Layer-wise Minimal Pair Probing Reveals Contextual Grammatical-Conceptual Hierarchy in Speech Representations

Linyang He* **Qiaolin Wang*** **Xilin Jiang** **Nima Mesgarani**
Department of Electrical Engineering, Columbia University
{linyang.he, qw2443, xj2289}@columbia.edu nima@ee.columbia.edu

Abstract

Transformer-based speech language models (SLMs) have significantly improved neural speech recognition and understanding. While existing research has examined how well SLMs encode shallow acoustic and phonetic features, the extent to which SLMs encode nuanced syntactic and conceptual features remains unclear. By drawing parallels with linguistic competence assessments for large language models, this study is the first to systematically evaluate the presence of contextual syntactic and semantic features across SLMs for self-supervised learning (S3M), automatic speech recognition (ASR), speech compression (codec), and as the encoder for auditory large language models (AudioLLMs). Through minimal pair designs and diagnostic feature analysis across 71 tasks spanning diverse linguistic levels, our layer-wise and time-resolved analysis uncovers that 1) all speech encode grammatical features more robustly than conceptual ones. 2) Despite never seeing text, S3M match or surpass ASR encoders on every linguistic level, demonstrating that rich grammatical and even conceptual knowledge can arise purely from audio. 3) S3M representations peak mid-network and then crash in the final layers, whereas ASR and AudioLLM encoders maintain or improve, reflecting how pre-training objectives reshape late-layer content. 4) Temporal probing further shows that S3Ms encode grammatical cues 500 ms before a word begins, whereas AudioLLMs distribute evidence more evenly—indicating that objectives shape not only *where* but also *when* linguistic information is most salient. Together, these findings establish the first large-scale map of contextual syntax and semantics in speech models and highlight both the promise and the limits of current SLM training paradigms.

1 Introduction

The past decade has witnessed transformative advancements in speech processing, driven by deep learning architectures that have largely supplanted traditional modular pipelines. Where early systems decomposed tasks into isolated stages of acoustic front-ends, phonetic decoders, and language models, modern approaches increasingly adopt end-to-end paradigms. Some architectures have emerged: (1) self-supervised speech models (S3Ms) that learn hierarchical representations from unlabeled audio through objectives like contrastive prediction (Mohamed et al., 2022); (2) automatic speech recognition (ASR) systems optimized for supervised speech-to-text mapping (Malik et al., 2021); (3) auditory large language models (AudioLLMs) that integrate speech processing with sophisticated language understanding (Wu et al., 2024a); and (4) discrete speech codec models—for example, EnCodec—that compress raw waveforms into compact, quantized codes, providing both efficient compression and rich features for downstream tasks (Défossez et al., 2022). These models achieve state-of-the-art performance across tasks ranging from speaker identification to speech recognition, yet fundamental questions persist about their linguistic encoding capabilities.

A critical unresolved question centers on whether modern speech models internalize high-level contextual syntax and semantics. Prior analyses have primarily focused on lower-level features—phonetics (Belinkov and Glass, 2017; Ma et al., 2021; Pasad et al., 2021; Wells et al., 2022; Abdullah et al., 2023; Choi and Yeo, 2022; Choi et al., 2024a), phonology (Martin et al., 2023), morphology (Pasad et al., 2024), speaker identity (Williams and King, 2019; Raj et al., 2019; de Seyssel et al., 2022; Raymondaud et al., 2024), word-level semantics (Pasad et al., 2021; Martin et al., 2023; Ashihara et al., 2023; Pasad et al.,

*Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35338–35353
November 4-9, 2025

2024; Choi et al., 2024b) and shallow syntax (Pasad et al., 2021; Martin et al., 2023; Shen et al., 2023; Mohebbi et al., 2023; Pasad et al., 2024).

Although these studies demonstrate that speech models effectively capture acoustic and lexical information, systematic investigations into context-dependent linguistic phenomena remain limited. Most existing benchmarks emphasize broad syntactic metrics (e.g., part-of-speech tagging, syntactic tree depth) and static, word-level semantic analyses. Such evaluations lack the granularity needed to probe long-range syntactic dependencies (e.g., island constraints, wh-movement) or semantic reasoning that depends on discourse context. Although many models achieve strong results on tasks like speech summarization and sentiment analysis, these contextual linguistic capabilities have not been subjected to rigorous, fine-grained scrutiny. This gap obscures our understanding of how speech models encode hierarchical linguistic structures beyond surface-level patterns.

This contrasts sharply with natural language processing (NLP), where minimal pair paradigms (Warstadt et al., 2020) have revealed how text-based models encode sophisticated syntactic dependencies (Hu et al., 2020; Gauthier et al., 2020; He et al., 2024b) and subtle semantic understanding (Misra et al., 2023). The absence of comparable methodologies in speech processing leaves a gap: *Can end-to-end speech systems, trained without explicit linguistic supervision, develop representations sensitive to complex grammatical phenomena and contextual meaning?*

To address this gap, we propose a new probing framework that transplants NLP's analytical rigor (He et al., 2024a) to speech model analysis. Our approach begins by synthesizing audio minimal pairs from two established NLP benchmarks: the BLIMP dataset (Warstadt et al., 2020) for grammaticality contrasts (e.g., filler gap dependency, negative polarity item licensing) and the COMPS dataset (Misra et al., 2023) for conceptual semantic relations. Through text-to-speech synthesis, we generate 116,300 audio pairs that isolate 67 syntactic phenomena and 4 semantic constructs via minimal-token substitutions—the first speech benchmark explicitly designed to evaluate hierarchical linguistic competence.

Subsequently, we systematically evaluate four speech model classes: S3Ms, ASR systems, AudioLLMs and codec. For each model, we extract layer-wise hidden states and train linear classifiers to distinguish acceptable vs. unacceptable constructs (e.g., "The key to the cabinets are..." vs. "The key to the cabinets is..."). Through this pipeline, we measure how sophisticated contextual syntax and semantics encoding evolve across layers, revealing whether speech models implicitly develop context-aware linguistic representations.

Our investigation yields several principal contributions:1) **How? form $\gg$ Meaning** – Across 16 models, syntactic and syntax–semantics interface contrasts are decoded 20 pp more accurately than conceptual contrasts, revealing a strong bias toward structural information. Syntax is already linearly separable in lower-middle layers, morphology emerges only in the top layers, and conceptual knowledge peaks around 65 %. 2) **Which? text-free models can win** – Self-supervised speech models outperform ASR encoders on every linguistic level—even though they never see text—demonstrating that rich grammar can arise from audio alone. 3) **Where? objective-specific layer tails** – S3M curves peak and then crash in the final layers, whereas ASR and AudioLLM encoders hold or improve: upper layers specialise for their training objective at the expense (S3M) or in favour (ASR/LLM) of linguistic encoding. 4) **When? temporal asymmetry** – despite bidirectional transformer design, temporal dynamics of S3Ms shows asymmetry. S3Ms' grammatical cues peaks at 500 ms *before* the critical word, while AudioLLMs accumulate evidence more evenly; thus training objectives determine not only where but also *when* information appears.

