

Frontal cortex organization supporting audiovisual processing during naturalistic viewing

Corresponding Author: Mr Faxin Zhou

Parts of this Peer Review File have been redacted as indicated to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Zhou et al. investigate changing representations of auditory and visual information in dynamic audiovisual naturalistic stimuli. Using two different approaches (NMF and mTRF) the authors demonstrate distinct auditory and visual clusters that map onto respective sensory regions as well as distinct electrodes in the frontal cortex. Tuning of frontal electrodes to auditory or visual information follows a ventral-dorsal gradient and is driven by high-level sensory features. The authors then show that this tuning of frontal contacts changes depending on the context, English versus foreign language condition. A distinct cluster of electrodes in the frontal cortex is associated with behavioral ratings of the modality.

The manuscript is well written. The analysis is clearly presented in the paper and supports the claims by the authors. Potential confounds (visual subtitles) are adequately controlled for. Overall, the paper seems suitable for the broad readership of Nature Communications. Some clarification would be helpful to support the specificity of effects to the frontal cortex.

Major comment

The context specific representation of visual or auditory information in the frontal cortex (Figure 5A) is convincing. It would be potentially insightful to look at this effect in regions connected to the frontal cortex. For example, it seems that both regions could be part of the frontoparietal network and the global modality variable showed responses in the temporo-parietal junction. A potential control analysis computing changes in mTRF r-values by context (as in Figure 5A,B) in the parietal cortex could support the specificity of the frontal cortex as the site mediating context switches.

The results showing the neural correlates of the global modality variable are interesting. However, the global context variable might relate to cognitive processes, such as attention, or potential differences in visual sampling through saccades depending on context. As many features in naturalistic conditions these could interact with sensory stimuli in ways that are not captured by the correlation to auditory and visual TRFs. It would be helpful to expand on the discussion of how the distinction of the global modality cluster from sensory clusters ties into its role in switching contexts.

Minor comments

Lines 86-88: A detail but since following the sentence “given recent evidence showing that high-gamma activity reflects local neuronal firing” Leszczynski et al. is cited, which shows that broadband-high frequency activity (or high-gamma as referred to in this manuscript) does not only reflect local neural firing, it would be appropriate to modify the sentence slightly, i.e., “... mostly reflects local neural firing, but see Leszczynski et al.” or something alike.

Lines 109-110: Would it be worth quantifying the number of electrodes? Or refer to Figure S6 if it refers to these results? Figure 2: To support the claim of modality specific responses in each cluster, it would be helpful to plot responses in both STG and occipital contacts for both clusters.

Related to the responses in Figure 2, it seems that the responses in both clusters extend past the time window of 1s. Also, considering previous work (e.g. Fedorenko et al., 2016; Hamilton, 2018 - which is cited already; Honey et al., 2012) that showed slow integration of information / meaning of speech over longer timescales, would it be worth extending the window of analysis or specifically testing if regions with slower dynamics show the same results? The results could otherwise be

biased towards areas with shorter responses (i.e. posterior versus anterior STG). If out of the scope of this analysis a brief mention of this literature might be helpful.

Figure 5G&H: The meaning of the colorbars was unclear. Also, why do these variables have to be z-scored?

The abrupt changes of the global modality variable look like they could be correlated with film cuts in the movies. Could this be controlled for?

Lines 441-445: Are the times for segments in each condition comparable? For the TRF analysis this could affect the reliability of responses.

Lines 447-448: "Trials shorter than one second were included to ensure data stability and robustness": is this a typo? Should that state "excluded"?

Since Figure S7 controls for a major potential confound of the analysis, wouldn't it be better to use this Figure in the main text as Figure 4?

Throughout the paper responses in the left and right hemisphere are combined. Are there any hemispheric differences in the main findings?

References:

Fedorenko, E., et al. "Neural correlate of the construction of sentence meaning." *Proceedings of the National Academy of Sciences* 113.41 (2016): E6256-E6262.

Honey, C. J. et al. "Slow Cortical Dynamics and the Accumulation of Information over Long Timescales." *Neuron* 76, 423-434 (2012).

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)

Zhou et al. employ a naturalistic stimulus (multilingual movie) to investigate iEEG neural activity, focusing on the spatial spread of electrodes discriminating audio- and visual-sensitive regions. They quantify high-frequency activity (70–150 Hz) and apply various unsupervised machine learning techniques to provide evidence for the dorso-ventral gradients in the PFC, with dorsal PFC processing visual and ventral PFC processing audio information. The sample size is reasonable for iEEG studies. The study is well-conceived with a clear analytical approach. However, the knowledge of dorso-ventral gradients relating to audio-visual information processing in the frontal cortex is not novel, with some precedent in neurobiological functional mapping. This study is probably the first to give such evidence with iEEG data. There are some methodological concerns that require justification or additional analyses, and also an improved discussion to synthesise the findings, which is currently very short and light. I list below some of my concerns, which will benefit the manuscript if addressed

Introduction

- 1) The dorsal-ventral gradient hypothesis is not new; prior fMRI (e.g., Calvert et al., 2001; Clarke et al., 1995) and electrophysiology work (e.g., Romanski, 2007) have established frontal audiovisual organization. The novelty here is iEEG evidence, but the manuscript misses some key literature (e.g., Braga et al., 2016; Abdallah et al., 2022) and overlooks structural connectivity explanations for this gradient.
- 2) Other important iEEG studies such as Honey et al., 2012, which have looked at macroscopic brain activity for movies, are missing. Other relevant papers include Nir et al., 2007, *Current Biology*.
- 3) The assertion that "iEEG enables investigation of neural dynamics fMRI/scalp EEG cannot assess" is misleading. Scalp EEG has successfully probed audiovisual processing (e.g., early N1/P2 components etc.).
- 4) Line 66: "Intracranial EEG directly records from cortex with excellent spatiotemporal resolution (millimeters and milliseconds) and high signal-to-noise ratio (SNR), enabling investigation of neural dynamics in audiovisual processing that cannot be assessed on fMRI or scalp EEG" — Intracranial EEG directly records from cortex and also from the subcortex.
- 5) Line 48–49: "Only two functional Magnetic Resonance Imaging (fMRI) studies have directly investigated audiovisual processing during naturalistic movie viewing" — This is a far-fetched claim; many studies have used audiovisual properties of stimuli in fMRI. In fact, one of the first studies to do so, Bartels & Zeki et al., 2004, used the audiovisual properties of the movie *Tomorrow Never Dies*. There have been quite a few studies investigating audiovisual properties in fMRI and also with iEEG.

Methods:

- 6) Movie presentation with a laptop at a distance (0.5–1 m) introduces too much variability, especially if the research question is investigating visual information. However, I understand the clinical setting sometimes precludes strict control. No fixation cross was mentioned, raising questions about baseline attention.

- 7) The audio was played via speakers. In clinical settings, there is a lot of noise disturbances — how did the authors control for this in the experiment? The paper does not clarify how the sound onset/offset was detected (amplitude threshold? manual annotation?).
- 8) In the FL condition, subtitles are present, introducing a visual linguistic element absent in EN and OS. This raises the possibility that observed neural differences reflect subtitle reading rather than foreign speech processing per se. The authors do control for these in the gradient analysis.
- 9) Baseline issues: "Silent black scenes" as a baseline may retain residual activity from prior scenes; a 30 s window would ideally be good as audiovisual information produces emotional experience in viewers.
- 10) Line 422: "Electrodes showing epileptiform activity, line noise artifacts, or large amplitude shifts were excluded." In the preprocessing, line noise removal is not mentioned. Furthermore, how was the line noise harmonic of 120 Hz removed in the logarithmic frequency bins for computing the high-frequency activity?
- 11) The rationale for including very short trials (<1 s) is noted ("data stability and robustness"), but such claims might need some supplementary analyses with different trial lengths. Furthermore, since these are all different segments of the movie, showing averaged time plots (Fig. 1 and 2) for trial-based averaging may not be the best choice. In this case, scatter bar plots with the mean activity would be more optimal. Although from the current plots audiovisual activity of electrodes is apparent.
- 12) The 1-second lag window (τ_{max}) is arbitrary, potentially missing early (<100 ms) or late (>1 s) responses. This fixed limit is not empirically justified. Naturalistic processing can span both faster dynamics (e.g., ~100 ms auditory evoked responses) and much slower ones (e.g., multi-second narrative integration). Such a window may miss sustained attention effects or late semantic processing.
- 13) The criterion of consecutive significant time points likely favours electrodes with prolonged activity (e.g., high-gamma during ongoing processing), potentially missing brief yet meaningful events such as early evoked potentials. Analyses such as peak amplitude detection may be useful.
- 14) Gabor filters are good for local motion but less sensitive to global motion (e.g., camera panning) or higher-order visual features. Some higher-level visual features of salience could also be used.
- 15) The 20 Hz–8 kHz spectrogram range is unnecessarily broad, extending into ultrasonic frequencies (>20 kHz), irrelevant for human hearing.
- 16) In Fig. 3, there seem to be a lot of common electrode contact positions assigned to both clusters. Can a map of common positions be shown where the contacts are active for both audio and visual processing?
- 17) For the functional clustering, I am not sure if I completely understand whether the clustering for electrode contacts accounts for contacts with multimodal tuning to both visual and auditory information. I understand the clustering is based on the weights, but does the clustering result then assume/imply that auditory and visual tuning are mutually exclusive?
- 18) Similarly, I am wondering if using mTRF r-values from models trained and tested on the same dataset may overfit? Furthermore, audio and visual mTRFs were trained separately, which will then ignore crossmodal integration effects and artificially segregate audiovisual activity. Would it be wise to then run a combined model with both feature sets entered simultaneously?
- 19) The need for a new coordinate system is unclear, as MNI coordinates would suffice for gradient-based analyses.
- 20) Demographics table of patients and their diagnosed epileptogenic zones and the contacts which were excluded, should be provided.

Results:

21) Figure 1 – Please provide the electrode contacts that were active for each condition with the same colour scheme as in panel C.

Discussion:

22) This section is extremely short with no synthesis of findings from the extensive results section. Furthermore, the limitations section is missing as well. One suggestion is to also remark on the structural connectivity from visual and auditory regions that might dictate the dorso-ventral gradients. Neuroanatomical and functional knowledge, such as the frontal eye fields, which are dorsal, also need to be discussed.

Minor:

- Abstract: "Here, we employed movie-watching to"... it should be something like "movie paradigm," as "watching" might indicate visual information only.
- Line 63 – iEEG is intracranial electroencephalography. ECoG is electrocorticography.

Reviewer #4

(Remarks to the Author)

This work relates cortical responses, especially in the frontal cortex, to various components of a naturalistic audiovisual movie. The dataset is valuable, but I was somewhat lost by the complexity of the analyses, and I am not sure whether the authors' interpretation of their findings is the most plausible one.

Major:

Large number of analyses: The paper employs a number of (if I understood correctly) independent analysis strategies: unsupervised clustering (Fig. 2), mTRF based on components of the auditory vs. visual components of the movie (Figs. 3 and 4), and an analysis that relies on moment-to-moment ratings of the movie's contents by distinct viewers (Fig. 5). I think I was able to follow rather well up until Figure 4 included. I don't think I understood what Figure 6 represents. More generally, I

had a hard time understanding exactly what each analysis was doing and what new information it brought with respect to the previous ones. This can likely be resolved by reformulating the Results section to provide more "plain-language" information.

Auditory or visual vs. speech-related signals: I watched the movie that was used as a stimulus. Most of the information therein relates to audiovisual speech (most of the shots are close-ups of the actors' faces); there are a handful of short, wordless scenes that help understand the stories but are not actually essential given all the close-ups and dialogues that precede or follow them. This might be what explains the relative low numbers of electrodes driven by the "other sound" or "silent" periods. More importantly in my opinion, I wonder whether the strong emphasis on close-up dialogue in the film, and the attending speech- and language-heavy contents, really allows the authors to generalize to "auditory" vs. "visual" in their interpretation of how components of the movie drive cortical responses. An alternative explanation, which seems at least as plausible to me, relates these responses to the relative attention directed by the patients to the auditory vs. visual components of speech stimuli depending on whether they could understand what was being said. This is supported by the authors' finding that visual stimulus components seem to be more important in driving cortical responses when the actors speak in a language that is unintelligible to the participants, whereas the auditory stimuli drive responses more when the actors speak English. Perhaps the authors have additional data collected with the same patients to resolve this confound? Additionally, cortical responses from patients who understood one of the "foreign" languages spoken in the movie would be useful.

Lack of a clear link between cortical activities and behavior: given that the patients did not perform any particular task while watching the movie, it is too much of a stretch to affirm that the frontal cortical activities described in the paper underlie multisensory perception.

Minor:

Flexible AV resource assignment (Fig. 5A): wouldn't this analysis make more sense if only cortical sites that are activated by both English and foreign languages were selected?

Silent periods: was a volume criterion used to determine when there was no sound during these periods?

Cognitive involvement (results, line 256): I'm not sure that, because viewers rated a movie snippet as more important for understanding the whole movie, it should mean that this snippet "demanded greater cognitive involvement".

Numbers of principal components (Fig. S4): given that only multiples of 5 PCs were explored, I suggest foregoing the continuous line and showing only the mean r-values at the actual counts of PCs investigated. Another question on PCs: was 15 the optimal number of PCs for both the auditory and the visual model? Is there anything of interest in such a coincidence?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have addressed most of our concerns and questions comprehensively. We have two additional comments that may be addressed at the author's discretion.

Re: answer 1

The analysis of flexible resource assignment in the parietal lobe is interesting. It underscores that this effect is particularly pronounced in the frontal cortex. Given that there was a significant interaction effect between language and modality conditions a possible alternative interpretation could be that audiovisual resource assignment does take place in the parietal cortex in some form. Though the presence of the main effect of modality suggests that sensory processing takes on a stronger role than in the frontal cortex. Acknowledging the significant interaction effect would allow the reader to interpret the data more clearly.

Re: answer 4

While the pie chart gives a good sense of how the electrodes are balanced between different anatomical regions, further quantification could be helpful. The total number of electrodes in each anatomical region would give a better sense of density of coverage. The ratio of active electrodes over the total number of electrodes in each area could give a sense of sensitivity of electrodes in each area to the stimulus features. In combination with figure R18 this could improve assessing coverage.

We congratulate the authors on their interesting and well-executed work.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)

The authors have satisfactorily addressed my concerns.

Reviewer #4

(Remarks to the Author)

I thank the authors for their thorough consideration of my (and the other reviewers') comments.

Regarding my second major point and the authors' A2: I'm not quite sure that we are talking about exactly the same thing. I did not mean to say that the movie segments that were intelligible were inherently more "attention-driving" than the others. Rather, I thought that, when the auditory component of the movie is intelligible speech, then it would make sense to think that it would recruit the inferior frontal cortex (and its speech- and language-related functions) more than when it is unintelligible. Thus, instead of representing differences in cortical engagement with auditory vs. visual components of the stimulus, the authors' findings could reflect more speech- and language-related processing vs. less. The additional analysis presented in Figure R19 does not alleviate my concern. In fact, if no non-speech and language AV stimuli are available in the same patients, and if no patient understood any of the "foreign" languages in the movie, I do not think that this potential confound can be addressed completely.

I think that the above confounds warrants a specific discussion of this particular point in the limitations sections -- unless I missed a subtle point of the authors' analytical approach, in which case I am happy to be guided again through it.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Frontal cortex organization supporting
audiovisual processing during naturalistic viewing
Response to reviewers

Faxin Zhou et.al.

November 19, 2025

Reviewer #1

Summary

Zhou et al. investigate changing representations of auditory and visual information in dynamic audiovisual naturalistic stimuli. Using two different approaches (NMF and mTRF) the authors demonstrate distinct auditory and visual clusters that map onto respective sensory regions as well as distinct electrodes in the frontal cortex. Tuning of frontal electrodes to auditory or visual information follows a ventral-dorsal gradient and is driven by high-level sensory features. The authors then show that this tuning of frontal contacts changes depending on the context, English versus foreign language condition. A distinct cluster of electrodes in the frontal cortex is associated with behavioral ratings of the modality.

