University of Essex

Research Repository

Decoding speech imagery or just noise?: a symptom of the replicability crisis

Accepted for publication in the Journal of Neural Engineering

Research Repository link: https://repository.essex.ac.uk/43210/

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the published version if you wish to cite this paper.
https://doi.org/10.1088/1741-2552/aea294

www.essex.ac.uk

University of Essex

Decoding Speech Imagery or Just Noise?: A Symptom of the Replicability Crisis

A. Tates, A. Matran-Fernandez, S. Halder, I. Daly

Brain-Computer Interfaces and Neural Engineering Lab, School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

E-mail: at18157@essex.ac.uk

July 2025

Abstract. *Objective:* Speech Imagery (SI) has emerged as a promising paradigm for Brain-Computer Interface (BCI) control, attracting growing interest due to its intuitive nature-—allowing users to interact with the system by internally saying a command. In this study, we investigate the replicability and reproducibility of SI decoding methods. These two aspects are critical in BCI research, where prior literature has highlighted that many studies suffer from incomplete methodological reporting or flawed evaluation procedures, making reproduction difficult. The inherent variability of brain signals further complicates the replication of results. Evaluating the reproducibility of SI decoding approaches is therefore essential to assess the true feasibility of SI as a viable BCI paradigm. *Approach.* To assess reproducibility, we selected two of the most widely used open-access SI datasets and attempted to reproduce four published decoding pipelines for each dataset. We followed each implementation step-by-step, documented missing or ambiguous information, detailed how we addressed it, and compared our decoding results to those originally reported. To assess replicability, we applied standard decoding pipelines across different time-frequency configurations to three open SI datasets and our own collected dataset. For context and validation, we conducted the same procedure on four publicly available and widely used motor imagery (MI) datasets. *Main Results.* All evaluated SI studies contained some form of missing methodological detail, and others did not include cross-validation procedures. Our reproduction attempts consistently yielded lower classification accuracies than originally reported, with discrepancies ranging from 2 to 39% ($\tilde{x} = 11.25 \pm 12.41\%$). In the replication analysis, we found no consistent time-frequency patterns across SI datasets. Furthermore, only 36% of SI participants achieved classification accuracies above statistical significance thresholds, compared to 91% of the participants in MI datasets. *Significance .* This is the first comprehensive assessment of both reproducibility and replicability in SI decoding. Our findings raise important concerns about the reliability of current SI research and suggest that the feasibility of SI as a practical BCI paradigm may have been overestimated.

1. Introduction

Brain-Computer Interfaces (BCIs) aim to build a direct communication channel with computers by decoding brain signals. In active BCI paradigms users perform an

instructed cognitive task while their brain activity is recorded [1]. Among different BCI paradigms, Speech Imagery (SI) has caught the attention of researchers as it represents an intuitive approach for BCI designs. An SI-BCI system requires the user to attempt speech without moving articulators or producing any sound. It is thought to be ideal to use a single word or sentence to control a BCI system [2]. For example, if the user imagines speaking direction commands this may be simpler than imagining moving limbs as with Motor Imagery (MI). Moreover, SI may be evoked endogenously, i.e., without the need of external stimulation as visual- or auditory-based BCIs [3]. SI-BCIs have the potential to vastly increase the possibilities for communication aid applications if a significant number of words can be accurately decoded [4, 5].

The current state-of-the-art for SI decoding looks promising, with a trend of new decoding attempts achieving significant classification accuracies [6]. Open-access electroencephalography (EEG) datasets [7, 8] have opened the possibility of trying different decoding techniques and, therefore, building a benchmark of how feasible it is to classify imagined speech units from EEG signals. However, SI decoding presents known intrinsic BCI challenges as well as paradigm-specific concerns that need to be addressed in order to unlock its full potential as a paradigm [6]. One important issue is replicability. Successful replication of other studies facilitates verification and validation of findings and highlights techniques that enable progress in the field.

Replicability is a major concern in BCI research, as it involves obtaining consistent results across studies that answer the same research question and is closely related to reproducibility, which means to obtain consistent results by analyzing the same input data [9]. BCI exists as a multidisciplinary junction that poses difficulties to knowledge sharing due to the domain-specific ways to share and report findings [10]. Broadly, there are two main sources on which to focus when addressing replicability in BCI: (1) procedures related to data collection that due to the complex nature of the brain result in high data variability between participants, experiment protocols and recording devices, and (2) the diversity of analysis/decoding techniques that include Machine Learning (ML) algorithms, usually with multiple parameters that affect their outputs [11].

The importance of a clear and complete description of all steps taken to decode EEG and how such decoding results are evaluated represents a key aspect for research reproducibility. Existing literature on BCI reproducibility suggests that a substantial proportion of published studies may yield results that cannot be reliably reproduced [12, 13]. Existing literature on SI decoding also suggests potential replicability issues, particularly due to the very low number of real-time decoding implementations. In our previous work, we found that only 6% of published SI decoding attempts were conducted in real-time settings [6]. In this study, we aim to investigate the reproducibility of SI decoding methods by focusing on three main aspects: (1) Completeness of methodological reporting, as prior research has shown that 32% of BCI studies lack key methodological details [14]; (2) Evaluation procedures, since improper evaluation methods—such as the absence of cross-validation—can result in inflated performance metrics. This is especially problematic in EEG studies, where data is typically

limited and models are prone to overfitting, as noted by Kinahan et al. [12]; (3) Reproducibility gap, by comparing our replication results with those reported in the original publications. For context, Menon et al. [13] failed to reproduce classification results in two EEG paradigms, with accuracy differences ranging from 10% to 20%.

We adopt a similar methodology to that of Menon et.al [13], attempting step-by-step reproduction of published SI decoding pipelines, documenting any missing or ambiguous information, and explaining how we addressed those gaps. We extend Menon's approach of a single reproduction attempt per dataset by adding 3 extra studies and widening our estimation on the reproducibility of SI. For this study, we selected two open-access datasets that, to the best of our knowledge, are the most widely cited in SI decoding. The first is the Kara One dataset [7], released in 2015, which contains EEG recordings of SI involving 4 words and 7 phonemes. The second is the dataset by Pressel Coretto et al. [8], released in 2017, which includes SI data of 6 words and 5 vowels. For each dataset, we attempted to replicate the original authors' approaches, as well as three additional decoding attempts that employed techniques currently trending in SI decoding [6, 15].

Our study also aims to address replicability by applying standard decoding pipelines and comparing the results from the two datasets with data from our SI experiments and a third open-access SI dataset [16].

In the next sections, we present the selection criteria and a summary of the decoding attempts, explaining the information that was missing for replication and the steps we took to overcome it. We also show how the results we obtained vary compared to the reported results in the literature. Finally, due to the high variability between the evaluated decoding results and our replication attempts, we investigate the apparent inconsistency of SI features and discover that informative features seem sparse across participants and datasets. To support our hypothesis, we run a replicability analysis on MI datasets by evaluating classification accuracies across a range of time-frequency configurations using different decoding pipelines and compare the performance consistency with results from SI datasets. This approach enables us to identify configurations that consistently capture informative features for MI datasets and assess whether any analogous consistency exists for SI-related features. Our study suggests that decoding SI from EEG signals may be more challenging than the current results reported from the literature suggest.

2. Study Selection

For the reproducibility study, we select two widely used open-access speech imagery datasets based on their considerable popularity and attempt to replicate the results reported by the original authors, as well as three additional decoding approaches. To select the additional approaches, we screened results from Google Scholar on "Speech Imagery decoding of Kara One OR Coretto dataset" and selected the first three from each that satisfied the inclusion criteria inclusion criteria: (1) Peer-reviewed published reports, (2) Record has been cited at least 10 times (3) Feature extraction methodologies

Table 1. Description of evaluated SI datasets.

	Kara One [7]	Coretto [8]	Nieto [16]	Our own
Classes	11 (4 words, 7 phonemes)	11 (6 words, 5 vowels)	4 words	2 words
Trials per class	13	50	50	25
Participants	14	15	10	16
Channels	64	6	64	128
Sampling rate (Hz)	1000	1024	1024	2048

belong to trends on SI decoding identified in [6, 15], including Common Spatial Patterns (CSP), Discrete Wavelet Decomposition (DWT) or Mel Frequency Cepstral Coefficients (MFCC) and the use of Convolutional Neural Networks (CNN).

Including the original authors' baseline models, our study systematically evaluates a total of eight distinct decoding approaches. This broad methodological scope provides a highly robust and representative assessment of the current state of reproducibility within the SI literature.

For the replicability study, we include a recently published SI dataset alongside a newly collected dataset of our own. The primary characteristics of all evaluated SI datasets are summarized in Table 1. We additionally incorporate four MI datasets. Because MI decoding has been successfully and consistently replicated using the Mother of All BCI Benchmarks (MOABB) framework [17], the inclusion of these MI datasets serves as a robust reliability test for our analysis. Therefore, this balanced four-versus-four comparison yields highly representative evidence regarding the replicability of these paradigms. Given the limited number of open-access SI datasets currently available in the literature [6], the evaluation of four distinct SI datasets provides a comprehensive and accurate assessment of the paradigm's current challenges.

Our newly collected dataset consists of 16 participants, representing the largest sample size among the SI datasets evaluated in this study. To ensure a balanced and comparable analysis across paradigms, considering that two of the MI datasets included recordings from over 50 participants [18, 19]. we limited our evaluation of the larger MI datasets to their first 16 participants. Furthermore, the single-task design utilizing the directional words 'left' and 'right' was chosen deliberately. These specific targets are standard in SI research and were explicitly selected to enable direct cross-dataset comparisons with the open-access Coretto and Nieto datasets, which also employ those words. The selected MI datasets were chosen based on their reproducible decoding results reported and being able to access them through MOABB. We summarize the evaluated MI datasets in Table 2. For a detailed description of each of these and the SI datasets, please refer to Appendix A.

Table 2. Description of evaluated MI datasets.

	Weibo [20]	Physionet [18]	Lee [19]	Schirrmeister [21]
Classes	7	4	2	4
Trials per class	80	23	100	120
Participants	10	109 (16 included)	54 (16 included)	14
Channels	64	64	62	128
Sampling rate (Hz)	200	512	1000	500

3. Methodology

We provide a summary of the decoding pipelines used in the Kara One and Coretto SI datasets, which we aim to replicate, in Tables 3 and 4, respectively. To simplify references throughout the report, we assign each decoding approach an identifier based on the initial of its dataset. We followed the step-by-step implementation of their pipelines using the following analysis tools: MNE Python library [22] was used to read and format the EEG data into epochs. MNE was used to apply filters and run independent component analysis (ICA) and downsample when required in the preprocessing steps; it was also used to apply common spatial patterns (CSP). Scikit-learn [23] was used to compute statistical features from the signals or from discrete wavelet transform (DWT) coefficients as well as fast Fourier transform (FFT) Coeficcients. DWT was implemented using the PyWaveletes library [24]. Support vector machine (SVM), linear discriminant analysis (LDA), and random forest (RF) were implemented from the scikit-learn library, and Convolutional Neural Networks (CNN) were implemented using the Keras library [25].

