Researching The Viability of Speech Imagery as a BCI Paradigm

Alberto Tates

School of Computer Science and Electronic Engineering
University of Essex

A thesis submitted for the degree of
Doctor of Philosophy

January 2026

Supervised by Dr. Ana Matran-Fernandez, Dr. Sebastian Halder and Dr. Ian Daly

Abstract

Speech Imagery (SI) is envisioned as a premier paradigm for Brain-Computer Interface (BCI) design. As an endogenous modality, it offers a highly intuitive control scheme—allowing users to command devices by mentally articulating words without the need for external sensory stimulation. While the field has gained significant momentum, resulting in a diverse body of literature, this thesis identifies a critical gap between experimental optimism and practical implementation. We aim to progress the consolidation of the SI paradigm by rigorously assessing its feasibility, reproducibility, and replicability.

Our analysis of 104 decoding attempts revealed a striking disparity: only 6% were conducted in real-time scenarios, despite over half involving original data collection. To investigate this "reproducibility crisis," we attempted to reproduce findings from two popular SI datasets and compared them against the well-established Motor Imagery (MI) paradigm. Our results uncovered significant challenges; reproduction performance was consistently lower than originally reported, with 50% of attempts showing performance drops of up to 40% due to incomplete methodological reporting and flawed evaluation procedures. Furthermore, a large-scale replicability analysis across 12 heterogeneous SI datasets painted a challenging picture, with the majority of datasets failing to meet practical BCI efficiency thresholds.

However, this comprehensive evaluation also identified a definitive path toward paradigm maturation. We discovered that rhythmic protocols in speech imagery lead to superior decoding results. Our meta-analysis suggests that rhythmic repetition may act as a organizational template, inducing more structured and consistent neural responses—characterised by lower covariance matrix entropy—which significantly facilitates decoding. We conclude by proposing a set of rigorous directives for the field, including standardised reporting frameworks and rhythmic task designs,

to transition Speech Imagery from an experimental curiosity to a consolidated and reliable BCI technology.

Acknowledgements

I want to express my deepest gratitude to my supervisors, Ana Matran-Fernandez, Sebastian Halder, and Ian Daly, for their outstanding scientific guidance, unwavering support, and kind compassion throughout my PhD journey. Their feedback elevated this thesis at every step.

Being part of the Brain-Computer Interfaces and Neural Engineering Laboratory has been an outstanding experience. I am grateful to have witnessed how a world-class research laboratory evolves and to have contributed to its mission.

I would like to thank my PhD colleagues for always being ready to lend a helping hand, as well as every friend at the University of Essex who generously spared their time to participate in my experiments.

I would like to thank my mother for her everlasting love and support, and my father for his goodwill and encouragement despite the distance. My thanks also go to every friend, uncle, aunt and cousin back in Ecuador who kept track of my progress and cheered me on.

Finally, I want to deeply thank my beloved partner, Daniela, who witnessed both the best and worst stages of this journey and always had the perfect words to keep me going.

Contents

List of Figures

List of Tables

List of Acronyms

BCI Brain-Computer Interface

BOLD Blood-Oxygen-Level-Dependent

CNN Convolutional Neural Network

CSP Common Spatial Patterns

CV Cross-Validation

DTCWT Dual-Tree Complex Wavelet Transform

DWT Discrete Wavelet Transform

ECoG Electrocorticography

EEG Electroencephalography

ERD Event-Related Desynchronisation

fMRI functional Magnetic Resonance Imaging

fNIRS functional Near-Infrared Spectroscopy

ICA Independent Component Analysis

IFG Inferior Frontal Gyrus

ITR Information Transfer Rate

LDA Linear Discriminant Analysis

LR Linear Regression

MAE Mean Absolute Error

MEG Magnetoencephalography

MFCC Mel-Frequency Cepstral Coefficients

MI Motor Imagery

MTG Middle Temporal Gyrus

PCA Principal Component Analysis

PSD Power Spectral Density

RF Random Forest

SAN Success-Anchored Normalisation

SD Speech Decoding

SEEG Stereotactic Electroencephalography

SI Speech Imagery

SNR Signal-to-Noise Ratio

SSVEP Steady-State Visual Evoked Potentials

STG Superior Temporal Gyrus

SVM Support Vector Machine

TS Tangent Space

TS+LR Tangent Space + Linear Regression

WER Word Error Rate

1 Introduction

This chapter introduces the motivation, research objectives and structure of this thesis.

1.1 Motivation

If you can read this paragraph silently, you are probably experiencing a voice inside your head. Where is this voice coming from? Do you perceive it as your ordinary voice, or does it have different characteristics? Is it the same as in your deepest moments of reasoning? Speech Imagery (SI), also referred to as auditory verbal imagery, is defined as the mental representation of sounds in the absence of external auditory stimuli; it evokes the kinesthetic experience of "hearing" speech inside our heads[1]. SI is closely related to Inner Speech (IS), and their definitional boundaries are thin. Inner Speech has long been a central topic in the intersecting fields of psychology, philosophy, and neuroscience. Once defined by Plato as "the voice within the mind," it is a relevant conscious experience linked to thought. SI also intersects with ongoing discussions in the fields of language and consciousness [2].

In contrast with other types of imagery as visual, gustatory or olfactory, that represent basic sensory mechanisms and perceptual functions, Motor Imagery (MI) and SI are higher brain functions that require both execution and perception

and encompass further cognitive processes [2, 3]. As a linguistic phenomenon, SI encompasses three main components: the auditory or phonological part, as the experience resembles hearing someone speak; a semantic character, as the meanings carried in sentences; and an articulatory component, which may involve articulator movement preparation or planning of its utterance. However, how exactly each of these elements relates to SI, and whether SI can happen without any of them, is still under debate [4].

Brain imaging studies have demonstrated significant overlap in neural activation between speech imagery (SI) and overt speech. These findings suggest that SI recruits core linguistic processes, such as phonetic decoding and syllabification. However, while overt speech culminates in articulatory movement, SI is thought to involve a simulatory process that generates the kinesthetic experience of an inner voice. Common activated regions include the inferior frontal cortex, sensorimotor cortex, and the temporal-parietal junction [1, 5–11]. Beyond these shared areas, some research indicates that SI uniquely engages hippocampal structures associated with memory retrieval [12, 13]. Furthermore, while auditory areas are often linked to the subjective experience of "hearing" during SI, their precise involvement remains a subject of ongoing debate [2, 14, 15].

Given that SI-related neural activity can be reliably captured through various neuroimaging techniques, leveraging it as a paradigm for BCI has become a logical progression in the field. Active BCI systems are designed to enable direct communication between a user's brain and a computer, typically by translating specific patterns of brain activity into commands [16]. For such systems to be effective, the brain activity must be voluntarily generated by the user. This requirement has led to the exploration of different cognitive strategies, including a strategy known as endogenous paradigms–approaches in which the relevant brain activity is internally generated without reliance on external stimuli. Within this context, mental imagery has emerged as a particularly compelling approach. Among

imagery paradigms, Motor Imagery (MI)–the mental simulation of movement–has gained significant attention and has become one of the most established and validated paradigms in BCI research [17, 18].

The idea of communication merely through thoughts or telepathy was in the mind of Hans Bengerg, who developed the EEG back in 1924 [19]. In modern terms, designing a BCI system that allows users to control a device simply by thinking of a command is not only intuitive but also highly appealing. In contrast to well-established BCI paradigms such as Steady-State Visual Evoked Potentials (SSVEP), which require continuous visual stimulation, SI can, in principle, be produced spontaneously, without external cues. While SSVEP applications—like visual spellers—are limited by the number of items that can be displayed on a screen [20], SI could theoretically offer a broader range of commands, depending on how many words or speech units can be accurately decoded. This potential for flexible, spontaneous communication makes SI a particularly promising paradigm for assistive BCI applications, especially in the context of speech synthesizers aimed at aiding communication. As such, SI stands out as a naturally compelling candidate for BCI research and development.

The first notable attempt to decode internally generated speech was made by DaSalla et al. (2009) [21], who recorded single-trial EEG responses as participants imagined articulating two English vowels: /a/ and /u/. Their classification results exceeded chance levels, paving the way for SI to be considered a viable BCI control mechanism. By 2016, the release of two open-access SI datasets [22, 23] led to a rapid increase in decoding studies, and literature grew with complex pipelines and encouraging accuracy reports. However, despite this recent interest, Speech Imagery has yet to be consolidated as a reliable BCI paradigm. One of the main challenges is the limited understanding of its underlying cognitive and neural mechanisms. Although several studies have reported SI-related neural activity [7, 24–26], no consistent neural signature has been identified, as responses vary widely across individuals. This contrasts with MI, where well-characterised patterns such as Event-Related Synchronisation and Desynchronisation (ERS/ERD) are documented

and understood. In addition to neurophysiological uncertainty, BCI inefficiency, also referred to as BCI illiteracy, remains a significant concern [27, 28]. This phenomenon—where some individuals are unable to produce brain signals that can be accurately decoded—affects all BCI paradigms to varying degrees. It is plausible that SI may be similarly affected, though its specific susceptibility is still unknown. Hypothetically, if SI were associated with a lower rate of BCI inefficiency compared to other paradigms, it could present a valuable advantage for BCI design. Conversely, a higher inefficiency rate would further challenge its feasibility. Psychological factors also complicate the use of SI in BCI applications. Studies in cognitive science and psychology highlight the difficulty of reliably assessing a person's capacity for inner speech. While some findings suggest that up to 50% of individuals frequently experience internal monologues, this estimate varies widely, and the subjective nature of inner speech makes it hard to quantify [29–31]. These considerations underscore the importance of accounting for individual differences when designing and interpreting SI-based experiments.

Second, no standardised experimental protocol exists for eliciting SI. There is ambiguity in how participants are instructed to perform the task, and SI can be easily confused with related but distinct cognitive processes. These include silent speech (covert articulation with minimal movement), silent reading, which has often been mistaken for SI [32, 33], and object identification tasks, which have also been misclassified as SI in some studies [22, 34]. Third, and importantly, despite the growing number of SI decoding studies and datasets, only a small number of real-time SI decoding implementations have been published. This lack of real-time applications may indicate underlying reproducibility issues within the field and raises concerns about the practical feasibility of SI-based BCIs.

Speech Imagery holds promise as a BCI paradigm; however, its complex nature has led to a body of research that is heterogeneous and lacks standardisation. A focused effort toward reproducibility and methodological consolidation is needed to assess its true feasibility. Advancing our understanding of SI could also deepen insights into inner speech, aid in identifying psychological profiles more likely to

produce decodable SI-related activity, and ultimately support the development of strategies for modulating SI activity and enhancing BCI learning.

1.2 Research Objectives

This thesis aims to explore the feasibility of speech imagery (SI) Brain-Computer Interfaces (BCIs) using electroencephalography (EEG). The primary objective is to identify methodologies that demonstrate consistent success in decoding SI by evaluating the reproducibility of existing results and their replication potential across open-access datasets and internally recorded data using tools like LaTeX. Addressing this objective involves answering the following research questions:

RQ1 To what extent does current literature support the feasibility of SI as a viable BCI paradigm?

RQ2 What is the current state of reproducibility in SI research, given the variability in available datasets?

RQ3 Can existing SI decoding pipelines be successfully reproduced using the most prominent open-access datasets?

RQ4 Is it possible to identify a decoding pipeline that achieves performance significantly above chance level across multiple SI datasets?

RQ5 Which recording protocols, if any, are associated with significantly higher decoding accuracies?

RQ6 How do decoding results from internally recorded data compare to those obtained from open-access datasets?

1.3 Thesis Structure

Chapter 2 presents a systematic literature review of speech imagery decoding across various neuroimaging modalities. It evaluates the different experimental designs and machine learning pipelines employed in the field, while highlighting

trends in feature extraction techniques and the recent exponential growth in SI research (RQ1). Additionally, it analyses the prevalence of real-time versus offline decoding approaches and pictures the current condition of the paradigm as a feasible BCI candidate (RQ2).

Chapter 3 introduces the experimental framework designed to test the reproducibility of prominent SI decoding techniques. First, the chapter describes the signal acquisition design along two variations of the protocol tested. Second, it describes the main decoding pipeline used in this thesis. Third, it describes the evaluation procedures used to compare decoding accuracies across different SI datasets.

Chapter 4 details a reproducibility and replicability study. It follows the methodologies of three significant decoding attempts using two of the most widely cited open-access SI datasets to further support RQ2 answer, comparing the reproduced results with those originally reported (RQ3). Furthermore, it compares the replicability of the SI paradigm against the Motor Imagery (MI) paradigm by evaluating various decoding pipelines across both modalities (RQ4).

Chapter 5 extends the analysis of replicability to identify whether specific variables within the datasets lead to significantly higher performance (RQ5). It provides an estimation of SI-BCI inefficiency within currently available SI data and includes a meta-analysis to identify metrics that may indicate lower-performing participants or datasets and distinctiveness of any protocol (RQ6).

Chapter 6 summarises the core findings of this research, offering conclusions and suggesting directions for future work in the neural decoding of speech imagery.

2

A Systematic Literature Review

This chapter presents a systematic literature review of speech imagery decoding from neural activity. This systematic literature review has been published in [35]

2.1 Introduction

Inner speech, along with inner seeing or feeling, is referred to as mental imagery activities. Neuroimaging techniques have shown clear brain activity elicited by these cognitive tasks [36–40]. These patterns of activity have been explained as perceptual internal representations reconstructed without perceptual processing of external stimulation [41]. Mental imagery has been proposed as a predictive process where the perceptual consequences can let us gain an advantage in different aspects of our human experience (such as motor control, decision-making, and language) [41, 42].

Brain-Computer Interface (BCI) systems provide an interaction channel to computers directly from brain activity and can help people who have lost control over their voluntary muscles by providing a new communication pathway [16]. Such systems work by decoding brain signals recorded via neuroimaging methods such as electroencephalography (EEG), stereotactic electroencephalography (SEEG), electrocorticogram (ECoG), or magnetoencephalography (MEG) for brain electromagnetic fields, and functional near-infrared spectroscopy (fNIRS) for blood-oxygen-

level-dependent (BOLD) signals. Each technology has different characteristics and limitations. EEG is one of the most used neuroimaging modalities for SI research because of its relatively lower cost and portability [43, 44].

BCI paradigms can be categorised based on whether the origin of brain activity is exogenous, wherein the recorded activity is generated by an external stimulus, or endogenous, wherein the recorded signals come from spontaneous activations related to the user's intention. Because additional devices are required for exogenous paradigms, endogenous paradigms may present a more comfortable and intuitive user experience. However, they can bring further challenges as they require user training and their performance can vary considerably over users [45, 46].

Motor Imagery (MI) is a category of mental imagery tasks that shares some properties with speech imagery when used as a BCI paradigm. MI has been broadly studied for BCI designs. The kinesthetic experience has been described with the use of internal forward models producing a simulation activity that has been denominated as efference copy [47]. Presence of efference copies of the motor cortex and other motor-related regions has been demonstrated in a variety of brain imaging studies [48–50]. MI-related activity can be measured by EEG and has been utilised to help design applications to control robotic limbs [51], communication interfaces, such as spellers [52], and videogames [53]. Like MI, SI is an attractive mental imagery-based BCI paradigm. Tian and Poepple [25] showed comparable brain activation processes for both paradigms and work by Wang et. al [54] demonstrated classifiable EEG signals in both SI and MI paradigms.

The idea of a perceptual representation of inner speech was presented by Tian and Poeppel [25] where activation in the auditory cortex was observed during speech imagery. Tian and Poeppel's experiments did not include auditory stimuli, so they explained this auditory cortex activation as being due to the presence of a perceptual efference copy. Work by Grandchamp et.al [55] supports the idea that this efference copy comes from motor commands that were inhibited due to the absence of articulatory onset during SI.

SI is an attractive paradigm for use in speech synthesis applications for people who have lost the ability to speak. It may have an advantage over MI in some control applications because it may be more intuitive for users to imagine command words rather than limb movements. Consequently, SI as a BCI paradigm has gained attention among researchers, therefore, so we aim to identify key aspects of SI decoding attempts with the following questions. What feature extraction techniques have been frequently used, do researchers agree with a most informative feature, is there an ideal SI experiment design, and what decoding results can be achieved with different modalities?

This paper reviews the existing literature on the decoding of imagined speech, with the goal of examining critical aspects of SI-BCI design. Specifically, we aim to identify methodological trends associated with successful decoding, as well as speech units that exhibit greater discriminability. Furthermore, we assess the proportion of offline versus online decoding implementations, in order to evaluate the extent to which reported methodologies have been validated in closed-loop settings. By analysing the literature, we aim to provide a comprehensive overview of the current state of the art in SI-BCI research. This analysis also serves to highlight ongoing challenges and propose potential directions for future research in this emerging field.

2.2 Speech and Speech Imagery Related Brain areas

Throughout years of research into the brain's language system, two regions have been identified as being most broadly associated with language understanding and generation: Wernicke's and Broca's areas of the temporal and frontal lobes. Broca's area was first linked with word production by Paul Broca in 1861 [56]. This posterior portion of the left inferior frontal gyrus (IFG) has been related to articulatory activity. However, further evidence shows that it is implicated in diverse cognitive processes such as action execution and music listening [14, 57]. Due to evidence of Broca area's involvement in speech production but not other oral-motor movements, its specific role during speech is still debated [58, 59]. Wernicke's

area was first identified as the brain's area that manages speech in 1874 by Carl Wernicke [60, 61]. The current knowledge about this anatomical site consisting of the pSTG (posterior Superior Temporal Gyrus) and the supramarginal gyrus (SMG) is that it represents a critical area for speech production, associated with the brain's representation of phonological information, a process named phonological retrieval [62]. However, further evidence suggests that the phonological processing involves a larger network including regions with the middle temporal gyrus (MTG) and angular gyrus (AG) [63].

Core speech operations consist of retrieving a word's phonological representation, translating it into articulatory code, and coordinating the motor instructions for vocal articulation [3]. An early definition of a speech production model proposed by Idefrey and Levelt [64] is that word production, regardless of its stimulation (reading, naming, or generating), starts with a conceptual preparation step, i.e., if we intend to name APPLE, the concept of a FRUIT is likely to first be activated. The second step is the lemma retrieval or lexical selection based on the activated concept, which involves retrieving the words's syntax (grammatical encoding), a task associated with the MTG. The next step is the form encoding and phonetic code retrieval involving Wernicke's area. This step is proposed to be responsible for morphological encoding, i.e., if the intended word is HIS, codes for /he/ and for possession /is/ are retrieved, as well as syllabification information. Finally, a step of phonetic encoding, that was originally attributed to Broca's area, makes syllable codes turn into motor action instructions that are forwarded to the ventral premotor and motor cortex for articulatory execution. The whole process is estimated to take around 600 ms until the beginning of articulation [14].

One influential model of speech production is the dual-stream model [65], which separates speech processing into semantic and sensorimotor parts, assigning two streams to it. According to this model, word production starts with a phonological network around the STG and diverges into parallel streams. In the ventral stream, information flows from lemma retrieval in the posterior MTG and posterior temporal sulcus (ITS) that communicates with the dorsal stream, these regions are consistent

with evidence for semantic decoding [66]. The dorsal stream is left-hemisphere dominant. In this stream, information is processed in the Sylvian parietal-temporal (Spt) area, considered a sensorimotor interface that forwards information to the articulatory network consisting of Broca's area and the primary left motor and premotor cortices. Further evidence about Broca's function in speech production suggests that its activity involves information mediation between the somatosensory representation of words coming from STG and their corresponding articulatory gestures sent to the motor cortex [67].

Broca's area, as well as the temporal cortex, has been shown to be active after participants performed imagined speech tasks. Early studies [68–70] have shown that these common areas are shared with with overt speech tasks without involvement of the motor cortex. However further fMRI studies comparing overt and covert speech found that covert speech also involves activation of the motor cortex with lower oxygenated blood level amplitude [71–73]. There are also additional reports that suggest more prominent Broca's activity in covert compared to overt speech [74, 75].

Based on the dual stream model of speech, the dual stream prediction model (DSPM) was proposed by Tian et. al [1, 76]. This model defines mental imagery as an internally generated quasi-perceptual experience, suggesting that, as in other imagery tasks, SI involves a forward model that simulates the somatosensory expected outputs and auditory representations as a predictive tool for possible speech mistakes. Such perceptual neural representations are produced in the auditory cortex. Work by Whitford et. al [15] collected evidence on the inhibition of the auditory response in favour of an efference copy produced in the auditory cortex by SI. However, both streams of DSPM may create the kinesthetic experience of SI and involve the STG region, these are referred to as the memory retrieval and the simulation estimation streams.

In the memory retrieval stream, auditory representations can be accessed from episodic memory involving hippocampal structures [12, 13] or from lexical and semantic information networks including the IFG, the posterior parietal lobe, and MTG regions. In the simulation-estimation stream, the perceptual consequences

of the auditory representations are predicted by simulation of the articulation and corresponding perceptual changes as an auditory feedback. Some researchers have suggested that SI involves a motor inhibition step by control signals of the intended articulatory execution [55, 77], such a process has been proposed for motor imagery [14, 15] allowing some to hypothesise involvement of event-related synchronisation activity during SI [25].

Solid evidence has settled the classic language regions (Broca's and Wernicke's areas) correlates of both convert and overt speech production. However, our understanding of all the underlying processes involving these two is still under debate and these higher-order functions require parallel activation of a wide array of cortical neural networks. Recent work highlighted the complexity of brain interactions involved in language-related tasks by analysing a collection of high-resolution fMRI studies that show evidence of the importance of subcortical structures [78].

Different studies hypothesise that SI evolves core speech processes such as phonetic retrieval and encoding, syllabification and articulation prompting with a suspected reference copy responsible for the kinesthetic experience of inner speech. Multiple studies agree on the existence of common areas for these speech variations which include the inferior frontal cortex, sensorimotor cortex (motor and premotor regions), and temporal parietal-junction with a left hemisphere predominant role [1, 5–11].

2.3 Literature Review Methods

To examine this topic, we followed the Preferred Reporting Items for Systematic review and Meta-Analysis (PRISMA) guidelines [79]. In this section we describe how the study selection process was carried out and introduce Information Transfer Rate (ITR) as the metric for evaluation. Due to the substantial differences in the data across studies, a direct and fair comparison of the decoding approaches is not feasible. Nevertheless, a partial comparison is proposed by grouping the studies according to the primary type of extracted features and ranking them according to their their reported or estimated ITR.

2.3.1 Study Selection

We searched within Google Scholar and PubMed databases to identify papers reporting imagined speech decoding attempts, the search was run from August 2023 to October 2024 with the following search queries:

1. “Speech Imagery"
2. “Speech Imagery" AND (Classification OR Decoding OR Recognition)
3. (“Speech Imagery" OR “Inner Speech") AND (decoding OR EEG OR ECoG OR MEG OR fNIRS OR fMRI OR BCI)
4. “Linguistics BCI"

We first screened each result from the databases including any paper describing work related to covert and overt speech decoding. We then filtered our results using the criteria set out in Table 2.1.

To further identify related articles that were not found via our initial search queries, we analysed cited references that described speech imagery decoding attempts or results, we ended up with a list of 104 articles which describe decoding pipelines for covert speech. Figure 2.1 shows a flowchart of the selection of records along with the number of studies identified during the screening process.

Information Transfer Rate

ITR is a widely accepted metric in BCI research; it quantifies the effective amount of information that a system can reliably transmit per unit of time. It is considered an optimal metric to report performance as it accounts for accuracy and a decoding time frame [80]. ITR uses the number of SI classes the decoder is attempting to label, the time window selected from the signals, and the reported decoding accuracy, as defined in [81] by:

$$B = \log_2 C + p \ \log_2 p + \log_2 \left(\frac{1-p}{C-1} \right) \tag{2.3.1.1}$$

Table 2.1: Screening criteria used for query results

Include	Exclude
1. Studies describing an attempt to develop and evaluate a model for imagined speech decoding in human participants 2. Studies clearly describe the methods used and the results in terms of accuracy/efficacy.	1. Studies researching neural representations of Speech Imagery without classification attempts 2. Studies describing decoding of perceived speech, auditory attention, overt speech or listening/-motor imagery 3. Papers that skip stimulus detail on experiment design 4. Papers on pre-print version without peer review. 5. Report is a review, position, or discussion articles

$$ITR = \frac{B}{T} \tag{2.3.1.2}$$

Where C denotes the number of classes, p denotes the classification accuracy and T the time window in seconds used for decoding. It is expressed in bits per minute in BCI evaluation.

ITR is based on the assumption of discrete choices and is well suited for applications involving a fixed set of predefined options. However, in the context of BCI communicastion systems, an ideal objective is the generation of continuous speech.

Word Error Rate WER has been proposed as BCI Speech Synthesizers performance measure and utilised to evaluate real-time decoders of attempted speech [82–84]. WER is a widely recognised metric to assess the performance of machine translation, quantifying the proportion of incorrect words relative to a reference sentence [85]. Nevertheless, we consider ITR better suited for our analysis, as the majority of decoding approaches examined in this study are offline and are designed to decode brain signals corresponding to discrete units of SI.

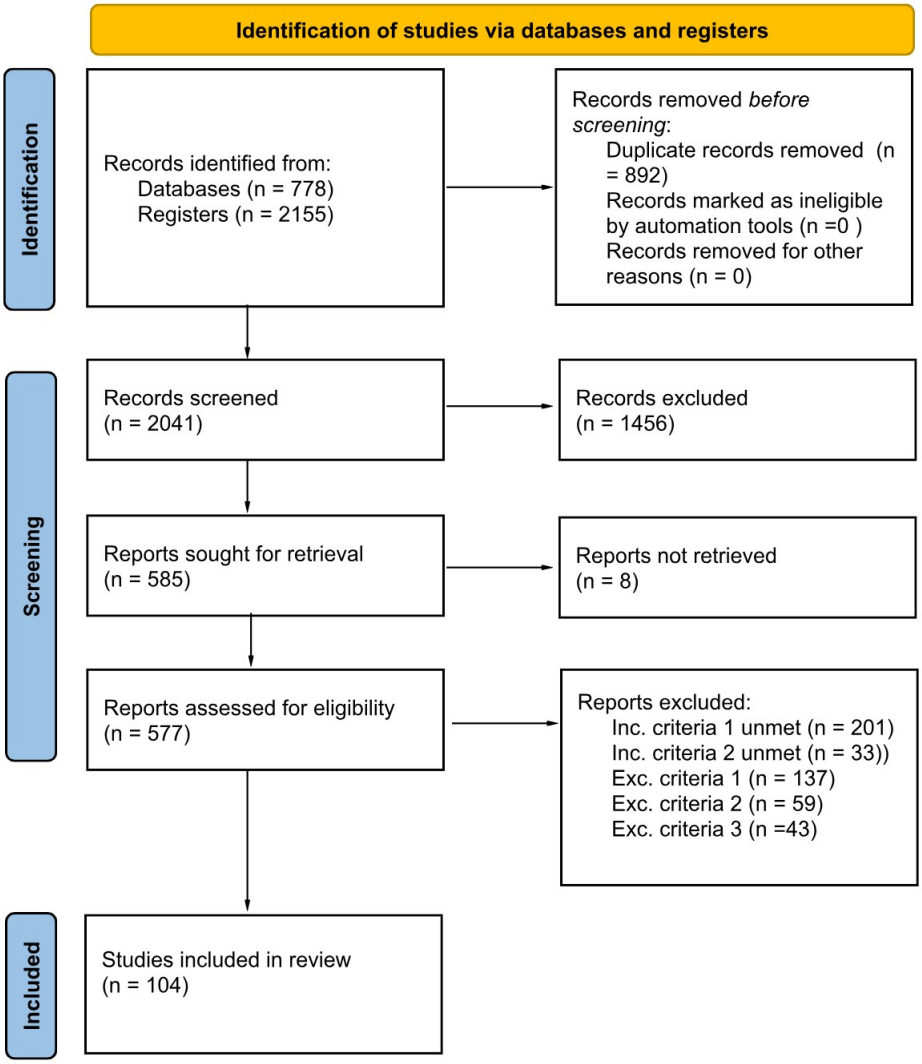

Figure 2.1: The flowchart describing the selection steps of the studies analysed in this review

2.4 Results

We first review the distribution of reports found across time to check for any changes in interest in SI. We then analyse the neuroimaging methods that have been employed to record SI-related signals. We enumerate available datasets that have been published and re-used. We analyse the experiment designs used to prompt participants and lastly, we report the signal processing and feature extraction techniques used in the attempts with a section dedicated to the prominent use of Deep Learning techniques.

Figure 2.2.a shows the time distribution of reports found, an important jump in

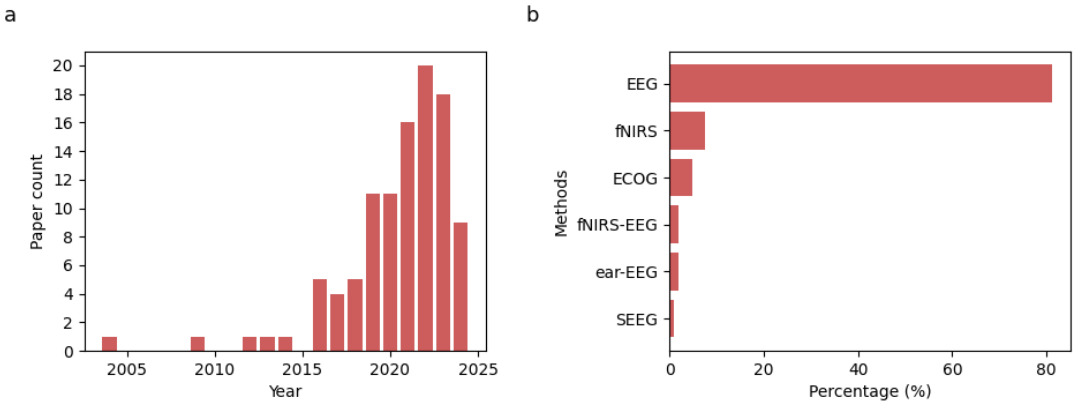

Figure 2.2: **a** Histogram of distributions of retrieved papers of reports attempting SI decoding over time. The first paper was found in 2004 and a significant increase in results is seen from 2019. Our retrieval was conducted in the middle of 2024, therefore there might be further papers published after our search. **b** Histogram of the selection frequency of each neuroimaging method, 82.26% of SI decoding reports used EEG.

the number of reports happened since 2018, several years after the first reported attempt at SI decoding back in 2004, this is because in 2015 and 2017 three open SI datasets were made available, since 2021 we have found a consistently larger number of publications. Our retrieval was conducted in the middle of 2024, therefore there might be further papers published after our search.

2.4.1 Neuroimaging methods

Five neuroimaging techniques (EEG, fNIRS, ECoG, and SEEG) and multiple different decoding approaches have been explored in the reports we retrieved that attempt to classify inner speech from brain activity.

