scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chisco: An EEG-based BCI dataset for decoding of imagined speech

Zihan Zhang[1], Xiao Ding[1]✉, Yu Bao[1], Yi Zhao[1], Xia Liang[2]✉, Bing Qin[1] & Ting Liu[1]

The rapid advancement of deep learning has enabled Brain-Computer Interfaces (BCIs) technology, particularly neural decoding techniques, to achieve higher accuracy and deeper levels of interpretation. Interest in decoding imagined speech has significantly increased because its concept akin to "mind reading". However, previous studies on decoding neural language have predominantly focused on brain activity patterns during human reading. The absence of imagined speech electroencephalography (EEG) datasets has constrained further research in this field. We present the *Chinese Imagined Speech Corpus* (Chisco), including over 20,000 sentences of high-density EEG recordings of imagined speech from healthy adults. Each subject's EEG data exceeds 900 minutes, representing the largest dataset per individual currently available for decoding neural language to date. Furthermore, the experimental stimuli include over 6,000 everyday phrases across 39 semantic categories, covering nearly all aspects of daily language. We believe that Chisco represents a valuable resource for the fields of BCIs, facilitating the development of more user-friendly BCIs.

Background & Summary

BCIs are a promising technology for improving the quality of life of people who have lost the capability to either communicate or interact with their environment[1]. A BCI provides an alternative way of interaction to such individuals, by decoding the neural activity and transforming it into control commands for triggering wheelchairs, prosthesis, spellers or any other virtual interface device[2].

In recent years, advancements in deep learning technology have made naturalistic paradigm BCIs[3] increasingly feasible. In this paradigm, participants are exposed to real-world-like stimuli, such as movies, novels, and images, and the content related to these stimuli is decoded from their neural responses. In contrast, classical paradigm BCIs, which utilize P300, Visual Evoked Potentials, Slow Cortical Potentials, and Motor Imagery, rely on meticulously designed and discrete stimuli presented to participants. Compared to classical paradigms, naturalistic paradigm BCIs are more user-friendly as they require minimal user training and offer significantly improved decoding speeds. Neural language decoding is a key area within this approach[4].

In the field of neural language decoding, considerable research has been conducted on the *decoding of voiced speech* due to its high potential for medical applications. Current studies predominantly focuses on implanting electrodes into the motor cortex of the brain of patients with Amyotrophic Lateral Sclerosis (ALS) through invasive surgeries to aid in restoring their speech motor mechanisms[5–7]. However, this approach is not suitable for healthy individuals, as in addition to the ethical concerns associated with invasive surgeries, speech recognition technology offers a more accurate and convenient alternative. Additionally, substantial research has been carried out on *decoding reading stimuli or auditory inputs*, such as ZuCo[8], ChineseEEG[9], BCCWJ-EEG[10] and Hsg[11], which represents initial attempts to decode neural language content using non-invasive methods. These experimental paradigms are elegant but as datasets for BCIs, they capture the brain's electrophysiological responses to stimuli. Consequently, these are only suitable as research materials for stimulus-based BCI studies, where users must receive external stimuli. This requirement limits the applicability of BCIs in real-life and long-term applications. Furthermore, BCIs for *decoding inner speech* are also not universally applicable, as not everyone experiences inner speech[12].

Therefore, the *decoding of imagined speech* is crucial for developing BCIs suitable for everyday use by healthy individuals. This involves collecting imagined speech directly from the brain using non-invasive devices and translating it into natural language. Despite its significance, prior research in this area has been hampered by a scarcity of datasets. The complexity of experimental paradigms for imagined speech, the difficulty in data

[1]Harbin Institute of Technology, Department of Computer Science, Harbin, 150000, China. [2]Harbin Institute of Technology, The School of Space Environment and Material Science, Harbin, 150000, China. ✉e-mail: xding@ir.hit.edu.cn; xia.liang@hit.edu.cn

Subject ID	Age	Gender	ANT(alerting)	ANT(orienting)	ANT(executive control)	ANT(mean)	HSK(Reading Section)
S1	26	female	15.3	35.6	83.3	44.7	96.0
S2	30	male	33.5	36.1	76.7	48.8	96.0
S3	22	male	22.6	25.3	100.4	49.4	92.0
S4	47	female	59.4	90.5	77.6	75.8	—
S5	40	female	48.8	29.1	129.8	69.2	—
S6	42	male	63.0	70.9	117.9	83.9	—
S7	28	female	20.1	46.0	132.2	66.1	—
S8	26	male	29.4	51.3	112.4	64.4	—
S9	26	male	44.6	30.2	88.4	54.4	—
S10	33	female	60.7	22.1	165.0	82.6	—
S11	27	male	40.5	32.8	81.0	51.4	—

Table 1. Details of all subjects in the study.

collection, and the limited availability of effective data are major challenges. Chisco aims to effectively mitigate these issues.

Although datasets for decoding imagined speech have previously existed[2,13–16], they focus on classification at the word level and do not involve the semantic reconstruction of complete sentences. To the best of our knowledge, our dataset represents the first attempt at sentence semantics reconstruction of imagined speech.

In our experimental paradigm, participants first read a passage of text, and then recall and imagine themselves speaking it. Therefore, each trial's data includes two parts: 5s of EEG during the reading phase and 3.3s of EEG during the imagined speech phase. Besides, to enhance the decoding performance in future research, we extended the experimental duration for each participant. We recruited three participants, each of whom completed 6,681 trials. Consequently, the EEG data for each participant exceeded 900 minutes, making it the largest dataset per individual currently available for decoding neural language. This large dataset for each individual ensures that the subject-specific decoders are more thoroughly trained, thereby promoting the development of more user-friendly BCIs.

Methods

Participants. EEG Data were collected from these three subjects: S1 (female, age 26), S2 (male, age 30) and S3 (male, age 22). No participant reported neurological or psychiatric history. All subjects were right-handed and had normal hearing and normal or corrected-to-normal vision. None of them had any previous BCI experience. All participants gave written consent for their participation and the reuse of the data prior to the start of the experiments. The study has been reviewed and approved by the Ethics Commission of Harbin Institute of Technology (Approval Number: HIT2024035) in accordance with the Declaration of Helsinki (2013)[17].

To enhance the quality of imagined speech EEG data, we first conducted a screening process for the participants. We initially recruited 11 volunteers (22-47 years old, Chinese native speakers, 6 males) for a five-day session of Integrative Body-Mind Training (IBMT)[18], a meditation technique aimed at improving concentration. Previous studies have demonstrated that IBMT significantly enhances participants' attention capacity and reduces fatigue levels. Furthermore, IBMT training has been associated with increased vigor, as measured by the Profile of Mood States scale, and a marked reduction in stress-related cortisol levels[18]. This training was intended to better prepare participants for our cognitively demanding experimental paradigm. Following the IBMT, we administered the Attention Network Test (ANT)[19] to all participants. The Attention Network Test (ANT) is a psychological assessment used to evaluate an individual's attentional functions. It measures three key indices—alerting, orienting, and executive control—to assess performance across different attentional tasks[19]. These indices correspond to the individual's preparedness for an impending event, the ability to focus attention on a specific spatial location, and the executive control when handling conflicting information, respectively. Consequently, lower ANT scores indicate greater attentional focus. We calculated the mean scores of all participants across the three ANT subtests and selected the three participants with the lowest ANT scores for subsequent experiments. Subsequently, the three selected participants underwent a Chinese language proficiency assessment (HSK Level 6 Reading Section) to ensure they could accurately comprehend the text displayed on the screen during the experiments. Details for each participant are presented in Table 1. No one withdrew during the experiment.

