Speech Imagery Brain-Computer Interfaces: A Systematic Literature Review

Speech Imagery Brain-Computer Interfaces: A Systematic Literature Review

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

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

E-mail: at18157@essex.ac.uk

February 2025

Abstract. Speech Imagery (SI) refers to the mental experience of hearing speech and may be the core of verbal thinking for people who undergo internal monologues. It belongs to the set of possible mental imagery states that produce kinesthetic experiences whose sensations are similar to their non-imagery counterparts. SI underpins language processes and may have similar building blocks to overt speech without the final articulatory outcome. The kinesthetic experience of SI has been proposed to be a projection of the expected articulatory outcome in a top-down processing manner. As SI seems to be a core human cognitive task it has been proposed as a paradigm for Brain Computer Interfaces (BCI). One important aspect of BCI designs is usability, and SI may present an intuitive paradigm, which has brought the attention of researchers to attempt to decode SI from brain signals. In this paper we review the important aspects of SI-BCI decoding pipelines. *Approach.* We conducted this review according to the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) guidelines. Specifically, we filtered peer-reviewed reports via a search of Google Scholar and PubMed. We selected a total of 104 reports that attempted to decode Speech Imagery from neural activity. *Main results.* Our review reveals a growing interest in SI decoding in the last 20 years, and shows how different neuroimaging modalities have been employed to record SI in distinct ways to instruct participants to perform this task. We discuss the signal processing methods used along with feature extraction techniques and found a high preference for Deep Learning models. We have summarized and compared the decoding attempts by quantifying the efficacy of decoding by measuring Information Transfer Rates. Notably, fewer than 6% of studies reported real-time decoding, with the vast majority focused on offline analyses. This suggests existing challenges of this paradigm, as the variety of approaches and outcomes prevents a clear identification of the field's current state-of-the-art. We offer a discussion of future research directions. *Significance* Speech Imagery is an attractive BCI paradigm. This review outlines the increasing interest in SI, the methodological trends, the efficacy of different approaches, and the current progress toward real-time decoding systems.

1. Introduction

What does your mind sound like at this precise moment? You are probably hearing your voice reading this paper with the same vocal characteristics you use when you speak but without any articulatory movement or external sound. Speech Imagery (SI) is a higher-order brain function related to thought, to think in verbal form is a natural expression from the human brain [1]. Some of the most reported content of inner speech is self-talk regarding self-valuations, emotions, physical appearance, and relationships [2]. Inner speech, along with inner seeing or feeling, is estimated to occur around 20% of the time in college students [3]. These inner experiences are referred to as mental imagery. Neuroimaging techniques have shown clear patterns of brain activity elicited by these cognitive tasks [4–8]. These patterns of activity have been explained as perceptual internal representations reconstructed without perceptual processing of external stimulation [9]. 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) [9, 10].

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 [11]. 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 [12, 13].

BCI paradigms can be categorized 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 [14, 15].

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 an efference copy [16]. Presence of efference copies of the motor cortex and other motor-related regions has been demonstrated in a variety of brain imaging studies [17–19]. MI-related activity can be measured by EEG and has been utilised to help design applications to control robotic limbs [20], communication interfaces, such as spellers [21], and videogames [22]. Like MI, SI is an attractive mental imagery-based BCI paradigm.

Tian and Poepple [23] showed comparable brain activation processes for both paradigms and work by Wang et. al [24] 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 [23] 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 [25] 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 analyzing 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. Literature Review Methods

To examine this topic, we followed the Preferred Reporting Items for Systematic review and Meta-Analysis (PRISMA) guidelines [26]. In this section we describe how the study selection process 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.

Table 1. Screening criteria used for query results

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

2.1. 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:

(i) "Speech Imagery"

(ii) "Speech Imagery" AND (Classification OR Decoding OR Recognition)

(iii) ("Speech Imagery" OR "Inner Speech") AND (decoding OR EEG OR ECoG OR MEG OR fNIRS OR fMRI OR BCI)

(iv) "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 1.

To further identify related articles that were not found via our initial search queries, we analyzed 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 1 shows a flowchart of the selection of records along with the number of studies identified during the screening process.

2.1.1. 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 [27]. ITR uses the number of SI classes the decoder

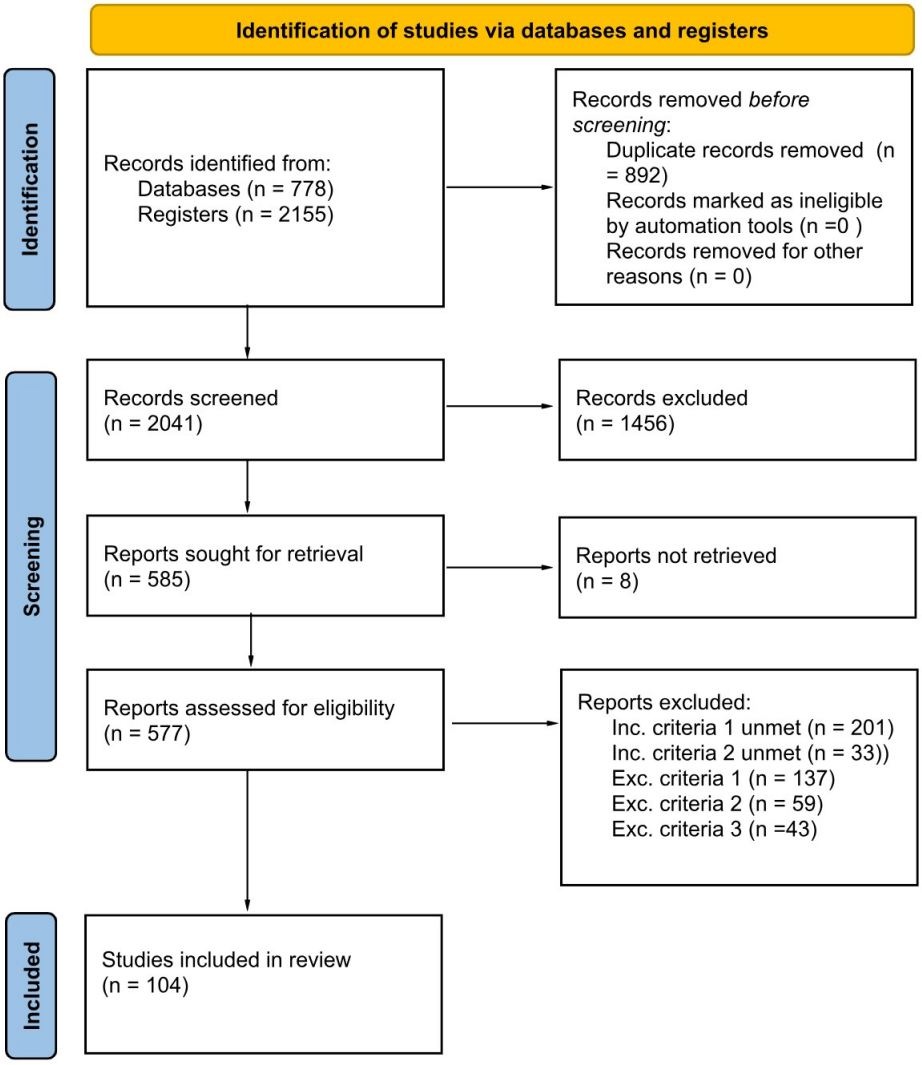

Figure 1. The flowchart describing the selection steps of the studies analyzed in this review

is attempting to label, the time window selected from the signals, and the reported decoding accuracy, as defined in [28] by:

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

$$ITR = B \times \frac{60}{T} \tag{2}$$

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 communication systems, an ideal objective is the generation of continuous speech.

