Novelty Assessment Report

Paper: Aligning the Brain with Language Models Through a Nonlinear and Multimodal Approach

PDF URL: https://openreview.net/pdf?id=cu6xWUNOzQ

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Self-supervised language and audio models effectively predict brain responses to speech. However, while nonlinear approaches have become standard in vision encoding, speech encoding models still predominantly rely on linear mappings from unimodal features. This linear approach fails to capture the complex integration of auditory signals with linguistic information across widespread brain networks during speech comprehension. Here, we introduce a nonlinear, multimodal prediction model that combines audio and linguistic features from pre-trained models (e.g., Llama, Whisper). Our approach achieves a 17.2% and 17.9% improvement in prediction performance (unnormalized and normalized correlation) over traditional unimodal linear models, as well as a 7.7% and 14.4% improvement over prior state-of-the-art models relying on weighted averaging of linear unimodal predictions. These substantial improvements not only represent a major step towards future robust in-silico testing and improved decoding performance, but also reveal distributed multimodal processing patterns across the cortex that support key neurolinguistic theories including the Motor Theory of Speech Perception, Convergence-Divergence Zone model, and embodied semantics. Overall, our work highlights the often neglected potential of nonlinear and multimodal approaches to speech encoding, paving the way for future studies to embrace these strategies in naturalistic neurolinguistics research.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Predicting Brain Responses to Naturalistic Speech from Multimodal Features**

A total of **39 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Multimodal Feature Integration for Brain Encoding**
- **Neural Tracking and Temporal Dynamics of Speech**
- **Spatial and Network-Level Organization**
- **Predictive Processing and Contextual Integration**
- **Clinical and Special Populations**
- **Methodological Advances and Datasets**
- **Multimodal Interaction and Occlusion Effects**
- **Behavioral Modeling and Listener Agent Systems**

Complete Taxonomy Tree

- Predicting Brain Responses to Naturalistic Speech from Multimodal Features Survey Taxonomy
- Multimodal Feature Integration for Brain Encoding
  - Nonlinear Multimodal Prediction Models ★ (3 papers)
  - [0] Aligning the Brain with Language Models Through a Nonlinear and Multimodal Approach (Anon et al., 2026) View paper
  - [10] Mind the Gap: Aligning the Brain with Language Models Requires a Nonlinear and Multimodal Approach (Cha Jiook, 2025) View paper
  - [20] A Multimodal Seq2Seq Transformer for Predicting Brain Responses to Naturalistic Stimuli (HE Qianyi, 2025) View paper
  - Linear and Weighted Averaging Approaches (2 papers)
  - [1] Interpretable prediction of brain activity during conversations from multimodal behavioral signals (Youssef Hmamouche, 2024) View paper
  - [24] Probing Multimodal Fusion in the Brain: The Dominance of Audiovisual Streams in Naturalistic Encoding (Abdollahi Hamid, 2025) View paper
  - Language Model Feature Representations (2 papers)
  - [3] Inducing brain-relevant bias in natural language processing models (Schwartz, 2019) View paper
  - [31] Do Feature Representations from Different Language Models Affect Accuracy of Brain Encoding Models' Predictions? (Muxuan Liu, 2024) View paper
- Neural Tracking and Temporal Dynamics of Speech
  - Acoustic and Linguistic Temporal Coupling (3 papers)
  - [14] Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations (Ariel Goldstein, 2023) View paper
  - [23] Brain activity reflects the predictability of word sequences in listened continuous speech (Miika Koskinen, 2020) View paper
  - [29] Speakerâ□□listener neural coupling correlates with semantic and acoustic features of naturalistic speech (Zhuoran Li, 2024) View paper
  - Multi-Timescale Representation and Prediction (2 papers)
  - [25] Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech (Shailee Jain, 2020) View paper