The manuscript is well written. The analysis is clearly presented in the paper and supports the claims by the authors. Potential confounds (visual subtitles) are adequately controlled for. Overall, the paper seems suitable for the broad readership of Nature Communications. Some clarification would be helpful to support the specificity of effects to the frontal cortex.

We sincerely appreciate your supporting and insightful comments that have greatly improved our work. We have addressed your concerns point-by-point below and incorporated your suggestions in our manuscript.

Comments

Major comments

The context specific representation of visual or auditory information in the frontal cortex (Figure 5A) is convincing. It would be potentially insightful to look at this effect in regions connected to the frontal cortex. For example, it seems that both regions could be part of the frontoparietal network and the global modality variable showed responses in the temporo-parietal junction. A potential control analysis computing changes in mTRF r -values by context (as in Figure 5A,B) in the parietal cortex could support the specificity of the frontal cortex as the site mediating context switches.

Answer 1 (A1): We thank you for this helpful suggestion. Given that the parietal cortex is largely connected with the frontal regions and has been implicated in audiovisual processing [1, 2], we agree that it would be valuable to examine parietal cortex activity to assess its functional specificity relative to frontal regions. To this end, we defined parietal regions using a region-of-interest (ROI) approach based on within-subject (FreeSurfer) anatomical labels, including the postcentral gyrus, inferior parietal lobe, and supramarginal gyrus. We examined and reported the spatial distribution and r values of significant electrodes within this ROI.

The results revealed that the parietal cortex exhibited a distinct pattern compared with the frontal cortex. Specifically, although we found a significant interaction effect between language and modality conditions (two-way ANOVA, language-by-modality interaction effect: $F(1, 122) = 55.178, p = 0.008, \eta^2 = 0.056$; Fig. 1B), we also observed a significant main effect of modality (two-way ANOVA, modality main effect: $F(1, 122) = 24.340, p < 0.001, \eta^2 = 0.166$; Fig. R1). This finding indicates that auditory encoding performance in the parietal cortex was greater than visual encoding performance across both EN and FL conditions, in contrast to the context-dependent modality switching observed in the frontal region. However, we note that the r values in the parietal cortex were relatively weak and the coverage of significant electrodes was sparse. We therefore remain cautious in this conclusion and look forward to future work that can validate this functional specificity within the frontoparietal network. We have added this analysis to the revised manuscript (line 237-242) and supplementary materials (Fig. S7):

“Notably, this effect was not observed in other areas typically implicated in audiovisual processing, such as the parietal regions [34, 35], in which auditory encoding performance exceeded visual encoding performance across both the EN and FL conditions (two-way ANOVA, modality main effect: $F(1, 122) = 24.340, p < 0.001, \eta^2 = 0.166$; Fig. S7). This result further supports the specificity of the frontal cortex as a key site in context switching.”

The results showing the neural correlates of the global modality variable are interesting. However, the global context variable might relate to cognitive processes, such as attention, or potential differences in visual sampling through saccades depending on context. As many features in naturalistic conditions these could interact with sensory stimuli in ways that are not captured by the correlation to auditory and visual TRFs. It would be helpful to expand on the discussion of how the distinction of the global modality cluster from sensory clusters ties into its role in switching contexts.

A2: We agree with your concerns. Naturalistic movie viewing involves multiple cognitive processes simultaneously,

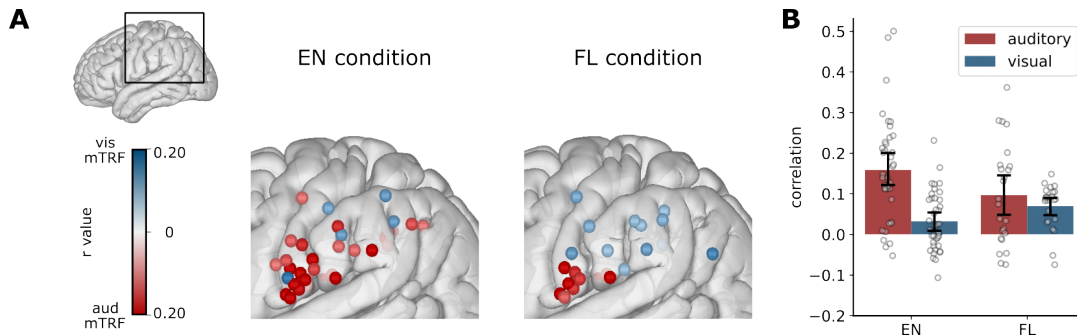


Figure R1 (A) The mTRF correlation maps in the parietal region for the English (EN) and Foreign Language (FL) conditions. (B) The bar plot of the auditory and visual mTRF models in the parietal cortex for the EN and FL conditions. Error bars represent 95% confidence interval (CI).

such as attention and visual sampling via eye movement, which could potentially confound the global modality variable (i.e., the modality assignment effect). We address these concerns with multiple approaches detailed below.

Attentional effect. During the “online movie-rating task”, we also collected continuous engagement ratings (Fig. R2A) by asking participants to frequently report how engaging the movie is during their first viewing (Fig. R2B). This measure has been previously applied to reflect the sustained attention during movie viewing [3]. We found that this attentional engagement was significantly correlated with the global context (one-sample t test; $T(25) = 11.987$, $p < 0.001$, $d = 2.351$), but not correlated with the local modality (one-sample t test; $T(25) = 2.068$, $p = 0.074$, $d = 0.406$) or global modality (one-sample t test; $T(25) = -0.608$, $p = 0.549$, $d = 0.119$; Fig. R2C). We further applied the mTRF model to this attentional engagement variable and observed a brain pattern distinct from that of the global modality ($r = 0.015$, $p = 0.431$; Fig. R2D). Moreover, mTRF brain maps trained with and without removing engagement yielded consistent global modality patterns, suggesting that attentional engagement does not confound our results ($r = 0.621$, $p < 0.001$; Fig. R2E, F).

Eye movement. Unfortunately, we were not able to collect eye-tracking data from our neurosurgical population due to the technical difficulty and time required for setup within the hospital environment. While we cannot directly assess this potential confound, previous studies have shown that film cuts are followed by transient increases in saccade frequency, and can strongly modulate eye movements [4]. Therefore, we investigated whether film cuts influence the global modality to implicitly assess whether eye movements act as a confounding factor. No significant correlation was observed between the film cuts and the global modality (please see response to your concern A8 for analytical details). Nevertheless, given the absence of direct eye-movement data, we acknowledge this as a limitation of our study (line 379-382):

“Second, the absence of behavioral measures from patients (e.g., eye-tracking recordings and direct assessments of movie comprehension) limited our ability to directly account for attention and eye movements.”

Taken together, our analyses suggest that the global modality cluster is not explained by the attentional effect (please see A2 to Reviewer #4 for more analyses) or by sudden visual transients induced by film cuts (please see A8). We therefore interpret the global modality cluster as a mechanism for context-dependent assignment of audiovisual modality, which in turn governs the relative weighting of auditory and visual information (please see also A3 to Reviewer #4 for more discussion). We have revised the manuscript to incorporate these additional analyses (line 275-289; Fig. S9):

“To evaluate the robustness of these findings, we conducted a series of control analyses. First, to rule out the potential confounding effect of attention, we applied the mTRF model to the attentional engagement variable obtained from the engagement rating task (Fig. S9A, B; see Methods: Task and procedure), which revealed a brain pattern distinct from that of the global modality ($r = 0.015$, $p = 0.431$; Fig. S9D). Moreover, the global modality mTRF brain maps trained with and without removing engagement yielded consistent patterns ($r = 0.621$, $p < 0.001$; Fig. S9E, F), suggesting that attentional engagement did not confound our results. Second, we examined whether film cuts could influence the results, as scene transitions might alter the assignment of audiovisual resources. To address this, we computed the first derivative of the global modality and compared it with the film cuts. We defined the film cuts at each abrupt change in movie scenes (i.e., the boundaries; Fig. S9G) [40]. No significant correlations were observed between the film cuts and changes in the global modality across raters (one-sample t test against 0: $T(25) = -1.817$, $p = 0.081$, $d = 0.356$; Fig. S9H), indicating that film cuts did not confound the observed effects.”

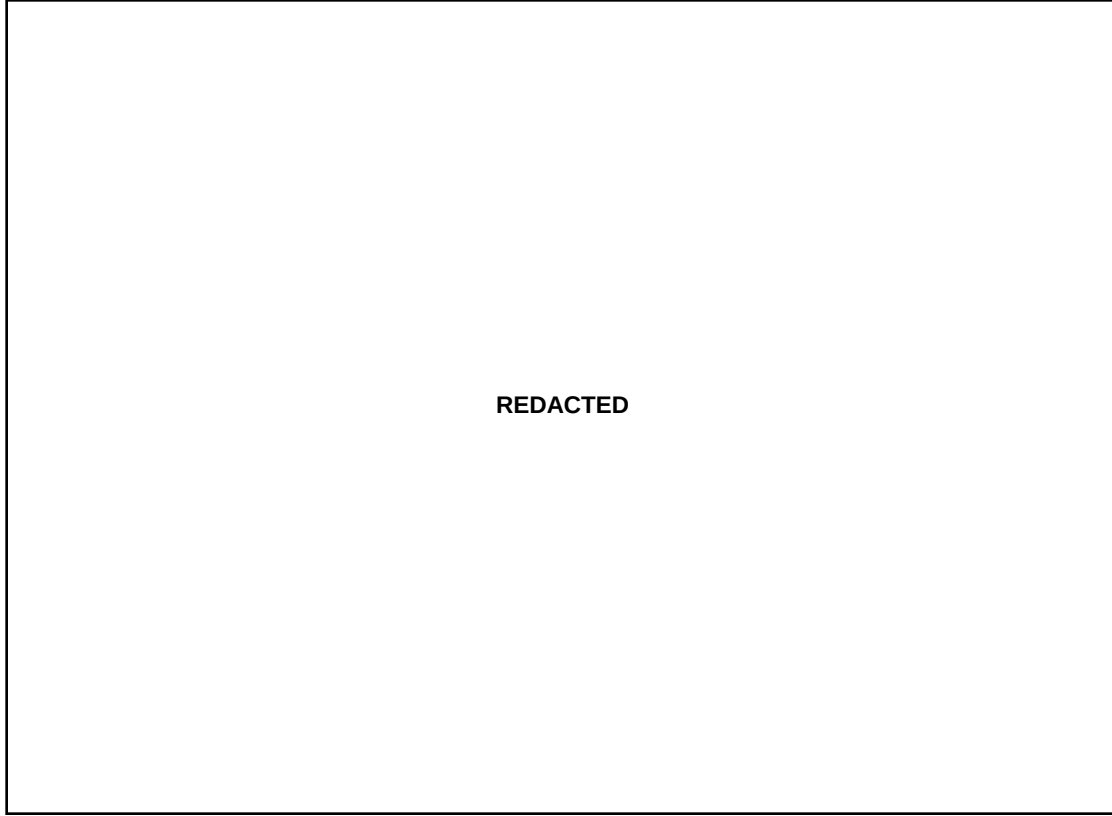


Figure R2 Control analysis of the attentional effect. (A) The attentional engagement ratings from all participants over time (upper panel) and the average time series across participants (lower panel). (B) Illustration of the engagement rating task conducted on the Amazon Mechanical Turk platform. (C) Correlations between attentional engagement ratings and global context, local modality, as well as global modality. Each dot represents an individual rater. (D) The mTRF result for the attentional engagement effect. (E) The mTRF results for the original global modality (left) and after removing the attentional effect (right). (F) Comparison of r values between the two models across all electrodes.

Minor comments

Lines 86-88: A detail but since following the sentence “given recent evidence showing that high-gamma activity reflects local neuronal firing” Leszczynski et al. is cited, which shows that broadband-high frequency activity (or high-gamma as referred to in this manuscript) does not only reflect local neural firing, it would be appropriate to modify the sentence slightly, i.e., “...mostly reflects local neural firing, but see Leszczynski et al.” or something alike.

A3: We apologize for the inappropriate citation. As you noted, Leszczyński and colleagues demonstrated that the high-frequency activity (BHA) reflects a mixture of neural action potentials and dendritic processes that can be dissociated from spiking activity. After careful evaluation, we believe that dendritic processes is not central to our current study, therefore we chose not to explicitly mention it. We have revised the manuscript accordingly (line 84-85) and replaced this reference with other, more directly relevant studies such as mukamel et al. (2005), Nir et al. (2007), and Flinker et al. (2010):

“In this study, we focused on the high-gamma (70 to 150 Hz) broadband field potentials, which are robust at the single-trial level [16], reflect local neuronal firing [17–19], and exhibit strong correlation with the fMRI blood-oxygenation level-dependent (BOLD) response [17, 20].”

Lines 109-110: Would it be worth quantifying the number of electrodes? Or refer to Figure S6 if it refers to these

results?

A4: We appreciate this constructive comment. We agree that quantifying the number of active electrodes is valuable and enhances the interpretability of the result. Accordingly, we have added a pie chart showing the anatomical distribution of all active electrodes (Fig. R3). We have also updated the manuscript (line 109-110) and the supplementary materials (Fig. S8C, which is the Fig. S6 in the original manuscript) accordingly:

The electrodes with significant responses were predominantly distributed in the superior temporal gyrus (STG), occipital regions, frontal cortex, and inferior parietal lobe (IPL; Fig. 1B, S1C, S8C; permutation test, $p < 0.05$, FDR corrected; see Methods: Active electrodes selection).

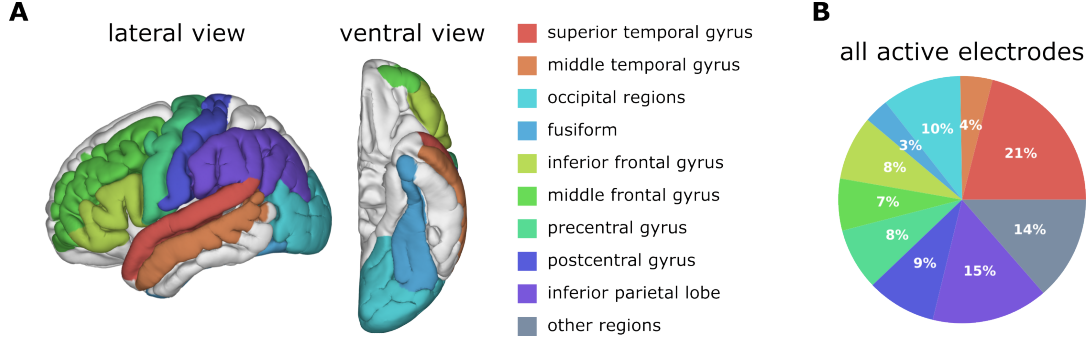


Figure R3 Nine ROIs illustrated on the brain (A) and the pie chart showing the anatomical distribution of the active electrodes (B).

Figure 2: To support the claim of modality specific responses in each cluster, it would be helpful to plot responses in both STG and occipital contacts for both clusters.

A5: We thank you for the helpful suggestion. We have plotted the responses in both the STG and occipital ROIs for each cluster. As expected, occipital electrodes in cluster 1 did not show notable changes across all four conditions (Fig. R4C). Likewise, STG electrodes in cluster 2 also did not exhibit discernible patterns. We have updated these results into Fig. 2 of the revised manuscript.

Related to the responses in Figure 2, it seems that the responses in both clusters extend past the time window of 1s. Also, considering previous work (e.g. Fedorenko et al., 2016; Hamilton, 2018 - which is cited already; Honey et al., 2012) that showed slow integration of information / meaning of speech over longer timescales, would it be worth extending the window of analysis or specifically testing if regions with slower dynamics show the same results? The results could otherwise be biased towards areas with shorter responses (i.e. posterior versus anterior STG). If out of the scope of this analysis a brief mention of this literature might be helpful.

A6: We thank you for raising this interesting and important topic. The timescale hierarchy in the brain has received increasing attention in recent years. Some studies (e.g., Fedorenko et al., 2016 and Hamilton et al., 2018) have shown that certain brain regions exhibit sustained activity over longer timescales [5, 6, 7]. Other studies (e.g., Honey et al., 2012) have approached this question from the perspective of working memory (WM) and proposed the concept of the temporal receptive window (TRW) [8, 9, 10], which describes how the brain integrate information from the recent past across multiple timescales. While a comprehensive investigation of the timescale hierarchy is beyond the scope of our study, we agree this is a highly interesting direction and have tentatively explored it with our movie-viewing data in detail below.

