BABA: Naturalistic fMRI and MEG recordings during viewing of a reality TV show

Authors:
Jixing Li (first and corresponding), Department of Linguistics and Translation, City University of Hong Kong, Hong Kong.
jixingli@cityu.edu.hk

Qiang Wang (co-first), Department of Otorhinolaryngology-Head & Neck Surgery, West China Hospital of Sichuan University, Chengdu, China
wangqiang1@stu.scu.edu.cn

Yike Wang, Department of Linguistics and Translation, City University of Hong Kong, Hong Kong.
yike.elaine.wang@my.cityu.edu.hk

Chengcheng Wang, Department of Linguistics and Translation, City University of Hong Kong, Hong Kong.
chengcheng.wang@my.cityu.edu.hk

Zhengwu Ma, Department of Linguistics and Translation, City University of Hong Kong, Hong Kong.
zhengwu.ma@my.cityu.edu.hk

Ruixuan Xu, Institute of Language Sciences, Shanghai International Studies University, Shanghai, China
ruixuan.xu@shisu.edu.cn

Shiyan Feng, Institute of Language Sciences, Shanghai International Studies University, Shanghai, China
fengshiyan@shisu.edu.cn

Xiaoming Jiang (corresponding), Institute of Language Sciences, Shanghai International Studies University, Shanghai, China
xiaoming.jiang@shisu.edu.cn

Zhaoli Meng (corresponding), Department of Otorhinolaryngology-Head & Neck Surgery, West China Hospital of Sichuan University, Chengdu, China
lucy-mengzhaoli@wchscu.cn

Abstract
Recent advances in the cognitive neuroscience of language have embraced naturalistic stimuli such as movies and audiobooks. However, most open-access neuroimaging datasets still focus on single-speaker scenarios, falling short of capturing the complexity of real-life, multi-speaker communication. To address this gap, we present the BABA fMRI and MEG dataset, collected while participants watched a 25-minute excerpt from a Chinese reality TV show featuring 11 speakers, including five fathers and their children. Set in a rural village, the show captures natural parent-child interactions with spontaneous, emotionally rich, and socially dynamic dialogue. Unlike scripted or isolated speech, it includes overlapping speech, speaker switches, and interruptions, offering a more ecologically valid stimulus. The combined use of fMRI and MEG allows researchers to explore both spatial and temporal aspects of language processing. This dataset provides a valuable resource for investigating the neural mechanisms underlying multi-talker comprehension, attentional shifts, and real-world social communication.

Background
Recent research in the cognitive neuroscience of language has shifted towards using a naturalistic paradigm, utilizing stimuli such as movies, podcasts, and audiobooks (Brennan et al., 2012; Hasson et al., 2004; Huth et al., 2016; Li et al., 2022; Malik-Moraleda et al., 2022; Momenian et al., 2024; Nastase et al., 2021; Wehbe et al., 2014). In contrast to traditional controlled experiments which often involves highly constrained and artificial linguistic inputs, naturalistic paradigms provide a richer, more ecologically valid window into how people comprehend language in everyday life. These approaches preserve the temporal dynamics, contextual richness, and social complexity inherent in natural communication, making them especially valuable for understanding language processing beyond the laboratory. However, most existing neuroimaging datasets have focused on single-speaker or two-speaker scenarios (Li et al., 2024; Zada et al., 2024; Zion Golumbic et al., 2013). In contrast, real-life communication frequently occurs in multi-speaker environments, such as group conversations and family interactions. Processing language in these contexts likely engages additional cognitive and neural mechanisms, including the allocation of attention across speakers, the monitoring of speaker identity, and the integration of multiple discourse streams (McDermott, 2009). Rapid speaker switching, a common feature of group dialogue, poses unique challenges for the brain. It requires listeners to update predictions dynamically, reorient attention, and maintain multiple mental representations of different speakers' perspectives and narrative threads (Mesgarani & Chang, 2012; Power et al., 2012). Understanding how the brain manages these demands is crucial for developing more comprehensive models of natural language comprehension and has implications for fields ranging from cognitive neuroscience of language, social cognition, developmental science, and computational modeling. Despite its importance, this area remains underexplored, highlighting the need for datasets and paradigms that reflect the true complexity of real-world communication.

Here, we present the BABA fMRI and MEG dataset, collected from native Chinese speakers as they watched a 25-minute excerpt from the Chinese reality TV show "Where Are We Going, Dad? (Season 1)" ("baba" means "dad" in Chinese, which inspired the name of the dataset). while undergoing neuroimaging. Originally aired in 2013, the show features unscripted interactions between five fathers and their young children as they travel to a rural village and engage in everyday activities, offering a rich and dynamic naturalistic stimulus that closely mirrors real-world communication. Participants completed five multiple-choice questions following the video to ensure their engagement and comprehension (see Table S1). In addition, we collected 15

minutes of resting-state data immediately after the video, during which participants lay still in the scanner and were instructed to recollect and reflect on the content of the video. This post-stimulus resting period provides a unique opportunity to examine memory consolidation, post-comprehension processes, and the lingering effects of complex language exposure on brain activity. The dataset includes recordings from both fMRI and MEG modalities, allowing researchers to investigate the spatial localization and fine-grained temporal dynamics of multi-speaker, emotionally nuanced, and socially embedded communication.

Validation analyses confirm the dataset's high quality and consistency across participants. Notably, we observed significant activation in the right temporoparietal junction (TPJ) during speaker-switching events, a region implicated in attentional reorienting to task-relevant, yet currently unattended stimuli (Corbetta & Shulman, 2002). These findings highlight the TPJ's central role in processing conversational dynamics and support the utility of this dataset for advancing our understanding of naturalistic, multi-talker language comprehension in the brain. The dataset also includes comprehensive annotations for the video, covering transcribed speech for each speaker and detailed word-level and phrase-level annotations, such as log frequency, part-of-speech (POS) tags, and the number of parser actions for each word derived from Stanford parser (Levy & Manning, 2003) based on bottom-up, top-down, and left-corner parsing strategies (Hale, 2014). These rich annotations enable fine-grained analysis of how specific acoustic and linguistic properties influence neural processing in complex auditory environments. All data and annotations are provided in standardized formats to ensure accessibility and reproducibility (see Fig. 1 for an overview of the data collection procedures, preprocessing steps, technical validation of the neuroimaging data, and annotation processes).

In sum, the BABA dataset offers several advantages for advancing research in language, social cognition, and auditory attention. It enables the investigation of neural mechanisms of multi-talker speech comprehension, including how the brain reallocates attentional resources across multiple interlocutors. The parent-child interactions also allow exploration of developmentally and emotionally grounded communication, which may engage unique neural circuits compared to adult-adult conversation.