We identify any missing or ambiguous information while following their published description of the decoding pipelines implementations and list the steps we took to attempt reproduction in Tables 5 and 6. However, we also explored how assuming missing parameters from methods influences in final results by running a grid-search for approaches where parameters were assumed, namely SVM (C and $gamma$) parameters and the ICA removed components. For a detailed description of each of the decoding approaches, please refer to the supplementary document provided with this publication.

It is important to note that Kara One decoding approaches present results using distinct numbers of participants; this is due to faulty data that we discuss further in the dataset definition (see Appendix A.1). We further explored how including all participants' data influences overall scores.

Additionally, we describe the time-frequency analysis used to assess replicability on SI and MI datasets below.

Table 3. Overview of SI decoding pipelines from the Kara One dataset, with source papers listed in the first column.

Report	Preprocessing	Feature Extraction	Classifier	Analysis tool	Evaluation
KO1[7]	Included 8 participants Filter 1-50 Hz. Laplacian Filter. Segment 500 ms windows with 250 ms overlap. Group trials in a superclass.	Statistical and wave features. Rank features with Pearson correlation and select the top 5.	SVM (RBF Kernel) DBN	Matlab EEGlab	Between participant Leave One Out
KO2[26]	Included 11 participants Filter 1-50 Hz ICA. Segment 500 ms windows with 250 ms overlap.	MFCC (13 coefficients, 9 filters). PCA (95% variance explained).	Linear SVM	Matlab EEGlab	Within participant 5-fold cross-validation
KO3[27]	Included 8 participants Notch filter at 60 Hz Segment 500 ms windows without overlap.	Compute FFT for each segment. Compute covariance matrices of the FFT coefficients.	CNN	not mentioned	Between participant split 50% training (25% val) - 50% testing
KO4[28]	Included 13 participants. No preprocessing	CSP	Ensemble Stacking Learning	Matlab	Within participant 10-fold cross-validation

Table 4. Overview of SI decoding pipelines from the Coretto dataset, with source papers listed in the first column.

Report	Preprocessing	Feature Extraction	Classifier	Analysis tool	Evaluation
CT1[8]	Downsample to 128 Hz. Filter (2 - 40 Hz).	DWT (D2 - D5, A5) Relative wavelet energy (RWE) from each level.	RF: 5 initial random features, 200 trees	Matlab	Within participant 10-fold cross-validation
CT2[29]	Downsample to 256 Hz.	DWT (D1 - D4, A5) Filter using CSP (4 components) for each decomposition level. Use the variance of each CSP component	SVM with RBF Kernel	Matlab fvtool dwtfilterbank	Within participant 5-fold cross-validation
CT3[30]	Downsample to 128 Hz Filter (2-40 Hz). ICA. Scaling.	Use Raw signal	CNN	Python Sckit-learn	Within participant 5-fold cross-validation
CT4[31]	Filter (2-40 Hz). ICA.	Use Raw signal	CNN	Matlab	Within participant 80% training - 10% testing - 10% validation

Table 5. Description of missing information in SI decoding pipelines applied to the Kara One dataset.

Report	Missing Information	Steps Taken
KO1[7]	Feature extraction: It is not clear how to get derivatives from the obtained features. If the derivative features are the same length as the feature vector, then 32×17 segments = 816 features per channel. However, 1197 features are mentioned.	Feature Extraction: We did not use derivative features due to uncertainty in their calculation. The number of features per channel we obtained is 221.
KO2[26]	Preprocessing: It is not mentioned what ICA algorithm was used, how the noise components were selected, and how many were removed. Classification: SVM regularization parameter C not specified	Preprocessing: We opted for the Picard ICA algorithm; we excluded one component associated with eye blinks, characterized by activity confined to frontal electrodes and the absence of 10–20 Hz power in its spectral profile. A second component was removed if lateral eye movements were evident, indicated by spatial lateralization of the component's topography. Classification: SVM regularisation parameter C was set to the default value from Scikit-learn
KO3[27]	Classification: The kernel sizes of the Convolution layers specification is ambiguous; the authors conclude that a kernel size equal to the input shape performed better. However, this changes CNN behaviour. Ambiguous data splitting procedures: authors first state a 50-50 split but then a cross-validation procedure is mentioned without specifying the number of folds. No model training specifications given.	Classification: We define the Convolutional layers' kernel sizes of (7x7) and (3x3). We applied a repeated 2-fold cross-validation. We train the model for 200 epochs and a batch size of 64.
KO4[28]	Feature Extraction: CSP features were extracted from each component and how many components were considered are not mentioned.	Feature Extraction: This approach is not replicable as critical information regarding CSP is missing.

Table 6. Description of missing information in SI decoding pipelines applied to Coretto's dataset.

Report	Missing Information	Steps Taken
CT1[8]	Preprocessing: It is not mentioned what ICA algorithm was used, how they selected the noise components, and how many were removed.	Preprocessing: Same as with the steps taken in KO2 (see Table 5)
CT2[29]	Classification: SVM regularization parameter C and Kernel Coeficient *gamma* not specified	Classification: SVM regularization parameter C set to 1 and Kernel Coeficient to $1/n$ where n is the number of features.
CT3[30]	Preprocessing: Not clear how to select noise components from ICA and how many were removed Feature Extraction: Not explicitly mentioning the raw signal as features or the shape of the input data for the CNN	Preprocessing: Same as with the steps taken in KO2 (see Table 5) Feature Extraction: Used the raw EEG signal (6 x 496) as input data for the CNN
CT4[31]	Preprocessing: Not clear how to select noise components from ICA and how many were removed Feature Extraction: Not explicitly mentioning the raw signal as feature or the shape of the input data of the CNN Classification: The report specifies 40 filters in the initial convolution layers, however, 20 filters are set in the source code	Preprocessing: Same as with the steps taken in KO2 Table 5 Feature Extraction: Used the raw EEG signal (6 x 496) as input data for the CNN Classification: We used 20 filters as specified in the source code, considering a smaller number of trainable parameters

3.1. Features Comparison Between Speech Imagery and Motor Imagery

To attempt a comparison of features from SI and MI paradigms, we use a Riemannian geometry-based pipeline that has demonstrated success in extracting MI features [17], mainly composed on a tangent space (TS) projection of regularised covariance matrices using Ledoit-Wolf estimation [32] and a logistic regressor (LR), covariance estimation and TS transformation was implemented using the PyRiemann library [33]. We refer to it as the TS+LR pipeline. Further details of this pipeline can be found in Appendix B.0.2.

We evaluated this pipeline using a binary classification approach for each motor imagery (MI) and speech imagery (SI) dataset. This analysis aimed to identify consistent features across participants within each dataset, regardless of the specific unit being classified. For all MI datasets, we classified left versus right hand imagery. For the SI datasets, we classified the imagined words 'left' versus 'right' (in the Coretto, Nieto, and our datasets) and 'pot' versus 'knew' (in the Kara One dataset). Additionally, we tested an extra pair of classes for the Coretto (/u/ vs. /a/), Nieto ('up' vs. 'down'), and Kara One (/iy/ vs. /m/) datasets to determine whether classification performance varied significantly based on the chosen phonetic or linguistic targets. We applied this pipeline to different frequency and time window settings and analysed the distribution of mean classification accuracies. The evaluation protocol was standardized to a within-participant basis to help identify common patterns across datasets. A 10-fold cross-validation procedure was used to enable meaningful comparisons of pipeline performance

Specifically, for each dataset, we tested all combinations of 16 frequency bands (from 0 to 130 Hz, each 10 Hz wide, with an 8 Hz overlap) and 7 time windows (ranging from 0 to 3.5 seconds, each 1.5 seconds long, with a 0.3-second step). For each combination, we calculated the median accuracy across participants and evaluated the results as a matrix of accuracies. This allowed us to visually identify any consistent configurations for both paradigms. To check for statistically significant median accuracy across participants on each dataset, we applied a Bonferroni-corrected one-sample t-test.

Thereafter, to check for the individual participant's highest-performing configurations, we check for significant median accuracies above the threshold of the binomial distribution of accuracy considering the number of trials [34] with 99% confidence ($\alpha = 1\%$) and evaluated the distribution across configurations for participants with statistically significant accuracies only.

4. Results

4.1. Reproduction of existing literature

The decoding accuracies obtained from our replication attempts were generally lower than those originally reported in the majority of the studies. To evaluate performance, we compared the mean reported accuracy across participants from the original studies with the accuracies obtained in our replication attempts.

For the Kara One dataset, our results were on average 22.4% lower than reported, with the exception of one condition in the KO1 approach, where our results were close to chance level. We were unable to replicate the KO4 approach due to missing essential information regarding the features extracted through CSP in the original publication. The results for Kara One are summarized in Table 7 as average accuracies, including our replication efforts and those originally reported in the relevant literature. The results on the influence of including data from the faulty participants show lower mean accuracies when including all participants, but the difference is not statistically significant. Results

Table 7. Summary of results (mean ± standard deviation of accuracy) obtained with the different decoding approaches for the Kara One dataset and the corresponding results from our replication attempts and the p-value from the Welch difference t-test. The KO4 report did not include the standard deviation of their results.

Decoding Approach	Condition	Results Reported in Literature (%)	Replication Results (%)	Welch T-test P-value
KO1	C/V with SVM	56.9 ± 14.1	49.3 ± 5.8	0.18
	C/V with DBN	88.6 ± 4.9	49.5 ± 1.2	< 0.01
	/uw/ with SVM	56.8 ± 13.6	59.1 ± 6.5	0.68
	/uw/ with DBN	80.1 ± 2.7	59.8 ± 3.1	< 0.01
KO2		20.45 ± 5.7	15.5 ± 6.5	0.09
KO3		31.6 ± 0.4	15.6 ± 0.08	< 0.01

Table 8. Summary of results (mean ± standard deviation of accuracy) obtained with the different decoding approaches for the Coretto dataset and the corresponding results from our replication attempts and the p-value from the Welch difference t-test. No standard deviation value was reported in the CT2 results.

Decoding Approach	Condition	Results Reported in Literature (%)	Replication Results (%)	Welch T-test P-value
CT1	vowels	22.3 ± 1.8	19.1 ± 8.1	0.29
	words	18.5 ± 1.47	15.9 ± 7.2	0.18
CT2	Ab–Iz	81.1	51.7 ± 10.7	
	At–Iz	80.3	48.4 ± 11.1	
CT3		35.2 ± 3.9	19.3 ± 7.1	<0.01
CT4		22.3 ± 1.81	19.1 ± 8.1	0.52

of this analysis are shown in table C3.

For the Coretto dataset, our replication results were on average 14.2% lower than those reported in the original studies. A detailed comparison of these results is provided in Table 8.

The influence of parameter selection showed that grid-search led to higher mean accuracies; however, the results were not statistically significant. Results on grid-search of SVM parameters are presented in table C1, and the results of grid-search of the number of ICA components is showed in table C2.