First, regarding your concern that our results might be biased toward areas with shorter responses due to the 1-second analysis window, we hold that our permutation approach (i.e., the consecutive time-point method) theoretically favors electrodes with prolonged activity rather than those with transient strong responses (please see A13 to Reviewer #3 for details). Consistently, we also showed that IFG electrodes with sustained activity also survived the permutation test despite having weaker amplitudes compared with STG electrodes (Fig. R5A, C). Furthermore, these sustained electrodes also survived our new T -statistic-based permutation approach, which balances sensitivity to both strong transient responses and weaker sustained responses (please see A13 to Reviewer #3). Therefore, we hold that our findings are not biased toward areas with shorter response and those electrodes with stable sustained activity can still be reliably detected within the 1-second window.

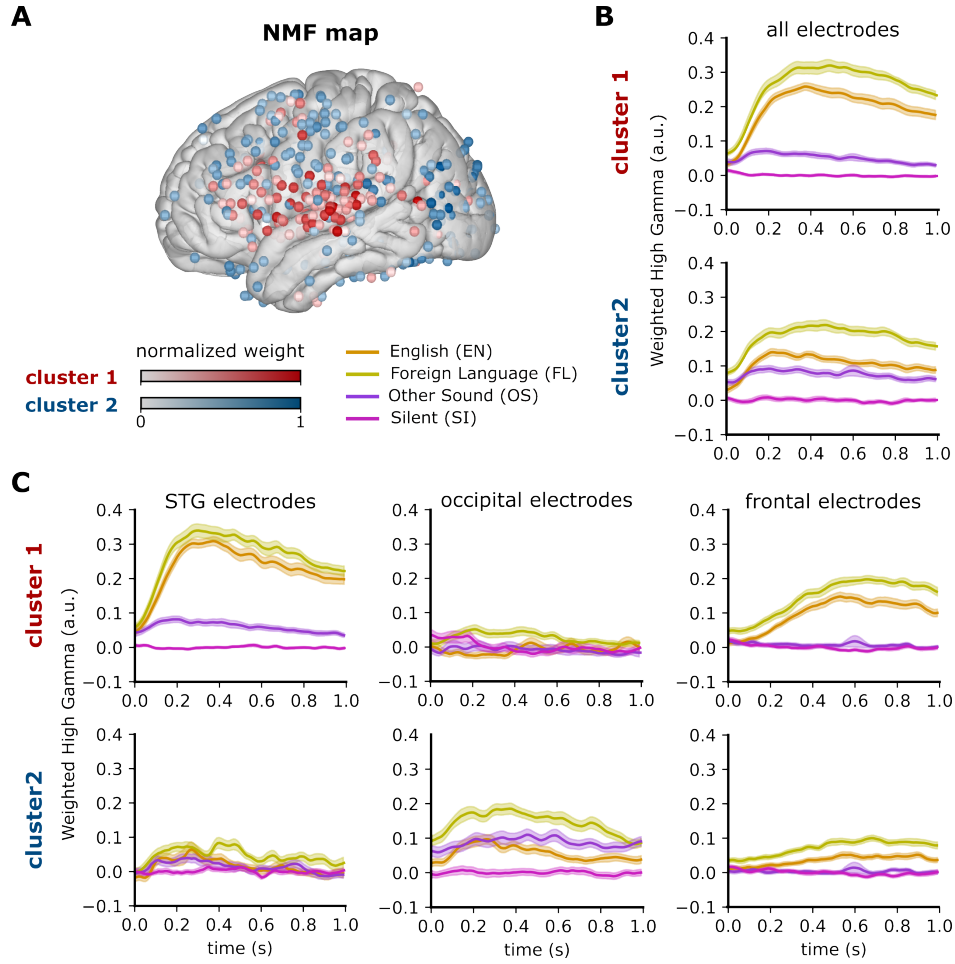


Figure R4 Functional clustering of the active responses. (A) Spatial map showing two identified clusters with the non-negative matrix factorization (NMF) approach. For both clusters, weights were normalized to the range 0-1, with higher values indicating greater contributions to each cluster. (B) Weighted average signals of the two clusters across all electrodes. (C) Weighted average signals across the electrodes in three specific ROIs (Fig. S1B): STG (left panel), occipital regions (middle panel), and frontal regions (right panel). Shaded areas around the curves represent the SEM.

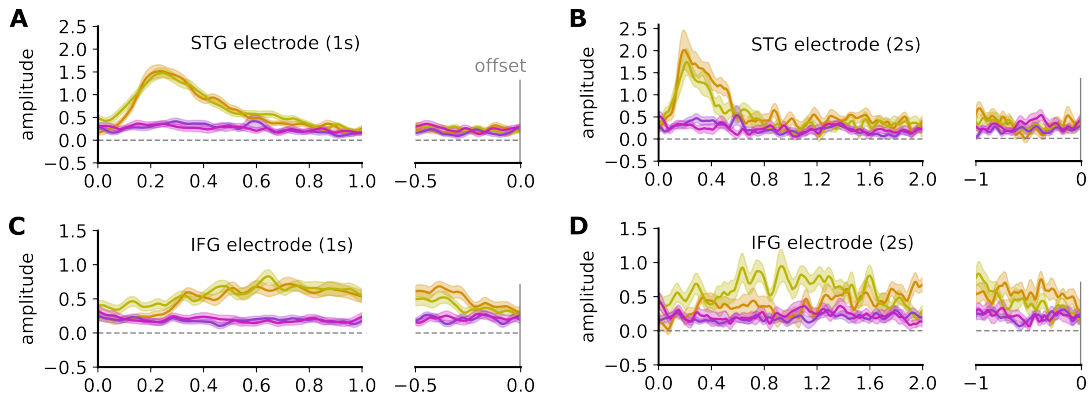


Figure R5 Averaged signals across the four conditions for an example STG electrode (A, B) and IFG electrode (C, D), with trials longer than 1 second (A, C) or 2 seconds (B, D). The left panels show signals aligned to trial onset and the right panels show signals aligned to offset. Shaded areas around the curves represent the SEM.

Next, we examined how long sustained activity could be maintained. Following your suggestion, we analyzed trials longer than 2 seconds across all conditions. However, the limited number of such trials (English: 10 trials; Foreign Language: 11 trials; Other Sound: 26 trials; Silent: 32 trials; please see also A9) resulted in noisy and unreliable signals (Fig. R5B, D). Further, we also aligned activity to trial offsets for each condition (Fig. R5A, C; right panels) and observed that STG activity typically diminished before the end of trials, whereas frontal activity persisted until trial completion (in the EN and FL conditions). These results suggest that sustained frontal activity is likely bounded by trial duration.

Previous studies have also reported that several other regions, including the precuneus and medial prefrontal cortex (mPFC), exhibit longer TRWs [8, 11]. Since most patients in our study lacked electrode coverage in these medial regions, we were unable to directly examine them. Nevertheless, we believe that investigating temporal hierarchies in the multisensory brain with iEEG data, especially under naturalistic conditions, is an important direction for future research. We look forward to studies with broader electrode coverage that can explore this question in greater depth. We have also briefly mentioned these findings in the Discussion section of the revised manuscript (line 335-337):

“Moreover, the sustained frontal activity during the language context (Fig. 2C) suggests its involvement in maintaining recent narrative information over time [50, 51].”

Figure 5G&H: The meaning of the colorbars was unclear. Also, why do these variables have to be z-scored?

A7: We apologize for the confusion. The colors in Fig. 5G-H correspond to the four conditions, consistent with the color scheme defined in Fig. 1A of the manuscript. Moreover, we agree with you that the global context variable should not be z-scored (Fig. R6). In addition, the global modality variable (Fig. 5H) is not z-scored; it is calculated as the product of global context (ranging from 0 to 1) and local modality (ranging from -1 to 1), resulting in values from -1 to 1. We have updated the figure and legend accordingly in the revised manuscript (Fig. 5G):

“The colors in figure G and H correspond to the four conditions, consistent with the color scheme defined in Fig. 1A.”

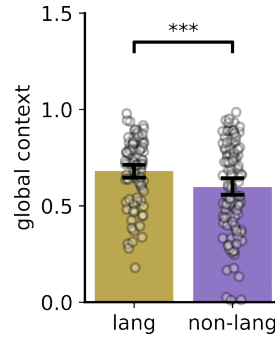


Figure R6 (A) The bar plot of the global context ratings (not z-scored) for the language (lang) condition, including the EN and FL conditions, and the non-language (non-lang) condition, including the OS and SI conditions. The dots represent trials in each condition.

The abrupt changes of the global modality variable look like they could be correlated with film cuts in the movies. Could this be controlled for?

A8: We understand your concern that film cuts could be a potential confounding factor. To address this, we computed the first derivative of the global modality and compared it with the film cuts. We defined the film cuts at each abrupt change in movie scenes (i.e., the boundaries; Fig. R7A) [12]. We did not observe significant correlations between the film cuts and changes in the global modality across raters (one-sample t test against 0: $T(25) = -1.817$, $p = 0.081$, $d = 0.356$; Fig. 7B). This result suggests that the film cuts do not confound the global modality.

Lines 441-445: Are the times for segments in each condition comparable? For the mTRF analysis this could affect the reliability of responses.

A9: We agree that the differences in segment lengths across conditions could potentially affect the mTRF results. To address this, we compared the segment durations across the four conditions and found no significant differences

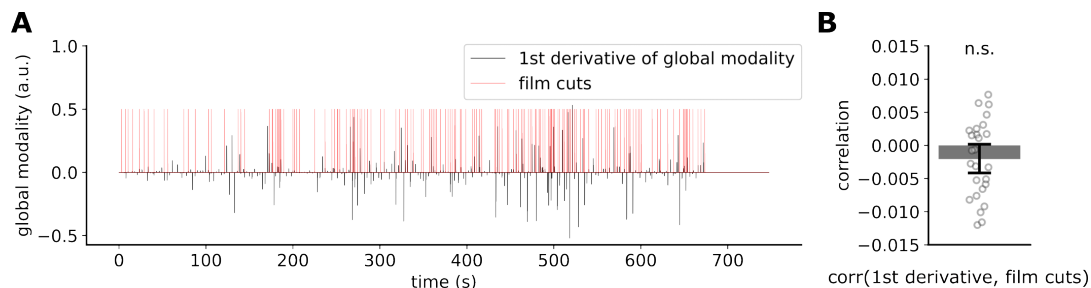


Figure R7 Control analysis of the film cuts. (A) Markers of abrupt changes in the averaged global modality (black; quantified as its first derivative) and the film cuts (red). (B) Correlations between the film cuts and the global modality changes across participants. Each dot represents an individual rater. Error bars represent 95% CI.

(one-way ANOVA: $F(3, 198) = 1.020$, $p = 0.385$, $\eta^2 = 0.061$; Fig. R8). This result indicates that segment lengths are comparable across all conditions, further supporting the reliability of our mTRF findings. This result has been incorporated into the revised manuscript (line 502-505; Fig. S3C):

“As a result, 44 EN trials, 46 FL trials, 58 OS trials, and 51 SI trials with comparable segment durations were included for analysis (one-way ANOVA: $F(3, 198) = 1.020$, $p = 0.385$, $\eta^2 = 0.061$; Fig. S3C).”

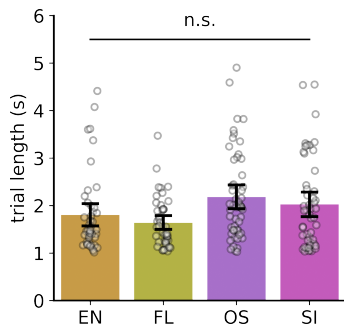


Figure R8 Bar plots of trial lengths across four conditions. Each dot represents an individual trial.

Lines 447-448: “Trials shorter than one second were included to ensure data stability and robustness”: is this a typo? Should that state “excluded”?

A10: We apologize for this typo. You are correct that the word “included” should be “excluded”. We have corrected this in the revised manuscript (line 503).

Since Figure S7 controls for a major potential confound of the analysis, wouldn’t it be better to use this Figure in the main text as Figure 4?

A11: We thank you for this useful suggestion. We have incorporated Fig. S7 into the Fig. 4 of the revised manuscript.

Throughout the paper responses in the left and right hemisphere are combined. Are there any hemispheric differences in the main findings?

A12: We thank you for raising this important question and believe that investigating the functional hemispheric asymmetries is valuable. However, due to the unbalanced electrode coverage between the two hemispheres (as shown in Fig. S1A), the left hemisphere contains substantially more electrodes than the right. As a result, the number of active and encoding electrodes in the right hemisphere is sparse, limiting our ability to explicitly test for hemispheric differences at this stage. We have listed this as a potential limitation (line 382-384):

“Finally, unbalanced electrode coverage across hemispheres (Fig. S1A) precluded an investigation of the potential

lateralization effects during movie viewing.”

References

Fedorenko, E., et al. “Neural correlate of the construction of sentence meaning.” *Proceedings of the National Academy of Sciences* 113.41 (2016): E6256-E6262.

Honey, C. J. et al. “Slow Cortical Dynamics and the Accumulation of Information over Long Timescales.” *Neuron* 76, 423–434 (2012).

Reviewer #2

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We sincerely thank you for your contributions in providing the valuable comments that helped us improve our manuscript.

Reviewer #3

Summary

Zhou et al. employ a naturalistic stimulus (multilingual movie) to investigate iEEG neural activity, focusing on the spatial spread of electrodes discriminating audio- and visual-sensitive regions. They quantify high-frequency activity (70–150 Hz) and apply various unsupervised machine learning techniques to provide evidence for the dorso-ventral gradients in the PFC, with dorsal PFC processing visual and ventral PFC processing audio information. The sample size is reasonable for iEEG studies. The study is well-conceived with a clear analytical approach. However, the knowledge of dorso-ventral gradients relating to audio-visual information processing in the frontal cortex is not novel, with some precedent in neurobiological functional mapping. This study is probably the first to give such evidence with iEEG data. There are some methodological concerns that require justification or additional analyses, and also an improved discussion to synthesise the findings, which is currently very short and light. I list below some of my concerns, which will benefit the manuscript if addressed

We sincerely thank you for your supportive and constructive comments. Below, we respond to each of your points in detail.

Comments

Introduction

1) The dorsal-ventral gradient hypothesis is not new; prior fMRI (e.g., Calvert et al., 2001; Clarke et al., 1995) and electrophysiology work (e.g., Romanski, 2007) have established frontal audiovisual organization. The novelty here is iEEG evidence, but the manuscript misses some key literature (e.g., Braga et al., 2016; Abdallah et al., 2022) and overlooks structural connectivity explanations for this gradient.

A1: We thank you for the additional literature cited, which has greatly helped us improve our manuscript. We have carefully read the studies you mentioned, revised the manuscript to better contextualize our work within the broader literature, and modified the Discussion section to explicitly relate our findings to these prior studies. Detailed responses are provided below.

The results of Clarke et al. (1995) and Romanski (2004, 2007) primarily demonstrated that distinct frontal neurons selectively responding to visual (e.g., faces) and auditory information (e.g., complex sounds), suggesting that the frontal cortex represents auditory and visual information separately [13, 14, 15]. We have incorporated these studies into Discussion section (line 321-324):

“Previous studies have shown that the frontal cortex is broadly involved in audiovisual processing during natural movie watching [13, 14] and distinct frontal subregions respond selectively to auditory and visual information in both humans [41–44] and non-human primates [6, 45, 46].”

Calvert and colleagues systematically investigated how the brain integrates audiovisual information and highlighted the crucial role of frontal cortex in this process [16, 17, 18]. By reviewing recent human neuroimaging and electrophysiological studies, they proposed a neurobiological model of multisensory processing, suggesting that the frontal cortex contributes to integrating newly acquired crossmodal associations [18]. This line of research has also been incorporated to the revised manuscript (line 371-373):

“The frontal cluster associated with the “modality assignment” also aligns with prior evidence implicating frontal regions in processing novel information during multimodal integration [68–70].”

