INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Wave to Syntax: Probing spoken language models for syntax

Gaofei Shen[1], *Afra Alishahi*[1], *Arianna Bisazza*[2], *Grzegorz Chrupała*[1]

[1]Tilburg University, the Netherlands
[2]University of Groningen, the Netherlands

g.shen@tilburguniversity.edu, a.alishahi@uvt.nl, a.bisazza@rug.nl, grzegorz@chrupala.me

Abstract

Understanding which information is encoded in deep models of spoken and written language has been the focus of much research in recent years, as it is crucial for debugging and improving these architectures. Most previous work has focused on probing for speaker characteristics, acoustic and phonological information in models of spoken language, and for syntactic information in models of written language. Here we focus on the encoding of syntax in several self-supervised and visually grounded models of spoken language. We employ two complementary probing methods, combined with baselines and reference representations to quantify the degree to which syntactic structure is encoded in the activations of the target models. We show that syntax is captured most prominently in the middle layers of the networks, and more explicitly within models with more parameters.[1]

Index Terms: speech recognition, syntax, computational linguistics

1. Introduction

State-of-the-art models of (spoken) language rely on deep learning architectures composed of various components and based especially on the Transformer model (e.g. [1, 2]). Evaluating the performance of these models via standard quantitative protocols is straightforward enough; but it is not always easy to understand the reasons for fine-grained patterns of behavior and failure modes, and not trivial to debug and iterate on design. One tool to aid in this process has been the analysis and interpretation of the representations learned by the models, as encoded in the activation patterns within the various components [3, 4].

For text-based models, numerous works have probed these activations for many types of information, with a special interest in syntactic structure [5, 6, 7]. By contrast, the focus on speech models has been on the encoding of acoustic information, speaker characteristics, phonetics and phonology (e.g. [8, 9, 10]).

However, as the current crop of speech models grows in size and sophistication, we ask whether they also learn to encode syntactic structure to any appreciable degree. If the knowledge of syntax is useful for optimizing a model's training objective, the expectation would be that, given enough data, the model will learn to represent it. As a simple example, consider the utterance *The authors of the book are French*: if the timesteps in the feature space corresponding to *are* are masked and the model needs to reconstruct them, then it can do it better if it encodes number agreement between subject and verb. In the current study, we evaluate the hypothesis that syntax information is in general represented in such models.

We use two established representation probing techniques [5, 7] in combination with carefully designed baselines and reference representations to quantify the encoding of syntactic structure in selected current models of spoken language trained via self-supervision objectives. We also apply our methodology to models trained via visual supervision (also known as visual grounding) and to a model trained via a combination of self- and visual supervision. We track the encoding of syntactic structure throughout the transformer layers of these target models.

Our findings show that syntax is captured by all these models, with the following caveats and details: Firstly, the encoding of syntax is generally weaker than in text-based models (such as BERT [11]) . Secondly, much of the syntactic structure that is captured may be encoded in lexical rather than purely syntactic form. Thirdly, self-supervised and combined objectives lead to less syntactic encoding in the final model layers, while the visually-supervised objective does not have this effect. Finally, increased model size is associated with the stronger encoding of syntactic information.

2. Related work

Within Natural Language Processing (NLP), there has been substantial interest in understanding the represen-

[1] Code: https://github.com/techsword/wave-to-syntax

10.21437/Interspeech.2023-679

tations emerging within text-based language models: surveys of such work include [3, 4, 12]. The predominant family of approaches relies on correlating the activation patterns in trained models to linguistic structures that are considered necessary for correctly processing natural language. For written language, these are often various types of word categories, constituency structures or syntactic and semantic dependencies, for example as proposed in [5, 6, 7].

