Beyond Decodability: Reconstructing Language Model Representations with an Encoding Probe

Anonymous ACL submission

Abstract

Probing is widely used to study which features can be decoded from language model representations. However, the common decoding probe approach has two limitations that we aim to solve with our new encoding probe approach: contributions of different features to model representations cannot be directly compared, and feature correlations can affect probing results. We present an ENCODING PROBE that reverses this direction and reconstructs internal representations of models using interpretable features. We evaluate this method on text and speech transformer models, using feature sets spanning acoustics, phonetics, syntax, lexicon, and speaker identity. Our results suggest that speaker-related effects vary strongly across different training objectives and datasets, while syntactic and lexical features contribute independently to reconstruction. These results show that the ENCODING PROBE provides a complementary perspective on interpreting model representations beyond decodability.

1 Introduction

Most current approaches to processing speech and text are based on deep learning architectures and especially large transformer models. While these systems work very well in many applications, they are large and complicated enough that understanding and controlling their function in detail becomes challenging. A standard approach to analyzing deep learning models is based on *probing* the representations which emerge in these systems. More specifically, a researcher would come up with a set of hypotheses about what kinds of information a model can be plausibly expected to encode and then attempt to systematically extract that information from activation patterns in different components of the system.

1.1 Decoding probe and its limitations

This classic probing approach, which we refer to in the rest of the paper as DECODING PROBE, can tell us whether a specific information type can be reliably extracted or *decoded* from a model's representations. This approach offers valuable insights, and can be indispensable, for example when we want to know whether a model stores sensitive information such as speaker identity that has privacy implications. However, this approach has two important limitations that we aim to solve with our new ENCODING PROBE approach.

Relative feature contributions. Suppose we are interested in the encoding of speaker identity versus phonetics in the activation patterns of a model of spoken language. We train two DECODING PROBES, one mapping activation patterns to speaker IDs and the other one mapping those same activation patterns to phone labels. We may then recover speaker IDs with (e.g.) 95% accuracy and phone labels with 58% accuracy. What can we conclude about the relative importance of these two features? Nothing definitive: the two numbers are not directly comparable, as they depend on the number of labels and their distribution in the two classification tasks.

Feature correlations. Consider a further problematic scenario: we are interested in the encoding of grammatical categories in a language model. We train a probing classifier which maps activation patterns to the grammatical category label for the current token. The classifier performs at 85% accuracy and we conclude that grammatical category is encoded in model representations. However, this conclusion may be misleading if we do not *control for correlated features*. Word identity and grammatical category are highly correlated: many words occur almost exclusively with a single grammatical category. As a result, the probe may rely primarily

on information about word identity present in the representations, and predict grammatical categories as a by-product of their correlation with specific words. The probe's performance may reflect the model's encoding of lexical identity rather than grammatical category itself.

1.2 Reversing the direction of probing

Within NLP, the terms *probe* and *probing classifier* have often been used in a narrow sense, as a classifier which maps an intermediate representation X from a target model to some (linguistic) property of interest Y (e.g., Belinkov, 2022). However, two assumptions in this definition can (and should) be dropped. First, the mapping need not take the form of a classifier since the targeted property need not be categorical. More fundamentally, the mapping need not be in the direction from model representation X to property of interest Y, $f : X \rightarrow Y$, but could be in the reverse direction $g : Y \rightarrow X$. Within neuroscience this reverse direction of probing, known as brain encoding, is widely used and well understood (Ivanova et al., 2021). Here we call the type of probe formulated as $f : X \rightarrow Y$ a DECODING PROBE, and the one formulated as $g : Y \rightarrow X$ an ENCODING PROBE.

We argue that the two directions of probing answer complementary sets of questions about the nature of neural representations, and they have quite different affordances in terms of how they can be implemented and applied. Below, we discuss the advantages of the ENCODING PROBE for addressing the two main challenges highlighted before.

1.2.1 Quantifying feature contributions

Consider again comparing the contribution of two separate features to the representations of a spoken language model. We can operationalize this comparison in terms of explaining the amount of variance in these representations. Using an ENCODING PROBE, we can fit a regression model to reconstruct activation patterns X based on a number of interpretable features Y, which include both speaker IDs Y_{sid} and phone labels Y_{ph}, formulated as $g : Y \rightarrow X$. We can now ablate the full probe, removing each feature of interest in turn: first, $g_1 : Y \setminus Y_{sid} \rightarrow X$, and second, $g_2 : Y \setminus Y_{ph} \rightarrow X$. Let $E(g)$ be the measure of reconstruction error for probe g. Then the increase in error due to ablating each feature, $E(g_i) - E(g)$, gives us a direct measure of the relative contribution of each feature to the model representation, over and above the other features. We may well find that the contribution of phone labels is higher than that of speaker labels, even if the latter are easier to decode.

1.2.2 Controlling for correlated features

Reconsider the question of whether grammatical categories are encoded in a language model. By employing an ENCODING PROBE, we can control for correlated features: our full model $g : Y \rightarrow X$ contains multiple, potentially correlated, sources of information, including word identities Y_w and grammatical category labels Y_g. We can ablate each feature separately, as well as both together, and measure the resulting change in reconstruction error. This allows us to assess the conditional contribution of each feature, while accounting for their correlation. We may find that removing Y_g does not increase reconstruction error, whereas removing Y_w does. This would indicate that Y_g does not provide additional information beyond what is already captured by Y_w, even though grammatical labels may be easy to recover from the model's representations due to their correlation with word identity. Conversely, if ablating Y_g does increase reconstruction error, this indicates that grammatical category contributes information to the representations that is not fully explained by word identity.

No doubt many readers recognize that in the above vignettes we are describing standard tools and insights of regression analysis, applied to the specific case of analyzing model representations. This connection to well-understood statistical methods is indeed one of the main attractions of the ENCODING PROBE formulation.

In the remainder of the paper we apply the ENCODING PROBE to a set of speech and text transformer models with controlled feature ablations. We focus on syntactic and lexical features in both modalities, and on speaker identity, phonetic, and acoustic features in speech models. Overall, the results demonstrate the value of the ENCODING PROBE for measuring relative contributions while controlling for correlated features.