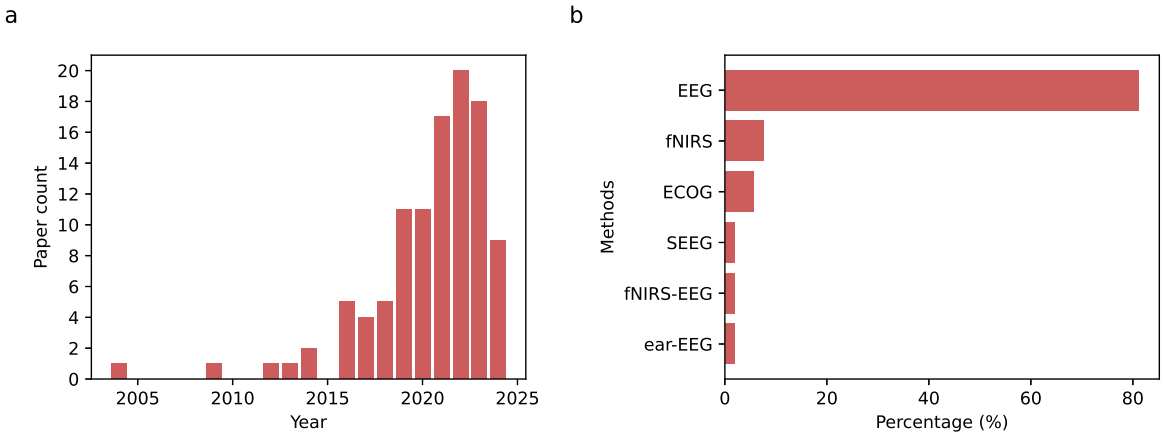

Figure 2. **a** Histogram of distributions of retrieved papers of reports attempting SI decoding over time. The first paper was found in 2004 and a significant increase in results is seen from 2019. Our retrieval was conducted in the middle of 2024, therefore there might be further papers published after our search. **b** Histogram of the selection frequency of each neuroimaging method, 82.26% of SI decoding reports used EEG.

2.1.2. Word Error Rate WER has been proposed as BCI Speech Synthesizers performance measure and utilized to evaluate real-time decoders of attempted speech [29–31]. WER is a widely recognized metric to assess the performance of machine translation, quantifying the proportion of incorrect words relative to a reference sentence [32]. 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.

3. Results

We first review the distribution of reports found across time to check for any changes in interest in SI. We then analyze the neuroimaging methods that have been employed to record SI-related signals. We enumerate available datasets that have been published and re-used. We analyze 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.a shows the time distribution of reports found, an important jump in 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.

3.1. 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 [13, 33, 34]. 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 3.2. Figure 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 [35] utilized 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. [36], 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. [37] applied a logarithmic transform to the high-gamma power of SEEG signals to develop a closed-loop BCI capable of synthesizing 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. [37] conducted research with a participant diagnosed with intractable epilepsy who had SEEG electrodes implanted for clinical monitoring. Likewise, the participants in Martin et al. [36] had subdural electrodes implanted as part of presurgical evaluation for epilepsy treatment.

fNIRS has also been demonstrated to allow decoding of SI-related activity [36, 38–41]. 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 [38] showed the capability of fNIRS for real-time SI

decoding of "yes" and "no" words with an average accuracy of $73\% \pm 9.4$.

Multimodal approaches have also been attempted to decode SI. For example, Rezazadeh et. al [42] combined fNIRS and EEG to classify SI of "yes" and "no", achieving an accuracy of $80.4\% \pm 19.1$ that proved significantly better than either of the modalities alone. Cooney et. al [43] classified 4 different words showing significant improvement when combining fNIRS and EEG.

3.2. 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:

(i) Wang et. al [44], 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.

(ii) KaraOne dataset [45], 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`

(iii) Coretto et. al [46], 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/`

(iv) Nguyen et. al [47], 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`

(v) 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/`

(vi) Nieto et. al [48], 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`

(vii) Liwicki et. al [49], 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

Table 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

dataset is available at `https://openneuro.org/datasets/ds004196/versions/2.0.2`

3.3. 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.

3.3.1. 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 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 [50, 51] suggesting these speech units may work well in experiments using this neuroimaging modality. Table 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 [46], Spanish vowels were selected because of their acoustic stationarity and lack of individual semantic meaning. DaSalla et. al [52] 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

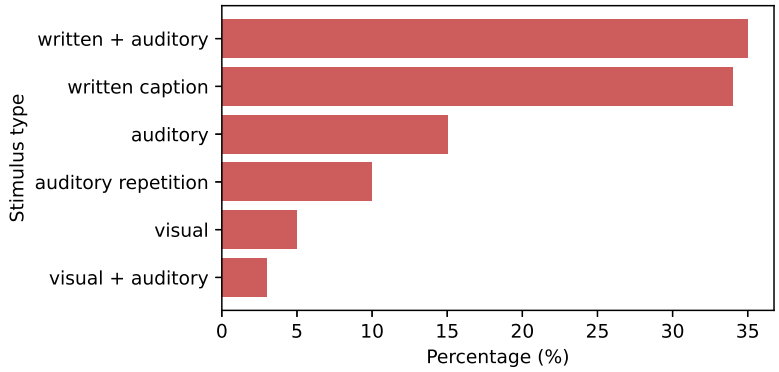

Figure 3. Histogram of selection frequencies of stimulus types, the written caption is the most preferred stimulus type in SI experiment designs, while visual stimuli that involve picture naming is the least preferred.

articulating these vowels. Gosh et. al (2022) [53] 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 [45] used the phonemes (/iy/, /uw/, /piy/, /tuy/, /diy/, /m/, /n/). Jahangiri and Sepulveda [54] used the phonemes (/ba/, /fo/, /le/, /ry/) that were abbreviations of commands (back, forward, left, right). Zhang et. al (2020) [55] used the phoneme /ba/ and asked their participants to utter it using 4 different tones (/bā/, /bá/,/bă/, /bà/), similarly, fNIRS work by Guo and Chen [39] 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 [56–58] 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 [38, 40, 42]. Simple common English words such as /hello/ or /thank you/ have also been employed [59–61].

Lastly, some ECoG approaches have asked participants to perform SI of full sentences to isolate all available English phonemes [29, 31].

3.3.2. 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 [43, 62]. 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 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 [57, 63, 64]. 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 [65].

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 [66].

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 [67]. 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 [45] used a text prompt and its corresponding audio utterance to ensure correct pronunciation. Nguyen et. al [47] 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) [54] showed images of arrows as stimuli and after a few seconds the imagery was cued by an auditory stimuli. Park and Lee [68] 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 [43] 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.

3.4. 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) [34, 69]. 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 [31, 35].

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 [70, 71]. 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 [54] 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 [72]. Kaongeon et. al [58] concluded that the gamma and delta bands had the highest F-score when classifying an imagery task against a resting state. Lee et. al [73] 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 [74] 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 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. [75] divided the gamma frequency band range into 4 sub-bands covering from 30–195 Hz or Willet et. al [31] 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

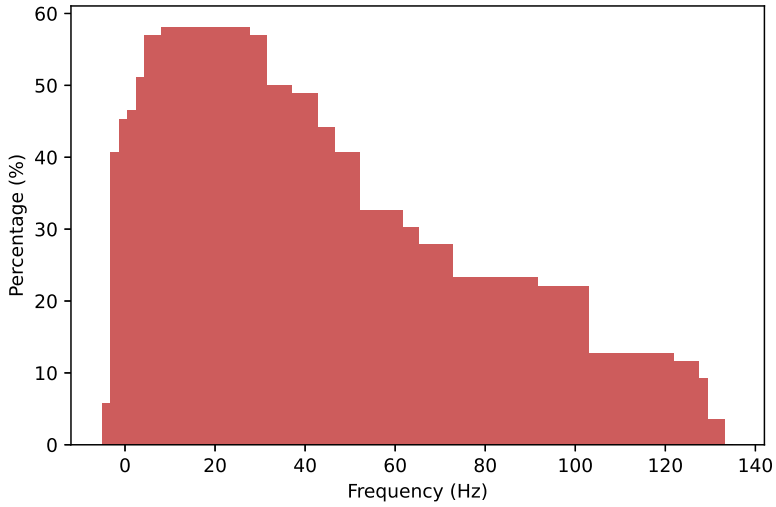

Figure 4. Distribution of focused frequencies in EEG. Most approaches focus on the alpha (8–12 Hz) and beta bands (12–30 Hz), but wider ranges have been investigated.