Most reports on SI decoding (82.6%) have used EEG data, due to EEG's good temporal resolution and portability, it is relatively cheaper and easier to use than other techniques, it is the most feasible neuroimaging device for labs to acquire even despite its drawbacks such as low spatial resolution and noise sensitivity [44, 86, 87]. EEG data is used in 84 studies we identified in our review, another reason for its popularity is due to open-access SI datasets recorded with EEG that we discuss in Section 2.4.2. Figure 2.2.b shows the distribution of the number of studies involving different neuroimaging techniques used in SI decoding approaches.

Of the other studies (7.3%) used invasive neuroimaging methods such as ECoG and SEEG. Because these techniques are based upon the implantation of recording electrodes within the brain, the signal-to-noise ratio of these techniques is considerably better than non-invasive neuroimaging techniques such as EEG.

ECoG has been employed in five studies (4.7%). For example, Wandelt et. al [88] utilised signals recorded during SI of six words and two pseudowords from participants implanted with microelectrode arrays in the supramarginal gyrus and somatosensory cortex. Invasive approaches, in general, have yielded promising results due to their superior spatial and temporal resolution. These methods often rely on relatively simple features that are sufficiently informative to produce significant decoding outcome. Unlike non-invasive approaches, which typically require more elaborate feature extraction techniques. For example, the study by Martin et al. [7], where the envelope of the high-gamma band derived from stereotactic EEG (SEEG) recordings showed clear and distinguishable modulations, enabling the decoding of two SI words with an accuracy of up to 88%. Similarly, Angrick et al. [24] applied a logarithmic transform to the high-gamma power of SEEG signals to develop a closed-loop BCI capable of synthesising speech, achieving a statistically significant correlation between the intended and generated output.

The relatively limited number of published invasive neuroimaging studies focused on SI can be attributed to the practical and ethical constraints associated with these modalities. Such studies typically involve participants who have undergone electrode implantation for clinical purposes. For instance, Ment et al. [24] conducted research with a participant diagnosed with intractable epilepsy who had SEEG electrodes implanted for clinical monitoring. Likewise, the participants in Martin et al. [7] had subdural electrodes implanted as part of presurgical evaluation for epilepsy treatment.

fNIRS has also been demonstrated to allow decoding of SI-related activity[7, 89–92]. fNIRS uses beams of light in the near-infrared spectrum to measure oxygenated and deoxygenated haemoglobin levels in the cortex via changes in the refracted and reflected light. For example, Hwang et. al [89] showed the

capability of fNIRS for real-time SI decoding of “yes" and “no" words with an average accuracy of 73% ± 9.4.

Multimodal approaches have also been attempted to decode SI. For example, Rezazadeh et. al [26] combined fNIRS and EEG to classify SI of “yes" and “no", achieving an accuracy of 80.4% ± 19.1 that proved significantly better than either of the modalities alone. Cooney et. al [93] classified 4 different words showing significant improvement when combining fNIRS and EEG.

2.4.2 Open datasets

Some researchers who acquired EEG data in covert speech studies have granted open access to their data allowing other research groups to attempt decoding and to evaluate different decoding methods. We found that 38 out of 84 EEG-SI decoding reports acquired their own data, 2 used internally shared data from their research group/lab and the remaining 44 reports made use of open-access datasets.

We have listed open datasets found in our review, and the corresponding data descriptors below:

1. Wang et. al [94], published in 2013, recorded data from 8 participants performing SI of 2 monosyllabic Chinese characters. The data set is publicly available upon request to the authors.

2. KaraOne dataset [23], published in 2015, includes data from 8 participants performing SI of 7 phonemes and 4 words. This dataset is publicly available at `https://www.cs.toronto.edu/~complingweb/data/karaOne/karaOne.html`

3. Coretto et. al [95], published in 2017, contains records from 15 participants imagining the pronunciation in Spanish of 5 vowels and multi-syllabic words. After further investigation, the data is publicly available at `https://sinc.unl.edu.ar/downloads/imagined_speech/`

4. Nguyen et. al [22], published in 2017 contains EEG recorded from 15 participants performing SI of 3 short words (monosyllabic), 2 long words (trisyllable)

and 3 vowels. The dataset is available at https://www.dropbox.com/s/01k9c75j0x3jfb9/dataset.zip?dl=0

5. The International BCI Competition in 2020, made available an SI dataset containing data from 15 participants who imagined five common English words. The dataset is available at https://osf.io/pq7vb/

6. Nieto et. al [96], published in 2022, includes EEG recorded via 136 channels from 10 participants performing SI of 4 Spanish words and a rest condition. To the best of our knowledge, no studies attempting decoding on this dataset have been published to date. The dataset is publicly available at https://openneuro.org/datasets/ds003626/versions/2.1.2/download

7. Liwicki et. al [32], published in 2023, reports the first open dataset considering a bimodal approach, that records SI of 8 words from 4 participants using fMRI and EEG. No studies have been published with decoding results using this dataset. The dataset is available at https://openneuro.org/datasets/ds004196/versions/2.0.2

2.4.3 Experiment design

Experimental design is a critical step in BCI research. The classes to decode are decided in this step along with other aspects of the data collection process that can result in different participant behavior and impact the data quality. This section discusses two important aspects of an SI experiment design: speech units and stimulus presentation.

Speech Units

SI decoding models aim to identify units of speech a person is imagining at a specific moment in time. Therefore, SI-BCI studies begin with an experiment designed to instruct participants to focus their attention on specific units of speech (e.g., syllables, words or phrases) during a specific, time-bound period while their neural activity is recorded. The recorded signal is then processed to isolate the signal

Table 2.2: Frequency of speech units used in SI experiment designs.

Count	Class	Prompts
38	words	hello, help me, thank you, yes, no, one
16	vowels	a, e, i, o, u
9	phonemes	m, n, ba, fo, le, ry, gi
7	commands	go, up, down, right, left, select, stop
7	words + phonemes	iy, uw, piy, tuy, diy, pat, pot, knew, gnaw
6	words + vowels	in, out, up, cooperate, independent
4	phrases	that is perfect, how are you, goodbye, I need help

of interest, extract and select discriminative features, and subsequently train a classification model and evaluate its performance based on how accurately the recorded samples can be labelled to the speech unit.

Different speech units have been used as prompts to design SI decoding experiments. These typically range from small units, such as phonemes, to long words and phrases. The experiment designs aim for speech units that have different phonemic characteristics that can be projected and decoded from brain signals. For example, ECoG studies have shown clear differences between phonemes in the ventral Sensory-Motor Cortex (vSMC) region of the brain [97, 98] suggesting these speech units may work well in experiments using this neuroimaging modality. Table 2.2 shows the frequency of use among the different categories of units of speech in the literature.

The five vowels (/a/,/e/,/i/,/o/,/u/) have been used in 16 SI studies. In the open dataset by Coretto et. al [95], Spanish vowels were selected because of their acoustic stationarity and lack of individual semantic meaning. DaSalla et. al [21] reported one of the first approaches to single-trial classification of SI, in which they chose to decode the /a/ and /u/ vowels due to their similarity of the muscle activations involved in articulating these vowels. Gosh et. al (2022) [99] used the Bengali version of these vowels, justifying the selection of these vowels due to their ease of utterance.

Syllables have been chosen for use in SI studies by considering differences in their phonetic characteristics and consequent combinations of language muscles. The KaraOne dataset [23] used the phonemes (/iy/, /uw/, /piy/, /tuy/, /diy/, /m/, /n/). Jahangiri and Sepulveda [100] used the phonemes (/ba/, /fo/, /le/, /ry/) that were abbreviations of commands (back, forward, left, right). Zhang

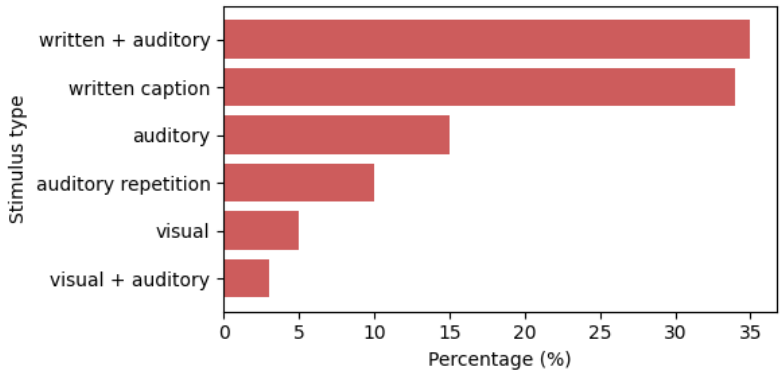

Figure 2.3: Histogram of selection frequencies of stimulus types, the written caption is the most preferred stimulus type in SI experiment designs, while visual stimuli that involve picture naming is the least preferred.

et. al (2020) [101] used the phoneme /ba/ and asked their participants to utter it using 4 different tones (/bā/, /bá/,/bǎ/, /bà/), similarly, fNIRS work by Guo and Chen [90] tried 4 different tones applied to the Chinese equivalents of the vowel phonemes (/a/,/e/,/i/,/o/,/u/).

Different words have also been explored as prompts for SI tasks, reports in [102–104] have used commands and directions (e.g., up, down, select, forward, etc.) for SI decoding. Participants were presented with closed questions, which they imagined answering with the words /yes/ or /no/ in [26, 89, 91]. Simple common English words such as /hello/ or /thank you/ have also been employed [105–107].

Lastly, some ECoG approaches have asked participants to perform SI of full sentences to isolate all available English phonemes [82, 84].

Stimulus delivery

The presentation of a stimulus to a participant causes differences in neural activity, and these differences can leak into the decoding window, which can mislead or confound the decoding of the activity of interest when the cue is included in the period of decoding. Such influences of the stimulus have not been widely investigated, and it is considered important that stimulus-evoked potentials, such as Event Related Potentials (ERPs), are handled to ensure they do not influence the decoding performance [93, 108]. Three modalities to cue imagery tasks have been

explored: text stimuli as written captions, visual stimuli such as object pictures, and auditory stimuli such as a natural voice uttering the intended unit of speech. Figure 2.3 shows the frequency of uses of the different types of stimuli modalities.

Images have been used as indirect representations of words when users have to perform picture-naming tasks [34, 103, 109]. Picture-naming-induced imagery tasks may be more effective as stimuli as they encompass a picture identification stage that could induce more prominent activation of learning and retrieval processes [110].

Audio stimuli are commonly used for SI experiments, as they have the practical advantage of demonstrating the intended pronunciation to the participants, as when not sufficiently practised participants may mispronounce the units of speech. However, this type of stimuli may also entail some downsides. Specifically, the brain's response to the auditory stimulus, which happens in the auditory cortex in a temporal region close to speech-related areas, may be included in the recorded signal of interest and mislead the analysis. Another consideration related to audio stimuli is that hearing someone else's voice may cause less natural speech imagery attempts [111].

Text stimuli have been the most frequently used within the reviewed articles. Text is perhaps the most practical way of presenting a stimulus as this may be done via a written caption displayed statically on a monitor. It is also more consistent than pictures, which could be subject to variable interpretation. As with any visual stimulus, text induces changes in activity in the occipital lobe. However, as this part of the brain has not been linked to speech production or comprehension this activity may not have a big impact on our signals of interest [112]. In the case of text stimuli, experiments have also been designed such that the task involves participants performing a silent reading or performing the imagery task a few seconds after the stimulus. With auditory or picture-naming, the task is specific to covert speech generation based on memory retrieval.

Some experiments have combined two stimuli modalities. For example, Zhao and Rudzicz [23] used a text prompt and its corresponding audio utterance to ensure correct pronunciation. Nguyen et. al [22] cued the imagery task with a written caption alongside a periodic beep to mark an activation rhythm.

A common approach for combined stimuli is the use of masking in which the intended imagery task is first prompted, then masked, before cuing the SI onset with another stimulus. For example, Jahangiri and Sepulveda (2017) [100] showed images of arrows as stimuli and after a few seconds the imagery was cued by an auditory stimuli. Park and Lee [113] prompted the participant to imagine the intended vowel via an audio cue, and after 1 second, cued the participant's imagery period with an auditory beep.

The use of combined stimuli for masking may bring the advantage of higher cognitive workload as a step of memory retrieval is involved, which may help isolate the imagery-induced activity from the prompt identification process, while a possible downside may be the risk of imagery task mismatch by the participant.

Cooney et. al [93] investigated different types of stimuli presentation. They found that presenting an image for participants to name led to the highest classification compared to auditory or text prompts.

2.4.4 Signal Preprocessing

Depending on the recording modality/ies used, different preprocessing methods have been employed to improve the signal-to-noise ratio by removing bad recording sections, referencing the data to neutral channels or applying average referencing, filtering frequencies of interest, and reducing the signal dimensionality.

For electromagnetic-based records (such as EEG), it is common to filter the signals in frequency ranges that are thought to capture cognitive-related activity. For example in the case of ECoG, filtering has focused on allowing high-frequency ranges above the gamma band (>70 Hz)[87, 114]. It is common not to see further preprocessing of ECoG signals other than windowing to isolate the SI-related potentials and due to the high signal-to-noise ratio of this signal it is common to feed the raw signal directly into classification models [84, 88].

Filtering is a broadly used preprocessing technique. Power spectrum density (PSD) analysis shows the power distribution of the signal over a range of frequencies. It is common for EEG signals to find MI-related activity in the same frequency

range as the mu (8–13 Hz) rhythm and a portion of frequencies in the beta rhythm (13–30 Hz) and harmonics of the mu rhythm [115, 116]. It is also common to find power peaks at 50 or 60 Hz that come from power noise, this noise is usually filtered out by a notch or band-pass filtering.

Our review revealed a wide selection of frequency ranges from which SI may be decoded ranging from 0.1 to 150 Hz for EEG signals. Some studies have experimented with the performance of their decoders focusing on different frequency bands. For example, Jahangiri and Sepulveda [100] studied the contribution of different frequencies in SI, and showed that the high gamma band activity leads to lower classification power but encompasses the highest number of features among the evaluated frequency bands. However, the gamma band led to the best classification accuracies in work by Min et. al [117]. Kaongeon et. al[104] concluded that the gamma and delta bands had the highest F-score when classifying an imagery task against a resting state. Lee et. al [118] tested three different classifiers with different groups of frequency bands, the wide gamma group (30–125 Hz) led to the best classification accuracy results. Kambale et. al [119] tried with 6 different frequency ranges to feed a deep learning model, their result suggested that the gamma range (30–100 Hz) gave the most informative features.

Figure 2.4 shows the frequency distributions among frequencies from 0 to 150 Hz reported in studies using EEG. Some of the studies (25 reports) used the whole frequency range or did not specify the most informative frequency band.

In intracranial studies, signal resolution due to direct cortex contact allows researchers to focus on higher frequency bands. For example, work by Meng et al. [11] divided the gamma frequency band range into 4 sub-bands covering from 30–195 Hz or Willet et. al [84] focuses on markedly high frequency bands, specifically filtered the signal from 200 to 5000 Hz.

Electromagnetic-based signals are very easily corrupted by electrical potentials generated by muscular movement. Electromyography (EMG) generated by muscles presents potentials in the order of several mV, which is easy to detect in EEG recordings (which are typically in the order of several μV). However, EMG may

often lie in the same frequency range as the EEG signals of interest and this presents a challenge when trying to remove EMG. Independent Component Analysis (ICA) is one of the most common source separation techniques used to identify and remove artifacts, especially eye blinks or head movements, we found that ICA was used as the main artifact reduction technique in 6 studies [93, 101, 120–123].

Another common preprocessing technique is down-sampling, this helps to reduce the data dimensionality thereby reducing the computational cost of processing the data. Liwicki et. al. [122] have highlighted the importance of down-sampling when applying deep learning methods, an EEG signal originally recorded at 1024 Hz was down-sampled to 128 Hz, leading to better classification performance via a convolutional neural network (CNN).

For fNIRS signals the use of light as the medium of measurement has the advantage of not being as sensitive to physiological, or motion artifacts in the way many other neuroimaging modalities are. The hemodynamic response has frequency content predominantly below 0.5 Hz, after converting raw optical intensity to a measure of haemoglobin, an increase in activity of around 1.6 Hz can be seen due to the person's heartbeat and at 1 Hz due to spontaneous oscillations in arterial blood pressure called Mayer waves [124]. Bandpass filtering may also be used to focus on frequencies from 0.05 to 0.7 Hz. Another common artifact is found in a range from 0.2 to 0.3 Hz due to respiration [26, 90, 93]. Hwang et. al[95] used common average reference (CAR) on (Oxygenated Haemoglobin) HbO, to help reduce this noise component.

2.4.5 Feature Extraction and Classification

Multivariate analysis methods are predominantly used, as modern neuroimaging systems allow multiple-channel recording. All the SI studies we identified are based on multichannel signals. These multivariate signals may be used to help find relations between different cortical regions.

The analysis of SI-related signals involves forming an array of features that uncover the representation of neural activity related to speech processes such as

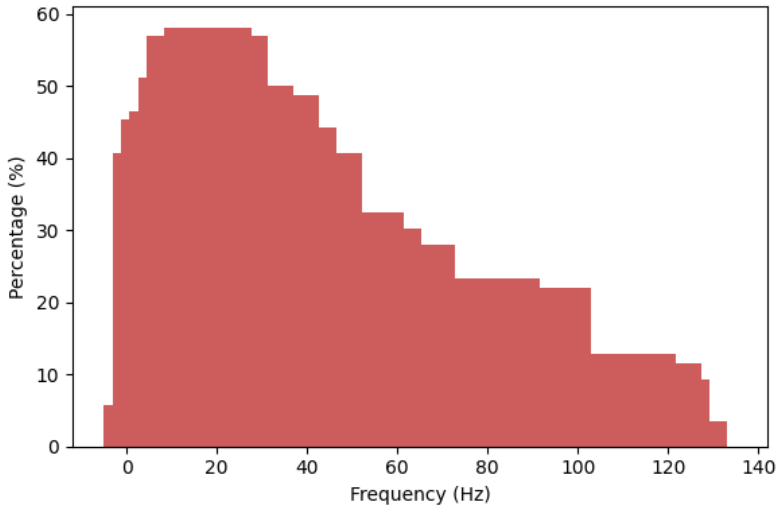

Figure 2.4: Distribution of focused frequencies in EEG. Most approaches focus on the alpha (8–12 Hz) and beta bands (12–30 Hz), but wider ranges have been investigated.

information retrieval, syllabification or articulation, among other cognitive tasks [3]. These features can be grouped based on what they represent, including temporal patterns like event-related potentials (ERPs) and oscillations, spatial details through cortical localisation, spectral characteristics such as oscillatory frequencies, and connectivity measures revealing brain network interactions.

After feature vectors are formed, the decoding pipelines usually involve machine learning models that aim to find patterns in these features to decode the speech imagery condition. In this section, we discuss the different types of features and the methods used to extract them from raw signals, along with the machine learning models employed for classification.

Spatial Features

Spatial features represent information about specific brain regions and their involvement during SI. These features can be extracted in different ways.

One way is to choose specific brain regions potentially active in SI. This could be done before the data collection takes place, as is the case when choosing the locations in which to implant of implant ECoG or SEEG electrodes, or by selecting a subset of channels that project from those regions, therefore restricting the dimensionality after data recording.

Table 2.3: Summary of reports with Common Spatial Patterns as the main feature extraction technique along with their estimated ITR in bits per minute.

Report	Number of classes	Record	Subject Dependent	Features and methods	Classification	ITR (bits/min)
[94]	2	EEG	SD	CSP	SVM	1.43
[109]	3	EEG	SD	Graph features	SVM	1.70
[21]	2	EEG	SD	CSP	SVM	3.49
[90]	2	fNIRS	SD	Common spatial, activation and connection	SVM	5.76
[125]	12	EEG	SD	Bank filters, CSP	SVM	15.30
[101]	4	EEG	SD	CSP	SVM	44.96

Another way is by applying a spatial filter that generates new channels that highlight the activity from regions of interest. Common Spatial Patterns (CSP) is a commonly used technique that produces a new filtered space based on the variance of activity between different conditions, it does so by solving the generalised eigenvalue problem where the covariance matrices are computed from the mean of trials of different imagery tasks [126].

Following its popularity in MI, CSP was first used in the context of SI by DaSalla et. al [21]. Due to its high performance it was used then in several further early attempts at SI decoding. Table 2.3 describes the reports that have used a CSP-based decoding pipeline. All these reports also use Support Vector Machines (SVM) as classifiers and make use of either the average power of the obtained CSP filters as features or compute additional statistical values from the CSP-derived feature set, such as reported in the fNIRS work by Guo et. al [90].

Spatial features have been chosen with the feature selection process by assigning weights to features from all channels and retaining the ones that are most relevant for classification or, alternatively, by iteratively testing different sets of vectors in order to optimise the classification accuracy.

Only 33 reports from those included in our review (33%), have mentioned which brain regions were most relevant for SI decoding, either during recording, channel selection, filtering, or feature selection. Figure 2.5 shows a spatial histogram of brain regions used for SI decoding. Based on the specific channels and brain locations

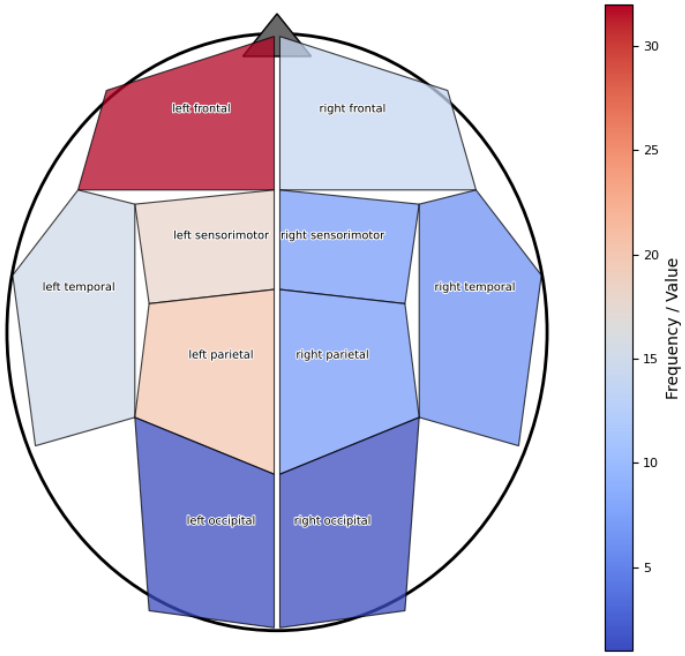

Figure 2.5: Spatial colour map of general brain regions showing the frequency of selection of features from each of these areas. Features from the left-frontal area of the brain were described in more than 20 reports as informative for SI decoding his area of the brain corresponds to Broca's area.

mentioned in the studies we have grouped the brain into 9 different regions (left and right frontal, temporal, sensorimotor, parietal, and occipital regions). Consistent with the literature, the map shows a predominance of the left hemisphere and Broca's area in SI decoding.

Spectral Features

These features describe the spectral properties of the brainwaves associated with the process of speech imagery. Some frequency bands (theta, alpha, beta, and gamma) have been linked with distinct conscious states of the brain. One direct method of extracting spectral features has been bandpass filtering, which is usually performed in the preprocessing step. As we can see in Figure 2.4 the majority of

Table 2.4: Reports with Frequency Decomposition Methods as the Main Feature Extraction Step along with their estimated ITRs in bits per minute.

Report	Number of classes	Record	Subject Dependent	Features and methods	Classification	ITR (bits/min)
[127]	5	EEG	SD	DWT	LDA	0.98
[128]	6	EEG	SD	DWT	SVM	1.73
[91]	3	fNIRS-EEG	SD	DWT, HbO	RLDA	2.90
[129]	2	EEG	SD	DWT, energy sum, waveform length	LDA	3.77
[130]	2	EEG	SD	FFT, amplitude of each frequency	SVM	4.16
[120]	5	EEG	SD	DWT, MaxLCor	SVM	4.80
[102]	3	EEG	SD	WPD	LightGBM	5.41
[131]	2	EEG	SD	CSP	kNN	5.66
[103]	2	EEG	SD	DWT	DNN	7.80
[123]	5	EEG	SD	DWT	Random forest	9.33
[132]	11	EEG	SD	MFCC	SVM	9.60
[133]	10	EEG	SI	FFT	RF	15.22
[134]	4	EEG	SD	DWT, CSP	ELM	21.92
[135]	8	EEG	SD	DWT, CSP	SVM	50.08
[136]	26	EEG	SI	DWT, CSP	SVM	91.49
[88]	300	ECoG	SD	FFT	LDA	139.15

EEG-related reports have focused on the alpha (8–12 Hz) and beta bands (12–30 Hz). However, multiple studies have also focused on the gamma band. The invasive studies emphasise rapid frequencies, as they have shown clear dynamical differences in the high gamma bands as in work by Angrick et.al [24] that used signals in the range 70–170Ḣz or Leuthardt et.al [137] that focused on singles from 40 to 160Ḣz. Table 2.4 groups the reports that have used a frequency decomposition technique as an important feature extraction step in their decoding pipelines.

For fNIRS approaches, the hemodynamics occur in a low-frequency range and, features are extracted from a narrow portion of the spectrum mainly below 0.5 Hz. Therefore, no other frequency decomposition techniques have been employed.

The Fast Fourier Transform (FFT) is a widely used technique for frequency decomposition, converting the time-domain signals into coefficients of frequency representation. Such coefficients have been directly used as features as in the ECoG based SI decoding approach reported by Mugler et. al [88] and Bejestani et. al [130].

The Mel Frequency Cepstral Coefficients (MFCC) were introduced for speech in audio processing, because of the $1/f$ property of sound, MFCC proposes scaling to balance high-low frequency amplitude contributions by applying a filterbank and logarithmic operations based on the human audio scale. MFCC has been proposed to extract EEG features and was applied by Mini et. al [138] to extract SI features. Rusnac et. al [139] used a slightly different version of the MFCC equation considering a lower dynamics scale.

The Discrete Gabor Transform (DGT) is a case of a short-time Fourier transform that uses a Gaussian function to obtain the frequency domain representation of a signal. It was used in work by Jahangiri et. al to decompose the signal into 2 Hz components to rank their classification power [100]

Wavelet Decomposition is another method widely used to decompose EEG signals, it addresses the limitations of FFT as decompositions include temporal information. Wavelet Decomposition uses a family of function wavelets that scale down the original signal by applying a convolution series. There have been different types of Wavelet Decomposition used to decode SI such as the Discrete Wavelet Transform (DWT). Continuous Wavelet Transform (CWT) and Wavelet Packet Decomposition (WPD) are two types of wavelet decomposition that differ in the scaling and type of wavelet usage. Some SI decoding reports have preferred the family of Dabeuchi 4 (db4) wavelets as it led to optimal performance [99, 102, 135, 140, 141]. However, Biorthogonal and Symlet wavelet families have also been explored for SI decoding.

Connectivity Features

These features refer to statistical dependencies between activity recorded from different parts of the brain, which are interpreted as a form of functional connectivity. They can be used to provide insights into how different parts of the brain coordinate to produce imagined speech patterns.

Connectivity features can be derived from a covariance matrix analysis, covariance matrices encode the inter-channel variability during the length of a trial.

Table 2.5: Summary of studies that used Riemannian geometry either for projecting SPD matrices or in the classifiers and estimated ITRs in bits per minute.

Report	Number of classes	Record	Subject Dependent	Features and methods	Classification	ITR (bits/min)
[22]	3	EEG	SD	Tangent projection of SPDs	SVM	1.70
[104]	4	ear-EEG	SD	Tangent space projection SPD	MLELM	1.78
[142]	4	EEG	SD	CSP, tangent projection of SPD	SVM	3.69
[143]	2	EEG	SD	Correntropy SPD	MDM	6.37
[144]	5	ear-EEG	SD	CSP, tangent projection of TSMBC	MLELM	36.00

Such matrices are Symmetric Positive Definite (SPD). If SPD matrices are placed as multidimensional points they lie in a Riemannian space or manifold. Therefore, Riemannian classifiers could have an accurate distance measure, and consequently, have been used in SI decoding attempts [145]. SPD matrices can also be projected into their corresponding tangent space to construct feature vectors whose distance can be approximately Euclidean for regular classifiers [22, 104]. The SPD property is also true for other estimators for matrices such as coherence matrices or cross-spectral density matrices [143, 146]. Table 2.5 groups the reports that have used Riemannian geometry in their approaches, either with tangent projections or Riemannian distance classifiers.

Phase connectivity features have been considered for the analysis of EEG signals. For example, Panachakel et. al [147] computed the Mean Phase Coherence (MPC) and Magnitude-Squared Coherence. The resulting statistical measures of phase synchronisation between channels resulted in two connectivity matrices that were used as inputs for a DL model, achieving an average result of 91% for binary classification. Phase and amplitude connectivity were used as a primary feature in an ECoG study by Proix et.al [148]. In their approach, phase-amplitude cross-frequency coupling was computed between the phase of one frequency range and the amplitude of a higher-frequency range and achieved a higher than chance classification accuracy with coupling of the Beta band and high gamma frequency band.

Guo et. al [149] used Pearson Correlation coefficients between pairs of HbO channels to measure synchronisation between channels over the motor and frontal cortices. They found that connections were stronger in Broca's area than in other regions, and consequently selected those channels for classification.

Chengaiyan et. al [36] employed phase synchronisation measures (EEG coherence, and Partial Direct Coherence) as well as Granger Causality measures (Direct Transfer Function) and entropy as feature vectors from 5 frequency bands to feed a DL model, which reached 79% average accuracy for binary classification.

Ilipoulos and Papasotiriou [150] developed an SI decoder by using operational architectonics, a neuroscience concept of brain function, to compute from EEG windows, the abrupt jumps in EEG amplitude named Rapid Transition processes (RTPs) to compare the relations between distinct areas and obtained measures (degree, strength, weighted global efficiency, density, weighted transitivity, eigenvector centrality) that formed the final feature vector. This vector was used to train a Naive Bayes Classifier, which achieved an average accuracy of 65% for 3 classes.

Figure 2.6.a summarises the most frequently selected feature extraction methods by the reports considered in this review. The DWT has been the most frequently selected through the years of SI research. We are also interested to see how preferences for feature selection change over time. In Figure 2.6.b we explore the cumulative sum of mentions for each method. We can see a recent increase of approaches using the raw signal due to the recent popularity of DL models. We cover the reports that used DL models to extract features from SI signals in Section 2.4.5.

Feature Selection

The high dimensionality of neural data compared with the limited amount of available samples presents a considerable challenge when training classification models. In the case of EEG, a large number of channels, frequency decompositions, and further characteristics extracted from those decompositions can result in very large feature vectors and may lead to under-fitting issues for classification algorithms.