In addition, Braga et al. (2017) applied both structural and functional connectivity analyses and consistently reported a dorsoventral gradient corresponding to an audio-to-visual transition in both the frontal and parietal cortices [19]. Their findings are highly consistent with our results, and we have incorporated their work into the revised manuscript (line 339-341):

“..... as well as studies combining structural and functional connectivity showing a dorsoventral frontal gradient corresponding to an auditory-to-visual transition [57].”

Finally, Abdallah and colleagues examined functional gradients in the human IPFC using a meta-analytic connectivity approach [20]. They identified a rostrocaudal and a dorsoventral gradient, which explained the largest and second-largest variance across the IPFC, respectively. This study supports a continuous topographic organization in the frontal cortex (line 347-352):

“By constructing a continuous ventrodorsal index from our coordinate system, we were able to quantify modality transition trends in a fully data-driven manner without imposing rigid anatomical assumptions. This suggests a spatially continuous coordinate system along the ventral–dorsal axis of the frontal cortex [58–60], reflecting a

functional topographic organization that supports audiovisual processing under naturalistic experiences.”

2) Other important iEEG studies such as Honey et al., 2012, which have looked at macroscopic brain activity for movies, are missing. Other relevant papers include Nir et al., 2007, Current Biology.

A2: We thank you for bringing these important studies to our attention. We have also revised the manuscript to incorporate these works. Specifically, Honey and colleagues investigated the timescale hierarchy of information accumulation during movie viewing [9], which has been added to the Discussion section (line 335-337):

“Moreover, the sustained frontal activity during the language context (Fig. 2C) suggests its involvement in maintaining recent narrative information over time [50, 51].”

In addition, Nir and colleagues demonstrated reliable coupling between neural firing, gamma LFP activity (40-130 Hz), and fMRI BOLD signals during both audiovisual movie-viewing and resting states [21]. We have incorporated this study into the Result section to further support our methodological choices (line 83-86):

“In this study, we focused on the high-gamma (70 to 150 Hz) broadband field potentials, which are robust at the single-trial level [16], reflect local neuronal firing [17–19], and exhibit strong correlation with the fMRI blood-oxygenation level-dependent (BOLD) response [17, 20].”

3) The assertion that “iEEG enables investigation of neural dynamics fMRI/scalp EEG cannot assess” is misleading. Scalp EEG has successfully probed audiovisual processing (e.g., early N1/P2 components etc.).

A3: We apologize for the potential confusion. We have removed this clause from the sentence accordingly. Please see our response to A4 for the revised sentence.

4) Line 66: “Intracranial EEG directly records from cortex with excellent spatiotemporal resolution (millimeters and milliseconds) and high signal-to-noise ratio (SNR), enabling investigation of neural dynamics in audiovisual processing that cannot be assessed on fMRI or scalp EEG” — Intracranial EEG directly records from cortex and also from the subcortex.

A4: We thank you for the reminder and apologize again for the misleading wording. We have change “cortex” to “brain” (line 65-67):

“Intracranial EEG directly records from the brain with excellent spatiotemporal resolution (millimeters and milliseconds) and high signal-to-noise ratio (SNR) [15].”

5) Line 48–49: “Only two functional Magnetic Resonance Imaging (fMRI) studies have directly investigated audiovisual processing during naturalistic movie viewing” — This is a far-fetched claim; many studies have used audiovisual properties of stimuli in fMRI. In fact, one of the first studies to do so, Bartels & Zeki et al., 2004, used the audiovisual properties of the movie Tomorrow Never Dies. There have been quite a few studies investigating audiovisual properties in fMRI and also with iEEG.

A5: We apologize for our overstated claim. We have carefully read the paper by Bartels & Zeki (2004) and conducted a comprehensive literature review on studies examining audiovisual properties using movies as stimuli.

Specifically, Bartels & Zeki (2004) investigated the functional specialization during movie viewing and focused primarily on visual attributes [22]. These attributes were derived from subjective ratings by human observers, including color, faces, language, and human bodies. Although the “language” feature seems to partially reflect the auditory modality, the instruction for this was rather modality-general: “indicate the degree to which you hear, read, and listen to language”. Therefore, it remains unclear how different sensory modalities are distinctly processed and organized in naturalistic conditions.

Furthermore, while several studies have examined audiovisual properties during movie viewing, most of them either primarily focused on a single modality (such as auditory [23, 24] or visual [22, 25]) or used audiovisual features to predict or classify other movie-related aspects (such as emotional states [26, 27] or movie genres [28]). Direct investigations of how auditory and visual movie features are represented and integrated in the brain remain relatively rare. We identified two main challenges in this regard: (1) in naturalistic movies, audio may be absent at times while visual information is always presented, leading to an imbalance of modality richness; (2) extracting comparable and comprehensive features across the two modalities, particularly at higher levels, is difficult. In the present study, we addressed these challenges by combining transformer-based deep learning models to extract audiovisual features and applying separate encoding models for different movie contexts (e.g., English and Foreign Language). Nonetheless, we acknowledge our original statement was too strong and have revised it as follows (line 48-49):

“Only a limited number of functional Magnetic Resonance Imaging (fMRI) studies have directly investigated audio-visual processing during naturalistic movie viewing.”

Methods

6) Movie presentation with a laptop at a distance (0.5–1 m) introduces too much variability, especially if the research question is investigating visual information. However, I understand the clinical setting sometimes precludes strict control. No fixation cross was mentioned, raising questions about baseline attention.

A6: We agree with your concerns and thank you for your understanding of our clinical difficulty. As you noted, the distance between participants and the laptop affects the visual angle, thus introducing the variability in visual cortical representations. However, since our study primarily investigates higher-level representations in the frontal cortex, we believe that this variability does not substantially affect our main conclusions.

Regarding the absence of “fixation cross”, you are correct that no fixation period was implemented before or after movie viewing. We acknowledge that this may raise concerns about baseline attention. To mitigate this issue, we used the “silent black scenes” as the baseline in the active electrodes analyses (please see also A9 for more details). In conjunction with the permutation procedure (please see also A13), we believe that although this baseline is not ideal, it does not substantially influence the overall pattern of our results.

We acknowledge these aspects as the limitations of our study and have incorporated them into the Discussion section (line 377-379):

“First, the constraints of bedside iEEG experiments in the hospital (e.g., variable distance between participants and the laptop) may introduce variability in cortical activity and reduce SNR.”

7) The audio was played via speakers. In clinical settings, there is a lot of noise disturbances — how did the authors control for this in the experiment? The paper does not clarify how the sound onset/offset was detected (amplitude threshold? manual annotation?).

A7: You are correct that the audio was played via speakers during movie viewing. The hospital ward door was fully closed throughout the experiment, and the experimenter remained silently at the bedside of the patients. Before the experiment, the experimenter verbally confirmed with each patient that the audio was clearly audible to ensure an optimal movie-viewing experience. For all included patients, the experiment was not intervened by clinicians or relatives. Therefore, we believe that ambient noise was minimal and did not significantly disturb the patients.

Regarding sound detection, we used the Praat software to manually annotate the onset and offset of each trial (Fig. R9). By zooming in on the waveform and spectrogram, the onsets and offsets were clearly marked across trials (please see also A5 to Reviewer #4). For a small number of trials with background sounds, an experimenter manually verified the onset and offset to ensure completeness and avoid redundancy. These procedures have been clarified in the Methods section of the revised manuscript (line 497-500):

“Specifically, for the conditions with sound (i.e., EN, FL, and OS), we manually annotated the trial onsets and offsets based on the sound waveform and spectrogram using the Praat software (<https://www.fon.hum.uva.nl/praat/>).”

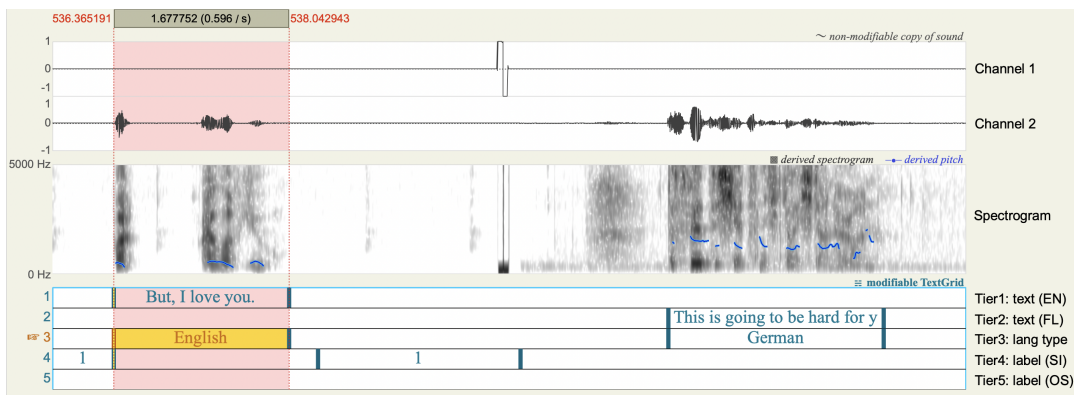


Figure R9 Screenshot of the Praat software used for movie annotation.

8) In the FL condition, subtitles are present, introducing a visual linguistic element absent in EN and OS. This raises

the possibility that observed neural differences reflect subtitle reading rather than foreign speech processing per se. The authors do control for these in the gradient analysis.

A8: We sincerely appreciate your acknowledgment of our approach to controlling for the subtitle effect.

9) Baseline issues: "Silent black scenes" as a baseline may retain residual activity from prior scenes; a 30 s window would ideally be good as audiovisual information produces emotional experience in viewers.

A9: We agree that using silent black scenes as the baseline may be biased by the residual activity from preceding content, and that a longer window would provide a more stable reference. Unfortunately, the movie did not include sufficiently long intervals for such a baseline, and we also did not insert a pre- or post-movie baseline period as in some related studies [8]. We have mentioned this limitation in the revised manuscript.

Nonetheless, we believe our results are unlikely to be substantially influenced by this baseline. For the active electrode selection and the NMF analysis, we employed a phase-randomization-based permutation procedure that preserves the power spectrum while disrupting temporal alignment [29], thereby minimizing baseline effects. For the mTRF encoding analysis, we z-scored the neural signals prior to training, ensuring encoding performance independent of baseline levels.

10) Line 422: "Electrodes showing epileptiform activity, line noise artifacts, or large amplitude shifts were excluded." In the preprocessing, line noise removal is not mentioned. Furthermore, how was the line noise harmonic of 120 Hz removed in the logarithmic frequency bins for computing the high-frequency activity?

A10: We thank you for this important comment. We agree that 60 Hz line noise contaminates neural signals and its harmonic at 120 Hz may affect the high-gamma (70–150 Hz) activity. To this aim, we first manually excluded electrodes showing obvious 60 Hz periodic patterns. We then applied the common average referencing (CAR) method to remove global noises shared across electrodes. We found that the 60 Hz line noise and its harmonics were strongly attenuated after CAR (Fig. R10A, B). This result is expected as line noise is typically the most common source of noise across electrodes.

To further validate our method, we applied the spectrotemporal receptive field (STRF) model to high-gamma activity based on two extraction approaches. In the original approach (as described in the manuscript), the high-gamma band was divided into 8 logarithmic frequency bins (70, 76.997, 84.693, 93.158, 102.469, 112.712, 123.978, 136.369, 150 Hz), with the analytic amplitude in each bin z-scored and averaged. In an alternative approach that explicitly excluded the 120 Hz component, we split the high-gamma band into two sub-ranges with proportional bins (70–119 Hz with 6 bins: 70, 76.473, 83.544, 91.269, 99.708, 108.928, 119 Hz; 121–150 Hz with 2 bins: 121, 134.722, 150 Hz). The analytic amplitude in each bin was also z-scored and averaged. STRF results from three example patients exhibited no significant differences in spatial patterns (Fig. R10C, compare the left and right panels) or prediction accuracy (Fig. R10D, r-values lying along the diagonal) between the two high-gamma extraction approaches.

Together, these findings indicate that (1) 60 Hz line noise and its 120 Hz harmonic were effectively attenuated by CAR, and (2) removing the 120 Hz component from the high-gamma activity does not significantly affect the encoding performance. We therefore did not explicitly remove the 120 Hz harmonic from the high-gamma activity in the revised manuscript, given this is not a standardized method in the field.

11) The rationale for including very short trials (<1 s) is noted ("data stability and robustness"), but such claims might need some supplementary analyses with different trial lengths. Furthermore, since these are all different segments of the movie, showing averaged time plots (Fig. 1 and 2) for trial-based averaging may not be the best choice. In this case, scatter bar plots with the mean activity would be more optimal. Although from the current plots audiovisual activity of electrodes is apparent.

A11: We apologize for the typo and the resulting confusion. The correct statement is: "Trials shorter than one second were **excluded** to ensure data stability and robustness" (please see also A10 to Reviewer #1). Consequently, all included trials are longer than one second, ensuring the use of line plots to display averaged neural activity within this time window is appropriate and optimal.

12) The 1-second lag window (τ_{max}) is arbitrary, potentially missing early (<100 ms) or late (>1 s) responses. This fixed limit is not empirically justified. Naturalistic processing can span both faster dynamics (e.g., ~ 100 ms auditory evoked responses) and much slower ones (e.g., multi-second narrative integration). Such a window may miss sustained attention effects or late semantic processing.

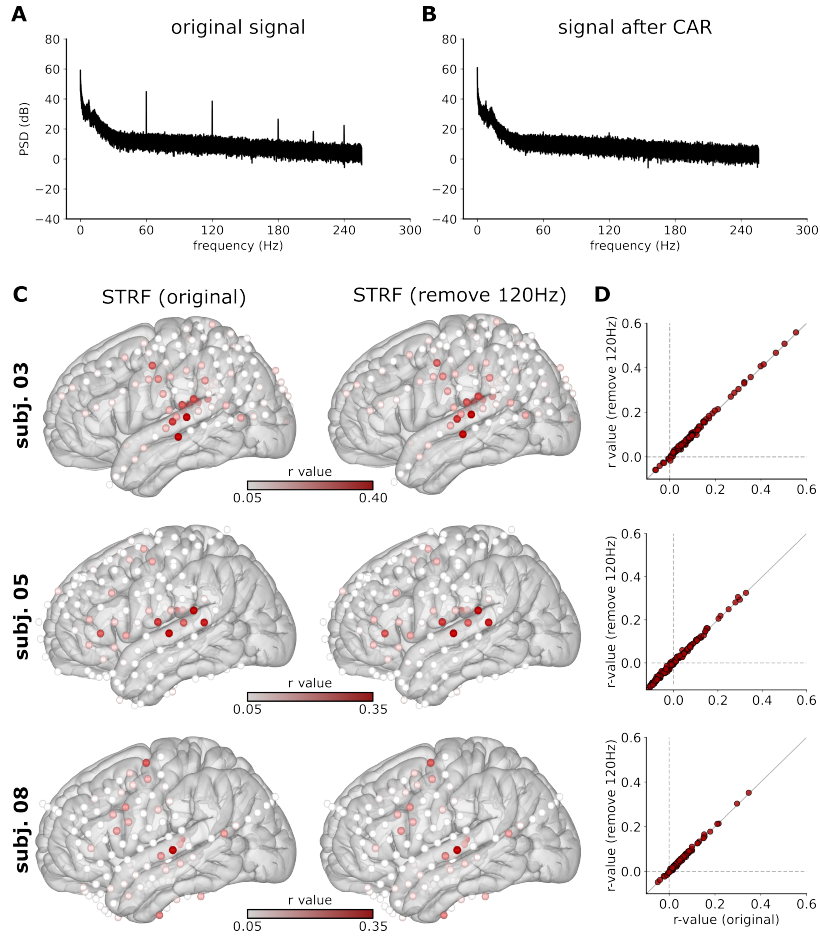


Figure R10 (A) The power spectral density (PSD) of the neural signal from an example electrode, estimated using Welch’s method to reduce noise. (B) The PSD of the same electrode after applying the common average referencing (CAR). (C) STRF brain maps from three example patients, comparing high-gamma activity that either include the 120 Hz harmonic (left panels) or exclude it (right panels). (D) Comparison of r values between the two STRF models.

