Neural Processing of Naturalistic Audiovisual Events in Space and Time

Yu Hu (yhu584@uwo.edu)
Western Institute for Neuroscience, Western University, 1151 Richmond Street
London, Ontario, Canada

Yalda Mohsenzadeh (ymohsenz@uwo.ca)
Department of Computer Science, Western University, 1151 Richmond Street
London, Ontario, Canada

Abstract

What we see and hear carry different physical properties, but our brain can integrate distinct information to form a coherent percept. However, when real-world audiovisual events are perceived, the specific brain regions and timings for processing and integrating different levels of information remain less investigated. To address that, we curated naturalistic videos and recorded fMRI and EEG data when participants viewed videos with accompanying sounds. We found the acoustic information was represented not only in auditory areas but also in early visual regions, suggesting the early cross-modal interaction and its role in combining acoustic features with visual features. However, the visual information was only represented in visual cortices, indicating that the early cross-modal interaction is asymmetrical. The visual and auditory features were processed with similar onset but different temporal dynamics. The high-level categorical and semantic information was identified in high-order and multi-modal areas and resolved later in time, demonstrating the late cross-modal integration and its distinct role in converging conceptual information. We further compared the neural representations with a two-branch deep neural network model and observed the mismatch in early cross-model interaction, suggesting the need for improvement to build a more biologically plausible model for audiovisual perception.

Keywords: Audiovisual Perception; Naturalistic Stimuli; Computational Models; Neural Representations and Dynamics

Introduction

Most visual scenes are associated with sounds. Thus, when we perceive them, two types of information are processed through different sensory channels and cerebral cortices, but eventually our brain is able to integrate the physically distinct information and create a coherent percept (Stein & Meredith, 1993; Ernst & Bülthoff, 2004). The cross-modal integration is observed in many brain regions including primary sensory areas and high-level cortical areas (Schroeder & Foxe, 2005; Ghazanfar & Schroeder, 2006). However, what functional roles each brain area plays during integration is still not well understood.

Many previous studies on audiovisual integration used simple stimuli like flash/tones (Shams, Kamitani, & Shimojo, 2000; Rohe & Noppeney, 2015; Cao, Summerfield, Park, Giordano, & Kayser, 2019) or image/sound pairs (Laurienti et al., 2003; Werner & Noppeney, 2010; Franzen, Delis, De Sousa, Kayser, & Philiastides, 2020), which are easy to manipulate the experimental conditions but lack ecological relevance. Therefore, the neural basis underlying the perception of real-world audiovisual events remains less investigated. To address that, we employed naturalistic video stimuli with sounds and aimed to investigate how different types of information are processed in the brain and across time with fMRI and EEG data as well as computational models.

Results

We curated 60 one-second videos with matching visuals and sounds for categories of animals, objects, and scenes. We recorded fMRI and EEG data separately while subjects (N=22) viewed the videos with accompanying sounds with an orthogonal oddball detection task to maintain their attention. Each video was presented 11 times for fMRI and 12-15 times for EEG.

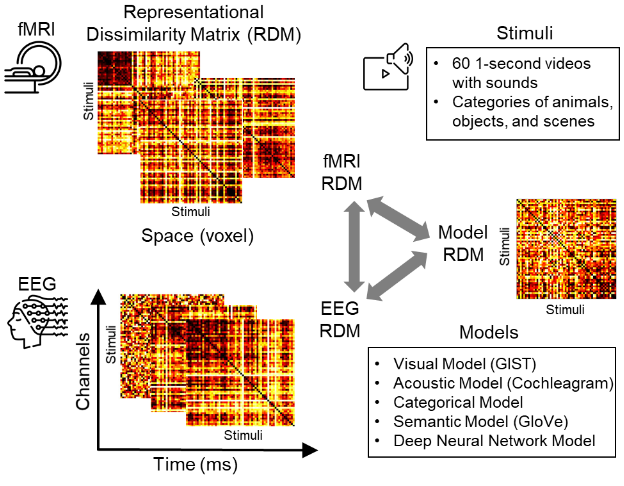

Figure 1: Experiment and data analysis scheme.