Text materials. To ensure our dataset is suitable for training BCIs for everyday use, our textual materials should encompass a broad range of daily language. To achieve this objective, we manually selected daily expressions from Chinese social media platform Weibo, public datasets ROCstory[20] and Dailydialog[21]. These expressions were initially categorized into 39 types using a combination of machine learning clustering algorithms and manual annotation by human experts. Subsequently, they were rephrased into sentences consisting of 6 to 15 Chinese characters through a crowdsourcing approach. Ultimately, we compiled a dataset containing 6,681 sentences of daily Chinese expressions. An overview is shown in Fig. 1, and specific contents can be found in the text labels provided by Chisco. The crowdsourcing method employed is outlined as follows:

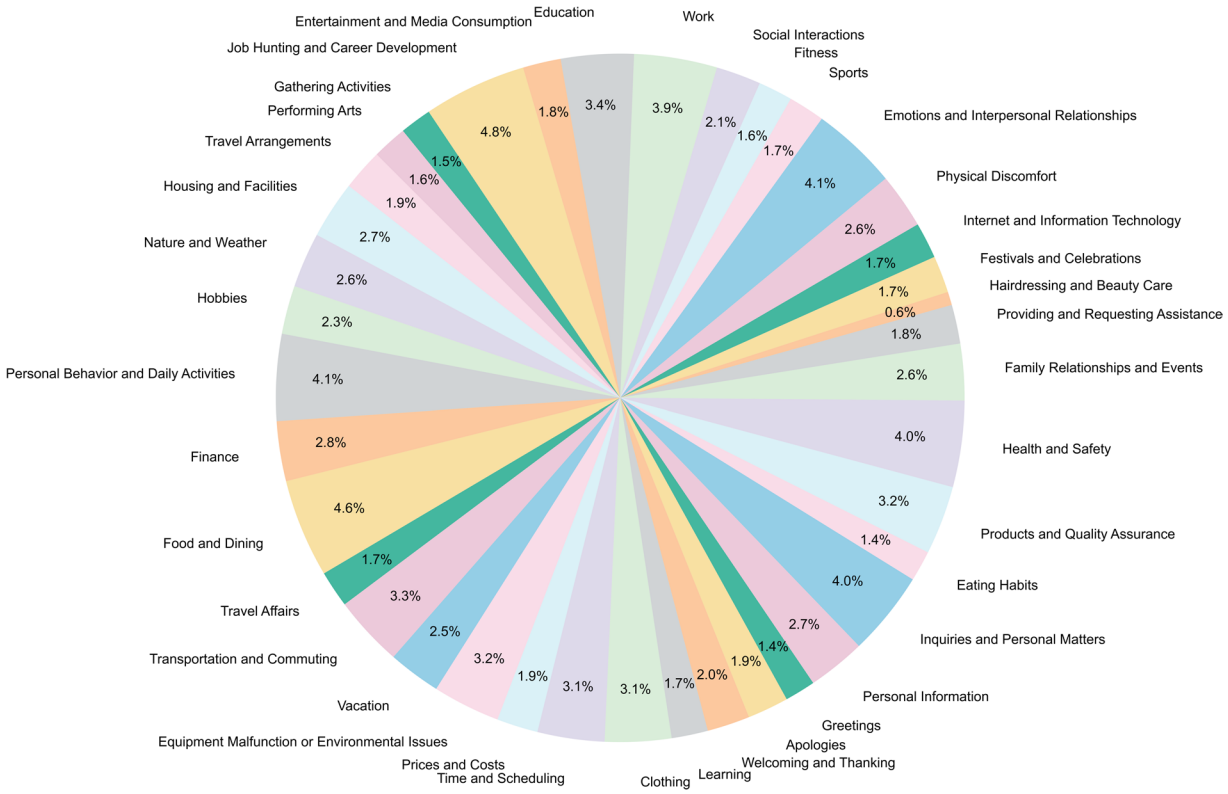

Fig. 1 Overview of text data. The proportion of each semantic category within the overall dataset, comprising a total of 39 categories.

- Each sentence must be independently understandable without the need for multiple segments to form a coherent discourse.
- Avoid containing clauses and indirect anaphora.
- Limit the nouns involved in the text to no more than three.
- Exclude sentences that are ambiguous.
- Prohibit any form of racial or gender discrimination.

The crowdsourcing method was designed to reduce syntactic complexity, thereby reducing the likelihood of errors during the imagined speech process. We provide English translations of several sentences and their categories as examples:

- "Today's dinner tasted great" - "Food and Dining"
- "She is going to listen to the concert" - "Performing Arts"
- "I am deeply sorry for my behavior" - "Apologies"

Experimental setup. In order to familiarize the participant with the experimental procedure and the room environment, a comprehensive explanation of all experimental steps was provided during the placement of the EEG cap and external electrodes. The setup process required approximately 30 minutes. Figure 2 shows the main experiment setup. In the experimental setup, a screen attached to PC1 presents stimuli to the subjects. The EEG cap captures the signals, which are then filtered and amplified by the headbox. These signals are further amplified by an amplifier and are tagged with markers derived from PC1 to enhance signal integrity. The marker signals facilitate the segmentation of the EEG data. Subsequently, the processed EEG signals are transmitted to the experimenter computer, PC2, for storage. The experimental operator monitors the experiment through PC2. To minimize interference with the subject, only the PC1 screen is placed inside the collection room, while other equipment is housed in the control room.

Experimental procedures. As described in the "Participants" subsection, based on the results of the ANT, we selected 3 participants from a pool of 11 applicants for the main experiment. The schedule for the main experiment is as follows:

1. Each participant engaged in a five-day experiment.
2. Nine experimental blocks were conducted daily.
3. Except for the final block, each experimental block consists of 150 trials (The total number of sentences is

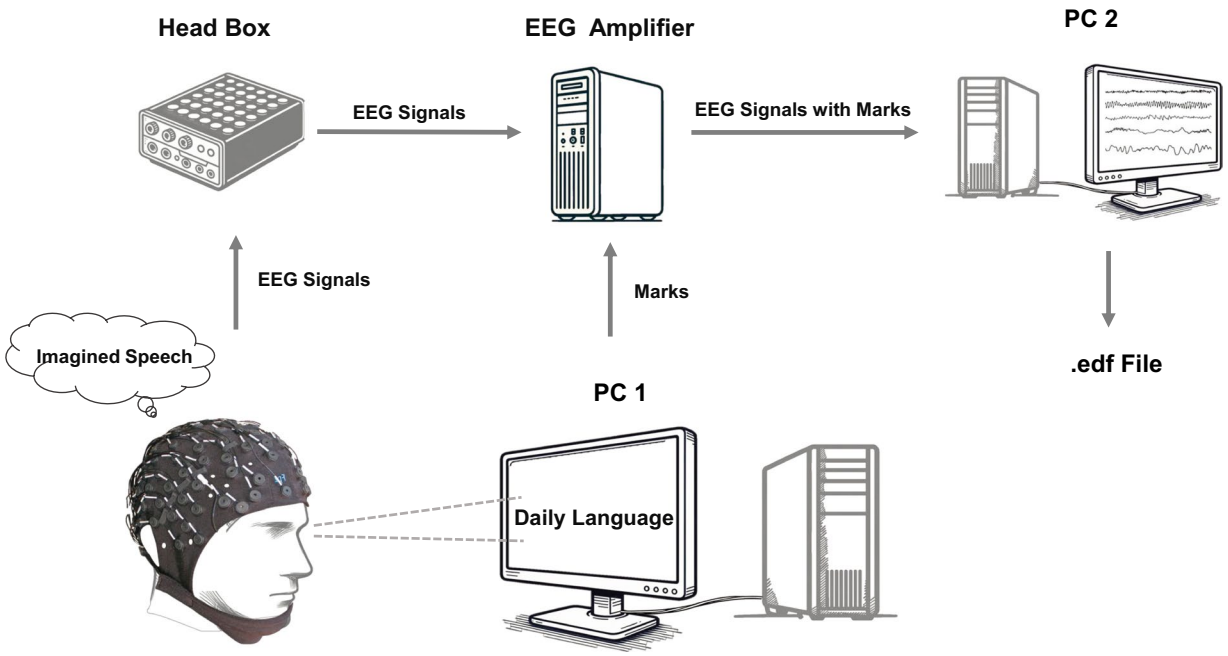

Fig. 2 Main experimental setup. PC1 was used to present stimuli and send markers, while the headbox was employed for preliminary filtering and amplification of EEG signals. The EEG amplifier further amplified and integrated the EEG signals, and PC2 was responsible for storing the EEG data and monitoring the entire experimental process.