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 [43, 55, 76–79].

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. [78] 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 [80]. 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 [39, 42, 43]. Hwang et. al [46] used common average reference (CAR) on (Oxygenated Haemoglobin) HbO, to help reduce this noise component.

Table 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)
[44]	2	EEG	SD	CSP	SVM	1.43
[64]	3	EEG	SD	graph features	SVM	1.70
[52]	2	EEG	SD	CSP	SVM	3.49
[39]	2	fNIRS	SD	common spatial, activation and connection	SVM	5.76
[82]	12	EEG	SD	bank filters,CSP	SVM	15.30
[55]	4	EEG	SD	CSP	SVM	44.96

3.5. 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 information retrieval, syllabification or articulation, among other cognitive tasks [81]. These features can be grouped based on what they represent, including temporal patterns like event-related potentials (ERPs) and oscillations, spatial details through cortical localization, 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.

3.5.1. 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.

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 [83].

Following its popularity in MI, CSP was first used in the context of SI by DaSalla et. al [52]. Due to its high performance it was used then in several further early attempts at SI decoding. Table 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 [39].

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 5 shows a spatial histogram of brain regions used for SI decoding. Based on the specific channels and brain locations 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.

3.5.2. 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 4 the majority of 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 emphasize rapid frequencies, as they have shown clear dynamical differences in the high gamma bands as in work by Angrick et.al [37] that used signals in the range 70–170$\dot{\text{H}}$z or Leuthardt et.al [94] that focused on singles from 40 to 160$\dot{\text{H}}$z. Table 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

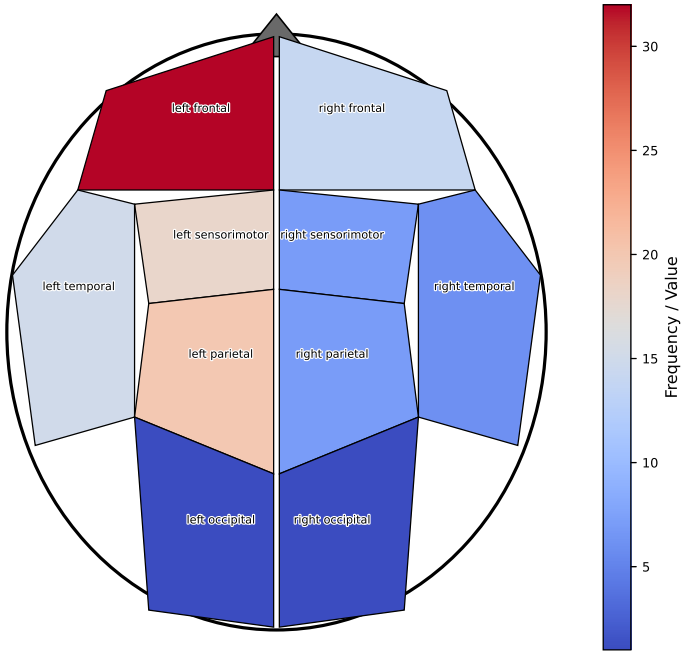

Figure 5. Spatial colour map of general brain regions showing the frequency of selection of features from each of these areas. Features from the left-frontal area of the brain were described in more than 20 reports as informative for SI decoding his area of the brain corresponds to Broca's area.

representation. Such coefficients have been directly used as features as in the ECoG based SI decoding approach reported by Mugler et. al [35] and Bejestani et. al [87].

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 [95] to extract SI features. Rusnac et. al [96] 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 [54]

Wavelet Decomposition is another method widely used to decompose EEG signals, it addresses the limitations of FFT as decompositions include temporal information.

Table 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)
[84]	5	EEG	SD	DWT	LDA	0.98
[85]	6	EEG	SD	DWT	SVM	1.73
[40]	3	fNIRS-EEG	SD	DWT, HbO	RLDA	2.90
[86]	2	EEG	SD	DWT, energy sum, waveform lenght	LDA	3.77
[87]	2	EEG	SD	FFT, amplitude of each frecuency	SVM	4.16
[76]	5	EEG	SD	DWT, MaxLCor	SVM	4.80
[56]	3	EEG	SD	WPD	LightGBM	5.41
[88]	2	EEG	SD	CSP	kNN	5.66
[57]	2	EEG	SD	DWT	DNN	7.80
[79]	5	EEG	SD	DWT	Random forest	9.33
[89]	11	EEG	SD	MFCC	SVM	9.60
[90]	10	EEG	SI	FFT	RF	15.22
[91]	4	EEG	SD	DWT, CSP	ELM	21.92
[92]	8	EEG	SD	DWT, CSP	SVM	50.08
[93]	26	EEG	SI	DWT, CSP	SVM	91.49
[35]	300	ECOG	SD	FFT	LDA	139.15

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 [53, 56, 92, 97, 98]. However, Biorthogonal and Symlet wavelet families have also been explored for SI decoding.

3.5.3. 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. 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

Table 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/minute)
[47]	3	EEG	SD	tangent projection of SPDs	SVM	1.70
[58]	4	ear-EEG	SD	tangent space projection SPD	MLELM	1.78
[99]	4	EEG	SD	CSP, tangent projection of SPD	SVM	3.69
[100]	2	EEG	SD	Correntropy SPD	MDM	6.37
[101]	5	ear-EEG	SD	CSP, tangent projection of TSMBC	MLELM	36.00

decoding attempts [102]. SPD matrices can also be projected into their corresponding tangent space to construct feature vectors whose distance can be approximately Euclidean for regular classifiers [47, 58]. The SPD property is also true for other estimators for matrices such as coherence matrices or cross-spectral density matrices [100, 103]. Table 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 [104] computed the Mean Phase Coherence (MPC) and Magnitude-Squared Coherence. The resulting statistical measures of phase synchronization 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 [105]. 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 [106] used Pearson Correlation coefficients between pairs of HbO channels to measure synchronization 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 [4] employed phase synchronization 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 [107] developed an SI decoder by using operational architectonics, a neuroscience concept of brain function, to compute from EEG windows,

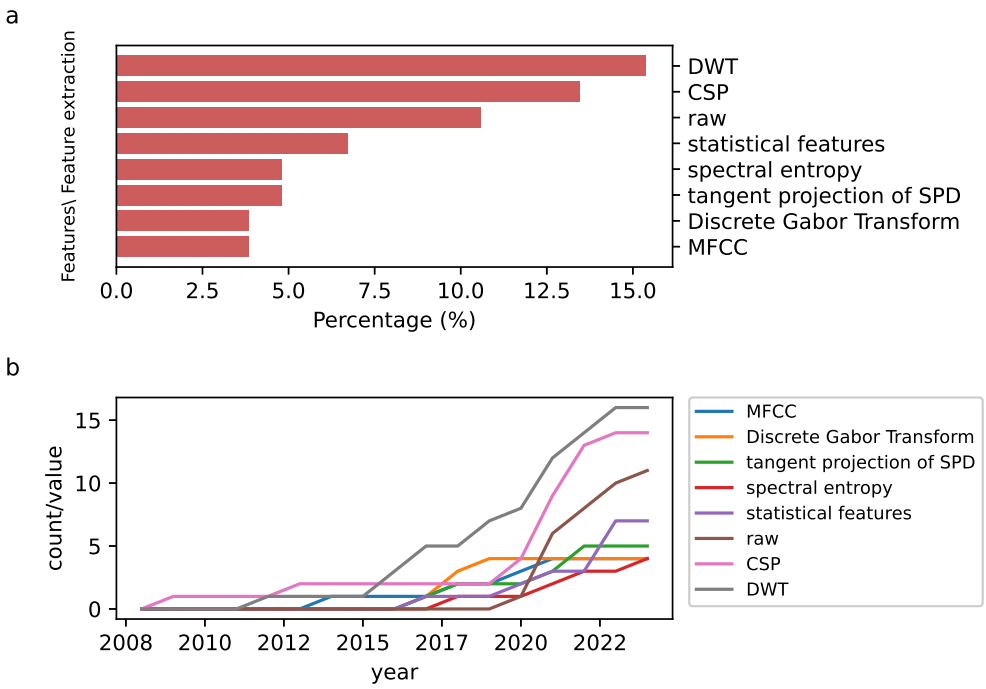

Figure 6. **a** Histogram of counts of the 7 most used feature extraction techniques among SI decoding attempts. **b** Timeline showing the cumulative sum of the most used feature extraction techniques over recent years, CSP has exhibited constant growth since its application in early approaches. The 'raw' feature describes approaches that use the unprocessed EEG signal to train DL models. DL approaches have become predominant in recent times.