These findings provide a comprehensive evidence that end-to-end speech models implicitly learn sophisticated linguistic representations without text supervision. By bridging NLP probing methodologies with speech analysis, our work offers actionable insights for model interpretability and informs future architectures seeking to integrate speech and language understanding. All code, synthesized audio datasets, and probing results will be publicly released to facilitate reproducibility.

2 Minimal Pairs

Why minimal pairs? The *minimal-pair* paradigm is inherited from psycholinguistic acceptability-judgment experiments, where it offers the strictest possible control over confounds by differing in a *single* locus of variation (Chomsky, 2002). Because both sentences share identical

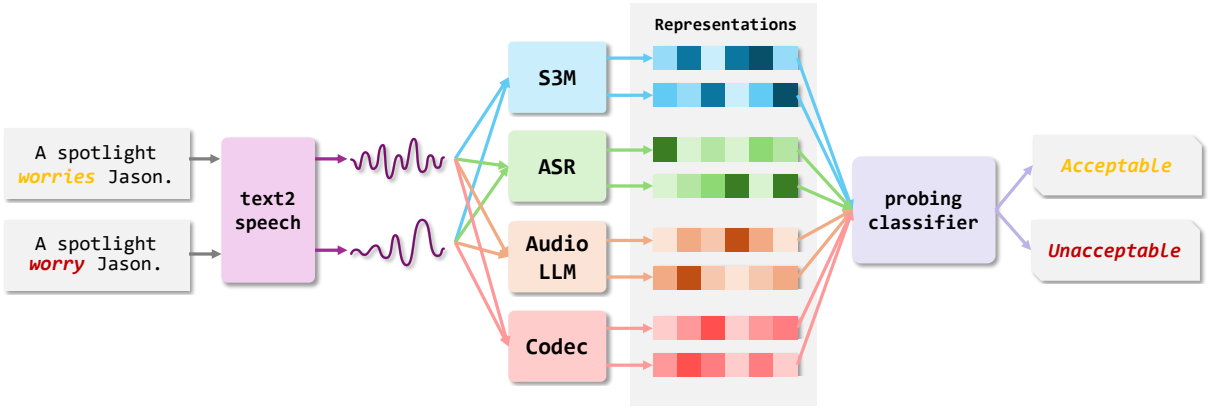

Figure 1: **Experimental design.** We synthesize speech minimal pairs from text-based benchmarks, convert them to audio, and probe speech models for contextual syntax and semantics.

lexical material and similar length, the probe must rely on the targeted grammatical constraint rather than surface n-gram statistics or memorised collocations. Consequently, minimal pairs have become the gold-standard diagnostic for whether a model truly internalises a rule, rather than merely approximating it through distributional heuristics (Linzen et al., 2016; Marvin and Linzen, 2018; Ettinger, 2020).

Adoption in NLP. A wide range of studies leverage minimal pairs to reveal systematic weaknesses that remain invisible to broad-coverage benchmarks: Linzen et al. (2016) show that LSTMs often fail long-distance subject–verb agreement; Marvin and Linzen (2018) extend this to hierarchical dependencies such as negative polarity licensing; Wilcox et al. (2018) and Gulordava et al. (2018) probe filler–gap dependencies and island constraints; McCoy et al. (2019) design HANS to expose models' lexical heuristics in natural-language inference; Warstadt et al. (2020) curate BLiMP, unifying 67 English phenomena into 67,000 minimal-pair templates; Misra et al. (2023) introduce the first set of *conceptual* minimal pairs and show that large language models (LLMs) encode conceptual knowledge far less consistently than they encode structural patterns. The paradigm has since been extended cross-lingually: Xiang et al. (2021) and Liu et al. (2024) target Chinese grammatical structure, while He et al. (2025) present X-COMPS, a suite of conceptual minimal pairs covering 17 languages to probe LLMs' multilingual generalization.

Together, this line of work shows that minimal-pair evaluations furnish a far more *fine-grained* lens on a model's linguistic competence than broad corpus-level scores: by isolating a single grammatical and conceptual contrast while holding all other lexical and length factors constant, they expose precisely which syntactic and compositional mechanisms the model has internalized—and which it still lacks.

Benefits for speech representations. For speech encoders, confounds multiply: local acoustic cues (e.g. coarticulation, energy) can correlate spuriously with grammaticality. Embedding-level minimal pairs neutralise such biases because the only acoustic difference is tied to the critical morpheme or word. In our setting, a single token replacement (e.g. *annoy* → *annoys*) can be synthesised with identical speaker, rate, and prosody, ensuring that any classification advantage stems from internal linguistic encoding rather than low-level acoustic artefacts.

Minimal pair details. Grammatical competence—including syntax, the syntax–semantics interface, and morphology—is probed with BLiMP (67,000 pairs) (Warstadt et al., 2020), whereas conceptual knowledge is evaluated using the COMPS suite (49,300 pairs) (Misra et al., 2023). Each item consists of an ACCEPTABLE sentence S^+ and an UNACCEPTABLE counterpart S^- that differ in exactly one token or morpheme, thereby targeting a single phenomenon. We cover four levels (* denotes unacceptable):

(i) **syntax** (e.g., filler-gap dependency)

a) Mark figured out that most governments appreciate Steve.

*b) *Mark figured out who most governments appreciate Steve.*

(ii) **syntax–semantics interface** (e.g., Negative Polarity Item (NPI) licensing)

a) Even Suzanne has really joked around.

*b) *Even Suzanne has ever joked around.*

(iii) **morphology** (e.g., subject verb agreement)

a) The hospital appreciates Claire.

*b) *The hospitals appreciates Claire.*

(iv) **conceptual meaning**

a) A kettle is used for boiling.

*b) *A hammer is used for boiling.*

These four linguistic levels cover across 67 grammatical tasks and 4 conceptual tasks, whichs yields 116,300 English pairs, each generated from linguist-crafted templates and validated to 96.4% (BLiMP) and 93.1%(COMPS) human agreement.

3 Experimental Setup

Model details could be found in the Appendix B.

Diagnostic probing with minimal pairs. To quantify the extent to which linguistic knowledge is encoded at different layers of pretrained models, we perform diagnostic probing on minimal pairs using a simple sentence-level classifier. For each sentence S, we extract hidden representations $\mathbf{h}^{(l)}(S) \in R^{T\times d}$ from every Transformer layer l of a frozen model, where T is the number of tokens (or audio frames) and d is the hidden size. To obtain a fixed-length representation, we apply mean pooling across the token dimension, i.e.,

$$\bar{\mathbf{h}}^{(l)}(S) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t^{(l)}(S),$$

which yields a single d-dimensional vector per sentence. This simple aggregation strategy has been shown to preserve sentence-level distinctions relevant to grammatical acceptability while avoiding task-specific tuning (Pasad et al., 2021, 2023).

We then train a logistic regression probe g on the pooled vector $\bar{\mathbf{h}}^{(l)}(S)$ to predict the acceptability label $z \in \{0, 1\}$ for each sentence in a minimal pair, where $z = 1$ indicates acceptable. All probes are trained using five-fold cross-validation, and evaluated using accuracy as the primary metric. By restricting the probe to a linear classifier and keeping the encoder frozen, we ensure that any performance gain reflects information already encoded in the model's representations, rather than introduced through fine-tuning. This setup enables a layer-wise analysis of where and how linguistic distinctions—defined by minimal contrasts—are captured in the model.