4.2. Replication results: Time-Frequency Comparison of SI and MI

Due to the large discrepancies observed in decoding performance, we further explored the nature of SI features by performing a grid search over configurations that yielded

statistically significant decoding accuracies. In the following subsection, we present results from a well-established MI decoding pipeline using a variety of time-frequency configurations on four different MI datasets and four SI datasets. As shown below, the results indicate that, unlike MI, EEG features associated with SI exhibit low consistency across participants and datasets for both tested class conditions.

We present the median accuracy of the TS+LR pipeline across participants for each time-frequency configuration in Figure 1. The results show that consistently high decoding performance was achieved with the MI datasets in the 0–20 Hz frequency range across most time windows, which aligns with the established literature on MI-based BCIs. Using a Bonferroni-corrected one-sample t-test, we identified several statistically significant accuracies for the MI datasets (Weibo: 6, Physio: 7, Lee: 7, Schirrmeister: 18).In contrast, the SI datasets do not exhibit any clear or consistent regions of high performance across frequency or time. Furthermore, no median accuracy reached statistical significance ($p < 0.05$, Bonferroni-corrected one-sample t-test) for either of the class pair conditions. Figure C1 also illustrates similar results for the extra pair of classes for Coretto (/u/ vs. /a/), Nieto ('up' vs. 'down'), and Kara One (/iy/ vs. /m/). Although the Kara One dataset displays several regions that appear to contain discriminative features for SI (e.g., 70–100 Hz for windows starting at 1.2 s), these patterns are not observed in the other SI datasets, and these apparent peaks do not deviate significantly from the dataset's overall accuracy distribution. This raises the question of whether the TS+LR pipeline generalizes well to SI features. To address this, we tested two other pipelines, one based on Common Spatial Patterns (CSP) and the other with a Convolutional Neural Network CNN model. All pipelines showed similar patterns: informative regions for MI were consistently found in lower frequencies, while SI-related features remained sparse and inconsistent (see Appendix C.2).

While considering the highest individual accuracies for each participant across the two paradigms, we found that on average, only 36.1% of participants in SI datasets achieved statistically significant decoding performance at a 99% confidence level ($\alpha = 0.01$), compared to 91.2% of participants in MI datasets. Figures 2.d and 2.e illustrate the proportion of participants in each dataset with significant results for both paradigms. Similar results are obtained with the extra pair of classes tested, while there was more participants with statistically significant accuracies in Coretto dataset (/u/ vs. /a/), the number of good performers decreased for Kara One (/iy/ vs. /m/) and Nieto ('up' vs. 'down') as shwon in Figure C2.

The distribution of the highest decoding scores across configurations shows a clear pattern for MI participants. Most high-performing configurations are concentrated in frequency bands between 0 and 25 Hz and time windows starting between 0 and 0.9 s. In contrast, SI participants do not show any consistent time-frequency region where high-decoding performance can be achieved. Figure 2.a displays the top-scoring configuration for each participant in both paradigms. We applied DBSCAN clustering to identify groups of at least five nearby samples within a 0.5 distance range. Clusters were found for participants in the MI datasets, but none were identified for SI. Figure 2.b shows

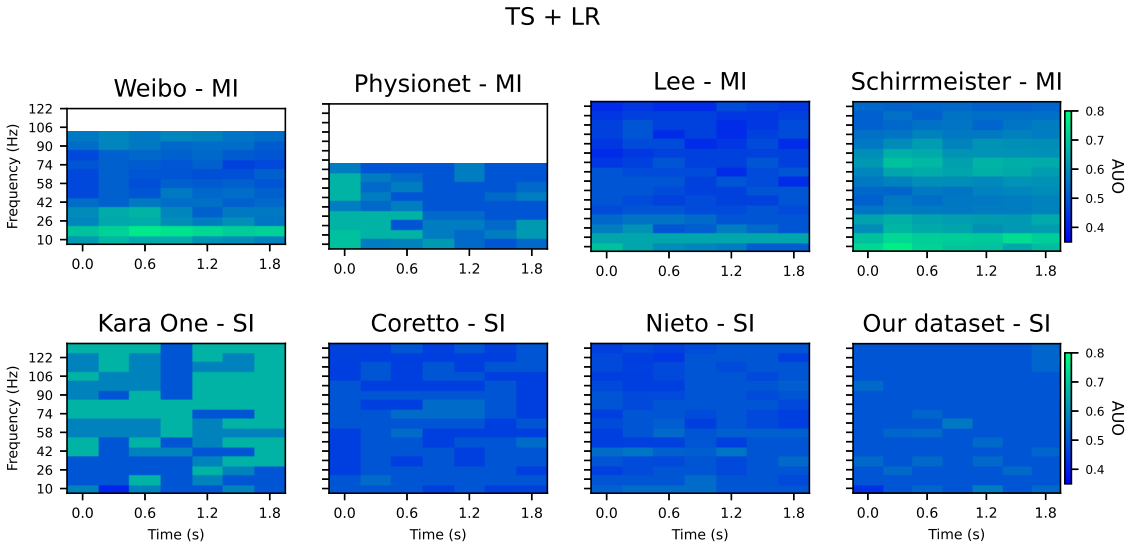


Figure 1. Heatmaps of median classification accuracies across participants for various time and frequency windows using the Tangent Space and Logistic Regressor (TS+LR) pipeline. The top row displays results for four Motor Imagery (MI) datasets, while the bottom row displays results for four Speech Imagery (SI) datasets. The evaluated tasks were left versus right hand imagery for all MI datasets, and the imagined words 'left' versus 'right' (Coretto, Nieto, and our dataset) or 'pot' versus 'knew' (Kara One) for the SI datasets. Evaluation was performed using 10-fold cross-validation. The results demonstrate that MI datasets exhibit consistently high, statistically significant performance in low-frequency ranges (0–20 Hz), whereas SI datasets lack any consistent regions of significant accuracy ($p < 0.05$)

a boxplot of frequency distributions for each participant, colour-coded by dataset. MI participants exhibit a tightly grouped distribution centred around 14 Hz, with only a few outliers at higher frequencies. In particular, all participants from the Lee dataset show optimal performance in the same frequency range. In contrast, the frequency distribution for SI participants is much more scattered, with a standard deviation of 27 Hz in contrast with the 16 Hz for MI participants. Figure 2.c presents the distribution of optimal time windows. For MI, most participants show peak decoding performance in time windows starting between 0 and 0.9 s with a median at 0.3 $s \pm 0.4$. SI participants again show broader variability with a median starting value at 0.9 $s \pm .6$.

Overall accuracies in SI are lower (t-test p-val > 0.05), with a median accuracy of 83.3%, compared to 76% for SI datasets. This comparison is shown in Figure 2.f, where each point represents the highest achieved accuracy with a colour-coded label for indicating the corresponding dataset. Notably, participants from the Kara One dataset with statistically significant results achieved the highest values among SI participants.

Interestingly, when we lower the statistical threshold to 95% confidence, the proportion of SI participants for whom decoding accuracy is above chance performance increases by 42.4% (from 13 to 30 participants). In contrast, the increase for MI participants is just 7.2% (from 51 to 55). However, even with this relaxed threshold,

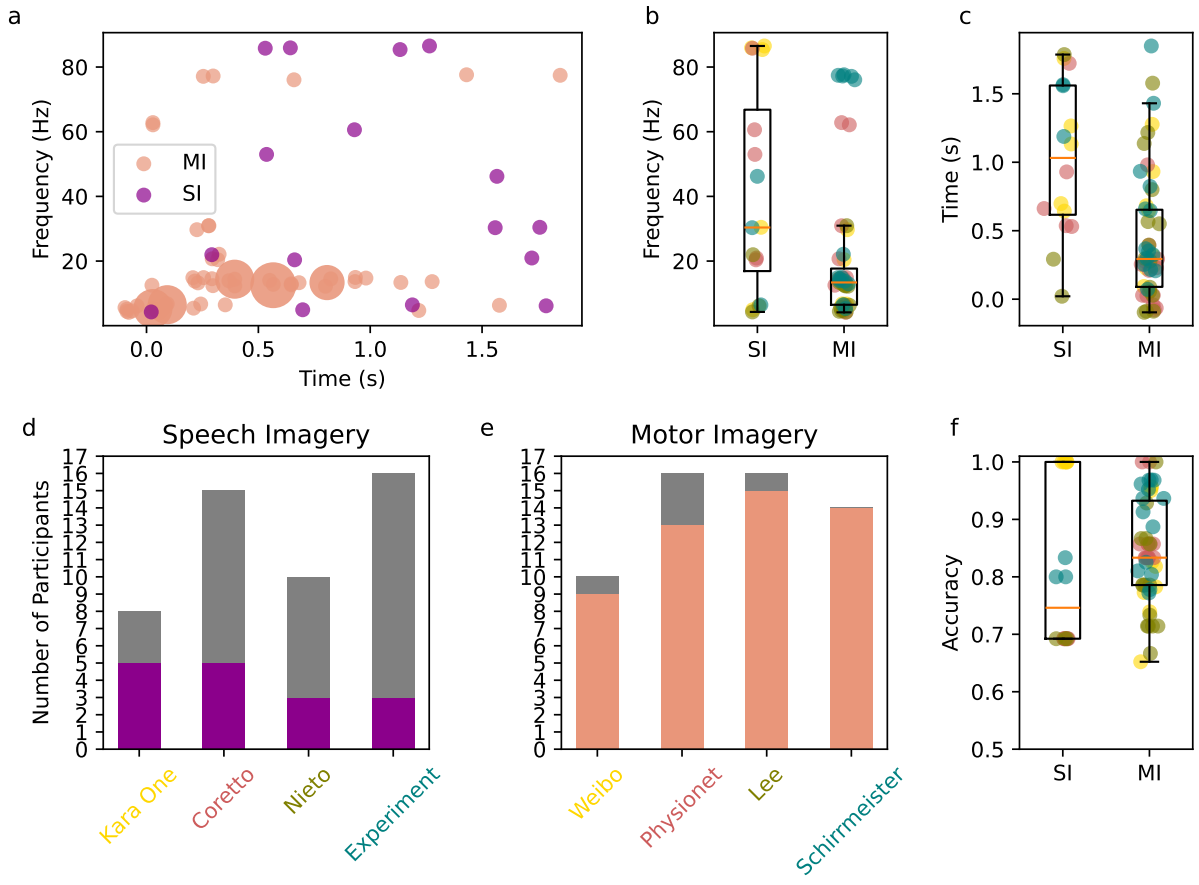


Figure 2. Distribution of the highest pair-wise decoding accuracies using the Tangent Space and Logistic Regressor (TS+LR) pipeline across different time-frequency configurations for the Motor Imagery (MI) and Speech Imagery (SI) datasets. The evaluated class pairs were left versus right hand imagery for all MI datasets. For the SI datasets, the tasks consisted of the imagined words 'left' versus 'right' (Coretto, Nieto, and our dataset) and 'pot' versus 'knew' (Kara One). Only participants with statistically significant performance at the 99% confidence level ($alpha = 0.01$) are included. *a.* Scatter plot of individual participant results across time-frequency space. Clusters of at least five points are found using DBSCAN. The size of the circles indicates the number of participants in each cluster; only MI participants show clustering, primarily within the 0–30 Hz range. *b.* Frequency distribution of peak decoding accuracies, with each point representing a participant, colour-coded by dataset. *c.* Distribution of best accuracies across time, where the Y-axis indicates the starting time of each 1.5 s decoding window. *d.*The proportion of participants in each SI dataset achieving significant accuracy, with non-grey bar segments indicating successful cases. *e.*Same as (d) but for MI datasets; all participants from the Schirrmeister dataset reached significant performance. *f* Overall accuracy distributions for SI and MI participants, with individual scores color-coded by dataset

we still do not observe consistent time-frequency patterns among SI participants. The extended results for the 95% confidence analysis are included in the Appendix C.1.

