
Listening to the Brain: Multi-Band sEEG Auditory Reconstruction via Dynamic Spatio-Temporal Hypergraphs

This supplement to our main paper, "Listening to the Brain: Multi-Band sEEG Auditory Reconstruction via Dynamic Spatio-Temporal Hypergraphs," provides an in-depth explanation of the dataset collection methods and includes a comprehensive data card. It also outlines the licensing information for the dataset and includes an author statement verifying compliance with these licensing terms. Furthermore, it addresses the societal implications, providing a Preliminary Assessment and Disposal Plan of Relevant Risks as well as discussing Ethical Issues and Countermeasures. Detailed descriptions of the methods implemented on the dataset, along with the datasheets, are also included.

Appendix A: Data Collection

In our study, subjects were exposed to auditory stimuli from three different classifications: 44 categories of Chinese Mandarin words, 10 categories of Chinese Mandarin digits, and 20 categories of English words in each round. The duration for listening was set to 2 seconds for each word. At the start of each round, each participant was given a 5-second interval to be ready, where a prompt "Please listen to the speech attentively" is played, which is followed by a "ding" sound to represent the start of the attended speech content.

A.1 Preliminary Assessment and Disposal Plan of Relevant Risks

To ensure the scientific property of the trial and the safety of the participants, we conducted a comprehensive assessment of the trial participants. Eligible trial participants were required to sign an informed consent form to understand the purpose, process, possible adverse reactions of the trial in detail, and clarify the relevant safety measures.

During the experiment, doctors and research teams worked together to ensure the safety and comfort of patients. If the patient felt tired during the trial, we would suspend the trial at any time to provide rest. In addition, we closely monitored any potential risks during the trial and be ready to respond to emergencies at any time to maximize the safety and legal rights of the subjects.

A.2 Ethical Issues and Countermeasures

- (1) Individuals participated in the study on a voluntary basis, and after ensuring that the subjects understand the relevant information, written informed consent were obtained from the subjects.
- (2) All measures have been taken to protect the privacy of the subjects and keep personal information confidential.
- (3) Each subject received sufficient information, including the purpose and methods of the study, any possible conflicts of interest, the researcher's organizational affiliation and potential risks, any discomfort that the study may cause, and any other information related to the study.
- (4) Each subject was informed of his or her right to refuse to participate in the study and the right to withdraw consent to withdraw from the study at any time

Appendix B: Dataset Structure

As illustrated in Figure. 1, we use a unified format to name the files of the seeg data, namely `index_wordName_LanguageID`, where word name represents the content of the words listened by the participant. For auditory data, we use similar format to name the files, namely `wordName_LanguageID`. For ease of use, we provide the preprocessed sEEG signal and mel-spectrogram, both stored in npy format. It contains the following data:

(1) sEEG: a data matrix representing sEEG signals, ending with `SEEG.npy`, in the shape of $T * F$, where T represents the time dimension and F is the number of features. For HGA and LFS, the number of features is the same as the number of sEEG channels, and for BBS, the number of features is twice the number of channels.

(2) Mel-Spectrogram: a data matrix representing the mel-spectrogram of audio signals, ending with `MEL.npy`, in the shape $T*80$, where T represents the time dimension and 80 represents the number of bin of the mel-spectrogram.

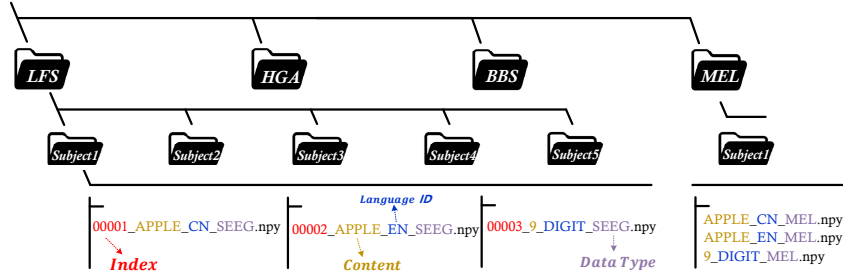


Figure 1: Dataset structure showing the organization of sEEG and auditory data, in npy format.

Appendix C: Societal Impact

As we point out in the paper, we publish a sEEG-audio dataset that is specifically designed for the study of auditory reconstruction from brain signals. The broad applicability of this dataset is crucial for explaining and predicting the neural mechanisms of human language. We not only confirm the quality and completeness of this dataset, but also verify the feasibility of sEEG-based auditory reconstruction. This technology provides new research paths at the intersection of neuroscience and artificial intelligence

Appendix D: Access to Dataset

The NeuroListen dataset, which is available on Zenodo at <https://zenodo.org/records/17426506> as a general-purpose open repository, is collected, updated, and maintained by our team members. The code for dataset creation and experiments can be accessed at <https://github.com/NeuroListen/NeuroListen>.

