Using Shapley interactions to understand how models use structure

Anonymous ACL submission

Abstract

Language is an intricately structured system, and a key goal of NLP interpretability is to provide methodological insights for understanding how language models internally represent this structure. In this paper, we use Shapley Taylor interaction indices (STII) in order to examine how language and speech models internally relate and structure their inputs. Pairwise Shapley interactions give us an attribution measure of how much two inputs work *together* to influence model outputs *beyond* if we linearly added their independent influences, providing a view into how models encode structural interactions between inputs. We relate the interaction patterns in models to three underlying linguistic structures: syntactic structure, non-compositional semantics, and phonetic interaction. We find that autoregressive text models encode interactions that correlate with the syntactic proximity of inputs, and that both autoregressive and masked models encode non-linear interactions in idiomatic phrases with non-compositional semantics. Our speech results show that inputs are more entangled for pairs where a neighboring consonant is likely to influence a vowel or approximant, showing that models encode the phonetic interaction needed for extracting discrete phonemic representations.

1 Introduction

How do language model features work *together* to influence prediction results? Do the internals of language models reflect the complex structure of language in how they combine features? Understanding feature attribution, how different model features (like inputs or neurons) influence output decisions, is a key question for understanding and interpreting neural models. One common approach to feature attribution is adapted from game theory scenarios, and treats features like agents in a cooperative game, attributing credit for the outcome to each feature (Lundberg and Lee, 2017). This credit value, or **Shapley value** (Shapley, 1952), quantifies the effect of each feature on the output, assuming that features act in a linearly independent manner on the output. The linearity assumption is not accurate for most deep learning scenarios: neural networks are non-linear, and features interact in complex ways inside model representations to influence output predictions.

What interactions between features do we miss when we assume this linear independence? To address this question, researchers have proposed methods to calculate residuals, how much information we lose when assuming linearity (Kumar et al., 2021), and **Shapley interactions**, accounting for how features have influence in pairs or groups on top of how they act independently (Agarwal et al., 2019).

In this paper, we investigate how Shapley interactions can enhance our interpretable understanding of the internal processes of language models. We ground our investigation in structural features that we know about the input data (like syntactic structure), and ask: what do Shapley interactions reveal about how the model uses the non-linear structure in language? By relating Shapley interactions to structural linguistic features, we showcase how different models use (or don't use) linguistic structural features in their internal representations. We run experiments on autoregressive and masked text models, as well as on automatic speech recognition models, and report the following findings:

- Autoregressive models (but not masked models) show a strong correlation between Shapley interaction and the syntactic proximity of features. This indicates that syntactic structure is encoded in non-linear interactions between model features (Section 3.2)
- Both autoregressive and masked models ex-

hibit stronger interactions between pairs of tokens in multiword expressions (MWEs) that have idiomatic non-compositional meaning (expressions like *I'll eat my hat*) (Section 3.3)

- In speech models, Shapley interactions are stronger between consonants and vowels than between pairs of consonants, in accordance with how sounds interact in speech: the acoustics of vowels are often shaped by the surrounding consonants, while consonants are more able to be interpreted in isolation (Rakerd, 1984) (Section 4.1). This finding also extends to more sonorant vowel-like consonants, which interact more with surrounding consonants than those produced with the vocal tract more closed (Section 4.2).

Understanding non-linearities and interactions in model internals is becoming a vital missing piece of the wider language model interpretability inquiry. Our work showcases how Shapley interactions are a powerful interpretability methodology for examining how language models use the structure in their inputs to organize their internal representations.

2 Background and related work

2.1 Shapley Interactions

Shapley values are used to attribute decisions to specific features in predictive models. The Shapley value of a set of features A is obtained by computing the difference in a model's output when A is included versus when it is excluded. If we take the set of all features, N, and remove A, we want to see how much value A adds to every possible subset $S \subseteq N \backslash A$. In our case, the value function v is the logit output of the model. The Shapley value is the weighted average of this marginal contribution over all S:

$$\phi(A) = \sum_{S \subseteq N \backslash A} w_S \left(v(S \cup A) - v(S) \right) \quad (1)$$

where the weight w_S for each subset is the number of possible subsets S of the same size:

$$w_S = \binom{|N| - |A|}{|S|} \quad (2)$$

If the interactions between features are linearly additive: $\phi(\emptyset) \approx \sum_{i \in S} v(\{i\})$. However, in scenarios where features are dependent and their composition is non-linear, Shapley values do not account for interacting effects between sets. Methods to understand and address this have been proposed by Owen (1972), Grabisch and Roubens (1999), Fumagalli et al. (2023), and Tsai et al. (2023). Here, we focus on the Shapley residual (Kumar et al., 2021), which calculates how much the Shapley linearity assumptions are violated:

$$r_i = \nabla_i \phi - \nabla \phi(\{i\}) \quad (3)$$

For simplicity, we consider the case of pairwise interactions: interaction between a pairs of feature sets A and B. To calculate pairwise Shapley interactions, we rely on the **Shapley Taylor interaction index** (STII) (Agarwal et al., 2019) to calculate second-order interactions using the discrete second-order derivative. Since our features are vectors, we calculate the scalar Shapley interaction value for each dimension individually, and take the norm of this Shapley vector for a scalar metric of interaction. Similar to Saphra and Lopez (2020), we scale the residual by the norm of the entire sequence with no feature ablations.