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 6.a summarizes 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 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 3.5.5.

3.5.4. 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

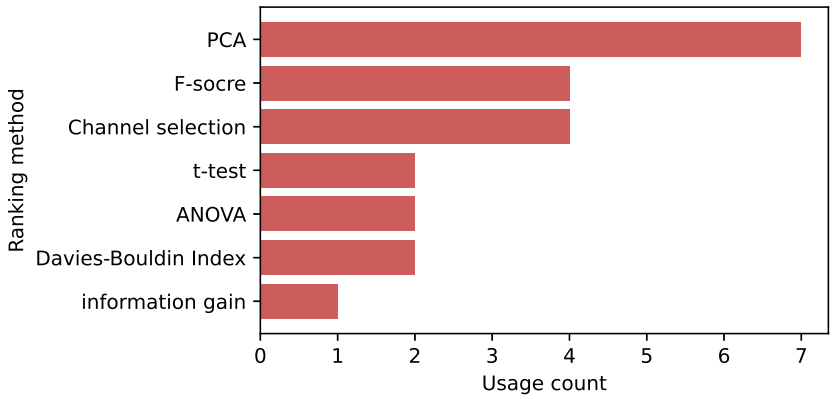

Figure 7. Histogram of counts of used feature selection techniques.

result in very large feature vectors and may lead to under-fitting issues for classification algorithms.

In Figure 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 [95] 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 [108] 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 [39] ranked HbO mean values using F-score to select the most important values and train an LDA. Other approaches, such as in work by Bajestani et. al [87], 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 [77] 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.

3.5.5. 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 [57] used raw EEG to train three Convolutional Neural Networks (CNNs), a shallow CNN whose temporal (2D convolution) and spatial (deepwise) convolutions are hypothesized 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 [109]. 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 [110]. 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 [72] 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 [103] 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 [111] 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 hypothesized to work better with covariance matrices as we discussed in Section 3.5.3

Table 6 groups the reports that have used DNN models as classifiers.

3.5.6. 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 [128] 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 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 [69] 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 [129] who concatenated EEG trials into a single vector to then generate a codebook based on K-NN clustering. With 250

Table 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)
[112]	5	EEG	SI	raw	CNN	0.50
[96]	7	EEG	SD	CNN	CNN	0.57
[78]	6	EEG	SI	ICA	CNN	1.02
[40]	2	EEG	SD	DWT	SNN	1.23
[57]	5	EEG	SI	spatiotemporal convolution	CNN	1.32
[113]	2	EEG	SD	CSP	Caps	2.63
[114]	2	EEG	SD	MPC	SNN	2.83
[111]	11	EEG	SD	raw	SPDNET	3.03
[74]	2	EEG	SI	SPWVD	CNN	3.58
[76]	5	EEG	SD	DWT	ELM	3.79
[115]	2	EEG	SI	statistical features	DNN	4.68
[116]	6	EEG	SI	instantaneous frequency and spectral entropy	CNN	4.73
[67]	2	EEG	SD	CSP	DNN	4.76
[95]	2	EEG	SD	MFCC	SNN	6.66
[117]	2	EEG	SI	raw	CNN	7.31
[57]	2	EEG	SD	DWT	DNN	7.80
[118]	6	EEG	SD	covariance	ELM	8.83
[119]	2	EEG	SI	raw	CNN	8.96
[120]	2	EEG	SD	spectogram	CNN	10.97
[77]	5	EEG	SI	DWT	Deep belief network	11.70
[121]	5	EEG	SD	raw	CNN	15.76
[68]	5	EEG	SD	MEMD	CNN	18.81
[122]	4	EEG	SI	spectrO-spatio-temporal convolution	CNN	19.30
[103]	6	EEG	SI	CCV	CNN	20.43
[97]	11	EEG	SD	daubecheis-4 (db4) wavelet	DNN	20.90
[91]	4	EEG	SD	CSP	ELM	21.92
[123]	3	EEG	SD	raw	LSTM	22.72
[30]	50	ECOG	SD	amplitude envelope	CNN	25.06
[72]	5	EEG	SD	sparse regression model	ELM	28.24
[124]	5	EEG	SI	DTCWT	CNN	28.54
[103]	8	EEG	SD	covariance matrixes	RNN	34.33
[108]	11	EEG	SD	statistical features	Caps	39.10
[63]	4	EEG	SD	spatio-temporal convolution	Fully connected layer	49.47
[125]	11	EEG	SI	Grammian transformation	DCNN	53.16
[126]	11	EEG	SI	raw	CNN	99.83
[127]	5	EEG	SI	raw	CNN	100.88

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 [64] 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 [130] 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 [131] 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 [132]. In their work, they performed subject-dependent binary classification of DDA features with an SVM classifier, achieving an average accuracy of 85%.

3.6. Summary

We analyzed 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 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.

Our analysis further revealed that 23% of the studies adopted an inter-participant approach, aiming to develop decoding pipelines that generalize 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 six studies (0.57%) reported real-time SI decoding attempts with user feedback, involving different modalities (2 ECoG, 1 SEEG, 1 fNIRS, 2 EEG) [37, 40, 101, 120, 133, 134].

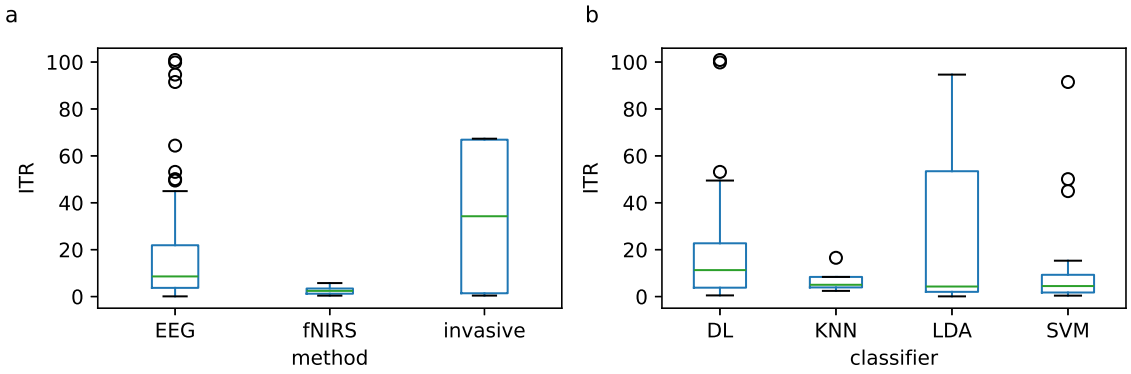

Figure 8. **a** Information Transfer Rate estimation for different recording modalities used to acquire SI. **b** ITR estimation for the different classifiers employed. The middle line of the box corresponds to the median ITR.

4. 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 conceptualization, promoting complex interaction of the brain's networks as discussed in different speech models [23, 135], 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 [136, 137], 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 analyzed 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 synthesize 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 3.1.

Moreover, in the context of neurorehabilitation, SI-based BCIs offer promising potential to provide users with feedback on their performance in speech imagery tasks relative to predefined targets. This has been demonstrated in online approaches where participants either read the decoding output [40, 134], hear the intended speech unit rendered as synthetic speech [37], or control an external device [101, 120]. Such closed-loop feedback mechanisms may support speech recovery and rehabilitation, opening new possibilities for assisting individuals with speech-related neurological impairment. However, the mentioned closed-loop approaches in our review did not study how the feedback improved the participants' SI performance.

