Concerns in evaluating hierarchical correlations between human brains and end-to-end automatic speech recognition models

Yi Wang (Wang.Yi@ed.ac.uk)
School of Informatics, University of Edinburgh

Peter Bell (Peter.Bell@ed.ac.uk)
School of Informatics, University of Edinburgh

Abstract

Recent neural encoding studies have attempted to compare the human brain speech perception network with artificial neural network models trained end-to-end (e2e) on automatic speech recognition (ASR), aiming to reveal the temporal dynamics of human speech processing. Multiple studies have reported a prominent correspondence between e2e ASR models and human brains in terms of the hierarchical encoding of linguistic features, from low-level acoustic features to high-level semantic features. While different types of e2e ASR models have been used to investigate this correspondence, there has not been a consensus on the most suitable ASR model type for such investigations. This extended abstract will discuss concerns regarding the use of three mainstream types of e2e ASR models when evaluating their hierarchical correlation with human speech perception network, including the recurrent neural network transducer, the attention-based encoder-decoder model using tokenizer (i.e. Whisper) and the self-supervised transformer model. We suggest that further caution is required when using these models in the hierarchical correlation studies, due to issues such as varying decoding latency, mismatched context window and difficulty in representation disentanglement inherent in each model type, respectively.

Keywords: speech perception; automatic speech recognition; neural coding; linguistic representations

Introduction

Recently developed automatic speech recognition (ASR) models based on artificial neural networks (ANN) and the back-propagation algorithm have achieved human parity for a few languages (Xiong et al., 2016). Such models could potentially serve as computational models for speech perception, considering that their computational weights are accessible for analyzing and manipulating. Studies (Li et al., 2023; Keshishian et al., 2025) have compared the neural activity in the speech processing pathway of human listeners with the layer representations of deep neural network (DNN) ASR models, and reported that the hierarchical information in ASR models prominently correlates with that in the ascending auditory pathway (Li et al., 2023). Keshishian et al. (2025) further reported that the temporal order of hierarchical encoding in an ASR model is similar to that in human brains.

Neural encoding studies have consistently chosen end-to-end (e2e) DNN ASR models for comparisons with human brains. This preference is likely due to the more unified internal structure of e2e systems than conventional modular systems, which better resembles the consistent biological structure of the human cortex. The major types of modern e2e ASR models under investigation (Li et al., 2023; Keshishian et al., 2025) include recurrent neural network transducer (RNN-T) models (Graves, 2012), and attention-based self-supervised (SSL) pre-trained model (Baevski, Zhou, Mohamed, & Auli, 2020; Hsu et al., 2021), etc. Additionally, other state-of-the-art e2e ASR model types, such as the attention-based encoder-decoder (AED) model (Dong, Xu, & Xu, 2018; Radford et al., 2023) hold potential for future investigation.

This paper connects with previous studies (Mahadeokar et al., 2021; Wagner, Thallinger, & Zusag, 2024; Pasad, Shi, & Livescu, 2023) that analyzed the characteristics of mainstream e2e ASR models. Our aim is to identify concerns in evaluating the hierarchical correlations between the human brain regions and e2e ASR model layers. Drawing attention to these concerns helps refine the modeling of encoding latency in such evaluations, particularly regarding context window estimation for both ASR representations and brain activities. We will discuss three specific concerns in such evaluations, each corresponding to a type of e2e ASR model.

- Noticeable and varying time lags in the prediction branch (i.e. decoder layer) of the RNN-T model (Mahadeokar et al., 2021), could lead to difficulties in modeling the speech and language neural encoding time lags in brains if using such representations as a reference;
- Unsupervised tokenization in supervised AED models: Models like Whisper (Radford et al., 2023), which use unsupervised tokenizers (Sennrich, Haddow, & Birch, 2015), generate tokens that can be temporally misaligned with fundamental human speech units (such as phonemes, syllables, or words) (Wagner et al., 2024). This misalignment can lead to significant inaccuracies in context window segmentation during correlation evaluations.
- Gradual evolving of encoding acoustic, phonetic, and word-level properties (Pasad et al., 2023) in representations of SSL model layers (Baevski et al., 2020; Hsu et al., 2021). This presents a challenge when deciding which SSL layer's representation is most appropriate for modeling specific auditory cortex brain region.

Concern case one: RNN-T

Keshishian et al. (2025) used a smart way to evaluate how the RNN-T ASR model maps to the human speech processing pathway, by fitting a single-lag regression model that predicts the brain neural activity from the layer activations in response to speech stimuli. Such regression models used a constant value to model the time lag of neural responses and that of ASR activations for each electrode and layer pair.

However, such an assumption on a relatively static time lag might not be valid for both the brain speech processing and the RNN-T model. The latency of speech neural encoding highly depends on the semantic contents and linguistic properties of the stimuli (Ding, Melloni, Zhang, Tian, & Poeppel, 2016; Yasmin, Irsik, Johnsrude, & Herrmann, 2023). The RNN-T models have shown that its time lag in the prediction branch (i.e. decoder), reflected by token (spoken word) emission delay, can be a large and varying duration compared to the length of the spoken word itself (Mahadeokar et al., 2021). Therefore, a regression model with more dynamic modeling on speech processing latency could be applied in the correlation evaluations.

Meanwhile, the RNN-T's architecture, whilst monotonic, allows the model to delay its output, pending future acoustic or lexical context to be processed before the incoming next output token predicted. The long token emission delay of RNN-T breaks the causal relationship in speech modeling. Consequently, the context window influencing the RNN-T decoder representation for a specific token (i.e. spoken word) might not be centered at the corresponding word center. When comparing RNN-T representations to the neural activity, the time window to select highly-correlated neural activity may require including future context.