Methods

Participants. Thirty participants (17 females, mean age=23.17±2.31 years) were recruited for the fMRI experiment at Shanghai International Studies University, Shanghai, China. An additional thirty participants (16 females, mean age=22.67±1.99 years) were recruited from the West China Hospital of Sichuan University, Chengdu, China. All participants were right-handed, had normal or corrected-to-normal vision, and reported no history of neurological disorders (see Table 1 for the participants' demographic information). Before the experiment, all participants provided written informed consent and were compensated for their participation.

Stimuli. The video stimulus was extracted from the first episode of the Chinese reality TV show "Where Are We Going, Dad? (Season 1)", which originally aired in 2013. The show features unscripted interactions between fathers and their child as they travel to a rural village and engage in daily activities. The selected excerpt has a total duration of 25 minutes and 19 seconds. The original video had a resolution of 640×368 pixels with a frame rate of 15 frames per second. It was presented in full-color (RGB) format, without embedded subtitles or captions.

Experimental design. The experimental procedures for both fMRI and MEG experiments were identical. Participants watched the video while inside the scanner. The video was presented via a mirror in the fMRI and MEG. Audio was delivered through MRI-compatible headphones (Sinorad, Shenzhen, China) during the fMRI experiment and MEG-compatible insert earphones (ComfortBuds 24, Sinorad, Shenzhen, China) during the MEG experiment. Following the video, participants were visually presented with 5 multiple-choice questions on the screen to assess their comprehension and ensure engagement with the stimuli (see Table S1). Participants responded using a button press, with a maximum response time of 10 seconds per question. If no response was recorded within this time, the experiment proceeded to the next question automatically. After the quiz, participants were instructed to close their eyes for 15 minutes without an explicit task. This period allowed for the recording of neural activity, capturing spontaneous mental replay of the video stimulus. The entire experimental procedure lasted approximately 45 minutes per participant. The fMRI experiment was approved by the Ethics Committee of Shanghai Key Laboratory of Brain-Machine Intelligence for Information Behavior (No. 2024BC028), and the MEG experiment was approved by the West China Hospital of Sichuan University Biomedical Research Ethics Committee (No. 2024[657]).

Data acquisition

The fMRI data was collected in a 3.0 T Siemens Prisma MRI scanner at Shanghai International Studies University, Shanghai. Anatomical scans were obtained using a Magnetization Prepared RApid Gradient-Echo (MP-RAGE) ANDI iPAT2 pulse sequence with T1-weighted contrast (192 single-shot interleaved sagittal slices with A/P phase encoding direction; voxel size=1×1×1 mm; FOV=256 mm; TR=2300 ms; TE=2.98 ms; TI=900 ms; flip angle=9°; acquisition time=6 min; GRAPPA in-plane acceleration factor=2). Functional scans were acquired using T2-weighted echo planar imaging (63 interleaved axial slices with A/P phase encoding direction, voxel size=2.5×2.5×2.5 mm; FOV=220 mm; TR=2000ms; TE=30 ms; acceleration factor=3; flip angle=60°).

MEG data were recorded at West China Hospital of Sichuan University using a 64-channel optically pumped magnetometer (OPM) MEG system (Quanmag, Beijing, China). OPM-MEG is a new type of MEG instrumentation that offers several advantages over conventional MEG systems. These include higher signal sensitivity, improved spatial resolution, and more uniform scalp coverage. Additionally, OPM-MEG allows for greater participant comfort and compliance, supports free movement during scanning, and features lower system complexity, making it a promising tool for more flexible and accessible neuroimaging (Brookes et al., 2022). The MEG Data were sampled at 1,000 Hz and bandpass-filtered online between 0 and 500 Hz. To facilitate source localization, T1-weighted MRI scans were acquired from the participants using a 3.0 T Siemens TrioTim MRI scanner at West China Hospital of Sichuan University (176 single-shot interleaved sagittal slices with A/P phase encoding direction; voxel size=1×1×1 mm; FOV=256 mm; TR=1900 ms; TE=2.3 ms; TI=900 ms; flip angle=9°; acquisition time=7 min).

Data preprocessing

All Digital Imaging and Communications in Medicine (DICOM) files of the raw fMRI data were first converted into the Brain Imaging Data Structure (BIDS) format using dcm2bids (v3.1.1; Boré et al., 2023) and subsequently transformed into Neuroimaging Informatics Technology Initiative (NIfTI) format via dcm2niix (v1.0.20220505). Facial features were removed from anatomical images using PyDeface (v2.0.2). Preprocessing was carried out with fMRIPrep (v20.2.0; Esteban

et al., 2019), following standard neuroimaging pipelines. For anatomical images, T1-weighted scans underwent bias field correction, skull stripping, and tissue segmentation into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). These images were then spatially normalized to the Montreal Neurological Institute (MNI) space using the MNI152NLin2009cAsym:res-2 template, ensuring consistent alignment across participants. Functional MRI preprocessing included skull stripping, motion correction, slice-timing correction, and co-registration to the T1-weighted anatomical reference. The blood-oxygen-level-dependent (BOLD) time series were resampled in both native and MNI space, and various confound regressors were computed to improve signal quality. Noise correction was applied to enhance the signal-to-noise ratio, and motion outliers were identified to mitigate potential artifacts in further analyses.

MEG data preprocessing was conducted using MNE-Python (v1.8.0; Gramfort et al., 2014). We first applied a bandpass filter (1–38 Hz) to remove low-frequency drifts and high-frequency noise. We then identified bad channels through visual inspection and cross-validated using PyPREP (v0.4.3; Bigdely-Shamlo et al., 2015), these bad channels were interpolated to maintain data integrity. To mitigate physiological artifacts, we performed independent component analysis (ICA) and removed components corresponding to heartbeat and eye movements. The data were then segmented into three task-related epochs, corresponding to the video watching, question answering, and post-task replay conditions, with each epoch defined strictly based on event markers without additional pre- or post-stimulus time windows. T1-weighted MRI data were converted to NIfTI format and processed with FreeSurfer (v7.3.2; Fischl, 2012) to reconstruct cortical surfaces and generate boundary element model (BEM) surfaces using a single-layer conductivity of 0.3 S/m. MEG-MRI coregistration was performed with fiducial points and refined via MNE-Python's graphical interface. A source space (resolution=5 mm) was generated using a fourth-order icosahedral mesh, and a BEM solution was computed to model head conductivity. A forward model was then created based on anatomical MRI and digitized head shape. Noise covariance matrices were estimated from raw MEG recordings, and inverse operators were constructed using minimum norm estimation. Source reconstruction employed dynamic statistical parametric mapping (dSPM) for noise-normalized estimates. Task-related epochs (video watching, question answering, post-task replay) were used to compute source estimates, which were morphed onto the FreeSurfer average brain template for group-level comparisons.

Annotations

We first used Whisper Large-V3 (Radford et al., 2022) to transcribe human speech into text with timestamps. We then manually reviewed and segmented the transcripts, resulting in 738 sentences and 2,930 words. Each sentence was annotated with one of 13 speaker labels, including 11 main speakers, a general "villagers" label for all villagers, and a general "staff" label for all staff members of the TV show (See example in Fig. 2a). Additionally, we identified and isolated background audio events without speech, categorizing them as 57 instances of natural ambient sounds ("sound") and 153 instances of artificial sound effects ("sound effect"). The demographic information for each speaker and their total speaking duration, total number of spoken words and speaking rate were shown in Fig. 2b.