2 Literature

Probing has been a cornerstone in the interpretability literature since the early days of its development (Alain and Bengio, 2017; Conneau et al., 2018; Hupkes and Zuidema, 2018). In transformer interpretability literature, the probing paradigm has been used to discover a variety of features encoded

in the hidden state activations of transformer models. However, Belinkov (2022) identifies several important caveats in probing: probe scores are often hard to interpret and compare across/within tasks; results can reflect probe design choices rather than properties of the target model; probes establish correlation rather than causation; outcomes are sensitive to dataset/task artefacts; and analyses are constrained by pre-selected properties.

To improve interpretability of probe results, Hewitt and Liang (2019) argue that a control task is needed to ascertain that a successful probe is uncovering linguistic structure from model internals rather than learning the task itself. In follow-up work, Hewitt et al. (2021) introduce *conditional probing* in order to control for correlated features. Their approach relies on the decoding probe, but compares decoding a property (such as POS label) from target representation only, versus decoding it from the representation paired with a control variable (such as word identity).

Our ENCODING PROBE, in contrast, controls for such effects simultaneously for multiple properties, while also quantifying their relative contribution to explaining the overall variance of neural representations. Our method does not address other caveats associated with probing. While enabling control for correlated features, the ENCODING PROBE remains observational and not interventional. Hence it cannot establish causal structure among the investigated variables or their role in the computations carried out by the model.

Complementary representational analysis methods such as Canonical Correlation Analysis (CCA) (Hotelling, 1936) and Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) are not directional and sometimes do not even require learnable parameters. Research using these methods often yields results corroborating those of probing on the encoding of specific features in the hidden layers of language models (Chrupała and Alishahi, 2019; Pasad et al., 2021; de Heer Kloots et al., 2025), but does not address the two main challenges we focus on here (i.e. comparison of different features and controlling of correlation between features).

2.1 Encoding of speaker identity

Probing experiments have shown that various features can be decoded from the internal representations of models such as wav2vec2 (Baevski et al., 2020), including phonetic (Ma et al., 2021; De La Fuente and Jurafsky, 2024; Cormac English et al., 2022) and prosodic features (Bentum et al., 2025) as well as speaker identity (Chiu et al., 2026).

Human listeners process speaker identity separately from linguistically relevant information (Lee et al., 2019), suggesting that speech models may also disentangle representations of these two information types. Liu et al. (2023) and Gubian et al. (2025) show evidence that while speaker identity can be decoded from the wav2vec2 model's internal hidden states, the subspace it occupies is orthogonal to that of phonetic and tonal information. Our ENCODING PROBE can provide further insight by directly comparing the relative contribution of phonetic and speaker features.

2.2 Encoding of syntactic structure

In the text domain, simple syntactic properties (such as POS labels, tree depth, constituent labels) have been probed for in neural representations since the early work of Adi et al. (2017) and Conneau et al. (2018). More complex syntactic and semantic structures have been shown to be recoverable from the hidden state activations of models such as BERT, relying on a variety of custom methods (Hewitt and Manning, 2019; Chrupała and Alishahi, 2019; Tenney et al., 2019). Using a probe formulated to specifically reveal relationships between tokens, Hewitt and Manning (2019) find that the structural relationships between two token representations are highly correlated with the structure of a syntax tree. Further experiments using the same structural probe with curated datasets show that while it is clear that there is some *syntactic* knowledge encoded by the BERT model, it is difficult to disentangle syntax from other confounding signals such as semantics or the closeness of two words in a sentence (Maudslay and Cotterell, 2021; Simon et al., 2025).

In the speech domain, pre-trained speech models such as wav2vec2 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) have been shown to encode syntactic information and word identity (Pasad et al., 2021; Shen et al., 2023; Pasad et al., 2024; Choi et al., 2024). However, Shen et al. (2023) note that the syntactic and lexical features are highly correlated, and the syntactic decodability shown could possibly be attributed to the underlying lexical representations in the model embeddings. They come to this conclusion by comparing their syntax probe when applied to the target model representation versus when applied to a bag-of-words reference rep-

Model	Training Objective
wav2vec2 (base)	Audio self-supervision
wav2vec2 (ASR)	Speech Recognition
wav2vec2 (SID)	Speaker identification
BERT (base)	Text self-supervision

Table 1: Target models tested in this paper.

resentation. The ENCODING PROBE offers a more general way of accounting for such confounds.

3 Methodology

While the ENCODING PROBE can be applied to any target model, in this paper we focus on transformer language models of text and speech as our main case studies. We compile a variety of interpretable feature sets that can be obtained from the speech audio waveform or text input. We then use subsets of these features to train multivariate ridge regression probes to reconstruct the hidden state activations of the target models.

3.1 Models

We focus on a number of classic Transformer models that have been extensively used in previous probing studies. We use the model versions listed in Table 1, which cover different modalities (text, speech) and training objectives (pre-trained only, ASR tuning, speaker identification tuning). Below we describe the basic architectures and variants.

Wav2vec2 (Baevski et al., 2020) is a popular speech model with 12 transformer layers and 768 hidden dimensions, pre-trained on 960 hours of the LibriSpeech dataset (Panayotov et al., 2015) in a self-supervised fashion. We also test the variant fine-tuned for automatic speech recognition. In addition to publicly available models, we also fine-tune a wav2vec2-base model for speaker identification using the 100-hour train split of LibriSpeech with the `transformers` library (Wolf et al., 2020).

BERT (Devlin et al., 2019) is a classic text transformer model. We use `bert-base-uncased`, pre-trained on 3.3B words from books and web data.

3.2 Feature sets

The following interpretable feature sets are used in our experiments, each representing aspects of human language in the audio and textual modalities.

Acoustics. Features directly extractable from the waveform signal (Eyben et al., 2010), including pitch, formants, and MFCCs[1].

Phonetics. Phonetic posteriorgram (PPG) vectors[2] (Hazen et al., 2009), representing the probability of each phone label at a given point in the waveform. We use the `ppgs` Python package (Churchwell et al., 2024) to automatically generate the PPGs from the waveform.

Speaker. Speaker identity is an utterance-level label. Within the speech signal, it is perceived as the quality of the voice or timbre. We one-hot encode the speaker label from the LibriSpeech dataset (Panayotov et al., 2015) into a feature vector of speaker identity.