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.

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.

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.

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 [138], 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 [139].

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 conceptualized 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 [37, 75, 134].

As discussed in Section 3.3 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 analyzing different types of speech units but no significant evidence to suggest that prompting specific speech units leads to higher classification accuracy. [46, 57, 78, 105]. 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 generalize 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 generalization 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 [139, 140].

Neurofeedback and BCI training remain largely unexplored within the context of SI. We suggest that future research should prioritize 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 [141, 142].

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 [30] 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 modeling to enhance accuracy. Similarly, Card et al. [29] 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 [31]. 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.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Conflict of interest

The authors declare they have no competing interests.

[1] Alderson-Day B and Fernyhough C 2015 *Psychological Bulletin* **141** 931–965 ISSN 0033-2909 URL http://dx.doi.org/10.1037/bul0000021

[2] Morin A, Uttl B and Hamper B 2011 *Procedia - Social and Behavioral Sciences* **30** 1714–1718 ISSN 1877-0428 URL http://dx.doi.org/10.1016/j.sbspro.2011.10.331

[3] Heavey C L and Hurlburt R T 2008 *Consciousness and Cognition* **17** 798–810 ISSN 1053-8100 URL http://dx.doi.org/10.1016/j.concog.2007.12.006

[4] Chengaiyan S and Anandan K 2022 *Cognitive Processing* **23** 593–618 ISSN 1612-4790 URL http://dx.doi.org/10.1007/s10339-022-01103-3

[5] Hemati S and Hossein-Zadeh G A 2018 *Frontiers in Human Neuroscience* **12** ISSN 1662-5161 URL http://dx.doi.org/10.3389/fnhum.2018.00515

[6] Bajaj S, Butler A J, Drake D and Dhamala M 2015 *NeuroImage: Clinical* **8** 572–582 ISSN 2213-1582 URL http://dx.doi.org/10.1016/j.nicl.2015.06.006

[7] Rathee D, Cecotti H and Prasad G 2017 *Journal of Neural Engineering* **14** 056005 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2552/aa785c

[8] Tullo M G, Almgren H, Van de Steen F, Sulpizio V, Marinazzo D and Galati G 2022 *Brain Structure and Function* **227** 1831–1842 ISSN 1863-2661 URL http://dx.doi.org/10.1007/s00429-022-02475-0

[9] Kosslyn S M, Ganis G and Thompson W L 2001 *Nature Reviews Neuroscience* **2** 635–642 ISSN 1471-0048 URL http://dx.doi.org/10.1038/35090055

[10] Moulton S T and Kosslyn S M 2009 *Philosophical Transactions of the Royal Society B: Biological Sciences* **364** 1273–1280 ISSN 1471-2970 URL http://dx.doi.org/10.1098/rstb.2008.0314

[11] Bci definition https://bcisociety.org/bci-definition/ accessed: 10 June 2025

[12] Reichert C, Dürschmid S, Heinze H J and Hinrichs H 2017 *Frontiers in Neuroscience* **11** ISSN 1662-453X URL http://dx.doi.org/10.3389/fnins.2017.00575

[13] Illman M, Laaksonen K, Liljeström M, Jousmäki V, Piitulainen H and Forss N 2020 *NeuroImage* **215** 116804 ISSN 1053-8119 URL http://dx.doi.org/10.1016/j.neuroimage.2020.116804

[14] Orban M, Elsamanty M, Guo K, Zhang S and Yang H 2022 *Bioengineering* **9** 768 ISSN 2306-5354 URL http://dx.doi.org/10.3390/bioengineering9120768

[15] Nicolas-Alonso L F and Gomez-Gil J 2012 *Sensors* **12** 1211–1279 ISSN 1424-8220 URL http://dx.doi.org/10.3390/s120201211

[16] Wolpert D M and Ghahramani Z 2000 *Nature Neuroscience* **3** 1212–1217 ISSN 1546-1726 URL http://dx.doi.org/10.1038/81497

[17] Latash M L 2021 *Journal of Neurophysiology* **125** 1079–1094 pMID: 33566734 (*Preprint* https://doi.org/10.1152/jn.00545.2020) URL https://doi.org/10.1152/jn.00545.2020

[18] Crammond D J 1997 *Trends in Neurosciences* **20** 54–57 ISSN 0166-2236 URL http://dx.doi.org/10.1016/S0166-2236(96)30019-2

[19] Anderson W S and Lenz F A 2011 *NeuroReport* **22** 939–942 ISSN 0959-4965 URL http://dx.doi.org/10.1097/WNR.0b013e32834ca58d

[20] Onose G, Grozea C, Anghelescu A, Daia C, Sinescu C J, Ciurea A V, Spircu T, Mirea A, Andone I, Spânu A, Popescu C, Mihăescu A S, Fazli S, Danóczy M and Popescu F 2012 *Spinal Cord* **50** 599–608 ISSN 1476-5624 URL http://dx.doi.org/10.1038/sc.2012.14

[21] Cao L, Xia B, Maysam O, Li J, Xie H and Birbaumer N 2017 *Frontiers in Human Neuroscience* **11** ISSN 1662-5161 URL http://dx.doi.org/10.3389/fnhum.2017.00274

[22] Paszkiel S 2016 *Journal of Automation, Mobile Robotics and Intelligent Systems* **10** 3–7 ISSN 2080-2145 URL http://dx.doi.org/10.14313/JAMRIS_4-2016/26

[23] Tian X 2010 *Frontiers in Psychology* **1** ISSN 1664-1078 URL http://dx.doi.org/10.3389/fpsyg.2010.00166

[24] Wang L, Zhang X and Zhang Y 2013 Extending motor imagery by speech imagery for brain-computer interface *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* pp 7056–7059

[25] Grandchamp R, Rapin L, Perrone-Bertolotti M, Pichat C, Haldin C, Cousin E, Lachaux J P,

Dohen M, Perrier P, Garnier M, Baciu M and Lœvenbruck H 2019 *Frontiers in Psychology* **10** ISSN 1664-1078 URL http://dx.doi.org/10.3389/fpsyg.2019.02019

[26] Page M J, Moher D, Bossuyt P M, Boutron I, Hoffmann T C, Mulrow C D, Shamseer L, Tetzlaff J M, Akl E A, Brennan S E, Chou R, Glanville J, Grimshaw J M, Hróbjartsson A, Lalu M M, Li T, Loder E W, Mayo-Wilson E, McDonald S, McGuinness L A, Stewart L A, Thomas J, Tricco A C, Welch V A, Whiting P and McKenzie J E 2021 *BMJ* **372** (*Preprint* https://www.bmj.com/content/372/bmj.n160.full.pdf) URL https://www.bmj.com/content/372/bmj.n160

[27] Fatourechi M, Mason S G, Birch G E and Ward R K 2006 Is information transfer rate a suitable performance measure for self-paced brain interface systems? *2006 IEEE International Symposium on Signal Processing and Information Technology* pp 212–216

[28] Billinger M, Daly I, Kaiser V, Jin J, Allison B Z, Müller-Putz G R and Brunner C 2012 *Is It Significant? Guidelines for Reporting BCI Performance* (Springer Berlin Heidelberg) p 333–354 ISBN 9783642297465 URL http://dx.doi.org/10.1007/978-3-642-29746-5_17

[29] Card N S, Wairagkar M, Iacobacci C, Hou X, Singer-Clark T, Willett F R, Kunz, Vahdati Nia M, Deo D R, Srinivasan A, Choi E Y, Glasser M F, Hochberg, Shahlaie K and Brandman 2023 URL http://dx.doi.org/10.1101/2023.12.26.23300110

[30] Moses D A, Metzger S L, Liu J R, Anumanchipalli G K, Makin J G, Sun P F, Chartier, Liu P M, Abrams, Ganguly K and Chang E F 2021 *New England Journal of Medicine* **385** 217–227 ISSN 1533-4406 URL http://dx.doi.org/10.1056/NEJMoa2027540