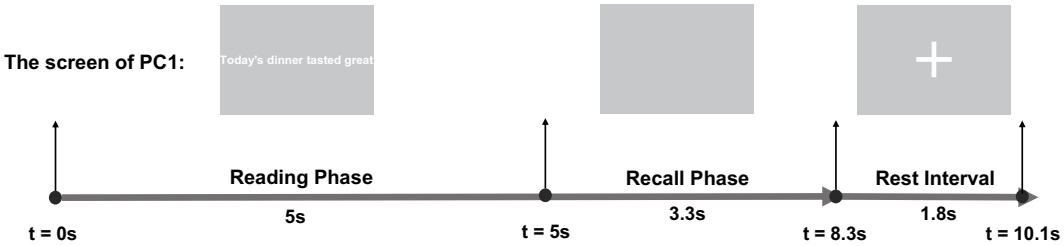

Fig. 3 Trial workflow. During the Reading Phase, participants read and memorized sentences. The Recall Phase involved a task where participants imagined the speech, and the Rest Interval provided time for rest.

6,681, which equals 150 multiplied by 44 plus 81. Consequently, the final block could not be fully completed), with each trial lasting 8.3s with an inter trial interval of 1.8s, totaling approximately 25 minutes.

Figure 3 describes the composition of each trial, together with the relative and cumulative times. Each trial started with a reading phase, during which the subjects were presented with a single sentence on the computer screen for 5,000ms and were instructed to saliently read it. The details of the presented sentences are specified in the "Text Materials" section. Following the reading phase, a blank screen was displayed for 3,300ms, during which the subjects were required to imagine the sentence in mind at a constant speed. During this stage, the subjects need to imagine themselves speaking and not making any movements in their throat or limbs. The third phase is a rest interval, during which participants are allowed a brief pause. The duration of all trials was fixed and did not require participants to control it via button presses, aiming to minimize the influence of movement-related cortical potentials (MRCP)[22], which could generate frontal lobe EEG activity and interfere with the decoding of imagined speech.

This experimental paradigm follows the classic design used in imagined content research, where imagination is guided through reading[2,13–16]. Moreover, the reading phase activates brain regions associated with language processing, such as the left temporal lobe and the inferior frontal gyrus, which have been shown to also be activated during imagined speech, with partial similarities in neural representations[23,24]. Thus, this classic paradigm offers two key advantages: first, it restricts participants' imagination content, avoiding dataset bias caused by free imagination; second, the activation of similar brain regions during both phases helps participants transition more quickly into the imagined state, improving the signal-to-noise ratio of the data. These factors contribute to more effective training of EEG decoders for BCI applications.

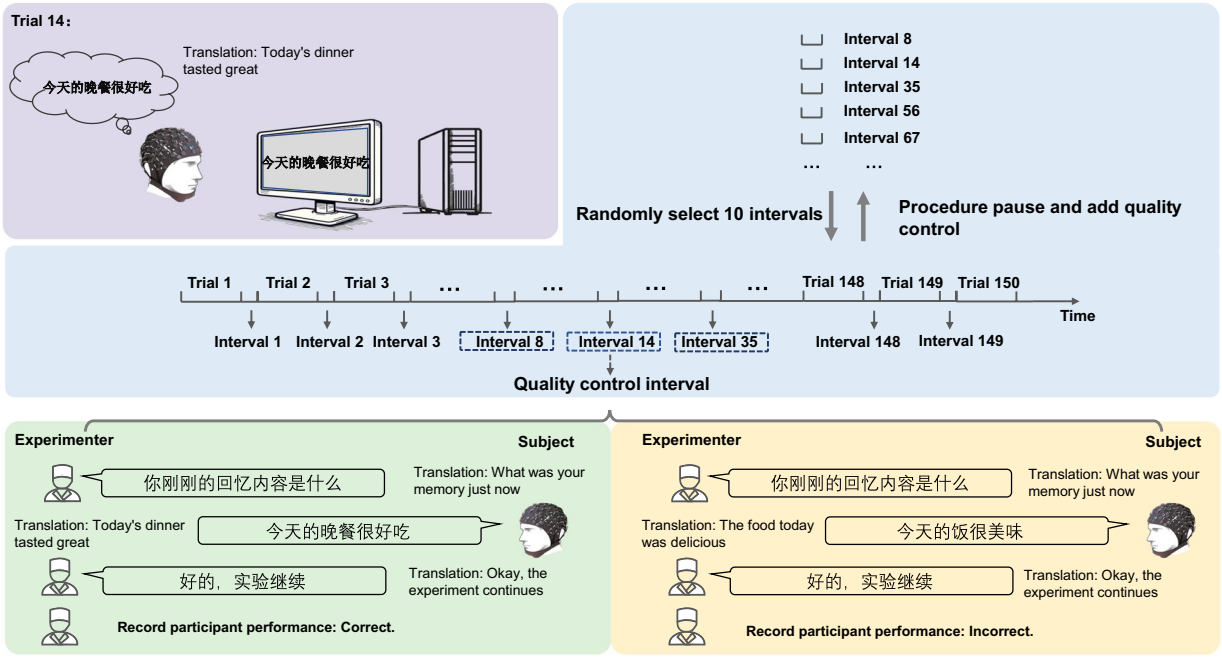

Fig. 4 An example of quality control. Each block consisted of 150 trials and 150 rest intervals, with 10 rest intervals randomly selected for quality control. For example, in one block, trial 14 was selected, with the correct answer being 'Today's dinner tasted great'. To minimize participant movement, researchers communicated orally with the participants to confirm the accuracy of their responses.

The duration for each phase was determined through a preliminary pilot experiment. We selected 20 sentences from the dataset, including both the shortest and longest sentences. In the reading phase, we initially considered reading times of 4,000 ms, 4,500 ms, 5,000 ms, 5,500 ms, and 6,000 ms based on an estimate of typical reading habits among Chinese individuals. Three participants each read these 20 sentences five times with varying durations. Two participants reported that 5,000 ms was just sufficient to read and memorize the sentences, while one participant reported that 4,500 ms was adequate. Therefore, we ultimately selected 5,000 ms as the reading time. For the longest sentence, this equates to 333 ms per character, which is very close to the 350 ms per character used in the ChineseEEG study[9]. Similarly, in the recall phase, we tested durations of 3,000 ms, 3,300 ms, 3,500 ms, 3,800 ms, and 4,000 ms to identify the most appropriate timing before commencing the experiment, and observed that 3300 ms was the optimal duration.

Our experimental paradigm presents a significant challenge for participants, primarily for three reasons. First, the requirement for precise memory necessitates that participants maintain continuous attention to minimize errors. Secondly, existing research on imagined speech has mainly concentrated on the recall of individual words[2,13–16]. In contrast, our experiment was designed to ask the subjects to recall a single sentences, thereby increasing their memory load. Thirdly, previous studies have indicated that even within a single subject, brain states can vary over time[25–27], introducing potential noise into language decoding tasks. To mitigate this issue, we arrange experiments for each participant in a condensed manner, with 9 blocks per day. This approach significantly reduces variance within the data set, thereby enhancing its quality. However, the high intensity of this schedule poses a challenge to subjects.

In this context, the pre-experimental training and selection of participants, as well as data quality control during the experiment, become particularly critical. To ensure participants were well-adapted to our experimental paradigm, in addition to receiving IBMT training before the experiment, participants were also required to recall text at a pace of approximately one character every 220ms. To facilitate this, we designed a software that displayed one character every 220ms, which participants were instructed to follow and read this, thereby acclimating to the rhythm.

In addition, to ensure the quality of data collected during the experiment, we evaluated the attention of participants during the experiment. We incorporated random attention checks into the experimental paradigm. In each block, we added a test at the end of 10 randomly selected trials, requiring participants to verbally recount the memorized content. This was evaluated by experimenters, and a discrepancy of more than four characters from the original text was considered a recall error, Fig. 4 provides a specific example. If a participant made two or more errors in these 10 tests, they were deemed to have lapses in concentration during that period. Subsequently, a rest period was provided to the participant to ensure attention restoration, followed by restarting the block.

Data acquisition. The study was conducted in an electrically shielded room. Participants were seated comfortably in front of a computer screen, positioned at a distance of 80 cm from the screen, directly facing the center.