$$\text{STII}_{A,B} = \frac{\|\phi(\emptyset) - \phi(A) - \phi(B) + \phi(A,B)\|_2}{\|\phi(\emptyset)\|_2} \quad (4)$$

Calculating the Shapley values for each coalition requires iterating over the powerset of N, requiring $O(2^{|N|})$ calculations. In high-dimensional input spaces, the exact calculation of Shapley residuals is therefore prohibitively expensive. We approximate Shapley values by using Monte Carlo Permutation Sampling (Castro et al., 2009).

2.2 Structure in language models

There is a huge and varied literature in NLP interpretability aimed at understanding how language models use and represent the structure in their linguistic input. Approaches include examining if the output probabilities of language models reflect structural rules (see for example Warstadt et al., 2018; Hu et al., 2024; Gauthier et al., 2020), as well as looking inside model representations. For the latter approaches, while many linguistic structural elements can be linearly extracted from the representations of text and speech models (see Hewitt and Manning, 2019; Belinkov, 2021; Pasad et al., 2024; Chrupała et al., 2020; Park et al., 2023, among many others), and attribution methods can relate the linear importance of different features in both text and speech models (Markert et al., 2021; Ethayarajh and Jurafsky, 2021; Yeh et al., 2020; Kokalj et al., 2021), the fact remains that neural

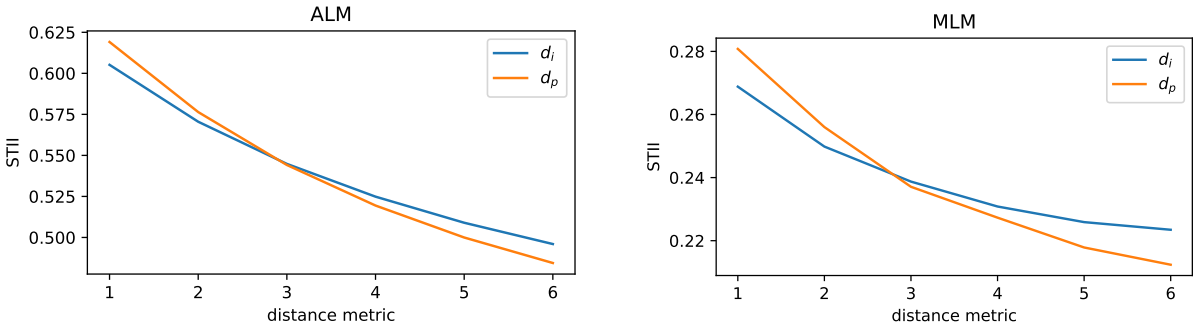


Figure 1: Our results for the experiments relating Shapley Interactions with a token's position in the sequence. We find that, for both autoregressive models (left) and masked models (right), STII decreases monotonically with distance. This holds both when we are measuring distance as the distance between the two elements in the interacting pair (d_i, blue line) and when we are measuring distance between the interacting pair and the token that the model is predicting (d_p, orange line). Our results indicate that models treat tokens that are far away from each other more like an unentangled bag-of-words, and that they treat pairs of tokens that are far away from the token being predicted as unentangled, no matter the distance between them.

models have complex nonlinearities in their internal processing.

How can we analyze the ways in which nonlinear interactions play out in model internals, and what they encode? Multiple papers have analyzed the difficulties of knowing what we can extract when using nonlinear probing methods (Voita and Titov, 2020; Pimentel and Cotterell, 2021; Hewitt et al., 2021), and others have proposed searching for causal effects which can be generally agnostic to whether the processing is linear (Geiger et al., 2021; Arora et al., 2024). Shapley interactions let us directly link features of the input to different *extents* of nonlinear processing. Prior work showing the utility of Shapley interactions in analyzing NLP models has focused on older architectures like LSTMs, and on models fine-tuned for simple text classification tasks (Saphra and Lopez, 2020; Jumelet and Zuidema, 2023; Chen et al., 2020; Singh et al., 2019). Our work builds on and generalizes these results by relating Shapley interactions to diverse forms of linguistic structure (syntactic, semantic, and phonetic) on models trained on domain-general language tasks (generation for text, and ASR for speech)