Probing positional representations. To further examine the impact of token aggregation strategy, we additionally evaluate a set of single-token probes, using the hidden state at fixed relative positions (0%, 25%, 50%, 75%, and 100%) along the token sequence. This allows us to assess whether specific frames disproportionately encode linguistic features, and to compare the robustness of mean pooling against position-specific representations.

Temporal probing. For each sentence we locate the onset of the *critical word*—the element that differentiates the acceptable and unacceptable member of a minimal pair—using WhisperX forced alignment (Bain et al., 2023). Around this onset we take a $\pm 1000ms$ window and, from the layer that yielded the best mean-pool accuracy, sample individual token embeddings at increasingly fine steps toward the centre. Each embedding is fed to a logistic-regression probe, allowing us to trace how well linguistic contrasts are encoded at each time point.

Although the encoders are bidirectional, this analysis still pinpoints *where* in the acoustic stream the model's internal state carries the strongest grammatical signal. In continuous speech—where token boundaries are fuzzy—such temporal profiles reveal how structural and conceptual information is distributed over time rather than *when* it is first detected, highlighting the regions the model relies on most for encoding linguistic distinctions.

Selection score. To reflect asymmetric effects of architectural bias, we define a selection score that adjusts the trained model's accuracy based on whether the untrained model performs above or below chance level:

$$Selec = Acc_{\text{trained}} \cdot \left(1 + \frac{0.5 - Acc_{\text{untrained}}}{0.5}\right)$$

This formulation rewards stronger gains when the untrained model performs below chance, and penalizes when it performs above chance. For speech models, although untrained networks contain no learned parameters, the input spectrograms often carry meaningful structure that can lead to above-chance performance. The selection score thus quantifies the improvement achieved by training, relative to the representational bias inherent in both the model architecture and the input features.

Confidence score. The confidence score is computed as follows: First, we define the set of correctly classified samples as $\mathcal{A} = \{i \mid y_{\text{true},i} = y_{\text{pred},i}\}$, where y_{true} represents the ground truth labels and y_{pred} represents the predicted labels. Next,

for each correctly classified sample, we extract the maximum predicted probability across all classes and compute the mean confidence score for these samples as

$$Conf = (1/|\mathcal{A}|) \sum_{i \in \mathcal{A}} \max_j P_{i,j}$$

, where $P_{i,j}$ is the probability of sample i belonging to class j. Measuring the confidence score allows us to assess how certain the model is when it makes correct predictions, offering insight into the calibration of its probabilistic outputs. A well-calibrated model should not only predict correctly but also assign high confidence to those correct predictions.

4 Results

4.1 Trained representations cluster by linguistic category and task

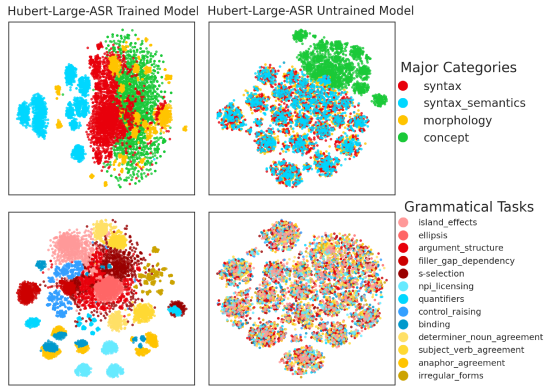

Figure 2: **t-SNE visualisation of minimal–pair Δ-embeddings in HuBERT-Large-ASR (layer 18).** For each sentence we compute the difference vector $\Delta\mathbf{h} = \bar{\mathbf{h}}(S^+) - \bar{\mathbf{h}}(S^-)$ between the acceptable and unacceptable member of a minimal pair, then project all $\Delta\mathbf{h}$ to two dimensions with t-SNE (perplexity = 50). *Top-left:* trained model coloured by major category (Syntax •, Syntax–Semantics Interface •, Morphology •, Concept •). *Top-right:* randomly initialised model. *Bottom-left:* trained model coloured by twelve sub-tasks within the first three categories. *Bottom-right:* corresponding untrained control.

The trained encoder forms well-defined clusters that are absent in the randomly-initialised network, indicating that self-supervised pre-training imbues the model with embeddings that systematically separate grammatical contrasts. Among the four major categories, Syntax, Syntax–Semantics interface, and Morphology occupy distinct regions, whereas Conceptual contrasts remain intermixed—mirroring the weaker probing accuracies reported earlier. Finer structure emerges in the bottom-left panel: sub-tasks such as island effects, ellipsis, and agreement each form coherent subclusters, showing that the model encodes nuanced distinctions within broader grammatical families. In the untrained baseline (bottom right) these patterns collapse into a homogeneous cloud, confirming that the observed structure arises from linguistic learning rather than architectural priors alone.

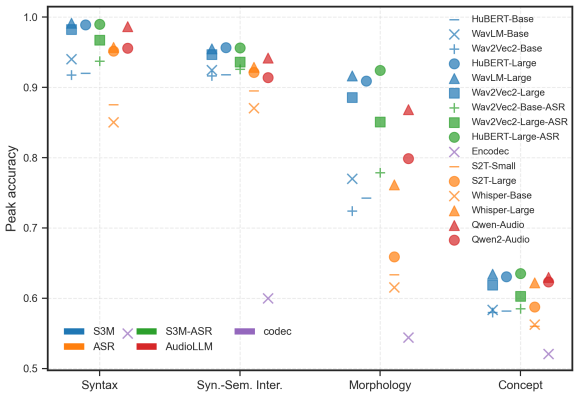

Figure 3: **Peak minimal-pair probing accuracy across linguistic levels.** For each model, we report the highest sentence-level accuracy attained by any layer on syntax, syntax–semantics interface, morphology, and concept tasks.

4.2 Speech models encode form better than meaning.

We first demonstrate a clear disparity in how effectively speech models capture linguistic "form" versus "meaning". As evident from Figure 3, grammatical minimal pairs (including syntax, syn.-sem. interface and morphology) are consistently distinguished better with a higher Acc score than conceptual pairs for all models. This performance difference in the structured nature of syntactic and morphological rules makes them more readily learnable from acoustic signals alone, while the abstraction required for conceptual understanding poses greater challenges. The robust encoding of linguistic form indicates that speech models implicitly acquire rule-based structural regularities even in the absence of explicit textual guidance.

4.3 Speech transformer layers show hierarchical linguistic features.

As shown in Figure 5, we observe a clear hierarchical emergence of linguistic representations across transformer layers in speech models. Syntactic and syntax–semantics interface features are learned earliest, showing robust gains beginning from the lower-middle layers and typically peaking around the upper-middle layers. Morphological features, in contrast, emerge much later—often only becoming linearly decodable in the final few layers.

Figure 4: **Peak accuracy for every individual minimal-pair task across models.** Rows list the 16 speech encoders; columns list the 67 grammatical tasks: syntax (red), syntax-semantics interface (blue), morphology (yellow) and 4 conceptual tasks (green).