[31] Willett F R, Kunz E M, Fan C, Avansino D T, Wilson G H, Choi E Y, Kamdar F, Glasser M F, Hochberg L R, Druckmann S, Shenoy K V and Henderson J M 2023 *Nature* **620** 1031–1036 ISSN 1476-4687 URL http://dx.doi.org/10.1038/s41586-023-06377-x

[32] Ali A and Renals S 2018 Word error rate estimation for speech recognition: e-wer *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Association for Computational Linguistics) URL http://dx.doi.org/10.18653/v1/P18-2004

[33] Hecht E and Stout D 2014 *Techniques for Studying Brain Structure and Function* (Springer International Publishing) p 209–224 ISBN 9783319085005 URL http://dx.doi.org/10.1007/978-3-319-08500-5_9

[34] Burle B, Spieser L, Roger C, Casini L, Hasbroucq T and Vidal F 2015 *International Journal of Psychophysiology* **97** 210–220 ISSN 0167-8760 on the benefits of using surface Laplacian (current source density) methodology in electrophysiology URL https://www.sciencedirect.com/science/article/pii/S0167876015001865

[35] Mugler E M, Patton J L, Flint R D, Wright Z A, Schuele S U, Rosenow J, Shih J J, Krusienski D J and Slutzky M W 2014 *Journal of Neural Engineering* **11** 035015 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2560/11/3/035015

[36] Martin S, Brunner P, Holdgraf C, Heinze H J, Crone N E, Rieger J, Schalk G, Knight R T and Pasley B N 2014 *Frontiers in Neuroengineering* **7** ISSN 1662-6443 URL http://dx.doi.org/10.3389/fneng.2014.00014

[37] Angrick M, Ottenhoff M C, Diener L, Ivucic D, Ivucic G, Goulis S, Saal J, Colon A J, Wagner L, Krusienski D J, Kubben P L, Schultz T and Herff C 2021 *Communications Biology* **4** ISSN 2399-3642 URL http://dx.doi.org/10.1038/s42003-021-02578-0

[38] Hwang H J, Choi H, Kim J Y, Chang W D, Kim D W, Kim K, Jo S and Im C H 2016 *Journal of Biomedical Optics* **21** 091303 ISSN 1083-3668 URL http://dx.doi.org/10.1117/1.JBO.21.9.091303

[39] Guo Z and Chen F 2022 *Biomedical Signal Processing and Control* **72** 103369 ISSN 1746-8094 URL http://dx.doi.org/10.1016/j.bspc.2021.103369

[40] Sereshkeh A R, Yousefi R, Wong A T and Chau T 2018 *Journal of Neural Engineering* **16** 016005 URL https://doi.org/10.1088%2F1741-2552%2Faae4b9

[41] Herff C, Heger D, Putze F, Guan C and Schultz T 2012 *Cross-Subject Classification of Speaking*

Modes Using fNIRS (Springer Berlin Heidelberg) p 417–424 ISBN 9783642344817 URL http://dx.doi.org/10.1007/978-3-642-34481-7_51

[42] Sereshkeh A R, Yousefi R, Wong and Chau T 2019 *Brain-Computer Interfaces* **6** 128–140 ISSN 2326-2621 URL http://dx.doi.org/10.1080/2326263X.2019.1698928

[43] Cooney C, Folli R and Coyle D 2022 *IEEE Transactions on Biomedical Engineering* **69** 1983–1994 ISSN 1558-2531 URL http://dx.doi.org/10.1109/TBME.2021.3132861

[44] Wang L, Zhang X, Zhong X and Zhang Y 2013 *Biomedical Signal Processing and Control* **8** 901–908 ISSN 1746-8094 URL http://dx.doi.org/10.1016/j.bspc.2013.07.011

[45] Zhao S and Rudzicz F 2015 Classifying phonological categories in imagined and articulated speech *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp 992–996

[46] Pressel Coretto G A, Gareis I E and Rufiner H L 2017 Open access database of eeg signals recorded during imagined speech *12th International Symposium on Medical Information Processing and Analysis* vol 10160 ed Romero E, Lepore N, Brieva J and Larrabide I (SPIE) p 1016002 ISSN 0277-786X URL http://dx.doi.org/10.1117/12.2255697

[47] Nguyen C H, Karavas G K and Artemiadis P 2017 *Journal of Neural Engineering* **15** 016002 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2552/aa8235

[48] Nieto N, Peterson V, Rufiner H L, Kamienkowski J E and Spies R 2022 *Scientific Data* **9** URL https://doi.org/10.1038/s41597-022-01147-2

[49] Simistira Liwicki F, Gupta V, Saini R, De K, Abid N, Rakesh S, Wellington S, Wilson H, Liwicki M and Eriksson J 2023 *Scientific Data* **10** ISSN 2052-4463 URL http://dx.doi.org/10.1038/s41597-023-02286-w

[50] Cheung C, Hamilton L S, Johnson K and Chang E F 2016 *eLife* **5** e12577 ISSN 2050-084X URL https://doi.org/10.7554/eLife.12577

[51] Bouchard K E, Mesgarani N, Johnson K and Chang E F 2013 *Nature* **495** 327–332 ISSN 1476-4687 URL http://dx.doi.org/10.1038/nature11911

[52] DaSalla C S, Kambara H, Sato M and Koike Y 2009 *Neural Networks* **22** 1334–1339 ISSN 0893-6080 URL http://dx.doi.org/10.1016/j.neunet.2009.05.008

[53] Ghosh R, Sinha N and Phadikar S 2022 *SN Computer Science* **3** ISSN 2661-8907 URL http://dx.doi.org/10.1007/s42979-022-01274-y

[54] Jahangiri A and Sepulveda F 2017 The contribution of different frequency bands in class separability of covert speech tasks for bcis *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE) pp 2093–2096 URL http://dx.doi.org/10.1109/EMBC.2017.8037266

[55] Zhang X, Li H and Chen F 2020 Eeg-based classification of imaginary mandarin tones *2020 42nd Annual International Conference of the IEEE Engineering in Medicine; Biology Society (EMBC)* (IEEE) p 3889–3892 URL http://dx.doi.org/10.1109/EMBC44109.2020.9176608

[56] Liu H P Z L C T L W F F 2022 *Cognitive Neurodynamics* **17** 373–384 ISSN 1871-4099 URL http://dx.doi.org/10.1007/s11571-022-09819-w

[57] Cooney C, Korik A, Folli R and Coyle D 2020 *Sensors* **20** 4629 ISSN 1424-8220 URL http://dx.doi.org/10.3390/s20164629

[58] Kaongoen N, Choi J and Jo S 2021 *Journal of Neural Engineering* **18** 016023 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2552/abd10e

[59] Lee B H, Kwon B H, Lee D Y and Jeong J H 2021 Speech imagery classification using length-wise training based on deep learning *2021 9th International Winter Conference on Brain-Computer Interface (BCI)* pp 1–5

[60] Abdulghani M M, Walters W L and Abed K H 2023 *Bioengineering* **10** 649 ISSN 2306-5354 URL http://dx.doi.org/10.3390/bioengineering10060649

[61] Ko W, Jeon E and Suk H I 2022 *Spectro-Spatio-Temporal EEG Representation Learning for Imagined Speech Recognition* (Springer International Publishing) p 335–346 ISBN 9783031024443 URL http://dx.doi.org/10.1007/978-3-031-02444-3_25

[62] Pei X, Barbour D L, Leuthardt E C and Schalk G 2011 *Journal of Neural Engineering* **8** 046028 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2560/8/4/046028

[63] Li F, Chao W, Li Y, Fu B, Ji Y, Wu H and Shi G 2021 *Journal of Neural Engineering* **18** 0460c4 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2552/ac13c0

[64] Einizade A, Mozafari M, Jalilpour S, Bagheri S and Hajipour Sardouie S 2022 *Neuroscience Informatics* **2** 100091 ISSN 2772-5286 URL http://dx.doi.org/10.1016/j.neuri.2022.100091

