrieved from <https://speechbox.linguistics.northwestern.edu/allstar>.
- Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. [Representations of language in a model of visually grounded speech signal](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622.
- Emmanuel Ferragne, Cédric Gendrot, and Thomas Pellegri. 2019. [Towards phonetic interpretability in deep learning applied to voice comparison](#). In *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 790–794, Melbourne, Australia.
- Calbert Graham. 2021. [L1 identification from l2 speech using neural spectrogram analysis](#). In *Interspeech 2021*, pages 3959–3963.

- Alex Graves. 2012. [Sequence transduction with recurrent neural networks](#). In *ICML 2012 Workshop on Representation Learning*.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. [Conformer: Convolution-augmented transformer for speech recognition](#). In *Interspeech 2020*, pages 5036–5040.
- Gregory Hickok and David Poeppel. 2007. [The cortical organization of speech processing](#). *Nature Reviews Neuroscience*, 8(5):393–402.
- Kate M. Knill, Diane Nicholls, Mark J. F. Gales, Mengjie Qian, and Pawel Stroinski. 2025. [Introducing the Speak & Improve Corpus 2025: an L2 English speech corpus for language assessment and feedback](#). In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 167–171.
- John Kominek and Alan W. Black. 2004. The CMU Arctic speech databases. In *5th ISCA Workshop on Speech Synthesis (SSW 5)*, pages 223–224.
- Kinan Martin, Jon Gauthier, Canaan Breiss, and Roger Levy. 2023. [Probing self-supervised speech models for phonetic and phonemic information: A case study in aspiration](#). In *Interspeech 2023*, pages 251–255.
- Murray J. Munro and Tracey M. Derwing. 1995. [Processing Time, Accent, and Comprehensibility in the Perception of Native and Foreign-Accented Speech](#). *Language and Speech*, 38(3):289–306.
- Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. [Layer-Wise Analysis of a Self-Supervised Speech Representation Model](#). In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.
- Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. [Comparative Layer-Wise Analysis of Self-Supervised Speech Models](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- David Poeppel, William J. Idsardi, and Virginie van Wassenhove. 2008. [Speech perception at the interface of neurobiology and linguistics](#). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1071–1086.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. [Do vision transformers see like convolutional neural networks?](#) In *Advances in Neural Information Processing Systems*, volume 34, pages 12116–12128. Curran Associates, Inc.
- Nathan Roll and Calbert Graham. 2025. [Scaling conformation bias in automatic speech recognition](#). In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 133–137.
- Steven Weinberger. 2015. Speech accent archive. George Mason University. Retrieved from <http://accent.gmu.edu>.
- Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. [L2-ARCTIC: A Non-native English Speech Corpus](#). In *Interspeech 2018*, pages 2783–2787.

A Experimental Details

A.1 Model Suite

Our analysis included 24 pre-trained models from the Transformer and Conformer families. Table 3 provides comprehensive architectural specifications.

A.2 Dataset Details

We aggregated probing results across seven speech corpora to ensure generalizability. Table 4 summarizes each dataset.

Results were computed per-dataset and then aggregated (mean peak positions across datasets) to ensure findings are not driven by idiosyncrasies of any single corpus. The L2-ARCTIC dataset (Zhao et al., 2018) was used for t-SNE visualizations (Figure 1) due to its balanced L1 representation.

A.3 Feature Extraction Details

A.3.1 Acoustic Features (24 probes)

Acoustic features were extracted using Praat (Boersma and Weenink, 2024) with a 10ms frame shift:

- **F0** (fundamental frequency): min, mean, median, max
- **F1, F2, F3** (formants): min, mean, median, max each
- **F3–F2** (formant dispersion): min, mean, median, max
- **Intensity**: min, mean, median, max

