Aligning Text/Speech Representations from Multimodal Models with MEG Brain Activity During Listening

Padakanti Srijith[1], **Khushbu Pahwa**[2*], **Radhika Mamidi**[1], **Raju Surampudi Bapi**[1], **Manish Gupta**[3], **Subba Reddy Oota**[4]
[1]IIIT Hyderabad, India; [2]Rice University, USA; [3]Microsoft, India
[4]Technische Universität Berlin, Germany
padakanti.srijith@research.iiit.ac.in, kp66@rice.edu,
radhika.mamidi@iiit.ac.in, raju.bapi@iiit.ac.in,
gmanish@microsoft.com, subba.reddy.oota@tu-berlin.de

Abstract

Although speech language models are expected to align well with brain language processing during speech comprehension, recent studies have found that they fail to capture brain-relevant semantics beyond low-level features. Surprisingly, text-based language models exhibit stronger alignment with brain language regions, as they better capture brain-relevant semantics. However, no prior work has examined the alignment effectiveness of text/speech representations from multimodal models. This raises several key questions: Can speech embeddings from such multimodal models capture brain-relevant semantics through cross-modal interactions? Which modality can take advantage of this synergistic multimodal understanding to improve alignment with brain language processing? Can text/speech representations from such multimodal models outperform unimodal models? To address these questions, we systematically analyze multiple multimodal models, extracting both text- and speech-based representations to assess their alignment with MEG brain recordings during naturalistic story listening. We find that text embeddings from both multimodal and unimodal models significantly outperform speech embeddings from these models. Specifically, multimodal text embeddings exhibit a peak around 200 ms, suggesting that they benefit from speech embeddings, with heightened activity during this time period. However, speech embeddings from these multimodal models still show a similar alignment compared to their unimodal counterparts, suggesting that they do not gain meaningful semantic benefits over text-based representations. These results highlight an asymmetry in cross-modal knowledge transfer, where the text modality benefits more from speech information, but not vice versa. We make the code publicly available[1].

1 Introduction

Despite the fact that Transformer-based language models are not explicitly trained on brain data, recent brain encoding studies have shown that both text and speech Transformer-based language models exhibit a high degree of alignment with brain activity when participants engage in reading or listening naturalistic stories (Toneva and Wehbe, 2019; Deniz et al., 2019; Schrimpf et al., 2021; Millet et al., 2022; Vaidya et al., 2022; Caucheteux and King, 2022). Recent studies further investigated which types of information within these models lead to stronger alignment in both text and speech modalities (Oota et al., 2024a). They found that text-based language models strongly align with brain language regions even after controlling for low-level features (e.g., number of characters, number of phonemes, phonological features, etc.), indicating that these models also capture brain-relevant semantics (i.e., aspects of the model's internal representations that align with how the brain encodes meaning during natural language comprehension). In contrast, the alignment of speech-based models is widely driven by low-level features, suggesting that they do not encode crucial brain-relevant semantics. However, prior research has primarily focused on text- or speech-based models in isolation and has not investigated the potential of multimodal models that integrate both modalities for enhanced linguistic knowledge transfer. In this paper, we explore whether multimodal models can benefit from joint text+speech understanding and capture crucial brain relevant semantics better than unimodal models, thereby improving their alignment with brain language comprehension.

Deep learning for natural language processing models has led to unimodal text models like BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), FLAN-T5 (Chung et al., 2024) and LLaMA-2 (Touvron et al., 2023a) which provide effec-

*Work was done prior to the current role at Amazon
[1]https://github.com/srijith9862/MEG_multimodal

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34458–34475
November 4-9, 2025

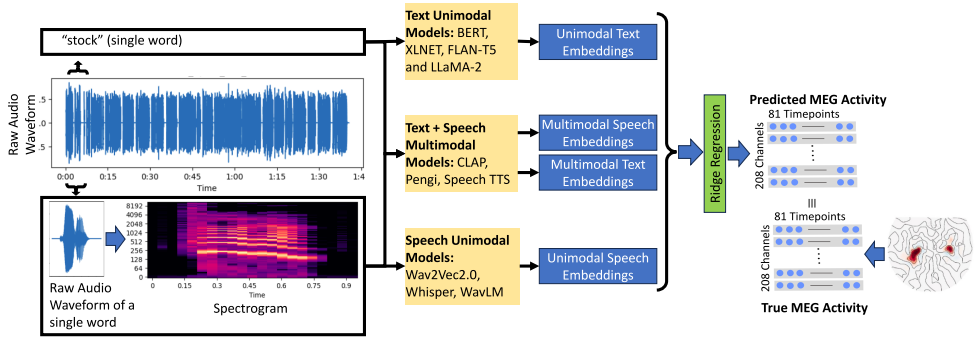


Figure 1: Methodology for studying the alignment of unimodal/multimodal text/speech encoding models for story listening with MEG brain activations. For unimodal alignment, we use representations from speech models or text models, where the input consists exclusively of either speech or text, respectively. For multimodal alignment, we leverage representations from text+speech aligned models, where the input consists of both text and speech.

tive text embeddings. Similarly, deep learning for speech processing has resulted in the development of effective unimodal speech models like Wav2Vec2.0 (Baevski et al., 2020), Whisper (Radford et al., 2023) and WavLM (Chen et al., 2022). To extract joint text+speech semantics, recent advances in AI systems have led to the development of multimodal models (like CLAP (Elizalde et al., 2023), Pengi (Deshmukh et al., 2023), and Speech TTS (Ao et al., 2022)) which are trained on massive interleaved text-audio data, to represent paired text+speech input. These multimodal models output separate text as well as speech embeddings based on their joint understanding of the input text and speech information. We refer to such representations from multimodal models as multimodal text and multimodal speech embeddings respectively.

In this work, we systematically address the following research questions, with respect to alignment with Magnetoencephalography (MEG) brain recordings during naturalistic story listening: **(RQ1)** Amongst unimodal text, unimodal speech, multimodal text, and multimodal speech, which embeddings best capture brain-relevant semantics? **(RQ2)** Is there asymmetric knowledge transfer across modalities in multimodal models, or do multimodal-text and multimodal-speech perform equally well? **(RQ3)** Do multimodal text/speech embeddings encode brain relevant semantics, i.e., are alignment patterns primarily driven by low-level auditory features or by higher-level, brain-relevant semantics?

Using MEG recordings of participants listening to naturalistic stories from MEG-MASC dataset (Gwilliams et al., 2023), we investigate the alignment between human brain language comprehension and unimodal/multimodal models. We selected MEG data over fMRI because, during naturalistic story listening, paired text-audio samples are available with MEG-a pairing that is difficult with fMRI recordings due to hemodynamic delayed response. Moreover, the high temporal resolution of MEG enables sub-word-level processing (Gwilliams et al., 2022; Oota et al., 2023b). Given these advantages of MEG, our work addresses the aforementioned research questions using naturalistic brain recordings alongside multimodal and unimodal models. For the purposes of this work, we focus on four unimodal text-based models, three unimodal speech-based models and three multimodal models as mentioned earlier. Overall, this research investigates the alignment of unimodal as well as multimodal representations to develop encoding models based on MEG responses (see Fig. 1).

Our analysis of multimodal models and brain alignment reveals several key conclusions: (1) Both multimodal and unimodal text embeddings show higher degree of brain predictivity beyond 350 ms in both temporal and frontal language regions, aligning with the time-frame associated with semantic word processing. In contrast, speech embeddings exhibit a sharp decline in performance around 350 ms, suggesting that text embeddings are more effective in processing semantic information compared to speech embeddings. (2) The improved alignment of multimodal text embeddings, particularly around the 200 ms window critical for auditory processing (primary auditory regions), indicates that integrating speech-derived features into text embeddings enhances their brain-relevant signal. When we removed unimodal speech features from multimodal text embeddings, we observed a substantial drop in activity which underscores the contribution of speech information. (3) While the text modality benefits from effective knowledge

transfer from the speech modality in multimodal models, the reverse is not observed. (4) Multimodal audio embeddings exhibit high MEG predictivity between 100 and 350 ms-consistent with early auditory processing and initial semantic integration-but lose much of their predictive power after 350 ms when low-level features are removed, indicating a lack of brain-relevant semantic representations. In contrast, while multimodal text embeddings also show a drop in early auditory predictivity without low-level features, their performance during later semantic processing remains largely intact, suggesting that they encode higher-level (beyond low-level features), brain-relevant semantic information.