Additionally, we noticed an influence of the number of trials used for classification.

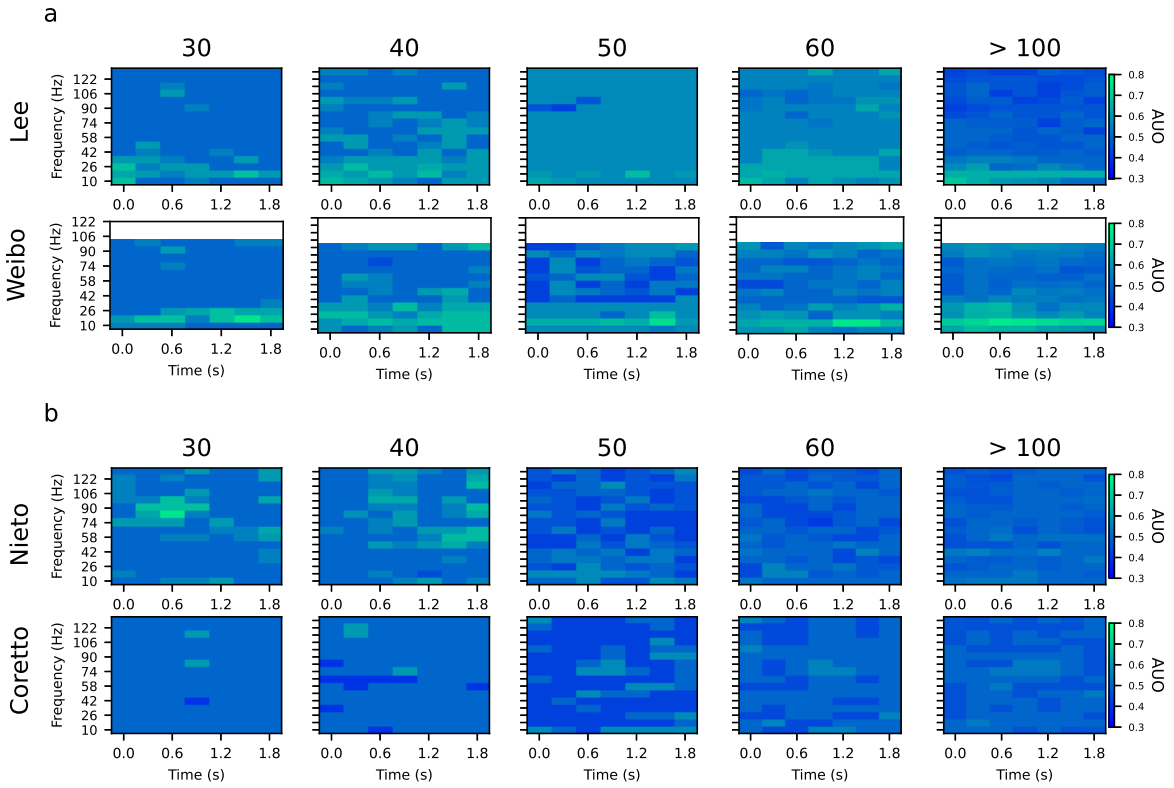


Figure 3. Effect of trial count on median classification accuracies across time and frequency windows using the Tangent Space and Logistic Regressor (TS+LR) pipeline. Heatmaps display cross-validated accuracy as the number of training trials sequentially increases (30, 40, 50, 60, and >100 trials) for (a) two Motor Imagery (MI) datasets (Lee and Weibo) and (b) two Speech Imagery (SI) datasets (Nieto and Coretto). The results indicate that while MI datasets develop consistent, robust discriminative patterns starting at 30 trials, SI datasets fail to yield reliable patterns, with any early high-accuracy configurations diminishing as more trials are added.

In Figure 3, we compare the decoding performance of Nieto and Coretto SI datasets with Lee and Weibo MI datasets at various consecutive trial counts used for the cross-validation evaluation. The MI datasets showed consistent patterns starting from 30 trials, and such patterns persist as more trials are added. For the Nieto SI dataset, using the first 30 trials led to seemingly consistent high scores in high-frequency configurations. However, this consistency diminished as more trials were added. Coretto's dataset decoding results do not present patterns of significantly high accuracies.

5. Discussion

The accuracy scores from our reproduction attempts were generally lower than those reported in the original studies, with discrepancies ranging from 8.1% to 36% for Kara One approaches and from 2.6% to 31.9% for Coretto approaches. These findings reveal

a significant and previously overlooked issue of reproducibility in SI decoding. Our analysis identified multiple factors that likely hinder reproducibility, consistent with the challenges highlighted in the BCI reproducibility literature discussed in Section 1 [12, 13]. We found that each of the evaluated studies lacked methodological details. Some missing elements – such as the number of ICA components removed during preprocessing or the classifier parameters (e.g., the SVM *gamma* value) – may be inferred based on common defaults or conventions in software tools. However, other omissions involved parameters with wide possible variations that could significantly affect results, such as the kernel sizes in deep learning convolutional layers. In study KO4, critical information regarding feature extraction was missing, and no source code or supplementary materials were provided with the publication. Furthermore, our attempts to contact the original authors were unsuccessful, which ultimately prevented us from attempting to reproduce this pipeline.‘ Importantly, our grid-search analysis showed no significant differences compared to the assumed parameters. This suggests that factors such as the SVM C and γ values or the number of ICA components may not substantially affect the reproducibility of decoding attempts. Instead, issues related to model evaluation and data leakage are likely the main causes of overestimated results.

We also found that the CT4 and KO3 approaches did not include cross-validation in their evaluation. As discussed by Kinahan et al. [12], this is particularly problematic given the small number of trials and the inherently high variability of EEG signals. Without appropriate validation methods, performance can be overestimated. Furthermore, only the approach CT4 made their code available, highlighting a significant barrier to reproducibility across the literature. It's important to note that we used different analysis tools than those reported in the original studies, and prior research suggests that software choices can influence decoding performance [35]. However, we believe that missing methodological details or flawed evaluation procedures likely had a greater impact on reproducibility than the choice of software itself.

We also evaluated replicability in SI decoding by testing the TS+LR pipeline, confirming its consistent performance across four MI datasets in line with MOABB findings [17]. While MI decoding demonstrated strong replicability across independent datasets, our analysis revealed highly variable outcomes for SI. Importantly, the highest decoding performances in MI were driven by consistent, shared frequency configurations across participants. These patterns were evident even in small cohorts, such as the 10-participant Weibo dataset. In contrast, no such feature consistency was observed across the SI datasets. Even with a larger cohort of 16 participants in our newly collected dataset, no shared representational patterns emerged. Furthermore, this lack of consistency persisted even when evaluating alternative speech units, confirming the absence of generalized SI features. Although several high-accuracy configurations were observed in the Kara One dataset, these did not deviate significantly from the overall distribution; given the low number of trials, these peaks could reflect overfitting. Ultimately, this demonstrates a fundamental challenge of the SI paradigm: the participant-dependent feature variability.

Our comparison of time-frequency decoding performance using the TS+LR pipeline further supports this point. Among MI participants, significant results clustered consistently in the 0–20 Hz range, reflecting informative features commonly attributed to event-related synchronization and desynchronization in motor imagery. In contrast, fewer SI participants achieved above-chance results, and their peak performance occurred across a wide range of time and frequency configurations, with no identifiable clusters in the time-frequency space. Even when the confidence threshold was relaxed to include more participants, the spatial and spectral distribution of SI features remained scattered and inconsistent.

We also observed an interesting effect related to the number of trials used for model training. It is well-documented that the quantity of trials can influence EEG signal quality, and that extended recording sessions may introduce fatigue and reduce data reliability [36]. In our experiments, MI-related patterns were evident even with a small number of trials (as few as 30) and became more robust as additional trials were included. In contrast, the SI datasets exhibited clustered patterns when using only the first 30 trials, but these patterns gradually faded as more trials were added. This suggests that SI-related features may deteriorate over time during extended recording sessions. Possible explanations for this effect are discussed in the following section.

6. Conclusion

We argue that the current state of speech imagery as a paradigm in EEG-based BCIs may be misleading due to non-reproducible decoding approaches with inflated performance scores. To advance SI as a viable paradigm, the field must align more strongly with reproducibility standards. Reproducibility is essential for enabling independent verification of findings and supporting progress toward real-time applications.

Summary Box: Key Recommendations for Reproducibility in SI research
Our findings highlight three key aspects that should be indispensable in SI decoding research:

- **Adhering to ML reporting frameworks** such as CRISP-DM [37] to ensure that no methodological details are missing.
- **Applying rigorous evaluation practices:** decoding accuracies marginally above chance are often reported as meaningful without proper statistical testing, particularly in small datasets. Researchers should use robust validation strategies, including cross-validation for participant-dependent models and leave-one-participant-out methods for participant-independent approaches.
- **Sharing code publicly** to allow verification of methods across different datasets.

SI-specific challenges may also stem from variability in the internal experience

of SI across individuals [38]. It is therefore critical that participants are thoroughly instructed and verified to engage in the intended cognitive task, whether auditory, articulatory, or otherwise. Exploring different SI modalities—such as internal monologue versus imagined dialogue—could help clarify the construct. Furthermore, participant screening with validated psychological instruments, such as the Varieties of Inner Speech Questionnaire (VISQ-R) [39], could enhance data quality by identifying individuals with distinct inner speech profiles. Our findings also suggest that SI-related signals may degrade over time, possibly due to participant fatigue during long sessions or to decoding pipelines that overfit on small datasets and fail to generalize to larger ones. We observed that decoding performance declined as more trials were used to train the classifiers, implying that prolonged sessions might reduce signal quality for SI more than for MI. Finally, while our results showed that significant SI decoding accuracies were achieved by only 36% of participants, a central challenge for future SI research lies in identifying these variant informative frequency ranges. We suggest that the field should adopt participant-dependent approaches, focusing on methods to detect and track informative frequencies and on studying how these may vary across trials or experimental sessions.