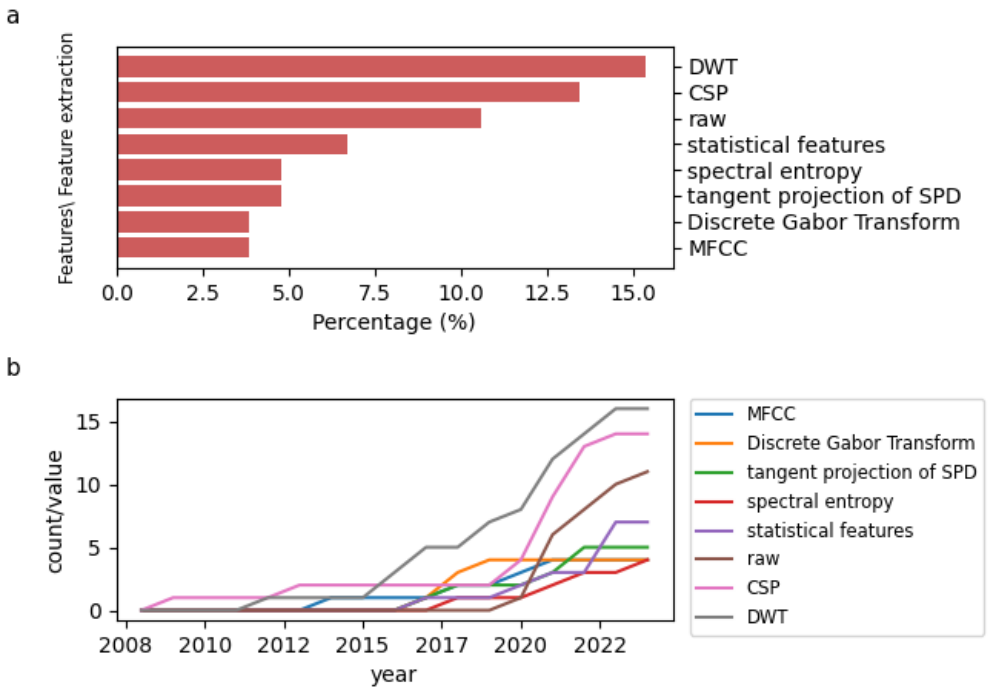

Figure 2.6: **a** Histogram of counts of the 7 most used feature extraction techniques among SI decoding attempts. **b** Timeline showing the cumulative sum of the most used feature extraction techniques over recent years, CSP has exhibited constant growth since its application in early approaches. The 'raw' feature describes approaches that use the unprocessed EEG signal to train DL models. DL approaches have become predominant in recent times.

In Figure 2.7 we have counted the feature selection algorithms used in papers selected for our review. Only 24 (22%) of the approaches we reviewed applied feature selection algorithms.

Some researchers have applied feature selection algorithms such as Principal Component Analysis (PCA) to help combine features based on their amount of variance, as used by Mini et. al [138] to derive uncorrelated components coming from MFCC coefficients to train a DL classifier. Analysis of Variance ANOVA has also been used to reduce dimensionality. For example, Macias-Macias [151] selected the most descriptive statistical features from the filtered EEG, allowing them to reduce the vector size from 264 to 30. Fisher Score has also been chosen to select features. For example, an fNIRS approach by Guo et. al [90] ranked HbO mean values using F-score to select the most important values and train an LDA. Other

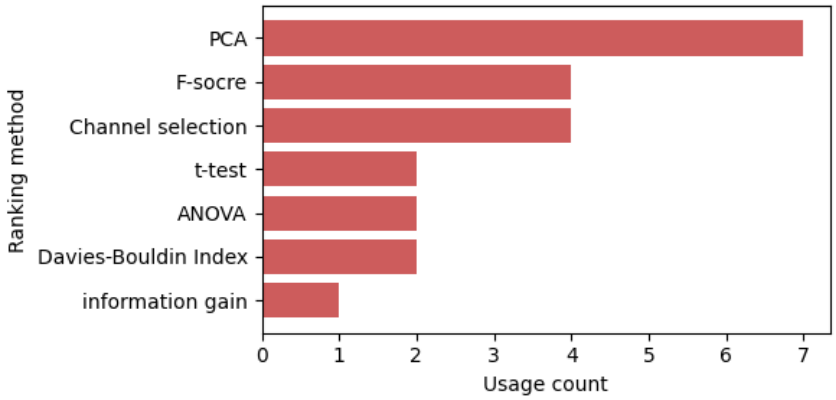

Figure 2.7: Histogram of counts of used feature selection techniques.

approaches, such as in work by Bajestani et. al [130], decided to select their features by choosing a subset of EEG channels by removing non-SI related channels based on literature, usually selecting temporal or central channels and removing frontal or occipital channels or work by Sree and Kavitha [121] that selected only channels from the left frontal and temporal hemisphere. However, these approaches do not compare against the effects of using the full set of EEG channels.

Deep Learning Approaches

Deep Learning has successfully solved numerous non-linear problems. Multilayer neural networks have also been employed for the classification of Speech imagery-related data.

Some studies have not used features fixed a-priori and instead have used the raw signals to train a neural network to extract either temporal, spectral or spatial features. For example, Cooney et. al [103] used raw EEG to train three Convolutional Neural Networks (CNNs), a shallow CNN whose temporal (2D convolution) and spatial (deepwise) convolutions are hypothesised to be analogous to bandpass and spatial filtering stages or a deeper version of the same CNN inspired by computer vision networks, is designed to extract broader features from EEG [152]. Without constraining the feature types, CNN has shown sensitivity to phase and amplitude features in the signal. Different CNN architectures have been proposed such as EEGNet [153]. This compact architecture has depth-wise and spatial convolution

layers that act as a filter bank approach, this architecture have been used to decode different EEG paradigms.

Min et. al [117] mentioned the overfitting phenomenon of EEG as the number of samples tends to be much smaller than the dimensionality of features, so they augmented their available data by dividing each imagery trial of 3 s into 30 time segments of 0.2 s length with a 0.1 s overlap, and then computed further statistical features from those segments. Additionally, they used a sparse regression model to select an ideal number of features that were then classified by a single hidden layer Neural Network.

Saha and Fels [146] state that deep learning techniques such as CNN, Recurrent Neural Network (RNN) or autoencoders fail to individually learn a complex representation of single-trial EEG data. Their investigation demonstrated that it is crucial to use multichannel features, so they used cross-covariance matrices as feature vectors to feed a parallel DL architecture with an RNN on one side and a CNN on the other. These parallel architectures reached an average accuracy of 83% for 3-class classification.

Multiple other architectures have been explored, and most of them prove to be able to learn from SI features. Rousis et. al [154] introduced the Symetric Positive Definite Network (SPDNet) model for SI decoding, they proposed combining EEGNet to extract frequency features into SPDNet after transforming EEGNet output to SPD matrices. SPDNet accounts for the Riemannian geometry in the network's forward and backward operations and it is hypothesised to work better with covariance matrices as we discussed in Section 2.4.5

Table 2.6 groups the reports that have used DNN models as classifiers.

Other Approaches

A few attempts at decoding SI have used other techniques, based on temporal properties of the signal or mapping representations. For example, Watanabe et. al [9] used first Dynamic Time Warping (DTW) to realign the signals from each trial to the envelope of each stimulus. This was then, used to compute the Euclidean

Table 2.6: Summary of reports using Deep Learning models as feature extraction techniques or classifiers and estimated ITRs in bits per minute

Report	Number of classes	Record	Subject Dependent	Features and methods	Classification	ITR (bits/minute)
[8]	5	EEG	SI	raw	CNN	0.50
[139]	7	EEG	SD	CNN	CNN	0.57
[122]	6	EEG	SI	ICA	CNN	1.02
[91]	2	EEG	SD	DWT	SNN	1.23
[103]	5	EEG	SI	spatiotemporal convolution	CNN	1.32
[155]	2	EEG	SD	CSP	Caps	2.63
[156]	2	EEG	SD	MPC	SNN	2.83
[154]	11	EEG	SD	raw	SPDNET	3.03
[119]	2	EEG	SI	SPWVD	CNN	3.58
[120]	5	EEG	SD	DWT	ELM	3.79
[157]	2	EEG	SI	statistical features	DNN	4.68
[158]	6	EEG	SI	instantaneous frequency and spectral entropy	CNN	4.73
[112]	2	EEG	SD	CSP	DNN	4.76
[138]	2	EEG	SD	MFCC	SNN	6.66
[33]	2	EEG	SI	raw	CNN	7.31
[103]	2	EEG	SD	DWT	DNN	7.80
[159]	6	EEG	SD	covariance	ELM	8.83
[160]	2	EEG	SI	raw	CNN	8.96
[161]	2	EEG	SD	spectogram	CNN	10.97
[121]	5	EEG	SI	DWT	Deep belief network	11.70
[162]	5	EEG	SD	raw	CNN	15.76
[113]	5	EEG	SD	MEMD	CNN	18.81
[10]	4	EEG	SI	spectro-spatio-temporal convolution	CNN	19.30
[146]	6	EEG	SI	CCV	CNN	20.43
[140]	11	EEG	SD	daubecheis-4 (db4) wavelet	DNN	20.90
[134]	4	EEG	SD	CSP	ELM	21.92
[163]	3	EEG	SD	raw	LSTM	22.72
[83]	50	ECoG	SD	amplitude envelope	CNN	25.06
[117]	5	EEG	SD	sparse regression model	ELM	28.24
[164]	5	EEG	SI	DTCWT	CNN	28.54
[146]	8	EEG	SD	covariance matrices	RNN	34.33
[151]	11	EEG	SD	statistical features	Caps	39.10
[34]	4	EEG	SD	spatio-temporal convolution	Fully connected layer	49.47
[165]	11	EEG	SI	Grammian transformation	DCNN	53.16
[166]	11	EEG	SI	raw	CNN	99.83
[167]	5	EEG	SI	raw	CNN	100.88

distance between the standardised test data and a template waveform of each class constructed by averaging the training data belonging to the specific class. This approach reached an average accuracy of 38.5% for 3 classes.

This process was inspired by the work of Martin et. al [114] who used ECoG signals which were time-aligned to their corresponding stimulus using DTW to compute the envelope from a high gamma band using the Hilbert transform. A further pair-wise classification of these features was then performed using SVM based on Euclidean similarity measures.

Another approach is reported by Garcia-Salinas [168] who concatenated EEG trials into a single vector to then generate a codebook based on K-NN clustering. With 250 clusters they could represent the signal via code-words where each epoch was represented as a histogram with the code-words count. The set of histograms was then fed into a Naive Bayes classifier, reaching an accuracy of 59% for 5 classes.

Einizade et. al [109] made use of graph-based features, these features are based on the structural connectivity of the signal where graphs are formed with the spatial information of the channels. A Laplacian matrix is obtained giving a 3D representation of the signal. The feature space is then reduced with the matrix eigenvalues to help identify important weights in the representation. The feature vector was then classified with an SVM, in a hierarchical model via a multiclass one-vs-all scheme, reaching an average accuracy of 50% for 3 classes.

Alizandeh and Omaranpour [169] proposed a CSP-based approach by combining One-vs-One and One-vs-All approaches. The filtered features were used to train an Ensemble Learning Classifier (ELC) compound composed of four different models. Logistic Regression, KNN, DT, and SVM. Their results proved better for the ELC than the individual classifiers.

Carvalho et. al [170] introduced Delay Differential Analysis (DDA) for SI data, this method has been proposed as a fast and robust feature extraction techniques capable of finding patterns in the raw EEG signals [171]. In their work, they performed subject-dependent binary classification of DDA features with an SVM classifier, achieving an average accuracy of 85%.

Real-time decoding approaches

The transition from offline analysis to online, closed-loop systems represents a critical milestone in SI-BCI research. Despite the vast majority of literature focusing on offline generalised models, a select subset of studies (5.7%) has successfully implemented real-time decoding with user feedback. These studies span various neuroimaging modalities and demonstrate the practical feasibility of SI in interactive environments.

Rezazadeh Sereshkeh [91] demonstrated a 3-class BCI utilising fNIRS. Capturing signals from 44 measurement channels across the frontal, parietal, and temporal cortices. The study involved 12 able-bodied participants who performed covert rehearsal of the words "yes" and "no" for 15 seconds, alongside an unconstrained rest condition. The experiment was divided into two sessions: the first session utilised 36 offline trials for initial model training, followed by two online blocks of 24 trials, while the second session comprised four online blocks of 24 trials. Immediately following the 15-second imagery window, participants received visual feedback on a screen displaying their detected answer. The decoding methodology extracted the mean value of the oxygenated hemoglobin concentration change across the time window, classifying these features using ??. By the final online block, 9 out of 12 participants performed above chance, with an average online 3-class accuracy of 64.1% over the last three blocks, and top performers reaching an accuracy of 83.8

Angrick et al.(2021) [24] achieved real-time synthesis of imagined speech using minimally invasive ??. Data was acquired from a single 20-year-old female epilepsy patient implanted with 119 electrode contacts across 11 shafts. The experiment began with an open-loop training run of 100 audibly spoken Dutch words, followed by closed-loop testing runs of 100 whispered and 100 imagined words. The speech units consisted of randomly drawn short words and numbers. In a major leap for feedback strategies, the system provided real-time, continuous auditory feedback by presenting the patient with a synthesised acoustic waveform over loudspeakers while she imagined speaking. The decoding pipeline extracted high-gamma band (70-170 Hz) log power features buffered with up to 200 ms of temporal context. These neural features were classified into nine discrete energy levels across 40 mel-scaled

frequency bins using individual ?? models, which were then synthesised into audio. The reconstructed audio for SI achieved a mean dynamic-time warping Pearson correlation coefficient of 0.32, functioning significantly above the chance level of 0.17.

Kaongoen et al.(2022)[144] introduced a highly wearable BCI system employing ear-EEG. Signals were acquired using a custom 3D-printed headphone device equipped with 8 channels positioned around the ears. The study involved 11 healthy participants who controlled an interactive simulated television. The training phase required participants to complete 50 offline trials for each of six commands ("power", "up", "down", "next", "back", and "rest"), taking roughly 20 minutes. This was followed by three online testing and calibration sessions of 60 trials each, culminating in a real-world continuous scenario task. Feedback was provided visually via a checkmark or cross mark, immediately followed by the state change of the simulated television. The decoding approach filtered the ear-EEG into multiple frequency bands (15-120 Hz), applied a multiclass CSP algorithm, and utilised Riemannian tangent space projections of the covariance matrices. These features were classified using a Multilayer Extreme Learning Machine (??). In the final online session, the average True Positive Rate across all participants reached 59%, with the highest-performing participant achieving an 85% success rate and a command delivery time of 3.79 seconds per command.

Jeong et al.(2023) [161] tackled the challenge of combining multiple neural languages into sentence-level commands using non-invasive 64-channel scalp EEG. Involving 11 healthy participants, the acquisition protocol required 50 calibration trials per word, later deployed in an online phase comprising 25 high-level collaborative tasks. The decoded speech units were categorised into subjects ("I", "partner"), verbs ("move", "have", "drink"), and objects ("box", "cup", "phone"). The feedback strategy was highly interactive: the decoded sentences were transmitted as neural commands to a prosthetic arm, allowing the user to perform real-time cooperative tasks—such as passing a cup—with a human partner. The decoding methodology, termed deep neurolinguistic learning, processed 30-125 Hz EEG signals through a CNN encoder,

utilising a Gated Recurrent Unit (??) to regress the features into target mel-spectrograms. The final classification was determined by calculating the Structural Similarity Index Measure (??) between the predicted and ground-truth spectrograms. Real-time classification accuracies averaged 69.9% for subjects, 72.2% for verbs, and 73.5% for objects, yielding an overall collaborative task success rate of 72.36%.

Wu et al.(2024)[172] specifically investigated how adaptive classification can overcome the non-stationarity of EEG during real-time control. Data was acquired via a 64-channel EEG system from 20 healthy participants across two separate online sessions on different days. Following an offline training phase of 90 trials, participants were tasked with the mental imagery of two distinct syllables ("/fo/" and "/gi/"). The feedback strategy utilised a dynamic visual display of an empty battery that filled up or depleted in real-time, updated every 50 ms based on the continuous probability output of the classifier. The decoding methodology extracted PSD features between 1 and 70 Hz. Crucially, an adaptive LDA classifier continuously updated its mean and inverse covariance matrix parameters during the online control phase as new trials were completed. BCI control performance with the adaptive classifier averaged 55.7%, demonstrating a statistically significant improvement over the 52.3% baseline accuracy achieved when using a traditional static LDA classifier.

2.4.6 Summary

We analysed 104 SI decoding pipelines, observing considerable variability in experimental setups, feature extraction methods, and classification algorithms. A comparison of ITR across different recording modalities and classifiers is presented in Figure 2.8. Invasive techniques exhibited a higher median ITR compared to EEG; however, the highest ITR values in our analysis were achieved using EEG data. The results also suggest that commonly used and relatively simple classifiers, such as SVMs and LDA, can achieve ITRs comparable to those obtained with more complex DL approaches. Nonetheless, DL methods achieved the highest ITRs among all classifiers evaluated.

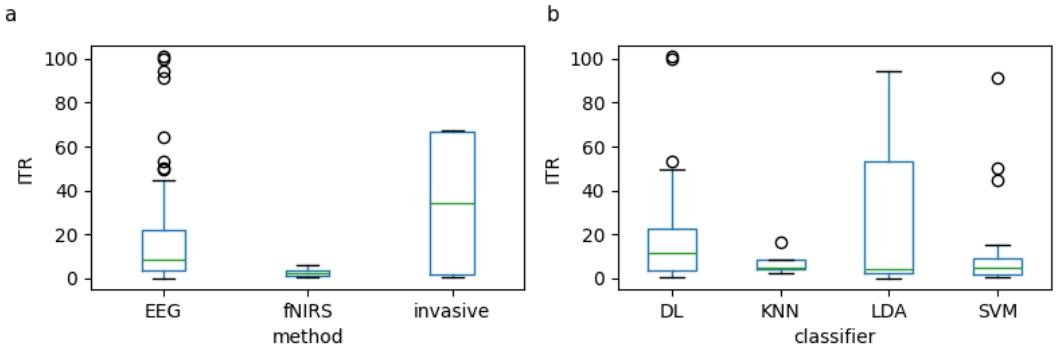

Figure 2.8: **a** Information Transfer Rate estimation for different recording modalities used to acquire SI. **b** ITR estimation for the different classifiers employed. The middle line of the box corresponds to the median ITR.

Our analysis further revealed that 23% of the studies adopted an inter-participant approach, aiming to develop decoding pipelines that generalise across participants. In contrast, 77% of the pipelines were evaluated on a participant-specific basis. Additionally, we examined dataset usage and found that only 63 studies (60%) conducted original experiments to collect SI data, while the remainder relied on previously available datasets

Furthermore, only five studies (5.7%) reported real-time SI decoding attempts with user feedback, involving different modalities (2 ECoG, 1 SEEG, 1 fNIRS, 2 EEG) [24, 91, 144, 161, 172]. A summary table of these datasets is described in 2.7.

2.5 Discusion

Speech imagery decoding has been investigated for over a decade, with several studies reporting promising results. However, as highlighted in this review, further research is necessary to consolidate these findings and to more clearly establish the maturity of SI as a viable BCI paradigm. Based on our analysis, we identify two key aspects of current SI research that may be critical for advancing the development and reliability of SI-based BCIs: Reproducibility and experiment design. Finally, we also comment on the decoding of attempted speech, a paradigm which may evoke activity in similar areas to SI.

Table 2.7: Summary of Real-Time Closed-Loop Speech Imagery Decoding Studies

Study	Acquisition & Cohort	Experimental Paradigm	Decoding Methodology	Online Performance
Rezazadeh Sereshkeh et al. (2019) [91]	**Modality:** fNIRS (44-ch) **Participants:** 12 Healthy **Data:** 36 offline train, 144 online test trials	**Speech Units:** Words ("yes", "no") and rest **Feedback:** Visual (immediate textual classification result)	**Features:** Mean [HbO] change **Classifier:** Regularised LDA (RLDA)	**Avg Accuracy:** 64.1% **Top Performer:** 83.8%
Angrick et al. (2021) [24]	**Modality:** sEEG (119-ch) **Participants:** 1 Epilepsy patient **Data:** 100 train (audible), 100 test (imagined)	**Speech Units:** Short Dutch words and numbers **Feedback:** Continuous auditory (synthesised acoustic waveform)	**Features:** High-gamma log power (70–170 Hz) **Classifier:** RLDA + Griffin-Lim audio synthesis	**DTW Correlation:** 0.32 (Chance: 0.17)
Kaongoen et al. (2022) [144]	**Modality:** Ear-EEG (8-ch) **Participants:** 11 Healthy **Data:** 50 train trials/class, 180 online test trials	**Speech Units:** "Power", "up", "down", "next", "back", rest **Feedback:** Visual (TV interface state change)	**Features:** CSP, Riemannian tangent space on covariance **Classifier:** Multilayer Extreme Learning Machine (MLELM)	**Avg TPR:** 59% **Top TPR:** 85%
Jeong et al. (2023) [161]	**Modality:** Scalp EEG (64-ch) **Participants:** 11 Healthy **Data:** 50 train trials/word, 25 online tasks	**Speech Units:** Subjects, verbs, objects (combined into sentences) **Feedback:** Visual/Physical (robotic arm movement)	**Features:** 30–125 Hz bands **Classifier:** CNN Encoder, GRU Regressor, SSIM thresholding	**Task Success Rate:** 72.36%
Wu et al. (2024) [172]	**Modality:** Scalp EEG (64-ch) **Participants:** 20 Healthy **Data:** 90 offline train, continuous online test	**Speech Units:** Syllables ("/fo/", "/gi/") **Feedback:** Visual (dynamic battery filling/depleting)	**Features:** PSD (1–70 Hz) **Classifier:** Adaptive LDA (continuously updated)	**Avg Accuracy:** 55.7%

2.5.1 Reproducibility

As identified in this review, the majority of SI decoding studies rely on offline pipelines that, in theory, hold promise for translation into online BCI applications. However, we found only six instances in which an online SI-BCI was implemented. Given that 57 studies reported conducting their own experiments, often achieving high offline decoding performance, the limited number of online implementations may point to challenges in reproducibility. Although certain feature extraction methods, such as DWT and CSP, have been applied across multiple studies, the resulting performance varies substantially, underscoring inconsistencies in implementation and outcome. Reproducibility remains a broader concern in the field of machine learning, yet we found no systematic efforts to address this issue within SI research specifically. Further development and evaluation of online SI-BCIs are needed to establish the paradigm's viability for real-world applications, particularly in the context of non-invasive approaches. Moreover, incorporating benchmarking frameworks and comparing SI with more established paradigms, such as MI, could be beneficial for its development.

2.5.2 Experiment designs

The inherently subjective nature of speech imagery may present challenges for its adoption as a robust and widely usable BCI paradigm. Estimates suggest that only 30–50% of individuals regularly experience inner speech [29], which could limit the consistency and generalizability of SI-based decoding approaches. Identifying participants who experience frequent inner dialogue may enhance the quality of data collection, particularly in early-stage studies. Psychological factors are known to influence BCI performance and should therefore be carefully considered in experimental design [173].

Effectively instructing participants to perform SI and assessing their comprehension of the task can be complex. Compounding this issue is the variability in how SI can be conceptualised and executed, ranging from visual imagery of words, auditory imagery, and motor imagery of articulatory movements to silent

naming. While several studies have explored these different strategies, no consensus has been reached regarding the most effective approach. Nonetheless, comparable neural activations and promising decoding performances have been reported across these methods [11, 24, 174].

As discussed in Section 2.4.3 there have been different approaches to SI designs, and each of them may generate a different outcome. However, based on the reported results, all different stimuli (auditory or visual) and types of SI tasks (naming, reading or generating) have been decoded with higher than chance accuracy. We found reports analysing different types of speech units but no significant evidence to suggest that prompting specific speech units leads to higher classification accuracy. [95, 103, 122, 148]. Further investigation to compare the performance of speech units may help in finding optimal prompts for SI paradigm.

It is known that machine learning algorithms may generalise better when data is abundant. We have identified that some of the most used databases in the field of SI decoding contain only a small number of trials per class, in some cases containing as little as 15 trials. Considering this small dataset size and the fact that a large number of channels are present in most datasets used in SI decoding studies, it is evident that decoding pipelines may under- or over-fit. However, another issue is that, to acquire more trials or SI classes, participants need to spend a long time in experiments, which can bring mental fatigue or stress and lead to lower data quality. One potential solution is the use of multi-session recordings, meaning that participants would need to repeat the experiment for more than one day, which in turn could also help to test model generalisation and increase the amount of data collected. Another design suggestion may be the gamification of the experiment, building something entertaining and adding a sense of purpose have been shown to be useful in BCI experiment designs [173, 175].

Neurofeedback and BCI training remain largely unexplored within the context of SI. We suggest that future research should prioritise the development of closed-loop paradigms, enabling participants to receive real-time feedback on their SI performance as a means to validate offline findings. Neurofeedback has been shown

to enhance BCI performance by facilitating users' ability to modulate their neural activity through learned self-regulation strategies [176, 177].

2.5.3 Attempted Speech

Notably, recent advances in speech neuroprosthesis have demonstrated the potential of BCIs for restoring communication in individuals with severe speech impairments. Unlike SI decoding paradigms developed with healthy participants, these approaches have primarily been applied to individuals with degenerative conditions resulting in the loss of speech. For instance, Moses et.al [83] demonstrated the feasibility of a BCI-based speech synthesizer, achieving a decoding rate of 15 words per minute with a 25% word error rate. This performance was made possible through the integration of SI decoding and language modelling to enhance accuracy. Similarly, Card et al. [82] reported a system capable of decoding a 125,000-word vocabulary with 90% accuracy, enabling a participant to engage in self-paced conversation at approximately 32 words per minute. A subsequent publication by the same group highlighted further system adaptations that positively impacted the participant's communicative experience [84]. To the best of our knowledge, these represent some of the earliest and most compelling demonstrations of speech neuroprosthesis, offering evidence that brain signals can be decoded into intelligible speech in real time. When compared to SI decoding, the superior performance observed in these attempted speech paradigms may reflect the unique motivation and engagement of participants for whom BCI systems offer a critical avenue for restoring communication.

2.6 Conclusion

Speech Imagery decoding holds significant promise for advancing our understanding of the brain's speech preparation processes and the relationship between speech imagery and cognitive thoughts. SI decoding enables the identification of covertly spoken speech units, shedding light on the neural mechanisms underlying this higher-order mental activity. Successful decoding systems open up a range of possible

applications, positioning SI decoding as a valuable tool in both neuroscience and linguistics research.

From a neuroscience perspective, SI decoding allows researchers to explore brain areas responsible for language tasks. Besides core functions such as syllabification and articulation, it also involves stages of memory retrieval and semantic conceptualisation, promoting complex interaction of the brain's networks as discussed in different speech models [25, 78], different SI models back up the responsibility in SI of overt speech-related areas. Additionally, SI may include an error correction step, akin to that which occurs in overt speech production [1, 76], all of which produce the kinesthetic inner experience of speech.

Furthermore, SI decoding holds considerable potential for practical applications, particularly as a foundation for BCIs designed for communication. In this review, decoding performance was primarily evaluated using ITR, as most of the analysed approaches attempted discrete offline classification. However, we also considered WER as a more appropriate metric for assessing speech-based BCIs, especially in the context of closed-loop systems designed to synthesise continuous speech.

We identified six real-time decoders reporting statistically significant results across different imaging modalities. These closed-loop systems demonstrated the ability to decode up to five SI classes using non-invasive techniques, and up to 100 words using invasive methods. Our findings indicate that invasive approaches not only achieve a higher median ITR but also benefit from the use of relatively simple and highly discriminative features. In contrast, non-invasive techniques typically require more complex decoding pipelines to extract informative features, as discussed in subsection 2.4.3.

Moreover, in the context of neurorehabilitation, real-time SI decoding can be coupled with neurofeedback to actively train and reinforce language-related neural pathways. By translating covert speech attempts into an explicit, perceptible representation, these systems enable users to consciously monitor their neural activity and iteratively adjust their cognitive strategies to better match predefined neural targets. This closed-loop interaction has been demonstrated in online

approaches where participants either read the decoding output [91, 174], hear the intended speech unit rendered as synthetic speech [24], or control an external device [144, 161]. Such feedback mechanisms may support operant conditioning for speech recovery and rehabilitation, opening new possibilities for assisting individuals with speech-related neurological impairments. However, it is important to note that the closed-loop approaches identified in our review focused primarily on system feasibility and did not systematically investigate how this feedback impacted or improved the participants' longitudinal SI proficiency.

Our results reflect a developing BCI paradigm that holds fundamental challenges, as evidenced by the diversity of decoding strategies and the variability in reported outcomes. One clear challenge evidenced in our results is reproducibility; we consider that the ratio of real-time decoding attempts is limited (6 studies) in contrast with the number of offline approaches, that carried data collection experiments (57 studies), suggest a gap between the experimental development and practical implementation possibly due to reproducibility challenges.

This literature review provides a comprehensive overview of the current state of SI decoding within the broader context of BCI research. While several studies report promising results, the field remains in a formative stage. The substantial variability in the studies, followed by the limited replication of findings, complicates efforts to determine the maturity of this field.

3 Methodology

This chapter presents the methodology employed to evaluate Speech Imagery. The chapter is structured to detail the complete research pipeline, beginning with the experimental design and data acquisition, followed by the core decoding pipeline used for SI classification. Subsequently, the cross-validation procedures and the thresholds established for statistical significance are described.

The experimental paradigm was designed to classify electroencephalography (EEG) signals distinguishing between two speech imagery words and a resting state. Furthermore, the protocol aimed to compare the effects of rhythmic imagery against single-word imagery. Utilising a masked stimulus paradigm, participants were instructed to internally articulate a prompted word immediately following a flash-like visual cue. The EEG data recorded during these experimental sessions serve as the foundation to address the primary research objectives explored in the subsequent chapters.

3.1 Experiment

3.1.1 Key choices

Chapter 2 reviewed studies that approached speech imagery decoding. Three key elements are considered when designing an experiment: the neuroimaging method,

the speech unit targeted for classification, and the stimulus presentation.

Neuroimaging methods

Apart from some invasive approaches, most of the approaches to SI decoding have been recorded using EEG (see Section 2.4.1). EEG is an appealing modality for SI decoding as it has the advantage of portability, and it also provides good temporal resolution.

Speech units

Speech imagery experiments have explored a range of speech units (see Section 2.4.3), from the smallest unit—a phoneme—to entire sentences. Unit selection aims to maximise differences in phonemic characteristics, although some studies have attempted classification based on word length. Phonemes have been tested either as vowels or consonants, while other approaches have selected units based on communicative intent. Our choice of the speech units was based on words that may be used as commands for a control application with phonetically disimilar characteristics. The words 'left', 'right', 'pinch', and 'stop' were selected; each has a different initial phoneme and manner of articulation, description of each is stated in table 3.1.

Table 3.1: Phonetic Characteristics of the chosen speech units for experiments.