After data preprocessing, we extracted fMRI and EEG pattern responses and constructed representational dissimilarity matrices (RDMs) for an fMRI voxel searchlight or an EEG time point using Pearson correlation distance. We used different computational models to capture low-level visual and acoustic features (GIST descriptors (Oliva & Torralba, 2001) and Cochleagram model (Brown & Cooke, 1994)) and high-level categorical and semantic features (GloVe word embeddings (Pennington, Socher, & Manning, 2014)). We also tested a two-branch audio-video deep neural network (DNN) model (Morgado, Vasconcelos, & Misra, 2021) and evaluated its similarity to the brain.

Early asymmetrical cross-modal interaction in early visual cortex and late integration in high-level areas

We found that the low-level visual areas showed significant correlations with the low-level visual model representations and the strength of the correlation decreased along the visual hierarchy (Figure 2). The low-level acoustic model correlated best with neural representations in the early auditory cortex, but also correlated with representations in the early visual cortex. This suggests the early cross-modal interaction in which auditory information is integrated as early as in V1. However, such interaction is not bidirectional, with no visual information represented in auditory regions.

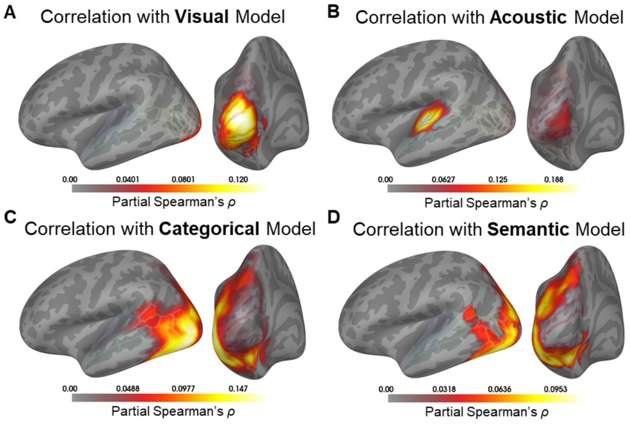

Figure 2: Subject-averaged significant partial correlation map between fMRI searchlight RDMs and model RDMs (1000 sign permutation test with cluster correction, cluster-definition $p < 0.001$, cluster $p < 0.01$).

As shown in Figure 3, the correspondence between EEG and visual model representations became significant at 55 ms and first peaked at 99 ms. The low-level acoustic feature representation emerged at 67 ms and first peaked at 92 ms, close to the first visual peak, suggesting that visual and acoustic information are processed almost simultaneously. The maximum peak for the auditory model was 193 ms, later than the maximum peak of the visual model at 133 ms, implying that the extraction of salient information from sounds may require more accumulated time than visual scenes.

The categorical and semantic information was mainly represented in high-order and multi-modal association areas, indicating their role in converging the high-level conceptual information. The categorical information emerged at 160 ms with a peak at 194 ms. The semantic representation was resolved with similar dynamics, with onset at 187 ms and peak at 232 ms. Together, we identified the role and timing of both early and late cross-modal interactions.

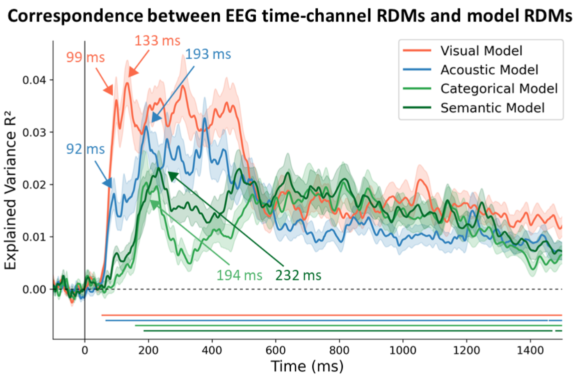

Figure 3: Adjusted explained variance was used as the correspondence measure when non-negative linear regressions were fitted between EEG time-channel RDMs and model RDMs. The bottom lines denote significant time windows (1000 sign permutation test with cluster correction, cluster-definition $p < 0.001$, cluster $p < 0.01$).