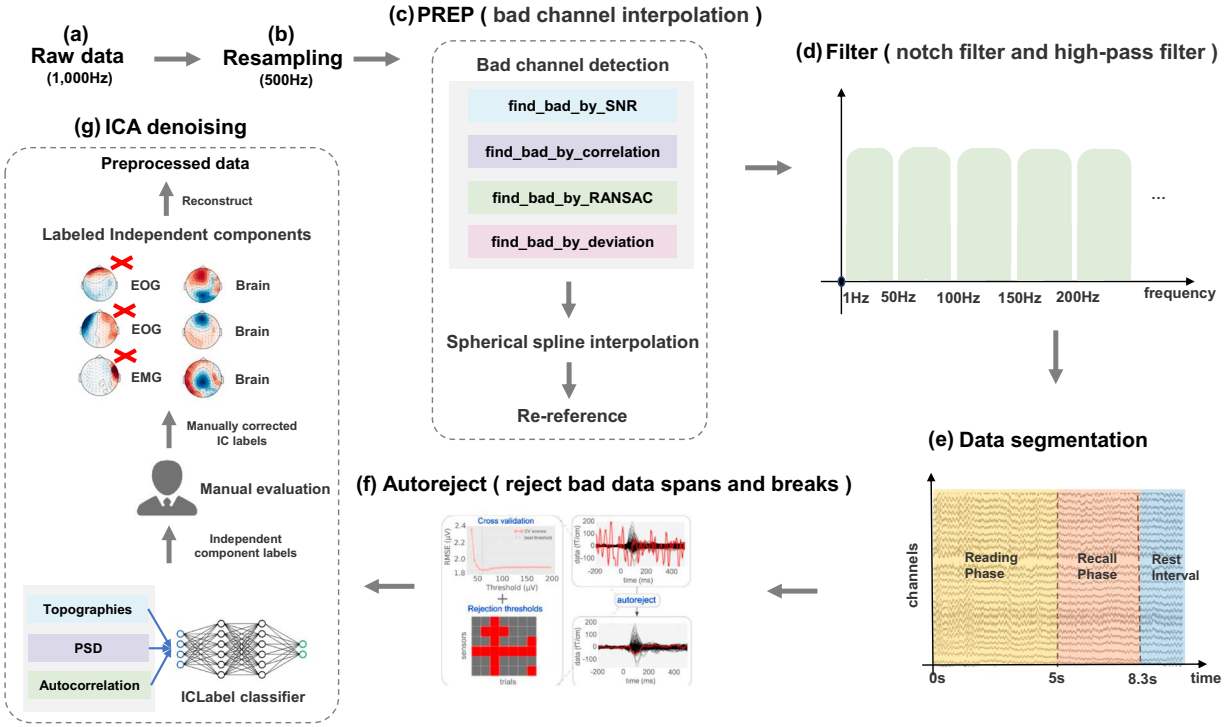

Fig. 5 EEG preprocessing pipeline (part (**f**) cites previous work[35]). (**a**) Raw data were recorded at a sampling rate of 1,000 Hz, stored in the original .edf files. (**b**) The raw data were then downsampled to 500 Hz. (**c**) The PREP algorithm was applied for bad channel detection and re-referencing. (**d**) Power line noise was removed, and high-pass filtering was performed. (**e**) The data were segmented according to the experimental paradigm. (**f**) Bad data spans and breaks were rejected using the Autoreject algorithm. (**g**) ICA algorithm was used to remove noise such as EOG and EMG.

The stimuli were presented on a 26-inch screen with a refresh rate of 60 Hz and a resolution of 1920 × 1080 pixels. Visual stimuli consisted of sentences drawn from daily language, ranging in length from 6 to 15 Chinese characters. These were displayed in 28-point Song typeface, on a single line, centrally placed on a grey background with white text to prevent dazzling and eye fatigue. Participants were also offered snacks and water during the breaks and were encouraged to rest.

Additionally, a personalized head case was designed for each participant with three objectives in mind. First, to control head movements and thereby reduce noise from electrode displacement. Second, the head case provides support to the participant's head, alleviating physical fatigue during prolonged experimental sessions. Third, since each participant needs to participate in the experiment for 5 days, the head case ensured that the positioning of electrodes on the scalp remained consistent each time the EEG cap was reapplied, thereby reducing the internal variance of the data.

EEG data were collected using the SynAmps-2 128-channel Amplifier and the 128-channels Quik-Cap, both manufactured by Compumedics Neuroscan. The Curry 9 software, also from Compumedics Neuroscan, was employed for hardware control and impedance checks before each experiment, ensuring impedances were below 40 kΩ. The gap between the scalp and the electrodes was filled with a conductive ZPTC® Easy Gel.

After each experimental block, PC2 generated and stored an EDF file containing continuous recordings from the 125 EEG channels, 6 external channels (utilized to record the vertical and horizontal electrooculography signals, as well as the voltages at the mastoids on both sides), and mark signals. The experiments were conducted with a sampling rate of 1 kHz using alternating current (AC) mode, with an accuracy of 3 nV/LSB. Stimulus presentation was managed using the open-source software package, PsychoPy[28] (v2023.2.3).

Data preprocessing. Preprocessing was implemented in Python, primarily using the MNE library[29]. We performed minimal preprocessing on the data to retain the maximum amount of valid information, allowing researchers to tailor further processing to their specific needs. The preprocessing pipeline is illustrated in Fig. 5, with detailed explanations of each step provided below.

Resampling. The raw data stored in the .edf file contains records of the complete EEG signals as well as the events, with a sample rate of 1,000 Hz. We resampled the data to 500 Hz. Since there is still debate about which frequency bands contain the information related to imagined speech[30–32], we employed a conservative downsampling approach. According to the Nyquist sampling theorem[33], our high-frequency sampling rate of 500 Hz fully captures EEG information below 250 Hz, which includes the ultra-high gamma band[32].

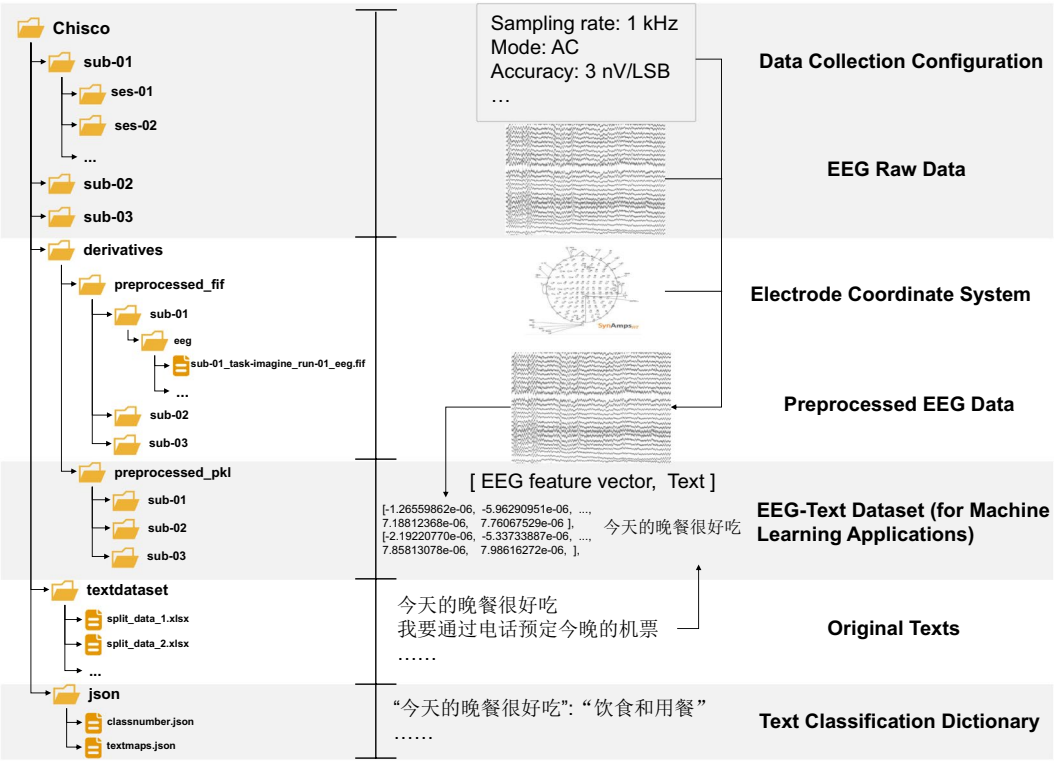

Fig. 6 Dataset structure. The dataset includes raw data from three participants, preprocessed data, text data, and JSON files formatted for use in machine learning programs.