3 Text models: Interactions between tokens

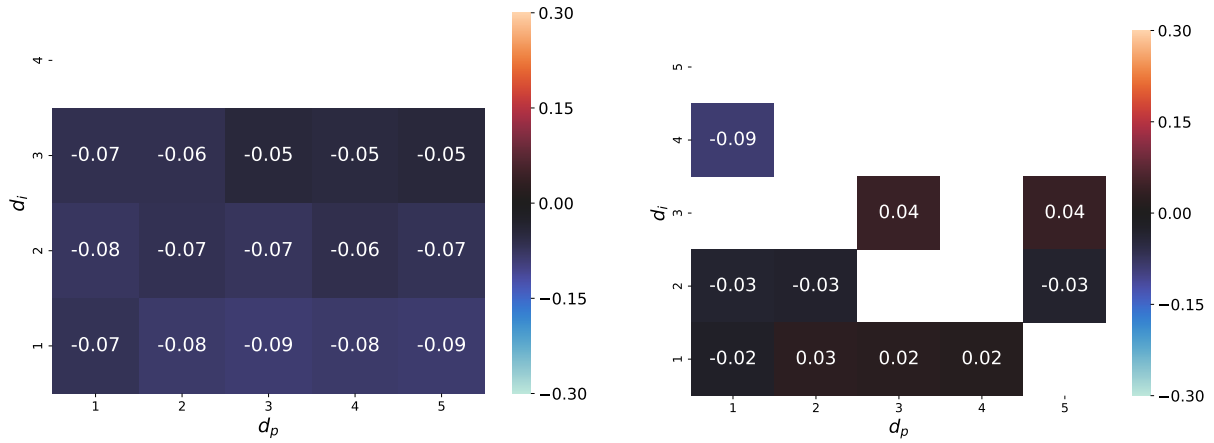
Our first experiments are on language models, measuring how known associations between tokens correlate with Shapley-based measures of feature interaction. We consider the influence of token position, idiomatic phrases, and syntax. We find that masked LMs and Autoregressive LMs differ in their interaction structure, especially in how they respond to syntax.

Models and Datasets We run all of our experiments on two models: the autoregressive model GPT-2 (Radford et al., 2019) and the masked language model BERT-base-uncased (Devlin et al., 2018). Each input sentence is unpadded and truncated to 20 tokens, and we apply softmax to the logit outputs to ensure that interactions across different examples are comparable.

All English language modeling experiments use `wikitext-2-raw-v1` (Merity et al., 2016) tokenized and dependency parsed (for syntax experiments) with spaCy (Honnibal et al., 2020). We resolve incompatibilities between the spaCy tokenizer and the model-specific tokenizers by assigning overlapping tokens a syntactic distance of zero. For the multiword expression experiments, we use the AMALGrAM supersense tagger (Schneider et al., 2014a), which identifies both strong and weak (Schneider et al., 2014b) MWEs.

3.1 Baseline: the effect of position

One potential factor influencing interactions between tokens is the positional distance between tokens Let's say that we are calculating the interaction between two tokens, x_{t_1} and x_{t_2} at positions t_1 and t_2. The token that the model is trying to predict (i.e. the next token in autoregressive models, and the masked token in masked models) is at position



(a) Autoregressive language model experiments

(b) Masked language model experiments

Figure 2: The results of our syntactic distance experiments (Section 3.2): how does syntactic distance correlate with STII, controlling for the effect of position? A negative correlation means that tokens closer in the parse tree (low syntactic distance) are more heavily entangled (high STII). Autoregressive models show a consistently negative correlation in all significant cells, meaning that syntax is encoded in Shapley interactions. We stratify our results by the two positional distance metrics in Section 3.1, so that we can calculate the effect of syntactic distance, marginalizing out the effect of positional distance. Each cell displays a correlation between syntactic distance and STII for a given interacting pair distance and prediction distance. We only provide results for cells where there exists at least one direct syntactic modifier pair separated by the positional distance d_i and the Spearman correlation given at that cell is statistically significance ($p < 0.05$). For our correlation calculation, we only include a syntactic distance if there are at least 50 data points with that syntactic distance in our data set.

t_{target}. There are two relevant positional distances that are likely to influence interaction.