A12: We understand your concerns regarding the lag window and thank you for the thoughtful comment. Determining an appropriate window length for mTRF models is non-trivial. On the one hand, this window should be long enough to capture most cognitive processes engaged during naturalistic processing, including early auditory-evoked and late semantic-related responses. On the other hand, increasing the window length raises the number of model parameters, which potentially leads to overfitting. Considering these two factors, we chose a 1-second lag window for two main reasons. First, only trials longer than 1 second were included (please see A11), ensuring sufficient data for model estimation within this time window while avoiding overfitting. Second, a 1-second window is sufficient to capture key neural processes related to audiovisual processing, including early sensory responses (50–200 ms) [30, 31], audiovisual integration (100–300 ms) [32, 33, 34], and semantic processing (~400 ms) [35]. Our timing analyses of the mTRF models also corroborate these temporal profiles (please see Fig. 3E and Fig. S10B in the manuscript).

Furthermore, We tentatively investigated the additional cognitive processes you mentioned:

Sustained attentional effect. We acknowledge that attention could be sustained throughout the movie and might act as a potential confounding factor. To address this, we conducted control analyses and revealed that the sustained attention did not significantly influence our findings (please see A2 to Reviewer #1 and A2 to Reviewer #4).

Late semantic processing. Given that late semantic processing involves accumulating information over longer time windows (resembling the concept of the TRW, please see also A7 to Reviewer #1), we hypothesized that extending the lag window would improve model performance if modality-specific semantics are sustained. To test this, we trained mTRF models with progressively increasing lag windows (0–600 ms, 0–800 ms, 0–1000 ms, and –100–1100 ms) for both auditory and visual modalities. The results showed no systematic changes across window sizes (r values clustered around the diagonal; Fig. R11), indicating that extending the window does not substantially

affect model performance. This suggests that modality-specific semantic responses are not sustained during movie viewing. Although this appears inconsistent with prior work on narrative integration over multi-second to minute-long timescales [10], we speculate that sustained processing is more likely to occur for the integrated audiovisual semantics across the narrative, rather than for the semantics in each separate modality. We believe this remains a valuable open question for naturalistic processing and temporal hierarchy.

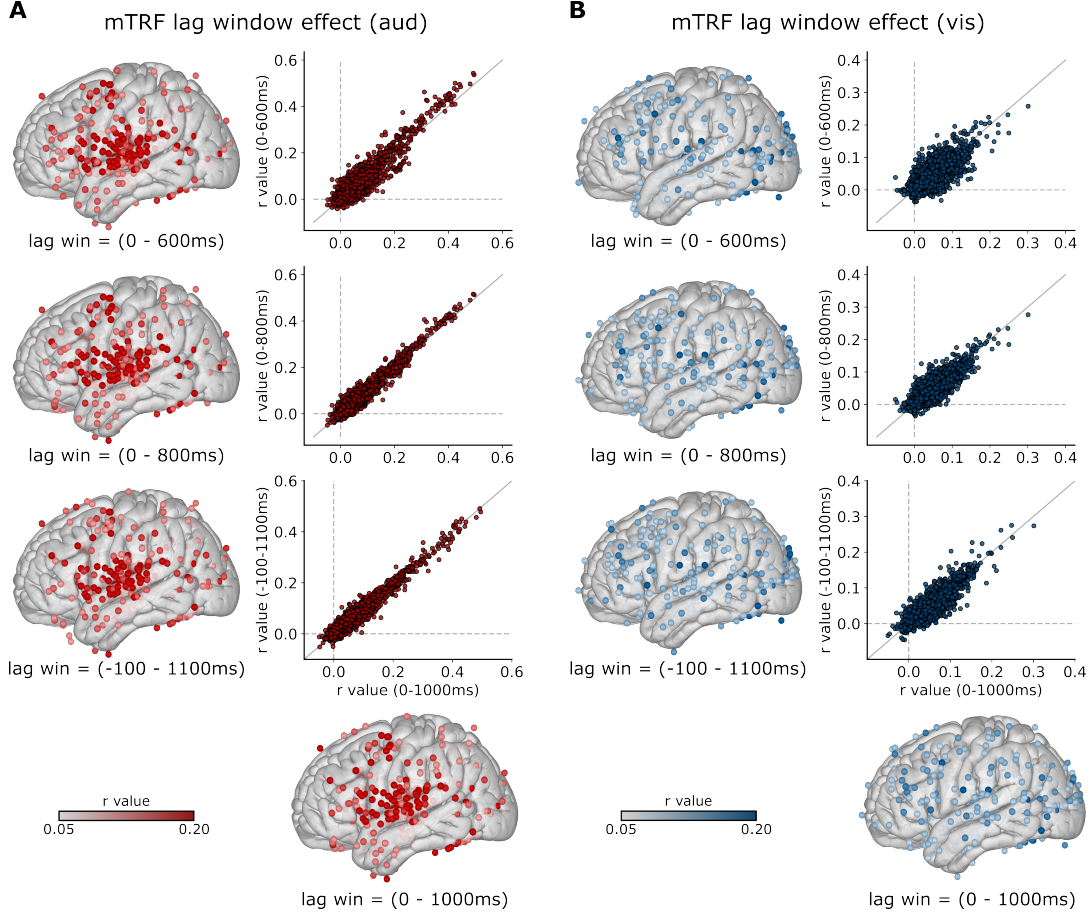


Figure R11 Brain maps and r -value comparisons across increasing lag windows (0–600 ms, 0–800 ms, 0–1000 ms, and –100–1100 ms) for both auditory (A) and visual (B) mTRF models. No permutation test or multiple correction were applied; a hard threshold of $r > 0.1$ was used for visualization.

13) The criterion of consecutive significant time points likely favours electrodes with prolonged activity (e.g., high-gamma during ongoing processing), potentially missing brief yet meaningful events such as early evoked potentials. Analyses such as peak amplitude detection may be useful.

A13: We agree with you that our original approach (i.e., the “consecutive time-point method”) may bias the detection of active electrodes toward those with prolonged activity, while potentially overlooking electrodes with transient but strong responses. To address this limitation, we adopted a T -statistic-based permutation procedure (i.e., the “ T -statistic-based method”), which considers the accumulated T -statistic (i.e., the sum of t -values) within a cluster rather than simply counting the number of consecutive significant time points [36]. Unlike the consecutive time-point method, this approach is capable of capturing both transient but strong responses (few larger T -values) and prolonged but relatively weaker responses (many smaller T -values). To assess significance, we randomized the phases of neural signals (1000 iterations) to generate a null distribution of accumulated T -values and computed p -values based on the position of the actual T -value value in this null distribution.

By comparing the active electrode maps derived from the two methods, we found that the distribution of significant electrodes were highly similar (Fig. R12A), and significant electrodes detected by both methods constituted the majority (80.6% for the consecutive method; 76.2% for the T -statistic method; Fig. R11B). Additionally, the active

maps for the EN and FL conditions remained largely unchanged (Fig. R12C, D, upper panels). In contrast, the numbers of significant electrodes in the OS and SI conditions were slightly reduced under the *T*-statistic-based method (Fig. R12C, D, lower panels). To assess robustness of our findings, we further compared the NMF results based on active maps derived from the two methods. The resulting patterns were also highly consistent (Fig. R13A, B). Taken together, these results indicate that the two permutation procedures produced no substantial differences. Therefore, we retained the original results in the revised manuscript (i.e., the consecutive time-point method) for consistency.

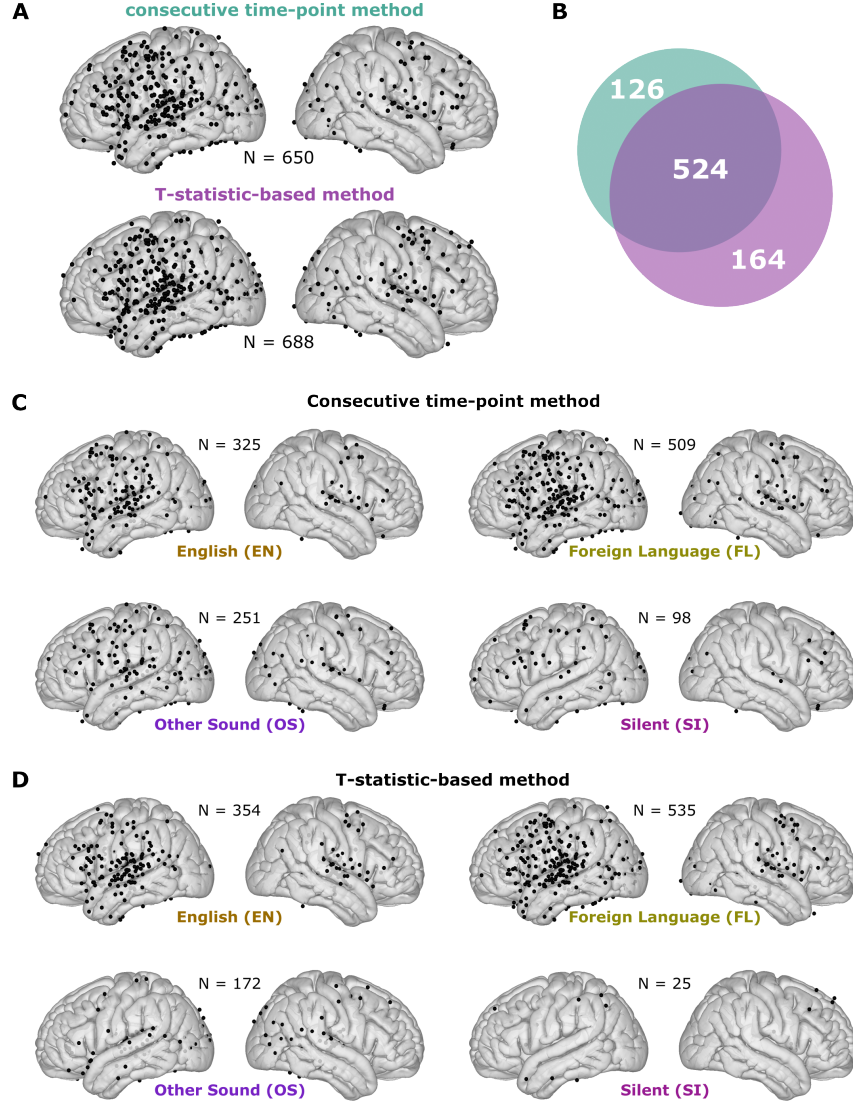


Figure R12 (A) The union brain maps of the active electrodes across all four conditions identified by the consecutive time-point method (upper panel) and the *T*-statistic-based method (lower panel). (B) Venn diagram showing the overlap of active electrodes between the two methods. (C, D) Active electrodes maps for the four conditions separately.

14) Gabor filters are good for local motion but less sensitive to global motion (e.g., camera panning) or higher-order visual features. Some higher-level visual features of salience could also be used.

A14: The motion energy features were derived from a pyramid of spatio-temporal Gabor filters applied across multiple spatial and temporal frequencies, motion directions, x-y positions, and filter sizes [37]. As the filters span both small and large receptive field sizes, it should be sensitive not only to local changes but also to global motions such as camera panning. To test this idea, we tentatively selected a movie clip containing camera panning (Fig. R14A) and

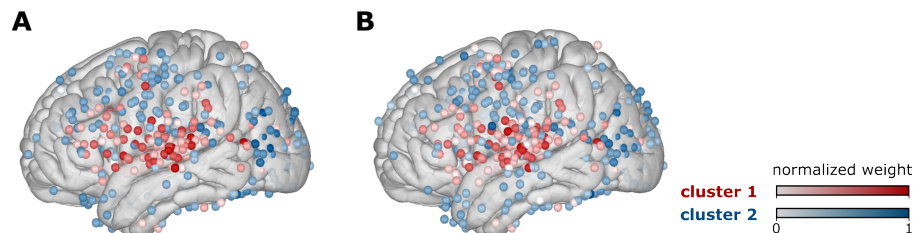


Figure R13 NMF results based on active brain maps derived from either the consecutive time-point method (A) or the T -statistic-based method (B).

visualized the motion energy features of these frames (Fig. R14B). We found that some dimensions of the features exhibited a gradual shift across frames, and the covariance matrix further showed that only adjacent frames were highly correlated (Fig. R14C). These results collectively suggest that the motion energy feature is sensitive to the global motion. We have clarified this point in the revised manuscript (line 551-552):

“This filter bank spans both small and large receptive field sizes, making it sensitive to both local changes and global motions.”

In the current study, motion energy embeddings were used to capture the low-level visual properties of the movie. Regarding higher-level features, we resorted to the vision transformers (ViT) model, a state-of-the-art deep learning model that has been widely recognized for its outstanding performance in tasks such as semantic segmentation [38, 39, 40]. We believe that combining motion energy with ViT-based features provides a comprehensive representation of the movie’s visual content.

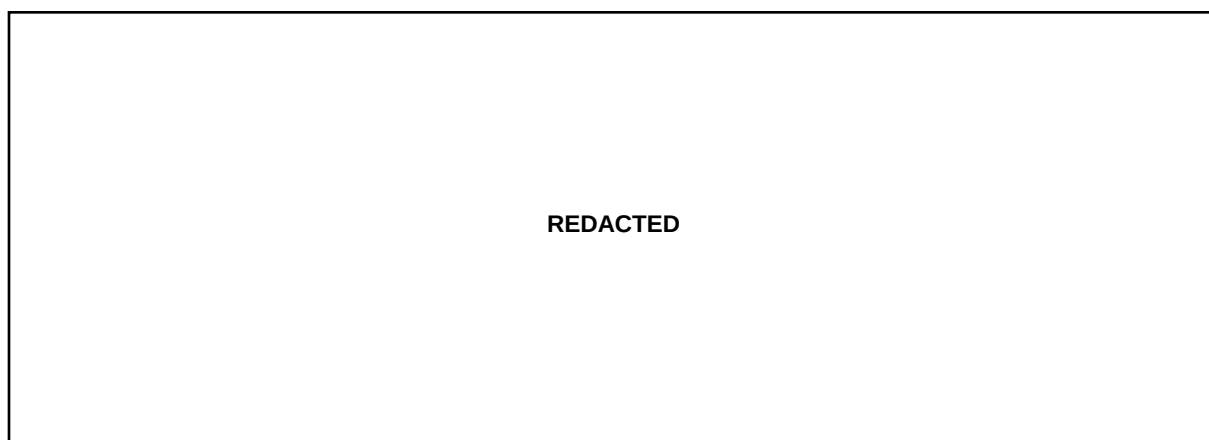


Figure R14 (A) Example frames from a movie snippet illustrating the camera panning. (B) Gabor-based motion energy features across these frames. (C) Covariance matrix of the motion energy features over time.

15) The 20 Hz–8 kHz spectrogram range is unnecessarily broad, extending into ultrasonic frequencies (>20 kHz), irrelevant for human hearing.

A15: We apologize for the confusion. The spectrogram range “20 Hz-8 kHz” represents “20-8000 Hz”, rather than “20-8 kHz”. Therefore, the spectrogram range we used does not extend into ultrasonic frequencies. We have updated the manuscript to explicitly state “20–8000 Hz” for improved clarity and readability (line 547).

16) In Fig. 3, there seem to be a lot of common electrode contact positions assigned to both clusters. Can a map of common positions be shown where the contacts are active for both audio and visual processing?

A16: We thank you for this helpful suggestion and have visualized the electrodes responsive to both visual and

auditory modalities (Fig. R15). In addition to the sparse representation in the frontal region, we also observed the overlapping electrodes in the middle STG. This finding is consistent with previous evidence indicating that superior temporal cortex serves as a pivotal hub for audiovisual processing [41, 42, 43]. We have incorporated this result into the revised manuscript (line 164-166, Fig. S4B):

“The proportion of electrodes tuned to both modalities was relatively low (14.483%) and these electrodes were primarily located in the middle STG, with some scattered across the frontal regions (Fig. S4B).”

