From Words to Waves: Analyzing Concept Formation in Speech and Text-Based Foundation Models

Asım Ersoy, Basel Mousi, Shammur Chowdhury, Firoj Alam, Fahim Dalvi, Nadir Durrani

Qatar Computing Research Institute, HBKU, Qatar,
shchowdhury@hbku.edu.qa, ndurrani@hbku.edu.qa

Abstract

The emergence of large language models has demonstrated that systems trained solely on text can acquire extensive world knowledge, develop reasoning capabilities, and internalize abstract semantic concepts–showcasing properties that can be associated with general intelligence. This raises an intriguing question: *Do such concepts emerge in models trained on other modalities, such as speech?* Furthermore, when models are trained jointly on multiple modalities: *Do they develop a richer, more structured semantic understanding?* To explore this, we analyze the conceptual structures learned by speech and textual models both individually and jointly. We employ Latent Concept Analysis, an unsupervised method for uncovering and interpreting latent representations in neural networks, to examine how semantic abstractions form across modalities. To support reproducibility, we have released our **code**[1] along with a curated **audio version of the SST-2 dataset**[2] for public access.

Index Terms: Multimodal Learning, Interpretability, Conceptual Abstractions

1. Introduction

Recent advances in artificial intelligence have led to the development of large neural models capable of processing and generating language, vision, and speech [1, 2, 3, 4, 5]. Among these, large language models (LLMs) have demonstrated emergent capabilities once thought to require human intelligence. From commonsense reasoning to medical diagnosis and legal analysis, these models have continuously pushed the boundaries of AI-driven understanding and decision-making [1, 6]. Their ability to internalize abstract concepts, perform multi-step reasoning, and apply knowledge in novel contexts has fueled discussions on their potential trajectory toward artificial general intelligence (AGI).

However, a fundamental question remains: *Are these emergent capabilities unique to text-based models, or do similar properties arise in models trained on other modalities, such as speech?* Furthermore, when models are trained on multiple modalities, *do they converge toward a shared semantic space that facilitates conceptual abstraction*, as proposed by the *Semantic Hub Hypothesis* [7]? To explore these questions, we analyze the conceptual structures learned by speech models and compare them to text-based models and jointly trained multimodal systems. We employ Latent Concept Analysis (LCA) [8], to uncover and compare the abstract concepts formed within these models. We align the discovered concepts with predefined taxonomies, enabling a structured comparison of semantic representations across modalities.

In our study, we investigate unimodal models (HuBERT for speech and BERT for text) and multimodal models (Seamless M4T and SpeechT5) to assess their ability to capture linguistic knowledge. We evaluate these models using core linguistic tasks like part-of-speech tagging, chunking, and semantic analysis, as well as sentiment analysis, to compare their ability to align with human-defined concepts and task-specific representations. Our study addresses the following research questions:

- **Question:** How do conceptual structures in speech models differ from those in text and multimodal models?
  Finding: The alignment patterns reveal that text models directly encode linguistic taxonomies from early layers, while speech models gradually transition from acoustic to linguistic representations. Multimodal models like SpeechT5 show unique alignment due to cross-modal training.
- **Question:** To what extent do different modalities yield shared or modality-specific semantic representations?
  Finding: Speech models allocate less capacity to linguistic and semantic taxonomies, focusing more on speech-specific features like phonetics. In contrast, text models, which operate on tokenized inputs, develop more structured and deep linguistic representations. This highlights the disparity in how each modality internalizes semantic structures.