The PREP pipeline. PREP is a standardized early-stage EEG processing procedure, which was applied after data resampling[34]. A summary of the PREP pipeline is:

1. Removing line noise without committing to a specific filtering strategy.
2. Robustly reference the signal relative to an estimate of the “true” average reterence.
3. Detect and interpolate bad channels relative to this reference.
4. Retain suffcient information to allow users to re-reference using another method or to undo interpolation of a particular channel.

Filtering. In this part, we only applied a 50 Hz notch filter to remove power line noise and a zero-phase high-pass finite impulse response filter with a lower cutoff at 1 Hz. We did not perform low-pass filtering for reasons detailed in the “Resampling” discussion.

Data segmentation. The continuous recorded data were segmented, dividing it into two kinds of segments based on the experimental paradigm: the phase from 0 to 5s after the event marker was designated as the reading segment, while the phase from 5 to 8.3s was defined as the imagined speech segment.

Autoreject. The Autoreject algorithm was employed to automatically identify and reject bad data spans and breaks[35]. This technique optimizes the rejection thresholds on a per-channel basis, ensuring that the cleaning process is adaptive and minimizes the loss of good data.

Independent components analysis (ICA). ICA is a standard and widely used blind source separation method for removing artifacts from EEG signals[36]. For our dataset, ICA processing was performed only on the EEG channels, using the MNE implementation of the extended infomax ICA. Due to the wide frequency range of our data, we set the number of independent components to 30, higher than in other neural language decoding studies[9], ensuring that they capture the majority of the information. For noise component identification, we combined MNE-ICALabel automatic annotation with manual evaluation, ensuring accuracy without excessively increasing the manual processing burden.

Data Records

All data files, including the raw recordings, can be accessed at the OpenNeuro repository and conform to BIDS standards[37]. All files are contained in a main folder called “Chisco”[38], structured as depicted in Fig. 6. Public data is distributed under the CC0 License[39].

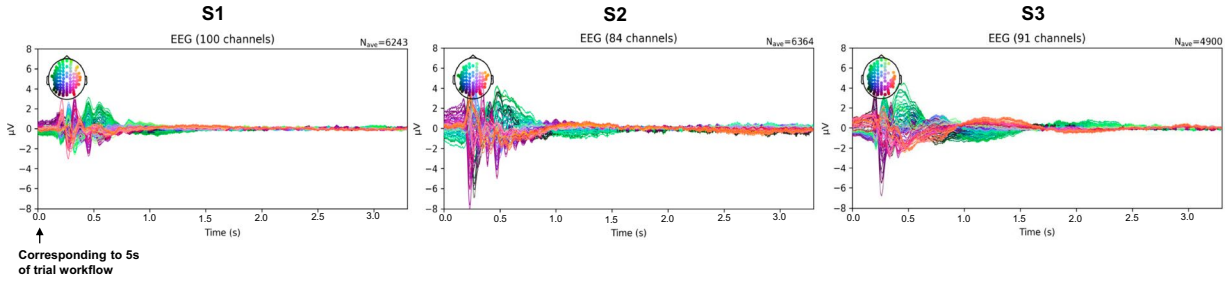

Fig. 7 ERPs of different subjects.

Subject ID	Reading Phase	Recall Phase	Rest Interval
S1	6.90	9.09	6.55
S2	11.92	15.99	10.91
S3	7.57	10.24	7.52

Table 2. Spontaneous eyeblink rate (SEBR) across different phases of the task. Values represent the average blink rate per minute across all available trials for each participant.

The open-source dataset includes raw data, preprocessed data and text data. The preprocessed data is available in two formats: .fif files suitable for neuroscience research and .pkl files designed for machine learning and BCI studies.

Raw data. The raw data are stored in the sub-01, sub-02, and sub-03 folders, corresponding to three subjects. Different ses identifiers represent data collected on different days. The raw data files contain continuous recordings of the entire session for all 125 channels. The .edf file includes all EEG data and tagged events, along with information about the recording sampling rate, channel names and recording filters. Raw events are extracted from the raw data file and contain tags synchronized with the recorded signals. Each event code with ID 65,380 marks the start of a trial, with a total of 20,043 trials acquired.

Preprocessed data. The preprocessed data are stored in the derivatives folder. The preprocessed_fif and preprocessed_pkl files correspond to two different data formats, but they were preprocessed in the same manner, as described in the "Data preprocessing" subsection of the "Methods".

Each of .fif files contains an MNE segment object, with the EEG data information of all trials in the corresponding session, including the sentences corresponding to each segment stored as metadata. The preprocessed data are also stored in .pkl format, which contains a list where each item is a dictionary entry in the form of {"text": sentence, "input_features": data}. Here, the "sentence" represents the text read or imagined, and the "data" is an ndarray containing the corresponding EEG segment.

Text data. The textdataset folder contains crowdsourced raw text data. The JSON files store this data in a dictionary format, facilitating the development of code for subsequent research based on this dataset.

Technical Validation

Event related potentials (ERPs) and Eye Blinks During Imagined Speech. It is well-known that ERPs can capture the brain's response to specific events and relate these responses to time, thus providing detailed insights into the dynamic processes of brain information processing. We aimed to identify how participants' brain activity was modulated by the experimental tasks. To this end, we calculated ERPs using all available trials, the results are shown in Fig. 7. To improve the precision of ERPs calculations, we computed the union of bad channels from all trials and excluded these channels from the final images. The results indicate that participants exhibit a significant ERP in the EEG upon entering the recall phase. This is consistent with the responses of a healthy human brain[40].

Furthermore, the spontaneous eyeblink rate (SEBR) has been demonstrated to serve as a measure of participants' responses to visual stimuli[41]. We analyzed the SEBR across different phases of the task cycle, as shown in Table 2. The results indicate that all participants exhibited a significantly lower blink rate during the reading phase and the resting interval, where a fixation cross was displayed on the screen, compared to the imagined speech recall phase. This suggests that participants actively engaged with the visual stimuli of the experimental paradigm, as evidenced by a reduced SEBR, reflecting their effort to gather more visual information. These findings demonstrate the participants' active cooperation with the experimental procedure.

EEG topographic mapping. As illustrated in Fig. 7, the activities of different brain regions exhibit markedly distinct patterns. To further analyze these differences, we used all available trials to separately calculate the brain region activity of the three participants throughout the imagined speech process. As shown in the Fig. 8, the brain activity of the three participants exhibited remarkable consistency during the imagined speech process.

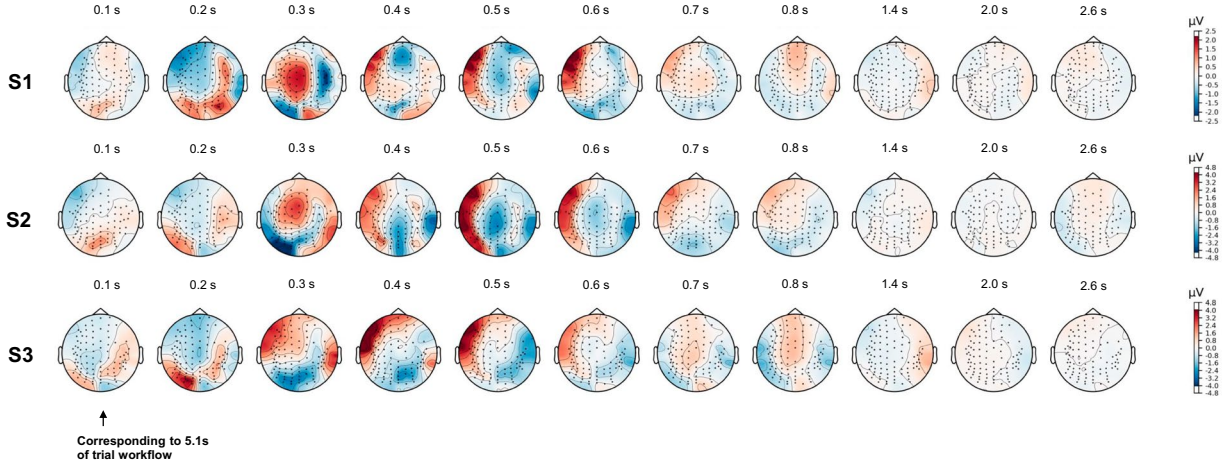

Fig. 8 Topographical maps of grand averaged imagined speech.