Model	Arch	Params	Layers	Hidden	Heads	Training
<i>Transformer Models (N=17)</i>						
whisper-tiny	Trans.	39M	4	384	6	Supervised, 680k hrs
whisper-tiny.en	Trans.	39M	4	384	6	Supervised, English-only
whisper-base	Trans.	74M	6	512	8	Supervised, 680k hrs
whisper-base.en	Trans.	74M	6	512	8	Supervised, English-only
whisper-small	Trans.	244M	12	768	12	Supervised, 680k hrs
whisper-small.en	Trans.	244M	12	768	12	Supervised, English-only
whisper-medium	Trans.	769M	24	1024	16	Supervised, 680k hrs
whisper-medium.en	Trans.	769M	24	1024	16	Supervised, English-only
whisper-large	Trans.	1.55B	32	1280	20	Supervised, 680k hrs
whisper-large-v2	Trans.	1.55B	32	1280	20	Supervised, 680k hrs
whisper-large-v3	Trans.	1.55B	32	1280	20	Supervised, 1M+ hrs
whisper-large-v3-turbo	Trans.	809M	32	1280	20	Supervised, distilled
wav2vec2-large-960h-lv60	Trans.	317M	24	1024	16	SSL + fine-tuned
hubert-large-ls960-ft	Trans.	317M	24	1024	16	SSL + fine-tuned
hubert-xl-large-ls960-ft	Trans.	1B	48	1280	16	SSL + fine-tuned
wavlm-large	Trans.	317M	24	1024	16	SSL
Phi-4-multimodal-instruct	Trans.	2.7B	40	3072	32	Supervised, multimodal
<i>Conformer Models (N=7)</i>						
wav2vec2-conformer	Conf.	430M	24	1024	16	SSL
canary-1b	Conf.	1B	24	1024	8	Supervised, multilingual
canary-1b-flash	Conf.	1B	24	1024	8	Supervised, distilled
canary-qwen-2.5b	Conf.	2.5B	40	1536	12	Supervised, Qwen decoder
parakeet-tdt-0.6b-v2	Conf.	600M	30	768	8	Supervised, TDT
granite-speech-3.3-2b	Conf.	3.3B	32	2048	16	Supervised, code-switched
speechbrain-loq	Conf.	135M	12	512	8	Supervised

Table 3: Comprehensive model specifications. Trans. = Transformer, Conf. = Conformer. SSL = self-supervised learning. All Whisper models from OpenAI, HuBERT/Wav2Vec2 from Meta, WavLM from Microsoft, Canary/Parakeet from NVIDIA, Granite from IBM, SpeechBrain from community.

Dataset	Utts.	Src. kHz	Description
L2-ARCTIC	3,750	44.1	Non-native English, 6 L1s
CMU ARC-TIC	4,500	16	Native English, regional accents
Common Voice	10,000+	48	Crowdsourced, accent-stratified
SAA	2,000+	44.1	Elicited passage, 100+ L1s
ALLSTAR	5,000+	44.1	L2 acquisition research
Cambridge	15,000+	16	Proficiency assessment
S&I Corpus 2025	3,000+	16	L2 learner English, proficiency assessment
Total	~50,000	—	

Table 4: Speech corpora used for probing. Utterance counts and source sampling rates are approximate. All corpora were MFA forced-aligned, and all audio was resampled to 16 kHz for model inference. MFA = Montreal Forced Aligner.

A.3.2 Demographic Features (2 probes)

- **Gender:** Binary classification (male/female).
- **Accent (L1 background):** 6-way classification (Arabic, Hindi, Korean, Mandarin, Spanish, Vietnamese). We standardized L1 labels to these categories based on speaker meta-

data. For corpora with self-reported accent labels (e.g., Common Voice), we mapped to the nearest L1 category. Corpora without L1 metadata were excluded from Accent probes.

For utterance-level labels (Gender, Accent), we applied the label to all frames within the utterance. Probes were trained at the frame level, but test accuracy was computed per-frame and then averaged per-utterance to avoid inflating sample sizes. We did not control for differences in frame stride across models, which is a limitation.

A.3.3 Linguistic Features (2 probes)

- **Phoneme category:** 39-way classification (CMU phoneme set, which groups allophones). Phoneme boundaries were obtained using the Montreal Forced Aligner (MFA) with the English pronunciation dictionary. For corpora lacking transcripts, we used Whisper-large-v3 for automatic transcription prior to alignment.
- **Duration:** Regression (utterance-level duration in ms)

A.4 Linear Probe Training

Linear probes were trained independently for each feature at each layer:

- **Classification tasks:** Cross-entropy loss (Gender, L1, Phoneme)
- **Regression tasks:** Mean squared error loss (Acoustic, Duration)
- **Optimizer:** Adam with learning rate 0.001
- **Batch size:** 32
- **Max epochs:** 50 with early stopping
- **Data split:** 80%/10%/10% train/validation/test. For demographic probes (Gender, Accent), we used speaker-disjoint splits where possible (L2-ARCTIC, CMU ARCTIC). For corpora without speaker metadata, we applied random splits; this may inflate early-layer accuracy for these features, and we note this as a limitation.