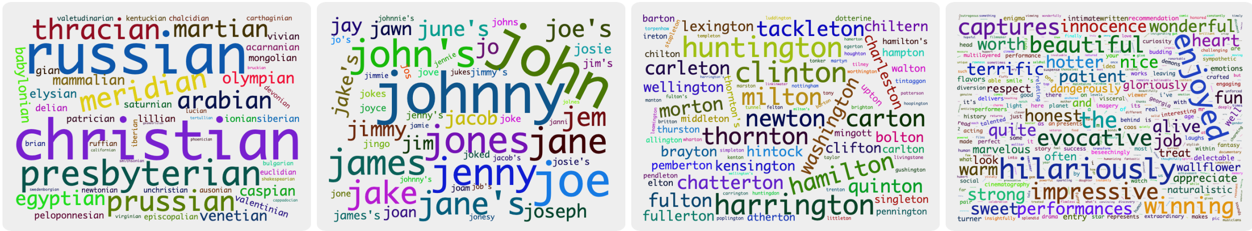
2. Methodology

Our methodology is designed to uncover and compare the latent conceptual structures emerging within speech, text, and multimodal foundation models. To achieve this, we employ Latent Concept Analysis (LCA) [8], an unsupervised approach that enables the discovery and interpretation of abstract representations learned by neural networks. Our approach consists of two stages: **concept discovery** and **concept alignment**.

2.1. Concept Discovery

Contextualized representations learned in foundation models capture latent conceptual structures that can be interpreted through clustering methods. Our investigation expands upon the work done in discovering latent ontologies in contextualized representations [9]. We extract contextualized representations from unimodal models and multimodal models, each with L layers $(l_1, l_2, \ldots, l_L)$, and cluster them to obtain encoded concepts. In this context, a concept refers to a set of linguistic units such as words, phonemes, or acoustic patterns, grouped based on lexical, semantic, syntactic, morphological, or phonetic relationships. Figure 1 showcases concepts within the latent space of the different models, wherein word representations are ar-

[1] https://github.com/shammur/MultimodalXplain
[2] https://huggingface.co/collections/QCRI/multimodalxplain-6839bbe6fc98a0b221dc42bb



(a) *Nationalities and Ethnicities* (b) *Names Starting with /dʒ/* (c) *NEs ending in the /tən/ sound* (d) *Positive Polarity Concept*

Figure 1: *Sample latent concepts from different models. Figures 1a & 1d show BERT concepts; 1b & 1c show HuBERT concepts.*

ranged based on distinct linguistic and task-specific concepts.

Textual Input. Given a textual utterance $\mathcal{U} = [w_1, \ldots, w_N]$, we extract contextual embeddings at layer l: $\mathcal{U} \xrightarrow{\mathrm{M}_t^l} \mathbf{\Phi}^l = [\phi_1^l, \ldots, \phi_N^l]$ where ϕ_i^l is the embedding of w_i at layer l.

Speech Input. For the corresponding speech utterance $\mathcal{X} = [x_1, x_2, \ldots, x_T]$, consisting of T frames, we extract frame-level representations $\mathbf{\Psi}^l$ at each layer. Since our analysis focuses on word-based concepts, we derive word-level representations[3] $\mathbf{\Psi}_w$ by averaging frame embeddings within the word boundary $[t_{\text{start}}, t_{\text{end}}]$ [10]. Word boundaries are obtained using the Montreal Forced Aligner.[4]

2.2. Concept Alignment

Encoded concepts capture latent relationships among words within a cluster, encompassing phonetic, lexical, syntactic, semantic, and task- or modality-specific patterns. To systematically interpret these concepts, we employ an alignment metric proposed by [11], which maps the discovered concepts to structured linguistic ontologies using predefined taxonomies.

Let $\mathcal{C}_\mathcal{L} = \{C_{l_1}, C_{l_2}, \ldots, C_{l_n}\}$ be the set of linguistic concepts (e.g., parts-of-speech tags of words), and $\mathcal{C}_\mathcal{E} = \{C_{e_1}, C_{e_2}, \ldots, C_{e_m}\}$ be the set of encoded concepts discovered within neural language models. We define their θ-alignment as follows:

$$\lambda_\theta(\mathcal{E}, \mathcal{L}) = \frac{1}{2}\left(\frac{\sum_\mathcal{E} \alpha_\theta(C_e)}{|\mathcal{C}_\mathcal{E}|} + \frac{\sum_\mathcal{H} \kappa_\theta(C_l)}{|\mathcal{C}_\mathcal{L}|}\right) \times 100$$

where alignment $\alpha_\theta(C_e)$ and coverage $\kappa_\theta(C_l)$ are defined as

$$\alpha_\theta(C_e) = \begin{cases} 1, & \text{if } \exists C_l \in \mathcal{C}_\mathcal{L} \text{ such that } \frac{|C_e \cap C_l|}{|C_e|} \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

$$\kappa_\theta(C_l) = \begin{cases} 1, & \text{if } \exists C_e \in \mathcal{C}_\mathcal{E} \text{ such that } \frac{|C_e \cap C_l|}{|C_e|} \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

The alignment term measures how many discovered concepts align with the categories of the underlying taxonomy, while the coverage term assesses how many linguistic concepts from a given taxonomy appear in the discovered clusters.