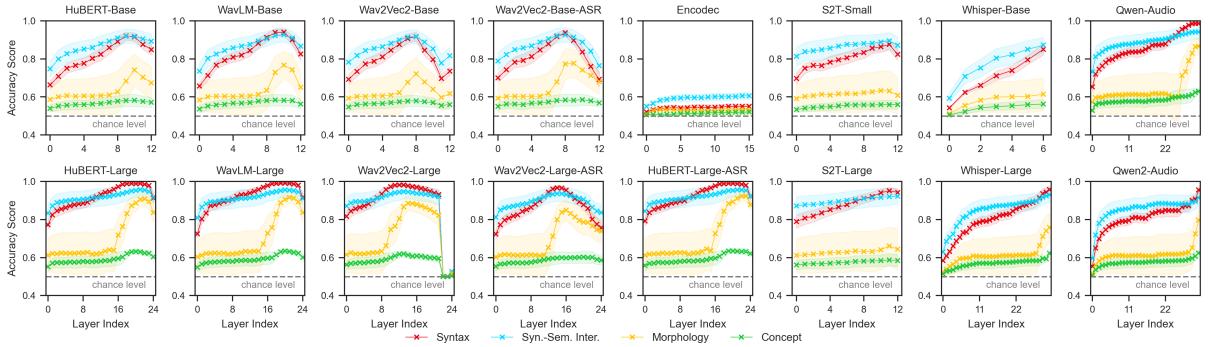

Figure 5: **Layer-wise trajectory of linguistic accuracy for each speech model.** For every encoder we plot sentence-level probing accuracy (y-axis) as a function of Transformer depth (layer index, x-axis). Shaded bands denote ±1 standard error over five cross-validation folds.

Conceptual features exhibit a different trend: while they become distinguishable earlier than morphology, their decodability remains consistently weak across all layers.

This pattern suggests that structural distinctions like syntax can be encoded relatively early from local phonological and prosodic cues, while morphological information requires more global context and fine-grained form recognition—hence emerging later. Conceptual meaning, on the other hand, likely requires a fundamentally different abstraction capacity that is not easily captured by current speech models, even at their deepest layers.

Interestingly, while larger models (e.g., Wav2Vec2-Large, HuBERT-Large, WavLM-Large) consistently outperform their base counterparts, we find that performance for most linguistic categories saturates before the final few layers. This supports the view that uppermost layers in self-supervised models are often specialized for pretraining objectives (e.g., contrastive or masked prediction), and that mid-to-upper layers more faithfully encode linguistic structure. These findings highlight a progressive and modular abstraction process in speech transformers, wherein linguistic features emerge at different depths according to their complexity and signal grounding.

4.4 Self-supervised speech models can learn linguistic knowledge without text

Self-supervised speech models (S3Ms), trained without textual supervision, exhibit remarkable ca-

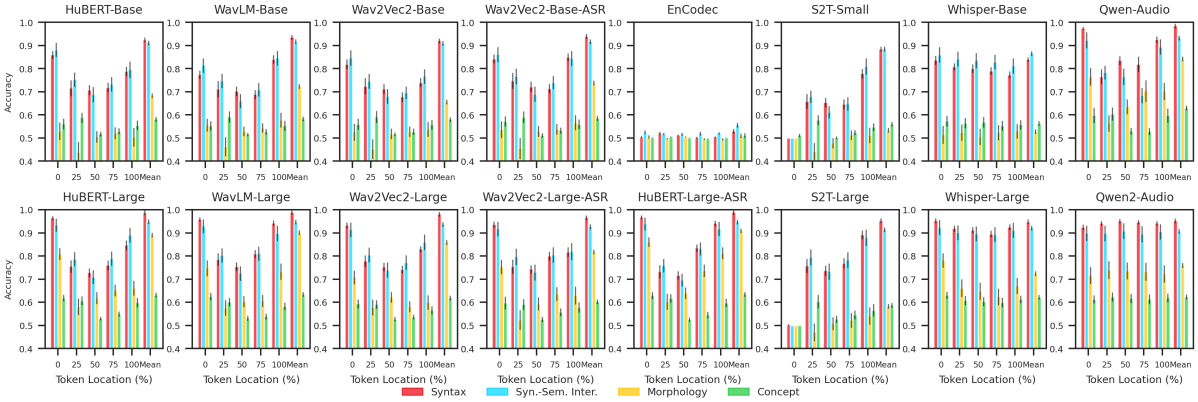

Figure 6: **Comparing single-token versus mean pooling across speech models.** We evaluate probing accuracy when sentence embeddings are extracted from a single token at five relative positions (0%, 25%, 50%, 75%, 100%) or via mean pooling across all tokens ("Mean"). Accuracy is measured on the layer that gave the best mean-pool score.

pabilities in acquiring linguistic knowledge solely from audio signals, as illustrated in Figure 3 and 5. Specifically, wav2vec 2.0, HuBERT, and WavLM achieve impressive performance across linguistic levels, with accuracy scores exceeding 90% on syntactic and syntax–semantics interface tasks, reaching 80–90% on morphology, and around 60% on conceptual distinctions. Surprisingly, these models outperform ASR encoders trained with explicit text supervision.

One plausible explanation is that ASR models, while grounded in linguistic output, are optimized primarily for surface-level transcription. This objective may bias them toward segmental and lexical patterns that are useful for decoding text but less aligned with abstract grammatical distinctions. In contrast, S3Ms benefit from pretraining objectives such as contrastive learning (wav2vec 2.0) and masked prediction (HuBERT, WavLM), which promote contextual abstraction and robustness beyond frame-level cues. These results highlight the promise of self-supervised acoustic pretraining as a pathway toward generalizable linguistic learning from raw audio.

4.5 Architectural and training differences shape linguistic depth

While ASR and AudioLLM models share similar encoder backbones (e.g., Transformer-based encoders with convolutional frontends), their training paradigms differ substantially. AudioLLMs (e.g., Qwen-Audio) are trained under multimodal generative objectives within large-scale LLM frameworks. As shown in Figure 3, their performance on syntactic and morphological tasks consistently exceeds that of ASR models, suggesting that joint language–audio modeling drives richer linguistic abstraction even in the encoder alone. This indicates that the nature of the downstream supervision—not just the model capacity—plays a critical role in shaping linguistic competence.

In addition, we observe a distinct architectural signature: across all S3M models, performance drops sharply in the final few layers. This effect is notably absent in ASR and AudioLLM encoders, where accuracy continues to rise or plateaus in the top layers. The late-layer drop in S3Ms likely reflects a shift in focus toward pretraining-specific representations—such as contrastive codebook prediction or masked target reconstruction—rather than general-purpose linguistic encoding. This divergence underscores a fundamental difference: while S3Ms produce general linguistic features as a byproduct of self-supervised objectives, ASR and LLM-based models are tuned explicitly for language-level outputs, preserving linguistic structure deeper into the network.