  - [30] Neural dynamics express syntax in the time domain during natural story listening (Cas W. Coopmans, 2024) View paper
  - Oscillatory Phase and Cortical Tracking (2 papers)
  - [2] Cortical tracking of continuous speech under bimodal divided attention (Zilong Xie, 2023) View paper
  - [36] The phase of cortical oscillations determines the perceptual fate of visual cues in naturalistic audiovisual speech. (RaphaÃ«l ThÃ©zÃ©, 2022) View paper
- Spatial and Network-Level Organization
  - Functional Connectivity and Network Coupling (2 papers)
  - [11] Cortical language areas are coupled via a soft hierarchy of model-based linguistic features (Ahmad Samara, 2025) View paper
  - [37] Neural dSCA: demixing multimodal interaction among brain areas during naturalistic experiments (Takagi, 2022) View paper
  - Regional Specialization and Modality Effects (3 papers)
  - [4] Neural representation of sensorimotor features in language-motor areas during auditory and visual perception (Yuanyi Zheng, 2025) View paper
  - [6] Fragmentation and multithreading of experience in the default-mode network (Fahd Yazin, 2025) View paper
  - [8] Modality-specific and amodal language processing by single neurons (Yair Lakretz, 2024) View paper
  - Cross-Modal Plasticity and Reorganization (1 papers)
  - [38] Cross-modal activity in adults with cochlear implants: A multimodal brain perspective (Fullerton, 2020) View paper
- Predictive Processing and Contextual Integration
  - Discourse Context and Hierarchical Prediction (1 papers)
  - [21] Discourse context and co-speech gestures jointly shape hierarchical prediction during the processing of a multimodal narrative (Yifei He, 2025) View paper
  - Co-Speech Gesture Integration (3 papers)
  - [5] The role of co-speech gestures in retrieval and prediction during naturalistic multimodal narrative processing (Sergio Osorio, 2024) View paper
  - [15] More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension (Ye Zhang, 2021) View paper
  - [18] The role of multimodal cues in second language comprehension (Ye Zhang, 2023) View paper
- Clinical and Special Populations
  - Neurodegenerative and Psychiatric Disorders (2 papers)
  - [12] Multimodal neurocognitive markers of naturalistic discourse typify diverse neurodegenerative diseases (A. Birba, 2022) View paper
  - [33] The processing of semantic complexity and co-speech gestures in schizophrenia: a naturalistic, multimodal fMRI study (Paulina Cuevas, 2021) View paper
  - Cochlear Implant Users and Autism Spectrum (2 papers)
  - [17] Impaired neural encoding of naturalistic audiovisual speech in autism (Theo Vanneau, 2025) View paper
  - [27] Neural responses to naturalistic audiovisual speech are related to listening demand in cochlear implant users (Bowen Xiu, 2022) View paper
- Methodological Advances and Datasets
  - Intracranial Electrophysiology Datasets and Benchmarks (4 papers)
  - [9] Alignment of Large Language Models and Brain Activity: Exploring Language Processing through sEEG in a Multimodal Syntactic Task (Gillioz, 2024) View paper
  - [16] The âPodcastâ ECoG dataset for modeling neural activity during natural language comprehension (Zaid Zada, 2025) View paper
  - [22] Neuroprobe: Evaluating Intracranial Brain Responses to Naturalistic Stimuli (Wang, 2025) View paper
  - [32] 450âA Shared Cortical Language Network for Multimodal Naming (Kathryn Snyder, 2023) View paper
  - Non-Invasive Multimodal Recording Techniques (5 papers)
  - [7] Decoding EEG brain activity for multi-modal natural language processing (Hollenstein, 2021) View paper
  - [13] Multisensory naturalistic decoding with high-density diffuse optical tomography (Kalyan Tripathy, 2025) View paper
  - [19] Generalizable EEG encoding models with naturalistic audiovisual stimuli (Maansi Desai, 2021) View paper
  - [28] Combining EEG and 3D-eye-tracking to study the prediction of upcoming speech in naturalistic virtual environments: A proof of principle. (E. Huizeling, 2023) View paper
  - [39] Multimodal imaging of language perception (Johanna Vartiainen, 2010) View paper
- Multimodal Interaction and Occlusion Effects (2 papers)
  - [34] Occlusion of lip movements impairs reconstruction of acoustic speech features and higher-level segmentational features in the presence of a distractor speaker (Chandra Leon Haider, 2021) View paper
  - [35] Electrophysiological Investigations of Attention and Audiovisual Integration During Naturalistic Speech Perception (Ahmed, 2023) View paper
- Behavioral Modeling and Listener Agent Systems (1 papers)
  - [26] An investigation on the effectiveness of multimodal fusion and temporal feature extraction in reactive and spontaneous behavior generative RNN models for listener â¦ (HH Huang, 2019) View paper

Narrative

Core task: predicting brain responses to naturalistic speech from multimodal features. The field has organized itself around several complementary perspectives. One major branch focuses on multimodal feature integration for brain encoding, exploring how acoustic, visual, and linguistic cues combine to drive neural activity—ranging from linear models to more sophisticated nonlinear architectures. A second branch examines neural tracking and temporal dynamics, investigating how the brain follows speech at multiple timescales and how oscillatory mechanisms support comprehension. Spatial and network-level organization studies map where different features are processed across cortical regions, while predictive processing frameworks ask how context and expectation shape neural responses. Additional branches address clinical and special populations (e.g., cochlear implant users, individuals with autism or schizophrenia), methodological advances including new datasets and recording techniques, multimodal interaction effects such as audiovisual occlusion, and even behavioral modeling of listener agents that simulate human-like responses.