A.5 Implementation Details

To extract hidden states consistently across models, we utilized the standard interface provided by the HuggingFace Transformers library. Specifically, models were invoked with `output_hidden_states=True`, returning a tuple containing the output of the embedding layer followed by the output of each Transformer or Conformer block. For models not in HuggingFace (e.g., some SpeechBrain or ESPnet models), we manually instrumented the forward pass to capture the equivalent intermediate representations.

B Detailed Statistical Results

B.1 Per-Model Peak Scores

Table 5 shows the maximum probing score achieved for each model across feature categories.

B.2 Complete Statistical Tests

B.2.1 Bootstrap Confidence Intervals

Feature	$\Delta\pi$	95% CI Low	95% CI High
Acoustic	+0.002	−0.061	+0.090
Gender	−0.126	−0.196	−0.046
Accent	−0.224	−0.321	−0.117
Phoneme	−0.286	−0.423	−0.146
Duration	+0.168	+0.029	+0.298

Table 6: Bootstrap 95% confidence intervals (10,000 resamples) for mean peak position differences between architectures.

B.2.2 Regression Controlling for Model Size

Feature	β_{arch}	p_{arch}	β_{size}	p_{size}
Acoustic	+0.016	0.729	+0.004	0.803
Gender	−0.105	0.116	+0.002	0.908
Accent	−0.214	0.041	−0.024	0.425
Phoneme	−0.257	0.096	−0.006	0.901
Duration	+0.146	0.240	+0.042	0.276

Table 7: Linear regression: $\pi_f = \beta_0 + \beta_{\text{arch}} \cdot \text{Conformer} + \beta_{\text{size}} \cdot \log(\text{params})$.

B.3 Architectural Classifier Details

The logistic regression classifier for predicting architecture achieved:

- **AUC:** 0.88 (95% CI [0.73, 1.00])
- **Leave-one-out accuracy:** 19/24 = 79.2%

Standardized coefficients:

- Accent position: $w_{\text{accent}} = -0.85$ (most discriminative)
- Gender position: $w_{\text{gender}} = -0.83$
- Duration position: $w_{\text{duration}} = +0.81$
- Acoustic position: $w_{\text{acoustic}} = +0.68$
- Phoneme position: $w_{\text{phoneme}} = -0.33$

B.4 Family-aware Sensitivity Analyses

Table 8 reports additional controls for the concentration of Whisper checkpoints among Transformers. Non-Whisper comparisons use Welch tests; mixed models include log parameter count and a random intercept for model family. The estimates use the retained per-model results rounded to three decimals.

B.5 Acoustic Feature Breakdown

Table 9 shows detailed statistics for individual acoustic features.

B.6 Whisper Model Family Analysis

Feature	Whisper	Non-Whisper	p
Gender	0.338	0.149	<0.001
Accent	0.669	0.342	<0.001
Duration	0.546	0.484	0.606

Table 10: Whisper (N=12) vs. non-Whisper Transformer (N=5) peak positions. Whisper models show significantly later Gender and Accent peaks.

Model	Acoustic	Gender	Accent	Phoneme	Duration
<i>Conformer Models</i>					
canary-1b	0.303	0.962	0.558	0.094	0.761
canary-1b-flash	0.305	0.960	0.574	0.079	0.834
canary-qwen-2.5b	0.302	0.959	0.557	0.107	0.748
granite-speech-3.3-2b	0.066	0.568	0.201	0.074	0.873
parakeet-tdt-0.6b-v2	0.307	0.951	0.554	0.079	0.727
speechbrain-loq	0.321	0.951	0.519	0.103	0.758
w2v2-conformer	0.311	0.962	0.461	0.168	0.786
<i>Transformer Models</i>					
Phi-4-multimodal	0.063	0.559	0.257	0.059	0.813
hubert-large	0.302	0.939	0.452	0.133	0.807
hubert-xlarge	0.302	0.938	0.466	0.102	0.823
wav2vec2-large	0.285	0.880	0.442	0.147	0.806
wavlm-large	0.276	0.883	0.514	0.118	0.785
whisper-base	0.308	0.952	0.467	0.083	0.913
whisper-base.en	0.301	0.917	0.464	0.082	0.901
whisper-large	0.339	0.943	0.592	0.095	0.917
whisper-large-v2	0.342	0.966	0.590	0.085	0.928
whisper-large-v3	0.343	0.959	0.582	0.063	0.929
whisper-large-v3-turbo	0.354	0.961	0.589	0.063	0.915
whisper-medium	0.337	0.952	0.587	0.088	0.900
whisper-medium.en	0.338	0.958	0.543	0.102	0.911
whisper-small	0.331	0.959	0.494	0.096	0.911
whisper-small.en	0.329	0.951	0.465	0.095	0.923
whisper-tiny	0.310	0.937	0.406	0.081	0.904
whisper-tiny.en	0.313	0.939	0.394	0.086	0.883