On the other hand, Keshishian et al. (2025) chose RNN-T to map to the human speech perception network because they believed that RNN-T processes speech in a causal and incremental manner. Given that the RNN-T causality might have been broken, future works may as well experiment on variations of RNN-T that have restrictions on time alignment and token emission delays (Mahadeokar et al., 2021).

Concern case two: ASR model using tokenizer

Li et al. (2023) explored cross-lingual hierarchical correlation evaluation on English and Chinese. If further studies aiming to investigate lower-resource languages using pre-trained ASR models that achieve human-level performance, the options become narrower. Whisper (Radford et al., 2023), a noise-robust speech processing model, is fully-supervised trained on 680,000 hours of multilingual speech, incorporating supervision on the transcription and translation tasks. This makes it a potentially convenient tool for correlation studies. However, the tokenizer and training objective used by Whisper may introduce a consistent delay in context window that corresponds to each representation frame, particularly in its decoder.

Whisper used Byte pair encoding (BPE) text tokenizers (Sennrich et al., 2015) during both training and inference, trained in an unsupervised manner on the transcription texts. Wagner et al. (2024) found that many tokens learned by BPE are prefixed with a space, and only about 13% of spaces are separated from the following tokens. This results in the token-level context window advancing ahead of the corresponding phoneme or word-level context window. Additionally, Whisper does not provide word-level timestamps. Determining the context window of high-level linguistic representations requires one of two ways. First, applying an additional force-alignment step to obtain the start and end time stamp of each spoken unit (phone or word). Second, inferring alignment from the cross-attention weights using Dynamic Time Warping (DTW) (Giorgino, 2009), which maps the context window of decoder representations back to the encoder representations that are more tightly aligned with the speech signal.

Using the first approach may lead to the pitfall of equating the phoneme or word time window with the ASR representation time window, without accounting for silence. Setting the start and end time with considerations on the space in the token and including the silence duration accordingly could reduce misalignment in evaluation. When comparing different ASR models in correlation studies, one potential solution is to use CrisperWhisper (Wagner et al., 2024), an extension of Whisper that is further trained on retokenized annotations where spaces are separated from word tokens, keeping the context window more closely with those of other ASR models.

Concern case three: SSL models

Li et al. (2023) modeled neural activities in the human auditory pathway with two SSL models, Wav2Vec 2 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), also by fitting ridge regression models to predict neural activity from different brain regions. By comparing normalized regression coefficients, they reported the SSL model hierarchy correlates with the ascending auditory pathway and selected the best layer predicting the neural activities in each pathway component.

However, according to the SSL representations analysis (Pasad et al., 2023) involving Wav2Vec 2 and HuBERT, the linguistic properties encoded in SSL models evolve very gradually through layers, reflected by plateau shapes in the layer-level canonical correlation analysis similarity (Hotelling, 1992) curves comparing the representations to the local spectrogram feature, the phone labels, and the word labels. This reduces the precision of using single-layer representations as a reference to dissect the functions of different brain regions. Such a challenge has been reflected in the analysis of intracranial cortical recordings from cortex components in the speech processing pathway (Li et al., 2023). Taking HuBERT as an example, 13 out of 14 layers were not statistically different in brain-prediction score from the best layer predicting the Heschl gyrus (HG) activities, 8 out of 14 for the superior temporal gyrus (STG). SSL representations that are more modularized and capable of separately encoding hierarchical linguistic representations may need to be developed to refine this kind of speech processing neural activity modeling.

Conclusion

We identified three concerns when examining the hierarchical correspondence between e2e ASR models and human brain activity: (1) the varying emission lag that disrupts strict causality in RNN-T, (2) the context window mismatch in models using token-level targets such as Whisper, and (3) the gradual layer-level evolution of representations in SSL models. Careful attention to time alignment, token segmentation, and SSL model representations may enhance the reliability of correlation studies.

Acknowledgements

This work was supported by the UKRI Centre for Doctoral Training (CDT) in Natural Language Processing, funded by the UKRI grant EP/S022481/1 and the University of Edinburgh.

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, *19*(1), 158–164.

Dong, L., Xu, S., & Xu, B. (2018). Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5884–5888).

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, *31*, 1–24.

Graves, A. (2012). Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution* (pp. 162–190). Springer.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460.

Keshishian, M., Mischler, G., Thomas, S., Kingsbury, B., Bickel, S., Mehta, A. D., & Mesgarani, N. (2025). Parallel hierarchical encoding of linguistic representations in the human auditory cortex and recurrent automatic speech recognition systems. *bioRxiv*, 2025–01.

Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, J., ... Chang, E. F. (2023). Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, *26*(12), 2213–2225.

Mahadeokar, J., Shangguan, Y., Le, D., Keren, G., Su, H., Le, T., ... Seltzer, M. L. (2021). Alignment restricted streaming recurrent neural network transducer. In *2021 ieee spoken language technology workshop (slt)* (pp. 52–59).

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. No. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Pasad, A., Shi, B., & Livescu, K. (2023). Comparative layer-wise analysis of self-supervised speech models. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Wagner, L., Thallinger, B., & Zusag, M. (2024). Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. *arXiv preprint arXiv:2408.16589*.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.

Yasmin, S., Irsik, V. C., Johnsrude, I. S., & Herrmann, B. (2023). The effects of speech masking on neural tracking of acoustic and semantic features of natural speech. *Neuropsychologia*, *186*, 108584.