For models of spoken language, most previous work has focused on acoustic, phonetic and phonemic structures as well as on speaker characteristics, which are the most plausible types of information a speech model is expected to learn. Works analyzing the encoding of phonemes in a variety of speech models ranging from basic CNN-based ASR models to current transformer-based self-supervised models [8, 9, 13, 14, 15, 16] tend to find a salient encoding of phonemes in some layers of the models analyzed. In [10] the authors check for the encoding of acoustic, phonemic, lexical and semantic information in the self-supervised wav2vec2 [1] using probes such as canonical correlation analysis and mutual information, finding an autoencoder-style behavior, where across the layers the representations first diverge from low-level input features and at the end approximate them again. The reverse is the case for higher-level information such as word identity and meaning. A few papers have focused on analyzing phonology and/or semantics in visually-grounded models [17], for example [18, 19], in general finding phonemes encoded in lower layers and semantics in higher layers.

Much less work has looked at syntactic structures in models of spoken language: one partial exception is [20], who probe wav2vec2 and Mockingjay [21] for the encoding of acoustic and linguistic information, including syntax tree depth. The results reported are quite surprising and even implausible in that the encoding of most linguistic features in the speech models is found to be stronger than in the text-based BERT.

In this paper, we focus exclusively on probing for syntactic structures, aiming to examine them in several large-scale speech models, while taking care to include two independent methods as well as all the appropriate baselines and sanity checks in order to quantify the encoding of syntax in a reliable way.

3. Methods

This study aims to reliably establish the presence of syntactic information in models of spoken language. We, therefore, use established methods for this type of analysis and focus on careful experimental design rather than technical novelty. We use two separate probing techniques [5, 7], with several target models trained and tested on two different datasets.

3.1. Datasets

We use two English audio datasets for the current study: LibriSpeech [22] and SpokenCOCO [23]. LibriSpeech consists of audiobook recordings from the LibriVox project with a total of 960 hours of audio. SpokenCOCO is a spoken version of the image caption dataset COCO [24] with more than 600,000 spoken utterances paired with text captions and images. We use the LibriSpeech train-100h split and the SpokenCOCO validation split in our probing experiment to reduce computational load. We filter out utterances longer than 52 words for LibriSpeech and those longer than 20 words for SpokenCOCO. Table 1 shows the details of the data used in our experiments.

Table 1: *Datasets used for this study.*

Name	#Utts.	# Filtered Utts.
LibriSpeech	24,766	24,592
SpokenCOCO	28,539	27,496

3.2. Target Models

For this study, we use the following model variants as specified in Table 2:

wav2vec2 [1] is pre-trained on LibriSpeech to discriminate masked time steps in feature encoder outputs. In addition to the base pre-trained model, we also include a fine-tuned base model and a fine-tuned large model. The large model[2] is fine-tuned for English ASR on the same dataset. Additionally, we fine-tuned the base model for English ASR on our experimental data for 10,000 steps, which enables us to check the effect of model size.

HuBERT [2] is similar to the wav2vec2 architecture but pre-trained on labels created off-line via clustering; also pre-trained on LibriSpeech.

FaST-VGS [25] is a visually grounded model based on the wav2vec2 architecture. It loads model weights from the pre-trained wav2vec2-base with randomly reinitialized final four layers and is further trained on SpokenCOCO to match images with the speech that describes them.

FaST-VGS+ [25] same as the previous model, but trained on both SpokenCOCO and LibriSpeech with a combination of the visually-grounded loss and the wav2vec2 self-supervised loss.

BERT [11] text-based language model included as a ceiling reference. BERT is pre-trained on 3,300M words of books and web content.

[2] `https://huggingface.co/jonatasgrosman/wav2vec2-large-english`

BoW a text-based bag-of-words representation constructed from the combined text from both datasets with all sentences containing non-Latin characters removed. This representation captures all the words in the utterance, but no word-order information.

Table 2: *Models investigated in this study. PT = Pre-Trained, FT = Fine-Tuned, SS = Self-Supervised, VS = Visually Supervised, AV = Audio-Visual.*

Model	Size	Train.	Loss	Mod.
wav2vec2	base	PT	SS	Audio
wav2vec2	base	FT	SS	Audio
wav2vec2	large[3]	FT	SS	Audio
HuBERT	base	PT	SS	Audio
FaST-VGS	base	PT	VS	AV
FaST-VGS+	base	PT	SS+VS	AV
BERT	base	PT	SS	Text
BoW				Text

3.3. TreeDepth Probe

The objective of this probe is to predict the maximum depth of the constituency tree of a given utterance from the activation pattern in each transformer layer of a model when processing this utterance.