Firstly, the **interacting pair distance**, d_i, is the distance between the two tokens, defined in Equation (5):

$$d_i(x_{t_1}, x_{t_2}, x_{t_{\text{target}}}) = t_2 - t_1 \quad (5)$$

Secondly, the **prediction distance**, d_p, is the distance between the pair of tokens that we are calculating the interaction of, and the target token that the model is trying to predict, defined in Equation (6):

$$d_p(x_{t_1}, x_{t_2}, x_{t_{\text{target}}}) = \min_{t \in \{t_1, t_2\}} \left| t_{\text{target}} - t \right| \quad (6)$$

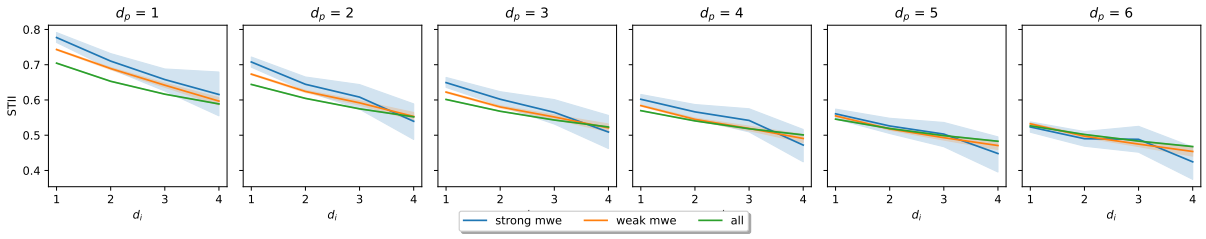
For our position baseline experiments, we test how both interacting pair distance and prediction distance influence the STII between the two tokens x_{t_1} and x_{t_2}

Results Our results are presented in Figure 1, confirming that distance has an effect on STII in both autoregressive and masked models. This holds whether we are measuring distance as distance between the interacting pair (interacting pair distance d_i) or distance between the last token in that pair and the target prediction token (prediction distance d_p). The dramatic decline of STII with increased prediction distance implies that when these models predict tokens, they treat the more distant context as a bag of words rather than as complex syntactic relations (Khandelwal et al., 2018). We also see that closer tokens interact more strongly with each other.

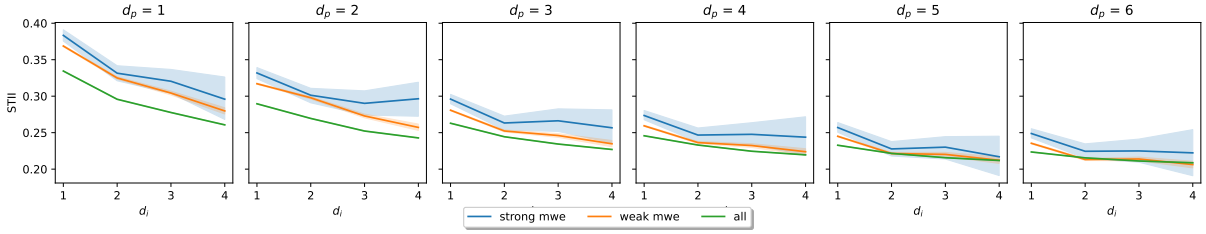
For the rest of our experiments, we will stratify samples by both d_i and d_p, so that we can measure the effects of linguistic structure *beyond* these position effects that we demonstrate here.

3.2 Syntactic structure

Syntactic structure can also influence an LM's predictions. If a model composed distant syntactic relations in a linear way, it would treat the wider context as though it were a bag of words. By instead exhibiting strong interactions between syntactically close tokens, the model would closely entangle the meaning of a modifier with its head. We measure **syntactic distance** by the number of dependency edges traversed to connect a pair of tokens, a metric encoded by projected representations in both masked (Hewitt and Manning, 2019) and autoregressive (Murty et al., 2022) models. We verify the role of modifier connections by the Spearman



(a) Autoregressive LM experiments



(b) Masked LM experiments

Figure 3: Results for our multiword expressions experiments: Shapley interactions are higher for tokens in multiword expressions than tokens that are not. The results are controlled for prediction distance d_p (different facets) and interacting pair distance d_i (x-axis). Within each facet for each x-axis value, we can see that the STIIs for tokens in Strong MWEs (blue) and Weak MWEs (orange) are significantly higher than the average over all pairs (green).

correlation between syntactic distance and STII, stratified by interacting pair distance and prediction distance.

Results Figure 2 shows correlation between syntactic distance and STII. Our analysis reveals that, for autoregressive language models, all statistically significant correlations are negative. In contrast, non-autoregressive language models exhibit both positive and negative correlations. This finding aligns with Saphra and Lopez (2020)'s research on LSTMs showing that syntax is handled more consistently in autoregressive models, and with Ahuja et al. (2024), who in a different setting show that autoregressive models are more predisposed to syntax-style generalizations.

The inconsistencies observed in non-autoregressive models may stem from their handling of positional proximity in less intuitive ways, complicating the relationship between syntactic and linear distance. The interaction between these two dimensions may be more difficult to manage in masked models, leading to the varied correlation outcomes.