Within the multimodal integration branch, a particularly active line of work contrasts linear versus nonlinear prediction models. Earlier efforts often relied on additive or simple weighted combinations of features, but recent studies reveal that nonlinear interactions—captured by neural networks or kernel methods—can substantially improve encoding accuracy. Nonlinear Brain Language Alignment[0] sits squarely in this nonlinear modeling cluster, emphasizing how deep architectures better capture the complex, context-dependent mappings between multimodal inputs and brain signals. It shares methodological kinship with Nonlinear Multimodal Gap[10], which similarly highlights the limitations of linear assumptions, and contrasts with more traditional approaches that treat modalities as independent additive components. Meanwhile, neighboring work such as Multimodal Seq2Seq Transformer[20] explores sequence-to-sequence architectures for similar prediction tasks, underscoring an ongoing shift toward flexible, data-driven models that can learn intricate feature interactions directly from naturalistic stimuli.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Mind the Gap: Aligning the Brain with Language Models Requires a Nonlinear and Multimodal Approach

Authors: Cha Jiook, Lee, Jay-Yoon | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Self-supervised language and audio models effectively predict brain responses to speech. However, traditional prediction models rely on linear mappings from unimodal features, despite the complex integration of auditory signals with linguistic and semantic information across widespread brain networks during speech comprehension. Here, we introduce a nonlinear, multimodal prediction model that combines audio and linguistic features from pre-trained models (e.g., LLAMA, Whisper). Our approach achi...

⚠ Similarity Notice

These papers appear to be highly similar or the same work, both presenting a nonlinear multimodal approach combining LLAMA and Whisper features for predicting brain responses to speech. They report identical performance improvements (17.2% and 17.9% over unimodal linear baselines, 7.7% and 14.4% over prior state-of-the-art), use the same dataset (LeBel et al., 2023), employ identical architectures (MLP with PCA preprocessing), and present the same core findings about multimodal integration and neurolinguistic theory alignment. The titles differ slightly, but the technical content, methodology, results, and contributions are essentially identical, strongly suggesting these are variants of the same paper.

2. A Multimodal Seq2Seq Transformer for Predicting Brain Responses to Naturalistic Stimuli

Authors: HE Qianyi, Qianyi He, Yuan Chang Leong | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The Algonauts 2025 Challenge called on the community to develop encoding models that predict whole-brain fMRI responses to naturalistic multimodal movies. In this submission, we propose a sequence-to-sequence Transformer that autoregressively predicts fMRI activity from visual, auditory, and language inputs. Stimulus features were extracted using pretrained models including VideoMAE, HuBERT, Qwen, and BridgeTower. The decoder integrates information from prior brain states and current stimuli via...

Relationship Analysis

Both papers belong to the Nonlinear Multimodal Prediction Models category, using nonlinear architectures to integrate multimodal features for predicting brain responses to naturalistic speech. They overlap in their use of pretrained models (LLaMA/Qwen for language, Whisper/HuBERT for audio) and nonlinear fusion mechanisms (MLP vs. Transformer) to capture complex audio-linguistic interactions. The key difference is that the original paper uses a single-hidden-layer MLP with PCA dimensionality reduction and focuses on variance partitioning to reveal distributed multimodal processing patterns, while the candidate employs a sequence-to-sequence Transformer with autoregressive decoding, cross-attention over narrative summaries, and subject-specific decoder heads to capture temporal dependencies and individual variability.

Contributions Analysis

Overall novelty summary. The paper introduces a nonlinear multimodal encoding model that combines audio and linguistic features from pre-trained models (Llama, Whisper) to predict brain responses to naturalistic speech. It resides in the 'Nonlinear Multimodal Prediction Models' leaf, which contains only three papers total, indicating a relatively sparse but emerging research direction. This leaf sits within the broader 'Multimodal Feature Integration for Brain Encoding' branch, which also includes linear and weighted averaging approaches as well as language model feature representation studies, suggesting the field is actively exploring different integration strategies.

The taxonomy reveals that the paper's immediate neighbors include linear and weighted averaging approaches (a separate leaf with two papers) and language model feature representation studies (another leaf with two papers). Beyond this branch, the field encompasses temporal dynamics research (acoustic/linguistic coupling, multi-timescale modeling, oscillatory tracking), spatial organization studies (connectivity, regional specialization), and predictive processing frameworks. The paper's focus on nonlinear integration distinguishes it from the linear methods in adjacent leaves, while its use of pre-trained models connects it to the language model feature representation work, though that leaf emphasizes architecture comparisons rather than nonlinear integration.