Acoustic features, including pitch (f0) and root-mean square (RMS) intensity, were extracted at 10 ms intervals using Praat-Parselmouth (v0.4.4), as shown in Fig. 2c. Additionally, a binary indicator was assigned to mark speaker switching between different speech segments. Linguistic annotations were applied to the manually segmented speech transcript, integrating

multiple levels of information (Fig. 2d). Lexical frequency was estimated by applying a log transformation to unigram frequency values obtained from the Google Books Ngram Viewer dataset. Part-of-speech (POS) tags were assigned using the spaCy Chinese pipeline (v3.7.5; Honnibal et al., 2020). To assess syntactic complexity, the number of parsing actions required to process each word within a sentence was derived from constituency trees generated by the Stanford Parser (Levy & Manning, 2003), following three parsing strategies: top-down, bottom-up, and left-corner.

Data Records

The dataset follows the BIDS format (v1.8.0), and is publicly available on the OpenNeuro repository (https://openneuro.org/datasets/ds005346). All data were fully anonymized by removing identifiable information and anatomical details. We present both raw and preprocessed fMRI and MEG data, accompanied by relevant metadata, including participant information, behavioral results, and text and audio annotations. Fig. 3 presents overview of the dataset's directory structure.

Annotation files

Location: annotation/transcripts.csv
annotation/wav_acoustic.csv
annotation/word_information.csv
File format: comma-separated value.
Annotation of acoustic and linguistic features for the audio and text of the stimuli.

Stimuli files

Location: stimuli/baba_video.mp4
File format: mp4 (MPEG-4 Part 14).
The video file corresponds to the first episode of the Chinese reality TV show "Where Are We Going, Dad? Season 1" (2013).

Quiz files

Location: quiz/quiz_questions.csv
quiz/behavioral_results.csv
File format: comma-separated value.
Five multiple-choice questions and participants' accuracy scores from the fMRI and MEG experiments.

Anatomical MRI files

Location: sub-<ID>/anat/sub-<ID>_rec-defaced_T1w.nii.gz
File format: NIfTI, gzip-compressed.
Sequence protocol: sub-<ID>/anat/sub-<ID>_rec-defaced_T1w.json
Preprocessed data:
derivatives/sub-<ID>/anat/sub-<ID>_desc-preproc_T1w.nii.gz

fMRI data files

Location: sub-<ID>/func/sub-<ID>_task-[baba, question, replay, rest]_bold.nii.gz
File format: NIfTI, gzip-compressed.

Sequence protocol: sub-<ID>/func/sub-<ID>_task-[baba, question, replay, rest]_bold.json
Preprocessed data:
derivatives/sub-<ID>/func/sub-<ID>_task-[baba, question, replay,rest]_desc-preproc_bold.nii.gz

MEG data files
Location: sub-<ID>/meg/sub-<ID>_task-[baba, question, replay]_meg.fif
File format:FIF
Sequence protocol: sub-<ID>/meg/sub-<ID>_task-[baba, question, replay]_meg.json
Preprocessed data:
derivatives/sub-<ID>/meg/sub-<ID>_task-[baba, question, replay]_desc-preproc_meg.fif

Anatomical MRI files from the participants involved in the MEG experiment
Location: sub-<ID>/anat/sub-<ID>_rec-defaced_T1w.nii.gz
File format: NIfTI, gzip-compressed.
Sequence protocol: sub-<ID>/anat/sub-<ID>_rec-defaced_T1w.json

Technical Validation

Behavioral results

The mean accuracy on the quiz questions was 79.33% ± 19.29% for the fMRI experiment and 85.33% ± 18.14% for the MEG experiment. The mean log-transformed reaction times (RTs) were 1.29 ± 0.29 seconds and 1.41 ± 0.27 seconds, respectively (see Figure S1). These results suggest that participants were attentive to the video content during scanning and responded to the quiz questions both accurately and promptly.

fMRI data quality

Data quality assessment was conducted using the MRI Quality Control (MRIQC; Esteban et al., 2017) on the preprocessed data, which quantifies factors such as signal fidelity, noise levels, and anatomical precision. We also computed inter-subject correlation (ISC; Hasson et al., 2004) on the fMRI data to examine the consistency of neural activity across participants. Additionally, whole-brain general linear modeling (GLM) was performed to explore neural responses to pitch and intensity of the audio and switching of speakers in the video.

MRIQC

Image-quality metrics (IQMs) results from MRIQC for the raw and preprocessed anatomical data and functional data were shown in Fig. 4a,b), indicating high overall data integrity. Low coefficient of joint variation (CJV) for gray and white matter, along with entropy-focus criterion (EFC), suggests minimal head motion and few artifacts. High contrast-to-noise ratio (CNR) and foreground-background energy ratio (FBER) indicate clear tissue differentiation and well-balanced image energy distribution. The white matter to maximum intensity ratio (WM2MAX) confirms that white matter intensity is within the expected range, while strong signal-to-noise ratio (SNR) for both gray and white matter further supports good data quality. The functional MRI data exhibited a high temporal signal-to-noise ratio (tSNR) and low temporal derivative of time courses of root mean square over voxels (DVARS), indicating strong temporal stability. Minimal head motion and artifacts were reflected in low average framewise displacement (FD) and AFNI's outlier ratio (AOR). Additionally, low mean full width at half maximum (FWHM) and AFNI's quality index (AQI) suggest a well-distributed and consistent image intensity across the brain. Low

global correction (GOR) values and a small number of dummy scans further confirm the effectiveness of correction procedures and the stability of the initial scan state.

ISC of fMRI timecourses

We computed ISC for each voxel's time series across participants during video watching. Specifically, each subject's voxel-wise time series was correlated with the mean time series of the same voxel across all other participants (leave-one-out approach). This analysis produced a group-level ISC map, highlighting regions with the strongest correlations, particularly in the left temporal and bilateral occipital lobes (see Fig. 5).

GLM analysis

We performed a GLM analysis to investigate the BOLD responses to pitch, intensity, and the switching of speakers in the video. The pitch and intensity features were extracted every 10 ms of the audio and were standardized via z-scoring. Speaker-switching is a binary regressor where the offset of switching of speaker was marked as 1 in the transcript. These regressors were convolved with the canonical hemodynamic response function (HRF) and regressed against each participant's voxel-wise BOLD activity. The computed beta maps were then subjected to a second-level GLM, where statistical significance at the group level was evaluated through cluster-based inference, applying a threshold of $p<0.05$ (see Fig. 6a for the GLM analysis pipeline). The GLM results showed significant bilateral activation in the superior temporal lobes for pitch and intensity, highly consistent with the activation map for the terms "pitch" (102 studies), "intensity" (428 studies) extracted from fMRI term-based meta-analysis from Neurosynth (Yarkoni et al., 2011; see Fig. 6b).