Syntax. Syntactic features are the grammatical and structural properties of individual segments of the input. We use spaCy (Honnibal et al., 2020) to automatically extract both categorical features (part-of-speech, dependency label) and numerical features (total syntax tree depth, depth of token within syntax tree, sequential position of the word, total number of words in the utterance). The categorical features are encoded into one-hot vectors.

Lexicon. Lexical information represents the form and meaning of the words in the utterance. We use FastText static word embeddings (Bojanowski et al., 2017) as a proxy for it. These embeddings are non-contextual and thus do not contain sentence-level structural information.

Full. The combination of all features listed above. The ablations involve removing features from the union of all the feature sets, e.g. *Full* \ *Syntax*.[3]

3.3 Token and frame embeddings

In classic text probing studies, token or word embeddings are the base units of information encoding for target models (Hewitt and Manning, 2019). When probing speech models, we work directly at the level of frames, which are the basic units extracted from the audio and which the transformer

[1]The complete list of low-level descriptors from the eGeMAPSv02 feature set (Eyben et al., 2016) can be found in Section A.1

[2]Different representations of the same feature may disproportionately affect the ENCODING PROBE more than the DECODING PROBE. See Section A.2 for an analysis of feature representation choice using phonetic features as an example.

[3]Results of correlation checks between feature sets can be found in Sections A.3 and A.4.

layers of speech models use as tokens. We randomly sample frames from a given utterance and look up the corresponding features for that given timestamp. The features themselves are computed on frames, words or whole utterances, as appropriate, but the model representations always correspond to single frames. This obviates the need to pool representations, and minimizes the imposition of text-based concepts on the analysis of models of spoken language.

3.4 The ENCODING PROBE

We fit a Ridge regressor with `scikit-learn` (Pedregosa et al., 2011) that takes the union of all feature sets *Full* to predict the model representation of the corresponding frame. Essentially, the probe learns to explain the variance within the model representations, and is evaluated on the test split of the experimental data. As our reconstruction error metric, we report the **unexplained variance** (UV) of the probe, defined as the ratio of residual sum of squares to the total sum of squares of the dependent variable:

$$\text{UV} = \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - R^2. \quad (1)$$

After a feature is ablated, a larger upward gap from the *Full* baseline indicates a greater relative contribution for that feature, irrespective of absolute UV values. The regularization strength α was tuned via cross-validation over the grid $\{10^n \mid n \in \mathbb{Z}, -3 \leq n \leq 5\}$. We refit the ridge regressor with the same hyperparameter search for every ablation.

We use the 100-hour train split of the LibriSpeech dataset (Panayotov et al., 2015) and employ a randomized 80:20 train-test split stratified by speaker ID. Stratification is applied only to speaker ID since it is the only utterance-level label we include in the probe's input. We select a maximum of 10 frames from each utterance in text models and a maximum of 15 from speech models. Silent frames are filtered out. The preprocessing pipeline results in ∼230k input-output pairs for speech models and ∼180k for text models. To quantify the impact of sampling frames, we ran the experiments for `wav2vec2-base` and `bert-base-uncased` with 10 additional random seeds $\{n \times 100 \mid n \in \mathbb{Z}, 1 \leq n \leq 10\}$ and found that the 95% confidence interval (CI) width of the mean differences of UV scores across all random seeds was < 0.002.

A built-in comparison point. We compare each probe to the analogous one with the complete set of features, *Full*, which by design yields the lowest Unexplained Variance (UV).

Parallel decoding probe analyses. We also implement parallel analyses with the DECODING PROBE alongside the ENCODING PROBE results. Our decoding analyses use `scikit-learn` Ridge regressors and classifiers that decode features of interest using the frame-level representations of the transformer models. We utilize the same features across both probe types, except that class labels are used instead of one-hot encoded vectors for the classification tasks. We conduct a hyperparameter search using the same parameter grid as above.

3.5 Design of case studies

We present two case studies to demonstrate the application of the ENCODING PROBE: the encoding of speaker identity in models of speech, and the encoding of syntactic features in models of text and speech. Both of these topics have been extensively investigated in prior research as described in Section 2, where the DECODING PROBE has been used to decode relevant features from various transformer models' inner representations. As discussed in Section 1.1, in both cases there is another, potentially highly correlated source of information (acoustics in the case of speaker identity and lexical features in the case of syntactic features) which might affect the previously reported findings. We design two sets of experiments for each case study, using a set of targeted transformer models to demonstrate how the ENCODING PROBE offers a complementary perspective that mitigates limitations of the DECODING PROBE.

4 Experiment 1: Speaker Identity

Here we investigate the relative contribution of speaker identity, phonetic, and acoustic features. We also look at the impact of the tuning objectives of the models on the encoding of these features in their internal representations.

4.1 Decoding speaker identity and phone label

Before applying the ENCODING PROBE for this purpose, we first use the DECODING PROBE to establish a point of comparison for speaker identity and phone labels. We fit linear models to decode speaker identity and phone label from the hidden-state representations of the speech models. As

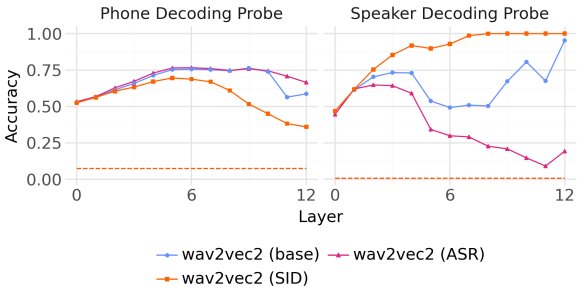


Figure 1: Accuracy of the DECODING PROBE on predicting speaker identity and phone label from frames of model internal hidden states. The dashed line represents the majority baseline of the classification task.

shown in the right panel of Figure 1, we find the performance of the speaker identity probe increases towards later layers in the base and SID models but decreases in the ASR model (with a dip in performance in the middle layers for base). The left panel of Figure 1 shows the phone label probe performance increases from the early to middle layers but drops in final layers for the base and SID models and remains high in the ASR model. However, the performance of all probes consistently exceeds the majority baseline for all layers in all models.