Overall, we make the following contributions to this paper. (1) To the best of our knowledge, we are the first to study the effectiveness of text/speech representations from multimodal language models for MEG brain encoding. (2) Besides experimenting with 3 multimodal models, we also evaluate the performance of several unimodal Transformer models (four text and three speech) and measure their brain alignment. (3) Additionally, using the residual approach proposed by Toneva et al. (2022); Oota et al. (2024b,a), we remove unimodal and low-level features from multimodal embeddings. This allows us to explore the impact on brain alignment before and after their removal, and check whether alignment is driven by low-level features or by higher-level semantics. We make the code publicly available[1].

2 Related Work

Our work closely relates to a growing literature that investigates the alignment between human brains and language models. A number of studies have used unimodal text-based language models to predict both text-evoked and speech-evoked brain activity to an impressive degree (Wehbe et al., 2014; Jain and Huth, 2018; Toneva and Wehbe, 2019; Caucheteux and King, 2022; Antonello et al., 2021; Oota et al., 2022a; Merlin and Toneva, 2024; Oota et al., 2024b,a; Chen et al., 2024). Similarly, the recent advances in Transformer-based models for speech (Chung et al., 2020; Baevski et al., 2020; Hsu et al., 2021) have motivated neuroscience researchers to test their brain alignment with speech-evoked brain activity (Millet et al., 2022; Vaidya et al., 2022; Tuckute et al., 2023; Oota et al., 2023c,a, 2024a). Our approach is complementary and can be used to further understand aligned multimodal text and speech models to understand the brain alignment of language models, particularly for MEG brain encoding.

Our work also connects with the growing body of literature on recent advances in multimodal models, offering insights into how embeddings from these models differ from those of unimodal models in the context of brain encoding. For instance, previous studies (Doerig et al., 2022; Wang et al., 2023; Oota et al., 2022b; Popham et al., 2021) have shown that multimodal models like CLIP (Radford et al., 2021) better predict neural responses in the high-level visual cortex compared to vision-only models. Additionally, Tang et al. (2024) used multimodal models in a cross-modal experiment to evaluate how well language encoding models predict movie-fMRI responses and how well vision encoding models predict narrative story-fMRI responses. Other recent studies (Dong and Toneva, 2023; Nakagi et al., 2024; Oota et al., 2025) have demonstrated that brain recordings for multimodal stimuli reveal distinct regions associated with different semantic levels, highlighting the need to model various levels of semantic content simultaneously. However, while these studies have primarily focused on visual stimuli and the alignment of multimodal models in embedding space for vision tasks, the alignment of text+speech multimodal models for language stimuli remains unexplored. In contrast, our work is the first to study text+speech alignment in multimodal models for MEG encoding, providing a comprehensive analysis of brain alignment during naturalistic story listening.

3 Dataset and Curation

We used data from 27 participants in the MEG-MASC dataset (Gwilliams et al., 2023), which consists of 15 females (age: M = 24.1 years, SD = 6.7 years) and 12 males (age = 25.25 years, SD = 6.21 years). The activity from 208 MEG sensors was recorded while each subject listened to naturalistic spoken stories selected from the Open American National Corpus (*"Cable spool boy"*, *"LW1"*, *"Black willow"* and *"Easy money"*).

MEG processing. We performed minimal processing steps as described in Gwilliams et al. (2023). On raw MEG data and for each subject separately, using *MNE-Python* default parameters, we: (i) bandpass filtered the MEG data between 0.5 and 30.0 Hz, (ii) temporally-decimated the data 10x, (iii) segmented these continuous signals between -200 ms and 600 ms after word onset, (iv) applied

a baseline correction between -200 ms and 0 ms, and (v) clipped the MEG data between the fifth and ninety-fifth percentile of the data across channels.

Text and Speech processing. Since MEG data are sampled at a higher rate ($1000Hz$) than word presentation, epoching and downsampling yield 81 time points for each word/audio and for each of the 208 sensors. In total, there are 8,567 words/audio samples in total across the four stories. In our experiments, for each word/audio sample, the model makes a prediction of MEG activity for all of the 208 sensors $\times$ 81 time points = 16848 values. Here, we synthesize each word into a raw wav file using a word synthesizer. We discuss data preprocessing in detail in Appendix A.

Estimated cross-subject predictions. To account for intrinsic noise in biological measurements and obtain a more accurate estimate of the model's performance for the MEG-MASC dataset, we estimate the cross-subject predictions approach proposed by Schrimpf et al. (2021); AlKhamissi et al. (2024). We first subsample the data with n participants into all possible combinations of s participants for all $s \in [2, n]$ (e.g. 2, 3, 4, ..., 27 for n=27), and use an encoding model to predict one participant's response from others. Note that the estimated cross-subject prediction accuracy is based on the assumption of a perfect model, which might differ from real-world scenarios, yet offers valuable insights into model's performance. We present the average cross-subject prediction accuracy (noise ceiling) across sensors for the *MEG-MASC* dataset across subjects in Appendix B.

4 Methodology

We experiment with 3 multimodal models, 4 unimodal text models and 3 unimodal speech models. We acknowledge that the models differ in architecture, training objectives, and data sources. However, as established in prior brain encoding studies (Schrimpf et al., 2021; Toneva and Wehbe, 2019; Antonello et al., 2021; Aw and Toneva, 2023; Oota et al., 2025), such diversity is a deliberate design choice aimed at capturing general trends in how different types of representation, regardless of specific implementation details, align with brain activity. Our goal is not to control for every architectural or dataset difference but to evaluate how effectively text, speech, and multimodal representations from state-of-the-art models, despite inherent differences, align with brain activity, and to identify consistent patterns in modality-specific brain alignment.

4.1 Multimodal Models

To analyse whether speech models benefit from the transfer of linguistic information via language models and have brain-relevant semantics, we use recent popular text+speech multimodal models and build the encoding models for MEG. To extract representations of the multimodal text-speech stimulus, we used three recent text-speech multimodal models: CLAP (Elizalde et al., 2023), Speech TTS (Ao et al., 2022), and Pengi (Deshmukh et al., 2023). Details of these models are reported in Table 1 in Appendix C.

Extracting multimodal representations: Building upon prior approaches for extracting multimodal representations (Oota et al., 2022b, 2025), we leveraged pretrained multimodal models to obtain joint hidden-state representations from both speech and text data. To align both modalities, we paired individual words from text stimuli with the corresponding spoken audio files. Specifically, each word w_i in a story of M words was linked with the corresponding speech file. The speech files were processed using librosa to ensure the appropriate format and sampling rate for model input. Both the text and corresponding speech were passed through the pretrained multimodal model, yielding aligned text+speech embeddings for each token pair.

4.2 Unimodal Models

To investigate the effectiveness of multimodal representations in comparison to representations for unimodal ones, we use the following methods to obtain embeddings for individual modalities.

Text-based language models. To extract representations of the text stimulus, we use four popular pretrained Transformer text-based models from Huggingface (Wolf et al., 2020): (1) BERT (Devlin et al., 2019), (2) LLaMA-2 (Touvron et al., 2023a,b), (3) XLNet (Yang et al., 2019) and (4) FLAN-T5 (Chung et al., 2024).

Speech-based language models. To extract representations of the speech stimulus, we use three popular pretrained Transformer speech-based language models from Huggingface (Wolf et al., 2020): (1) Wav2vec-2.0 (Baevski et al., 2020), (2) WavLM (Chen et al., 2022) and (3) Whisper (Radford et al., 2023).

Unimodal feature extraction. Details of the unimodal text and speech models are reported in the Appendix C. To obtain text embeddings from unimodal text models, we extracted individual words from the text stimuli and passed them through pretrained text-based language models, resulting in an embedding for each word. Similarly, for speech embeddings, we processed corresponding audio files through pretrained speech-based language models, yielding an embedding for each spoken word.

4.3 Rationale for Choice of Unimodal and Multimodal Models

To comprehensively investigate the alignment between language models and brain activity during naturalistic speech comprehension, we selected a diverse set of models spanning unimodal and multimodal architectures. For multimodal text-speech models, we included CLAP, SpeechT5, and Pengi, which are designed to learn joint representations across text and audio, making them ideal candidates for probing cross-modal semantic transfer.

For unimodal text models, we chose BERT, LLaMA-2, XLNet, and FLAN-T5, representing a spectrum of transformer-based architectures with varying pretraining objectives and inductive biases, allowing us to assess how different textual representations align with brain semantics. On the speech side, Wav2Vec, WavLM, and Whisper were selected for their strong performance in learning speech representation and their ability to capture low-level acoustic features and higher-level linguistic cues. This curated selection enables a systematic comparison across modalities and model types, helping us isolate the contributions of multimodal integration versus unimodal specialization in capturing brain-relevant semantics. We discuss the details of the rationale for the choice of unimodal and multimodal models in Appendix C.