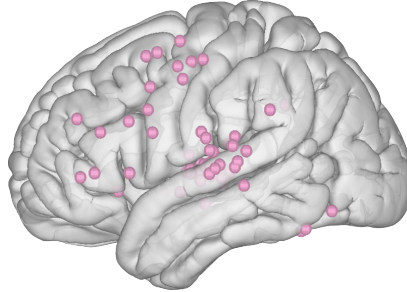


Figure R15 Brain map showing significant electrodes for both audio and visual mTRF models.

17) For the functional clustering, I am not sure if I completely understand whether the clustering for electrode contacts accounts for contacts with multimodal tuning to both visual and auditory information. I understand the clustering is based on the weights, but does the clustering result then assume/imply that auditory and visual tuning are mutually exclusive?

A17: We understand your concern. Theoretically, NMF is a dimensionality reduction method that decomposes the data into part-based representations due to its non-negativity constraints [44]. This method does not necessarily guarantee the orthogonality, and thus the clusters are not mutually exclusive (but see other NMF variants such as ONMF [45]). To address this issue, we visualized the two clusters and found that they were largely independent (Fig. R16). Electrodes with higher weights in one cluster generally exhibited lower weights in the other (e.g., the STG electrodes). This pattern empirically supports using NMF weights for functional clustering, and we have clarified this in the Supplementary Materials and according Methods section (line 537-539; Fig. S3D):

“Further, we observed that the two factorized clusters were largely independent with minimal overlap (Fig. S3D). Therefore, electrodes were assigned to clusters based on their maximum contributions in the weighting matrix $\mathbf{W}$, and were normalized (0-1) for visualization purposes (Fig. 2A).”

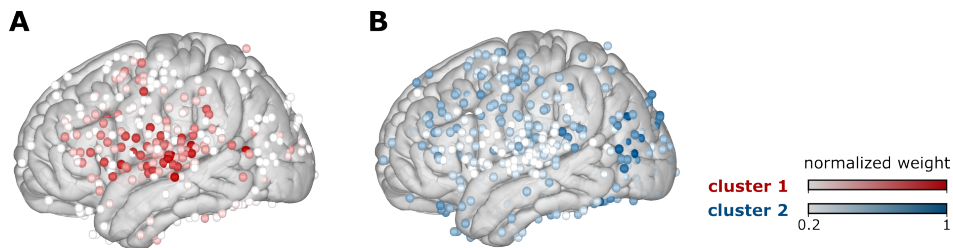


Figure R16 The NMF brain maps for auditory (A) and visual (B) clusters are shown separately.

18) Similarly, I am wondering if using mTRF r -values from models trained and tested on the same dataset may overfit? Furthermore, audio and visual mTRFs were trained separately, which will then ignore crossmodal integration effects and artificially segregate audiovisual activity. Would it be wise to then run a combined model with both feature sets entered simultaneously?

A18: We fully agree with you that preventing overfitting in the mTRF model is essential for robust and stable results. To this end, we did not use r values from models trained and tested on the same dataset. Instead, we applied a 4-fold cross-validation (CV) procedure in which the data were split into four groups—three used for training and the remaining one for testing. This procedure ensures that training and testing were performed on independent data, thereby minimizing overfitting while efficiently utilizing the neural data.

Additionally, we also greatly appreciate your thoughtful suggestion to consider a combined model incorporating both auditory and visual features. One potential advantage of such a model is that it might capture the audiovisual integration. However, after careful examination of prior studies, we believe that this claim is not well supported and somewhat speculative. Originally, this idea stems from classical multimodal experiments that typically include three conditions: audio-only (A), visual-only (V), and audiovisual (AV). Then, integration effect can be isolated using contrasts between conditions (such as $AV > A + V$) [46]. This approach is justified because task performances often improved in the AV conditions, indicating that crossmodal integration is involved in the brain [2, 34]. However, this comparison is not applicable to the corresponding three model types (i.e., M_A , M_V , and M_{AV}). Mathematically, M_{AV} is the linear combination of M_A and M_V , with its performance constrained such that $r_{AV} \leq r_A + r_V$ (the equality only achieved when the auditory and visual features are uncorrelated). In other words, r_{AV} cannot exceed the sum of r_A and r_V and therefore does not reflect integration effects (please see also A3 to Reviewer #4 for more discussion).

Another potential benefit of a combined model is to derive unique auditory or visual brain maps by controlling for the other modality. However, one of our main objectives was to characterize the full involvement of brain regions associated with each modality during naturalistic movie viewing. Since auditory and visual features often exhibit certain correlations in such complex stimuli, regressing out one modality likely overlooks meaningful crossmodal representations, such as visual cortex responding to sound and vice versa [47, 48]. Additionally, we believe that introducing a combined model (M_{AV}) would not substantially enhance the frontal gradient analysis either. In practice, we have introduced an audiovisual index to quantify the modality selectivity ($I_{AV} = (r_V - r_A)/(r_V + r_A)$), where r_V and r_A are derived from separate auditory and visual models. We consider this index a stricter measure, because the numerator ($r_V - r_A$) reflects relative differences between modalities and is unaffected by r_{AV} , while the denominator ($r_V + r_A$) provides an upper bound independent of r_{AV} .

Taken together, while a combined model is conceptually appealing, we reason that it would likely not yield additional insights beyond those provided by our current approach. We hope our rationale clarifies our methodological choice and thank the reviewer again for raising this valuable point.

19) The need for a new coordinate system is unclear, as MNI coordinates would suffice for gradient-based analyses. A19: We agree that a clear justification for our new coordinate system is necessary. Our motivation for this system was to quantify the ventrodorsal extent in the frontal cortex. A common segmentation approach was based on cytoarchitecture and connectivity (Fig. R17A), where dlPFC corresponds to Brodmann areas (BA) 46 and 9/46 and vlPFC corresponds to BA 45 and 47/12 [49, 50]. Anatomically, this separation aligns with the division between the inferior frontal gyrus (IFG) and middle frontal gyrus (MFG), separated by the inferior frontal sulcus (IFS). Importantly, this boundary is not strictly cardinal (i.e., horizontal or vertical) but instead forms an oblique line, with regions above assigning to dlPFC and those below to vlPFC (Fig. R17B). As a result, relying solely on either the y or z MNI coordinate cannot adequately capture the ventrodorsal distinction, whereas combining y and z provides a more accurate index.

This motivated our construction of a polar coordinate system in the frontal cortex. Conceptually, the method is straightforward: by defining an appropriate origin (y_0, z_0), we quantify the ventrodorsal position for each electrode by computing its radius based on the MNI coordinate (y_1, z_1) within the frontal sector (Fig. R17C). Therefore, this new coordinate system can be viewed as a reparameterization of the standard MNI system that better captures the ventrodorsal gradient. Practically, we approximated the origin at the intersection of the precentral sulcus and Sylvian fissure (origin MNI: $[-55, 15, -8]$; the x coordinate was not critical in this analysis).

An additional advantage of this approach is that we did not restrict the analysis to predefined anatomical areas (e.g., vlPFC or dlPFC). Prior evidence has shown that auditory information can be represented in dlPFC (e.g., speech syntax [51]), while visual information can be localized in vlPFC (e.g., facial information [14, 15]). Thus, functional gradients in frontal cortex do not always align with anatomical divisions. By constructing a continuous ventrodorsal index from our coordinate system, we are able to quantify modality transition trends in a data-driven manner without imposing rigid anatomical assumptions. We therefore believe this approach is particularly well suited for our analysis and have modified the Discussion (line 347-350) and Methods (lines 660-674) sections accordingly.

20) Demographics table of patients and their diagnosed epileptogenic zones and the contacts which were excluded, should be provided.

A20: We thank you for this valuable comment. We have expanded the demographic table in the Supplementary Materials to include each patient’s seizure onset zone (SOZ) based on the clinicians’ reports (Table R1). In addition,

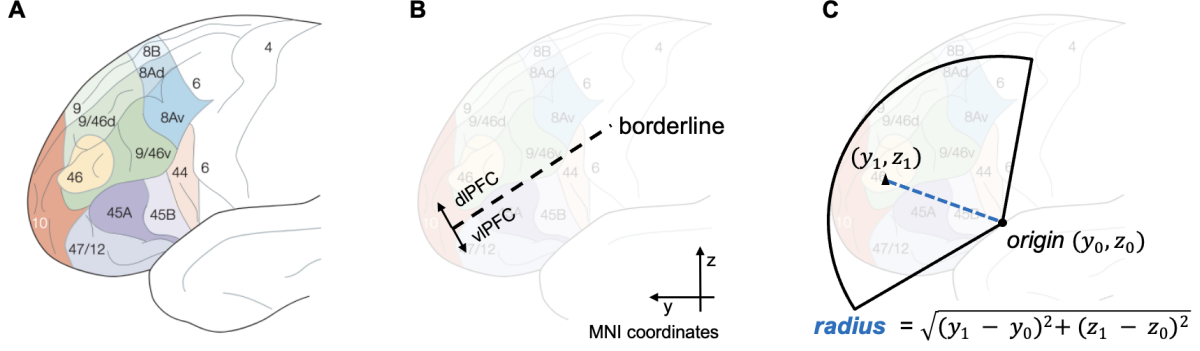


Figure R17 (A) Brodmann area (BA) map for the lateral prefrontal cortex, reprinted from [52]. (B) A putative borderline distinguishing the vIPFC and dIPFC. (C) The polar coordinate system we employed to quantify the continuous ventrodorsal gradient.

we provided brain maps showing the SOZ electrodes (Fig. R18A) and the included electrodes excluding the SOZ electrodes (Fig. R18B). These updates have been incorporated into the revised manuscript (Fig. S1A and table R1).

Table R1: Patient information

ID	Coverage	Gender	Handedness	Age at Surgery (y)	SOZ
1	Right	Male	Right	29	temporal
2	Left	Female	Right	45	insula/frontal
3	Left	Male	Right	24	parietal
4	Left	Female	Right	30	temporal
5	Left	Female	Right	56	temporal
6	Bilateral	Male	Right	26	temporal
7	Right	Female	Right	21	frontal
8	Left	Male	Right	22	temporal
9	Left	Male	Right	44	temporal
10	Left	Female	Right	33	insula/temporal
11	Left	Female	Right	23	temporal
12	Left	Female	Right	18	occipital
13	Left	Female	Right	61	temporal
14	Bilateral	Female	Right	33	temporal
15	Left	Female	Right	27	temporal/frontal
16	Right	Male	Right	28	insula/temporal/frontal
17	Left	Male	Right	26	parietal
18	Bilateral	Female	Right	33	insula/temporal
19	Bilateral	Female	Right	30	insula/temporal/frontal

Results

21) Figure 1 – Please provide the electrode contacts that were active for each condition with the same colour scheme as in panel C.

A21: We appreciate your suggestion. In fact, we also considered rendering the electrodes by condition in Fig. 1B. However, some electrodes responded to multiple conditions, requiring more than four colors in practice and potentially making the figure overly complex and less interpretable. For this reason, we opted not to color the electrodes in the main figure and instead present the condition-specific activation maps in the Supplementary Materials (Fig. S1).

Discussion

22) This section is extremely short with no synthesis of findings from the extensive results section. Furthermore, the limitations section is missing as well. One suggestion is to also remark on the structural connectivity from visual and auditory regions that might dictate the dorso-ventral gradients. Neuroanatomical and functional knowledge, such as

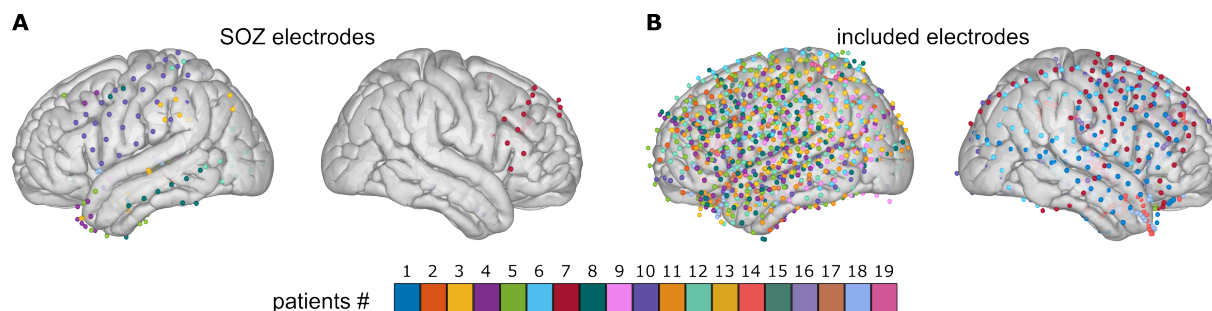


Figure R18 Brain maps showing (A) all SOZ electrodes and (B) all included electrodes in the current study without the SOZ electrodes across 19 patients.

the frontal eye fields, which are dorsal, also need to be discussed.

A22: We agree that the Discussion section should be expanded. Based on your and other reviewer’s constructive comments, we have revised this section accordingly (line 320-376) and added a paragraph outlining the limitations (line 377-384). The revised parts have been marked in purple in the Discussion section.

Minor

Abstract: “Here, we employed movie-watching to”... it should be something like “movie paradigm,” as “watching” might indicate visual information only.

A23: We thank you for raising this question. We have revised the text to use the term “movie-viewing paradigm”, aligning with the terminology and style used throughout the manuscript.

Line 63 – iEEG is intracranial electroencephalography. ECoG is electrocorticography.

A24: We apologize for this obvious mistake and have modified the manuscript accordingly (line 64):

“To bridge these gaps, we examined neural responses in a cohort of neurosurgical patients using intracranial electroencephalography (iEEG) while they watched a multilingual naturalistic movie.”

Reviewer #4

Summary

This work relates cortical responses, especially in the frontal cortex, to various components of a naturalistic audiovisual movie. The dataset is valuable, but I was somewhat lost by the complexity of the analyses, and I am not sure whether the authors' interpretation of their findings is the most plausible one.

We sincerely thank you for recognizing the value of our dataset and for carefully engaging with our work. We apologize if parts of our analyses or interpretations appeared overly complex or unclear. In our revised manuscript, we have clarified the rationale for our analytical choices and provided additional explanations to guide the reader through the main results. We hope these modifications make our interpretations to results more transparent and accessible.

Comments

Major

Large number of analyses: The paper employs a number of (if I understood correctly) independent analysis strategies: unsupervised clustering (Fig. 2), mTRF based on components of the auditory vs. visual components of the movie (Figs. 3 and 4), and an analysis that relies on moment-to-moment ratings of the movie's contents by distinct viewers (Fig. 5). I think I was able to follow rather well up until Figure 4 included. I don't think I understood what Figure 6 represents. More generally, I had a hard time understanding exactly what each analysis was doing and what new information it brought with respect to the previous ones. This can likely be resolved by reformulating the Results section to provide more "plain-language" information.

A1: We apologize for the lack of clarity in our analyses and for any confusion this may have caused. To improve readability, we have reformulated the Results section with clearer, plain-language explanations.

Regarding the rationale for Fig. 6: after observing the dynamic frontal representation across conditions (Fig. 5A), we tentatively proposed the "weight-based modality assignment" principle, which describes that the brain adjusts the relative importance (or weighting) of auditory and visual inputs depending on context (please see A3 for more details). We further reasoned that if the weight-based assignment principle is applicable to audiovisual processing during naturalistic movie viewing, this modality assignment effect (i.e., the global modality feature; Fig. 5) should also be detectable at the neural level. Therefore, we applied an mTRF encoding model to the global modality feature and identified a distinct set of frontal electrodes that appear to carry out the weight-based assignment, separate from those involved in encoding audiovisual features (Fig. 6). We hope this explanation clarifies our motivations, and we have revised the manuscript accordingly to improve accessibility and flow (line 268-272):

"Further, we reasoned that if this weight-based assignment hypothesis is applicable to audiovisual processing during naturalistic movie viewing, this modality assignment effect should also be detectable at the neural level. To this end, we fitted the mTRF models to the audiovisual assignment variable (i.e., global modality; see Methods: Encoding modeling procedure)."