3. Experimental Setup

Models. We investigate both unimodal and multimodal models, focusing on HuBERT, BERT, Seamless M4T, and SpeechT5. HuBERT [12] is a self-supervised speech model that excels at learning speech representations through masked prediction of quantized acoustic features. BERT [13] is a widely used pre-trained text model that captures rich contextual information in language through bidirectional transformers. We explore two multimodal models: Seamless M4T [14], which combines speech and text in a shared decoder, and SpeechT5 [15], an extension of T5 that jointly learns speech and text for tasks like speech-to-text and text-to-speech. These models represent a spectrum of approaches to analyze representations learned across different modalities.

Tasks. We conducted experiments using traditional taxonomies designed to capture core linguistic concepts. These include word morphology, represented by part-of-speech tagging [16]; syntax, explored through chunking tagging [17]; and semantics, examined through Parallel Meaning Bank annotations [18]. We trained sequence taggers for each of these tasks and annotated the corresponding training data. Each core linguistic task represents a human-defined concept, which we align with encoded representations to assess how linguistic knowledge is structured in the model's latent space. We also used the shallow lexical concept such as suffixation.

We also compared encoded concepts with task-specific concepts in sentiment analysis [19] using the alignment function to measure affinity. For the *positive sentiment concept*, we define $C_{sst}(+ve)$ as the set of words appearing exclusively in positively labeled sentences. An encoded concept C_e is considered aligned with $C_{sst}(+ve)$ if a threshold θ of its words also appear in positive sentences, with each word represented by its contextualized embedding. The same process applies to $C_{sst}(-ve)$ to identify negative polarity concepts.

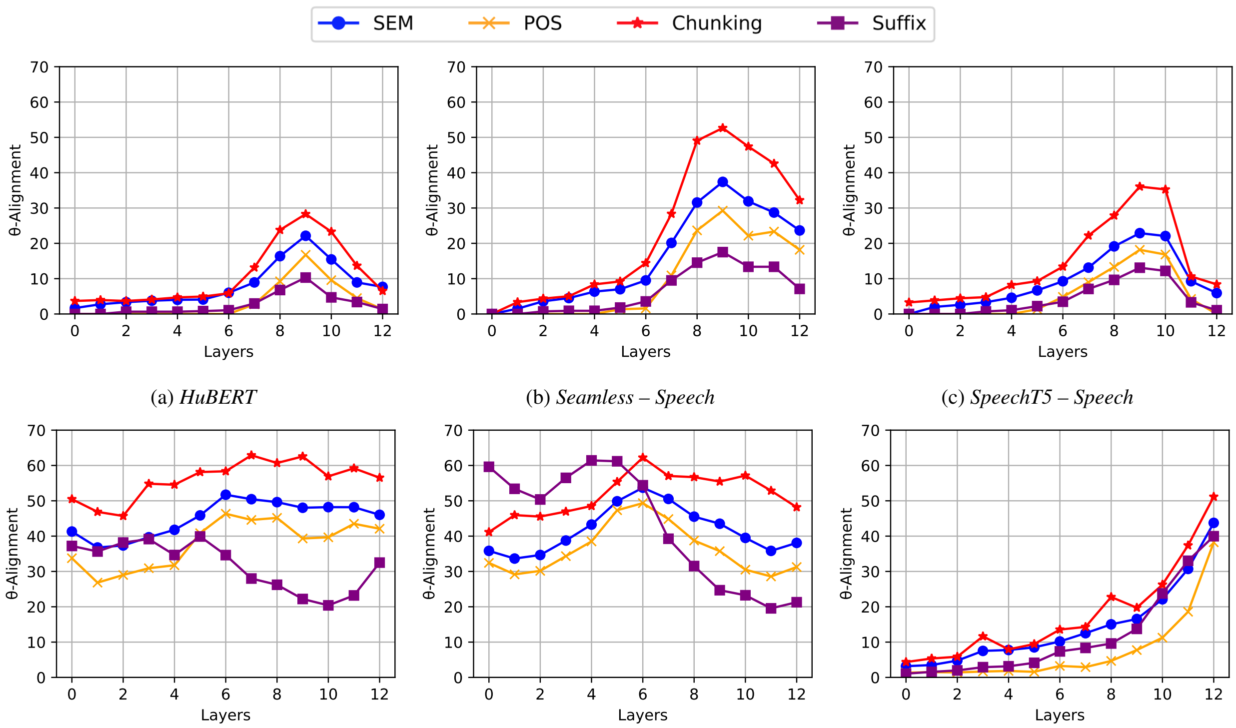
Data. We used two datasets for concept analysis: LibriSpeech [20], a large-scale read-speech corpus with diverse speakers, and Stanford Sentiment Treebankv2 (SST2) [19], a text-only sentiment classification dataset. **We extended SST2 into the speech domain (SST2-audio) using XTTSv2**,[5] a controllable TTS model, to generate high-quality audio. SST2-audio preserves sentiment polarity while minimizing speaker and environmental variability, enabling analysis of semantic concepts in speech models.

