Bridging Acoustics and Semantics: Native Language Experience and Hierarchical Temporal Integration in the Human Brain

Zhao Xu (xuz22@m.fudan.edu.cn)[1], **Chunhe Li (chunheli@fudan.edu.cn)**[3]
Jianfeng Feng (jianfeng64@gmail.com)[1,2] **& Yuwei Jiang (yuwjiang@fudan.edu.cn)**[1,2,*]
[1]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China
[2]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Shanghai, China
[3]Shanghai Center for Mathematical Sciences and School of Mathematical Sciences, Fudan University, Shanghai, China

Abstract

Deciphering the real-time transformation from acoustic signals to semantic meaning remains a challenge in cognitive neuroscience. Leveraging high-temporal-resolution magnetoencephalography data from native Chinese speakers and the Whisper computational framework, we investigated the neural mechanisms underlying speech processing. We demonstrated a robust scaling law, whereby larger models increasingly align with neural activity. Crucially, a model fine-tuned on native language experience outperformed generic multilingual models, underscoring the brain's sensitivity to language-specific statistics. Furthermore, variable-context analyses revealed a hierarchical temporal architecture: cortical integration of low-level acoustic features occurred within a rapid 40 ms window, whereas high-level speech representations required an integration window of approximately 400 ms. These findings provide a quantitative account of how native language experience and hierarchical temporal integration shape the neural encoding of human speech comprehension.

Keywords: speech; auditory; neural encoding; language; MEG

Introduction

Language is the foundation and core of human cognition. It is not merely a tool for communication but a sophisticated cognitive system that enables humans to transform continuous sensory streams into structured symbolic meanings. Understanding how the human brain achieves this remarkable feat—processing speech from the low-level acoustic fluctuations to high-level semantic integration—remains one of the most profound challenges in neuroscience.

For decades, modeling this acoustic-to-semantic transformation relied on linear mapping of handcrafted features. Pioneering studies demonstrated that the auditory cortex faithfully tracks low-level acoustic properties, such as the speech envelope and spectro-temporal features (Desai et al., 2021; Di Liberto et al., 2015; Hamilton et al., 2021). However, these low-level descriptors fail to capture the rich, compositional nature of high-level language comprehension, e.g., long-range contextual dependencies. The recent development of deep neural networks (DNNs), particularly large language models (LLMs), has triggered a paradigm shift. Emerging evidence suggests a remarkable convergence between artificial and biological neural networks. Studies utilizing functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) have shown that models trained on vast text corpora, such as GPT and Llama, have generated internal representations that linearly align with human brain activity (Caucheteux et al., 2023; Goldstein, Ham, et al., 2025), following predictable "scaling law" where larger models yield better neural alignment (Raugel et al., 2025).

Despite these advances, a critical disconnect still remains: the majority of current "brain-model-alignment" research are based on text LLMs (e.g., GPT series) or self-supervised speech representation models (e.g., wav2vec 2.0). These approaches often decouple the acoustics and text modalities from semantic supervision, failing to fully capture the end-to-end transformation from "hearing" to "understanding" that characterizes natural speech perception. Recently, Whisper (Radford et al., 2023), a weakly supervised automatic speech recognition (ASR) model, has emerged as a unified computational framework connecting acoustic, speech, and linguistic embedding spaces (Goldstein, Wang, et al., 2025). Unlike pure text or audio models, Whisper is trained to transcribe massive datasets of audio directly to text, effectively bridging the acoustic and semantic domains in a biologically plausible manner.

However, even within this unified framework, three fundamental computational mechanisms remain elusive. First, the role of native language experience is debated. While some theories propose universal neural computations across languages (Zada et al., 2025), it remains unclear whether the brain's encoding efficiency is sharpened by specific statistical exposure to one's native language. Does a model fine-tuned on native (e.g., Chinese) data align better with the brain than a generic multilingual model? Second, while context is essential for disambiguation, the temporal scale of integration is often treated as a fixed hyperparameter. We lack a precise quantitative description of how the brain's context window expands across the language hierarchy. Third, the spatiotemporal dynamics of representational shifts remain poorly defined, particularly regarding how the brain transitions from tracking acoustic features to encoding language-specific units (e.g., single characters vs. bigrams).

