Transformer-based Model Captures Neural Representation of Audio-visual Speech in Natural Scenes

Wenyuan Yu, Zhoujian Sun, Yunyi Qi, and Cheng Luo*
({wenyuanyu, sunzj, qiyunyi, luo_cheng}@zhejianglab.com)

Zhejiang Lab
Hangzhou, Zhejiang, 311121 China

Abstract:

Natural scenes comprise both visual and auditory information, which are integrated to form a coherent perception. The multi-modal encoding of the auditory and visual information has been extensively investigated in biological brains and various advanced deep neural networks (DNNs) respectively. However, little is known about the underlying relationship between information representations in biological brains and DNN models. In the current study, we investigate whether humans and DNNs represent the auditory and visual information in a comparable way during audio-visual speech recognition (AVSR). For humans, we used electroencephalography (EEG) recordings to analyze neural activity of auditory and visual features when participants engaged in a speech recognition task within audio-visual scenes. For DNNs, we analyzed hidden layers' embeddings from a transformer-based model, i.e., AV-HuBERT, which achieves state-of-the-art performance in AVSR tasks. We observed significant representational similarity between the EEG responses and model embeddings. Further analysis revealed that the model embeddings from lower hidden layers exhibited the greater similarity with the neural encoding of visual and auditory features. These results suggest that DNNs can naturally evolve human-like information representations, and their hidden layers' embeddings effectively capture auditory and visual patterns in neural representations of humans.

Keywords: EEG; AVSR; information representation

Introduction

The brain continuously receives and integrates these sensory inputs, forming a coherent perception of the external world (Senkowski et al., 2008). Nevertheless, the linear model seems insufficient to capture the neural encoding of multi-modal information in complex brain networks. Recently, transformer-based deep neural network (DDN) models have shown promising results in multi-modality tasks (Vaswani et al., 2017; Devlin et al., 2018; Lan et al., 2019). These models encode information input as contextual vector representations known as model embeddings. These embeddings effectively capture the structure and semantics of data, enabling these models to achieve human-level performance in tasks such as natural language processing (NLP), computer vision (CV), and automatic speech recognition (ASR). Given that the artificial neural networks are inspired by the biophysical properties and cognitive functions of the brain, it is a key area of research to explore the underlying relationship of information representations between biological brains and DNN models (Cox et al., 2014; Li et al., 2023). In the current study, we focus on transformer-based DNN models for audio-visual speech recognition (AVSR) tasks, and investigate whether the model embeddings reflect the neural representations of visual and auditory features as observed in human brains.

The present study aims to reveal the shared information representations between human brains and transformer-based DNN models. Such comparison is challenging due to the fundamentally divergent encoding methods used by biological and artificial neural networks. To address this, we employed representation similarity analysis (RSA), a multivariate technique that compares the similarities between data types based on the shared structure of their distance matrices (Kriegeskorte et al., 2008). The study makes two primary contributions. First, it introduces an effective method for comparing information representations between human brains and DNN models. Second, it provides direct evidence of a correlation between neural responses and model embeddings during audio-visual speech recognition tasks.

Results

Ten English native speakers were recruited for the experiment. They were asked to listen to sixteen speech segments, each approximately one minute long, with background noise in natural scenes (i.e., gym and road). While listening, participants simultaneously viewed corresponding videos of the natural scenes. Their EEG signals were recorded by 64-channel Biosemi ActiveTwo system. The audio and video stimuli employed in the EEG experiment were processed through a Transformer-based model, AV-HuBERT, with state-of-the-art performance in AVSR tasks (Shi et al., 2022). We extracted model embeddings from the eleven hidden layers of the AV-HUBERT model.

* Corresponding author: Cheng Luo

Spreading Representational similarity across channels and layers

We conducted a representational similarity analysis (RSA) on the multivariate neural activity at the channel-level. Specifically, we calculated the channel-level neural representation dissimilarity matrix (RDM) averaged across participants, and the model RDM for each layer. We then computed the representational similarity between the neural and model RDM for each channel. The results revealed consistent topographic patterns of representational similarity across layers. Moreover, this similarity was notably more pronounced in the central and occipital electrodes (Fig. 1).

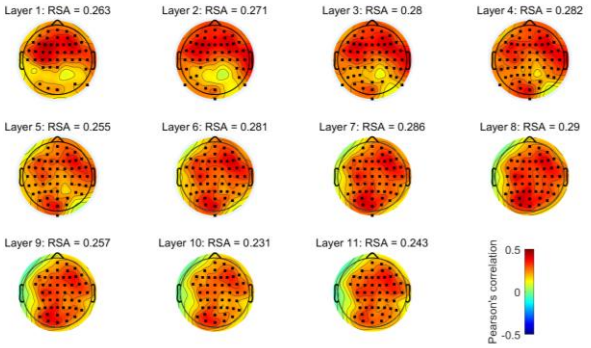

Fig 1. Representational similarity between EEG responses and model embeddings in each hidden layer. Black dots denote the channels where the representational similarity is significant (p< 0.05, FDR-corrected).