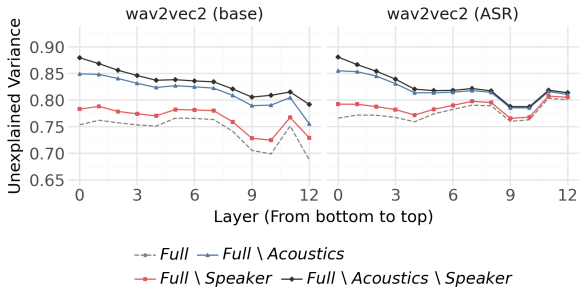
These results show that the decodability of speaker identity and phone labels varies depending on the training/tuning objectives of the model. The tuning objective of the ASR model pushes the model to normalize across different speakers while maintaining its ability to identify the correct phone label for the CTC decoding task, whereas the SID model's objective encourages it to maximize its ability to differentiate speakers, resulting in a larger drop in the decodability of phone labels.

These patterns in the current findings confirm previous research. However, the results only show that there is enough information in the model's internal hidden states for the probe to decode these features. Due to differences in the baselines and the levels of difficulty of the tasks, the results of these two probes are not directly comparable against each other, and the relative contribution of each feature to model representations cannot be quantified.

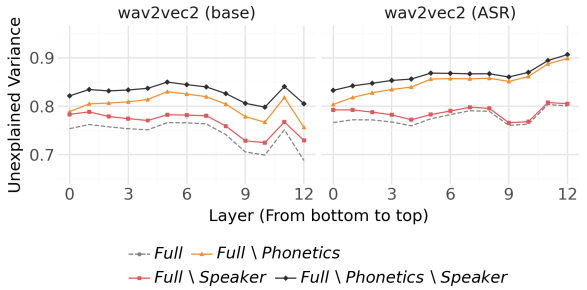
4.2 Encoding to reconstruct representations

Now we apply the ENCODING PROBE to reconstruct representations of the same speech models (base, ASR-tuned, and SID-tuned variants of wav2vec2)[4]. In addition to the *Full* probe, we also

[4]The experimental results for an extended set of models can be found in Section A.5.



(a) Acoustics and speaker identity



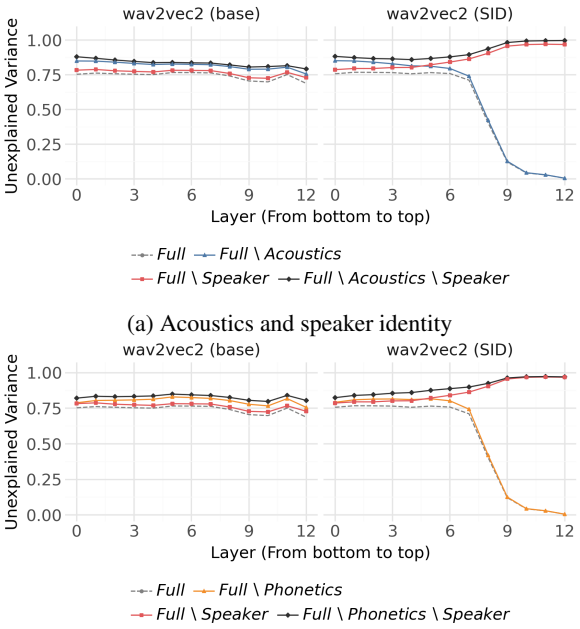
(b) Phonetics and speaker identity

Figure 2: Unexplained Variance (UV) for the base and ASR models (larger gap from *Full*/dashed = greater contribution).

construct ablated feature sets with each of the targeted feature sets (*Acoustics, Phonetics, Speaker*) and joint ablations of two sets.

Figures 2 and 3 show the ENCODING PROBE results for the speaker identity experiments (note that the two sets of graphs use different scales on their y-axes). A larger gap between the grey dashed line (i.e. the *Full* probe) and a colored solid line (i.e. an ablated probe) means that the ablated feature has a larger relative contribution to the reconstruction of the representation.

We first focus on Figure 2, which shows the impact of speaker identity, phonetic, and acoustic features in the base and ASR-tuned models. The top row shows that the UV of the *Full* \ *Acoustics* probe is substantially higher than that of the *Full* \ *Speaker* probe, for both models and all layers (blue line is always above red line). The bottom row shows that *Full* \ *Phonetics* also scores higher than *Full**Speaker*. These combined results indicate that both acoustic and phonetic features have a larger relative feature contribution than speaker identity in the base and ASR-tuned models. Comparing the results between the left and right columns, we see a smaller gap between *Full* \ *Speaker* and *Full*. This indicates that the relative contribution of speaker identity is smaller in the ASR-tuned model than in the base model. Additionally, in the upper lay-



(a) Acoustics and speaker identity

(b) Phonetics and speaker identity

Figure 3: Unexplained Variance (UV) for the base and SID models (larger gap from *Full*/dashed = greater contribution).

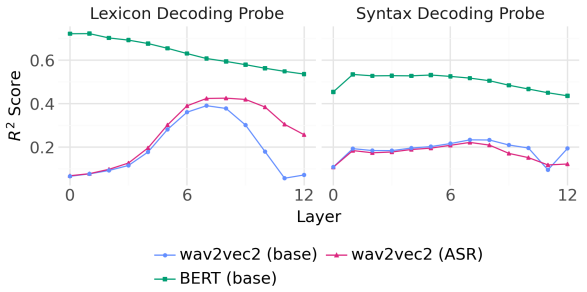


Figure 4: R^2 of the DECODING PROBE on predicting syntactic and lexical feature vectors from the frames of target model representations.

ers, we can see a larger effect for *Full* \ *Phonetics* for the ASR-tuned model compared to the base model, whereas *Full* \ *Acoustics* shows the opposite relationship, indicating that phonetic features are encoded more prominently in the internal hidden states of the ASR-tuned model than in those of the base model.

For the comparison between base and SID-tuned models shown in Figure 3, the situation is similar in the bottom layers, but changes drastically in the upper layers. First, for layers 9-12 of the SID-tuned model, *Full* achieves a very low UV, i.e. it accounts for most of the variance in the target model representations. Furthermore, from around layer 7 of the SID model onward, both *Full* \ *Acoustics* and *Full* \ *Phonetics* barely show any increase above *Full*. In the same layers the UV of the *Full**Speaker* probe, on the other hand, is close to one. This profile of results indicates that the top layers of the SID-tuned model are almost exclusively dedicated to encoding speaker identity to the exclusion of any other information.