To address these issues, we took advantages of SMN4Lang (Wang et al., 2022), a public dataset collecting MEG data from native speakers listening to 60 naturalistic Chinese discourses, and employed the computational framework based on Whisper. Our study introduced three specific contributions to the field. First, we systematically probed the "scaling law" and native language effect by comparing Whisper models of varying sizes and, crucially, comparing the generic model

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

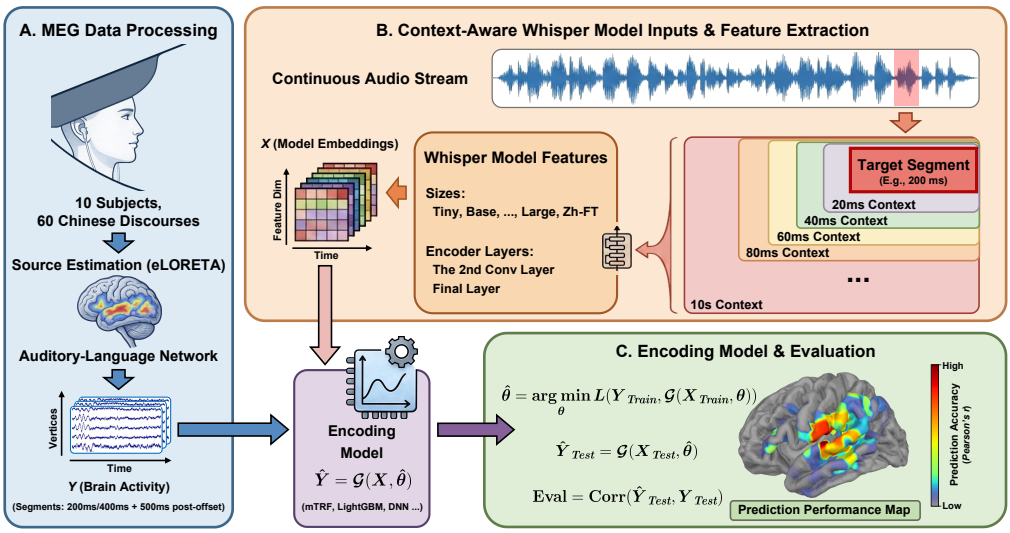

Figure 1: **Experimental pipeline and hierarchical encoding framework.** Schematic overview of the data processing and modeling workflow. (A) MEG signals from 10 subjects listening to 60 Chinese discourses were source-localized to the auditory–language network using eLORETA, yielding time-resolved cortical activity. (B) Continuous speech stimuli were processed by Whisper encoder models of varying parameter scales (Tiny to Large) as well as a Chinese fine-tuned variant (Zh-FT). Hierarchical speech representations were extracted from the second convolutional layer (acoustics-level) and the final encoder layer (speech-level). For each target segment, model features were computed with varying amounts of preceding context, ranging from 0 to 10 s. (C) Encoding models, including linear (mTRF) and nonlinear (LightGBM, DNN) regressors, were trained to predict neural responses at different temporal granularities (200 ms and 400 ms, approximately corresponding to character- and bigram-level units). Prediction performance was evaluated using the Pearson correlation between predicted and observed neural activity, yielding cortical performance maps.

with a Chinese fine-tuned version. This allowed us to isolate the specific contribution of native language statistics to neural alignment. Second, we implemented a systematic context-window-variable analysis, incrementally expanding the contextual length of the audio input from 0 ms to 10,000 ms to quantitatively determine the brain's effective temporal integration window. Third, leveraging the high temporal resolution of MEG, we dissected the fine-grained time course of acoustics versus speech processing, specifically examining how input granularity triggers shifts in neural processing modes.

Our results demonstrated that a model fine-tuned on native language experience significantly outperforms generic models, highlighting the experience-dependent nature of the human auditory-language network. Furthermore, our variable context analysis revealed a distinct temporal hierarchy: the brain integrates low-level acoustic features within a rapid window of approximately 40 ms, whereas high-level speech representations require a longer integration window of approximately 400 ms—corresponding to the timescale of lexical units—to achieve semantic stability.