Concept Discovery and Annotation. We perform a forward pass through the models to generate contextualized feature vectors using NeuroX toolkit [21]. Subsequently, we apply K-means clustering to the feature vectors, yielding K clusters (also referred to as encoded concepts) for both base and fine-tuned models. We set K = 600 and filter out representations that appear at least 10 times, following the settings prescribed by [11]. We consider an encoded concept to be aligned with the linguistic concept, if it has at least 90% ($\theta = 0.9$) match in the number of words.

[3] Also known as the acoustic word embeddings [10]

[4] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[5] https://docs.coqui.ai/en/latest/models/xtts.html



(a) *HuBERT* (b) *Seamless – Speech* (c) *SpeechT5 – Speech*

(d) *BERT* (e) *Seamless – Text* (f) *SpeechT5 – Text*

Figure 2: *Layer-wise concept alignment. Y-axis shows % of aligned concepts (using θ=0.9) per encoder layer (X-axis).*

4. Findings and Analysis

4.1. Comparing Modalities

In Figure 2, we illustrate how concepts learned by text, speech, and multimodal models align with the linguistic taxonomies studied in this paper. The alignment patterns across layers reveal distinct processing strategies: speech and text models handle linguistic information differently. Specifically, speech models do not encode word-based linguistic taxonomies in early layers; instead, linguistic structures emerge in the middle layers and peak in the upper layers. This trend is consistent with SpeechT5, which has been jointly trained with text.

Unlike speech models, which gradually transition from acoustic to linguistic representations, text models operate directly on tokenized inputs, allowing them to encode linguistic taxonomies from the initial layers. Text models capture subword-level information such as suffixation early on, while morphology (POS), syntax (chunking), and semantics (SEM) develop progressively, peaking in the middle layers. This pattern suggests that text models incrementally build linguistic understanding, with deeper layers focusing more on integrating syntax and semantics.

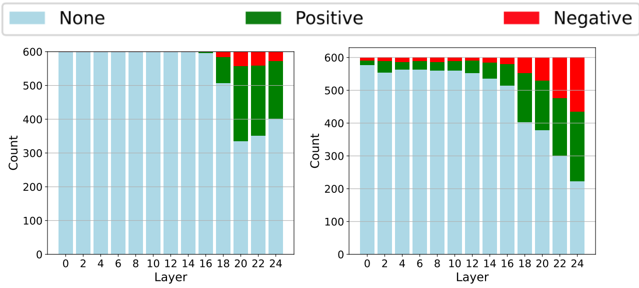
Both speech and text models exhibit a falling pattern in alignment in the upper most layers. This decline likely reflects the increasing abstraction of learned representations. In speech models, this suggests a shift from phonetic and word-level patterns to more holistic cues such as paralingusitic representations. In text models, the decrease in alignment indicates a transition toward contextual abstraction and task-specific reasoning, where representations become more specialized for downstream tasks rather than directly reflecting linguistic taxonomies.