As our video stimulus features multiple speakers, it provides a unique opportunity to examine the neural mechanisms underlying multi-talker conversation processing. To this end, we specifically analyzed brain activity at moments when the speaker switched during the video. The results revealed significant activation in the right temporoparietal junction (TPJ; see Fig. 6b). The TPJ is widely recognized as a key component of the ventral attention network, which is responsible for reorienting attention to task-relevant but currently unattended stimuli (see Corbetta & Shulman, 2002 for a review). To further validate our findings, we extracted the activation map for the term "switching" (based on 193 studies) from Neurosynth. The meta-analytic results showed prominent clusters in the bilaterial TPJ, which partially overlap with the speaker-switching activations observed in our data. This further suggests that speaker changes during naturalistic conversation engage domain-general cognitive control mechanisms, particularly those associated with task switching and attentional reorientation.

MEG data quality

To assess the quality and reliability of the MEG data, we calculated the proportion of bad channels and removed artifact-related ICA components for all participants. We then computed ISC on the MEG sensor time series to evaluate the consistency of neural responses across individuals during video viewing. Additionally, we conducted a regression analysis to identify significant spatiotemporal clusters associated with pitch, intensity, and speaker switching, providing insights into how low-level acoustic features and conversational dynamics are represented in the brain during naturalistic viewing.

Portion of bad channels and excluded ICA components

Fig 7a illustrates the MEG sensor layout. Fig 7b (left) summarizes the distribution of bad channels, indicating overall good signal quality. Eight participants had no bad channels, while for seventeen participants, Channel 19 and Channel 48 were consistently identified as bad due to equipment-related issues. Figure 7b (right) presents the number of excluded components following ICA, the mean number of excluded components is 2.03±1.67 channels across participants.

ISC of MEG timecourses

To establish a group-level reference, a median time series was computed for each sensor across all participants. Pearson correlation was computed between each subject's sensor-level time series and the corresponding group median time series. A spatiotemporal cluster-based permutation test with 10,000 permutations was then performed to identify statistically significant clusters. Significant ISC patterns were visualized using topographic maps, highlighting sensors in that exhibited the highest across-subject synchronization in neural activity (Fig. 7c).

MEG regression results

To examine MEG responses to pitch, intensity, and speaker switching, we performed a linear regression analysis using these three features as regressors to predict MEG activity during word-level epochs. Specifically, we extracted -100 to 300 ms epochs time-locked to the offset of each word in the audio stream. For each word, we used the maximum pitch and intensity values within the word interval as representative features. We performed the regression at 20 ms intervals across the 400 ms epoch at each sensor for every participant, yielding three beta time courses corresponding to the three regressors: pitch, intensity, and speaker switching. At the group level, we z-transformed these beta coefficients and conducted statistical analysis using a spatiotemporal cluster-based permutation t-test (Maris & Oostenveld, 2007) with 10,000 permutations to identify sensors and time points where the beta coefficients for pitch, intensity, and speaker switching is significantly different from 0. Our results are shown in Fig. 7d. We found significant cluster in the right temporal regions for pitch at around 240-280 ms after word offsets ($t=2.88$, $p=0.05$). Intensity was associated with sensors in the left temporal regions at around 40-80 ms before word offsets ($t=2.42$, $p=0.05$). These time windows were consistent with prior literature on the processing of acoustic features during naturalistic listening (e.g., Li et al., 2024). The speaker switching resulted in two marginally significant clusters in the left parietal regions at around 260-300 ms ($t=2.81$, $p=0.09$) and the right temporal regions at around 260-300 ms ($t=3.51$, $p=0.1$) after word offset, respectively. These results provide additional insight into the temporal dynamics of speaker switching during multi-talker speech conversation, complementing the spatial information revealed by the fMRI results (see Fig. 6b).

Usage Notes

This dataset provides a valuable resource for investigating the neural mechanisms underlying naturalistic audiovisual processing, multi-talker speech comprehension, and developmentally and emotionally grounded communication in both fMRI and MEG modalities. However, there are several limitations and potential bottlenecks in its usage.

Annotation bottleneck

Frame-by-frame speaker annotations in the video were not labeled in the current dataset, limiting the precision with which speaker-specific neural responses can be analyzed. Future work could incorporate automated annotation pipelines using multimodal large language models (LLMs)

capable of processing audiovisual inputs. These models can jointly analyze visual frames and audio tracks to detect and label speaker identities, speaking turns, and facial expressions in real time.

Analysis bottleneck

More advanced analytical techniques—such as multivariate approaches or computational models capable of representing group conversations—may be employed to further uncover the neural mechanisms underlying. These methods can offer a more nuanced understanding of how the brain navigates and interprets overlapping speech signals in complex, real-world listening environments.

Code availability

Scripts used for data preprocessing and analysis are available on GitHub (https://github.com/compneurolinglab/baba).

Acknowledgements

This work was supported by the CityU Start-up Grant 7020086 and CityU Strategic Research Grant 7200747 (LJ).

Author contributions

JL designed the study. QW and ZM collected the MEG data. RX, SF and XJ collected the fMRI data. YW, CW, ZM and JL analysed the data. JL wrote the paper.

Conflict of interest

The authors declare no competing interests.

References

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The prep pipeline: Standardized preprocessing for large-scale eeg analysis. *Frontiers in Neuroinformatics*, *9*.

Boré, A., Guay, S., Bedetti, C., Meisler, S., & GuenTher, N. (2023). *Dcm2Bids* (3.1.1). Zenodo.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*(2), 163–173.

Brookes, M. J., Leggett, J., Rea, M., Hill, R. M., Holmes, N., Boto, E., & Bowtell, R. (2022). Magnetoencephalography with optically pumped magnetometers (OPM-MEG): The next generation of functional neuroimaging. *Trends in Neurosciences*, *45*(8), 621–634.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215.

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). Mriqc: Advancing the automatic prediction of image quality in mri from unseen sites. *PLOS ONE*, *12*(9), e0184661.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). Fmriprep: A robust preprocessing pipeline for functional mri. *Nature Methods*, *16*(1), 111–116.

Fischl, B. (2012). Freesurfer. *NeuroImage*, *62*(2), 774–781.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460.

Hale, J. (2014). *Automaton theories of human sentence comprehension*. CSLI Publications.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *Spacy: Industrial-strength natural language processing in python*.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Levy, R., & Manning, C. D. (2003). Is it harder to parse Chinese, or the Chinese treebank? *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 439–446.

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R. N., Brennan, J. R., Yang, Y., Pallier, C., & Hale, J. (2022). Le Petit Prince multilingual naturalistic fMRI corpus. *Scientific Data*, *9*(1), Article 1.