We generated hidden-state outputs and applied mean-pooling along the time axis to generate utterance vectors for all transformer layers of each model. We used the Stanza parser [26] to generate constituency trees for all utterances and calculate their tree depth. We fit a ridge regression model on embeddings generated for both LibriSpeech and SpokenCOCO. As controls, wordcount, and bag-of-words (BoW) model representation and their combinations with the embeddings were used in training the regression model as well. A 75:25 train-test split was used, and the model selected via 10-fold cross-validation was then evaluated on the test split and its score reported.

3.4. TreeKernel Probe

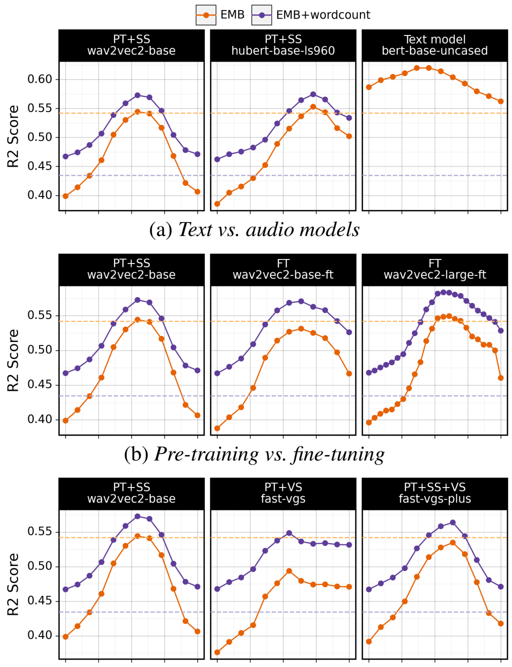
Representational Similarity Analysis (RSA) [27] is a method which correlates the similarity structures of two representational spaces. $\text{RSA}_{\text{regress}}$ [7] introduces a trainable version of RSA, where two input and output spaces are set up in terms of vectors of similarity to a held-out set of anchor points, and then a multivariate regression model is fit to map between them.

As in the original paper, we used cosine similarity between vectors in the input space as metric, and tree kernels between pairs of syntax trees in the output space. Tree kernel is a way of measuring similarities between syntactic trees by efficiently computing the proportion of shared tree segments. After obtaining the constituency trees, we delexicalized the trees to ensure the tree kernel is only based on structure and not word overlap. We used the algorithms introduced in [28, 29] for computing the normalized tree kernel; we closely followed the specific details in [7], using parameter $\lambda = \frac{1}{2}$ and two hundred anchor sentences. The score of the model selected via 10-fold cross-validation is reported.

For both probes, the hyperparameter tuned was regularization strength α for values $\{10^n \mid n \in \{-3, -2, -1, 0, 1, 2\}\}$.

4. Results

Figures 1 and 2 show the results for the TreeDepth and TreeKernel probing tasks, respectively. For clarity, Figure 1 only shows results on Librispeech.