In summary, our findings highlight two distinct challenges in SI decoding: methodological issues within current literature and intrinsic challenges of the paradigm itself. Regarding the literature, our reproduction attempts consistently yielded much lower classification accuracies than originally reported, often approaching chance level. Regarding the intrinsic nature of SI, our replicability analysis revealed that informative features are highly participant-dependent, and performance tends to degrade as more data is introduced. These combined findings raise concerns about the reliability of SI-related EEG signals for practical BCI applications. They suggest that the proportion of participants whose SI signals can be successfully decoded may be below 50%, likely due to the inherent difficulty in consistently evoking or detecting SI-related neural activity. Importantly, we are not suggesting that the Speech Imagery paradigm is inherently unfeasible or without merit. Rather, our evidence firmly indicates that prior studies have significantly overestimated its current practical readiness and feasibility.

Appendix A. Evaluated datasets

The following speech imagery and motor imagery datasets were used in this investigation. Tables 1 and 2 summarize the datasets.

Appendix A.1. SI Dataset: Kara One

Data was collected from 12 participants, using a 64-channel SynAmps RT to record the signals. 7 phonemic/syllabic prompts (/iy/, /uw/, /piy/, /tiy/, /uw/, /piy/, /tiy/, /m/, /n/) and 4 phonetically-similar pairs (pat, pot, knew and gnaw) were prompted to the participants. Twelve trials were recorded for each one of these SI units. The

experiment consisted on a 5-second rest state, a stimulus state where the prompt text appeared on the screen along with its associated auditory utterance played over speakers, this was followed by a 2-second preparation period for a 5-second SI state and a final speaking state. Further details can be found in the author's publication [7]. This dataset presents a notable challenge for any decoding effort, as several recordings were affected by unstable sensor connections. According to the original authors, recordings from 4 out of 12 participants were discarded due to poor signal quality; however, participant identifiers corresponding to these exclusions were not specified. Compounding this issue, the dataset includes 14 folders, each containing data from a different participant, leading to a clear inconsistency. We contacted the original authors for clarification, but they were unable to resolve the discrepancy due to information being lost over time. Two out of three peer-reviewed decoding studies appear to overlook this issue and include the corrupted data. These faulty signals are distinguishable through visual inspection, characterised by rapid semi-regular voltage drifts from a non-physiological source. To mitigate this, we excluded participants exhibiting more than 30% corrupted trials, ultimately retaining nine participants for our analysis. Full details on the discarded participants and trials are provided in the code included with this publication.

Appendix A.2. SI Dataset: Coretto

SI signals were recorded from 15 participants, using the 10-20 channel position system, they placed 6 active electrodes in locations F3, F4, C3, C4, P3 and P4, using an analogue amplifier, which performs an band-pass filter for each channel at a lower and upper cutoff frequencies of 0.3 and 35 Hz, respectively. The 5 Spanish vowels /a/, /e/, /i/, /o/, and /u/ along with the corresponding translation from the command words "up", "down", "right", "left", "forward" and "backwards" were used as prompts. 50 trials per word/vowel were recorded for each participant. The experiment consisted of a 2-second preparation state, a 2-second stimulus state where they textually presented the word to imagine, a 4-second imagine interval and a 4-second rest state. Further details can be found in the author's publication [8].

Appendix A.3. SI Dataset: Nieto

EEG signals were acquired using a BioSemi ActiveTwo with 128 active EEG channels. The 4 directional words in Spanish, equivalent to "up", "down", "right" and "left" were used to prompt SI. For each participant and each word, 50 trials were recorded. The experiment consisted of a 0.5-second preparation state, a stimulus state where participants were shown an arrow pointing towards the direction intended for imagery, followed by a cue to instruct the imagery interval for 2.5 seconds and final relaxation state with a time-variant interval. Further details can be found in the author's publication [16]. We considered the 64 channels equivalent to a 64-electrode setup in order to reduce the dimensionality.

Appendix A.4. SI Dataset: Our dataset

Sixteen right-handed able-bodied participants (8 female) between the ages of 20 and 35 ($\mu = 25.65, \sigma = 8.3$) were recruited from the student population of the University of Essex. Participants received a compensation voucher worth £10 (GBP) for their time. All volunteers read, understood and signed the consent form based on the recommendations of the Ethical Committee of the University of Essex in January 2023 (Reference Number ETH2223-0220). EEG was recorded using a 64-channel Biosemi Active-Two system. Electrode placement was done via the international 10-20 system, plus one electrode close to the pterion after each eyebrow for electrooculography (EOG) and one electrode behind each ear on the mastoids for electromyography (EMG) recording. Data were recorded at a sampling rate of 2048 Hz, unaffected by hardware cut-off. Participants were seated in a comfortable chair facing a 52-inch screen. A graphical user interface developed with PsychToolbox 9.0 [40] in Matlab R2022 was used to display the prompts over a plain grey background. We used a stimulus-masking approach where we first showed the imagery prompt and then had a visual cue presented as a circle in the middle of the screen that remained for 300 ms, creating a flash-like effect. Participants were asked to perform the speech imagery of the words "left" and "right" as soon as they saw the cue stimulus. 25 trials per class for each participant were recorded. We first presented a fixation cross for 2 seconds followed by the imagery prompt for 6 seconds and a time-variant (1.5–2 s) fixation cross before the cue. We cued our participants with the described flash stimulus and proceeded to leave a plain screen for 5 seconds until the "relax" prompt was shown.

Appendix A.5. MI Dataset: Weibo

The MI trials were recorded from 10 participants, using 64 electrodes acquired with a Neuroscan SynAmps2 system that applies a band-pass filter in the range 0.5–100 Hz. Each trial started with a white circle for 2 seconds, followed by a preparation stage of marked by a red circle in the screen for 1 second. Then the prompting or the character ("Left Hand", "Left Hand and Right Foot", et al) was presented on the screen for 4 seconds, during which the participants were asked to perform kinesthetic motor imagery rather than a visual type of imagery while avoiding any muscle movement. After 7 seconds, "Rest" was presented for 1 second before next trial. Further details of the experiment can be found at [20].

Appendix A.6. MI Dataset: Physionet

EEG data were recorded from 109 participants using a 64-channel setup with a Brainproducts amplifier, recorded with BCI2000 system. In the experiment, the participants were shown a mark on the left or right side of the screen to cue the left or right fist opening and closing imagery, they also recorded foot movement imagery. 25 trials per class were recorded. Further details can be found in the authors' publication

[18]. For practical reasons and to match the reduced number of participants in SI datasets, only EEG data from the first 16 participants were included in our analysis.

Appendix A.7. MI Dataset: Lee

EEG signals were collected with 62 Ag/AgCl electrodes from 54 participants with a BrainAmp amplifier. The experiment procedures started with 3 seconds of a black fixation cross as a preparation stage. Then the participant performed the imagery task of grasping with the appropriate hand for 4 seconds when the right or left arrow appeared as a visual cue. After each task, the screen remained blank for 6 s ($\pm$ 1.5 s). The experiment consisted of training and test phases, each consisting of 100 trials with balanced right- and left-hand imagery tasks. More details about the experiment can be found in the authors' publication [19]. For practical reasons and to match the reduced number of participants in SI datasets, only EEG data from the first 16 participants were included in our analysis.

Appendix A.8. MI Dataset: Schirrmeister

The MI data was recorded from a 128-electrode headset from 14 participants. The considered movement classes were left hand, right hand, both feet and rest. 260 trials of each class and rest were recorded with BCI2000 system. The experiment consisted of a grey arrow pointing either up (for the relax condition), down (repetitively clenching their toes), left (finger tapping with left hand) or right (finger tapping with right hand). Further details from the experiment can be found in the authors' publication [21]. We considered the 64 channels equivalent to a 64-electrode setup in order to reduce the dimensionality.

Appendix B. Decoding pipelines for time-frequency testing

We evaluated three decoding pipelines with different feature extraction strategies to assess whether any particular type of representation is better suited for SI classification. Specifically, we tested: (1) a spatial filtering approach using CSP combined with Linear Discriminant Analysis (LDA); (2) a Riemannian geometry-based pipeline employing Tangent Space (TS) projection followed by a Linear Regressor (LR); and (3) a deep learning model based on CNNs. We detail each of these approaches below.

Appendix B.0.1. CSP + LDA CSP is a well-established spatial filtering technique commonly used in BCI research. It derives spatial filters that maximize variance differences between two classes, enhancing discriminative information in the data [41]. We selected the first two and last two spatial components obtained from the CSP projection and computed the mean power of each as feature inputs. These features were then classified using single value decomposition LDA.

Appendix B.0.2. TS + LR Riemannian geometry-based methods have demonstrated strong performance and robustness for EEG decoding tasks [17]. In this approach, each trial is represented by a covariance matrix, which belongs to the space of Symmetric Positive Definite (SPD) matrices. When treated as a point, SPD matrices lay on a Riemannian space [42]. To enable the use of standard machine learning algorithms, these matrices are projected to a Euclidean space via a logarithmic mapping known as the tangent space projection. Once mapped, a linear regressor was employed for classification.

Appendix B.0.3. CNN CNNs have been widely applied to EEG decoding problems in recent years [6]. We implemented the CNN architecture proposed by Wimpff et al. [43], which includes five convolutional layers and an attention mechanism designed to amplify signals from informative EEG channels. Prior to input into the network, EEG signals were downsampled to 250 Hz to reduce computational complexity. We trained the model, setting a batch size of 32 samples and 160 epochs.

Appendix C. Extra Results

Appendix C.1. Time-Frequency Decoding results of TS+LR on SI and MI

To determine whether the choice of speech unit influenced the consistency of our SI results, we evaluated an alternative set of class pairs. Figure C2 compares the distributions of MI and SI datasets using these alternative pairs: /u/ versus /a/ for Coretto, 'up' versus 'down' for Nieto, and /iy/ versus /m/ for Kara One. Consistent with our primary analysis, no reliable time-frequency region emerged across participants in these datasets, with no configurations achieving statistical significance (p ¡ 0.05, Bonferroni-corrected one-sample t-test). Furthermore, when examining the highest individual accuracies, while the number of participants with statistically significant performance increased in the Coretto dataset, the number of successful performers actually decreased for Kara One and Nieto. As observed with the initial class pair comparisons, the peak accuracies of the best performers remained scattered across widely different time-frequency configurations as seen in Figure C1, further reinforcing the conclusion that SI features are highly participant-dependent.

If we evaluate classification accuracies using a binomial distribution threshold based on the number of trials [34] at a relaxed 95% confidence level ($\alpha = 0.05$) instead of 99% ($\alpha = 0.01$), the proportion of participants with significant results increases by 42.4% for SI, but only by 7.2% for MI. However, even with this more inclusive 95% threshold, no clear consistency in SI features emerges. This dramatic increase in "successful" SI participants at the lower statistical threshold indicates that their decoding accuracies are marginal—hovering just above the upper limit of chance—rather than demonstrating robust, highly significant class separability. Figure C3 shows the time-frequency analysis results for decoding accuracies that reached this 95% significance threshold.