Li, J., Wang, Q., Zhou, Q., Yang, L., Shen, Y., Huang, S., Wang, S., Pylkkänen, L., & Huang, Z. (2024). Multi-talker speech comprehension at different temporal scales in listeners with normal and impaired hearing. *eLife*, *13*.

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), Article 8.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

McDermott, J. H. (2009). The cocktail party problem. *Current Biology*, *19*(22), R1024–R1027.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236.

Momenian, M., Ma, Z., Wu, S., Wang, C., Brennan, J., Hale, J., Meyer, L., & Li, J. (2024). Le Petit Prince Hong Kong (LPPHK): Naturalistic fMRI and EEG data from older Cantonese speakers. *Scientific Data*, *11*(1), 992.

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., … Hasson, U. (2021). The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*(1), 250.

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*(9), 1497–1503.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision* (arXiv:2212.04356). arXiv.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS One*, *9*(11), e112575.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), Article 8.

Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Devore, S., Flinker, A., Devinsky, O., Nastase, S. A., & Hasson, U. (2024). A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron*, *112*.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, *77*(5), 980–991.

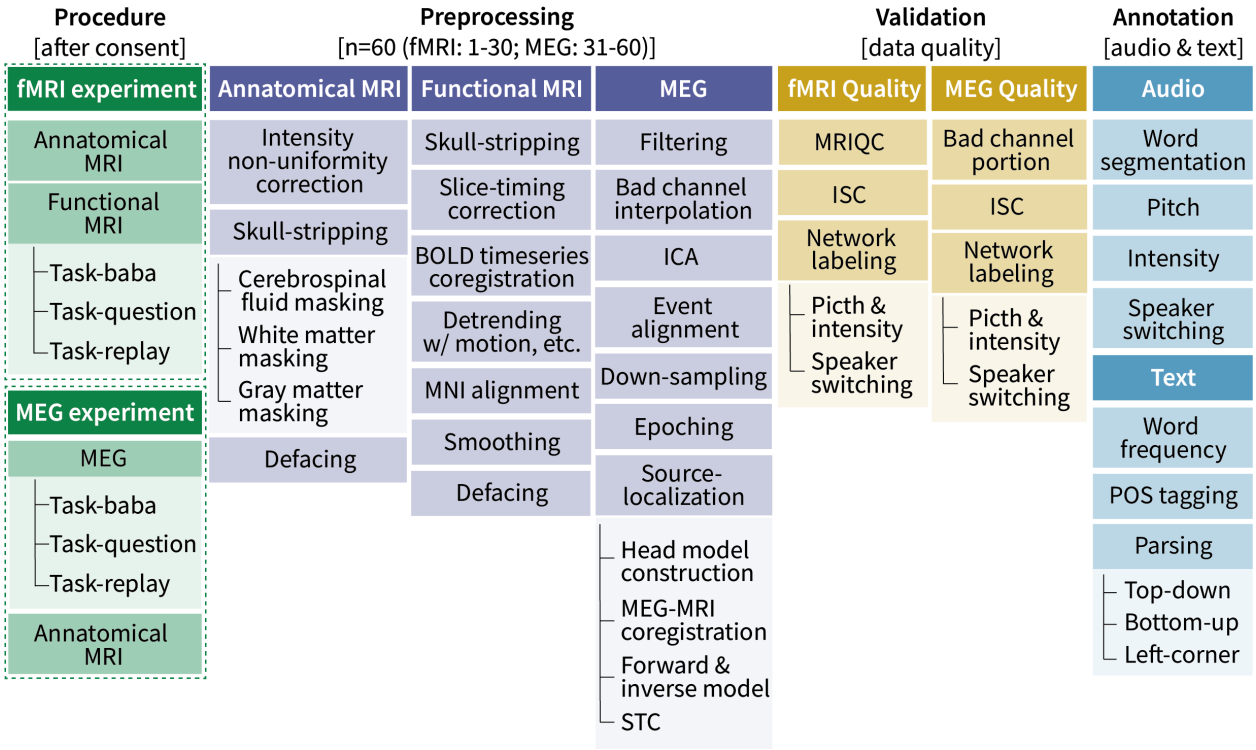


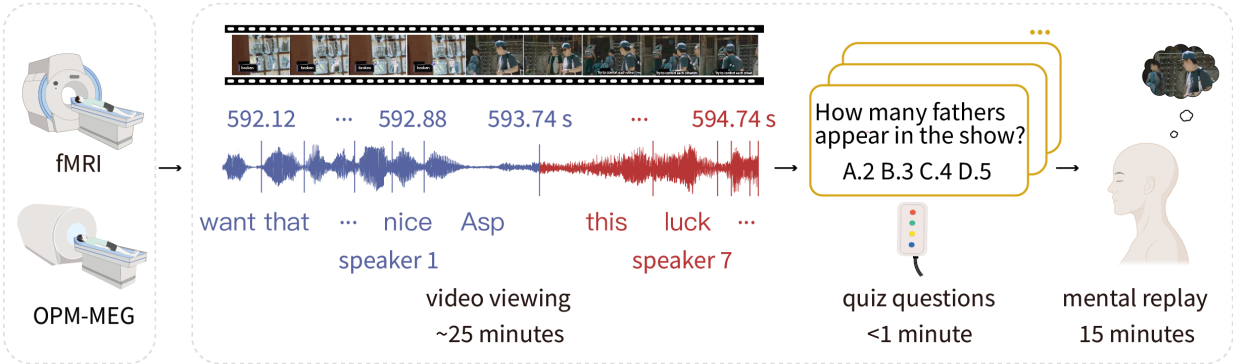
Fig. 1 Overview of the fMRI and MEG data collection and analysis procedures. Different participants performing identical tasks in the fMRI and MEG scanners (green). Both fMRI and MEG data underwent preprocessing (purple) and were evaluated using quality control methods (yellow). Audio and transcribed speech from the video stimuli were annotated with acoustic and linguistic features (blue).

fMRI experiment			**MEG experiment**		
Participant ID	**Sex**	**Age**	**Participant ID**	**Sex**	**Age**
sub-01	F	24	**sub-31**	F	24
sub-02	F	20	**sub-32**	M	24
sub-03	F	22	**sub-33**	F	27
sub-04	M	25	**sub-34**	F	22
sub-05	F	25	**sub-35**	F	21
sub-06	F	24	**sub-36**	F	22
sub-07	F	24	**sub-37**	F	21
sub-08	F	25	**sub-38**	F	21
sub-09	F	21	**sub-39**	F	24
sub-10	F	25	**sub-40**	F	26
sub-11	M	25	**sub-41**	F	24

sub-12	F	23	**sub-42**	M	21
sub-13	F	19	**sub-43**	M	20
sub-14	F	22	**sub-44**	M	19
sub-15	F	22	**sub-45**	M	21
sub-16	M	19	**sub-46**	F	24
sub-17	M	21	**sub-47**	M	23
sub-18	M	24	**sub-48**	F	25
sub-19	F	25	**sub-49**	M	20
sub-20	M	20	**sub-50**	M	21
sub-21	F	23	**sub-51**	F	22
sub-22	F	21	**sub-52**	F	22
sub-23	M	24	**sub-53**	F	22
sub-24	M	24	**sub-54**	M	24
sub-25	F	21	**sub-55**	F	21
sub-26	M	25	**sub-56**	M	22
sub-27	M	24	**sub-57**	M	22
sub-28	M	24	**sub-58**	M	24
sub-29	M	24	**sub-59**	M	26
sub-30	M	30	**sub-60**	M	25