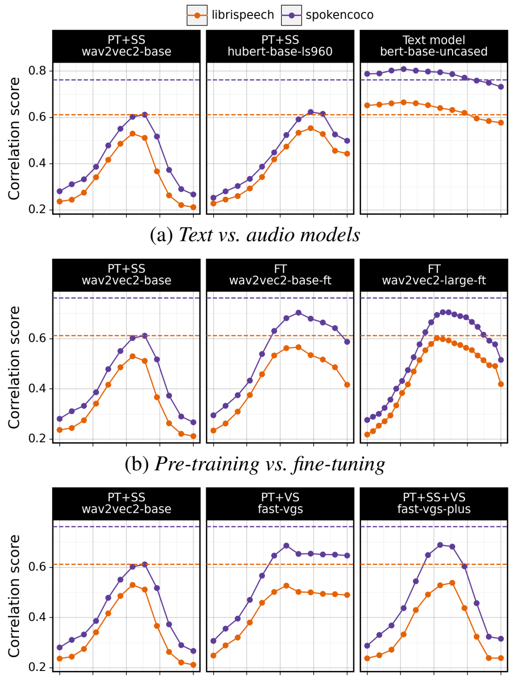
(a) *Text vs. audio models*

(b) *Pre-training vs. fine-tuning*

(c) *Self vs. visual supervision*

Figure 1: R^2 *scores for predicting TreeDepth from embeddings (Librispeech). X-axis = transformer layer from shallow to deep, orange dashed line = BoW reference, purple dashed line = Word count reference. See Table 2 for panel headings.*

[3] The base models have 12 transformer layers and a hidden size of 768; the large model has 24 layers and a hidden size of 1024.



(a) *Text vs. audio models*

(b) *Pre-training vs. fine-tuning*

(c) *Self vs. visual supervision*

Figure 2: *R^2 scores for embedding distances and TreeKernel. X-axis = transformer layer from shallow to deep. Dashed lines = BoW reference. See Table 2 for panel headings.*

Syntax encoding in spoken vs. text models. Figure 1a shows that spoken language models feature the strongest encoding of tree depth in the middle to deep layers. Concatenating wordcount features to embeddings in the TreeDepth task improves the probing results in the middle layers and outperforms the BoW reference. Figure 2a shows a similar trend for the TreeKernel probe.

As expected, both probing tasks show that syntactic information is encoded more strongly and consistently in text models across all layers.[4] However, TreeKernel results for the middle layers of the spoken language models come close to the text models, with wav2vec2-large-ft having the highest correlation score. Overall, these results suggest that spoken language models encode syntactic structure to a moderate degree. Comparison with the BoW reference suggests that most of the syntactic information encoded in speech models is entangled with lexical representations, rather than being abstract.

Pre-training vs. fine-tuning Figure 1b shows that models fine-tuned on ASR achieve higher scores in the TreeDepth probe than the pre-trained models. The pre-trained models have a large dip at the final layers with the scores go back to those for the first layer. For the fine-tuned models, the final layer dip is less pronounced. The pre-training objective is to reconstruct/discriminate audio features, whereas the fine-tuning objective is to output well-formed transcriptions, which requires more syntactic information to be encoded in the activation patterns, resulting in the contrast between the final layer scores.

Comparing the panels in Figure 2b, we can see that while fine-tuning increases the amount of syntactic information encoded by the model, the size of the model also matters. The larger hidden size and deeper model architecture prove useful in encoding extra information. Wav2vec2-large-ft was also fine-tuned on significantly more data than wav2vec2-base-ft.

Self- vs. Visual supervision As shown in Figures 1c and 2c, the score curve from FaST-VGS+ has a similar shape as wav2vec2-base. In addition to visual supervision, the plus variant also uses the same masked language modeling loss as wav2vec2, and therefore the two models behave similarly. In comparison, the final layer dip is absent for the FaST-VGS model, likely due to the fact that FaST-VGS does not use the self-supervised objective, and thus not display a decrease in syntax encoding in the final layers.

5. Conclusions

We use two established probing techniques to assess the amount of syntactic information encoded by several spoken language models. The results from both probes confirm that spoken language models encode a moderate level of syntactic information. We see that different training objectives considerably affect the degree of syntax encoded in each layer of the models and so does model size, with text-based training and larger model size leading to higher syntactic probe scores.

Our study only looks at sentence-level representations: it would also be interesting to extend our experiments to sub-sentence level probing [6]. Additionally, while we only use English datasets in this work, future studies could compare the ability of large-scale spoken language models to encode syntactic structures across different languages.

6. Acknowledgements

This publication is part of the project *InDeep: Interpreting Deep Learning Models for Text and Sound* (with project number NWA.1292.19.399) of the National Research Agenda (NWA-ORC) program.

[4] Our results contradict those of [20]: possibly due to them report probing scores directly on probe-training data: the description in the paper is unclear on this point.

7. References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NIPS 2020*, 2020.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, vol. 29, pp. 3451–3460, 2021.