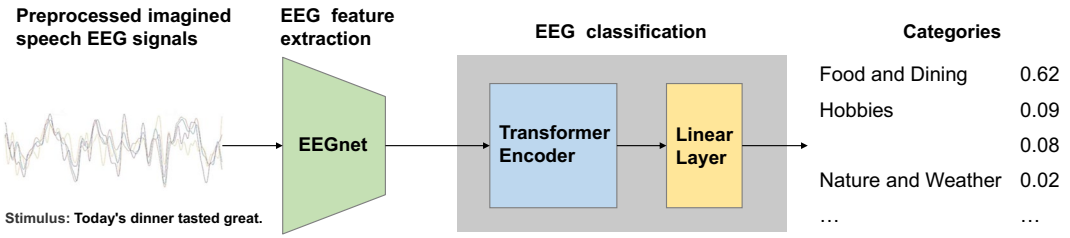

Fig. 9 Model architecture. First, EEGNet was employed to extract primary EEG features from the preprocessed EEG data. Subsequently, self-attention mechanism was applied using a transformer encoder combined with a fully connected layer for classification, ultimately yielding the classification results.

Since all three participants were right-handed, the left hemisphere of their brains is dominant. Initially, around 0.1s, the occipital lobe showed significant activation, which is associated with the processing of visual information[42], indicating that participants first received instructions to perform the recall phase through visual input. Subsequently, at around 0.2s, the posterior part of the temporal lobe and inferior parietal lobule was activated, indicating that participants were generating sentence semantics[43]. At 0.3s, the left hemisphere's parietal lobe and temporal lobe were activated, consistent with previous studies[44] that show these regions are engaged during creative tasks. This marks the transition to the next stage of sentence generation for the participants.

For S1 and S2, the left temporal regions showed significant activation. This indicates that the participants were constructing the grammatical structure of sentences[45], a process that starting from 0.4s and lasting approximately 0.3s, and S3 completed this process at 0.6s. Following this, the frontal and parts of the parietal lobes became the most active brain regions, which are involved in logical thinking and motor control[46,47]. This suggests that participants were managing the imagined speech process and to suppress any muscle movements.

Additionally, around 1.4s, significant activation was observed in the right hemisphere of the brain, associated with emotion and visual imagination[48]. This implies that during the imagined speech, participants might have had positive or negative connotations and visual imagery related to the sentence semantics.

Semantic information in EEG. To demonstrate that our imagined speech dataset contains effective semantic information and to provide a baseline for future work based on this dataset, we constructed a deep learning model to classify imagined speech EEG signals. Our model is built on EEGNet[49] and Transformer Encoder[50] architectures. The input to the model is preprocessed imagined speech EEG signals, and the output is the semantic category of the sentence corresponding to the imagined speech, as annotated in the "Text materials" subsection of "Methods", encompassing 39 categories. The model architecture is illustrated in Fig. 9.

The mathematical formulation of the model can be formally expressed as follows:

$$\hat{y} = \mathrm{MLP}(\mathrm{Transformer_encoder}(\mathrm{EEGNet}(x))).$$

The training objective of the model is:

$$\min \mathscr{L},\ \mathscr{L} = -\sum_{i=1}^{39} y_i \log\left(\hat{y}_i\right).$$

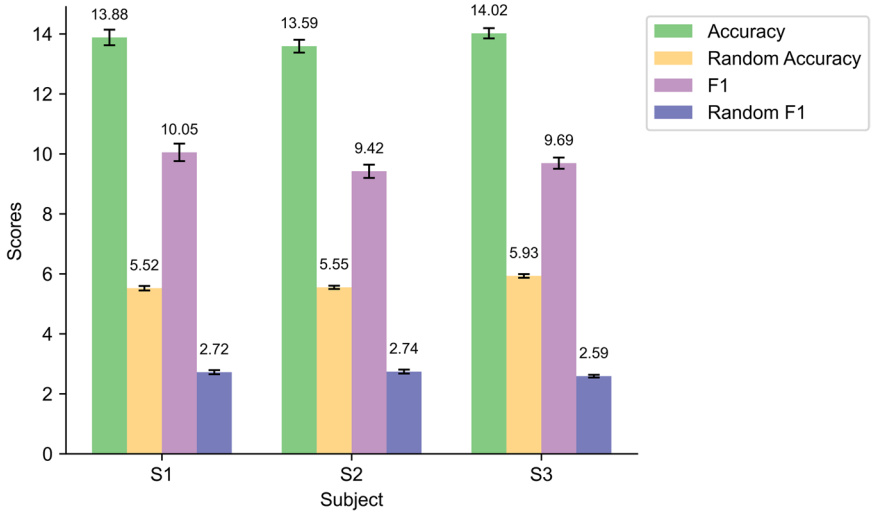

Fig. 10 The results of semantic classification. The accuracy and F1 scores of the three participants and their corresponding random baseline values.

where x represents the preprocessed imagined speech EEG signals, EEGNet(x) denotes the feature extraction layer using the EEGNet architecture, Transformer_encoder($\cdot$) indicates the application of the Transformer Encoder to the extracted features, MLP($\cdot$) refers to the Linear Layer used to classify the output from the Transformer Encoder, $\hat{y}$ is the predicted probability distribution over the 39 semantic categories, y is the true probability distribution over the 39 semantic categories and is a one-hot vector, $\mathscr{L}$ is the cross-entropy loss.

The final classification results are shown in the Fig. 10, with all three subjects achieving significantly higher than random classification accuracy and F1, demonstrating active participation in the experiment. In this method, we obtain random values by disrupting the correspondence between the input EEG signals x and their associated labels y. This approach prevents the model from learning how to compute the correct labels from the EEG signals, while still allowing it to learn from the label distribution. The experimental results indicate that the Chisco EEG signals contain valid semantic information.

Moreover, to assess the consistency of neural language representations across different subjects, we referenced a previously established method to calculate inter-subject correlation (ISC)[51]. This approach involves applying spatial filters to EEG signals, followed by computing Pearson correlation coefficients on the filtered data. In our experiment, we first randomly combined training data from three subjects to train the classifier in Fig. 9, subsequently utilizing the trained EEGNet as a spatial filter. The mathematical formulation is as follows:

$$r(A,\,B) = \frac{1}{m}\sum_{i=1}^{m}\frac{\sum_{j=1}^{n}\left(a_{ij} - \frac{1}{n}\sum_{j=1}^{n} a_{ij}\right)\left(b_{ij} - \frac{1}{n}\sum_{j=1}^{n} b_{ij}\right)}{\sqrt{\sum_{j=1}^{n}\left(a_{ij} - \frac{1}{n}\sum_{j=1}^{n} a_{ij}\right)^2}\sqrt{\sum_{j=1}^{n}\left(b_{ij} - \frac{1}{n}\sum_{j=1}^{n} b_{ij}\right)^2}}$$

where m denotes the number of trials each subject completed, and n represents the dimensionality of the signal for each trial post-EEGNet filtering. Specifically, a_{ij} refers to the j-th dimension of the filtered EEG signal from the i-th trial of subject A, while b_{ij} represents the corresponding filtered EEG signal from subject B. Our results yielded correlations of $r(S1, S2) = 0.126$, $r(S2, S3) = 0.147$, and $r(S1, S3) = 0.168$, closely aligning with prior experimental ranges (0.1 - 0.25)[51].

Comparison of decoding between imagined speech and reading stimuli. To compare with previous work, we applied the same method to the Zuco dataset[8], which contains EEG recordings of human reading tasks. However, since Zuco does not include sentence categories, we used k-means clustering and determined, using the elbow method[52], that the Zuco dataset contains 22 semantic categories. As illustrated in the Fig. 11, the x-axis represents the number of clusters, while the y-axis denotes the SSE (Sum of Squared Errors). The elbow point, indicating the optimal number of clusters, is observed at 22. Given the relatively small number of categories in the Zuco dataset, the classification task is comparatively simple. To ensure a fair comparison between Zuco and Chisco, we employed the Cohen's Kappa Coefficient (CKC)[53] for comparison, defined by the following formula:

$$CKC = \frac{p_o - p_e}{1 - p_e},$$

where p_o represents the observed accuracy, p_e denotes the expected accuracy by chance, and CKC can represent the difficulty of classification tasks, the lower the CKC value, the more challenging the decoding task becomes. As depicted in the Table 3, decoding EEG during imagined speech is more challenging than decoding EEG during reading. This aligns with our intuition and previous research[54], as the brain's response during reading is more robust compared to the self-initiated brain activity associated with imagined speech.