While SpeechT5-Speech follows the observed trend in speech models, SpeechT5-Text exhibits a different alignment pattern. Unlike BERT and Seamless-text, which refine linguistic representations throughout the network, SpeechT5's shared encoder allocates less capacity to explicit linguistic taxonomies in its deeper layers. This is due to its multimodal training objective, which aligns speech and text representations for speech-to-text and text-to-speech tasks. Unlike BERT, optimized for masked language modeling and hierarchical linguistic abstraction, and Seamless, where text and speech do not share the latent space, SpeechT5's encoder is optimized for cross-modal consistency, leading to representations that do not follow typical patterns observed in other models.

Overall, our findings suggest that speech models allocate less capacity to learning linguistic and semantic taxonomies than text models, likely because they must also account for speech-specific features like phonetics, prosody, and speaker variability, which consume much of the network's representational capacity. As a result, linguistic structures emerge later in speech models and remain less prominent throughout their layers. We speculate that this constraint limits the ability of speech models to encode higher-level conceptual abstractions as effectively as text models, which operate directly on symbolic representations and can dedicate more capacity to linguistic and semantic processing. This distinction could explain why speech models often struggle with tasks requiring deep linguistic reasoning or structured understanding compared to their text-based counterparts [22, 23].

4.2. Task-specific Concepts

In the previous section, we hypothesized that the observed decrease in alignment in the final layers reflects a shift toward task-specific reasoning, where the model's representations become more specialized for downstream tasks. To test this hypothesis, we compared the BERT and HuBERT large mod-



(a) *HuBERT–SST* (b) *BERT–SST*

Figure 3: *Alignment with task-specific polarity concepts*

els trained for sentiment classification. Using SST2-text and our SST2-audio, we extracted activation vectors and underlying concepts from the models and aligned them with the output classes: positive and negative. Our results, presented in Figure 3, indicate that polarity concepts begin to emerge in the final layers of both the fine-tuned text and speech models.

A closer analysis reveals a notable asymmetry between text and speech-based models in how they encode sentiment. Specifically, the speech-based SST models predominantly rely on signals from positive polarity concepts, whereas textual models exhibit a more balanced distribution of positive and negative concepts. This suggests that speech models may struggle to capture negative sentiment as effectively as textual models, potentially due to the inherent differences in how sentiment is conveyed in speech versus text. In text, explicit negations and sentiment-laden words provide clear cues for polarity, whereas speech-based sentiment relies more on prosodic features such as intonation, pitch, among others which may be harder to disentangle in the absence of large-scale speech-specific supervision.

This observation aligns with task performance: HuBERT underperformed in predicting negative sentiment (accuracy of 87.48% versus BERT's 93.21%), while both models performed comparably when classifying positive sentiment (93.31% versus 94.98%). The disparity in negative sentiment prediction highlights a potential limitation of current speech-based sentiment classifiers, which may require additional fine-tuning or explicit modeling of prosodic features to achieve performance parity with text-based models. Overall, our results reinforce the idea that final layer representations are shaped by task-specific requirements [24, 25], but they also highlight potential gaps in how different modalities encode sentiment. Addressing these gaps could involve integrating additional linguistic or prosodic cues in speech-based models, leveraging multi-modal learning strategies, or exploring alternative architectures that better capture the nuances of spoken sentiment.

4.3. Qualitative Analysis

In our qualitative analysis, we explored how the models learn and represent concepts by leveraging GPT for annotation. While we identified concepts that align with established linguistic taxonomies, it's important to note that these models may not always strictly conform to human-defined concepts. To interpret and analyze these concepts more effectively, we can annotate them for further exploration. Following [26], we employ ChatGPT in a zero-shot setting, prompting the model with structured instructions, reported in Listing 1.