4.4 Low-level Features

We use the DisVoice library[2] to the following low-level speech features from spoken audio files corresponding to each word: (1) phonological, (2) phonation, and (3) articulation features. Further, we use the Self-Supervised Speech Pre-training and Representation Learning (S3PRL) toolkit[3] to compute features like (4) filter banks (fbanks), (5) Mel Spectrogram and (6) MFCC. Details of each low-level feature are reported in Appendix C.

[2] https://github.com/jcvasquezc/DisVoice

[3] https://github.com/s3prl/s3prl

4.5 Encoding Model

We use the extracted features for each stimulus word with an encoding model to predict brain responses. MEG encoder models attempt to predict brain responses associated with each MEG sensor and each time point when given speech stimuli (spoken words in our case). Let L denote the number of MEG sensors (208 in our case) and T represent the time dimension of the brain activity (81 in our case). Then for each spoken word, the goal is to predict a vector of length $L \times T$. We trained a model per subject separately. Following the literature on brain encoding, we chose to use ridge regression to train our encoding model. The ridge regression objective function is $f(X_s) = \min_{W_s} \|Y_b - X_s W_s\|_F^2 + \lambda \|W_s\|_F^2$. Here, Y_b denotes the actual brain activity of size $L \times T$, $W_s \in \mathbb{R}^{F_s LT}$ are the learnable weight parameters, F_s denotes the number of features (dimensionality) of stimulus representation, where s denotes the current word whose representation is being constructed, $\|.\|_F$ denotes the Frobenius norm, and $\lambda > 0$ is a tunable hyper-parameter representing the regularization weight. λ was tuned on a small disjoint validation set obtained from the training.

Cross-Validation. We follow 4-fold (K=4) cross-validation. All the data samples from K-1 folds (3 stories data) were used for training, and the model was tested on samples of the left-out fold (1 story). Model details and hyper-parameter settings are in Appendix E.

Removal of unimodal modality/low-level features from multimodal representations. To understand the contribution of unimodal or low-level features to multimodal model representations, we remove them from multimodal model representations by employing the direct or residual approach previously proposed by Toneva et al. (2022); Oota et al. (2024b); Dong and Toneva (2023); Oota et al. (2024a, 2025). This method estimates the linear contribution of specific modality or low-level features to the multimodal representations on the alignment between the model and brain recordings by comparing the alignment before and after computationally removing the targeted modality features from the multimodal representations. In our setting, we perform the removal by training a ridge regression model, where the unimodal feature (text or speech) or low-level feature vector is considered as input and the multimodal representation serves as the target. We compute the residuals by subtract-

ing the predicted feature representations from the actual features, resulting in the (linear) removal. Since our encoding model (ridge regression) is also a linear function, this linear removal limits the contribution of features for the particular modality to the eventual brain alignment. The mathematical notation of residual approach is described in Appendix D.

Residual contamination is possible; that is, some speech-related information may persist in the "text-only" residuals after linear removal, especially when modalities are deeply entangled within a shared latent space. However, prior work supports the effectiveness of this approach in capturing modality interactions. For example, Oota et al. (2024a) report that removing phoneme-level features from text models results in a significant drop in brain alignment in the early auditory cortex. Similarly, in speech models, removing letter- or word-level features impacts early auditory processing. These findings suggest that linguistic and acoustic features are often correlated across modalities, and removing one can impact the other. In our case, we hypothesize that removing unimodal embeddings from multimodal representations eliminates shared components.

4.6 Evaluation Metrics

We evaluate our models using normalized predictivity which is a popular brain encoding evaluation metric. To compute Normalized Predictivity, we first compute Pearson correlation coefficient (PCC) between real and predicted MEG activity to measure prediction performance for each sensor location and each time point within epochs. Then, **Normalized Predictivity** is computed as neural model predictivity values normalized by their respective subject ceiling values. The final measure of a model's performance ('normalized predictivity') on a dataset is thus PCC between model predictions and neural recordings divided by the estimated cross-subject predictions and averaged across sensor locations and participants.

4.7 Statistical Significance

We check statistical significance of PCC scores using a permutation test. We permute blocks of MEG predictions and compute PCC scores between permuted predictions and real data 5000 times to estimate an empirical distribution of chance performance and corresponding p-values. Finally, the Benjamini-Hochberg False Discovery Rate (FDR) correction (Benjamini and Hochberg, 1995) is applied globally, across all models, participants, and sensors, on all tests to control the type I error rate.

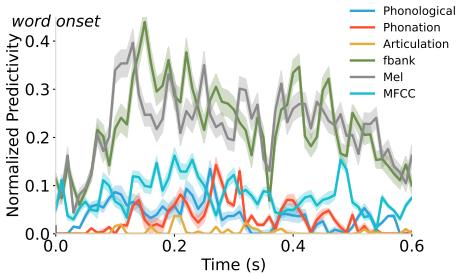


Figure 2: Avg. Normalized predictivity score for basic speech features between predicted and real MEG activity, across sensors and subjects. Word onset is at 0ms.

5 Results

5.1 Encoding Performance of Low-level Features

Fig. 2 reports the average normalized predictivity across subjects and sensors in the time dimension. From Fig. 2, we make the following observations: (i) FBank and Mel Spectrogram features perform the best across all time points. The activity is very high across the time duration from 50ms-550ms with a peak around 200ms, and a sudden drop around 350ms. This is in line with our understanding that auditory processing continues until around 350ms after which the semantic word processing happens (Dikker et al., 2020). (ii) It is difficult to capture articulation features from speech related to single words. Hence, those features seem to perform the worst. (iii) Phonation and phonological features both peak at similar time points, although activity for phonological features seems to start earlier.

5.2 Do multimodal models have better brain alignment over unimodal models?

Fig. 3 presents the average normalized predictivity for text and speech embeddings from multimodal models (aggregated across subjects) as well as unimodal text and speech embeddings. Similar results for individual models are shown in Fig. 9 in Appendix F. We make the following observations from Fig. 3: (i) Text embeddings from both multimodal and unimodal models significantly outperform speech embeddings from these models. (ii) Multimodal text embeddings exhibit a peak around 200ms, suggesting that they benefit from speech embeddings, with heightened activity during this time period. Specifically, the topographical maps

reveal that the higher normalized brain predictivity in primary auditory regions is unique to multimodal text embeddings and is not observed in unimodal text models (see Appendix Fig. 8). (iii) Both multimodal and unimodal text embeddings remain active beyond 350ms, aligning with the timeframe associated with semantic word processing. In contrast, speech embeddings experience a sharp decline in performance around 350ms. Topographical maps in Appendix Fig. 8 show that temporal, frontal, and parietal language regions exhibit higher predictivity during this period, whereas speech embeddings display low predictivity for both multimodal and unimodal models.

Impact of extended input context and prolonged MEG signal on brain predictivity. We further explored the impact of varying the input context length on model performance, as shown in Appendix G. In this analysis, we expanded the MEG recording window from 800 ms to 3 seconds. For example, given a story comprised of multiple speech files (with each file representing a single word) and a context length of 5, we input the sequence (w_1, w_2, w_3, w_4, w_5) and use the representation of the last word (w_5) as the target. From Fig. 10, we observe that-similar to the single-word context-an increase in context length further demonstrates that both multimodal text and unimodal text embeddings exhibit a higher degree of brain predictivity than speech embeddings across the extended temporal window. Our findings show that extending the input context not only reinforces the robustness of text embeddings but also underscores their superior alignment with prolonged brain activity compared to speech embeddings.

Lastly, multimodal text embeddings perform almost similar to unimodal text embeddings except during the peak around 200ms timepoint when the auditory information processing reaches heightened activity.

5.3 Is there asymmetric knowledge transfer across modalities in multimodal models?

To answer this question, we employ a residual approach by removing unimodal text embeddings from multimodal speech embeddings and unimodal speech embeddings from multimodal text embeddings. This method helps determine whether removing a specific modality affects the additional knowledge gained in multimodal models. Fig. 4 presents MEG encoding performance after removing unimodal information from multimodal embeddings for all three multimodal models. From the figure, we make the following observations: (1) **Impact of Removing Speech Embeddings from Multimodal Text:** Removing unimodal speech embeddings significantly impacts multimodal text embeddings, particularly around 200ms. This suggests that multimodal text embeddings incorporate additional speech-derived information, contributing to a higher peak at 200ms, as shown in Figure 3. (2) **Impact of Removing Text Embeddings from Multimodal Speech:** Removing unimodal text embeddings from multimodal speech embeddings primarily affects MEG predictivity during semantic information processing ($>$ 350 ms), while no significant impact is observed in auditory processing around 200 ms. This implies that speech embeddings from multimodal models contain additional information beyond unimodal text embeddings, but only during auditory information processing (around 200ms).