Table 1. Participants' demographic information.

a Experiment procedure



b Annotations of speakers in the video stimuli

group	speaker	sex	age	total duration of speech	total number of words	speaking rate (s)
adults (mean age=38.33±4.89 years; mean speaking rate=0.33±0.07 s)	1	M	43	185.56	415	0.45
	3	M	31	116.28	466	0.25
	5	M	39	176.17	617	0.29
	7	M	40	59.16	163	0.36
	9	M	34	53.06	170	0.31
	11	M	43	89.23	267	0.33
children (mean age=5±1 years; mean speaking rate=0.5±0.11 s)	2	M	6	65.05	156	0.42
	4	M	6	83.67	217	0.39
	6	M	4	36.35	59	0.62
	8	F	4	53.29	86	0.62
	10	F	5	65.24	140	0.47

c Acoustic features

time	pitch	intensity
...	...	...
0.08	117.19	73.66
0.09	117.02	73.54
0.1	116.89	73.48
...	...	...

d Linguistic features

word	speaker	switching	onset	offset	duration	logfreq	pos	td	bu	lc
want	1	1	591.9	591.12	0.03	18.55	VERB	4	2	1
that	1	0	592.12	592.37	0.025	17.79	PRON	2	2	1
...	...	...	...	...	...	...	...	...	...	...
this	7	1	593.74	594.25	0.51	19.56	PRON	4	3	2
luck	7	0	594.25	594.55	0.3	11.45	NOUN	2	3	3
...	...	...	...	...	...	...	...	...	...	...

Fig. 2 Experimental procedure and the annotations of the audio. (a) Participants watched the video while undergoing neuroimaging. The video frames, audio waveforms and corresponding transcripts are shown, with different colors indicating different speakers. They then answered 5 multiple-choice quiz questions and rested for 15 minutes in the scanner. (b) Acoustic features including pitch (f0) and root-mean square (RMS) intensity at every 10 ms of the audio were extracted. (c) Each word was annotated with speaker identity, timing (onset, offset, duration), log unigram frequency, part-of-speech (POS) tags, and syntactic complexity measures from different

parsing strategies. (d) Information of speakers and the duration, total number of words and speaking rate of their speech output in the video stimuli.

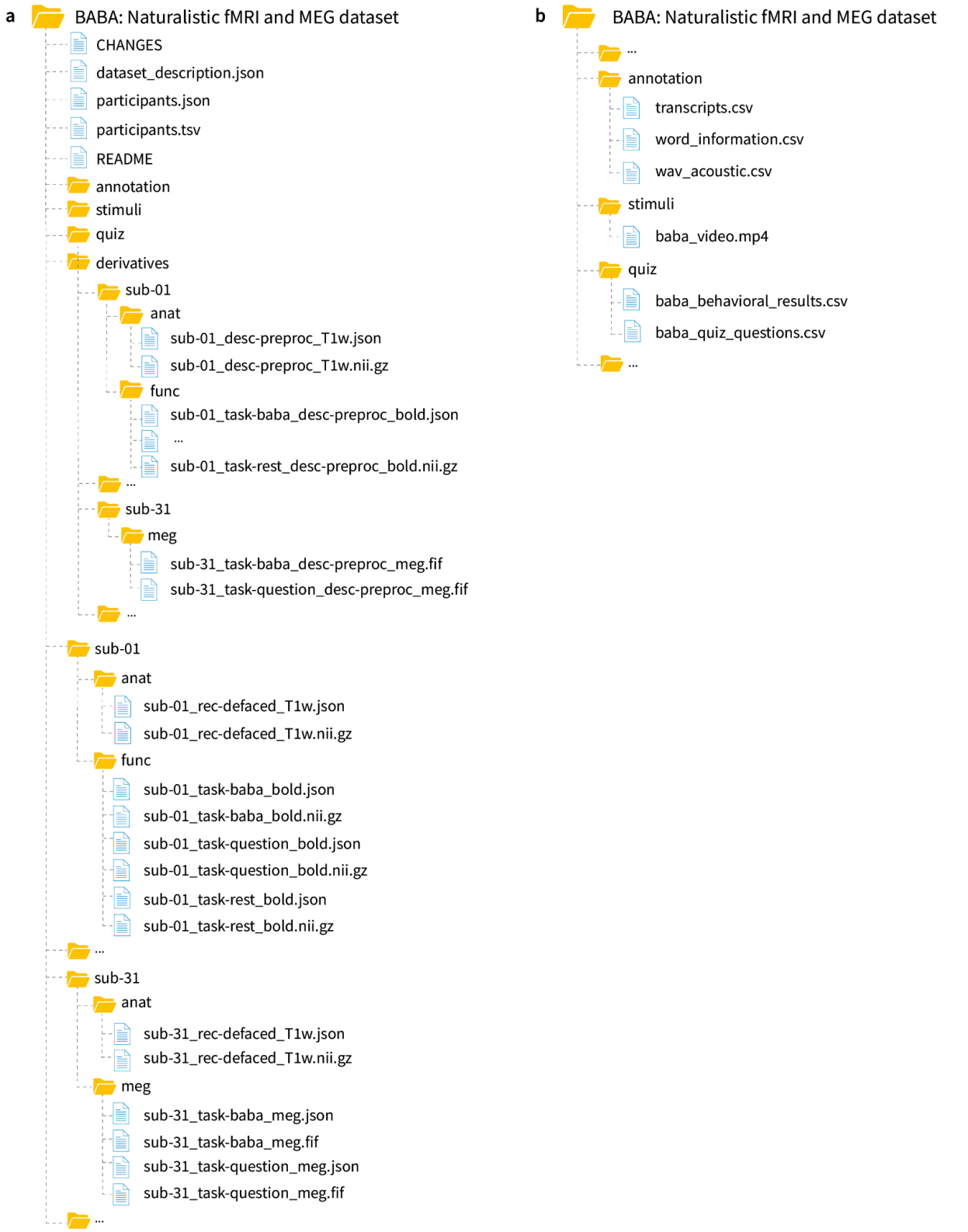


Fig. 3 Structure of the dataset directories. (a) Folder layout for both raw and processed fMRI and MEG data. (b) Folder layout for audio and text annotations, video stimuli, and behavioral results for the quiz questions.