4.6 Mean pooling outperforms single-token alternatives in speech models

A common practice in large language model research is to extract sentence-level representations using the final token embedding—particularly in autoregressive (unidirectional) architectures. To evaluate whether similar heuristics apply to speech models, we compare mean pooling with single-token probing at five fixed relative positions (0%, 25%, 50%, 75%, and 100%) along the token sequence. As shown in Figure 6, mean pooling consistently yields higher probing accuracy across all linguistic categories and models, confirming its robustness for speech-based sentence embedding.

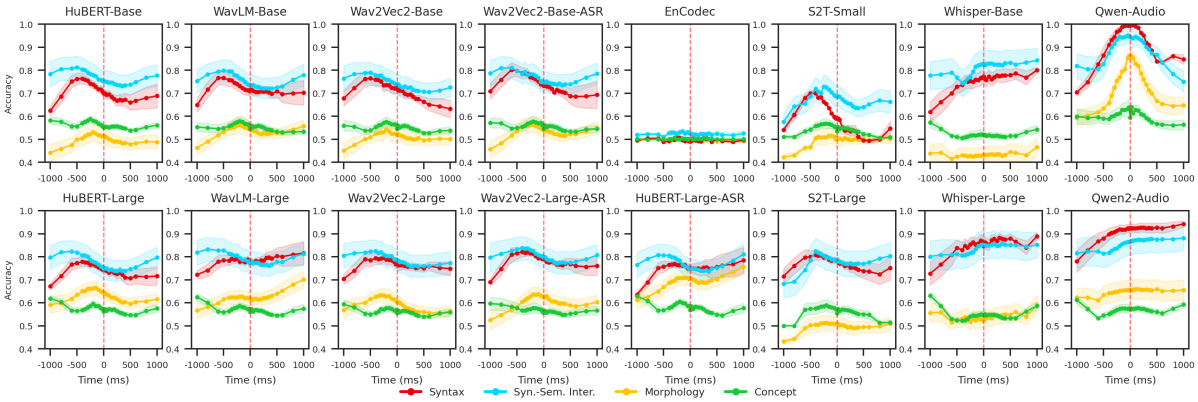

Figure 7: **Temporal dynamics of linguistic information around the critical word.** For each model we plot probing accuracy as a function of time relative to the onset of the critical word (vertical dashed line = 0 ms). Accuracy is measured on the layer that gave the best mean-pool score.

Surprisingly, although most speech models are based on bidirectional Transformers and thus in principle have symmetric access to context, we consistently find that the first token (0%) encodes more linguistic information than the final token (100%). This asymmetry is especially evident in models like HuBERT, Wav2Vec2, and WavLM. One possible explanation is that the first token may serve as an implicit global aggregator, especially in architectures without explicit [CLS] tokens. During training, early tokens might become more sensitive to contextual summaries or more stable across time, while final tokens are often affected by alignment noise or signal boundary effects. This positional imbalance suggests that even in bidirectional encoders, representational asymmetry can emerge as a byproduct of training dynamics and audio processing pipelines. While mid-sequence tokens (e.g., 25%, 50%) sometimes approach mean pooling performance, no single-token baseline reliably matches it, reinforcing the importance of full-sentence integration in capturing hierarchical linguistic features in speech.

4.7 Temporal asymmetry despite bidirectional context

As Figure 7 shows, although all encoders are bidirectional, their temporal profiles are strikingly *asymmetric*. In the three self-supervised families (HuBERT, WavLM, Wav2Vec2) and in the S2T ASR models, syntactic and interface accuracy peaks -600 to -500 ms **before** the critical word begins, then decays after onset. A plausible explanation is that these models learn to anticipate upcoming phone sequences from co-articulatory cues: masked-prediction and CTC-style contrastive objectives reward early integration of left-context information, causing pre-onset tokens to act as forward predictors.

By contrast, Whisper and Qwen-Audio—whose encoders are trained inside multimodal generative frameworks—display a flatter, gradually rising curve that culminates around or shortly after onset. These models appear to accumulate evidence more evenly across the word, perhaps because their objectives (sequence-to-sequence transcription or audio-text alignment) emphasise complete lexical identification rather than early prediction.

Morphological information emerges later and more slightly symmetrically, reflecting the need for full word forms, while conceptual distinctions remain weak throughout the window. EnCodec stays at chance, confirming the absence of linguistically relevant structure in a purely compression-oriented codec. Together, these results show that speech encoders encode grammatical distinctions at different *timescales*: S3M and S2T models exploit anticipatory acoustic cues, whereas AudioLLM encoders rely on longer integration windows—revealing how training objectives shape not only *where* but also *when* linguistic knowledge is realised in the representation.

5 Conclusion

By first introducing minimal pairs into speech probing, we show that contemporary speech transformers internalise a structured linguistic hierarchy. Our results suggest that evaluation of spoken-language understanding should move beyond coarse benchmarks to fine-grained, temporally resolved diagnostics.

Limitation

Language and audio diversity

All experiments use English sentences rendered by a single, studio-quality TTS voice. Consequently, the reported hierarchy may be specific to English morpho-syntax and to clean, single-speaker recordings. Languages with richer inflection (e.g., Turkish) or tonal contrasts (e.g., Mandarin), as well as spontaneous, multi-speaker, or noisy speech, might yield very different layer-wise and temporal patterns.

Scope of the linguistic tests

Our 116,300 minimal pairs target core morpho-syntactic phenomena but exclude discourse-level skills such as anaphora, presupposition, or pragmatic implicature. The strong form–meaning gap we observe therefore says little about how models handle context-dependent or world-knowledge-rich interpretations that dominate real conversation.

Bidirectionality and temporal claims

The encoders are bidirectional, so pre-onset peaks do not prove real-time anticipation; they merely locate where the internal state most strongly distinguishes the contrast. True causal timing would require strictly streaming models or masking future context during probing.

Model and parameter range

We analyse 16 open-weight encoders while larger proprietary AudioLLMs and causal streaming ASR systems are absent, and scaling laws above the tested range remain unexplored.

Absence of speech-specific minimal pairs

Our benchmark manipulates linguistic content while holding the acoustic signal constant, but it does not include *phonetic* or *prosodic* minimal pairs—e.g. vowel length contrasts, stress-shift alternations, or intonation-driven meaning changes. As a result, we cannot assess whether the same hierarchical pattern extends to speech-specific cues such as co-articulation, pitch accent, or rhythm, nor can we determine how lexical and segmental information interact in the encoder. A companion suite of acoustically controlled minimal pairs would be required to map the boundary between linguistic and purely phonetic knowledge in self-supervised speech models.

No neuron-level dissection

All conclusions are drawn from layer-global probes; we do not inspect individual neurons or sparse subnetworks that might carry specialised features. Prior work in text LLMs has shown that single units can act as high-precision detectors for concepts like negation or syntax islands. Without analogous neuron-level analyses, we cannot say whether speech encoders store linguistic knowledge in distributed patterns, in a handful of specialised channels, or in attention heads. Such fine-grained dissection could reveal bottlenecks, redundancies, or pathways amenable to targeted editing.