Methodology

Figure 1 illustrated the complete experimental and analytical pipeline used in this study.

Stimulus Processing and Alignment

We utilized the auditory stimuli and corresponding transcripts from the SMN4Lang dataset (Wang et al., 2022), which comprises MEG recordings from twelve native Chinese speakers as they listened to sixty Chinese discourses. Two participants were excluded from subsequent analyses due to alignment issues between their neural recordings and the auditory stimuli during data preprocessing. To establish precise temporal mappings between the continuous speech stream and linguistic units, we employed the Montreal Forced Aligner (MFA; McAuliffe et al., 2017). Text transcripts were preprocessed by removing punctuation and converting characters to simplified forms, followed by tokenization using a probability-based segmentation tool (Jieba). MFA was then used to generate hierarchical alignments at the character, word, and phoneme levels. We filtered the resulting annotations to retain only entries with sufficient preceding context (minimum 10 s) to ensure valid history for feature extraction. For word-level analyses, we controlled for lexical length by restricting the dataset to disyllabic words.

MEG Preprocessing and Source Localization

Magnetoencephalography data underwent standard preprocessing procedures using the MNE-Python framework (Gramfort et al., 2013). The raw signals were notch-filtered to remove line noise (50 Hz and harmonics), and resampled to

match the frame rate of the Whisper encoder (50 Hz). Noise covariance matrices were estimated from pre-stimulus periods using an automated regularization method.

Anatomical reconstruction was performed on T1-weighted MRI scans using FreeSurfer (Fischl, 2012). We computed a single-layer boundary element model and created a surface-based source space with icosahedral subdivision (ico5). MEG sensors were co-registered with the MRI coordinate system via an iterative closest point (ICP) algorithm. We solved the inverse problem using the exact Low Resolution Brain Electromagnetic Tomography (eLORETA; Pascual-Marqui et al., 2011) method with loose orientation constraints (0.2) and depth weighting (0.8). The resulting source estimates were spatially morphed to the fsaverage template. For subsequent analyses, we extracted source time courses from a predefined region of interest (ROI) encompassing the auditory and language networks (Fedorenko et al., 2024).

Feature Extraction

We extracted auditory and linguistic representations using the OpenAI Whisper automatic speech recognition model. To investigate hierarchical processing, features were derived from two distinct layers of the encoder:

1) Acoustics-level: the second convolutional layer.

2) Speech-level: the final hidden layer.

To systematically vary the historical context available to the model, we cropped the input audio to different lengths with context window lengths ranging from 0 to 10,000 ms, before feeding it into the model. All features were then temporally aligned to the neural data by cropping the resulting continuous representations to the corresponding time windows of the target character or word (200 ms and 400 ms, respectively).

Neural Encoding Models

We formulated the encoding task as a regression problem to predict source-localized neural activity from stimulus features. Here, we implemented and compared three modeling frameworks:

Multivariate Temporal Response Function (mTRF) As a baseline, we trained a mTRF using Ridge regression. This linear model estimates the impulse response that maps stimulus features to neural responses at specific time lags. To mitigate the curse of dimensionality, particularly for high-dimensional Whisper features, we optionally applied principal component analysis (PCA) to the feature space, retaining 95% of the variance prior to regression.

Gradient Boosting Regression To capture non-linear dependencies, we employed LightGBM (Ke et al., 2017), a highly efficient gradient boosting decision tree algorithm. Separate regressors were trained for each source vertex. The models were optimized using the GBDT (Friedman, 2001) boosting type with L2 regularization. We utilized early stopping based on validation set performance to determine the optimal number of boosting iterations.

Deep Neural Encoding Network We developed a Transformer-based DNN, denoted as TF, to capture complex temporal dynamics. The architecture consists of: (1) a linear projection layer to map input features to a latent dimension; (2) a sinusoidal position-encoded Transformer encoder (4 blocks) to process temporal dependencies within the feature sequence; and (3) a 1-dimensional convolutional regression head to map the latent sequence to the multivariate neural signal. The model was trained using adaptive moment estimation with a mean squared error loss. Training incorporated learning rate scheduling and early stopping.