Model embeddings in the lower layers reflect the neural encoding of sensory features

The pronounced representational similarity observed in the central and occipital electrodes suggests that the model embeddings capture neural representations of auditory and visual features. To identify which layers' model embeddings represent low-level auditory and visual features, we used the temporal response function (TRF) method (Ding & Simon, 2012) to remove the neural responses to these features. We then examined how such manipulation altered the representational similarity between neural activity and the model embeddings across different layers.

We extracted the sound envelope of speech and the optical flow of the video as low-level auditory and visual features, respectively. TRFs were derived from these features based on time-aligned EEG signals. We simulated the neural responses to these features by convolving the sound envelope and optical flow with their corresponding TRFs. Then, we subtracted the simulated neural responses from the original EEG responses. Finally, we conducted RSA on the EEG responses after excluding the neural response to the sound envelope and optical flow.

We found that the exclusion of neural responses to sensory features resulted in decreased representational similarity between the EEG responses and model embeddings across all model layers (Fig. 2). Notably, such exclusion induced more substantial changes in representational similarity patterns within the first three layers (Fig. 3).

Taken together, our findings revealed the representational similarity between EEG responses and model embeddings during speech processing in audio-visual natural scenes. In addition, model lower layers of the model primarily encode sensory features, thereby exhibiting greater similarity with the neural encoding of sensory features in the brain.

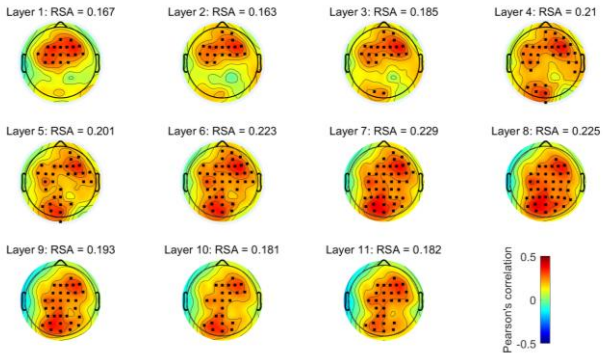

Fig 2. Representational similarity between EEG activity excluding the neural responses to sound envelope and optical flow, and model embedding in each hidden layer. Black dots denote the channel where the representational similarity is significant (p< 0.05, FDR-corrected).

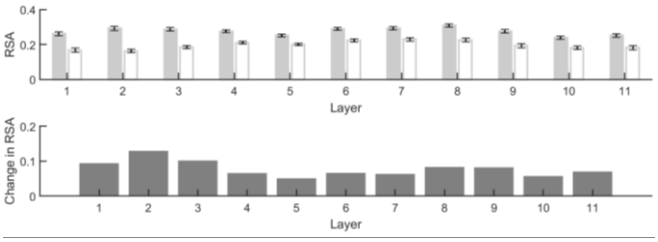

Fig 3. Top: Representational similarity averaged across channel before (gray bars) and after (white bars) excluding the neural responses to sound envelope and optical flow. The differences between each pair of bars are significant (p< 0.05, FDR-corrected). Error-bars denotes ±*SE*. Bottom: The difference between representational similarity before and after the exclusion of neural responses to sound envelope and optical flow.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2021ZD0201501).

References

Shi B., Hsu W. N., Lakhotia K., Mohamed A. (2022). Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv: 2201.02184.

Cox, D. D., & Dean, T. (2014). Neural Networks and Neuroscience-Inspired Computer Vision. Current Biology, vol. 24, no. 18, pp. R921-R929.

Devlin, M. W., Chang, K. L., Toutanova K. (2018). BERT: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

Ding, N., Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proceedings of the National Academy of Sciences of the United States of America, 109 (29), 11854–11859.

Kriegeskorte N., Mur M., Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 2: 249.

Lan, Z., Chen, M., Goodman S., Gimpel K., Sharma, P., Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. arxiv preprint arxiv:1909.11942.

Li Y., Anumanchipalli G. K., Mohamed A., et al. (2023). Dissecting neural computations in the human auditory pathway using deep neural networks for speech. Nature Neuroscience, 26(12): 2213-2225.

Senkowski, D., Schneider, T.R., Foxe, J.J., Engel, A.K. (2008). Crossmodal binding through neural coherence: Implications for multisensory processing. Trends Neuroscience 31 (8),401-409.

Vaswani A., Shazeer N., Parmar, N. et al. (2017). Attention Is All You Need. Advances in neural information processing systems, 30.