Unanalysed attention structure

Transformer attention maps provide a window into how models route information, yet our study treats them as a black box. We do not test whether attention heads focus on the critical word, respect syntactic dependencies, or exhibit long-range patterns akin to filler–gap tracking. Consequently, we cannot link the probing results to specific computational mechanisms inside the model. Future work combining attention diagnostics with our temporal and layer-wise probes could show *how*—not just *where*—speech transformers build the observed hierarchy of linguistic representations.

References

Badr M Abdullah, Mohammed Maqsood Shaik, Bernd Möbius, and Dietrich Klakow. 2023. An information-theoretic analysis of self-supervised discrete representations of speech. *arXiv preprint arXiv:2306.02405*.

Takanori Ashihara, Takafumi Moriya, Kohei Matsuura, Tomohiro Tanaka, Yusuke Ijima, Taichi Asami, Marc Delcroix, and Yukinori Honma. 2023. Speechglue: How well can self-supervised speech models capture linguistic knowledge? *arXiv preprint arXiv:2306.08374*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Yonatan Belinkov, Ahmed Ali, and James Glass. 2019. Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition. *arXiv preprint arXiv:1907.04224*.

Yonatan Belinkov and James Glass. 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems. *Advances in Neural Information Processing Systems*, 30.

Xuankai Chang, Brian Yan, Kwanghee Choi, Jee-Weon Jung, Yichen Lu, Soumi Maiti, Roshan Sharma, Jiatong Shi, Jinchuan Tian, Shinji Watanabe, et al. 2024. Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11481–11485. IEEE.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Kwanghee Choi, Jee-weon Jung, and Shinji Watanabe. 2024a. Understanding probe behaviors through variational bounds of mutual information. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5655–5659. IEEE.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024b. Self-supervised speech representations are more phonetic than semantic. *arXiv preprint arXiv:2406.08619*.

Kwanghee Choi and Eun Jung Yeo. 2022. Opening the black box of wav2vec feature encoder. *arXiv preprint arXiv:2210.15386*.

Noam Chomsky. 2002. *Syntactic structures*. Mouton de Gruyter.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Yu-An Chung, Yonatan Belinkov, and James Glass. 2021. Similarity analysis of self-supervised speech representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3040–3044. IEEE.

Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. 2022. Probing phoneme, language and speaker information in unsupervised speech representations. *arXiv preprint arXiv:2203.16193*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R Brennan. 2024a. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497.

Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Adrian Florea, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, et al. 2025. Xcomps: A multilingual benchmark of conceptual minimal pairs. *arXiv preprint arXiv:2502.19737*.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. 2024b. Large language models as neurolinguistic subjects: Discrepancy in performance and competence for form and meaning. *arXiv preprint arXiv:2411.07533*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. 2022. istftnet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6207–6211. IEEE.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36:19594–19621.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, et al. 2024. Zhoblimp: a systematic assessment of language models with linguistic minimal pairs in chinese. *arXiv preprint arXiv:2411.06096*.

Danni Ma, Neville Ryant, and Mark Liberman. 2021. Probing acoustic representations for phonetic properties. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 311–315. IEEE.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.

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. *arXiv preprint arXiv:2306.06232*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2920–2941.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023. Homophone disambiguation reveals patterns of context mixing in speech transformers. *arXiv preprint arXiv:2310.09925*.

Anna Ollerenshaw, Md Asif Jalal, and Thomas Hain. 2022. Insights on neural representations for end-to-end speech recognition. *arXiv preprint arXiv:2205.09456*.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

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.

Archiki Prasad and Preethi Jyothi. 2020. How accents confound: Probing for accent information in end-to-end speech recognition systems. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3739–3753.

Yao Qian, Ximo Bianv, Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao, and Michael Zeng. 2021. Speech-language pre-training for end-to-end spoken language understanding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7458–7462. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur. 2019. Probing the information encoded in x-vectors. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 726–733. IEEE.

Quentin Raymondaud, Mickael Rouvier, and Richard Dufour. 2024. Probing the information encoded in neural-based acoustic models of automatic speech recognition systems. *arXiv preprint arXiv:2402.19443*.

Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah. 2021. What all do audio transformer models hear? probing acoustic representations for language delivery and its structure. *arXiv preprint arXiv:2101.00387*.

Gaofei Shen, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2023. Wave to syntax: Probing spoken language models for syntax. *arXiv preprint arXiv:2305.18957*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R

Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Dan Wells, Hao Tang, and Korin Richmond. 2022. Phonetic analysis of self-supervised representations of english speech. In *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*, pages 3583–3587. ISCA.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Jennifer Williams and Simon King. 2019. Disentangling style factors from speaker representations. In *20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language*, pages 3945–3949. ISCA.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-wei Chang, Ho-Lam Chung, Alexander H Liu, and Hung-yi Lee. 2024a. Towards audio language modeling–an overview. *arXiv preprint arXiv:2402.13236*.

Junkai Wu, Xulin Fan, Bo-Ru Lu, Xilin Jiang, Nima Mesgarani, Mark Hasegawa-Johnson, and Mari Ostendorf. 2024b. Just asr+ llm? a study on speech large language models' ability to identify and understand speaker in spoken dialogue. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1137–1143. IEEE.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. Climp: A benchmark for chinese language model evaluation. *arXiv preprint arXiv:2101.11131*.

Shu-wen Yang, Andy T Liu, and Hung-yi Lee. 2020. Understanding self-attention of self-supervised audio transformers. *arXiv preprint arXiv:2006.03265*.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. 2025. Speech self-supervised representations benchmarking: a case for larger probing heads. *Computer Speech & Language*, 89:101695.

A Related Work

A.1 Low-Level Acoustic and Phonetic Probing

Early investigations focused mainly on how neural models capture fundamental acoustic signals and phonetic details. The analyses of end-to-end ASR systems by (Belinkov and Glass, 2017; Belinkov et al., 2019) revealed that phonetic and graphemic information is distributed in a layer-dependent manner, highlighting the role of early layers in the capture of fine-grained details. Extending these insights to self-supervised settings, (Ma et al., 2021) and (de Seyssel et al., 2022) demonstrated that models such as Contrastive Predictive Coding effectively capture phoneme-level distinctions. Researchers have employed techniques such as canonical correlation analysis and mutual information measures to trace the evolution of these representations. These methods reveal that early layers are particularly sensitive to subtle cues—including accent variations and phonetic nuances—as shown by studies such as (Prasad and Jyothi, 2020; Raymondaud et al., 2024). Comparative analyses of transformer architectures further illustrate diverse learning patterns for low-level acoustic features (Chung et al., 2021; Yang et al., 2020; Shah et al., 2021), underscoring the foundational role of these representations in speech processing.

A.2 Linguistic Feature Probing

Beyond basic acoustic and lexical feature analyses, several studies have probed deeper linguistic encodings in speech models. Layer-wise investigations revealed that phonology and morphology emerge predominantly in lower to mid layers (Pasad et al., 2021, 2024), while speaker-identity probes demonstrated partial disentanglement of speaker-specific traits from linguistic content (Williams and King, 2019; Raj et al., 2019). Word-level semantic evaluations—most notably via SpeechGLUE—have assessed static lexical semantics, indicating that many models still prioritize phonetic similarity over deeper meaning (Ashihara et al., 2023). Shallow syntactic metrics (e.g., part-of-speech tags, dependency tree depth) have also been applied, suggesting that basic syntactic cues are learnable but without distinguishing long-range dependencies (Martin et al., 2023; Pasad et al., 2021; Shen et al., 2023).