Model Training and Evaluation Data were partitioned into training (70%), validation (10%), and held-out test sets (20%). Model performance was evaluated on the test set using Pearson's correlation coefficient (r) between the predicted and ground-truth neural signals for each vertex.

Iterative Analyses of Context and Latency

We conducted a series of ablation studies to dissect the spatiotemporal dynamics of speech processing:

1) Model Scale and Layer: We compared encoding performance across different Whisper model sizes (from "tiny" to "large") and feature layers to identify the optimal representation of speech. A Chinese fine-tuned version (Zh-FT; BELLEGroup, 2023) of large-scale Whisper model was also included in this analysis.

2) Contextual Window: We trained separate models using features extracted with varying context lengths (ranging from 0 to 10,000 ms) to determine the temporal integration window of the auditory-language network.

3) Temporal Dynamics: We performed a time-lag analysis by shifting the target neural window relative to the stimulus onset (0–500 ms) with a stride of 20 ms, to characterize the latency of neural responses to acoustic and linguistic features.

Results

Scaling Law and Non-linear Encoding Efficiency

We first evaluated the impact of model scales and encoding algorithms on brain alignment (Figure 2). Our results demonstrated a clear scaling law: as the Whisper model size increases from "tiny" to "large", the encoding performance improves consistently.

First, we sought to determine the optimal mapping function between high-dimensional computational features and biological neural signals. As shown in Figure 1, we compared the encoding performance of linear models (mTRF, mTRF-PCA) against non-linear models (TF, LightGBM). We found that the tree-based gradient boosting model outperformed all other methods across all model sizes. Notably, LightGBM achieved a significantly higher encoding performance compared to TF (paired t-test, $p<0.0001$), suggesting that for limited neuroimaging data samples, tree-based models offer superior robustness and alignment with the non-linear manifold

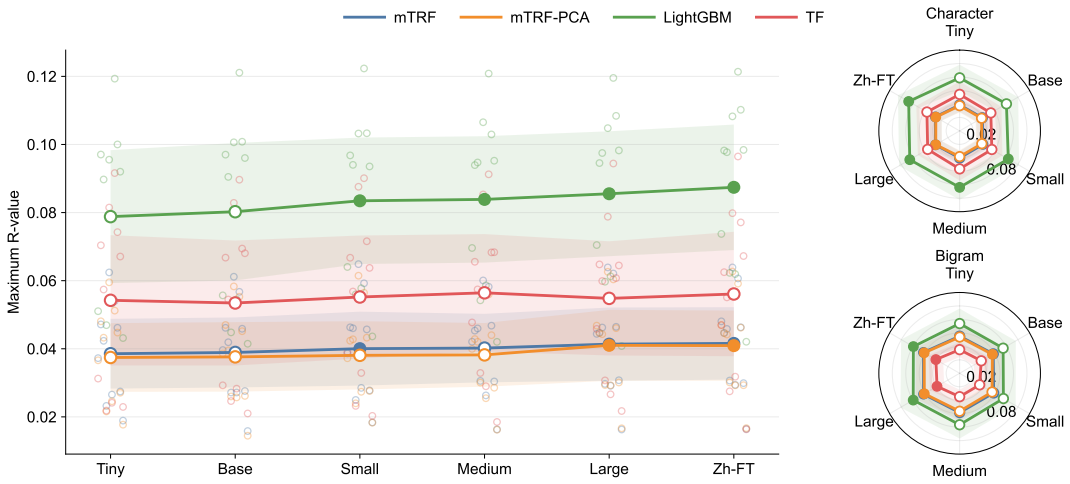

Figure 2: **Predictive performances across model scales and regressors.** Comparison of maximum Pearson correlation coefficients (r-values) obtained with different Whisper model scales and encoding methods. (Left) Line plot shows encoding performances for character-level segments (200 ms). (Right) Radar plots illustrate performance patterns across character- and bigram-level granularities for different Whisper scales. Solid markers indicate significantly higher performance compared to the corresponding "Tiny" Whisper baseline within the same regressor ($p < 0.05$). Shaded areas denote 95% confidence intervals.

of neural representations. Consequently, LightGBM was employed for all subsequent analyses.