Table 5: Peak probing scores for each model. Acoustic/Duration values are R^2 ; Gender/Accent/Phoneme values are accuracy.

Feature	Δ_{NW}	p_{NW}	β_{Mixed}	p_{Mixed}
Acoustic	+0.038	0.536	+0.056	0.084
Gender	+0.008	0.880	+0.028	0.608
Accent	+0.005	0.947	+0.012	0.818
Phoneme	-0.108	0.570	-0.139	0.309
Duration	+0.212	0.120	+0.152	0.070

Table 8: Family-aware sensitivity analyses. Δ_{NW} is Conformer minus non-Whisper Transformer peak position. β_{Mixed} is the architecture coefficient in a model-family-random-intercept model. Only Wav2Vec2 spans both architecture labels, so the mixed-effects estimates are weakly identified.

Feature	Transformer		Conformer	
	Pos	Score	Pos	Score
F0 mean	0.167	0.662	0.067	0.651
F0 median	0.186	0.708	0.075	0.693
F0 min	0.286	0.179	0.283	0.187
F0 max	0.248	0.113	0.267	0.098
F1 mean	0.119	0.286	0.183	0.226
F1 median	0.138	0.283	0.200	0.239
F2 mean	0.154	0.280	0.170	0.346
F3 mean	0.138	0.362	0.114	0.346
F3 median	0.180	0.466	0.113	0.463
Intensity max	0.088	0.640	0.209	0.348
Intensity mean	0.152	0.532	0.154	0.376

Table 9: Selected acoustic feature positions and scores by architecture. Pos = mean peak layer position; Score = mean peak R^2 .

C Additional Metric Definitions

C.1 Peak Width

To measure whether features are sharply localized or broadly distributed, we computed **peak width** w_f as the proportion of layers where performance exceeds 70% of peak strength:

$$w_f = \frac{1}{L+1} \sum_{l=0}^L \mathbb{I}[\text{score}_f(l) \geq 0.7 \cdot s_f]$$

Low values indicate sharp, localized peaks; high values indicate distributed encoding.

C.2 Layer-wise Entropy

To measure the concentration of feature information, we computed the entropy of the normalized performance distribution across layers:

$$H_f = - \sum_{l=0}^L q_f(l) \log q_f(l)$$

where

$$q_f(l) = \frac{\text{score}_f(l)}{\sum_{l'} \text{score}_f(l')}$$

is the normalized performance at layer l . High entropy indicates distributed encoding; low entropy

indicates concentration in specific layers.

D Robustness Checks

D.1 Sensitivity Analysis: Robustness to Outliers

Given the small Conformer sample size ($N=7$) and the previously noted granite-speech-3.3-2b anomaly, we conducted a leave-one-out sensitivity analysis to assess whether our conclusions depend on this outlier.

Excluding granite-speech: With $N=6$ Conformers, the pooled differences become *stronger*, not weaker:

- **Gender:** $\Delta\pi$ changes from -0.126 to -0.154 ; p improves from 0.011 to 0.003 , now surviving Bonferroni correction
- **Accent:** $\Delta\pi$ remains stable at -0.23 ; $p = 0.007$
- **Phoneme:** $\Delta\pi$ remains stable at -0.29 ; $p = 0.015$

These leave-one-out results show that the pooled “Categorize Early” pattern is not driven by the granite-speech outlier. Because family-aware controls do not isolate architecture from model family and training regime, we interpret the strengthened difference as a robust pooled association rather than an architecture-specific effect. Indeed, granite-speech’s late Gender peak (0.32 vs. Conformer mean 0.13) was *attenuating* the pooled Conformer–Transformer contrast.