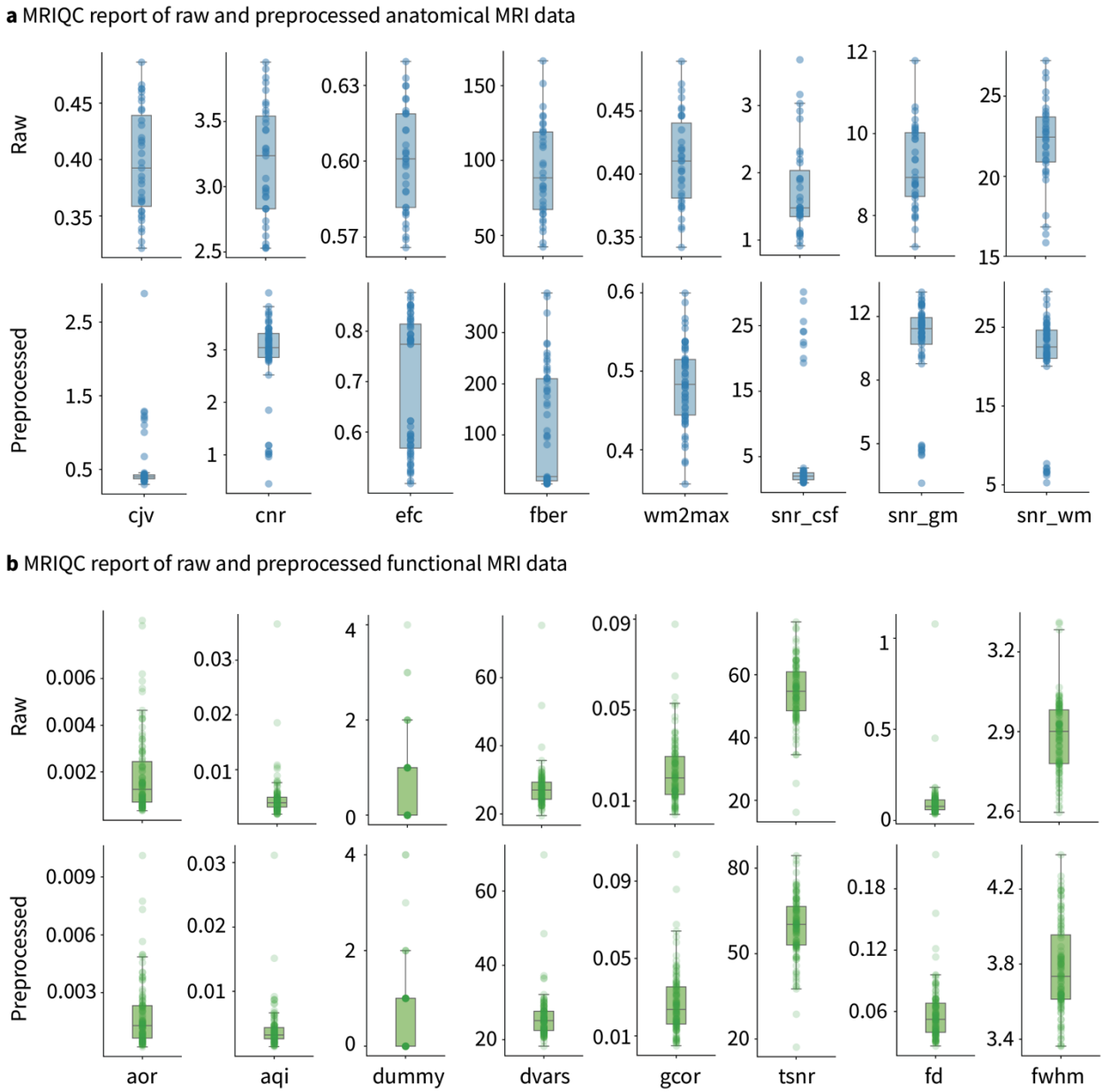


Fig. 4 Quality assessment of fMRI data. (a) Group-level Image-quality metrics (IQMs) of anatomical MRI data through MRIQC. (b) Group-level IQMs of functional MRI data through MRIQC.

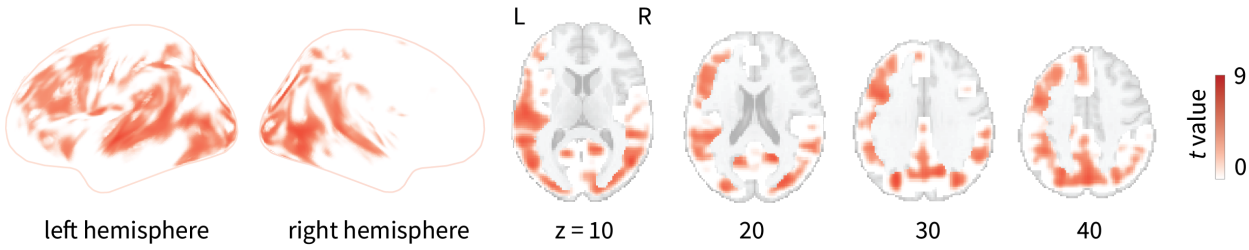


Fig. 5 Brain regions exhibited significant inter-subject correlation (ISC) during video viewing in the fMRI scanner.