[65] Defeyter M A, Russo R and McPartlin P L 2009 *Cognitive Development* **24** 265–273 ISSN 0885-2014 URL http://dx.doi.org/10.1016/j.cogdev.2009.05.002

[66] Cooney C, Folli R and Coyle D 2022 *Neuroscience; Biobehavioral Reviews* **140** 104783 ISSN 0149-7634 URL http://dx.doi.org/10.1016/j.neubiorev.2022.104783

[67] Panachakel J T and G R A 2021 Classification of phonological categories in imagined speech using phase synchronization measure *2021 43rd Annual International Conference of the IEEE Engineering in Medicine; Biology Society (EMBC)* (IEEE) pp 2226–2229 URL http://dx.doi.org/10.1109/EMBC46164.2021.9630699

[68] Park H j and Lee B 2023 *Frontiers in Human Neuroscience* **17** ISSN 1662-5161 URL http://dx.doi.org/10.3389/fnhum.2023.1186594

[69] Martin S, Brunner P, Iturrate I, Millán J d R, Schalk G, Knight R T and Pasley B N 2016 *Scientific Reports* **6** ISSN 2045-2322 URL http://dx.doi.org/10.1038/srep25803

[70] Ray W J and Cole H W 1985 *Science* **228** 750–752 ISSN 1095-9203 URL http://dx.doi.org/10.1126/science.3992243

[71] Yu H, Ba S, Guo Y, Guo L and Xu G 2022 *Brain Sciences* **12** 194 ISSN 2076-3425 URL http://dx.doi.org/10.3390/brainsci12020194

[72] Min B, Kim J, Park H j and Lee B 2016 *BioMed Research International* **2016** 1–11 ISSN 2314-6141 URL http://dx.doi.org/10.1155/2016/2618265

[73] Lee S H, Lee M and Lee S W 2020 *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 9268982* **28** 2647–2659

[74] Kamble A, Ghare P H, Kumar V, Kothari A and Keskar A G 2023 *IEEE Transactions on Instrumentation and Measurement* **72** 1–9

[75] Meng K, Grayden D B, Cook M J, Vogrin S and Goodarzy F 2021 Identification of discriminative features for decoding overt and imagined speech using stereotactic electroencephalography *2021 9th International Winter Conference on Brain-Computer Interface (BCI)* pp 1–6

[76] Pawar D and Dhage S 2022 Imagined speech classification using eeg based brain-computer interface *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)* pp 662–666

[77] Sree R A and Kavitha A 2017 Vowel classification from imagined speech using sub-band eeg frequencies and deep belief networks *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)* pp 1–4

[78] Liwicki F S, Gupta V, Saini R, De K and Liwicki M 2022 *NeuroSci* **3** 226–244 ISSN 2673-4087 URL http://dx.doi.org/10.3390/neurosci3020017

[79] Torres-Garcia A A, Reyes-Garciaa and GarcÃa-Aguilar G 2016 *Expert Systems with Applications* **59** 1–12 ISSN 0957-4174 URL http://dx.doi.org/10.1016/j.eswa.2016.04.011

[80] Luke R, Shader M J and McAlpine D 2021 *Neurophotonics* **8** ISSN 2329-423X URL http://dx.doi.org/10.1117/1.NPh.8.4.041001

[81] Levelt W J 1999 *Trends in Cognitive Sciences* **3** 223–232 ISSN 1364-6613 URL http://dx.doi.org/10.1016/S1364-6613(99)01319-4

[82] Lee S H, Lee M and Lee S W 2020 *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **28** 2647–2659

[83] Ramoser H, Muller-Gerking J and Pfurtscheller G 2000 *IEEE Transactions on Rehabilitation Engineering* **8** 441–446

[84] Garcia A A T, Garcia C and Villasellor-Pineda L *International Conference on Bio-*

inspired Systems and Signal Processing URL https://pdfs.semanticscholar.org/1e9e/620ce36718a13b29af69dabb59e1e3ec8c30.pdf

[85] Varshney Y V and Khan A 2022 *Frontiers in Signal Processing* **2** ISSN 2673-8198 URL http://dx.doi.org/10.3389/frsip.2022.760643

[86] Idrees B M and Farooq O 2016 Vowel classification using wavelet decomposition during speech imagery *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)* pp 636–640

[87] Asghari Bejestani M R, Mohammad Khani G R, Nafisi V R and Darakeh F 2022 *BioMed Research International* **2022** 1–20 ISSN 2314-6133 URL http://dx.doi.org/10.1155/2022/8333084

[88] Riaz A, Akhtar S, Iftikhar S, Khan A A and Salman A 2014 Inter comparison of classification techniques for vowel speech imagery using eeg sensors *The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014)* pp 712–717

[89] Cooney C, Folli R and Coyle D 2018 Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from eeg *2018 29th Irish Signals and Systems Conference (ISSC)* pp 1–7

[90] Hossain A, Das K, Khan P and Kader M F 2023 *Machine Learning with Applications* **13** 100486 ISSN 2666-8270 URL http://dx.doi.org/10.1016/j.mlwa.2023.100486

[91] Yongsheng Zhao Ying Liu Y G 2021 *Journal of Beijing Institute of Technology* **30** 44 (pages 7) URL https://journal.hep.com.cn/jbit/EN/abstract/article_32401.shtml

[92] Biswas S and Sinha R 2021 *IET Signal Processing* **16** 92–105 URL https://doi.org/10.1049%2Fsil2.12059

[93] Agarwal P and Kumar S *International Journal of Imaging Systems and Technology* URL https://doi.org/10.1002/ima.22655

[94] Leuthardt E C, Schalk G, Wolpaw J R, Ojemann J G and Moran D W 2004 *Journal of Neural Engineering* **1** 63 URL https://dx.doi.org/10.1088/1741-2560/1/2/001

[95] Mini P, Thomas T and Gopikakumari R 2021 *Biomedical Signal Processing and Control* **68** 102625 ISSN 1746-8094 URL http://dx.doi.org/10.1016/j.bspc.2021.102625

[96] Rusnac A L and Grigore O 2020 Generalized brain computer interface system for eeg imaginary speech recognition *2020 24th International Conference on Circuits, Systems, Communications and Computers (CSCC)* pp 184–188

[97] Ananthapadmanabha J T P A R T 2019 Decoding imagined speech using wavelet features and deep neural networks *2019 IEEE 16th India Council International Conference (INDICON)* pp 1–4

[98] Pawar D and Dhage S 2020 *Biomedical Engineering Letters* **10** 217–226 ISSN 2093-985X URL http://dx.doi.org/10.1007/s13534-020-00152-x

[99] Choi J, Kaongoen N and Jo S 2022 Investigation on effect of speech imagery eeg data augmentation with actual speech *2022 10th International Winter Conference on Brain-Computer Interface (BCI)* pp 1–5

[100] Bakhshali M A, Khademi M and Ebrahimi-Moghadam 2020 *Biomedical Signal Processing and Control* **59** 101899 ISSN 1746-8094 URL http://dx.doi.org/10.1016/j.bspc.2020.101899

[101] Kaongoen N, Choi J and Jo S 2022 *Computer Methods and Programs in Biomedicine* **224** 107022 ISSN 0169-2607 URL http://dx.doi.org/10.1016/j.cmpb.2022.107022

[102] Lotte F Y M B F 2017 *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **25** 1753–1762 ISSN 1558-0210 URL http://dx.doi.org/10.1109/TNSRE.2016.2627016

[103] Saha P and Fels S 2019 *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 10019â10020 ISSN 2159-5399 URL http://dx.doi.org/10.1609/aaai.v33i01.330110019

[104] Nunna M A G R K S J T P N N V 2020 *An Improved EEG Acquisition Protocol Facilitates Localized Neural Activation* (Springer Singapore) pp 267–281 ISBN 9789811539923 URL http://dx.doi.org/10.1007/978-981-15-3992-3_22