D.2 Quantifying the Multilingual Effect

To isolate whether the “Whisper Anomaly” reflects multilingual training data rather than other factors, we performed a paired comparison of multilingual vs. English-only (.en) Whisper models at matched scales (tiny, base, small, medium; $N=4$ pairs).

Results: Surprisingly, multilingual training shows *no significant effect* on peak positions:

- Gender: $\Delta = -0.004$, $p = 0.93$ (no difference)
- Accent: $\Delta = -0.002$, $p = 0.97$ (no difference)
- Phoneme: $\Delta = -0.115$, $p = 0.30$ (trend, not significant)

- Duration: $\Delta = -0.007$, $p = 0.85$ (no difference)

Interpretation: The delayed Gender and Accent peaks in Whisper models are not attributable to multilingual versus monolingual training status. Candidate explanations shared across Whisper models include (1) the scale of training data (680K+ hours), (2) the multitask objective (transcription, translation, and language identification), and (3) implementation details specific to the model family. The present comparison does not distinguish among these factors.

D.3 Performance-weighted Center of Mass

As a robustness measure less sensitive to a single maximum, we compare the primary peak positions with the performance-weighted center of mass c_f defined in Section 2.5. We set the no-skill baseline to 0 for R^2 probes, 0.5 for binary Gender, $1/6$ for Accent, and $1/39$ for Phoneme category. We compute feature-level values within corpus, average features into the five categories, and then average corpora within model, matching the primary peak-position aggregation. Table 11 reports model-level architecture-group means.

Center of mass agrees with the peak-position direction for Gender, Accent, Phoneme, and Duration. Acoustic changes from a near-zero positive peak contrast ($+0.004$) to a near-zero negative center contrast (-0.026); neither interval excludes zero. Thus, the only directional disagreement concerns a null pooled effect.

Feature	Conf.	Trans.	Δc	95% CI	p
Acoustic	.378	.403	-.026	[-.062, .023]	.326
Gender	.287	.367	-.080	[-.133, -.015]	.035
Accent	.382	.547	-.166	[-.228, -.099]	< .001
Phoneme	.369	.520	-.151	[-.232, -.073]	.008
Duration	.579	.515	.064	[-.035, .095]	.006

Table 11: Performance-weighted center-of-mass depth. Δc is Conformer minus Transformer; confidence intervals use 10,000 model bootstraps and p values are Welch tests.

D.4 Gold-transcript-only Replication

We repeat the Phoneme analysis using only CMU ARCTIC and L2-ARCTIC, whose alignments use gold transcripts. Table 12 retains an earlier pooled Conformer depth under both peak-position and center-of-mass metrics. This replication rules out automatic transcription as the source of the pooled Phoneme direction, but it does not remove model-family confounding.

Metric	Conf.	Trans.	Δ	95% CI	p
Peak position	.131	.383	-.252	[-.406, -.101]	.005
Center of mass	.393	.483	-.090	[-.161, -.020]	.039

Table 12: Gold-transcript-only phoneme replication. Differences are Conformer minus Transformer; confidence intervals use 10,000 model bootstraps and p values are Welch tests.

D.5 Leave-Cambridge-out Replication

We repeat the five pooled peak-position comparisons after excluding the private Cambridge Assessment corpus. Table 13 retains the original direction for all five categories. Gender, Accent, Phoneme, and Duration intervals exclude zero; Acoustic remains null. The headline pooled directions therefore do not depend on the private corpus.

Feature	Conf.	Trans.	$\Delta\pi$	95% CI	p
Acoustic	.191	.173	.018	[-.052, .120]	.731
Gender	.164	.288	-.124	[-.204, -.040]	.011
Accent	.337	.583	-.246	[-.349, -.126]	< .001
Phoneme	.169	.487	-.318	[-.458, -.177]	< .001
Duration	.703	.484	.220	[.073, .354]	.008

Table 13: Peak-position comparison after excluding Cambridge Assessment. $\Delta\pi$ is Conformer minus Transformer. Confidence intervals use 10,000 model bootstraps and p values are Welch tests.

D.6 Layer-truncation WER Experiment

We directly test the proposed early-exit prediction using Wav2Vec2-Large (Transformer) and Wav2Vec2-Conformer-Large checkpoints. Both are 24-layer, 1024-hidden English CTC systems fine-tuned on LibriSpeech-960h; the Transformer was pretrained on Libri-Light and the Conformer on LibriSpeech-960h, so this is a depth/task-matched deployment check rather than an isolated architecture experiment. Exact model and dataset revisions are pinned in the public manifest.