Using LightGBM, we examined the "scaling law" by evaluating a suite of Whisper models of increasing parameter size (tiny, base, small, medium, large). We observed a monotonic increase in prediction accuracy as model parameter size increased. Specifically, the small-, medium-, and large-scale models demonstrated significantly superior encoding performance relative to the tiny-scale baseline, as indicated by the solid markers in Figure 2 ($p < 0.05$). This trajectory indicates that as computational models scale up to capture more complex statistical regularities of speech, their internal representations converge towards the biological representations found in the human auditory-language network.

A critical question in cognitive neuroscience is whether the brain's encoding of speech is driven by universal acoustic features or language-specific statistical priors. To test this, we compared the standard large-scale model (trained on multilingual data) with Zh-FT (a model of identical architecture but fine-tuned on Chinese data). Strikingly, despite having the same number of parameters, Zh-FT significantly outperformed the large-scale model in predicting neural activity. This advantage was robust across both single character and bigram (i.e., disyllabic words) granularities (Figure 2, radar plots; paired-t test, $p<0.05$). The specific enhancement of Zh-FT suggests that the human auditory-language network is not merely a general-purpose sound processor but is finely tuned to the statistical properties of the native language. The brain-model alignment is maximized when the computational model shares the same "linguistic experience" (i.e., exposure to the native language distribution) as the human subjects.

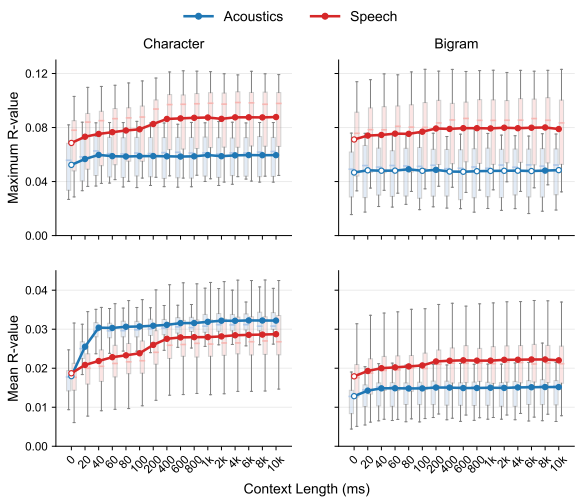

Figure 3: **Temporal context dependency of neural representation alignment.** Solid markers indicate context lengths at which performance differs significantly from the 0 ms preceding context condition ($p < 0.05$). Error bars denote variability across subjects.

Hierarchical Dissociation of Temporal Integration Windows

Having established Zh-FT as the optimal model, we investigated how different layers of the model—representing different levels of processing—interact with temporal context. We extracted representations from the early encoder layer (acoustics-level) and the final encoder layer (speech-level) under context windows ranging from 0 ms to 10,000 ms. The

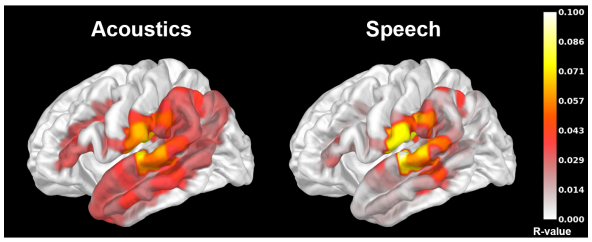

Figure 4: **Anatomical distribution of acoustics and speech encoding performance.** The spatial distribution of encoding performance (single character granularity) was mapped onto the cortical surface using the Zh-FT model with a fixed 1,000 ms context and a 300 ms time lag.