Our findings suggest that concepts in the models are often organized in compositional hierarchies, where the model initially groups concepts based on a primary objective and then refines them according to semantic relations. For example, the model may first group all names starting with the sound dʒ before distinguishing between other linguistic or semantic categories (see the concept in Figure 1b). This hierarchical structure highlights the model's ability to form complex, layered associations among concepts, which may offer insights into how it processes and organizes knowledge.

```
Assistant is a large language model trained by
OpenAI.
Instructions:
Give a short and concise label that best
describes the following list of words:
["w_1", "w_2", ..., "w_N"]
```

Listing 1: *Prompt for label assignment.*

5. Related Work

The discovery and interpretation of latent concepts in deep models remain crucial challenges in NLP, particularly in speech processing. Recent studies have focused on understanding the internal representations learned by these models, with an emphasis on layer-wise analysis and latent concept discovery [27, 28, 9, 29, 30]. These foundational works have explored how hierarchical linguistic features–such as surface properties, syntax, and semantics are distributed across layers. The findings suggest a progression from lexical and syntactic features in the lower and middle layers to more abstract semantic representations in the higher layers.

There are a handful of studies on how information is encoded in speech models [31, 32, 33, 34, 35]; however, latent concept discovery and the evolution of representations across layers remain largely underexplored compared to text-based models. Studies such as [36, 37, 38] indicate that phonetic and phonemic distinctions emerge in the early layers, often overshadowing semantic information in speech models [39]. Previous analyses show a strong alignment between HuBERT's latent units and linguistic structures, particularly phonetic categories [38], and syllabic [37]. Additionally, studies like [40] demonstrate that model size and training objectives significantly influence the distribution of linguistic information. Despite these insights, in-depth research on speech, multimodal, and encoder-decoder models remains underexplored. To address these gaps, our study focuses on understanding how semantic concepts emerge in deep speech models.

6. Conclusion

In this study, we compared speech, text, and multimodal models to understand how they represent linguistic concepts. Our findings suggest that text models, such as BERT, directly encode linguistic structures from early layers, while speech models, like HuBERT, gradually develop linguistic representations from acoustic features. Multimodal models like SpeechT5 exhibit unique alignment patterns due to cross-modal training. We observed that speech models allocate less capacity to linguistic taxonomies, focusing more on speech-specific features like phonetics. In task-specific tests, such as sentiment analysis, speech models showed weaker performance in capturing negative sentiment compared to text models. Our results emphasize the different ways these models process and internalize language, with text models offering richer, more structured linguistic representations.

7. References

[1] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of artificial general intelligence: Early experiments with GPT-4," Tech. Rep., 2023.

[2] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv:2302.13971*, 2023.

[4] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[5] J. Zhao, Q. Yang, Y. Peng, D. Bai, S. Yao, B. Sun, X. Chen, S. Fu, X. Wei, L. Bo *et al.*, "Humanomni: A large vision-speech language model for human-centric video understanding," *arXiv preprint arXiv:2501.15111*, 2025.

[6] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, J. Ricke, and M. Ingrisch, "Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports," 2022.

[7] K. E. Patterson, P. J. Nestor, and T. T. Rogers, "Where do you know what you know? the representation of semantic knowledge in the human brain," *Nature Reviews Neuroscience*, 2007.

[8] F. Dalvi, A. R. Khan, F. Alam, N. Durrani, J. Xu, and H. Sajjad, "Discovering latent concepts learned in BERT," in *Proc. of ICLR*, 2022.

[9] J. Michael, J. A. Botha, and I. Tenney, "Asking without telling: Exploring latent ontologies in contextual representations," in *Proc. of EMNLP*, 2020.

[10] R. Sanabria, O. Klejch, H. Tang, and S. Goldwater, "Acoustic word embeddings for untranscribed target languages with continued pretraining and learned pooling," in *Proc. of Interspeech*, 2023.

[11] M. Hawasly, F. Dalvi, and N. Durrani, "Scaling up discovery of latent concepts in deep NLP models," in *Proc. of EACL*, 2024.