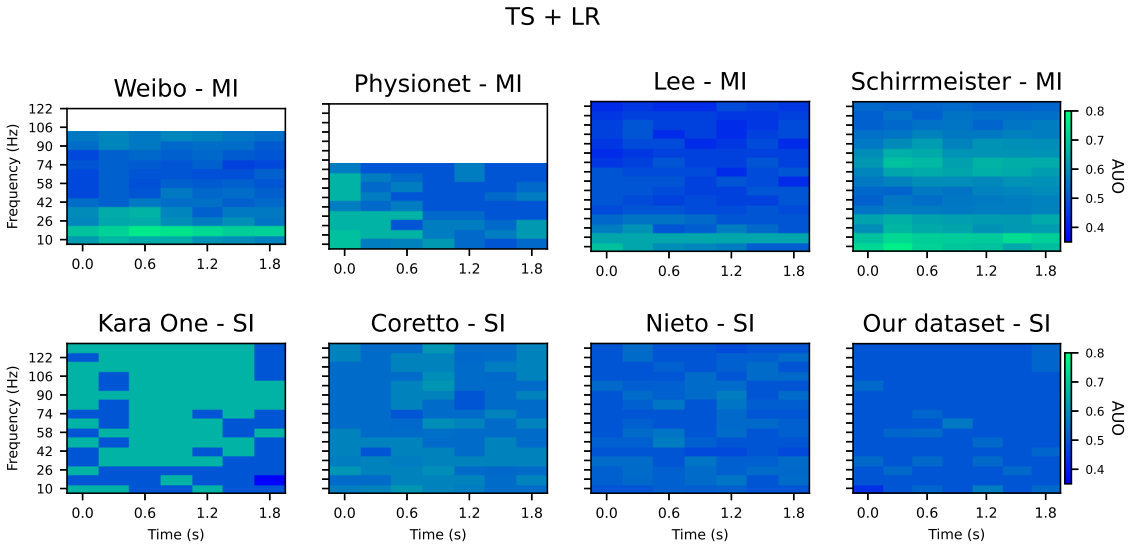


Figure C1. Heatmaps of median classification accuracies across participants for various time and frequency windows using the Tangent Space and Logistic Regressor (TS+LR) pipeline. The top row displays results for four Motor Imagery (MI) datasets, while the bottom row displays results for four Speech Imagery (SI) datasets. For this analysis, the evaluated SI targets were the imagined phonemes /u/ versus /a/ (Coretto) and /iy/ versus /m/ (Kara One), and the imagined words 'up' versus 'down' (Nieto) and 'left' versus 'right' (our dataset). MI tasks remained left versus right hand imagery. Evaluation was performed using 10-fold cross-validation. SI features continued to lack any consistent regions of significant accuracy ($p < 0.05$).

Appendix C.2. Time-Frequency Decoding results from other decoding pipelines

The decoding results across different time-frequency configurations for the other classifiers do reflect a similar difference between the paradigms as with TS+LR pipeline, where accuracies for MI dataset show a pattern in lower frequency regions, but no pattern is seen for SI datasets. Our results also suggest that TS+LR scores outperform the CNN and CSP+LDA pipeline. CSP+LDA accuracy heatmap is presented in Figure C4 and CNN results in Figure C5.

Appendix C.3. Results on parameter selection

Table C1 shows the results of grid-search on SVM parameters C and *gamma* for reproduction attempts of K02 and CT2 reports. Mean accuracies across participants of reports with assumed parameters are lower; however, differences in accuracies are not statistically significant. Table C2 shows the results of the grid search in the number of ICA components to remove for reproduction attempts of KO2 and CT4 reports. The mean accuracy across participants with 1 removed component is lower than the grid-search for the optimal number of components; however, the difference in accuracy distributions is not statistically significant.

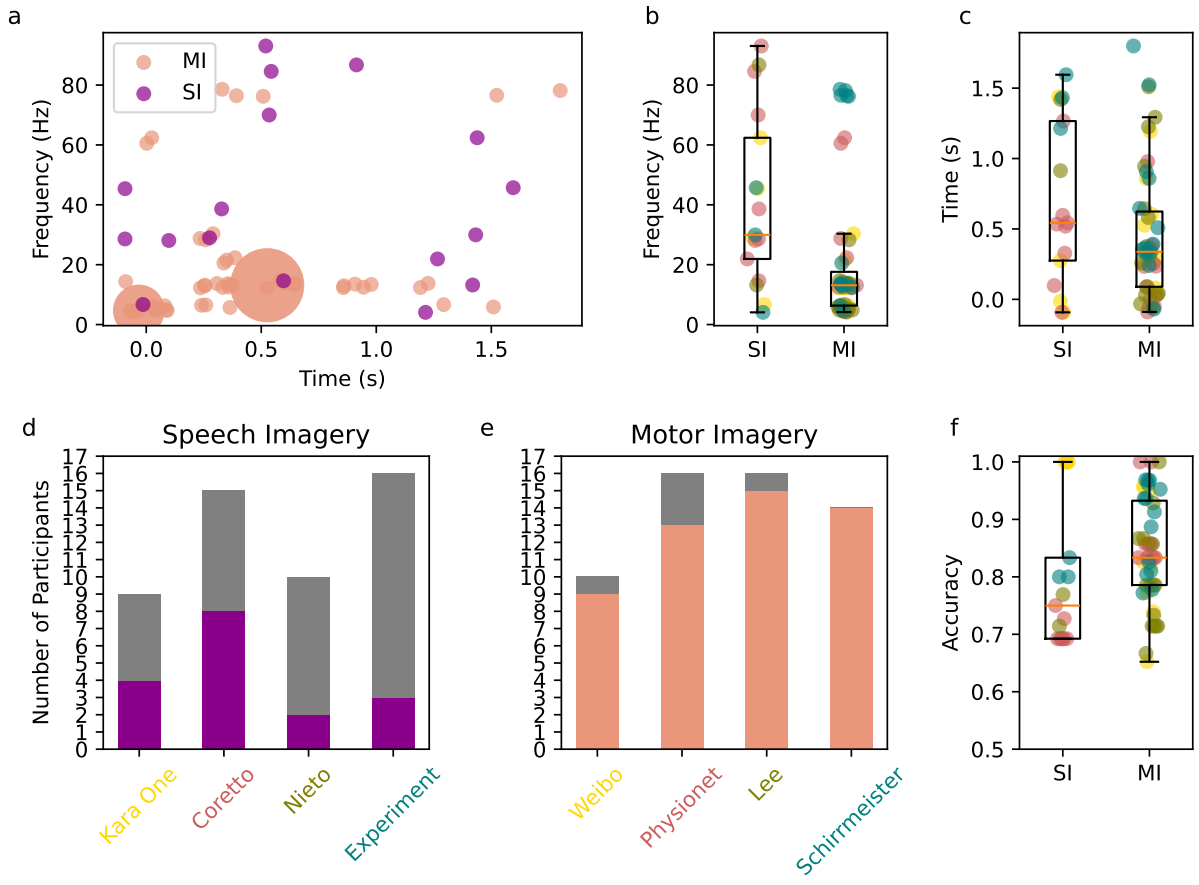


Figure C2. Distribution of the highest pair-wise decoding accuracies using the Tangent Space and Logistic Regressor (TS+LR) pipeline across different time-frequency configurations for the Motor Imagery (MI) and Speech Imagery (SI) datasets. The evaluated class pairs were left versus right hand imagery for all MI datasets. For the SI datasets, the evaluated targets were the imagined phonemes /u/ versus /a/ (Coretto) and /iy/ versus /m/ (Kara One), and the imagined words 'up' versus 'down' (Nieto) and 'left' versus 'right' (our dataset). Only participants with statistically significant performance at the 95% confidence level ($\alpha = 0.05$) are included. *a.* Scatter plot of individual participant results across time-frequency space. Clusters of at least five points are found using DBSCAN. The size of the circles indicates the number of participants in each cluster; only MI participants show clustering, primarily within the 0–30 Hz range, whereas significant SI features remained sparse across the time-frequency space. *b.* Frequency distribution of peak decoding accuracies, with each point representing a participant, colour-coded by dataset. *c.* Distribution of best accuracies across time, where the Y-axis indicates the starting time of each 1.5 s decoding window. *d.* The proportion of participants in each SI dataset achieving significant accuracy, with non-grey bar segments indicating successful cases. Notably, under these alternative SI conditions, the proportion of significant performers decreased for the Kara One and Nieto datasets, but increased for the Coretto dataset. *e.* Same as (d) but for MI datasets; all participants from the Schirrmeister dataset reached significant performance. *f.* Overall accuracy distributions for SI and MI participants, with individual scores color-coded by dataset.