Among the three contributions analyzed, the first two—the nonlinear multimodal encoding model and the demonstration of nonlinear multimodal interactions—each examined ten candidates and found one refutable prior work, suggesting some overlap with existing literature within the limited search scope of thirty total candidates. The third contribution, RED-based clustering analysis for spatiotemporal tracking, examined ten candidates with none appearing to refute it, indicating this methodological component may be more novel. The analysis explicitly notes this is based on top-K semantic search plus citation expansion, not an exhaustive review, so these findings reflect the most semantically similar work rather than the entire field.

Given the limited search scope and the sparse population of the target leaf (three papers), the work appears to advance an emerging direction in brain encoding research. The contribution-level statistics suggest the core modeling approach has some precedent among the thirty candidates examined, while the spatiotemporal analysis method shows less overlap. The taxonomy structure indicates this sits at the intersection of multiple active research threads—multimodal integration, nonlinear modeling, and naturalistic speech processing—where methodological innovation is ongoing.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Nonlinear multimodal encoding model for naturalistic speech

Description: The authors introduce a nonlinear encoding model that combines audio features from Whisper and semantic features from language models like Llama using a single-hidden-layer MLP with PCA preprocessing. This approach achieves substantial improvements (17.2% and 17.9% in unnormalized and normalized correlation) over traditional linear unimodal baselines and reveals distributed multimodal processing patterns across the cortex.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. RobinNet: A Multimodal Speech Emotion Recognition System With Speaker Recognition for Social Interactions

URL: View paper

Brief Assessment

RobinNet Emotion Recognition[44] focuses on speech emotion recognition using RoBERTa and Inception-ResNet-v2 for classification tasks, not brain encoding models that predict cortical activity from naturalistic speech stimuli.

2. Thinking with Sound: Audio Chain-of-Thought Enables Multimodal Reasoning in Large Audio-Language Models

URL: View paper

Brief Assessment

Audio Chain of Thought[43] focuses on audio-language models for audio understanding tasks (speech translation, audio Q&A) with acoustic tool integration, not on brain encoding models that predict fMRI responses from speech stimuli using nonlinear multimodal architectures.

3. Text-Infused Audio-Visual Video Parsing with Semantic-Aware Multimodal Contrastive Learning

URL: View paper

Brief Assessment

Text Infused Parsing[41] addresses audio-visual video parsing for event recognition in video segments, not brain encoding models for naturalistic speech comprehension. The candidate's multimodal approach combines audio, visual, and text modalities for video event parsing tasks, which is fundamentally different from the original paper's fMRI-based speech encoding framework that predicts cortical activity from audio and semantic features.

4. Multimodal fusion for multimedia analysis: a survey

URL: View paper

Brief Assessment

Multimodal Fusion Survey[47] is a general survey paper on multimodal fusion techniques for multimedia analysis. It does not present specific nonlinear encoding models for naturalistic speech processing or brain activity prediction, which is the core contribution of the original paper.

5. TD-PLC: A Semantic-Aware Speech Encoding for Improved Packet Loss Concealment

URL: View paper

Brief Assessment

TD-PLC[45] focuses on packet loss concealment in real-time communication networks by integrating semantic information with audio for transmission purposes, not on predicting brain responses to naturalistic speech using nonlinear encoding models.

6. Flow-SLM: Joint Learning of Linguistic and Acoustic Information for Spoken Language Modeling

URL: View paper

Brief Assessment

Flow-SLM[40] focuses on generative modeling of speech tokens using flow matching for spoken language generation, not on encoding brain activity from naturalistic speech stimuli using nonlinear multimodal models.

7. Mind the Gap: Aligning the Brain with Language Models Requires a Nonlinear and Multimodal Approach

URL: View paper

Prior Art Analysis

Nonlinear Multimodal Gap[10] demonstrates that a nearly identical approach was published prior to the original paper's submission. Both papers introduce nonlinear encoding models combining audio features from Whisper and semantic features from Llama using single-hidden-layer MLPs with PCA preprocessing for naturalistic speech. The candidate reports improvements of 17.2% and 17.9% in unnormalized and normalized correlation over linear unimodal baselines, matching the original's exact performance claims. Both papers use the same dataset (Lebel et al., 2023), identical feature extraction methods (Llama and Whisper models), the same PCA dimensionality reduction (512 components), and the same MLP architecture (single hidden layer, 256 units). The extensive overlap in methodology, results, and even specific numerical values strongly suggests this is prior work that refutes the novelty claim.

Evidence

Evidence 1 - **Rationale**: Both papers claim identical improvements (17.2% and 17.9%) using the same approach (nonlinear multimodal model combining Llama and Whisper features), demonstrating that Nonlinear Multimodal Gap[10] presented this contribution before the original paper. - **Original**: we introduce a nonlinear, multimodal prediction model that combines audio and linguistic features from pre-trained models (e.g., llama, whisper). our approach achieves a 17.2% and 17.9% improvement in prediction performance (unnormalized and normalized correlation) over traditional unimodal linear m... - **Candidate**: we introduce a nonlinear, multimodal prediction model that combines audio and linguistic features from pre-trained models (e.g., llama, whisper). our approach achieves a 17.2% and 17.9% improvement in prediction performance (unnormalized and normalized correlation) over traditional unimodal linear m...