A two-branch audiovisual deep neural network captures the hierarchical processing, but not the early cross-modal interaction

We compared the neural representations with a two-branch audiovisual deep neural network (Morgado et al., 2021), which has a similar structure of separate visual and auditory cortices (Figure 4). The model was trained on Audioset (Gemmeke et al., 2017) through contrastive training to learn cross-modal agreement of video and audio.

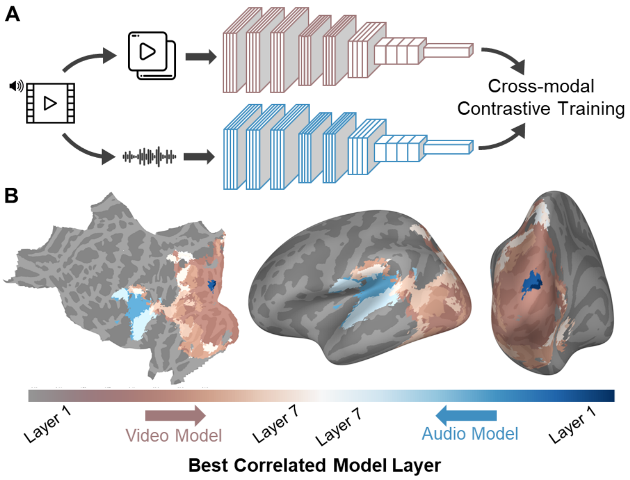

Figure 4: (A) Schematic illustration of a two-branch DNN model (Morgado et al., 2021) trained on audio-video stimuli with contrastive learning. (B) Model representations of seven blocks of each branch were extracted and compared with fMRI RDMs (1000 sign permutation test with cluster correction, cluster-definition $p < 0.001$, cluster $p < 0.01$). The best correlated model layer was visualized on the whole-brain map.

We observed that the model exhibited modality correspondence and hierarchical progression: early layers correlated with early regions, while later layers correlated with higherlevel regions. However, some voxels in early visual areas correlated best with early layers of the audio model, suggesting the model fails to capture the early cross-modal interaction. Therefore, including early integration in the DNN model is needed to build a more biologically plausible computational model and potentially improve model performance (Mo & Morgado, 2023).

Conclusion

In summary, our results revealed the spatiotemporal dynamics of information processing during the perception of naturalistic audiovisual stimuli and suggested two stages of cross-modal interactions with distinct roles. We also provided insights on how to build a more biologically plausible model of audiovisual processing.

Acknowledgments

This study was generously supported by the Canada First Research Excellence Fund (CFREF) through a BrainsCAN grant to Y.M., and a Western Strategic Support Grant to Y.M., and a Vector Institute Research Grant to Y.M.

References

Brown, G. J., & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech & Language*, *8*(4), 297–336.

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal inference in the multisensory brain. *Neuron*, *102*(5), 1076–1087.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in cognitive sciences*, *8*(4), 162–169.

Franzen, L., Delis, I., De Sousa, G., Kayser, C., & Philiastides, M. G. (2020). Auditory information enhances post-sensory visual evidence during rapid multisensory decision-making. *Nature communications*, *11*(1), 5440.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 776–780).

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in cognitive sciences*, *10*(6), 278–285.

Laurienti, P. J., Wallace, M. T., Maldjian, J. A., Susi, C. M., Stein, B. E., & Burdette, J. H. (2003). Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices. *Human brain mapping*, *19*(4), 213–223.

Mo, S., & Morgado, P. (2023). Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. *arXiv preprint arXiv:2312.01017*.

Morgado, P., Vasconcelos, N., & Misra, I. (2021). Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12475–12486).

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, *42*, 145–175.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Rohe, T., & Noppeney, U. (2015). Cortical hierarchies perform bayesian causal inference in multisensory perception. *PLoS biology*, *13*(2), e1002073.

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level,'unisensory'processing. *Current opinion in neurobiology*, *15*(4), 454–458.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*(6814), 788–788.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. MIT press.

Werner, S., & Noppeney, U. (2010). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *Journal of Neuroscience*, *30*(7), 2662–2675.