4.3 Decoding versus encoding results

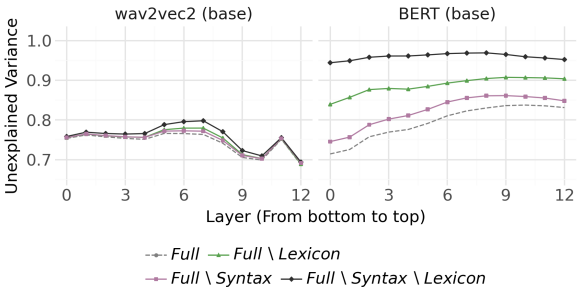
To highlight the different insights obtainable via the DECODING PROBE and ENCODING PROBE, we focus on the especially striking results for the tuned models. We can decode phonetics from the representations in the top layers of the SID model with non-trivial accuracy, while at the same time phonetic features barely contribute to the reconstruction of these representations. Similarly, we can decode speaker labels from the top layers of the ASR-tuned model with above-baseline accuracy, but the contribution of speaker identity to the reconstruction of these same layers is very minor. In other words, DECODING PROBE detects phonetic structure sufficient for classification, but the ENCODING PROBE reveals that this structure accounts for a negligible fraction of the total representational variance in these layers. The fact that decodability and relative contribution can diverge substantially underscores the value of the ENCODING PROBE.

5 Experiment 2: Syntactic Information

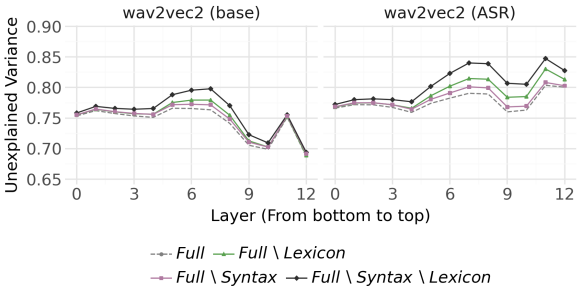
In this section we quantify the contributions of syntactic versus lexical information to the reconstruction of target models' representations, while controlling for correlations between these features, which have been highlighted in previous work. We examine both text and speech transformer models.

5.1 Decoding syntactic and lexical features

We first apply a DECODING PROBE to establish the decodability of syntactic and lexical features separately. Figure 4 shows that syntactic and lexical features are much more decodable in BERT (base) than in the wav2vec2 models. However, while the highest R^2 score of the probes for both the base and ASR-tuned models appears around layer 7, the pattern from layers 9-12 is slightly different. The lexical probe for the ASR-tuned model has a smaller decrease in R^2 score compared to the probe for the base model, and the syntax probes show smaller differences between the two models.



(a) Speech versus text



(b) Base versus ASR-tuned

Figure 5: Unexplained Variance (UV) $(1 - R^2)$ of the ENCODING PROBE for the text base, speech base, and ASR models (larger gap from *Full*/dashed = greater contribution).

5.2 Reconstructing representations

Figure 5 shows the results of applying the ENCODING PROBE to reconstruct the representations of one text model, BERT (base), and two speech models (base and ASR variants of wav2vec2)[5]. We re-use the *Full* probe and report additional ablated feature sets with each of the targeted features (*Syntax, Lexicon*) as well as joint ablation of two sets. We discuss the main observations below.

First, the increase in UV due to ablating lexical features is higher than the increase due to ablating syntactic features: the line corresponding to *Full* \ *Lexicon* is consistently above *Full* \ *Syntax*, for all three target models.

Second, the ablated probes always have higher UV scores than the *Full* probe, which indicates that lexical and syntactic features each make an independent contribution in the probe's reconstruction of model activation patterns. The hypothesis that syntactic decodability is entirely attributable to lexical correlations is not supported: the ablation of syntactic features produces an increase in UV even when lexical features are present in the probe input.

Regarding the impact of the target model modality, the effect of the ablations is consistently larger in BERT (base) than in the speech models, meaning that lexical and syntactic features make a more prominent contribution to the reconstruction of the representations of BERT (base) compared to those of the speech models.

Finally, fine-tuning on ASR leads to a target model whose representations are somewhat harder to reconstruct in absolute terms, but where lexical and syntactic features contribute more to the reconstruction compared with what they contribute to the base model.

5.3 Decoding versus encoding results

The DECODING PROBE shows that lexical and syntactic features are decodable with moderate success, but only the ENCODING PROBE allows quantitative comparison of their relative contributions while controlling for correlations. This case illustrates that the DECODING PROBE shows whether information is present, and the ENCODING PROBE reveals how much of the representational variance it explains.

6 Discussion and Conclusion

We present the ENCODING PROBE, which reconstructs representations in the layers of a target model based on multiple interpretable features. Our probe enables direct comparison of the relative contributions of different features through a feature ablation analysis under a shared regression objective, while controlling for correlated features.

We design two case studies to showcase the ENCODING PROBE. In the first, speaker identity is decodable across all wav2vec2 variants, confirming earlier findings using decoding probes (Gubian et al. (2025)), yet its contribution to reconstruction is modest in base and ASR models and substantial only with speaker-ID tuning. In the second, lexical and syntactic features make independent contributions to reconstruction in both text and speech models, with lexical features accounting for a higher fraction of the variance, confirming that these models encode syntactic information beyond what lexical features alone can explain.

Such observations are only possible because the multiple regression analysis framework underlying the ENCODING PROBE enables us to control for correlations between features by design. This framework likewise leads to a direct quantification of relative feature contributions.

[5]The experimental results for an extended set of models can be found in Figure 9 in the Appendix.