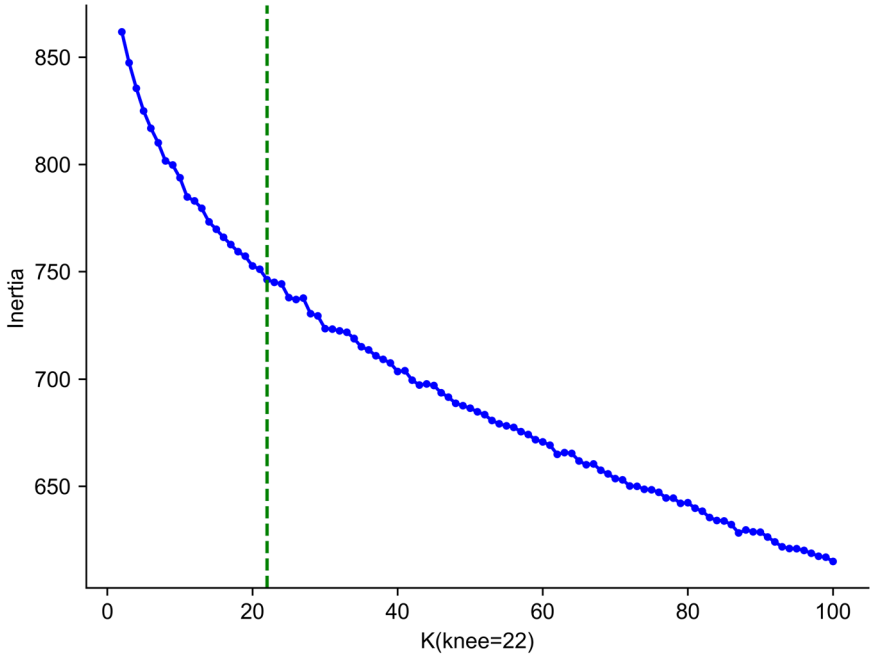

Fig. 11 Utilizing the elbow method with k-means clustering on the Zuco dataset. The green line corresponds to the elbow point, where the curve transitions from a steep decline to a more gradual slope, indicating the onset of diminishing returns.

Subject ID	Accuracy	Random Accuracy	CKC
S1	13.88	5.52	8.85
S2	13.59	5.55	8.51
S3	14.02	5.93	8.60
Zuco[8]	27.80	12.87	17.14

Table 3. Comparing semantic information in EEG during imagined speech and reading tasks.

Limitations and future work. Given the specificity of EEG signals across different subjects, we extended the experimental duration for each participant to facilitate subsequent research on language decoding BCI based on Chisco. Each participant completed 6,681 trials, with the experimental stimuli covering nearly all aspects of everyday language use, resulting in over 900 minutes of EEG recording time, with more than 360 minutes dedicated to imagined speech EEG recordings. To the best of our knowledge, this represents the largest dataset per individual currently available for decoding neural language.

Therefore, we have reason to believe that the 900 minutes of EEG recordings capture brain activity associated with a broad range of everyday language use. Additionally, compared to shorter durations, 900 minutes of EEG data is beneficial, as machine learning models have limited capacity to learn from noisy data. As shown in Fig. 12, we reduced the dataset size to train the EEG semantic classifier discussed in the previous section, and the classifier's performance improved with increasing data. Each value in the figure represents the average of five repetitions. However, despite the extended recording time, the dataset remains insufficient, as the curve indicates that even with all available data, performance has yet to reach a plateau. According to the scaling laws of machine learning, the curve shows no clear trend of diminishing marginal returns[55]. Therefore, effectively expanding the dataset and efficiently utilizing existing data will be key directions for future research in imagined speech decoding.

Furthermore, exploring the similarities and differences in neural representations of imagined speech at the sentence level across different subjects is of significant interest. This endeavor will require future studies to be conducted with larger participant cohorts. We believe that this study provides a valuable reference point for such future research.

Usage Notes

To analyze and process the raw EEG signals, we recommend modifying our data preprocessing script. Our data preprocessing script was developed in Python 3.7, utilizing the MNE-Python package v1.6.0, MNE-icalabel v0.6.0, Pyprep v0.4.3, and Autoject v0.4.3.

For those interested in using our EEG signals for BCI development or machine learning research, we recommend using the preprocessed .pkl files available in Chisco. These files can be read using Python's pickle package

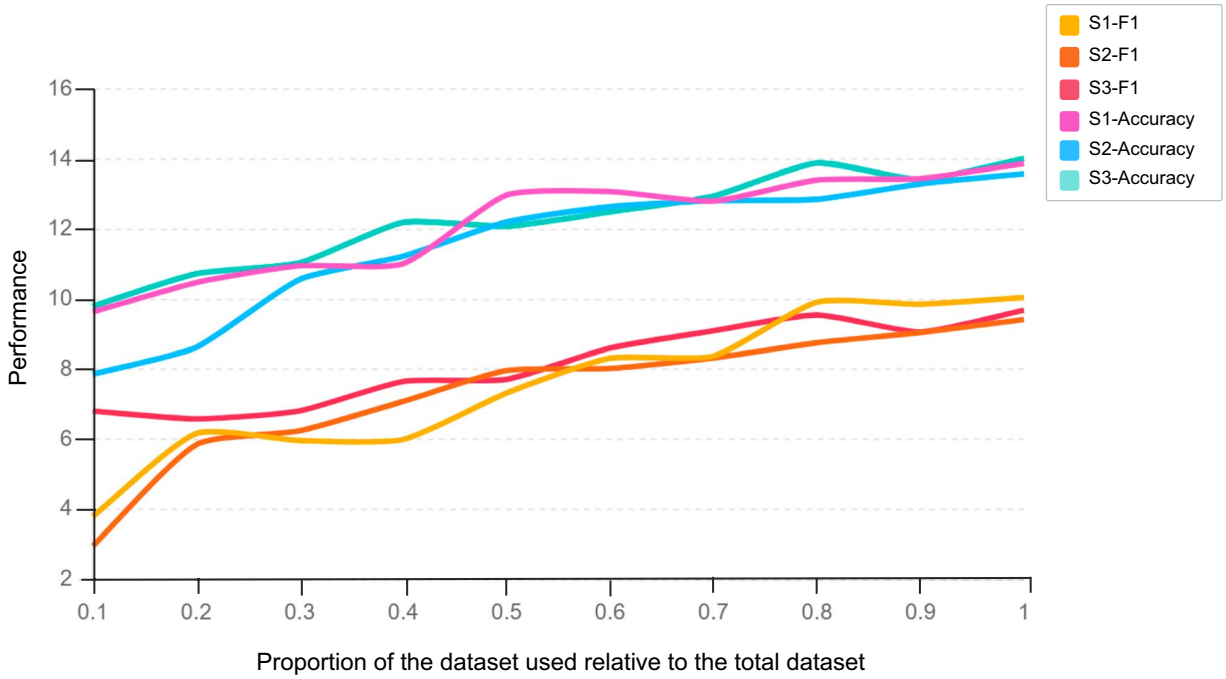

Fig. 12 Performance of the classifier with varying data scales.

and are compatible with various versions of PyTorch. The code provided in "Semantic information in EEG" can serve as an example of a machine learning application and can be used as a baseline for future research. In our example, we used Python 3.7 along with the following libraries: Scikit-learn v1.3.2, Torch v2.3.1, Numpy v1.23.1, and Transformers v4.41.0.

For analyzing the preprocessed EEG signals, we recommend using our preprocessed .fif files. As a specialized file format for storing neuroscience data, .fif has broader applicability and can be used with tools such as FieldTrip and MNE.

Code availability

In line with the philosophy of reproducible research, all codes used in this paper, including those for stimulus presentation, data preprocessing, and technical validation, are accessible at https://github.com/zhangzihan-isgood/Chisco. The usage instructions and parameter settings for all codes can also be found at this link.