Appendix E: Licence

We publish all data under CC-BY-4.0 licence. We include detailed instructions on how to obtain our data and provide preprocessing scripts in our GitHub repository. This dataset is intended for research purposes only and not for clinical usage.

Appendix F: Implementation Details

F.1 Experimental Parameter

In our experiments, to ensure uniformity and fairness across all experimental setups, we applied identical hyperparameter configurations for all comparison tests. Each model was trained over 300 epochs to guarantee convergence in every experiment. Specifically, we set the batch size to 16 and chose an initial learning rate of 0.0625. Utilizing the Adam optimizer with betas parameters of 0.9

and 0.98 allowed us to regulate the exponential moving average of both the gradient and its squared form, aiming to achieve a balance between training stability and speed. Additionally, we implemented a gradient clipping threshold of 1.0 to effectively mitigate the risk of gradient explosion. Additionally, we implemented a warm-up strategy to stabilize the training process.

F.2 Evaluation Metrics

F.2.1 Pearson Correlation Coefficient

PCC (Pearson Correlation Coefficient) is a statistical indicator used to measure the strength and direction of the linear relationship between two variables. PCC is the most commonly used metric in the field of sEEG-based speech decoding(1; 2; 3; 4). The value range of this indicator is between -1 and 1, where:

- If PCC is equal to 1, it means that the two variables are completely positively correlated, that is, when one variable increases, the other variable also increases, and the relationship between the two is linear.
- If PCC is equal to -1, it means that the two variables are completely negatively correlated, that is, when one variable increases, the other variable decreases, which is also a linear relationship.
- If PCC is equal to 0, it means that there is no linear relationship between the two variables.

F.2.2 Mel Cepstral Distortion

MCD (Mel Cepstral Distortion) is a measure of the difference between two speech signals in the mel-cepstral domain, which is more perceptually relevant than raw spectral comparison. The MCD between two mel-cepstral vectors \mathbf{c} and \mathbf{c}' of dimension D is calculated as:

$$\text{MCD} = 10 \cdot \frac{\sqrt{2}}{\ln 10} \cdot \sqrt{\sum_{i=1}^D (c_i - c'_i)^2} \quad (1)$$

F.2.3 Root Mean Square Error

RMSE (Root Mean Square Error) measures the difference between the reconstructed auditory signal \mathbf{y} and the original signal $\hat{\mathbf{y}}$, which is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (2)$$

where N is the number of samples in the signals. A lower RMSE indicates higher accuracy in decoding, meaning the reconstructed speech is closer to the original.

F.2.4 Short-Time Objective Intelligibility

STOI (Short-Time Objective Intelligibility) is a metric designed to predict the intelligibility of speech signals degraded by noise, reverberation, or coding. It assesses the preservation of short-term spectral features between the processed signal \mathbf{y} and the clean signal $\hat{\mathbf{y}}$:

$$\text{STOI} = \frac{1}{K} \sum_{k=1}^K \frac{\mathbf{s}_k^T \hat{\mathbf{s}}_k}{\|\mathbf{s}_k\| \|\hat{\mathbf{s}}_k\|} \quad (3)$$

where \mathbf{s}_k and $\hat{\mathbf{s}}_k$ are short-time spectral vectors of the processed and clean signals, respectively, and K is the number of time-frequency segments. STOI values range from 0 to 1, with higher values indicating better predicted intelligibility (5).

F2.5 SMOS

System Interface One reference audio and one audio to be evaluated (containing target text).

Questionnaire Compare the audio to be evaluated with the reference audio, and rate the similarity of the evaluated audio in terms of accuracy.

Scoring Criteria 5 (Excellent. The speech is almost completely identical to the reference audio), 4 (Good. The speech is highly consistent with the reference audio), 3 (Fair. The speech is somewhat similar to the reference audio but has minor differences), 2 (Poor. The speech has obvious differences from the reference audio), 1 (Bad. The speech is completely different from the reference audio).

F2.6 CMOS

System Interface One audio to be evaluated (containing target text).

Questionnaire Evaluate the clarity of the audio to be evaluated, focusing on whether the pronunciation is clear, the sound quality is intelligible, and whether the vocabulary is easy to recognize.