Word	Initial sound	Manner of articulation
left	/l/	Tongue up
right	/r/	Tongue back
pinch	/p/	Lips close
stop	/s/	Tongue on palate

Mental Task

Two main approaches to SI have been explored: single speech imagery and imagery repetition. Imagery repetition is intended to strengthen the SI-related signal by having participants internally repeat a word throughout the imagery period. The repetition rhythm may be externally paced using auditory cues [22] or left to

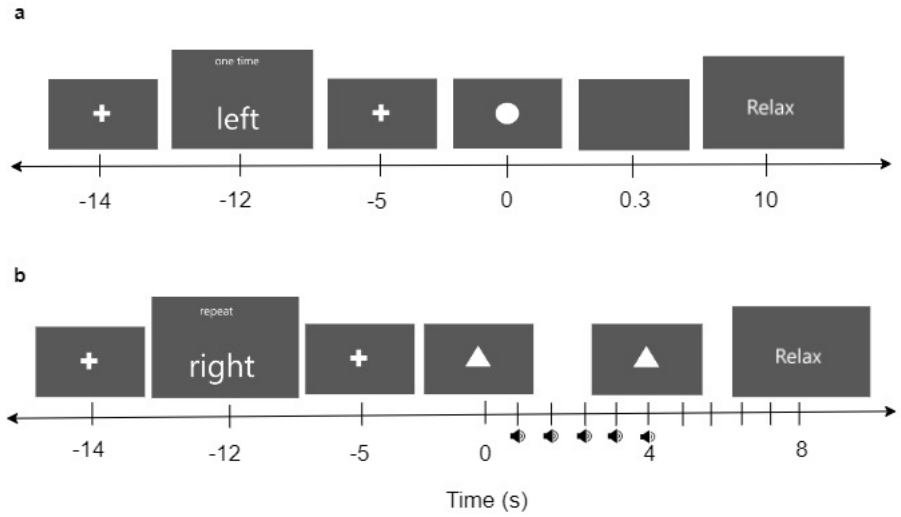

Figure 3.1: a. Timeline of the experiment for single-time imagery, b. Timeline of experiment for rhythmic imagery

the participant's own discretion [106, 149]. In this work, we investigate whether either of these mental tasks can elicit neural activity that leads to improved decoding performance.

Stimuli

Three main types of stimuli have been used to cue SI (see Section 2.4.3): visual cues, such as written captions of the target speech unit or pictures of objects to be named during the imagery period; auditory cues, such as a synthetic voice uttering the speech unit or rhythmic beeps to guide repetition; and mixed modalities that combine both. The choice of stimulus is critical, as different presentation modes elicit distinct neural responses, which can affect the time windows of interest [93, 108]. In this study, we adopted a stimulus masking approach: participants were prompted with the intended speech unit in advance, but the imagery itself was triggered by a brief flash-like visual cue. This ensured that neural responses to the prompt did not contaminate the signal or otherwise lead to inflated decoding results [178]. It also ensured that the mental task was generated deliberately by the participants.

3.2 Experimental design

Participants were seated in a comfortable chair facing a 52-inch screen. A graphical user interface developed with PsychToolbox 9.0 [179] in Matlab R2022 was used to display the prompts over a plain grey screen. Each SI trial began with a fixation cross to prompt participants to prepare for the task. This was followed by the imagery prompt, displaying the words "one time" or "repeat" to indicate single or repeated imagery, respectively, along with a large-text display of the target imagery word. The speech units were shown randomly for each trial, or each block in the rhythmic variation, for a total of 25 times per unit. Next, another fixation cross appeared to give participants time to memorise the speech unit and prepare mentally. The cue stimulus—a circle displayed for 0.3 s and perceived as a brief flash—then signalled the start of imagery.

For single imagery, the task period lasted 2 s. For repeated imagery, participants were cued simultaneously with visual and auditory stimuli marking an 800 ms rhythm for five repetitions, which formed one block. They then continued the rhythmic imagery for an additional 4 s without auditory cues before entering a rest period. The rest period varied randomly between 5–10 s. The full experimental timeline is shown in Figure 3.1.

3.2.1 Participants

Sixteen right-handed able-bodied participants (9 female, 15 male) 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).

3.2.2 Instrumentation

We collected SI-related brain activity simultaneously from EEG on each device's acquisition software and synchronised the recordings via hardware triggering.

3.2.3 EEG

The signal was recorded using a 64-channel Biosemi Active-Two system. Electrode placement was done via the international 10-20 system, plus 2 electrodes close to each eyebrow for Electrooculography (EOG) and 2 electrodes around the mastoids for Electromyography (EMG) recording. Data was recorded at a sampling rate of 2048 Hz unaffected by hardware cut-off.

3.3 Riemann Tangent Space Projection and Logistic Regression Pipeline

The tangent space projection with logistic regression (TS+LR) is a Riemannian geometry–based method that has proven successful in motor imagery classification [18]. We extended this approach by incorporating a filter bank step, defining 20 frequency bands (2–128 Hz) with 8 Hz bandwidths and 6 Hz moving windows.

For each frequency band, we computed the spatial covariance matrix for each EEG trial. To ensure the matrices were well-conditioned, especially given the small sample size relative to channel count, we applied Ledoit–Wolf shrinkage [180] to obtain regularised covariance matrices. These calculations and the subsequent mapping of the Symmetric Positive Definite (SPD) covariance matrices into the Euclidean tangent space were performed using the `pyRiemann` library [181]. This projection maps the manifold structure onto a tangent space at the geometric mean of the training data, allowing the matrices to be represented as feature vectors.

Specifically, we used a vectorisation operator to extract the upper triangular elements of the projected matrices, preserving the Riemannian metric by applying a $\sqrt{2}$ weight to off-diagonal elements. The resulting feature vectors from all frequency bands were concatenated to form the final input for the classifier.

We employed a Logistic Regression (LR) model with an L_1 penalty (Lasso), implemented via the `scikit-learn` library [182]. L_1 regularisation adds a penalty equal to the absolute value of the magnitude of coefficients, which effectively performs feature selection by shrinking the coefficients of less informative features

to zero. This induces sparsity in the model, promoting robustness and mitigating overfitting when dealing with the high-dimensional concatenated feature space. The model was trained with a maximum of 600 iterations to ensure convergence.

3.4 Evaluation

Evaluating EEG-based Brain-Computer Interfaces requires rigorous statistical frameworks, primarily due to the inherent non-stationarity of neural signals, low signal-to-noise ratios, and characteristically high inter-subject variability. When working with relatively small sample sizes, relying on simple train-test splits often leads to over-fitted models. To ensure robust, unbiased estimates of decoding performance and to mitigate the effects of variant values across sessions, we implemented cross-validation evaluation procedures.

3.4.1 Stratified Cross-Validation

For standard evaluations, we utilised a Stratified K-Fold cross-validation strategy. This procedure divides the dataset into K mutually exclusive subsets while strictly preserving the percentage of samples for each class in every fold. Maintaining the original class distribution is crucial to prevent the classifier from developing a bias toward the majority class. The implementation was carried out using the `StratifiedKFold` module from the Scikit-Learn Python library [182].

3.4.2 Group Cross-Validation

In experimental designs where trials are recorded in distinct blocks or runs, standard cross-validation methods risk artificially inflating decoding accuracy due to data leakage. Consecutive trials within the same block inherently share temporal correlations, background noise, and baseline physiological states. To control for this, we implemented Group Cross-Validation using Scikit-Learn's `GroupKFold` [**Pedregosa2011**]. By leveraging the dataset's block structure to define groups, this method ensures that all trials from a single experimental block are assigned

to either the training set or the test set, providing a realistic estimate of how the classifier performs on unseen data.

3.4.3 Pooled Accuracy

To aggregate performance across all folds, we reported the Pooled Accuracy. This calculates the overall performance by dividing the total number of correct predictions across all folds by the total number of evaluated samples:

$$Accuracy_{pooled} = \frac{\sum_{i=1}^{K} c_i}{\sum_{i=1}^{K} n_i} \tag{3.4.3.1}$$

where K is the total number of folds, c_i is the number of correct predictions in fold i, and n_i is the total number of samples in fold i.

3.4.4 Statistical Significance Threshold

To determine whether decoding accuracy was significantly better than random chance, we utilised a threshold based on the binomial distribution, specifically tailored for BCI applications [183]. We computed the adjusted Wald confidence interval (Agresti-Coull) to account for small sample sizes.

For a two-class problem ($p_0 = 0.5$), the adjusted probability $\tilde{p}$ given n trials is:

$$\tilde{p} = \frac{n \times 0.5 + 2}{n + 4} \tag{3.4.4.1}$$

The critical threshold for statistical significance at a 99% confidence level ($\alpha = 0.01$) is:

$$Threshold = \tilde{p} + z_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} \tag{3.4.4.2}$$

where $z_{1-\alpha/2} \approx 2.576$. Any pooled accuracy exceeding this threshold was deemed statistically significant.

4

Decoding Speech Imagery or just noise: A symptom of the reproducibility crisis

4.1 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 [16]. 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 [184]. 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 [185]. SI-BCIs have the potential to vastly increase the possibilities for communication aid applications if a significant number of words can be accurately decoded [148, 186].

The current state-of-the-art for SI decoding looks promising, with a trend of new decoding attempts achieving significant classification accuracies [35]. Open-access electroencephalography (EEG) datasets [23, 95] 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 [35]. 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 analysing the same input data [187]. BCI exists as a multidisciplinary junction that poses difficulties to knowledge sharing due to the domain-specific ways to share and report findings [188]. 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 [189].

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 [190, 191]. 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 [35]. 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 [192]; (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. [190]; (3) Reproducibility gap, by comparing our replication results with those reported in the original publications. For context, Menon et al. [191] 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 [191], 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 [23], 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. [95], 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 [35, 193].

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 [96].

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.

4.2 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 belong to trends on SI decoding identified in [35, 193], 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 summarised in Table 4.1. We additionally incorporate four MI datasets. Because MI decoding has been successfully and consistently replicated using the Mother of All BCI Benchmarks (MOABB) framework [18], 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 [35], 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 [194, 195]. we limited our evaluation of the larger MI datasets to their first 16 participants. Furthermore, the single-task design utilising 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 summarise the evaluated MI datasets in Table 4.2. For a detailed description of each of these and the SI datasets, please refer to A.1.

Table 4.1: Description of evaluated SI datasets.

	Kara One [23]	Coretto [95]	Nieto [96]	Our own
Classes	11	11	4	2
Trials/class	13	50	50	25
N	14	15	10	16
Ch	64	6	64	128
Fs (Hz)	1000	1024	1024	2048

Table 4.2: Description of evaluated MI datasets.

	Weibo [196]	Physionet [194]	Lee [195]	Schirrmeister [152]
Classes	7	4	2	4
Trials/class	80	23	100	120
N	10	109 (16)	54 (16)	14
Ch	64	64	62	128
Fs (Hz)	200	512	1000	500

4.3 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 4.4 and 4.3, 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 [197] 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 [182] 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 [198]. 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 [199].

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 4.5 and 4.6. 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 A.1.1). 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.

4.3.1 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 [18], mainly composed on a tangent space (TS) projection of regularised covariance

Table 4.3: 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[95]	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[135]	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[200]	Downsample to 128 Hz Filter (2-40 Hz). ICA. Scaling.	Use Raw signal	CNN	Python Sckit-learn	Within participant 5-fold cross-validation
CT4[122]	Filter (2-40 Hz). ICA.	Use Raw signal	CNN	Matlab	Within participant 80% training - 10% testing - 10% validation

Table 4.4: 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[95]	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[135]	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[200]	Downsample to 128 Hz Filter (2-40 Hz). ICA. Scaling.	Use Raw signal	CNN	Python Sckit-learn	Within participant 5-fold cross-validation
CT4[122]	Filter (2-40 Hz). ICA.	Use Raw signal	CNN	Matlab	Within participant 80% training - 10% testing - 10% validation

Table 4.5: Description of missing information in SI decoding pipelines applied to the Kara One dataset.

Report	Missing Information	Steps Taken
KO1[23]	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 compute the derivative of the vector composed of all the obtained feature values. The number of features per channel we obtained is 221.
KO2[132]	Preprocessing: It is not mentioned what ICA algorithm was used, how the noise components were selected, and how many were removed. Classification: SVM regularisation parameter C not specified	Preprocessing: We opted for the Picard ICA algorithm; we excluded one component associated with eye blinks, characterised 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 lateralisation of the component's topography. Classification: SVM regularisation parameter C was set to the default value from Scikit-learn
KO3[166]	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[169]	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 4.6: Description of missing information in SI decoding pipelines applied to Coretto's dataset.

Report	Missing Information	Steps Taken
CT1[95]	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 4.5)
CT2[135]	Classification: SVM regularisation parameter C and Kernel Coeficient *gamma* not specified	Classification: SVM regularisation parameter C set to 1 and Kernel Coeficient to $1/n$ where n is the number of features.
CT3[200]	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 4.5) Feature Extraction: Used the raw EEG signal (6 x 496) as input data for the CNN
CT4[122]	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 4.5 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

matrices using Ledoit-Wolf estimation [180] and a logistic regressor (LR), covariance estimation and TS transformation was implemented using the PyRiemann library [181]. We refer to it as the TS+LR pipeline. Further details of this pipeline can be found in A.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 standardised 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 [183] with 99% confidence ($\alpha = 1\%$) and evaluated the distribution across configurations for participants with statistically significant accuracies only.

4.4 Results

4.4.1 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

Table 4.7: 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

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 summarised in Table 4.7 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 of this analysis are shown in table A.3.

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 4.8.

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 A.1, and the results of grid-search of the number of ICA components is showed in table A.2.

4.4.2 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

Table 4.8: 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

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 4.1. 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 A.1 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 generalises well to SI features. To address this, we tested

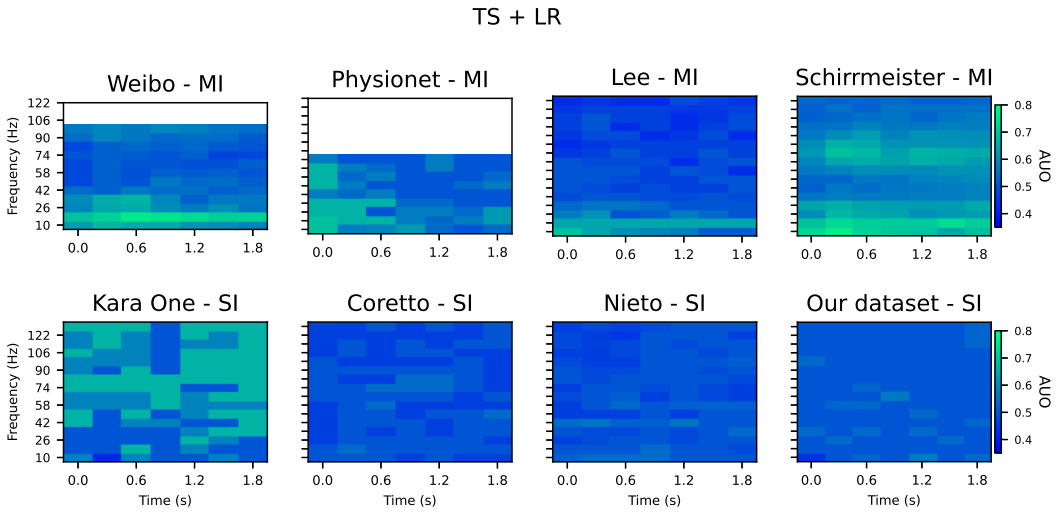

Figure 4.1: Heatmaps of median classification accuracies across participants for various time and frequency windows using the TS and LR pipeline. The top row displays results for four Motor Imagery (MI) datasets, while the bottom row displays results for four Speech Imagery (SI) datasets. The evaluated 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 ($p < 0.05$)

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 A.4.2).

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 4.2.d and 4.2.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 A.2.

The distribution of the highest decoding scores across configurations shows a clear

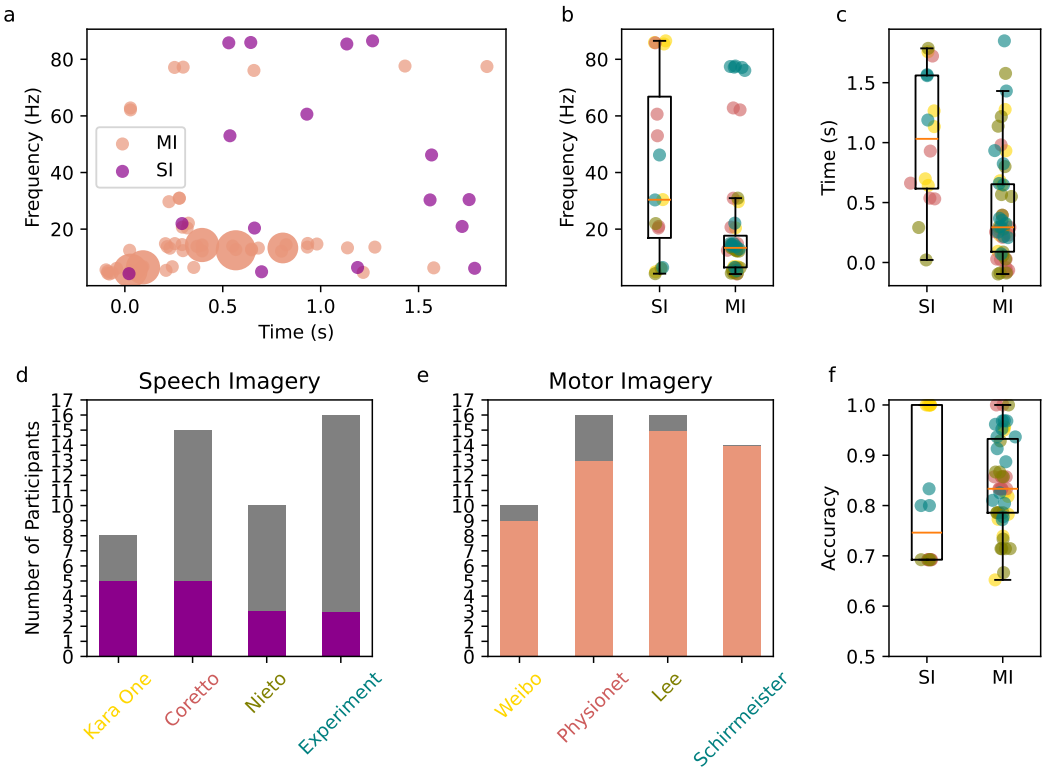

Figure 4.2: Distribution of the highest pair-wise decoding accuracies using the TS and LR pipeline across different time-frequency configurations for the Motor Imagery (MI) and Speech Imagery (SI) datasets. The evaluated class pairs were left versus right hand imagery for all MI datasets. For the SI 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

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 4.2.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 4.2.b shows 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 4.2.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 4.2.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, we still do not observe consistent time-frequency patterns among SI participants. The extended results for the 95% confidence analysis are included in the A.4.1.

Additionally, we noticed an influence of the number of trials used for classification. In Figure 4.3, 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

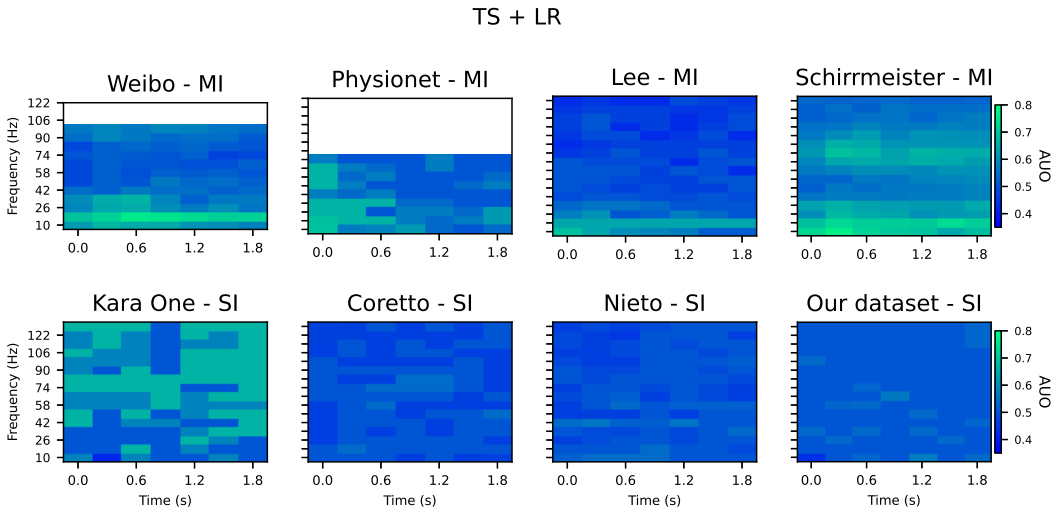

Figure 4.3: Effect of trial count on median classification accuracies across time and frequency windows using the TS and LR pipeline. Heatmaps display cross-validated accuracy as the number of training trials sequentially increases (30, 40, 50, 60, and >100 trials) for (a) two Motor Imagery (MI) datasets (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.

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.

4.5 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 4.1 [190, 191]. 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. [190], 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 [201]. 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 [18]. 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 generalised 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 synchronisation and desynchronisation 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 [202]. 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.

4.6 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 [203] 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 [29]. 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) [204], 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 generalise 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.

5

Consolidating the Speech Imagery Paradigm: Evidence that Rhythmic Protocols Drive Superior Decoding Accuracy

5.1 Introduction

Speech Imagery (SI) is an emerging and compelling paradigm within Brain–Computer Interface (BCI) research. It is considered intuitive to instruct participants to internally articulate a word to control a device, and it is naturally envisioned by researchers as an optimal candidate for speech neuro-synthesis applications [184]. SI research has gained increasing attention in recent years, and the release of open-access datasets has greatly facilitated the development and testing of decoding pipelines [35]. However, the large variability in decoding results and the limited number of real-time implementations raise uncertainty about the actual feasibility of decoding speech imagery from electroencephalography (EEG) [35, 170, 172].

One of the most trusted ways to verify the feasibility of a BCI paradigm is via successful replication and validation in real-time scenarios [205]. Our previous investigation into the replicability of SI 4 revealed that, in contrast to the well-established endogenous paradigm of Motor Imagery (MI), only 39% of participants

exhibited SI-related activity that could be classified with statistically significant accuracy in a binary decoding task—compared with 91% in MI datasets. These results raise the question whether SI-related neural signals can be reliably produced only by a limited portion of the population. A fundamental issue of EEG-based BCI research is the low performance of some users with BCIs, known as BCI inefficiency. It is estimated that, for sensorimotor responses, this may affect between 15–30% of users [17, 206].

As BCI inefficiency appears more prevalent in speech imagery, we argue that it is pressing to study this phenomenon to consolidate the paradigm. However, it is crucial to consider that acquisition protocols play a critical role in BCI performance and could lead to a reduction in inefficiency [176, 207]. In contrast to other mature paradigms—such as MI, where there is consensus regarding the kinesthetic strategy (feeling movement from a first-person perspective) [208] and established acquisition procedures like the Graz protocol, with defined temporal markers and feedback interfaces [209]; or in Steady-State Visual Evoked Potentials (SSVEPs), where stimulus presentation is fixed based on standard display frequencies and harmonics [210]—Speech Imagery presents a strongly heterogeneous landscape filled with diverse decoding strategies and often contradictory results [35].

The current study aims to quantify BCI inefficiency in SI across 12 heterogeneous datasets to evaluate how data acquisition protocols influence performance. Building on our previous findings regarding pipeline stability [211], we employ tangent space projection of covariance matrices to ensure a robust baseline for accuracy distribution. To standardise the definition of "efficiency" across varying sample sizes, we utilise a dual-threshold approach: a statistical significance marker derived from binomial distributions [183] and a 70% practical utility threshold established for clinical viability [212, 213]. Additionally, we aim to evaluate the influence of cross-validation (CV) procedures on performance estimation, particularly within rhythmic experimental designs where consecutive trials may exhibit temporal dependencies that can inflate the decoding accuracies [214].

5. Consolidating the Speech Imagery Paradigm: Evidence that Rhythmic Protocols Drive Superior Decoding Accuracy

Furthermore, we conduct a multi-stage meta-analysis to investigate the physiological and algorithmic drivers of success, aiming to isolate the factors underlying SI-BCI inefficiency. We analysed spectral modelling features previously employed in motor imagery research [17], while also examining features from the covariance matrices' structures. Finally, we evaluate model feature coefficients to identify specific importance mappings that characterise high-performing participants. Establishing whether these structural characteristics correlate with accuracy is essential for identifying why certain experimental variations yield superior results, representing a critical step toward a cohesive pathway for SI-BCI research.

5.2 Methods

We identified 9 open-access datasets and included our own acquired data. Two of the datasets have protocol variants, giving a total of 12 datasets. We evaluated our pipeline per participant across all 12 datasets and analysed the distribution of the obtained accuracies on each dataset.

5.2.1 Datasets

We summarise the employed datasets in table 5.1 and give a full description of the employed dataset in the appendix section B.1.

5.2.2 Preprocessing

Data from each source was read and formatted into epochs using MNE Python library v1.11.0 [197]. The EEG signal was re-referenced using common average referencing to mitigate common-mode noise. We analysed the PSD distribution and applied a notch filter at 50 or 60 Hz, depending on the dataset, where prominent line noise was identified. If the sampling rate exceeded 256 Hz, the signal was resampled to 256 Hz to ensure even resolution across datasets. ICA using the Picard algorithm was employed to identify eye-blink and lateral eye-movement artifacts. Based on the PSD and component distribution across epochs, one to two components were removed, and the signal was subsequently reconstructed. No

Table 5.1: Summary of the 12 evaluated datasets, encompassing diverse acquisition protocols, stimulus modalities and trial structures used to evaluate replicability in Speech Imagery

Dataset	Prompt	Cue	Protocol	Time(s)	Trials	N
BCI Competition [215]	auditory	blank	rhythmic	2	70	15
Coretto [95]	text	text	single	4	50	15
Kara One [23]	audio+text	ns	single	5	26	14
Liwicki [32]	text	on prompt	single	2	20	4
Malta [216]	visual	on prompt	single	4	40	5
Nguyen [22]	visual	auditory	rhythmic	1	100	6
Nieto [96]	visual	visual	single	2.5	50	10
Ours	text	visual	single	1.5	25	16
Ours rhythm	text	auditory	rhythmic	0.8	100	16
Rekrut [217]	no prompt	fix cross	game	2	80	15
Tec [218]	text	text+audio	rhythmic	1.4	30	15
Tec game [218]	no prompt	visual	game	1.4	30	15
Note: Time = time window for imagery; Trials = trials per class; N = number of participants; ns = not specified; fix cross = fixation cross.						

preprocessing was applied to the Coretto dataset, as the published data was already filtered (2–40 Hz) and cleaned of eye blinks via ICA. Similarly, the Nguyen dataset was used as published, having been bandpass filtered (8–80 Hz) and cleaned of ocular artifacts using electrooculography channels.

5.2.3 Filter-Bank TS+LR Decoding pipeline

The tangent space projection with logistic regression (TS+LR) is a Riemannian geometry–based method that has proven successful in motor imagery replication studies [18]. We extended this approach with a filter bank step. Filterbanks were defined with an 8 Hz bandwidth and a 6$\dot{\text{H}}$z moving window, covering 2–128$\dot{\text{H}}$z (20 bands) for most datasets, 2–40$\dot{\text{H}}$z for Coretto, and 2–80$\dot{\text{H}}$z for Nguyen.

For each frequency band, the filtered signal was transformed into regularised covariance matrices using Ledoit–Wolf estimation [180]. These matrices were then projected into the tangent space, converting them into Euclidean feature vectors that preserve their Riemannian structure. Each resulting vector has a dimensionality of $n(n+1)/2$, where n is the number of channels.

Finally, the tangent space vectors from all frequency bands were concatenated and used to train an LR classifier with an L1 penalty and a maximum of 600 iterations.

5.2.4 Evaluation

To evaluate decoding performance while accounting for the experimental structure of each task, we implemented two distinct cross-validation (CV) schemes based on the paradigm type:

- **Non-Rhythmic Paradigms:** Performance was evaluated using a standard 10-fold stratified random cross-validation procedure.
- **Rhythmic Paradigms:** To account for the dependent structure of consecutive trials, we employed a Group 10-fold CV approach. Folds were constructed by grouping trials based on the specific repetitions performed within a block, ensuring that all trials from a single repetition group were kept together to prevent data leakage during validation.

For both validation schemes, we reported the pooled accuracy across folds. Pooled accuracy provides a more statistically justified measure than the median or mean accuracy, especially in small datasets, as it accounts for fold-size weighting [219].

5.2.5 Meta-analysis and Statistical Evaluation

To compare results across heterogeneous datasets, we performed a meta-analysis of features derived from the imagery task window of all recorded trials. Participants were categorised into "Efficient" and "Inefficient" groups based on a performance threshold (≥ 0.7). To ensure comparability across diverse recording conditions, we applied the SAN.

Statistical significance between groups was assessed using Welch's t-test—to account for unequal variances and group sizes—complemented by Cohen's d to determine effect sizes. Furthermore, we investigated the sparsity of the learned models by analysing the distribution LR coefficients assigned to the TS vectors. Specifically, we computed the proportion of features with near-zero weights to evaluate whether successful paradigms (e.g., rhythmic) promote more concentrated and efficient feature representations compared to non-rhythmic tasks.

Power Spectral Density (PSD) Features

Spectral parameter modelling was performed using the *SpecParam* library [220] to decompose the power spectrum into aperiodic and periodic components. Features were averaged across electrodes (C3, C4, P3, P4). The primary features retained for meta-analysis were:

- **Alpha Prominence:** The amplitude of the periodic peak within the alpha band (8–13 Hz) relative to the aperiodic signal, serving as a measure of oscillatory strength.
- **Spectral Error (MAE):** The MAE of the model fit, representing the deviation of the actual power spectrum from the idealised modelled components.

Covariance Matrix Dynamics

To quantify the stability and complexity of the EEG signal structure, we calculated metrics based on the eigenvalues of the Riemannian mean covariance matrices for each trial:

- **Covariance Entropy Std:** The standard deviation across frequency bands of the Shannon entropy of the normalised eigenvalue distribution. This reflects the trial-to-trial variability in signal complexity.
- **Log Condition Number Std:** The standard deviation of the base-10 logarithm of the matrix condition number (the ratio of the largest to smallest eigenvalue). This measures the stability and of the spatial covariance.

Linear Regressor Weight Analysis

To evaluate the "efficiency" of the learned models, we analysed the structural distribution of the absolute values of the LR classifier coefficients.

- **Sparsity (Prop. Near Zero):** The proportion of TS features assigned weights near zero. This metric evaluates the model's ability to selectively identify key biomarkers while discarding non-informative noise, particularly in rhythmic versus non-rhythmic paradigms.