We retain the first 6, 12, 18, or all 24 encoder blocks together with each checkpoint’s original final normalization and CTC head. The deterministic evaluation slice contains the first 40 L2-ARCTIC examples in dataset order for each of six L1 labels (240 utterances; 2,257 reference words). Audio is resampled to 16 kHz; references and predictions are lowercased and stripped to ASCII letters, digits, and apostrophes. We use 10,000 paired utterance bootstraps for degradation relative to each checkpoint’s full depth.

Table 14 shows that both unmodified checkpoints collapse below full depth. Contrary to

the prediction, absolute Conformer WER degradation is larger at every truncated depth. The Conformer-minus-Transformer degradation differences are +.153 (95% CI [.112, .196]) at 25% depth, +.854 ([.713, 1.007]) at 50%, and +.712 ([.619, .813]) at 75%. Early linear accessibility is therefore insufficient evidence that an unmodified checkpoint can exit early; trained early-exit heads or streaming objectives require direct evaluation.

Depth	WER		Δ vs. full	
	Trans.	Conf.	Trans.	Conf.
25%	1.000	1.164	+.880	+1.034
50%	1.000	1.865	+.880	+1.735
75%	1.000	1.722	+.880	+1.592
100%	.120	.130	.000	.000

Table 14: WER after retaining the first fraction of 24 encoder blocks. WER can exceed 1 because insertions count as errors.

E Feature Definitions

This appendix summarizes the feature taxonomy used for probing. Full dataset-specific extraction parameters, preprocessing, and label construction are included in Appendix A.

- **Acoustic:** F0, formants (F1–F3), intensity, and related summary statistics.
- **Demographic:** Gender (binary) and Accent (categorical L1 background).
- **Phonetic:** phoneme identity (categorical).
- **Temporal:** utterance-level duration (continuous).

F Additional Results and Outliers

We provide expanded discussion of representational-strength trends, entropy/concentration summaries, and model-family heterogeneity. This includes additional detail on models that depart from canonical hierarchies and on objective-driven modulation (e.g., multitask training), complementing the main-text “hierarchy of controls” framing.

G Outliers and Training-Regime Analyses

Several models illustrate within-family heterogeneity in the pooled Transformer–Conformer comparison. For example, Whisper models delay Gender and Accent peaks relative to SSL Transformers, consistent with modulation by model

family or training regime. We also observe a small number of outliers (e.g., multimodal Phi-4; hubert-large; granite-speech) with atypical peak patterns or reduced probe strength. We retain these models to avoid selective exclusion and report robustness checks in Appendix D.

H Reproducibility

All analyses were conducted using Python 3.9 with NumPy 2.0, SciPy 1.13, and scikit-learn 1.6. Probe training used PyTorch 2.0. Code, analysis scripts, and redistributable outputs are available at <https://github.com/Nathan-Roll1/categorize-early-asr>. For the five architectural comparison t-tests (one per feature), we applied Bonferroni correction, yielding a corrected significance threshold of $\alpha = 0.05/5 = 0.01$.

The appendix robustness analyses use seed 20260829. Their pinned environment is Python 3.11.15, NumPy 2.0.2, Pandas 2.3.3, SciPy 1.13.1, PyTorch 2.8.0, Transformers 4.57.1, Datasets 3.0.1, and JiWER 4.0. The public repository includes the analysis scripts, exact input hashes and model/dataset revisions, aggregate result tables, and reproduction commands; private Cambridge-derived source records are not redistributed.

H.1 Software Versions

- Python 3.9
- NumPy 2.0.2
- Pandas 2.3.3
- SciPy 1.13.1
- Scikit-learn 1.6.1
- Statsmodels 0.14.6
- Matplotlib 3.9.2
- PyTorch 2.0

H.2 Computational Resources

Probing experiments were conducted on NVIDIA A100 GPUs. Each model’s full layer-wise probing (all features, all layers) required approximately 2–8 hours depending on model size.

H.3 Data Availability

The L2-ARCTIC dataset is publicly available. Cambridge Assessment is private and cannot be redistributed. Model weights are available

from HuggingFace Hub under their respective licenses (Apache 2.0 for most models). Code and redistributable probing outputs are available at <https://github.com/Nathan-Roll1/categorize-early-asr>.