Evidence 2 - **Rationale**: Both papers employ the same architectural choices: PCA preprocessing combined with a single-hidden-layer MLP, demonstrating identical methodological approaches that were already published in the candidate. - **Original**: with only pca and a single-hidden-layer mlp, our nonlinear multimodal encoder improves prediction accuracy by17.2%(unnormalized) and17.9%(normalized) over the standard semantic linear baseline - **Candidate**: • multi-layer perceptron (mlp): mlp with a single hidden layer of 256 units. we predicted pca-reduced fmri representations, rather than the full voxel space... in detail, we applied pca to the aggregate fmri response matrix yorg ∈ rntrxnvoxels , reducing its dimensionality to ypca ∈ rntrxnpca

Evidence 3 - **Rationale**: Both papers use the exact same dataset with identical specifications (3 subjects, 20 hours, 95 stories, 33,000 time points), confirming that the candidate paper already applied this approach to the same data. - **Original**: we used a public fmri dataset (lebel et al., 2023) of three subjects listening to 20 hours of english podcast. training data included 95 stories across 20 scanning sessions (33,000 time points). - **Candidate**: we used a publicly available fmri dataset (lebel et al., 2023; tang et al., 2023) of three subjects listening to approximately 20 hours of english podcast. the training data comprised 95 stories across 20 scanning sessions (approximately 33,000 time points).

8. Separating the â□□Chirpâ□□ from the â□□Chatâ□□: Self-supervised Visual Grounding of Sound and Language

URL: View paper

Brief Assessment

Chirp Chat Grounding[48] focuses on audio-visual grounding in videos for sound localization and speech-prompted segmentation, not on fMRI brain encoding or predicting neural responses to speech stimuli.

9. Speech recognition and intelligent translation under multimodal humanâ□□computer interaction system

URL: View paper

Brief Assessment

Intelligent Translation System[46] focuses on speech-to-text translation with multimodal gating fusion for translation tasks, not on nonlinear encoding models that predict brain activity from audio and semantic features during naturalistic speech comprehension.

10. Multi-modal multi-channel target speech separation

URL: View paper

Brief Assessment

Multi Channel Separation[42] focuses on target speech separation from overlapped audio using visual and spatial modalities in a signal processing context, not on brain encoding models predicting fMRI responses to naturalistic speech stimuli.

Contribution 2: Demonstration that nonlinear multimodal interactions drive encoding improvements

Description: The authors systematically compare linear models, reduced-rank linear models (MLLinear), and delayed interaction MLPs (DIMLP) to isolate the contribution of nonlinearity versus dimensionality reduction. They show that linear models fail to capture complex interactions between audio and language information, whereas nonlinear encoders model these interactions more effectively with fewer parameters.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Predicting 2-year neurodevelopmental outcomes in preterm infants using multimodal structural brain magnetic resonance imaging with local connectivity

URL: View paper

Brief Assessment

Preterm Connectivity Prediction[63] focuses on predicting neurodevelopmental outcomes in preterm infants using brain MRI, not on multimodal neural encoding models for speech comprehension. The technical domains and research objectives are fundamentally different.

2. Multimodal Brain Growth Patterns: Insights from Canonical Correlation Analysis and Deep Canonical Correlation Analysis with Auto-Encoder

URL: View paper

Brief Assessment

Canonical Correlation Growth[62] focuses on brain development patterns using neuroimaging (T1/diffusion MRI), comparing linear CCA versus nonlinear DCCAE for gray/white matter growth analysis. This is fundamentally different from the original paper's speech encoding task using audio-language models to predict fMRI responses.

3. Reconstructing nonlinear dynamical systems from multi-modal time series

URL: View paper

Brief Assessment

Nonlinear Dynamical Reconstruction[64] focuses on reconstructing dynamical systems from multi-modal time series using RNNs with different observation models (Gaussian, categorical), not on comparing linear vs. nonlinear multimodal interactions in neural encoding models for speech/language processing.

4. Linearâ□□Nonlinear Feature Reconstruction Network for Emotion Recognition From Brain Functional Connectivity

URL: View paper

Brief Assessment

Linear Nonlinear Reconstruction[61] focuses on EEG-based emotion recognition using brain functional connectivity features, not on multimodal speech encoding with audio and language models. The technical domains and methodologies are fundamentally different.