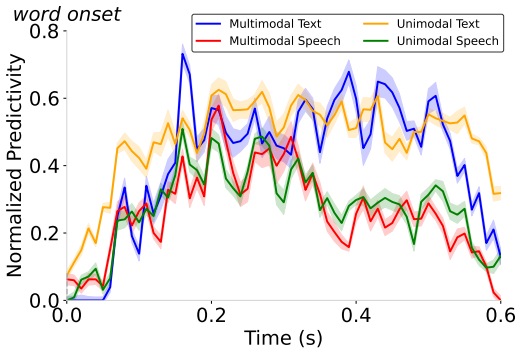


Figure 3: The average normalized predictivity of both text and speech embeddings from multimodal models, and unimodal text and speech representations, is measured by comparing predicted and real MEG activity across sensors and subjects. The word onset is at 0ms.

5.4 Do multimodal text/speech embeddings encode brain relevant semantics beyond low-level features?

To further examine whether text/speech embeddings from multimodal models predict MEG activity beyond low-level features-potentially capturing brain-relevant semantics-we adopt a residual approach, as discussed in Appendix D. Specifically, we remove low-level features from multimodal speech and text embeddings and assess whether this removal affects MEG brain predictivity. We follow the same feature removal method described in Section 4.5. Fig. 5 shows normalized predictivity over time (word onset is at 0 ms) for both multimodal text and speech embeddings before and after removal of low-level features.

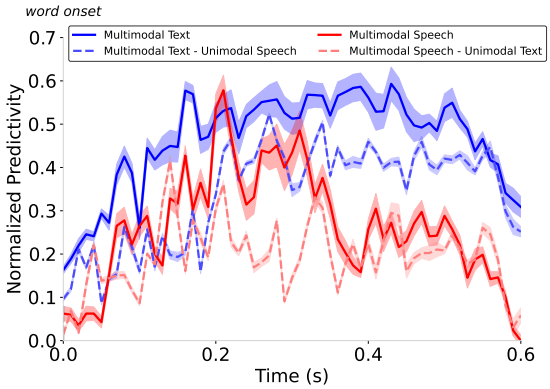


Figure 4: Residual Analysis: The average normalized brain predictivity was calculated by comparing predicted and real MEG activity across sensors and participants, both before and after the removal of unimodal speech and text embeddings from multimodal speech and text embeddings. The word onset is set at 0 ms. "-" symbol represents residuals.

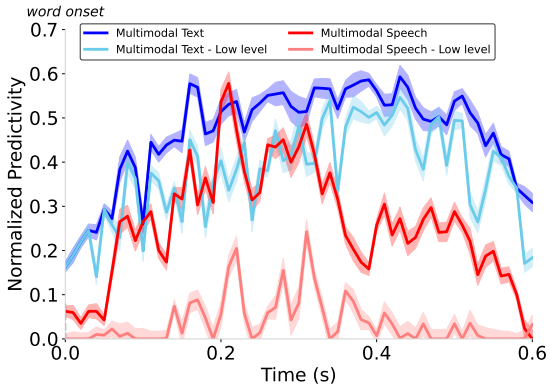


Figure 5: Residual Analysis for low-level features: The average normalized brain predictivity was calculated by comparing predicted and real MEG activity across sensors and participants, both before and after the removal of low-level features from multimodal speech and text embeddings. The word onset is set at 0 ms. "-" symbol represents residuals.

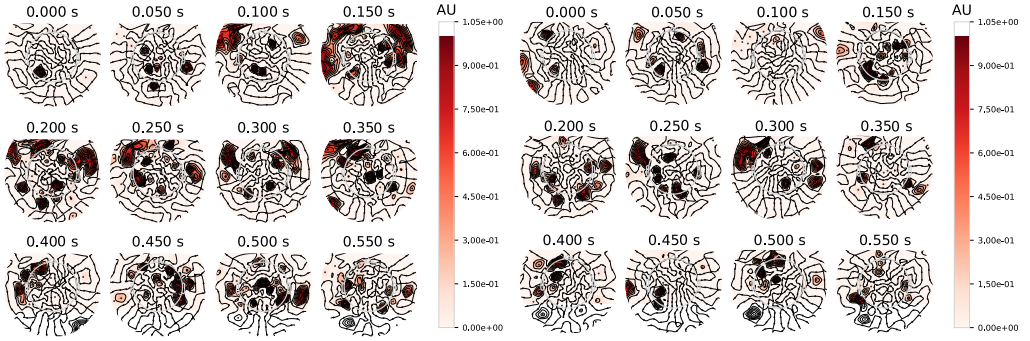
Removal of low-level features from multimodal speech embeddings. The speech embeddings from multimodal models generally exhibit higher MEG predictivity, particularly between 100 to 350 ms, where auditory information is predominantly processed, followed by semantic information. However, the removal of low-level features from the multimodal speech embeddings leads to a significant drop in prediction performance. Notably, after 350 ms, the prediction becomes negligible. This suggests that the predictive power of multimodal speech embeddings is primarily driven by low-level features, indicating a lack of brain-relevant semantic representations. In contrast, the removal of low-level features does not fully diminish MEG predictivity between 100 to 350 ms, implying that speech embeddings retain some additional information beyond low-level features during this window.

Removal of low-level features from multimodal text embeddings. Multimodal text embeddings exhibit higher MEG predictivity, particularly between 100 ms and 350 ms, a time window associated with auditory processing. This suggests that multimodal text embeddings incorporate additional information beyond standard textual representations, likely benefiting from speech-derived semantic and acoustic cues. The removal of low-level features from multimodal text embeddings results in a significant drop in MEG predictivity, especially during auditory processing windows. In contrast, the drop in predictivity after 350 ms, corresponding to semantic processing, is marginal and not statistically significant. These findings suggest that the predictive power of multimodal text embeddings extends beyond low-level features, indicating that these embeddings contain brain-relevant semantics.

Topographic analysis of low-level feature residuals. Fig. 6 displays the topographical maps of the residuals after the removal of low-level features from multimodal text (left) and speech embeddings (right). These maps highlight the MEG sensor locations that are significantly predicted across subjects with word onset at 0 ms. We observe that removing low-level features does not significantly affect the predictivity of text embeddings across the frontal and temporal regions between 150 and 550 ms. In contrast, for speech embeddings, the removal leads to a substantial decline in activity beyond 350 ms, although significant predictivity remains between 150 and 300 ms.

6 Discussion and Conclusions

Using speech and text embeddings from multimodal models, we evaluated how these representations predict MEG brain activity when participants engage in naturalistic story listening. Additionally, we compared multimodal and unimodal representations to assess their predictivity over time across sensors and participants. Furthermore, we examined which modality exhibits better knowledge transfer by removing information related to unimodal stimulus features (text and speech) from the multimodal embeddings and analyzing how this perturbation affects alignment with MEG brain recordings during story listening. Finally, to determine whether speech embeddings from multimodal models contain brain-relevant semantics, we removed low-level features and investigated whether



(a) Multimodal Text - Lowlevel (b) Multimodal Speech - Lowlevel

Figure 6: Residual Analysis: Topomaps display the average normalized predictivity for MEG activity (across subjects) after removing low-level features from multimodal embeddings. Word onset is at 0 ms. The left plot (a) shows the residual predictivity for multimodal text embeddings, while the right plot (b) presents that for multimodal speech embeddings.

the speech embeddings could still predict MEG activity beyond low-level auditory information. Our analysis of speech and text embeddings from multimodal models in relation to MEG brain alignment yields several key insights.

Towards RQ1, compared to speech models, both multimodal and unimodal text embeddings show higher degree of brain predictivity beyond 350 ms, aligning with the time-frame associated with semantic word processing. In contrast, speech embeddings exhibit a sharp decline in performance around 350 ms, suggesting that text embeddings are more effective in processing semantic information compared to speech embeddings.

Towards RQ2, using residual approach, we observe a clear asymmetry in the knowledge transfer across modalities in multimodal models. The improved alignment of multimodal text embeddings, particularly around the 200 ms window critical for auditory processing, indicates that integrating speech-derived features into text embeddings enhances their brain-relevant signal. When unimodal speech features are removed, activity in this window drops, underscoring the contribution of speech information. However, speech embeddings fail to show similar gains from text integration, particularly in semantic processing regions.