Auditory or visual vs. speech-related signals: I watched the movie that was used as a stimulus. Most of the information therein relates to audiovisual speech (most of the shots are close-ups of the actors' faces); there are a handful of short, wordless scenes that help understand the stories but are not actually essential given all the close-ups and dialogues that precede or follow them. This might be what explains the relative low numbers of electrodes driven by the "other sound" or "silent" periods. More importantly in my opinion, I wonder whether the strong emphasis on close-up dialogue in the film, and the attending speech- and language-heavy contents, really allows the authors to generalize to "auditory" vs. "visual" in their interpretation of how components of the movie drive cortical responses. An alternative explanation, which seems at least as plausible to me, relates these responses to the relative attention directed by the patients to the auditory vs. visual components of speech stimuli depending on whether they could understand what was being said. This is supported by the authors' finding that visual stimulus components seem to be more important in driving cortical responses when the actors speak in a language that is unintelligible to the participants, whereas the auditory stimuli drive responses more when the actors speak English. Perhaps the authors have additional data collected with the same patients to resolve this confound? Additionally, cortical responses from patients who understood one of the "foreign" languages spoken in the movie would be useful.

A2: We sincerely thank you for taking the time to view our experimental stimulus and for providing such insightful comments regarding the role of language in audiovisual processing during naturalistic movie viewing. In response to

your valuable feedback, we provide clarifications and additional analyses below.

First, we agree that the wordless scenes (i.e., the OS and SI conditions) surrounding the dialogue segments (i.e., the EN and FL conditions) are relatively trivial for understanding the storyline. This interpretation is supported by our behavioral results showing that global context ratings (i.e., the importance for understanding the entire movie) were significantly higher for the language conditions compared with the non-language conditions (Fig. 5G). As you have pointed out, this finding is also consistent with our NMF and mTRF results, both of which revealed stronger cortical responses in the language conditions (Fig. 2B and Fig. S5). We believe this pattern is intuitive, as language and dialogue constitute the core components of our movie, conveying the most informative content.

Second, we fully understand your concern that the language-dominant nature of the movie could bias attentional engagement through language intelligibility, thereby limiting the generalizability of our conclusions regarding how auditory and visual information drive cortical responses. Your proposed explanation was indeed one of our initial considerations. After conducting the NMF analysis and observing distinct auditory and visual clusters (Fig. 2), we questioned whether this separation reflected (1) attentional effects induced by language bias or (2) representational differences intrinsic to each modality. To address this, we performed an mTRF analysis using audiovisual stimulus features and observed similar brain patterns as the NMF results (Fig. 3C-D), suggesting that the cortical responses primarily reflect modality-specific representations rather than language-driven attentional effects. To further examine the potential influence of the attention, we applied the mTRF model to identify the brain regions associated with the attentional engagement (Fig R2D; please see also A2 to Reviewer #1) and compared their spatial distribution with audiovisual feature encoding across three ROIs (including the frontal regions, STG, and occipital regions as shown in Fig S1; Fig R19A). We found the r -values from the attention models did not correlate with either auditory or visual r -values in any of the three ROIs (Fig R19B; Pearson correlation; FDR corrected). These results further confirm that the audiovisual representations in the brain are independent from attentional effects.

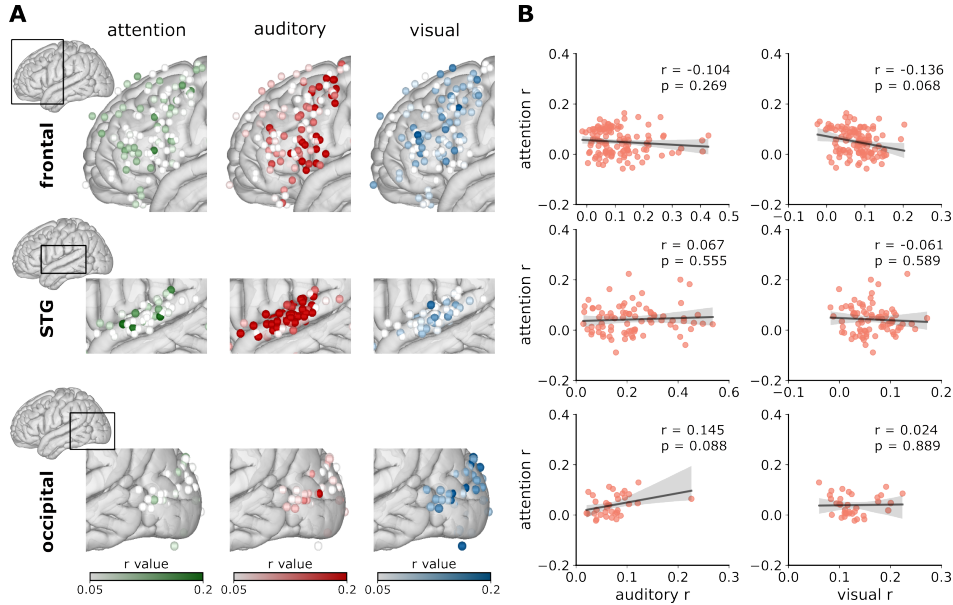


Figure R19 Control analyses for attentional effects. (A) The attention (green), auditory (red), and visual (blue) mTRF results in three ROIs, including the frontal regions (upper panels), STG (middle panels), and occipital regions (lower panels). (B) The r -values from the attentional models did not correlate with either auditory or visual r -values in any of the ROIs.

Nevertheless, we acknowledge that the attention induced by language intelligibility could modulate the frontal audiovisual representations during the movie viewing (Fig. 5A-B) and may be distinct from our attentional behavioral rating. This is consistent with previous evidence showing that selective attention could influence the spectrotemporal representation of the speech [53, 54, 55]. Beyond the attentional effect, we believe that such changes in frontal audiovisual representation could also reflect the multisensory integration effect (please see A3 for more discussion). Unfortunately, our current dataset does not include patients who understood "foreign" languages for further exploration, as patient recruitment in iEEG experiments is highly constrained and cannot be explicitly controlled. We have noted this as a limitation in the revised manuscript (line 379-382) and look forward to future studies employing more

diverse materials (e.g., natural documentaries with or without narration) to further test and extend our conclusions.

Lack of a clear link between cortical activities and behavior: given that the patients did not perform any particular task while watching the movie, it is too much of a stretch to affirm that the frontal cortical activities described in the paper underlie multisensory perception.

A3: We fully understand your concern and acknowledge that our study lacks direct behavioral measures to demonstrate multisensory perception. Previous studies have typically characterized the integration effect during multisensory perception by comparing task performances (e.g., reaction time or accuracy) across audio-only (A), visual-only (V), and audiovisual (AV) conditions (please see also A18 to Reviewer #3) [56, 57]. However, implementing these three conditions in an iEEG experiment, especially under naturalistic processing, is challenging. Using the same movie across multiple conditions risks familiarity effects in later sessions, while employing different movies introduces confounds such as semantic complexity and content richness. Moreover, variability in electrode coverage across patients also makes a between-subject design infeasible. Additionally, unlike in well-controlled psychological experiments, it is difficult to establish objective metrics for evaluating audiovisual integration during movie viewing. However, we believe developing such tools and approaches to quantify the multimodal integration and track audiovisual weighting in real time represents a valuable direction for future research.

In light of these limitations, we have been cautious in our interpretation. Based on the observed dynamic frontal representation across conditions (Fig. 5A in the Manuscript), we postulated that this effect likely reflects either (1) the allocation of attention to each modality (as discussed in A2) or (2) an integration process that fuses the bimodal information. Given that it is difficult to disentangle these two cognitive processes effectively with our current dataset, we intentionally avoided making definitive claims about multisensory integration. Instead, we framed our findings in terms of a “weight-based modality assignment” process (please see also A1) that describes how auditory and visual inputs are differentially weighted depending on context. Computationally, this hypothesis could reflect both the fusion process during multisensory integration [58, 59, 60, 1, 61] and the “gain adjustment” or “re-weighting” associated with the attentional modulation [62, 63, 64], without presupposing whether the underlying mechanism reflects integration, attention, or both. Nonetheless, we believe that distinguishing these cognitive components remains an important direction for future work, particularly under naturalistic scenarios. We have also noted this as a potential limitation in the revised manuscript (line 379-382):

“Second, the absence of behavioral measures from patients (e.g., eye-tracking recordings and direct assessments of movie comprehension) limited our ability to directly account for attention and eye movements.”

Minor

Flexible AV resource assignment (Fig. 5A): wouldn’t this analysis make more sense if only cortical sites that are activated by both English and foreign languages were selected?

A4: We thank you for the suggestion. In practice, only 17 electrodes (20.48%) showed significant encoding in both EN and FL conditions (please see also A16 to Reviewer #3 and Fig. R15), which we consider too few to support a reliable analysis. Moreover, significant electrodes in the auditory mTRF model were primarily localized in the EN condition, whereas those in the visual model were predominantly observed in the FL condition (Fig. S5). Since this effect is driven by both the movie context and modality, we believe it is reasonable that the frontal encoding maps in the two conditions largely differ. We therefore included all significant electrodes for this analysis.

Silent periods: was a volume criterion used to determine when there was no sound during these periods?

A5: We identified the silent and sound periods by leveraging the Praat software (Fig. R20). By inspecting the waveform and spectrogram in the software, the onsets and offsets of the sound period were clearly identifiable (please see also A7 to Reviewer #3). The remaining intervals were then labeled as the silent period. To minimize the potential residual activity, we marked the onsets of silent periods 200-500 ms after the preceding sound offsets. The abrupt change observed in “channel 1” served as the trigger for aligning neural signals and was not explicitly considered during labeling. This information has been incorporated to the Methods section (line 500-502):

“The remaining intervals were then labeled as the silent period. To minimize the potential residual activity, we marked the onsets of silent periods 200-500 ms after the preceding sound offsets.”

Cognitive involvement (results, line 256): I’m not sure that, because viewers rated a movie snippet as more important

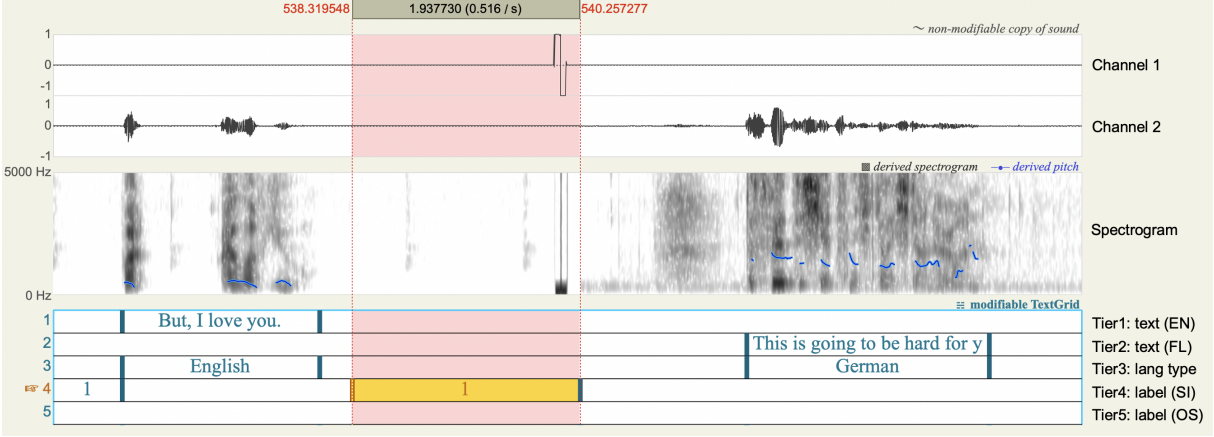


Figure R20 Screenshot of the Praat software used for movie annotation.

for understanding the whole movie, it should mean that this snippet "demanded greater cognitive involvement".

A6: We agree that the "cognitive involvement" may not be the most precise term. We have removed this jargon and rephrased the sentence (line 261-263):

"This result is consistent with our NMF and mTRF findings showing stronger responses for the language conditions (i.e., English and Foreign Language; Fig. 2B and S5)."

Numbers of principal components (Fig. S4): given that only multiples of 5 PCs were explored, I suggest foregoing the continuous line and showing only the mean r -values at the actual counts of PCs investigated. Another question on PCs: was 15 the optimal number of PCs for both the auditory and the visual model? Is there anything of interest in such a coincidence?

A7: Thank you for this helpful comment. In response to your first suggestion, we have revised the Fig. S4A by replacing the continuous line plot with the error-bar plot that displays the mean r -values at the actual counts of PCs (Fig. R21A).

Regarding your second question, we analyzed the auditory and visual mTRF models separately using the incremental PCs test. Interestingly, both models showed elbow points around 15 PCs (Fig. R21B, C). We believe, however, this is not a coincidence. A possible explanation is that the capacity of the mTRF model is constrained by several practical factors (such as the size of training data, the noise level of the neural activity, the length of the lag window, etc.), which collectively impose an upper bound on the feature dimensions that can stably capture the possible effects before overfitting occurs. Therefore, despite the features from the two modalities differs, both models appear to converge on a similar dimensionality constraint.

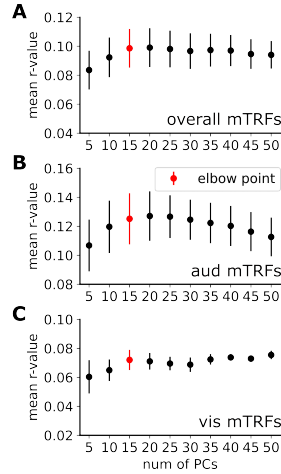


Figure R21 Results of the incremental principal components (PCs) test for all the mTRF models (A), auditory mTRF models (B), and visual mTRF models (C). The red error bars indicate the elbow points identified using kneed toolbox.

References

- [1] Tim Rohe and Uta Noppeney. Cortical hierarchies perform bayesian causal inference in multisensory perception. 13(2):e1002073.
- [2] Ambra Ferrari and Uta Noppeney. Attention controls multisensory perception via two distinct mechanisms at different levels of the cortical hierarchy. 19(11):e3001465.
- [3] Hayoung Song, Emily S. Finn, and Monica D. Rosenberg. Neural signatures of attentional engagement during narratives and its consequences for event memory. 118(33):e2021905118.
- [4] Maximilian Nentwich, Marcin Leszczynski, Brian E. Russ, Lukas Hirsch, Noah Markowitz, Kaustubh Sapru, Charles E. Schroeder, Ashesh D. Mehta, Stephan Bickel, and Lucas C. Parra. Semantic novelty modulates neural responses to visual change across the human brain. 14(1):2910.
- [5] Evelina Fedorenko, Terri L. Scott, Peter Brunner, William G. Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. Neural correlate of the construction of sentence meaning. 113(41):E6256–E6262.
- [6] L. S. Hamilton, E. Edwards, and E. F. Chang. A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. 28(12):1860–+.
- [7] Sam V. Norman-Haignere, Laura K. Long, Orrin Devinsky, Werner Doyle, Ifeoma Irobunda, Edward M. Merriks, Neil A. Feldstein, Guy M. McKhann, Catherine A. Schevon, Adeen Flinker, and Nima Mesgarani. Multi-scale temporal integration organizes hierarchical computation in human auditory cortex. 6(3):455–469.
- [8] Y. Lerner, C. J. Honey, L. J. Silbert, and U. Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. 31(8):2906–2915.
- [9] Christopher J. Honey, Thomas Thesen, Tobias H. Donner, Lauren J. Silbert, Chad E. Carlson, Orrin Devinsky, Werner K. Doyle, Nava Rubin, David J. Heeger, and Uri Hasson. Slow cortical dynamics and the accumulation of information over long timescales. 76(2):423–434.
- [10] U. Hasson, J. Chen, and C. J. Honey. Hierarchical process memory: memory as an integral component of information processing. 19(6):304–313.
- [11] Y. Yeshurun, M. Nguyen, and U. Hasson. Amplification of local changes along the timescale processing hierarchy. 114(35):9475–9480.