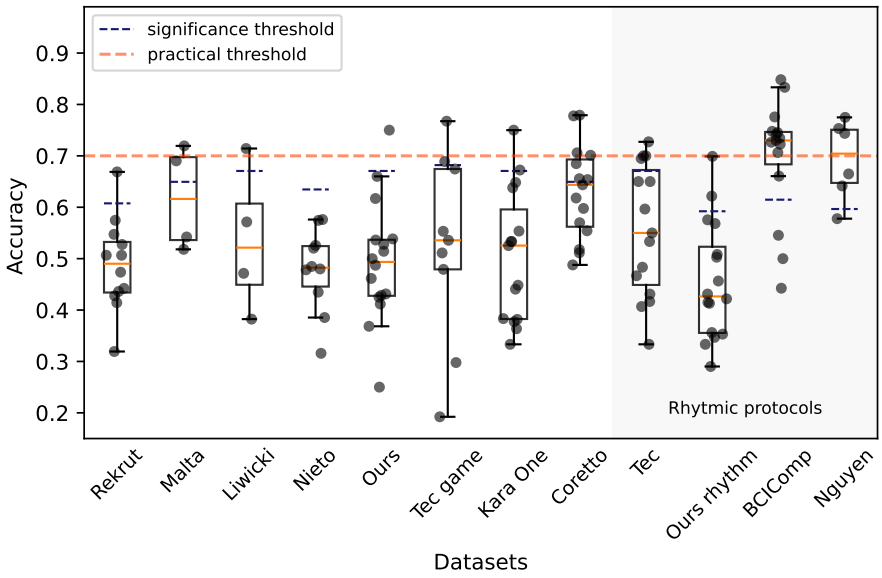

Figure 5.1: Pooled accuracy distributions for each dataset for the filter-bank TS+LR pipeline. The datasets recorded with a rhythmic protocol are grouped on the right side. The blue dashed lines indicate the statistically significant accuracy thresholds with 95% confidence, while the orange-red dashed line marks the practical threshold of 70%.

Success-Anchored Z-score Normalisation

To analyse feature distributions from the perspective of proficient users, we implemented a SAN. For each dataset, features were normalised by centring the data around the mean (μ) of the high-performing subset (accuracy ≥ 0.7). However, the scaling factor (σ) was derived from the standard deviation of the entire dataset. In instances where a dataset contained no participants meeting the performance threshold, standard dataset-wide Z-score normalisation was applied as a fallback.

Standard normalisation centres the feature space around the global population mean, which, in heterogeneous datasets with many low-performing users, can be dominated by noisy or non-informative patterns [221]. By using the efficient-performer distribution as the reference, we obtain a metric that quantifies how far each participant's features deviate from the feature space associated with good BCI performance.

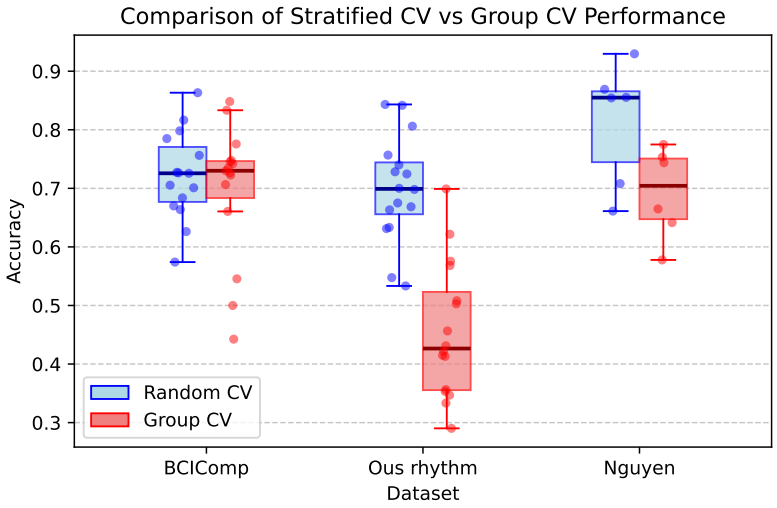

Figure 5.2: Comparison of decoding accuracy between cross-validation (CV) evaluation schemes across three rhythmic datasets: *BCIComp*, *Ours–rhythm*, and *Nguyen*. The blue boxplots represent Random (Stratified) CV performance, while the red boxplots represent Group CV performance. Horizontal lines within the boxes indicate the median values, dots show individual participant accuracies.

5.3 Results

The proportion of participants achieving statistically significant decoding accuracies varied substantially across the evaluated datasets, ranging from 0% to nearly 90%. Figure 5.1 illustrates these results for the optimal SI class pairs.

- **Significance and Protocol:** Only two datasets, *Nguyen* and *BCIComp*, saw more than 80% of participants reach statistical significance. While both utilised rhythmic protocols, this trend was not universal; the rhythmic *Tec* dataset reached only 40% significance, and our rhythmic data (*Ours–rhythm*) reached 12.5%. Notably, the raw performance of our internal data remained relatively stable between rhythmic and non-rhythmic sessions. However, the significance threshold (blue dashed line) for the rhythmic session is lower due to the number of available trials.

- **Practical Utility:** When applying the more stringent practical threshold (70% accuracy), *BCIComp* and *Nguyen* were the only datasets where over 50% of the cohort achieved practically viable results, as indicated by their median

lines sitting above the 0.7 mark. In contrast, the majority of other datasets—regardless of protocol—remained centred near or below the significance boundary.

The decoding results for all available SI class pairs are presented in appendix section B.2. The diversity of speech units and the inconsistent outcomes preclude the identification of specific SI classes that reliably lead to better performance across studies. It is noteworthy, however, that in one of the best-performing datasets (BCIComp), the percentage of statistically significant results remained above 70% even for its least discriminative class pair. In contrast, for low-performing datasets, accuracies remained low across all tested pairs. The Coretto dataset exhibited the greatest performance variability; for example, classifying a word against a consonant led 60% of participants to achieve a significant score, while classifying between words yielded no significant results.

5.3.1 Cross-Validation Strategy Comparison

Our results show the impact in decoding performance of the different CV procedures. The distribution of accuracies in figure 5.2 shows different degrees of impact. For BCIComp, the distribution of decoding accuracies remained highly stable between the two procedures. The median accuracy for Random CV was 0.73, matching the median accuracy observed for Group CV, though the Group CV distribution contained three lower outliers extending down to 0.44. In Ours–rhythm, a distinct divergence in performance was observed between the validation methods. While Random CV yielded a median decoding accuracy of 0.72, the median accuracy fell to 0.43 under the Group CV scheme, with the entire interquartile range shifting downwards. For Nguyen, similarly, a downward shift in performance was visible between the conditions. The median decoding accuracy decreased from approximately 0.85 in Random CV to approximately 0.70 in the Group CV implementation.

5.3.2 Significant Meta-Features for Efficiency

Statistical evaluation identified key biomarkers that differentiate efficient and inefficient performers. The most significant features were predominantly associated with alpha/mu band prominence and spectral model fit. Specifically, *C4 Alpha Prominence* and *C3 Alpha Prominence emerged* as features with the strongest group divergence ($d > 0.5$). Additionally, features related to the *SpecParam* model fit, including *C4* R^2 and *C4 Spectral Error MAE*, showed consistent differences between the two performance groups.

Analysis of the normalised feature distributions, shown in figure 5.3, further clarifies these distinctions: the efficient group exhibited a statistically significant positive shift in P4 Alpha Prominence and Log Condition Number Std, Covariance Entropy Std and Spectral Error MAE compared to the inefficient group.

Spectral power analysis in the parietal right region (2–60 Hz) provided further context for the previously identified Alpha Prominence biomarkers as shown in figure 5.4. Datasets demonstrating higher individual performance, specifically *BCIComp* and *Nguyen*, exhibited a larger alpha peak centred at 10 Hz in the relative power spectrum. Conversely, the aggregated "Others" group—which showed lower overall Alpha Prominence—exhibited a lower spectral profile in the alpha band. Furthermore, this "Others" group maintained a consistently higher relative power floor in the higher frequency bands (> 30 Hz) compared to the rhythmic datasets, which showed a more pronounced decay in spectral power at higher frequencies.

5.3.3 Classifier Coefficient Sparsity

Beyond the raw spectral features, we evaluated the structural composition of the LR coefficients to determine how the models weighted the TS vectors. We assessed the relationship between decoding accuracy and model sparsity—defined as the proportion of features assigned near-zero weights—to determine if efficient performers utilised more selective feature representations, as shown in figure 5.5. In Non-Rhythmic Paradigms, model accuracy showed no linear dependency on coefficient sparsity, yielding a non-significant correlation ($R = 0.05$, $p = 0.679$). In

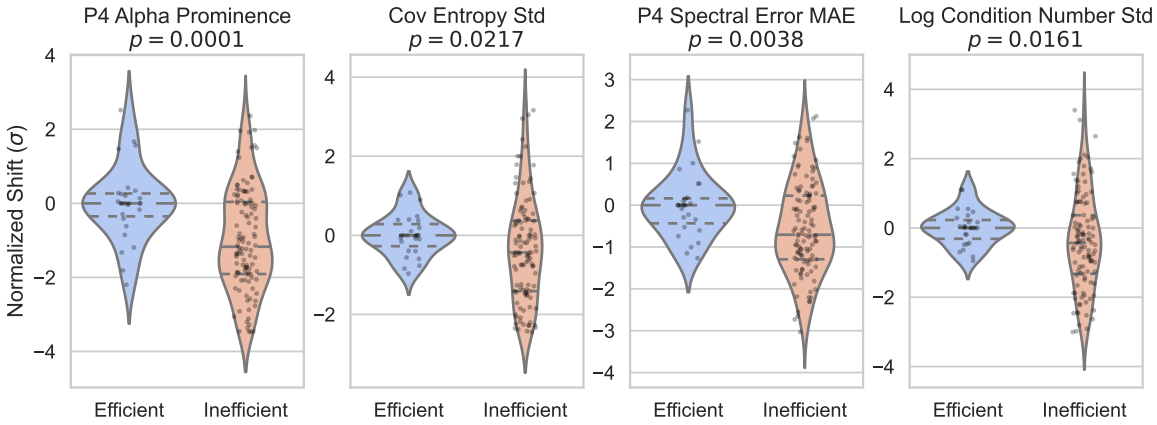

Figure 5.3: Distribution of the top four statistically significant meta-features differentiating the *Efficient* and *Inefficient* user groups across all datasets. From left to right, the panels display: P4 Alpha Prominence, Covariance Entropy Standard Deviation (Cov Entropy Std), P4 Spectral Error Mean Absolute Error (MAE), and Log Condition Number Standard Deviation (Std). The vertical axis measures the feature values as a normalised shift in units of standard deviation (σ) derived via Success-Anchored Normalisation. Individual participant values are overlaid as shaded dots within each violin plot, and horizontal dashed lines indicate the median and interquartile ranges for each distribution. Corresponding p-values from the statistical group comparisons are reported above each subfigure panel.

contrast, rhythmic paradigms exhibited a significant positive correlation between model accuracy and the proportion of near-zero weights ($R = 0.60$, $p < 0.005$).

5.4 Discussion

The substantial variation in the proportion of participants achieving statistical significance underscores the influence of data acquisition protocols on SI decoding. While two rhythmic datasets (*Nguyen* and *BCIComp*) yielded superior outcomes with over 80% of participants reaching significance, this trend was not universal; our internal rhythmic variation did not show a corresponding performance increase. Furthermore, gamified protocols (*Tec game* and *Rekrut*) demonstrated no clear advantage. Although analysis window lengths varied from 0.8 to 5 seconds across the datasets, this parameter did not correlate linearly with performance, though the most favorable results consistently aligned with shorter windows of up to 2 seconds.

The effect of evaluation procedure is critical; inappropriate CV schemes significantly inflate decoding scores. In block designs, consecutive trials share strong

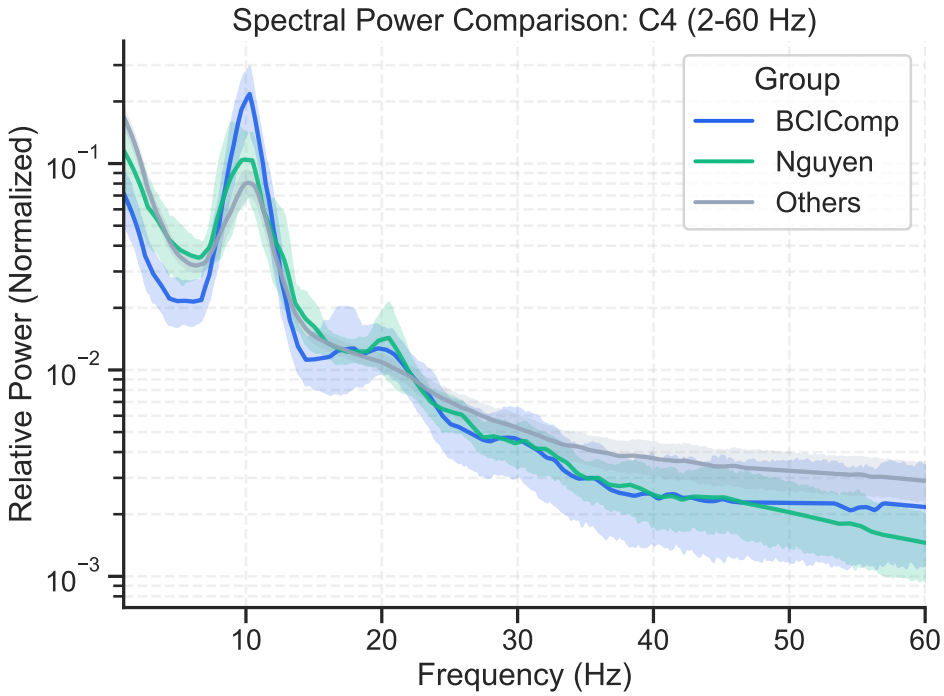

Figure 5.4: Normalised relative power spectrum during speech imagery trials recorded at the C4 electrode channel. Group-level averages and corresponding variance intervals are plotted for the top-performing rhythmic datasets—*BCIComp* (blue) and *Nguyen* (green)—alongside an aggregated ensemble of all remaining datasets (*Others*, gray). Individual subject PSDs were cleaned of 50 and 60 Hz powerline noise, linearly resampled to a common frequency grid, and normalised to yield relative power values. High-performing rhythmic groups display a distinct narrow-band peak centred at 10 Hz, whereas the *Others* group exhibits a flatter profile and a higher relative power floor across high frequencies (> 30 Hz).

temporal correlations, allowing a data-driven classifier to fit localised experimental drift rather than genuine neural features. Notably, *BCIComp* displayed minimal sensitivity to the validation strategy, which may be attributed to longer time intervals between imagery trials. However, because three participants in this dataset experienced a sharp decrease in accuracy under Group CV, there remains a possibility that the original trial sequencing was altered in the public repository, making perfect epoch grouping challenging.

The difficulty in comparing these highly heterogeneous datasets highlights the critical need for paradigm consolidation in SI-BCI research. Although a wide variety of speech units (phonemes, words, phrases) were evaluated, classification performance remained highly variable. Interestingly, within *BCIComp*, the per-

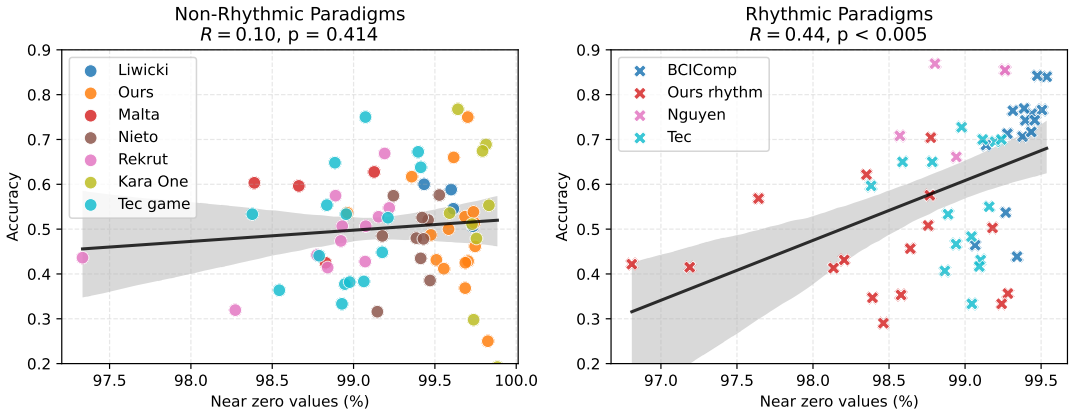

Figure 5.5: Correlation between decoding accuracy and classifier coefficient sparsity across individual participants, separated by paradigm category. The left panel displays results for *Non-Rhythmic Paradigms* using circular markers, showing a non-significant correlation ($R = 0.10$, $p = 0.414$) across seven datasets (Liwicki, Ours, Malta, Nieto, Rekrut, Kara One, Tec game). The right panel displays results for *Rhythmic Paradigms* using cross markers, demonstrating a statistically significant positive correlation ($R = 0.44$, $p < 0.005$) across four datasets (BCIComp, Ours rhythm, Nguyen, Tec). In both panels, solid black lines represent linear regression fits, the surrounding shaded areas denote the 95% confidence intervals, and the horizontal and vertical axes show decoding accuracy and the percentage of near-zero coefficient values, respectively.

centage of proficient participants only dropped from 100% to 70% when shifting from the best-performing to the worst-performing class pairs. This implies that the underlying signal acquisition protocol plays a more pivotal role in ensuring decoding viability than the specific linguistic token selected.

The identified meta-features align closely with sensorimotor BCI efficiency and proficiency frameworks [17, 222]. The efficient group exhibited a prominent positive shift in P4 Alpha Prominence, confirming that a pronounced oscillatory component provides more discriminable spatial covariance topologies. This finding presents a compelling neurophysiological nuance; classically, alpha suppression (ERD) is viewed as the primary indicator of active cognitive processing and cortical engagement. However, the bilateral positive shift across parietal channels observed in efficient performers can be reconciled through the lens of the inhibition-gating hypothesis [223, 224]. Under this framework, elevated parietal alpha power represents the active functional inhibition of task-irrelevant visuospatial networks, suggesting that proficient users more effectively shield internal speech generation from sensory

distraction. Methodologically, this mechanism acts as a critical driver of structural SNR. Rather than reflecting cortical idling, a distinct and well-organised alpha peak provides a stable baseline that anchors the spatial covariance configurations across trials. This is structurally supported by the *SpecParam* metrics; efficient users clustered tightly around the baseline for *P4 Spectral Error MAE* and *Covariance Entropy Std*. This indicates a stable, predictable, and non-random power spectral distribution that adheres well to parameterised fits rather than collapsing into an unstructured $1/f$ noise floor. Combined with a positive shift in the *Log Condition Number Std*, these features reflect well-conditioned, structured covariance matrices. Conversely, inefficient performers exhibited high signal disorder and feature variance, which disrupts the mapping stability within the tangent space.

Post-hoc analysis of the LR coefficients demonstrated that the LR+TS pipeline successfully drives feature pruning, dropping near-zero weights by up to 95% (reducing a vector of nearly 20,000 components down to roughly 100 key features). Crucially, a strong positive correlation emerged between coefficient sparsity and decoding accuracy within rhythmic paradigms. This suggests that isolating a compact, highly structured set of neural biomarkers while rejecting background noise is essential for successful classification in rhythmic contexts—a behavior not observed in non-rhythmic paradigms, where accuracy fluctuated independently of model sparsity.

Finally, to estimate a representative baseline for BCI-SI inefficiency, we isolate only the top-performing acquisition protocols, arguing that suboptimal recording designs artificially skew the failure rate. Defining efficiency as the ability to surpass the 70% clinical viability threshold, the true BCI-SI inefficiency rate is estimated to lie between 30% and 50%.

From a practical standpoint, deploying this rhythmic speech imagery protocol in a real-time application yields specific operational implications:

- **Calibration and Retraining:** To mitigate day-to-day signal non-stationarities and maintain classifier stability, a brief daily calibration phase of a few minutes remains necessary to adjust the baseline spatial covariance matrices.

5. Consolidating the Speech Imagery Paradigm: Evidence that Rhythmic Protocols Drive Superior Decoding Accuracy

- **Real-Time Implementation and Pacing:** An online multi-class system (e.g., a 4-class architecture for mouse cursor navigation) would operate under a synchronous framework. The system would provide an auditory or visual metronome stimulus to pace the user's repetitive speech imagery cycles.

- **Impact of Rhythmic SI on Online Performance:** Because individual trials are inherently noisy, the highly structured nature of rhythmic SI allows the classifier to accumulate decoding evidence across three or more consecutive imagery repetitions. A command is only translated into action once a system-confidence threshold is satisfied, which would minimise false-positive activations compared to single-trial decoding.

- **Asynchronous Gating:** Given the current limitations in multi-class classification and the reliance on active pacing, early practical iterations would require a dedicated gating mechanism (such as a simple button or gaze activation) to allow the user to manually initiate or terminate the decoding phase.

5.5 Conclusion

This study paints a concerning picture of the Speech Imagery paradigm's feasibility for EEG-based BCIs. The landscape is challenging, underscoring the paradigm's immaturity. Our results show that SI critically struggles with replicability, with data acquisition protocols having a decisive impact on performance. Beyond acquisition design, we identified the choice of evaluation procedure—specifically the cross-validation scheme—as a critical new pitfall that can lead to severe performance overestimation. In continuous or block-based rhythmic paradigms, consecutive trials share strong temporal correlations, allowing a data-driven classifier to fit localised experimental drift rather than genuine neural features if Group CV is not rigorously applied. Even when optimal protocols are used, the viability remains low: while 2 out of 12 datasets initially showed widespread statistical significance, proficiency drops considerably when a practical accuracy threshold ($> 70\%$) is enforced. Ultimately, if SI-related activity cannot be reliably produced by a large

portion of the population—as our 30–60% inefficiency estimate suggests—then this paradigm may remain unfeasible for creating universally accessible BCI systems.

Nevertheless, our findings do illuminate a necessary path forward. We found that an important difference between efficient and non-efficient participants is linked to signal stability. Our meta-analysis shows that high-performers, especially those using rhythmic protocols, produce a more structured brain signal with lower entropy in the covariance matrix space. This is an insightful finding because it suggests that the rhythmic approach helps the brain create a clearer pattern that leads to better classification. We also found that in good performers, the classifier consistently pruned the features reducing up to a 95% of features. For the low performers, the classifier may get lost in the noise and select weight equivalently more features allaround.

These results should be taken as a clear signal to shift research strategies. We found that established pipelines can extract SI features, suggesting the field should now prioritise testing signal acquisition protocols and effective SI units, rather than the decoding procedures that are currently abundant yet often lack reproducibility. Our study specifically indicates that rhythmic paradigms require deeper investigation because they seem to "force" the brain into a more consistent state that is easier to decode. Furthermore, at this early stage, research should focus on maximising accuracy by adhering to pairwise classification to validate the core concepts before attempting more complex tasks. Finally, while we utilise the 70% accuracy threshold based on established literature, future work should focus on identifying the optimal practical threshold dictated by the demands of real-time applications and further refining the characterisation of BCI-SI inefficiency.

From an optimistic perspective, the paradigm could be refined for the participant subset that shows promise, aiming to enhance their performance and understand the neuro-cognitive profiles that enable it. By focusing on how these "efficient" users organise their brain activity, we might find ways to train others to do the same.

6
Discussion

This thesis has critically examined the current state of the art of Speech Imagery (SI) as a paradigm for Brain–Computer Interfaces (BCIs). Throughout this research, we have addressed our initial research questions and identified significant concerns regarding the field's maturity and methodological rigour. While the initial literature review portrayed a highly feasible paradigm with promising decoding capabilities, our subsequent investigations revealed an over optimistic landscape lacking the formal direction necessary for genuine paradigm consolidation.

The scarcity of real-time decoding approaches in the literature—contrasted with the abundance of offline studies—suggests a field currently characterised by low reproducibility. Our investigations in Chapters 4 and 5 confirmed this substantial gap; we highlighted that significant discrepancies of up to 40% in reproduction attempts often stem from missing methodological details, flawed evaluation procedures, and inflated performance scores. Furthermore, we identified that SI-related signals are inherently more difficult to produce consistently across trials than established paradigms like Motor Imagery (MI). While we identified stable, high-performing time-frequency configurations for MI, the configurations for SI remained highly variable and inconsistent. However, our evidence also suggests a potential solution: rhythmic recording protocols may help induce more consistent neural responses, offering a clearer path for future development.

The following sections summarise our key findings and provide a roadmap for guiding SI research toward scientific maturity.

6.1 The reproducibility crisis

Our analysis of a vast array of decoding approaches revealed that only 6 out of 104 decoding attempts were conducted in a real-time scenario. Given that 57 of these studies involved original data collection, there is a gap between offline decoding attempts and real-time implementations. This disparity highlights a lack of reproducibility within the field; the optimistic results often reported in offline decoding have not been consistently demonstrated in real-time environments. Consequently, the practical feasibility of these paradigms remains to be verified, necessitating further formal research to consolidate and validate the progress made to date.

These literature-based concerns were empirically evidenced in Chapter 4, where we attempted to reproduce findings from eight decoding publications across two popular datasets. Our results revealed substantial discrepancies: reproduction performance was consistently lower than in the original reports, with four of the eight attempts showing significant performance drops of up to 40%. Furthermore, we identified that three of these reports failed to provide a complete description of either the detailed decoding methodology or the evaluation procedures. Incomplete documentation often obscures flawed procedures, further increasing the barriers to reproducibility. These findings align with the broader reproducibility crisis observed in other areas of BCI and served as the foundation for the evaluative approach adopted in Chapter 4.

6.2 The replicability crisis

Our literature review revealed an extensive range of approaches toward Speech Imagery (SI) paradigms. Experimental designs have explored diverse methods for instructing and cueing imagery—ranging from written captions to object

recognition—as well as various speech units, including phonemes, words, and full phrases. Furthermore, studies utilised vastly different imagery window lengths and instructions (e.g., repeated vs. single imagery). Crucially, there is no prevailing consensus on which techniques produce consistent results. These findings suggest that the high degree of variation in SI research, while attempting to answer the same fundamental questions, leads to fluctuating results that naturally cast doubt on the replicability of the paradigm.

The evidence for this crisis was further exposed in Chapter 4, where SI was compared against the well-established Motor Imagery paradigm. While decoding configurations produced equivalent and stable results for MI across four datasets—reflecting the known, reliable EEG–based MI response—-the results for SI were highly variable. Chapter 4 also demonstrated that SI-related prominent frequency responses are participant-specific. This suggests that only a small subset of the population may produce decodable SI signals, posing a significant challenge for the generalizability of the technology.

Chapter 5 presents an even more concerning picture of replication failure in SI. Through the exploration of 12 heterogeneous SI datasets, we found that only three produced average decoding accuracies above chance levels. By proposing a practical threshold to assess BCI efficiency, we discovered that seven of these datasets had fewer than two participants achieving usable scores. Remarkably, in none of the 12 datasets did all participants achieve a classification accuracy above 0.7. Ultimately, the findings in Chapter 5 highlight an urgent need for the consolidation of the SI paradigm, given the current landscape of methodological heterogeneity and the challenge of achieving optimal, replicable performance.

6.2.1 Determinants of Success: Rhythmic Protocols and Covariance Stability

Chapter 5 underscored the profound heterogeneity of Speech Imagery (SI) research, yet it also provided a path toward paradigm consolidation. Our analysis revealed that datasets recorded during the rhythmic performance of SI led to significantly higher

decoding scores. We identified a clear distinction in the features extracted from the tangent space projection decoding pipeline when comparing "good" and "bad" performers across datasets. High-performing participants consistently exhibited lower average entropy in the covariance matrices of their signals across frequency bands. Mathematically, this indicates a lower noise floor and a more robust signal structure within those specific experimental setups.

We attribute this to the participant's ability to produce consistent, repeatable neural activity. Specifically, a rhythmic task instruction appears to entrain the brain, forcing a consistent neural response throughout the recording procedure. This structure-led performance was further validated by a meta-analysis of the weights assigned by the logistic regression model; we found that the correlation between model coefficients across frequencies was strongly predictive of the final classification score.

The findings in Chapter 5 provide a foundational framework for the consolidation of SI. They suggest that rhythmic protocols lead to superior results regardless of the specific cues used or the speech units prompted. We further validated this through our own empirical data collection, where the same participants demonstrated a significant performance improvement when performing rhythmic imagery compared to single-trial (discrete) imagery. This evidence suggests that the SI-BCI inefficiency often reported in SI may be a byproduct of suboptimal task design rather than an inherent limitation of the paradigm itself. However, it must be noted that even with optimised rhythmic protocols, the percentage of SI-BCI inefficiency appears notably higher than that observed in the Motor Imagery (MI) paradigm. This indicates that while design improvements can help to bridge the gap, Speech Imagery remains a more complex and challenging modality for achieving universal user proficiency.

6.2.2 Future work

This thesis has revealed that while Speech Imagery (SI) remains a promising BCI paradigm, it requires urgent methodological consolidation to achieve genuine progress. Our findings suggest that much of the existing literature has presented an

overly optimistic view of SI, often failing to withstand the rigors of reproduction or real-time application. However, our large-scale replicability analysis offers a clear path forward, specifically through the implementation of rhythmic protocols. To facilitate the maturation of SI as a viable BCI modality, we propose the following recommendations based on the findings of this research:

- **Prioritising Pairwise Classification** At this stage of paradigm development, research into offline SI decoding should prioritise high-fidelity pairwise classification over complex multi-class attempts. It is more valuable for the field to achieve robust, optimal decoding accuracies between two states than to report results that only marginally exceed chance levels in multi-class setups.

- **Standardisation of Rhythmic Protocols** Following our evidence that rhythmic instructions may "entrain" neural activity and force a more consistent response, future studies should adopt and validate rhythmic protocols.

- **Accounting for Psychological Factors** Given our findings on participant-specific responses, future work must consider the user's psychological profile. Utilising instruments such as the VISQ-R questionnaire [225] to assess inner speech traits may help identify neural correlates of BCI inefficiency and allow for the development of participant-tailored decoding models.

- **Adherence to Reporting Frameworks** To mitigate the reproducibility crisis identified in Chapter 4, SI decoding attempts must follow strict reporting frameworks, such as CRISP-DM [203]. Crucially, researchers should make their preprocessing pipelines and decoding code available alongside publications to ensure that optimistic offline results can be verified by the broader community.

- **Transition to Real-time Validation:** At the current stage of SI development, research should prioritise identifying the factors that optimise offline performance—such as rhythmic protocols and speech unit selection—while refining the estimation of SI-BCI inefficiency. Once practical and consistent decoding accuracies are established offline, the next logical step is to validate

these findings through real-time implementation. This progression is essential to bridge the gap between experimental development and practical utility.