Towards RQ3, multimodal speech embeddings show high MEG predictivity between 100 and 350 ms-consistent with early auditory processing and subsequent semantic integration. However, the removal of low-level features from the multimodal speech embeddings leads to a significant drop in prediction performance, especially after 350 ms, suggests that the predictive power of these embeddings is primarily driven by low-level features, indicating a lack of brain-relevant semantic representations. In contrast, the MEG predictivity between 100 and 350 ms is only partially reduced when these features are removed, it suggests that the embeddings still capture additional, higher-level information during this time. These observations are consistent with similar findings in unimodal speech language models reported in recent work (Oota et al., 2024a). Although removing low-level features from multimodal text embeddings leads to a significant drop in MEG predictivity during early auditory processing, the impact on later (semantic) processing windows is minimal. This indicates that these embeddings encode brain-relevant semantic information that extends beyond basic auditory features-a characteristic that is also observed in unimodal text models in recent studies (Oota et al., 2024a).

Together, these results deepen our understanding of how multimodal models process and represent language-related information, bridging the gap between low-level auditory features and high-level semantic processing as observed in brain dynamics. Multimodal integration improves text-based representations, enhancing brain alignment in early auditory ($\sim$200 ms) processing window. However, speech-derived representations within multimodal models fail to exhibit comparable improvements, remaining constrained by low-level auditory features and lacking robust brain-relevant semantics. These findings extend prior work on unimodal models by demonstrating that multimodal integration enhances text-side brain alignment but does not yet overcome limitations on the speech side. Asymmetric knowledge transfer reflects current gaps in multimodal model design, highlighting the need for further research, including brain-informed model development.

7 Limitations

One limitation of our approach is that the differences observed between the brain alignment of text- and speech-based embeddings may be influenced by factors beyond the stimulus modality, such as variations in training data amounts and objective functions across the underlying encoder models. To mitigate this concern, we tested multiple models from each category-varying in objective functions and training data-and found that our results generalize across these models. However, differences in architectural variability and pretraining methods could still contribute to performance discrepancies, suggesting that future work should involve more tightly controlled comparisons (like parameter-matched unimodal checkpoints trained under identical settings) to isolate these effects.

Further, as our study utilizes brain recordings and stimuli in English with models predominantly trained on English data, the findings may not generalize to other languages, highlighting an important avenue for future research.

Recent brain-tuning studies have demonstrated that integrating neural brain data into speech-based language models enables these unimodal systems to capture semantic information that extends beyond low-level features (Moussa et al., 2025). Building on this insight, incorporating brain data into multimodal models could further enhance their end-to-end language processing capabilities. We believe that this brain-tuning approach could act as a multimodal convergent buffer, effectively integrating semantic-level language information to enrich the processing of speech-language semantics. This is something that can be explored in the future.

8 Ethics Statement

We did not create any new MEG data as part of this work. We used the MEG-MASC dataset which is publicly available without any restrictions. MEG-MASC dataset can be downloaded from https://osf.io/ag3kj/. Please read their terms of use[4] for more details.

We do not foresee any harmful uses of this technology.

[4] https://osf.io/ag3kj/metadata

References

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2024. Brain-like language processing via a shallow untrained multihead attention network. *arXiv preprint arXiv:2406.15109*.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. 2021. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in Neural Information Processing Systems*, 34:8332–8344.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, and 1 others. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738.

Khai Loong Aw and Mariya Toneva. 2023. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 33:12449–12460.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134.

Catherine Chen, Tom Dupré la Tour, Jack L Gallant, Daniel Klein, and Fatma Deniz. 2024. The cortical representation of language timescales is shared between reading and listening. *Communications Biology*, 7(1):284.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *J. Selected Topics in Signal Processing*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Yu-An Chung, Hao Tang, and James Glass. 2020. Vector-quantized autoregressive predictive coding. *Interspeech*, pages 3760–3764.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. 2019. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *J. Neuroscience*, 39(39):7722–7736.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Suzanne Dikker, M Florencia Assaneo, Laura Gwilliams, Lin Wang, and Anne Kösem. 2020. Magnetoencephalography and language. *Neuroimaging Clinics*, 30(2):229–238.

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. 2022. Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*.

Dota Tianai Dong and Mariya Toneva. 2023. Vision-language integration in multimodal video transformers (partially) aligns with the brain. *arXiv preprint arXiv:2311.07766*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. 2023. Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific data*, 10(1):862.

Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. 2022. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature communications*, 13(1):6606.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *TASLP*, 29:3451–3460.

Shailee Jain and Alexander G Huth. 2018. Incorporating context into language encoding models for fmri. In *NIPS*, pages 6629–6638.

Gabriele Merlin and Mariya Toneva. 2024. Language models and brain alignment: beyond word-level semantics and prediction. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18431–18454.

Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, and 1 others. 2022. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35:33428–33443.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. 2025. Improving semantic understanding in speech language models via brain-tuning. In *The Thirteenth International Conference on Learning Representations*.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. 2024. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20313–20338.

Subba Reddy Oota, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Raju Surampudi Bapi. 2023a. Speech taskonomy: Which speech tasks are the most predictive of fmri brain activity? In *INTERSPEECH 2023-24th INTERSPEECH Conference*, pages 5167–5171.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. 2022a. Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3220–3237.

Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. 2022b. Visio-linguistic brain encoding. In *COLING*, pages 116–133.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. 2024a. Speech language models lack important brain-relevant semantics. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8503–8528. Association for Computational Linguistics.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. 2024b. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36.

Subba Reddy Oota, Trouvain Nathan, Frederic Alexandre, and Xavier Hinaut. 2023b. Meg encoding using word context semantics in listening stories. In *Interspeech*.

Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Manish Gupta, and Bapi S Raju. 2023c. Neural architecture of speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Subba Reddy Oota, Khushbu Pahwa, mounika marreddy, Maneesh Kumar Singh, Manish Gupta, and Bapi Raju Surampudi. 2025. Multi-modal brain encoding models for multi-modal stimuli. In *The Thirteenth International Conference on Learning Representations*.

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. 2021. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. *Image*, 2:T2.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *PNAS*, 118(45).

Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. 2024. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, 36.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. 2022. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. 2023. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366.

Aditya R Vaidya, Shailee Jain, and Alexander Huth. 2022. Self-supervised models of audio effectively explain human cortical responses to speech. In *International Conference on Machine Learning*, pages 21927–21944. PMLR.

Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. 2023. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32.

Overview of Appendix Sections

- Section A: Details of Data Preprocessing.
- Section B: Cross-subject Prediction Accuracy.
- Section C: Multimodal and Unimodal Models.
- Section D: Residual Approach.
- Section E: Implementation details for reproducibility.
- Section F: Encoding Performance of Multimodal Models vs. Unimodal Models.
- Section G: Impact of extended input context and prolonged MEG signal on brain predictivity.

A Details of Data Preprocessing

The MEG-MASC dataset (Gwilliams et al., 2023) already provides word-level aligned MEG recordings and corresponding onset and offset information for each word in the story listening task. Each word is associated with a pre-defined epoch of MEG activity, already aligned to the word onset, following the standardized preprocessing protocol detailed in the MEG-MASC release (Gwilliams et al., 2023).

MEG Preprocessing and Word-Aligned Epoching. The MEG data used in our study comes from the MEG-MASC dataset, which includes standardized preprocessing steps applied consistently across all participants, including the following.

- Epoching of MEG signals from -200 ms to +600 ms relative to each word onset.
- Temporal decimation by a factor of 10 (i.e. downsampling).
- Baseline correction using the -200 ms to 0 ms pre-stimulus window.

Thus, each word is associated with a fixed-duration MEG segment (800 ms), independent of the actual spoken word length. This standardized window allows for consistent temporal alignment of brain responses across all words.

Downsampling to 81 Time Points and Predictivity Curve Computation. Following the MEG-MASC protocol, the 800 ms word-aligned MEG segments were downsampled to 81 time points, ensuring consistent temporal resolution across all words, regardless of word duration in the audio stimulus. Our encoding model predicts MEG activity for each of these 81 time points $\times$ 208 sensors, based on the corresponding model embeddings. At each time point, we compute predictivity scores (e.g., correlation or explained variance) between predicted and actual MEG activity. This yields time-resolved curves showing how brain alignment evolves over time, relative to word onset, capturing both early auditory responses and later semantic processing stages.