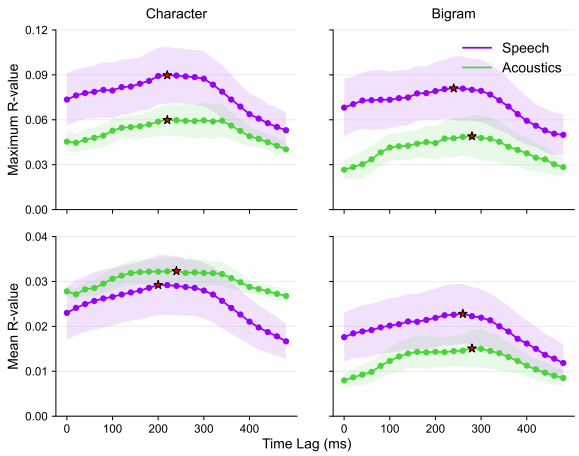

Figure 5: **Chronometry of encoding performances across time lags.** Time-lagged encoding between Whisper embeddings and MEG signals (20 ms stride). Red stars denote the peak r-value for each curve. Shaded areas denote 95% confidence intervals.

results (Figure 3) revealed a profound dissociation in temporal processing strategies between these two distinct features.

For acoustics-level representations (blue lines), a minor and quick improvement was observed within the first 40 ms, but extending context beyond this window yielded no significant gain. This suggests that brain regions tracking low-level acoustic properties (e.g., pitch and envelope) might operate on a localized temporal scale, processing information in near-instantaneous fragments. By contrast, the speech-level representations (red lines) showed a relatively slow and sustained increase in prediction accuracy as context length increased. Performance improved gradually from 0 ms to approximately 400 ms, and reached a plateau.

In addition, a "crossover" effect was observed in the mean r-value analysis: at short context lengths in the single character condition, acoustics-level features outperformed speech-level features; however, as the linguistic unit transitioned to bigrams, speech-level features demonstrated superior average predictive power. In all instances, the maximum r-value for speech-level features remained higher than that of acoustics-level features across the entire range of context lengths.

Anatomical Distribution of Acoustic and Speech Representations

To ground these computational findings in neuroanatomy, we projected the prediction performance onto the cortical surface using the Zh-FT model with a fixed 1,000 ms context and a 300 ms time lag (Figure 4; single character granularity). Both acoustics- and speech-level encoding models achieved peak performances around the primary auditory cortex, covering the Heschl's gyrus (HG) and superior temporal gyrus (STG).

However, a critical spatial dissociation explained the performance differences observed in Figure 3, where representations at the acoustics level showed higher mean r-values, whereas those at the speech level exhibited higher maximum r-values. Acoustic features (Figure 4, left) engaged a widespread cortical territory, suggesting a broad-tuning mechanism whereby a large population of neurons collectively track fundamental auditory attributes, resulting in a higher average encoding performance. Speech features (Figure 4, right), conversely, exhibited a highly focal distribution, with peak prediction accuracy concentrated in specific subregions of the STG. This pattern suggests a narrow-tuning mechanism for high-level semantics: specialized neural populations function as computational hubs for linguistic integration. These hubs aligned extremely well with the model's deep representations, yielding high maximum r-value, despite the surrounding general auditory-language cortex exhibiting lower correspondence.

Temporal Lag Dependence of Neural Encoding

Finally, we utilized a time-lag analysis spanning 0 to 500 ms to resolve the temporal dynamics underlying neural processing (Figure 5). Lag profiles were calculated using increments of 20 ms, corresponding to the 50 Hz sampling frequency of the Whisper embeddings.

The encoding performance peaked at a latency of approximately 200-300 ms following stimulus onset for both character and bigram segments, aligning with the established temporal window of auditory-linguistic processing. Specifically, for character-level segments, the highest r-values for both speech and acoustic features were observed at a lag of around 220 ms. In contrast, for the bigram-level segments, the peak latencies demonstrated a temporal shift: the speech-level feature attained its maximal predictive accuracy at approximately 240 ms, while the acoustics-level feature exhibited a later peak at about 280 ms.

Discussion

This study employed the large-scale ASR model Whisper alongside MEG data to systematically investigate the dynamic progression of acoustic and semantic representations during Chinese speech processing.