This finding suggests that we can interpret feature interaction as a distinctly syntactic alternative to the inherent distance encoding found in autoregressive architectures (Haviv et al., 2022). In these models, the degree of interaction is learned to prioritize syntactic relationships rather than depending solely on positional information within the language modeling objective. This highlights a fundamental difference in how these models integrate syntactic structure and distance.

3.3 Multiword expressions

While semantics is often treated as compositional (the meaning of a sentence can be composed by rules, following the syntax and the meaning of each individual word), language is also characterized by non-compositional, or idiomatic, phrases. These are groups of words whose meaning can only be derived when looking at the entire group rather than the individual words. These word groups, known as **multiword expressions** (MWEs), include idioms like *break a leg*, where the isolated meaning of each of the component words *break*, *a*, and *leg* fail to compose the meaning of the entire expression. Higher interaction values for the tokens in the idiom would indicate a less compositional treatment of the whole phrase.

In these experiments, we compare interactions between arbitrary pairs of tokens to interactions between tokens contained within an MWE. The extreme case where there is no Shapley residual would imply perfect compositionality—after all, linear addition is compositional—so our hypothesis is that MWEs have a larger than average residual.

Results Figure 3 compares the STII between tokens that belong to the same MWE to the average STII between all tokens, stratified by interacting

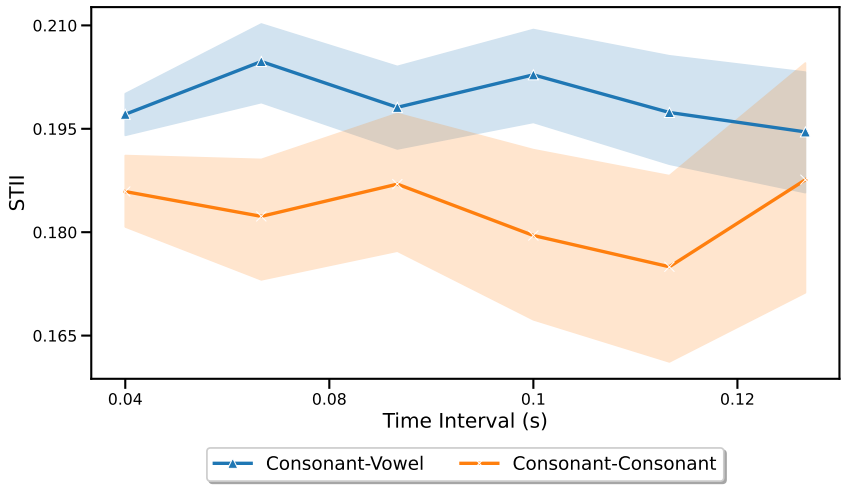


Figure 4: Vowel-consonant interactions are higher than consonant-consonant interactions when comparing adjacent inputs. There isn't a clear relationship between interaction and the size of the interval around the phoneme boundary. Confidence intervals are provided by bootstrap.

pair distance d_i and prediction distance d_p. For both the autoregressive models (Figure 3a) and masked models (Figure 3b), STII is higher when the interacting pair is in a MWE: the blue and orange MWE lines are overall higher in STII than the green baseline. The effect is consistent across positional distances and more pronounced when predicting nearby tokens.

4 Speech models: Interactions between phones

Do speech models represent phonetic interactions? Consonants influence the realization of vowels, and in order to be able to separate vowels into a consistent discrete system a listener has to take these interactions into account (Rakerd, 1984; Rosner, 1994). Vowels are produced in a continuous space, without clear boundaries that delineate which vowel a specific vocal tract positioning refers to (for example, a speaker can glide on the continuum between [i] and [e], but there is no clear analog of a continuum between [p] and [k]). The realization of vowels is influenced by the consonants that surround them. Despite the continuous nature of vowel phonetics, listeners perceive vowels as belonging to a few discrete classes of vowel phonemes. To derive this discrete phonological representation, a listener — or predictive speech model — would need to represent the structure of consonant-vowel interaction in order to be able to take this into account. We use Shapley interactions to

Since the inputs of speech models are not cleanly tokenized into phones, and the transition between phones is continuous and without a well-defined boundary, we measure interaction by taking the average pairwise interaction within a time interval that includes a transition. For a given interval length, we measure STII between all temporally consecutive features p_{t_1} and p_{t_2} when predicting the immediate next sound p_{t_3}. Formally, the interaction N between different phonemes over a temporal interval within range δ of the approximated phone boundary time t_b is:

$$\bar{r}_\delta = \sum_{t_1=t_b-\delta}^{t_b+\delta} \text{STII}_{p_{t_1},p_{t_2}} \quad (7)$$

Note, however, that in the case where no acoustic feature is sampled at exactly $t_b - \delta$, we instead start the summation with t_1 at the earliest timestamp such that $t_1 \geq t_b - \delta$. Since all interaction pairs are consecutive, the confounder of positional distance is automatically removed for these experiments.