Limitations

The ENCODING PROBE addresses some of the issues with probing highlighted by Belinkov (2022), but new limitations arise and some are not addressed. The probe remains observational rather than interventional: it reveals statistical associations between interpretable features and model representations, but it does not establish causal mechanisms. While the ENCODING PROBE framework in itself supports the use of any regression model, our implementation here relies on ridge regression, which may fail to capture nonlinear relationships in the transformer representations. While nonlinear probes might achieve lower absolute reconstruction error, linear probes offer better interpretability and are standard in the probing literature. The encoding results are also conditional on the feature sets we chose to operationalize the input space (i.e., acoustics, phonetics, syntax, lexicon, and speaker identity). Factors not included in our analysis can still contribute to the probe's reconstruction performance. While the frame-level setup allows for a faithful analysis of the representations of speech models, it potentially under-represents long-range structural information. Our speech experiments use read audiobook data; results may differ for spontaneous or noisy speech. Finally, observed feature contributions and interaction patterns may vary with dataset composition and alignment quality. For these reasons, the reported ablation effects should be read as a case study, rather than as an exhaustive characterization of internal representations of the target models tested.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *International Conference on Learning Representations*.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. International Conference on Learning Representations.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Martijn Bentum, Louis ten Bosch, and Tomas O. Lentz. 2025. Word stress in self-supervised speech models: A cross-linguistic comparison. In *Proc. Interspeech 2025*, pages 251–255.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Aemon Yat Fei Chiu, Kei Ching Fung, Roger Tsz Yeung Li, Jingyu Li, and Tan Lee. 2026. A Large-Scale Probing Analysis of Speaker-Specific Attributes in Self-Supervised Speech Representations. *Preprint*, arXiv:2501.05310.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-Supervised Speech Representations are More Phonetic than Semantic. In *Proc. Interspeech 2024*, pages 4578–4582.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating Neural and Symbolic Representations of Language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. Association for Computational Linguistics.

Cameron Churchwell, Max Morrison, and Bryan Pardo. 2024. High-Fidelity Neural Phonetic Posteriorgrams. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 823–827, Seoul, Korea, Republic of. IEEE.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen. 2022. Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 83–91, Seattle, Washington. Association for Computational Linguistics.

Marianne de Heer Kloots, Hosein Mohebbi, Charlotte Pouw, Gaofei Shen, Willem Zuidema, and Martijn Bentum. 2025. What do self-supervised speech models know about Dutch? Analyzing advantages of language-specific pre-training. In *Proc. Interspeech 2025*, pages 256–260.

Anton De La Fuente and Dan Jurafsky. 2024. A layer-wise analysis of Mandarin and English suprasegmentals in SSL speech models. In *Interspeech 2024*, pages 1290–1294. ISCA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462, Firenze Italy. ACM.

Michele Gubian, Ioana Krehan, Oli Liu, James Kirby, and Sharon Goldwater. 2025. Analyzing the relationships between pretraining language, phonetic, tonal, and speaker information in self-supervised speech models. *Preprint*, arXiv:2506.10855.

Timothy J. Hazen, Wade Shen, and Christopher White. 2009. Query-by-example spoken term detection using phonetic posteriorgram templates. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 421–426.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: Measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Harold Hotelling. 1936. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377.

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.

Dieuwke Hupkes and Willem Zuidema. 2018. Visualisation and 'Diagnostic Classifiers' Reveal how Recurrent and Recursive Neural Networks Process Hierarchical Structure (Extended Abstract). pages 5617–5621.

Anna A Ivanova, John Hewitt, and Noga Zaslavsky. 2021. Probing artificial neural networks: Insights from neuroscience. In *ICLR 2021 Workshop "How Can Findings about the Brain Improve AI Systems?"*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Yoonjeong Lee, Patricia Keating, and Jody Kreiman. 2019. Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, 146(3):1568–1579.

Oli Danyi Liu, Hao Tang, and Sharon Goldwater. 2023. Self-supervised Predictive Coding Models Encode Speaker and Phonetic Information in Orthogonal Subspaces. In *Proc. Interspeech 2023*, pages 2968–2972.

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.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131, Online. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

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, Cartagena, Colombia. IEEE.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Gaofei Shen, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2023. Wave to Syntax: Probing spoken language models for syntax. In *INTERSPEECH 2023*, pages 1259–1263.

Pablo J. Diego Simon, Emmanuel Chemla, Jean-Remi King, and Yair Lakretz. 2025. Probing syntax in large language models: Successes and remaining challenges. In *Second Conference on Language Modeling*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

A.1 Acoustic descriptor inventory

This section outlines the feature names for the complete eGeMAPSv02(Eyben et al., 2016) feature set extracted by openSMILE (Eyben et al., 2010). We use the low-level descriptors that extract each descriptor at 20 ms intervals to best align with the frame sizes of the speech models (25 ms for wav2vec2).

#	Acoustic descriptor
1	Loudness_sma3
2	alphaRatio_sma3
3	hammarbergIndex_sma3
4	slope0-500_sma3
5	slope500-1500_sma3
6	spectralFlux_sma3
7	mfcc1_sma3
8	mfcc2_sma3
9	mfcc3_sma3
10	mfcc4_sma3
11	F0semitoneFrom27.5Hz_sma3nz
12	jitterLocal_sma3nz
13	shimmerLocaldB_sma3nz
14	HNRdBACF_sma3nz
15	logRelF0-H1-H2_sma3nz
16	logRelF0-H1-A3_sma3nz
17	F1frequency_sma3nz
18	F1bandwidth_sma3nz
19	F1amplitudeLogRelF0_sma3nz
20	F2frequency_sma3nz
21	F2bandwidth_sma3nz
22	F2amplitudeLogRelF0_sma3nz
23	F3frequency_sma3nz
24	F3bandwidth_sma3nz
25	F3amplitudeLogRelF0_sma3nz

Table 2: Full openSMILE eGeMAPSv02 low-level descriptor inventory used for the acoustic feature set in the probe input ensemble.

A.2 Different feature representations

To show the importance of feature representation type, we test three different ways of representing phonetic information in both the ENCODING PROBE and DECODING PROBE settings. We transform the PPG probabilities into one-hot encoding as well as phone IDs. The results for the probes are reported in Figure 6. The differences in the ENCODING PROBE R^2 score for Phone ID and the other multi-dimensional representations of the phone are substantial. The differences between the Phone ID and original PPGs are likely due to the additional representational power of a higher dimensional vector compared to a 1-D vector. The one-hot encoded PPGs also suffer from loss of information from the removal of the probabilities of other phones appearing in the same timestamp. On the other hand, the DECODING PROBE results show much smaller differences in the overall trend of the decodability of phonetic features from model internal hidden states. Choosing the right feature representation of the features has a larger impact on the success of the ENCODING PROBE than on the DECODING PROBE.