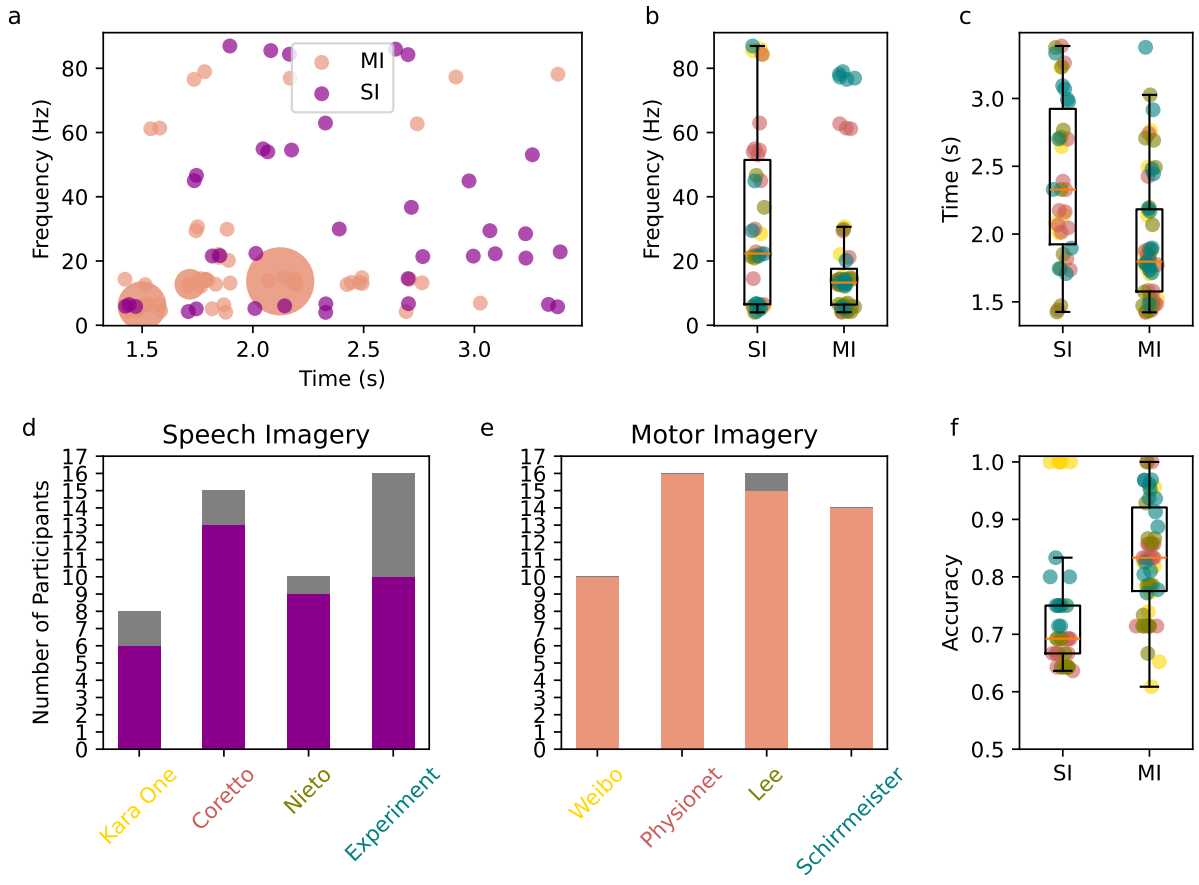


Figure C3. Distribution of the highest pair-wise decoding accuracies using the Tangent Space and Logistic Regressor (TS+LR) pipeline across different time-frequency configurations for the Motor Imagery (MI) and Speech Imagery (SI) datasets. The evaluated class pairs were left versus right hand imagery for all MI datasets. For the SI datasets, the tasks consisted of the imagined words 'left' versus 'right' (Coretto, Nieto, and our dataset) and 'pot' versus 'knew' (Kara One). Only participants with statistically significant performance at the 95% confidence level ($alpha = 0.05$) are included. *a.* Scatter plot of individual participant results across time-frequency space. Clusters of at least five points are found using DBSCAN. The size of the circles indicates the number of participants in each cluster; only MI participants show clustering, primarily within the 0–30 Hz range. *b.* Frequency distribution of peak decoding accuracies, with each point representing a participant, colour-coded by dataset. *c.* Distribution of best accuracies across time, where the Y-axis indicates the starting time of each 1.5 s decoding window. *d.*The proportion of participants in each SI dataset achieving significant accuracy, with non-grey bar segments indicating successful cases. *e.*Same as (d) but for MI datasets; all participants from the Schirrmeister dataset reached significant performance. *f* Overall accuracy distributions for SI and MI participants, with individual scores color-coded by dataset.

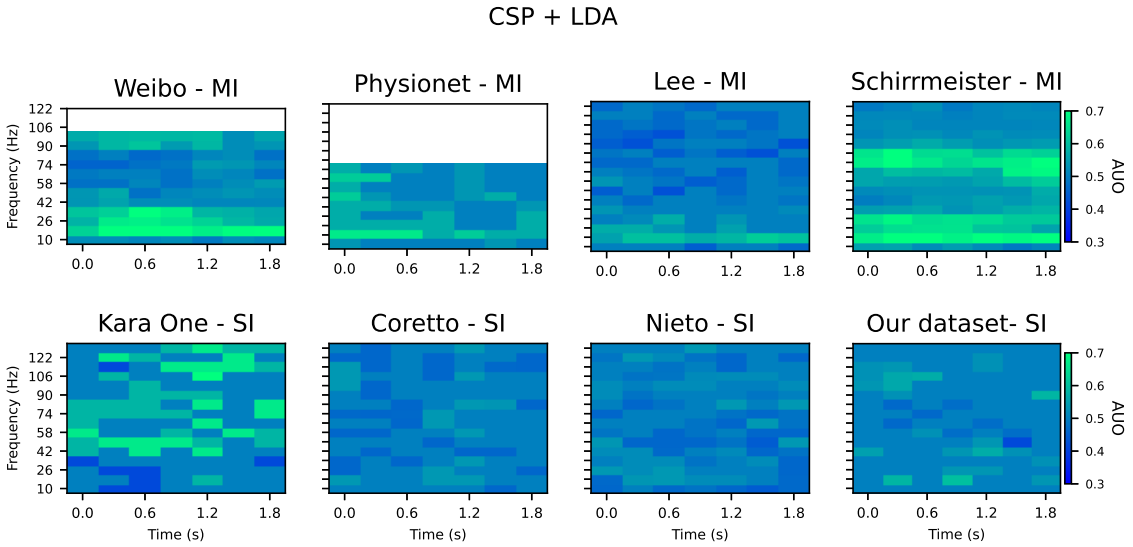


Figure C4. Heatmaps of median classification accuracies across participants for various time and frequency windows using the Tangent Space and Common Spatial Patterns and Linear Discriminant Analysis (CSP+LDA) pipeline. The top row displays results for four Motor Imagery (MI) datasets, while the bottom row displays results for four Speech Imagery (SI) datasets. Evaluation was performed using 10-fold cross-validation. The results demonstrate that MI datasets exhibit consistently high, statistically significant performance in low-frequency ranges (0–20 Hz), whereas SI datasets lack any consistent regions of significant accuracy ($p < 0.05$)

Table C1. Results of replication attempt with grid-search of SVM parameters in comparison with assumed values due to missing information

Decoding Approach	Results Reported in Literature (%)	Results with assumed values (%)	Results with grid-search
KO2	20.45 ± 5.7	15.5 ± 6.5	15.9 ± 6.2
CT2	81.1	51.7 ± 10.7	52.1 ± 11.3

Appendix C.4. Results on the use of participants with faulty data

Table C3 shows the results of reproduction attempts of Kara One dataset (KO2 and KO3) using all participants' data, in contrast to using only the 9 participants with non-faulty data. When adding the faulty data, the mean accuracies decrease, but the difference is not statistically significant.

Data availability statement

All data that support the findings in this study are publicly available.

Zhao and Rudzicz (2015), Kara One, `https://www.cs.toronto.edu/~complingweb/`

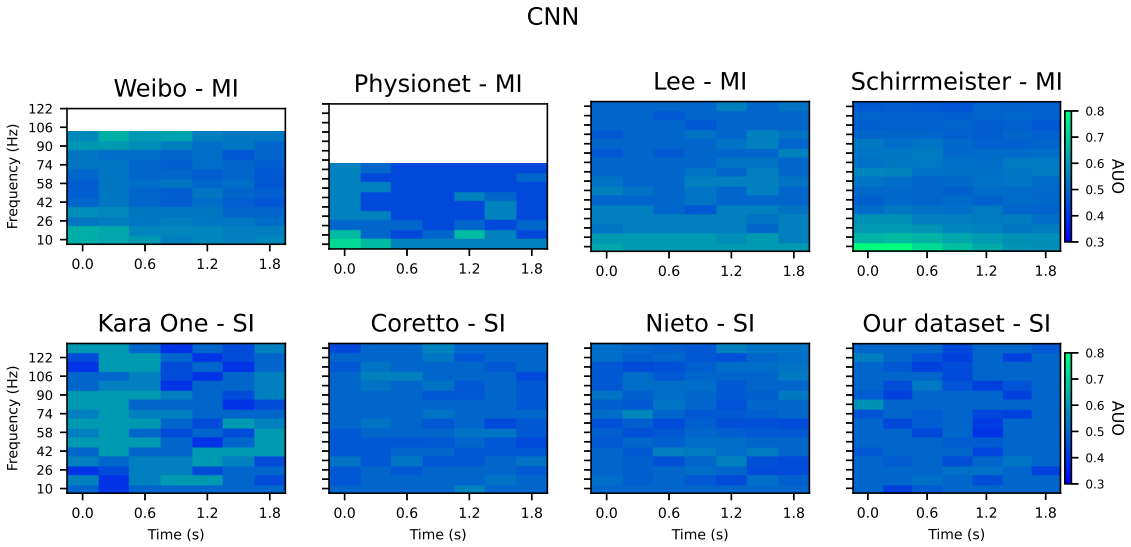


Figure C5. Heatmaps of median classification accuracies across participants for various time and frequency windows using the Tangent Space and Common Spatial Patterns and Convolutional Neural Networks (CNN) pipeline. The top row displays results for four Motor Imagery (MI) datasets, while the bottom row displays results for four Speech Imagery (SI) datasets. Evaluation was performed using 10-fold cross-validation. The results demonstrate that MI datasets exhibit consistently high, statistically significant performance in low-frequency ranges (0–20 Hz), whereas SI datasets lack any consistent regions of significant accuracy ($p < 0.05$)

Table C2. Results of replication attempt with grid-search of ICA components in comparison with assumed values due to missing information

Decoding Approach	Results Reported in Literature (%)	Results with assumed values (%)	Results with grid-search
KO2	20.45 ± 5.7	15.5 ± 6.5	16.1 ± 7.8
CT4	22.3 ± 1.81	19.1 ± 8.1	21.1 ± 7.1

Table C3. Results of replication attempt with different number of participants in Kara One approaches

Decoding Approach	Results reported in literature	Results with non-faulty data (%)	Results with all participants data
KO2	20.45 ± 5.7	15.5 ± 6.5	14.1 ± 7.5
KO3	31.6 ± 0.4	15.6 ± 0.08	13.1 ± 7.5

data/karaOne/karaOne.html.

Coretto et.al (2017), https://sinc.unl.edu.ar/downloads/imagined_speech/

Nieto N, Inner Speech Dataset, https://openneuro.org/datasets/ds003626/versions/2.1.2

All Motor Imagery datasets evaluated in this study can be accessed via Moabb https://moabb.neurotechx.com/docs/paper_results.html

The code from our replication attempts can be accessed on the github repository https://github.com/donalbertico/speech-imagery-replicability

Conflict of interest

The authors declare they have no competing interests.

References

[1] *BCI Definition*. https://bcisociety.org/bci-definition/. Accessed: 5 May 2026. BCI Society.

[2] Chang-Hee Han et al. "Electroencephalography-based endogenous brain–computer interface for online communication with a completely locked-in patient". In: *Journal of NeuroEngineering and Rehabilitation* 16.1 (Jan. 2019). ISSN: 1743-0003. DOI: 10.1186/s12984-019-0493-0. URL: http://dx.doi.org/10.1186/s12984-019-0493-0.

[3] Natasha Padfield et al. "A Comprehensive Review of Endogenous EEG Based BCIs for Dynamic Device Control". In: *Sensors* 22.15 (Aug. 2022), p. 5802. ISSN: 1424-8220. DOI: 10.3390/s22155802. URL: http://dx.doi.org/10.3390/s22155802.

[4] Kinkini Bhadra, Anne-Lise Giraud, and Silvia Marchesotti. "Learning to operate an imagined speech Brain-Computer Interface involves the spatial and frequency tuning of neural activity". In: *Communications Biology* 8.1 (Feb. 2025). ISSN: 2399-3642. DOI: 10.1038/s42003-025-07464-7. URL: http://dx.doi.org/10.1038/s42003-025-07464-7.

[5] Timothée Proix et al. "Imagined speech can be decoded from low- and cross-frequency intracranial EEG features". In: *Nature Communications* 13.1 (Jan. 2022). ISSN: 2041-1723. DOI: 10.1038/s41467-021-27725-3. URL: http://dx.doi.org/10.1038/s41467-021-27725-3.

[6] A Tates et al. "Speech imagery brain–computer interfaces: a systematic literature review". In: *Journal of Neural Engineering* 22.3 (June 2025), p. 031003. ISSN: 1741-2552. DOI: 10.1088/1741-2552/ade28e. URL: http://dx.doi.org/10.1088/1741-2552/ade28e.