B Cross-subject Prediction Accuracy

Fig. 7 displays the mean estimated cross-subject prediction accuracy across MEG sensor locations for different subjects. We observe that the average cross-subject prediction accuracy (Pearson correlation) across sensor locations is different across subjects. Some subjects (12, 13, 19, and 26) have higher Pearson correlation scores, while several other subjects have lower Pearson correlation scores.

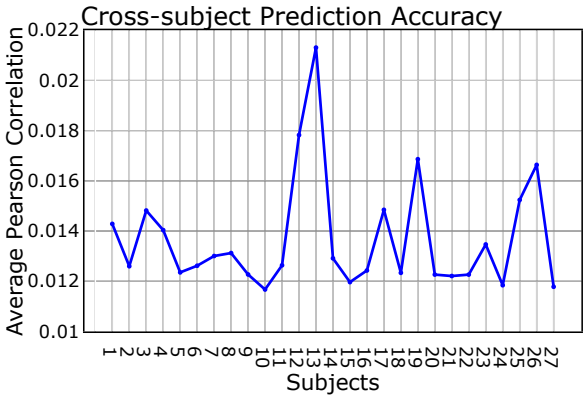


Figure 7: Estimated Cross-subject prediction accuracy for each subject in the MEG-MASC dataset.

C Multimodal and Unimodal Models

C.1 Multimodal models

CLAP: The CLAP model ("Contrastive Language-Audio Pretraining"), specifically the "laion/clap-htsat-unfused" checkpoint from Hugging Face which consists of 12 layers and produces 512-dimensional embeddings, is a multimodal model designed for contrastive learning between audio and text. It leverages a Hierarchical Token-Semantic Audio Transformer (HTSAT) for encoding audio, paired with a text encoder. The model aligns audio and textual representations in a shared latent space, enabling cross-modal retrieval and classification tasks. The unfused architecture ensures that audio and text features are processed separately before being aligned, making it efficient for a range of audio-text applications.

SpeechTTS: We use the "microsoft/speecht5_tts" model from Hugging Face, a state-of-the-art text-to-speech (TTS) model designed to generate natural, high-quality speech from text. SpeechT5 integrates both speech and text modalities through a shared Transformer architecture, enabling it to produce expressive speech outputs. The model leverages a large pretrained network with multi-layered attention mechanisms to convert text into 512-dimensional speech embeddings, which are then decoded into waveform outputs. Its robust architecture ensures clear and accurate speech synthesis, making it well-suited for various TTS applications.

Pengi: Pengi is an advanced Audio Language

Model that utilizes transfer Learning to frame various audio tasks as text-generation problems. It processes both audio recordings and text inputs to produce free-form text outputs. Its unified architecture supports a wide range of tasks, from open-ended to close-ended, without needing additional fine-tuning or specific task extensions.

C.2 Unimodal Text Models

BERT (Bidirectional Encoder Representations from Transformers): BERT, developed by Devlin et al. (2019), is a transformer-based model designed for natural language understanding tasks. It utilizes bidirectional training to capture context from both left and right directions. BERT's base model consists of 12 layers, 768 hidden units, and 12 attention heads, producing 768-dimensional representations. Pre-trained on massive corpora, BERT is widely used for tasks like text classification, question answering, and language inference.

XLNet: XLNet, introduced by Yang et al. (2019), is a generalized autoregressive pretraining method that captures bidirectional contexts by maximizing the expected likelihood over all permutations of the input sequence. XLNet's base model consists of 12 layers, 768 hidden units, and 12 attention heads, producing 768-dimensional representations.

LLaMA-2: LLAMA-2 is an autoregressive language model (decoder-based) (Touvron et al., 2023b), designed for extensive language understanding and generation tasks. The model consists of 32 layers, and it employs 4096-dimensional representations.

FLAN-T5: FLAN-T5 (Chung et al., 2024) builds upon the original T5 (Text-to-Text Transfer Transformer) by fine-tuning it with instruction-based datasets for better generalization on a variety of tasks. The base model consists of 12 transformer layers with 768 hidden units and 12 attention heads, producing 512-dimensional representations. FLAN-T5 excels in translation, summarization, and question-answering tasks, offering robust performance across many domains.

C.3 Unimodal Speech Models

WavLM: WavLM is a self-supervised model designed for full-stack speech processing, handling tasks like ASR, speaker identification, and speech enhancement. It combines masked speech prediction and denoising in pre-training, capturing both content and non-ASR features. Using gated relative position bias in its Transformer architecture, WavLM excels at sequence modeling. Trained on 94,000 hours of data, it achieves state-of-the-art performance across benchmarks like SUPERB. We utilized the "wavlm-libri-clean-100h-base-plus" model checkpoint which consists of 12 layers and produces 768-dimensional representations.

OpenAI Whisper: Whisper is OpenAI's speech recognition model designed for automatic speech-to-text transcription. It utilizes an encoder-decoder transformer architecture with 1.6 billion parameters to process audio and generate transcriptions. Trained on a large multilingual corpus, Whisper excels in various transcription tasks, including multi-language speech recognition, speaker identification, and noise-robust processing.

Wav2Vec: Wav2Vec is a self-supervised model for speech recognition that pre-trains on raw audio data. The model learns contextualized representations by solving a contrastive task over masked audio segments. The base model consists of 12 transformer layers with 512 hidden units and 8 attention heads. Wav2Vec achieves state-of-the-art performance in ASR tasks by learning robust speech representations with minimal supervision.

We present details of these pretrained unimodal and multimodal models in Table 1.

C.4 Low-level Features

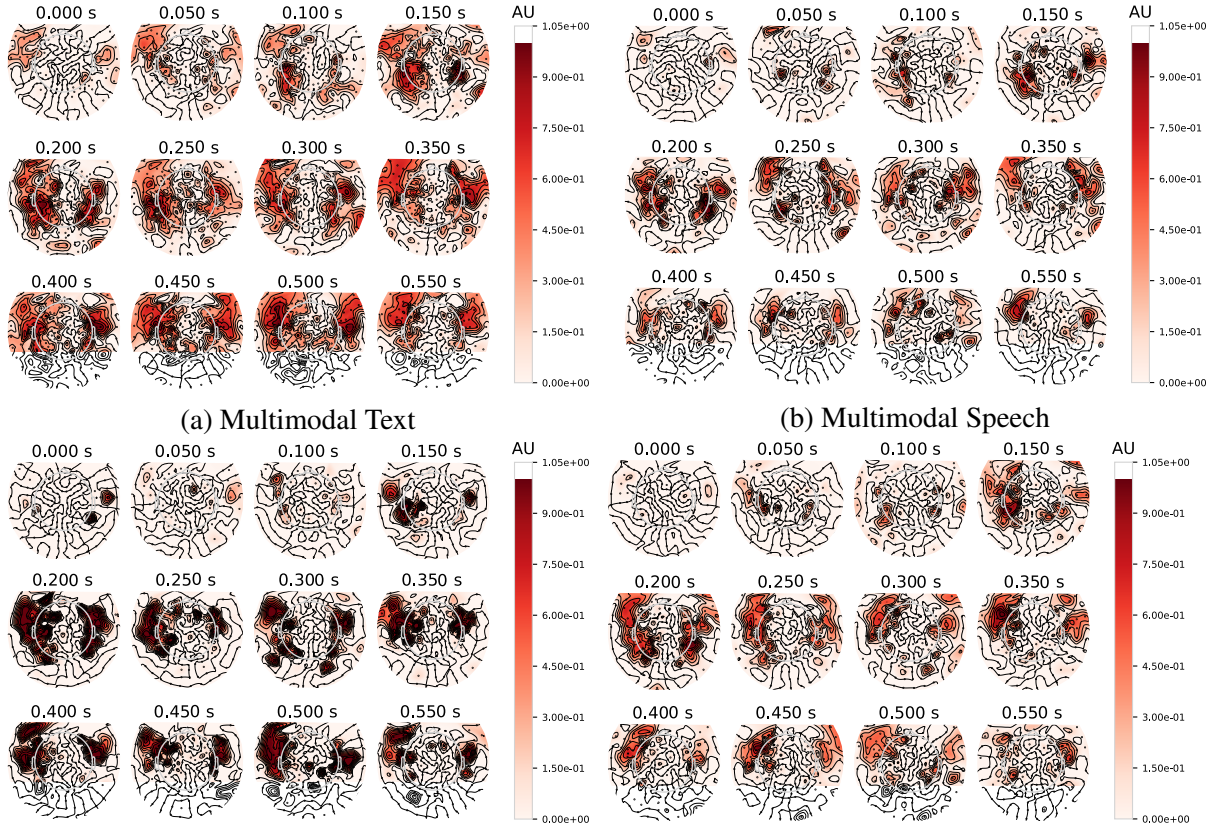
FBank: Filter bank separates the raw audio signal into multiple components (each one carrying a single frequency sub-band of the original signal) using a bandpass filter. Each raw audio signal results in 104 sized FBank vector.