5. Intrinsic dimension correlation: uncovering nonlinear connections in multimodal representations

URL: View paper

Brief Assessment

Intrinsic Dimension Correlation[65] focuses on measuring nonlinear correlations between high-dimensional manifolds using intrinsic dimension as a metric, not on comparing linear versus nonlinear multimodal encoding models for neural activity prediction. The technical domains differ fundamentally: correlation measurement versus neural encoding architecture comparison.

6. Neural Mixed Effects for Nonlinear Personalized Predictions

URL: View paper

Brief Assessment

Neural Mixed Effects[67] focuses on personalized prediction with nonlinear person-specific parameters in sequential tasks (e.g., mood prediction), not on multimodal neural encoding models for brain activity prediction from speech stimuli.

7. Mind the Gap: Aligning the Brain with Language Models Requires a Nonlinear and Multimodal Approach

URL: View paper

Prior Art Analysis

Nonlinear Multimodal Gap[10] systematically demonstrates that nonlinear multimodal interactions drive encoding improvements through the same comparative framework as the original paper. The candidate introduces and compares MLLinear (reduced-rank linear) and DIMLP (delayed interaction MLP) architectures to isolate nonlinearity from dimensionality reduction, showing that linear models fail to capture complex audio-language interactions. The candidate reports that DIMLP yields a 2.0% gain over linear models (4.10% to 4.18% r^2), while full MLP achieves an additional 2.6% gain (4.18% to 4.29%), demonstrating that cross-modal nonlinear interactions are most significant. This matches the original's methodology and conclusions, establishing prior work on this specific contribution.

Evidence

Evidence 1 - **Rationale**: Both papers demonstrate that linear models cannot capture complex audio-language interactions and that nonlinear encoders are more effective, with the candidate establishing this experimental design and conclusion first. - **Original**: linear models fail to capture the complex interactions between audio and language information in llm embeddings, whereas our nonlinear encoders model these interactions more effectively and with fewer parameters. - **Candidate**: to assess the role of nonlinear cross-modal interactions, we developed a delayed interaction mlp (dimlp), which processes audio and semantic features separately before a final linear fusion stage. this contrasts with mlp, which allows full nonlinear interactions across modalities. this design enable...

Evidence 2 - **Rationale**: The candidate reaches the same conclusion through systematic comparison: nonlinear multimodal interactions are essential and drive improvements over linear models, demonstrating prior establishment of this finding. - **Original**: Through systematic comparisons, we show nonlinear multimodal interactions drives these improvements.linear models fail to capture the complex interactions between audio and language information in llm embeddings - **Candidate**: this suggests that nonlinear interactions between audio and semantic features are essential for modeling the complex, distributed neural representations underlying speech comprehension

8. Nonlinear fusion is optimal for a wide class of multisensory tasks

URL: View paper

Brief Assessment

Nonlinear Multisensory Fusion[66] focuses on multisensory integration tasks in animals (vision/hearing) using ideal observers and spiking neural networks, not on neural encoding models for speech comprehension using fMRI data with audio-language features from pre-trained models.

9. LinBridge: A Learnable Framework for Interpreting Nonlinear Neural Encoding Models

URL: View paper

Brief Assessment

LinBridge Interpretation[59] focuses on interpreting nonlinear neural encoding models through Jacobian analysis in visual cortex, not on comparing linear versus nonlinear multimodal interactions in speech encoding with audio and language features.

10. Simple but Effective Raw-Data Level Multimodal Fusion for Composed Image Retrieval

URL: View paper

Brief Assessment

Raw Data Fusion[60] focuses on composed image retrieval with vision-language models, not neural encoding of brain responses to speech. The domains, tasks, and evaluation contexts are fundamentally different.

Contribution 3: RED-based clustering analysis for spatiotemporal neural response tracking

Description: The authors propose Relative Error Difference (RED) as a metric that preserves temporal dynamics alongside spatial patterns, enabling joint analysis of spatiotemporal organization. This approach achieves superior functional clustering compared to linear encoders and standard connectivity analysis, revealing previously hidden patterns of brain organization and language processing dynamics.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity

URL: View paper

Brief Assessment

Continuous Phoneme Decoding[50] focuses on speech recognition from neural signals using temporal dynamics modeling, not on clustering analysis methods for tracking neural responses across space and time.

2. Spatiotemporal dynamics of activation in motor and language areas suggest a compensatory role of the motor cortex in second language processing

URL: View paper

Brief Assessment

Motor Cortex Compensation[49] focuses on MEG-based cluster permutation testing for motor vs. language ROI activation patterns in L1/L2 processing, not on developing spatiotemporal clustering methods for tracking neural responses across space and time.

3. Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension

URL: View paper

Brief Assessment

Speech Source Dynamics[52] focuses on computing spatio-temporal response functions for speech stimuli but does not describe a RED (Relative Error Difference) metric or clustering methodology comparable to the original paper's approach for tracking neural responses.

4. Spatiotemporal dynamics of word processing in the human brain

URL: View paper

Brief Assessment

Word Processing Dynamics[53] focuses on electrocorticography (ECoG) recordings during word processing tasks, using high gamma activity to track cortical activation sequences. The original paper's RED metric and clustering approach for fMRI speech encoding represents a distinct methodological contribution not addressed in this candidate.

5. Combinatorics At-a-Glance: On the Spatiotemporal Dynamics of Temporally Unstructured Language

URL: View paper

Brief Assessment

Combinatorics Spatiotemporal Dynamics[51] mentions ANOVA for spatiotemporal clustering but does not provide sufficient methodological detail to demonstrate prior work on RED-based metrics or joint spatiotemporal analysis approaches comparable to the original paper's contribution.

6. The time-course and spatial distribution of brain activity associated with sentence processing

URL: View paper

Brief Assessment

Sentence Processing Distribution[57] focuses on identifying spatio-temporal patterns of brain activity during sentence processing using MEG and eye-tracking, not on developing clustering analysis methods for tracking neural responses over time and space as proposed in the original paper's RED metric.

7. Spatio-temporal analysis of electric brain activity during semantic and phonological word processing

URL: View paper

Brief Assessment

Semantic Phonological Analysis[58] focuses on electric field analysis for temporal dynamics in language processing, not on clustering methods for spatiotemporal neural response tracking. The candidate does not present a comparable clustering methodology.

8. Mapping, learning, visualization, classification, and understanding of fMRI data in the NeuCube evolving spatiotemporal data machine of spiking neural networks

URL: View paper

Brief Assessment

NeuCube Mapping[56] focuses on fMRI voxel mapping and STDP-based connectivity learning in spiking neural networks, not on RED metrics or clustering methods for tracking temporal neural dynamics during language processing.

9. Spatiotemporal Contributions to Pre-speech Semantic and Syntactic Processing

URL: View paper

Brief Assessment

Pre Speech Processing[55] focuses on intracranial EEG analysis of semantic/syntactic processing during word reading tasks, not on fMRI-based spatiotemporal clustering methods for tracking neural responses to naturalistic speech.

10. Recreating Neural Activity During Speech Production with Language and Speech Model Embeddings

URL: View paper

Brief Assessment

Speech Production Recreation[54] focuses on reconstructing neural activity during speech production using language/speech model embeddings with elasticnet regression. It does not propose clustering methods for spatiotemporal neural response tracking or introduce metrics like RED for preserving temporal dynamics alongside spatial patterns.

Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Mind the Gap: Aligning the Brain with Language Models Requires a Nonlinear and Multimodal Approach

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Aligning the Brain with Language Models Through a Nonlinear and Multimodal Approach View paper
- [1] Interpretable prediction of brain activity during conversations from multimodal behavioral signals View paper
- [2] Cortical tracking of continuous speech under bimodal divided attention View paper
- [3] Inducing brain-relevant bias in natural language processing models View paper
- [4] Neural representation of sensorimotor features in language-motor areas during auditory and visual perception View paper
- [5] The role of co-speech gestures in retrieval and prediction during naturalistic multimodal narrative processing View paper
- [6] Fragmentation and multithreading of experience in the default-mode network View paper
- [7] Decoding EEG brain activity for multi-modal natural language processing View paper
- [8] Modality-specific and amodal language processing by single neurons View paper
- [9] Alignment of Large Language Models and Brain Activity: Exploring Language Processing through sEEG in a Multimodal Syntactic Task View paper
- [10] Mind the Gap: Aligning the Brain with Language Models Requires a Nonlinear and Multimodal Approach View paper
- [11] Cortical language areas are coupled via a soft hierarchy of model-based linguistic features View paper
- [12] Multimodal neurocognitive markers of naturalistic discourse typify diverse neurodegenerative diseases View paper
- [13] Multisensory naturalistic decoding with high-density diffuse optical tomography View paper
- [14] Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations View paper
- [15] More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension View paper
- [16] The â□□Podcastâ□□ ECoG dataset for modeling neural activity during natural language comprehension View paper
- [17] Impaired neural encoding of naturalistic audiovisual speech in autism View paper
- [18] The role of multimodal cues in second language comprehension View paper
- [19] Generalizable EEG encoding models with naturalistic audiovisual stimuli View paper
- [20] A Multimodal Seq2Seq Transformer for Predicting Brain Responses to Naturalistic Stimuli View paper
- [21] Discourse context and co-speech gestures jointly shape hierarchical prediction during the processing of a multimodal narrative View paper
- [22] Neuroprobe: Evaluating Intracranial Brain Responses to Naturalistic Stimuli View paper
- [23] Brain activity reflects the predictability of word sequences in listened continuous speech View paper
- [24] Probing Multimodal Fusion in the Brain: The Dominance of Audiovisual Streams in Naturalistic Encoding View paper
- [25] Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech View paper
- [26] An investigation on the effectiveness of multimodal fusion and temporal feature extraction in reactive and spontaneous behavior generative RNN models for listener â□¦ View paper
- [27] Neural responses to naturalistic audiovisual speech are related to listening demand in cochlear implant users View paper
- [28] Combining EEG and 3D-eye-tracking to study the prediction of upcoming speech in naturalistic virtual environments: A proof of principle. View paper
- [29] Speakerâ□□listener neural coupling correlates with semantic and acoustic features of naturalistic speech View paper
- [30] Neural dynamics express syntax in the time domain during natural story listening View paper
- [31] Do Feature Representations from Different Language Models Affect Accuracy of Brain Encoding Models' Predictions? View paper
- [32] 450â□□A Shared Cortical Language Network for Multimodal Naming View paper