[12] 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," 2021.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL*, 2019.

[14] S. Communication, "Seamlessm4t: Massively multilingual & multimodal machine translation," 2023.

[15] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," 2022.

[16] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Computational Linguistics*, 1993.

[17] E. F. Tjong Kim Sang and S. Buchholz, "Introduction to the CoNLL-2000 shared task chunking," in *Proc. of CoNLL and the Second Learning Language in Logic Workshop*, 2000.

[18] L. Abzianidze, J. Bjerva, K. Evang, H. Haagsma, R. van Noord, P. Ludmann, D.-D. Nguyen, and J. Bos, "The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations," in *Proc. of EACL*, 2017.

[19] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. of ICASSP*, 2015.

[21] F. Dalvi, A. Nortonsmith, A. Bau, Y. Belinkov, H. Sajjad, N. Durrani, and J. Glass, "Neurox: A toolkit for analyzing individual neurons in neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 9851–9852, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5063

[22] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. of Blackbox NLP*, 2018.

[23] S. Shon, A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu, and K. J. Han, "Slue: New benchmark tasks for spoken language understanding evaluation on natural speech," 2022.

[24] A. Merchant, E. Rahimtoroghi, E. Pavlick, and I. Tenney, "What happens to BERT embeddings during fine-tuning?" in *Proc of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Nov. 2020.

[25] N. Durrani, F. Dalvi, and H. Sajjad, "Discovering salient neurons in deep nlp models," *Journal of Machine Learning Research*, vol. 24, no. 362, pp. 1–40, 2023.

[26] B. Mousi, N. Durrani, and F. Dalvi, "Can llms facilitate interpretation of pre-trained language models?" in *Proc. of EMNLP*, 2023.

[27] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proc. of ACL*, 2019.

[28] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Proc. of ACL*, 2019.

[29] H. Sajjad, N. Durrani, F. Dalvi, F. Alam, A. Khan, and J. Xu, "Analyzing encoded concepts in transformer language models," in *Proc. of NAACL*, 2022.

[30] N. Durrani, H. Sajjad, F. Dalvi, and F. Alam, "On the transformation of latent space in fine-tuned nlp models," in *Proc. of EMNLP*, 2022.

[31] S. A. Chowdhury, N. Durrani, and A. Ali, "What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis," *Computer Speech & Language*, 2024.

[32] A. Waheed, H. Atwany, B. Raj, and R. Singh, "What do speech foundation models not learn about speech?" *arXiv preprint arXiv:2410.12948*, 2024.

[33] Y. El Kheir, A. Ali, and S. A. Chowdhury, "Speech representation analysis based on inter-and intra-model similarities," in *Proc. of ICASSP Workshop*, 2024.

[34] H. Lee, D. Liu, S. Sinhamahapatra, and J. Niehues, "How do multimodal foundation models encode text and speech? an analysis of cross-lingual and cross-modal representations," *arXiv preprint arXiv:2411.17666*, 2024.

[35] M. de Heer Kloots and W. Zuidema, "Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in wav2vec2. 0," in *Proc. INTERSPEECH*, 2024.

[36] K. Martin, J. Gauthier, C. Breiss, and R. Levy, "Probing self-supervised speech models for phonetic and phonemic information: A case study in aspiration," in *Proc. Interspeech*, 2023.

[37] C. J. Cho, A. Mohamed, S.-W. Li, A. W. Black, and G. K. Anumanchipalli, "SD-HuBERT: Sentence-level self-distillation induces syllabic organization in hubert," in *Proc. of ICASSP*, 2024.

[38] D. Wells, H. Tang, and K. Richmond, "Phonetic analysis of self-supervised representations of english speech," in *Proc. of Interspeech*, 2022.

[39] K. Choi, A. Pasad, T. Nakamura, S. Fukayama, K. Livescu, and S. Watanabe, "Self-supervised speech representations are more phonetic than semantic," *arXiv preprint arXiv:2406.08619*, 2024.

[40] A. Pasad, C.-M. Chien, S. Settle, and K. Livescu, "What do self-supervised speech models know about words?" *TACL*, 2024.