A.3 Direct correlations between speaker ID, phonetic, and acoustic features

We directly estimate the linear correlations between speaker identity and phonetic or acoustic features.

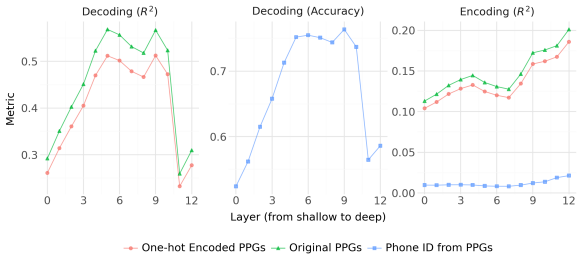


Figure 6: Comparison between ENCODING PROBE and DECODING PROBE results using different representations of the same phonetic feature.

We fit a ridge classifier using phonetic, acoustic and the combined features to predict the speaker identity. The results in Table 3 show that speaker identity is only weakly decodable from the explicit acoustic and phonetic features with a linear classifier. While it is not surprising that there is no correlation between phonetics and speaker identity, we also could not decode speaker label directly from acoustic features. Our acoustic features consist of highly localized acoustic descriptors based on a 20 ms waveform segment. The results show that there are almost no speaker characteristics that can be extracted from the local acoustic features.

	Accuracy
Acoustics	0.061
Phonetics	0.008
Phonetics + Acoustics	0.071

Table 3: Accuracy of a linear classifier using acoustics, phonetics, and combined acoustics + phonetics to predict speaker identity. The majority baseline is around 0.004.

A.4 Direct correlations between syntactic and lexical features

We also directly estimate the linear correlations between syntactic features and word embeddings. We fit ridge linear regressors/classifiers to predict individual syntactic features from the static fastText word embeddings. Results reported in Table 4 indicate that syntactic category labels (e.g. part-of-speech) are more decodable from the word embeddings than the other syntactic features. Overall, these results indicate that static word embeddings are not directly correlated with most of the syntactic features we include in this study.

	Metric	Score
POS	Accuracy	0.628
Dependency label	Accuracy	0.417
Tree depth	R^2	0.030
Word position	R^2	0.012
Total tree depth	R^2	0.012
Total word count	R^2	0.009

Table 4: Linear probe results on decoding syntactic features from fastText word embeddings.

A.5 All model results

We also run the ENCODING PROBE analysis across an extended set of speech and text Transformer-based models. We cover different architectures (HuBERT, WavLM, RoBERTa, ModernBERT), sizes (12-layer, 24-layer), and training objectives (speaker identification model from the SUPERB task). The complete list of the extended model set is presented in Table 5. The speaker identification model from the main text is omitted to improve readability of the results. Since the *Full* probe of the SID model reaches a near-zero UV score in the final layers and dominates the scale, we omit the SID model's results here.

All extended-model runs use the same probing setup, feature sets, sampling procedure, and train/test split described in Section 3, so differences across models are attributable to representational differences rather than pipeline changes. We extract hidden states layer-by-layer (including the embedding/output layer, indexed as layer 0 in our plots), and apply identical ablation-based encoding analyses as in the main experiments. Layer indices are normalized to $[0, 1]$ for comparability across different model sizes. WavLM, which uses a joint-speech enhancement objective, follows a broadly similar trend to wav2vec2 and HuBERT base models.

In all figures below we report Unexplained Variance (UV = $1 - R^2$). The grey dashed line shows the *Full* probe without ablations; a larger upward gap from this baseline means the ablated feature contributes more to reconstructing the layer's representations. Only the gaps between curves—not the absolute UV values—are the quantities of interest.

A.5.1 Speaker Identity

We run the speaker identity experiment on the extended set of speech models following the exact configuration in Section 4. The results are shown

Model ID	Training
wav2vec2-base	speech / PT
wav2vec2-base-960h	speech / ASR
wav2vec2-base-superb-sid	speech / SID
wav2vec2-large	speech / PT
wav2vec2-large-960h	speech / ASR
hubert-base-ls960	speech / PT
hubert-large-ll60k	speech / PT
hubert-large-ls960-ft	speech / ASR
wavlm-base	speech / PT
bert-base-uncased	text / PT
roberta-base	text / PT
ModernBERT-base	text / PT

Table 5: Extended model inventory used for additional probing runs. PT = Pre-training, ASR = Automatic Speech Recognition, SID = Speaker Identification.

in Figures 7 and 8. The results from the base and large variants of wav2vec2 follow a very similar trend, except in the final two layers where the large 24-layer model has a large drop in Unexplained Variance. At the same time, the HuBERT base and large variants mostly differ in the early layers where the gap between the *Full* probe and *Full* \ *Acoustics* \ *Speaker* is much larger for the 24-layer model than the 12-layer model. Our observation on the ASR tuning objective in Section 4 holds across different model architectures and sizes. Curiously, the SID-SUPERB model does not seem very different from the base model it was based on. This could be due to two factors: SID-superb is fine-tuned on the VoxCeleb dataset, which is out-of-distribution for the current task, and the fine-tuning process may differ between SID-superb and the SID model from the main text.

A.5.2 Syntactic Information

Turning to syntactic information, we again see a similar trend in different model sizes: there are differences at the last three layers for wav2vec2 base and large models and first six layers for HuBERT base and large models. We also confirm that ASR-tuning increases the contribution from both lexical and syntactic features in the large variants of both wav2vec2 and HuBERT. At the same time, the probe results from the text models are quite stable and similar across all three architectures.