A.3 Probing in Downstream Tasks and Integrated Systems

The practical utility of speech representations is underscored by their impact on downstream applications, where the quality of internal representations directly influences system performance. Early efforts in this domain include the unified speech language pre-training framework proposed by (Qian et al., 2021), which integrates an ASR encoder with a language model to improve spoken language understanding. Subsequent studies by (Chang et al., 2024) and (Abdullah et al., 2023) have explored the use of discrete speech units to compress input sequences while maintaining robust performance in ASR, translation, and comprehension tasks. Benchmark studies (Zaiem et al., 2025) further highlight that the architectural design of probing heads can critically affect model rankings and task outcomes. Statistical analyzes (Ollerenshaw et al., 2022) linking layer-wise properties to recognition accuracy underscore the direct relationship between internal representations and practical performance. More recent evaluations of emerging SpeechLLMs (Wu et al., 2024b) point to challenges such as limited speaker awareness, even as these models demonstrate impressive capabilities in context-based dialogue tasks. This line of inquiry emphasizes that the effectiveness of speech models in real-world applications depends on a delicate balance between capturing low-level signals and encoding higher-level linguistic and contextual features.

B Models Details

B.1 Probed Models

Self-Supervised Speech Models (S3Ms) Modern self-supervised learning paradigms learn robust speech representations through pre-training objectives that do not require transcribed data:

- **HuBERT** (Hsu et al., 2021): HuBERT employs an iterative clustering procedure combined with a masked prediction objective to jointly model acoustic and linguistic characteristics. We analyze both the `facebook/hubert-large-ll60k` and the `facebook/hubert-base-ls960` variants in our experiments.
- **wav2vec 2.0** (Baevski et al., 2020): This model employs latent space masking with contrastive learning, where representations are learned jointly with

a quantized codebook. Our study includes the `facebook/wav2vec2-base` and `facebook/wav2vec2-large-lv60` models.

- **WavLM** (Chen et al., 2022): WavLM extends self-supervised learning by integrating a denoising objective with masked speech prediction, thereby enhancing its robustness across a variety of speech tasks. We consider both the `microsoft/wavlm-base-plus` and the `microsoft/wavlm-large` variants.

Automatic Speech Recognition (ASR) Models ASR models typically achieve high transcription accuracy while embedding additional linguistic biases. Our experiments probe only the encoder components of these models, extracting representations before the decoder stage:

- **Speech-to-Text Models** (Wang et al., 2020): The S2T model provides transformer-based architectures for end-to-end speech recognition and translation, and integrates acoustic modeling with language model fusion. We study both the `facebook/s2t-small-librispeech-asr` and `facebook/s2t-large-librispeech-asr` configurations to assess how model complexity influences downstream performance.

- **Whisper** (Radford et al., 2023): Trained in a weakly supervised extensive corpus, Whisper demonstrates robust performance in various speech conditions. Our experiments involve the `openai/whisper-base` and `openai/whisper-large` variants.

Auditory Large Language Models (AudioLLMs) AudioLLMs fuse the strengths of acoustic modeling with the advanced reasoning capabilities of large-language models. Similar to our ASR experiments, we probe only the audio encoder components of these models, allowing direct comparison with other speech encoders:

- **Qwen Audio** (Chu et al., 2023): A multimodal large language model that accepts speech, natural sounds, music and text, producing textual outputs. It employs a multi-task learning framework that enables knowledge sharing across more than 30 tasks while avoiding one-to-many interference.

- **Qwen2 Audio** (Chu et al., 2024): An evolution of the original Qwen Audio. It integrates a dedicated audio front-end (initialized from Whisper) with the Qwen-7B language model backbone. It simplifies pre-training by using natural language prompts rather than hierarchical tags.

EnCodec (Défossez et al., 2022) Encodec is a neural audio codec that provides real-time, high-fidelity compression through a streaming encoder-decoder architecture with quantized latent space. It employs a multiscale spectrogram adversary to reduce artifacts and a novel loss balancer mechanism. Our experiments utilize the `facebook/encodec_24khz` model variant. For probing EnCodec, we extract representations from each quantizer by passing audio through the encoder and accessing the discrete codes at each quantizer. This allows us to analyze how linguistic information is encoded at different levels of the quantization hierarchy.

B.2 Text2Speech Generation

We generated speech minimal pairs by synthesizing speech from minimal pairs in text using the Kokoro-82M text-to-speech (TTS) model. Kokoro is a state-of-the-art human-level TTS model based on the StyleTTS 2 (Li et al., 2023) architecture and ISTFTNet (Kaneko et al., 2022) vocoder. The output 24 KHz speech waveforms were subsequently resampled to 16 kHz, aligning with the sampling rate of speech models. Additionally, we manually reviewed synthesized speech samples to ensure they contained no acoustic glitches and that the incorrect sentences were not unintentionally auto-corrected.

C Ablation: probing with matched random embeddings.

To ensure that our probing results reflect meaningful structure in model representations—rather than spurious correlations exploited by the linear classifier—we conduct a control experiment using randomly generated embeddings.

Specifically, for each sentence, we generate a random vector sampled from a Gaussian distribution that matches the mean and standard deviation of the original sentence embeddings from the best-performing layer of the model. To preserve the pairing structure of the data, we assign the same random vector to both minimal-pair variants if they are based on the same audio clip. These random

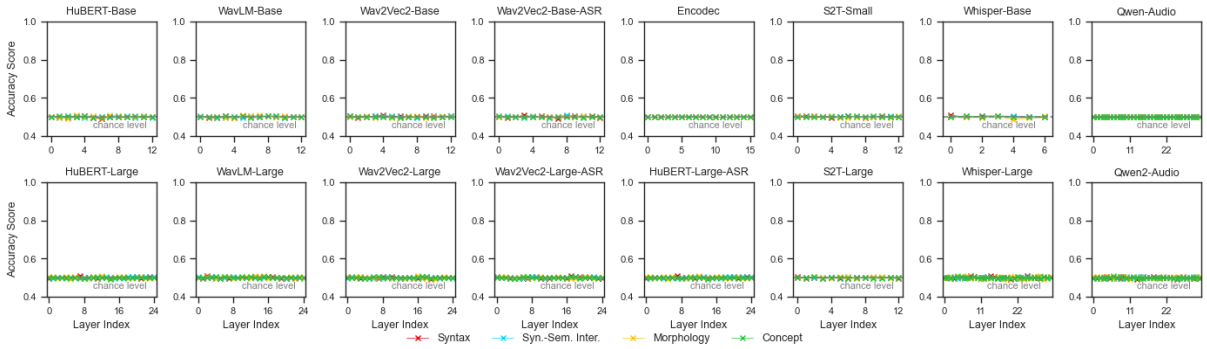

Figure 8: **Control experiment with random embeddings.** To verify that probe accuracy reflects genuine linguistic encoding rather than dataset biases, we conduct an ablation in which each speech input is replaced with a random embedding sampled from a Gaussian distribution matching the original mean and standard deviation of that model's sentence embeddings.

embeddings are then passed through the same probing pipeline, including five-fold logistic regression and accuracy evaluation.