Scoring Criteria 5 (Very clear. The speech is extremely clear, with all vocabulary easily recognizable and no ambiguity), 4 (Slightly clear. Mostly clear, with most vocabulary easily recognizable and almost no difficulty in identification), 3 (Moderate clarity. Generally clear speech, but with a few minor ambiguities or difficult - to - recognize words), 2 (Unclear. Speech is blurry, with some vocabulary difficult to recognize and requiring concentrated attention to identify), 1 (Completely unclear. Speech is unintelligible, with almost no recognizable vocabulary and poor sound quality).

Appendix G: Authorstatement

As the authors, we solemnly assure that we accept full responsibility for any possible infringements regarding the data compilation or related proceedings, and commit to promptly taking necessary steps - such as data removal - when dealing with such issues.

Appendix H: Hypergraph Neural Network Theory

H.1 Hypergraph Structure and Spectral Representation

A hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ extends the concept of a conventional graph by allowing hyperedges to connect an arbitrary number of vertices. Here, $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ denotes the set of vertices, $\mathcal{E} = \{e_1, e_2, \dots, e_M\}$ denotes the set of hyperedges, and $W \in \mathbb{R}^{M \times M}$ is a diagonal matrix representing hyperedge weights. The hypergraph structure is encoded by the incidence matrix $H \in \{0, 1\}^{N \times M}$, where $H(i, m) = 1$ indicates that vertex v_i is connected to e_m , and $H(i, m) = 0$ otherwise.

The degree of a vertex $v_i \in \mathcal{V}$ and the degree of a hyperedge $e_m \in \mathcal{E}$ are defined respectively as:

$$d(v_i) = \sum_{m=1}^M w_m H(i, m), \quad \delta(e_m) = \sum_{i=1}^N H(i, m), \quad (4)$$

and their corresponding diagonal matrices are denoted $D_v \in \mathbb{R}^{N \times N}$ and $D_e \in \mathbb{R}^{M \times M}$. Based on this structure, the normalized hypergraph Laplacian is formulated as:

$$\mathcal{L} = I - D_v^{-1/2} H W D_e^{-1} H^T D_v^{-1/2}, \quad (5)$$

which generalizes the classical graph Laplacian to high-order relations. This operator governs signal propagation on hypergraphs by enforcing spectral smoothness over vertex features and has been widely adopted in spectral hypergraph learning frameworks. It enables the aggregation of information across shared hyperedges, thereby capturing semantic structures beyond local pairwise interactions.

H.2 Theoretical Comparison with Graph Neural Networks

While both hypergraph neural networks (HGNNs) and graph neural networks (GNNs) aim to learn vertex-level representations from relational data, they diverge in their structural modeling assumptions

and theoretical expressivity. In GNNs, the underlying graph structure is defined by pairwise edges. A prototypical example is the graph convolutional network (GCN), whose update rule is:

$$X^{(l+1)} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X^{(l)} \Theta^{(l)} \right), \quad (6)$$

where $\tilde{A} = A + I$ is the adjacency matrix with self-loops, \tilde{D} is the corresponding degree matrix, $\Theta^{(l)}$ is a learnable transformation matrix, and $\sigma(\cdot)$ is a non-linear activation function. This formulation restricts message passing to immediate neighbors and only captures first-order topological dependencies.

HGNNs generalize this model by enabling information propagation across arbitrary-sized vertex groups connected via hyperedges. The corresponding convolution operation is:

$$X^{(l+1)} = \sigma \left(D_v^{-\frac{1}{2}} H W D_e^{-1} H^T D_v^{-\frac{1}{2}} X^{(l)} \Theta^{(l)} \right), \quad (7)$$

where $H \in \{0, 1\}^{N \times M}$ is the vertex-hyperedge incidence matrix, $W \in \mathbb{R}^{M \times M}$ is a diagonal hyperedge weight matrix (set to the identity in the unweighted case), and D_v, D_e are the vertex and hyperedge degree matrices. This operation can be interpreted as a two-step propagation process: first aggregating vertex features within each hyperedge, then redistributing the aggregated features back to the vertex domain.

Alternatively, the same operation can be expressed in spectral form using the normalized hypergraph Laplacian \mathcal{L} :

$$X^{(l+1)} = \sigma \left((I - \mathcal{L}) X^{(l)} \Theta^{(l)} \right). \quad (8)$$

This formulation emphasizes smoothness over hypergraph structure and facilitates spectral filtering of signals across higher-order relations.