Received: 2 July 2024; Accepted: 7 November 2024;
Published online: 21 November 2024

References

1. Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G. & Vaughan, T. M. Brain–computer interfaces for communication and control. *Clinical neurophysiology* **113**, 767–791 (2002).
2. Nieto, N., Peterson, V., Rufiner, H. L., Kamienkowski, J. E. & Spies, R. Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition. *Scientific Data* **9**, 52 (2022).
3. Lincoln, Y. S. & Guba, E. G.*Naturalistic inquiry* (sage, 1985).
4. Cooney, C., Folli, R. & Coyle, D. Neurolinguistics research advancing development of a direct-speech brain-computer interface. *IScience* **8**, 103–125 (2018).
5. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
6. Willett, F. R. *et al.* A high-performance speech neuroprosthesis. *Nature* **620**, 1031–1036 (2023).
7. Metzger, S. L. *et al.* A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620**, 1037–1046 (2023).
8. Hollenstein, N. *et al.* Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data* **5**, 1–13 (2018).
9. Mou, X. *et al.* Chineseeeg: A chinese linguistic corpora eeg dataset for semantic alignment and neural decoding. *Scientific Data* **11**, 550 (2024).
10. Oseki, Y. & Asahara, M. Design of bccwj-eeg: Balanced corpus with human electroencephalography. In *LREC*, 189–194 (2020).
11. Brennan, J. R. & Hale, J. T. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one* **14**, e0207741 (2019).
12. Nedergaard, J. S. & Lupyan, G. Not everybody has an inner voice: Behavioral consequences of anendophasia. *Psychological Science* 09567976241243004 (2024).
13. Wandelt, S. K. *et al.* Representation of internal speech by single neurons in human supramarginal gyrus. *Nature human behaviour* 1–14 (2024).
14. Wilson, H., Golbabaee, M., Proulx, M. J., Charles, S. & O'Neill, E. Eeg-based bci dataset of semantic concepts for imagination and perception tasks. *Scientific Data* **10**, 386 (2023).
15. Coretto, G. A. P., Gareis, I. E. & Rufiner, H. L. Open access database of eeg signals recorded during imagined speech. In *12th International Symposium on Medical Information Processing and Analysis*, vol. 10160, 1016002 (SPIE, 2017).
16. Zhao, S. & Rudzicz, F. Classifying phonological categories in imagined and articulated speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 992–996 (IEEE, 2015).

17. Association, W. M. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA* **310**, 2191–2194, https://doi.org/10.1001/jama.2013.281053 (2013).
18. Tang, Y.-Y. *et al*. Short-term meditation training improves attention and self-regulation. *Proceedings of the national Academy of Sciences* **104**, 17152–17156 (2007).
19. Fan, J., McCandliss, B. D., Sommer, T., Raz, A. & Posner, M. I. Testing the efficiency and independence of attentional networks. *Journal of cognitive neuroscience* **14**, 340–347 (2002).
20. Mostafazadeh, N. *et al*. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2016).
21. Li, Y. *et al*. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)* (2017).
22. Shakeel, A. *et al*. A review of techniques for detection of movement intention using movement-related cortical potentials. *Computational and mathematical methods in medicine* **2015**, 346217 (2015).
23. Tian, X. & Poeppel, D. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in psychology* **1**, 7029 (2010).
24. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nature reviews neuroscience* **8**, 393–402 (2007).
25. Finn, E. S. *et al*. Can brain state be manipulated to emphasize individual differences in functional connectivity? *Neuroimage* **160**, 140–151 (2017).
26. Pfurtscheller, G. Functional brain imaging based on erd/ers. *Vision research* **41**, 1257–1260 (2001).
27. Bergmann, T. O. Brain state-dependent brain stimulation. *Frontiers in psychology* **9**, 422698 (2018).
28. Peirce, J. W. *et al*. Psychopy2: experiments in behavior made easy. *Behavior Research Methods* **51**, 195–203 (2019).
29. Gramfort, A. *et al*. Mne software for processing meg and eeg data. *neuroimage* **86**, 446–460 (2014).
30. Lee, S.-H., Lee, M. & Lee, S.-W. Neural decoding of imagined speech and visual imagery as intuitive paradigms for bci communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **28**, 2647–2659 (2020).
31. Proix, T. *et al*. Imagined speech can be decoded from low-and cross-frequency intracranial eeg features. *Nature communications* **13**, 48 (2022).
32. Crone, N. E., Sinai, A. & Korzeniewska, A. High-frequency gamma oscillations and human brain mapping with electrocorticography. *Progress in Brain Research* **159**, 275–295 (2006).
33. Nyquist, H. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers* **47**, 617–644 (1928).
34. Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M. & Robbins, K. A. The prep pipeline: standardized preprocessing for large-scale eeg analysis. *Frontiers in neuroinformatics* **9**, 16 (2015).
35. Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F. & Gramfort, A. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage* **159**, 417–429 (2017).
36. Hyvärinen, A. Independent component analysis by minimization of mutual information (1997).
37. Gorgolewski, K. J. *et al*. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data* **3**, 160044 (2016).
38. Zhang, Z., Zhao, Y., Bao, Y. & Ding, X. "chisco" https://doi.org/10.18112/openneuro.ds005170.v1.0.1 (2024).
39. Creative Commons. Creative commons public domain dedication. https://creativecommons.org/publicdomain/zero/1.0 (2010).
40. Beres, A. M. Time is of the essence: A review of electroencephalography (eeg) and event-related brain potentials (erps) in language research. *Applied psychophysiology and biofeedback* **42**, 247–255 (2017).
41. Doughty, M. J. Consideration of three types of spontaneous eyeblink activity in normal humans: during reading and video display terminal use, in primary gaze, and while in conversation. *Optometry and vision science* **78**, 712–725 (2001).
42. Rehman, A. & Al Khalili, Y. Neuroanatomy, occipital lobe (2019).
43. Saur, D. *et al*. Ventral and dorsal pathways for language. *Proceedings of the national academy of Sciences* **105**, 18035–18040 (2008).
44. Liu, S. *et al*. Brain activity and connectivity during poetry composition: Toward a multidimensional model of the creative process. *Human brain mapping* **36**, 3351–3372 (2015).
45. Sakai, K. L., Noguchi, Y., Takeuchi, T. & Watanabe, E. Selective priming of syntactic processing by event-related transcranial magnetic stimulation of broca's area. *Neuron* **35**, 1177–1182 (2002).
46. Fuster, J. M. Frontal lobe and cognitive development. *Journal of neurocytology* **31**, 373–385 (2002).
47. Bisley, J. W. & Goldberg, M. E. Attention, intention, and priority in the parietal lobe. *Annual review of neuroscience* **33**, 1–21 (2010).
48. Demaree, H. A., Everhart, D. E., Youngstrom, E. A. & Harrison, D. W. Brain lateralization of emotional processing: historical roots and a future incorporating "dominance". *Behavioral and cognitive neuroscience reviews* **4**, 3–20 (2005).
49. Lawhern, V. J. *et al*. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering* **15**, 056013 (2018).
50. Vaswani, A. *et al*. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
51. Dmochowski, J. P., Sajda, P., Dias, J. & Parra, L. C. Correlated components of ongoing eeg point to emotionally laden attention–a possible marker of engagement? *Frontiers in human neuroscience* **6**, 112 (2012).
52. Thorndike, R. L. Who belongs in the family? *Psychometrika* **18**, 267–276 (1953).
53. Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**, 37–46 (1960).
54. Tomasino, B., Fink, G. R., Sparing, R., Dafotakis, M. & Weiss, P. H. Action verbs and the primary motor cortex: a comparative tms study of silent reading, frequency judgments, and motor imagery. *Neuropsychologia* **46**, 1915–1926 (2008).
55. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM* **64**, 107–115 (2021).

Acknowledgements

We gratefully acknowledge the financial support provided by the National Natural Science Foundation of China under Grants U22B2059 and 62176079, and the Natural Science Foundation of Heilongjiang Province under Grant Y02022F005.

Author contributions

Z.Z., X.D., Y.Z. and X.L. conceived the experiments, Z.Z., Y.Z. and Y.B. conducted the experiments, Z.Z., Y.Z. and Y.B. analysed the results, Z.Z., X.D., X.L., B.Q. and T.L. contributed to the writing of this manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.D. or X.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024