- **Reporting Participant Proficiency Distributions:** To accurately estimate SI-BCI inefficiency, researchers should go beyond reporting average accuracies and explicitly state the percentage of participants who achieve a "functional utility" threshold. This threshold represents the accuracy required for reliable, impactful control rather than mere statistical significance. While a 70% benchmark was adopted in this investigation based on existing literature, future real-time studies are essential to empirically define this level by determining which specific performance scores translate into a meaningful result for the user.

Appendices

A

Chapter 4: Supplementary Materials

A.1 Evaluated Datasets

The following speech imagery and motor imagery datasets were used in Chapter 4.

A.1.1 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 [23]. 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.

A.1.2 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 [95].

A.1.3 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 s and final relaxation state with a time-variant interval. Further details can be found

in the author's publication [96]. We considered the 64 channels equivalent to a 64-electrode setup in order to reduce the dimensionality.

A.1.4 SI Dataset: Our dataset

Sixteen right-handed able-bodied participants (8 female, 8 male) 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 [226] 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 s until the "relax" prompt was shown.

A.1.5 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 s. Then the prompting or the character ("Left Hand", "Left Hand and Right Foot", et al) was presented on the screen for 4 s, 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 s before next trial. Further details of the experiment can be found at [196].

A.1.6 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 [194]. 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.

A.1.7 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 s of a black fixation cross as a preparation stage. Then the participant performed the imagery task of grasping with the appropriate hand for 4 s 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 [195]. 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.

A.1.8 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 [152]. We considered the 64 channels equivalent to a 64-electrode setup in order to reduce the dimensionality.

A.2 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.

CSP + LDA

CSP is a well-established spatial filtering technique commonly used in BCI research. It derives spatial filters that maximise variance differences between two classes, enhancing discriminative information in the data [227]. 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.

TS + LR

Riemannian geometry-based methods have demonstrated strong performance and robustness for EEG decoding tasks [18]. 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 [228]. 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.

CNN

CNNs have been widely applied to EEG decoding problems in recent years [35]. We implemented the CNN architecture proposed by Wimpff et al. [229], 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.ochs.chs..chs.

A.3 Evaluated Decoding Approaches

We describe the decoding pipelines applied to Kara One and Coretto SI datasets that we attempted to reproduce, we identify missing or ambiguous information in their method explanations and state the steps we took to attempt reproduction.

A.3.1 KO1 [23] approach

The Kara One authors' approach to decoding SI consisted of grouping the classes based on the phonemic characteristics in vowel vs consonant (C/V) and the presence of a high-back vowel (/uw/). Therefore, they attempted a two-class classification approach.

Preprocessing

The authors applied a band-pass filter between 1–50 Hz. No specification about type or filter order was given; they also applied a small Laplacian filter to the data using the adjacent channels.

We applied a FIR hamming window band-pass filter on the described frequencies.

Feature extraction and selection

For each EEG segment, the signal was windowed to 10% of the segment, with a 50% overlap, resulting in 17 segments of 500 ms for each channel. From this segment, various features were obtained: mean, median, standard deviation, variance, maximum, minimum, maximum ± minimum, sum, spectral entropy, and energy. Also, the mean maximum and minimum, sum and difference of the maximum and minimum for the absolute value of the segments. Furthermore, the authors mentioned they computed the first and second derivatives of the above features, resulting in a final 1197×1 feature vector. Due to the high dimensionality of the features, they ranked the features by the Pearson correlation with each class and selected the top 5 features.

We found that the description of the procedure for computing the derivative of the initial feature set is incomplete and may result in incorrect implementation. We reached out to the authors, but the information couldn't be clarified due to the time passed since then. Furthermore, the authors may have employed additional features not explicitly described, as our reconstruction yields a total of 32 features per segment, resulting in 816 features overall, which conflicts with the 1,197 features reported in their publication. In our reproduction attempt, we excluded the derivative-based features due to the ambiguity in their computation, resulting in a feature vector of size 221×1.

Classification

A participant-independent leave-one-out cross-validation procedure was used to evaluate two classifiers, a Deep Belief Network (DBN) and a Support Vector Machine

(SVM) with quadratic and radial basis function kernels. The DBN was set with one hidden layer, whose bottleneck size is 25% of the input size. Training was done over 10 epochs. For both SVMs, we allow 90% of the data to violate the Karush-Kuhn-Tucker conditions [230].

A.3.2 KO2 decoding [132] approach

The authors compared between different types of features, linear features, non-linear features, and Mel Frequency Cepstral Coefficients (MFCC). It was found that MFCC performed best when classifying the 11 SI classes in a participant-dependent approach. Therefore, we attempted to reproduce the MFCC pipeline. We raise a concern with this approach as they ignored the faulty data we mentioned in section A.1.1 and used the data of 11 participants without participant identifiers.

Preprocessing

EEG signals were filtered between 1–50 Hz, then a small Laplacian filter was applied to each channel. The signal of each channel was segmented as in the original approach, using 500 ms windows with 250 ms overlap. The authors state that, Independent Component Analysis (ICA) was performed to remove noise artifacts. However, there is no detail on the removal criteria for the artifacts or whether the ICA was applied before or after the segmentation.

We applied ICA to the channels before segmenting and removed 1–2 ocular components based on their location.

Feature Extraction and Selection

Thirteen cepstral coefficients were obtained from each window from a filterbank of nine filters. Each of the 13 MFCC was calculated on all 62 channels and 17 windows, resulting in a total of 13702 features. To reduce the dimensionality, Principal Component Analysis (PCA) was applied and set to keep components that explain 95% of the total variance.

Classification

Using 5-fold cross-validation procedure was used for each participant to evaluate the performance of an SVM and decision tree classifiers. The authors did not specify the regularisation parameter C of the SVM. We attempt the reproduction of these results with an SVM and set C to 1.

A.3.3 KO3 decoding [166] approach

The authors approached SI decoding with deep learning, they investigated different architectures of Convolutional Neural Networks (CNNs) and evaluated their performance on classifying 11 SI-classes by extracting and comparing temporal and frequency features in a participant-independent approach. The study concludes that frequency features perform better than temporal features. The authors took into account the faulty data and kept only 8 participants. However, the participants' identifications were not mentioned.

Preprocessing

The signal was filtered using a notch filter to remove 60 Hz artifacts and harmonics smaller than the Nyquist frequency.

Feature Extraction

Each channel's signal was divided into different sections of 0.25, 0.5, and 1 s without overlapping in order to identify an optimal window. The authors treated each of these windows as a new sample. Each signal was transformed using FFT, and then a covariance matrix was computed for each window, resulting in a (channel x channel) shaped matrix.

Classification

The authors tried different CNN architectures and concluded that an architecture with 2 convolutional layers and 1 dense layer reached the best performance. However, the specification about the kernel sizes of the convolutional layers is ambiguous as they conclude that a kernel size equal to the input shape performed better;

such a kernel changes the expected behaviour of the CNN. To evaluate their model, the authors state that they randomly split the windowed data into 50% for training and 50% for testing. We assume that the random partition did not lead to windows of the same trial being leaked into both splits. The authors did not use cross-validation procedures.

We used (7x7) and (3x3) kernels for the convolution layers.

A.3.4 KO4 decoding [169] approach

The authors explored Common Spatial Patterns (CSP) in a 1-vs-all approach for multiclass classification and used an Ensemble Stacking Learning classifier. The authors did not mention if they addressed the issue of the faulty data, and used data from 13 participants. Unfortunately, the feature extraction description is substantially incomplete, as we detail below, and we failed to reproduce their approach.

Preprocessing

No preprocessing steps are mentioned.

Feature Extraction

The authors re-labelled the trials in order to perform 1-vs-all CSP. There is missing information about how many CSP components they used and what features were extracted from such the components.

This missing information is critical for the reproduction of the findings, as these choices in the feature extraction steps have a strong influence in results. Therefore, we were unable to reproduce this approach.

A.3.5 CT1 decoding [95] approach

The Coretto dataset authors' approach attempts classification of SI by grouping the trials into vowels and words.

Preprocessing

The EEG signal was filtered with cut-off frequencies of 2 and 40 Hz. With low and high-pass FIR filters of orders 372 and 1204 respectively. The signal was downsampled to 128 Hz. ICA was used to remove the blinking artifacts, but the authos did not specify which version of the ICA algorithm they used.

In our replication attempt, we used the Picard ICA algorithm.

Feature Extraction

The authors used the Discrete Wavelet Transform (DWT) to extract frequency features; they decomposed the signal using the Daubechies wavelet family. From the decomposition levels generated, they selected detail coefficients from levels 1–5 and the approximation coefficients from level 5. From each of these levels, they computed the Relative Wavelet Energy (RWE), resulting in 6 features per channel.

Classification

Two classifiers were explored, a random forest (RF) with 6 features and 50-200 trees were tested and an SVM with a linear kernel. The pipelines were evaluated in a participant-dependent manner using a 10-fold cross-validation procedure.

A.3.6 CT2 decoding [135] approach

This classification attempt on SI combined DWT and CSP to decode 2 SI classes.

Preprocessing

Signals were downsampled to 256 Hz.

Feature Extraction

The signal was composed into 4 levels using the Daubechies wavelet family. Among the decomposition levels, detail coefficients 1–4 and approximation coefficients from level 4 were extracted. From each of these levels, CSP was applied so the full signal was filtered down to 4 components per level, and the logarithm of the

normalised variance of each component was computed as the final feature. Resulting in a feature vector of shape (20 x 1).

Classification

An SVM with a Radial Basis Function (RBF) Kernel was used as a classifier. The authors did not specify the regularisation C or Kernel coefficient *gamma* parameters used for the SVM. The model was evaluated in a participant-dependent manner following a 5-fold cross-validation procedure.

For our reproduction attempt, we set SVM $C = 1$ and $gamma = 1/n$, where n is the number of features.

A.3.7 CT3 decoding [166] approach

The authors attempted speech decoding using Correto's dataset and a CNN. They investigated capabilities for transfer learning. They proposed a CNN architecture where the first layers could be used as a pretrained model for other users. We attempted to reproduce their initial within-participant findings without the transfer learning approach.

Preprocessing

The signal was downsampled to 128 Hz and artifacts were removed using ICA. The authors did not specify which criteria were used to select noisy ICA components or how many were removed. The signal was finally standardised by centring to the median and scaling accordingly before feeding the data to the CNN.

In our reproduction attempt, we removed 1–2 ICA components related to eye artifacts based on the scalp location.

Feature Extraction

The paper does not specify any feature to be computed from the raw signal, suggesting that they directly fed the raw signal of shape (6 channels x 496 times) as input to the model.

Classification

A CNN architecture consisting of 7 convolutional layers connected to a final linear classifier unit. The authors detailed the dropouts and pooling layer specifications after each convolutional layer. Further details on the architecture can be found in [166]. The model was evaluated using a 5-fold cross-validation procedure.

A.3.8 CT3 decoding [122] approach

This study emphasised the need of reproducibility as a SI decoding approach criteria; they also approached this by using CNN and investigated the impact of downsampling the signal in temporal space. The code for their approach was made available online. Unfortunately, we found inconsistencies between the shared code and the report specifications. Therefore, we followed the described methods in the report as closely as possible.

Preprocessing

The data were downsampled to 128 Hz. ICA was applied to remove noisy components. However, the study does not mention how many components were removed. In our reproduction attempt we remove 1 IC per participant based on its projected scalp location.

Feature Extraction

The paper does not specify any feature to be computed from the raw signal, so we used as input the raw signal of shape (6 channels x 496 times).

Classification

A CNN architecture was proposed to classify the signals, it consisted of 7 convolutional layers and a final Softmax layer as output. The normalisation, pooling and dropout specification is given in detail for all the layers. A complete description of the CNN architecture can be found in [122]. The model was evaluated with an 80/10/10 training/validationn/testing split. No cross-validation procedures were applied. The author's code, however, specifies an architecture with 20 filters

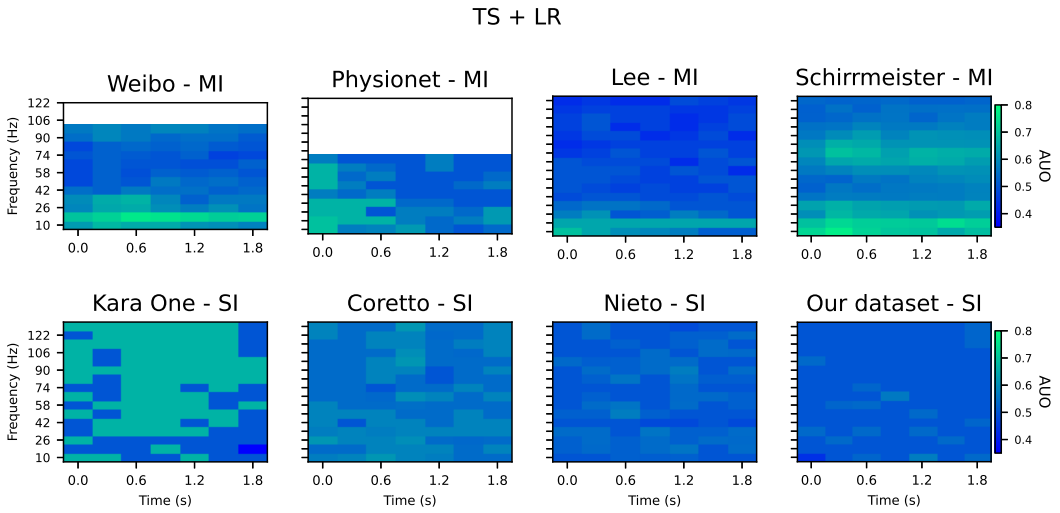

Figure A.1: Heatmaps of median classification accuracies across participants for various time and frequency windows using the TS and LR pipeline. The top row displays results for four Motor Imagery (MI) datasets, while the bottom row displays results for four Speech Imagery (SI) datasets. For this analysis, 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 ($p < 0.05$).

in the initial convolutional layers in contrast with the 40 reported. The number of filters increases significantly the parameters to be tuned, which could play an important role in performance. In our decoding attempt, we used the parameters from the publication rather than the code.

A.4 Extra Results

A.4.1 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 A.2 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

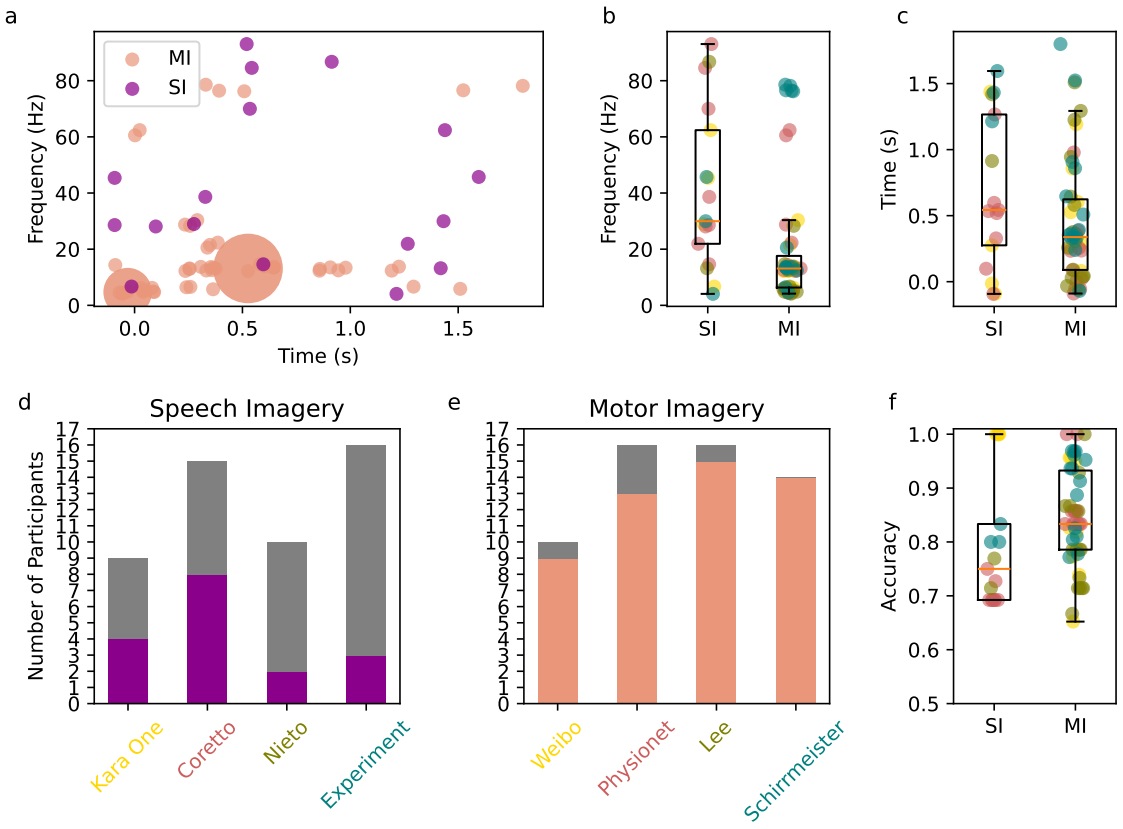

Figure A.2: Distribution of the highest pair-wise decoding accuracies using the Tangent Space and Logistic Regressor (TS+LR) pipeline across different time-frequency configurations for the Motor Imagery (MI) and Speech Imagery (SI) datasets. The evaluated class pairs were left versus right hand imagery for all MI datasets. For the SI datasets, the 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.

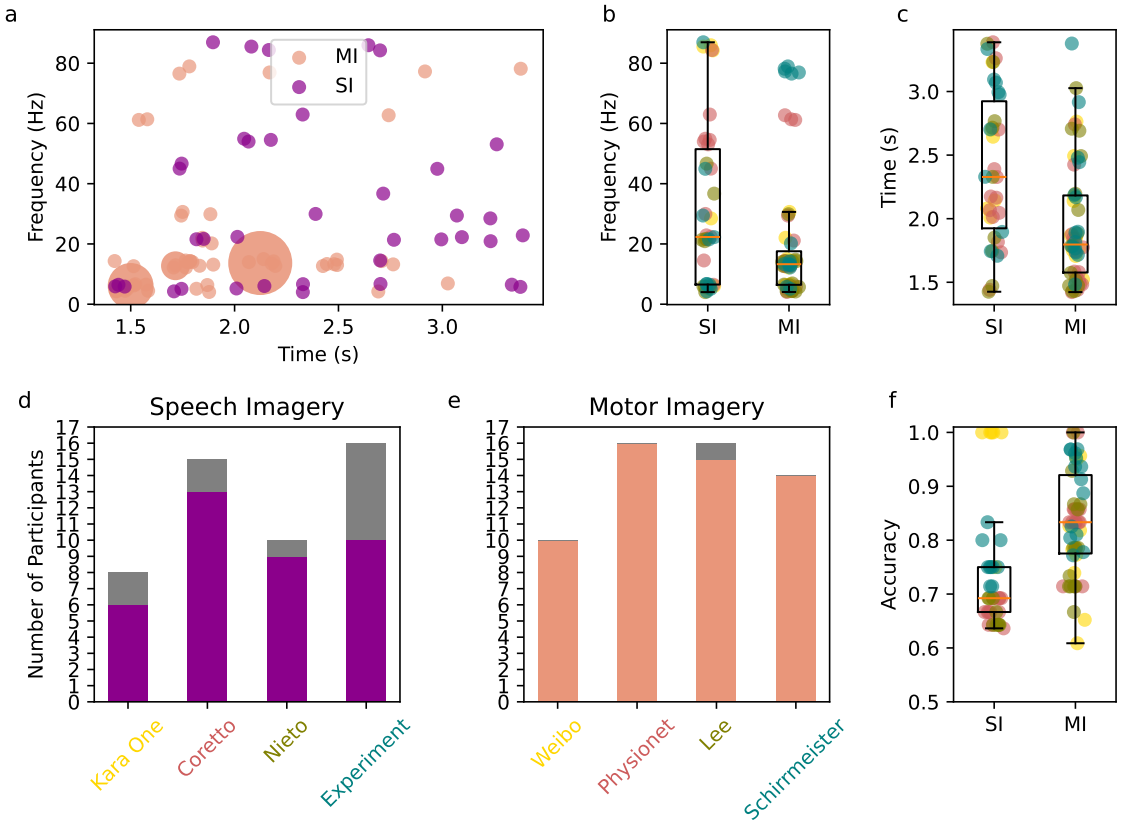

Figure A.3: Distribution of the highest pair-wise decoding accuracies using the Tangent Space and Logistic Regressor (TS+LR) pipeline across different time-frequency configurations for the Motor Imagery (MI) and Speech Imagery (SI) datasets. The evaluated class pairs were left versus right hand imagery 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.

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 A.1, 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 [183] 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 A.3 shows the time-frequency analysis results for decoding accuracies that reached this 95% significance threshold.

A.4.2 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 A.4 and CNN results in Figure A.5.

A.4.3 Results on parameter selection

Table A.1 shows the results of grid-search on SVM parameters C and *gamma* for reproduction attempts of K02 and CT2 reports. Mean accuracies across participants

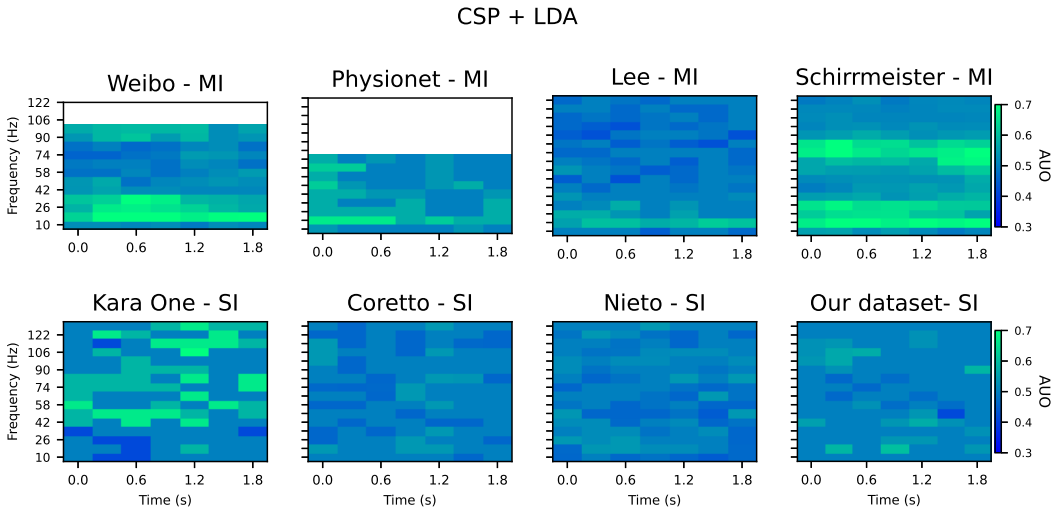

Figure A.4: Heatmaps of median classification accuracies across participants for various time and frequency windows using the Tangent Space and Common Spatial Patterns and Linear Discriminant Analysis (CSP+LDA) pipeline. The top row displays results for four Motor Imagery (MI) datasets, while the bottom 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 ($p < 0.05$)

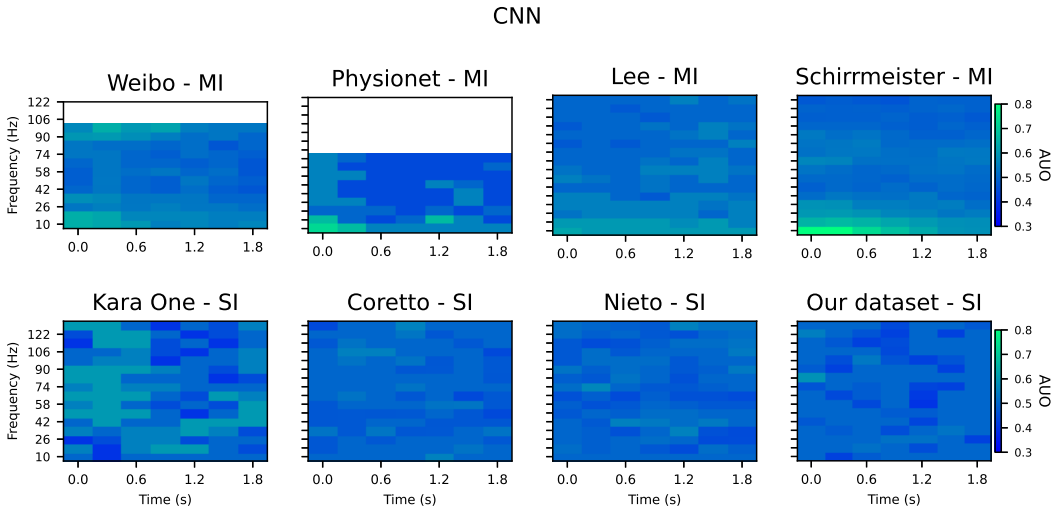

Figure A.5: Heatmaps of median classification accuracies across participants for various time and frequency windows using the Tangent Space and Common Spatial Patterns and Convolutional Neural Networks (CNN) pipeline. The top row displays results for four Motor Imagery (MI) datasets, while the bottom 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 ($p < 0.05$)

of reports with assumed parameters are lower; however, differences in accuracies are not statistically significant. Table A.2 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.

A.4.4 Results on the use of participants with faulty data

Table A.3 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.

Table A.1: 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

Table A.2: 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 A.3: 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

B

Chapter 5: Supplementary Materials

B.1 Evaluated Datasets

BCI Competition

This dataset is part of BCI competition 2020 [215], data was recorded from 15 participants using a 64 electrode set-up and BrainAmp amplifier. Five phrases for basic communication were recorded: 'hello', 'help me', 'stop', 'thank you', and 'yes'. Participants were instructed to imagine the silent pronunciation of the given word as if it were performed in covert speech, without making any sound or moving articulators.

An auditory cue of the five words/phrases was randomly presented for 2 s, followed by 0.8–1.2 s of across mark. The subjects were instructed to perform imagined speech of the given cue as soon as the cross mark disappears on the screen. Four times of cross mark (0.8–1.2 s) and imagined speech phase (2 s) were followed in a row per random cue. After performing four times of imagined speech, 3 s of the relaxation phase was given to clear up the mind for the next word/phrase. A total of 60 trials per class were recorded.

B.1.1 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 [95].

B.1.2 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 presented 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 [23]. 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.

B.1.3 Liwicki

Data was recorded from 4 participants using a BioSemi Active2 system with a sampling rate of 512Ḣz. A BioSemi EEG 64 channels electrode cap was employed. Two categories, social and number, with four words each, were selected. The two selected categories were mapped into different brain areas, and the selected words appear to have a high word co-occurrence frequency. The social category contained the words child, daughter, father, and wife. The number category contained the words four, three, ten, and six. The textual representation of the words was presented randomly on the screen in front of the participant. During the rest period, the participants were allowed to relax and prepare for the next trial. The total duration of the recording which contained 320 repetitions. Further details on the publication can be found in the author's publication [32].

B.1.4 Malta

The data was recorded using the BioSemi ActiveTwo EEG recording equipment, at a sampling frequency of 2048Ḣz. 24 channels of EEG data from the 10-20 system are available in the dataset. The SI tasks were to perform the words: The data was recorded using the BioSemi ActiveTwo electroencephalogram (EEG) recording equipment, at a sampling frequency of 2.048Ḣz. 24 channels of EEG data from the 10-20 system are available in the dataset. At the start of a run, subjects are given one minute to settle down before the cued trials begin. First, a fixation cross appears on-screen, indicating to the subject to remain relaxed but aware that the next trial is forthcoming. The cue then appears in the form of an arrow, with its direction being associated with a particular task. The subject starts executing the task as soon as they see the cue, and continues even when it has disappeared, until

the fixation cross appears again. The cues consist of a left-facing arrow for 'left', a right-facing arrow for 'right', an upward-facing arrow for 'up' and a corresponding image for 'down'. Each trial, therefore, lasted 4 seconds and 40 trials per class. Further details on the publication can be found in the author's publication [216].

B.1.5 Nguyen

Signal was obtained from 15 subjects that were split into groups to perform different imagined speech tasks, namely short words, long words and vowels. The group of short words included the words 'in', 'out' and 'up', while the group of long words consisted of 'cooperate' and 'independent' and vowels consisted in /a/, /i/ and /u/. The subjects were instructed to pronounce these words internally in their minds and avoid any overt vocalisation or muscle movements. One session comprised of 1000 trials per class. During each trial, the subject would hear a beep sound that was repeated at a period 1.4. This helped create the rhythm that subjects should imagine pronouncing the words or phonemes. At the beginning of the trial, the subject was also prompted with a visual cue indicating the desired word to be imagined. The cue lasted for 7x1.4 s. The subject was instructed to perform speech imagery at each beep sound and continue at the same rhythm until the visual cue disappeared. Data was recorded using a BrainProducts ActiCHamp amplifier system from 64 electrodes and recorded at a sampling rate of 1000Ḣz. Further details can be found in the authro's publication [22].

B.1.6 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 [96]. We considered the 64 channels equivalent to a 64-electrode setup in order to reduce the dimensionality.

B.1.7 Ours

Data was recorded from 16 participants, participants signed a consent form based on recommendations of the Ethical Committee of the University of Essex in January 2023 (Reference Number ETH2223-0220). Data was recorded from a 64-electrode cap using a BioSemi Amplifier system at 2048Ḣz. Each SI trial began with a fixation cross to prompt participants to prepare for the task. This was followed by the imagery prompt, displaying the words "one time". Directional words 'left' and 'right'were employed. Next, another fixation cross appeared to give participants time to memorise the speech unit and prepare mentally. The cue stimulus—a circle displayed for 0.3 s and perceived as a brief flash—then signalled the start of imagery. Task period was 2 s and a total of 25 trials per class.

B.1.8 Ours rhythmic

The same participants and EEG setup were employed as in our previous setup. For the rhythm marking, participants were cued simultaneously with visual and auditory stimuli marking an 800 ms rhythm for five repetitions. We employed 'stop' and 'pinch' words for this task. The rest period varied randomly between 5–10 s. A total of 100 trials were recorded.

B.1.9 Rekrut

Data was recorded from 15 participants using a wireless 64-channel EEG with Brain Products Live Amp 64 amplifier at 500Ḣz. 80 repetitions. Participants were seated in a chair and controlled the simulated robot on a screen in a game-like setup through a maze. They were presented with a birds-view of the robots' surroundings with the robot in the middle. Participants had to decide on its next step and interact for one part of the study via overt and in the second part via imagined speech. The interaction consisted of moving the robot in 3 different directions

resulting in the command words ”left”, ”right” and ”up” and picking up screws and pushing boxes out of the way by the words ”pick” and ”push”. Whenever the user had made a decision about the next command they could press the spacebar to indicate the desire for interaction. After the spacebar was pressed, the screen turned black for 2 s to give the participant time to prepare the input. After the 2 s, a fixation cross appeared, which indicated to start speaking or producing imagined speech of the desired command, depending on the current condition. After 2 s, the fixation cross disappeared, and the few switched back to the robot with its updated position. We recorded 80 repetitions per word and paradigm, meaning for the 5 words. Further detail on the experiment can be found on aouthros publication [231].