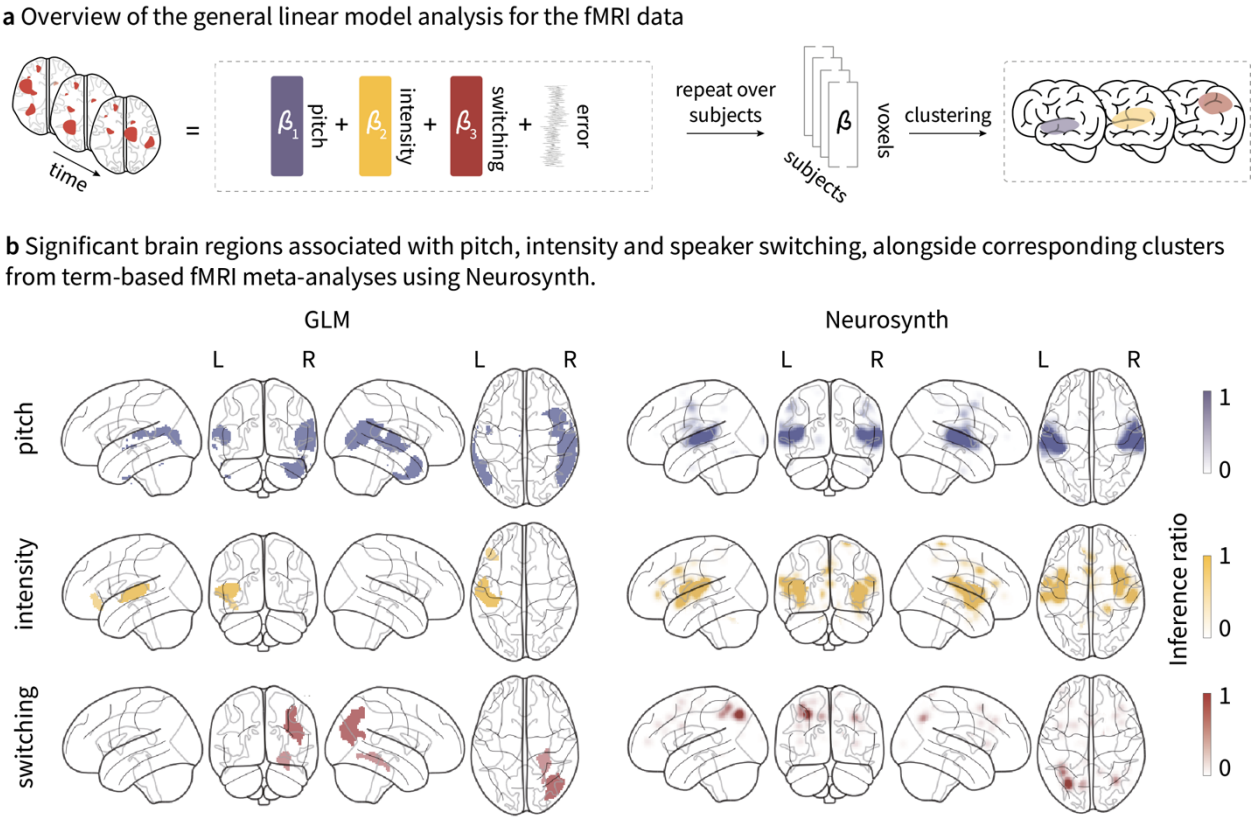


Fig. 6 GLM analyses for localizing brain responses to word rate, pitch, intensity, and speaker switching. (a) Overview of the GLM methods. (b) Brain maps showing significant activation clusters associated with each feature, alongside corresponding clusters from term-based fMRI meta-analyses using Neurosynth (Yarkoni et al., 2011).

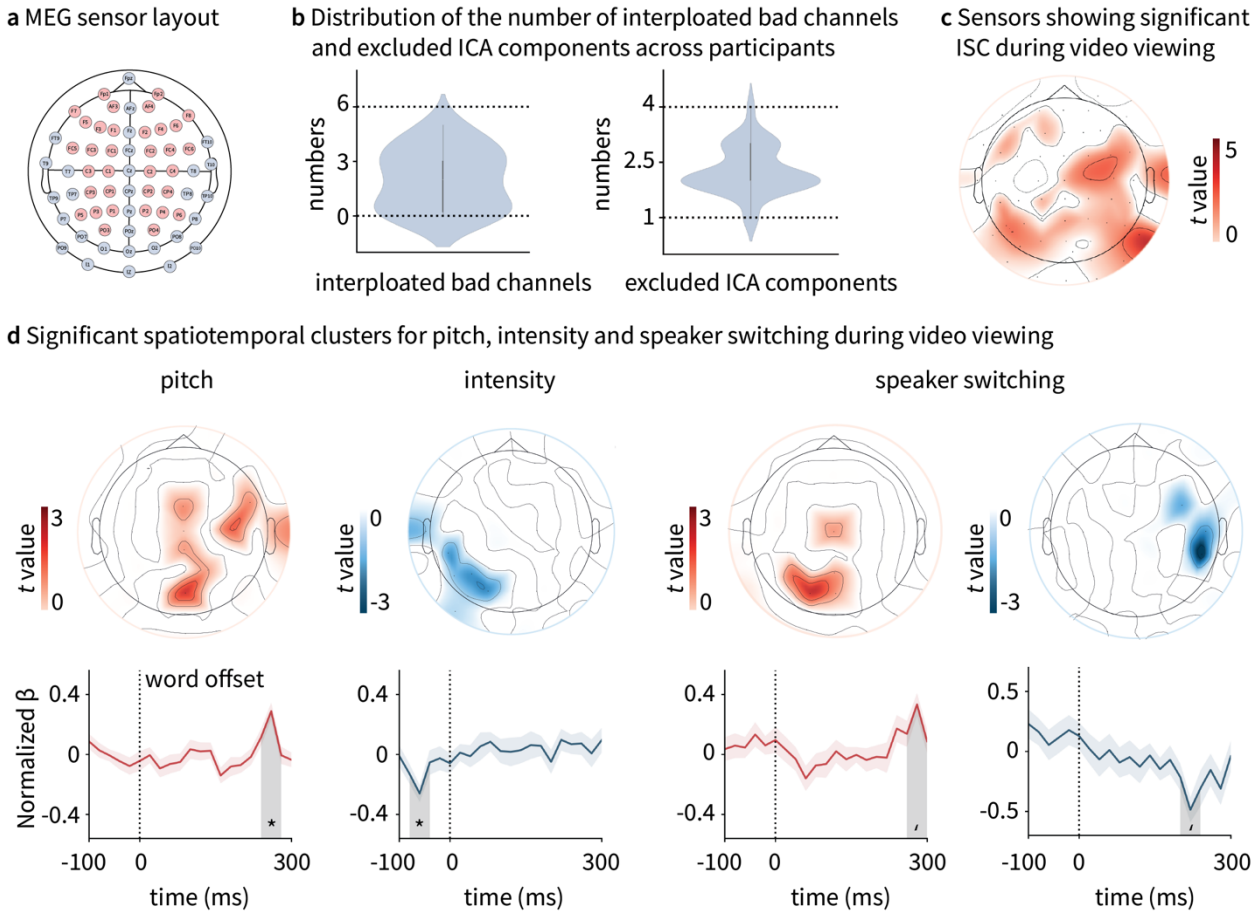
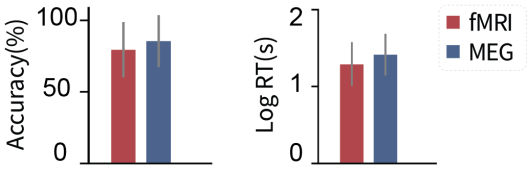


Fig. 7 Quality assessment of the MEG data. (a) MEG sensor layout. (b) Number of interpolated channels and excluded ICA components for all participants. (c) Sensors showing significant ISC during video viewing. (d) Significant spatiotemporal clusters for pitch, intensity and speaker switching.

Supplementary information

ID	Question	Choice A	Choice B	Choice C	Choice D	Answer
1	How many fathers appear in the film?	2	3	4	5	D
2	How many children in the film are girls?	1	2	3	4	B
3	How many children cried during the toy handover scene?	1	2	3	4	B
4	In the clip of allocating houses, which room has the worst conditions?	2	3	4	5	B
5	In the dinner-making segment, how many groups were the children divided into to look for cooking tools and ingredients?	1	2	3	4	B

Supplementary Table 1. The five multiple-choice questions after video viewing.



experiment	fMRI		MEG	
	mean	sd	mean	sd
accuracy (%)	79.33	19.29	85.33	18.14
log RT (s)	1.29	0.29	1.41	0.27

Supplementary Figure 1. Behavioral results of the quiz questions.