[7] Shunan Zhao and Frank Rudzicz. "Classifying phonological categories in imagined and articulated speech". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 992–996. DOI: 10.1109/ICASSP.2015.7178118.

[8] Germán A. Pressel Coretto, Iván E. Gareis, and H. Leonardo Rufiner. "Open access database of EEG signals recorded during imagined speech". In: *12th International Symposium on Medical Information Processing and Analysis*. Ed. by Eduardo Romero et al. Vol. 10160. SPIE, Jan. 2017, p. 1016002. DOI: 10.1117/12.2255697. URL: http://dx.doi.org/10.1117/12.2255697.

[9] Robert A. McDougal, Anna S. Bulanova, and William W. Lytton. "Reproducibility in Computational Neuroscience Models and Simulations". In: *IEEE Transactions on Biomedical Engineering* 63.10 (Oct. 2016), pp. 2021–2035. ISSN: 1558-2531. DOI: 10.1109/tbme.2016.2539602. URL: http://dx.doi.org/10.1109/TBME.2016.2539602.

[10] Avinash Kumar Singh et al. "Editorial: Advances and challenges to bridge computational intelligence and neuroscience for brain-computer interface". In: *Frontiers in Neuroergonomics* 5 (Aug. 2024). ISSN: 2673-6195. DOI: 10.3389/fnrgo.2024.1461494. URL: http://dx.doi.org/10.3389/fnrgo.2024.1461494.

[11] Simanto Saha et al. "Progress in Brain Computer Interface: Challenges and Opportunities". In: *Frontiers in Systems Neuroscience* 15 (Feb. 2021). ISSN: 1662-5137. DOI: 10.3389/fnsys.2021.578875. URL: http://dx.doi.org/10.3389/fnsys.2021.578875.

[12] Sean Kinahan et al. "Achieving Reproducibility in EEG-Based Machine Learning". In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. ACM, June 2024, pp. 1464–1474. DOI: 10.1145/3630106.3658983. URL: http://dx.doi.org/10.1145/3630106.3658983.

[13] Parthiv Menon, Vignesh Sekaran, and Garima Bajwa. "Reproducibility in Brain-Computer Interface Research: A Replication-Based Analysis". In: *2022 IEEE 18th International Conference on e-Science (e-Science)*. 2022, pp. 462–467. DOI: 10.1109/eScience55777.2022.00083.

[14] Güliz Demirezen, Tuğba Taşkaya Temizel, and Anne-Marie Brouwer. "Reproducible machine learning research in mental workload classification using EEG". In: *Frontiers in Neuroergonomics* 5 (Apr. 2024). ISSN: 2673-6195. DOI: 10.3389/fnrgo.2024.1346794. URL: http://dx.doi.org/10.3389/fnrgo.2024.1346794.

[15] Ke Su and Liang Tian. "Systematic review: progress in EEG-based speech imagery brain-computer interface decoding and encoding research". In: *PeerJ Computer Science* 11 (June 2025), e2938. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.2938. URL: http://dx.doi.org/10.7717/peerj-cs.2938.

[16] Nicolás Nieto et al. "Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition". In: *Scientific Data* 9.1 (Feb. 2022). ISSN: 2052-4463. DOI: 10.1038/s41597-022-01147-2. URL: http://dx.doi.org/10.1038/s41597-022-01147-2.

[17] Sylvain Chevallier et al. "The largest EEG-based BCI reproducibility study for open science: the MOABB benchmark". In: (Apr. 2024). working paper or preprint. URL: https://universite-paris-saclay.hal.science/hal-04537061.

[18] G. Schalk et al. "BCI2000: A General-Purpose Brain-Computer Interface (BCI) System". In: *IEEE Transactions on Biomedical Engineering* 51.6 (June 2004), pp. 1034–1043. ISSN: 0018-9294. DOI: 10.1109/tbme.2004.827072. URL: http://dx.doi.org/10.1109/TBME.2004.827072.

[19] Min-Ho Lee et al. "EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy". In: *GigaScience* 8.5 (Jan. 2019). ISSN: 2047-217X. DOI: 10.1093/gigascience/giz002. URL: http://dx.doi.org/10.1093/gigascience/giz002.

[20] Weibo Yi et al. "Evaluation of EEG Oscillatory Patterns and Cognitive Process during Simple and Compound Limb Motor Imagery". In: *PLoS ONE* 9.12 (Dec. 2014). Ed. by Natasha M. Maurits, e114853. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0114853. URL: http://dx.doi.org/10.1371/journal.pone.0114853.

[21] Robin Tibor Schirrmeister et al. "Deep learning with convolutional neural networks for EEG decoding and visualization". In: *Human Brain Mapping* 38.11 (Aug. 2017), pp. 5391–5420. ISSN: 1097-0193. DOI: 10.1002/hbm.23730. URL: http://dx.doi.org/10.1002/hbm.23730.

[22] Eric Larson, Alexandre Gramfort, and .. Engemann. *MNE-Python*. 2025. DOI: 10.5281/ZENODO.15928841. URL: https://zenodo.org/doi/10.5281/zenodo.15928841.

[23] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[24] Gregory Lee et al. "PyWavelets: A Python package for wavelet analysis". In: *Journal of Open Source Software* 4.36 (Apr. 2019), p. 1237. ISSN: 2475-9066. DOI: 10.21105/joss.01237. URL: http://dx.doi.org/10.21105/joss.01237.

[25] François Chollet et al. *Keras*. https://keras.io. 2015.

[26] Ciaran Cooney, Rafaella Folli, and Damien Coyle. "Mel Frequency Cepstral Coefficients Enhance Imagined Speech Decoding Accuracy from EEG". In: *2018 29th Irish Signals and Systems Conference (ISSC)*. 2018, pp. 1–7. DOI: 10.1109/ISSC.2018.8585291.

[27] Ana-Luiza Rusnac and Ovidiu Grigore. "CNN Architectures and Feature Extraction Methods for EEG Imaginary Speech Recognition". In: *Sensors* 22.13 (June 2022), p. 4679. ISSN: 1424-8220. DOI: 10.3390/s22134679. URL: http://dx.doi.org/10.3390/s22134679.

[28] Danial Alizadeh and Hesam Omranpour. "EM-CSP: An efficient multiclass common spatial pattern feature method for speech imagery EEG signals recognition". In: *Biomedical Signal Processing and Control* 84 (July 2023), p. 104933. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2023.104933. URL: http://dx.doi.org/10.1016/j.bspc.2023.104933.

[29] Sukanya Biswas and Rohit Sinha. "Wavelet filterbank-based EEG rhythm-specific spatial features for covert speech classification". In: *IET Signal Processing* 16.1 (Sept. 2021), pp. 92–105. ISSN: 1751-9683. DOI: 10.1049/sil2.12059. URL: http://dx.doi.org/10.1049/sil2.12059.

[30] Ciaran Cooney, Raffaella Folli, and Damien Coyle. "Optimizing Layers Improves CNN Generalization and Transfer Learning for Imagined Speech Decoding from EEG". In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2019, pp. 1311–1316. DOI: 10.1109/SMC.2019.8914246.

[31] Foteini Simistira Liwicki et al. "Rethinking the Methods and Algorithms for Inner Speech Decoding and Making Them Reproducible". In: *NeuroSci* 3.2 (Apr. 2022), pp. 226–244. ISSN: 2673-4087. DOI: 10.3390/neurosci3020017. URL: http://dx.doi.org/10.3390/neurosci3020017.

[32] Olivier Ledoit and Michael Wolf. "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection". In: *Journal of Empirical Finance* 10.5 (Dec. 2003), pp. 603–621. ISSN: 0927-5398. DOI: 10.1016/s0927-5398(03)00007-0. URL: http://dx.doi.org/10.1016/S0927-5398(03)00007-0.

[33] Alexandre Barachant et al. *pyRiemann/pyRiemann: v0.3*. 2022. DOI: 10.5281/ZENODO.7547583. URL: https://zenodo.org/record/7547583.

[34] Gernot Mueller-Putz et al. "Better than random? A closer look on BCI results". In: 2008. URL: https://api.semanticscholar.org/CorpusID:15418438.

[35] Aya Kabbara et al. "Successful reproduction of a large EEG study across software packages". In: *Neuroimage: Reports* 3.2 (June 2023), p. 100169. ISSN: 2666-9560. DOI: 10.1016/j.ynirp.2023.100169. URL: http://dx.doi.org/10.1016/j.ynirp.2023.100169.

[36] Megan A. Boudewyn et al. "How many trials does it take to get a significant ERP effect? It depends". In: *Psychophysiology* 55.6 (Dec. 2017). ISSN: 1469-8986. DOI: 10.1111/psyp.13049. URL: http://dx.doi.org/10.1111/psyp.13049.

[37] Pete Chapman et al. "CRISP-DM 1.0: Step-by-step data mining guide". In: *SPSS inc* 9.13 (2000), pp. 1–73.

[38] Russell T. Hurlburt et al. "Exploring the Ecological Validity of Thinking on Demand: Neural Correlates of Elicited vs. Spontaneously Occurring Inner Speech". In: *PLOS ONE* 11.2 (Feb. 2016). Ed. by Frederic Dick, e0147932. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0147932. URL: http://dx.doi.org/10.1371/journal.pone.0147932.

[39] Ben Alderson-Day et al. "The varieties of inner speech questionnaire – Revised (VISQ-R): Replicating and refining links between inner speech and psychopathology". In: *Consciousness and Cognition* 65 (Oct. 2018), pp. 48–58. ISSN: 1053-8100. DOI: 10.1016/j.concog.2018.07.001. URL: http://dx.doi.org/10.1016/j.concog.2018.07.001.

[40] David H. Brainard. "The Psychophysics Toolbox". In: *Spatial Vision* 10.4 (1997), pp. 433–436. ISSN: 1568-5683. DOI: 10.1163/156856897x00357. URL: http://dx.doi.org/10.1163/156856897X00357.

[41] Johannes Müller-Gerking, Gert Pfurtscheller, and Henrik Flyvbjerg. "Designing optimal spatial filters for single-trial EEG classification in a movement task". In: *Clinical Neurophysiology* 110.5 (May 1999), pp. 787–798. ISSN: 1388-2457. DOI: 10.1016/s1388-2457(98)00038-8. URL: http://dx.doi.org/10.1016/S1388-2457(98)00038-8.

[42] Florian Yger, Maxime Berar, and Fabien Lotte. "Riemannian Approaches in Brain-Computer Interfaces: A Review". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (2017), pp. 1753–1762. DOI: 10.1109/TNSRE.2016.2627016.

[43] Martin Wimpff et al. "EEG motor imagery decoding: a framework for comparative analysis with channel attention mechanisms". In: *Journal of Neural Engineering* 21.3 (May 2024), p. 036020. ISSN: 1741-2552. DOI: 10.1088/1741-2552/ad48b9. URL: http://dx.doi.org/10.1088/1741-2552/ad48b9.