Unlike the *randomly initialized model* condition—which still processes structured spectrograms through a frozen encoder—this ablation entirely removes content information by bypassing the encoder and directly replacing model representations with noise. As shown in Figure 8, probing accuracy drops to chance for all models and linguistic categories, confirming that the original signal arose from linguistic encoding in the representations—not from classifier bias or residual acoustic confounds. This further validates the diagnostic precision of our minimal-pair probing setup.

D More Layer-wise Analysis

Untrained models' performance as baseline. As shown in Figure 9, despite the absence of learned weights, many models achieve above-chance performance, particularly for syntax and interface tasks. This suggests that input representations—derived from log-mel spectrograms or early convolutional layers—already encode structure-correlated acoustic cues. The mild upward trend in some shallow models (e.g., S2T, Whisper) further highlights that even without pretraining, architectural priors combined with spectrotemporal patterns can give rise to latent linguistic signals. In contrast, semantic and conceptual tasks remain near chance throughout, indicating their reliance on abstract, model-learned representations.

Selection score As show in Figure 10, relative to the raw accuracies in Figure 5, selection scores sharpen the contrast between layers: (i) self-supervised models show a steady climb, peaking in the upper-middle layers before tapering, indicating that the diagnostic signal becomes both stronger and more concentrated with depth; (ii) ASR-fine-tuned encoders plateau earlier and exhibit a late-layer drop, consistent with a redistribution of capacity toward transcription; (iii) AudioLLM encoders (Whisper, Qwen) maintain rising curves to the final layer, suggesting continued accumulation of structurally relevant information; and (iv) the EnCodec baseline remains at chance, confirming that its representations lack discriminative linguistic content even after normalization.

Confidence score As shown in Figure 11, confidence increases consistently with depth across nearly all models, mirroring trends seen in accuracy and selection score. However, models vary in their *calibration*: self-supervised models and AudioLLMs exhibit sharp late-layer confidence gains, whereas ASR-fine-tuned encoders often peak earlier and display more modest increases, especially for conceptual and morphological tasks. EnCodec shows low, flat confidence across the board, consistent with its poor accuracy. Overall, confidence dynamics highlight not just where linguistic features emerge, but how decisively they are encoded in the internal representation.

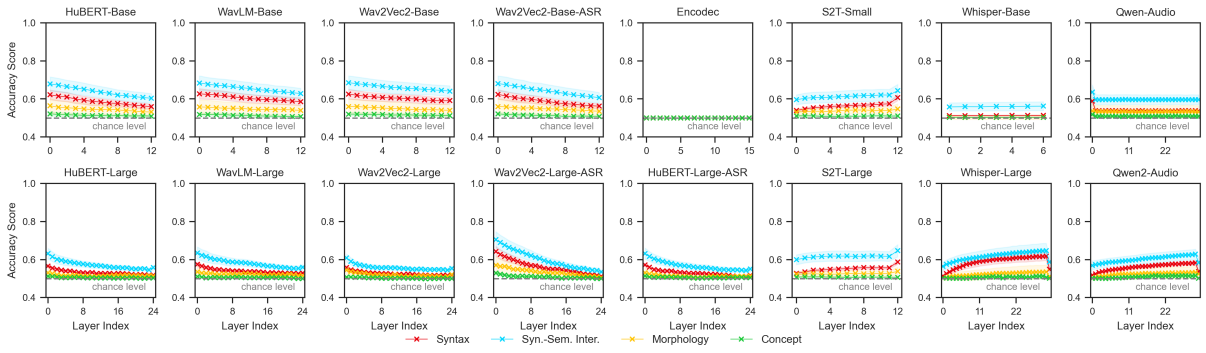

Figure 9: Layer-wise accuracy of randomly initialized (untrained) speech models across linguistic categories. We evaluate probing accuracy at each layer of speech models whose weights are randomly initialized (i.e., no pretraining).

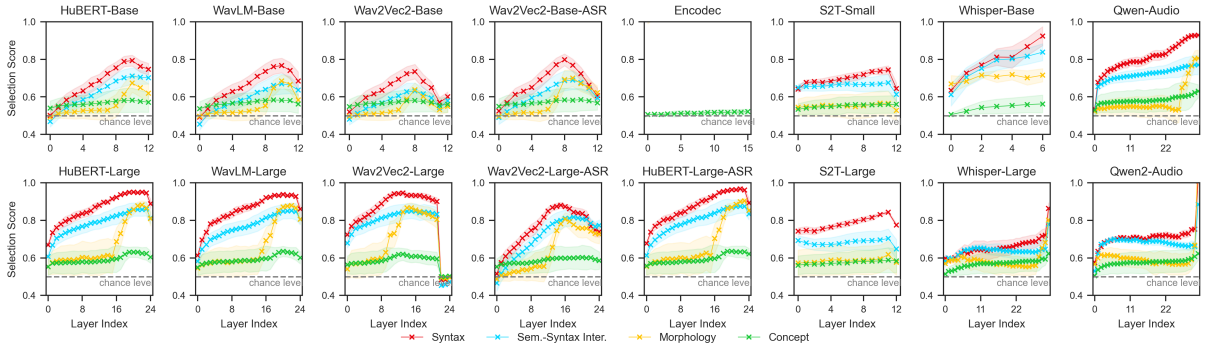

Figure 10: Layer-wise *selection scores* reveal where each linguistic property is *most informative*.

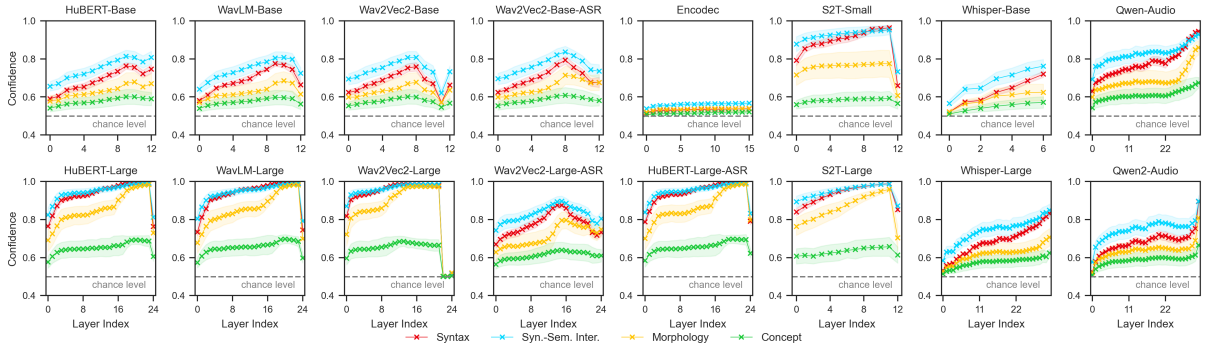

Figure 11: Layer-wise *confidence scores* reveal how strongly models commit to correct predictions.