Mel Spectrogram: It is computed by applying a Fourier transform on the raw audio signal to analyze a signal's frequency content and convert it to the mel-scale, yielding an 80-dimensional feature vector.

MFCC: MFCCs are Mel-frequency cepstral coefficients obtained by taking the Discrete Cosine Transform (DCT) of the spectral envelope obtained from Logarithmic filter bank outputs.

Phonation features: These are the parts of speech sounds that are related to the vibration of the vocal folds and the modification of the airstream by the larynx. We compute 29 phonation features consisting of (seven descriptors)$\times$(4 functionals: mean, std, skewness, kurtosis) + degree of Unvoiced segments.

Phonological features: These are the smallest units of distinction between any two phonemes. Phonemes are the basic sounds or signs that convey



(a) Multimodal Text

(b) Multimodal Speech

(c) Unimodal Text

(d) Unimodal Speech

Figure 8: Topo maps showing the average normalized predictivity across four different settings for MEG activity: (a) Multimodal Text, (b) Multimodal Speech, (c) Unimodal Text, and (d) Unimodal Speech. Word onset is at 0 ms.

Model Name	Pretraining data modality	# Parameters	Dataset	Layers	Backbone
CLAP	Speech+Text	154M	LAION-Audio-630K	12	SWINTransformer, RoBERTa-base
SpeechTTS	Speech+Text	150M	LibriSpeech audio	12	Wav2Vec2.0-base, Transformer
Pengi	Speech+Text	250M (90M Trainable)	3.4M audio+text pairs	12	SWINTransformer,RoBERTa-base
BERT	Text	110M	BooksCorpus, English Wikipedia	12	Transformer encoder
XLNet	Text	110M	BERT dataset, ClueWeb, Giga5	12	Transformer encoder
LLaMA-2	Text	7B	2T tokens	32	Transformer decoder
FLAN-T5	Text	250M	C4, 1,800 NLP tasks	12	T5
WavLM	Speech	94.7M	LibriSpeech	12	HuBERT
Whisper	Speech	74M	680K hours data	12	Transformer encoder-decoder
Wav2Vec	Speech	95M	LibriSpeech	12	CNN+Transformer encoder

Table 1: Multimodal and Unimodal Text and Speech Models

linguistic meaning in spoken or signed languages. We compute 108 phonological features consisting of (18 descriptors)×(6 functionals: mean, std, skewness, kurtosis, max, min).

Articulation features: Articulation features in speech refer to the physical properties of speech sounds, including how they are produced, how long they last, and their loudness. This feature set is formed with 488 features = (122 descriptors)×(4 functionals: mean, std, skewness, kurtosis).

Phonation, phonological, and articulation features are passed to speech and multimodal models to create embeddings and extract the feature representations.

C.5 Rationale for choice of unimodal and multimodal models

We experiment with 3 multimodal models, 4 unimodal text models and 3 unimodal speech models. We acknowledge that the models differ in architecture, training objectives, and data sources. However, as established in prior brain encoding studies (Schrimpf et al., 2021; Toneva and Wehbe, 2019; Antonello et al., 2021; Aw and Toneva, 2023; Oota et al., 2025), such diversity is a deliberate design choice aimed at capturing general trends in how different types of representations, regardless of specific implementation details, align with brain activity. Our goal is not to control for every architectural or dataset difference but to evaluate how effectively text, speech, and multimodal rep-

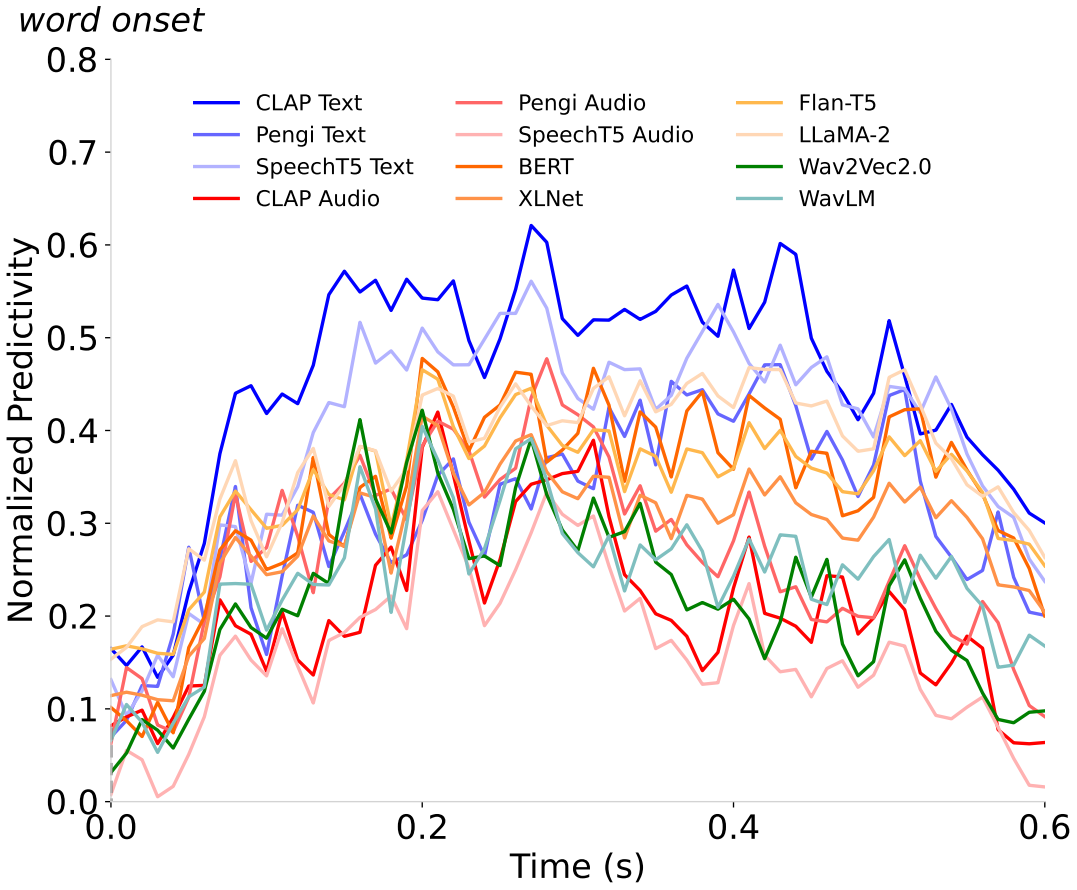


Figure 9: Individual Model Analysis: The average normalized predictivity of both text and speech embeddings from multimodal models, as well as unimodal text and speech representations, is measured by comparing predicted and real MEG activity across sensors and subjects. The word onset occurs at 0 ms.

resentations from state-of-the-art models, despite inherent differences, align with brain activity, and to identify consistent patterns in modality-specific brain alignment.

We deliberately selected BERT, LLaMA-2, XL-Net, and FLAN-T5 to represent a diverse and representative cross-section of modern text models with varying architectural and training paradigms. BERT and XLNet are well-established in the literature and offer contrasting pretraining strategies, masked language modeling versus permutation-based modeling, making them ideal for probing different semantic encoding mechanisms. LLaMA-2 brings in the perspective of large-scale autoregressive models optimized for instruction-following, while FLAN-T5 represents a strong encoder-decoder model fine-tuned on a wide range of tasks. We intentionally limited the scope to these models to maintain interpretability and avoid redundancy, as many newer models are architectural variants or scale-ups of these foundational types. Including every available model would dilute the clarity of our comparative analysis without necessarily adding new insights into brain alignment. Our goal was to strike a balance between diversity, interpretability, and relevance to current trends in both NLP and cognitive modeling.

We selected Wav2Vec, WavLM, and Whisper as our core unimodal speech models due to their strong representational capabilities and complementary design philosophies. Wav2Vec is a foundational self-supervised model that has demonstrated robust performance in learning contextualized speech representations from raw audio, making it a natural choice for probing low- and mid-level auditory features. WavLM extends this by incorporating masked prediction and speech-specific pretraining objectives, enabling it to capture richer prosodic and phonetic cues. Whisper, on the other hand, is a large-scale, multitask model trained on diverse multilingual and noisy data, offering a broader perspective on speech understanding that includes robustness to real-world variability.