Models and Datasets Our experiments are run on the Wav2Vec 2.0 model wav2vec2-base-960h (Baevski et al., 2020), which is trained on 960 hours of English audio to predict the next sound in a recording. When computing Shapley values, ablated acoustic features are replaced with silence.

For all experiments, we use the Common Voice dataset (Ardila et al., 2020) of English language voice recordings, which are contributed by volunteers around the world and comprise 92 hours of recorded speech. This compilation is characterized by its rich diversity, featuring a total of 1,570 unique voices. We preprocess the dataset by align-

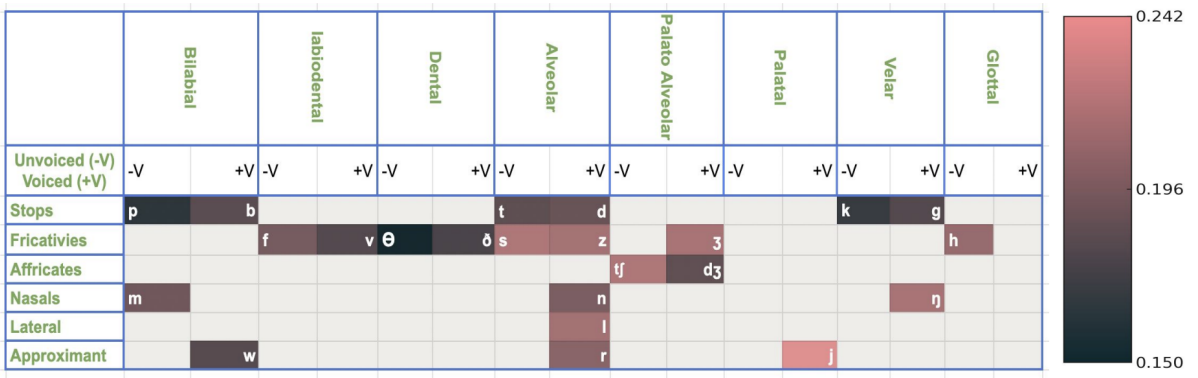


Figure 5: Consonant chart with a heat map indicating average interaction with acoustic features from adjacent phonemes (samples from 0.1s around the phoneme boundary). Columns indicate the place of articulation while rows indicate the manner of articulation. Only interactions for acoustic features within 0.1s range around the phoneme boundary are considered. Consonants with more vowel-like articulations (lower down in the chart) tend to have higher interactions with surrounding phonemes.

ing the audio recordings with their corresponding phonemes using p2fa_py3 [1], an implementation of the Penn Phonetics Lab Forced Aligner (Yuan et al., 2008), which uses acoustic models to map the audio recordings to their corresponding phonemes. We preprocess all audio files to a WAV and standard sampling rate and then use p2fa_py3 to detect and align phonemes within the speech to their corresponding timeframes in the recordings, marking the start and end of each phoneme. It is important to note, as a caveat to the following results, that identifying the exact duration of a phoneme is not only challenging but undefined in practice, as the vocal tract is in a state of continuous transition between phonemes throughout an utterance.

4.1 Interactions between consonants and vowels

Vowels are formed with an open vocal tract that produces no turbulent airflow, with the specific position of each part of that anatomy largely determined by the surrounding consonants. Therefore, it is harder to map vowel sounds in isolation to their corresponding discrete phoneme than it is to map consonants (Rakerd, 1984). In Figure 4, we compare the interactions over consonant-vowel boundaries and consonant-consonant boundaries, and find that interactions are significantly higher in the consonant-vowel case. This implies that the model is taking this entanglement into account, which is necessary for reaching a discrete phonological analysis of the input similar to human phonological perception.

[1] https://github.com/jaekookang/p2fa_py3

4.2 The effect of consonant manner of articulation

Not all consonants are equally stable in their capacity to be interpreted in isolation. In describing consonants, the *manner of articulation* refers to a hierarchy of vocal tract occlusion, ranging from the stops (consonants like [p], formed by briefly blocking all air through the vocal tract) to the approximants (consonants like [j] as in “universe”, that produce only slightly more turbulent airflow than vowels). Therefore, some consonants in practice behave more like vowels, and we expect them to exhibit more nonlinear interactions across phoneme boundaries, as vowels do.