Under specific structural conditions, HGNN reduces to GCN. Specifically, if the hypergraph is 2-uniform (i.e., each hyperedge connects exactly two vertices), the weights are uniform ($W = I$), and the hyperedge degrees satisfy $D_e = 2I$, the HGNN update simplifies to:

$$X^{(l+1)} = \sigma \left(\frac{1}{2} (I + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) X^{(l)} \Theta^{(l)} \right), \quad (9)$$

which is mathematically equivalent to a GCN layer. This illustrates that GCN is a special case of HGNN and validates the latter's compatibility with conventional graph structures.

Beyond this equivalence, HGNNs offer a substantially richer modeling capacity. One important distinction lies in the granularity of message passing. In GCNs, information aggregation is limited to one-hop neighbors connected by edges. In contrast, HGNNs enable vertices to simultaneously aggregate features from entire hyperedges, representing semantically coherent vertex groups. This group-wise aggregation mechanism provides a structural prior that naturally supports modeling of co-activation or co-functionality patterns among multiple vertices.

Furthermore, even when hyperedge weights are uniform and the incidence structure is binary, the induced propagation dynamics are generally asymmetric. Specifically, the influence that vertex v_i receives from another vertex v_j through shared hyperedges depends not only on direct co-membership but also on the size of each hyperedge $\delta(e)$. As a result, each vertex receives a normalized group-level message, where contributions are diluted over the number of participating vertices. This leads to a form of implicit attention on local structure size, which cannot be captured by fixed pairwise adjacency matrices.

Appendix I: Inference Procedure of HyperSpeech

The inference pipeline of HyperSpeech is presented in Algorithm 1, detailing the sequential computation from multi-band sEEG input to waveform reconstruction. Each stage corresponds to a core

functional block in the proposed framework and is aligned with the equations defined in our main paper.

Algorithm 1 HyperSpeech Inference Pipeline

Require: Raw sEEG signals $X^{(f)} \in \mathbb{R}^{C \times d}$ for frequency band $f \in \{\text{HGA}, \text{LFS}\}$, number of electrode shafts N , number of time windows T , number of neighbors K , number of convolution layers L

- 1: **for** $f \in \{\text{HGA}, \text{LFS}\}$ **do**
- 2: $X_1^{(f)} \leftarrow$ intra-shaft channel fusion via Main Eq. (1)
- 3: $X_2^{(f)} \leftarrow$ spatio-temporal feature extraction via Main Eq. (2)
- 4: $X_3^{(f)} \leftarrow$ windowed representation
- 5: **for** $t = 1$ to T **do**
- 6: $\mathcal{G}_{spatial}^t \leftarrow$ spatial hypergraph construction via Main Eq. (3)
- 7: $X_{spatial}^{t(l)} \leftarrow$ spatial hypergraph convolution via Main Eq. (4)
- 8: **end for**
- 9: $\mathcal{G}_{temporal} \leftarrow$ temporal hypergraph construction
- 10: $X_5^{(f)} \leftarrow$ temporal hypergraph convolution via Main Eq. (5)
- 11: **end for**
- 12: $X_6 \leftarrow$ multi-band feature fusion via Main Eq. (6)
- 13: $X_7 \leftarrow$ Bi-LSTM modeling via Main Eq. (7)
- 14: $X_{\text{mel}} \leftarrow$ mel-spectrogram projection
- 15: $X_{\text{wave}} \leftarrow$ waveform generation (HiFi-GAN)

Ensure: Reconstructed speech waveform X_{wave}

References

- [1] M. Verwoert, M. C. Ottenhoff, S. Goulis, A. J. Colon, L. Wagner, S. Tousseyn, J. P. Van Dijk, P. L. Kubben, and C. Herff, “Dataset of speech production in intracranial electroencephalography,” *Scientific data*, vol. 9, no. 1, p. 434, 2022.
- [2] S. Duraivel, S. Rahimpour, C.-H. Chiang, M. Trumpis, C. Wang, K. Barth, S. C. Harward, S. P. Lad, A. H. Friedman, D. G. Southwell *et al.*, “High-resolution neural recordings improve the accuracy of speech decoding,” *Nature communications*, vol. 14, no. 1, p. 6938, 2023.
- [3] M. Angrick, M. C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski *et al.*, “Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity,” *Communications biology*, vol. 4, no. 1, p. 1055, 2021.
- [4] X. Chen, R. Wang, A. Khalilian-Gourtani, L. Yu, P. Dugan, D. Friedman, W. Doyle, O. Devinsky, Y. Wang, and A. Flinker, “A neural speech decoding framework leveraging deep learning and speech synthesis,” *Nature Machine Intelligence*, pp. 1–14, 2024.
- [5] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.