We intentionally focused on these models because they represent the state-of-the-art across different axes: self-supervised learning, task general-

ization, and robustness to noise and multilinguality. Including additional models such as HuBERT, DeepSpeech, or traditional ASR systems would have introduced architectural or performance redundancies without significantly expanding the diversity of representational strategies. Our goal was to maintain a balance between coverage, comparability, and interpretability, ensuring that the models we analyzed were both representative and distinct in their approach to speech representation learning.

We selected CLAP, SpeechT5, and Pengi as our multimodal models because they represent distinct and innovative approaches to learning joint representations across text and speech, making them particularly well-suited for investigating cross-modal semantic alignment with brain activity. CLAP (Contrastive Language-Audio Pretraining) is designed to align audio and text through contrastive learning, enabling it to capture rich semantic correspondences between modalities. SpeechT5 adopts a unified encoder-decoder framework that supports multiple speech and text tasks, offering flexibility and strong performance in both modalities. Pengi, a recent model from Google, is trained on large-scale multimodal data and optimized for general-purpose speech understanding, making it a strong candidate for capturing nuanced semantic features.

We focused on these models because they are among the few that explicitly integrate text and speech in a way that supports bidirectional representation learning, which is critical for studying cross-modal transfer effects. Other multimodal models either focus on vision-language tasks or lack sufficient granularity in speech-text alignment. By choosing CLAP, SpeechT5, and Pengi, we ensured coverage of contrastive, generative, and instruction-tuned paradigms, allowing us to systematically assess how different multimodal learning strategies influence alignment with MEG brain responses.

While PaLM-Audio and MM-Whisper are promising multimodal models, we chose not to include them in our study due to practical and methodological considerations. PaLM-Audio, being part of a larger and more complex family of models, is not publicly available in a form that allows fine-grained extraction of intermediate text and speech representations needed for brain alignment analysis. Its architecture also integrates multiple modalities beyond speech and text, which could introduce confounding factors when isolating cross-modal interactions specific to language processing. MM-Whisper, although built on the robust Whisper backbone, is primarily optimized for multilingual and multitask ASR performance rather than joint semantic representation learning across modalities. In contrast, CLAP, SpeechT5, and Pengi offer clearer and more accessible pathways for extracting aligned embeddings from both text and speech, making them more suitable for systematic comparison in the context of MEG-based brain encoding. Our selection prioritizes models with transparent architecture, accessible embeddings, and explicit multimodal training objectives, ensuring interpretability and reproducibility in cognitive neuroscience research.

D Residual Approach

Specifically, given an input feature vector $\mathbf{L}_i$ with dimension $\mathbf{N} \times \mathbf{d}$ for input feature i, and target neural model representations $\mathbf{W} \in \mathbb{R}^{\mathbf{N} \times \mathbf{D}}$, where $\mathbf{N}$ denotes the number of samples, and d and D denote the dimensionality of unimodal/low-level and neural model representations, respectively, the ridge regression objective function is $f(\mathbf{L}_i) = \min_{\theta_i} \|\mathbf{W} - \mathbf{L}_i\theta_i\|_F^2 + \lambda\|\theta_i\|_F^2$ where θ_i denotes the learned weight coefficient for embedding dimension $\mathbf{D}$ for the input feature i, $\|.\|_F^2$ denotes the Frobenius norm, and $\lambda > 0$ is a tunable hyper-parameter representing the regularization weight for each feature dimension. Using the learned weight coefficients, we compute the residuals as follows: $r(\mathbf{L}_i) = |\mathbf{W} - \mathbf{L}_i\theta_i|$.

E Implementation details for reproducibility

All experiments were conducted on a machine with 1 NVIDIA GeForce-GTX GPU with 16GB GPU RAM. We used cross-validated ridge-regression with MSE loss function; L2-decay (λ) varied from 10^1 to 10^3. Best λ was chosen by tuning on validation data that comprised a randomly chosen 10% subset from train set used only for hyper-parameter tuning.

F Encoding Performance of Multimodal Models vs. Unimodal Models

Multimodal text embeddings exhibit a peak around 200ms, suggesting that they benefit from speech embeddings, with heightened activity during this time period. Specifically, the topographical maps shown in Fig. 8 reveal that the higher normalized

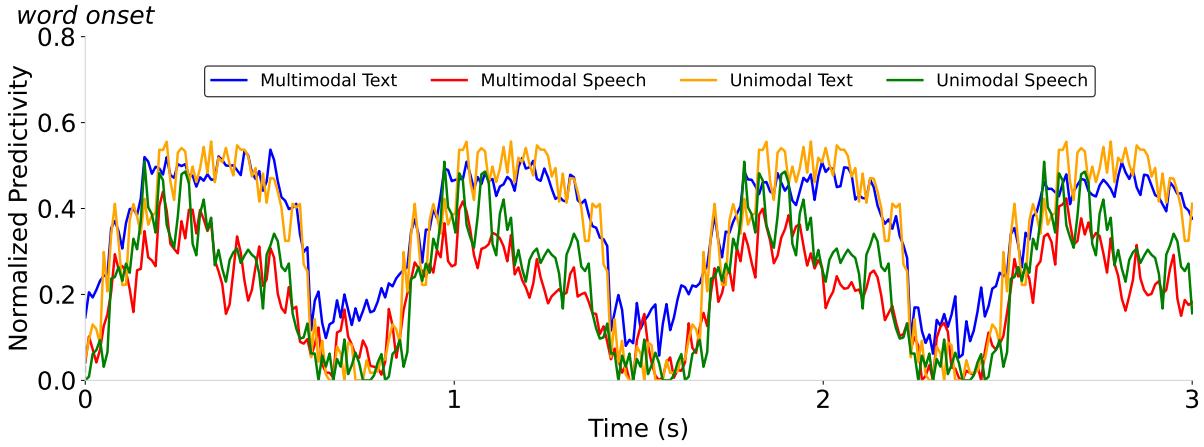


Figure 10: Extended input context and prolonged MEG signal: The average normalized predictivity of both text and speech embeddings from multimodal models, as well as unimodal text and speech representations, is measured by comparing predicted and real MEG activity across sensors and subjects. The word onset occurs at 0 ms.

brain predictivity in primary auditory regions is unique to multimodal text embeddings and is not observed in unimodal text models. Both multimodal and unimodal text embeddings remain active beyond 350ms, aligning with the timeframe associated with semantic word processing. In contrast, speech embeddings experience a sharp decline in performance around 350ms. Topographical maps in Fig. 8 show that temporal, frontal, and parietal language regions exhibit higher predictivity during this period, whereas speech embeddings display low predictivity for both multimodal and unimodal models.

Fig. 9 presents the average normalized predictivity for text and speech embeddings from individual multimodal models (aggregated across subjects) as well as unimodal text and speech embeddings. For clarity, we use a distinct color scheme to differentiate: Multimodal text embeddings (blue palette), Multimodal speech embeddings (red palette), Unimodal text embeddings (yellow), and Unimodal speech embeddings (green). From Fig. 9, we observe that, consistent with the averaged trends in Fig. 3. Multimodal text embeddings (CLAP Text, Pengi Text, SpeechT5 Text) exhibit consistently higher normalized predictivity across the entire time window compared to multimodal speech, unimodal text, and unimodal speech models. Among these, CLAP Text and Speech-T5 Text align with enhanced brain alignment during early auditory processing, as highlighted in our main results.

G Impact of extended input context and prolonged MEG signal on brain predictivity

We further explored the impact of varying the input context length on model performance, as shown in Fig. 10. The primary goal of this experiment is to assess whether providing a longer temporal context to the models improves their alignment with brain activity during naturalistic speech comprehension, particularly in time windows associated with higher-level semantic processing.

In this analysis, we expanded the MEG recording window from 800 ms to 3 seconds. For example, given a story comprised of multiple speech files (with each file representing a single word) and a context length of 5, we input the sequence (w_1, w_2, w_3, w_4, w_5) and use the representation of the last word (w_5) as the target.

From Fig. 10, we observe that, similar to the single-word context, an increase in context length further demonstrates that both multimodal text and unimodal text embeddings exhibit a higher degree of brain predictivity than speech embeddings across the extended temporal window. Our findings indicate that extending the input context not only reinforces the robustness of text embeddings but also underscores their superior alignment with prolonged brain activity compared to speech embeddings. This experiment also provides additional evidence for asymmetric knowledge transfer across modalities: text embeddings integrate contextual and cross-modal information effectively, while speech embeddings do not, even with access to both modalities during training.