B.1.10 Tec

Data was recorded from 15 participants with a 24-channel recording cap and mBrainTrain Smarting amplifier. Used prompts are Spanish command words “avanzar,” “retroceder,” “derecha,” and “izquierda” (which correspond to “advance”, “backwards”, “right” and “left”, respectively). Each trial began with a black screen, followed by the presentation of a visual cue (a written word displayed on a monitor) and an auditory cue (a beep sound delivered through the headphones). The visual cue remained on screen for seven intervals, while the auditory cue was repeated four additional times at a fixed rhythm (period T=1.4 s), establishing the pacing of the task. Participants were instructed to imagine pronouncing the displayed word each time they heard a beep. After the final beep, they continued the task for two additional imagined repetitions at the same rhythm, but without auditory guidance. Only these last three instances of each trial were recorded. 30 trials per class were recorded. Further details on the dataset can be found in the author’s publication [218]

B.1.11 Tec game

Data was recorded from the same participants and equipment setup as the previous description. The paradigm is designed as a game; it consists of a character that

can move in four directions, in Spanish equivalent to: forward, backwards, left, and right. The goal was to escape a maze by visiting all checkpoints in numerical order. The main character of this video game was an animated dog with an identifiable face and tail, which allowed for indicating where it was facing. Cues on this paradigm were given by color changes in the maze borders, indicating user action. White borders indicated no specific action, green borders indicated imagined speech, and blue borders indicated vocalised speech. A total of 30 trials per class was recorded. Further details on the dataset can be found in the author's publication [218]

B.2 Pairwise comparison of statistically and practically significant accuracies

In this section, we present the comprehensive results of all pairwise classification accuracy percentages for datasets containing more than two available classes.

Table B.1: Percentage (%) of participants from the BCIComp dataset achieving statistically significant classification accuracies per class pair.

	hello	help me	stop	thank you	yes
hello	–	73.33	86.67	80.00	100.00
help me	73.33	–	86.67	86.67	100.00
stop	86.67	86.67	–	100.00	100.00
thank you	80.00	86.67	100.00	–	93.33
yes	100.00	100.00	100.00	93.33	–

Table B.2: Percentage (%) of participants from the BCIComp dataset achieving practically significant classification accuracies per class pair.

	hello	help me	stop	thank you	yes
hello	–	33.33	33.33	40.00	46.67
help me	33.33	–	20.00	33.33	46.67
stop	33.33	20.00	–	53.33	60.00
thank you	40.00	33.33	53.33	–	66.67
yes	46.67	46.67	60.00	66.67	–

Table B.3: Percentage (%) of participants from the Liwicki dataset achieving statistically significant classification accuracies per class pair.

	child	daughter	father	four	six	ten	three	wife
child	–	25.00	0.00	0.00	0.00	0.00	0.00	0.00
daughter	25.00	–	25.00	25.00	25.00	0.00	25.00	25.00
father	0.00	25.00	–	0.00	25.00	0.00	25.00	25.00
four	0.00	25.00	0.00	–	25.00	0.00	25.00	25.00
six	0.00	25.00	25.00	25.00	–	0.00	0.00	0.00
ten	0.00	0.00	0.00	0.00	0.00	–	50.00	0.00
three	0.00	25.00	25.00	25.00	0.00	50.00	–	0.00
wife	0.00	25.00	25.00	25.00	0.00	0.00	0.00	–

Table B.4: Percentage (%) of participants from the Liwicki dataset achieving practically significant classification accuracies per class pair.

	child	daughter	father	four	six	ten	three	wife
child	–	25.00	0.00	0.00	0.00	0.00	0.00	0.00
daughter	25.00	–	0.00	25.00	25.00	0.00	0.00	0.00
father	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00
four	0.00	25.00	0.00	–	25.00	0.00	25.00	0.00
six	0.00	25.00	0.00	25.00	–	0.00	0.00	0.00
ten	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00
three	0.00	0.00	0.00	25.00	0.00	0.00	–	0.00
wife	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–

Table B.5: Percentage (%) of participants from the Malta dataset achieving practically significant classification accuracies per class pair.

	down	left	right	up
down	–	0.00	25.00	0.00
left	0.00	–	25.00	25.00
right	25.00	25.00	–	25.00
up	0.00	25.00	25.00	–

Table B.6: Percentage (%) of participants from the Nguyen dataset achieving statistically significant classification accuracies per class pair.

	a	i	in	out	corporate
a	–	50.00	83.33	80.00	83.33
i	50.00	–	0.00	0.00	76.66
in	83.33	0.00	–	0.00	100.00
out	80.00	0.00	0.00	–	76.66
corporate	83.33	76.66	100.00	76.66	–

Table B.7: Percentage (%) of participants from the Nguyen dataset achieving practically significant classification accuracies per class pair.

	a	i	in	out	corporate
a	–	0.00	52.22	23.33	42.22
i	0.00	–	0.00	0.00	23.33
in	52.22	0.00	–	0.00	53.33
out	23.33	0.00	0.00	–	43.66
corporate	42.22	23.33	53.33	43.66	–

Table B.8: Percentage (%) of participants from the Nieto dataset achieving statistically significant classification accuracies per class pair.

	down	left	right	up
down	–	30.00	20.00	30.00
left	30.00	–	0.00	0.00
right	20.00	0.00	–	0.00
up	30.00	0.00	0.00	–

Table B.9: Percentage (%) of participants from the Nieto dataset achieving practically significant classification accuracies per class pair.

	down	left	right	up
down	–	0.00	10.00	10.00
left	0.00	–	0.00	0.00
right	10.00	0.00	–	0.00
up	10.00	0.00	0.00	–

Table B.10: Percentage (%) of participants from the Rekrut dataset achieving statistically significant classification accuracies per class pair.

	Left	Pick	Push	Right	Up
Left	–	8.33	8.33	0.00	0.00
Pick	8.33	–	8.33	16.67	0.00
Push	8.33	8.33	–	8.33	8.33
Right	0.00	16.67	8.33	–	0.00
Up	0.00	0.00	8.33	0.00	–

Table B.11: Percentage (%) of participants from the Coretto dataset achieving statistically significant classification accuracies per class pair.

	a	back	down	e	fwd	i	left	o	right	u	up
a	–	46.67	40.00	13.33	53.33	0.00	0.00	6.67	53.33	6.67	0.00
back	46.67	–	0.00	40.00	0.00	46.67	6.67	40.00	0.00	0.00	6.67
down	40.00	0.00	–	0.00	0.00	26.67	0.00	40.00	6.67	46.67	0.00
e	13.33	40.00	0.00	–	0.00	0.00	60.00	0.00	46.67	13.33	0.00
fwd	53.33	0.00	0.00	0.00	–	53.33	6.67	0.00	0.00	46.67	0.00
i	0.00	46.67	26.67	0.00	53.33	–	60.00	20.00	0.00	6.67	0.00
left	0.00	6.67	0.00	60.00	6.67	60.00	–	60.00	0.00	66.67	6.67
o	6.67	40.00	40.00	0.00	0.00	20.00	60.00	–	0.00	0.00	40.00
right	53.33	0.00	6.67	46.67	0.00	0.00	0.00	0.00	–	46.67	6.67
u	6.67	0.00	46.67	13.33	46.67	6.67	66.67	0.00	46.67	–	0.00
up	0.00	6.67	0.00	0.00	0.00	0.00	6.67	40.00	6.67	0.00	–

Table B.12: Percentage (%) of participants from the Coretto dataset achieving practically significant classification accuracies per class pair.

	a	back	down	e	fwd	i	left	o	right	u	up
a	–	13.33	13.33	6.67	33.33	0.00	0.00	0.00	13.33	0.00	0.00
back	13.33	–	0.00	6.67	0.00	6.67	0.00	6.67	0.00	0.00	6.67
down	13.33	0.00	–	0.00	0.00	20.00	0.00	13.33	0.00	6.67	0.00
e	6.67	6.67	0.00	–	0.00	0.00	20.00	0.00	6.67	0.00	0.00
fwd	33.33	0.00	0.00	0.00	–	6.67	0.00	0.00	0.00	13.33	0.00
i	0.00	6.67	20.00	0.00	6.67	–	20.00	0.00	0.00	0.00	0.00
left	0.00	0.00	0.00	20.00	0.00	20.00	–	13.33	0.00	33.33	0.00
o	0.00	6.67	13.33	0.00	0.00	0.00	13.33	–	0.00	0.00	26.67
right	13.33	0.00	0.00	6.67	0.00	0.00	0.00	0.00	–	6.67	0.00
u	0.00	0.00	6.67	0.00	13.33	0.00	33.33	0.00	6.67	–	0.00
up	0.00	6.67	0.00	0.00	0.00	0.00	0.00	26.67	0.00	0.00	–

Table B.13: Percentage (%) of participants from the Tec game dataset achieving statistically significant classification accuracies per class pair.

	AVANZAR	DERECHA	IZQUIERDA	RETROCEDER
AVANZAR	–	13.33	6.67	6.67
DERECHA	13.33	–	0.00	20.00
IZQUIERDA	6.67	0.00	–	6.67
RETROCEDER	6.67	20.00	6.67	–

Table B.14: Percentage (%) of participants from the Tec game dataset achieving practically significant classification accuracies per class pair.

	AVANZAR	DERECHA	IZQUIERDA	RETROCEDER
AVANZAR	–	13.33	0.00	6.67
DERECHA	13.33	–	0.00	6.67
IZQUIERDA	0.00	0.00	–	0.00
RETROCEDER	6.67	6.67	0.00	–

Table B.15: Percentage (%) of participants from the Tec dataset achieving statistically significant classification accuracies per class pair.

	AVANZAR	DERECHA	IZQUIERDA	RETROCEDER
AVANZAR	–	33.33	33.33	33.33
DERECHA	33.33	–	40.00	13.33
IZQUIERDA	33.33	40.00	–	13.33
RETROCEDER	33.33	13.33	13.33	–

Table B.16: Percentage (%) of participants from the Tec dataset achieving practically significant classification accuracies per class pair.

	AVANZAR	DERECHA	IZQUIERDA	RETROCEDER
AVANZAR	–	33.33	6.67	26.67
DERECHA	33.33	–	20.00	13.33
IZQUIERDA	6.67	20.00	–	13.33
RETROCEDER	26.67	13.33	13.33	–

Bibliography

[1] Xing Tian, Jean Mary Zarate and David Poeppel. 'Mental imagery of speech implicates two mechanisms of perceptual reactivation'. In: *Cortex* 77 (Apr. 2016), pp. 1–12. DOI: 10.1016/j.cortex.2016.01.002.

[2] Sharon Geva. *Inner Speech and Mental Imagery*. Oxford University Press, Oct. 2018. DOI: 10.1093/oso/9780198796640.003.0005.

[3] Willem J. M. Levelt, Ardi Roelofs and Antje S. Meyer. 'A theory of lexical access in speech production'. In: *Behavioral and Brain Sciences* 22.1 (Feb. 1999), pp. 1–38. DOI: 10.1017/s0140525x99001776.

[4] Peter Langland-Hassan and Agustín Vicente. *Introduction*. Oxford University Press, Oct. 2018. DOI: 10.1093/oso/9780198796640.003.0001.

[5] Franziska Stephan, Henrik Saalbach and Sonja Rossi. 'The Brain Differentially Prepares Inner and Overt Speech Production: Electrophysiological and Vascular Evidence'. In: *Brain Sciences* 10.3 (Mar. 2020), p. 148. DOI: 10.3390/brainsci10030148.

[6] Xiaopeng Si et al. 'Imagined speech increases the hemodynamic response and functional connectivity of the dorsal motor cortex'. In: *Journal of Neural Engineering* 18.5 (Oct. 2021), p. 056048. DOI: 10.1088/1741-2552/ac25d9.

[7] StÃ©phanie Martin et al. 'Decoding spectrotemporal features of overt and covert speech from the human cortex'. In: *Frontiers in Neuroengineering* 7 (May 2014). DOI: 10.3389/fneng.2014.00014.

[8] Md Sultan Mahmud, Mohammed Yeasin and Gavin M Bidelman. 'Data-driven machine learning models for decoding speech categorisation from evoked brain responses'. In: *Journal of Neural Engineering* 18.4 (Mar. 2021), p. 046012. DOI: 10.1088/1741-2552/abecf0.

[9] Hiroki Watanabe et al. 'Synchronisation between overt speech envelope and EEG oscillations during imagined speech'. In: *Neuroscience Research* 153 (Apr. 2020), pp. 48–55. DOI: 10.1016/j.neures.2019.04.004.

[10] Jae Moon, Silvia Orlandi and Tom Chau. 'A comparison and classification of oscillatory characteristics in speech perception and covert speech'. In: *Brain Research* 1781 (Apr. 2022), p. 147778. DOI: 10.1016/j.brainres.2022.147778.

[11] Kevin Meng et al. 'Identification of discriminative features for decoding overt and imagined speech using stereotactic electroencephalography'. In: *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*. 2021, pp. 1–6. DOI: 10.1109/BCI51272.2021.9385355.

[12] Margaret F Carr, Shantanu P Jadhav and Loren M Frank. 'Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval'. In: *Nature Neuroscience* 14.2 (Jan. 2011), pp. 147–153. DOI: 10.1038/nn.2732.

[13] Howard Eichenbaum et al. 'Towards a functional organisation of episodic memory in the medial temporal lobe'. In: *Neuroscience; Biobehavioral Reviews* 36.7 (Aug. 2012), pp. 1597–1608. DOI: 10.1016/j.neubiorev.2011.07.006.

[14] Peter Indefrey. 'The Spatial and Temporal Signatures of Word Production Components: A Critical Update'. In: *Frontiers in Psychology* 2 (2011). DOI: 10.3389/fpsyg.2011.00255.

[15] Thomas J Whitford et al. 'Neurophysiological evidence of efference copies to inner speech'. In: *eLife* 6 (Dec. 2017). DOI: 10.7554/elife.28197.

[16] *BCI Definition*. https://bcisociety.org/bci-definition/. Accessed: 30th May 2026.

[17] Benjamin Blankertz et al. 'Neurophysiological predictor of SMR-based BCI performance'. In: *NeuroImage* 51.4 (July 2010), pp. 1303–1309. DOI: 10.1016/j.neuroimage.2010.03.022.

[18] Sylvain Chevallier et al. 'The largest EEG-based BCI reproducibility study for open science: the MOABB benchmark'. In: (Apr. 2024). working paper or preprint.

[19] Richard Jung and Wiltrud Berger. 'Hans Bergers Entdeckung des Elektrenkephalogramms und seine ersten Befunde 1924?1931: His first records in 1924?1931'. In: *Archiv fr Psychiatrie und Nervenkrankheiten* 227.4 (Dec. 1979), pp. 279–300. DOI: 10.1007/bf00344814.

[20] Felix Gembler, Piotr Stawicki and Ivan Volosyak. 'Exploring the possibilities and limitations of multitarget SSVEP-based BCI applications'. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Aug. 2016, pp. 1488–1491. DOI: 10.1109/embc.2016.7590991.

[21] Charles S. DaSalla et al. 'Single-trial classification of vowel speech imagery using common spatial patterns'. In: *Neural Networks* 22.9 (Nov. 2009), pp. 1334–1339. DOI: 10.1016/j.neunet.2009.05.008.

[22] Chuong H Nguyen, George K Karavas and Panagiotis Artemiadis. 'Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features'. In: *Journal of Neural Engineering* 15.1 (Nov. 2017), p. 016002. DOI: 10.1088/1741-2552/aa8235.

[23] 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.

[24] Miguel Angrick et al. 'Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity'. In: *Communications Biology* 4.1 (Sept. 2021). DOI: 10.1038/s42003-021-02578-0.

[25] Xing Tian. 'Mental imagery of speech and movement implicates the dynamics of internal forward models'. In: *Frontiers in Psychology* 1 (2010). DOI: 10.3389/fpsyg.2010.00166.

[26] Alborz Rezazadeh Sereshkeh et al. 'Development of a ternary hybrid fNIRS-EEG brain-computer interface based on imagined speech'. In: *Brain-Computer Interfaces* 6.4 (Oct. 2019), pp. 128–140. DOI: 10.1080/2326263x.2019.1698928.

[27] Margaret C. Thompson. 'Critiquing the Concept of BCI Illiteracy'. In: *Science and Engineering Ethics* 25.4 (Aug. 2018), pp. 1217–1233. DOI: 10.1007/s11948-018-0061-1.

[28] Brendan Z. Allison and Christa Neuper. 'Could Anyone Use a BCI?' In: *Brain-Computer Interfaces*. Springer London, 2010, pp. 35–54. DOI: 10.1007/978-1-84996-272-8_3.

[29] 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. DOI: 10.1371/journal.pone.0147932.

[30] Alan Tonnies Moore and Eric Schwitzgebel. 'The experience of reading'. In: *Consciousness and Cognition* 62 (July 2018), pp. 57–68. DOI: 10.1016/j.concog.2018.03.011.

[31] Ronald P. Endicott. 'Inner speech and the body error theory'. In: *Frontiers in Psychology* 15 (Mar. 2024). DOI: 10.3389/fpsyg.2024.1360699.

[32] Foteini Simistira Liwicki et al. 'Bimodal electroencephalography-functional magnetic resonance imaging dataset for inner-speech recognition'. In: *Scientific Data* 10.1 (June 2023). DOI: 10.1038/s41597-023-02286-w.

[33] Sahil Datta and Nikolaos V. Boulgouris. 'Recognition of grammatical class of imagined words from EEG signals using convolutional neural network'. In: *Neurocomputing* 465 (Nov. 2021), pp. 301–309. DOI: 10.1016/j.neucom.2021.08.035.

[34] Fu Li et al. 'Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network'. In: *Journal of Neural Engineering* 18.4 (Aug. 2021), p. 0460c4. DOI: 10.1088/1741-2552/ac13c0.

[35] A Tates et al. 'Speech imagery brain–computer interfaces: a systematic literature review'. In: *Journal of Neural Engineering* 22.3 (June 2025), p. 031003. DOI: 10.1088/1741-2552/ade28e.

[36] Sandhya Chengaiyan and Kavitha Anandan. 'Effect of functional and effective brain connectivity in identifying vowels from articulation imagery procedures'. In: *Cognitive Processing* 23.4 (July 2022), pp. 593–618. DOI: 10.1007/s10339-022-01103-3.

[37] Sobhan Hemati and Gholam-Ali Hossein-Zadeh. 'Distinct Functional Network Connectivity for Abstract and Concrete Mental Imagery'. In: *Frontiers in Human Neuroscience* 12 (Dec. 2018). DOI: 10.3389/fnhum.2018.00515.

[38] Sahil Bajaj et al. 'Brain effective connectivity during motor-imagery and execution following stroke and rehabilitation'. In: *NeuroImage: Clinical* 8 (2015), pp. 572–582. DOI: 10.1016/j.nicl.2015.06.006.

[39] Dheeraj Rathee, Hubert Cecotti and Girijesh Prasad. 'Single-trial effective brain connectivity patterns enhance discriminability of mental imagery tasks'. In: *Journal of Neural Engineering* 14.5 (Aug. 2017), p. 056005. DOI: 10.1088/1741-2552/aa785c.

[40] Maria Giulia Tullo et al. 'Individual differences in mental imagery modulate effective connectivity of scene-selective regions during resting state'. In: *Brain Structure and Function* 227.5 (Mar. 2022), pp. 1831–1842. DOI: 10.1007/s00429-022-02475-0.

[41] Stephen M. Kosslyn, Giorgio Ganis and William L. Thompson. 'Neural foundations of imagery'. In: *Nature Reviews Neuroscience* 2.9 (Sept. 2001), pp. 635–642. DOI: 10.1038/35090055.

[42] Samuel T. Moulton and Stephen M. Kosslyn. 'Imagining predictions: mental imagery as mental emulation'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1521 (May 2009), pp. 1273–1280. DOI: 10.1098/rstb.2008.0314.

[43] Christoph Reichert et al. 'A Comparative Study on the Detection of Covert Attention in Event-Related EEG and MEG Signals to Control a BCI'. In: *Frontiers in Neuroscience* 11 (Oct. 2017). DOI: 10.3389/fnins.2017.00575.

[44] Mia Illman et al. 'Comparing MEG and EEG in detecting the 20-Hz rhythm modulation to tactile and proprioceptive stimulation'. In: *NeuroImage* 215 (July 2020), p. 116804. DOI: 10.1016/j.neuroimage.2020.116804.

[45] Mostafa Orban et al. 'A Review of Brain Activity and EEG-Based Brain–Computer Interfaces for Rehabilitation Application'. In: *Bioengineering* 9.12 (Dec. 2022), p. 768. DOI: 10.3390/bioengineering9120768.

[46] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. 'Brain Computer Interfaces, a Review'. In: *Sensors* 12.2 (Jan. 2012), pp. 1211–1279. DOI: 10.3390/s120201211.

[47] Daniel M. Wolpert and Zoubin Ghahramani. 'Computational principles of movement neuroscience'. In: *Nature Neuroscience* 3.S11 (Nov. 2000), pp. 1212–1217. DOI: 10.1038/81497.

[48] Mark L. Latash. 'Efference copy in kinesthetic perception: a copy of what is it?' In: *Journal of Neurophysiology* 125.4 (2021). PMID: 33566734, pp. 1079–1094. DOI: 10.1152/jn.00545.2020. eprint: https://doi.org/10.1152/jn.00545.2020.

[49] Donald J Crammond. 'Motor imagery: never in your wildest dream'. In: *Trends in Neurosciences* 20.2 (Feb. 1997), pp. 54–57. DOI: 10.1016/s0166-2236(96)30019-2.

[50] William S. Anderson and Frederick A. Lenz. 'Review of motor and phantom-related imagery'. In: *NeuroReport* 22.17 (Dec. 2011), pp. 939–942. DOI: 10.1097/wnr.0b013e32834ca58d.

[51] G Onose et al. 'On the feasibility of using motor imagery EEG-based brain–computer interface in chronic tetraplegics for assistive robotic arm control: a clinical test and long-term post-trial follow-up'. In: *Spinal Cord* 50.8 (Mar. 2012), pp. 599–608. DOI: 10.1038/sc.2012.14.

[52] Lei Cao et al. 'A Synchronous Motor Imagery Based Neural Physiological Paradigm for Brain Computer Interface Speller'. In: *Frontiers in Human Neuroscience* 11 (May 2017). DOI: 10.3389/fnhum.2017.00274.

[53] Szczepan Paszkiel. 'Control Based on Brain-Bomputer Interface Technology for Video-Gaming With Virtual Reality Techniques'. In: *Journal of Automation, Mobile Robotics and Intelligent Systems* 10.4 (Dec. 2016), pp. 3–7. DOI: 10.14313/jamris_4-2016/26.

[54] Li Wang, Xiong Zhang and Yu Zhang. 'Extending motor imagery by speech imagery for brain-computer interface'. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2013, pp. 7056–7059. DOI: 10.1109/EMBC.2013.6611183.

[55] Romain Grandchamp et al. 'The ConDialInt Model: Condensation, Dialogality, and Intentionality Dimensions of Inner Speech Within a Hierarchical Predictive Control Framework'. In: *Frontiers in Psychology* 10 (Sept. 2019). DOI: 10.3389/fpsyg.2019.02019.

[56] Yosef Grodzinsky and Andrea Santi. 'The battle for Broca's region'. In: *Trends in Cognitive Sciences* 12.12 (Dec. 2008), pp. 474–480. DOI: 10.1016/j.tics.2008.09.001.

[57] Luciano Fadiga, Laila Craighero and Alessandro D'Ausilio. 'Broca's Area in Language, Action, and Music'. In: *Annals of the New York Academy of Sciences* 1169.1 (July 2009), pp. 448–458. DOI: 10.1111/j.1749-6632.2009.04582.x.

[58] Evelina Fedorenko and Idan A. Blank. 'Broca's Area Is Not a Natural Kind'. In: *Trends in Cognitive Sciences* 24.4 (Apr. 2020), pp. 270–284. DOI: 10.1016/j.tics.2020.01.001.

[59] Brenda Rapp, Adam Buchwald and Matthew Goldrick. 'Integrating accounts of speech production: the devil is in the representational details'. In: *Language, Cognition and Neuroscience* 29.1 (Oct. 2013), pp. 24–27. DOI: 10.1080/01690965.2013.848991.

[60] Bogen GM Bogen JE. 'WERNICKE'S REGION–WHERE IS IT'. In: (1976).

[61] K. Goldstein. *Language and Language Disturbances: Aphasic Symptom Complexes and Their Significance for Medicine and Theory of Language*. Grune & Stratton, 1948.

[62] W.J.M. Levelt. *Speaking: From Intention to Articulation*. ACL-MIT Series in Natural Language Processing. MIT Press, 1993.

[63] Jeffrey R. Binder. 'The Wernicke area: Modern evidence and a reinterpretation'. In: *Neurology* 85.24 (Dec. 2015), pp. 2170–2175. DOI: 10.1212/wnl.0000000000002219.

[64] P Indefrey and W.J.M Levelt. 'The spatial and temporal signatures of word production components'. In: *Cognition* 92.1–2 (May 2004), pp. 101–144. DOI: 10.1016/j.cognition.2002.06.001.

[65] Gregory Hickok and David Poeppel. 'The cortical organisation of speech processing'. In: *Nature Reviews Neuroscience* 8.5 (Apr. 2007), pp. 393–402. DOI: 10.1038/nrn2113.

[66] James L. McClelland and Timothy T. Rogers. 'The parallel distributed processing approach to semantic cognition'. In: *Nature Reviews Neuroscience* 4.4 (Apr. 2003), pp. 310–322. DOI: 10.1038/nrn1076.

[67] Adeen Flinker et al. 'Redefining the role of Broca's area in speech'. In: *Proceedings of the National Academy of Sciences* 112.9 (Feb. 2015), pp. 2871–2875. DOI: 10.1073/pnas.1414491112.

[68] Kyuya Kogure and Takashi Yoshimoto. *XVIth International Symposium on Cerebral Blood Flow and Metabolism: Sendai International Centre, Sendai, Japan, May 22-28, 1993*. Journal of cerebral blood flow and metabolism v. 13, no. 1. Raven Press, 1993.

[69] P. K. McGuire et al. 'The Neural Correlates of Inner Speech and Auditory Verbal Imagery in Schizophrenia: Relationship to Auditory Verbal Hallucinations'. In: *British Journal of Psychiatry* 169.2 (Aug. 1996), pp. 148–159. DOI: 10.1192/bjp.169.2.148.

[70] S. S. SHERGILL et al. 'A functional study of auditory verbal imagery'. In: *Psychological Medicine* 31.2 (Feb. 2001), pp. 241–253. DOI: 10.1017/s003329170100335x.

[71] Howard J. Rosen et al. 'Comparison of Brain Activation during Word Retrieval Done Silently and Aloud Using fMRI'. In: *Brain and Cognition* 42.2 (Mar. 2000), pp. 201–217. DOI: 10.1006/brcg.1999.1100.

[72] Erica D. Palmer et al. 'An Event-Related fMRI Study of Overt and Covert Word Stem Completion'. In: *NeuroImage* 14.1 (July 2001), pp. 182–193. DOI: 10.1006/nimg.2001.0779.

[73] L SHUSTER and S LEMIEUX. 'An fMRI investigation of covertly and overtly produced mono- and multisyllabic words'. In: *Brain and Language* 93.1 (Apr. 2005), pp. 20–31. DOI: 10.1016/j.bandl.2004.07.007.

[74] Jie Huang, Thomas H. Carr and Yue Cao. 'Comparing cortical activations for silent and overt speech using event-related fMRI'. In: *Human Brain Mapping* 15.1 (Oct. 2001), pp. 39–53. DOI: 10.1002/hbm.1060.

[75] Lingxi Lu et al. 'Common and distinct neural representations of imagined and perceived speech'. In: *Cerebral Cortex* 33.10 (Dec. 2022), pp. 6486–6493. DOI: 10.1093/cercor/bhac519.

[76] Xing Tian and David Poeppel. 'Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation'. In: *Frontiers in Human Neuroscience* 6 (2012). DOI: 10.3389/fnhum.2012.00314.

[77] Ladislas Nalborczyk et al. 'The Role of Motor Inhibition During Covert Speech Production'. In: *Frontiers in Human Neuroscience* 16 (Mar. 2022). DOI: 10.3389/fnhum.2022.804832.

[78] Stephanie J. Forkel and Peter Hagoort. 'Redefining language networks: connectivity beyond localised regions'. In: *Brain Structure and Function* 229.9 (Nov. 2024), pp. 2073–2078. DOI: 10.1007/s00429-024-02859-4.

[79] Matthew J Page et al. 'PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews'. In: *BMJ* 372 (2021). DOI: 10.1136/bmj.n160. eprint: https://www.bmj.com/content/372/bmj.n160.full.pdf.

[80] Mehrdad Fatourechi et al. 'Is Information Transfer Rate a Suitable Performance Measure for Self-paced Brain Interface Systems?' In: *2006 IEEE International Symposium on Signal Processing and Information Technology*. 2006, pp. 212–216. DOI: 10.1109/ISSPIT.2006.270799.

[81] Martin Billinger et al. 'Is It Significant? Guidelines for Reporting BCI Performance'. In: *Towards Practical Brain-Computer Interfaces*. Springer Berlin Heidelberg, 2012, pp. 333–354. DOI: 10.1007/978-3-642-29746-5_17.

[82] Nicholas S. Card et al. 'An accurate and rapidly calibrating speech neuroprosthesis'. In: (Dec. 2023). DOI: 10.1101/2023.12.26.23300110.

[83] David A. Moses et al. 'Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria'. In: *New England Journal of Medicine* 385.3 (July 2021), pp. 217–227. DOI: 10.1056/nejmoa2027540.

[84] Francis R. Willett et al. 'A high-performance speech neuroprosthesis'. In: *Nature* 620.7976 (Aug. 2023), pp. 1031–1036. DOI: 10.1038/s41586-023-06377-x.

[85] Ahmed Ali and Steve Renals. 'Word Error Rate Estimation for Speech Recognition: e-WER'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2018. DOI: 10.18653/v1/p18-2004.

[86] Erin Hecht and Dietrich Stout. 'Techniques for Studying Brain Structure and Function'. In: *Human Paleoneurology*. Springer International Publishing, July 2014, pp. 209–224. DOI: 10.1007/978-3-319-08500-5_9.

[87] Borís Burle et al. 'Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view'. In: *International Journal of Psychophysiology* 97.3 (2015). On the benefits of using surface Laplacian (current source density) methodology in electrophysiology, pp. 210–220. DOI: https://doi.org/10.1016/j.ijpsycho.2015.05.004.

[88] Emily M Mugler et al. 'Direct classification of all American English phonemes using signals from functional speech motor cortex'. In: *Journal of Neural Engineering* 11.3 (May 2014), p. 035015. DOI: 10.1088/1741-2560/11/3/035015.