Our results show that native language experience exerts a

significant influence even within a unified acoustic-to-speech modeling framework. The critical point is not merely that language-specific tuning enhances model performance from an engineering perspective, but that a Chinese-tuned model more accurately predicts cortical activity in native Mandarin speakers compared to an architecture-matched multilingual model. Interpreted cautiously, this result suggests that exposure to native language statistics refines the representational structure most pertinent to neural speech encoding. Thus, the present work operationalizes native language experience in a computationally explicit manner, rather than treating it solely as a behavioral or developmental background factor.

This point further elucidates the theoretical contribution of the present study. Our results demonstrate that sensitivity to native language statistical patterns emerges within a unified model linking acoustic input to higher-level speech representations, thereby extending prior work showing that listeners are attuned to native-language statistics. Hence, the current results not only align with experience-dependent accounts of speech perception, but also show how such experience can be instantiated within a brain-aligned computational framework.

Hierarchical Transitions from Acoustic Units to Semantic Constructs

The current results support a hierarchical transition from the analysis of local acoustic properties to more abstract and context-dependent speech representations. A key finding of this study is the representational inversion observed in the neural processing of Chinese speech. At the level of individual characters, neural activity within the auditory cortex is primarily driven by the underlying acoustic representations. However, when processing moves to the disyllabic word level, the explanatory power of higher-level speech representations surpasses that of acoustic features. This observed transition suggests that the auditory-language network does not rely on a singular representational regime during natural speech comprehension. Rather, cortical encoding appears to shift gradually from processing relatively localized acoustic structures toward the integration of higher-level representations that encompass larger linguistic units.

Prior research has proposed a "two-phase abstraction process" theory, suggesting a shared mechanism between LLMs and the human brain (Cheng & Antonello, 2024). This theory posits that language processing involves a transition from component combination to conceptual abstraction. Our findings further specified the temporal dynamics of this transition: in Chinese, a speech segment of approximately 400ms (the duration of a bigram) is sufficient to elicit semantic closure. Moreover, by comparing single-character and bigram conditions, we revealed that STG functions not merely as a passive feature mapping region, but as a hub with dynamic computational flexibility. Consistent with recent studies highlighting the STG's enhanced encoding of lexical boundaries in the native language (Bhaya-Grossman et al., 2025), the representational inversion observed here provided direct evidence that lexicalization prompts a re-weighting of acoustic and semantic computational contributions within the brain.

More broadly, our findings also resonate with previous work on Chinese speech processing. Prior fMRI studies suggest that spoken Chinese sentence comprehension engages a distributed temporal network (Liu & Chen, 2020), and that sentence context can facilitate lexical access even when tonal cues are degraded (Xu et al., 2013). However, these studies have largely addressed specific components of speech processing in relative isolation—such as regional involvement during sentence comprehension or the compensatory role of contextual information—rather than characterizing how the brain moves from lower-level acoustic analysis to higher-level speech representations within a single framework. The present study extends earlier work by quantifying this transition within a unified computational framework that spans acoustics- and speech-level representations, rather than examining acoustic, phonological, or lexical processes in isolation.

Temporal Characteristics of Context-Dependent Gain and Time Delays

This study quantitatively examined the temporal receptive window involved in Chinese speech processing, revealing that contextual facilitation reaches saturation at approximately 40 ms for single characters and 400 ms for bigrams. This striking dissociation in effective context length suggests a fundamental difference in the processing timescales of short-timescale acoustic information and lexically meaningful units in Chinese, providing quantitative support for a hierarchical temporal architecture of speech comprehension.

Notably, in the single-character condition, although acoustic features predominantly influence processing, an early and modest peak in speech-related features was observed at a latency of 200 ms (Figure 5, bottom left). This finding may reflect the hierarchical structure of predictive coding (Caucheteux et al., 2023): even prior to the full integration of acoustic input, the brain leverages pre-existing linguistic priors, modeled here by high-level representations derived from Whisper, to generate preliminary semantic predictions. In the bigram condition, this early predictive signal converged with the subsequent semantic integration signal, resulting in an enhanced performance peak. These results indicated that Chinese speech processing operates as a continuous, multilayered, and real-time alignment mechanism.