- [33] The processing of semantic complexity and co-speech gestures in schizophrenia: a naturalistic, multimodal fMRI study View paper
- [34] Occlusion of lip movements impairs reconstruction of acoustic speech features and higher-level segmentational features in the presence of a distractor speaker View paper
- [35] Electrophysiological Investigations of Attention and Audiovisual Integration During Naturalistic Speech Perception View paper
- [36] The phase of cortical oscillations determines the perceptual fate of visual cues in naturalistic audiovisual speech. View paper
- [37] Neural dSCA: demixing multimodal interaction among brain areas during naturalistic experiments View paper
- [38] Cross-modal activity in adults with cochlear implants: A multimodal brain perspective View paper
- [39] Multimodal imaging of language perception View paper
- [40] Flow-SLM: Joint Learning of Linguistic and Acoustic Information for Spoken Language Modeling View paper
- [41] Text-Infused Audio-Visual Video Parsing with Semantic-Aware Multimodal Contrastive Learning View paper
- [42] Multi-modal multi-channel target speech separation View paper
- [43] Thinking with Sound: Audio Chain-of-Thought Enables Multimodal Reasoning in Large Audio-Language Models View paper
- [44] RobinNet: A Multimodal Speech Emotion Recognition System With Speaker Recognition for Social Interactions View paper
- [45] TD-PLC: A Semantic-Aware Speech Encoding for Improved Packet Loss Concealment View paper
- [46] Speech recognition and intelligent translation under multimodal humanâcomputer interaction system View paper
- [47] Multimodal fusion for multimedia analysis: a survey View paper
- [48] Separating the âChirpâ from the âChatâ: Self-supervised Visual Grounding of Sound and Language View paper
- [49] Spatiotemporal dynamics of activation in motor and language areas suggest a compensatory role of the motor cortex in second language processing View paper
- [50] Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity View paper
- [51] Combinatorics At-a-Glance: On the Spatiotemporal Dynamics of Temporally Unstructured Language View paper
- [52] Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension View paper
- [53] Spatiotemporal dynamics of word processing in the human brain View paper
- [54] Recreating Neural Activity During Speech Production with Language and Speech Model Embeddings View paper
- [55] Spatiotemporal Contributions to Pre-speech Semantic and Syntactic Processing View paper
- [56] Mapping, learning, visualization, classification, and understanding of fMRI data in the NeuCube evolving spatiotemporal data machine of spiking neural networks View paper
- [57] The time-course and spatial distribution of brain activity associated with sentence processing View paper
- [58] Spatio-temporal analysis of electric brain activity during semantic and phonological word processing View paper
- [59] LinBridge: A Learnable Framework for Interpreting Nonlinear Neural Encoding Models View paper
- [60] Simple but Effective Raw-Data Level Multimodal Fusion for Composed Image Retrieval View paper
- [61] LinearâNonlinear Feature Reconstruction Network for Emotion Recognition From Brain Functional Connectivity View paper
- [62] Multimodal Brain Growth Patterns: Insights from Canonical Correlation Analysis and Deep Canonical Correlation Analysis with Auto-Encoder View paper
- [63] Predicting 2-year neurodevelopmental outcomes in preterm infants using multimodal structural brain magnetic resonance imaging with local connectivity View paper
- [64] Reconstructing nonlinear dynamical systems from multi-modal time series View paper
- [65] Intrinsic dimension correlation: uncovering nonlinear connections in multimodal representations View paper
- [66] Nonlinear fusion is optimal for a wide class of multisensory tasks View paper
- [67] Neural Mixed Effects for Nonlinear Personalized Predictions View paper