Our hypothesis is largely confirmed in Figure 5, modeled on a International Phonetic Alphabet consonant chart where row indicates the manner of articulation. Although the pattern is not perfect, the figure shows high cross-phoneme STII for more sonorant consonants on the lower rows, which are articulated like vowels with a more open oral cavity. Sibilants ([s], [z], [ʒ], [ʃ], [tʃ], [dʒ]) also show high cross-phoneme STII, which is also expected as they are known to lie on a continuum (a continuous space of where the tongue articulates on the roof of the mouth) where boundaries are influenced by surrounding phonemes (Mann and Repp, 1980; Fleischer et al., 2013). Notable exceptions to the pattern include [w] (a possible reason being that, even though it is an approximant, it is articulated in two places simultaneously: the lips and the velum in the back of the mouth), and [h], which is marked as a fricative in the IPA chart for varied rea-

sons (Laufer, 1991) but is articulated largely like an approximant (Ladefoged, 1990).

5 Future Work

Our primary objective in this work has been to showcase the versatility of Shapley interactions in showing the ways that language models encode linguistic structure. Understanding structural representation, and especially how this can be nonlinear, is a long-standing problem and inquiry in NLP interpretability. This work suggests a number of open questions and follow-up problems, in addition to having the potential to be applied as is to different types of annotated linguistic structure.

Speech has multiple layers of structure, as it comprises both an acoustic signal and the language structure underlying the utterance. Our investigation of feature interactions is limited to the phonetic level, but future work may find the degree to which these multiple layers of linguistic structure affect nonlinear feature interactions. Do these speech models exhibit similar interaction patterns to the autoregressive language models we also analyze? Speech, often neglected in interpretability research, is ripe with open problems.

While we compare the behavior of the models trained on the masked and autoregressive objectives, we do not compare any models that are trained on the same objective with different architectures. The inductive bias and function of a given architecture are matters of great interest to many researchers in machine learning, and we believe that measuring nonlinear interactions can provide many insights into how specific models are similar and different.

This work focuses on pairwise interactions, and so has not taken full advantage of the versatility of Shapley residuals as a tool. Higher order Shapley interactions (Sundararajan et al., 2020) provide a method of hierarchical clustering on features and introduce yet more nuance into approximations of linear and nonlinear behavior in neural networks. We also do not consider interactions of internal model features. We suggest that future work in the area should incorporate knowledge about the underlying semantics of the input as well as the model architecture.

Finally, and most crucially, we believe that followup work in this area should be interdisciplinary. Speech, language, image processing, and other areas that can benefit from interpretability are all well-studied, with decades or even centuries of scientific research. By collaborating with specialists in these data domains, we can potentially contribute not only to the understanding of artificial models, but also to the understanding of the natural phenomena in question. Interpretability is an important new area in the emerging field of AI for scientific understanding and discovery, and we encourage others to start future work by finding domain experts to choose questions worth asking.

6 Conclusions

In accordance with The Bitter Lesson (Sutton, 2019), researchers and engineers typically apply machine learning methods generically, incorporating as little explicit data structure as possible. However, The Bitter Lesson does not apply to *interpretability*. Instead, meaningful interpretations of representational and mechanistic structures at scale should be informed by the underlying structure of data. Our results show how to use constituents, phones, and object boundaries to build a scientific understanding that goes beyond intuitions about n-grams, acoustic features, and pixels.

These results have spanned modality and task. By measuring feature interaction in language models, we present a novel way of describing how the hierarchy of syntactic structure and the encoding of non-compositional semantics both function in model internal representations. In speech prediction models, we show that consecutive acoustic features near a phone transition have more nonlinear interactions if the transition is between a consonant and vowel, rather than between two consonants. We also see that in this sense, sonorant consonants behave more like vowels.

These studies do not focus on individual data samples, but on patterns in the structure underlying the data. Understanding these general patterns requires greater domain expertise than is often required for sample-level interpretability research. We hope to inspire future interdisciplinary work with phonology, syntax, visual perception, and other sciences that characterize corpus-wide structural phenomena.

7 Limitations

The work in this paper shows correlations between pairwise Shapley interactions and structural relationships between two inputs. Both the pairwise aspect, and the fact that we only do correlational

analyses, are limitations. There are two ways to expand the analysis to make it more descriptive and informative about the internal processing of models. Firstly, we could look beyond pairwise interactions, creating a hierarchy of interaction: single feature, pairwise, groups of three features, etc. This hierarchy of interaction could be related to more subtle and hierarchical features. While currently we're limited to pairwise features like syntactic proximity, we could more fully analyze complex tree structure if we had a hierarchy of interaction effects. The second way in which this analysis could be made stronger would be to go beyond looking at correlations, and investigate the causal predictive power of Shapley interactions, and the ways in which they change the structural processing and effects of language models.