- [12] Jie Zheng, Andrea G. P. Schjetnan, Mar Yebra, Bernard A. Gomes, Clayton P. Mosher, Suneil K. Kalia, Taufik A. Valiante, Adam N. Mamelak, Gabriel Kreiman, and Ueli Rutishauser. Neurons detect cognitive boundaries to structure episodic memories in humans. 25(3):358–368.
- [13] J. M. Clarke, E. Halgren, J.-M. Scarabin, and P. Chauvel. Auditory and visual sensory representations in human prefrontal cortex as revealed by stimulus-evoked spike-wave complexes. 118(2):473–484.
- [14] L. M. Romanski. Domain specificity in the primate prefrontal cortex. 4(4):421–429.
- [15] L. M. Romanski. Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex. 17:i61–i69.
- [16] Gemma A. Calvert, Ruth Campbell, and Michael J. Brammer. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. 10(11):649–657.
- [17] Gemma A. Calvert, Peter C. Hansen, Susan D. Iversen, and Michael J. Brammer. Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. 14(2):427–438.
- [18] G. A. Calvert. Crossmodal processing in the human brain: Insights from functional neuroimaging studies. 11(12):1110–1123.
- [19] R. M. Braga, P. J. Hellyer, R. J. S. Wise, and R. Leech. Auditory and visual connectivity gradients in frontoparietal cortex.
- [20] Majd Abdallah, Gaston E Zanitti, Valentin Iovene, and Demian Wassermann. Functional gradients in the human lateral prefrontal cortex revealed by a comprehensive coordinate-based meta-analysis. 11:e76926.
- [21] Yuval Nir, Lior Fisch, Roy Mukamel, Hagar Gelbard-Sagiv, Amos Arieli, Itzhak Fried, and Rafael Malach. Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. 17(15):1275–1285.
- [22] Andreas Bartels and Semir Zeki. Functional brain mapping during free viewing of natural scenes. 21(2):75–85.
- [23] Camilla Bellini, Nicola Del Maschio, Marco Gentile, Gianpaolo Del Mauro, Rita Franceschini, and Jubin Abutalebi. Original language versus dubbed movies: Effects on our brain and emotions. 253:105424.
- [24] Leonardo Muller-Rodriguez. Neural association between musical features and shared emotional perception while movie-watching: fMRI study.
- [25] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. Intersubject synchronization of cortical activity during natural vision. 303(5664):1634–1640.
- [26] Junwei Han, Xiang Ji, Xintao Hu, Lei Guo, and Tianming Liu. Arousal recognition using audio-visual features and FMRI-based brain response. 6(4):337–347.
- [27] Yoshiaki Tsushima, Koharu Nakayama, Teruhisa Okuya, Hiroko Koiwa, Hiroshi Ando, and Yoshiaki Watanabe. Brain activities in the auditory area and insula represent stimuli evoking emotional response. 14(1):21335.
- [28] Pouya Ghaemmaghami, Mojtaba Khomami Abadi, Seyed Mostafa Kia, Paolo Avesani, and Nicu Sebe. Movie genre classification by exploiting MEG brain signals. In Vittorio Murino and Enrico Puppo, editors, *Image Analysis and Processing — ICIAP 2015*, volume 9279, pages 683–693. Springer International Publishing. Series Title: Lecture Notes in Computer Science.
- [29] S. A. Nastase, V. Gazzola, U. Hasson, and C. Keysers. Measuring shared responses across subjects using intersubject correlation. 14(6):669–687.
- [30] Liberty S Hamilton, Yulia Oganian, Jeffery Hall, and Edward F Chang. Parallel and distributed encoding of speech across human auditory cortex. 184:1–14.
- [31] Iris I. A. Groen, Giovanni Piantoni, Stephanie Montenegro, Adeen Flinker, Sasha Devore, Orrin Devinsky, Werner Doyle, Patricia Dugan, Daniel Friedman, Nick F. Ramsey, Natalia Petridou, and Jonathan Winawer. Temporal dynamics of neural responses in human visual cortex. 42(40):7562–7580.

- [32] Stephanie C. Boyle, Stephanie J. Kayser, and Christoph Kayser. Neural correlates of multisensory reliability and perceptual weights emerge at early latencies during audio-visual integration. 46(10):2565–2577.
- [33] Máté Aller and Uta Noppeney. To integrate or not to integrate: Temporal dynamics of hierarchical bayesian causal inference. 17(4):e3000210.
- [34] Tim Rohe, Ann-Christine Ehlis, and Uta Noppeney. The neural dynamics of hierarchical bayesian causal inference in multisensory perception. 10(1):1907.
- [35] M. Kutas and K. D. Federmeier. Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (ERP). In S. T. Fiske, D. L. Schacter, and S. E. Taylor, editors, *Annual Review of Psychology*, volume 62 of *Annual Review of Psychology*, pages 621–647.
- [36] E. Maris and R. Oostenveld. Nonparametric statistical testing of EEG- and MEG-data. 164(1):177–190.
- [37] Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. 21(19):1641–1646.
- [38] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, and Yifan Liu. SegViT: Semantic segmentation with plain vision transformers.
- [39] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. On the effectiveness of ViT features as local semantic descriptors. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, volume 13804, pages 39–55. Springer Nature Switzerland. Series Title: Lecture Notes in Computer Science.
- [40] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. 126:106669.
- [41] Michael S Beauchamp. See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. 15(2):145–153.
- [42] Michael S Beauchamp, Kathryn E Lee, Brenna D Argall, and Alex Martin. Integration of auditory and visual information about objects in superior temporal sulcus. 41(5):809–823.
- [43] Francesca Setti, Giacomo Handjaras, Davide Bottari, Andrea Leo, Matteo Diano, Valentina Bruno, Carla Tinti, Luca Cecchetti, Francesca Garbarini, Pietro Pietrini, and Emiliano Ricciardi. A modality-independent proto-organization of human multisensory areas.
- [44] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. 401(6755):788–791.
- [45] Amirhossein Khalilian-Gourtani, Ran Wang, Xupeng Chen, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. A corollary discharge circuit in human speech. 121(50).
- [46] Chuanji Gao, Jessica J Green, Xuan Yang, Sewon Oh, Jongwan Kim, and Svetlana V Shinkareva. Audiovisual integration in the human brain: a coordinate-based meta-analysis. 33(9):5574–5584.
- [47] David Brang, John Plass, Aleksandra Sherman, William C. Stacey, Vibhangini S. Wasade, Marcia Grabowecky, EunSeon Ahn, Vernon L. Towle, James X. Tao, Shasha Wu, Naoum P. Issa, and Satoru Suzuki. Visual cortex responds to sound onset and offset during passive listening. 127(6):1547–1563.
- [48] Ganesan Karthik, Cody Zhewei Cao, Michael I. Demidenko, Andrew Jahn, William C. Stacey, Vibhangini S. Wasade, and David Brang. Auditory cortex encodes lipreading information through spatially distributed activity. 34(17):4021–4032.e5.
- [49] Clayton E. Curtis and Mark D’Esposito. Persistent activity in the prefrontal cortex during working memory. 7(9):415–423.
- [50] E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. 24:167–202.

- [51] Adam M Morgan, Orrin Devinsky, Werner K Doyle, Patricia Dugan, Daniel Friedman, and Adeen Flinker. Hybrid spatial organization and magnitude-independent coding of linguistic information during sentence production. 5(3):184–194.
- [52] N. Ramnani and A. M. Owen. Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. 5(3):184–194.
- [53] N. Mesgarani and E. F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception. 485(7397):233–U118.
- [54] J. A. O’Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. 25(7):1697–1706.
- [55] Adam Bednar and Edmund C. Lalor. Where is the cocktail party? decoding locations of attended and unattended moving sound sources using EEG. 205:116283.
- [56] Michael J. Crosse, John S. Butler, and Edmund C. Lalor. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. 35(42):14195–14204.
- [57] David Meijer and Uta Noppeney. Metacognition in the audiovisual McGurk illusion: perceptual and causal confidence.
- [58] Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. 415(6870):429–433.
- [59] J. M. Hillis, M. O. Ernst, M. S. Banks, and M. S. Landy. Combining sensory information: Mandatory fusion within, but not between, senses. 298(5598):1627–1630.
- [60] Christopher R Fetsch, Alexandre Pouget, Gregory C DeAngelis, and Dora E Angelaki. Neural correlates of reliability-based cue weighting during multisensory integration. 15(1):146–154.
- [61] Yinan Cao, Christopher Summerfield, Hame Park, Bruno Lucio Giordano, and Christoph Kayser. Causal inference in the multisensory brain. 102(5):1076–1087.e8.
- [62] Durk Talsma, Daniel Senkowski, Salvador Soto-Faraco, and Marty G. Woldorff. The multifaceted interplay between attention and multisensory integration. 14(9):400–410.
- [63] Grace W. Lindsay. Attention in psychology, neuroscience, and machine learning. 14:29.
- [64] Uta Noppeney. Perceptual inference, learning, and attention in a multisensory world. 44(1):449–473.

Frontal cortex organization supporting
audiovisual processing during naturalistic viewing
Response to reviewers

Faxin Zhou et.al.

February 5, 2026

Reviewer #1

The authors have addressed most of our concerns and questions comprehensively. We have two additional comments that may be addressed at the author’s discretion.

Comments

Re: Answer 1

The analysis of flexible resource assignment in the parietal lobe is interesting. It underscores that this effect is particularly pronounced in the frontal cortex. Given that there was a significant interaction effect between language and modality conditions a possible alternative interpretation could be that audiovisual resource assignment does take place in the parietal cortex in some form. Though the presence of the main effect of modality suggests that sensory processing takes on a stronger role than in the frontal cortex. Acknowledging the significant interaction effect would allow the reader to interpret the data more clearly.

Answer 1 (A1): We agree that presenting the comprehensive ANOVA results would improve the clarity and interpretability of this analysis. Explicitly reporting these effects could also provide useful evidence on potential functional similarities and specializations within the frontoparietal network. We have therefore revised the manuscript to report the complete statistical results accordingly (lines 239-246):

“Notably, this pattern was not observed in other areas typically implicated in audiovisual processing, such as the parietal cortex [34, 35], in which auditory encoding performance exceeded visual encoding performance across both the EN and FL conditions, alongside a weaker but significant interaction effect (two-way ANOVA, modality main effect: $F(1, 122) = 24.340, p < 0.001, \eta^2 = 0.166$; language main effect: $F(1, 122) = 0.448, p = 0.505, \eta^2 = 0.004$; language-by-modality interaction effect: $F(1, 122) = 7.262, p = 0.008, \eta^2 = 0.056$; Fig. S7).”

Re: Answer 4

While the pie chart gives a good sense of how the electrodes are balanced between different anatomical regions, further quantification could be helpful. The total number of electrodes in each anatomical region would give a better sense of density of coverage. The ratio of active electrodes over the total number of electrodes in each area could give a sense of sensitivity of electrodes in each area to the stimulus features. In combination with figure R18 this could improve assessing coverage.

A2: We appreciate this valuable suggestion. To complement the pie charts (Fig. S8) with more quantitative information, we now provide the exact number of electrodes in each anatomical region to better characterize the density of coverage (Fig. S1A). In addition, we report both the number and proportion of the significant electrodes in each region across different models (columns in the table), providing an estimate of regional sensitivity to different stimulus features. We have incorporated this table into the Supplementary Materials accordingly (Table S2).

Table R1: Electrode anatomical coverage and significant responses

ROI	total	active	NMF (aud)	NMF (vis)	mTRF (aud)	mTRF (vis)	mTRF (assign.)
STG	146	72 (49.32%)	53 (36.39%)	12 (8.22%)	64 (43.84%)	36 (24.66%)	24 (16.44%)
MTG	136	20 (14.71%)	7 (5.15%)	6 (4.41%)	19 (13.97%)	14 (10.29%)	17 (12.50%)
OCC	59	32 (54.24%)	3 (5.08%)	29 (49.15%)	2 (3.39%)	34 (57.63%)	6 (10.17%)
FUS	22	10 (45.45%)	3 (13.64%)	7 (31.82%)	2 (9.09%)	9 (40.91%)	3 (13.64%)
IFG	90	29 (32.22%)	15 (16.67%)	11 (12.22%)	17 (18.89%)	22 (24.44%)	25 (27.78%)
MFG	145	24 (16.55%)	5 (3.45%)	16 (11.03%)	12 (8.28%)	17 (11.72%)	22 (15.17%)
prCG	101	28 (27.72%)	11 (10.89%)	14 (13.86%)	16 (15.84%)	21 (20.79%)	11 (10.89%)
psCG	106	36 (33.96%)	10 (9.43%)	18 (16.98%)	19 (17.92%)	13 (12.26%)	10 (9.43%)
IPL	208	51 (24.52%)	14 (6.73%)	33 (15.87%)	21 (10.10%)	32 (15.38%)	24 (11.54%)
Other regions	293	49 (16.72%)	9 (3.07%)	33 (11.26%)	24 (8.19%)	48 (16.38%)	35 (11.95%)
Other (WM & UNK)	1150	299 (26%)	35 (3.04%)	19 (1.65%)	123 (10.70%)	44 (3.83%)	129 (11.22%)

“ROI” corresponds to the anatomical regions shown in Fig. S8, and “Other (WM & UNK)” indicates other electrodes labeled as white matter (WM) and unknown (UNK). The column “active” reports the number (and percentage) of active electrodes that survived the permutation test (see Fig. 1B). Columns “NMF (aud)” and “NMF (vis)” report the number (and percentage) of electrodes assigned to the auditory and visual functional clusters, respectively. Columns “mTRF (aud)”, “mTRF (vis)”, and “mTRF (assign.)” report the number (and percentage) of electrodes showing significant encoding for the auditory, visual, and audiovisual assignment models, respectively.

We congratulate the authors on their interesting and well-executed work.

We sincerely thank you and Reviewer #2 for your encouraging comments and thoughtful evaluation of our work. We believe that these constructive suggestions and recommendations has substantially strengthened the manuscript.

Reviewer #3

The authors have satisfactorily addressed my concerns.

We sincerely thank you for your constructive feedback on our work.

Reviewer #4

I thank the authors for their thorough consideration of my (and the other reviewers') comments.

We are grateful for your kind acknowledgment and for providing valuable feedback that has guided the improvements in the revised manuscript.

Comments

Regarding my second major point and the authors' A2: I'm not quite sure that we are talking about exactly the same thing. I did not mean to say that the movie segments that were intelligible were inherently more "attention-driving" than the others. Rather, I thought that, when the auditory component of the movie is intelligible speech, then it would make sense to think that it would recruit the inferior frontal cortex (and its speech- and language-related functions) more than when it is unintelligible. Thus, instead of representing differences in cortical engagement with auditory vs. visual components of the stimulus, the authors' findings could reflect more speech- and language-related processing vs. less. The additional analysis presented in Figure R19 does not alleviate my concern. In fact, if no non-speech and language AV stimuli are available in the same patients, and if no patient understood any of the "foreign" languages in the movie, I do not think that this potential confound can be addressed completely.

I think that the above confounds warrants a specific discussion of this particular point in the limitations sections – unless I missed a subtle point of the authors' analytical approach, in which case I am happy to be guided again through it.

We sincerely appreciate your clarification on this important point and apologize for the misunderstanding on our part in the previous version. We agree that the recruitment of the inferior frontal cortex (IFC) in the auditory domain may reflect language-related processing during naturalistic viewing, which seems partially supported by our results. For example, our study also included the other sound (OS) condition, which involves auditory processing in the absence of linguistic content. In this condition, we neither observed frontal activation (please see Fig. 2C in the manuscript) nor detected significant encoding performance in frontal electrodes (please see Fig. S5 in the supplementary materials).

Unfortunately, as you noted we agree that we cannot fully resolve the question with the current dataset. We did carefully inspect all available data from these patients but did not identify any datasets that include appropriate non-speech audiovisual stimuli (our typical audiovisual stimuli include speech, and non-language paradigms were either visual or auditory alone). Following the reviewer's suggestion, we have explicitly discussed this limitation (line 388-393) and hope that future work with more targeted experimental designs could address this important and valuable question more directly:

"Lastly, vIPFC responses were primarily observed during intelligible speech, raising the possibility that the frontal audiovisual gradient may be influenced by language-related processes rather than purely audiovisual. The present study did not include naturalistic stimuli lacking linguistic content, thus we could not rule out this interpretation, and we believe that the relationship between language and audiovisual processing warrants further investigation."