[105] Proix T, Delgado Saa J, Christen A, Martin S, Pasley B N, Knight R T, Tian X, Poeppel D, Doyle W K, Devinsky O, Arnal L H, Mégevand P and Giraud A L 2022 *Nature Communications* **13** ISSN 2041-1723 URL http://dx.doi.org/10.1038/s41467-021-27725-3

[106] Guo Z and Chen F 2022 *Journal of Neural Engineering* **19** 066007 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2552/ac9e1d

[107] Papasotiriou A I I 2022 *Neuroscience* **484** 98–118 ISSN 0306-4522 URL http://dx.doi.org/10.1016/j.neuroscience.2021.11.045

[108] MacÃas-MacÃas J M, RamÃrez-Quintana J A, ChacÃ³n-MurguÃa M I, Torres-GarcÃa A A and Corral-MartÃnez L F 2023 *Computers in Biology and Medicine* **159** 106909 ISSN 0010-4825 URL http://dx.doi.org/10.1016/j.compbiomed.2023.106909

[109] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger, Hutter F, Burgard W and Ball T 2017 *Human Brain Mapping* **38** 5391–5420 ISSN 1097-0193 URL http://dx.doi.org/10.1002/hbm.23730

[110] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 *Journal of Neural Engineering* **15** 056013 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2552/aace8c

[111] Rousis G, Kalaganis F P, Nikolopoulos S, Kompatsiaris I and Petrantonakis P C 2024 1531–1535

[112] Mahmud M S, Yeasin M and Bidelman G M 2021 *Journal of Neural Engineering* **18** 046012 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2552/abecf0

[113] Manuel Macías-Macías J, Alberto Ramírez-Quintana J, Ramírez-Alonso G and Ignacio Chacón-Murguía M 2020 Deep learning networks for vowel speech imagery *2020 17th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)* pp 1–6

[114] Tsukahara A, Yamada M, Tanaka K and Uchikawa Y 2019 Analysis of eeg frequency components and an examination of electrodes localization during speech imagery *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE) pp 4698–4702 URL http://dx.doi.org/10.1109/EMBC.2019.8857047

[115] Sharon R A and Murthy H A 2020 (*Preprint* 2011.02195) URL http://arxiv.org/pdf/2011.02195

[116] Lee D Y, Lee M and Lee S W 2020 Classification of imagined speech using siamese neural network *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE) p 2979â2984 URL http://dx.doi.org/10.1109/SMC42975.2020.9282982

[117] Datta S and Boulgouris N V 2021 *Neurocomputing* **465** 301–309 ISSN 0925-2312 URL http://dx.doi.org/10.1016/j.neucom.2021.08.035

[118] Qureshi M N I, Min B, Park H j, Cho D, Choi W and Lee B 2018 *IEEE Transactions on Biomedical Engineering* **65** 2168–2177

[119] Wang J and Wang L 2022 Parallel convolutional neural network based on multi-band brain networks for eeg classification *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* pp 49–53

[120] Jeong J H, Cho J H, Lee B H and Lee S W 2023 *IEEE Transactions on Cybernetics* **53** 7469–7482

[121] Lee B H, Kwon B H, Lee D Y and Jeong J H 2021 Speech imagery classification using length-wise training based on deep learning *2021 9th International Winter Conference on Brain-Computer Interface (BCI)* pp 1–5

[122] Moon J, Orlandi S and Chau T 2022 *Brain Research* **1781** 147778 ISSN 0006-8993 URL http://dx.doi.org/10.1016/j.brainres.2022.147778

[123] Patel J and Umar S A 2021 *Detection of Imagery Vowel Speech Using Deep Learning* (Springer Singapore) p 237â247 ISBN 9789811614767 URL http://dx.doi.org/10.1007/978-981-16-1476-7_23

[124] Hernandez-Galvan A, Ramirez-Alonso G and Ramirez-Quintana J 2023 *Biomedical Signal Processing and Control* **86** 105154 ISSN 1746-8094 URL http://dx.doi.org/10.1016/j.bspc.2023.105154

[125] Islam M M and Shuvo M M H 2019 Densenet based speech imagery eeg signal classification using gramian angular field *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)* pp 149–154

[126] Rusnac A L and Grigore O 2022 *Sensors* **22** 4679 ISSN 1424-8220 URL http://dx.doi.org/10.3390/s22134679

[127] Nitta T, Horikawa J, Iribe Y, Taguchi R, Katsurada K, Shinohara S and Kawai G 2023 *Frontiers in Human Neuroscience* **17** ISSN 1662-5161 URL http://dx.doi.org/10.3389/fnhum.2023.1163578

[128] Watanabe H, Tanaka H, Sakti S and Nakamura S 2020 *Neuroscience Research* **153** 48–55 ISSN 0168-0102 URL http://dx.doi.org/10.1016/j.neures.2019.04.004

[129] Garcia-Salinas J S, Villaseor-Pineda L, Reyes-Garcia C A and Torres-Garcia A A 2019 *Biomedical Signal Processing and Control* **50** 151–157 ISSN 1746-8094 URL http://dx.doi.org/10.1016/j.bspc.2019.01.006

[130] Alizadeh D and Omranpour H 2023 *Biomedical Signal Processing and Control* **84** 104933 ISSN 1746-8094 URL http://dx.doi.org/10.1016/j.bspc.2023.104933

[131] Carvalho V R, Mendes, Sejnowski T J, Comstock L and Lainscsek C 2024 *Frontiers in Human Neuroscience* **18** ISSN 1662-5161 URL http://dx.doi.org/10.3389/fnhum.2024.1398065

[132] Lainscsek C, Weyhenmeyer J, Cash S S and Sejnowski T J 2017 *Neural Computation* **29** 3181–3218 ISSN 1530-888X URL http://dx.doi.org/10.1162/neco_a_01009

[133] Wu S, Bhadra K, Giraud A L and Marchesotti S 2024 *Brain Sciences* **14** 196 ISSN 2076-3425 URL http://dx.doi.org/10.3390/brainsci14030196

[134] de Borman A, Wittevrongel B, Dauwe I, Carrette E, Meurs A, Van Roost D, Boon P and Van Hulle M M 2024 *Communications Biology* **7** ISSN 2399-3642 URL http://dx.doi.org/10.1038/s42003-024-06518-6

[135] Forkel S J and Hagoort P 2024 *Brain Structure and Function* **229** 2073–2078 ISSN 1863-2661 URL http://dx.doi.org/10.1007/s00429-024-02859-4

[136] Tian X, Zarate J M and Poeppel D 2016 *Cortex* **77** 1–12 ISSN 0010-9452 URL http://dx.doi.org/10.1016/j.cortex.2016.01.002

[137] Tian X and Poeppel D 2012 *Frontiers in Human Neuroscience* **6** ISSN 1662-5161 URL http://dx.doi.org/10.3389/fnhum.2012.00314

[138] Hurlburt R T, Alderson-Day B, Kühn S and Fernyhough C 2016 *PLOS ONE* **11** e0147932 ISSN 1932-6203 URL http://dx.doi.org/10.1371/journal.pone.0147932

[139] Jeunet C, Jahanpour E and Lotte F 2016 *Journal of Neural Engineering* **13** 036024 ISSN 1741-2552 URL http://dx.doi.org/10.1088/1741-2560/13/3/036024

[140] Mohamed Selim A, Rekrut M, Barz M and Sonntag D 2024 Speech imagery bci training using game with a purpose *Proceedings of the 2024 International Conference on Advanced Visual Interfaces* AVI '24 (New York, NY, USA: Association for Computing Machinery) ISBN 9798400717642 URL https://doi.org/10.1145/3656650.3656654

[141] Lotte F, Larrue F and Mühl C 2013 *Frontiers in Human Neuroscience* **7** ISSN 1662-5161 URL http://dx.doi.org/10.3389/fnhum.2013.00568

[142] Alkoby O, Abu-Rmileh A, Shriki O and Todder D 2018 *Neuroscience* **378** 155–164 ISSN 0306-4522 URL http://dx.doi.org/10.1016/j.neuroscience.2016.12.050