Acknowledgments

We would like to thank the anonymous reviewers for commenting on the initial submission of this manuscript. This work was supported by the grant from the National Natural Science Foundation of China (No. 32471091 to Y.J.).

Code Availability

The analysis code corresponding to the present study is available at https://github.com/Solaris12138/YWJLab-zh-whisper-brain-alignment.

References

BELLEGroup. (2023). Belle: Be everyone's large language model engine. https://github.com/LianjiaTech/BELLE

Bhaya-Grossman, I., Leonard, M. K., Zhang, Y., Gwilliams, L., Johnson, K., Lu, J., & Chang, E. F. (2025). Shared and language-specific phonological processing in the human temporal lobe. *Nature*, 1–12. https://doi.org/10.1038/s41586-025-09748-8

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, *7*(3), 430–441. https://doi.org/10.1038/s41562-022-01516-2

Cheng, E., & Antonello, R. J. (2024). Evidence from fmri supports a two-phase abstraction process in language models. *arXiv preprint arXiv:2409.05771*.

Desai, M., Holder, J., Villarreal, C., Clark, N., Hoang, B., & Hamilton, L. S. (2021). Generalizable eeg encoding models with naturalistic audiovisual stimuli. *Journal of Neuroscience*, *41*(43), 8946–8962. https://doi.org/10.1523/JNEUROSCI.2891-20.2021

Di Liberto, G. M., O'sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*(5), 289–312. https://doi.org/10.1038/s41583-024-00802-4

Fischl, B. (2012). Freesurfer. *Neuroimage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Goldstein, A., Ham, E., Schain, M., Nastase, S. A., Aubrey, B., Zada, Z., Grinstein-Dabush, A., Gazula, H., Feder, A., Doyle, W., et al. (2025). Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature communications*, *16*(1), 10529. https://doi.org/10.1038/s41467-025-65518-0

Goldstein, A., Wang, H., Niekerken, L., Schain, M., Zada, Z., Aubrey, B., Sheffer, T., Nastase, S. A., Gazula, H., Singh, A., et al. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature human behaviour*, 1–15. https://doi.org/10.1038/s41562-025-02105-9

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. (2013). Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, *7*, 267.

Hamilton, L. S., Oganian, Y., Hall, J., & Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell*, *184*(18), 4626–4639. https://doi.org/10.1016/j.cell.2021.07.019

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Liu, H., & Chen, S. A. (2020). The neural correlates of spoken sentence comprehension in the chinese language: An fmri study. *Psychology Research and Behavior Management*, 641–652.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Interspeech*, *2017*, 498–502.

Pascual-Marqui, R. D., Lehmann, D., Koukkou, M., Kochi, K., Anderer, P., Saletu, B., Tanaka, H., Hirata, K., John, E. R., Prichep, L., et al. (2011). Assessing interactions in the brain with exact low-resolution electromagnetic tomography. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1952), 3768–3784. https://doi.org/10.1098/rsta.2011.0081

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International conference on machine learning*, 28492–28518.

Raugel, J., d'Ascoli, S., Rapin, J., Wyart, V., & King, J.-R. (2025). Scaling and context steer llms along the same computational path as the human brain. *arXiv preprint arXiv:2512.01591*.

Wang, S., Zhang, X., Zhang, J., & Zong, C. (2022). A synchronized multimodal neuroimaging dataset for studying brain language processing. *Scientific Data*, *9*(1), 590. https://doi.org/10.1038/s41597-022-01708-5

Xu, G., Zhang, L., Shu, H., Wang, X., & Li, P. (2013). Access to lexical meaning in pitch-flattened chinese sentences: An fmri study. *Neuropsychologia*, *51*(3), 550–556.

Yu, K., Wang, R., Li, L., & Li, P. (2014). Processing of acoustic and phonological information of lexical tones in mandarin chinese revealed by mismatch negativity. *Frontiers in Human Neuroscience*, *8*, 729.

Zada, Z., Nastase, S. A., Li, J., & Hasson, U. (2025). Brains and language models converge on a shared conceptual space across different languages. *arXiv preprint arXiv:2506.20489*.