[89] Han-Jeong Hwang et al. 'Toward more intuitive brain–computer interfacing: classification of binary covert intentions using functional near-infrared spectroscopy'. In: *Journal of Biomedical Optics* 21.9 (Apr. 2016), p. 091303. DOI: 10.1117/1.jbo.21.9.091303.

[90] Zengzhi Guo and Fei Chen. 'Idle-state detection in motor imagery of articulation using early information: A functional Near-infrared spectroscopy study'. In: *Biomedical Signal Processing and Control* 72 (Feb. 2022), p. 103369. DOI: 10.1016/j.bspc.2021.103369.

[91] Alborz Rezazadeh Sereshkeh et al. 'Online classification of imagined speech using functional near-infrared spectroscopy signals'. In: *Journal of Neural Engineering* 16.1 (Nov. 2018), p. 016005. DOI: 10.1088/1741-2552/aae4b9.

[92] Christian Herff et al. 'Cross-Subject Classification of Speaking Modes Using fNIRS'. In: *Neural Information Processing*. Springer Berlin Heidelberg, 2012, pp. 417–424. DOI: 10.1007/978-3-642-34481-7_51.

[93] Ciaran Cooney, Raffaella Folli and Damien Coyle. 'A Bimodal Deep Learning Architecture for EEG-fNIRS Decoding of Overt and Imagined Speech'. In: *IEEE Transactions on Biomedical Engineering* 69.6 (June 2022), pp. 1983–1994. DOI: 10.1109/tbme.2021.3132861.

[94] Li Wang et al. 'Analysis and classification of speech imagery EEG for BCI'. In: *Biomedical Signal Processing and Control* 8.6 (Nov. 2013), pp. 901–908. DOI: 10.1016/j.bspc.2013.07.011.

[95] 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.

[96] 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). DOI: 10.1038/s41597-022-01147-2.

[97] Connie Cheung et al. 'The auditory representation of speech sounds in human motor cortex'. In: *eLife* 5 (Mar. 2016). Ed. by Barbara G Shinn-Cunningham, e12577. DOI: 10.7554/eLife.12577.

[98] Kristofer E. Bouchard et al. 'Functional organisation of human sensorimotor cortex for speech articulation'. In: *Nature* 495.7441 (Feb. 2013), pp. 327–332. DOI: 10.1038/nature11911.

[99] Rajdeep Ghosh, Nidul Sinha and Souvik Phadikar. 'Classification of Silent Speech in English and Bengali Languages Using Stacked Autoencoder'. In: *SN Computer Science* 3.5 (July 2022). DOI: 10.1007/s42979-022-01274-y.

[100] Amir Jahangiri and Francisco Sepulveda. 'The contribution of different frequency bands in class separability of covert speech tasks for BCIs'. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2017, pp. 2093–2096. DOI: 10.1109/embc.2017.8037266.

[101] Xinyu Zhang, Hua Li and Fei Chen. 'EEG-based Classification of Imaginary Mandarin Tones'. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine; Biology Society (EMBC)*. IEEE, July 2020, pp. 3889–3892. DOI: 10.1109/embc44109.2020.9176608.

[102] Hongguang Pan Zhuoyi Li Chen Tian Li Wang Fu Fei Liu. 'The LightGBM-based classification algorithm for Chinese characters speech imagery BCI system'. In: *Cognitive Neurodynamics* 17.2 (June 2022), pp. 373–384. DOI: 10.1007/s11571-022-09819-w.

[103] Ciaran Cooney et al. 'Evaluation of Hyperparameter Optimisation in Machine and Deep Learning Methods for Decoding Imagined Speech EEG'. In: *Sensors* 20.16 (Aug. 2020), p. 4629. DOI: 10.3390/s20164629.

[104] Netiwit Kaongoen, Jaehoon Choi and Sungho Jo. 'Speech-imagery-based brain–computer interface system using ear-EEG'. In: *Journal of Neural Engineering* 18.1 (Feb. 2021), p. 016023. DOI: 10.1088/1741-2552/abd10e.

[105] Byeong-Hoo Lee et al. 'Speech Imagery Classification using Length-Wise Training based on Deep Learning'. In: *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*. 2021, pp. 1–5. DOI: 10.1109/BCI51272.2021.9385347.

[106] Mokhles M. Abdulghani, Wilbur L. Walters and Khalid H. Abed. 'Imagined Speech Classification Using EEG and Deep Learning'. In: *Bioengineering* 10.6 (May 2023), p. 649. DOI: 10.3390/bioengineering10060649.

[107] Wonjun Ko, Eunjin Jeon and Heung-Il Suk. 'Spectro-Spatio-Temporal EEG Representation Learning for Imagined Speech Recognition'. In: *Pattern Recognition*. Springer International Publishing, 2022, pp. 335–346. DOI: 10.1007/978-3-031-02444-3_25.

[108] Xiaomei Pei et al. 'Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans'. In: *Journal of Neural Engineering* 8.4 (July 2011), p. 046028. DOI: 10.1088/1741-2560/8/4/046028.

[109] Aref Einizade et al. 'Neural decoding of imagined speech from EEG signals using the fusion of graph signal processing and graph learning techniques'. In: *Neuroscience Informatics* 2.3 (Sept. 2022), p. 100091. DOI: 10.1016/j.neuri.2022.100091.

[110] Margaret Anne Defeyter, Riccardo Russo and Pamela Louise McPartlin. 'The picture superiority effect in recognition memory: A developmental study using the response signal procedure'. In: *Cognitive Development* 24.3 (July 2009), pp. 265–273. DOI: 10.1016/j.cogdev.2009.05.002.

[111] Ciaran Cooney, Raffaella Folli and Damien Coyle. 'Opportunities, pitfalls and trade-offs in designing protocols for measuring the neural correlates of speech'. In: *Neuroscience; Biobehavioral Reviews* 140 (Sept. 2022), p. 104783. DOI: 10.1016/j.neubiorev.2022.104783.

[112] Jerrin Thomas Panachakel and Ramakrishnan A G. 'Classification of Phonological Categories in Imagined Speech using Phase Synchronisation Measure'. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine; Biology Society (EMBC)*. IEEE, Nov. 2021, pp. 2226–2229. DOI: 10.1109/embc46164.2021.9630699.

[113] Hyeong-jun Park and Boreom Lee. 'Multiclass classification of imagined speech EEG using noise-assisted multivariate empirical mode decomposition and multireceptive field convolutional neural network'. In: *Frontiers in Human Neuroscience* 17 (Aug. 2023). DOI: 10.3389/fnhum.2023.1186594.

[114] Stephanie Martin et al. 'Word pair classification during imagined speech using direct brain recordings'. In: *Scientific Reports* 6.1 (May 2016). DOI: 10.1038/srep25803.

[115] William J. Ray and Harry W. Cole. 'EEG Alpha Activity Reflects Attentional Demands, and Beta Activity Reflects Emotional and Cognitive Processes'. In: *Science* 228.4700 (May 1985), pp. 750–752. DOI: 10.1126/science.3992243.

[116] Hongli Yu et al. 'Effects of Motor Imagery Tasks on Brain Functional Networks Based on EEG Mu/Beta Rhythm'. In: *Brain Sciences* 12.2 (Jan. 2022), p. 194. DOI: 10.3390/brainsci12020194.

[117] Beomjun Min et al. 'Vowel Imagery Decoding toward Silent Speech BCI Using Extreme Learning Machine with Electroencephalogram'. In: *BioMed Research International* 2016 (2016), pp. 1–11. DOI: 10.1155/2016/2618265.

[118] Seo-Hyun Lee, Minji Lee and Seong-Whan Lee. 'Neural Decoding of Imagined Speech and Visual Imagery as Intuitive Paradigms for BCI Communication'. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 9268982* 28.12 (2020), pp. 2647–2659. DOI: 10.1109/TNSRE.2020.3040289.

[119] Ashwin Kamble et al. 'Spectral Analysis of EEG Signals for Automatic Imagined Speech Recognition'. In: *IEEE Transactions on Instrumentation and Measurement* 72 (2023), pp. 1–9. DOI: 10.1109/TIM.2023.3300473.

[120] Dipti Pawar and Sudhir Dhage. 'Imagined Speech Classification using EEG based Brain-Computer Interface'. In: *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)*. 2022, pp. 662–666. DOI: 10.1109/CSNT54456.2022.9787644.

[121] R. Anandha Sree and A. Kavitha. 'Vowel classification from imagined speech using sub-band EEG frequencies and deep belief networks'. In: *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*. 2017, pp. 1–4. DOI: 10.1109/ICSCN.2017.8085710.

[122] 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. DOI: 10.3390/neurosci3020017.

[123] Alejandro A. Torres-Garcia, Reyes-Garciaa and Gregorio GarcÃa-Aguilar. 'Implementing a fuzzy inference system in a multi-objective EEG channel selection model for imagined speech classification'. In: *Expert Systems with Applications* 59 (Oct. 2016), pp. 1–12. DOI: 10.1016/j.eswa.2016.04.011.

[124] Robert Luke, Maureen J. Shader and David McAlpine. 'Characterisation of Mayer-wave oscillations in functional near-infrared spectroscopy using a physiologically informed model of the neural power spectra'. In: *Neurophotonics* 8.04 (Dec. 2021). DOI: 10.1117/1.nph.8.4.041001.

[125] Seo-Hyun Lee, Minji Lee and Seong-Whan Lee. 'Neural Decoding of Imagined Speech and Visual Imagery as Intuitive Paradigms for BCI Communication'. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.12 (2020), pp. 2647–2659. DOI: 10.1109/TNSRE.2020.3040289.

[126] H. Ramoser, J. Muller-Gerking and G. Pfurtscheller. 'Optimal spatial filtering of single trial EEG during imagined hand movement'. In: *IEEE Transactions on Rehabilitation Engineering* 8.4 (2000), pp. 441–446. DOI: 10.1109/86.895946.

[127] A. A. T. Garcia, C. Garcia and Luis Villasellor-Pineda. 'Toward a Silent Speech Interface based on Unspoken Speech'. In: *International Conference on Bio-inspired Systems and Signal Processing* ().

[128] Yash V. Varshney and Azizuddin Khan. 'Imagined Speech Classification Using Six Phonetically Distributed Words'. In: *Frontiers in Signal Processing* 2 (Mar. 2022). DOI: 10.3389/frsip.2022.760643.

[129] Basil M. Idrees and Omar Farooq. 'Vowel classification using wavelet decomposition during speech imagery'. In: *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*. 2016, pp. 636–640. DOI: 10.1109/SPIN.2016.7566774.

[130] M. R. Asghari Bejestani et al. 'EEG-Based Multiword Imagined Speech Classification for Persian Words'. In: *BioMed Research International* 2022 (Jan. 2022). Ed. by Yue Zhang, pp. 1–20. DOI: 10.1155/2022/8333084.

[131] Anaum Riaz et al. 'Inter comparison of classification techniques for vowel speech imagery using EEG sensors'. In: *The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014)*. 2014, pp. 712–717. DOI: 10.1109/ICSAI.2014.7009378.

[132] 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.

[133] Arman Hossain et al. 'A BCI system for imagined Bengali speech recognition'. In: *Machine Learning with Applications* 13 (Sept. 2023), p. 100486. DOI: 10.1016/j.mlwa.2023.100486.

[134] Yunlong Gao Yongsheng Zhao Ying Liu. 'Analysis and Classification of Speech Imagery EEG Based on Chinese Initials'. In: *Journal of Beijing Institute of Technology* 30.zk, 44 (2021), p. 44. DOI: 10.15918/j.jbit1004-0579.20095.

[135] 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. DOI: 10.1049/sil2.12059.

[136] P. Agarwal and Sandeep Kumar. 'Electroencephalography based imagined alphabets classification using spatial and timeâdomain features'. In: *International Journal of Imaging Systems and Technology* ().

[137] Eric C Leuthardt et al. 'A brain–computer interface using electrocorticographic signals in humans*'. In: *Journal of Neural Engineering* 1.2 (June 2004), p. 63. DOI: 10.1088/1741-2560/1/2/001.

[138] P.P. Mini, Tessamma Thomas and R. Gopikakumari. 'EEG based direct speech BCI system using a fusion of SMRT and MFCC/LPCC features with ANN classifier'. In: *Biomedical Signal Processing and Control* 68 (July 2021), p. 102625. DOI: 10.1016/j.bspc.2021.102625.

[139] Ana-Luiza Rusnac and Ovidiu Grigore. 'Generalised Brain Computer Interface System for EEG Imaginary Speech Recognition'. In: *2020 24th International Conference on Circuits, Systems, Communications and Computers (CSCC)*. 2020, pp. 184–188. DOI: 10.1109/CSCC49995.2020.00040.

[140] Jerrin Thomas Panachakel, A.G. Ramakrishnan and T.V. Ananthapadmanabha. 'Decoding Imagined Speech using Wavelet Features and Deep Neural Networks'. In: *2019 IEEE 16th India Council International Conference (INDICON)*. 2019, pp. 1–4. DOI: 10.1109/INDICON47234.2019.9028925.

[141] Dipti Pawar and Sudhir Dhage. 'Multiclass covert speech classification using extreme learning machine'. In: *Biomedical Engineering Letters* 10.2 (Mar. 2020), pp. 217–226. DOI: 10.1007/s13534-020-00152-x.

[142] Jaehoon Choi, Netiwit Kaongoen and Sungho Jo. 'Investigation on Effect of Speech Imagery EEG Data Augmentation with Actual Speech'. In: *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*. 2022, pp. 1–5. DOI: 10.1109/BCI53720.2022.9735108.

[143] Mohamad Amin Bakhshali, Morteza Khademi and Ebrahimi-Moghadam. 'EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density'. In: *Biomedical Signal Processing and Control* 59 (May 2020), p. 101899. DOI: 10.1016/j.bspc.2020.101899.

[144] Netiwit Kaongoen, Jaehoon Choi and Sungho Jo. 'A novel online BCI system using speech imagery and ear-EEG for home appliances control'. In: *Computer Methods and Programs in Biomedicine* 224 (Sept. 2022), p. 107022. DOI: 10.1016/j.cmpb.2022.107022.

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

[146] Pramit Saha and Sidney Fels. 'Hierarchical Deep Feature Learning for Decoding Imagined Speech from EEG'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), 10019â10020. DOI: 10.1609/aaai.v33i01.330110019.

[147] Maanvi Angarai Ganesan Ramakrishnan Kanishka Sharma Jerrin Thomas Panachakel Nandagopal Netrakanti Vinayak Nunna. 'An Improved EEG Acquisition Protocol Facilitates Localised Neural Activation'. In: *Advances in Communication Systems and Networks*. Springer Singapore, 2020, pp. 267–281. DOI: 10.1007/978-981-15-3992-3_22.

[148] Timothée Proix et al. 'Imagined speech can be decoded from low- and cross-frequency intracranial EEG features'. In: *Nature Communications* 13.1 (Jan. 2022). DOI: 10.1038/s41467-021-27725-3.

[149] Zengzhi Guo and Fei Chen. 'Decoding lexical tones and vowels in imagined tonal monosyllables using fNIRS signals'. In: *Journal of Neural Engineering* 19.6 (Nov. 2022), p. 066007. DOI: 10.1088/1741-2552/ac9e1d.

[150] A.C. Iliopoulos I. Papasotiriou. 'Functional Complex Networks Based on Operational Architectonics: Application on EEG-based Brain-computer Interface for Imagined Speech'. In: *Neuroscience* 484 (Feb. 2022), pp. 98–118. DOI: 10.1016/j.neuroscience.2021.11.045.

[151] JosÃ© M. MacÃas-MacÃas et al. 'Interpretation of a deep analysis of speech imagery features extracted by a capsule neural network'. In: *Computers in Biology and Medicine* 159 (June 2023), p. 106909. DOI: 10.1016/j.compbiomed.2023.106909.

[152] Robin Tibor Schirrmeister et al. 'Deep learning with convolutional neural networks for EEG decoding and visualisation'. In: *Human Brain Mapping* 38.11 (Aug. 2017), pp. 5391–5420. DOI: 10.1002/hbm.23730.

[153] Vernon J Lawhern et al. 'EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces'. In: *Journal of Neural Engineering* 15.5 (July 2018), p. 056013. DOI: 10.1088/1741-2552/aace8c.

[154] Georgios Rousis et al. 'Combining EEGNet with SPDNet Towards an End To End Architecture for Imagined Speech Decoding'. In: (2024), pp. 1531–1535. DOI: 10.23919/EUSIPCO63174.2024.10715364.

[155] José Manuel Macías-Macías et al. 'Deep Learning Networks for Vowel Speech Imagery'. In: *2020 17th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*. 2020, pp. 1–6. DOI: 10.1109/CCE50788.2020.9299143.

[156] Akihiko Tsukahara et al. 'Analysis of EEG Frequency Components and an Examination of Electrodes Localisation during Speech Imagery'. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2019, pp. 4698–4702. DOI: 10.1109/embc.2019.8857047.

[157] Rini A Sharon and Hema A Murthy. 'Correlation based Multi-phasal models for improved imagined speech EEG recognition'. In: (2020). arXiv: 2011.02195 [astro-ph.IM].

[158] Dong-Yeon Lee, Minji Lee and Seong-Whan Lee. 'Classification of Imagined Speech Using Siamese Neural Network'. In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Oct. 2020, 2979â2984. DOI: 10.1109/smc42975.2020.9282982.

[159] Muhammad Naveed Iqbal Qureshi et al. 'Multiclass Classification of Word Imagination Speech With Hybrid Connectivity Features'. In: *IEEE Transactions on Biomedical Engineering* 65.10 (2018), pp. 2168–2177. DOI: 10.1109/TBME.2017.2786251.

[160] Jing Wang and Li Wang. 'Parallel Convolutional Neural Network Based on Multi-Band Brain Networks for EEG Classification'. In: *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*. 2022, pp. 49–53. DOI: 10.1109/AEMCSE55572.2022.00016.

[161] Ji-Hoon Jeong et al. 'Real-Time Deep Neurolinguistic Learning Enhances Noninvasive Neural Language Decoding for Brain–Machine Interaction'. In: *IEEE Transactions on Cybernetics* 53.12 (2023), pp. 7469–7482. DOI: 10.1109/TCYB.2022.3211694.

[162] Byeong-Hoo Lee et al. 'Speech Imagery Classification using Length-Wise Training based on Deep Learning'. In: *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*. 2021, pp. 1–5. DOI: 10.1109/BCI51272.2021.9385347.

[163] Jigar Patel and Syed Abudhagir Umar. 'Detection of Imagery Vowel Speech Using Deep Learning'. In: *Advances in Energy Technology*. Springer Singapore, July 2021, 237â247. DOI: 10.1007/978-981-16-1476-7_23.

[164] Alan Hernandez-Galvan, Graciela Ramirez-Alonso and Juan Ramirez-Quintana. 'A prototypical network for few-shot recognition of speech imagery data'. In: *Biomedical Signal Processing and Control* 86 (Sept. 2023), p. 105154. DOI: 10.1016/j.bspc.2023.105154.

[165] Md. Monirul Islam and Md. Maruf Hossain Shuvo. 'DenseNet Based Speech Imagery EEG Signal Classification using Gramian Angular Field'. In: *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*. 2019, pp. 149–154. DOI: 10.1109/ICAEE48663.2019.8975572.

[166] 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. DOI: 10.3390/s22134679.

[167] Tsuneo Nitta et al. 'Linguistic representation of vowels in speech imagery EEG'. In: *Frontiers in Human Neuroscience* 17 (May 2023). DOI: 10.3389/fnhum.2023.1163578.

[168] Jess S. Garcia-Salinas et al. 'Transfer learning in imagined speech EEG-based BCIs'. In: *Biomedical Signal Processing and Control* 50 (Apr. 2019), pp. 151–157. DOI: 10.1016/j.bspc.2019.01.006.

[169] 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. DOI: 10.1016/j.bspc.2023.104933.

[170] Vinícius Rezende Carvalho et al. 'Decoding imagined speech with delay differential analysis'. In: *Frontiers in Human Neuroscience* 18 (May 2024). DOI: 10.3389/fnhum.2024.1398065.

[171] Claudia Lainscsek et al. 'Delay Differential Analysis of Seizures in Multichannel Electrocorticography Data'. In: *Neural Computation* 29.12 (Dec. 2017), pp. 3181–3218. DOI: 10.1162/neco_a_01009.

[172] Shizhe Wu et al. 'Adaptive LDA Classifier Enhances Real-Time Control of an EEG Brain–Computer Interface for Decoding Imagined Syllables'. In: *Brain Sciences* 14.3 (Feb. 2024), p. 196. DOI: 10.3390/brainsci14030196.

[173] Camille Jeunet, Emilie Jahanpour and Fabien Lotte. 'Why standard brain-computer interface (BCI) training protocols should be changed: an experimental study'. In: *Journal of Neural Engineering* 13.3 (May 2016), p. 036024. DOI: 10.1088/1741-2560/13/3/036024.

[174] Aurélie de Borman et al. 'Imagined speech event detection from electrocorticography and its transfer between speech modes and subjects'. In: *Communications Biology* 7.1 (July 2024). DOI: 10.1038/s42003-024-06518-6.

[175] Abdulrahman Mohamed Selim et al. 'Speech Imagery BCI Training Using Game with a Purpose'. In: *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*. AVI '24. Arenzano, Genoa, Italy: Association for Computing Machinery, 2024. DOI: 10.1145/3656650.3656654.

[176] Fabien Lotte, Florian Larrue and Christian Mühl. 'Flaws in current human training protocols for spontaneous Brain-Computer Interfaces: lessons learned from instructional design'. In: *Frontiers in Human Neuroscience* 7 (2013). DOI: 10.3389/fnhum.2013.00568.

[177] O. Alkoby et al. 'Can We Predict Who Will Respond to Neurofeedback? A Review of the Inefficacy Problem and Existing Predictors for Successful EEG Neurofeedback Learning'. In: *Neuroscience* 378 (May 2018), pp. 155–164. DOI: 10.1016/j.neuroscience.2016.12.050.

[178] Milan Rybář, Riccardo Poli and Ian Daly. 'Using data from cue presentations results in grossly overestimating semantic BCI performance'. In: *Scientific Reports* 14.1 (Nov. 2024). DOI: 10.1038/s41598-024-79309-y.

[179] Brainard and. 'The Psychophysics Toolbox'. In: *Spatial vision* 10.4 (1997).

[180] 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. DOI: 10.1016/s0927-5398(03)00007-0.

[181] Alexandre Barachant et al. *pyRiemann/pyRiemann: v0.3*. 2022. DOI: 10.5281/ZENODO.7547583.

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

[183] Gernot Mueller-Putz et al. 'Better than random? A closer look on BCI results'. In: 2008.

[184] 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). DOI: 10.1186/s12984-019-0493-0.

[185] Natasha Padfield et al. 'A Comprehensive Review of Endogenous EEG-Based BCIs for Dynamic Device Control'. In: *Sensors* 22.15 (Aug. 2022), p. 5802. DOI: 10.3390/s22155802.

[186] 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). DOI: 10.1038/s42003-025-07464-7.

[187] 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. DOI: 10.1109/tbme.2016.2539602.

[188] 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). DOI: 10.3389/fnrgo.2024.1461494.

[189] Simanto Saha et al. 'Progress in Brain Computer Interface: Challenges and Opportunities'. In: *Frontiers in Systems Neuroscience* 15 (Feb. 2021). DOI: 10.3389/fnsys.2021.578875.

[190] Sean Kinahan et al. 'Achieving Reproducibility in EEG-Based Machine Learning'. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAcc 24. ACM, June 2024, pp. 1464–1474. DOI: 10.1145/3630106.3658983.

[191] 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.

[192] 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). DOI: 10.3389/fnrgo.2024.1346794.

[193] 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. DOI: 10.7717/peerj-cs.2938.

[194] 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. DOI: 10.1109/tbme.2004.827072.

[195] 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). DOI: 10.1093/gigascience/giz002.

[196] 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. DOI: 10.1371/journal.pone.0114853.

[197] Eric Larson et al. *MNE-Python*. 2025. DOI: 10.5281/ZENODO.15928841.

[198] Gregory Lee et al. 'PyWavelets: A Python package for wavelet analysis'. In: *Journal of Open Source Software* 4.36 (Apr. 2019), p. 1237. DOI: 10.21105/joss.01237.

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

[200] Ciaran Cooney, Raffaella Folli and Damien Coyle. 'Optimising Layers Improves CNN Generalisation 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.

[201] Aya Kabbara et al. 'Successful reproduction of a large EEG study across software packages'. In: *Neuroimage: Reports* 3.2 (June 2023), p. 100169. DOI: 10.1016/j.ynirp.2023.100169.

[202] 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). DOI: 10.1111/psyp.13049.

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

[204] 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. DOI: 10.1016/j.concog.2018.07.001.

[205] A Sawyer et al. 'Building consensus on clinical outcome assessments for BCI devices. A summary of the 10th BCI society meeting 2023 workshop'. In: *Journal of Neural Engineering* 22.1 (Jan. 2025), p. 010201. DOI: 10.1088/1741-2552/ad7bec.

[206] Thorsten Dickhaus et al. 'Predicting BCI performance to study BCI illiteracy'. In: *BMC Neuroscience* 10.S1 (July 2009). DOI: 10.1186/1471-2202-10-s1-p84.

[207] Jelena Mladenovic. 'Standardisation of protocol design for user training in EEG-based Brain-Computer Interface'. In: *Journal of Neural Engineering* (Nov. 2020). DOI: 10.1088/1741-2552/abcc7d.

[208] D Martinez-Peon et al. 'Characterisation and classification of kinesthetic motor imagery levels'. In: *Journal of Neural Engineering* 21.4 (July 2024), p. 046024. DOI: 10.1088/1741-2552/ad5f27.

[209] G.R. Müller-Putz et al. 'From classic motor imagery to complex movement intention decoding'. In: *Brain-Computer Interfaces: Lab Experiments to Real-World Applications*. Elsevier, 2016, pp. 39–70. DOI: 10.1016/bs.pbr.2016.04.017.

[210] Jing Mu et al. 'Frequency set selection for multi-frequency steady-state visual evoked potential-based brain-computer interfaces'. In: *Frontiers in Neuroscience* 16 (Dec. 2022). DOI: 10.3389/fnins.2022.1057010.

[211] Alberto Tates et al. 'Decoding Speech Imagery or Just Noise?: A Symptom of the Replicability Crisis'. In: *ResearchGate* (July 2025). Preprint. DOI: 10.13140/RG.2.2.26172.55684.

[212] Febo Cincotti et al. 'Non-invasive brain–computer interface system: Towards its application as assistive technology'. In: *Brain Research Bulletin* 75.6 (Apr. 2008), pp. 796–803. DOI: 10.1016/j.brainresbull.2008.01.007.

[213] YUYA SAITO et al. 'Review of Performance Improvement of a Noninvasive Brain-computer Interface in Communication and Motor Control for Clinical Applications'. In: *Juntendo Medical Journal* 69.4 (2023), pp. 319–326. DOI: 10.14789/jmj.jmj23-0011-r.

[214] Anne Porbadnigk et al. 'EEG-based Speech Recognition - Impact of Temporal Effects'. In: *BIOSIGNALS 2009 - Proceedings of the International Conference on Bio-inspired Systems and Signal Processing, Porto, Portugal, January 14-17, 2009*. Ed. by Pedro Encarnação and António P. Veloso. INSTICC Press, 2009, pp. 376–381.

[215] BCI Committee. '2020 International BCI Competition'. In: (). DOI: 10.17605/OSF.IO/PQ7VB.

[216] Natasha Padfield et al. 'Motor and Speech Imagery EEG Dataset'. In: (Nov. 2023). DOI: 10.60809/drum.24465871.v1.

[217] Maurice Rekrut, Abdulrahman Mohamed Selim and Antonio Krüger. 'Improving Silent Speech BCI Training Procedures Through Transfer from Overt to Silent Speech'. In: *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2022, pp. 2650–2656. DOI: 10.1109/SMC53654.2022.9945447.

[218] Edgar Aguilera-Rodríguez et al. 'An EEG-based Imagined Speech Database for comparing Paradigm Designs'. In: *Scientific Data* 12.1 (Oct. 2025). DOI: 10.1038/s41597-025-05926-5.

[219] Tzu-Tsung Wong and Po-Yang Yeh. 'Reliable Accuracy Estimates from k-Fold Cross Validation'. In: *IEEE Transactions on Knowledge and Data Engineering* 32.8 (2020), pp. 1586–1594. DOI: 10.1109/TKDE.2019.2912815.

[220] Thomas Donoghue et al. 'Parameterising neural power spectra into periodic and aperiodic components'. In: *Nature Neuroscience* 23.12 (Nov. 2020), pp. 1655–1665. DOI: 10.1038/s41593-020-00744-x.

[221] Bureau de coopération interuniversitaire. *The R Score: What It Is and What It Does*. Technical Report. Detailed mathematical definition and justification of the Cote R (R Score) calculation for non-homogeneous group ranking. Bureau de coopération interuniversitaire (BCI), Sept. 2020.

[222] Claudia Sannelli et al. 'A large scale screening study with a SMR-based BCI: Categorisation of BCI users and differences in their SMR activity'. In: *PLOS ONE* 14.1 (Jan. 2019). Ed. by Hasan Ayaz, e0207351. DOI: 10.1371/journal.pone.0207351.

[223] Wolfgang Klimesch, Paul Sauseng and Simon Hanslmayr. 'EEG alpha oscillations: The inhibition–timing hypothesis'. In: *Brain Research Reviews* 53.1 (Jan. 2007), pp. 63–88. DOI: 10.1016/j.brainresrev.2006.06.003.

[224] Ole Jensen and Ali Mazaheri. 'Shaping Functional Architecture by Oscillatory Alpha Activity: Gating by Inhibition'. In: *Frontiers in Human Neuroscience* 4 (2010). DOI: 10.3389/fnhum.2010.00186.

[225] Ben Alderson-Day and Charles Fernyhough. 'Inner speech: Development, cognitive functions, phenomenology, and neurobiology.' In: *Psychological Bulletin* 141.5 (Sept. 2015), pp. 931–965. DOI: 10.1037/bul0000021.

[226] David H. Brainard. 'The Psychophysics Toolbox'. In: *Spatial Vision* 10.4 (1997), pp. 433–436. DOI: 10.1163/156856897x00357.

[227] 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. DOI: 10.1016/s1388-2457(98)00038-8.

[228] 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.

[229] 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. DOI: 10.1088/1741-2552/ad48b9.

[230] Hsien-Chung Wu. 'The Karush–Kuhn–Tucker optimality conditions in multiobjective programming problems with interval-valued objective functions'. In: *European Journal of Operational Research* 196.1 (July 2009), pp. 49–60. DOI: 10.1016/j.ejor.2008.03.012.

[231] Maurice Rekrut and Abdulrahman Mohamed Selim. *EEG data recorded during spoken and imagined speech interaction with a simulated robot*. en. 2025. DOI: 10.5281/ZENODO.15516012.