The analyses in this paper are not on model sizes close to the order of magnitude of state-of-the-art production models, meaning that the specifics of our results might not be relevant to the models that are having the most effect on the world at the moment. Our paper is meant to showcase the applicability of STIIs to relating model internals to structure in the input, and like all interpretability methods introduced on smaller models, we hope that the viewpoint and methodologies of this paper can be applied to larger models in the future as the field and our understanding develops.

References

Ashish Agarwal, Kedar Dhamdhere, and Mukund Sundararajan. 2019. A new interaction index inspired by the taylor series. *CoRR*, abs/1902.05622.

Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A Smith, Navin Goyal, and Yulia Tsvetkov. 2024. Learning syntax without planting trees: Understanding when and why transformers generalize hierarchically. *arXiv preprint arXiv:2404.16367*.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. Causalgym: Benchmarking causal interpretability methods on linguistic tasks. *arXiv preprint arXiv:2402.12560*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Preprint*, arXiv:2102.12452.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Comput. Oper. Res.*, 36:1726–1730.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. *Preprint*, arXiv:2004.02015.

Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi. 2020. Analyzing analytical methods: The case of phonology in neural models of spoken language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4146–4156, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Kawin Ethayarajh and Dan Jurafsky. 2021. Attention flows are shapley value explanations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 49–54.

David Fleischer, Michael Wagner, and Meghan Clayards. 2013. A following sibilant increases the ambiguity of a sibilant continuum. In *Proceedings of Meetings on Acoustics*, volume 19. AIP Publishing.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. 2023. Shap-iq: Unified approximation of any-order shapley interactions. *Preprint*, arXiv:2303.01179.

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.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Michel Grabisch and Marc Roubens. 1999. "an axiomatic approach to the concept of interaction among players in cooperative games". *International Journal of Game Theory*, 28:547–565.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer language models without positional encodings still learn positional information. *Preprint*, arXiv:2203.16634.

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 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.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Jaap Jumelet and Willem Zuidema. 2023. Feature interactions reveal linguistic structure in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8697–8712, Toronto, Canada. Association for Computational Linguistics.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.

Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. Bert meets shapley: Extending shap explanations to transformer-based classifiers. In *Proceedings of the EACL hackashop on news media content analysis and automated report generation*, pages 16–21.

Indra Kumar, Carlos Scheidegger, Suresh Venkatasubramanian, and Sorelle Friedler. 2021. Shapley residuals: Quantifying the limits of the shapley value for explanations. In *Advances in Neural Information Processing Systems*, volume 34, pages 26598–26608. Curran Associates, Inc.

Peter Ladefoged. 1990. Some proposals concerning glottal consonants. *Journal of the International Phonetic Association*, 20(2):24–25.

Asher Laufer. 1991. The 'glottal fricatives'. *Journal of the International Phonetic Association*, 21(2):91–93.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Virginia A Mann and Bruno H Repp. 1980. Influence of vocalic context on perception of the []-[s] distinction. *Perception & Psychophysics*, 28(3):213–228.

Karla Markert, Romain Parracone, Mykhailo Kulakov, Philip Sperl, Ching-Yu Kao, and Konstantin Böttinger. 2021. Visualizing automatic speech recognition - means for a better understanding? In *2021 ISCA Symposium on Security and Privacy in Speech Communication*. ISCA.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D. Manning. 2022. Characterizing intrinsic compositionality in transformers with tree projections. *Preprint*, arxiv:2211.01288 [cs].

Guillermo Owen. 1972. Multilinear extensions of games. *Management Science*, 18(5):P64–P79.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

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.

Tiago Pimentel and Ryan Cotterell. 2021. A bayesian framework for information-theoretic probing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2869–2887. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Brad Rakerd. 1984. Vowels in consonantal context are perceived more linguistically than are isolated vowels: Evidence from an individual differences scaling study. *Perception & psychophysics*, 35:123–136.

BS Rosner. 1994. Vowel perception and production.

Naomi Saphra and Adam Lopez. 2020. LSTMs compose—and Learn—Bottom-up. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2797–2809, Online. Association for Computational Linguistics.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014a. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 455–461. European Language Resources Association (ELRA).

Lloyd S. Shapley. 1952. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA.

Chandan Singh, W. James Murdoch, and Bin Yu. 2019. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. 2020. The shapley taylor interaction index. In *International conference on machine learning*, pages 9259–9268. PMLR.

Richard Sutton. 2019. The bitter lesson. *Incomplete Ideas (blog)*, 13(1).

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. 2023. Faith-shap: The faithful shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. *arXiv preprint arXiv:2003.12298*.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *CoRR*, abs/1805.12471.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565.

Jiahong Yuan, Mark Liberman, et al. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878.