[3] Y. Belinkov and J. R. Glass, "Analysis methods in neural language processing: A survey," *TACL*, vol. 7, pp. 49–72, 2018.

[4] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *TACL*, vol. 8, pp. 842–866, 2020.

[5] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties," in *ACL 2018*, vol. 1: Long Papers, Jul. 2018, pp. 2126–2136.

[6] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," in *NAACL*, 2019.

[7] G. Chrupała and A. Alishahi, "Correlating Neural and Symbolic Representations of Language," in *ACL 2019*. ACL, 2019, pp. 2952–2962. [Online]. Available: https://aclanthology.org/P19-1283

[8] A. Krug, R. Knaebel, and S. Stober, "Neuron activation profiles for interpreting convolutional speech recognition models," in *NeurIPS-IRASL*, 2018.

[9] G. Chrupała, B. Higy, and A. Alishahi, "Analyzing analytical methods: The case of phonology in neural models of spoken language," in *ACL 2020*. ACL, 2020, pp. 4146–4156.

[10] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *IEEE ASRU 2021*, 2021, pp. 914–921.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019*, vol. Long and Short Papers, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[12] H. Sajjad, N. Durrani, and F. Dalvi, "Neuron-level interpretation of deep nlp models: A survey," *TACL*, vol. 10, pp. 1285–1303, 2021.

[13] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *Proc. Interspeech*, 2015.

[14] Y. Belinkov and J. R. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *NIPS*, 2017.

[15] Y. Belinkov, A. Ali, and J. Glass, "Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 81–85.

[16] M. de Seyssel, M. Lavechin, Y. Adi, E. Dupoux, and G. Wisniewski, "Probing phoneme, language and speaker information in unsupervised speech representations," in *Proc. Interspeech*, 2022.

[17] G. Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," *J. Artif. Intell. Res.*, vol. 73, pp. 673–707, 2021.

[18] A. Alishahi, M. Barking, and G. Chrupala, "Encoding of phonology in a recurrent neural model of grounded speech," in *Proc. 21st CoNLL 2017*. ACL, 2017, pp. 368–378. [Online]. Available: https://aclanthology.org/K17-1037

[19] K. Khorrami and O. Räsänen, "Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? - a computational investigation," *LDR*, vol. 1, no. 1, 2021. [Online]. Available: http://ldr.lps.library.cmu.edu/article/id/434/

[20] Y. K. Singla, J. Shah, C. Chen, and R. R. Shah, "What do audio transformers hear? probing their representations for language delivery & structure," in *ICDMW 2022*, 2022, pp. 910–925.

[21] A. T. Liu, S. wen Yang, P.-H. Chi, P.-C. Hsu, and H. yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *ICASSP 2020*, pp. 6419–6423, 2019.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP 2015*, 2015, pp. 5206–5210.

[23] W.-N. Hsu, D. Harwath, T. Miller, C. Song, and J. Glass, "Text-Free Image-to-Speech Synthesis Using Learned Segmental Units," in *ACL-IJCNLP 2021*, vol. 1: Long Papers. ACL, 2021, pp. 5284–5300. [Online]. Available: https://aclanthology.org/2021.acl-long.411

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/

[25] P. Peng and D. Harwath, "Self-supervised representation learning for speech using visual grounding and masked language modeling," in *AAAI-SAS 2022*, 2022.

[26] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages," in *ACL 2020: System Demonstrations*. ACL, 2020, pp. 101–108. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-demos.14

[27] N. Kriegeskorte, M. Mur, and P. A. Bandettini, "Representational similarity analysis-connecting the branches of systems neuroscience," *Frontiers in systems neuroscience*, p. 4, 2008.

[28] A. Moschitti, "Making Tree Kernels Practical for Natural Language Learning," in *EACL 2006*. ACL, 2006, pp. 113–120. [Online]. Available: https://aclanthology.org/E06-1015

[29] M. Collins and N. Duffy, "Convolution Kernels for Natural Language," in *NIPS 2001*, vol. 14. MIT Press, 2001.