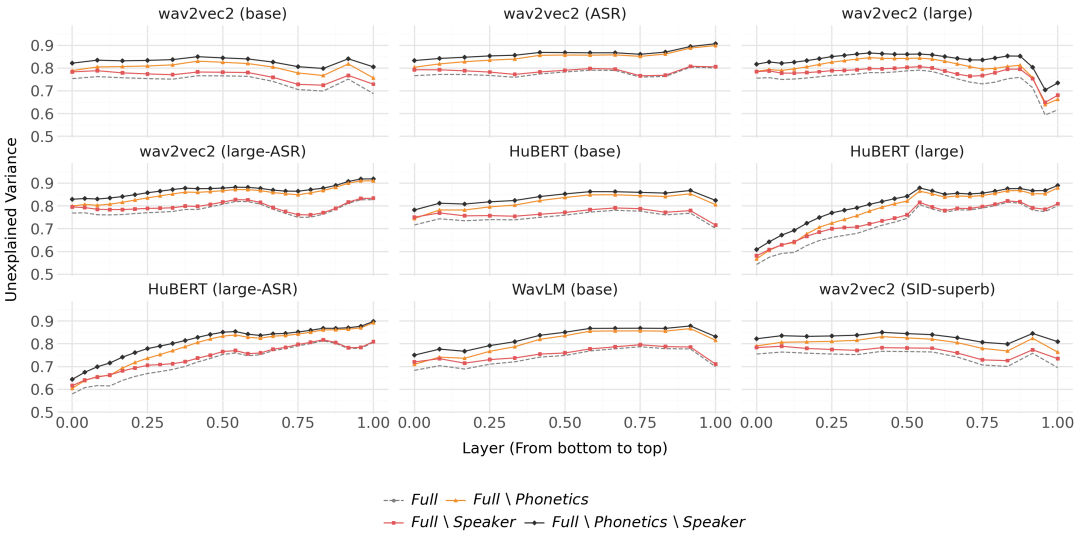


Figure 7: Encoding probe UV $(1 - R^2)$ score results of phonetics and speaker identity for all models (larger gap from *Full*/dashed = greater contribution).

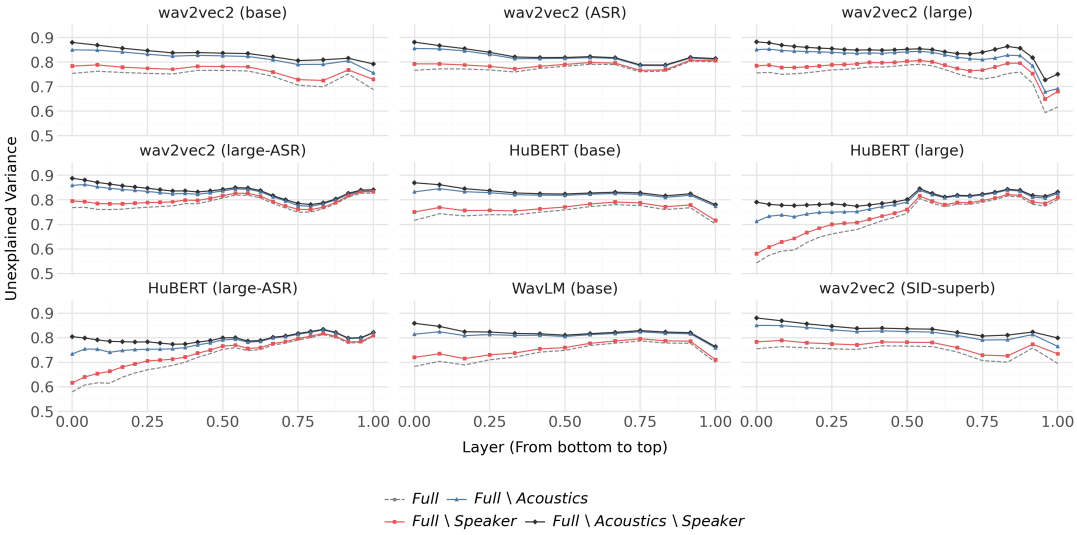


Figure 8: Encoding probe UV $(1 - R^2)$ score results of acoustics and speaker identity for all models (larger gap from *Full*/dashed = greater contribution).

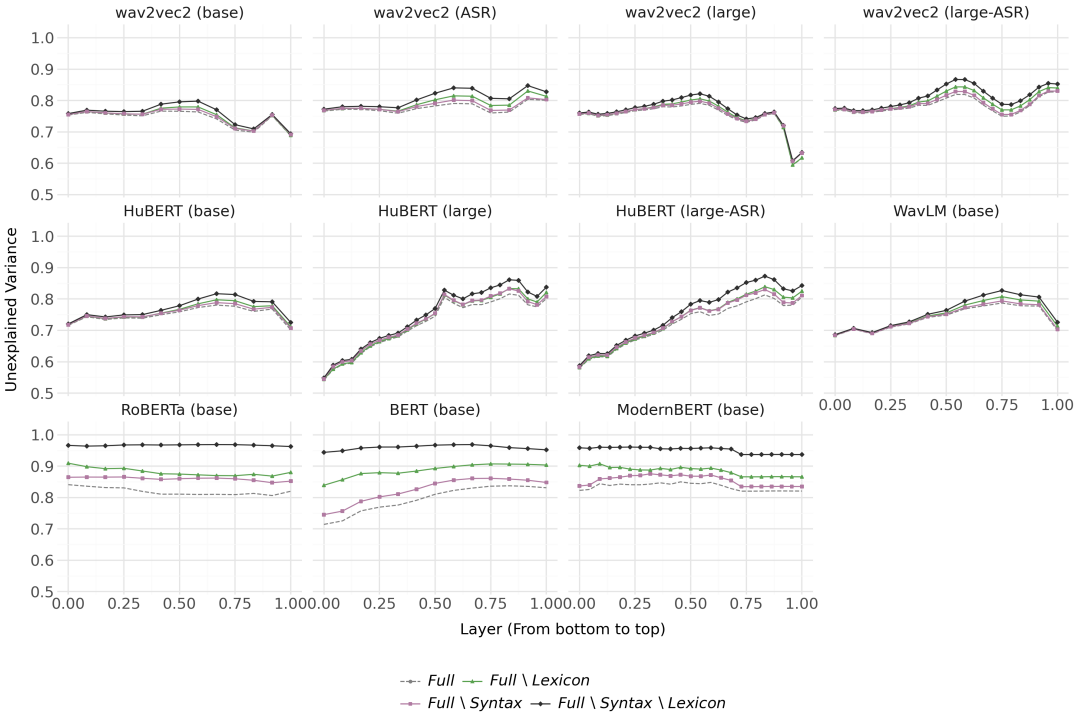


Figure 9: Encoding probe UV $(1 - R^2)$ score results of syntax and lexicon for all models (larger gap from *Full*/dashed = greater contribution).